中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛

论文成果汇编

中国科学院大学前沿交叉科学学院

二零二五年十一月

目录

一、人工智能

报告人	报告题目
邓竞成	大语言模型中的隐式推理
房泽锐	Agent4S：大语言模型视角下的科研范式变革
冯一纯	用多智能体协作的强化学习系统打造主动问诊的AI医生
胡明	半监督医学影像分割
李翀	基于词元对齐的高效大语言模型词表替换方法
刘硕	基于多模态大模型的时空轨迹数据挖掘
梁奥	多智能体时空一致的多模态生成式世界模型研究
倪雨嫣	基于直线扩散模型的高效 3D 分子生成方法 Straight-Line Diffusion Model for Efficient 3D Molecular Generation
谭敏	Safe navigation for robotic digestive endoscopy via human intervention-based reinforcement learning
完颜宇洋	Look Before You Leap: A GUI-Critic-R1 Model for Pre-Operative Error Diagnosis in GUI Automation
王梦斐	基于交叉频域特征的光流估计
熊兵	USIGAN: Unbalanced Self-Information Feature Transport for Weakly Paired Image IHC Virtual Staining
闫雪	大语言模型引导的高效的强化学习
郑金兰	人工智能基础模型与算法
周昊龙	基于频率滤波的校正误差鲁棒立体匹配方法

目录

二、集成电路

报告人	报告题目
龚帅宇	人工反铁磁斯格明子VCMA-MRAM器件及连续型阵列设计
李维娜	基于极紫外光刻的卤代金属有机框架材料纳米级无光刻胶图案化
李旭帆	原子尺度调控的高热稳定性铟镓锌氧化物半导体器件的仿真及工艺实现
李岩	特征值问题的并行轨道更新方法的数值分析
廖天浩	In2O3基晶体管的热稳定性作用机理与结构调控研究
卢宇鹏	抑制IGZO可变电容提升2T0C DRAM稳定性
秦善天	面向边缘高通量应用的数据流架构软硬协同调度优化研究
芮定海	基于掩模热-力耦合效应的高精度套刻误差补偿方法
汤瑞丰	Region-Selective Oxygen Vacancy Engineering for Ferroelectric HZO Thin Films Processed at 300 ° C
张程	一种SiC MOSFET的三维电热耦合模型
赵天成	面向极高功率散热的金刚石基氧化镓异质集成材料与器件
周晓莉	高功率密度多结级联VCSEL设计与制备

目录

三、量子科技	
报告人	报告题目
韩英卓	转角双层石墨烯中赝能隙态的实空间分布研究
李雅楠	mK温区连续制冷的多级绝热去磁制冷机
王小妮	基于超导纳米线的量子相滑移器件研究
王一诺	面向量子精密传感的单模单偏振VCSEL器件研究
杨洁	利用电荷中性点处的拓扑平带增强催化活性
张钧翔	开放量子系统亚稳态在固态体系中的观测

目录

四、生物制造	
报告人	报告题目
郭沛林	数据驱动工程化益生菌创制及其用于溃疡性结肠炎治疗的研究
贾坦	Construction of Linear Heart Tubes via Coaxial Printing
李欣然	肿瘤中染色体外DNA的复制和维持与DNA损伤间双向调控关系
刘伟松	生物电催化二氧化碳还原
尹若鹏	益生菌细胞壁改善酒精性肝病
张鹏鹏	改造亚胺还原酶催化轴手性化合物的合成
钟晓彤	人工智能助力解锁黏细菌天然产物生物合成新机制

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：邓竞成
方向：人工智能
单位：中国科学院计算技术研究所

题目：大语言模型中的隐式推理

LATENT REASONING IN LLMS AS A VOCABULARY-SPACE SUPERPOSITION

Jingcheng Deng[1,2], **Liang Pang**[*1], **Zihao Wei**[1,2], **Shichen Xu**[1,2], **Zenghao Duan**[1,2], **Kun Xu** , **Yang Song** , **Huawei Shen**[1,2], **and Xueqi Cheng**[1,2]

[1]**State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences**
[2]**University of Chinese Academy of Sciences**
{dengjingcheng23s, pangliang}@ict.ac.cn

arXiv:2510.15522v1 [cs.CL] 17 Oct 2025

ABSTRACT

Large language models (LLMs) demonstrate strong reasoning abilities with chain-of-thought prompting, but explicit reasoning introduces substantial computational overhead. Recent work on latent reasoning reduces this cost by reasoning in latent space without explicit supervision, but performance drops significantly. Our preliminary experiments suggest that this degradation stems from the unstructured latent space, which makes fitting latent tokens difficult. To address this, we restrict the latent space to the column space of the LLM vocabulary, treating latent reasoning as a superposition over vocabulary probabilities. Once latent reasoning concludes, it collapses into an eigenstate of explicit reasoning to yield the final answer. Based on this idea, we propose Latent-SFT, a two-stage learning framework. In the first stage, we design two specialized attention masks to guide the Latent Token Encoder in generating latent tokens, allowing the LLM to produce the correct answer conditioned on them. In the second stage, the Latent Token Encoder is discarded, and the LLM is directly trained to generate these latent tokens autonomously for latent reasoning, optimized with KL and CE losses. Latent-SFT sets a new state of the art on GSM8k, matching explicit SFT performance while cutting reasoning chains by up to 4× and outperforming prior latent methods. On Math500 and AIME24, lexical probability–based latent reasoning also clearly surpasses hidden-state–based approaches. Our metrics of effective compression rate and effective global parallelism further show that latent reasoning is both the compression of a single path and the superposition of multiple paths. Our code, datasets, and models are publicly available at: `https://github.com/DJC-GO-SOLO/Latent-SFT`.

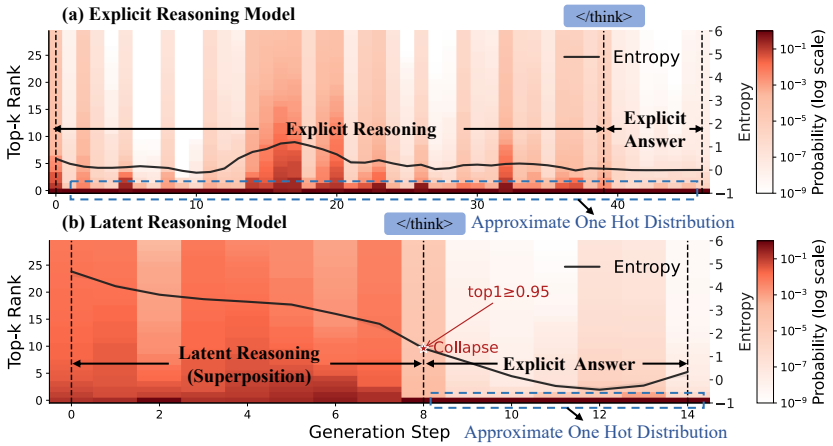


Figure 1: Comparison of Latent and Explicit Reasoning. Latent reasoning operates in a vocabulary superposition state, completing the reasoning process and reaching the correct answer with fewer tokens. Each step of explicit reasoning can be regarded as approximating a one-hot distribution.

[*]Corresponding author.

1 INTRODUCTION

Large language models (LLMs) have shown strong reasoning abilities (Zhao & Zhang, 2024) across diverse tasks (Cao & Zhao, 2025; Liu et al., 2025; Deng et al., 2025b;a; Wei et al., 2025a). A key driver of this success is chain-of-thought prompting (Wei et al., 2022; Xu et al., 2024; Yao et al., 2023), which solves complex problems by generating intermediate steps in natural language. However, current reasoning models (Team, 2025; DeepSeek-AI & et al., 2025) rely on long chains of tokens (Xu et al., 2025), leading to redundancy and high computational cost. Existing solutions, such as token pruning (Xia et al., 2025; Wei et al., 2025b) and reinforcement learning with length penalties (Arora & Zanette, 2025), reduce token usage but still constrain reasoning to the natural language space. In this work, we explore Latent Reasoning, a framework that enables reasoning in continuous latent space.

Unlike explicit reasoning, latent reasoning retains the last hidden state from the latent layer of the LLM, bypassing token sampling and directly feeding it back as the next-step representation. This hidden state is expected to carry richer information than a single token, effectively increasing bandwidth and enabling exploration of more reasoning paths with fewer generation steps (Zhu et al., 2025). While recent latent reasoning methods have succeeded in reducing token usage, they often suffer from substantial performance degradation (Hao et al., 2024).

To enhance latent reasoning, we analyze its challenges by comparing it with explicit reasoning. Firstly, supervision is difficult. Explicit reasoning leverages natural language as an anchor, enabling supervision through annotation or sampling data, with labels transferable across models. Latent reasoning, however, lacks natural labels; even if latent tokens are generated, they are often model-specific and hard to reuse. Secondly, learning is difficult. Explicit labels are discrete distributions, easy to optimize, and aligned with pre-training objectives. By contrast, latent tokens are typically the hidden states of the LLM's last layer—high-dimensional, unconstrained, and usually optimized via MSE or cosine distance, which diverges from the pre-training task. In summary, effective latent reasoning requires a principled definition of latent tokens that supports model learning.

In Section 3, we present preliminary experiments leading to two key observations. Observation 1: the hidden state distribution of the LLM's last layer is entirely inconsistent with the token embedding distribution. Since LLMs are trained with token embeddings as inputs, directly feeding hidden states introduces out-of-distribution values, preventing the model from interpreting their semantics and degrading performance. Observation 2: the effective rank of token embeddings in current LLMs is not full, indicating that the semantic space is inherently low-dimensional. Without constraining latent tokens, the model risks losing semantic consistency and struggles to learn meaningful representations. Based on these findings, we define latent tokens as elements within the column space of token embeddings. This ensures that latent and explicit tokens share consistent distributions and reside in the same low-rank space, thereby addressing the two central challenges revealed by our experiments.

Building on our definition of latent tokens, we introduce Latent-SFT, a two-stage learning framework. Stage 1: Latent token generation. We construct an encoder–decoder architecture using the same LLM, where the encoder is forced to produce latent tokens that enable the decoder to generate the correct answer. Specifically, the Latent Token Induction Mask (LTIM) constrains special tokens in the encoder to attend only to a sub-segment of the explicit reasoning chain. Latent tokens are then derived by linearly combining the vocabulary probabilities associated with these tokens. The Latent Token Supervision Mask (LTSuM) further enforces that each latent token decodes the subsequent explicit reasoning segment and the final answer. Together, these mechanisms ensure that latent tokens remain semantically compact, compatible, and correct. Stage 2: Latent token learning. The decoder learns to generate latent tokens independently. Specifically, KL loss is used to learn latent tokens, while CE loss is applied to optimize the final explicit answer.

As illustrated in Figure 1, our model performs latent reasoning in the column space of the vocabulary, analogous to wavefunction superposition in Hilbert space. Once reasoning is complete, entropy collapses into an eigenstate, producing the final natural language answer. Experiments confirm that Latent-SFT surpasses other latent baselines and achieves performance comparable to explicit SFT models on low-difficulty mathematical reasoning datasets. On high-difficulty datasets, latent tokens defined via soft embeddings significantly outperform those defined by hidden states. Further

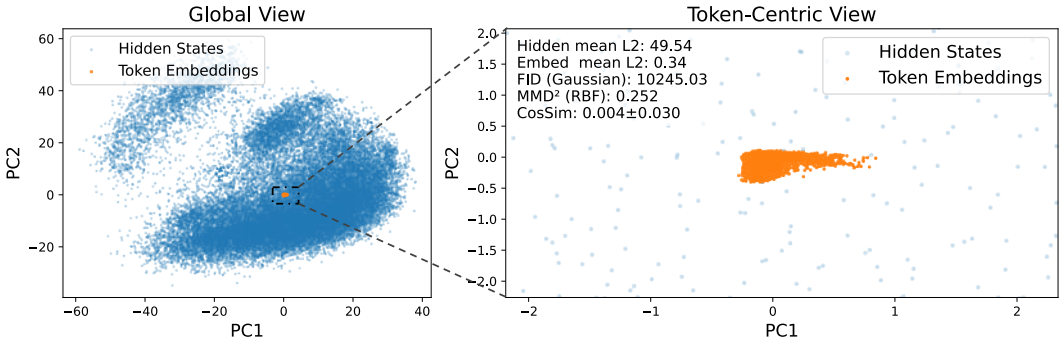


Figure 2: Visualization of last-layer hidden states and token embeddings of the LLaMA-3.2-1B-Instruct model on the GSM8k dataset. Results are shown from both global and token-specific perspectives to highlight the substantial distributional differences. In addition, we report the statistical differences between the two distributions using the Fréchet Inception Distance (FID) and the squared Maximum Mean Discrepancy (MMD^2). We also estimate cosine similarity via random sampling.

analyses reveal that its latent reasoning process is not merely a compression of a single reasoning chain but also a superposition of multiple chains.

2 Related Work

Our work centers on latent inference, and we categorize related research into two groups: training-based methods and training-free methods.

Training-based Latent Reasoning. COCONUT (Hao et al., 2024) first proposed using the last hidden state as the input for the next token without applying the language model head, laying the foundation for latent reasoning. However, it relied on a fixed number of latent steps and did not address the misalignment between hidden (latent) states and token embeddings, leading to weak performance. CCOT (Cheng & Durme, 2024) introduced continuous contemplation tokens and extended this idea to variable-length latent reasoning, achieving modest gains. CODI (Shen et al., 2025) improved alignment by adding an MLP to map hidden states into the token embedding space, and further adopted a teacher–student distillation strategy, forcing the latent model to match hidden states across all layers up to the final token of the explicit model—significantly boosting performance. PCCOT (Wu et al., 2025b) enhanced CODI by applying Jacobian iteration to improve both training and inference efficiency. CoLaR (Tan et al., 2025) advanced this line by introducing a latent head to predict compressed embedding distributions, maintaining alignment between latent and token spaces. Combined with reinforcement learning, CoLaR achieved state-of-the-art latent reasoning, though still trailing explicit SFT. Unlike these approaches, we restrict latent reasoning to the column space of the LLM vocabulary, directly addressing misalignment between latent and token spaces.

Training-free Latent Reasoning. Several works have explored training-free latent reasoning, including Soft Thinking (Zhang et al., 2025) and Mixture-of-Input (MoI) (Zhuang et al., 2025). Like our approach, they use next-token probability distributions to linearly combine token embeddings, thereby linking hidden and embedding spaces. However, Wu et al. (2025a) showed that this remains a form of greedy decoding, which limits performance. In contrast, our method can be seen as a trained variant of Soft Thinking, enabling the model to explore multiple reasoning paths.

3 Preliminary Experiments: Definition of Latent Token

To gain a deeper understanding ofing of the latent reasoning process, we conducted two preliminary experiments addressing the core question of what the representation of latent tokens should be.

3.1 Distributional Discrepancy Between Final-Layer Hidden States and Token Embeddings

Most existing studies (Shen et al., 2025; Wu et al., 2025b) follow the COCONUT design, using the last-layer hidden state directly as the representation of the next latent token rather than sampling from the explicit token embedding matrix. However, a statistical analysis of the last-layer hidden states and token embeddings in the representative LLMs reveals the following observation.

Observation 1. *The distribution of last-layer hidden states in the LLM is entirely inconsistent with that of the token embeddings.* As shown in Figure 2 and Figure 7, the distributions of hidden states and token embeddings after PCA dimensionality reduction differ substantially. Statistically, their means, variances, and inter-distribution distances are all large. From a modeling perspective, shallow parameters (e.g., RMSNorm) encountered during pre-training are exposed only to the token embedding distribution. Feeding vectors from an unfamiliar distribution into the model can thus induce distribution shift or other unpredictable effects, which is one of the reasons why methods such as COCONUT perform poorly.

3.2 Non-Full Effective Rank of Token Embeddings in Large Language Models

Building on Observation 1, we posit that the latent token representations should ideally align with the distributional geometry of the token embeddings. To examine this property, we analyze the singular value spectra of token embedding matrices across different LLMs.

Observation 2. *The Effective Rank of LLM Token Embeddings Is Not Full.* The resulting curves in Figure 3 exhibit a rapid decay rather than a flat plateau, indicating that only a subset of singular directions contributes substantially to the embedding space. The effective rank of word embeddings is far below the theoretical maximum, indicating that although the embedding matrix has high nominal dimensionality, it occupies only a low-dimensional subspace. Thus, the semantic space of LLMs can be regarded as inherently low-dimensional.

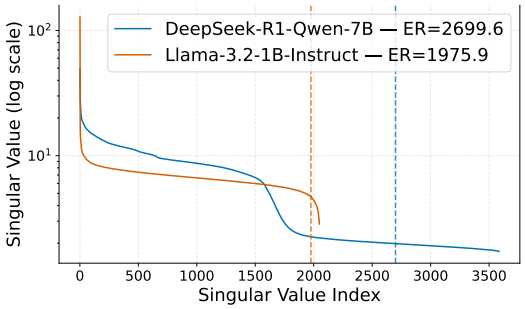


Figure 3: Decay curves of singular values for embedding matrices in various LLMs. The legend indicates the effective rank of each embedding matrix.

These findings suggest that latent tokens should not be treated as unconstrained hidden vectors, but rather as structured elements residing within the embedding subspace. Guided by this insight, we formally define latent tokens.

Definition 1 (*Soft Embedding*)**.** Formally, a latent token $z \in \mathbb{R}^d$ can be expressed as a linear combination of token embeddings:

$$z = \sum_{i=1}^{V} \alpha_i e_i, \quad \text{with } \alpha_i \in \mathbb{R}, e_i \in \mathbb{R}^d, \tag{1}$$

where $\{e_i\}_{i=1}^{V}$ denotes the embedding vectors of the vocabulary, and the coefficients $\{\alpha_i\}$ determine the semantic mixture. This formulation ensures that latent tokens remain aligned with the intrinsic semantic geometry of the LLM's vocabulary space, providing a principled interpretation of latent tokens as interpolation within the embedding manifold.

4 Latent-SFT: Equipping Large language Models with Latent Reasoning

With **Definition 1**, we introduce Latent-SFT to bring latent reasoning into LLMs. As a prerequisite, latent tokens must be generated (Definition 1), since they provide the labels needed for supervision. Thereafter, the model is trained, in analogy to explicit reasoning, to generate these latent tokens and perform latent reasoning. The overall training framework is illustrated in Figure 4.

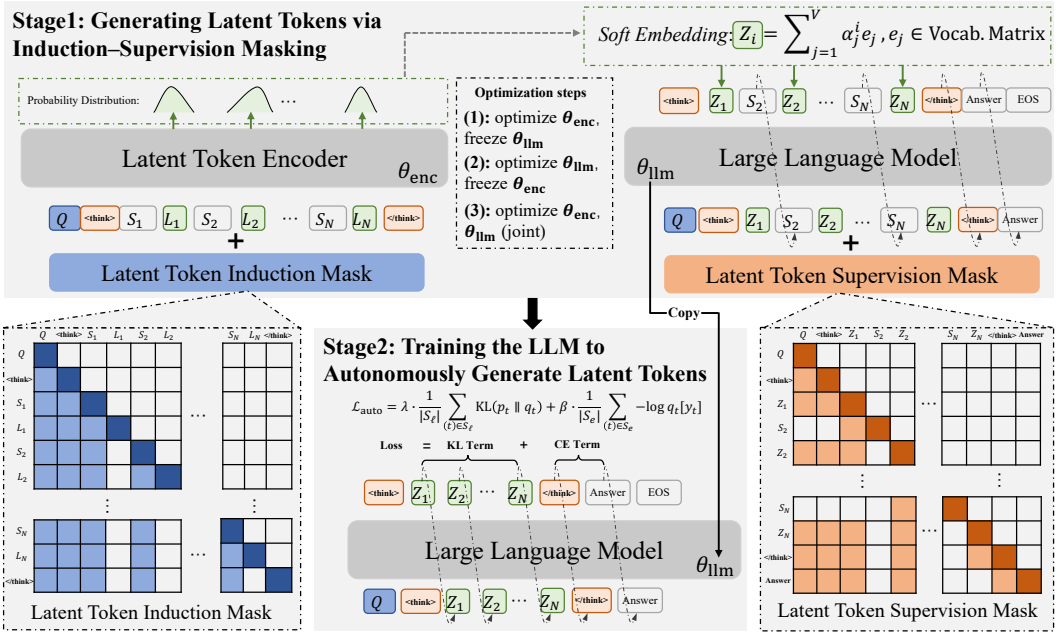


Figure 4: Overview of Latent-SFT. The training process consists of two phases: (a) Generating Latent Tokens via Induction–Supervision Masking. In the mask sub-figure, components such as Q, S_i and Answer (which may span multiple tokens) are implemented via a standard autoregressive attention mechanism. The Latent Token Encoder shares both structure and initialization with the base LLM. (b) Training the LLM to Autonomously Generate Latent Tokens. In this phase, the LLM is trained to perform latent reasoning independently, using a weighted combination of KL loss and CE loss.

4.1 Generating Latent Tokens via Induction-Supervision Masking

Latent tokens are designed to be: (1) **semantic compactness**, replacing multiple explicit tokens; (2) **semantic compatibility**, constrained to the semantic space of explicit tokens (Section 3); and (3) **semantic correctness**, capable of producing the right answer. We implement these properties within an encoder–decoder framework through three key steps: information compression, representation alignment, and supervised decoding.

Information Compression Process (Semantic Compactness). We insert special tokens into the explicit reasoning sequence to segment and capture intermediate reasoning information. Given an input $X = [Q, \text{<think>}, T, \text{</think>}]$, we define a segmentation function $\mathrm{Seg}(\cdot)$ that divides T (explicit reasoning sequence) into N subsegments, i.e., $T = \{S_i\}_{i=1}^{N}$. This segmentation can be performed either by a fixed token length (compression ratio) r or by grouping semantically continuous thoughts. After each S_i , we insert a special token L_i and apply the Latent Token Induction Mask (LTIM) to restrict each L_i to attend only to the explicit reasoning information preceding it (see Figure 4). This design ensures that the semantic content encoded by latent tokens grows with reasoning completeness, facilitating subsequent model learning.

Representation Alignment Process (Semantic Compatibility). Given the inserted special tokens $\{L_i\}_{i=1}^{N}$, we first obtain their hidden states $h_i \in \mathbb{R}$ from the (LTIM-masked) model. To align each latent token with the LLM's vocabulary manifold, we project h_i onto the vocabulary simplex and then reconstruct its embedding as a mixture of token embeddings, in accordance with Definition 1. Concretely, let $E = [e_1, ..., e_V] \in \mathbb{R}^{d\times V}$ denote the (input) embedding matrix and $W \in \mathbb{R}^{d\times V}$ the LM head (often $W = E$) under weight tying). We compute vocabulary logits and a probability vector

$$\boldsymbol{\ell}_i = W^\top \mathbf{h}_i, \qquad \boldsymbol{\alpha}_i = \mathrm{softmax}(\boldsymbol{\ell}_i/T) \in \Delta^{V-1}, \tag{2}$$

where $T > 0$ is an optional temperature. The representation of the i-th latent token is then obtained by linear interpolation within the embedding space:

$$\mathbf{z}_i \;=\; E\,\boldsymbol{\alpha}_i \;=\; \sum_{v=1}^{V} \alpha_{i,v}\,\mathbf{e}_v. \tag{3}$$

which instantiates Definition 1 (*Soft Embedding*) with coefficients given by the model-implied vocabulary distribution. This construction guarantees z_i lies in the span (and, with softmax, the convex hull) of $\{e_v\}_{v=1}^{V}$, thereby enforcing **semantic compatibility**: latent tokens remain confined to the intrinsic geometry of the LLM's vocabulary space. In training Stage 2, α_i serves as the supervision signal for latent positions (KL term), while explicit positions are optimized with CE; optionally, one may prune α_i to top-k and renormalize to reduce computation without departing from the embedding manifold.

Supervised Decoding Process (Semantic Correctness). Given the latent representations $\{z_i\}_{i=1}^{N}$ from the alignment step, we use each z_i to decode the content from the $(i+1)$-th explicit reasoning subsegment through to the final answer. Let the prefix up to step i be $\Pi_i = [Q, \text{<think>}, z_1, S_2, z_2, ..., z_i]$. Under the Latent Token Supervision Mask (LTSuM), tokens in $\mathcal{Y}_i = [S_{i+1}, ..., S_N, \text{</think>}, \text{Answer}]$ are decoded conditioned only on Π_i (in particular on $[Q, \text{<think>}, z_1, ..., z_i]$), while attention to future latent tokens and to past explicit reasoning steps is blocked. Denoting by $\mathcal{J}_i$ the index set of tokens in $\mathcal{Y}_i$, the supervised decoding objective is

$$\mathcal{L}_{\text{sup}} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{J}_i|}\sum_{t\in\mathcal{J}_i}\Big(-\log p_\theta\big(x_t \,\big|\, \Pi_i;\,\text{LTSuM}\big)\Big). \tag{4}$$

Training Schedule. In practice, the strong masking constraints (LTIM/LTSuM) make naive SFT optimization difficult. We therefore adopt an EM-style alternating schedule. Let the complete set of parameters be $\theta = (\theta_{\text{enc}}, \theta_{\text{llm}})$. First, with the LLM frozen, we optimize the encoder only, $\theta_{\text{enc}}^{t+1} = \arg\min_{\theta_{\text{enc}}} \mathcal{L}_{\text{sup}}(\theta_{\text{enc}}, \theta_{\text{llm}}^{t})$. Next, with the encoder frozen, we optimize the LLM, $\theta_{\text{llm}}^{t+1} = \arg\min_{\theta_{\text{llm}}} \mathcal{L}_{\text{sup}}(\theta_{\text{enc}}^{t+1}, \theta_{\text{llm}})$. Finally, we jointly fine-tune both components by minimizing, $(\theta_{\text{enc}}, \theta_{\text{llm}}) \leftarrow \arg\min_{\theta_{\text{llm}},\theta_{\text{enc}}} \mathcal{L}_{\text{sup}}(\theta_{\text{enc}}, \theta_{\text{llm}})$. This alternating–joint procedure stabilizes training under the constraints while steadily improving convergence.

4.2 Training the LLM to Autonomously Generate Latent Tokens

After generating latent tokens with the three desired characteristics, we concatenate them with the explicit tokens. The resulting sequence $X = [Q, \text{<think>}, z_1, ..., z_n, \text{</think>}, \text{Answer}]$ then serves as the input for Stage 2 (in Figure 4(b)). Denote the index sets of latent and explicit positions by S_{lat}, S_{exp}, respectively. We use $p_t = \alpha_t$ as the soft label for the latent position, and ground-truth tokens y_t for explicit positions. Let $q_t = \text{softmax}(W^T h_t)$ be the decoder's predictive distribution at position t. We optimize the decoder with a KL term on latent slots and a CE term on explicit slots:

$$\mathcal{L}_{\text{auto}}(\theta_{\text{llm}}) = \lambda \cdot \frac{1}{|S_{\text{lat}}|}\sum_{t\in S_{\text{lat}}} \text{KL}\big(p_t \,|\, q_t\big) \;+\; \beta \cdot \frac{1}{|S_{\text{exp}}|}\sum_{t\in S_{\text{exp}}} \big(-\log q_t[\,y_t\,]\big). \tag{5}$$

Optionally, a temperature $T > 0$ can be applied to both teacher and student distributions (with the usual T^2 scaling) and p_t may be pruned to top-k then renormalized. Minimizing $\mathcal{L}_{\text{auto}}$ trains the decoder to autonomously generate latent token sequences $[z_1, ..., z_N]$ along with the final answer.

5 Experiments

5.1 Experimental setup

Datasets. To comprehensively assess the effectiveness of our method, we conducted experiments on both low-difficulty mathematical reasoning datasets (GSM8k-Aug (Deng et al., 2024), GSM-Hard (Gao et al., 2023), SVAMP (Patel et al., 2021), and MultiArith (Roy & Roth, 2015)) and high-difficulty datasets (Math500 (Hendrycks et al., 2021) and AIME24). For the low-difficulty tasks, we

Table 1: Experimental results on four low-difficulty datasets. The number in brackets represents the compression ratio r. " " marks the best result, and " " marks the second best. - Hidden State refers to removing the Representation Alignment step and directly using the last-layer hidden state as the representation of latent tokens. - w/o LTIM and - w/o LTSuM denote removing the Latent Token Induction Mask and Latent Token Supervision Mask, respectively, reducing the model to a standard autoregressive attention mechanism.

	GSM8k-Aug		GSM-Hard		SVAMP		MultiArith		Average		
	Pass@1	# L	Pass@1	# L	Pass@1	# L	Pass@1	# L	Pass@1	# L	Pass@1/# L
CoT-SFT	49.4$_{\pm.7}$	25.6$_{\pm.1}$	**11.9**$_{\pm.2}$	34.2$_{\pm.1}$	**59.8**$_{\pm.3}$	12.1$_{\pm.0}$	93.2$_{\pm.5}$	13.7$_{\pm.1}$	**53.6**	21.4	2.50
iCoT	19.8$_{\pm.2}$	0.00$_{\pm.0}$	3.87$_{\pm.2}$	0.00$_{\pm.0}$	36.4$_{\pm.5}$	0.00$_{\pm.0}$	38.2$_{\pm.7}$	0.00$_{\pm.0}$	24.6	0.00	-
COCONUT	23.1$_{\pm.3}$	6.00$_{\pm.0}$	5.49$_{\pm.3}$	6.00$_{\pm.0}$	40.7$_{\pm.7}$	6.00$_{\pm.0}$	41.1$_{\pm.2}$	6.00$_{\pm.0}$	27.6	6.00	4.60
CODI	13.3$_{\pm.6}$	6.00$_{\pm.0}$	2.97$_{\pm.2}$	6.00$_{\pm.0}$	21.7$_{\pm.7}$	6.00$_{\pm.0}$	19.2$_{\pm.8}$	6.00$_{\pm.0}$	14.3	6.00	2.38
CoLaR-5	26.8$_{\pm.2}$	5.57$_{\pm.0}$	5.87$_{\pm.1}$	6.53$_{\pm.0}$	48.4$_{\pm.5}$	2.95$_{\pm.0}$	86.4$_{\pm.4}$	3.21$_{\pm.0}$	41.7	4.57	9.12
CoLaR-2	40.1$_{\pm.2}$	12.7$_{\pm.0}$	9.08$_{\pm.0}$	14.0$_{\pm.0}$	54.9$_{\pm.2}$	6.11$_{\pm.0}$	91.3$_{\pm.1}$	7.35$_{\pm.0}$	48.8	10.0	4.88
Latent-SFT(4)	44.6$_{\pm.0}$	6.64$_{\pm.0}$	10.2$_{\pm.0}$	7.68$_{\pm.0}$	55.6$_{\pm.1}$	3.33$_{\pm.0}$	93.6$_{\pm.0}$	4.11$_{\pm.0}$	51.0	5.44	**9.37**
- Hidden State	36.7$_{\pm.4}$	6.61$_{\pm.1}$	8.41$_{\pm.2}$	7.77$_{\pm.0}$	44.4$_{\pm.2}$	3.22$_{\pm.0}$	90.3$_{\pm.1}$	4.06$_{\pm.0}$	45.0	5.41	8.32
- w/o LTIM	38.8$_{\pm.1}$	6.59$_{\pm.1}$	8.67$_{\pm.1}$	8.01$_{\pm.1}$	48.9$_{\pm.1}$	3.34$_{\pm.1}$	90.8$_{\pm.0}$	4.36$_{\pm.0}$	46.8	5.58	8.39
- w/o LTSuM	42.6$_{\pm.0}$	6.70$_{\pm.0}$	9.98$_{\pm.1}$	7.60$_{\pm.0}$	53.9$_{\pm.1}$	3.36$_{\pm.0}$	92.6$_{\pm.0}$	4.27$_{\pm.0}$	49.7	5.48	9.07
Latent-SFT(2)	**50.4**$_{\pm.1}$	12.4$_{\pm.0}$	10.9$_{\pm.0}$	13.9$_{\pm.0}$	57.8$_{\pm.1}$	5.98$_{\pm.0}$	**93.8**$_{\pm.0}$	7.21$_{\pm.0}$	53.2	9.87	5.39
- Hidden State	41.1$_{\pm.3}$	12.8$_{\pm.0}$	9.01$_{\pm.1}$	14.2$_{\pm.1}$	49.6$_{\pm.2}$	6.17$_{\pm.0}$	91.1$_{\pm.0}$	7.29$_{\pm.0}$	47.7	10.1	4.72
- w/o LTIM	41.3$_{\pm.1}$	12.9$_{\pm.1}$	9.32$_{\pm.0}$	14.0$_{\pm.0}$	53.6$_{\pm.1}$	6.25$_{\pm.1}$	91.5$_{\pm.0}$	7.96$_{\pm.0}$	48.9	10.3	4.75
- w/o LTSuM	49.0$_{\pm.0}$	12.1$_{\pm.0}$	10.3$_{\pm.0}$	13.7$_{\pm.0}$	56.5$_{\pm.2}$	6.22$_{\pm.0}$	93.0$_{\pm.0}$	7.23$_{\pm.0}$	52.2	9.81	5.32

trained on the GSM8k-Aug training split to ensure fair comparison with prior work. For the high-difficulty tasks, due to limited computational resources, we used a subset of Open-R1 (Hugging Face, 2025) containing explicit reasoning chains shorter than 4k tokens, resulting in 57,544 training samples. This restriction inevitably constrained the performance of our method.

Baselines and Models. The baselines considered in our experiments include COT-SFT (Wei et al., 2022), iCOT (Deng et al., 2024), COCONUT (Hao et al., 2024), CODI (reproduced) (Tan et al., 2025), and CoLaR (Tan et al., 2025). For low-difficulty tasks, we adopt LLaMA-3.2-1B-Instruct (Dubey & et al., 2024), while for high-difficulty tasks, we employ DeepSeek-Distill-Qwen-7B (DeepSeek-AI & et al., 2025).

Evaluation Metrics and Details. Our evaluation measures both accuracy (Pass@N) and efficiency (# L, the number of tokens in the reasoning chain). For the low-difficulty datasets, we conducted five runs with different random seeds. For the high-difficulty datasets, Math500 was evaluated four times, while AIME24 was evaluated 64 times. We report Pass@1 and # L as the mean ± 95% confidence interval (CI). To jointly assess accuracy and efficiency, we also report their ratio, Pass@1/# L. Additional implementation details are provided in Appendix B.

5.2 Main Results

Performance on Low-Difficulty Tasks. Table 1 compares Latent-SFT with state-of-the-art baselines under different compressed ratio r settings. Relative to the latent reasoning baseline, Latent-SFT consistently achieves higher accuracy, shorter reasoning chains, and substantial gains in the Pass@1 metric. These results confirm that defining latent tokens as linear combinations of the vocabulary matrix—thus aligning them with the same semantic space as explicit tokens—is more effective than the traditional last-hidden-state definition. Furthermore, Latent-SFT demonstrates strong out-of-distribution generalization, achieving state-of-the-art performance on all three out-of-distribution datasets and even surpassing COT-SFT on MultiArith. This advantage arises from our latent token definition, which is more compatible with explicit tokens and thus better leverages token representations learned during pre-training, leading to improved generalization.

It is noteworthy that Latent-SFT(2) surpasses explicit COT-SFT in Pass@1 on both GSM8k-Aug and MultiArith, while reducing inference chain length by roughly one quarter on average. This finding demonstrates that explicit reasoning does not represent the upper bound of latent reasoning methods. With an appropriate definition and learning strategy for latent tokens, latent reasoning can exceed explicit reasoning models in accuracy. When considering both Pass@1 and #L, Latent-SFT delivers nearly a fourfold improvement over traditional COT-SFT, demonstrating strong overall performance. The real inference example is provided in the Figure 10.

Table 2: Experimental results on high-difficulty datasets. We primarily examine the performance gap between *Soft Embedding* and hidden state representations. Long inference chains amplify error accumulation in latent reasoning, while training restricted to lengths of 4k further limits performance. Consequently, the accuracy drop on AIME24 is acceptable. Since prior latent baselines were neither trained nor evaluated on high-difficulty datasets, their results are not reported.

Model	MATH-500		AIME24	
	Pass@1	# L	Pass@1	# L
COT-SFT	$92.6_{\pm 1.19}$	$3616_{\pm 232}$	$54.4_{\pm 1.74}$	$12992_{\pm 346}$
Latent-SFT(Hidden State)	$67.8_{\pm 0.94}$	$698.1_{\pm 22.1}$	$7.78_{\pm 2.08}$	$1509_{\pm 114}$
Latent-SFT(*Soft Embedding*)	$79.8_{\pm 1.35}$ (12.0% ↑)	$759.3_{\pm 36.9}$	$19.2_{\pm 1.60}$ (11.4% ↑)	$1484_{\pm 420}$

Performance on High-Difficulty Reasoning Tasks. Experimental results on high-difficulty mathematical tasks are shown in Table 2. We found that the accuracy of Latent-SFT on the AIME24 dataset dropped sharply compared to the COT-SFT model. We speculate that this is primarily due to **the increased difficulty of learning latent reasoning models with long reasoning chains**. Simply put, as reasoning chains length increases, it becomes more difficult to learn the three features required for latent tokens, resulting in a deterioration in latent token representation. This deterioration in latent token representation leads to cumulative errors in the latent reasoning stage, ultimately resulting in a drop in accuracy. Furthermore, due to training resource limitations, our model's **training examples have reasoning chains no longer than 4k in length**, further hindering its generalization to the AIME24 dataset, which requires an average of 12k inference chains. (As evidence, our model performs significantly better on the MATH500 dataset, which requires shorter reasoning chains, than on AIME24, which requires longer chains.) Therefore, adapting latent reasoning to high-difficulty mathematical reasoning datasets remains a challenging task.

In terms of inference chain length, Latent-SFT remains effective, achieving substantial reductions on both datasets. More importantly, *Soft Embedding*, the core definition proposed in this work, consistently outperforms Hidden State even on high-difficulty datasets, thereby validating the generalizability of **Definition 1**.

5.3 Ablation Study

To comprehensively assess the effectiveness of Latent-SFT, we perform an ablation study focusing on its three core components for latent token generation: LTIM, LTSuM, and *Soft Embedding*.

Replacing LTIM and LTSuM with Standard Autoregressive Attention. LTIM is employed during the information compression process to restrict each special token to attend only to preceding explicit reasoning steps, while preventing attention among special tokens themselves. When replaced with standard autoregressive attention, the hidden states of later special tokens may be influenced by earlier ones. In the early stages of training, before convergence, this effect can be particularly detrimental. As shown in Table 1, across Latent-SFT models with different r values, removing LTIM leads to a substantial drop in Pass@1 (about 5% on average), underscoring the necessity of LTIM.

LTSuM is applied in the supervised decoding process to ensure that each latent token can decode the subsequent explicit reasoning steps, thereby maintaining semantic correctness. Replacing it with standard autoregressive attention—i.e., removing all explicit reasoning sub-segments and allowing only latent tokens to generate the final answer—eliminates intermediate supervision during latent token learning and may cause semantic entanglement. As shown in Table 1, removing LTSuM leads to a substantial drop in accuracy, demonstrating its necessity.

Replace Latent Token Representation from *Soft Embedding* to the Last Hidden State Soft Embedding serves as our core definition of latent tokens. While Table 1 shows that it already outperforms traditional last-hidden-state methods such as COCONUT and CODI, for a fair comparison we replace Soft Embedding with the last hidden state in our framework while keeping all other settings unchanged. As reported in Tables 1 and Table 2, Soft Embedding achieves substantially better performance than the last hidden state on both low- and high-difficulty mathematical reasoning tasks, thereby empirically validating its effectiveness.

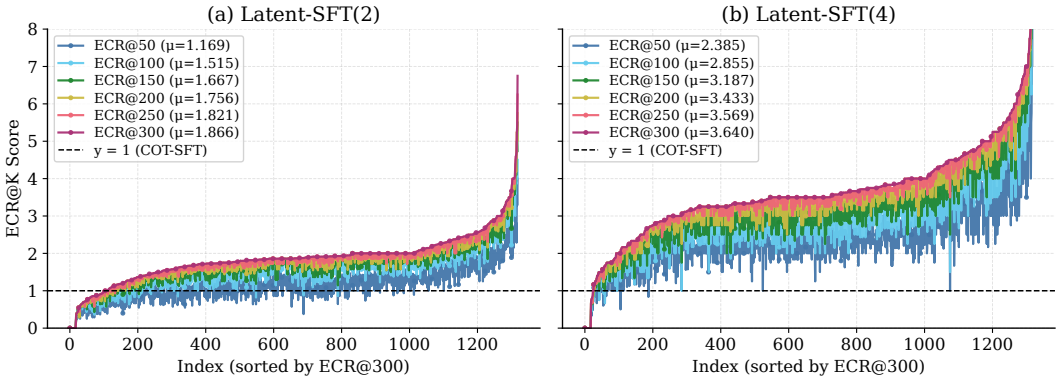


Figure 5: Latent-SFT's ECR@K values on GSM8k-Aug test samples. Each curve represents the per-sample ECR@K, with samples sorted along the x-axis by their ECR@300 values.

6 Analysis: Understanding the Essence of Latent Reasoning

In order to probe the fundamental nature of latent reasoning, we raise a central question: *What, precisely, does latent reasoning capture during the reasoning process?* Inspired by COCONUT's perspective that latent reasoning explores multiple reasoning paths simultaneously, we propose the following two hypotheses.

Hypothesis 1. Latent Reasoning Represents the Compression of a Single Reasoning Path. To quantify how much explicit information is preserved at each step of latent reasoning, we introduce the **Top-K Effective Compression Ratio (ECR@K)**. This metric measures the proportion of explicit chain tokens that fall within the Top-K of the probability distribution at the corresponding latent step. Larger values indicate stronger horizontal compression across the reasoning chain, while values greater than one reflect true compression of the explicit chain. The formal definition is provided in the Appendix D.2.

The experimental results are presented in Figure 5. Across models trained with different compression ratios r and varying K values, the ECR scores exceed 1 for almost all samples. This quantitatively validates our hypothesis that latent reasoning compresses a single reasoning path. From the training perspective, the model indeed performs information compression along a single reasoning path. Moreover, comparing Figure 5(a) and (b), doubling the training compression ratio nearly doubles the ECR score during inference, further demonstrating the effectiveness of our training strategy.

Hypothesis 2. Latent Reasoning Represents the Superposition of Multiple Reasoning Paths. To examine whether latent reasoning corresponds to the superposition of multiple paths, we define the effective global parallelism (N_{eff}), which quantifies global evidence for the presence of multiple parallel reasoning paths. A larger value reflects a higher degree of parallelism. In addition, we introduce the Top-2 Score, defined as the ratio of the probabilities of the two most likely paths. Formal definitions are provided in the Appendix D.3. To assess the parallel reasoning capability of latent models, we construct multiple alternative reasoning chains corresponding to a single explicit reasoning path within GSM8k-Aug dataset (Multi-Chain GSM8k-Aug dataset in Appendix C).

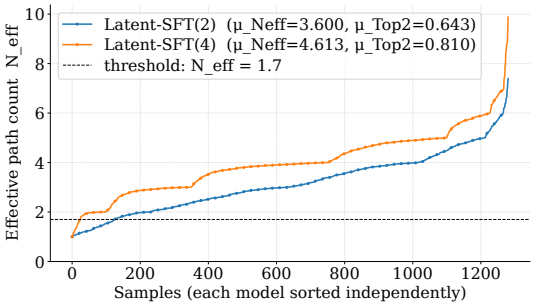


Figure 6: N_{eff} values of Latent-SFT on the Multi-Chain GSM8k-Aug dataset. The legend reports the mean values of both N_{eff} and the Top-2 Score.

Figure 6 presents the analysis results. Across models with different compression ratios, most samples exceed the conservative threshold for significant parallelism (1.7), indicating that latent reasoning reflects the superposition of multiple reasoning chains. On average, evidence is distributed

across roughly 3–4 reasoning paths rather than concentrated on a single path. The average Top-2 ratio exceeds 0.6, suggesting strong "two-way competition" and parallel reasoning. Furthermore, as the compression ratio increases, the degree of parallelism also rises. Case studies are provided in the Figure 9.

7 Conclusion

In this work, we propose Latent-SFT, a method that enables latent reasoning in LLMs by defining latent tokens within the column space of the vocabulary. Experimental results show that on low-difficulty mathematical reasoning datasets, Latent-SFT achieves performance comparable to explicit COT-SFT while significantly shortening reasoning chains. On high-difficulty datasets, it outperforms methods based on last hidden state representations. Furthermore, our analysis demonstrates that latent reasoning embodies both the compression of a single reasoning chain and the superposition of multiple chains.

8 Ethics statement and Reprodicibility statement

This work does not raise ethical concerns. The code, datasets, and models of our method are available in an GitHub repository `https://github.com/DJC-GO-SOLO/Latent-SFT`.

References

Daman Arora and Andrea Zanette. Training language models to reason efficiently. *CoRR*, abs/2502.04463, 2025. doi: 10.48550/ARXIV.2502.04463. URL `https://doi.org/10.48550/arXiv.2502.04463`.

Linbo Cao and Jinman Zhao. Pretraining on the test set is no longer all you need: A debate-driven approach to qa benchmarks, 2025. URL `https://arxiv.org/abs/2507.17747`.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *CoRR*, abs/2412.13171, 2024. doi: 10.48550/ARXIV.2412.13171. URL `https://doi.org/10.48550/arXiv.2412.13171`.

DeepSeek-AI and et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL `https://doi.org/10.48550/arXiv.2501.12948`.

Jingcheng Deng, Zhongtao Jiang, Liang Pang, Liwei Chen, Kun Xu, Zihao Wei, Huawei Shen, and Xueqi Cheng. Following the autoregressive nature of LLM embeddings via compression and alignment. *CoRR*, abs/2502.11401, 2025a. doi: 10.48550/ARXIV.2502.11401. URL `https://doi.org/10.48550/arXiv.2502.11401`.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. Everything is editable: Extend knowledge editing to unstructured data in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL `https://openreview.net/forum?id=X5rO5VyTgB`.

Yuntian Deng, Yejin Choi, and Stuart M. Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *CoRR*, abs/2405.14838, 2024. doi: 10.48550/ARXIV.2405.14838. URL `https://doi.org/10.48550/arXiv.2405.14838`.

Abhimanyu Dubey and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL `https://doi.org/10.48550/arXiv.2407.21783`.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: program-aided language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799. PMLR, 2023. URL `https://proceedings.mlr.press/v202/gao23f.html`.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *CoRR*, abs/2412.06769, 2024. doi: 10.48550/ARXIV.2412.06769. URL https://doi.org/10.48550/arXiv.2412.06769.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Sheng Liu, Qiang Sheng, Danding Wang, Yang Li, Guang Yang, and Juan Cao. Forewarned is forearmed: Pre-synthesizing jailbreak-like instructions to enhance LLM safety guardrail to potential attacks. *CoRR*, abs/2508.20038, 2025. doi: 10.48550/ARXIV.2508.20038. URL https://doi.org/10.48550/arXiv.2508.20038.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 2080–2094. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.168. URL https://doi.org/10.18653/v1/2021.naacl-main.168.

Subhro Roy and Dan Roth. Solving general arithmetic word problems. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 1743–1752. The Association for Computational Linguistics, 2015. doi: 10.18653/V1/D15-1202. URL https://doi.org/10.18653/v1/d15-1202.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: compressing chain-of-thought into continuous space via self-distillation. *CoRR*, abs/2502.21074, 2025. doi: 10.48550/ARXIV.2502.21074. URL https://doi.org/10.48550/arXiv.2502.21074.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Ruihua Song. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. *CoRR*, abs/2505.16552, 2025. doi: 10.48550/ARXIV.2505.16552. URL https://doi.org/10.48550/arXiv.2505.16552.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. Mlake: Multilingual knowledge editing benchmark for large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 4457–4473. Association for Computational Linguistics, 2025a. URL https://aclanthology.org/2025.coling-main.301/.

Zihao Wei, Liang Pang, Jiahao Liu, Jingcheng Deng, Shicheng Xu, Zenghao Duan, Jingang Wang, Fei Sun, Xunliang Cai, Huawei Shen, and Xueqi Cheng. Stop spinning wheels: Mitigating LLM overthinking via mining patterns for early reasoning exit. *CoRR*, abs/2508.17627, 2025b. doi: 10.48550/ARXIV.2508.17627. URL https://doi.org/10.48550/arXiv.2508.17627.

Chünhung Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua Wu. Llms are single-threaded reasoners: Demystifying the working mechanism of soft thinking, 2025a. URL https://arxiv.org/abs/2508.03440.

Haoyi Wu, Zhihao Teng, and Kewei Tu. Parallel continuous chain-of-thought with jacobi iteration. *CoRR*, abs/2506.18582, 2025b. doi: 10.48550/ARXIV.2506.18582. URL https://doi.org/10.48550/arXiv.2506.18582.

Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *CoRR*, abs/2502.12067, 2025. doi: 10.48550/ARXIV.2502.12067. URL https://doi.org/10.48550/arXiv.2502.12067.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee (eds.), *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pp. 1362–1373. ACM, 2024. doi: 10.1145/3589334.3645363. URL https://doi.org/10.1145/3589334.3645363.

Shicheng Xu, Liang Pang, Yunchang Zhu, Jia Gu, Zihao Wei, Jingcheng Deng, Feiyang Pan, Huawei Shen, and Xueqi Cheng. Distilling the implicit multi-branch structure in llms' reasoning via reinforcement learning. *CoRR*, abs/2505.16142, 2025. doi: 10.48550/ARXIV.2505.16142. URL https://doi.org/10.48550/arXiv.2505.16142.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *CoRR*, abs/2505.15778, 2025. doi: 10.48550/ARXIV.2505.15778. URL https://doi.org/10.48550/arXiv.2505.15778.

Jinman Zhao and Xueyan Zhang. Large language model is not a (multilingual) compositional relation reasoner. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=wLQ3I0F1oj.

Ruijie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, Zhenhe Wu, Yongchi Zhao, Tianyu Liu, Jian Yang, Wangchunshu Zhou, Chujie Zheng, Chongxuan Li, Yuyin Zhou, Zhoujun Li, Zhaoxiang Zhang, Jiaheng Liu, Ge Zhang, Wenhao Huang, and Jason Eshraghian. A survey on latent reasoning. *CoRR*, abs/2507.06203, 2025. doi: 10.48550/ARXIV.2507.06203. URL https://doi.org/10.48550/arXiv.2507.06203.

Yufan Zhuang, Liyuan Liu, Chandan Singh, Jingbo Shang, and Jianfeng Gao. Text generation beyond discrete token sampling. *CoRR*, abs/2505.14827, 2025. doi: 10.48550/ARXIV.2505.14827. URL https://doi.org/10.48550/arXiv.2505.14827.

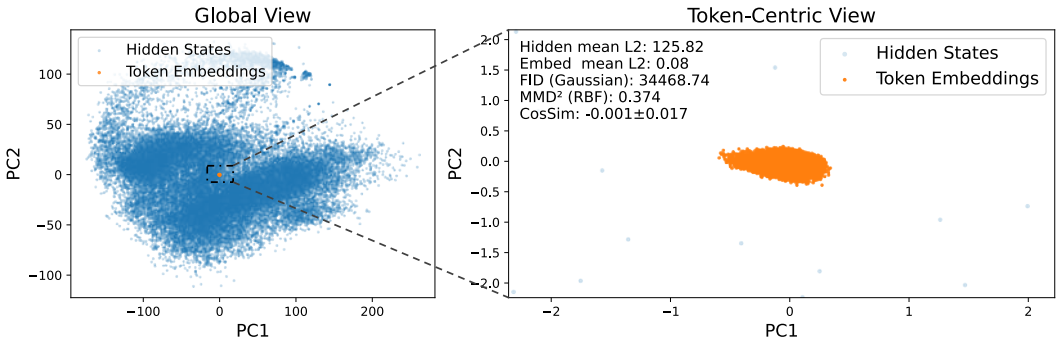


Figure 7: Visualization of last-layer hidden states and token embeddings of the Deepseek-Distill-Qwen-7B model on the Math500 dataset.

A Use of LLMs

We employ GPT-5 to polish the language of our manuscripts. Specifically, we first draft the text, then use GPT-5 to refine it into a more coherent form. After verifying its logical and semantic consistency, we incorporate the polished version into the main text.

B Implementation Details

Since Latent-SFT requires weighting with respect to the vocabulary matrix, we freeze the vocabulary parameters during training. All models are trained using LoRA with rank 64. Following COCONUT, we initialize all methods from the COT-SFT model to accelerate convergence. In practice, we find that a fixed-length segmentation function performs better than a semantic-based segmentation. Therefore, all results reported in this paper are based on fixed-length segmentation functions. For LLaMA-3.2-1B-Instruct, we set the learning rate to 1×10^{-4} in Step 1 and 4×10^{-4} in Step 2. For DeepSeek-Distill-Qwen-7B, the learning rates are 5×10^{-5} (Step 1) and 1×10^{-5} (Step 2). Following DeepSeek-AI & et al. (2025), we use a decoding temperature of 0.6 and top-p of 0.95. We did not perform parameter tuning for λ and β; setting both to 1 yielded satisfactory performance.

During evaluation, low-difficulty mathematical reasoning datasets were inferred using the Transformers library with a maximum sequence length of 128. High-difficulty datasets were evaluated with the improved SGlang library, supporting a maximum token length of 32,768. All training and inference experiments were conducted on 8×A100 or 8×H100 GPUs.

C Construction of the Multi-Chain GSM8k-Aug Dataset

We carefully designed prompts and employed the advanced GPT-5 model to generate preliminary candidate reasoning chains. Given the simplicity and limited number of questions, we manually inspected and removed examples that either failed to follow the reasoning format or contained incorrect logic. To further maximize diversity, we retained only reasoning chains with pairwise similarity below 0.9 based on edit distance. An example prompt call is illustrated in Figure 8. Among the 1,319 data points, 1,310 contained multiple reasoning paths.

D Analytical Metrics

D.1 Notation and Preliminaries

Let $\mathcal{V}$ be the vocabulary, $|\mathcal{V}| = V$. For a given problem, fix a reference explicit chain (from an explicit CoT solution)

$$\boldsymbol{x} = (x_1, \ldots, x_{L_{\text{exp}}}), \quad x_i \in \mathcal{V}. \tag{6}$$

Instruction:

```
"Given a GSM8k problem, solution, and answer, identify alternative
valid reasoning paths. Enclose each path within \\boxed{}. The paths
must not contain any natural language, and should follow the same
style as the provided example. Ensure the length of the solution and
the number of reasoning steps <<>> remain consistent with the
standard solution. Each reasoning path must strictly adhere to the
problem's logical structure—do not generate unrelated or fabricated
paths. If no alternative reasoning path exists, output \\boxed{False}.

{"problem": "Janet's ducks lay 16 eggs per day. She eats three for
breakfast every morning and bakes muffins for her friends every day
with four. She sells the remainder at the farmers' market daily for
$2 per fresh duck egg. How much in dollars does she make every day at
the farmers' market?", "solution": "<<16-3-4=9>> <<9*2=18>>",
"answer": "18"}"
```

Response:

```
"\boxed{<<<<16-(3+4)=9>> <<9*2=18>>}
\boxed{<<3+4=7>> <<(16-7)*2=18>>}
\boxed{<<16-3=13>> <<(13-4)*2=18>>}
\boxed{<<16-4=12>> <<(12-3)*2=18>>}
\boxed{<<2*16=32>> <<32-2*3-2*4=18>>}"
```

✗ Rejected (edit-distance criterion against the solution violated).

✗ Rejected (edit-distance criterion against the previous item violated).

Figure 8: An example of calling GPT5 and post-processing.

A latent model generates a latent reasoning chain consisting of T latent steps, each with a vocabulary distribution

$$p_t \in \Delta^{V-1} \quad (t = 1, \ldots, T), \tag{7}$$

where Δ^{V-1} is the probability simplex. To quantify the extent of explicit information captured by each latent step (whether from a single chain or multiple chains), we introduce:

Assumption 1. During training, if each latent token is constrained to receive information from a fixed number of explicit tokens, then at inference time the amount of explicit information (in terms of token granularity) captured by each latent step is uniform.

This assumption is reasonable, as the model is expected to behave consistently during training and inference. So we align the two sequences by a ratio $r \in \mathbb{N}$: every r explicit tokens are grouped to one latent step. Define the number of aligned steps

$$T' = \min\big(T, \lceil L_{\text{exp}}/r \rceil\big), \tag{8}$$

and the explicit token set aligned to latent step t:

$$S_t = \big\{x_{(t-1)r+1}, \ldots, x_{\min(tr, L_{\text{exp}})}\big\}, \qquad |S_t| \le r. \tag{9}$$

Let $\mathcal{T}_K(p_t) \subset \mathcal{V}$ be the set of the K highest-probability tokens under p_t.

D.2 ECR@K (Effective Compression Rate)

To verify whether latent reasoning corresponds to the compression of a single chain, we examine, *on average, how many aligned explicit tokens from the reference chain appear within the Top-K probabilities at each latent step*. Thus, we define

$$\text{ECR@K} \;=\; \frac{1}{T'} \sum_{t=1}^{T'} \big|S_t \cap \mathcal{T}_K(p_t)\big| \;\in\; [0, r]. \tag{10}$$

A value greater than 1 indicates genuine compression of a single reasoning chain, meaning that each latent step, on average, covers more than one token from the reference explicit chain.

D.3 N_{eff} AND TOP-2 SCORE

Given M candidate explicit chains $\{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$ (with lengths L_m), we align each chain to the T latent steps using the same ratio r as above. Let the explicit tokens aligned to step t on chain m be $S_{m,t} \subseteq \mathcal{V}$ (at most r tokens), and let $\mathcal{T}_K(p_t)$ be the Top-K tokens under p_t. Define the per-step *mass* of chain m as

$$\text{mass}_{m,t}^{(K)} \;=\; \sum_{x \in S_{m,t} \cap \mathcal{T}_K(p_t)} p_t[x], \tag{11}$$

and aggregate it across steps to obtain a sequence-level evidence score

$$\text{score}_m \;=\; \frac{1}{T'} \sum_{t=1}^{T'} \log\big(\text{mass}_{m,t}^{(K)} + \varepsilon\big), \tag{12}$$

where $T' = \min\big(T, \lceil L_m/r \rceil\big)$, $\varepsilon > 0$ is a small constant, and K controls robustness (setting $K = V$ recovers the full-mass version). We convert $\{\text{score}_m\}$ to a posterior over paths via a temperatured softmax

$$P_m \;=\; \frac{\exp(\text{score}_m/\tau)}{\sum_{j=1}^{M} \exp(\text{score}_j/\tau)} \quad (\tau = 1). \tag{13}$$

The **effective number of supported paths** is

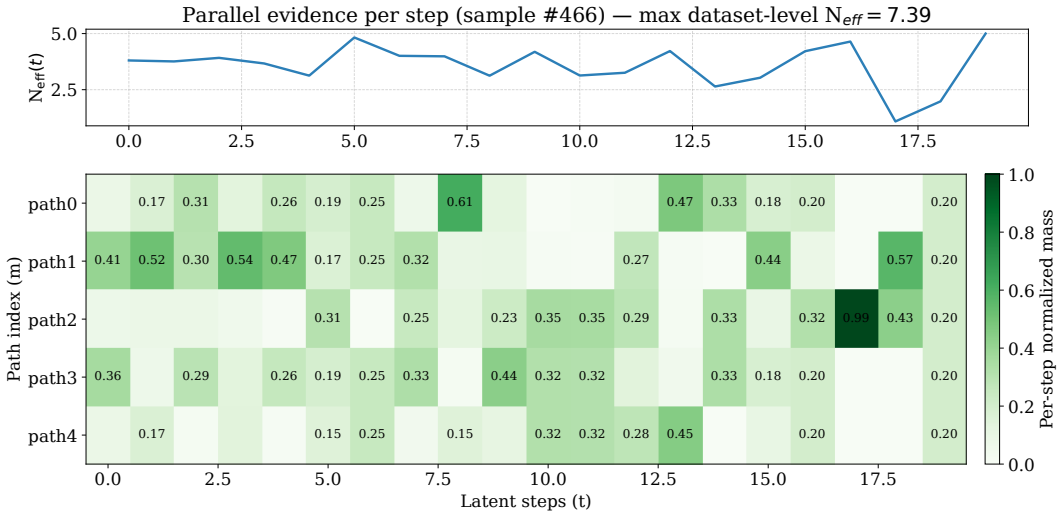
$$\text{N}_{\text{eff}} \;=\; \exp\Big(-\sum_{m=1}^{M} P_m \log P_m \Big), \tag{14}$$

and the **Top-2 score** is the ratio between the two largest posteriors,

$$\text{Top-2} \;=\; \frac{P_{(2)}}{P_{(1)}} \quad \text{with} \quad P_{(1)} \geq P_{(2)}. \tag{15}$$

Intuitively, a larger score_m means the model consistently places high probability on chain m along the sequence; a larger N_{eff} and higher Top-2 indicate stronger parallel support across multiple chains rather than collapse to a single path.

When only a single path exists in parallel, $\text{N}_{\text{eff}} = 1$. However, since different paths may partially overlap in their prefixes, we adopt a conservative baseline of $\text{N}_{\text{eff}} = 1.7$ (corresponding to approximately 70% probability mass on path 1 and 30% on path 2). Values exceeding this threshold are taken as evidence of parallel inference. In addition, to mitigate the influence of tail noise, we set $K = 100$.



Path0: <<20*0.45=9>> <<9-1=8>> <<20*0.5=10>> <<(10*10)/100=1>> <<10-1=9>> <<9-8=1>>
Path1: <<45*20=900>> <<900-100=800>> <<50*20=1000>> <<1000-100=900>> <<900-800=100>> <<100/100=1>>
Path2: <<0.50-0.45=0.05>> <<0.05*20=1>> <<20*0.50=10>> <<(10*10)/100=1>> <<1-1=0>> <<1+0=1>>
Path3: <<(45/100)*20=9>> <<9-1=8>> <<(50/100)*20=10>> <<10*(1/10)=1>> <<10-1=9>> <<9-8=1>>
Path4: <<20*0.45=9>> <<9-1=8>> <<20*0.50=10>> <<10*9=90>> <<90/10=9>> <<9-8=1>>

Figure 9: An example of the maximum N_{eff} value for the Latent-SFT(2) model on the Multi-Chain GSM8k-Aug dataset. At each step, the N_{eff} value exceeds 1.7. Path0 denotes the original reasoning chain, while the remaining paths are additional candidate chains. The heatmap illustrates the fraction of paths supported at each latent step, clearly revealing the phenomenon of parallel reasoning.

Question:

"Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?"

Reference Answer:

"<<16-3-4=9>> <<9*2=18>>\\boxed{18}",

Latent-SFT(2) Response:

"47 <<000=>>9218 <<18\\boxed{18}"

Top-1 Tokens in Latent Reasoning

Figure 10: An example of Latent-SFT(2) reasoning on the GSM8k-Aug test set. The reasoning chain length is reduced by nearly half.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 房泽锐

方向： 人工智能

单位： 中国科学院大学前沿交叉科学学院

题目： Agent4S：大语言模型视角下的科研范式变革

arXiv:2506.23692v1 [cs.AI] 30 Jun 2025

Agent4S: The Transformation of Research Paradigms from the Perspective of Large Language Models

Boyuan Zheng[1,2*†], Zerui Fang[3,4†], Zhe Xu[1,5†], Rui Wang[3,1,2], Yiwen Chen[1,2], Cunshi Wang[6,7], Mengwei Qu[8,9], Lei Lei[10], Zhen Feng[11,12], Yan Liu[13], Yuyang Li[6,14*], Mingzhou Tan[15*], Jiaji Wu[16*], Jianwei Shuai[17,12*], Jia Li[4*], and Fangfu Ye[1,18,12*]

[1]Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing, China

[2]Beijing Gongyu Zhiyan Technology Co., Ltd, Beijing, China

[3]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Science, Beijing, China

[4]Smart Sensing Chip and System R&D Center, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China

[5]Key Laboratory of Material Physics, Ministry of Education, School of Physics and Microelectronics, Zhengzhou University, Zhengzhou, China

[6]Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China

[7]College of Astronomy and Space Sciences, University of Chinese Academy of Sciences, Beijing, China

[8]State Key Laboratory of Isotope Geochemistry, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou, Guangdong China

[9]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China

[10]Alibaba Cloud, Hangzhou, Zhejiang, China

[11]College of Information and Engineering, Wenzhou Medical University, Wenzhou, Zhejiang, China

[12]Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, Zhejiang, China

[13]Department of science and development, Chinese academy of sciences, Beijing, China

[14]College of Astronomy and Space Sciences, University of Chinese Academy of Sciences, Beijing, China

[15]Emotion Machine (Beijing) Technology Co., Ltd., Beijing, China

[16]School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China

[17]Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), Wenzhou, Zhejiang, China

[18]School of Physical Sciences, University of Chinese Academy of Sciences, Beijing, China

[*]Corresponding authors are as follows: Fangfu Ye(fye@iphy.ac.cn), Jia Li(lijia@ime.ac.cn), Yuyang Li(liyuyang22@mails.ucas.ac.cn), Jianwei Shuai(jianweishuai@xmu.edu.cn), Mingzhou Tan(SwanLab.cn), Jiaji Wu(wujj@mail.xidian.edu.cn), Boyuan Zheng(zhengboyuan17@mails.ucas.ac.cn)

[†]Contributed equally (Co-first authorship)

Abstract

While AI for Science (AI4S) serves as an analytical tool in the current research paradigm, it doesn't solve its core inefficiency. We propose "Agent for Science" (Agent4S)—the use of LLM-driven agents to automate the entire research workflow—as the true Fifth Scientific Paradigm. This paper introduces a five-level classification for Agent4S, outlining a clear roadmap from simple task automation to fully autonomous, collaborative "AI Scientists." This framework defines the next revolutionary step in scientific discovery.

1 The Basic Characteristics and Development History of Scientific Research

The essence of scientific research is to establish a set of methods to discover laws from data (phenomena), and then use these laws to guide production and transform the world.

The four scientific paradigm revolutions[1] shown in Table 1 are essentially transformations in methods of data acquisition and data processing.

The first scientific paradigm is the empirical paradigm, where the research paradigm of this stage is characterized by observation and induction. Data acquisition mainly relied on naked-eye observation and measurements with simple tools, such as the experimental methodology systematically expounded by Bacon in Novum Organum (1620)[2] and Galileo's telescopic observations[3]. Data processing involved manual recording and simple statistics, establishing causal relationships through repeated observations and emphasizing the reproducibility and intuitiveness of experience.

The second scientific paradigm is the theoretical paradigm. With the development of precision instruments and mathematical tools, scientific research achieved a leap from qualitative to quantitative analysis. Data acquisition shifted from simple observation to controlled experiments and precise measurements, such as Newton's prism experiments on the dispersion of light spectra in 1665–1666[4] and Cavendish's torsion balance experiment to measure the gravitational constant in 1797–1798[5]. Mathematical formulas and equation solving were introduced into data processing, enabling the derivation of universal laws from limited experimental data through mathematical abstraction and establishing a "mathematical" scientific language system.

The third scientific paradigm is the computational science paradigm. With the advent of electronic computers, scientific research broke through the physical limitations of traditional experiments. Data acquisition shifted from physical observation to computer-generated data, generating massive simulated data through mathematical models—examples include nuclear weapons test simulations, weather forecasting models, and molecular dynamics simulations. Data processing underwent a revolutionary shift from manual calculations to numerical computations: the introduction of high-performance computing (HPC) techniques made simulation research on complex systems feasible.

The fourth scientific paradigm is the data-intensive scientific paradigm[6]. With the development of sensor networks, the internet, and artificial intelligence technologies, scientific research has entered an era of "big data-driven" inquiry. Data acquisition now enables massive, high-dimensional, real-time collection through sensors, automated devices, and internet platforms—examples include gene

Table 1: Revolution of the four scientific paradigms

Paradigm	Key Characteristics	Data Acquisition Methods	Data Processing Methods
Empirical Paradigm	Observation-Induction	Naked-Eye Observation/Simple Instruments	Oral/Textual Records + Naive Induction
Theoretical Paradigm	Experimental-Quantitative Theory	Controlled Experiments and Precision Instruments	Mathematical Abstraction + Equation Solving
Computational Paradigm	Numerical Simulation	Computer-Generated Data	Numerical + High-Performance Computing
Data-Driven Paradigm	Big Data	Massive High-Dimensional Data from Sensors/Automated Devices	Machine Learning/Deep Learning

sequencing, radio telescope arrays, and social network data streams. Data processing has evolved from statistical learning to machine learning and now deep learning, endowing computers with the ability to directly extract patterns from high-dimensional data spaces and achieving a fundamental shift from hypothesis-driven to data-driven research.

Each transformation has brought about corresponding scientific research processes and gradually improved the systematic methodologies of scientific research. The current scientific research process primarily involves: formulation of scientific questions, definition of research content, design of research plans, experimentation, and data processing with iteration.

With the continuous improvement of productivity, the rate of new data generation and the volume of knowledge derived from existing data have been growing exponentially. Under the framework of the fourth scientific research paradigm, a past contradiction was the mismatch between the increasing data dimensionality and insufficient data analysis methods—the curse of dimensionality[7]. The development of deep learning has alleviated this contradiction to some extent[8]. However, more fundamentally, from the perspective of productivity, a deeper contradiction lies in the growing volume of scientific research information and the inefficiency of existing research paradigms.

2 Understanding AI4S from the Perspective of Scientific Data Processing

2.1 Definition of AI4S

For the first contradiction, namely the mismatch between increasing data dimensionality and insufficient data analysis methods, AI algorithms—especially deep neural networks—as computational methods in the big data era that can handle higher-dimensional data compared to statistical learning and classical machine learning (or "high-order function fitting methods"), are capable of processing

higher-dimensional data and uncovering more patterns within them. Their application in scientific research constitutes the classical definition of AI4S[9]. For example, AlphaFold leverages the Evoformer algorithm to extract patterns from massive protein data on how amino acid interactions affect their distances, while deep learning potentials (DPMD) use neural networks to learn the relationships between molecular structures and potential energies.

2.2 Development History and Current Status of AI4S

Since expert systems first emerged in the 1980s and deep learning breakthroughs in image/speech domains around 2010 were rapidly applied to scientific computing, AI4S has undergone three critical leaps:

1. In the symbolism stage, relying on expert-defined rules, it faced limitations in modeling complex systems;

2. The statistical machine learning stage focused on low-dimensional feature engineering (e.g., support vector machines, random forests)—while enabling automatic modeling, it struggled with exponentially increasing data dimensions;

3. The deep representation learning stage introduced "high-order function fitters" like convolutional networks, attention mechanisms, and graph neural networks into research scenarios, achieving breakthroughs in protein folding (AlphaFold 2/Evoformer), molecular potential energy surfaces (DPMD), inverse materials design (GraphGPT-Materials), etc.

3 Agent4S

3.1 Definition of Agent4S

Starting from the development of Agent technology and combined with the existing paradigms of scientific research, we propose five levels of Agent4S, as shown in Table 2.

The earliest form of agents originated from the LLM + Prompt mechanism, supplemented by Function Calling (FC) to map language outputs to API calls. This has been widely deployed in industry for customer service Q&A and content generation; in scientific research, it manifests as the intelligent encapsulation of single scientific tools, such as literature retrieval, database query, image annotation, etc. (specific cases and references needed, note: the implementation must be a single agent). We define this as Level 1 (L1) of Agent4S—Automation of a Single Scientific Tool—referring to the automation of a simple task in research.

With the maturation of Task-/Workflow Orchestration frameworks (Airflow, Dagster, Ray Serve, etc.), LLM Agents can now maintain state and dependency relationships across multi-step tasks. Integrated "instruction-execution-callback" loops have been widely deployed in industrial scenarios such as robotic process automation (RPA), A/B testing, and advertising. The most typical scientific research practice is the "end-to-end data pipeline"—for example, integrating four steps of sequencing quality control, alignment, quantification, and statistics into a single-trigger automated

Table 2: Agent4S Hierarchy

Level	Definition	Agent Hierarchy	Underlying Agent Technology	Agent Technology Status	Research Phase	Paradigm Transformation Mode	Current Research Applications	Implementation Challenges
L1	Automation of a Single Tool	AI Agents (AI Proxies)	Prompt Engineering + FC + Workflow	Mature	Scientific Data Generation/Acquisition	Automated Tools for Fixed Processes	Preliminary Cases with Clear Development Trajectory	Hardware Digitization
L2	Automation of Complex Processes							Data Transmission Robustness
L3	Intelligence of a Single Process	Agentic AI (Intelligent Agent AI)	Reasoning + Context Engineering + MCP	Early Development	Scientific Data Analysis/Processing	Intelligent Collaborative Partner for Innovative Processes	No Clear Pattern Established	MCP-Enabled Research Tools and Context Engineering
L4	Full-Process Intelligence (Single Laboratory)							Laboratory Hardware/Software Integration
L5	Collaboration of Multiple Intelligent Processes (Multiple Laboratories)	Multi-Agent System	A2A (Agent-to-Agent)	Early Development	Interdisciplinary Research	Intelligent Network for Cross-Disciplinary Collaboration	No Clear Pattern Established	Breaking Disciplinary (Laboratory) Barriers

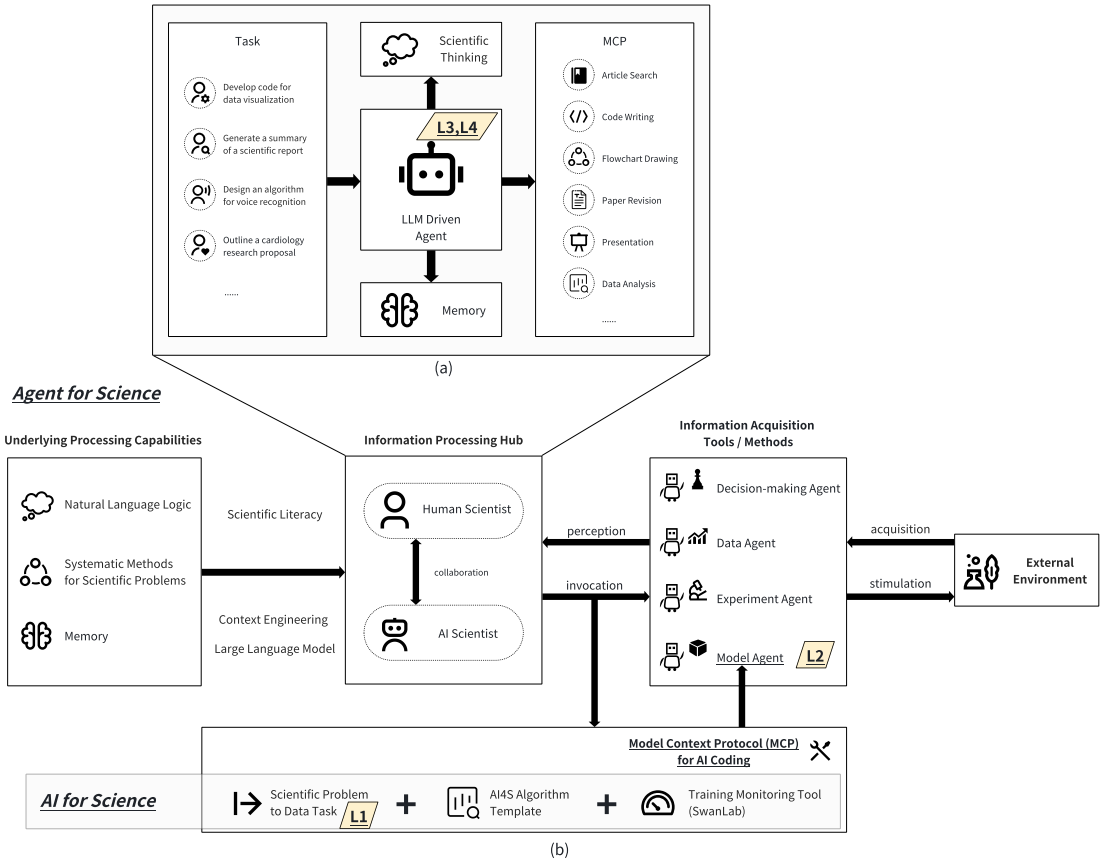


Figure 1: Technical framework

task flow; or chaining structure generation, first-principles calculations, and database storage into one-click scripts for high-throughput materials computation (specific cases and references needed, note: implementation must be workflow-driven multi-agent, where workflows can be linear, parallel, or autonomously planned). We define this as Level 2 (L2) of Agent4S—Automation of Complex Scientific Pipelines—referring to orchestrating multiple L1 tools via workflow methods into reusable, low-human-intervention research pipelines to automate a complex fixed process in scientific research.

Currently, with the development of Reasoning Frameworks (ReAct, Tree-of-Thought, Graph-of-Thought) and Long-Context Memory Engineering, LLM Agents have evolved from "pipeline executors" to master agents capable of chain reasoning, real-time decision-making, and self-reflection. They achieve intelligent tool invocation through the Model-Context-Protocol (MCP), marking the transition from AI Agents to Agentic AI [10]. This paradigm has validated the feasibility of "observation-planning-iteration" closed-loop decision-making in industrial scenarios like AIOps, real-time risk control, and AI coding. Although AI-related technologies are newly developed and have limited applications in scientific research, it is foreseeable that the next stage will involve an AI Scientist with autonomous planning and long-term memory capabilities—capable of invoking various L2 tools via MCP to realize a closed-loop of "planning-tool usage-data analysis" for specific research

processes, enabling intelligent iteration. We define this as Level 3 (L3) of Agent4S—Intelligent Single-Flow Research.

Looking ahead, as Agent capabilities in memory length, multi-step planning, and MCP invocation continue to advance—alongside the development of embodied intelligence and further integration/data connectivity across all hardware-software components in laboratory workflows—the L3-level AI Scientist will be able to participate in the full "hypothesis-experiment-analysis" closed loop of research projects and enable intelligent iteration. We define this as Level 4 (L4) of Agent4S—Lab-Scale Closed-Loop Autonomy.

Ultimately, with the advancement of full-process intelligence in each laboratory and relying on the A2A protocol, multiple super-agents will be able to communicate with each other, enabling the formation of an intelligent network for cross-disciplinary collaboration driven by multi-agent systems in scientific research. This constitutes the Level 5 (L5) stage of Agent4S and represents the final form of the fifth scientific paradigm driven by artificial intelligence.

From the perspective of general research processes, AI4S represents a more advanced method within the data analysis phase; L1 of Agent4S refers to the automation of a simple tool used in a specific research sub-task (e.g., literature retrieval or data annotation), involving single-agent operations without cross-component integration. L2 involves the automation of an entire research sub-process (e.g., a data pipeline combining sequencing QC and statistical analysis), differing from L1 in featuring a multi-agent-driven complete data closed loop orchestrated via workflow frameworks. L3 denotes the intelligence of a specific research workflow, encompassing autonomous planning, tool invocation, data processing, and result collation. Its key distinction from L2 lies in the presence of a super-agent with context engineering capabilities that autonomously plans, orchestrates, and iterates the workflow. L4 represents full-process intelligence in scientific research, covering the entire lifecycle from scientific question formulation, research design, hypothesis generation, experimental simulation, to data interpretation. It differs from L3 in enabling end-to-end participation across the entire research process rather than individual workflows.

In terms of future human-computer interaction models, as illustrated in Figure 1: Levels L1 and L2 involve the automation of fixed processes, primarily applied to the historically machine-driven data generation/acquisition stages. Fundamentally, they remain tools within the research workflow. Levels L3 and L4 represent the intelligence of innovative processes with stochastic and emergent characteristics, focusing on data analysis and interpretation. These constitute super-agents analogous to scientists—i.e., AI Scientists—capable of autonomous reasoning. A plausible future interaction model is scientists collaborating with AI Scientists as partners, jointly invoking various tools (L3/L4), which themselves may integrate numerous agents (L1/L2). Level L5 refers to interactions among multiple AI Scientists, forming an intelligent network for cross-disciplinary research.

In terms of research organizational forms, as shown in Figure 2: L1 and L2 represent the automation of tools used by scientists in specific research sub-tasks, where agents remain as passive instruments within the workflow. L3 marks a shift: agents evolve from "tools" to collaborative assistants with defined roles, capable of participating in targeted workflow stages under human guidance. L4 elevates agents to the role of central coordinators in full experimental cycles—analogous

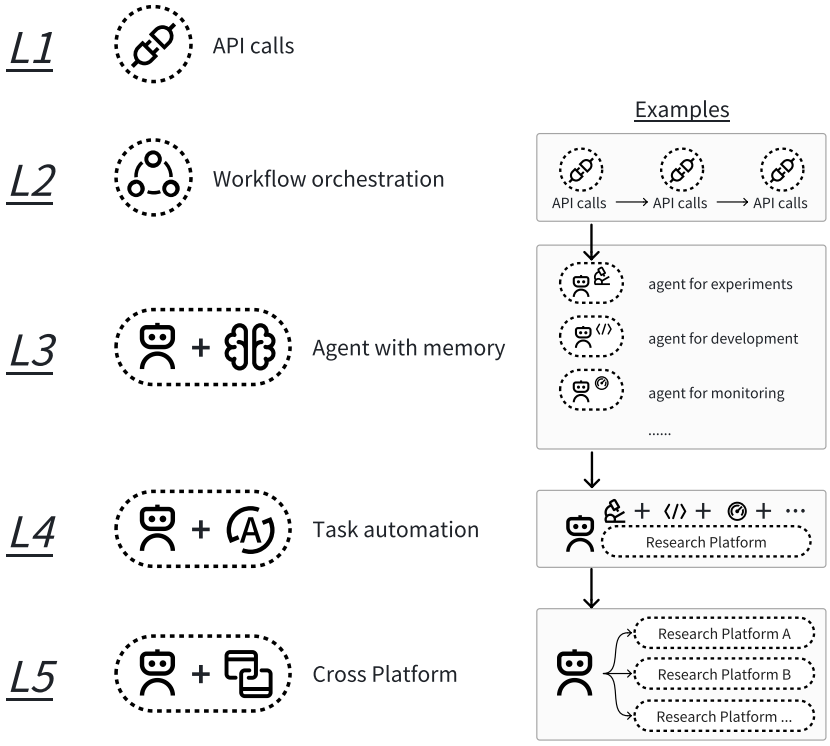


Figure 2: Five levels of Agent4S

to an AI project leader/laboratory director that oversees the entire research lifecycle, from proposal development to completion, with full situational awareness and autonomous decision-making capabilities. L5 enables cross-disciplinary collaboration across multiple laboratories, where information flows through a network of super-agents, facilitating seamless interaction and knowledge exchange between distributed intelligent entities.

4 Significance

More than just a nominal definition: In AI4S, "AI" denotes a data analysis methodology—specifically, using AI algorithms to address high-dimensional data challenges in scientific research. In Agent4S, "Agent" represents a new productivity tool—Agent4S signifies agent-driven automation and intelligence in scientific research, emerging as a transformative productivity paradigm for data acquisition and processing. The connection between AI4S and Agent4S in the entire scientific research process: Classical AI4S serves as the fundamental building blocks of algorithmic agents in the data processing stage within the Agent4S paradigm. Both fall under the category of applying AI to scientific research. This analysis, from an AI technical perspective, provides concrete implementation and development frameworks for intelligent scientific research—moving beyond mere macro-level visions of the past. For AI practitioners, it clarifies how to advance AI technologies and integrate them with scientific workflows; for domain scientists, it identifies clear entry points for AI integration in their

fields. Additionally, by outlining the scientific challenges at each developmental level of the Agent4S hierarchy, it guides researchers on aligning their work with AI advancements to systematically elevate the intelligence levels of research processes.

This work resolves the confusion between algorithms and intelligence: fundamentally, the technical essence of AlphaFold differs from that of cutting-edge intelligent laboratories. In terms of scope, the AI4S algorithms behind AlphaFold—at their core computational methods—represent a single component within the scientific research workflow and constitute a subset of the future Agent4S framework. Regarding the classification of the Fifth Scientific Paradigm, numerous prior taxonomies existed but primarily focused on the degree of research automation, serving more as automation ratings rather than providing actionable technical roadmaps or trend analyses. Our classification innovatively integrates the technological evolution of Agents with the levels of research automation/intelligence, clearly delineating future development trajectories and critical challenges. This approach fundamentally advances the transformation toward the Fifth Scientific Paradigm by bridging theoretical visions with practical implementation frameworks.

In fact, the concept of the Fifth Scientific Paradigm was proposed early on [11], which outlined its characteristics:

1. Full integration of artificial intelligence into scientific, technological, and engineering research, enabling knowledge automation and intelligence across the entire research process;

2. Human-machine symbiosis, where emergent machine intelligence becomes an integral part of research, giving rise to tacit knowledge and machine-generated hypotheses;

3. Focus on complex systems as primary research objects, effectively addressing combinatorial explosion problems with extremely high computational complexity;

4. Orientation toward non-deterministic problems, where probabilistic and statistical reasoning play an expanded role in research;

5. Cross-disciplinary collaboration as the mainstream research mode, achieving integration of the first four paradigms—especially the convergence of first-principles model-driven and data-driven approaches;

6. Heavy reliance on large-scale platforms characterized by large models, with close integration between scientific research and engineering implementation.

These characteristics are incisive, yet the paradigm lacked a strict definition at the time—due to the underdevelopment of agent technologies, many past AI algorithms for data processing were essentially refinements within the data-driven Fourth Scientific Paradigm but were erroneously categorized as part of the Fifth. From the Agent4S perspective, we can now naturally distinguish between "improvements within the data-driven paradigm" and "scientific paradigm shifts brought by new productivity tools."

In summary, both AI4S and Agent4S represent applications of AI in scientific research—where AI4S positions "AI" as a data analysis methodology, using AI algorithms to address high-dimensional

data challenges and resolve the historical contradiction between data dimensionality and computational methods within the Fourth Scientific Paradigm. In contrast, Agent4S defines "Agent" as a new productivity tool, enabling agent-driven automation and intelligence in scientific research to tackle the contradiction between information richness and the productivity limitations of past paradigms.

It refers to using Agent technology to drive automation and intelligent scientific research, addressing the contradiction between information richness and productivity under past paradigms.

More than a nominal definition: In AI4S, "AI" denotes a data analysis methodology—specifically, using AI algorithms to address high-dimensional data challenges in scientific research. In Agent4S, "Agent" represents a new productivity tool—Agent4S signifies agent-driven automation and intelligence in research, emerging as a transformative paradigm for data acquisition and processing.

References

1. Hey T, Tansley S, Tolle K, and Gray J. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009. URL: `https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/`.
2. Bacon F. Instauratio Magna (Novum Organum). Publisher Unknown, 1620.
3. Taton R, Wilson C, and Hoskin M. Planetary Astronomy from the Renaissance to the Rise of Astrophysics, Part A, Tycho Brahe to Newton. General History of Astronomy. Cambridge University Press, 2003. URL: `https://books.google.com/books?id=hMgXh8jMSGgC`.
4. Newton I. A letter of Mr. Isaac Newton, Professor of the Mathematicks in the University of Cambridge; containing his new theory about light and colors: sent by the author to the publisher from Cambridge, Febr. 6. 1671/72; in order to be communicated to the R. Society. Philosophical Transactions of the Royal Society of London 1672;6:3075–87.
5. Cavendish H. XXI. Experiments to determine the density of the earth. Philosophical Transactions of the Royal Society of London 1798;88:469–526.
6. Gray J and Szalay A. eScience-A transformed scientific method. Presentation to the Computer Science and Technology Board of the National Research Council 2007.
7. Hammer P. Adaptive control processes: a guided tour (R. Bellman). 1962.
8. Bach F. Breaking the curse of dimensionality with convex neural networks. Journal of Machine Learning Research 2017;18:1–53.
9. Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence. Nature 2023;620:47–60.
10. Sapkota R, Roumeliotis KI, and Karkee M. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenge. arXiv preprint arXiv:2505.10468 2025.
11. LI G. AI4R: The fifth scientific research paradigm. Bulletin of Chinese Academy of Sciences (Chinese Version) 2024;39:1–9.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 冯一纯

方向： 人工智能

单位： 中国科学院大学前沿交叉科学学院

题目： 用多智能体协作的强化学习系统打造主动问诊的AI医生

DoctorAgent-RL: A Multi-Agent Collaborative Reinforcement Learning System for Multi-Turn Clinical Dialogue

Yichun Feng[1,2,3†], Jiawei Wang[4†], Lu Zhou[3], Yikai Zheng[5], Zhen Lei[1,2*], Yixue Li[3,6,7,8,9*]

[1*]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, 380, Huaibei Town, Huairou District, Beijing, 100049, Beijing, China.
[2]Institute of Automation, Chinese Academy of Sciences, No. 95, Zhongguancun East Road, Beijing, 100190, China.
[3]Guangzhou National Laboratory, No. 9 XingDaoHuanBei Road, Guangzhou, 510005, Guangdong, China.
[4]Department of EEIS, University of Science and Technology of China, No. 443, Huangshan Road, Shushan District, Hefei, 230026, Anhui, China.
[5]Department of EEIS, University of Science and Technology of China, No. 443, Huangshan Road, Shushan District, Hefei, 230026, Anhui, China.
[6]GZMU-GIBH Joint School of Life Sciences, The Guangdong-Hong Kong-Macau Joint Laboratory for Cell Fate Regulation and Diseases, Guangzhou Laboratory, Guangzhou Medical University, Guangzhou, 511436, China.
[7]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China.
[8]Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, 200030, China.
[9]Bioland Laboratory, Guangzhou, 510005, China.

*Corresponding author(s). E-mail(s): zhen.lei@ia.ac.cn; li_yixue@gzlab.ac.cn;

Contributing authors: fengyichun22@mails.ucas.ac.cn; wangjiawei@mail.ustc.edu.cn; zhou_lu@gzlab.ac.cn; iiiauthor@gmail.com;
†These authors contributed equally to this work.

Abstract

Large language models (LLMs) struggle in real-world clinical consultations. Single-turn consultation systems require patients to describe all symptoms at once, which often leads to unclear complaints and vague diagnoses. Traditional dialogue models, constrained by static supervised learning, are limited to superficially imitating existing dialogue patterns and lack the ability to actively construct understanding in dynamic interactions, thus failing to achieve genuine clinical reasoning. To address these challenges, we propose DoctorAgent-RL, a reinforcement learning (RL)-based multi-agent collaborative framework. Our work fundamentally reframes the role of agent from a passive knowledge reproducer to an active knowledge constructor. Within this framework, a medical consultation is modeled as a dynamic decision-making process under uncertainty. The core intelligence of the agent is shifted from knowing the answer to learning and mastering a questioning methodology aimed at achieving an optimal diagnosis. Through strategic questioning, it guides the progressive emergence of key patient information in multi-turn dialogues. To support this high-fidelity simulation of the real diagnostic process, we constructed MTMedDialog, a novel English multi-turn medical consultation dataset designed for dynamic, interactive training. To validate its real-world effectiveness, rigorous evaluations including blinded human assessments and trials with real patients were conducted. DoctorAgent-RL outperformed frontier models and achieved a 70% exact diagnostic match rate, confirming its potential as a collaborative tool. By handling initial screenings, it can free clinicians to focus on complex cases, thereby addressing critical issues like physician shortages and misdiagnosis risks while alleviating the strain on healthcare resources.

Keywords: Large Language Models, Medical Multi-Turn Consultation, Reinforcement Learning, Multi-Agent System

1 Introduction

Large language models (LLMs) like ChatGPT [1], LLaMA [2], and ChatGLM [3] show significant potential for assisting in clinical decision-making [4]. This capability is urgently needed in the context of a global doctor shortage and high-pressure healthcare environments [5]. In these settings, time pressure on clinicians often leads to rapid diagnoses based on incomplete information, increasing misdiagnosis risks [6], while routine consultations consume resources that could be allocated to more critical cases [7]. Therefore, developing an intelligent system to improve primary care efficiency

holds significant practical importance for optimizing resource allocation and alleviating healthcare strain [8].

The limitations of current systems are evident on multiple levels. First, single-turn question-answering systems, such as MedAlpaca [9] and BioMistral [10], require patients to provide a comprehensive description of their condition all at once [11]. This contradicts the clinical reality of complex situations where chief complaints are unclear and information is vague [12], often leading to overly broad or even risky diagnostic suggestions [13]. Second, even traditional models that support multi-turn dialogue, like Bianque [14] and DialoGPT [15], rely heavily on static supervised learning, which amounts to a superficial imitation of existing dialogue transcripts [16]. These models lack a dynamic decision-making mechanism and cannot intelligently adjust their questioning strategies based on the real-time progress of the conversation to probe for key information. Even systems that integrate knowledge graphs [17–19] fundamentally rely on predefined dialogue paths, failing to achieve genuine clinical reasoning. This fundamentally solidifies the role of LLM as a passive knowledge reproducer, whose intelligence remains limited to reproducing existing knowledge rather than actively exploring the unknown.

In recent years, research has begun to explore more dynamic approaches. For instance, systems like AMIE [20] optimize the diagnostic dialogue itself through self-play simulations, though their adaptability is limited by a reliance on static training data. Other multi-agent systems, such as Agent Hospital [21] and AI Hospital [22], expand this simulation to create entire hospital environments with multi-role collaboration protocols. However, they also face challenges, including the risk of factual inconsistencies and a failure to account for the diversity of real-world treatment options. Concurrently, the application of RL in the medical field is transitioning from static decision-making to dynamic interaction paradigms. MedVLM-R1 [23] and Med-R1 [24] employ the Group Relative Policy Optimization (GRPO) framework [25], enhancing diagnostic accuracy but exhibiting limited adaptability to new modalities. HuatuoGPT-o1 [26] enhances clinical reasoning through verifiable question generation but is constrained by its reliance on converting data to multiple-choice formats. MedRIA [27] employs actor-critic frameworks to optimize inquiry efficiency in emergencies but requires manual feature engineering for complex cases , while PPME [28] leverages clinical experience replay at the cost of poor generalization. These studies collectively indicate that RL requires further refinement in dynamic interaction objective modeling and clinically oriented evaluation frameworks to address the complexity and uncertainty inherent in real-world medical practice

To address these challenges, we propose DoctorAgent-RL—a multi-agent collaborative reinforcement learning (RL) framework that reformulates clinical reasoning as a Markov Decision Process (MDP), as shown in Figure 1. Our work aims to transform the role of LLM from a passive knowledge reproducer to an active knowledge constructor. Within this framework: (1) A high-fidelity patient agent based on LLMs, which generates pathologically consistent responses while mimicking the diversity of real-world communication; (2) A doctor agent initialized by cloning medical actions from real consultation records and refined through reinforcement learning (RL) to master

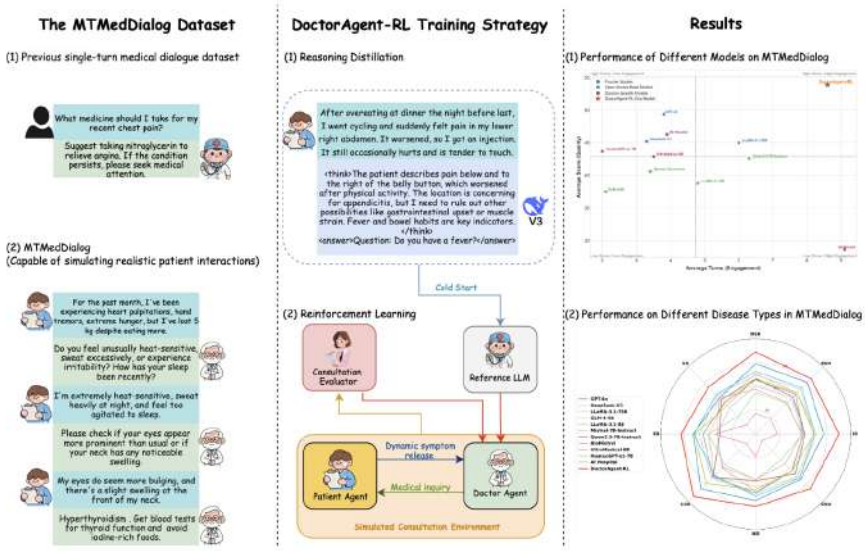


Fig. 1: Overview of this work. DoctorAgent-RL achieves state-of-the-art on the MTMedDialog task.

effective questioning strategies; and (3) A consultation evaluator that provides multi-dimensional rewards based on diagnostic accuracy, patient information responsiveness, and the standardization of questions. Here, the core intelligence of the doctor agent no longer lies in knowing the answer, but rather, through RL training, it learns and masters a questioning methodology aimed at achieving optimal diagnosis. By employing strategic questioning, it guides the gradual emergence of key patient information throughout multi-turn dialogues.

While existing datasets provide valuable examples of multi-turn medical conversations, they consist of static transcripts. This limits their use for training agents that need to learn adaptive questioning strategies through interactive simulation. To address this gap, we constructed MTMedDialog, the first English dataset specifically designed to support dynamic, stateful patient simulations. Each dialogue is grounded in a comprehensive patient profile, allowing for realistic, turn-by-turn information revelation based on the clinician agent's diagnostic reasoning process. This high-fidelity simulation of the actual diagnostic process is an essential prerequisite for training a knowledge constructor, as it enables the agent to learn and validate its questioning strategies in a near-realistic environment. DoctorAgent-RL redefines the diagnostic process through reward-based strategic policy optimization: The doctor agent continuously optimizes its questioning strategy within a RL framework, obtaining immediate feedback through multi-turn interactions with the patient agent, and dynamically adjusting the information-gathering path based on the comprehensive rewards provided by the consultation evaluator. This end-to-end RL fine-tuning mechanism enables LLMs to autonomously evolve interaction strategies that align

with clinical reasoning logic, rather than simply imitating surface patterns in existing dialogue data.

To validate the effectiveness of our proposed new framework in real-world scenarios, we conducted a series of rigorous evaluations. Experimental results demonstrate that DoctorAgent-RL is superior to existing models in both multi-turn questioning capability and final diagnostic performance. Specifically, we conducted in-depth validations that included blinded human assessments and direct interactions with real patients. Notably, our agent achieved a 70% exact diagnostic match rate in real-world consultations, which strongly confirms its powerful practical capabilities and potential for clinical application. This research aims to address real-world problems such as the global physician shortage and the risk of misdiagnosis in high-pressure environments. DoctorAgent-RL is positioned as an efficient collaborative tool rather than a replacement for human doctors. By handling tasks like initial screening and information gathering, it frees clinicians from repetitive work to enhance the overall efficacy of the human-AI collaborative system, offering a viable technical path to alleviate the strain on healthcare resources.

2 Results

2.1 Comparative Performance Evaluation of Multi-Model Approaches on MTMedDialog

The experimental results are shown in Figure 2(a). The left Y-axis, Score (%), represents Diagnosis and Recommendation Accuracy, a metric that uses a large language model (Qwen2.5-32B-Instruct) to assess the semantic consistency (on a 0-5 scale) between the model's output and the gold standard, which is then converted to a 100-point scale; the right Y-axis, Avg. Turns, represents the average interaction turns required for the model to reach a final diagnosis. As shown in the figure, DoctorAgent-RL achieved a comprehensive average score of 53.9%, significantly outperforming all frontier, open-source base, and domain-specific models. More importantly, its box plot not only shows the highest median score, but its interquartile range is also positioned entirely above all competitors, indicating that our model performs more stably and with less variance across different disease types. This stable advantage on diseases that require in-depth consultation demonstrates its superior ability to simulate the clinical consultation process of doctors.

The figure clearly illustrates the relationship between performance and interaction frequency. For instance, the domain-specific model BioMistral has the highest average number of turns, yet its score box plot is at the lowest level and shows the widest distribution, indicating that its numerous questions are repetitive and ineffective. In contrast, some models (e.g., GLM-4, HuatuoGPT-o1-7B), despite having decent scores or very few turns, failed to acquire sufficient information through effective questioning or improperly combined multiple questions, which hindered the diagnosis of complex conditions. While frontier models strictly adhere to instructions, their lack of professional knowledge leads to questions that fail to address critical points. These findings provide strong evidence that effective diagnosis stems from the quality of questions, not the mere quantity of interactions.

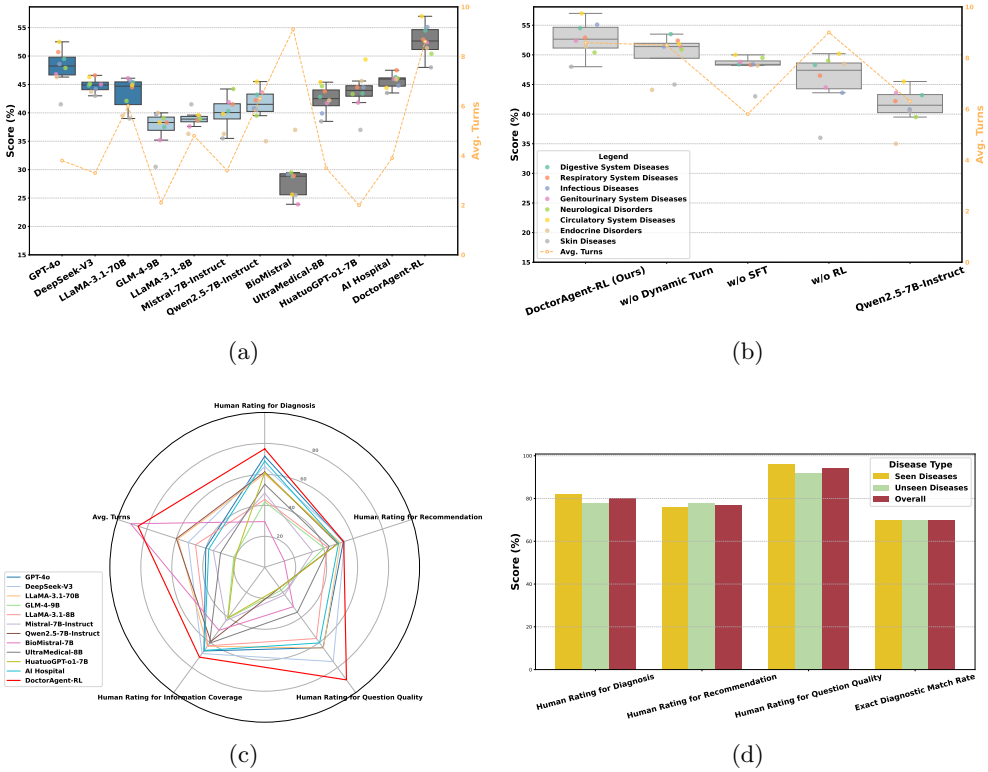


Fig. 2: (a) Main results by disease category on MTMedDialog dataset, showing the average of diagnostic accuracy and recommendation accuracy scores. Dark blue boxes represent Frontier Models, light blue boxes represent Open-Source Base Models, and dark gray boxes represent Domain-Specific Models. (b) Performance comparison of different fine-tuning methods for Qwen2.5-7B-Instruct on MTMedDialog. (c) Human randomly selected 100 data samples to compare the performance of different models on the MTMedDialog dataset. (d) Performance of DoctorAgent-RL in real-world interactive scenarios with actual patients.

The success of DoctorAgent-RL is attributed to its phased training strategy. The initial fine-tuning equips the model with systematic consultation knowledge, while the subsequent reinforcement learning allows it to move beyond imitating static patterns to truly learn and master a questioning methodology aimed at achieving an optimal diagnosis. The core of this methodology is not simply to increase the number of questions, but to actively guide the progressive emergence of key patient information in multi-turn dialogues through strategic, layered questioning. The relatively high Avg. Turns of DoctorAgent-RL is not ineffective or repetitive like in other models, but a direct reflection of its execution of this rigorous questioning methodology. Each interaction turn serves the final diagnostic goal, and therefore, this seemingly high interaction cost ultimately translates into the highest performance score among

all models. This strongly proves that DoctorAgent-RL has learned a truly intelligent and efficient interaction strategy, and its interaction frequency is entirely valuable and reasonable.

2.2 Adaptive Fine-Tuning Strategies for Task-Specific Optimization

Experimental results demonstrate the significant advantages of our proposed two-stage optimization framework (SFT + RL) in medical dialogue tasks. As shown in Figure 2(b), DoctorAgent-RL achieves a superior average score of 53.9, outperforming all baseline models. The key advantage stems from our phased strategy: (1) SFT on doctor-patient dialogues establishes reasonable questioning baselines through behavioral cloning, followed by (2) RL optimization that enables dynamic adjustment of questioning strategies for high-value information acquisition. Ablation studies on three critical components further validate our design:

***w/o* Dynamic_Turn**: When trained with fixed budget of turns during RL, the model shows only a 1.2% performance drop in matched scenarios but reveals strategy rigidity during inference—it mechanically adheres to the training budget of turns regardless of specified variations, proving impractical for real-world deployment.

***w/o* SFT**: Direct RL training without SFT initialization causes a 5.5% average score degradation with the lowest average turns. While capable of planning effective questions for information gathering, the model demonstrates insufficient initiative in question generation due to the absence of behavioral cloning foundations.

***w/o* RL**: SFT-only training results in a 6.5% lower average score despite having the highest turn count. The model memorizes question sequences from training data without truly understanding diagnostic logic, leading to mechanical questioning rather than strategic information acquisition.

The comprehensive performance of DoctorAgent-RL confirms each component's necessity: SFT establishes reliable behavioral baselines, RL injects dynamic decision-making capability, and adaptive turn mechanisms ensure strategic flexibility—together forming a reproducible paradigm for task-oriented medical dialogue optimization.

2.3 Human Evaluation and In-depth Analysis of Multi-Model Approaches on MTMedDialog

To conduct a more detailed evaluation of the comprehensive performance of different models in simulating real-world clinical consultation scenarios, we performed a human evaluation. Specifically, we randomly selected 100 samples from the MTMedDialog test set and then anonymized and shuffled the dialogues generated by all models for these samples. Evaluators, blinded to the model origins, independently scored the dialogues against the ground truth references. The scores were then mapped back to the respective models using their IDs to ensure the objectivity and fairness of the evaluation. Evaluation dimensions included Diagnostic Accuracy, Recommendation Reasonableness, Question Quality, and Information Coverage. The detailed results are presented in Figure 2 (c).

The human evaluation results clearly indicate that our proposed DoctorAgent-RL outperforms all other models, achieving the highest scores in the three core metrics of Diagnostic Accuracy, Question Quality, and Information Coverage. It comprehensively surpassed frontier closed-source models, open-source base models, and other domain-specific models. This demonstrates that our method not only excels in the final diagnostic task but also holds a significant advantage in the strategic effectiveness of the consultation process.

In contrast, other models generally showed significant deficiencies in instruction following and interaction quality. Frontier models, representing the state-of-the-art in large language models, demonstrated strong overall capabilities. Among them, GPT-4o achieved high scores across the board, making it the best alternative to our model. DeepSeek-V3 excelled in question quality but sometimes violated the one-question-per-turn instruction. LLaMA-3.1-70B showed stable performance but lagged slightly behind the other two, similarly posing multiple questions in a single turn. While this increased information coverage, it also highlighted its weakness in instruction following. Open-source base models commonly suffered from poor instruction adherence and limited medical knowledge. For instance, Qwen2.5-7B-Instruct tended to provide a diagnosis directly rather than gathering information through questions and repeatedly violated the single-question rule. Mistral-7B-Instruct had the worst performance in instruction following, often failing to ask questions or producing repetitive, irrelevant output. Although LLaMA-3.1-8B could formulate relatively professional questions, its diagnostic accuracy was low due to insufficient internal knowledge. The advantages of DoctorAgent-RL were particularly prominent among domain-specific models. Other models in this category exhibited numerous issues. HuatuoGPT-o1-7B and BioMistral-7B displayed severe problems with instruction following, even exhibiting "role reversal" by mimicking patient symptom descriptions. HuatuoGPT-o1-7B rarely asked questions and often included its overthinking process in the answers. BioMistral-7B had the highest average number of turns, but this was due to incessant, repetitive, and irrelevant questioning, leading to inefficiency and the lowest scores across all metrics. This phenomenon strongly proves that the key to effective multi-turn dialogue is question quality, not quantity. UltraMedical-8B severely violated the single-question instruction by asking five to six questions at once, resulting in too few turns for in-depth, progressive information gathering. AI Hospital, a relatively high-performing domain-specific model, achieved performance close to that of frontier models but still showed a clear gap compared to our model.

In summary, our DoctorAgent-RL model, through SFT and RL, has learned to conduct high-quality, multi-turn interactions while adhering to the one question at a time instruction. The model is able to act like a real doctor, effectively gathering key information through a progressive and logical chain of questions, thereby achieving the highest diagnostic accuracy and information coverage.

2.4 Performance Validation in Real Clinical Scenarios

Automated evaluations and testing based on LLM-simulated patient datasets struggle to fully capture a model's actual performance in real, dynamic conversational

interactions. To comprehensively assess the end-to-end capabilities of DoctorAgent-RL in real clinical scenarios, we designed and conducted a real-world human-machine interaction evaluation. For this purpose, we recruited 20 real patients, ranging in age from their 20s to 80s, to engage in real-time online consultations with our model regarding their actual health conditions, which covered 15 common disease types. Subsequently, professional evaluators scored the complete dialogue process across multiple dimensions.

As detailed in Figure 2 (d), DoctorAgent-RL demonstrated exceptional performance in these real-world interaction tests. The model achieved consistently high scores for both diagnostic accuracy and recommendation rationality, affirming its capacity to make reliable clinical judgments based on dynamically gathered information. Furthermore, the high scores in question quality, coupled with an efficient number of average dialogue turns, indicate that the model's path to achieving high accuracy is both clinically logical and effective.

To further scrutinize the model's clinical precision, we calculated the Exact Diagnostic Match Rate(ED), which measures the percentage of diagnoses that perfectly matched the ground truth. DoctorAgent-RL achieved a rate of 70%, signifying that in the vast majority of interactions, it not only provided a reasonable assessment but correctly identified the patient's specific condition. This high rate of exact matches, combined with the strong overall performance scores, confirms that DoctorAgent-RL possesses the ability to stably, efficiently, and accurately complete clinical dialogue tasks in real-world interactive scenarios.

2.5 Generalization Assessment on HealthBench

To further evaluate the generalization capability of DoctorAgent-RL on unseen medical data distributions, we conducted comprehensive evaluations on HealthBench [29], a standardized benchmark for medical AI systems covering eight evaluation dimensions: Average Score, Emergency Referrals, Health Data Interpretation, Communication Skills, Global Health Knowledge, Hedging Appropriateness, Context-Seeking Behavior, and Complex Response Handling. Following the official HealthBench evaluation protocol, all models were assessed using GPT-4.1 as the automated scoring mechanism to ensure methodological consistency. Notably, both the Qwen2.5-7B-Instruct and our DoctorAgent-RL were evaluated under identical conditions using the official HealthBench testing infrastructure.

The experimental results are presented in Figure 4(a). DoctorAgent-RL achieved an average score of 22.3%, ranking first among all evaluated open-source small-scale models. This strong performance on unseen data distributions is particularly noteworthy given that DoctorAgent-RL was specifically trained on MTMedDialog, a multi-turn medical consultation dataset emphasizing active information gathering through strategic questioning and progressive diagnosis. The substantial improvement over the base model and competitive results compared to models trained on broader medical corpora demonstrate that our reinforcement learning framework enables the agent to learn transferable clinical reasoning patterns rather than merely memorizing consultation templates. The model's ability to maintain robust performance across diverse evaluation dimensions—ranging from emergency triage to nuanced communication

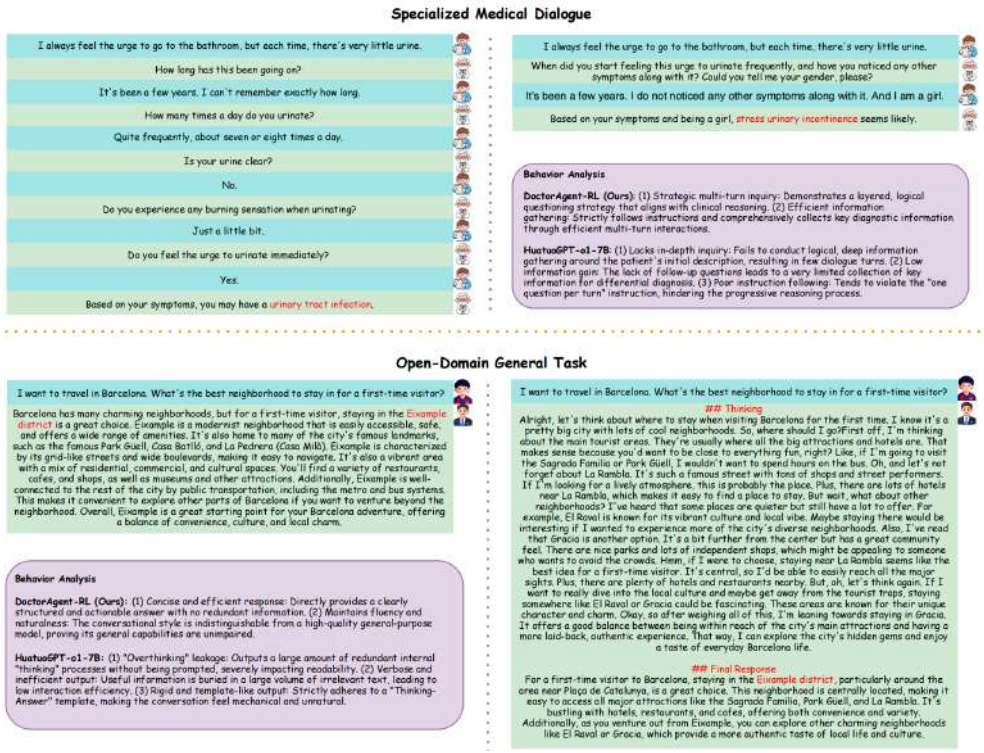


Fig. 3: Qualitative comparison of dialogue outputs from DoctorAgent-RL (left column) and HuatuoGPT-o1-7B (right column) during interactions with a human user. The top row showcases performance on a specialized medical task, while the bottom row demonstrates capabilities on an open-domain general task. This comparison highlights the differences in dialogue quality, fluency, and the ability to maintain general-purpose skills after domain-specific training. In the dialogues, blue boxes indicate the human user's input, and green boxes indicate the LLM's output.

skills—suggests that the strategic information-gathering methodology acquired during training generalizes effectively beyond the specific consultation scenarios encountered during development. These results provide compelling evidence that training medical AI systems to actively construct knowledge through dynamic interaction, rather than passively reproducing existing knowledge, leads to more generalizable and clinically versatile capabilities even when evaluated on data distributions substantially different from the training corpus.

2.6 In-depth Analysis of Generalization and General-Purpose Capabilities

Having validated the model's overall performance in real-world scenarios, we proceeded with an in-depth analysis focusing on two dimensions that are critical for evaluating

the maturity of a specialized LLM: its generalization ability and its general-purpose capabilities. The former assesses the model's capacity to handle unseen clinical cases, while the latter evaluates whether its foundational conversational skills remain intact after domain-specific training.

We began by analyzing the performance discrepancies between disease types the DoctorAgent-RL had seen versus those it had not been exposed to during its training phase. As illustrated in Figure 2(d), the results for both groups of cases showed negligible differences across core metrics. This suggests that the model has acquired a transferable clinical reasoning framework rather than relying on rote memorization of disease-specific patterns, thereby demonstrating strong generalization capabilities.

To further evaluate the robustness and general-purpose capabilities of DoctorAgent-RL, we conducted a qualitative analysis where human evaluators interacted with multiple models across a range of scenarios, covering specialized medical inquiries, open-domain tasks (e.g., travel planning), and role-playing exercises. In our tests, we observed significant behavioral differences across different training paradigms. For instance, domain-specific SFT LLMs like BioMistral exhibited severe instruction-following deficiencies, leading to unsatisfactory performance in both medical and general tasks; its effectiveness is likely confined to a narrow distribution highly similar to its training data. In contrast, a domain-specific RL LLM like HuatuoGPT-o1-7B, despite performing well on the medical task, showed clear signs of being over-tuned, resulting in rigid, template-like outputs and a severe degradation of its general-purpose conversational abilities. In contrast, DoctorAgent-RL successfully avoids this issue. It not only excels in specialized medical dialogues but also fully preserves the powerful general-purpose conversational skills of its base model, remaining fluent and efficient in non-medical tasks. The model is also robust in terms of role consistency; when instructed to switch roles and act as a patient, it can describe symptoms naturally without any confusion from its core doctor training. Figure 3 shows a set of representative dialogues that vividly illustrates the behavioral differences between our model and HuatuoGPT-o1-7B on both a medical and a travel planning task. Based on this qualitative evaluation, we conclude that DoctorAgent-RL is an intelligent agent that achieves deep domain specialization while fully preserving its general-purpose capabilities.

2.7 Impact Analysis of Budget of Turns on Diagnosis and Recommendation Performance

The experimental results shown in Figure 4(b) demonstrate that as the budget of turns increases, the average performance of diagnosis and recommendation exhibits a distinct two-phase characteristic: in the initial phase (low-turn range), performance rises rapidly with additional turns, primarily due to the multi-turn dialogue mechanism enabling the LLM to progressively collect and refine patient information through iterative questioning; in the intermediate phase (medium-to-high turn range), the performance curve's slope noticeably flattens as the valuable information patients can provide gradually becomes saturated, making it difficult for the LLM to extract more meaningful information through additional questioning. Nevertheless, on the

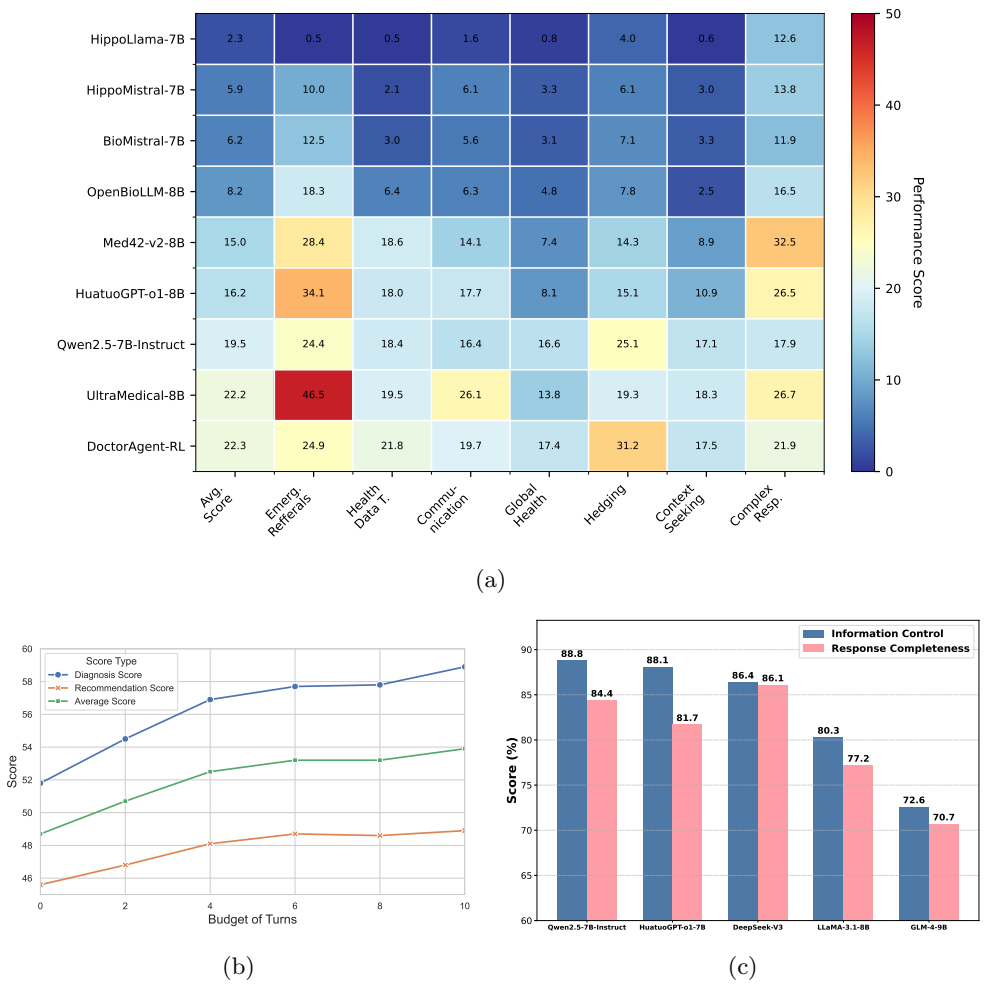


Fig. 4: (a) Performance comparison on HealthBench benchmark across eight evaluation dimensions. Partial experimental data sourced from [30]. (b) Comparative analysis of Diagnosis Score, Recommendation Score, and the resulting Average Score of DoctorAgent-RL on MTMedDialog at different levels of the Budget of Turns. (c) Performance comparison of different models in simulating patient agent on MTMedDialog.

whole, a larger budget of turns still leads to better performance, as additional interaction opportunities can capture potential subtle information differences. It is worth noting that diagnostic performance consistently outperforms recommendation performance, as diagnosis tasks can progressively confirm symptom characteristics through multi-turn interactions, while recommendation tasks rely more on established medical knowledge bases, leaving relatively limited room for performance improvement.

2.8 Patient Agent Behavior Analysis on MTMedDialog

Figure 4 (c) presents the evaluation results of the effectiveness of the patient agent in simulating real-world medical interactions. The assessment employs a scoring system based on DeepSeek-V3, evaluating two key metrics: (1) Information Control, which quantifies unsolicited information disclosure with a baseline score of 1.0 (indicating perfect compliance) and deducts 0.2 points per unauthorized disclosure instance; and (2) Response Completeness, which evaluates critical information omission against doctor queries using an initial score of 1.0, penalized by 0.2 points per essential detail omission. Both metrics are ultimately converted to a percentage-based rating. In the dimension of information control, Qwen2.5-7B demonstrates optimal compliance performance, precisely constraining output content and effectively avoiding the leakage of irrelevant information. In terms of response completeness, while DeepSeek-V3 performs best, Qwen2.5-7B still maintains near-optimal levels of key information density required for doctor-patient dialogue. Considering both model performance and cost-effectiveness in training, we ultimately selected Qwen2.5-7B as the implementation solution for the patient agent. For a more detailed description of the dialogue flow and rating prompts, please refer to Supplementary Material.

3 Discussion

This study successfully designed and validated a multi-agent reinforcement learning framework named DoctorAgent-RL, aimed at addressing the core challenges faced by current large language models in real-world, multi-turn clinical consultations. Our central contribution lies in achieving a fundamental paradigm shift: reframing the role of medical AI from a passive knowledge reproducer to an active knowledge constructor. By modeling clinical diagnosis as a dynamic decision-making process and training the agent to master a "questioning methodology" for optimal diagnosis, our model has demonstrated exceptional performance across multiple evaluations.

Our most significant findings come from a multi-faceted and rigorous evaluation system. In both automated evaluations based on our custom-built MTMedDialog dataset and in blinded human assessments, DoctorAgent-RL significantly outperformed all baselines, including frontier closed-source and domain-specific models, on key metrics such as diagnostic accuracy, recommendation rationality, and question quality. Critically, in real-time interaction tests with 20 real patients, our agent achieved a 70% exact diagnostic match rate. This result provides powerful evidence that our framework is not only effective in simulated environments but, more importantly, possesses the robust practical capability to stably and efficiently complete clinical dialogue tasks in unpredictable real-world scenarios, successfully bridging the critical sim-to-real gap.

Compared to existing work, the advantages of DoctorAgent-RL are evident. Traditional supervised fine-tuning models, such as BioMistral [10], exhibit poor performance despite a high number of interaction turns due to their inability to comprehend diagnostic logic, confirming the limitations of superficial imitation [16]. In contrast to other frontier multi-agent or reinforcement learning systems (e.g., AMIE, HuatuoGPT-01-7B) [20, 26], our two-stage training paradigm (SFT+RL) and the MTMedDialog

dataset, designed specifically for dynamic interaction, enable the model to learn a transferable clinical reasoning framework rather than merely "overfitting" to specific disease patterns. Our qualitative analysis also indicates that DoctorAgent-RL successfully preserves the general-purpose conversational abilities of its base model while acquiring deep domain expertise, thus avoiding the "catastrophic forgetting" common in many domain-specific models.

Despite these encouraging results, we also recognize the limitations of this study. First, although the MTMedDialog dataset is innovative in its support for dynamic interaction, it currently covers only eight major disease categories. This may limit the model's generalizability across a broader spectrum of diseases and different cultural contexts. Second, our real-patient validation study, while a crucial step in verifying the model's practical capabilities, has a relatively small sample size (20 patients). Larger and more diverse clinical trials will be necessary to further confirm its safety and efficacy. Furthermore, while our high-fidelity patient agent can effectively simulate dialogues, it cannot fully replicate the complex psychological states, emotional variations, and descriptive inconsistencies of real human patients. Lastly, the current framework does not yet integrate multimodal information (e.g., medical imaging), which is essential for real-world diagnosis.

Looking forward, our work opens up several research directions for developing next-generation clinical decision support systems. First, efforts can be dedicated to expanding and enriching the MTMedDialog dataset to include more disease types, languages, and cultural backgrounds. Second, conducting larger-scale, prospective clinical trials is a necessary step for deploying this model in real clinical settings. Future research should also explore the integration of multimodal data inputs into the model's decision-making loop. Additionally, enabling the model to explain its thought process to clinicians to enhance transparency and trustworthiness is another important topic for in-depth investigation.

In summary, DoctorAgent-RL not only technically surpasses existing medical dialogue systems but, more importantly, provides a clear and validated path for building a medical AI capable of dynamic, strategic information gathering. Our findings suggest that by positioning AI as a collaborative tool to augment, rather than replace, physicians' capabilities, we can hope to develop intelligent systems that can genuinely alleviate the pressure on global healthcare resources, reduce the risk of misdiagnosis, and ultimately improve the quality of patient care.

4 Methods

4.1 Dataset creation process

To facilitate a high-fidelity simulation of the real diagnostic process and provide an ideal environment for agents requiring dynamic, interactive training, we present MTMedDialog. It is a novel benchmark dataset specifically designed for training and evaluating AI agents in a multi-turn medical consultation setting. A key distinction from previous datasets, which consist of static scripts, is that each dialogue in MTMedDialog is grounded in a comprehensive, hidden patient profile. This structure enables a patient simulator to dynamically reveal relevant symptoms and medical history in

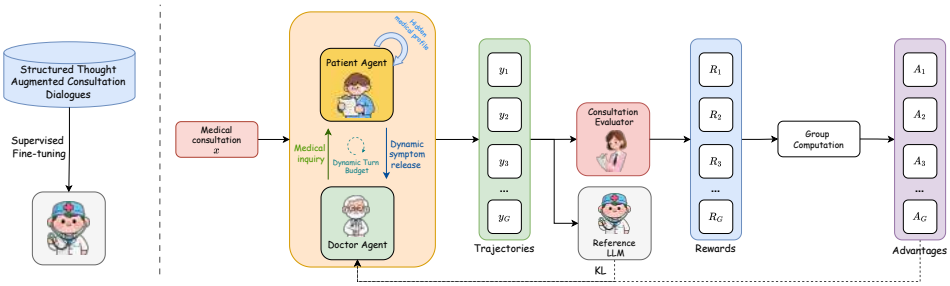


Fig. 5: The multi-agent collaborative reinforcement learning framework for DoctorAgent-RL. During the rollout stage, multi-turn interactions are conducted between the doctor agent and the patient agent.

response to a doctor agent's questions, rather than following a fixed script. This design is crucial for the development and evaluation of agents with advanced, adaptive inquiry strategies, allowing them to learn and evolve through realistic interaction. The dataset comprises 8,086 training samples and 2,082 test samples derived from three Chinese benchmark datasets: IMCS21 [31], CHIP-MDCFNPC [32], and MedDG [33]. The test set covers 8 major disease categories, with detailed distribution statistics provided in Supplementary Material.

As the source datasets were collected from real doctor-patient conversations, we implemented a two-stage denoising strategy: (1) filtering shallow dialogues with less than three turns through exact turn-count matching, and (2) removing noisy segments containing consecutive meaningless responses using DeepSeek-V3 [34]. We strictly preserved the original data partitioning protocol during this cleaning process to ensure evaluation reliability. The retained dialogues were then translated into English using DeepSeek-V3. Finally, we performed a manual inspection and refinement of the test set.

The test set serves dual evaluation purposes: 1) assessing doctor agent's questioning and diagnostic capabilities using complete dialogue trajectories with gold-standard diagnosis labels; 2) evaluating patient agent's response quality through a randomly selected subset of 500 samples.

4.2 Task Formulation

We model the multi-round medical consultation process as a multi-agent collaborative RL system, where the doctor agent serves as the main agent for strategy optimization, the patient agent acts as a collaborative counterpart to form a dynamic game relationship with it, and the consultation evaluator functions as a neutral arbiter that coordinates doctor-patient interactions and guides the optimization process of the doctor agent's strategy through well-designed reward mechanisms.

4.2.1 Doctor Agent

As the decision-making doctor agent, its state space $s_t \in \mathcal{S}$ encompasses the dialogue history H_t, providing a comprehensive record of the consultation. The agent's actions are drawn from the action space $\mathcal{A} = \{a_{query}, a_{diagnose}\}$, which includes two distinct behaviors: generating medical inquiries and executing diagnostic decisions.

Environmental dynamics are governed by the patient agent through state transitions:

$$s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

Here, P represents the transition probability function, which quantifies the likelihood of moving to a new state s_{t+1} given the current state s_t and the action a_t taken by the doctor agent.

The consultation continues until either a predefined limit on conversation turns is reached or the doctor agent provides a final diagnosis, at which point a complete dialogue history H_T forms the consultation trajectory.

The reward signal $R \in \mathbb{R}$ for each trajectory is provided by an independent consultation evaluator (detailed in Section 4.2.3), considering diagnostic accuracy, information acquisition efficiency, and protocol compliance as metrics.

To enhance the stability of policy optimization and eliminate the requirement for an additional value function approximation, we propose Group Relative Policy Optimization (GRPO) [25] as our policy gradient algorithm. Unlike Proximal Policy Optimization (PPO) [35], GRPO employs the average reward of multiple sampled outputs as a baseline instead of relying on a learned value function. Specifically, for each patient's consultation x, GRPO samples a set of trajectories $y_1, y_2, \ldots, y_G$ through the interaction between the doctor agent π_D and the patient agent π_{p}. The doctor agent, as the policy model, is then optimized by maximizing the following objective function:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\substack{x\sim\mathcal{D} \\ \{y_i\}_{i=1}^{G}\sim\pi_{\mathrm{D_{old}}}(\cdot|x;\pi_{\mathrm{p}})}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{\sum_{t=1}^{|y_i|} I(y_{i,t})} \sum_{\substack{t=1 \\ I(y_{i,t})=1}}^{|y_i|} \right.$$
$$\left. \min\left(\frac{\pi_{\mathrm{D}}(y_{i,t}|x, y_{i,<t})}{\pi_{\mathrm{D_{old}}}(y_{i,t}|x, y_{i,<t})}\hat{A}_{i,t},\ \mathrm{clip}\left(\frac{\pi_{\mathrm{D}}(y_{i,t}|x, y_{i,<t})}{\pi_{\mathrm{D_{old}}}(y_{i,t}|x, y_{i,<t})}, 1-\epsilon, 1+\epsilon \right)\hat{A}_{i,t} \right) - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_{\mathrm{D}}||\pi_{\mathrm{D_{ref}}}\right] \right] \tag{1}$$

where ϵ and β are hyperparameters, and $\hat{A}_{i,t}$ represents the advantage, computed based on the relative rewards of outputs within each group. Here, $I(y_{i,t}) = 1$ indicates that the token $y_{i,t}$ is generated by π_{D}. Since doctors cannot predict patients' symptoms in advance, during training, the responses of the patient agent are masked.

4.2.2 Patient Agent

The patient agent is implemented using Qwen2.5-7B-Instruct [36] and is incorporated into a two-phase dialogue simulation framework via carefully crafted prompt engineering.

In the first phase, the system combines patient self-reports and multi-turn dialogue content to create case descriptions. Additionally, it augments potential symptom features using standard diagnostic results, thereby forming a more comprehensive hidden medical profile. This design effectively mitigates the symptom coverage issues stemming from incomplete doctor inquiries in traditional datasets.

In the second phase, the patient agent utilizes dynamic symptom release strategies in response to the real-time queries of the doctor agent. It maintains strict pathological consistency while mimicking the natural variability in patients' symptom description granularity and the order of their complaints. The detailed prompt designs for both phases are presented in Supplementary Material. By retaining complete hidden case data, the patient agent ensures that its natural language responses adhere to clinical standards and are generated solely based on the dialogue history.

4.2.3 Consultation Evaluator

In our RL framework, guiding the doctor agent to master essential clinical diagnostic skills is paramount. To achieve this, we've designed a sophisticated Consultation Evaluator, acting as a multi-faceted reward system that assesses the agent's performance across critical dimensions of a medical consultation. This evaluator comprises three core components, each contributing to a comprehensive assessment of the agent's diagnostic capabilities and consultation conduct.

Diagnostic Accuracy Reward. The first pillar of our Consultation Evaluator focuses on the agent's diagnostic precision and treatment recommendations. To ensure the reliability of this evaluation and prevent any potential reward hacking, we employ a rule-based reward mechanism. This mechanism meticulously calculates the word-level F1 score between the doctor agent's predicted diagnosis and the gold-standard diagnosis, as well as for the recommended treatments. The formulation for this reward is as follows:

$$R_{\text{accuracy}} = 5 \times {}^{\bullet}\text{F1}_{\text{diagnosis}} + \text{F1}_{\text{recommendation}}) \tag{2}$$

The coefficient 5 serves to adjust the relative weight of this crucial reward signal within the overall evaluation. This design ensures a balanced and robust assessment of both the diagnostic output and the quality of suggested interventions, providing a stable foundation for learning.

Information Acquisition Efficiency Reward. A proficient doctor knows how to ask the right questions efficiently. To foster this skill in our doctor agent, the Consultation Evaluator incorporates a dynamic reward mechanism that promotes valuable questioning while discouraging repetitive or unhelpful queries. This reward is directly tied to the patient agent's feedback after each dialogue turn and accumulates throughout the interaction:

$$R^t_{\text{information}} = \begin{cases} 1, & \text{if the patient agent answers normally} \\ -2, & \text{if the patient agent refuses to answer} \end{cases} \tag{3}$$

$$R_{\text{information}} = \sum_t R^t_{\text{information}} \tag{4}$$

Through this feedback loop, the model learns to optimize its questioning strategy, prioritizing the acquisition of diagnostically relevant information and refining its inquiry process.

Protocol Compliance Reward. Adherence to established clinical interview protocols is a hallmark of professional medical practice. To instill this discipline, our Consultation Evaluator includes a compliance reward. This component penalizes deviations from predefined norms and ensures that the agent completes the diagnostic process within specified limits:

$$R^t_{\text{compliance}} = \begin{cases} -2, & \text{Question Format Violation} \\ -5, & \text{No Diagnosis within Allowed Turns} \\ 0, & \text{otherwise (i.e., protocol is followed)} \end{cases} \tag{5}$$

$$R_{\text{compliance}} = \sum_t R^t_{\text{compliance}} \tag{6}$$

This mechanism reinforces the learning of structured interview processes and ensures the timely completion of a diagnosis, closely mirroring the practical constraints and expectations of real-world clinical environments.

By combining these three critical components, the total consultation evaluation score, or reward, received by the doctor agent at each time step t is calculated as:

$$R = R_{\text{accuracy}} + R_{\text{information}} + R_{\text{compliance}} \tag{7}$$

This sophisticated, multi-dimensional Consultation Evaluator not only guides the model toward superior diagnostic accuracy but also actively encourages the development of efficient information-gathering strategies and professional, compliant clinical behavior, ultimately aiming to achieve diagnostic capabilities that closely resemble those of expert human doctors.

4.3 Training Procedure

As illustrated in Figure 5, our training framework for the doctor agent is built upon Qwen2.5-7B-Instruct, following the DeepSeek-R1 training paradigm [37]. The approach employs a two-stage training pipeline, integrating SFT and RL to cultivate clinical reasoning capabilities. Detailed experimental settings are illustrated in Supplementary Material.

Specifically, we randomly select 1,000 multi-turn consultation dialogues from the training corpus. Each doctor's query in the sampled dialogues is augmented with structured thought processes using DeepSeek-V3. These include hypothesis generation, evidence evaluation, and differential diagnosis steps derived from context. The doctor agent is fine-tuned on this enriched dataset to activate core capabilities in question-asking, diagnostic reasoning, and recommendation generation.

Following SFT, we refine the agent's decision-making under interaction constraints using the policy optimization algorithm detailed in the Task Formulation section. To enhance robustness and mimic real-world clinical scenarios, we introduce a **Dynamic Turn Budget Training Strategy**. Each training episode is assigned a random dialogue turn budget (2–10 turns). The model is explicitly reminded of the remaining turns after each interaction step, encouraging efficient information gathering. This two-stage approach ensures the agent first internalizes clinical reasoning patterns via supervised learning, then refines its strategy through interactive reward optimization. Ablation studies comparing alternative training strategies (direct SFT, direct RL, fixed-turn training) are presented in the experimental section.

5 Data Availability

The codes and datasets are available for access at https://github.com/JarvisUSTC/DoctorAgent-RL.git.

6 Competing Interests

The authors declare no competing interests.

7 Funding

This work was supported in part by funds from the National Key R&D Program (No. 2022YFF1202101, 2023YFF1204701); the CAS Research Fund (No. XDB38050200); the Self-supporting Program of Guangzhou National Laboratory (No. SRPG22007); the Startup Program of Guangzhou National Laboratory (No. YW-YFYJ0101).

8 Author contributions

Conceptualization, Y.F. and J.W.; methodology, Y.F. and J.W.; formal analysis, Y.F.; investigation, Y.F. and J.W.; resources, Y.L.; data curation, Y.F. and J.W.; writing—original draft preparation, Y.F. and J.W.; writing—review and editing, Y.F., J.W., L.Z, Y.Z, Z.L. and Y.L.; visualization, Y.F.; supervision, Y.L. and Z.L.; All authors have read and agreed to the published version of the manuscript.

9 Additional information

9.1 Supplementary information

Supplementary information is available at Supplementary material.pdf.

9.2 Corresponding authors

Correspondence to Zhen Lei, Yixue Li.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[2] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[3] GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al.: Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793 (2024)

[4] Kopka, M., Kalckreuth, N., Feufel, M.A.: Accuracy of online symptom assessment applications, large language models, and laypeople for self–triage decisions. npj Digital Medicine **8**(1), 178 (2025)

[5] Agyeman-Manu, K., Ghebreyesus, T.A., Maait, M., Rafila, A., Tom, L., Lima, N.T., Wangmo, D.: Prioritising the health and care workforce shortage: protect, invest, together. The Lancet Global Health **11**(8), 1162–1164 (2023)

[6] ALQahtani, D.A., Rotgans, J.I., Mamede, S., ALAlwan, I., Magzoub, M.E.M., Altayeb, F.M., Mohamedani, M.A., Schmidt, H.G.: Does time pressure have a negative effect on diagnostic accuracy? Academic Medicine **91**(5), 710–716 (2016)

[7] Arndt, B.G., Micek, M.A., Rule, A., Shafer, C.M., Baltus, J.J., Sinsky, C.A.: More tethered to the ehr: Ehr workload trends among academic primary care physicians, 2019-2023. The Annals of Family Medicine **22**(1), 12–18 (2024)

[8] Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A.V., Al Muhanna, D., Al-Muhanna, F.A.: A review of the role of artificial intelligence in healthcare. Journal of personalized medicine **13**(6), 951 (2023)

[9] Han, T., Adams, L.C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressem, K.K.: Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247 (2023)

[10] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373 (2024)

[11] Wu, C., Qiu, P., Liu, J., Gu, H., Li, N., Zhang, Y., Wang, Y., Xie, W.: Towards evaluating and building versatile large language models for medicine. npj Digital Medicine **8**(1), 58 (2025)

[12] Laban, P., Hayashi, H., Zhou, Y., Neville, J.: Llms get lost in multi-turn conversation. arXiv preprint arXiv:2505.06120 (2025)

[13] Auerbach, A.D., Lee, T.M., Hubbard, C.C., Ranji, S.R., Raffel, K., Valdes, G., Boscardin, J., Dalal, A.K., Harris, A., Flynn, E., *et al.*: Diagnostic errors in hospitalized adults who died or were transferred to intensive care. JAMA Internal Medicine **184**(2), 164–173 (2024)

[14] Chen, Y., Wang, Z., Xing, X., Xu, Z., Fang, K., Wang, J., Li, S., Wu, J., Liu, Q., Xu, X., et al.: Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. arXiv preprint arXiv:2310.15896 (2023)

[15] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., Dolan, B.: Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536 (2019)

[16] Liu, R., Xue, K., Zhang, X., Zhang, S.: Interactive evaluation for medical llms via task-oriented dialogue system. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 4871–4896 (2025)

[17] Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., Liu, T.: Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975 (2023)

[18] Zhang, K., Zeng, S., Hua, E., Ding, N., Chen, Z.-R., Ma, Z., Li, H., Cui, G., Qi, B., Zhu, X., *et al.*: Ultramedical: Building specialized generalists in biomedicine. Advances in Neural Information Processing Systems **37**, 26045–26081 (2024)

[19] Feng, Y., Zhou, L., Ma, C., Zheng, Y., He, R., Li, Y.: Knowledge graph–based thought: a knowledge graph–enhanced llm framework for pan-cancer question answering. GigaScience **14**, 082 (2025)

[20] Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Cheng, Y., et al.: Towards conversational diagnostic artificial intelligence. Nature, 1–9 (2025)

[21] Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.-Q., Ma, W., et al.: Agent hospital: A simulacrum of hospital with evolvable medical agents. arXiv preprint arXiv:2405.02957 (2024)

[22] Fan, Z., Tang, J., Chen, W., Wang, S., Wei, Z., Xi, J., Huang, F., Zhou, J.: Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. arXiv preprint arXiv:2402.09742 (2024)

[23] Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. arXiv preprint arXiv:2502.19634 (2025)

[24] Lai, Y., Zhong, J., Li, M., Zhao, S., Yang, X.: Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. arXiv preprint arXiv:2503.13939 (2025)

[25] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)

[26] Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., Wang, B.: Huatuogpt-o1, towards medical complex reasoning with llms. arXiv preprint arXiv:2412.18925 (2024)

[27] Zou, X., He, W., Huang, Y., Ouyang, Y., Zhang, Z., Wu, Y., Wu, Y., Feng, L., Wu, S., Yang, M., *et al.*: Ai-driven diagnostic assistance in medical inquiry: Reinforcement learning algorithm development and validation. Journal of Medical Internet Research **26**, 54616 (2024)

[28] Sun, Z., Liu, Z., Luo, C., Chu, J., Huang, Z.: Improving interactive diagnostic ability of a large language model agent through clinical experience learning. arXiv preprint arXiv:2503.16463 (2025)

[29] Arora, R.K., Wei, J., Hicks, R.S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., et al.: Healthbench: Evaluating large language models towards improved human health. arXiv preprint arXiv:2505.08775 (2025)

[30] Lai, Y., Liu, K., Wang, Z., Ma, W., Liu, Y.: Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. arXiv preprint arXiv:2510.04284 (2025)

[31] Chen, W., Li, Z., Fang, H., Yao, Q., Zhong, C., Hao, J., Zhang, Q., Huang, X., Peng, J., Wei, Z.: A Benchmark for Automatic Medical Consultation System: Frameworks, Tasks and Datasets. Bioinformatics (2022)

[32] Zhang, N., Chen, M., Bi, Z., Liang, X., Li, L., Shang, X., Yin, K., Tan, C., Xu, J., Huang, F., et al.: Cblue: A chinese biomedical language understanding evaluation benchmark. arXiv preprint arXiv:2106.08087 (2021)

[33] Liu, W., Tang, J., Cheng, Y., Li, W., Zheng, Y., Liang, X.: Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In: CCF International Conference on Natural Language Processing and Chinese Computing, pp. 447–459 (2022). Springer

[34] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)

[35] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[36] Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)

[37] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名： 胡明

方向： 人工智能

单位： 中国科学院西安光学精密机械研究所

题目： 半监督医学影像分割

This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

β-FFT: Nonlinear Interpolation and Differentiated Training Strategies for Semi-Supervised Medical Image Segmentation

Ming Hu[1,2] Jianfu Yin[1,2] Zhuangzhuang Ma[3*] Jianheng Ma[3*] Feiyu Zhu[1,2]
Bingbing Wu[1,2] Ya Wen[4] Meng Wu[5] Cong Hu[5,6†] Bingliang Hu[1†] Quan Wang[1†]
[1] Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences [3] Xidian University [4] Xi'an University of Technology
[5] Zhongnan Hospital of Wuhan University [6] The First Affiliated Hospital of Guangxi Medical University

Abstract

Co-training has achieved significant success in the field of semi-supervised learning(SSL); however, the homogenization phenomenon, which arises from multiple models tending towards similar decision boundaries, remains inadequately addressed. To tackle this issue, we propose a novel algorithm called β-FFT from the perspectives of data processing and training structure. In data processing, we apply diverse augmentations to input data and feed them into two sub-networks. To balance the training instability caused by different augmentations during consistency learning, we introduce a nonlinear interpolation technique based on the Fast Fourier Transform (FFT). By swapping low-frequency components between variously augmented images, this method not only generates smooth and diverse training samples that bridge different augmentations but also enhances the model's generalization capability while maintaining consistency learning stability. In training structure, we devise a differentiated training strategy to mitigate homogenization in co-training. Specifically, we use labeled data for additional training of one model within the co-training framework, while for unlabeled data, we employ linear interpolation based on the Beta(β) distribution as a regularization technique in additional training. This approach allows for more efficient utilization of limited labeled data and simultaneously improves the model's performance on unlabeled data, optimizing overall system performance. Code is available at the following link. https://github.com/Xi-Mu-Yu/beta-FFT.

1. Introduction

As manually annotating medical images such as CT, MRI, and pathology images is both costly and labor-intensive, this process becomes increasingly challenging for radiologists and other medical professionals as data volumes continue to grow, leading to scalability issues. Consequently, semi-supervised semantic segmentation has become particularly important in medical image analysis.

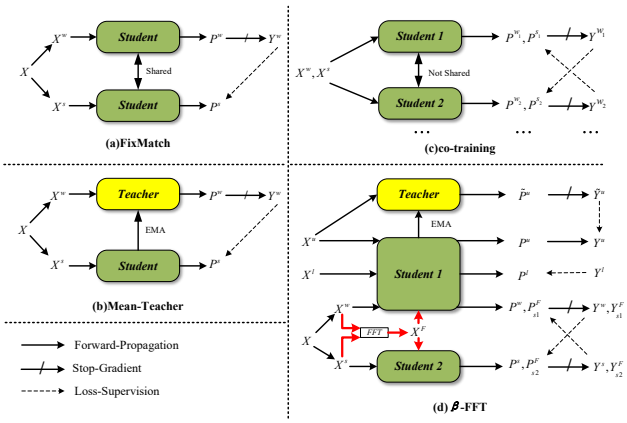


Figure 1. Illustrating the architectures for (a) FixMatch, (b) Mean-Teacher , (c) co-training , and (d) our approach β-FFT. X represents the dataset, P is the confidence map, Y denotes the labels encoded in one-hot format. More details in the method section.

The research on semi-supervised learning(SSL) began with the self-training method [11, 30], which enhances model learning by leveraging unlabeled data. Initially, researchers aimed to achieve this by self-generating labels, but this often led to unreliable results when handling unlabeled data[1]. To address this issue, researchers introduced consistency regularization methods[24, 31], usually applied by enforcing perturbations on unlabeled data in an online fashion, helping the model maintain stability when facing data variations.

As research progressed, data augmentation's role in bolstering model robustness has become evident. The **Fix-**

*Equal contribution
†Corresponding author.

Match method[32] optimizes semi-supervised learning by leveraging pseudo-labeling and consistency regularization on augmented unlabeled data (Fig.1a). Another notable approach in medical image semi-supervised segmentation is the **Mean Teacher** architecture[33], which employs an exponential moving average (EMA) of the student model's weights for the teacher network(Fig.1b). Inspired by this framework, various methods have been developed to enhance semi-supervised segmentation. For example, the UA-MT framework[45] utilizes uncertainty information to guide the student model in learning reliable targets. Verma et al.[34] introduced an **Interpolation Consistency Training Framework** to ensure that the predictions of interpolated unlabeled points align with actual data points. URPC[20] guarantees consistent predictions across scales. The method **BCP**[3] further enhances the Mean Teacher architecture by bi-directionally copying labeled and unlabeled data, allowing the unlabeled data to learn shared semantics from the labeled data and addressing the empirical mismatch in semi-supervised medical image segmentation.

While these approaches significantly enhance semi-supervised learning, they also pose challenges. The close coupling between teacher and student models[13] can hinder effective knowledge transfer, limiting the teacher's capacity to convey valuable insights. This coupling may lead to **confirmation bias**, causing models to excessively rely on existing biases and overlook potential new information. To overcome the limitations associated with the coupling in teacher-student models, researchers introduced the **co-training** method[13, 25]. Co-training leverages the complementary characteristics of multiple models to enhance knowledge sharing and transfer, enabling each model to gain additional knowledge from others, thus effectively improving the performance of SSL (Fig.1c).

Although co-training has achieved significant success in SSL, the risk of **model homogenization** remains a critical challenge[16]. Specifically, multiple models tend to converge to similar decision boundaries[1, 16], leading to homogenization. This phenomenon reduces the diversity of learned representations, thereby limiting the models' generalization capacity in semi-supervised settings. **This raises a crucial question: Can we introduce additional information or corrective mechanisms to maintain model accuracy while reducing homogenization between models in co-training, thereby enhancing overall performance?**

To mitigate the issue of model homogenization in co-training, we introduced improvements from two aspects: data processing and network training strategies.

Firstly, in terms of data processing, we incorporated diverse data augmentation techniques. By generating augmented samples with different intensities (including strong and weak augmentations) and training different subnetworks separately, we achieved model differentiation at the data level. Meanwhile, to prevent training instability caused by varying augmentation strengths, we employed the Fast Fourier Transform (FFT) to exchange low-frequency information between strongly and weakly augmented images, thereby creating new samples that lie between the two. This approach not only provides the model with new data perspectives but also effectively alleviates the instability caused by strong and weak augmentations.

Secondly, in terms of training strategies, we applied an additional training step to one of the subnetworks, while the other followed the original training process. We utilized labeled data for extra training on the selected model and introduced unlabeled data, generated through linear interpolation based on Beta distribution sampling, as a regularization term in the training process. This strategy not only helps maintain diversity in decision boundaries between the two models and reduces their homogenization but also further enhances the effectiveness of collaborative learning.We performed additional supervised training on Student Model 1 using labeled data, which not only enhances the performance of Student Model 1 but also helps to increase its independence. This approach aids in mitigating the confirmation bias between Student Model 1 and the Teacher network. To summarize, we make the following contributions:

1. **Nonlinear Interpolation Strategy**: By using the Fast Fourier Transform to exchange low-frequency information between weakly and strongly augmented images, we effectively mitigated the training instability caused by using different augmented data to address model homogenization. At the same time, this approach enriched the diversity of data samples and improved the model's generalization ability.
2. **Dehomogenization during training**: Student model 1 undergoes additional training, while Student model 2 does not. This differentiated training approach generates a unique loss for each model, effectively reducing homogenization between the two student models and preserving the diversity of their decision boundaries.

2. Related Work

2.1. Semi-Supervised Medical Image Segmentation

Previous methods can be broadly categorized into self-training methods[6, 49] and consistency regularization methods[8, 13]. Self-training algorithms are considered the fundamental prototype of pseudo-labeling methods[28], where a model is pre-trained on a labeled dataset and iteratively retrained or fine-tuned using predictions from unlabeled data. [23] matched these pseudo-labels by synthesizing new images rather than optimizing them. Within the framework of consistency regularization, [12] employed strong augmentation and weak augmentation to handle unlabeled data. Some studies have explored consistent data

transformations, such as patch shuffling data transformation [15], cut-and-paste augmentation [44], and copy-paste [3]. ABD[9] effectively integrates multiple perturbations through an adaptive bidirectional displacement mechanism, enhancing the quality of consistency learning. AD-MT[48] reduces confirmation bias and enhances model performance under limited labeled data by employing random periodic alternation and a counteracting disturbance module.

2.2. Frequency Domain Enhancement Techniques

Fourier domain processing techniques enhance model generalization, robustness, and adaptability in computer vision tasks, particularly for domain adaptation and data augmentation. Fourier Domain Adaptation (FDA) [43] replaces the low-frequency amplitude spectrum of source images with that of target images, enabling model adaptation to new domains while preserving structural information. The Fourier-based Domain Generalization Framework [41] systematically investigates the roles of amplitude and phase spectra in domain shifts, revealing that the amplitude spectrum captures domain-specific style information, while the phase spectrum retains structural content. This insight underpins frequency-based augmentation strategies for improved generalization. FreMix [40] performs frequency-based augmentation by mixing amplitude spectra of different images, thereby enhancing domain generalization.

2.3. Research Status of Homogenization

In the co-training framework of semi-supervised learning, the issue of model homogenization has become a core challenge that restricts performance improvement. To address this problem, researchers have proposed systematic solutions from three levels: data augmentation, model architecture, and training strategies. On the level of data augmentation, consistency training based on strong-weak augmentation combinations[9, 16, 27, 32] generates multi-view samples through differentiated disturbances, while the contrastive learning framework[14] further utilizes graph structures to dynamically allocate samples, enhancing complementarity among models. Recent work [26] also minimizes mutual information to constrain the independence of view features, reducing redundancy. On the model architecture level, heterogeneous network designs (such as combinations of CNN and Transformer[16, 22]) and model parameter diversification[10] are used to force models to focus on different feature patterns. On the training strategy level, dynamic optimization methods[39], asymmetric learning mechanisms (such as alternating training[48]) have been proven to effectively prevent model convergence.

3. Method

In semi-supervised segmentation, we aim to train a model using both labeled and unlabeled data. The labeled dataset $\mathcal{D}^l = \left\{(X_i^l, Y_i^l)\right\}_{i=1}^N$ contains N labeled images, where X_i^l is the image and Y_i^l is its corresponding segmentation label. The unlabeled dataset $\mathcal{D}^u = \left\{X_j^u\right\}_{j=1}^M$ consists of M unlabeled images, where X_j^u has no associated label. Typically, $N \ll M$, meaning the number of labeled images is much smaller than the unlabeled ones.

In our approach, we employ a single **teacher model** alongside two **student models**. The parameters of the teacher model are updated using an Exponential Moving Average (EMA) mechanism, specifically tuned based on the parameters of **Student Model 1**. The update process for the teacher model at each iteration can be expressed as:

$$\theta_T^{(t)} = \lambda\theta_T^{(t-1)} + (1-\lambda)\theta_{S1}^{(t)} \tag{1}$$

Here, $\theta_T^{(t)}$ denotes the parameters of the teacher model at the t-th iteration, $\theta_{S1}^{(t)}$ represents the parameters of Student Model 1 at the same iteration, and $\lambda \in [0, 1]$ serves as a smoothing factor that balances the influence of previous teacher parameters against those of Student Model 1.

3.1. Overview

1. **Data Augmentation and Teacher Network Training via Copy-Paste** We apply both weak and strong data augmentations to the data, enhancing its diversity through simple transformations and advanced techniques. Additionally, we utilize a Copy-Paste method to train a teacher network, ensuring high-quality pseudo-labels for the unlabeled data. This approach effectively improves the accuracy of the generated pseudo-labels.
2. **Nonlinear Interpolation:** To reduce homogenization in collaborative training and increase data diversity, we input data with varying degrees of enhancement (strong and weak) into different sub-models. We also exchange low-frequency components between weakly and strongly enhanced images to reduce the instability of collaborative training.
3. **Differentiated Training:** One model in the co-training framework undergoes additional training, while the other does not, reducing homogeneity between them.

3.2. Data Augmentation and Teacher Network Training via Copy-Paste

We begin by applying both strong(s) and weak(w) augmentations to the data. Weak augmentations include simple transformations such as rotation and flipping, while strong augmentations build on these with techniques like Cutout [7] and ColorJitter[42]. Inspired by BCP [3], we employ a Copy-Paste technique for further data enhancement. Specifically, the Copy-Paste process can be expressed as follows:

$$X_{w/s}^{in} = M \odot X_{w/s}^l + (1-M) \odot X_{w/s}^u, \tag{2}$$

$$X_{w/s}^{out} = M \odot X_{w/s}^u + (1-M) \odot X_{w/s}^l, \tag{3}$$

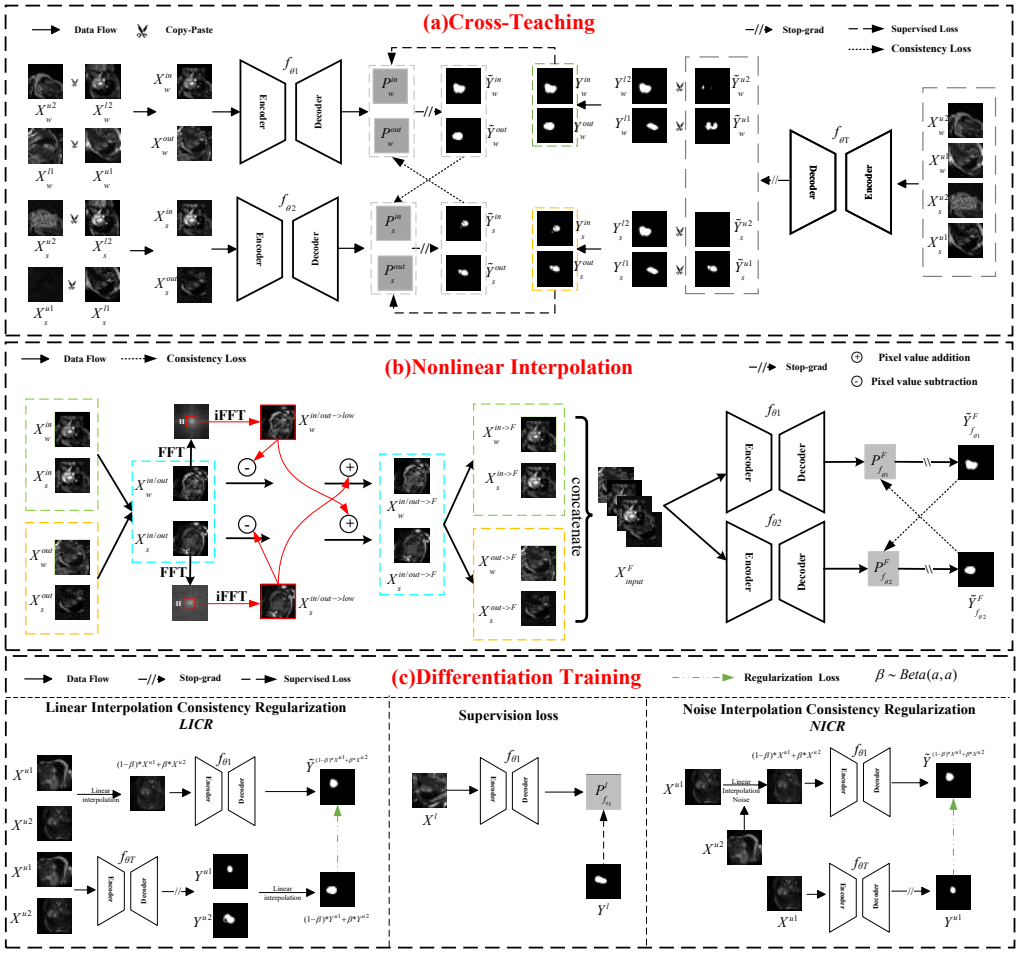


Figure 2. Overview of our β-FFT framework.In the figure, X represents the data, l represents labeled data, u represents unlabeled data, w denotes weak augmentation, and s denotes strong augmentation. P represents the confidence map obtained from the data through the model, $\tilde{Y}$ represents the class prediction map corresponding to the confidence map obtained through the model, and Y represents the corresponding ground truth labels.FFT represents the Fast Fourier Transform, and iFFT represents the Inverse Fast Fourier Transform.

Here, $X^l_{w/s}$ represents labeled data, and $X^u_{w/s}$ represents unlabeled data. We utilize a mask $M \in \{0,1\}^{W\times H}$ to perform the bidirectional copy-pasting operation, controlling the blending between the images. The mask defines a zero-value region of size $\eta H \times \eta W$, where $\eta \in (0,1)$ governs the proportion of the foreground region.

To ensure high-quality predictions for the unlabeled data, we pre-train a teacher model f_{θ_T} using two different labeled data, l_1 and l_2, as follows:

$$X^{Teacher}_{w/s} = M \odot X^{l_1}_{w/s} + (1-M) \odot X^{l_2}_{w/s}, \quad (4)$$

$$Y^{Teacher}_{w/s} = M \odot Y^{l_1}_{w/s} + (1-M) \odot Y^{l_2}_{w/s}. \quad (5)$$

Thus, the labels corresponding to the unlabeled data $X^u_{w/s}$ are given by:

$$Y^u_{w/s} = \underset{c\in C}{argmax}(f_{\theta_T}(X^u_{w/s}), c). \quad (6)$$

Accordingly, the labels for $X^{in}_{w/s}$ and $X^{out}_{w/s}$ can be expressed as:

$$Y^{in}_{w/s} = M \odot Y^l_{w/s} + (1-M) \odot Y^u_{w/s}, \quad (7)$$

$$Y^{out}_{w/s} = M \odot Y^u_{w/s} + (1-M) \odot Y^l_{w/s}. \quad (8)$$

3.3. Non-linear Interpolation

We utilize a non-linear interpolation technique based on the exchange of low-frequency components to enhance the diversity of data samples, particularly for image augmentation. Our approach begins by decomposing an image I into its frequency components using the FFT:

$$I = F^{-1}(F(I)) \quad (9)$$

where F denotes the FFT and F^{-1} represents its inverse, the Inverse Fast Fourier Transform (iFFT). Subsequently,

we extract the low-frequency component $I^{\rightarrow \text{low}}$ of the image, defined as:

$$I^{\rightarrow \text{low}} = F^{-1}(F(I) \cdot H) \tag{10}$$

H is a low-pass filter employed to isolate low-frequency information.We conducted experimental investigations on the setting of H in the experimental section.

To enhance the diversity of the augmented images, we then perform non-linear interpolation by swapping the low-frequency components of the weakly augmented image I_w with those of the strongly augmented image I_s. This generates two new images $I_w^{->F}$ and $I_s^{->F}$:

$$I_{w/s}^{->F} = I_{w/s} - I_{w/s}^{->low} + I_{s/w}^{->low} \tag{11}$$

Next, to further increase the diversity of the augmented images, we perform non-linear interpolation by swapping the low-frequency components of the weakly augmented images X_w^{in} and X_s^{in} with those of the strongly augmented images X_w^{out} and X_s^{out}. This process generates four new images, defined as:

$$X_{w/s}^{\text{in/out}\rightarrow F} = X_{w/s}^{\text{in/out}} - X_{w/s}^{\text{in/out}\rightarrow \text{low}} + X_{s/w}^{\text{in/out}\rightarrow \text{low}} \tag{12}$$

We aim for different models to produce similar outputs across various samples to ensure they learn consistent feature representations. This consistency enables the models to maintain strong performance on unseen samples, thereby enhancing their generalization ability. To strengthen this effect, we concatenate the four images generated through non-linear interpolation into a new set X_{input}^F:

$$X_{\text{input}}^F = \text{Concat}\left[X_w^{\text{in}\rightarrow F}, X_s^{\text{in}\rightarrow F}, X_w^{\text{out}\rightarrow F}, X_s^{\text{out}\rightarrow F}\right] \tag{13}$$

3.4. Differentiated Training of Student Models

Student Model 1 is additionally trained using **labeled data**, whereas **Student Model 2** does not undergo this supplementary training. To enhance the robustness of Student Model 1, we incorporate two regularization techniques: linear interpolation consistency regularization and noise interpolation consistency regularization.

3.4.1. Linear Interpolation Consistency Regularization

Student Model 1 employs a pixel-wise data perturbation strategy along with consistency regularization that leverages unlabeled data. Given two unlabeled data points, X^{u_1} and X^{u_2}, we generate an interpolated data point $M_\beta(X^{u_1}, X^{u_2})$, defined as follows:

$$M_\beta(X^{u_1}, X^{u_2}) = \beta X^{u_1} + (1-\beta)X^{u_2} \tag{14}$$

In this equation, the hyperparameter β is sampled from a **Beta distribution**,following the setup in Mixup[46]. We apply the linear interpolation consistency regularization, which compares the output at the interpolated data point $f_{\theta_1}(M_\beta(X^{u_1}, X^{u_2}))$ with the outputs of the original data points:

$$M_\beta(f_{\theta_1}(X^{u_1}), f_{\theta_1}(X^{u_2})) \approx f_{\theta_1}(M_\beta(X^{u_1}, X^{u_2})) \tag{15}$$

3.4.2. Noise Interpolation Consistency Regularization

We also introduce a noise interpolation consistency constraint by reformulating equation 14 as follows:

$$\begin{aligned} M_\beta(X^{u_1}, X^{u_2}) &= \beta X^{u_1} + (1-\beta)X^{u_2} \\ &= X^{u_1} + (1-\beta)\cdot(X^{u_2} - X^{u_1}) \end{aligned} \tag{16}$$

In this formulation, we interpret $(1-\beta)\cdot(X^{u_2} - X^{u_1})$ as noise interference. The noise consistency loss can then be expressed as:

$$M_\beta(f_{\theta_1}(X^{u_1}), f_{\theta_1}(X^{u_2})) \approx f_{\theta_1}(M_\beta(X^{u_1})) \tag{17}$$

This noise consistency constraint ensures that Student Model 1 produces outputs that closely resemble those of the teacher model, even when the input data is slightly perturbed. By assessing the difference between the output at the interpolated point $M_\beta(X^{u_1}, X^{u_2})$ and the teacher model's output at X^{u_1}, this loss term encourages Student Model 1 to utilize information from more than just a single data point X^{u_1}. This approach effectively enhances the model's generalization ability.

In contrast, **Student Model 2** does not leverage labeled data for training. It employs cross pseudo-supervision and applies cross-consistency loss derived from the nonlinear interpolation process. Unlike Student Model 1, which utilizes both labeled and unlabeled data for consistency regularization, Student Model 2 focuses on leveraging shared learning signals between the two student models. This results in markedly different training trajectories for Student Model 1 and Student Model 2, promoting diversity in learning signals throughout the overall learning process.

Since Linear Interpolation Consistency Regularization and Noise Interpolation Consistency Regularization serve as two distinct additional training processes for labeled data, we refer to the training process with linear interpolation as **LICR**, and the one with noise interpolation as **NICR**.

4. Loss Functions

The overall loss function is comprised of three main components: **Cross Teaching Loss**, **Nonlinear Interpolation Loss**, and **Differentiation Loss**.The notation $\mathcal{L}_{\text{ce}}$ represents the Cross Entropy Loss, and $\mathcal{L}_{\text{dice}}$ represents the Dice Loss.

4.1. Cross Teaching Loss

Cross Teaching Loss leverages pseudo-labels from one model to supervise the other. It consists of a supervised loss and a cross pseudo-supervised loss. The supervised

loss ensures effective learning from ground truth labels by combining cross-entropy and Dice losses on both strongly and weakly augmented inputs:

$$\begin{aligned}\mathcal{L}_{sup}^{aug} = &\frac{1}{2}\left(\mathcal{L}_{ce,dice}(f_{\theta_1}(X_w^{in}), Y_w^{in}) + \mathcal{L}_{ce,dice}(f_{\theta_1}(X_w^{out}), Y_w^{out})\right) \\ &+\frac{1}{2}\left(\mathcal{L}_{ce,dice}(f_{\theta_2}(X_s^{in}), Y_s^{in}) + \mathcal{L}_{ce,dice}(f_{\theta_2}(X_s^{out}), Y_s^{out})\right)\end{aligned} \tag{18}$$

To enforce consistency, the cross pseudo-supervised loss aligns predictions with pseudo-labels from the other model:

$$\begin{aligned}\mathcal{L}_{cps}^{aug} = &\mathcal{L}_{dice}\left(f_{\theta_1}(X_w^{in}), \underset{c\in C}{argmax}(f_{\theta_2}(X_s^{in}), c)\right) \\ &+ \mathcal{L}_{dice}\left(f_{\theta_1}(X_w^{out}), \underset{c\in C}{argmax}(f_{\theta_2}(X_s^{out}), c)\right) \\ &+ \mathcal{L}_{dice}\left(f_{\theta_2}(X_s^{in}), \underset{c\in C}{argmax}(f_{\theta_1}(X_w^{in}), c)\right) \\ &+ \mathcal{L}_{dice}\left(f_{\theta_2}(X_s^{out}), \underset{c\in C}{argmax}(f_{\theta_1}(X_w^{out}), c)\right)\end{aligned} \tag{19}$$

4.2. Nonlinear Interpolation Consistency Loss

The Nonlinear Interpolation Consistency Loss ensures that the model outputs are consistent across interpolated data points. This is represented as follows:

$$\begin{aligned}\mathcal{L}_{consistency}^{Nonlinear} = &\mathcal{L}_{dice}\left(f_{\theta_1}(X_{input}^F), \underset{c\in C}{argmax}(f_{\theta_2}(X_{input}^F), c)\right) \\ &+ \mathcal{L}_{dice}\left(f_{\theta_2}(X_{input}^F), \underset{c\in C}{argmax}(f_{\theta_1}(X_{input}^F), c)\right)\end{aligned} \tag{20}$$

4.3. Differentiation Loss

Differentiated training primarily focuses on Student Model 1, incorporating supervised training with labeled data and a regularization term for unlabeled data. Consider labelled samples $(X_i^l, Y_i^l) \sim D^l$ from joint distribution $P(X, Y)$ and unlabelled samples $(X_i^u, X_j^u) \sim D^u$ from borderline distribution $P(X) = \frac{P(X,Y)}{P(X|Y)}$. Using SGD for every iteration t, the encoder-decoder parameter θ is updated minimising the objective function:

$$\mathcal{L}_{student1} = \mathcal{L}_{student1}^l + r(t) \cdot \mathcal{L}_{student1}^u \tag{21}$$

where $\mathcal{L}_{student1}^l$ is the cross entropy loss and dice loss applied over the labelled data D^l .Therefore, $\mathcal{L}_{student1}^l$ can be expressed as:

$$\mathcal{L}_{student1}^l = \frac{1}{2}\left(\mathcal{L}_{ce}(f_{\theta_1}(X^l), Y^l) + \mathcal{L}_{dice}(f_{\theta_1}(X^l), Y^l)\right) \tag{22}$$

$\mathcal{L}_{student1}^u$ is the interpolation consistency regularization loss applied over the unlabelled data D^u, $r(t)$ is the ramp function adjusting the weight of $\mathcal{L}_{student1}^u$ after every iteration. $\mathcal{L}_{student1}^u$ has two options: one is the LICR, and the other is the NICR.

LICR is calculated over (X_i^u, X_j^u) of sampled mini-batches and the pseudo labels $f_{\theta_T}(X_i^u)$ and $f_{\theta_T}(X_j^u)$. Next, interpolation $M_\beta(X_i^u, X_j^u)$ and model prediction $f_{\theta_1}(M_\beta(X_i^u, X_j^u))$ are computed updating θ to bring model prediction closer to the interpolation of the pseudo labels, $M_\beta(f_{\theta_T}(X_i^u), f_{\theta_T}(X_j^u)$. The deviation in model prediction and the interpolation of the pseudo labels is penalised using the mean squared loss. LICR can be expressed as:

$$\mathcal{L}_{LICR}^u = \mathbb{E}_{X_i^u, X_j^u}\left[\left\|f_{\theta_1}\left(M_\beta(X_i^u, X_j^u)\right) - M_\beta\left(f_{\theta_T}(X_i^u), f_{\theta_T}(X_j^u)\right)\right\|^2\right] \tag{23}$$

According to Equations 16 and 17, we can transform LICR to obtain a new loss, NICR, which is defined as:

$$\mathcal{L}_{NICR}^u = \mathbb{E}_{X_i^u, X_j^u}\left[\left\|f_{\theta_1}(M_\beta(X_i^u, X_i^u)) - f_{\theta_T}(X_i^u)\right\|^2\right] \tag{24}$$

5. Experiments

5.1. Dataset and Evaluation Metrics

ACDC Dataset: The ACDC dataset [5] consists of 200 annotated short-axis cardiac cine-MR images from 100 patients across four classes. 2D segmentation is more common than 3D [2]. Evaluation metrics include *Dice Similarity Coefficient* (DSC), *Jaccard*, *95% Hausdorff Distance* (95HD), and *Average Surface Distance* (ASD). Following the methods of BCP[3] and ABD[9], the input size was set to 256 × 256, with a batch size of 24 for training.

PROMISE12 Dataset: The PROMISE12 dataset [17] was introduced in the MICCAI 2012 prostate segmentation challenge, comprising MRI scans of 50 patients. All 3D scans are converted into 2D slices. DSC and ASD are used for evaluation. Following ABD[9], the input size was set to 224 × 224 with a batch size of 16 for training.

MS-CMRSeg 2019: The MS-CMRSeg 2019 dataset [35, 51] from the MICCAI 2019 challenge includes 45 multi-sequence cardiac MRI scans of cardiomyopathy patients. Evaluation metrics are DSC, Jaccard, 95HD, and ASD. Following the DiffRect[19] approach, the input size was 256 × 256. The training batch size was set to 8.

5.2. Comparison with SOTA Methods

Compared to SOTA methods on the ACDC test set, β**-FFT** demonstrates superior performance, particularly when using 10% labeled data, where the model achieves Dice and Jaccard scores of 90.50±0.04% and 83.12±0.12%, outperforming many recent methods, such as AD-MT and ABD.

On the PROMISE12 test set, β-FFT also outperforms existing methods with 20% labeled data, achieving a Dice score of 83.75±0.65% and an ASD of 1.20±0.07, surpassing AD-MT and ABD.

On the MS-CMRSEG 2019 dataset, β-FFT achieves a Dice score of 87.79±0.04% and a Jaccard index of 78.60±0.06% significantly outperforming popular semi-supervised approaches and approaching fully supervised performance with less labeled data.

Figure 3 presents a visual comparison of our method with other similar approaches.

Table 1. Comparisons with other methods on the ACDC test set.

Method	Scans used		Metrics			
	Labeled	Unlabeled	DSC↑	Jaccard↑	95HD↓	ASD↓
U-Net (MICCAI'2015) [29]	3(5%)	0	47.83	37.01	31.16	12.62
	7(10%)	0	79.41	68.11	9.35	2.70
	70(All)	0	91.44	84.59	4.30	0.99
DTC (AAAI'2021) [20]	3(5%)	67(95%)	56.90	45.67	23.36	7.39
URPC (MICCAI'2021) [21]			55.87	44.64	13.60	3.74
MC-Net (MICCAI'2021) [36]			62.85	52.29	7.62	2.33
SS-Net (MICCAI'2022) [38]			65.83	55.38	6.67	2.28
SCP-Net (MICCAI'2023) [47]			87.27	-	-	2.65
Cross Teaching (MIDL'2022) [22]			65.60	-	16.2	-
BCP (CVPR'2023) [3]			87.59	78.67	1.90	0.67
DiffRect (MICCAI'2024) [19]			82.46	71.76	7.18	1.94
ABD (CVPR'2024) [9]			88.96	80.70	1.57	0.52
AD-MT (ECCV'2024) [48]			88.75	80.41	**1.48**	**0.50**
Ours-β-FFT			**89.46**±0.12	**81.46**±0.22	1.78±0.32	0.55±0.10
DTC (AAAI'2021) [20]	7(10%)	63(90%)	84.29	73.92	12.81	4.01
URPC (MICCAI'2021) [21]			83.10	72.41	4.84	1.53
MC-Net (MICCAI'2021) [36]			86.44	77.04	5.50	1.84
SS-Net (MICCAI'2022) [38]			86.78	77.67	6.07	1.40
Cross Teaching (MIDL'2022) [22]			86.45	77.02	6.30	1.86
SCP-Net (MICCAI'2023) [47]			89.69	-	-	0.73
PLGCL (CVPR'2023) [4]			89.1	-	4.98	1.80
BCP (CVPR'2023) [3]			88.84	80.62	3.98	1.17
DiffRect (MICCAI'2024) [19]			89.27	81.13	3.85	1.00
ABD (CVPR'2024) [9]			89.81	81.95	**1.46**	0.49
AD-MT (ECCV'2024) [48]			89.46	81.47	1.51	**0.44**
Ours-β-FFT			**90.50**±0.04	**83.12**±0.12	2.38±0.87	0.62±0.13

Table 2. Comparisons with state-of-the-art semi-supervised segmentation methods on the PROMISE12 test set.

Method	Scans used		Metrics	
	Labeled	Unlabeled	DSC↑	ASD↓
U-Net [29]	7(20%)	0	60.88	13.87
	35(100%)	0	84.76	1.58
CCT [25]	7(20%)	28(80%)	71.43	16.61
URPC [21]			63.23	4.33
SS-Net [38]			62.31	4.36
SLC-Net [18]			68.31	4.69
SCP-Net [47]			77.06	3.52
ABD [9]			82.06	1.33
AD-MT [48]			79.82	1.77
Ours-β-FFT			**83.75**±0.65	**1.20**±0.07

Table 3. Segmentation results on MS-CMRSEG 2019 with 20% data labeled.

Method	Dice ↑	Jaccard↑	HD95↓	ASD↓
UAMT [45]	84.27	73.69	12.15	4.18
FixMatch [32]	84.31	73.57	17.79	4.81
CPS [8]	83.66	73.03	15.01	4.30
ICT [34]	83.66	73.06	17.24	4.85
MCNetV2 [37]	83.93	73.45	13.10	3.39
INCL [50]	84.33	73.92	9.95	2.61
DiffRect [19]	86.78	77.13	6.39	1.85
ABD [9]	87.25	77.77	11.74	4.25
AD-MT [48]	86.30	76.39	**3.56**	**1.21**
Ours-β-FFT	**87.79**±0.04	**78.60**±0.06	3.75±0.36	1.62±0.20
Supervised [29]	88.19	79.28	4.21	1.32

5.3. Ablation Study Analysis

The baseline method we use is an improved version of the BCP[3] from the ABD[9], combined with the Cross-Teaching framework[22]. Its structure consists of a mean-teacher framework with two student models and one teacher model.**All experiments are conducted on the ACDC dataset, with 10% of the data labeled**.

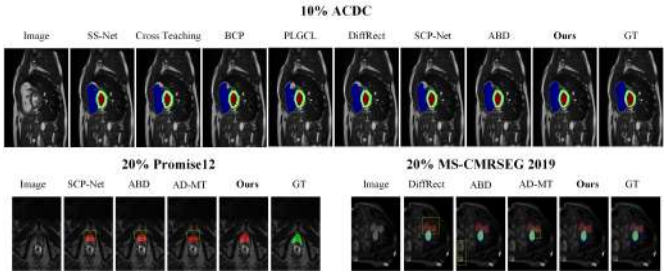


Figure 3. Visualization of segmentation results on ACDC dataset with 10% labeled data, PROMISE12 dataset with 20% labeled data,and MS-CMRSEG 2019 dataset with 20% labeled data .

5.3.1. Effect of Non-linear Interpolation Strategy

The experimental results demonstrate that the low-frequency component enhancement method, utilizing low-pass filters, effectively improves model performance on the ACDC dataset. Specifically, the 20x20 filter size achieved the best results on the validation set, while the 30x30 filter size excelled on the test set. This finding highlights the importance of selecting an appropriate filter size to balance detail and global information, ultimately optimizing the model's generalization ability.

Table 4. Effect of Low-Pass Filter Size on Model Performance.

H	**ACDC Validation dataset**		**ACDC Test dataset**	
	Dice↑	**Jaccard↑**	**Dice↑**	**Jaccard↑**
None	89.53	81.86	89.60	81.70
20x20	**90.23**	**82.94**	88.96	80.76
30x30	89.77	82.27	**89.96**	**82.29**
40x40	89.93	82.58	89.65	81.80
50x50	89.81	82.38	89.49	81.55

5.3.2. Impact of Differentiated Training Strategies on Model Performance

In this analysis, we use Student Model 1 as an example to validate the impact of differentiated training on model performance across the ACDC validation and test datasets.

The table 5 indicates that standardized training is conducted jointly for Student Model 1 and Student Model 2, while the differentiated training section represents additional training specifically for Student Model 1. Here, w denotes the use of weakly augmented data, and s denotes the use of strongly augmented data. The results show that by introducing the two differentiated training strategies, LICR and NICR, in the baseline model, the Dice value of the model on the test set significantly improves, with LICR(w) achieving a Dice value of 90.24 and NICR(w) reaching 90.23, highlighting the effectiveness of differentiated training in enhancing model performance.

Employing the Non-linear interpolation strategy, the

Table 5. Ablation study of Differentiated Training on ACDC Validation and Test Datasets.

Standardized Training		Differentiated Training				ACDC Validation		ACDC Test	
Baseline	Non-linear	LICR(w)	NICR(w)	LICR(s)	NICR(s)	Dice↑	Jaccard↑	Dice ↑	Jaccard ↑
✓						89.53	81.86	89.60	81.70
✓		✓				89.95	82.58	90.24	82.73
✓			✓			90.34	82.83	90.23	82.69
✓				✓		90.25	82.97	90.26	82.70
✓					✓	90.38	83.23	90.26	82.74
✓		✓		✓		90.09	82.79	90.13	82.47
✓		✓			✓	90.10	82.70	90.10	82.54
✓	✓					89.77	82.27	89.96	82.29
✓	✓	✓				90.53	83.54	90.27	82.76
✓	✓		✓			90.21	83.01	90.34	82.84
✓	✓			✓		90.30	83.14	90.38	82.95
✓	✓				✓	90.31	83.02	89.97	82.32
✓	✓	✓		✓		**90.77**	**83.81**	**90.54**	**83.23**

model's performance as measured by the Dice value on the test set reaches 90.54 when both LICR(w) and LICR(s) are applied simultaneously, confirming the effectiveness of our differentiated training approach. This outcome indicates that optimizing training strategies can significantly enhance model performance. Visualizing the training process, as shown in Figure 4, further illustrates that the performance of the two sub-models in β-FFT is significantly better than the Baseline, highlighting the benefits of our proposed method.

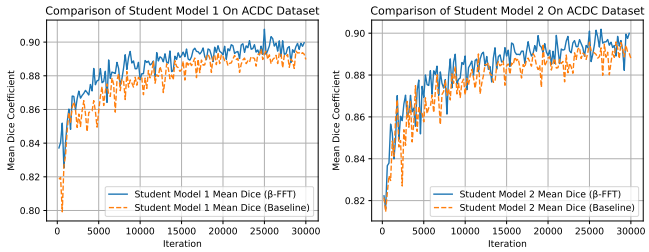


Figure 4. Comparison of the training process between β-FFT and Baseline methods.

5.3.3. Exploring the De-Homogenization Effects of Differentiated Training

To further investigate the impact of differentiated training on model homogenization, we first applied LICR(w) to Student Model 1 and LICR(s) to Student Model 2. This strategy aligns with the initial data augmentation strengths (strong and weak augmentation) assigned to the two student models, constituting a synchronized operation and enabling collaborative training. Experimental results demonstrate that this approach indeed enhances model performance. However, when we removed LICR(w) from Student Model 1 or LICR(s) from Student Model 2, the model performance further improved, indicating that applying LICR to a single sub-model is more effective than applying it to both simultaneously.

Further analysis reveals that when LICR(w) and LICR(s) are applied exclusively to Student Model 1, the model achieves its highest performance. As shown in **Table 6**, the Dice and Jaccard scores on the ACDC validation and test datasets reach **90.77 / 83.81** and **90.54 / 83.23** . In contrast,

Table 6. Ablation study results demonstrating the performance of two models after applying Non-Linear Interpolation, followed by differentiated training using LICR(w) and LICR(s) on the ACDC dataset.The gray row indicates that both Student Model 1 and Student Model 2 undergo simultaneous training, rather than differentiated training.

Differentiation Training		ACDC Validation dataset		ACDC Test dataset	
LICR(w)	LICR(s)	Dice↑	Jaccard↑	Dice ↑	Jaccard ↑
		89.77	82.27	89.96	82.29
Student 1		90.53	83.54	90.27	82.76
Student 1	**Student 2**	90.19	83.01	90.10	82.47
	Student 2	90.20	83.00	90.06	82.47
Student 1	**Student 1**	**90.77**	**83.81**	**90.54**	**83.23**

when both student models undergo simultaneous LICR(w) and LICR(s) training (as indicated by the gray row), performance slightly decreases. This further confirms that differentiated training effectively enhances model performance and mitigates homogenization issues.

5.3.4. Effect of Beta Distribution Sampling in Differentiated Training

We investigated the impact of Beta distribution parameter (a, a) on sample mixing and model performance. By adjusting a, we controlled sample diversity and feature complexity. Ablation results in Figure 5 show that smaller Beta parameters improve Dice and Jaccard scores, peaking at **Beta(0.1, 0.1)** with validation/test scores of 90.77 / 83.81 and 90.54 / 83.23, respectively. This indicates that lower Beta values enhance model generalization and robustness.

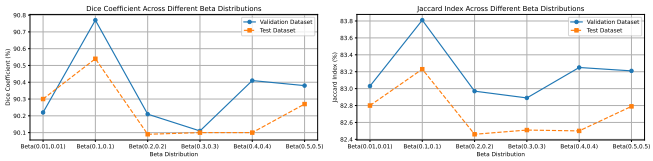


Figure 5. Differentiated training was conducted on Student 1 under the conditions of simultaneously applying nonlinear interpolation and both LICR(w) and LICR(s).

6. Conclusion

In this study, we address the issue of homogenization in co-training from both data and structural perspectives. We distinguish different sub-models using strong and weak augmentations and introduce a nonlinear interpolation method based on the Fast Fourier Transform (FFT) to generate more diverse training samples, thereby enhancing the model's generalization ability. Furthermore, we implement differentiated training by applying additional training to one of the models, effectively reducing homogenization. Extensive ablation experiments validate the effectiveness of our approach, with results demonstrating that β-FFT outperforms current state-of-the-art (SOTA) methods on three public medical image datasets.

Acknowledgment. This research was supported by: The Outstanding Award for Talent Project of the Chinese Academy of Sciences [Grant Number 29J20-052-III]; The Shaanxi Province Technological Innovation Guidance Special Project: Regional Science and Technology Innovation Center, Strategic Scientific and Technological Strength Category (No.2024QY-SZX-26); The Key Project for Teaching Research of the Medical Department of Wuhan University [Grant Number 2024ZD21]; The Key R&D Project of Hubei Province [Grant Number 2023BCB024].

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020. 1, 2

[2] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017. 6

[3] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11514–11524, 2023. 2, 3, 6, 7

[4] Hritam Basak and Zhaozheng Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19786–19797, 2023. 7

[5] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37 (11):2514–2525, 2018. 6

[6] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning. *Advances in Neural Information Processing Systems*, 35:32424–32437, 2022. 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[8] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 2, 7

[9] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4070–4080, 2024. 3, 6, 7

[10] Kaiwen Cui, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Fangneng Zhan, and Shijian Lu. Genco: Generative co-training for generative adversarial networks with limited data. In *Proceedings of the AAAI conference on artificial intelligence*, pages 499–507, 2022. 3

[11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004. 1

[12] Nastassya Horlava, Alisa Mironenko, Sebastian Niehaus, Sebastian Wagner, Ingo Roeder, and Nico Scherf. A comparative study of semi-and self-supervised semantic segmentation of biomedical microscopy data. *arXiv preprint arXiv:2011.08076*, 2020. 2

[13] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6728–6736, 2019. 2

[14] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9475–9484, 2021. 3

[15] Xiaoxu Li, Yu Peng, and Min Xu. Patch-shuffle-based semi-supervised segmentation of bone computed tomography via consistent learning. *Biomedical Signal Processing and Control*, 80:104239, 2023. 3

[16] Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. *arXiv preprint arXiv:2308.09281*, 2023. 2, 3

[17] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 6

[18] Jinhua Liu, Christian Desrosiers, and Yuanfeng Zhou. Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–150. Springer, 2022. 7

[19] Xinyu Liu, Wuyang Li, and Yixuan Yuan. Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation. *arXiv preprint arXiv:2407.09918*, 2024. 6, 7

[20] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, number 10, pages 8801–8809, 2021. 2, 7

[21] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *Medical Image*

Computing and Computer Assisted Intervention, pages 318–329. Springer, 2021. 7

[22] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International Conference on Medical Imaging with Deep Learning*, pages 820–833. PMLR, 2022. 3, 7

[23] Fei Lyu, Mang Ye, Jonathan Frederik Carlsen, Kenny Erleben, Sune Darkner, and Pong C Yuen. Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. *IEEE Transactions on Medical Imaging*, 42(3):797–809, 2022. 2

[24] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1

[25] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 2, 7

[26] Qianyao Qiang, Bin Zhang, Feiping Nie, and Fei Wang. Multi-view semi-supervised learning with adaptive graph fusion. *Neurocomputing*, 557:126685, 2023. 3

[27] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018. 3

[28] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018. 2

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 7

[30] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *IEEE Workshop on Applications of Computer Vision*, 2005. 1

[31] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 1

[32] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2, 3, 7

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2

[34] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022. 2, 7

[35] Fuping Wu and Xiahai Zhuang. Minimizing estimated risks on unlabeled data: A new formulation for semi-supervised medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6021–6036, 2022. 6

[36] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *Medical Image Computing and Computer Assisted Intervention*, pages 297–306. Springer, 2021. 7

[37] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022. 7

[38] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 34–43. Springer, 2022. 7

[39] Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical image analysis*, 65:101766, 2020. 3

[40] Yang Xiu, Xinyi Zheng, Linlin Sun, and Zhuohao Fang. Fremix: Frequency-based mixup for data augmentation. *Wireless Communications and Mobile Computing*, 2022(1): 5323327, 2022. 3

[41] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14383–14392, 2021. 3

[42] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022. 3

[43] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020. 3

[44] Boon Peng Yap and Beng Koon Ng. Cut-paste consistency learning for semi-supervised lesion segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6160–6169, 2023. 3

[45] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen,*

China, October 13–17, 2019, proceedings, part II 22, pages 605–613. Springer, 2019. 2, 7

[46] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[47] Zhenxi Zhang, Ran Ran, Chunna Tian, Heng Zhou, Xin Li, Fan Yang, and Zhicheng Jiao. Self-aware and cross-sample prototypical learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2305.16214*, 2023. 7

[48] Zhen Zhao, Zicheng Wang, Longyue Wang, Dian Yu, Yixuan Yuan, and Luping Zhou. Alternate diverse teaching for semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pages 227–243. Springer, 2025. 3, 7

[49] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Time-consistent self-supervision for semi-supervised learning. In *International conference on machine learning*, pages 11523–11533. PMLR, 2020. 2

[50] Ye Zhu, Jie Yang, Si-Qi Liu, and Ruimao Zhang. Inherent consistent learning for accurate semi-supervised medical image segmentation. *arXiv preprint arXiv:2303.14175*, 2023. 7

[51] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12): 2933–2946, 2018. 6

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：李翀
方向：人工智能
单位：中国科学院自动化研究所

题目：基于词元对齐的高效大语言模型词表替换方法

TokAlign: Efficient Vocabulary Adaptation via Token Alignment

Chong Li, Jiajun Zhang*, Chengqing Zong
State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, CAS, Beijing, China
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
lichong2021@ia.ac.cn, {jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

Tokenization serves as a foundational step for Large Language Models (LLMs) to process text. In new domains or languages, the inefficiency of the tokenizer will slow down the training and generation of LLM. The mismatch in vocabulary also hinders deep knowledge transfer between LLMs like token-level distillation. To mitigate this gap, we propose an efficient method named **TokAlign** to replace the vocabulary of LLM from the token co-occurrences view, and further transfer the token-level knowledge between models. It first aligns the source vocabulary to the target one by learning a one-to-one mapping matrix for token IDs. Model parameters, including embeddings, are rearranged and progressively fine-tuned for the new vocabulary. Our method significantly improves multilingual text compression rates and vocabulary initialization for LLMs, decreasing the perplexity from 3.4e^2 of strong baseline methods to 1.2e^2 after initialization. Experimental results on models across multiple parameter scales demonstrate the effectiveness and generalization of TokAlign, which costs as few as 5k steps to restore the performance of the vanilla model. After unifying vocabularies between LLMs, token-level distillation can remarkably boost (+4.4% than sentence-level distillation) the base model, costing only 235M tokens. [1]

1 Introduction

Large language models (Touvron et al., 2023a; OpenAI, 2023; Yang et al., 2024) first tokenize text input into several tokens during inference and training, which compresses text and addresses the out-of-vocabulary problem (Sennrich et al., 2016; Wu et al., 2016; Kudo, 2018). However, the low compression rate of vanilla tokenizers on new languages or domains decelerates the training and inference process. As shown in Figure 1, the compression rate of capable large language models like LLaMA3 (Meta, 2024) on low-resource languages still largely lags behind the others. For example, Armenian text is 3.95x longer in tokens than English text under the same byte size with the LLaMA3 tokenizer. On the other hand, each LLM has specific strengths and weaknesses, which arise from its pre-training corpus and method. The mismatch in the vocabulary impedes the deep knowledge transfer between them like token-level distillation and ensemble (Xu et al., 2024; Lu et al., 2024). Considering the huge cost of re-training LLM for a new tokenizer, it is important to investigate efficient vocabulary adaptation methods.

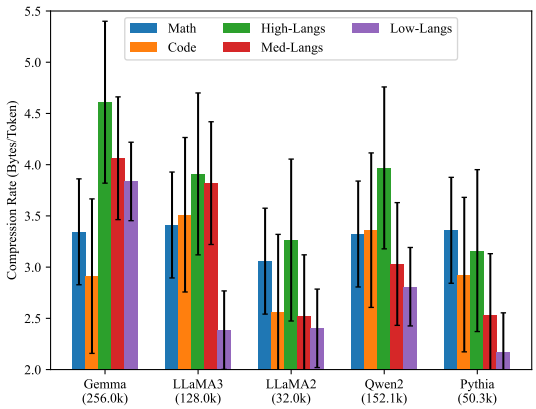


Figure 1: The compression rates of tokenizers across different domains and languages, which are still low in the code domain and low-resource languages for most of tokenizers. Refer to Table 6 in Appendix B.1 for more details.

To address the problems above, we introduce a novel method called **TokAlign** for large language models from a view of token-token co-occurrences. It is motivated by the general process of training an LLM: the pre-training corpus is first tokenized into tokens, and then input into the model. Given the same pre-training corpus, different tokenizers result in various sequences of token IDs, while the semantic and syntactic information is

*Corresponding author.
[1] Our codes and model are available at https://github.com/ZNLP/TokAlign

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4109–4126
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

preserved in the token-token co-occurrence. Therefore, TokAlign strives to align token IDs from the original vocabulary and the target ones based on the global token-token co-occurrence matrix (Pennington et al., 2014) and learns a token-token alignment matrix. We further propose two metrics to evaluate the performance of the token-token alignment matrix based on text matching and semantic similarity. Given the learned alignment matrix, the new target embedding and language modeling head of LLM ("lm_head" in the Transformers (Wolf, 2019)) are initialized from the parameters of the most similar source token. Further vocabulary adaptation process is divided into a progressive two-stage procedure to improve the stability of convergence.

Given a target multilingual vocabulary for substitution, the model trained on the English corpus obtains a good initialization, decreasing the perplexity from 3.4e^2 to 1.2e^2, and improves 29.2% compression rates across 13 languages on average. The training process of TokAlign is 1.92x faster than strong baseline methods, and does not require additional hundreds of GPU hours to train a hypernetwork for embedding initialization (Minixhofer et al., 2024). Experimental results on models across different scales show that as few as 5k steps are needed for our method to recover the performance of vanilla models on the general domain. Moreover, unifying vocabulary between models further facilitates the token-level distillation, which is 4.4% better than the sentence-level distillation on the same corpus. The performance of the 1B model is comparable with the vanilla 7B model after token-level distillation from a capable LLM. In summary, our contributions are as follows:

- We propose an unsupervised method to align token IDs between two vocabularies and replace the vocabulary of LLMs from the token-token co-occurrence view.
- We introduce two metrics to evaluate the performance of the token-level alignment matrix learned, which are proportional to the initial loss of pre-training.
- Experimental results on ten datasets show that our method promotes the cross-lingual knowledge transfer among multiple languages and deep knowledge transfer between models like token-level distillation.

2 Related Works

Our work is related to word representation, large language models, and vocabulary adaption, which will be briefly introduced below.

Word Representation Based on the distributional semantic hypothesis, Bengio et al. (2003) introduced the neural probabilistic language model to learn word representation. Researchers mainly focus on improving the effectiveness during learning word representations (Mikolov et al., 2013a,b; Bojanowski et al., 2017; Li et al., 2017; Wang et al., 2018), which provide a good initialization for neural networks like LSTM and GRU (Hochreiter, 1997; Chung et al., 2014). GloVe (Pennington et al., 2014) provides a method to train word representations from a view of global word-word co-occurrence matrix decomposition. It motivates us to train a word representation for each token and align tokens from statistical co-occurrence information in the pre-training corpus.

Large Language Model Through scaling in the parameters and pre-training corpus (Kaplan et al., 2020; Hoffmann et al., 2022), large language models like GPT-4 and LLaMA3 (Radford et al., 2018, 2019; Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b; Meta, 2024; GLM et al., 2024) demonstrate impressive performance across multiple tasks. However, the mismatch in the vocabulary greatly hinders the deep knowledge transfer between different models. We aim to mitigate this problem by introducing an efficient method to replace the tokenizer of a large language model.

Vocabulary Adaption is investigated mainly in the multilingual domain, especially the cross-lingual knowledge transfer problem (Scao et al., 2023; Muennighoff et al., 2023; Yang et al., 2023; Zhu et al., 2023; Üstün et al., 2024; Li et al., 2024; Liu et al., 2024; Minixhofer et al., 2024; Yamaguchi et al., 2024; Mundra et al., 2024; Balde et al., 2024). It aims to improve the encoding effectiveness of tokenizer on corpora from new languages or domains, and is often implemented by extending the original vocabulary (Tran, 2020; Chau et al., 2020; Minixhofer et al., 2022; Dobler and de Melo, 2023; Downey et al., 2023). Most methods, like Focus (Dobler and de Melo, 2023), rely on the tokens belonging to both source vocabulary and target vocabulary to initialize the other new tokens in the target vocabulary. Our method differs from

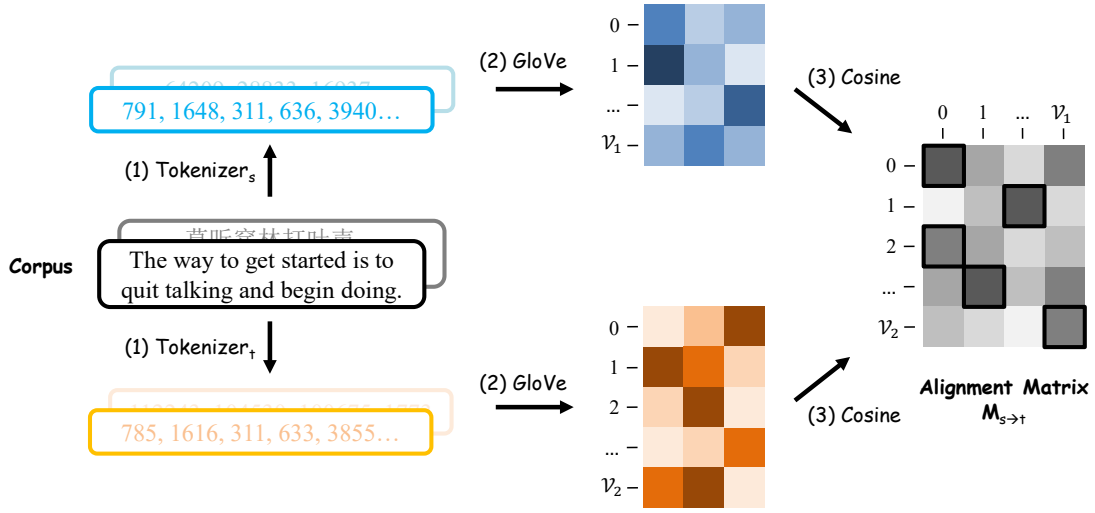


Figure 2: Illustration of TokAlign to align token IDs from different vocabularies. We train token representations on the tokenized corpus, and align token IDs by the cosine similarity. It is noted that the IDs of tokens belonging to both vocabularies are directly replaced without alignment.

these studies for the whole replacement of vocabulary and does not rely on the tokens in both source vocabulary and target vocabulary.

The pipeline of TokAlign to adapt vocabulary is similar to WECHSEL(Minixhofer et al., 2022), while the main difference lies in the representation and alignment of tokens. WECHSEL requires a bilingual dictionary and word representation to align tokens and calculates the similarity between tokens by tokenizing all words in the dictionary and linearly composing word representations. In contrast, TokAlign conducts token representation learning and alignment in an unsupervised way, which can apply to languages without bilingual dictionaries.

3 Method: TokAlign

3.1 Vocabulary Alignment

As shown in Figure 2, there are three steps for TokAlign to align two vocabularies from the token-token co-occurrence information. We denote the source tokenizer as Tokenizer$_s$, which has $\mathcal{V}_s$ tokens, and the target tokenizer as Tokenizer$_t$ with $\mathcal{V}_t$ tokens, correspondingly.

Step 1: Tokenization The comprehensiveness of the pre-training corpus is important to obtain a well-trained token representation. An unbalanced corpus makes it hard to learn the representation of tokens in the tail of vocabulary. Thus, the corpus used in this work is empirically composed of multilingual corpus "CulturaX" [40%] (Nguyen et al., 2024), code corpus "The Stack" [30%] (Kocetkov et al., 2023), and math corpus "Proof-Pile-2" [30%] (Azerbayev et al., 2024). We tokenize the mixed corpus using various tokenizers and obtain multiple sequences of token IDs for the same corpus. The default amount of tokens used in this step is 1B, which is investigated in Appendix B.2.

Step 2: Token Representation Learning We adopt GloVe (Pennington et al., 2014) to learn the representation of tokens from the first step. The main reason is that GloVe considers more global statistical information than those slide window methods like CBOW and FastText (Mikolov et al., 2013a,b; Bojanowski et al., 2017). The details of training settings for GloVe vectors refer to Appendix A.

Step 3: Token Alignment Based on the assumption that token representations capture the semantic information in the token, we align token IDs using the pair-wise cosine similarity of learned token representations. It should be noted that the IDs of tokens belonging to both vocabularies are directly replaced without the need to align. $M_{s\to t}$ denotes the learned token-token alignment matrix, which records the pair-wise similarity of each source token and target token. It can serve as the one-to-one mapping function for each source/target token to find the most similar token from the target/source vocabulary.

3.2 Alignment Evaluation

Figure 3(a) illustrates our metrics to evaluate the performance of alignment matrix $M_{s\to t}$. We first

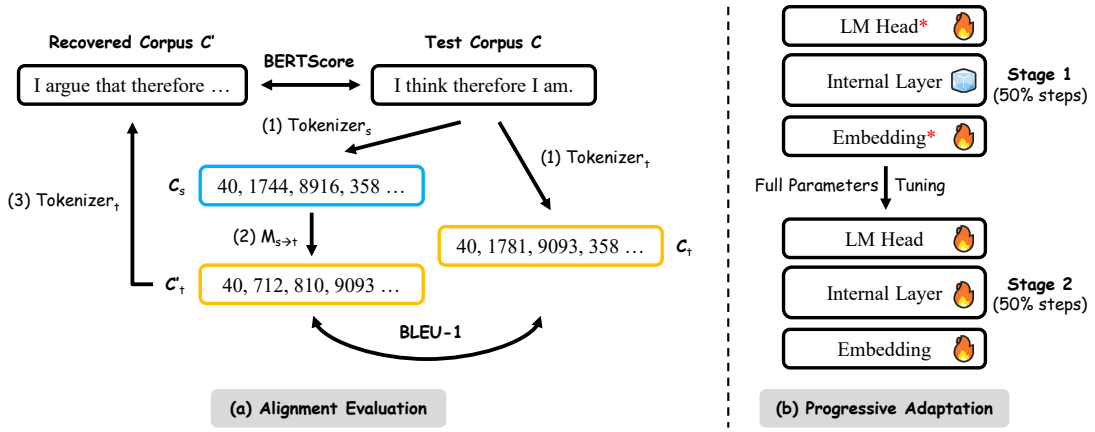


Figure 3: (a) We choose BLEU-1 and BERTScore to evaluate the performance of alignment matrix $M_{s\to t}$ (b) Embedding and lm_head are tuned at the first half part of the process, followed by full parameter tuning. * indicates the parameter of each target token is first initialized from the most similar source token by alignment matrix $M_{s\to t}$.

tokenize the test corpus $\mathcal{C}$ using different tokenizers, which results in $\mathcal{C}_s$ and $\mathcal{C}_t$. The token ID corpus $\mathcal{C}_s$ from the source tokenizer is converted to its most similar target token ID by alignment matrix $M_{s\to t}$, and comes to the corpus $\mathcal{C}_t^{'}$. From the view of token ID matching, the higher BLEU-1 score between $\mathcal{C}_t^{'}$ and the corpus $\mathcal{C}_t$ from the Tokenizer$_t$, the better alignment matrix $M_{s\to t}$ is.

We further propose a semantic evaluation metric: It de-tokenizes the target token ID corpus $\mathcal{C}_t^{'}$ using Tokenizer$_t$ into the recovered text corpus $\mathcal{C}^{'}$, and evaluates the semantic similarity between $\mathcal{C}^{'}$ and original corpus $\mathcal{C}$ using BERTScore. The better alignment matrix $M_{s\to t}$ learned preserves more semantics in the test corpus $\mathcal{C}$, bringing higher BERTScore of the recovered $\mathcal{C}^{'}$ and $\mathcal{C}$.

3.3 Progressive Adaptation

Given the alignment matrix $M_{s\to t}$, the parameters of each token in the target vocabulary are initialized from the ones of the most similar source token. We find that these re-arranged embeddings and lm_head provide a good initialization for the new model (Section 4.2.1). Figure 3(b) illustrates the two-stage tuning for an LLM to adapt to the new vocabulary. The re-arranged embedding and lm_head are tuned first to avoid loss spike and improve the training stability (Figure 6). The other parameters of internal layers are further tuned together in the last half-part process.

4 Experiments

4.1 Experiments Settings

Large Language Models We adopt the fully open-source language model series Pythia (Biderman et al., 2023) as base models in this work. It is noted that we do not intend to achieve state-of-the-art large language model performance but rather investigate an efficient method to replace the English-centric tokenizer like Pythia. To transfer token-level knowledge from other capable large language models, tokenizers and vocabularies of Gemma (Team et al., 2024), Qwen2 (Yang et al., 2024), LLaMA2 (Touvron et al., 2023b), and LLaMA3 (Meta, 2024) are selected as the target to replace. We report hyper-parameters in Appendix A.

Corpus To reduce the risk of distribution shift from the training data, we choose the vanilla pre-training corpus Pile (Gao et al., 2020) of Pythia in the fine-tuning process. We also investigate the robustness of the corpus used in the vocabulary alignment by replacing it with Slimpajama (Soboleva et al., 2023). Corpora of downstream tasks and multiple languages are applied in cross-lingual and cross-model knowledge transfer experiments (Section 4.2.1 and 4.2.2).

Evaluation Tasks Following the common practices to evaluate large language models (Lin et al., 2022; Biderman et al., 2023; Zhang et al., 2024), there are 10 datasets, including commonsense reasoning (Clark et al., 2018; Mihaylov et al., 2018; Zellers et al., 2019; Ponti et al., 2020; Bisk et al., 2020; Sakaguchi et al., 2020) and reading comprehension (Clark et al., 2019) tasks, used in this work. To avoid the randomness from the prompt and evaluation method, we adopt the default prompt from the commonly used language model evaluation harness framework (Gao et al., 2024). Further information about the evaluation tasks is reported in Appendix C.

Model	High					Medium					Low			
	ar	de	en	ja	zh	bn	ko	th	uk	vi	ta	te	ur	Avg ↓
$\text{Qwen2}_{1.5B}$	4.7	11.1	15.7	6.0	4.6	2.4	3.3	2.6	5.7	3.3	2.8	3.4	4.0	5.3
Pythia_{1B}	7.6	15.4	**21.7**	9.9	13.2	3.4	5.6	4.3	6.7	6.3	2.9	3.3	5.8	8.2
w/ Focus Init.	$4.1e^3$	$1.7e^5$	$1.8e^6$	$2.1e^4$	$9.6e^2$	$6.5e^4$	$1.0e^3$	$5.6e^3$	$1.6e^6$	$8.4e^2$	$5.0e^4$	$1.9e^5$	$1.9e^5$	$3.1e^5$
+ LAT	8.3	27.1	59.7	14.0	14.0	3.6	5.9	3.8	7.3	5.9	3.5	3.6	4.3	12.4
w/ ZeTT Init.	$3.0e^2$	$4.2e^2$	$1.3e^2$	$1.2e^3$	$2.4e^2$	$3.0e^2$	$2.4e^2$	$3.3e^2$	$2.5e^2$	$2.0e^2$	$2.4e^2$	$1.8e^2$	$4.7e^2$	$3.4e^2$
+ LAT	7.1	15.7	26.4	10.0	10.3	2.8	5.0	3.6	5.9	4.9	2.6	2.7	4.2	7.8
w/ TokAlign Init.	$1.2e^2$	$2.2e^2$	$1.0e^2$	$3.6e^2$	$1.2e^2$	46.5	60.1	70.8	$1.5e^2$	49.2	61.0	$1.1e^2$	50.9	$1.2e^2$
+ LAT	**6.3**	**13.9**	23.6	**8.9**	**9.0**	**2.4**	**4.4**	**3.2**	**5.2**	**4.4**	**2.3**	**2.4**	**3.7**	**6.9**
Qwen2_{7B}	3.9	8.1	11.8	4.9	3.8	2.1	2.9	2.3	3.8	2.9	2.3	2.6	3.3	4.2
$\text{Pythia}_{6.9B}$	5.9	10.8	**16.7**	7.9	9.9	3.0	4.6	3.7	4.9	4.9	2.6	2.9	4.8	6.3
w/ Focus Init.	$6.9e^3$	$1.6e^5$	$1.2e^6$	$2.4e^4$	$1.3e^3$	$2.5e^4$	$7.2e^2$	$3.3e^3$	$1.9e^6$	$7.9e^2$	$1.7e^4$	$1.5e^5$	$1.2e^5$	$2.8e^5$
+ LAT	6.8	17.6	39.3	10.8	11.1	2.5	5.0	3.3	5.2	4.8	2.3	2.5	3.7	8.8
w/ TokAlign Init.	$1.2e^2$	$1.9e^2$	81.4	$3.7e^2$	$1.3e^2$	52.5	53.3	66.2	$1.4e^2$	49.2	46.4	92.1	48.7	$1.1e^2$
+ LAT	**5.2**	**9.9**	17.8	**7.4**	**7.9**	**2.1**	**3.8**	**2.8**	**4.0**	**3.7**	**2.1**	**2.1**	**3.1**	**5.5**
Δ Length (%) ↓	−44.5	−13.1	−0.8	−32.4	−50.0	−22.2	−52.2	−46.1	−15.5	−51.7	−20.3	−2.9	−28.5	−29.2

Table 1: The normalized perplexity on the valid corpus of CulturaX. The perplexity is normalized to the vocabulary of Pythia following Wei et al. (2023). "**High**", "**Medium**", and "**Low**" indicates the available amount of linguistic resources. "w/ xxx Init." denotes the performance of the model after initialization without any tuning steps.

Baselines We introduce the following vocabulary adaptation methods as baseline methods in this work:

- **Random Initialization** for each token $t \in \{\mathcal{V}_t \setminus (\mathcal{V}_t \cap \mathcal{V}_s)\}$ employs the default initialization method of huggingface Transformers and reuses the parameters of token $t \in \{\mathcal{V}_t \cap \mathcal{V}_s\}$, which belongs to both vocabularies.
- **Random Permutation** initializes each token $t \in \{\mathcal{V}_t \setminus (\mathcal{V}_t \cap \mathcal{V}_s)\}$ using the parameter of randomly chosen token from the source vocabulary. The parameters of shared tokens are also reused.
- **Multivariate** initializes each token $t \in \{\mathcal{V}_t \setminus (\mathcal{V}_t \cap \mathcal{V}_s)\}$ by sampling from the multivariate Gaussian distribution with the mean and covariance of source embedding E_s.
- **Mean** use the mean of source embedding E_s to initialize all tokens $t \in \{\mathcal{V}_t \setminus (\mathcal{V}_t \cap \mathcal{V}_s)\}$.
- **WECHSEL** (Minixhofer et al., 2022) linearly transfers embeddings of source tokens into target tokens by tokenizing and recomposing additional word embeddings W^s and W^t, which are aligned with a bilingual dictionary.
- **OFA** (Liu et al., 2024) factorizes the embeddings of source model E_s into the primitive embedding P and source coordinate F_s that is further re-composed by multilingual word embedding W to the target coordinate F_t. The assembled primitive embedding P and target coordinate F_t yield the target embedding E_t.
- **Focus** (Dobler and de Melo, 2023) initializes the embedding parameters of token $t \in \{\mathcal{V}_t \setminus (\mathcal{V}_t \cap \mathcal{V}_s)\}$ using the weighted sum of the ones from the token $t \in \{\mathcal{V}_t \cap \mathcal{V}_s\}$. It largely depends on the size of $\|\mathcal{V}_t \cap \mathcal{V}_s\|$, and performs poorly when the overlapping percentage of $\mathcal{V}_t$ and $\mathcal{V}_s$ is low.
- **ZeTT** (Minixhofer et al., 2024) trains an additional hypernetwork H_θ to generate the parameters for each token $t \in \mathcal{V}_t$. The added hypernetwork brings a lot of training costs.

4.2 Main Results

We first report the final results of two applications after replacing vocabulary: cross-lingual transfer (Section 4.2.1) and cross-model knowledge transfer (Section 4.2.2), then show vocabulary adaptation results of methods (Section 4.3).

4.2.1 Cross-lingual Transfer

When applied to new domains or languages, tokenizers with higher compression rates can speed up the learning and inference of large language models. From the view of token co-occurrence, tokens from other languages can be aligned and initialized by the tokens with similar semantics in the source vocabulary, which can boost the cross-lingual knowledge transfer. Therefore, we replace the English-centric tokenizer of Pythia with the one of Qwen2 to evaluate the performance on cross-lingual transfer settings.

As shown in Table 1, the perplexity of Pythia initialized using TokAlign ($1.2e^2$) is significantly

Model	XNLI							PAWS-X					XCOPA			XStoryCloze				Avg
	en	de	zh	ar	th	vi	ur	de	en	ja	ko	zh	th	vi	ta	en	zh	ar	te	
Pythia$_{1B}$	**51.0**	37.8	42.6	35.9	34.8	37.0	34.7	49.6	49.3	54.8	**54.9**	52.9	54.0	53.2	55.4	**64.3**	48.6	48.0	52.9	48.0
w/ Focus Init.	32.8	32.2	33.6	33.6	33.5	32.0	32.8	44.8	44.9	45.7	44.8	44.7	52.4	48.6	**57.0**	45.9	47.8	**48.8**	46.5	42.2
+ LAT	46.0	35.1	34.9	32.9	32.5	35.4	34.7	50.6	45.5	55.9	53.4	**55.3**	53.8	52.6	55.4	55.8	48.8	47.6	50.4	46.1
w/ ZeTT Init.	45.9	34.6	32.9	32.8	33.5	33.6	34.5	51.5	50.3	54.8	51.5	53.5	52.6	48.2	55.6	53.2	46.9	46.9	48.1	45.3
+ LAT	48.6	38.6	40.6	36.9	36.0	39.3	35.1	53.0	51.0	55.8	53.8	**55.3**	**55.8**	50.8	54.0	60.3	49.3	47.2	52.1	48.1
w/ TokAlign Init.	49.9	36.6	33.2	31.8	33.2	34.4	34.4	52.4	**52.1**	**56.1**	54.7	**55.3**	53.6	48.0	55.2	61.0	47.6	47.1	51.0	46.7
+ LAT	50.9	**39.3**	**42.7**	**37.4**	**37.4**	**40.3**	**35.7**	**54.6**	50.2	55.9	**54.9**	**55.3**	55.2	**53.6**	53.6	64.0	**51.1**	47.8	**53.5**	**49.1**
Pythia$_{6.9B}$	54.4	**39.0**	**46.2**	39.3	39.8	39.3	36.4	43.8	40.2	50.2	54.2	50.2	**56.2**	54.4	52.2	**70.4**	53.9	**50.3**	53.8	48.6
w/ Focus Init.	31.5	31.3	33.0	32.6	33.4	32.2	32.6	44.8	42.4	52.7	45.5	44.7	52.2	48.6	**55.6**	44.5	47.1	47.8	47.1	42.1
+ LAT	52.6	34.9	36.6	35.1	33.6	39.0	34.5	**51.1**	43.8	55.9	55.3	55.4	54.2	52.4	53.8	61.0	48.7	47.7	53.7	47.3
w/ TokAlign Init.	53.3	36.3	35.0	34.6	34.6	33.0	33.8	48.8	44.6	**56.2**	55.7	55.3	54.6	52.2	54.6	66.8	48.6	47.7	50.0	47.1
+ LAT	**55.2**	35.8	43.5	**40.4**	**40.2**	**43.0**	**37.1**	43.2	**45.8**	55.8	**55.8**	**55.5**	54.6	**57.0**	54.6	70.2	**54.4**	49.3	**53.9**	**49.7**

Table 2: Zero-shot in-context learning results of cross-lingual transfer. Refer to Table 8 for few-shot results.

Model	ARC-E		BoolQ		HellaSwag		OpenbookQA		PIQA		WinoGrande		Avg	
	0	5	0	5	0	5	0	5	0	5	0	5	0	5
Pythia$_{1B}$	56.82	58.71	60.43	57.37	37.68	37.66	18.80	19.00	70.40	71.49	53.20	52.01	49.55	49.37
+ Direct tuning	57.49	55.64	70.70	72.11	41.24	41.60	25.40	28.40	69.04	70.08	54.70	54.78	53.10	53.77
+ Sentence distill	52.27	53.41	67.49	67.06	39.03	39.08	21.80	22.80	66.97	68.99	51.85	52.17	49.90	50.58
w/ Gemma$_{7B}$	55.39	56.99	67.19	69.69	36.53	37.26	19.00	22.80	68.82	69.21	52.33	53.51	49.88	51.58
w/ Qwen2$_{7B}$	62.33	63.17	70.18	72.54	41.58	42.21	22.00	**28.20**	**73.01**	73.18	55.01	55.56	54.02	55.81
w/ LLaMA3$_{8B}$	**64.02**	**64.56**	**73.91**	**74.19**	**42.11**	**42.34**	**24.20**	27.60	72.74	**73.83**	**55.49**	**56.43**	**55.41**	**56.49**
Pythia$_{6.9B}$	65.99	69.23	62.84	62.02	47.56	47.64	25.00	27.00	74.65	75.41	60.46	62.43	56.08	57.29
+ Direct tuning	66.25	66.20	79.30	78.87	52.21	53.39	33.20	33.00	72.91	74.48	62.90	61.72	61.13	61.28
+ Sentence distill	61.70	65.36	76.64	76.88	48.98	51.33	28.20	30.40	70.18	71.55	58.96	62.19	57.44	59.62
w/ Gemma$_{7B}$	67.59	68.94	76.06	75.66	47.83	48.36	28.40	31.40	73.78	75.52	59.04	64.17	58.78	60.67
w/ Qwen2$_{7B}$	**71.72**	**73.27**	**79.85**	**80.00**	**50.78**	**51.12**	**29.20**	**34.00**	**77.26**	**77.91**	**61.33**	**64.56**	**61.69**	**63.48**
w/ LLaMA3$_{8B}$	67.05	69.78	77.83	78.78	48.83	50.15	26.00	32.00	74.21	76.22	60.22	60.93	59.02	61.31

Table 3: The main results of token-level distillation on six downstream tasks with only 235M tokens. "+Sentence distill" denotes the sentence-level distillation results with Qwen2$_{7B}$(Yang et al., 2024), which fine-tunes on the output from Qwen2$_{7B}$ given questions as prompt.

better than other two strong baseline methods Focus ($2.9e^5$) and ZeTT($3.4e^2$). The length of tokens after text tokenization has reduced by 29.2% on average across these languages. After only 2k steps of **L**anguage **A**daptation **T**uning ("+LAT"), TokAlign improved 14.5% over the vanilla model on average, while Focus still performed worse. It is noted that the performance of Pythia using TokAlign on three low-resource languages even outperforms the ones of Qwen2 with a similar parameter amount.

Table 2 and 8 in Appendix B.5 further report zero-shot and few-shot in-context learning results on four multilingual datasets. We can find that TokAlign brings a better-initialized model than the baseline method Focus (+4.4%), and transfers the knowledge into other languages like Japanese (ja, +2.3%) and Vietnamese (vi, +2.2%).

It is interesting to find that the perplexity of Pythia$_{1B}$ initialized by TokAlign reaches $1.2e^2$, while the in-context learning results are comparable with the ones of Focus after adapting on the multilingual corpus. We argue that it arises from the reserved English ability with TokAlign (54.2%), which significantly outperforms Focus (40.8%).

4.2.2 Cross-model Transfer

Unifying vocabulary with capable LLMs enables token-level distillation and transfers the knowledge learned into smaller models to decrease inference costs. In this section, training samples from downstream tasks and the corpus of Pile are used in the token-level distillation experiments. The logit of each token from the teacher model is taken as the soft label for Pythia to learn. Specifically, we add the KL-divergence loss between the logit from the teacher and student models to the original next token prediction loss on the training samples. The proportion of training samples is empirically set to 15% to avoid a significant degradation in language modeling performance (Wei et al., 2023). There are two baseline methods: "+ Direct tuning", where models directly fine-tune on the training samples, and "+ Sentence distill" for comparison, where models fine-tune on the output text from the teacher model given the question as a prompt.

Table 3 reports the results of two baseline methods and token-level distillation from three teacher models using 235M tokens. It can be found that token-level distillation is significantly better than sentence-level distillation. In the neural machine

Model	#GPU Hour	ARC-E		BoolQ		HellaSwag		OpenbookQA		PIQA		WinoGrande		Avg	
		0	5	0	5	0	5	0	5	0	5	0	5	0	5
Pythia_{1B}	–	56.82	58.71	60.43	57.37	37.68	37.66	18.80	19.00	70.40	71.49	53.20	52.01	49.55	49.37
w/ Rand. Init.	99.70	31.36	31.61	37.83	49.11	26.35	26.40	14.00	12.60	54.57	55.33	49.17	49.17	35.55	37.37
w/ Rand. Perm.	99.70	31.69	32.95	37.77	54.80	26.43	26.39	14.00	12.60	55.50	55.98	47.04	50.67	35.40	38.90
w/ Multivariate	99.70	32.79	34.18	45.08	49.72	27.67	27.87	15.20	16.20	56.09	57.83	50.51	50.12	37.89	39.32
w/ Mean	99.70	44.87	46.97	53.39	55.20	31.59	31.67	16.20	17.00	61.32	62.46	49.25	51.85	42.77	44.19
w/ OFA	99.70	38.17	37.79	55.14	52.35	28.29	28.62	14.40	12.20	58.43	58.54	49.96	50.99	40.73	40.08
w/ WECHSEL	99.70	43.35	45.33	56.61	54.34	32.53	32.41	14.80	16.20	61.70	62.89	52.01	52.72	43.50	43.98
w/ Focus	99.70	46.55	48.95	56.21	**55.78**	32.27	32.46	19.20	18.00	63.82	64.80	51.70	51.78	44.96	45.29
w/ ZeTT	418.94	47.14	49.03	57.06	53.70	34.06	34.06	18.40	19.40	64.15	65.34	52.09	51.22	45.48	45.46
w/ TokAlign	99.70	**54.46***	**56.86***	58.90*	52.26	36.16*	36.27*	**21.00***	**20.20***	**67.74***	**68.50***	52.25*	50.91	48.42	47.50
w/ SlimPajama	99.70	53.54	55.68	57.55	53.85	36.10	35.99	19.40	**20.20**	67.03	67.52	52.09	51.22	47.62	47.41
+ Align Rep.	99.70	54.25	56.65	**59.33**	54.68	**37.08**	**36.91**	20.20	19.40	67.36	68.17	**54.38**	**52.80**	**48.77**	**48.10**
$\text{Pythia}_{2.8B}$	–	63.80	67.00	63.91	65.14	45.32	45.04	24.00	25.20	74.05	74.43	58.64	60.77	54.95	56.26
w/ Rand. Init.	194.78	30.47	32.91	38.20	51.07	26.46	26.69	14.40	13.20	55.17	55.06	48.30	50.51	35.50	38.24
w/ Rand. Perm.	194.78	31.48	31.86	37.83	50.46	26.48	26.49	13.60	14.40	54.03	54.95	50.20	48.86	35.60	37.84
w/ OFA	194.78	50.13	54.12	60.89	61.47	36.39	36.88	18.00	19.00	65.18	64.80	54.06	54.85	47.44	48.52
w/ WECHSEL	194.78	52.48	54.92	59.42	56.76	36.79	37.30	19.20	20.80	64.04	64.25	56.43	55.72	48.06	48.29
w/ Focus	194.78	54.29	58.16	61.44	62.84	38.38	39.09	20.00	20.20	68.44	68.28	54.62	56.04	49.53	50.77
w/ ZeTT	855.96	57.15	59.42	61.68	62.05	42.17	42.25	21.80	23.60	71.11	71.16	56.59	59.19	51.75	52.95
w/ TokAlign	194.78	61.62*	65.15*	63.82*	65.47*	43.13*	43.18*	**23.40***	**25.80***	72.14*	72.42*	**58.17***	**61.17***	53.71	55.53
+ Align Rep.	194.78	**61.66**	**65.66**	**64.56**	**65.66**	**43.97**	**44.09**	22.40	25.00	**73.01**	**73.23**	58.09	60.54	**53.95**	**55.70**

Table 4: The main results of replacing the vocabulary of Pythia to Gemma. The best performance among the eight methods is displayed in **bold**. * indicates statistically significant improvements of 5% level. "+Align Rep." denotes the GloVe embeddings for tokens are converted into relative representations using 300 common tokens in both vocabularies before alignment following (Mosca et al., 2023).

translation domain, token-level distillation outperforms sentence-level distillation when using larger student models, simpler texts, and abundant decoding information (Kim and Rush, 2016; Wei et al., 2024). Given the same teacher model Qwen2_{7B}, the improvement of Pythia over the sentence-level distillation result reaches 4.4%. The performance of Pythia_{1B} is even comparable with the vanilla Pythia_{7B} after token-level distillation. It is also noted that the knowledge transfer between models will be constrained in sentence-level distilling without unifying vocabulary, which further demonstrates the importance of unifying tokenizers between models.

4.3 Vocabulary Adaptation Results

We show experimental results of replacing the Pythia vocabulary (50.3k) with the Gemma vocabulary (256.0k) using all methods in Table 4. Given the same amount of tokens to fine-tune, it can be found that TokenAlign performs better than other baseline methods. The average improvement of TokenAlign over the strong baseline method ZeTT reaches 2.4%, and 97.6% performance of the vanilla model is reserved after vocabulary replacement. ZeTT requires more computation to train a hypernetwork for the parameters prediction, e.g., 661.2 GPU hours for $\text{Pythia}_{2.8B}$, while our method only costs less than two hours on a CPU server with 128 cores to train GloVe embeddings and align tokens. Replace the corpus to train the GloVe embedding with 1B SlimPajama (Soboleva et al., 2023) tokens brings comparable results (the "w/ SlimPajama" row). It demonstrates the robustness of our method on the pre-training corpus for token embedding and alignment matrix. Following Moschella et al. (2023), we also evaluate the method that converts token representations into relative ones using 300 common tokens in both vocabularies as anchors before calculating the alignment matrix $M_{s \to t}$, which brings better performance.

4.4 Analysis

The loss curves of $\text{Pythia}_{2.8B}$ with different methods during the first 2.5k steps are shown in Figure 4. We find that TokAlign brings a better initialization and decreases the first-step training loss from 17.8 (Focus) to 9.5. Moreover, the training process with TokAlign is faster than other methods, which reaches 2.75 at the 1.3k step and is 1.92x (2.5/1.3) speed up than Focus.

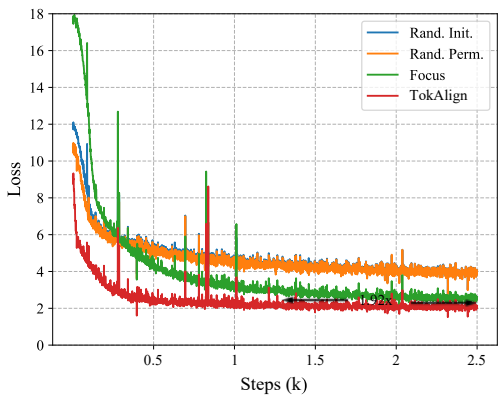
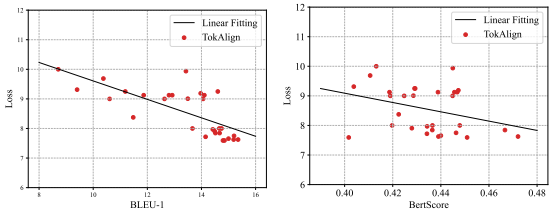


Figure 4: The training loss of $\text{Pythia}_{2.8B}$.

Model	$\#\mathcal{V}$ (k)	ARC-E		BoolQ		HellaSwag		OpenbookQA		PIQA		WinoGrande		Avg	
		0	5	0	5	0	5	0	5	0	5	0	5	0	5
Pythia$_{1B}$	50.3	56.82	58.71	60.43	57.37	37.68	37.66	18.80	19.00	70.40	71.49	53.20	52.01	49.55	49.37
→ Gemma	256.0	54.46	56.86	**58.90**	52.26	36.16	36.27	**21.00**	20.20	67.74	68.50	52.25	50.91	48.42	47.50
→ Qwen2	152.1	54.46	57.07	54.80	49.79	37.18	37.04	19.20	18.40	68.44	**70.24**	53.35	52.80	47.91	47.56
→ LLaMA2	32.0	49.45	52.02	58.32	55.75	35.38	35.45	18.80	17.80	66.32	66.65	53.91	50.91	47.03	46.43
→ LLaMA3	128.0	**54.63**	**57.28**	55.84	**53.70**	**37.34**	**37.43**	20.20	**20.40**	**69.04**	70.18	**54.46**	**53.43**	**48.59**	**48.74**
Pythia$_{2.8B}$	50.3	63.80	67.00	63.91	65.14	45.32	45.04	24.00	25.20	74.05	74.43	58.64	60.77	54.95	56.26
→ Gemma	256.0	61.62	65.15	63.82	**65.47**	43.13	43.18	23.40	**25.80**	72.14	72.42	58.17	**61.17**	53.71	**55.53**
→ Qwen2	152.1	**62.54**	**66.04**	62.35	63.55	44.46	44.39	23.20	24.60	**73.50**	**73.56**	**59.04**	59.59	54.18	55.29
→ LLaMA3	128.0	61.83	64.60	**64.40**	63.94	**44.62**	**44.59**	**23.80**	25.60	73.45	73.29	57.54	58.72	**54.27**	55.12
Pythia$_{6.9B}$	50.3	65.99	69.23	62.84	62.02	47.56	47.64	25.00	27.00	74.65	75.41	60.46	62.43	56.08	57.29
→ Gemma	256.0	65.40	68.35	62.39	59.57	45.75	45.86	22.00	25.60	73.39	74.10	60.38	61.17	54.89	55.77
→ Qwen2	152.1	65.57	**68.43**	**64.07**	57.61	46.84	46.91	**25.60**	25.40	73.45	74.65	61.17	63.14	56.12	56.02
→ LLaMA3	128.0	**66.46**	68.35	63.79	**60.64**	**47.28**	**47.31**	**25.60**	**28.20**	**74.48**	**75.84**	**61.48**	**63.30**	**56.52**	**57.27**

Table 5: The benchmark results of replacing different tokenizers using TokAlign. The overlapping ratio between the vocabulary of Pythia and other models are 6.23% (Gemma), 26.92% (Qwen2), 28.10% (LLaMA2), 32.85% (LLaMA3).



(a) BLEU-1($\mathcal{C}_t, \mathcal{C}_t^{'}$) (b) BERTScore($\mathcal{C}, \mathcal{C}^{'}$)

Figure 5: The relationship between initial training loss and BLEU-1 (a) or BERTScore (b) for Pythia$_{1B}$.

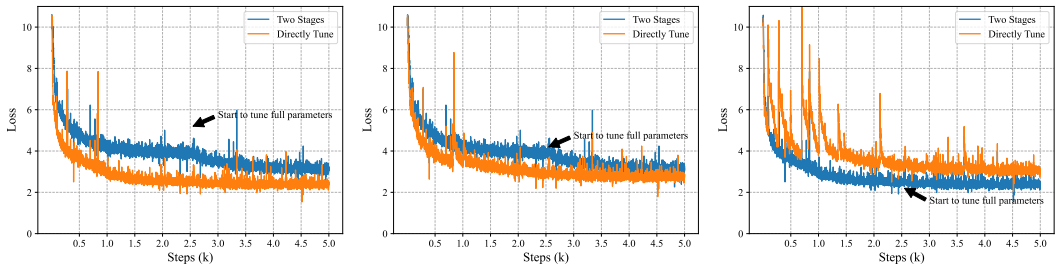
Better alignment brings better initialization. We further investigate the impact of the learned alignment matrix $M_{s\to t}$ by changing the hyper-parameters of GloVe. It is noted that different alignment matrices $M_{s\to t}$ bring different initial parameters, and also result in different BLEU-1 scores on the same evaluation corpus. Figure 5(a) illustrates the negative relationship between the first-step training loss and BLEU-1. The sentence embedding model named "all-mpnet-base-v2" (Song et al., 2020) is adopted in the BERTScore evaluation. As shown in Figure 5(b), it also shows a clear negative relationship with the initial training loss. In other words, the higher the BLEU-1 score or BERTScore for the alignment matrix $M_{s\to t}$, the better the initial parameter is.

More overlapping comes to faster convergence and higher performance. TokAlign is further applied to the other three target tokenizers: Qwen2, LLaMA2, and LLaMA3. Table 5 reports the performance of models after replacing vocabulary on six datasets. TokAlign recovers 98.0% performance of the base model on average with only 5k steps. Given a target vocabulary with more tokens than the one of Pythia (50.3k), it can be found that a higher overlapping ratio brings a better performance of model replaced (97.6% for Gemma to 99.1% for LLaMA3). The zero-shot in-context learning results for Pythia$_{6.9B}$ with LLaMA3 vocabulary even surpass the vanilla base model. The results of Pythia$_{1B}$ with LLaMA2 vocabulary are only 94.5%, which is inferior to the average result. We argue that it may come from the missing 75.0M parameters (7.4% for Pythia$_{1B}$) after switching to a 32.0k vocabulary from the 50.3k vocabulary.

Figure 9 in Appendix B.3 shows the training loss curve. The replacing process of the Gemma tokenizer is the slowest, which may come from the only 6.23% overlapping ratio between two vocabularies. It is in line with the result of random initialization in Figure 10. Appendix B.3 reports more quantitative results by shuffling the alignment matrix, which further demonstrates the importance of token alignment.

Two-stage tuning brings a more stable convergence. To replace the tokenizer and keep the performance of the vanilla model, we only fine-tune the vocabulary-related parameters at the first stage. The main reason for two-stage tuning is to take these parameters as the adapters of different tokenizers and avoid the well-trained parameters of the internal layer being distracted by the new initialized parameters.

Figure 6 illustrates that our two-stage tuning method makes the convergence more stable under a high learning rate like 6.4e^{-4}, which comes to better performance after vocabulary adaptation. It is noted that the loss spike also occurs at the first stage, fine-tuning vocabulary-related parameters only, under such a high learning rate like 2.56e^{-3} in Figure 7.



(a) Learning rate 2e^{-5} (b) Learning rate 8e^{-5} (c) Learning rate 6.4e^{-4}

Figure 6: The loss curve of Pythia$_{1B}$ under two-stage tuning or direct full parameters tuning.

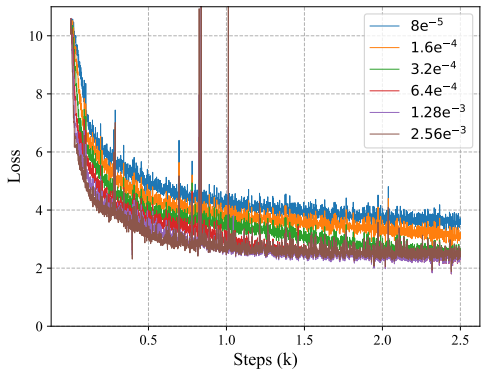


Figure 7: The training loss curve of Pythia$_{1B}$ for learning rate used during replacing to the Gemma tokenizer.

5 Conclusion and Future Work

In this paper, we introduce a method named TokAlign to replace the tokenizer of large language models from a token-token co-occurrence view. Extensive experiments demonstrate that TokAlign restores the performance of vanilla models after vocabulary adaptation, which enables cross-lingual knowledge transfer and deep knowledge transfer between models like token-level distillation.

Beyond replacing the vocabulary of large language models, our method can be extended to replace the vocabulary of multi-modal models by aligning different modal tokens. The other direction is to develop a faster method, e.g., incorporating meta-learning in the two-stage tuning method to speed up the convergence.

Limitations

The first limitation comes from the assumption that the pre-training data distribution is available. We conduct experiments on Pythia with different parameter amounts, which provide public model weights and pre-training corpus. Due to the limited computation resource budget, open-source language models with unknown pre-training corpus like Mistral (Jiang et al., 2023) are not investigated in this work. However, the pre-training corpus distribution of open-weighted large language models can be roughly inferred by the BPE vocabulary (Hayase et al., 2024). It can re-construct a similar pre-training corpus to conduct replacing tokenizer experiments.

Another limitation is the additional 5k steps for vocabulary adaptation to replace a tokenizer. From the loss curve of TokAlign (Figure 9), we find that the start of full parameters tuning can be faster, which may result in a better balance between performance and computational budget. Appendix B.4 reports a preliminary result with only 2k steps, where TokAlign also shows a promising result.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful discussions and valuable comments. The research work was supported by the National Key R&D Program of China (No. 2022ZD0160602) and the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA04080400).

References

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2024. Medvoc: vocabulary adaptation for fine-tuning pre-trained language models on medical text summarization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pages 16344–16359. Curran Associates, Inc.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

C.m. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. 2023. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 268–281, Singapore. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, and et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Jonathan Hayase, Alisa Liu, Yejin Choi, Sewoong Oh, and Noah A Smith. 2024. Data mixture inference: What do bpe tokenizers reveal about their training data? *arXiv preprint arXiv:2407.16607*.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Denis Kocetkov, Raymond Li, Loubna Ben allal, Jia LI, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro Von Werra, and Harm de Vries. 2023. The stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.

Haoran Li, Jiajun Zhang, and Chengqing Zong. 2017. Implicit discourse relation recognition for english and chinese with multiview modeling and effective representation learning. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(3):1–21.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *Preprint*, arXiv:2407.06089.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Qwen blog*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2018. Mixed precision training. In *International Conference on Learning Representations*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of

subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. 2024. Zero-shot tokenizer transfer. *Preprint*, arXiv:2405.07883.

Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. 2023. Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the LLM era. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 190–207, Toronto, Canada. Association for Computational Linguistics.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Nandini Mundra, Aditya Nanda Kishore Khandavally, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M Khapra. 2024. An empirical comparison of vocabulary expansion and initialization approaches for language models. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 84–104, Miami, FL, USA. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

BigScience Workshop: Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, and Alexandra Sasha Luccioni et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Alexey Tikhonov and Max Ryabinin. 2021. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3534–3546, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ke Tran. 2020. From english to foreign languages: Transferring pre-trained language models. *arXiv preprint arXiv:2002.07306*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2018. Learning multimodal word representation via dynamic fusion methods. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5973–5980. AAAI Press.

Jingxuan Wei, Linzhuang Sun, Yichong Leng, Xu Tan, Bihui Yu, and Ruifeng Guo. 2024. Sentence-level or token-level? A comprehensive study on knowledge distillation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 6531–6540. ijcai.org.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. 2023. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Yangyifan Xu, Jinliang Lu, and Jiajun Zhang. 2024. Bridging the gap between different vocabularies for LLM ensemble. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7140–7152, Mexico City, Mexico. Association for Computational Linguistics.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024. An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6760–6785, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Hyper-parameters

GloVe Training We empirically train GloVe vectors with 1B tokens, which covers most tokens from Gemma (95.10%), Qwen2 (93.40%), LLaMA2 (99.35%), and LLaMA3 (98.04%). The dimension size is set to 300. The max training iteration and the size of the slide window are 15.

Model Tuning The optimizer adopted in this work is AdamW (Loshchilov and Hutter, 2019), where $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate for baseline methods is set to 5e-5 to reduce the loss spike in Figure 6(b) and Figure 6(c). We adopt bf16 mixed precision training, ZeRO-1, and flash-attention to save GPU memory cost and speed up the training process (Micikevicius et al., 2018; Rasley et al., 2020; Dao et al., 2022). Following Biderman et al. (2023), the batch size is set to 2M tokens and the max sequence length is 2048.

B Additional Results

B.1 Tokenizer Compression Rate

Table 6 reports detailed compression rates of tokenizers across different domains and languages. We randomly sample 10 subsets or languages from vanilla datasets (Azerbayev et al., 2024; Kocetkov et al., 2023) to estimate the compression rate. Following Lai et al. (2023), the division of languages between "**High**", "**Medium**" and "**Low**" is determined by the available amount resource on CommonCrawl.

B.2 GloVe Vectors

We show the effects of different token amounts for the GloVe vectors training in Figure 8. It can be found that 1B tokens used in this work provide a high vocabulary coverage (>90%) and better initialization for Pythia_{1B}. Due to the limited computation budget, experiments with more than 1B tokens are not conducted.

B.3 Convergence Analysis

To investigate the effect of overlapping rate between two tokenizers to the convergence of training, we plot Figure 10 for the random initialization baseline method. The convergence of Gemma tokenizer is slower than the other tokenizers and comes to worse results, which are similar to the case in Figure 9.

Moreover, we randomly shuffle the alignment matrix learned in TokAlign to imitate the case that other worse methods rather than cosine similarity to calculate the alignment matrix. Figure 11 shows that the higher percentage of randomly shuffle comes to higher initial training loss and slower convergence.

B.4 Fast Vocabulary Adaptation Results

We further investigate a challenge condition that fine-tunes only 2B tokens to adapt the target vocabulary. To meet the requirement, we reduce the batch size to 1M tokens and set the number of fine-tuning steps to 2k. Table 7 shows the results of adapting to the other 3 tokenizers using TokAlign. It can be found that 95.66% performance of the vanilla model is recovered on average, which further demonstrates the effectiveness of our method.

B.5 In-context Learning Results during Cross-lingual Transfer

Table 2 and 8 report the 0-shot and 5-shot in-context learning results on 4 multilingual datasets. The average improvement over the baseline method Focus is 2.35% after language adaptation pre-training. We can find that the model initialized by TokAlign is comparable to the one of Focus after language adaptation pre-training, which mainly comes from the strong English performance preserved by TokAlign.

Case study of multilingual token alignment. Table 9 provides nine new tokens from three languages with their top 3 tokens in the source vocabulary for qualitative analyses. In most cases, a clear semantic relationship between two aligned tokens cannot be found. We argue that it may come from the following two reasons:

- BPE algorithm (Sennrich et al., 2016) divides words into the sub-word units, also called tokens, from the statistical co-occurrence information. There may be less superficial semantic information in the tokens divided compared with words in the natural language.

Domain	Subset / Language	Tokenizer				
		Gemma	LLaMA3	LLaMA2	Qwen2	Pythia
Math (Azerbayev et al., 2024)	*ArXiv*	2.8561	2.7765	2.7040	2.7445	2.8489
	Textbooks	4.0883	4.3270	3.6500	4.2899	3.9464
	Wikipedia	3.1753	3.2049	2.8792	3.0312	3.2898
	ProofWiki	2.7538	2.8115	2.5996	2.7900	2.7363
	StackExchange	3.2062	3.2814	3.0094	3.2107	3.2222
	WebPages	3.9885	4.0655	3.5070	3.8720	4.1136
Code (Kocetkov et al., 2023)	*Python*	3.3401	4.1331	3.0072	4.0339	3.2328
	Java	3.7175	4.4900	3.2193	4.4141	3.4914
	Go	2.9274	3.4797	2.5189	3.3870	2.8542
	VHDL	2.1038	2.4814	1.8724	2.2961	2.1395
	ActionScript	3.3470	3.9717	2.7852	3.9180	3.2949
	Scheme	2.7178	3.3045	2.4586	2.9713	2.9326
	Haml	3.2423	3.8429	2.9588	3.8002	3.1016
	Xbase	2.8739	3.4325	2.3300	3.3475	2.7837
	Mako	3.4387	4.0746	3.1238	4.0311	3.2844
	EmberScript	1.4104	1.9017	1.3819	1.4082	2.1540
High-Langs (Nguyen et al., 2023)	*English*	4.4971	4.6042	3.8647	4.4875	4.4505
	Russian	6.7529	5.8131	4.9275	5.3559	3.5802
	Spanish	4.6068	3.8416	3.4517	3.8330	3.3655
	German	4.4605	3.6314	3.4417	3.6041	3.1096
	French	4.2258	3.7378	3.4445	3.7243	3.3565
	Chinese	3.7378	3.2373	1.8434	3.9859	1.9896
	Italian	4.2211	3.4952	3.3320	3.4573	3.1928
	Portuguese	4.2731	3.6030	3.2031	3.5850	3.2022
	Polish	3.5583	2.8548	2.6639	2.9464	2.4333
	Japanese	5.7640	4.2796	2.4701	4.7059	2.9326
Medium-Langs (Nguyen et al., 2023)	*Czech*	3.3402	3.2875	2.5978	2.4490	2.3884
	Vietnamese	4.5376	4.2766	1.9699	4.2877	2.0382
	Persian	5.6465	5.3015	1.7938	3.1923	2.3707
	Hungarian	3.2337	2.6008	2.6311	2.5500	2.3878
	Greek	4.4691	4.5671	1.8544	2.1225	3.0283
	Romanian	3.5558	3.0566	2.8355	3.0083	2.8981
	Swedish	3.7087	3.1398	2.9214	3.0977	2.9620
	Ukrainian	5.5141	5.5985	4.5904	3.6179	3.0702
	Finnish	3.2659	2.6748	2.4176	2.6473	2.6112
	Korean	3.3556	3.6957	1.5977	3.3330	1.5667
Low-Langs (Nguyen et al., 2023)	*Hebrew*	4.0487	1.8592	1.7875	4.3773	2.0380
	Serbian	4.8596	3.9234	4.2642	3.6267	2.9896
	Tamil	5.6161	2.0279	2.2615	2.4759	1.9765
	Albanian	2.8919	2.6536	2.2945	2.6037	2.3631
	Azerbaijani	2.8585	2.4857	2.0407	2.3797	2.1534
	Kazakh	3.8172	2.9176	3.0869	2.9263	2.3236
	Urdu	4.4364	2.8462	1.7260	2.7174	1.9458
	Georgian	3.8237	1.4828	2.5595	2.6951	2.2077
	Armenian	3.2133	1.1658	1.7000	1.8531	1.3922
	Icelandic	2.7964	2.4860	2.3050	2.4330	2.3185

Table 6: The compression rates (bytes/token) of different tokenizers.

Model	$\#\mathcal{V}$ (k)	ARC-E		BoolQ		HellaSwag		OpenbookQA		PIQA		WinoGrande		Avg	
		0	5	0	5	0	5	0	5	0	5	0	5	0	5
Pythia_{1B}	50.3	56.82	58.71	60.43	57.37	37.68	37.66	18.80	19.00	70.40	71.49	53.20	52.01	49.55	49.37
→ Gemma	256.0	51.09	52.44	53.12	52.35	35.00	35.05	20.20	18.60	64.80	65.83	53.12	51.62	46.22	45.98
→ Qwen2	152.1	53.41	55.47	53.52	55.81	36.12	36.38	20.80	18.00	68.50	68.88	54.38	52.80	47.79	47.89
→ LLaMA3	128.0	51.73	55.09	59.05	55.08	36.42	36.52	19.40	19.60	67.68	68.34	53.43	53.75	47.95	48.06

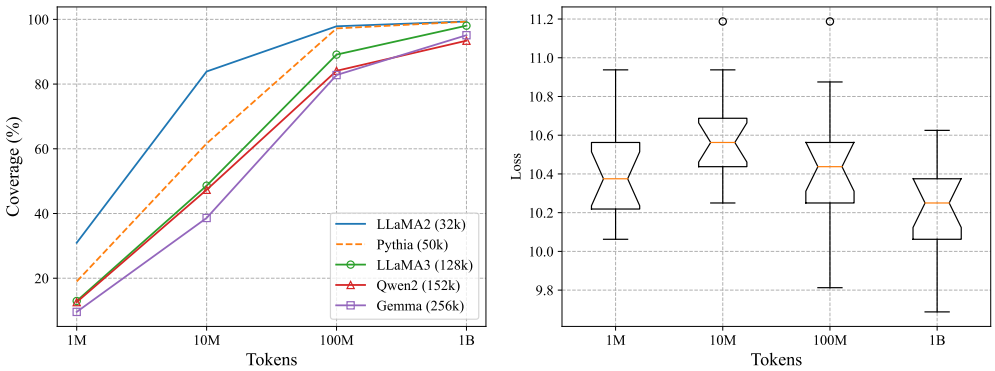
Table 7: The main results of replacing the vocabulary of Pythia for TokAlign using 2B tokens from the Pile corpus.

Model	XNLI							PAWS-X					XCOPA			XStoryCloze				Avg
	en	de	zh	ar	th	vi	ur	de	en	ja	ko	zh	th	vi	ta	en	zh	ar	te	
Pythia_{1B}	46.2	38.6	**38.9**	**36.9**	35.2	**38.9**	34.9	48.9	48.3	52.9	**53.3**	54.1	53.4	52.6	55.4	**65.3**	48.6	48.2	52.2	**47.5**
w/ Focus Init.	32.8	32.2	33.6	33.6	33.5	32.0	32.8	44.8	46.0	48.9	44.8	44.7	51.4	47.6	**55.6**	45.9	48.6	**48.5**	46.8	42.3
+ LAT	**47.0**	36.7	35.4	34.3	33.5	35.1	33.9	51.5	48.6	53.7	51.2	54.0	**54.4**	51.6	**55.6**	55.8	48.7	47.5	50.4	46.3
w/ TokAlign Init.	44.9	37.4	34.0	32.8	**35.3**	35.2	34.5	50.2	**50.3**	52.0	53.1	**54.4**	**54.4**	50.0	54.4	61.2	48.3	47.6	50.0	46.3
+ LAT	44.4	**39.0**	38.7	35.6	35.1	37.8	**35.5**	**51.9**	49.3	**54.7**	53.1	50.6	54.2	**54.0**	52.8	64.7	**50.8**	48.0	**52.4**	**47.5**
$\text{Pythia}_{6.9B}$	**53.0**	40.7	**41.7**	**38.9**	37.3	41.3	35.1	49.4	47.1	52.9	52.2	52.4	**55.0**	53.6	53.6	**73.1**	**54.6**	**49.9**	**53.9**	49.2
w/ Focus Init.	31.5	31.3	33.0	32.6	33.4	32.2	32.6	44.8	46.4	52.3	51.2	54.5	52.4	47.4	**56.0**	44.9	47.3	48.5	47.6	43.1
+ LAT	45.1	37.7	35.3	33.4	35.0	38.1	33.8	49.5	49.0	52.6	54.5	55.3	52.0	51.2	53.8	61.5	48.3	47.3	53.4	46.7
w/ TokAlign Init.	50.8	39.1	34.4	34.5	33.9	34.6	**35.2**	50.0	47.7	**53.9**	54.3	55.2	53.2	51.2	53.2	68.0	48.5	47.8	50.2	47.1
+ LAT	49.2	**41.5**	37.8	36.9	**38.7**	**41.9**	34.7	**51.2**	**49.5**	53.5	**54.8**	**55.4**	53.4	**59.8**	52.8	73.0	53.9	49.2	53.6	**49.5**

Table 8: Five-shot in-context learning results of cross-lingual transfer.

Top-3	French			Chinese			Korean		
	dire(speak)	aller(go)	oui(are)	吃(eat)	科学(science)	智能(intelligence)	능(competence)	집(house)	왜(why)
Qwen2 (Target Tokenizer)									
1	ada	Ġsta	Ġsalv	allel	Ġantagon	_{[	Si	ĠBart	bst
2	ays	ĠÃ¨	Ġvas	Ġindicator	Ġign	liquid	uria	ĠPAT	rains
3	Ġ-	Ġdetermin	Ġexplos	Ġbasic	Ġcritic	Layer	ost	ĠEdgar	irc
Gemma (Target Tokenizer)									
1	Ġj	Cor	Tools	kernel	ĠLed	Ġcommittee	Ġmang	Ġcru	Ġcholesterol
2	Ġdar	Ġequality	directed	sentence	COUNT	ĠUND	ial	Ġcal	Ġmolecule
3	ba	Lex	afx	messages	Ġglycine	Ġfactors	Ġrebut	Ġmalt	apor

Table 9: The case study of new tokens from other languages in the target vocabulary with top-3 source tokens aligned. The language family of French, Chinese, and Korean are Indo-European, Sino-Tibetan, and Koreanic, respectively.



(a) Vocabulary coverage

(b) Initial loss with Gemma tokenizer

Figure 8: The average vocabulary coverage (a) and initial training loss of Pythia_{1B} (b) under different amount tokens to train the GloVe vector.

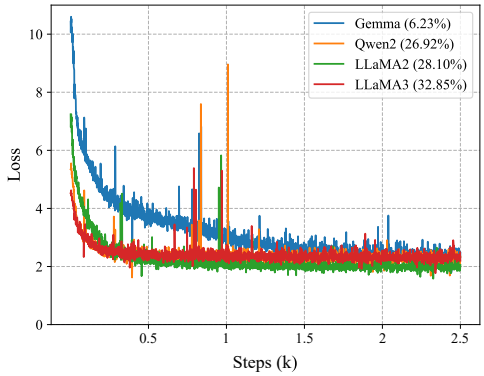


Figure 9: The training loss curve of Pythia_{1B} for different overlapping ratios.

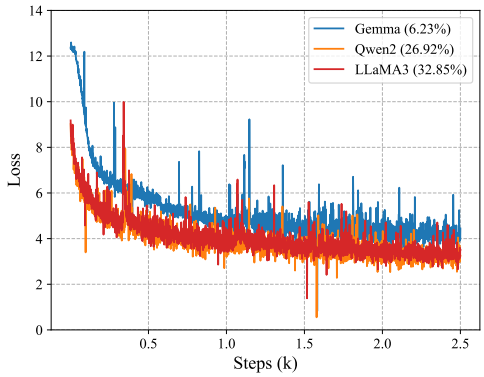


Figure 10: The training loss to different tokenizers using random initialization baseline.

- The GloVe vector for each token is obtained from the token-token co-occurrence information. These aligned tokens often appear together, e.g., 科学(science) and "Ġcritic", 왜(why) and "rains".

Therefore, it is better to choose a matric to quantify the performance of the alignment matrix learned, for example, the BLEU-1 score or BERTScore in Section 3.2.

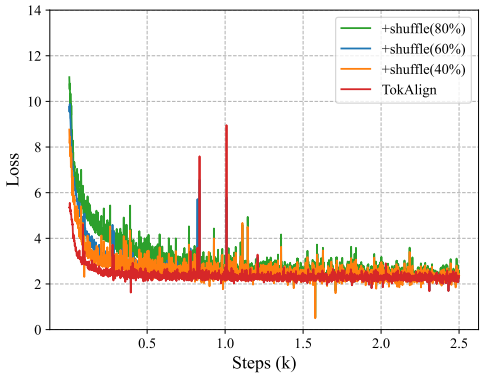


Figure 11: The training loss of Pythia_{1B} when replacing tokenizer to Qwen2 under different percentages of shuffling.

C Evaluation Tasks

We report the statistics of evaluation tasks used in Table 10. Here are the descriptions of these evaluation tasks:

Natural Language Inference aims to determine the semantic relationship (Entailment, neural, or contradiction) between the premise and hypothesis (Conneau et al., 2018).

Paraphrase Detection requires the model to evaluate whether the second sentence is a paraphrase of the first sentence in this task (Yang et al., 2019).

Commonsense Reasoning is a task for the model to reason the gold answer based on the semantic coherence and physic rules (Clark et al., 2018; Mihaylov et al., 2018; Zellers et al., 2019; Ponti et al., 2020; Bisk et al., 2020; Sakaguchi et al., 2020; Tikhonov and Ryabinin, 2021).

Task	Dataset	#Lang	#Class	Data Curation	#Train	#Dev	#Test
Natural Language Inference	XNLI	15	3	Translation	–	2,490	5,010
Paraphrase Detection	PAWS-X	7	2	Aligned	–	2,000	2,000
Reasoning	ARC-Easy	1	4	–	2,251	570	2,376
	HellaSwag	1	4	–	39,905	10,042	10,003
	OpenbookQA	1	4	–	4,957	500	500
	PIQA	1	2	–	16,000	2,000	3,000
	XCOPA	12	2	Translation	33,810	100	500
	XStoryCloze	11	2	Translation	361	–	1,511
	WinoGrad	1	2	–	40,398	1,267	1,767
Reading Comprehension	BoolQ	1	2	–	9,427	3,270	–

Table 10: Statistic of evaluation datasets used.

Reading Comprehension needs the model to infer whether the given passage can answer the query (Clark et al., 2019).

D Language Codes

We provide details of languages involved in Table 11. Following Lai et al. (2023), languages are divided by the data ratios in CommomCrawl: High (>1%), Medium (>0.1%), and Low (>0.01%).

ISO 639-1	Language	Family
AR	Arabic	Afro-Asiatic
BN	Bengali	Indo-European
DE	German	Indo-European
EN	English	Indo-European
JA	Japanese	Japonic
KO	Korean	Koreanic
TA	Tamil	Dravidian
TE	Telugu	Dravidian
TH	Thai	Kra-Dai
UR	Urdu	Indo-European
VI	Vietnamese	Austroasiatic
ZH	Chinese	Sino-Tibetan

Table 11: Details of language codes in this work.

E Licenses of Scientific Artifacts

We follow and report the licenses of scientific artifacts involved in Table 12.

Name	License
Transformers	Apache 2.0 license
lm-evaluation-harness	MIT license
matplotlib	PSF license
Focus	MIT license
WECHSEL	MIT license
Pythia	Apache 2.0 license
LLaMA3	Meta LLaMA 3 community license
Qwen2	Tongyi Qianwen license
Gemma	Gemma license
The Pile	MIT license

Table 12: Licenses of scientific artifacts involved in this work.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 刘硕

方向： 人工智能

单位： 中国科学院大学前沿交叉科学学院

题目： 基于多模态大模型的时空轨迹数据挖掘

Traj-MLLM: Can Multimodal Large Language Models Reform Trajectory Data Mining?

Shou Liu
liushuo22s@ict.ac.cn
University of Chinese Academy of Sciences
China

Di Yao†
yaodi@ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences
China

Yan Lin
lyan@cs.aau.dk
Department of Computer Science, Aalborg University
Denmark

Gao Cong
gaocong@ntu.edu.sg
College of Computing and Data Science, Nanyang Technological University
Singapore

Jingping Bi
bjp@ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences
China

arXiv:2509.00053v1 [cs.MM] 25 Aug 2025

Abstract

Building a general model capable of analyzing human trajectories across different geographic regions and different tasks becomes an emergent yet important problem for various applications. However, existing works suffer from the generalization problem, *i.e.*, they are either restricted to train for specific regions or only suitable for a few tasks. Given the recent advances of multimodal large language models (MLLMs), we raise the question: can MLLMs reform current trajectory data mining and solve the problem? Nevertheless, due to the modality gap of trajectory, how to generate task-independent multimodal trajectory representations and how to adapt flexibly to different tasks remain the foundational challenges. In this paper, we propose Traj-MLLM, which is the first general framework using MLLMs for trajectory data mining. By integrating multiview contexts, Traj-MLLM transforms raw trajectories into interleaved image-text sequences while preserving key spatial-temporal characteristics, and directly utilizes the reasoning ability of MLLMs for trajectory analysis. Additionally, a prompt optimization method is proposed to finalize data-invariant prompts for task adaptation. Extensive experiments on four publicly available datasets show that Traj-MLLM outperforms state-of-the-art baselines by 48.05%, 15.52%, 51.52%, 1.83% on travel time estimation, mobility prediction, anomaly detection and transportation mode identification, respectively. Traj-MLLM achieves these superior performances without requiring any training data or fine-tuning the MLLM backbones.

Keywords

Trajectory Data Mining, Multimodal Large Language Models

1 Introduction

Technical advances in GPS positioning enable the collection of world-wide human trajectory data derived from mobile devices [44], ride-hailing services [27], *etc.*. Trajectory data mining has become an increasingly important research topic, providing tools for numerous tasks such as travel time estimation [10, 34, 47], trajectory prediction [6, 47], anomaly detection [9, 23, 29], transportation mode identification [3, 11, 19], *etc.*. However, most works [6, 9, 23, 29] are tailored to specific tasks and regions hindering their usage in real-world scenarios. Building a unified model capable of generalizing across different geographic regions and trajectory data mining tasks remains an emergent yet important problem.

In recent years, many efforts have been made to address this problem. These works can be categorized into two groups, *i.e.* trajectory foundation models and large language model (LLM)-based models. Methods in the first group [10, 31, 34] pre-train general trajectory representations on large-scale dataset and employ several learnable task heads for task adaptation. Due to the geographical heterogeneity, the learned representation models are only effective in trained regions and hard to extend to other data-scarce regions. For LLM-based models [30, 37, 41, 47], researchers attempt to utilize the generalization ability of LLMs for trajectory data mining. The LLMs are either treated as well-trained encoders directly feeding raw trajectories to obtain the task results [41] or fine-tuned to enhance traditional trajectory foundation models [30, 37, 47]. Nevertheless, LLMs struggle with handling numerical data and performing arithmetic operations[17], which limits their performances to encode raw trajectories. The commonly fine-tuned LLM is GPT-2[37, 47] which has only 1.5B parameters and exhibits limited generalization capability compared to modern, larger-scale models. Therefore, the generalization problem of trajectory data mining remains unsolved.

Notably, multimodal large language models (MLLMs) have shown remarkable reasoning ability, excelling in integrating multimodal information to enhance the performance of various tasks. The recent advances of MLLMs raise the question: can MLLMs reform current trajectory data mining and solve the generalization problem? Raw trajectories, recording as the sequences of GPS coordinates, are the simplified spatial-temporal expression of human movements. By projecting trajectories on the map, visual contextual information, such as POIs, road networks and traffic lights, can be represented in the visual modality. MLLMs' natural language processing (NLP) capability allows us to capture the regional information, human knowledge and task description with text. Moreover, well-trained MLLMs integrate rich common-sense knowledge and have strong generalization ability across different scenarios. It has significant

† Corresponding authors.

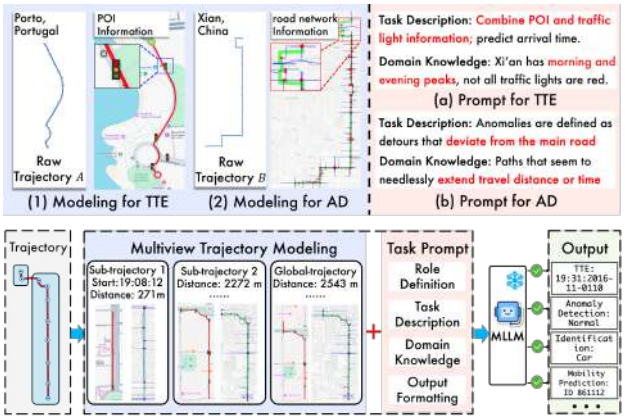


Figure 1: The motivation of `Traj-MLLM`.

potential to understand trajectories from visual and textual modalities without extra training and extend to analyze trajectories in unseen regions.

However, due to the modality gap of trajectory, two foundational challenges should be addressed to achieve region-agnostic and task-adaptive trajectory data mining with MLLMs:

- **Task-independent Trajectory Modeling.** The spatial-temporal characteristics vary a lot in trajectories and the required contextual information is heterogeneous for different tasks, leading to the challenge for modeling cross-region trajectories in a task-independent way. As shown in Figure 1, trajectories A and B are in different regions and A is much shorter than B. The information required for travel time estimation(TTE) is also different from anomaly detection(AD). Directly compressing all factors in one image would obscure real useful information leading to suboptimal performance. The trajectory modeling should not only capture the heterogeneity of all spatial-temporal and contextual information, but also extract region-agnostic features useful across different geographical regions.
- **Flexible Task Adaptation.** Instead of using trainable task heads for adaptation [47], we argue that textual prompts are more effective to activate the multimodal reasoning ability of MLLMs. The prompts can easily integrate human knowledge and define the reasoning steps for tasks clearly. As illustrated in Figure 1, the task descriptions, reasoning steps and output formats of TTE and AD are totally different. Nonetheless, these prompts should be easy to reference the information contained in trajectory modeling and data-invariant to meet the requirements of various tasks across different trajectory datasets. Therefore, it is challenging to construct task-specified prompts and achieve flexible task adaptation.

To address these challenges, we propose `Traj-MLLM` which is the first general framework that employs MLLMs for trajectory data mining. For unified trajectory modeling, we design a map-anchored tokenization mechanism which segments trajectories into sub-trajectories based on semantic completeness on the map. Both the visual and text representations are generated on the granularity of sub-trajectories. Taking these sub-trajectories as inputs, `Traj-MLLM` constructs different views of spatial scales and contextual factors that contain all required information for different tasks. For example, the trajectory A in Figure 1 is described at global and sub-trajectory scales. The POIs and road networks information are modeled as the images and texts for each scale. Utilizing the multimodal reasoning ability of MLLMs, the temporal constraints are captured with the interleaved image-text sequences of visual and text representations. To achieve flexible task adaptation, we proposed a prompt optimization method that employs few seed trajectories and performs multi-round interactions with MLLMs to finalize the data-invariant task prompts. In usage, an arbitrary trajectory can be transformed into multimodal inputs of MLLMs to generate the analyzing results of various tasks.

By doing so, `Traj-MLLM` has three attractive benefits: (1) **Training-free.** `Traj-MLLM` relies on the reasoning ability of MLLMs to conduct trajectory data mining and does not need any data for model training or fine-tuning. This allows `Traj-MLLM` to be suitable for handling trajectories in arbitrary regions. (2) **Extensible.** Both the trajectory modeling and prompt optimization are general and independent, which makes `Traj-MLLM` extensible to new tasks by involving other information views and task prompts. (3) **Interpretable.** As the by-products, `Traj-MLLM` also outputs the reasoning process, which can be used to interpret the results.

The key contributions of this paper are summarized as follows:

- We propose a novel framework, namely `Traj-MLLM`, which employs MLLMs for region-agnostic and task-adaptive trajectory data mining. To the best of our knowledge, this is the first work to employ the reasoning ability of MLLMs to analyze human movement trajectories.
- To model trajectories in a unified way, we design a map-anchored mechanism to segment trajectories into sub-trajectories, and represent them in visual and text modalities, considering multiple views. For adapting to different tasks, a prompt optimization method is proposed to finalize data-invariant task prompts.
- Extensive experiments on four datasets demonstrate `Traj-MLLM` achieves significant performance improvements, *i.e.* 48.05%, 15.52%, 51.52%, 1.83% on four representative tasks, compared with state-of-the-art baselines. These results are achieved without any training.
- We release a dataset that contains the multimodal responses of four trajectory mining tasks on seven latest MLLM backbones. This dataset can shed light on future trajectory data mining research, such as fine-tuning a new MLLM specially used for trajectory data.

2 Related Works

Trajectory Data Mining. Existing research in trajectory data mining can be broadly divided into three categories. Early approaches focus on task-specific models designed for solving one specific task, such as travel time estimation [10, 31, 34, 47], trajectory prediction [6, 47], anomaly detection [9, 23, 29], or classification [3, 11, 19]. These works embed assumptions in model design, limiting their generalization ability for other tasks. To address this, trajectory foundation models [10, 18, 33–36, 39, 42, 45] are proposed to learn a unified model for various tasks. However, large-scale trajectories are required to pretrain these foundation models, leading to suboptimal performance on trajectories of unseen regions. Witnessing the progress of LLMs, several works try to solve the generalization problem with LLMs. These works treat LLMs as encoders [30, 37, 47],

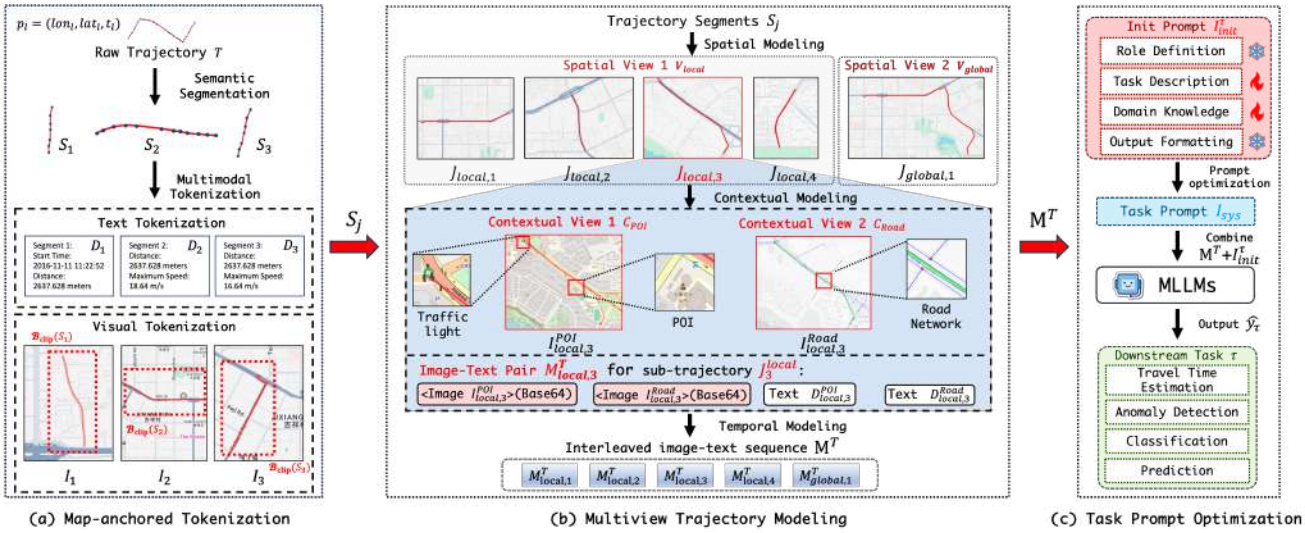


Figure 2: Overview of Traj-MLLM. Instead of direct model training, Traj-MLLM transforms raw trajectories into interleaved image-text sequences through map-anchored tokenization and multiview trajectory modeling. This enables us to reframe downstream tasks as multimodal reasoning problems, which can be directly solved by MLLMs in a training-free manner.

enhancer [41] and agents [5] to boost trajectory data mining. Nevertheless, the used LLMs are language models failing to handle image data, which is suitable for modeling the spatial and contextual information of trajectory. These researches inspire us to propose Traj-MLLM which is the first general framework utilizing MLLMs for trajectory data mining.

Multimodal Large Language Models. MLLMs have become a hot topic of recent AI research. Numerous works are proposed to either align visual modality to text[14, 15, 25] or enhance the multimodal reasoning ability[4, 12, 43]. The recent advances of MLLMs raise many geographical applications[13, 16, 28, 32, 40]. For example, MLLMs are now capable of inferring the geographic location of images, such as those shared on social media or captured from street views, based solely on visual content.[16, 28] However, most of current works are designed for analyzing static images, *e.g.* remote sensing pictures. Trajectory contains rich spatial-temporal and contextual information. Thus, current works cannot be directly used for trajectory data mining.

3 Preliminaries

In this section, we first define the problem of MLLM-based trajectory data mining, followed by a brief overview of Traj-MLLM.

Problem Definition: Assuming that a raw trajectory T consists of a sequence of points, $T = \{p_1, \ldots, p_l, \ldots, p_L\}$, where each point p_l consists of a GPS coordinate pair and a timestamp, *i.e.*, $(\text{lon}_l, \text{lat}_l, t_l)$. Trajectory data mining tasks are denoted as a set $\mathcal{T}$, where $\tau \in \mathcal{T}$ is one specific task, such as travel time estimation, anomaly detection, or trajectory prediction. The problem of MLLMs-based trajectory data mining can be formally defined as follows:

Taking raw trajectory T and a task τ as inputs, there is a multimodal mapping function $\mathcal{M}$ that first transforms T into visual and text modalities, denoted as $\mathbf{I}^T$ and $\mathbf{D}^T$, and combines them to obtain the interleaved image-text sequence $\mathbf{M}^T$. There is another prompt construction function $\mathcal{P}$ that generates the task prompt P^τ for task τ. Note that $\mathcal{M}$ and $\mathcal{P}$ are independent of each other, which makes the problem general for different geographical regions and tasks. The goal of MLLMs-based trajectory data mining is to generate the analysis result $\hat{y}_\tau = \text{MLLM}(\mathbf{M}^T, P^\tau)$.

Overview of Traj-MLLM: As illustrated in Figure 2, Traj-MLLM consists of three key modules, *i.e.*, map-anchored tokenization, multiview trajectory modeling, and task prompt optimization. The first two modules are designed to build the multimodal mapping function $\mathcal{M}$, and the third module is proposed for $\mathcal{P}$. To bridge the modality gap of trajectory data, raw trajectory T is segmented into sub-trajectories by map-anchored tokenization which encourages records having complete semantic information in one sub-trajectory. To model all the information required for different tasks, spatial, contextual, and temporal views of these sub-trajectories are constructed in the multiview trajectory modeling module to obtain an interleaved image-text sequence $\mathbf{M}^T$ of T. Finally, the task prompt optimization module is designed to obtain $\mathcal{P}$ and achieve flexible task adaptation.

4 Methodology

In this section, we elaborate the three key modules in Traj-MLLM, *i.e.*, map-anchored tokenization, multiview trajectory modeling and task prompt optimization in Section 4.1, 4.2 and 4.3, respectively. As the whole method is formulated for one trajectory, we omit the subscript T in the notations of following sections.

4.1 Map-Anchored Tokenization

Raw trajectory data consists of numerical coordinates and timestamps, lacking clear semantic structure, which makes it difficult to understand by MLLMs directly. Therefore, this module decomposes trajectories into semantic coherent sub-trajectories, and transform them into visual and text tokens.

4.1.1 Semantic Segmentation. A trajectory may contain semantically distinct phases, such as high-speed driving on highways, low-speed urban navigation, or stationary periods at intersections. Instead of feeding raw trajectory, sub-trajectories with coherent semantics can better capture fine-grained details and unleash the reasoning power of MLLMs. Therefore, `Traj-MLLM` conducts semantic segmentation at first and formulate it as an optimization problem.

Given a raw trajectory T, semantic segmentation aims to find a globally optimal segmentation $\mathbf{S} = \{S_1, \ldots, S_n, \ldots, S_N\}$ by minimizing the sum of segment costs. We first project T on the map and define a cost function $\text{cost}(a, b)$ containing E map-anchored factors to quantify the semantic consistency of sub-trajectory from p_a to p_b:

$$\text{cost}(a, b) = \sum_{e=1}^{E} f_e(a, b) \tag{1}$$

In this paper, we utilize three factors, *i.e.* f_{speed}, f_{road} and f_{len} that capture motion, routing, and structural properties respectively. $f_{speed}(a, b)$ measures speed consistency that encourages sub-trajectory with stable motion, such as cruising at a constant speed or being stationary. $f_{road}(a, b)$ represents the number of road type changes, which prevents a sub-trajectory having different road types, $f_{len}(a, b)$ is a length regularization term that penalizes breaking the trajectory into too many short sub-trajectories. More details can be found in Appendix A.1.

To solve the optimization problem, we employ dynamic programming. Let $\text{DP}[h]$ denote the minimum cumulative cost for an optimal segmentation of the prefix trajectory $\{p_1, \ldots, p_h\}$. The recurrence relation is defined by:

$$\text{DP}[h] = \min_{1 \le d < h} \left(\text{DP}[d] + \text{cost}(d + 1, h)\right) \tag{2}$$

Here, h denotes the current endpoint of the subtrajectory for which the segmentation cost is being computed, and d indicates the candidate previous segmentation point. By computing $\text{DP}[h]$, we obtain the globally minimal cost for the entire trajectory. Then, by backtracking the DP table, the specific segment boundary $\{S_1, \ldots, S_N\}$ that produces the optimal cost can be recovered. This approach effectively balances computational efficiency and segmentation quality, making it suitable for large-scale trajectory data analysis.

4.1.2 Multimodal Tokenization. After segmenting the trajectory into sub-trajectories, we convert each sub-trajectory $S_n \in \mathbf{S}$ into multimodal tokens suitable for MLLM inference. Specifically, we generate two complementary representations: structured text tokens (D_n) that summarizes meta features and basic statistics; map-projected visual tokens (I_n) that combine spatial and contextual information. Here, we only provide a general method to generate these tokens. More details to obtain the interleaved image-text sequence of T are specified in Section 4.2.

Text Tokenization. For each sub-trajectory S_n, multiple statistical features are extracted from raw trajectory to generate a structured text description D_n.

$$D_n = \text{Concat}\big(\text{Time}(S_n), \text{Dist}(S_n), \text{Speed}(S_n), \ldots\big) \tag{3}$$

The components here include $\text{Time}(\cdot)$ for the start and end timestamps, $\text{Dist}(\cdot)$ for the total distance traveled, and $\text{Speed}(\cdot)$ for summarizing dynamic metrics such as average speed and maximum speed. These features are formatted to structured natural language sentences, providing semantic descriptions of S_n. More details can be found in Appendix A.2.

Visual Tokenization. To complement the text tokens, we generate visual tokens I_n which is the map-projected image. The image generation process contains three steps:

(1) *Trajectory Rendering:* Each sub-trajectory S_n is rendered as a visually distinct polyline on a map tile service (e.g., OpenStreetMap). The start and end points of the segment are marked with unique icons to indicate direction.

(2) *Map Clipping:* To focus the image on the relevant map area while preserving surrounding context, we dynamically calculate a minimal bounding box that encompasses all points within S_n. This bounding box is then expanded by a predefined padding factor Δ to obtain the clipping area.

(3) *Image Capturing:* Leveraging a headless browser engine (e.g., Puppeteer), we capture a screenshot of the rendered map within the computed clipping area $\mathcal{B}_{\text{clip}}(S_n)$.

Note that the method above is a general version to clarify the basic idea of multimodal tokenization. Next, we will detail how to integrate spatial, contextual, and temporal information with the tokenization method.

4.2 Multiview Trajectory Modeling

After trajectory tokenization, a core challenge remains: how to uniformly integrate sufficient information for meeting the diverse demands of tasks and model the trajectory in a unified way. Given a sequence of sub-trajectories $\mathbf{S} = \{S_1, \ldots, S_n, \ldots, S_N\}$ of trajectory T, multiview trajectory modeling aims to construct a multi-view interleaved image-text sequence M^T which contains different views of spatial scales $\mathbf{V} = \{V_1, \ldots, V_k, \ldots, V_K\}$ and contextual information $\mathbf{C} = \{C_1, \ldots, C_z, \ldots, C_K\}$.

Algorithm 1 Multiview Trajectory Modeling

Input: A sequence of sub-trajectories of a trajectory T, denoted as $\mathbf{S} = \{S_1, \ldots, S_n, \ldots, S_N\}$.
Output: Multiview interleaved image-text sequence $\mathbf{M}^T$.

```
M^T ← []
V = {V_1, ..., V_K} ← GenerateSpatialViews(S, G)
C = {C_1, ..., C_K} ← GenerateContextualViews()
for each spatial view V_k ∈ V do
    for each sub-trajectory J_{k,i} ∈ J_k do
        for each contextual view C_z ∈ C do
            I^z_{k,i} ← GenerateVisual(J_{k,i}, C_z) {Using the context information provided by C_j, generate image based on view J_{k,i}}
            D^z_{k,i} ← GenerateText(J_{k,i}, C_z) {Generate statistical text description for J_{k,i}}
            M^T.append((I^z_{k,i}, D^z_{k,i}))
        end for
    end for
end for
return M^T
```

4.2.1 Spatial Information Modeling. To model coarse-to-fine spatial information of T, a set of K spatial views, $\mathbf{V} = \{V_1, \ldots, V_k, \ldots, V_K\}$

is constructed to describe the trajectory at different scales. Each spatial view V_k represents a scale of trajectory T and can be obtained by combining sub-trajectories with a specific coarse-grained strategy g_k. For trajectory T, we construct a set of coarse-grained strategies $\mathcal{G} = \{g_1, \dots, g_k, \dots, g_K\}$ which convert multimodal representations for different spatial scales.

Formally, by applying g_k on the sub-trajectories $\mathbf{S}$, we can obtain a partition of $\mathbf{S}$, denote as $J_k = \{J_{k,1}, \dots, J_{k,i}, \dots, J_{k,I}\}$, where $J_{k,i}$ is a combination of successive sub-trajectories, *e.g.*, $J_{k,1} = \{S_1, S_2, \dots\}$ and $J_{k,i} = \{S_{n-1}, S_n, \dots\}$. As illustrated in Figure 2(b), we construct two coarse-grained strategies, *i.e.*, g_{global} and g_{local}. g_{global} takes T as a whole to generate the global view V_{global}, and g_{local} treats each sub-trajectory independently to generate the local view V_{local}. Notice that T contains four sub-trajectories, the spatial information modeling of T can be formalized as follows: $V_{local} = \{J_{local,1}, J_{local,2}, J_{local,3}, J_{local,4}\}$, and $V_{global} = J_{global,1}$.

To obtain multimodal representation, each $J_{k,i}$ can also be viewed as a sub-trajectory of T. We can get the structured text tokens $D_{k,i}$ and visual tokens $I_{k,i}$ for each sub-trajectory $J_{k,i}$ by applying the **Map-Anchored Tokenization** module described in Section 4.1. The combination of multimodal representations of these spatial views captures the coarse-to-fine spatial semantics of the trajectory.

4.2.2 Contextual Information Modeling. Traj-MLLM enriches the visual tokens of spatial view partitions to integrate various contextual information. We introduce a set of contextual views $\mathbf{C} = \{C_1, \dots, C_z, \dots, C_Z\}$, each representing a different type of contextual information relevant to the trajectory, such as POIs, road networks, and traffic lights. Given an arbitrary $J_{k,i} \in J_k$, we extend the visual tokenization method described in Section 4.1.2 and generate image $\{I^z_{k,i}\}_{z \in 1 \dots Z}$ for each context view $C_z \in \mathbf{C}$.

As illustrated in Figure 2(b), we enrich the spatial views by POIs and road networks, and obtain $I^{POI}_{k,i}$ and $I^{road}_{k,i}$. We employ two key strategies to ensure the resulting images are informative yet clear.

Layer Decomposition. Directly overlaying all available contextual data on a single image would obscure the useful information in visual clutter. To address this, we decompose contextual information into distinct visual map layers. Each map layer can be treated as a contextual view of T. Thus, the map layer decomposition produces multiple contextual views for a single spatial $J_{k,i}$. As shown, $I^{POI}_{k,i}$ and $I^{road}_{k,i}$ are generated for T.

Context Filtering. However, due to the large amount of contextual data available in a region, it is necessary to filter the most relevant information for the trajectory. Specifically, we filter the context elements rendered in each view based on their relevance to the sub-trajectory $J_{k,i}$. Let $\mathcal{E}^z_{k,i}$ denote the full set of context elements associated with view type C_z. We define a general scoring function $\text{dis}(e, J_{k,i})$ that measures how closely a context element $e \in \mathcal{E}^z_{k,i}$ relates to the sub-trajectory represented in $J_{k,i}$. We retain only those elements below a threshold θ which represents the relevance of the trajectory:

$$\mathcal{E}^{z*}_{k,i} = \{e \in \mathcal{E}^z_{k,i} \mid \text{dis}(e, J_{k,i}) \leq \theta\} \tag{4}$$

In practice, we realize the relevance function primarily based on spatial proximity. For example, in a POI-centric view, the element set $\mathcal{E}^{\text{POI}}_{k,i}$ consists of all POIs in the region. The filtered subset is defined as:

$$\mathcal{E}^{\text{POI}*}_{k,i} = \{e \in \mathcal{E}^{\text{POI}}_{k,i} \mid \text{dis}^{POI}(e, J_{k,i}) \leq \theta_{\text{POI}}\}, \tag{5}$$

where $\text{dis}^{POI}(e, J_{k,i})$ is the shortest Euclidean distance from the POI e to $J_{k,i}$ and the θ_{POI} is the distance threshold.

Furthermore, to visually highlight the trajectories, we present them clearly enough to distinguish trajectories from the contextual data. As shown in Figure 2, we use thick red lines in the POI-centric view or thick green lines in the road network view. These vision and filtering strategies can effectively guide MLLMs to understand critical contextual cues.

4.2.3 Temporal Information Modeling. Temporal order is a fundamental property of trajectories. Instead of using a trainable temporal encoder, Traj-MLLM exploits the sequential modeling capability of MLLMs to capture the temporal information. By arranging the multimodal tokens of all generated spatial and contextual views, we can obtain an interleaved image-text sequence for T.

Formally, given a set of multimodal tuples $\{(I^z_{k,i}, D^z_{k,i})\}$ where $k \in (1, K)$, $i \in [1, I]$ and $z \in [1, Z]$, we assemble them into an ordered interleaved image-text sequence $\mathbf{M}^T = [M^T_1, \dots, M^T_k, \dots, M^T_K]$ to capture temporal dependencies. Each M^T_k corresponds to a spatial view V_k. Each M^T_k is further composed of I spatial sub-views, denoted as $M^T_k = [M^T_{k,1}, \dots, M^T_{k,i}, \dots, M^T_{k,I}]$. Each sub-view $M^T_{k,i}$ consists of Z multimodal pairs derived from different contextual views $\mathbf{C} = \{C_1, \dots, C_z, \dots, C_Z\}$. The overall process is summarized in Algorithm 1.

$$M^T_{k,i} = \left[(I^1_{k,i}, D^1_{k,i}), \dots, (I^z_{k,i}, D^z_{k,i})\right], \tag{6}$$

where $(I^z_{k,i}, D^z_{k,i})$ denotes the image-text pair in contextual view C_z.

As shown in Figure 2(b), the sequence $M^T_{\text{local},3}$ refers to the third sub-trajectory $J_{local,3}$ of the local spatial view V_{local}. It contains multimodal pairs from two contextual views: the POI-centric view C_{POI} and the road network view C_{road}, denoted as $(I^{\text{POI}}_{\text{local},3}, D^{\text{POI}}_{\text{local},3})$ and $(I^{\text{road}}_{\text{local},3}, D^{\text{road}}_{\text{local},3})$, respectively.

The images are encoded as base64 to fit the inputs of MLLM prompt. The corresponding text description both provides a semantic interpretation of the image and serves as a text anchor (e.g., "POI image of segment 1: <image>"). This structured and sequential presentation enables MLLMs to naturally interpret trajectories as temporally coherent narratives, effectively capturing time-dependent relationships without additional training.

4.3 Task Prompt Optimization

To achieve flexible task adaptation, we propose to model the task-specified information in the task prompt instead of training task heads. However, both the reasoning steps and used information of different tasks vary significantly. It is challenging to construct a data-invariant task prompt that can be easily understood by MLLMs and suitable for different trajectories. This module aims to address this challenge with a few labeled seed trajectories (e.g., less than 10) and conducts multi-rounds interaction with MLLMs to optimize the task prompts. Next, we first introduce the construction of task prompt template, and then describe the optimization process.

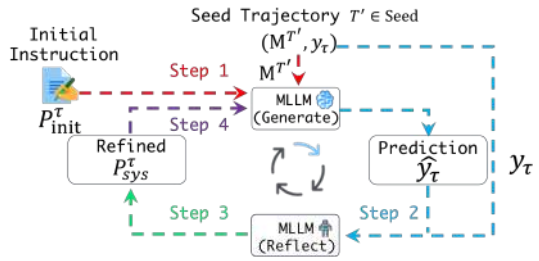


Figure 3: Task Prompt Optimization.

4.3.1 Task Prompt Template. For each trajectory mining task τ, we generate the task prompt with a general prompt template. The task prompt P_{sys} consists of four components:

$$P_{\text{sys}} = (P_{\text{role}}, P_{\text{task}}, P_{\text{know}}, P_{\text{format}}) \tag{7}$$

Here, the Role Definition (P_{role}) and Output Format (P_{format}) are fixed for each task. The optimization focuses on refining the Task Description (P_{task}) and Domain Knowledge (P_{know}).

4.3.2 Prompt Optimization. In this module, `Traj-MLLM` takes an initial prompt P^τ_{init} predefined by human experience and a few seed trajectories as inputs. We aim to optimize P^τ_{init} into a more effective prompt P^τ_{sys} through an automated, MLLM refinement loop. As shown in Figure 3, the optimization proceeds as follows: (1) We first select a set of seed trajectories **Seed** and generate their corresponding interleaved image-text sequence set $\mathbb{M} = \{\mathbf{M}^{T'} \mid T' \in \mathcal{S}\}$; (2) For each $T' \in$ **Seed**, we feed its interleaved image-text sequence $\mathbf{M}^{T'}$ and the initial system prompt P^τ_{init} into the MLLM to obtain the output y; (3) We construct a feedback loop, *i.e.*, MLLM reflects on its output y in light of the true label $\hat{y}$, generating suggestions to improve the prompt. Based on the feedback from disagreement samples, the MLLM generates a refined prompt P^τ_{sys}, which is subsequently used to re-evaluate other seed trajectories; (4) Repeat (2)-(3) several times, allowing MLLM to self-improve the prompt based on feedback. The process concludes when the refined prompt P^τ_{sys} demonstrates satisfactory performance on **Seed**.

Notice that once P^τ_{sys} is finalized, it can be directly reused across different regions by simply updating some location-specific parts in P_{know}, such as the population and city scale.

Model Inference. Once P^τ_{sys} is finalized, inferring any new trajectory T is straightforward. The trajectory is first transformed into an interleaved image-text sequence $\mathbf{M}^T$ (see Sections 4.1 and 4.2). This sequence, together with the optimized prompt P^τ_{sys}, is used as input to the MLLM, which generates a textual response. The output is then parsed according to P_{format} to extract structured results.

5 Experiment

We conduct experiments to answer the following questions:
RQ1: How does `Traj-MLLM` perform on different tasks?
RQ2: How does `Traj-MLLM` perform on different MLLM backbones?
RQ3: What is the efficiency of `Traj-MLLM`?
RQ4: What are the influences of the proposed components?
RQ5: How is the reasoning ability of `Traj-MLLM`?
RQ6: What are the influences of hyper-parameter θ?

5.1 Experimental Settings

In this section, we briefly introduce the datasets, tasks, evaluation metrics, and baselines. All the datasets and codes of `Traj-MLLM` are released [1]. More detailed settings can be found in Appendix A.3.

Datasets. We conduct extensive experiments on four public trajectory datasets, Xian, Chengdu[2], Porto[3] and Geolife[44]. The map layers, road networks and POIs are obtained by OpenStreetMap[4]. Due to budget constraints, we select over 80,000 trajectories and perform four tasks to obtain about 320,000 MLLMs responses.

Tasks and Evaluation Metrics. We choose four tasks that are widely studied in trajectory data mining, *i.e.*, travel time estimation (TTE) [31, 45], anomaly detection (AD) [26, 29], mobility prediction (MP) [47] and transportation mode identification (TMI) [46], to evaluate the effectiveness of `Traj-MLLM`. We follow existing works[29, 31, 46] to choose the same metrics. For TTE, we report mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). For AD, we evaluate the performance with precision-recall area under the curve (PR-AUC). For MP, we choose Top-1 Accuracy (ACC@1) and Top-5 Accuracy (ACC@5). For TMI, we report Accuracy, Macro-F1, and Weighted-F1. More details can be found in Appendix A.3.2.

Baselines. For each task, we choose three kinds of baselines, *i.e.*, task-specified models, trajectory foundation models and LLM-based models. The baselines for TTE and MP include Traj2Vec [36], T2Vec [18], Trember [7], CTLE [22], Toast [2], TrajCL [1], MMTEC [21], JGRM [24], START [10], UniTR[42], BigCity[37] and BLUE[45], PLM4Traj [47] and Path-LLM[30] and MM-Path [31]. For AD, we compare `Traj-MLLM` against iBAT [38], GM-VSAE [23], ATROM [8], DeepTEA [9], MST-OATD [29], and FOTraj [26]. For TMI, `Traj-MLLM` is compared with SECA [3], TrajODE [20], TrajFormer [19], and RED [46].

5.2 Effectiveness of `Traj-MLLM` (RQ1)

To answer **RQ1**, we compare `Traj-MLLM` against multiple baselines across four tasks on four datasets. The results show that `Traj-MLLM` consistently achieves the best performance on all tasks and regions. Notice that `Traj-MLLM` achieves these results without any training or fine-tuning.

For TTE, we report the results in Table 1 and derive three observations. (1) `Traj-MLLM` outperforms all baselines across all three datasets. For instance, compared to the strongest multimodal baseline MM-Path, `Traj-MLLM` reduces the MAE by over 29% on Xian, 15% on Porto, and 21% on Chengdu. This consistent improvement demonstrates the effectiveness of `Traj-MLLM`. (2) Although MM-Path considers the image modality, it is also inferior to `Traj-MLLM` because multiview trajectory modeling provides richer contextual information to inspire the reasoning ability of MLLMs. (3) `Traj-MLLM` achieves over 34.9% performance improvements on MAPE compared to the LLM-based methods, *i.e.* PLM4Traj and Path-LLM. This result demonstrates that `Traj-MLLM` performs better than fine-tuning a small LLM, *e.g.*, GPT-2.

The results of MP are summarized in Table 1. MP requires capturing fine-grained spatial-temporal information of the trajectory. Similar to TTE, we observe over 15% improvements of ACC@1

[1] https://anonymous.4open.science/r/Traj-MLLM/
[2] https://outreach.didichuxing.com/
[3] https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory
[4] https://www.openstreetmap.org/

Table 1: Performance on Travel Time Estimation and Mobility Prediction.

Methods	Xian					Porto					Chengdu				
	Travel Time Est.			Mobility Pred.		Travel Time Est.			Mobility Pred.		Travel Time Est.			Mobility Pred.	
	MAE	RMSE	MAPE	ACC@1	ACC@5	MAE	RMSE	MAPE	ACC@1	ACC@5	MAE	RMSE	MAPE	ACC@1	ACC@5
Traj2Vec[36]	230.98	382.31	28.18	42.31	72.99	210.12	362.76	26.61	43.77	73.98	238.92	390.11	29.11	42.12	71.88
T2Vec[18]	228.84	377.63	27.28	43.21	74.01	208.22	355.47	26.23	44.22	75.27	234.52	378.24	28.87	44.01	73.23
Trember[7]	223.73	377.23	25.82	49.72	79.17	204.49	347.66	24.56	49.57	79.82	231.55	377.15	25.92	49.72	79.54
CTLE[22]	232.29	377.34	26.82	50.83	80.14	205.53	355.82	26.25	50.22	80.65	228.52	388.31	25.88	50.23	81.02
Toast[2]	226.65	378.92	26.18	50.57	80.07	204.24	353.43	26.11	50.09	80.21	228.12	386.78	25.81	49.97	80.22
TrajCL[1]	218.12	365.12	25.44	49.81	79.92	198.23	340.44	24.11	49.96	79.81	222.13	368.33	25.65	49.82	79.87
START[10]	203.11	339.90	24.01	49.71	80.29	195.49	354.22	23.84	51.98	81.37	196.44	314.14	24.12	53.22	86.47
MMTEC[21]	201.23	332.18	23.91	52.25	81.16	194.21	352.34	23.67	51.99	81.13	195.32	318.98	24.02	53.12	86.52
UniTR[42]	204.93	345.33	25.14	51.67	79.39	199.41	347.59	23.82	50.11	80.14	208.44	339.23	24.82	52.87	84.10
JGRM[24]	202.19	332.88	23.98	52.28	81.02	194.43	352.45	23.78	52.01	81.29	195.34	314.62	24.01	53.11	86.42
BLUE[45]	186.34	298.13	19.35	49.45	80.44	171.22	298.11	19.32	48.43	81.72	169.22	264.93	20.11	57.90	83.24
MM-Path[31]	197.41	317.85	22.85	48.76	81.67	181.16	320.66	22.12	48.16	80.92	187.03	279.40	22.76	58.88	84.17
BigCity[37]	198.78	325.98	23.69	50.12	80.33	186.91	329.39	22.72	51.44	81.01	188.42	301.11	23.18	55.23	81.76
PLM4Traj[47]	196.72	334.42	23.61	52.76	82.15	189.99	344.64	22.89	52.33	81.44	189.09	298.24	23.21	59.50	86.27
Path-LLM[30]	192.76	313.88	23.54	50.85	80.73	184.42	323.83	22.67	50.15	80.87	184.67	294.55	22.97	56.44	82.45
Traj-MLLM	**131.02**	**154.86**	**15.36**	**60.72**	**83.25**	**144.63**	**173.54**	**15.14**	**60.45**	**82.33**	**133.43**	**154.67**	**16.34**	**60.12**	**87.22**
Improvement	*29.69%*	*48.05%*	*20.62%*	*15.09%*	*1.34%*	*15.53%*	*41.79%*	*21.64%*	*15.52%*	*0.75%*	*21.15%*	*41.62%*	*18.75%*	*1.04%*	*0.81%*

Table 2: Performance on Anomaly Detection (PR-AUC score).

Method	Xian					Porto					Chengdu				
	Detour			Switch		Detour			Switch		Detour			Switch	
	Low	Medium	High	$\mu = 0.3$	$\mu = 0.5$	Low	Medium	High	$\mu = 0.3$	$\mu = 0.5$	Low	Medium	High	$\mu = 0.3$	$\mu = 0.5$
iBAT[38]	0.439	0.487	0.694	0.199	0.262	0.273	0.313	0.501	0.201	0.241	0.450	0.497	0.691	0.201	0.254
GM-VSAE[23]	0.854	0.921	0.944	0.521	0.612	0.374	0.722	0.910	0.871	0.893	0.858	0.919	0.947	0.547	0.609
ATROM[8]	0.816	0.897	0.901	0.532	0.601	0.386	0.709	0.882	0.845	0.881	0.841	0.872	0.921	0.535	0.611
DeepTEA[9]	0.871	0.932	0.940	0.576	0.654	0.408	0.736	0.921	0.861	0.872	0.875	0.931	0.935	0.576	0.655
MST-OATD[29]	0.902	0.951	0.971	0.598	0.691	0.739	0.892	0.970	0.899	0.912	0.907	0.949	0.969	0.604	0.677
FOTraj[26]	0.911	0.954	0.979	0.625	0.733	0.741	0.901	0.974	0.927	0.946	0.913	0.954	0.971	0.649	0.656
Traj-MLLM	**0.967**	**0.979**	**0.981**	**0.947**	**0.956**	**0.968**	**0.980**	**0.981**	**0.971**	**0.981**	**0.967**	**0.976**	**0.980**	**0.966**	**0.983**
Improvement	*6.15%*	*2.62%*	*0.20%*	*51.52%*	*30.42%*	*30.63%*	*8.77%*	*0.72%*	*4.75%*	*3.70%*	*5.91%*	*2.31%*	*0.93%*	*48.84%*	*45.19%*

Table 3: Performance on Transportation Mode Identification.

Method	Geolife		
	Accuracy	Macro-F1	Weighted-F1
SECA[3]	78.91	78.78	79.02
TrajODE[20]	85.48	84.98	85.17
TrajFormer[19]	85.56	85.12	85.71
RED[46]	85.89	85.23	86.11
Traj-MLLM	**87.46**	**86.71**	**87.45**
Improvement	*1.83%*	*1.74%*	*1.56%*

on Xian and Porto compared to the strongest baseline PLM4Traj, demonstrating that Traj-MLLM can guide MLLMs to understand complex and fine-grained mobility patterns.

The results of anomaly detection are presented in Table 2. From these results, we observe: (1) Traj-MLLM achieves the best performance among all baselines, consistently outperforming on all datasets, anomaly types, and perturbation levels. (2) Traj-MLLM is robust on both subtle detour and complex switch anomalies, maintaining stable performance under varying perturbation magnitudes. For instance, the PR-AUC score increases from 0.741 to 0.968 for low–magnitude detour anomalies. This improvement can be attributed to the *multiview trajectory modeling*, which provides multi-scale spatial information that encourages the model to capture diverse anomalous patterns. (3) Traj-MLLM achieves high performance without requiring any model training or explicit anomaly examples, enabling to generalize Traj-MLLM to unseen regions.

Moreover, we present the transportation mode identification results in Table 3. As illustrated, Traj-MLLM improves accuracy by 1.83% over RED[46]. By feeding multimodal information to the MLLM, Traj-MLLM is able to combine visual cues from the map (e.g., the presence of railway tracks) and statistical information from the text prompts (e.g., velocity and acceleration) to infer the transportation mode. In summary, across the four tasks, Traj-MLLM can leverage the MLLMś capacity for both fine-grained spatial analysis (as in TTE and MP) and high-level semantic reasoning (as in AD and TMI). according to these results, we argue that Traj-MLLM has strong generalization ability to achieve region-agnostic and task-adaptive trajectory data mining without the requirement of any data.

5.3 Performance of different MLLMs (RQ2)

To answer **RQ2**, we explore the performance of Traj-MLLM under seven MLLM backbones on the Porto dataset across two tasks, *i.e.* TTE and AD (detour anomalies). The results are shown in Figure 4. According to the results, we have three observations: (1) Within the same model family, closed-source models tend to outperform open-source models. For example, the closed-source qwen-vl-max achieves better results on both tasks than the open-source model qwen-2.5-vl-72b, which may be due to its architecture or proprietary optimizations during training. (2) Model scale is not the sole determinant of performance. Notably, the open-source gemma-3-27b-it performs better than the larger qwen-2.5-vl-72b and achieves a PR-AUC score nearly on par with the powerful closed-source gemini-2.5-pro. This highlights the significant potential and bright

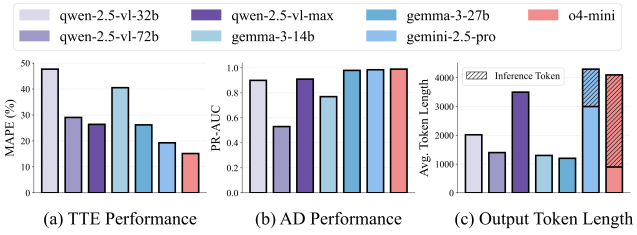


Figure 4: Performance with different MLLMs.

Table 4: Efficiency of Different MLLMs.

Model	Latency	Throughput	AvgTime	Cost
gemma-3-14b	14.1	178.6	0.079	$0.002
gemma-3-27b	11.2	268.1	0.042	$0.004
qwen2.5-vl-32b	15.2	189.9	0.080	$0.006
qwen-vl-max	10.4	277.3	0.038	$0.017
qwen2.5-vl-72b	19.3	149.2	0.129	$0.018
gemini-2.5-pro	32.0	89.9	0.356	$0.024
o4-mini	9.23	312.5	0.030	$0.014

future of open-source models in trajectory data mining. (3) As shown in Figure 4(c), o4-mini achieves the best results on both tasks while generating the shortest output responses. Its superior performance and efficiency can be attributed to its longest inference tokens. Given its dominant performance, we select o4-mini as the default MLLM backbone to report our performances. We show the output differences of different MLLMs in Appendix A.5.2

5.4 Efficiency of `Traj-MLLM` (RQ3)

To answer **RQ3**, we evaluate the efficiency of `Traj-MLLM` by measuring inference latency, throughput, average time, and cost on different MLLM backbones. The throughput and average time are calculated based on the maximum allowed for the API provider's highest access tier (e.g., o4-mini has a throughput of 150M tokens per minute (TPM)). The results are summarized in Table 4.

As shown in the table, we have the following observations: (1) o4-mini is the most efficient, achieving the lowest latency and the highest throughput. In contrast, gemini-2.5-pro is the slowest model due to its significantly longer response time. (2) In terms of cost, gemma-3-14b is the most economical choice, which can be attributed to its smaller parameter size. The gemini-2.5-pro is the most expensive model owing to its large number of output tokens. (3) With the large scale deployment of MLLMs, the throughput increases and average processing time of each trajectory drops sharply, *i.e.*, 0.030s per trajectory. These results highlight the trade-offs between inference speed, throughput, and cost to use MLLMs.

Table 5: Ablation Study Results on the Chengdu Dataset.

Method	Chengdu				
	Travel Time Estimation			Anomaly Detection	
	MAE ↓	RMSE ↓	MAPE ↓	F1-score ↑	PR-AUC ↑
w/o image	350.15	392.33	35.97	-	-
w/o seg	243.26	357.02	28.44	0.788	0.821
w/o order	330.23	368.12	32.87	0.952	0.953
w/o POI	194.06	247.79	19.92	0.940	0.942
w/o road	234.91	304.88	24.75	0.862	0.871
`Traj-MLLM`	**133.43**	**154.67**	**16.34**	**0.980**	**0.980**

5.5 Ablation Study (RQ4)

We compare `Traj-MLLM` with five ablations on two downstream tasks to analyze the effectiveness of the proposed components. We obtain **w/o image**, **w/o segment**, **w/o order**, **w/o POI**, and **w/o road**, by removing the visual modality, replacing the principled segmentation with a naive strategy, breaking the correspondence between textual descriptions and their visual components, removing the POI-centric view, and removing the road-network-centric view, respectively. The results are summarized in Table 5.

Removing the image modality (**w/o image**) causes the most significant performance drop on both tasks. This indicates that constructing visual representations is crucial for MLLMs to understand trajectory details. Breaking the order of visual-text pairs (**w/o order**), which removes temporal correspondence, also leads to a notable performance decrease. This means that temporal order is crucial to capture trajectory dynamics accurately. Replacing our semantic segmentation approach with a naive segmentation strategy (**w/o segment**) also leads to large performance degradation, demonstrating that the local detail view obtained through semantically coherent segmentation can provide crucial information. Furthermore, the lack of POI-centric view and road network-centric view also significantly reduces the performance. In a word, these ablation results support that the designed components of `Traj-MLLM` have positive contributions to the performance of `Traj-MLLM`.

5.6 Reasoning Ability Analysis (RQ5)

To answer **RQ5**, we analyze the reasoning ability of `Traj-MLLM`. Due to the space limit, we show the case figures and more detailed analysis in Appendix A.5. `Traj-MLLM` demonstrates strong reasoning abilities in two key aspects: (1) For **interpretable result**, `Traj-MLLM` provides detailed, step-by-step reasoning. For example, in the anomaly detection task, it not only identifies an anomalous trajectory but also precisely locates the anomalous segments and explains its conclusion by referencing the multimodal information. (2) For **multi-round reasoning**, the model can adapt its predictions based on real-time contextual updates. For instance, in the TTE task, after being notified of traffic congestion on a specific road, `Traj-MLLM` incorporates this new information to revise its initial travel time estimation. These capabilities show that `Traj-MLLM` functions as an interactive reasoning engine, not just a static prediction model.

5.7 Sensitivity Analysis of θ (RQ6)

We conduct a sensitivity analysis on θ to investigate how different distance thresholds between context elements and the trajectory affect downstream task performance. According to the results presented in Figure 5, we have two key observations: (1) For all datasets, task performance consistently drops when the distance threshold exceeds 100m, indicating that larger thresholds introduce excessive irrelevant contextual information, which negatively impacts downstream tasks. (2) The optimal threshold differs by dataset: for Porto, the best results are achieved at 50m, while for Chengdu and Xian, 100m yields the highest performance. This difference reflects the variations in urban layouts—Chengdu and Xian have more regular road networks, making 100m a balanced choice between informativeness and noise. In contrast, Porto's dense and irregular road

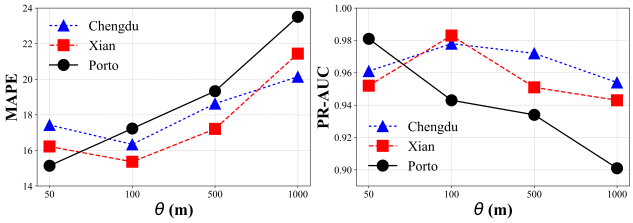


Figure 5: Sensitivity Analysis.

structure favors a smaller 50m threshold to avoid interference from unrelated elements. More details are in Appendix A.4.

6 Conclusion

In this paper, we propose a novel training-free framework Traj-MLLM which is the first work employing the reasoning ability of MLLMs for trajectory data mining. By effectively bridging the semantic gap of raw trajectory, Traj-MLLM integrates multiview contexts and transforms trajectories into multimodal sequences, enabling the use of the reasoning ability of MLLMs directly. Experiments on multiple representative tasks and cross-city datasets show that Traj-MLLM achieves performance improvements while maintaining strong generalization ability without any training or fine-tuning. We also release all the data responded by MLLMs, which can be used as the foundation for future research in training MLLMs specialized for trajectories and analyzing other spatial-temporal data.

References

[1] Yanchuan Chang, Jianzhong Qi, Yuxuan Liang, and Egemen Tanin. 2023. Contrastive trajectory similarity learning with dual-feature attention. In *2023 IEEE 39th International conference on data engineering (ICDE)*. IEEE, 2933–2945.

[2] Yile Chen, Xiucheng Li, Gao Cong, Zhifeng Bao, Cheng Long, Yiding Liu, Arun Kumar Chandran, and Richard Ellison. 2021. Robust road network representation learning: When traffic patterns meet traveling semantics. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 211–220.

[3] Sina Dabiri, Chang-Tien Lu, Kevin Heaslip, and Chandan K Reddy. 2019. Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data. *IEEE Transactions on Knowledge and Data Engineering* 32, 5 (2019), 1010–1023.

[4] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 9062–9072.

[5] Yuwei Du, Jie Feng, Jie Zhao, and Yong Li. 2024. Trajagent: An agent framework for unified trajectory modelling. *arXiv e-prints* (2024), arXiv–2410.

[6] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 world wide web conference*. 1459–1468.

[7] Tao-Yang Fu and Wang-Chien Lee. 2020. Trembr: Exploring road networks for trajectory representation learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 1 (2020), 1–25.

[8] Qiang Gao, Xiaohan Wang, Chaoran Liu, Goce Trajcevski, Li Huang, and Fan Zhou. 2023. Open anomalous trajectory recognition via probabilistic metric learning. In *IJCAI*. International Joint Conferences on Artificial Intelligence Organization.

[9] Xiaolin Han, Reynold Cheng, Chenhao Ma, and Tobias Grubenmann. 2022. DeepTEA: Effective and efficient online time-dependent trajectory outlier detection. *Proceedings of the VLDB Endowment* 15, 7 (2022), 1493–1505.

[10] Jiawei Jiang, Dayan Pan, Houxing Ren, Xiaohan Jiang, Chao Li, and Jingyuan Wang. 2023. Self-supervised trajectory representation learning with temporal regularities and travel semantics. In *2023 IEEE 39th international conference on data engineering (ICDE)*. IEEE, 843–855.

[11] Xiang Jiang, Erico N de Souza, Ahmad Pesaranghader, Baifan Hu, Daniel L Silver, and Stan Matwin. 2017. Trajectorynet: An embedded gps trajectory representation for point-based classification using recurrent neural networks. *arXiv preprint arXiv:1705.02636* (2017).

[12] Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Harry Chao. 2024. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. *Advances in Neural Information Processing Systems* 37 (2024), 28798–28827.

[13] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27831–27840.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[16] Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Forty-first International Conference on Machine Learning*.

[17] Wenbin Li, Di Yao, Ruibo Zhao, Wenjie Chen, Zijie Xu, Chengxue Luo, Chang Gong, Quanliang Jing, Haining Tan, and Jingping Bi. 2025. Stbench: Assessing the ability of large language models in spatio-temporal analysis. In *Companion Proceedings of the ACM on Web Conference 2025*. 749–752.

[18] Xiucheng Li, Kaiqi Zhao, Gao Cong, Christian S Jensen, and Wei Wei. 2018. Deep representation learning for trajectory similarity computation. In *2018 IEEE 34th international conference on data engineering (ICDE)*. IEEE, 617–628.

[19] Yuxuan Liang, Kun Ouyang, Yiwei Wang, Xu Liu, Hongyang Chen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. 2022. TrajFormer: Efficient trajectory classification with transformers. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1229–1237.

[20] Yuxuan Liang, Kun Ouyang, Hanshu Yan, Yiwei Wang, Zekun Tong, and Roger Zimmermann. 2021. Modeling Trajectories with Neural Ordinary Differential Equations.. In *IJCAI*. 1498–1504.

[21] Yan Lin, Huaiyu Wan, Shengnan Guo, Jilin Hu, Christian S Jensen, and Youfang Lin. 2023. Pre-training general trajectory embeddings with maximum multi-view entropy coding. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2023), 9037–9050.

[22] Yan Lin, Huaiyu Wan, Shengnan Guo, and Youfang Lin. 2021. Pre-training context and time aware location embeddings from spatial-temporal trajectories for user next location prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4241–4248.

[23] Yiding Liu, Kaiqi Zhao, Gao Cong, and Zhifeng Bao. 2020. Online anomalous trajectory detection with deep generative sequence modeling. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 949–960.

[24] Zhipeng Ma, Zheyan Tu, Xinhai Chen, Yan Zhang, Deguo Xia, Guyue Zhou, Yilun Chen, Yu Zheng, and Jiangtao Gong. 2024. More than routing: Joint GPS and route modeling for refine trajectory representation learning. In *Proceedings of the ACM Web Conference 2024*. 3064–3075.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[26] Wei Shao, Ziquan Fang, lu Chen, and yunjun Gao. 2025. Towards Trajectory Anomaly Detection: A Fine-Grained and Noise-Resilient Framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[27] Jiahui Sun, Haiming Jin, Zhaoxing Yang, Lu Su, and Xinbing Wang. 2022. Optimizing long-term efficiency and fairness in ride-hailing via joint order dispatching and driver repositioning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 3950–3960.

[28] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems* 36 (2023), 8690–8701.

[29] Chenhao Wang, Lisi Chen, Shuo Shang, Christian S Jensen, and Panos Kalnis. 2024. Multi-Scale Detection of Anomalous Spatio-Temporal Trajectories in Evolving Trajectory Datasets. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2980–2990.

[30] Yongfu Wei, Yan Lin, Hongfan Gao, Ronghui Xu, Sean Bin Yang, and Jilin Hu. 2025. Path-LLM: A Multi-Modal Path Representation Learning by Aligning and Fusing with Large Language Models. In *Proceedings of the ACM on Web Conference 2025*. 2289–2298.

[31] Ronghui Xu, Hanyin Cheng, Chenjuan Guo, Hongfan Gao, Jilin Hu, Sean Bin Yang, and Bin Yang. 2025. Mm-path: Multi-modal, multi-granularity path representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 1703–1714.

[32] Shixiong Xu, Chenghao Zhang, Lubin Fan, Gaofeng Meng, Shiming Xiang, and Jieping Ye. 2024. Addressclip: Empowering vision-language models for city-wide image address localization. In *European Conference on Computer Vision*. Springer, 76–92.

[33] Peilun Yang, Hanchen Wang, Ying Zhang, Lu Qin, Wenjie Zhang, and Xuemin Lin. 2021. T3s: Effective representation learning for trajectory similarity computation. In *2021 IEEE 37th international conference on data engineering (ICDE)*. IEEE, 2183–2188.

[34] Sean Bin Yang, Jilin Hu, Chenjuan Guo, Bin Yang, and Christian S Jensen. 2023. Lightpath: Lightweight and scalable path representation learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2999–3010.

[35] Di Yao, Gao Cong, Chao Zhang, and Jingping Bi. 2019. Computing trajectory similarity in linear time: A generic seed-guided neural metric learning approach. In *2019 IEEE 35th international conference on data engineering (ICDE)*. IEEE, 1358–1369.

[36] Di Yao, Chao Zhang, Zhihua Zhu, Jianhui Huang, and Jingping Bi. 2017. Trajectory clustering via deep representation learning. In *2017 international joint conference on neural networks (IJCNN)*. IEEE, 3880–3887.

[37] Xie Yu, Jingyuan Wang, Yifan Yang, Qian Huang, and Ke Qu. 2024. BIGCity: A universal spatiotemporal model for unified trajectory and traffic state data analysis. *arXiv preprint arXiv:2412.00953* (2024).

[38] Daqing Zhang, Nan Li, Zhi-Hua Zhou, Chao Chen, Lin Sun, and Shijian Li. 2011. iBAT: detecting anomalous taxi trajectories from GPS traces. In *Proceedings of the 13th international conference on Ubiquitous computing*. 99–108.

[39] Hanyuan Zhang, Xingyu Zhang, Qize Jiang, Baihua Zheng, Zhenbang Sun, Weiwei Sun, and Changhu Wang. 2020. Trajectory similarity learning with auxiliary supervision and optimal matching. (2020).

[40] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. 2024. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–20.

[41] Zheng Zhang, Hossein Amiri, Zhenke Liu, Liang Zhao, and Andreas Züfle. 2024. Large language models for spatial trajectory patterns mining. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection*. 52–55.

[42] Jie Zhao, Chao Chen, Yuanshao Zhu, Mingyu Deng, and Yuxuan Liang. 2025. UniTR: A Unified Framework for Joint Representation Learning of Trajectories and Road Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 13348–13356.

[43] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems* 36 (2023), 5168–5191.

[44] Yu Zheng, Hao Fu, Xing Xie, Wei-Ying Ma, and Quannan Li. 2011. *Geolife GPS Trajectory Dataset – User Guide*. Technical Report. Microsoft Research Asia.

[45] Silin Zhou, Yao Chen, shuo Shang, lisi Chen, Bingsheng He, and Shibasaki Ryosuke. 2025. Blurred Encoding for Trajectory Representation Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[46] Silin Zhou, Shuo Shang, Lisi Chen, Christian S Jensen, and Panos Kalnis. 2024. RED: Effective Trajectory Representation Learning with Comprehensive Information. *arXiv preprint arXiv:2411.15096* (2024).

[47] Zeyu Zhou, Yan Lin, Haomin Wen, Shengnan Guo, Jilin Hu, Youfang Lin, and Huaiyu Wan. 2024. PLM4Traj: Cognizing Movement Patterns and Travel Purposes from Trajectories with Pre-trained Language Models. *arXiv preprint arXiv:2405.12459* (2024).

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名：梁奥
方向：人工智能
单位：中国科学院沈阳自动化研究所

题目：多智能体时空一致的多模态生成式世界模型研究

LiDARCrafter: Dynamic 4D World Modeling from LiDAR Sequences

Anonymous submission

Abstract

Generative world models have become essential data engines for autonomous driving, yet most existing efforts focus on videos or occupancy grids, overlooking the unique LiDAR properties. Extending LiDAR generation to dynamic 4D world modeling presents challenges in controllability, temporal coherence, and evaluation standardization. To this end, we present **LiDARCrafter**, a unified framework for 4D LiDAR generation and editing. Given free-form natural language inputs, we parse instructions into ego-centric scene graphs, which condition a tri-branch diffusion network to generate object structures, motion trajectories, and geometry. These structured conditions enable diverse and fine-grained scene editing. Additionally, an autoregressive module generates temporally coherent 4D LiDAR sequences with smooth transitions. To support standardized evaluation, we establish a comprehensive benchmark with diverse metrics spanning scene-, object-, and sequence-level aspects. Experiments on the nuScenes dataset using this benchmark demonstrate that LiDARCrafter achieves state-of-the-art performance in fidelity, controllability, and temporal consistency across all levels, paving the way for data augmentation and simulation. Code and benchmark will be released to the community.

1 Introduction

Generative world models are rapidly advancing the synthesis and interpretation of large-scale sensor data for autonomous driving (Hu et al. 2023; Mei et al. 2024). Recent work has concentrated on structured modalities such as video and occupancy grids, whose dense, regular structure fits image and voxel pipelines (Gao et al. 2023; Wang et al. 2024a). LiDAR, though indispensable for metric 3D geometry and all-weather perception, remains underexplored. Its point clouds are sparse, unordered, and irregular, so techniques designed for images or grids do not transfer cleanly. Early efforts, such as LiDARGen, project full 360° scans to range images and borrow pixel-based methods (Zyrianov, Zhu, and Wang 2022). Later approaches improve single-frame fidelity but stop short of dynamic sequences (Nakashima and Kurazume 2024; Ran, Guizilini, and Wang 2024). Multimodal systems like UniScene and GENESIS rely on occupancy or video as intermediaries, limiting LiDAR independence and increasing computation (Li et al. 2025; Guo et al. 2025). A dedicated 4D LiDAR world model is therefore still absent.

Bridging this gap requires progress on three fronts. First, *user controllability*. Text prompts are the most accessible interface, yet they lack the spatial detail to specify LiDAR scenes, whereas precise inputs such as 3D boxes, HD maps, or trajectories demand costly annotation (Hu et al. 2023; Bian et al. 2025). Second, *temporal consistency*. A single frame cannot reveal occlusion patterns or object kinematics for downstream utilization, necessitating a full 4D generation model. Conventional cross-frame attention strategies neglect the geometric continuity inherent in point clouds. Third, *standardized evaluation*. Existing video world models benefit from established benchmarks that score visual quality, perception accuracy, and closed-loop applicability (Huang et al. 2024). LiDAR still lacks standard protocols for evaluating fidelity and consistency across perspectives.

To close these gaps, we introduce **LiDARCrafter**, a unified framework for controllable 4D LiDAR sequence generation. The key is an explicit, object-centric 4D layout that captures geometry and motion while achieving *accessible yet precise control*. In **Text2Layout**, a large language model (LLM) lifts the user prompt to an ego-centric scene graph, and a tri-branch diffusion network predicts object boxes, trajectories, and shape priors, yielding the full 4D layout. **Layout2Scene** feeds the layout to a range-image diffusion network and produces a high-fidelity first scan. The explicit layout modeling also enables fine-grained control like insertion, deletion, and dragging. **Scene2Seq** then generates the remaining frames *autoregressively*, warping past points with motion priors to keep temporal coherence. To standardize 4D LiDAR world models assessment, we release an *evaluation suite* that scores scene, object, and sequence quality.

Experiments on the nuScenes dataset (Caesar et al. 2020) demonstrate that our method achieves top performance in both single-frame fidelity and sequence coherence, while offering intuitive control and editing capabilities. This work establishes a new benchmark for LiDAR-based dynamic world modeling in autonomous driving.

In summary, the core contributions of this work are:

- We present **LiDARCrafter**, the first 4D generative world model dedicated to LiDAR data, with superior controllability and spatiotemporal consistency.
- We introduce a tri-branch 4D layout conditioned pipeline that turns language into an editable 4D layout and uses it to guide temporally stable LiDAR synthesis.

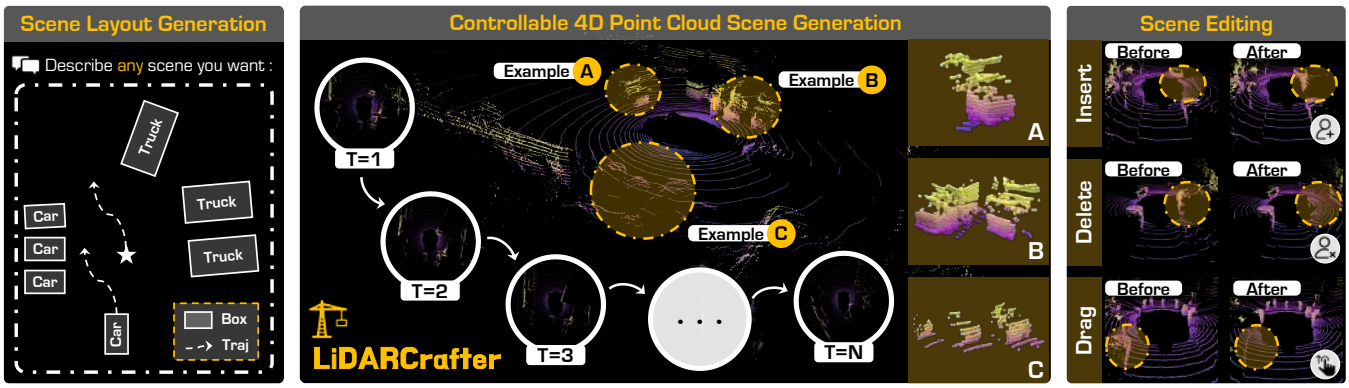


Figure 1: We propose **LiDARCrafter**, a 4D LiDAR-based generative world model that supports controllable point cloud layout generation (**left**), dynamic sequential scene generation (**center**), and rich scene editing applications (**right**). Our framework enables intuitive *"what you describe is what you get"* LiDAR-based 4D world modeling.

- We propose a comprehensive evaluation suite for LiDAR sequence generation, encompassing scene-level, object-level, and sequence-level metrics.
- We demonstrate best single-frame and sequence-level LiDAR point cloud generation performance on nuScenes, with improved foreground quality over existing methods.

2 Related Work

Driving Generative World Models. Generative world models aim to simulate environmental dynamics to aid autonomous driving decisions. Recent works primarily use video and occupancy representations. Video-based methods like GAIA-1 (Hu et al. 2023), Dreamforge (Mei et al. 2024), and Epona (Zhang et al. 2025) employ autoregressive modeling and richer conditioning for improved synthesis. MagicDrive (Gao et al. 2023) utilizes a bird's-eye view (BEV) for temporal consistency (Wang et al. 2024b; Zhao et al. 2025; Ni et al. 2025a). Occupancy-based methods, such as OccWorld (Zheng et al. 2024a), GaussianWorld (Zuo et al. 2025b), QuadricFormer (Zuo et al. 2025a), OccSora (Wang et al. 2024a), and DynamicCity (Bian et al. 2025), provide structured spatial representations beneficial for downstream tasks. Multimodal models like UniScene (Li et al. 2025) and GENESIS (Guo et al. 2025) align cross-modal features for consistency. Despite its importance, LiDAR remains underexplored, mostly limited to forecasting-oriented methods (Liu, Zhao, and Rhinehart 2025; Zhang et al. 2023).

LiDAR Point Cloud Generation. Early LiDAR generative methods, including LiDARGen (Zyrianov, Zhu, and Wang 2022), projected LiDAR point clouds into range images. Subsequent approaches like RangeLDM (Hu, Zhang, and Hu 2024), Text2LiDAR (Wu et al. 2024), and WeatherGen (Wu et al. 2025) introduced latent diffusion models (Rombach et al. 2022) or diverse conditioning signals. R2DM (Nakashima and Kurazume 2024) and R2Flow (Nakashima et al. 2024) adopted single-stage diffusion models (Ho, Jain, and Abbeel 2020) for geometric precision. UltraLiDAR (Xiong et al. 2023) utilized BEV-based VQ-VAE (Van Den Oord, Vinyals et al. 2017) for rich scene editing, further extended by OpenDWM (Ni et al. 2025b). X-Drive (Xie et al. 2024) and UniScene (Li et al. 2025) leveraged cross-modal consistency. Yet, dynamic sequence generation and fine-grained controllability remain unresolved.

Controllability in Scene Synthesis. Fine-grained control is crucial in generative modeling. Current methods depend on detailed inputs such as BEV semantic maps (Gao et al. 2023; Mei et al. 2024), HD maps (Swerdlow, Xu, and Zhou 2024), and 3D bounding boxes (Zhang et al. 2024; Yang et al. 2024), but preparing these inputs is labor-intensive. Text-based methods (Wu et al. 2024; Hu et al. 2023; Mei et al. 2024) offer simpler control but typically lack spatial precision. Two-stage indoor scene synthesis methods (Zhai et al. 2024; Yang et al. 2025b) use intermediate scene graphs to improve spatial control, recently extended to outdoor static occupancy grids (Liu et al. 2025; Yang et al. 2025a). Mature methods for controllable generation and editing of dynamic 4D LiDAR scenes, however, are not yet established.

3 LiDARCrafter: 4D LiDAR World Model

We propose **LiDARCrafter**, the first LiDAR world model that turns free-form user instructions into temporally coherent 4D point cloud sequences with object-level control. The cornerstone is an explicit 4D foreground layout that bridges the descriptive power of language and the geometric rigor required by LiDAR. As shown in Fig. 2, our framework adopts a three-stage process. In the **Text2Layout** stage (*cf.* Sec. 3.1), an LLM converts the instruction into an egocentric scene graph, and a tri-branch diffusion sampler generates *object boxes, trajectories, and shape priors*, which serve as the conditioning layout signal. In the **Layout2Scene** stage (*cf.* Sec. 3.2), a range-image diffusion model turns the layout into a high-fidelity static scan. In the **Scene2Seq** stage (*cf.* Sec. 3.2), the static cloud is autoregressively warped and inpainted to yield drift-free frames. Finally, our **EvalSuite** (*cf.* Sec. 3.4) adds metrics for object semantics, layout soundness, and motion fidelity, giving the first comprehensive benchmark for 4D LiDAR generation.

3.1 Text2Layout: 4D Layout Generation

Natural-language prompts alone lack the spatial precision needed for complex world modeling. We therefore introduce a scene graph as an intermediate, explicit encoding of object geometry and relations. LLMs can parse text into such

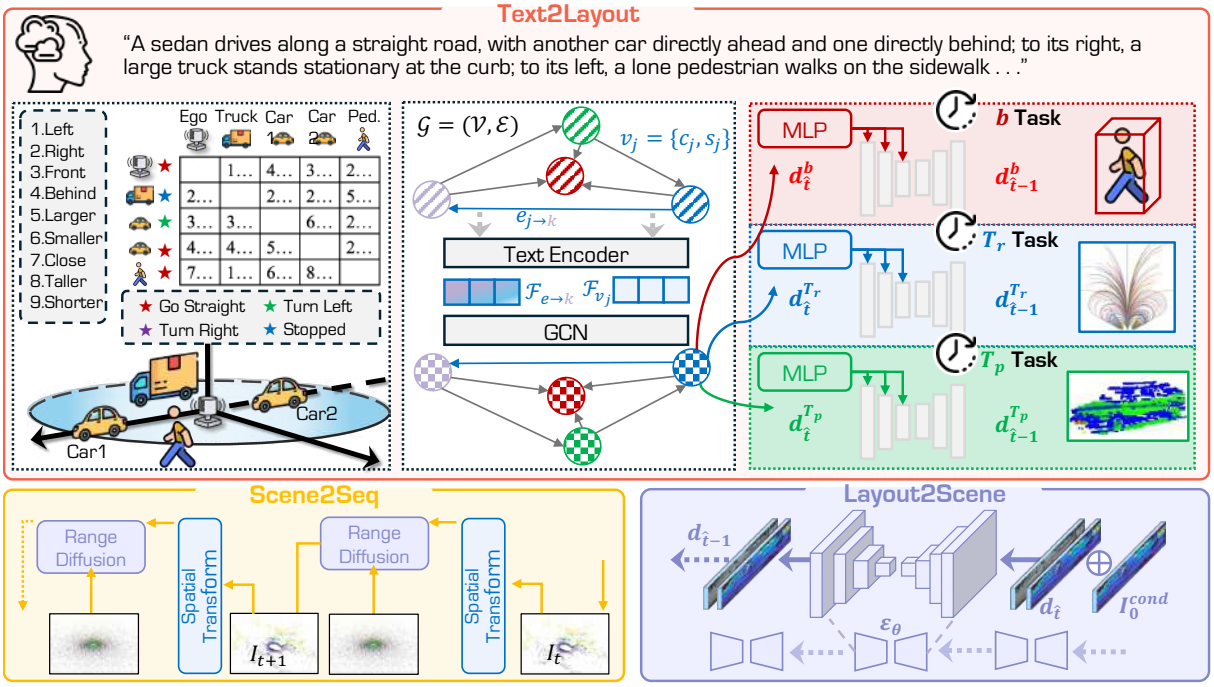


Figure 2: Framework of **LiDARCrafter**. In the **Text2Layout** stage (*cf.* Sec. 3.1), the natural-language instruction is parsed into an ego-centric scene graph, and a tri-branch diffusion network generates 4D conditions for bounding boxes, future trajectories, and object point clouds. In the **Layout2Scene** stage (*cf.* Sec. 3.2), a range-image diffusion model uses these conditions to generate a static LiDAR frame. In the **Scene2Seq** stage (*cf.* Sec. 3.3), an autoregressive module warps historical points with ego and object motion priors to generate each subsequent frame, producing a temporally coherent LiDAR sequence.

graphs, a strategy proven effective for scene synthesis (Feng et al. 2023; Yang et al. 2025b). **LiDARCrafter** extends this idea to dynamic outdoor settings. The LLM first builds a 4D scene graph from the prompt. A diffusion decoder then transforms this graph into a detailed layout of object boxes, trajectories, and shape priors that guides the subsequent LiDAR sequence generation process.

Language-Driven Graph Construction. Given a textual instruction, we prompt an LLM (Achiam et al. 2023) to build an ego-centric scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. A tailored query enumerates all foreground objects, producing the node set $\mathcal{V} = \{v_0, \ldots, v_M\}$, where v_0 denotes the ego vehicle and the remaining M nodes represent dynamic objects. Each node v_i is annotated with its semantic class c_i and a motion state phrase s_i (*e.g.*, "go straight"). For every ordered pair (i, j) with $i \neq j$, a directed edge $e_{i\to j} \in \mathcal{E}$ encodes their spatial relation (*e.g.*, "in front of", "lager than", details in Fig. 2). Unlike prior work (Liu et al. 2025), including the ego node yields a structurally complete scene graph that fully conditions downstream layout generation.

Scene-Graph Lifting. Given a textual scene graph, we aim to infer for each node v_i a 4D layout tuple $\mathcal{O}_i = \big(\mathbf{b}_i, \boldsymbol{\delta}_i, \mathbf{p}_i\big)$, where $\mathbf{b}_i = (x_i, y_i, z_i, w_i, l_i, h_i, \psi_i)$ is the 3D bounding box capturing the 3D center, size, and yaw angle of object i. $\boldsymbol{\delta}_i = \{(\Delta x_i^t, \Delta y_i^t)\}_{t=1}^T$ records planar displacements over T future frames, and $\mathbf{p}_i \in \mathbb{R}^{N\times 3}$ stores N canonical foreground points that sketch the shape of object. This tuple captures where, how, and what for every node and serves as the target of our denoiser during the diffusion process.

Graph-Fusion Encoder. To obtain context-aware priors for every tuple, following the method in the indoor area (Zhai et al. 2024), we process the scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with an L-layer TripletGCN (Johnson, Gupta, and Fei-Fei 2018). We first embed nodes and edges with a frozen CLIP text encoder (Radford et al. 2021) to bring richer semantics:

$$\begin{aligned} \mathbf{h}_{v_i}^{(0)} &= \texttt{concat}(\text{CLIP}(c_i),\, \text{CLIP}(s_i),\, \boldsymbol{\omega}_i) \\ \mathbf{h}_{e_{i\to j}}^{(0)} &= \text{CLIP}(e_{i\to j}), \end{aligned} \tag{1}$$

where $\boldsymbol{\omega}_i$ is a learnable positional code. At layer ℓ, we update triplets with two lightweight MLPs: Φ_{edge} for edge reasoning and Φ_{agg} for neighborhood aggregation as follows:

$$\begin{aligned} (\tilde{\mathbf{h}}_{v_i}^{(\ell)}, \mathbf{h}_{e_{i\to j}}^{(\ell+1)}, \tilde{\mathbf{h}}_{v_j}^{(\ell)}) &= \Phi_{\text{edge}}(\mathbf{h}_{v_i}^{(\ell)}, \mathbf{h}_{e_{i\to j}}^{(\ell)}, \mathbf{h}_{v_j}^{(\ell)}) \\ \mathbf{h}_{v_i}^{(\ell+1)} &= \tilde{\mathbf{h}}_{v_i}^{(\ell)} + \Phi_{\text{agg}}\Big(\texttt{avg}\big(\tilde{\mathbf{h}}_{v_j}^{(\ell)} \mid v_j \in N_{\mathcal{G}}(v_i)\big)\Big). \end{aligned} \tag{2}$$

After L hops, each node feature $\mathbf{h}_{v_i}^{(L)}$ encodes both global semantics and local geometry, providing a strong semantic-geometric prior for LiDAR layout generation.

Layout Diffusion Decoder. The final node embeddings condition a tri-branch diffusion decoder (Rombach et al. 2022), one branch per element of $\mathcal{O}_i$. Let $\mathbf{d}_\tau^o$ be the noisy sample of modality $o \in \mathcal{O}_i$ at timestep τ. Each branch minimizes

$$\mathcal{L}^o = \mathbb{E}_{\tau,\mathbf{d}^o,\varepsilon}\big\|\varepsilon - \varepsilon_\theta^o(\mathbf{d}_\tau^o, \tau, c^o)\big\|_2^2, \tag{3}$$

sharing a common noise schedule. Boxes and trajectories are denoised using a lightweight 1D U-Net (Ronneberger, Fischer, and Brox 2015), while object shapes are synthesised

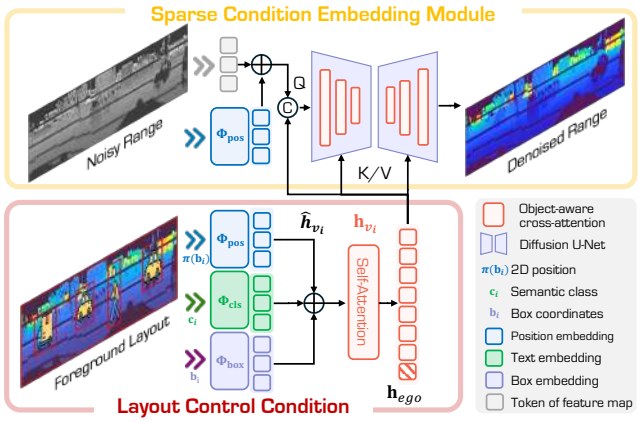


Figure 3: Details of our range-image diffusion model.

with a point-based U-Net (Zheng et al. 2024b). Unlike LOGen (Yan et al. 2025), we match only the LiDAR sampling distribution, not the exact foreground points, which eliminates the heavy DiT cost (Peebles and Xie 2023) yet still delivers plausible inputs for later refinement. More details about the denoiser are given in the Appendix.

3.2 Layout2Scene: Controlled LiDAR Generation

LiDARCrafter ensures generation fidelity by using a unified range-image diffusion backbone that generates LiDAR point clouds end to end. Given the scene graph $\mathcal{G}$ and the decoded layout $\mathcal{O}_i$, the network denoises Gaussian noise into the clean range frame $\mathbf{I}^0$, thereby bootstrapping the LiDAR sequence $\mathcal{P} = \{\mathbf{P}^t\}_{t=0}^{T}$ while following the pose, trajectory, and coarse shape of each object. The range-view representation preserves native LiDAR geometry while remaining convolution-friendly (Nakashima and Kurazume 2024).

Sparse Object Conditioning. Directly projecting all foreground points into the range image, as in OLiDM (Yan et al. 2025), inadequately represents small or distant objects (*e.g.*, a car at 15m may occupy only a few dozen pixels). To address this, we condition the model on sparse object representations that encode semantics, pose, and coarse shape, thereby enabling the model to hallucinate fine structure, as shown in Fig. 3. For each node, we aggregate its features

$$\hat{\mathbf{h}}_{v_i} = \Phi_{\text{pos}}\big(\pi(\mathbf{b}_i)\big) + \Phi_{\text{cls}}(c_i) + \Phi_{\text{box}}(\mathbf{b}_i), \tag{4}$$

where $\pi(\mathbf{b}_i)$ is the 3D box projected to image coordinates, Φ_{pos} is a positional embedder, and $\Phi_{\text{cls}}, \Phi_{\text{box}}$ are learned MLPs. A lightweight self-attention layer diffuses contextual cues across tokens (Vaswani et al. 2017), producing the refined vector $\mathbf{h}_{v_i}$. The ego token is further compressed by an MLP to form a scene-level vector $\mathbf{h}_{\text{ego}}$.

During denoising step τ, the noisy range map $\mathbf{d}_\tau$ is concatenated with a sparse conditioning map $\mathbf{I}_{\text{cond}}$ as model input, which is obtained by projecting all layout points $\{\mathbf{p}_i\}_{i=0}^{M}$ onto the image plane. The global context is formed by summing the scene-level vector, a time embedding, and a CLIP embedding of the ego state:

$$\mathbf{h}_{\text{cond}} \;=\; \mathbf{h}_{\text{ego}} + \Phi_{\text{time}}(\tau) + \text{CLIP}(s_0). \tag{5}$$

A transformer-based U-Net then predicts the clean signal, progressively sharpening geometry and semantics.

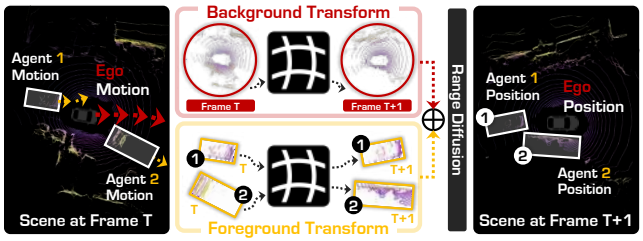


Figure 4: Details of the foreground and background warp.

Layout-Driven Scene Editing. As each frame is anchored by an explicit layout, we can edit objects without disturbing the static background, which is crucial for testing planners. After the original scene $\mathbf{d}_0^{\text{orig}}$ is synthesized, a user may alter the layout tuple. We then rerun the reverse diffusion, preserving pixels whose 2D projections remain unchanged following (Lugmayr et al. 2022). At every denoising step,

$$\mathbf{d}_{\tau-1} \;=\; (1-\mathbf{m}) \odot \tilde{\mathbf{d}}_{\tau-1} \;+\; \mathbf{m} \odot \hat{\mathbf{d}}_{\tau-1}, \tag{6}$$

where $\hat{\mathbf{d}}_{\tau-1}$ is the freshly denoised sample, $\tilde{\mathbf{d}}_{\tau-1} \sim \mathcal{N}\big(\sqrt{\bar{\alpha}}\,\mathbf{d}_0^{\text{orig}},\, (1-\bar{\alpha})\mathbb{I}\big)$ is a Gaussian-perturbed copy of the original scene, and the binary mask $\mathbf{m}$ marks pixels affected by the edited boxes. The blend locks untouched regions and resynthesizes only the modified objects, delivering instant, artifact-free edits for closed-loop simulation.

3.3 Scene2Seq: Autoregressive LiDAR Synthesis

A core innovation of **LiDARCrafter** is its ability to generate the LiDAR stream autoregressively. In RGB video, textures and lighting change constantly, whereas a LiDAR sweep sees a mostly static environment, with only the ego vehicle and annotated objects moving. We exploit this stability by warping previously observed points to create a strong prior, as shown in fig. 4. Concretely, we back-project the first range image $\mathbf{I}^0$ to a point cloud $\mathbf{P}^0$, then split it with the layout boxes into background $\mathbf{B}^0$ and foreground sets $\{\mathbf{F}_i^0\}_{i=1}^{M}$. In later frames, we warp $\mathbf{B}^0$ with the ego pose, and update each $\mathbf{F}_i^0$ by its own motion prior, providing a strong drift-free geometric prior for the diffusion model at every denoising step.

Static-Scene Warp. We update the background points with the ego pose. Taking frame 0 as the world origin, the ego translation at step t is $\mathbf{u}_0^t = [\Delta x_0^t, \Delta y_0^t, z_0]^\top$, with z_0 is the fixed sensor height, and its incremental yaw is

$$\psi_0^t = \text{atan2}(\Delta y_0^t - \Delta y_0^{t-1},\, \Delta x_0^t - \Delta x_0^{t-1}). \tag{7}$$

We form the homogeneous ego pose matrix $\mathbf{G}_0^t \in \text{SE}(3)$ with the rotation matrix $\mathbf{R}_z(\psi_0^t)$ and translation $\mathbf{u}_0^t$, and compute the relative motion $\Delta\mathbf{G}_0^t = \mathbf{G}_0^t(\mathbf{G}_0^{t-1})^{-1}$, then we propagate the static cloud via $\mathbf{B}^t = \Delta\mathbf{G}_0^t\mathbf{B}^{t-1}$.

Dynamic-Object Warp. For each object i, we first shift its box center by its own cumulative planar offsets $(\Delta x_i^t, \Delta y_i^t)$, giving the world-frame position $\mathbf{u}_i^t = [x_i + \Delta x_i^t,\; y_i + \Delta y_i^t,\; z_i]^\top$. Its heading change ψ_i^t is obtained exactly as in eq. (7). To express the box in the current ego frame, we apply the inverse ego transform: first translate by $-\mathbf{u}_0^t$ and then rotate by $-\psi_0^t$. The same rigid transform maps the stored foreground points $\mathbf{F}_i^0$ to $\mathbf{F}_i^t$. These warped foreground object points, combined with the updated background points, supply a strong geometric prior for the later timestep.

Table 1: Evaluations of **scene-level fidelity** for LiDAR generation on the *nuScenes* dataset. MMD values are reported in 10^{-4} and JSD in 10^{-2}. Lower is better for all metrics (↓).

Base	Method	Venue	Range		Points		BEV	
			FRD↓	MMD↓	FPD↓	MMD↓	JSD↓	MMD↓
Voxel	UniScene	CVPR'25	–	–	976.47	29.06	31.55	13.61
	OpenDWM	CVPR'25	–	–	714.19	21.95	20.17	5.61
	OpenDWM-DiT	CVPR'25	–	–	381.91	12.46	19.90	5.73
Range	LiDARGen	ECCV'22	759.65	1.71	159.35	35.52	5.74	2.39
	LiDM	CVPR'24	495.54	0.18	210.20	8.45	5.86	0.73
	RangeLDM	ECCV'24	–	–	–	–	5.47	1.92
	R2DM	ICRA'24	243.35	1.40	33.97	1.62	3.51	0.71
	LiDARCrafter	**Ours**	**194.37**	**0.08**	**8.64**	**0.90**	**3.11**	**0.42**

Table 2: Comparison of **foreground object quality** using FDC (↑), which reflects detector confidence on generated scenes. #Box is the average number of boxes per frame.

Base	Method	Venue	Car↑	Ped↑	Truck↑	Bus↑	#Box
Uncond.	LiDARGen	ECCV'22	0.57	0.29	0.42	0.38	0.364
	LiDM	CVPR'24	0.65	0.22	0.45	0.31	0.28
	R2DM	ICRA'24	0.54	0.29	0.39	0.35	0.53
Cond.	UniScene	CVPR'25	0.53	0.28	0.35	0.25	0.98
	OpenDWM	CVPR'25	0.74	0.30	0.51	0.44	0.54
	OpenDWM-DiT	CVPR'25	0.78	0.32	**0.56**	0.51	0.64
	LiDARCrafter	**Ours**	**0.83**	**0.34**	0.55	**0.54**	**1.84**

Autoregressive Generation. At every timestep $t > 0$ we build a condition range map by projecting and combining,

$$I^t_{\text{cond}} = \Pi\big(\mathbf{B}^{0\to t} \cup \mathbf{B}^{t-1\to t} \cup \{\mathbf{F}_i^{t-1\to t}\}_{i=1}^{M}\big), \qquad (8)$$

where $\Pi(\cdot)$ denotes spherical projection, and the superscript indicates the warp between two timestamps. Including the first frame background warp $\mathbf{B}^{0\to t}$ eliminates accumulated drift. We concatenate I^t_{cond} with the noisy sample and feed it into the diffusion backbone to generate the next range image, iterating until the whole sequence is synthesized.

3.4 EvalSuite: Temporal & Semantic Scoring

Existing LiDAR generation metrics like FRD judge only static realism. They ignore object semantics, layout validity, and motion coherence, which are essential for a controllable 4D world model. Our EvalSuite adds targeted scores for each facet. **Object metrics** (FDC, CDA, CFCA, CFSC) verify that generated foreground clouds carry the right labels, box geometry, and detector confidence. **Layout metrics** (SCR, MSCR, BCR, TCR) measure spatial and trajectory consistency while penalizing box or path collisions. **Temporal metrics** (TTCE, CTC) track frame-to-frame transform accuracy and sequence smoothness. Together, these metrics give a complete, 4D-aware assessment. More details are given in the supplementary material.

4 Experiments

4.1 Experimental Settings

We benchmark **LiDARCrafter** on the widely used nuScenes dataset (Caesar et al. 2020), whose 32-beam LiDAR sweeps, 3D boxes, instance IDs, and HD maps provide rich conditioning cues. Evaluation combines classic static scores (FRD, FPD, JSD, MMD) with our object-, layout-, and motion-centric metrics from section 3.4. Training uses six NVIDIA A40s, the three-branch layout diffuser runs $1M$ steps at batch 64, and the range-image model (32×1024) runs $500K$ steps at batch 32. All diffusers adopt 1024 denoise steps for training and 256 for sampling. More experimental settings appear in the supplementary material.

Table 3: Comparisons of **foreground object perception** accuracy using CDA (↑), which measures 3D detection average precision (AP) on generated scenes.

Method	Venue	$\mathbf{AP}^{\text{R11}}_{\text{BEV}}$	$\mathbf{AP}^{\text{R11}}_{\text{3D}}$	$\mathbf{AP}^{\text{R40}}_{\text{BEV}}$	$\mathbf{AP}^{\text{R40}}_{\text{3D}}$
UniScene	CVPR'25	0.19	0	0.02	0
OpenDWM	CVPR'25	17.07	9.09	11.84	1.03
OpenDWM-DiT	CVPR'25	16.37	11.27	10.62	1.89
LiDARCrafter	**Ours**	**23.21**	**15.24**	**18.27**	**8.26**

Table 4: Evaluation of **object-level fidelity** for LiDAR generation. MMD is reported in 10^{-4}, and JSD in 10^{-2}.

Base	Method	Venue	FPD↓	P-MMD↓	JSD↓	MMD↓
Uncond.	LiDARGen	ECCV'22	1.39	0.15	0.20	16.22
	LiDM	CVPR'24	1.41	0.15	0.19	13.49
	R2DM	ICRA'24	1.40	0.15	0.17	12.76
Cond.	UniScene	CVPR'25	1.19	0.18	0.23	16.65
	OpenDWM	CVPR'25	1.49	0.19	0.16	9.11
	OpenDWM-DiT	CVPR'25	1.48	0.18	**0.15**	9.02
	LiDARCrafter	**Ours**	**1.03**	**0.13**	**0.15**	**5.48**

4.2 Scene-Level LiDAR Generation

This section evaluates the overall quality of generated LiDAR scenes, focusing on both *whole-scene fidelity* and the accuracy of synthesized *foreground objects*.

Whole-Scene Fidelity. Table 1 compares **LiDARCrafter** against recent LiDAR generation baselines in terms of whole-scene fidelity. Our method outperforms all existing methods, achieving the lowest FRD (194.37, a 20% improvement over R2DM) and FPD (8.64), indicating more accurate reconstruction of both 2D range structures and 3D point distributions. As shown in Figure 5, our method produces scans that closely resemble the ground truth, with well-preserved foreground structures, while other methods suffer from background noise or blurring.

Foreground Object Accuracy. To evaluate the fidelity of synthesized foregrounds, we apply a pre-trained VoxelRCNN detector (Deng et al. 2021) to the generated scenes. We report Foreground Detection Confidence (FDC), which reflects the detector's confidence (Table 2), and Conditioned Detection Accuracy (CDA), which measures the average precision (AP) of detected boxes (Table 3). LiDARCrafter achieves the highest FDC scores across most categories, indicating superior object fidelity, and also attains the best detection AP, demonstrating stronger alignment between synthesized structures and the given conditions.

4.3 Object-Level LiDAR Generation

This section evaluates the quality of individual object generation, focusing on both *fidelity* and *semantic and geometric consistency* under box-level conditioning.

Object-Wise Fidelity. To assess instance-level fidelity, we extract 2,000 *Car* objects from each method and compute object-level metrics (Table 4). LiDARCrafter achieves the lowest FPD (1.03) and MMD (5.48), significantly outper-

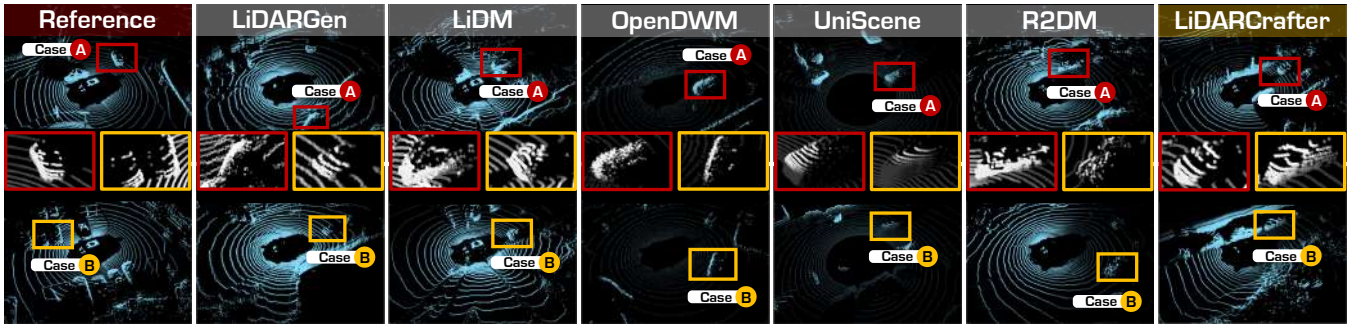


Figure 5: Single-frame LiDAR point cloud generation results. LiDARCrafter produces the pattern closest to the ground truth, with notably superior foreground quality compared to other methods. Best viewed at high resolution.

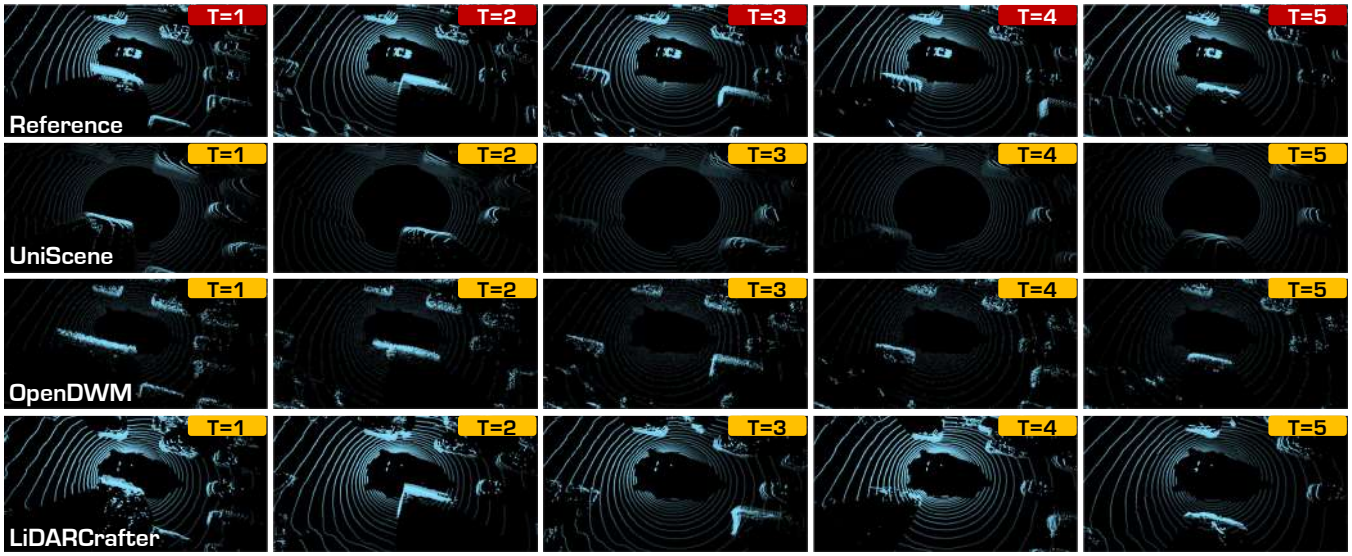


Figure 6: Sequence point cloud generation results. LiDARCrafter maintains temporal consistency while producing patterns closest to the ground truth. Frames are arranged in temporal order from left to right. Best viewed at high resolution.

Table 5: Comparison of **object generation consistency** using CFCA (↑) and CFSC (↑). CFCA measures classification accuracy on generated points using a PointMLP trained on real data. CFSC assesses geometric consistency by regressing boxes from generated points and computing IoU; the number indicates the point count within each box.

Method	Venue	CFCA↑	CFSC↑		
			< 150	150–300	> 300
Original	–	92.49	0.50	0.61	0.72
UniScene	CVPR'25	34.25	0.14	0.17	0.23
OpenDWM	CVPR'25	62.35	0.17	0.21	0.26
OpenDWM-DiT	CVPR'25	70.65	0.31	0.32	0.34
LiDARCrafter	**Ours**	**73.45**	**0.35**	**0.36**	**0.42**

forming OpenDWM and demonstrating better reconstruction of fine-grained geometry.

Semantic and Geometric Consistency. To further evaluate object quality under conditioning, we introduce two metrics: CFCA for semantic fidelity and CFSC for geometric consistency (Table 5). For *semantic fidelity*, we apply a PointMLP (Ma et al. 2022) classifier trained on real data to classify generated instances, yielding a CFCA score of 73.48% for LiDARCrafter, indicating strong alignment with real-world categories. For *geometric consistency*, we use a conditional variational autoencoder to regress bounding boxes from generated point clouds, and compute the mean IoU with ground truth. LiDARCrafter achieves the highest IoU across all point count settings, demonstrating superior adherence to geometric constraints.

Table 6: Comparison of **temporal consistency** in 4D LiDAR generation. TTCE measures transformation error from point cloud registration; CTC computes Chamfer Distance between adjacent frames. Numbers indicate frame intervals.

Method	Venue	TTCE↓		CTC↓			
		3	4	1	2	3	4
UniScene	CVPR'25	2.74	3.69	0.90	1.84	3.64	**3.90**
OpenDWM	CVPR'25	2.68	3.65	1.02	2.02	3.37	5.05
OpenDWM-DiT	CVPR'25	2.71	3.66	**0.89**	**1.79**	3.06	4.64
LiDARCrafter	**Ours**	**2.65**	**3.56**	1.12	2.38	**3.02**	4.81

4.4 Autoregressive 4D LiDAR Generation

Temporal Consistency. We evaluate temporal consistency in 4D LiDAR generation in Table 6. TTCE measures the error between the predicted and ground-truth transformation matrices obtained via point cloud registration, while CTC computes the Chamfer Distance between consecutive frames. LiDARCrafter achieves the lowest TTCE scores

Table 7: Ablation on **foreground conditioning methods** for LiDAR generation. 2D masks are projected from 3D boxes.

No.	Type	Variant	Scene		Object		
			FRD↓	FPD↓	FPD↓	CFCA↑	CFSC↑
1	Baseline	–	243.35	33.97	1.40	–	–
2	Dense	*w/* 2D mask	237.17	33.21	1.35	61.22	0.24
3		*w/* Obj mask	217.83	24.02	1.20	64.54	0.27
4	Dense+Sparse	*w/* E_{pos}	205.27	15.97	1.08	72.46	0.40
5		*w/* $E_{pos} + E_{cls}$	**193.27**	10.52	1.05	**75.27**	0.40
6		*w/* All	194.37	**8.64**	**1.03**	73.45	**0.42**

Table 8: Ablation on **generation paradigm** (E2E vs. AR) and **historical conditioning** for 4D LiDAR generation.

No.	Type	Intensity	Depth	TTCE↓		CTC↓		Scene	
				3	4	3	4	FRD↓	FPD↓
1	E2E	-	-	3.21	4.36	5.68	7.41	477.21	182.36
2	AR	-	-	3.31	4.84	4.31	6.21	311.27	90.10
3		✓	✓	2.96	3.87	3.24	4.85	254.39	22.20
4		✓	✗	3.21	4.21	3.42	5.19	364.27	154.21
5		✗	✓	**2.65**	**3.56**	**3.02**	**4.81**	**194.37**	**8.64**

across both frame intervals and maintains competitive CTC performance at all intervals, demonstrating strong temporal coherence. Qualitative comparisons in Figure 6 further show that LiDARCrafter produces sequences with consistent structure and fine geometric detail, whereas other methods often suffer from degraded fidelity over time.

4.5 Ablation Study

We conduct ablations on **foreground generation** (necessity and conditioning) and **4D consistency** (generation paradigm and historical priors) to validate key designs.

Ablation on the necessity of foreground generation. Table 7 shows that introducing a 2D foreground mask projected from 3D boxes *(No.2)* notably improves scene generation, particularly for foreground objects. Further incorporating the foreground generation branch *(No.3)* that produces fine-grained object masks leads to a lower FPD, showing the benefit of detailed geometric and depth supervision.

Ablation on foreground conditioning mechanism. Foreground objects are inherently sparse, often occupying only a few pixels, making dense mask-only conditioning insufficient. As shown in Table 7, our proposed **sparse conditioning modules** are crucial: embedding 2D box features alone *(No.4)* reduces FRD, while adding semantic and geometric attributes *(No.5 and No.6)* yields the best FPD and further improves FRD. These results underscore the benefits of richer, object-centric conditioning.

Ablation on generation paradigm in 4D generation. Unlike RGB videos, where appearance varies due to lighting and texture changes, LiDAR sequences capture largely static environments, with dynamics introduced only by ego-motion and moving agents. We exploit this stability by warping previously observed points using ego and object trajectories, providing strong priors for **autoregressive generation**. As shown in Table 8, our inpainting-based autoregressive framework *(No.2)* outperforms the end-to-end baseline *(No.1)* on temporal metrics, demonstrating that the autoregressive design naturally aligns with the relatively static nature and limited temporal variation of LiDAR sequences.

Ablation on historical conditioning in 4D generation. Table 8 investigates the impact of different historical priors on

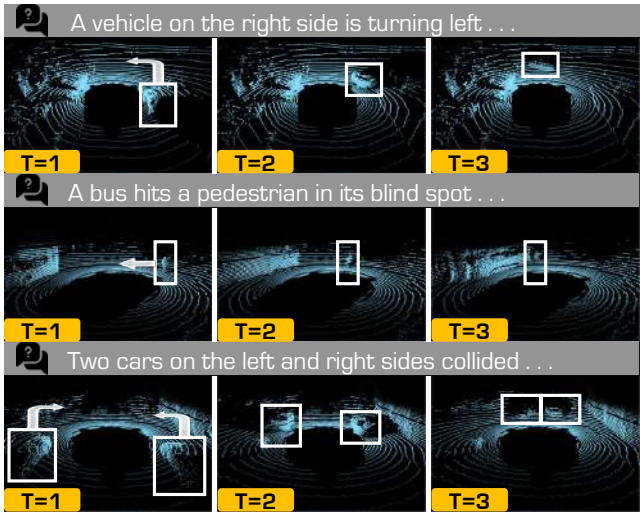


Figure 7: **Diverse corner cases** generated by LiDARCrafter with object-centric controllability. Best viewed at high resolution. Frames are arranged sequentially from left to right.

4D LiDAR generation. Using both depth and intensity features as conditioning inputs *(No.3)* significantly improves performance over the baseline without historical guidance *(No.2)*. Notably, excluding the depth prior *(No.4)* leads to substantial error accumulation (FRD increases by 109.88 compared to *No.3*), while using depth alone *(No.5)* achieves the best FRD. These results indicate that depth cues are more reliable and crucial for maintaining temporal consistency, whereas intensity features are harder to model effectively.

4.6 Applications

Leveraging its object-centric generation capability, **LiDARCrafter** enables the synthesis of rare and diverse corner-case scenarios, which are valuable for data augmentation and for evaluating the robustness of downstream algorithms. As illustrated in Figure 7, our method can generate challenging situations such as a vehicle cutting into the driving lane, a pedestrian emerging in a bus blind spot, a collision between two adjacent cars, and an overtaking maneuver. These results demonstrate that our method not only produces the desired scenarios on demand, but also maintains strong temporal coherence throughout the sequence. Additional qualitative examples are provided in the supplementary material.

5 Conclusion

We presented **LiDARCrafter**, a unified framework for controllable 4D LiDAR sequence generation and editing. By leveraging scene graph descriptors, the multi-branch diffusion model, and an autoregressive generation strategy, our approach achieves fine-grained controllability and strong temporal consistency. Experiments on nuScenes demonstrate clear improvements over existing methods in fidelity, coherence, and controllability. Beyond high-quality data synthesis, LiDARCrafter enables the creation of safety-critical scenarios for robust evaluation of downstream autonomous driving systems. Future work will explore multi-modal extensions and further efficiency improvements.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bian, H.; Kong, L.; Xie, H.; Pan, L.; Qiao, Y.; and Liu, Z. 2025. DynamicCity: Large-Scale 4D Occupancy Generation from Dynamic Scenes. In *International Conference on Learning Representations*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; and Li, H. 2021. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI Conference on Artificial Intelligence*, 1201–1209.

Feng, W.; Zhu, W.; Fu, T.-j.; Jampani, V.; Akula, A.; He, X.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36: 18225–18250.

Gao, R.; Chen, K.; Xie, E.; Hong, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2023. MagicDrive: Street View Generation with Diverse 3D Geometry Control. In *International Conference on Learning Representations*.

Guo, X.; Wu, Z.; Xiong, K.; Xu, Z.; Zhou, L.; Xu, G.; Xu, S.; Sun, H.; Wang, B.; Chen, G.; et al. 2025. Genesis: Multimodal Driving Scene Generation with Spatio-Temporal and Cross-Modal Consistency. *arXiv preprint arXiv:2506.07497*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; and Corrado, G. 2023. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*.

Hu, Q.; Zhang, Z.; and Hu, W. 2024. RangeLDM: Fast realistic LiDAR point cloud generation. In *European Conference on Computer Vision*, 115–135. Springer.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1219–1228.

Li, B.; Guo, J.; Liu, H.; Zou, Y.; Ding, Y.; Chen, X.; Zhu, H.; Tan, F.; Zhang, C.; Wang, T.; et al. 2025. UniScene: Unified occupancy-centric driving scene generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11971–11981.

Liu, T.; Zhao, S.; and Rhinehart, N. 2025. Towards foundational LiDAR world models with efficient latent flow matching. *arXiv preprint arXiv:2506.23434*.

Liu, Y.; Li, X.; Zhang, Y.; Qi, L.; Li, X.; Wang, W.; Li, C.; Li, X.; and Yang, M.-H. 2025. Controllable 3D outdoor scene generation via scene graphs. *arXiv preprint arXiv:2503.07152*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Ma, X.; Qin, C.; You, H.; Ran, H.; and Fu, Y. 2022. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. *arXiv preprint arXiv:2202.07123*.

Mei, J.; Hu, T.; Yang, X.; Wen, L.; Yang, Y.; Wei, T.; Ma, Y.; Dou, M.; Shi, B.; and Liu, Y. 2024. Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003*.

Nakashima, K.; and Kurazume, R. 2024. Lidar data synthesis with denoising diffusion probabilistic models. In *IEEE International Conference on Robotics and Automation*, 14724–14731. IEEE.

Nakashima, K.; Liu, X.; Miyawaki, T.; Iwashita, Y.; and Kurazume, R. 2024. Fast LiDAR Data Generation with Rectified Flows. *arXiv preprint arXiv:2412.02241*.

Ni, J.; Guo, Y.; Liu, Y.; Chen, R.; Lu, L.; and Wu, Z. 2025a. MaskGWM: A Generalizable Driving World Model with Video Mask Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22381–22391.

Ni, J.; Guo, Y.; Liu, Y.; Chen, R.; Lu, L.; and Wu, Z. 2025b. OpenDWM: Open Driving World Models. Https://github.com/SenseTime-FVG/OpenDWM.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PmLR.

Ran, H.; Guizilini, V.; and Wang, Y. 2024. Towards realistic scene generation with LiDAR diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14738–14748.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241. Springer.

Swerdlow, A.; Xu, R.; and Zhou, B. 2024. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 9(4): 3578–3585.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L.; Zheng, W.; Ren, Y.; Jiang, H.; Cui, Z.; Yu, H.; and Lu, J. 2024a. OccSora: 4D occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*.

Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; Zhu, J.; and Lu, J. 2024b. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, 55–72. Springer.

Wu, Y.; Zhang, K.; Qian, J.; Xie, J.; and Yang, J. 2024. Text2LiDAR: Text-guided LiDAR point cloud generation via equirectangular transformer. In *European Conference on Computer Vision*, 291–310. Springer.

Wu, Y.; Zhu, Y.; Zhang, K.; Qian, J.; Xie, J.; and Yang, J. 2025. WeatherGen: A Unified Diverse Weather Generator for LiDAR Point Clouds via Spider Mamba Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17019–17028.

Xie, Y.; Xu, C.; Peng, C.; Zhao, S.; Ho, N.; Pham, A. T.; Ding, M.; Tomizuka, M.; and Zhan, W. 2024. X-Drive: Cross-modality consistent multi-sensor data synthesis for driving scenarios. *arXiv preprint arXiv:2411.01123*.

Xiong, Y.; Ma, W.-C.; Wang, J.; and Urtasun, R. 2023. UltraLiDAR: Learning compact representations for LiDAR completion and generation. *arXiv preprint arXiv:2311.01448*.

Yan, T.; Yin, J.; Lang, X.; Yang, R.; Xu, C.-Z.; and Shen, J. 2025. OLiDM: Object-aware LiDAR Diffusion Models for Autonomous Driving. In *AAAI Conference on Artificial Intelligence*, 9121–9129.

Yang, X.; Wen, L.; Ma, Y.; Mei, J.; Li, X.; Wei, T.; Lei, W.; Fu, D.; Cai, P.; Dou, M.; et al. 2024. Drivearena: A closed-loop generative simulation platform for autonomous driving. *arXiv preprint arXiv:2408.00415*.

Yang, Y.; Liang, A.; Mei, J.; Ma, Y.; Liu, Y.; and Lee, G. H. 2025a. X-Scene: Large-Scale Driving Scene Generation with High Fidelity and Flexible Controllability. *arXiv preprint arXiv:2506.13558*.

Yang, Z.; Lu, K.; Zhang, C.; Qi, J.; Jiang, H.; Ma, R.; Yin, S.; Xu, Y.; Xing, M.; Xiao, Z.; et al. 2025b. Mmgdreamer: Mixed-modality graph for geometry-controllable 3d indoor scene generation. In *AAAI Conference on Artificial Intelligence*, 9391–9399.

Zhai, G.; Örnek, E. P.; Chen, D. Z.; Liao, R.; Di, Y.; Navab, N.; Tombari, F.; and Busam, B. 2024. Echoscene: Indoor scene generation via information echo over scene graph diffusion. In *European Conference on Computer Vision*, 167–184. Springer.

Zhang, J.; Sheng, H.; Cai, S.; Deng, B.; Liang, Q.; Li, W.; Fu, Y.; Ye, J.; and Gu, S. 2024. PerlDiff: Controllable street view synthesis using perspective-layout diffusion models. *arXiv preprint arXiv:2407.06109*.

Zhang, K.; Tang, Z.; Hu, X.; Pan, X.; Guo, X.; Liu, Y.; Huang, J.; Yuan, L.; Zhang, Q.; Long, X.-X.; et al. 2025. Epona: Autoregressive Diffusion World Model for Autonomous Driving. *arXiv preprint arXiv:2506.24113*.

Zhang, L.; Xiong, Y.; Yang, Z.; Casas, S.; Hu, R.; and Urtasun, R. 2023. Copilot4D: Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*.

Zhao, G.; Wang, X.; Zhu, Z.; Chen, X.; Huang, G.; Bao, X.; and Wang, X. 2025. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *AAAI Conference on Artificial Intelligence*, 10412–10420.

Zheng, W.; Chen, W.; Huang, Y.; Zhang, B.; Duan, Y.; and Lu, J. 2024a. OccWorld: Learning a 3D occupancy world model for autonomous driving. In *European Conference on Computer Vision*, 55–72. Springer.

Zheng, X.; Huang, X.; Mei, G.; Hou, Y.; Lyu, Z.; Dai, B.; Ouyang, W.; and Gong, Y. 2024b. Point cloud pre-training with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22935–22945.

Zuo, S.; Zheng, W.; Han, X.; Yang, L.; Pan, Y.; and Lu, J. 2025a. QuadricFormer: Scene as Superquadrics for 3D Semantic Occupancy Prediction. *arXiv preprint arXiv:2506.10977*.

Zuo, S.; Zheng, W.; Huang, Y.; Zhou, J.; and Lu, J. 2025b. GaussianWorld: Gaussian world model for streaming 3D occupancy prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6772–6781.

Zyrianov, V.; Zhu, X.; and Wang, S. 2022. Learning to generate realistic LiDAR point clouds. In *European Conference on Computer Vision*, 17–35. Springer.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no)

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no)

2.4. Proofs of all novel claims are included (yes/partial/no)

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no)

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no)

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA)

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA)

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-

dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 倪雨嫣

方向： 人工智能

单位： 中国科学院数学与系统科学研究院

题目： 基于直线扩散模型的高效 3D 分子生成方法 Straight-Line Diffusion Model for Efficient 3D Molecular Generation

Straight-Line Diffusion Model for Efficient 3D Molecular Generation

Yuyan Ni[1,4 * †] **Shikun Feng**[2 *] **Haohan Chi**[3] **Bowen Zheng**[5] **Huan-ang Gao**[3]
Wei-Ying Ma[3] **Zhi-Ming Ma**[1] **Yanyan Lan**[3,6,7 ‡]
[1] Academy of Mathematics and Systems Science, Chinese Academy of Sciences
[2] Zhongguancun Institute of Artificial Intelligence, China
[3] Institute for AI Industry Research (AIR), Tsinghua University
[4] University of Chinese Academy of Sciences
[5]Huazhong University of Science and Technology
[6]Beijing Frontier Research Center for Biological Structure, Tsinghua University
[7]Beijing Academy of Artificial Intelligence

Abstract

Diffusion-based models have shown great promise in molecular generation but often require a large number of sampling steps to generate valid samples. In this paper, we introduce a novel Straight-Line Diffusion Model (SLDM) to tackle this problem, by formulating the diffusion process to follow a linear trajectory. The proposed process aligns well with the noise sensitivity characteristic of molecular structures and uniformly distributes reconstruction effort across the generative process, thus enhancing learning efficiency and efficacy. Consequently, SLDM achieves state-of-the-art performance on 3D molecule generation benchmarks, delivering a 100-fold improvement in sampling efficiency.[1]

1 Introduction

3D molecular generation is an essential task in drug discovery, material science, and molecular engineering. The goal is to computationally design 3D molecular structures that not only capture intricate physical and chemical constraints but also fulfill specific properties.

Recently, diffusion models have been widely applied in this field, inspired by their remarkable success in image synthesis [Dhariwal and Nichol, 2021, Rombach et al., 2022, Peebles and Xie, 2023], and other domains [Brooks et al., 2024, Abramson et al., 2024]. Methods like EDM [Hoogeboom et al., 2022], EDM-Bridge [Wu et al., 2022] and GeoLDM [Xu et al., 2023] have demonstrated the potential of diffusion-based frameworks to generate chemically valid 3D molecular structures. However, these direct applications of diffusion methods usually require a large number of sampling steps to produce valid molecules.

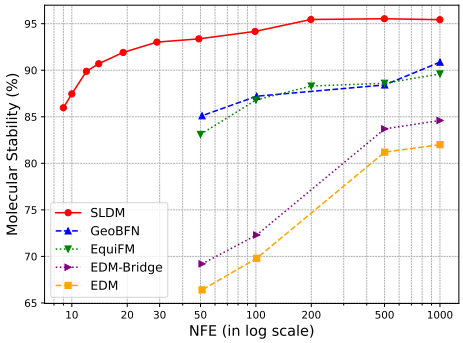


Figure 1: Comparison of molecule stability (↑) across diffusion-based molecular generation models on QM9 unconditional generation, evaluated by the number of function evaluations (NFE) during the sampling process.

*Equal contribution. † Work was done during Yuyan's internship at AIR. ‡ Corresponding author: Yanyan Lan (lanyanyan@air.tsinghua.edu.cn).

[1]The code is open-sourced at https://github.com/fengshikun/SLDM

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Taking EDM as an example, it requires approximately 1000 steps of function evaluations to generate molecules with around 82% stability, which is a key metric for assessing sample quality by quantitatively measuring whether the molecule satisfies chemical constraints. Reducing the sampling steps significantly degrades the molecule stability.

To improve sampling efficiency, EquiFM [Song et al., 2024] and GeoBFN [Song et al., 2023a] have been proposed to utilize the Flow Matching (FM) framework [Tong et al., 2023] and Bayesian Flow Networks (BFN) [Graves et al., 2023] for molecular generation. The use of these advanced generative AI models enables a speedup of 5× and 20×, respectively, compared to EDM. However, they still require a large number of sampling steps (e.g. 1000) to achieve high molecule stability (e.g. 90%), as shown in Figure 1.

To understand why existing methods suffer from low efficiency, we analyze the issue through the lens of truncation error in sampling. We begin by establishing a unified perspective on previous diffusion-based methods, including diffusion models, flow matching methods, and Bayesian flow networks. Specifically, their noise corrupting process can be generalized as $\boldsymbol{x}_t = \mu(t)\boldsymbol{x}_0 + \sigma(t)\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$, where $\boldsymbol{x}_0$ represents the clean data, $\boldsymbol{x}_t$ is the noise corrupted data at time $t \in [0, T]$, $\mu(t)$ and $\sigma(t)$ define the schedule of the process. In this framework, all these processes can be equivalently framed as continuous-time Ordinary Differential Equations (ODEs), even though they may employ stochastic sampling in practice. This viewpoint allows the sampling process to be interpreted as a numerical approximation of the solution trajectory of the underlying ODE. Crucially, most existing methods rely on first-order estimation, whose truncation error is governed by the second-order term $\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2}(\Delta t)^2$. We observe that $\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2}$ in these approaches can be large, requiring small step sizes Δt to reduce the truncation error, which results in a large number of sampling steps.

To address this issue, we propose a novel diffusion process called Straight-Line Diffusion Model (SLDM). The key idea is to minimize truncation error by striving to achieve a linear sampling trajectory, i.e. $\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2} = 0$. This approach allows the diffusion dynamics to tolerate larger step sizes without sacrificing accuracy, leading to a substantial improvement in sampling efficiency. Building on this objective, we theoretically prove that when $\mu(t) = 1 - t/T$ and σ is a small constant, the process guarantees a near-linear trajectory. Intuitively, this process features a linearly decreasing mean term and a consistently small variance term, representing a smooth linear progression from the origin point to the data.

Notably, SLDM strikes a good balance between efficiency and efficacy by using our straight-line schedule. Firstly, this strategy aligns well with the inductive bias of molecular generation, preventing the introduction of chemically implausible conformations. Unlike images, molecular structures are much more sensitive to noise, and even small perturbations can lead to unrealistic structures that violate chemical principles. This challenge requires a slower signal-to-noise ratio (SNR) decay during the noise-adding process, as suggested in Song et al. [2023a]. By naturally satisfying a slower SNR decay, our method maintains chemical information of the intermediate states, enhancing computational efficiency compared to traditional methods such as EDM. Secondly, our strategy achieves a more balanced generative process, significantly improving the model's learning efficacy. In methods like GeoBFN, minimal perturbations are applied in the later stages, shifting most of the reconstruction burden to earlier stages. Although this reduces the computational load in the later stages, it creates an uneven distribution of effort, limiting the model's learning capacity. In contrast, our approach evenly distributes the reconstruction effort across the entire process, enabling the model to learn effectively at each stage. This balance results in a more stable and efficient learning process, enhancing the robustness and accuracy of the generated molecular structures.

We conduct extensive experiments to demonstrate the potential of straight-line diffusion in 3D molecular generation and other domains. As shown in Figure 1, using only at most 10 or 15 sampling steps, SLDM surpasses EDM or EquiFM, GeoBFN with 1000 sampling steps, achieving up to **100- or 70-fold** improvement in sampling efficiency. In terms of generation quality, SLDM with 200 sampling steps achieves 95% molecular stability, significantly outperforming the best baseline, GeoBFN, which requires 1000 steps to reach 90% molecular stability. We also observe that similar improvements can be achieved when applying SLDM to the conditional generation task, i.e. generating molecules with a desired property, highlighting its potential to enable more practical and controllable molecular design in future applications.

2 Analysis on Sampling Efficiency

We begin by theoretically analyzing the underlying factors contributing to the sampling efficiency issue. In particular, we first present a unified framework for diffusion-based methods, including their SDE and ODE formulations. We then examine the sampling truncation error from the ODE perspective, highlighting the critical role of the process's second-order derivative in improving sampling efficiency.

We denote the data as $\boldsymbol{x} \in \mathbb{R}^N$, where for molecules, it refers to a 3D point cloud comprising atomic coordinates and potentially other atomic features. According to Karras et al. [2022], Xue et al. [2024a], various diffusion-based models, including DDPM [Ho et al., 2020], DDIM [Song et al., 2021a], VE [Song et al., 2021b], FM [Lipman et al., 2023], and BFN [Graves et al., 2023], can be formulated as a unified form with the noise corrupting process defined as:

$$\boldsymbol{x}_t = \mu(t)\boldsymbol{x}_0 + \sigma(t)\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N), t \in [0, T] \tag{1}$$

where $\boldsymbol{x}_0, \boldsymbol{x}_t$ are clean data and noise corrupted data respectively, $\mu(t)$ and $\sigma(t)$ define the schedule of the process. Specifically, as detailed in Appendix A.2, the schedule parameters are summarized as follows: $\mu_{\text{DDPM(EDM)}} = 1-(t/T)^2$, $\mu_{\text{VE}} = \mu_{\text{DDIM}} = 1$, $\mu_{\text{BFN}} = 1 - \sigma_{min}^{2(1-t/T)}$, $\mu_{\text{FM}} = 1-t/T$; $\sigma_{\text{DDPM(EDM)}} = \sqrt{1-(1-(t/T)^2)^2}$, $\sigma_{\text{VE}} = \sqrt{t}$, $\sigma_{\text{DDIM}} = t$, $\sigma_{\text{BFN}} = \sqrt{\mu_{\text{BFN}}(1-\mu_{\text{BFN}})}$, $\sigma_{\text{FM}} = t/T + (1-t/T)\sigma_{min}$, where DDPM(EDM) uses the approximated DDPM schedule given in EDM [Hoogeboom et al., 2022]. σ_{min} are defined as small constants to ensure $\mu(0) \approx 1$ and $\sigma(0) \approx 0$. T is typically chosen to be sufficiently large so that $\boldsymbol{x}_T$ approximates a known distribution.

Extending a similar theoretical technique from Karras et al. [2022] to the unified form, we can prove that equation 1 is the solution to the following linear stochastic differential equation (SDE):

$$\mathrm{d}\boldsymbol{x}_t = \frac{\dot{\mu}(t)}{\mu(t)}\boldsymbol{x}_t \mathrm{d}t + \sqrt{2\sigma(t)\dot{\sigma}(t) - 2\sigma(t)^2\frac{\dot{\mu}(t)}{\mu(t)}}\mathrm{d}w_t, \tag{2}$$

where $\mu(t)$ and $\sigma(t)$ should adhere to the constraints $\mu(0) = 1$, $\sigma(0) = 0$, with $\mu(t) \geq 0$ being monotone non-increasing, $\sigma(t) \geq 0$, and $\sigma(t)/\mu(t)$ being monotone non-decreasing. The detailed proof is elaborated in Appendix A.1.

According to the ODE and SDE relations revealed in Song et al. [2021b], we can derive the equivalent ODE that follows the same marginal probability densities as the above SDE:

$$\mathrm{d}\boldsymbol{x}_t = \left[\frac{\dot{\mu}(t)}{\mu(t)}\boldsymbol{x}_t - \left(\sigma(t)\dot{\sigma}(t) - \sigma(t)^2\frac{\dot{\mu}(t)}{\mu(t)}\right)\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t)\right]\mathrm{d}t. \tag{3}$$

Consequently, most existing diffusion-based methods widely used in 3D molecular generation, including DDPM, DDIM, VE, FM, and BFN, can be interpreted from a continuous-time perspective, where their diffusion process can equivalently be viewed as an SDE in equation 2 or an ODE in equation 3. Therefore, we can use the ODE formulation as a valuable perspective to analyze the sampling efficiency issue.

Specifically, the diffusion sampling process can be interpreted as a numerical approximation of the backward solution trajectory of the underlying ODE in equation 3 [2]. This numerical approximation inherently introduces truncation errors at each step. For example, under Euler's method, the simplest and widely-used numerical scheme, $x(t-\Delta t)$ is approximated by $x(t) - \frac{\mathrm{d}x(t)}{\mathrm{d}t}\Delta t$ when solving the ODE backward in time. However, the true value can be derived from the Taylor expansion:

$$x(t-\Delta t) = x(t) - \frac{\mathrm{d}x(t)}{\mathrm{d}t}\Delta t + \frac{1}{2}\frac{\mathrm{d}^2x(t)}{\mathrm{d}t^2}(\Delta t)^2 + O((\Delta t)^3), \tag{4}$$

where Δt is a small step size. Thus, the truncation error primarily arises from the second-order term, governed by $\frac{\mathrm{d}^2x(t)}{\mathrm{d}t^2}$.

To minimize this truncation error, traditional numerical analysis primarily focuses on developing higher-order solvers to estimate higher-order terms, assuming the ODE structure is fixed. EquiFM

[2] Connections between sampling of baseline methods and first-order ODE sampling are discussed in appendix A.3.

[Song et al., 2024] applied such a technique in 3D molecular generation to speedup the sampling process. However, higher-order solvers face a trade-off between the number of function evaluations (NFE) and accuracy, limiting their ability to achieve even smaller NFEs. As a result, the issue remains partially unsolved. In contrast, we take a different approach: rather than focusing on the sampling algorithm, we reformulate the diffusion process itself to minimize the truncation error, which directly influences the sampling efficiency.

However, our unified formulation of diffusion introduces a key flexibility: the ability to modify the schedule of the process, thereby altering the ODE structure itself to reduce these errors. Therefore, we emphasize that the key to minimizing the truncation error is to reduce the second-order derivative of the process.

3 Straight-Line Diffusion Process

To reduce the truncation error of the sampling process, we aim to design a new diffusion process whose inherent ODE exhibits minimal $\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2}$. By achieving this, even a basic Euler iteration can deliver low truncation error without resorting to complex solvers that require multiple function evaluations. This novel perspective advances the Pareto frontier of efficiency and accuracy in diffusion sampling.

3.1 Derivation of SLDM

Given the intractability of the score function for general data distributions, we begin by examining a simplified case where the initial distribution is a delta distribution. This setup provides a tractable backward ODE, enabling a more straightforward analysis. In this case, the only solution that ensures $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}$ remains constant is for $\sigma(t)$ to be constant and $\mu(t)$ to be a linear function of t. Details can be found in Appendix A.4.

Given the boundary condition that x_0 approximates data distribution and $x_{t=1}$ approximates a known distribution, the only feasible choice is to set $\sigma(t)$ to a small poistive constant and $\mu(t) = 1 - t/T$, which satisfies all the schedule constraints noted under equation 2, except $\sigma(0) = 0$. This is because setting $\sigma(0) = 0$ exactly would result in a trivial denoising objective, To avoid this, we instead approximate $\sigma(0)$ by a small value, which keeps the denoising task non-trivial while still satisfying the boundary condition approximately. In practice, we use a value for σ that is two orders of magnitude smaller than the data scale, yielding good results.

Rescaling time to the interval $[0, 1]$, the resulting diffusion process is:

$$\boldsymbol{x}_t = (1 - t)\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}, t \in [0, 1], \tag{5}$$

which ensures a straight-line trajectory under the delta data distribution assumption.

We then analyze the above diffusion process for general data distribution and demonstrate that a near-linear trajectory could be achieved by setting a small constant value for σ, as shown in the following theorem.

Theorem 3.1 (Near-linear Trajectory of SLDM)**.** *For a general data distribution and the schedule in equation 5, the following inequality holds for each dimension i:*

$$P(|\frac{\mathrm{d}\boldsymbol{x}_t^{(i)}}{\mathrm{d}t} + \frac{\boldsymbol{x}_t^{(i)}}{1-t}| \geq \delta) \leq \frac{\sigma^2}{\delta^2(1-t)^2}. \tag{6}$$

When $\sigma \to 0$, $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} + \frac{\boldsymbol{x}_t}{1-t}$ converges to zero in probability, where the solution to equation $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} + \frac{\boldsymbol{x}_t}{1-t} = 0$ is that $\frac{\boldsymbol{x}_t}{1-t}$ is constant, which corresponds to a linear trajectory.

In other words, for $t \in [0, 1 - \frac{\sigma}{\delta \cdot \epsilon}]$, $|\frac{\mathrm{d}\boldsymbol{x}_t^{(i)}}{\mathrm{d}t} + \frac{\boldsymbol{x}_t^{(i)}}{1-t}| < \delta$ holds with probability at least $1 - \epsilon^2$, indicating that setting σ to a small value ensures that the trajectory remains close to a straight line for most timesteps. This aligns with our empirical observations shown in Figure 6: the overall SLDM trajectory is straighter compared to other diffusion processes, and the sampling trajectory deviates more from a straight line in the initial steps but becomes increasingly linear later.

To further improve the sampling quality, we aim to mitigate the impact of early-stage truncation errors. This is achieved by our stochastic sampling design, which exhibits robustness to such early-stage errors, as supported by empirical evidence provided in Appendix C.6. As a result, despite the

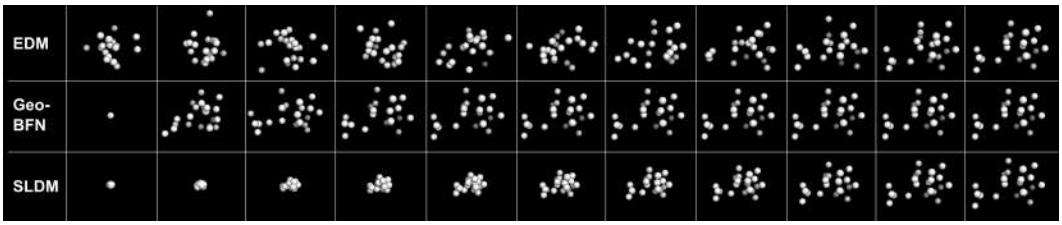


Figure 2: The diffusion process of atomic coordinates in EDM, GeoBFN and SLDM.

challenges in minimizing the early-stage truncation error, our method manage to achieve high-quality samples with few sampling steps.

3.2 Comparisons with Previous Diffusion Processes

Previous studies [Nichol and Dhariwal, 2021, Rombach et al., 2022] have underscored the importance of amortizing the generative difficulty through the diffusion process to improve sample quality. These works typically customize signal-to-noise ratio (SNR) schedules to adapt to varying data characteristics. Notably, as pointed out in Song et al. [2023a], the point cloud representation of molecular structures is far more sensitive to noise compared to images. Consequently, for molecular generation, the SNR needs to be decreased at a much slower pace than in the image generation domain.

Our proposed method aligns remarkably well with these insights. As shown in Figure 3, the SNR in our method decreases significantly slower than that of previous methods. In addition, our schedule results in a smoother and more stable generative process for molecules, as demonstrated in Figure 2. The process unfolds uniformly from the origin, preserving the relative spatial relationships of atomic coordinates in intermediate states and retaining critical chemical information throughout the generative trajectory.

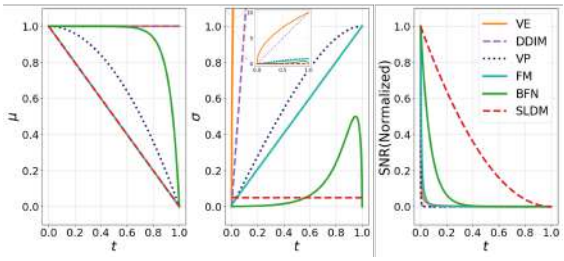


Figure 3: Comparison of schedule parameters across diffusion-based models. Here SNR refers to $(\mu/\sigma)^2$. For SLDM, we set $\sigma = 0.05$, consistent with experiments. The cutoff values σ_{max} and σ_{min} used in other baselines are provided in Appendix A.2.

In contrast, traditional diffusion models like EDM, which follows the process schedule of DDPM, also known as VP, involve a noise variance that increases monotonically during the forward process. This increasing noise quickly disrupts the spatial structure of molecules, rendering much of the diffusion process chemically meaningless. As a result, many computational resources are wasted on steps that attempt to reconstruct signal-less states, leading to a severe reduction in overall model efficiency. Though GeoBFN adopts a low-noise regime for most timesteps, it keeps $\mu \approx 1$ and $\sigma \approx 0$ for over a half of the process, making little changes to the molecular structure. Therefore, GeoBFN shifts the majority of reconstruction difficulty to the earlier stages, creating a severely unbalanced generative process. More illustrations of generation process are provided in Appendix C.7

To sum up, compared to existing diffusion-based models, the new schedule in SLDM aligns better with the inductive biases of molecular data. The approach of distributing the reconstruction difficulty more evenly across the entire diffusion process helps facilitate more effective learning, improving the model's efficiency and efficacy.

3.3 Sampling Strategy of SLDM

Here we introduce the specific sampling strategy of SLDM. First, we need to derive the reverse process of equation 2 or equivalently equation 3. Notably, previous work [Xue et al., 2024b] (Equation 6) provides a unified form of the reverse-time stochastic differential equation (SDE), which shares the same marginal distribution as the forward process, thereby ensuring that the sampling process has the ability to reconstruct the data distribution. We apply the parameter μ and σ of SLDM to this equation and then discretize it using the Euler-Maruyama method, yielding the following discretized

Algorithm 1 SLDM Training

Input: data $\boldsymbol{x}_0$ with dimension N, neural network ϕ, constant σ

repeat

  Sample $t \sim U([0,1])$

  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I_N)$

  Compute $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}$

  Minimize $||\boldsymbol{\epsilon} - \phi(\boldsymbol{x}_t, t)||^2$

until *converged*;

return ϕ

Algorithm 2 SLDM Sampling

Input: trained network ϕ, same N and σ as training, sampling steps T, temperature annealing rate ν

Sample $\boldsymbol{x} \sim \sigma \cdot \mathcal{N}(\mathbf{0}, I_N)$

for $i = T-1$ ***to*** 1 **do**

  $\boldsymbol{x} \leftarrow \frac{T-(i-1)}{T-i} \cdot (\boldsymbol{x} - \sigma \cdot \phi(\boldsymbol{x}, t))$

  if $i > 1$ **then**

    $\boldsymbol{x} \leftarrow \boldsymbol{x} + (\frac{i}{T})^{\nu}\sqrt{2} \cdot \sigma \cdot \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I_N)$

  end

end

return $\boldsymbol{x}$

reverse-time SDE:

$$\boldsymbol{x}_{t-\Delta t} = \underbrace{\boldsymbol{x}_t + \frac{\Delta t}{1-t}\left(\boldsymbol{x}_t + \sigma^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\right)}_{\text{ODE Sampling of equation 3}} \underbrace{+\beta(t)\frac{\Delta t}{1-t}\sigma^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t) + \sqrt{2\beta(t)\frac{\Delta t}{1-t}}\sigma\boldsymbol{\epsilon}}_{\text{Langevin dynamics}}, \tag{7}$$

where $\beta(t)$ is any non-negative bounded function.

Similar to previous works [Xue et al., 2024b, Karras et al., 2022], this sampling can be interpreted as a combination of ODE sampling and Langevin dynamics. The ODE component drives the denoising process along deterministic trajectories, while the Langevin dynamics introduce stochastic corrections. Specifically, we choose $\beta(t) = (1-t)/\Delta t$, to ensure the expectation of the RHS of equation 7 is $\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t]$, as it provides the optimal estimation of $\boldsymbol{x}_{t-\Delta t}$ in terms of minimizing the mean squared error (MSE). Details are demonstrated in Appendix A.5.

Thus the sampling algorithm becomes:

$$\boldsymbol{x}_{t-\Delta t} = \frac{1-(t-\Delta t)}{1-t}(\boldsymbol{x}_t + \sigma^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)) + \sqrt{2}\sigma\boldsymbol{\epsilon}. \tag{8}$$

Note that $1-t$ appears in the denominator of the above equation. To avoid potential division by zero, we choose to skip the $t = 1$ sampling step and find that it works well. This might be because the early sampling steps are less critical to the final result, likely due to Langevin dynamics not strictly requiring a specific prior, which imparts an inherent error-correction capability to the algorithm. This property is further supported by our observation that the sampling process remains relatively robust even when initialized from an invalid distribution, as demonstrated in Appendix C.6.

The above sampling algorithm provides a principled way to approximately sample from the input data distribution, but its practical application in molecular datasets presents unique challenges. In particular, the molecular stability of the dataset is not guaranteed to be 100%. To enhance stability and mitigate the impact of data noise, low-temperature sampling can be employed to generate samples that preserve essential properties of the input data. For Langevin dynamics, the sampling temperature can be controlled by scaling the stochastic term with a constant. However, prior work has demonstrated that such temperature control in diffusion model sampling often fails to effectively balance diversity and fidelity [Dhariwal and Nichol, 2021]. Additionally, as observed in our toy dataset experiments, conventional low-temperature sampling suffers from mode collapse, failing to fully cover all high-density regions of the target distribution.

To address these limitations, we propose a time-annealing temperature schedule:

$$\boldsymbol{x}_{t-\Delta t} = \frac{1-(t-\Delta t)}{1-t}(\boldsymbol{x}_t + \sigma^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)) + t^{\nu}\sqrt{2}\sigma\boldsymbol{\epsilon}, \tag{9}$$

where ν controls the decay rate of temperature over time. This approach allows for higher stochasticity during the initial stages of sampling, enabling the exploration of distant probability density maxima. As the temperature decreases in later stages, the process converges to the probability density maxima. Further theoretical explanations and toy data illustrations are provided in Appendix A.6. With the help of the temperature control, our sampling strategy can reduce the impact of noise in the data, and generate molecules with enhanced stability. The complete training and sampling procedure of

straight-line diffusion are given in algorithm 1 and 2. We further discussed the relation to relevant techniques used in general generative approaches in Related Work (Appendix D.2), including noise scheduling, flow-based methods, other straight-line trajectory models, etc.

Table 1: Unconditional molecular generation results on QM9 and GEOM-Drugs datasets. For all diffusion-based models, T denotes sampling steps. Metrics are calculated with 10000 samples generated from each model. Higher values indicate better performance.

	QM9				GEOM-Drugs	
#Metrics	Atom sta(%)	Mol sta(%)	Valid(%)	V*U(%)	Atom sta(%)	Valid(%)
Data	99.0	95.2	97.7	97.7	86.5	99.9
E-NF	85.0	4.9	40.2	39.4	-	-
G-Schnet	95.7	68.1	85.5	80.3	-	-
EDM (T=1000)	98.7	82.0	91.9	90.7	81.3	92.6
GDM (T=1000)	97.6	71.6	90.4	89.5	77.7	91.8
EDM-Bridge (T=1000)	98.8	84.6	92.0	90.7	82.4	92.8
GeoLDM (T=1000)	98.9	89.4	93.8	92.7	84.4	99.3
EquiFM (T=200)	98.9	88.3	94.7	93.5	84.1	98.9
GeoBFN (T=1000)	99.08	90.87	95.31	92.96	85.60	92.08
END (T=1000)	98.9	89.1	94.8	92.6	87.0	89.2
SLDM (T=1000)	**99.43**	**95.42**	**97.07**	90.42	88.30	**99.95**
SLDM (T=50)	99.30	93.37	96.24	**93.63**	**89.03**	99.57

4 Experiment

To validate the advantages of our method in molecular generation, we evaluate its overall performance and sampling efficiency in both unconditional and conditional generation scenarios.

4.1 Setup

Datasets We evaluate our model using two widely adopted datasets for unconditional molecular generation, with all dataset splitting strictly following baseline settings [Hoogeboom et al., 2022, Song et al., 2024, 2023a]. QM9 [Ruddigkeit et al., 2012, Ramakrishnan et al., 2014] contains approximately 134,000 small organic molecules with up to nine heavy atoms. It is split into training (100K), validation (18K), and test (13K) sets. GEOM-Drugs [Axelrod and Gomez-Bombarelli, 2022] focuses on drug-like molecules, comprising around 430,000 molecules with sizes ranging up to 181 atoms and an average of 44.4 atoms per molecule. Its larger size and greater diversity make it more challenging for generative models. The dataset is randomly divided into training, validation, and test sets using an 8:1:1 ratio.

For conditional molecular generation, we adopt the QM9 dataset with the same setup as prior work [Hoogeboom et al., 2022, Song et al., 2024, 2023a]. The QM9 training partition is split into two halves, each containing 50K samples. Specifically, the QM9 training set is divided into two halves of 50K samples each. The first half is used to train a classifier for ground-truth property labels, while the second half is used to train the conditional generative model.

Implementation The molecule is represented by atomic coordinates and atom types, $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{h})$, where $\boldsymbol{x} \in \mathbb{R}^{3M}$ denotes the atomic coordinates, M is the number of atoms, and $\boldsymbol{h}$ encodes the atom type information. Thus for molecule generation, the model needs to generate both coordinates and atom types. A key requirement for the molecular coordinates is ensuring the SE(3) invariance of the probability distribution, meaning that the probability of generating two molecular conformations should be identical if they only differ by translation or rotation. Following Xu et al. [2022], Hoogeboom et al. [2022], we tailor the SLDM algorithm to satisfy equivariance. Specifically, to ensure translation invariance, we constrain the coordinates to a zero Center of Mass (CoM) space, while rotation invariance is preserved by employing equivariant neural networks to predict the noise. To ensure a fair comparison of generative algorithms, we use EGNN [Satorras et al., 2021] as the backbone model, consistent with the baseline methods [Garcia Satorras et al., 2021, Hoogeboom et al., 2022, Wu et al., 2022, Xu et al., 2023, Song et al., 2024, 2023a]. We prove that the generated data distribution satisfies SE(3) invariance, as provided in Appendix A.7. For atom types, we follow

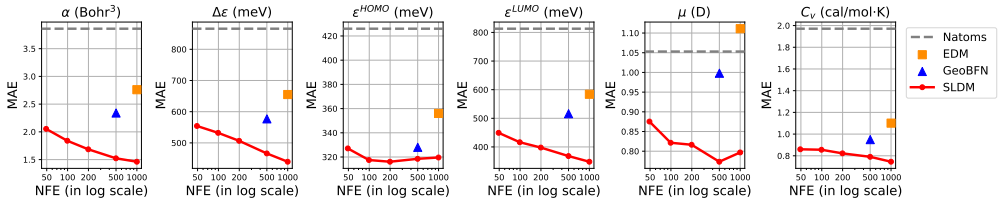


Figure 4: Comparison of performance in conditional generation (↓), with respect to the number of function evaluations (NFE) during the sampling process.

UniGEM [Feng et al., 2024] to predict atom types based on the generated coordinates. The SLDM algorithms tailored for molecular generation are provided in Appendix B. Hyperparameters are summarized in Appendix E. An introduction to the baseline models is included in Section D.1.

Baselines To ensure a fair comparison, we select competitive baselines that also focus on generative modeling and have the same neural architectures (e.g., EGNN) and input information (e.g., coordinates and types). We also acknowledge that some studies propose strategies orthogonal to generative algorithms, as discussed in Related Work (Appendix D.1). However, these strategies are not directly comparable to our work, as they do not aim to replace or improve the generative model itself but rather focus on improving network expressivity and augmenting input information. We believe that substituting their generative algorithms with ours could lead to improvements, and exploring this will be part of the future work.

For conditional molecular generation, Two additional basic baselines: Random and N_{atoms} follow EDM [Hoogeboom et al., 2022]. The Random baseline involves shuffling property labels in the training data and evaluating the property classifier on the shuffled data. The N_{atoms} baseline uses the property classifier network to predict the property based solely on the number of atoms in the molecule.

Metrics It is important to note that evaluation protocols differ across the literature, as discussed in section D.1. To ensure consistency, we strictly adhere to the evaluation methods used in our baselines. We sample 10,000 molecules and predict the bond type (single, double, triple, or non-existent) based on the distances between each pair of atoms, as in Hoogeboom et al. [2022]. Atom stability is computed as the proportion of atoms with correct valency, and molecule stability is the fraction of generated molecules in which all atoms are stable. Validity is evaluated using RDKit by checking whether the 3D molecular structures can be successfully parsed into SMILES format. Uniqueness is determined as the ratio of distinct molecules among all valid samples, indicating the diversity of generated molecules. V*U means the ratio of valid and unique molecules.

4.2 Unconditional Molecular Generation

Unconditional generation assesses the model's ability to learn the underlying molecular data distribution, aiming to generate chemically valid and structurally diverse molecules. The results, summarized in Table 1, show that our method significantly outperforms the baselines across both quality and diversity metrics for the generated molecules. Of particular note is the substantial improvement in the stability of the generated molecules, indicating that our method better satisfies chemical constraints. This validates our hypothesis that our low-noise dynamic is well-suited for generating molecular data. A visualization of generated molecules is provided in Appendix C.7.

Additionally, our approach can achieve superior results with significantly fewer generative steps. A detailed comparison, presented in Figure 1, demonstrates that SLDM achieves 100× faster sampling than the baseline EDM and around 70× speedup compared to GeoBFN and EquiFM. Since the methods use the same network architecture, NFE can reflect practical sampling time. It is important to note that EquiFM utilizes a variety of advanced and efficient ODE solvers, and the results we present correspond to the best performance reported in their paper for different step sizes. Besides, EDM and EDM-Bridge results are from the EDM-Bridge paper while GeoBFN results are reproduced from the official codebase. This observation underscores the advantages of our straight-line diffusion process: while sophisticated solvers play an important role, the diffusion process design may offer more fundamental improvements. Furthermore, our handcrafted sampling strategies can be seamlessly combined with modern sampling techniques, such as optimal time discretization and advanced solvers, which we leave for future exploration.

Table 2: Conditional generation on QM9 dataset, evaluated by MAE(↓) between the property condition and the properties of generated molecules predicted by a pretrained EGNN classifier. SLDM uses $T = 1000$.

Property	α	$\Delta\epsilon$	ϵ^{HOMO}	ϵ^{LUMO}	μ	C_v
Units	Bohr^3	meV	meV	meV	D	$\frac{\text{cal}}{\text{mol K}}$
Data	0.10	64	39	36	0.043	0.040
Random	9.01	1470	645	1457	1.616	6.857
N_{atoms}	3.86	866	426	813	1.053	1.971
EDM	2.76	655	356	584	1.111	1.101
GeoLDM	2.37	587	340	522	1.108	1.025
EquiFM	2.41	591	337	530	1.106	1.033
GeoBFN	2.34	577	328	516	0.998	0.949
SLDM	**1.46**	**440**	**320**	**348**	**0.797**	**0.745**

Table 3: Comparison of generative methods within the UniGEM framework, marked by $_U$, for unconditional generation on the QM9 dataset using small sampling steps T. Higher values indicate better performance.

T	Model	Atom sta(%)	Mol sta(%)	Valid(%)	V*U(%)
50	EDM$_U$	98.55	85.73	93.29	91.78
50	GeoBFN$_U$	98.28	87.16	93.97	91.82
50	SLDM	**99.30**	**93.37**	**96.24**	**93.63**
30	EDM$_U$	97.58	78.75	89.39	87.96
30	GeoBFN$_U$	96.74	81.05	90.93	87.47
30	SLDM	**99.30**	**93.02**	**96.20**	**92.76**

4.3 Conditional Molecule Generation

Conditional generation evaluates the model's capability to produce molecules with desired properties. Following the baseline approaches, we incorporate property values as additional inputs during training and sample them from a prior distribution during inference. Descriptions of baselines and evaluations are provided in Appendix E.

The results in Table 2 demonstrate that our method consistently outperforms baseline models across all metrics. Moreover, as illustrated in Figure 4, our approach achieves a 20-fold acceleration over previous state-of-the-art methods. This result showcases its potential for application in a wide range of controllable generation scenarios.

4.4 Ablation Study

Previous approaches differ in their methods for atom type generation. EDM represents atom types as one-hot vectors, generating them simultaneously with coordinates through diffusion. In contrast, GeoBFN treats atom types as atomic numbers and generates them by the BFN algorithm for discretized data, which incorporates a binning technique to convert continuous probabilities into discrete probabilities. UniGEM, on the other hand, generates only coordinates via diffusion and predicts atom types based on the generated coordinates. We adopt the UniGEM framework due to its superior performance.

To compare generative methods without the influence of atom type generation differences, we integrated EDM and GeoBFN coordinate generation algorithms into the UniGEM framework and compared them with our approach. The results, shown in Table 3 with sampling steps $T = 50$ and $T = 30$, demonstrate that our method exhibits a clear advantage in smaller steps, confirming the efficiency benefits of our generative algorithm. A stability comparison w.r.t various sampling steps is provided in Figure 7. Besides, we also conduct an extensive comparison of sampling efficiency of SLDM under joint diffusion framework, i.e. without UniGEM framework with GeoBFN and EDM in Appendix C.3 and an ablation study about temperature control in Appendix C.5.

4.5 Evaluation on Toy Dataset

We provide a toy dataset experiment in Appendix C.4, demonstrating the superior generative capabilities of our approach in faithfully modeling complex data distributions. These results suggest that SLDM holds promise for generalization to broader application domains.

5 Conclusion

This paper proposes the Straight-Line Diffusion Model (SLDM), a novel generative method that ensures a near-linear diffusion trajectory, effectively reducing truncation error during sampling.

The proposed process schedule naturally aligns with the characteristics of point cloud molecular data and effectively balances the generative difficulty. As a result, our method achieves significant improvements in both sampling efficiency and quality, as demonstrated in both unconditional and conditional generation settings, paving the way for large-scale, controllable molecular generation in practice.

Several challenges remain open for future investigation. On the theoretical side, refining the schedule to fully satisfy the boundary conditions, establishing principled guidelines for selecting the noise level σ, such as analyzing its impact on training stability and sampling accuracy, and deriving optimal discretization schemes could strengthen the theoretical foundations of our approach and further accelerate sampling. On the practical side, applying our method to diverse controllable molecular generation tasks and extending it to broader domains presents exciting opportunities for future advancements.

Acknowledgment

This work is supported by Beijing Academy of Artificial Intelligence and Beijing Frontier Research Center for Biological Structure Fundings.

References

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. Diffusion-based molecule generation with informative prior bridges. *Advances in Neural Information Processing Systems*, 35:36533–36545, 2022.

Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pages 38592–38610. PMLR, 2023.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuxuan Song, Jingjing Gong, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified generative modeling of 3d molecules with bayesian flow networks. In *The Twelfth International Conference on Learning Representations*, 2023a.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian FATRAS, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023. URL `https://openreview.net/forum?id=HgDwiZrpVq`.

Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks. *arXiv preprint arXiv:2308.07037*, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Kaiwen Xue, Yuhao Zhou, Shen Nie, Xu Min, Xiaolu Zhang, JUN ZHOU, and Chongxuan Li. Unifying bayesian flow networks and diffusion models through stochastic differential equations. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=1jHiq640y1.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Shuchen Xue, Mingyang Yi, Weijian Luo, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhi-Ming Ma. Sa-solver: Stochastic adams solver for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

Victor Garcia Satorras, Emiel Hoogeboom, Fabian Fuchs, Ingmar Posner, and Max Welling. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems*, 34:4181–4192, 2021.

Shikun Feng, Yuyan Ni, Yan Lu, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. Unigem: A unified approach to generation and property prediction for molecules. *arXiv preprint arXiv:2410.10516*, 2024.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yuyan Ni, Shikun Feng, Wei-Ying Ma, Zhi-Ming Ma, and Yanyan Lan. Revisiting sampling strategies for molecular generation. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025. URL https://openreview.net/forum?id=ID6GjrJaJx.

Zihan Zhou, Xiaoxue Wang, and Tianshu Yu. Generating physical dynamics under priors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eNjXcP6C0H.

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

Youzhi Luo and Shuiwang Ji. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International conference on learning representations (ICLR)*, 2022.

Ameya Daigavane, Song Eun Kim, Mario Geiger, and Tess Smidt. Symphony: Symmetry-equivariant point-centered spherical harmonics for 3d molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MIEnYtlGyv.

Changyou Chen, Chunyuan Li, Liqun Chen, Wenlin Wang, Yunchen Pu, and Lawrence Carin Duke. Continuous-time flows for efficient inference and density estimation. In *International Conference on Machine Learning*, pages 824–833. PMLR, 2018.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pages 5361–5370. PMLR, 2020.

François Cornet, Grigory Bartosh, Mikkel Schmidt, and Christian Andersson Naesseth. Equivariant neural diffusion for molecule generation. *Advances in Neural Information Processing Systems*, 37: 49429–49460, 2024.

Lei Huang, Hengtong Zhang, Tingyang Xu, and Ka-Chun Wong. Mdm: Molecular diffusion model for 3d molecule generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5105–5112, 2023a.

Zian Li, Cai Zhou, Xiyuan Wang, Xingang Peng, and Muhan Zhang. Geometric representation condition improves equivariant molecule generation. *arXiv preprint arXiv:2410.03655*, 2024.

Chenqing Hua, Sitao Luan, Minkai Xu, Zhitao Ying, Jie Fu, Stefano Ermon, and Doina Precup. Mudiff: Unified diffusion for complete molecule generation. In *Learning on Graphs Conference*, pages 33–1. PMLR, 2024.

Tuan Le, Julian Cremer, Frank Noé, Djork-Arné Clevert, and Kristof Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. 2024a.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Efficient 3d molecular generation with flow matching and scale optimal transport. In *ICML 2024 AI for Science Workshop*, 2024. URL https://openreview.net/forum?id=CxAjGjdkqu.

Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2d & 3d diffusion models for complete molecule generation. *arXiv preprint arXiv:2305.12347*, 2023b.

Giangiacomo Mercatali, Yogesh Verma, Andre Freitas, and Vikas Garg. Diffusion twigs with loop guidance for conditional graph generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 137741–137767. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f90338be8ad676ee02cc32d239fc40e7-Paper-Conference.pdf.

Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. *International conference on machine learning*, 2023.

Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 560–576. Springer, 2023.

Tuan Le, Julian Cremer, Frank Noé, Djork-Arné Clevert, and Kristof Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. *International Conference on Learning Representations*, 2024b.

Shuchen Xue, Zhaoqiang Liu, Fei Chen, Shifeng Zhang, Tianyang Hu, Enze Xie, and Zhenguo Li. Accelerating diffusion sampling with optimized time steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8292–8301, 2024c.

Lijiang Li, Huixia Li, Xiawu Zheng, Jie Wu, Xuefeng Xiao, Rui Wang, Min Zheng, Xin Pan, Fei Chao, and Rongrong Ji. Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7105–7114, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=TIdIXIpzhoI`.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023b.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Nikita Maksimovich Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=kqmucDKVcU`.

Grigory Bartosh, Dmitry P Vetrov, and Christian Andersson Naesseth. Neural flow diffusion models: Learnable forward process for improved diffusion modelling. *Advances in Neural Information Processing Systems*, 37:73952–73985, 2024.

Appendix

Table of Contents

A Supplementary Theoretical Results **22**
- A.1 The Derivation of A Unified Formulation of Diffusion Models 22
- A.2 A Summary of Previous Process Schedules 23
- A.3 Connection Between Sampling of Baseline Methods and First-Order ODE Sampling 24
- A.4 Diffusion Schedule with Straight-line Trajectory 24
- A.5 Supplementary Proof for Sampling 26
- A.6 Temperature Control 27
- A.7 Modeling Invariant Probability Density for 3D Coordinate Generation 27

B Molecular Generation Algorithms **29**

C Supplementary Illustrations and Results **29**
- C.1 Sampling Trajectory of Diffusion-based Models 29
- C.2 Sampling Efficiency Comparisons under UniGEM Framework 30
- C.3 Sampling Efficiency Under Joint Diffusion Framework 30
- C.4 Toy Data Results 31
- C.5 Ablation Study for Temperature Control 32
- C.6 Robustness to Different Prior Distributions 33
- C.7 Visualizations of Generated Molecules and Sampling Process 34

D Related Work **34**
- D.1 An overview of 3D Molecular Generation 34
- D.2 Related Diffusion-Based Studies and Key Differences with SLDM 35

E Implementation Detail **36**

A Supplementary Theoretical Results

A.1 The Derivation of A Unified Formulation of Diffusion Models

The stochastic process of the diffusion-based model can be formulated in general as a linear stochastic differential equation (SDE):

$$\mathrm{d}X_t = f(t)X_t\mathrm{d}t + g(t)\mathrm{d}W_t, \tag{10}$$

where X_t characterizes the noise corrupting process of the data in a diffusion model, $f(t) \leq 0$ and $g(t) \geq 0$ are measurable functions defined on the interval $[0, \infty)$, W_t represents a standard Wiener process. The process X_t is a time rescaled Ornstein-Uhlenbeck process whose law converges exponentially fast to the standard Gaussian distribution [Chen et al., 2023].

Under the assumption that all the relevant integrals exist, the solution of the above SDE is given by:

$$X_t = X_0 \cdot \exp(\int_0^t f(\xi)\mathrm{d}\xi) + \int_0^t \exp(\int_s^t f(\xi)\mathrm{d}\xi)g(s)\mathrm{d}W_s, t \in [0, \infty) \tag{11}$$

A quick proof is as follows:

Proof. By Itô's formula, and applying equation 10:

$$\begin{aligned} \mathrm{d}(X_t \cdot \exp(-\int_0^t f(\xi)\mathrm{d}\xi)) &= (\mathrm{d}X_t - X_t f(t)\mathrm{d}t) \cdot \exp(-\int_0^t f(\xi)\mathrm{d}\xi) \\ &= \exp(-\int_0^t f(\xi)\mathrm{d}\xi) \cdot g(t)\mathrm{d}W_t \end{aligned} \tag{12}$$

Next, we integrate both sides and obtain:

$$X_t \cdot \exp(-\int_0^t f(\xi)\mathrm{d}\xi)) - X_0 = \int_0^t \exp(-\int_0^s f(\xi)\mathrm{d}\xi) \cdot g(s)\mathrm{d}W_s. \tag{13}$$

Finally, we multiply both sides of the equation by $\exp(\int_0^t f(\xi)\mathrm{d}\xi))$ and rearrange the terms to obtain equation 11. □

By the property that the stochastic integral with respect to a Wiener process, X_t is a normal random variable. The mean and variance of X_t are calculated as follows:

$$\mathbb{E}[X_t] = X_0 \cdot exp(\int_0^t f(\xi)\mathrm{d}\xi) \tag{14}$$

$$\mathbb{COV}[X_t] = E[(\int_0^t exp(\int_s^t f(\xi)\mathrm{d}\xi)g(s)\mathrm{d}W_s)^2]\boldsymbol{I} = \int_0^t (exp(\int_s^t f(\xi)d\xi)g(s))^2\mathrm{d}s\boldsymbol{I}. \tag{15}$$

The covariance computation uses the Itô isometry property.

To fit the process X_t to the general form

$$x_t = \mu(t)x_0 + \sigma(t)\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{16}$$

we compare it with our solution in equation 11, and identify that

$$\mu(t) = exp(\int_0^t f(\xi)\mathrm{d}\xi), \quad \sigma(t) = \mu(t)\sqrt{\int_0^t g(s)^2/\mu(s)^2\mathrm{d}s}. \tag{17}$$

Furthermore, we can express $f(t)$ and $g(t)$ in terms of $\mu(t)$ and $\sigma(t)$:

$$\begin{aligned} f(t) &= \frac{\mathrm{d}(\ln \mu(t))}{\mathrm{d}t} = \frac{\dot{\mu}(t)}{\mu(t)}, \\ g(t)^2 &= 2\sigma(t)\mu(t)\frac{\mathrm{d}(\sigma(t)/\mu(t))}{\mathrm{d}t} = 2\sigma(t)\dot{\sigma}(t) - 2\sigma(t)^2\frac{\dot{\mu}(t)}{\mu(t)}. \end{aligned} \tag{18}$$

$\mu(t)$ and $\sigma(t)$ satisfy the initial conditions $\mu(0) = 1$, $\sigma(0) = 0$ and additional conditions: $\mu(t) \geq 0$ is monotone non-increasing, $\sigma(t) \geq 0$, $\frac{\mathrm{d}(\sigma(t)/\mu(t))}{\mathrm{d}t} \geq 0$, i.e. $\sigma(t)/\mu(t)$ is monotone non-decreasing.

Substituting the expressions for $f(t)$ and $g(t)$ back into the original SDE equation 10, we obtain:

$$\mathrm{d}X_t = \frac{\dot{\mu}(t)}{\mu(t)} X_t \mathrm{d}t + \sqrt{2\sigma(t)\dot{\sigma}(t) - 2\sigma(t)^2 \frac{\dot{\mu}(t)}{\mu(t)}} \mathrm{d}W_t, \tag{19}$$

This result is conceptually equivalent to that presented in Appendix B of Karras et al. [2022]. But they adopt a different definition of the schedule compared to equation 16, which leads to a different outcome compared to equation 19.

A.2 A Summary of Previous Process Schedules

This section provides a comprehensive overview of several widely adopted diffusion-based models and their corresponding process schedules.

The Denoising Diffusion Probabilistic Model (DDPM) [Ho et al., 2020] is characterized by a process schedule that exhibits the **Variance Preserving (VP)** property, which can be expressed as: $\mu(t)^2 + \sigma(t)^2 = 1$. This formulation ensures the preservation of variance across timesteps, assuming the data has unit variance. In the domain of 3D molecular unconditional generation, the Equivariant Diffusion Model (EDM) [Hoogeboom et al., 2022] employs a schedule closely resembling the cosine noise schedule introduced by Nichol and Dhariwal [2021], albeit with a simplified notation:

$$\boldsymbol{x}_t = (1 - t^2)\boldsymbol{x}_0 + \sqrt{1 - (1 - t^2)^2}\boldsymbol{\epsilon}, t \in [0, 1], \tag{20}$$

The **Variance Exploding (VE)** schedule, initially proposed in the context of Denoising Score Matching [Song and Ermon, 2019], can also be categorized as a denoising diffusion model with a distinct process schedule [Karras et al., 2022, Song et al., 2021b], which is given by:

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + \sqrt{t}\boldsymbol{\epsilon}, t \in [0, T_{max}]. \tag{21}$$

By rescaling time, we derive the following expression:

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + \sqrt{t}\sigma_{max}\boldsymbol{\epsilon}, t \in [0, 1], \tag{22}$$

where σ_{max} needs to be sufficiently large to ensure that x_1 approximates a uniform distribution. For illustrative clarity, we set $\sigma_{max} = 10$ for Figure 3 and $\sigma_{max} = 20$ for Figure 6.

Denoising Diffusion Implicit Model (DDIM) offers an accelerated sampling process for diffusion models. As proved by Karras et al. [2022], DDIM employs the following schedule:

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + t\boldsymbol{\epsilon}, t \in [0, T_{max}]. \tag{23}$$

By rescaling time, we derive the following expression:

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + t\sigma_{max}\boldsymbol{\epsilon}, t \in [0, 1], \tag{24}$$

For illustrative clarity, we set $\sigma_{max} = 10$ for Figure 3 and Figure 6.

Flow Matching (FM) [Lipman et al., 2023] proposes a linear interpolation between the data distribution and a standard Gaussian, with a neural network learning the corresponding vector field. This noise-adding process can also be viewed as defining a diffusion process schedule, given by:

$$\boldsymbol{x}_t = (1 - t)\boldsymbol{x}_0 + (t + (1 - t)\sigma_{min})\boldsymbol{\epsilon}, t \in [0, 1], \tag{25}$$

where σ_{min} needs to be set sufficiently small to ensure that $\boldsymbol{x}_t$ aligns with the data distribution at $t = 0$. Besides, smaller values of σ_{min} have been reported to yield better results for FM [Tong et al., 2023]. Accordingly, we set $\sigma_{\text{min}} = 0.001$ for both Figure 3 and Figure 6.

Bayesian Flow Networks (BFN) [Graves et al., 2023] is a generative model based on Bayesian inference that accommodates both continuous and discrete variables. For continuous variables, the data is parameterized by Gaussian distributions. In this scenario, the generative algorithms can be interpreted as a denoising diffusion model [Xue et al., 2024a] with a process schedule given by:

$$\boldsymbol{x}_t = (1 - \sigma_{min}^{2t})\boldsymbol{x}_0 + \sqrt{(1 - \sigma_{min}^{2t})\sigma_{min}^{2t}}\boldsymbol{\epsilon}, t \in [0, 1], \tag{26}$$

where σ_{min} needs to be set small to satisfy x_t align with the data distribution when $t = 0$. Following the settings in [Song et al., 2023a], we set $\sigma_{min} = 0.001$ for Figure 3 and Figure 6.

A.3 Connection Between Sampling of Baseline Methods and First-Order ODE Sampling

In section 2, we analyzed the error of first-order ODE discretization methods and claim that the baseline sampling methods primarily rely on first-order ODE discretization. In this section, we provide further clarification on this matter: EquiFM [Song et al., 2024] directly utilizes ODE-based sampling which includes first-order ODE discretization. For methods like EDM [Hoogeboom et al., 2022] and GeoBFN [Song et al., 2023a], while they resemble first-order methods due to requiring only a single function evaluation per iteration, they incorporate random sampling. This distinction necessitates an explanation of how their random sampling processes relate to ODEs. Specifically, EDM uses the same sampling method as DDPM, which is proved as a first-order discretization to the reverse-time SDE of DDPM in Appendix E of Song et al. [2021b]. Similarly, GeoBFN adopts the same sampling method as BFN [Graves et al., 2023], which is proved as a first-order discretization to the reverse-time SDE of BFN in Proposition 4.2 in Xue et al. [2024a]. Please note that both of the reverse-time SDE can be decomposed as an ODE and langevin dynamics:

$$\begin{aligned} \mathrm{d}\boldsymbol{x}_t &= [f(t)\boldsymbol{x}_t - g^2(t)\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t \\ &= \underbrace{[f(t)\boldsymbol{x}_t - \frac{g^2(t)}{2}\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t)]\mathrm{d}t}_{\text{reverse time ODE in equation 3}} - \underbrace{\frac{g^2(t)}{2}\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t}_{\text{Langevin dynamics}} \end{aligned} \tag{27}$$

where $f(t) = \frac{\dot{\mu}(t)}{\mu(t)}$ and $g(t)^2 = 2\sigma(t)\dot{\sigma}(t) - 2\sigma(t)^2\frac{\dot{\mu}(t)}{\mu(t)}$. Thus the sampling processes of EDM and GeoBFN can be effectively approximated as a first-order discretization of an ODE augmented by Langevin dynamics. By isolating and analyzing the discretization error of this ODE component, we gain valuable insights into the limitations of methods with first-order discretization, including baseline approaches in molecular generation such as EDM, GeoBFN, and EquiFM.

A.4 Diffusion Schedule with Straight-line Trajectory

As illustrated in section 3.1, we aim to reduce the truncation error during sampling by minimizing the second-order derivative of the trajectory. To this end, we first consider a simple case where the initial distribution is a delta distribution $\boldsymbol{x}_0 \sim \delta_{\boldsymbol{a}}(\boldsymbol{x})$, $\boldsymbol{a} \in \mathbb{R}^N$. In this scenario, $\boldsymbol{x}_t$ follows a normal distribution $\mathcal{N}(\mu(t)\boldsymbol{a}, \sigma(t)^2\boldsymbol{I}_N)$ according to equation 1. The score function is then tractable as $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t) = -\frac{\boldsymbol{x}-\mu(t)\boldsymbol{a}}{\sigma(t)^2}$. Substituting this into equation 3 gives:

$$\frac{d\boldsymbol{x}}{dt} = \frac{\dot{\sigma}(t)}{\sigma(t)}\boldsymbol{x} + \dot{\mu}(t)\boldsymbol{a} - \frac{\dot{\sigma}(t)}{\sigma(t)}\mu(t)\boldsymbol{a} \tag{28}$$

We aim to keep $\frac{d\boldsymbol{x}}{dt}$ constant, which requires $\frac{\dot{\sigma}(t)}{\sigma(t)} = 0$ and $\dot{\mu}(t)$ to be constant. This implies that $\sigma(t)$ must be constant, and $\mu(t)$ must be a linear function of t.

Given the boundary condition that $x_{t=0}$ approximates data distribution and $x_{t=T}$ approximates a known distribution, the only feasible choice is to set $\sigma(t)$ to a constant $\mu(t) = 1 - t/T$. This satisfy all the schedule constraints noted under equation 2, except $\sigma(0) = 0$, which is approximately satisfied by choosing σ to be a small value. In practice, we use a value for σ that is two orders of magnitude smaller than the data scale, yielding good results. We can also rescale the time as in $[0, 1]$, leading to the following schedule for the diffusion process:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}, t \in [0, 1], \tag{29}$$

which ensures a straight-line trajectory under a special data distribution assumption.

Next, we extend the result into general data distribution in the following theorem.

Theorem A.1. *For a general data distribution and our process schedule $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}$, $t \in [0, 1]$, the following inequality holds for each data dimension i:*

$$P(|\frac{d\boldsymbol{x}_t^{(i)}}{dt} + \frac{\boldsymbol{x}_t^{(i)}}{1-t}| \geq \delta) \leq \frac{\sigma^2}{\delta^2(1-t)^2}. \tag{30}$$

As $\sigma \to 0$, the term $\frac{d\boldsymbol{x}_t}{dt} + \frac{\boldsymbol{x}_t}{1-t}$ converges to zero in probability. More specifically, the solution to the equation $\frac{d\boldsymbol{x}_t}{dt} + \frac{\boldsymbol{x}_t}{1-t} = 0$ is that $\frac{\boldsymbol{x}_t}{1-t}$ becomes a constant, which corresponds to a linear trajectory.

In practice, we set σ to be small and when $t \ll 1-\sigma$, the trajectory is approximately linear. The proof of the theorem needs two following lemmas.

Lemma A.2. *For a general data distribution $f(\boldsymbol{x}_0)$ and a general diffusion process with the schedule $\boldsymbol{x}_t = \mu(t)\boldsymbol{x}_0 + \sigma(t)\boldsymbol{\epsilon}$, its ODE description in equation 3 can be rewritten as*

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \dot{\mu}(t)\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] + \frac{\dot{\sigma}(t)}{\sigma(t)}(\boldsymbol{x}_t - \mu(t)\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]). \tag{31}$$

Proof of Lemma A.2. For the general schedule, we have $p_t(\mathbf{x}_t|\mathbf{x}_0) \sim \mathcal{N}(\mu(t)\mathbf{x}_0, \sigma(t)^2\mathbf{I})$. We start by considering the score function

$$\begin{aligned}
\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t) &= \frac{\int f(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}} p(\boldsymbol{x}_t|\boldsymbol{x}_0) d\boldsymbol{x}_0}{p(\boldsymbol{x}_t)} \\
&= \frac{\int f(\boldsymbol{x}_0) p(\boldsymbol{x}_t|\boldsymbol{x}_0)(-\frac{\boldsymbol{x}_t - \mu(t)\boldsymbol{x}_0}{\sigma(t)^2}) d\boldsymbol{x}_0}{p(\boldsymbol{x}_t)} \\
&= -\frac{\boldsymbol{x}_t}{\sigma(t)^2} + \frac{\mu(t)}{\sigma(t)^2}\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t].
\end{aligned} \tag{32}$$

Substituting this into the ODE form of the diffusion process, we have:

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} &= \frac{\dot{\mu}(t)}{\mu(t)}\boldsymbol{x}_t - \left(\sigma(t)\dot{\sigma}(t) - \sigma(t)^2\frac{\dot{\mu}(t)}{\mu(t)}\right)(-\frac{\boldsymbol{x}_t}{\sigma(t)^2} + \frac{\mu(t)}{\sigma(t)^2}\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]) \\
&= \dot{\mu}(t)\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] + \frac{\dot{\sigma}(t)}{\sigma(t)}(\boldsymbol{x}_t - \mu(t)\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]).
\end{aligned} \tag{33}$$

□

Lemma A.3. *Let $\boldsymbol{a}$ be a random variable that satisfies $p(\boldsymbol{a}|\boldsymbol{x}_0) \sim \mathcal{N}(\boldsymbol{x}_0, \sigma(t)^2/\mu(t)^2\boldsymbol{I})$. Define $\boldsymbol{y} = \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{a}] - \boldsymbol{a}$. Then, the following properties hold:*

1. $\mathbb{E}[\boldsymbol{y}] = \mathbf{0}$,
2. $\mathrm{Var}[y^{(i)}] \leq \sigma(t)^2/\mu(t)^2$, $i = 1, ..., N$, *where $y^{(i)}$ is the ith component of $\boldsymbol{y}$.*

Proof of Lemma A.3. The expectation of $\boldsymbol{y}$ is given by:

$$\mathbb{E}\boldsymbol{y} = \mathbb{E}[\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{a}] - \boldsymbol{a}] = \mathbb{E}_a\mathbb{E}_{\boldsymbol{x}_0|\boldsymbol{a}}[\boldsymbol{x}_0 - \boldsymbol{a}] = \mathbb{E}_{\boldsymbol{x}_0}\mathbb{E}_{\boldsymbol{a}|\boldsymbol{x}_0}[\boldsymbol{x}_0 - \boldsymbol{a}] = 0 \tag{34}$$

For the i-th component $y^{(i)}$, we compute the variance:

$$\begin{aligned}
\mathrm{Var}[y^{(i)}] &= \mathbb{E}[(y^{(i)})^2] = \mathbb{E}[(\mathbb{E}_{x_0|a}[x_0^{(i)} - a^{(i)}])^2] \\
&\leq \mathbb{E}[\mathbb{E}_{x_0|a}[(x_0^{(i)} - a^{(i)})^2]] = \mathbb{E}_{x_0}\mathbb{E}_{a|x_0}[(x_0^{(i)} - a^{(i)})^2] = \sigma(t)^2/\mu(t)^2
\end{aligned} \tag{35}$$

The inequality follows from Jensen's inequality. □

Proof of A.1. For our specific schedule $\sigma(t) = \sigma$ and $\mu(t) = 1 - t$, equation 31 in lemma A.2 reduces to:

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = -\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]. \tag{36}$$

For simplicity of notation, we define $\mathbf{a} = \mathbf{x}_t/\mu(t)$, satisfying $p(\mathbf{a}|\mathbf{x}_0) \sim \mathcal{N}(\mathbf{x}_0, \sigma(t)^2/\mu(t)^2\mathbf{I})$. By applying lemma A.3, the defined $\boldsymbol{y} = \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{a}] - \boldsymbol{a}$ satisfies $\mathbb{E}[\boldsymbol{y}] = \mathbf{0}$ and $\mathrm{Var}[y^{(i)}] \leq \sigma(t)^2/\mu(t)^2 = \frac{\sigma^2}{(1-t)^2}$, for all dimension i.

Applying Chebyshev's inequality, we get:

$$P(|y^{(i)}| \geq \delta) \leq \frac{\mathrm{Var}[y^{(i)}]}{\delta^2} \leq \frac{\sigma^2}{\delta^2(1-t)^2}. \tag{37}$$

Thus, we have the inequality:

$$P(|\frac{d\boldsymbol{x}_t^{(i)}}{dt} - (-\boldsymbol{x}_t^{(i)}/(1-t))| \geq \delta) \leq \frac{\sigma^2}{\delta^2(1-t)^2}. \tag{38}$$

□

A.5 Supplementary Proof for Sampling

Lemma A.4 (Conditional Expectation Minimizes MSE)**.** *The conditional expectation $\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t]$ is the estimator of $\boldsymbol{x}_{t-\Delta t}$ given $\boldsymbol{x}_t$ that minimizes the mean squared error (MSE). That is, for any estimator $h(\boldsymbol{x}_t)$, the following inequality holds:*

$$\mathbb{E}[(\boldsymbol{x}_{t-\Delta t} - h(\boldsymbol{x}_t))^2] \geq \mathbb{E}[(\boldsymbol{x}_{t-\Delta t} - \mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t])^2]. \tag{39}$$

Proof. We begin by expanding the conditional squared error term. By the properties of conditional expectation, the cross-term vanishes:

$$\begin{aligned}
&\mathbb{E}[(\boldsymbol{x}_{t-\Delta t} - h(\boldsymbol{x}_t))^2|\boldsymbol{x}_t] = \mathbb{E}[(\boldsymbol{x}_{t-\Delta t} - \mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] + \mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] - h(\boldsymbol{x}_t))^2|\boldsymbol{x}_t] \\
&= \mathbb{E}[(\boldsymbol{x}_{t-\Delta t} - \mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t])^2|\boldsymbol{x}_t] + \mathbb{E}[(\boldsymbol{x}_{t-\Delta t} - \mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t])(\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] - h(\boldsymbol{x}_t))|\boldsymbol{x}_t] \\
&\quad + \mathbb{E}[(\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] - h(\boldsymbol{x}_t))^2|\boldsymbol{x}_t] \\
&= \mathbb{E}[(\boldsymbol{x}_{t-\Delta t} - \mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t])^2|\boldsymbol{x}_t] + \mathbb{E}[(\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] - h(\boldsymbol{x}_t))^2|\boldsymbol{x}_t] \\
&\geq \mathbb{E}[(\boldsymbol{x}_{t-\Delta t} - \mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t])^2|\boldsymbol{x}_t]
\end{aligned} \tag{40}$$

Taking the expectation with respect to $\boldsymbol{x}_t$, we get equation 39, which completes the proof.

□

Proposition A.5. *For the straight-line diffusion schedule defined in equation 5, we have*

$$\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] = \frac{1-(t-\Delta t)}{1-t}\left(\boldsymbol{x}_t + \sigma^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\right) \tag{41}$$

Proof. From equation 32, the conditional expectation of $\boldsymbol{x}_0$ given $\boldsymbol{x}_t$ is:

$$\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] = \frac{1}{\mu(t)}\left(\boldsymbol{x}_t + \sigma(t)^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\right) \tag{42}$$

For $\boldsymbol{x}_{t-\Delta t}$, the conditional expectation is derived as follows:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] &= \int \boldsymbol{x}_{t-\Delta t} p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t) \mathrm{d}\boldsymbol{x}_{t-\Delta t} \\
&= \int \boldsymbol{x}_{t-\Delta t} \int p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_0) p(\boldsymbol{x}_0|\boldsymbol{x}_t) \mathrm{d}\boldsymbol{x}_0 \mathrm{d}\boldsymbol{x}_{t-\Delta t} \\
&= \int \left(\int \boldsymbol{x}_{t-\Delta t} p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_0) \mathrm{d}\boldsymbol{x}_{t-\Delta t} \right) p(\boldsymbol{x}_0|\boldsymbol{x}_t) \mathrm{d}\boldsymbol{x}_0 \\
&= \int \mu(t-\Delta t) \boldsymbol{x}_0 p(\boldsymbol{x}_0|\boldsymbol{x}_t) \mathrm{d}\boldsymbol{x}_0 \\
&= \mu(t-\Delta t) \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t].
\end{aligned} \tag{43}$$

Here, we make the assumption that x_t and $x_{t-\Delta t}$ are independent given x_0. In fact, the specific temporal dependency among the random variables $(x_t)_{t=0,\cdots,1}$ defines a particular sampling method within the broader family of valid samplers characterized by equation 7. Our subsequent work Ni et al. [2025] generalizes this framework beyond the SLDM schedule and introduces the sampling approach used in this paper as StoMax (maximal stochasticity) sampling, which corresponds to the variant with the highest degree of randomness within the family.

Substituting $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ from equation 42, we have:

$$\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] = \frac{\mu(t-\Delta t)}{\mu(t)}\left(\boldsymbol{x}_t + \sigma(t)^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\right) \tag{44}$$

Adopting the straight-line diffusion schedule defined in equation 5, we produce equation 41. □

As proved in [Xue et al., 2024b], there are a family of reverse processes that share the same marginal probability distributions as equation 2 and equation 3. When applied to our process schedule and apply Euler-Maruyama method discretization, we obtain the following iterative algorithm:

$$\boldsymbol{x}_{t-\Delta t} = \frac{1-t+\Delta t}{1-t}\boldsymbol{x}_t + (1+\beta(t))\frac{\Delta t}{1-t}\sigma^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t) + \sqrt{2\beta(t)\frac{\Delta t}{1-t}\sigma^2}\boldsymbol{\epsilon}. \tag{45}$$

equation 45 uses a Gaussian distribution to model the backward probability $p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t)$, whose expectation is $\frac{1-t+\Delta t}{1-t}\boldsymbol{x}_t + (1+\beta(t))\frac{\Delta t}{1-t}\sigma^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t)$. According to Lemma A.4, the optimal iterative step that that minimizes MSE should satisfy $\frac{1-t+\Delta t}{1-t}\boldsymbol{x}_t + (1+\beta(t))\frac{\Delta t}{1-t}\sigma^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t) = \frac{1-(t-\Delta t)}{1-t}\left(\boldsymbol{x}_t + \sigma^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t)\right)$. This result in $\beta(t) = \frac{1-t}{\Delta t}$.

A.6 Temperature Control

For Langevin dynamics, the sampling temperature can be controlled by scaling the stochastic term with a constant τ.

$$d\boldsymbol{x} = \frac{1}{2}g(t)^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})dt + \tau g(t)\mathrm{d}w'. \tag{46}$$

where τ is the temperature parameter. Then $\pi(\boldsymbol{x}) \propto p(\boldsymbol{x})^{\frac{1}{\tau^2}}$ is the stationary distribution of the process in equation 46, as proved as follows:

Proof. The marginal probability density $\pi_t(x)$ evolves according to Fokker-Planck equation

$$\frac{\partial \pi_t(\boldsymbol{x})}{\partial t} = -\nabla\cdot\left[\frac{1}{2}g(t)^2\nabla_{\boldsymbol{x}}\log p_t(x)\pi_t(\boldsymbol{x})\right] + \frac{1}{2}\nabla\cdot\nabla\cdot\left[\tau^2 g(t)^2\pi_t(\boldsymbol{x})\right] \tag{47}$$

For the stationary distribution, the probability density becomes time-independent, i.e., $\frac{\partial \pi_t(x)}{\partial t} = 0$. Thus, we solve:

$$\nabla\cdot\left[\frac{1}{2}g(t)^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})\pi(\boldsymbol{x})\right] = \frac{1}{2}\nabla\cdot\nabla\cdot\left[\tau^2 g(t)^2\pi(\boldsymbol{x})\right] \tag{48}$$

We can easily validate that $\pi(\boldsymbol{x}) \propto p(\boldsymbol{x})^{\frac{1}{\tau^2}}$ satisfies the stationary equation. □

Thus, higher temperatures ($\tau \to \infty$) increase diversity, with $\pi(\boldsymbol{x})$ approaching a uniform distribution. Conversely, lower temperatures ($\tau \to 0$) enhance fidelity, with $\pi(\boldsymbol{x})$ converging to a δ-distribution at the global maximum of $p(\boldsymbol{x})$.

However, prior work has shown that this low-temperature sampling in diffusion models fails to effectively balance diversity and fidelity in image generation, often resulting in blurred and overly smoothed outputs [Dhariwal and Nichol, 2021]. We also notice that this low-temperature sampling approach applied to our straight-line diffusion fail to fully cover all high-density regions of the target distribution, as shown in Figure 5a.

To address this, we propose an empirically designed time-annealing schedule that introduces higher stochasticity during the initial stages of sampling, allowing for the exploration of distant probability density maxima:

$$\boldsymbol{x}_{t-\Delta t} = \frac{1-t+\Delta t}{1-t}(\boldsymbol{x}_t + \sigma^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})) + t^{\nu}\sqrt{2}\sigma\boldsymbol{\epsilon}, \tag{49}$$

where ν controls the decay rate of temperature over time. Larger ν enhances molecule stability and enables the model to cover all modes, as shown in Figure 5b. The low-temperature sampling helps mitigate the disturbance caused by noise in the training data, and thereby can improve the quality of the samples. The default value of ν for molecular generation is analyzed through an ablation study in Section C.5.

A.7 Modeling Invariant Probability Density for 3D Coordinate Generation

In this section, we aim to prove that the probability density of the generated atomic coordinates, as produced by Algorithms 3 and 4, is invariant to both translations and rotations. Formally, we aim to establish that for the Probability Density modeled by our model $p(\boldsymbol{x}_0) = p(\boldsymbol{x}_0 + \boldsymbol{t})$ and $p(\boldsymbol{x}_0) = p(\boldsymbol{R}\boldsymbol{x}_0)$, where $\boldsymbol{t}$ is a translation vector and $\boldsymbol{R}$ is an orthogonal matrix representing a rotation. The proof is in the same spirit of that in Hoogeboom et al. [2022], Xu et al. [2022]. The result can be seen as a special case of Theorem 1 and Theorem 2 in Zhou et al. [2025].

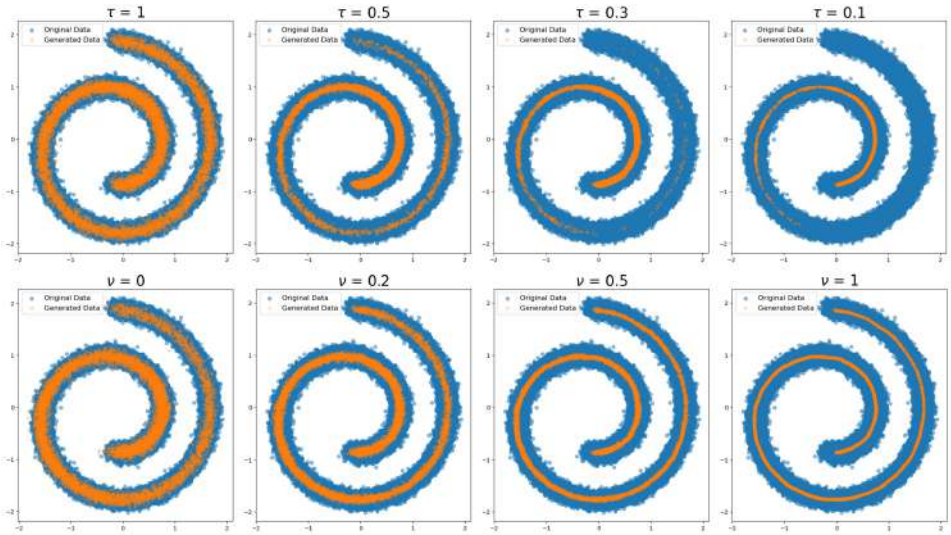


Figure 5: **a.** Generation results from straight-line diffusion with vanilla temperature control as defined in equation 46. Training data (black points) represent 2D Swiss roll coordinates with added Gaussian noise, where the highest density region lies in the interior of the spiral. Diffusion-generated samples (yellow points) exhibit reduced diversity as τ decreases but fail to cover all high-density regions, favoring those near the origin. **b.** Generation results from straight-line diffusion with our annealing temperature control as described in equation 49. Faster temperature decay (larger ν) leads to concentrated samples in high-density regions, successfully covering all local maxima.

To ensure translation invariance, the generative process is defined in the quotient space of translations, specifically the zero Center of Mass (CoM) space. This is achieved through two key operations. First, noise is sampled from a CoM-restricted Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}_{\text{CoM}}(\mathbf{0}, I_{3M})$, which is sampled by first sampling a standard Gaussian noise vector $\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, I_{3M})$ and then subtracting its center of mass: $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}' - \frac{1}{3M}\sum_{i=1}^{3M} \epsilon'_i$. Second, the network's output at each step is projected into the zero CoM space. These operations ensure that all intermediate coordinates $\boldsymbol{x}_t, t = 0, \cdots, 1$ remain strictly within the zero CoM space throughout the generative process.

To establish rotation invariance, we rely on two key properties. First, we use an equivariant neural network satisfying $\phi^{(x)}(\boldsymbol{R}\boldsymbol{x}_t, t) = \boldsymbol{R}\phi^{(x)}(\boldsymbol{x}_t, t)$. Second, we utilize the rotational invariance of the zero mean isotropic Gaussian distributions. This property can also be extended to the CoM-restricted Gaussian distribution, whose probability density function is: $f_{\mathcal{N}_{\text{CoM}}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{3(M-1)/2}} exp(\frac{1}{2}||\boldsymbol{x}||^2)$. Then, we can verify $f_{\mathcal{N}_{\text{CoM}}}(\boldsymbol{x}) = f_{\mathcal{N}_{\text{CoM}}}(\boldsymbol{R}\boldsymbol{x})$ for any orthogonal rotation matrix $\boldsymbol{R}$. Now we prove the rotation invariance of the generation probability as follows:

At each iterative step of the generative process, we have:

$$\boldsymbol{x}_{t-\Delta t} \sim \mathcal{N}_{\text{CoM}}\left(\frac{1-(t-\Delta t)}{1-t} \cdot (\boldsymbol{x}_t - \sigma \cdot \phi(\boldsymbol{x}_t, t)), 2t^{2\nu}\sigma^2 I_{3M}\right) \tag{50}$$

$$\begin{aligned} p(\boldsymbol{R}\boldsymbol{x}_{t-\Delta t}|\boldsymbol{R}\boldsymbol{x}_t) &= f_{\mathcal{N}_{\text{CoM}}}\left(\frac{\boldsymbol{R}\boldsymbol{x}_{t-\Delta t} - \frac{1-(t-\Delta t)}{1-t} \cdot (\boldsymbol{R}\boldsymbol{x}_t - \sigma \cdot \phi(\boldsymbol{R}\boldsymbol{x}_t, t))}{\sqrt{2}t^{\nu}\sigma}\right) \\ &= f_{\mathcal{N}_{\text{CoM}}}\left(\boldsymbol{R}\frac{\boldsymbol{x}_{t-\Delta t} - \left(\frac{1-(t-\Delta t)}{1-t} \cdot (\boldsymbol{x}_t - \sigma \cdot \phi(\boldsymbol{x}_t, t))\right)}{\sqrt{2}t^{\nu}\sigma}\right) \\ &= f_{\mathcal{N}_{\text{CoM}}}\left(\frac{\boldsymbol{x}_{t-\Delta t} - \left(\frac{1-(t-\Delta t)}{1-t} \cdot (\boldsymbol{x}_t - \sigma \cdot \phi(\boldsymbol{x}_t, t))\right)}{\sqrt{2}t^{\nu}\sigma}\right) = p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t) \end{aligned} \tag{51}$$

In the second equality, we apply the equivariance property of the neural network. The third equality follows from the rotational invariance of the isotropic Gaussian distribution.

Additionally, the initial distribution $p(\boldsymbol{x}_1) = \mathcal{N}_{\text{CoM}}(\mathbf{0}, \sigma^2 I_{3M})$ is rotation invariant. Combining these facts, we can propagate rotation invariance across the generative process. Thus, for the final generated distribution:

$$\begin{aligned} p(\boldsymbol{R}\boldsymbol{x}_0) &= \int \cdots \int \prod_{t=\Delta t}^{1} p(\boldsymbol{R}\boldsymbol{x}_{t-\Delta t} | \boldsymbol{R}\boldsymbol{x}_t) p(\boldsymbol{R}\boldsymbol{x}_1) \mathrm{d}\boldsymbol{x}_1 \cdots \mathrm{d}\boldsymbol{x}_{\Delta t} \\ &= \int \cdots \int \prod_{t=\Delta t}^{1} p(\boldsymbol{x}_{t-\Delta t} | \boldsymbol{x}_t) p(\boldsymbol{x}_1) \mathrm{d}\boldsymbol{x}_1 \cdots \mathrm{d}\boldsymbol{x}_{\Delta t} = p(\boldsymbol{x}_0). \end{aligned} \tag{52}$$

In the second equality, we use the rotational invariance of both the transition probabilities and the initial distribution. This proves that the final probability density of the generated data is rotation invariant.

B Molecular Generation Algorithms

Algorithm 3 Straight-Line Diffusion Training for Molecules (UniGEM)

Input: 3D molecular data $\boldsymbol{z}_0 = [\boldsymbol{x}_0, \boldsymbol{h}_0]$ with M atoms, $\boldsymbol{x}_0 \in$ CoM space, neural network ϕ, variance constant σ, nucleation time $t_n \in [0, 1]$

repeat
- Sample $t \sim \frac{1}{2}U([0, t_n]) + \frac{1}{2}U([t_n, 1])$
- Sample $\boldsymbol{\epsilon} \sim \mathcal{N}_{\text{CoM}}(\mathbf{0}, I_{3M})$
- Compute $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}$
- Minimize $||\boldsymbol{\epsilon} - \phi^{(x)}(\boldsymbol{x}_t, t)||^2 + 1_{t \le t_n} |\boldsymbol{h}_0 - \phi^{(h)}(\boldsymbol{x}_t, t)|$

until *converged*;

Algorithm 4 Straight-Line Diffusion Sampling for Molecules (UniGEM)

Input: Number of atoms M, neural network ϕ, variance constant σ, nucleation time t_n, sampling steps T, temperature annealing rate ν

Sample $\boldsymbol{x}_1 \sim \sigma \cdot \mathcal{N}_{\text{CoM}}(\mathbf{0}, I_{3M})$

for $i = T-1$ ***to*** 1 **do**
- $t = i/T$, $\Delta t = 1/T$
- $\boldsymbol{x}_{t-\Delta t} = \frac{1-(t-\Delta t)}{1-t} \cdot (\boldsymbol{x}_t - \sigma \cdot \phi(\boldsymbol{x}_t, t))$
- Projecting $\boldsymbol{x}_{t-\Delta t}$ into zero CoM space
- **if** $i > 1$ **then**
  - Sample $\boldsymbol{\epsilon} \sim \mathcal{N}_{\text{CoM}}(\mathbf{0}, I_{3M})$
  - $\boldsymbol{x}_{t-\Delta t} = \boldsymbol{x}_{t-\Delta t} + t^{\nu}\sqrt{2} \cdot \sigma \cdot \boldsymbol{\epsilon}$
- **end**
- **else**
  - $\boldsymbol{h}_0 = \phi^{(h)}(\boldsymbol{x}_0, 0)$
- **end**

end

return $\boldsymbol{z}_0 = [\boldsymbol{x}_0, \boldsymbol{h}_0]$

C Supplementary Illustrations and Results

C.1 Sampling Trajectory of Diffusion-based Models

To evaluate whether SLDM exhibits a near-linear trajectory under general data distributions, as suggested by Theorem 3.1, we visualize the ODE trajectory and compare it with other diffusion-based models. We consider a scenario where the data distribution is a one-dimensional Gaussian mixture, as this setup offers a tractable score function and serves as a representative example of general distributions. The resulting trajectories are shown in Figure 6. Our method maintains a trajectory that is consistently closer to a straight line compared to other approaches, leading to smaller truncation

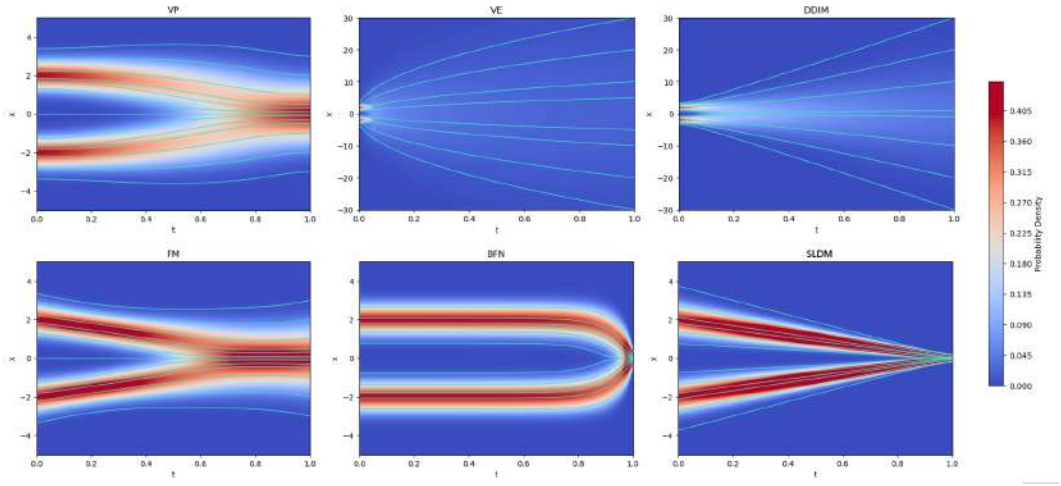


Figure 6: The trajectory of the ODE in equation 3 of various diffusion-based models. The data distribution is a mixture of Gaussians in one dimension, defined as $x_0 \sim 0.5 \cdot \mathcal{N}(2, 1/4) + 0.5 \cdot \mathcal{N}(-2, 1/4)$. The background color indicates the value of probability density $p(x_t)$. During sampling, the timestep t decreases from 1 to 0.

errors under first-order numerical discretization. As suggested in Theorem 3.1, the trajectory slightly deviates from a straight line in the early stages (i.e., as t approaches 1), which could increase the sampling error in theory. However, our empirical observations indicate that the sampling process is relatively robust to such errors during these initial steps. This robustness may be attributed to the incorporation of Langevin dynamics in the sampling algorithm, which can mitigate the impact of early-stage inaccuracies. As a result, SLDM achieves an overall lower sampling error.

In comparison, DDIM demonstrates a relatively linear trajectory during the early stages but exhibits significant curvature in later stages, where accurate sampling is more crucial. This could potentially degrade its performance. For the FM method, where the initial distribution is Gaussian and no explicit prior-data joint distribution is predefined as in Tong et al. [2023], it fails to achieve a straight-line trajectory. Other methods also demonstrate an evident curved trajectory. Besides, we can also see from the illustration that the distribution remains nearly static during the later stages of BFN sampling, which aligns with the discussion in section 3.2 and suggests potential inefficiencies of time scheduling.

C.2 Sampling Efficiency Comparisons under UniGEM Framework

We complement the ablation study in Section 4.4 by incorporating results with diverse sampling steps. The results demonstrate that while UniGEM enhances the performance of EDM and BFN when sampling step is abundant, these methods still struggle to achieve satisfactory molecular stability in the low sampling step scenario.

C.3 Sampling Efficiency Under Joint Diffusion Framework

A natural question arises: can SLDM achieve strong performance without the UniGEM framework? Specifically, can it jointly generate atom types and coordinates using diffusion models, following the approach of EDM and GeoBFN? To address this, we introduce a variant called *SLDM (joint diffusion)* as a preliminary exploration, where atom types are generated using an SLDM model tailored for discretized data, replacing UniGEM's prediction strategy. The training and sampling procedures of SLDM (joint diffusion) for molecule generation are detailed in Algorithms 5 and 6, respectively. Unlike UniGEM, we follow GeoBFN to represent atom types by their atomic numbers (positive integers) and adopt the binning technique from BFN to handle discretized data. Here, $P_{\hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\sigma}}_h}(b)$ denotes the discretized cumulative distribution function (CDF), defined as: $F(b) = \frac{1}{2}[1 + erf(\frac{b-\mu}{\sqrt{2}\sigma})]$, with $P(b) = F(b)$ for $-1 < b < 1$, $P(b) = 0$ for $b \leq -1$, and $P(b) = 1$ for $b \geq 1$. "NearestCenter" decodes the generated continuous continuous normalized atomic number into an atomic type by

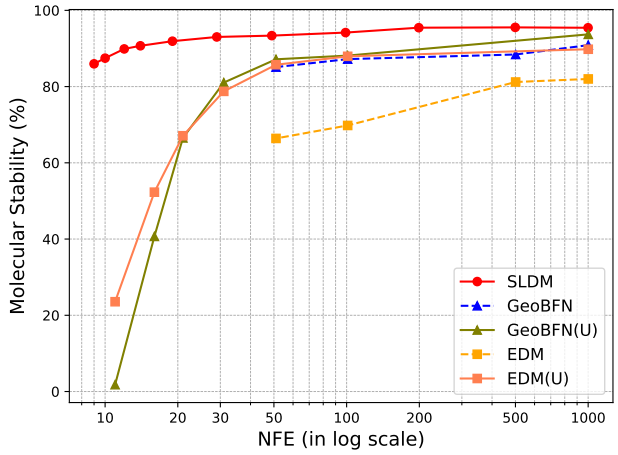


Figure 7: Comparison of molecule stability (↑) across diffusion-based molecular generation models with UniGEM framework on QM9 unconditional generation, evaluated with respect to the Number of Function Evaluations (NFE) during the sampling process.

finding the nearest center of bins (normalized atomic numbers), then unnormalizing it and mapping it to the corresponding atomic type. Notably, there remains substantial room for further exploration regarding the handling of diffusion for discretized data, which we leave for future work.

Algorithm 5 SLDM Training (joint diffusion)

Input: Number of bins $K \in \mathbb{N}$ (equal to the largest atomic number), fixed variance σ, 3D molecular data $\boldsymbol{z}_0 = [\boldsymbol{x}_0, \boldsymbol{h}_0]$ with M atoms, $\boldsymbol{x}_0 \in$ CoM space, normalized atom number $\tilde{\boldsymbol{h}}_0 = (2 * \boldsymbol{h}_0 - 1)/K - 1 \in \left[\frac{1}{K} - 1, 1 - \frac{1}{K}\right]^D$

repeat

Sample $t \sim \mathcal{U}(0,1)$, $\boldsymbol{\epsilon}^x \sim \mathcal{N}_{\text{CoM}}(\mathbf{0}, I_{3M})$, $\boldsymbol{\epsilon}^h \sim \mathcal{N}\mathbf{0}, I_{3M})$

Compute $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}^x$

Compute $\boldsymbol{h}_t = (1-t)\tilde{\boldsymbol{h}}_0 + \sigma\boldsymbol{\epsilon}^h$

Network output: $\hat{\boldsymbol{\epsilon}}^x, \hat{\boldsymbol{\mu}}_h^\epsilon, \ln \hat{\boldsymbol{\sigma}}_h^\epsilon = \Phi(\boldsymbol{x}_t, \boldsymbol{h}_t, t)$

$\hat{\mu}_h = \frac{\boldsymbol{h}_t - \sigma\hat{\boldsymbol{\mu}}_h^\epsilon}{1-t}$, $\hat{\boldsymbol{\sigma}}_h = \frac{\sigma \exp(\ln \hat{\boldsymbol{\sigma}}_h^\epsilon)}{1-t}$

Minimize denoise loss: $\|\boldsymbol{\epsilon}^x - \hat{\boldsymbol{\epsilon}}^x\|^2 + \left\|(\hat{\boldsymbol{\mu}}_h^\epsilon - \boldsymbol{\epsilon}^h) - \frac{1-t}{\sigma}\sum_{i=1}^K P_{\hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\sigma}}_h}(b_i)(b_i - \hat{\boldsymbol{\mu}}_h)\right\|^2$

until *converged*;

Results are presented in Table 4. The proposed temperature annealing mechanism endows the model with greater flexibility, enabling a balanced trade-off between stability/validity and uniqueness/diversity. We manually tuned the temperature annealing rates for coordinates (ν_x) and atomic numbers (ν_h), and empirical observations indicate that setting $\nu_x > \nu_h$ yields better performance in achieving such balance. It should be noted that we did not conduct exhaustive hyperparameter searches, leaving room for potential further improvements.SLDM (joint diffusion) outperforms EDM across all metrics and exhibits superior stability and validity with slightly lower uniqueness compared to GeoBFN when the number of sampling steps T is limited. Notably, the overall performance of this variant is still inferior to the original SLDM which employs the UniGEM framework. This finding reaffirms that the UniGEM framework may be a more optimal approach for generating small-molecule.

C.4 Toy Data Results

We evaluate our model on several 2D toy data to test its generality. The datasets included swissroll and moons to represent continuous data distributions, as well as chessboard to simulate discretized data distributions. The dataset consists of 100,000 samples. For the moons and swissroll datasets,

Algorithm 6 SLDM Sampling (joint diffusion)

Input: Number of atoms M, neural network ϕ, variance constant σ, nucleation time t_n, sampling steps T, temperature annealing rate ν_x and ν_h

Sample $\boldsymbol{x}_1 \sim \sigma\mathcal{N}_{\text{CoM}}(\mathbf{0}, I_{3M})$, $\boldsymbol{h}_1 \sim \sigma\mathcal{N}(\mathbf{0}, I_{3M})$

for $i = T-1$ ***to*** 1 **do**

  $t = i/T$, $\Delta t = 1/T$

  Network output: $\hat{\boldsymbol{\epsilon}}^x, \hat{\boldsymbol{\mu}}_h^\epsilon, \ln \hat{\boldsymbol{\sigma}}_h^\epsilon = \Phi(\boldsymbol{x}_t, \boldsymbol{h}_t, t)$

  $\hat{\mu}_h = \frac{\boldsymbol{h}_t - \sigma\hat{\boldsymbol{\mu}}_h^\epsilon}{1-t}$, $\hat{\boldsymbol{\sigma}}_h = \frac{\sigma \exp(\ln \hat{\boldsymbol{\sigma}}_h^\epsilon)}{1-t}$

  Update $\boldsymbol{x}_{t-\Delta t} = \frac{1-(t-\Delta t)}{1-t} \cdot (\boldsymbol{x}_t - \sigma \cdot \hat{\boldsymbol{\epsilon}}^x)$

  Update $\hat{\boldsymbol{h}}_0 = \sum_{i=1}^{K} P_{\hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\sigma}}_h}(b_i) b_i$, $\boldsymbol{h}_{t-\Delta t} = (1-(t-\Delta t))\hat{\boldsymbol{h}}_0$

  Projecting $\boldsymbol{x}_{t-\Delta t}$ into zero CoM space

  if $i > 1$ **then**

    Sample $\boldsymbol{\epsilon}^x \sim \mathcal{N}_{\text{CoM}}(\mathbf{0}, I_{3M})$, $\boldsymbol{\epsilon}^h \sim \mathcal{N}\mathbf{0}, I_{3M})$

    $\boldsymbol{x}_{t-\Delta t} = \boldsymbol{x}_{t-\Delta t} + t^{\nu_x}\sqrt{2} \cdot \sigma \cdot \boldsymbol{\epsilon}^x$

    $\boldsymbol{h}_{t-\Delta t} = \boldsymbol{h}_{t-\Delta t} + t^{\nu_h}\sqrt{2} \cdot \sigma \cdot \boldsymbol{\epsilon}^h$

end

return $\boldsymbol{z}_0 = [\boldsymbol{x}_0, \textit{NearestCenter}(\boldsymbol{h}_0)]$

Table 4: Sampling efficiency comparison under joint diffusion framework on QM9 dataset unconditional generation task.

Method	Atom Sta (%)	Mol Sta (%)	Valid (%)	V×U (%)
EDM (T=1000)	98.70	82.00	91.90	90.70
GeoBFN (T=1000)	**99.08**	90.87	**95.31**	**92.96**
SLDM (joint diffusion) (T=1000)	99.01	**91.57**	95.20	91.57
GeoBFN (T=500)	98.78	88.42	93.35	91.78
SLDM (joint diffusion) (T=500)	**99.89**	**90.57**	**94.00**	91.11
GeoBFN (T=100)	98.64	87.21	93.03	**91.53**
SLDM (joint diffusion) (T=100)	**98.92**	**90.31**	**93.84**	89.36
GeoBFN (T=50)	98.28	85.11	92.27	**90.72**
SLDM (joint diffusion) (T=50)	**98.70**	**88.09**	**92.84**	89.53

we performed training with 40 diffusion steps and 100 epochs, and for the chessboard dataset, the training was extended to 600 epochs with 100 diffusion steps to ensure convergence for all models. Other settings are kept the same for all datasets: The model is a 5-layer MLP. The batch size was set to 2048, and the optimizer used was Adam with a learning rate of 0.001. No temperature control is used during sampling. These settings follow https://github.com/albarji/toy-diffusion/, and are kept consistent across all generative algorithms to ensure a fair comparison. The results are provided in Figure 8. The data generated by our model show the closest alignment to the original distributions. Specifically, our model produced samples with fewer outliers and achieved good coverage of the data distributions. These findings highlight the superior generative capabilities of our approach in faithfully modeling complex data distribution.

C.5 Ablation Study for Temperature Control

Theoretically, a larger temperature annealing rate v corresponds to a faster cooling scheme. The results in Table 5 demonstrate that adjusting the temperature can effectively enhance molecular stability. However, extreme values of v may significantly reduce diversity. We selected $v = 0.5$ as the default setting for our model, as it achieves the highest $U \times V$ score, indicating the optimal balance between diversity and fidelity.

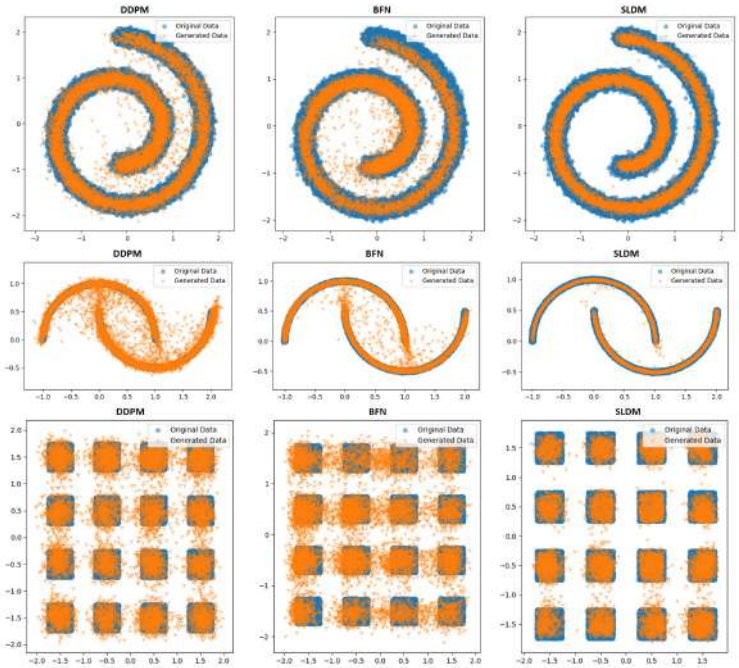


Figure 8: Generative performance comparison of Straight-Line Diffusion, DDPM, and BFN on 2D toy datasets.

Table 5: Impact of temperature annealing rate evaluated on the QM9 unconditional generation with sampling step $T = 50$. Higher values indicate better performance.

	Atom sta(%)	Mol sta(%)	Valid(%)	$U \times V$(%)
SLDM(v=0)	99.14	91.74	95.40	92.76
SLDM(v=0.5)	99.28	93.03	96.20	**93.28**
SLDM(v=1)	99.27	93.46	96.08	92.42
SLDM(v=3)	99.37	94.34	96.84	90.08
SLDM(v=5)	99.42	94.83	97.33	86.10
SLDM(v=10)	99.53	96.06	97.98	75.13

C.6 Robustness to Different Prior Distributions

Empirically, we find that SLDM maintains high validity and stability even when the initial distribution during sampling deviates from the theoretical valid one, i.e., Gaussian distribution with standard deviation σ, as shown in Table 6. This robustness helps mitigate the impact of deviations from the ideal linear trajectory in the early sampling steps, as discussed under the Theorem 3.1.

Table 6: Robustness of SLDM to different initial sampling distributions on the QM9 unconditional generation task. Here $\epsilon \sim \mathcal{N}(0, I)$ and the sampling step $T = 50$.

Prior Distribution	**Mol_stab (%)**	**Atom_stab (%)**	**Validity (%)**	**V*U (%)**
$\sigma \cdot \epsilon$ (theoretical)	93.43	99.32	96.10	93.31
$2 \cdot \sigma \cdot \epsilon$	91.81	99.12	95.00	92.79
$0.1 \cdot \sigma \cdot \epsilon$	93.31	99.29	96.07	93.46
$0.01 \cdot \sigma \cdot \epsilon$	92.94	99.24	96.06	93.06

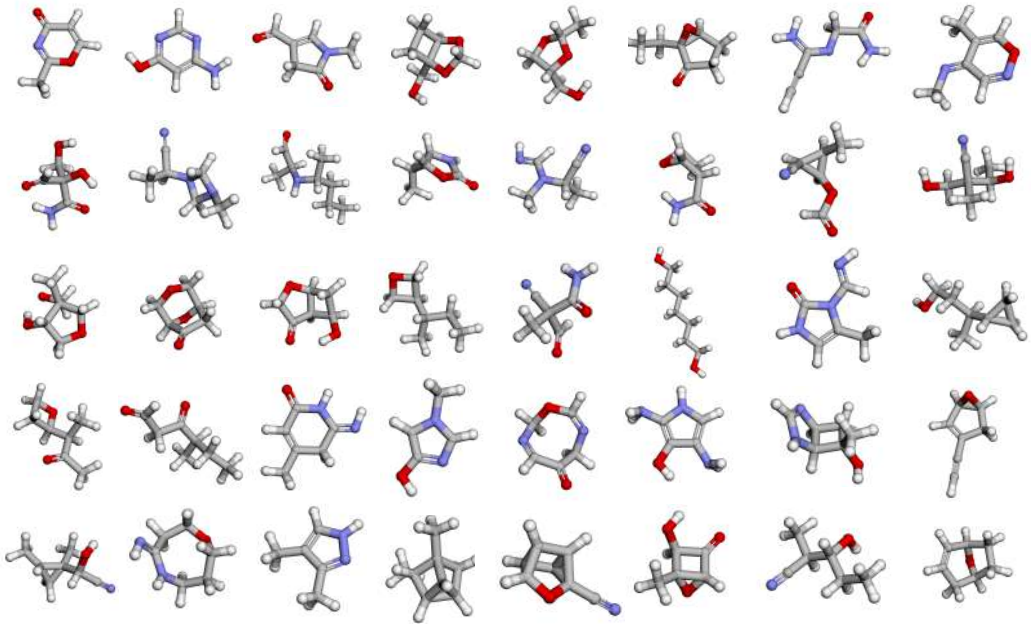


Figure 9: 3D molecules generated by the SLDM model trained on the QM9 dataset.

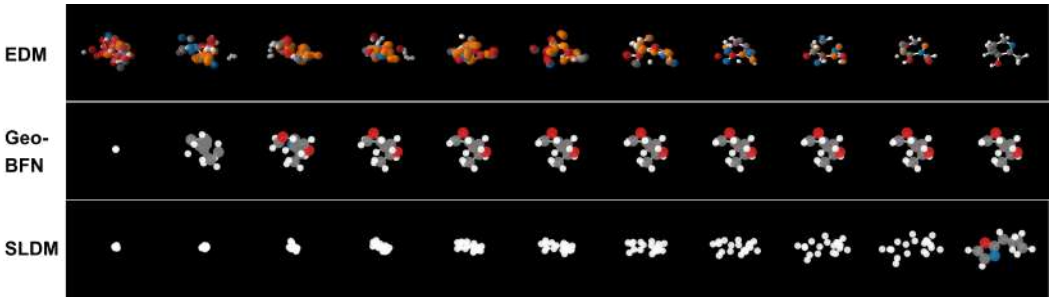


Figure 10: The generative process of EDM, GeoBFN and SLDM models.

C.7 Visualizations of Generated Molecules and Sampling Process

We have randomly generated 40 molecules using our trained models on QM9 in the unconditional setting, and visualized their 3D structures using py3Dmol, as shown in Figure 9. These examples clearly demonstrate the structural plausibility and diversity of our generated molecules.

In Figure 2, we compare the diffusion processes of different models by adding noise to the same molecule's coordinates, enabling a straightforward visual comparison of their noise schedules. As a complement, we present a comparison of the practical sampling processes of the three models in Figure 10. Although generating the exact same molecule across models is not feasible in this case, we still arrive at conclusions analogous to those in Section 3.2 of the main text. Additionally, unlike the noise-adding process, the generation process of SLDM demonstrates more pronounced randomness.

D Related Work

D.1 An overview of 3D Molecular Generation

Recent advances in 3D molecular generation can be categorized based on the underlying **generative algorithms**: Autoregressive methods generate molecules step by step, progressively connecting atoms or molecular fragments. G-SchNet [Gebauer et al., 2019] and G-SphereNet [Luo and Ji, 2022] are early examples that use this strategy to model 3D molecular structures. Building on

these, Symphony [Daigavane et al., 2024] incorporates higher-degree E(3)-equivariant features and message-passing to improve the modeling of molecular geometries. Normalizing flow [Chen et al., 2018] has also been applied to molecule generation by E-NF [Garcia Satorras et al., 2021] and Köhler et al. [2020]. Diffusion-based models have gained prominence for 3D molecular generation recently. Hoogeboom et al. [Hoogeboom et al., 2022] introduced the Equivariant Diffusion Model (EDM) to jointly generate atomic coordinates and atom types. Extensions of EDM include EDM-Bridge [Wu et al., 2022], which enhances performance through prior bridges, and GeoLDM [Xu et al., 2023], which performs diffusion in a latent space using an autoencoder. EquiFM [Song et al., 2024] employs flow matching for efficient molecule generation, and GeoBFN [Song et al., 2023a] combines Bayesian Flow Networks with distinct generative algorithms tailored for discretized charges and continuous coordinates. END [Cornet et al., 2024] proposes a learnable data- and time-dependent noise schedule of the diffusion process, and achieves improved sampling efficiency.

Several studies offer **complementary strategies** to generative algorithms. First, representation-conditioned generation has been explored by MDM [Huang et al., 2023a], which conditions on VAE representations, and GeoRCG [Li et al., 2024], which extends EDM by leveraging pretrained molecular representations. Second, some studies propose advanced network architectures to enhance molecular generation [Hua et al., 2024, Huang et al., 2023a, Le et al., 2024a, Irwin et al., 2024][Huang et al., 2023b, Mercatali et al., 2024]. Third, additional input information has been introduced to the molecular generation. For instance, MolDiff [Peng et al., 2023] explicitly predicts bonds during generation. MiDi [Vignac et al., 2023], JODO [Huang et al., 2023b], EQGAT-diff [Le et al., 2024a], and SemlaFlow [Irwin et al., 2024] further extend generation to include bonds and formal charges, enriching the molecular inputs. Twigs [Mercatali et al., 2024], on the other hand, integrates additional molecular properties into the diffusion training process, leading to enhanced conditional generation.

It is worth noting that **evaluation strategies** for 3D molecular generation can be different across methods. EDM [Hoogeboom et al., 2022] employs strict rules for bond definitions based on interatomic distances, implicitly enforcing constraints on bond lengths and steric hindrance. In this framework, the metric "stability" is rigorously defined, requiring correct valency and neutral atomic charges. Our method, along with most diffusion-based approaches [Wu et al., 2022, Xu et al., 2023, Song et al., 2023a], adheres to this evaluation standard. In contrast, another line of studies [Vignac et al., 2023, Le et al., 2024a, Irwin et al., 2024] infers bonds and formal charges directly from the model, allowing atoms to have non-zero formal charges. Therefore, the bond inference imposes no constraints on bond lengths or steric hindrance. Further, the "stability" metric is defined more loosely, permitting discrepancies between valency and the number of covalent bonds. This relaxed evaluation framework makes it challenging to directly compare these two approaches. We advocate future efforts to establish more equitable evaluation methods to fairly assess the strengths of both paradigms.

D.2 Related Diffusion-Based Studies and Key Differences with SLDM

Noise Scheduling It is important to note that our proposed process schedule differs fundamentally from previous works on noise scheduling [Karras et al., 2022, Nichol and Dhariwal, 2021, Le et al., 2024b], time discretization strategies [Xue et al., 2024c, Li et al., 2023], and adaptive step size methods [Lu et al., 2022]. These approaches can be interpreted as applying a time transformation $\mathcal{T}(\cdot)$ that jointly rescales the diffusion process schedule as $\mu(\mathcal{T}(t))$ and $\sigma(\mathcal{T}(t))$. Notably, such transformations preserve the monotonicity and endpoints of the schedule functions. In contrast, our method decouples μ and σ, and fundamentally alters the monotonicity and endpoint of $\sigma(t)$.

Linear $\boldsymbol{\mu}(\mathbf{t})$. Although our method is derived from the diffusion perspective, its process schedule shares some similarities with flow matching (FM) algorithms, such as FM-OT [Lipman et al., 2023] and conditional flow matching (CFM) [Tong et al., 2023]. Both methods employ a linear $\mu(t)$, and CFM introduces additional low-scale noise. From this perspective, our process schedule can also be interpreted as a variant of CFM. Specifically, our approach employs a prior distribution modeled as a small-scale Gaussian distribution centered at the origin. However, FM and diffusion differ fundamentally in their perspectives: FM models the generative process as an ODE, while diffusion models a stochastic process. This core distinction leads to differences in both the learning targets and sampling methods. Therefore, unlike FM that learns the velocity field and relies on ODE-based sampling, our approaches focus on learning the noise and employ stochastic sampling.

Straight Sampling Trajectory. Moreover, similar to our approach, flow-based methods also aim at flows with straight trajectories. However, these methods rely on predefined prior-data joint distribution

to produce straighter paths. This distribution is procured by solving an optimal transport (OT) problem during the training of flow matching [Tong et al., 2023, Song et al., 2024], and solving such problem is often challenging. When the OT solution is unavailable, achieving straight-line trajectories typically requires additional distillation steps or solving optimization problems, as studied in rectified flow [Liu et al., 2022], progressive distillation [Salimans and Ho, 2022], consistency model [Song et al., 2023b, Luo et al., 2023] and optimal flow matching [Kornilov et al., 2024]. NFDM [Bartosh et al., 2024] proposes a learnable forward process to adapt to align with the reverse process and introduces penalties on the curvature of the reverse process's trajectories. In contrast to these existing methods, our approach offers a more straightforward solution for achieving a straight trajectory, by designing a novel diffusion process that minimizes the second-order derivative of the trajectory.

E Implementation Detail

The hyperparameter settings for molecular generation are detailed in Table 7. Settings follow UniGEM [Feng et al., 2024], with two additional tunable hyperparameters introduced by our generative algorithm: the noise variance σ and the temperature annealing rate ν.

Table 7: Network and training hyperparameters.

Network Hyperparameters	Value
Embedding size	256 for unconditional generation, 192 for conditional generation
Layer number	9 for QM9, 4 for Geom-Drugs
Shared layers	1
Training Hyperparameters	**Value**
Batch size	64 for QM9, 128 for Geom-Drugs
Train epoch	3000 for QM9, 32 for Geom-Drugs
Learning rate	1.00×10^{-4}
Optimizer	Adam
Sample steps T	$10 \sim 1000$
Nucleation time	10
Oversampling ratio	0.5 for each branch
Loss weight	1 for each loss term
Generative Algorithm Hyperparameters	**Value**
Noise Variance σ	0.05 for unconditional generation, 0.1 for conditional generation
Temperature Annealing Rate ν	0.5 for unconditional generation, 3 for conditional generation
Non-uniform Discretization	False if $T > 13$

For QM9, it takes approximately 10 days on a single A100 GPU. For GEOM-drugs, it takes approximately 16 days on four A100 GPUs. While the training epochs are longer compared to the baselines, this extended duration is not mandatory. Although longer training leads to slight performance gains, we found that the mol stable and atom stable metrics remain stable around 2000 epochs on QM9 and 13 epochs on GEOM-Drugs during the training process.

For sampling steps greater than 13, the geometric straight-line diffusion use a uniform time discretization like GeoBFN and EDM. However, according to theorem 3.1, our trajectory exhibits a larger second-order derivative at the beginning of sampling. Therefore, a more efficient discretization strategy is to use fine-grained discretization for larger t values. We manually set an empirical discretization strategy that yields a 1% to 10% improvement in Mol Stable when $T \leq 13$. For sampling steps greater than 13, the impact on the results is less significant ($< 1\%$). We leave the exploration of the optimal discretization strategy for straight-line diffusion for future work.

For conditional molecular generation, we follow the baseline approaches [Hoogeboom et al., 2022] by incorporating property values as additional inputs during training and sampling them from a prior distribution during inference. The results are measured using a property classification network trained on the first half of the QM9 dataset, while the remaining portion is used for training the generative model. In addition to generative model baselines, we include two basic baselines: Random and N_{atoms}. The Random baseline involves shuffling property labels in the training data and evaluating

the property classifier on the shuffled data. The N_{atoms} baseline uses the property classifier network to predict the property based solely on the number of atoms in the molecule. The property classification network is used for evaluating the generated molecules. The metric is the MAE between predicted properties and target properties. As for the conditioning machanism, property values are concatenated to atom types as EGNN's invariant input, which is in line with other baselines.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：谭敏
方向：人工智能
单位：中国科学院深圳先进技术研究院

题目：Safe navigation for robotic digestive endoscopy via human intervention-based reinforcement learning

Expert Systems With Applications 294 (2025) 128841

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Safe navigation for robotic digestive endoscopy via human intervention-based reinforcement learning ★

Min Tan [a,b], Yushun Tao [a,b], Boyun Zheng [c], Gaosheng Xie [a], Lijuan Feng [d], Zeyang Xia [e], Jing Xiong [a,b,*]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China
[b] University of Chinese Academy of Sciences, Beijing, 101400, China
[c] Department of Electronic Engineering, Chinese University of Hong Kong, 999077, Hong Kong SAR, China
[d] Department of Gastroenterology and Hepatology, Shenzhen University General Hospital, 518055, Shenzhen, China
[e] School of Mechanical Engineering, Shanghai Jiao Tong University, 200240, Shanghai, China

ARTICLE INFO

Keywords:
Robotic endoscopy
Reinforcement learning
Safe navigation

ABSTRACT

With the increasing application of automated robotic digestive endoscopy (RDE), ensuring safe and efficient navigation in the unstructured and narrow digestive tract has become a critical challenge. Existing automated reinforcement learning navigation algorithms often result in potentially risky collisions due to the absence of essential human intervention, which significantly limits the safety and effectiveness of RDE in actual clinical practice. To address this limitation, we proposed a Human Intervention (HI)-based Proximal Policy Optimization (PPO) framework, dubbed HI-PPO, which incorporates expert knowledge to enhance RDE's safety. Specifically, HI-PPO combines Enhanced Exploration Mechanism (EEM), Reward-Penalty Adjustment (RPA), and Behavior Cloning Similarity (BCS) to address PPO's exploration inefficiencies for safe navigation in complex gastrointestinal environments. Comparative experiments were conducted on a simulation platform, and the results showed that HI-PPO achieved a mean ATE (Average Trajectory Error) of 8.02 mm and a Security Score of 0.862, demonstrating performance comparable to human experts. http://siat-medical-robot-hippo.index.

1. Introduction

Clinical digestive endoscopy examinations are crucial medical procedures, regarded as the gold standard in early diagnosis of colorectal cancer (Chen et al., 2017; Martin et al., 2020). Traditional endoscopy carries the risk of endoscopic looping and, more seriously, a perforation rate of 0.1 % to 0.3 % during insertion (Ciuti et al., 2016; Huang et al., 2021; Manfredi, 2021; Pore et al., 2023). These complications can cause patient pain and potential tissue damage (Chen & Rex, 2007; Rex et al., 2015). In contrast, automated Robotic Digestive Endoscopy (RDE), including magnetic endoscope robots (Huang et al., 2021; Martin et al., 2020) and vision-based autonomous navigation robots (Mahmoud et al., 2017; Prendergast et al., 2018) are designed to overcome the problems of manually navigating the endoscope, such as the non-intuitive control and the perforation risk. However, as the use of RDE systems increases, achieving safe and efficient navigation through the unstructured and narrow digestive tract has become a critical challenge.

Currently, several medical endoscopic robots have adopted reinforcement learning (RL) for navigation (Corsi et al., 2023; Lazo et al., 2022; Pore et al., 2022; Trovato et al., 2010; Zhang et al., 2022). This approach excels in learning complex control strategies through environmental interactions, making it well-suited for dynamically changing and irregularly structured environments like human body lumens. Relevant to our work, Davide et al. Corsi et al. (2023) introduced Constrained Proximal Policy Optimization (C-PPO). By incorporating formal verification techniques, this method reduces the likelihood of agents entering extreme regions, thereby minimizing risks such as tissue damage or perforation. However, C-PPO's reliance on four discretized actions paired with sparse image blocks neglects continuous steering motions, introducing safety risks that limit its clinical applicability.

★ This work was supported in part by the National Natural Science Foundation of China (12426305, 62073309) and in part by the Guangdong Basic and Applied Basic Research Foundation (2022B1515020042). (Min Tan and Yushun Tao contributed equally to this work.)

* Corresponding author.
E-mail addresses: min.tan@siat.ac.cn (M. Tan), ys.tao@siat.ac.cn (Y. Tao), byzheng@link.cuhk.edu.hk (B. Zheng), gs.xie@siat.ac.cn (G. Xie), fenglj@szu.edu.cn (L. Feng), zxia@sjtu.edu.cn (Z. Xia), jing.xiong@siat.ac.cn (J. Xiong).

https://doi.org/10.1016/j.eswa.2025.128841
Received 29 March 2025; Received in revised form 13 June 2025; Accepted 28 June 2025
Available online 30 June 2025
0957-4174/© 2025 Elsevier Ltd. All rights reserved, including those for text and data mining, AI training, and similar technologies.

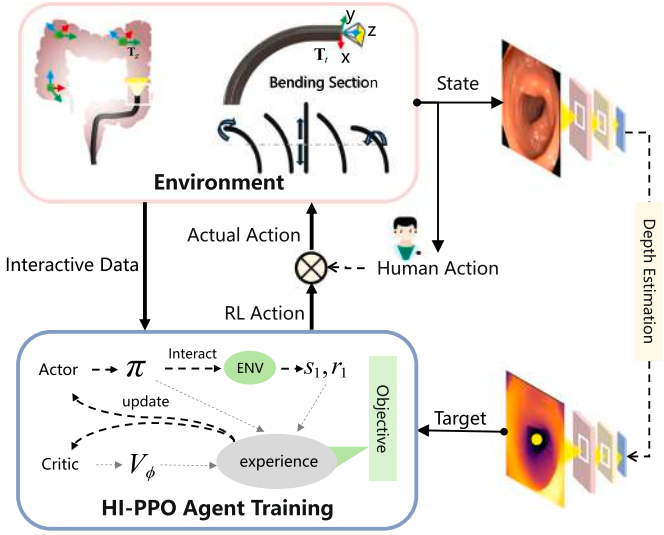


Fig. 1. Overview framework of the proposed HI-PPO.

Automated RL-based navigation algorithms lacking human intervention (HI) may lead to unsafe collisions (Corsi et al., 2023; Ng et al., 2024). To address this, recent studies (Saunders et al., 2018; Wu et al., 2023) have explored integrating HI into RL frameworks. For instance, Ameya et al. Pore et al. (2022) developed a vision-based RL control system that uses human-in-the-loop supervision to replace risky agent actions with safe human demonstrations during emergencies. A more effective strategy involves imitating human behavior. Vecerik et al. (2017) combined Behavioral Cloning (BC) with Deep Deterministic Policy Gradient (DDPG), outperforming conventional RL approaches by incorporating human demonstration data.

Despite these advancements, HI-augmented RL navigation faces two critical challenges: 1) RL systems relying solely on post-training human supervision fail to utilize prior knowledge, leading to high training costs; 2) Existing HI methods focus on isolated algorithm components rather than providing a unified integration framework, hindering safe and efficient navigation in narrow and irregular digestive tracts.

To address these gaps, we propose HI-PPO, a Human Intervention-enhanced Proximal Policy Optimization framework, as shown in Fig. 1. HI-PPO improves robotic digestive endoscopy safety and efficiency by 1) Using behavioral cloning to mimic expert actions during training and 2) Incorporating HI to enhance exploration efficiency. Unlike prior work (Corsi et al., 2023; Pore et al., 2022; Turan et al., 2019), our method models the distal bending section of commercial colonoscopes, accounting for complex turning motions. Navigation targets are derived from pre-trained network depth estimates of endoscopic images.

Our main contributions are as follows:

- First modeling of commercial colonoscope distal bending sections using HI-augmented RL, improving RDE safety through expert knowledge integration.
- A unified framework combining Enhanced Exploration Mechanism (EEM), Reward-Penalty Adjustment (RPA), and Behavior Cloning Similarity (BCS) to address PPO's exploration inefficiencies for safe navigation in complex gastrointestinal environments.
- Simulation-based validation demonstrating HI-PPO's superiority over existing methods, achieving human-expert-level performance.

2. Related work

2.1. RL-based autonomous navigation for endoscopes

Current research widely employs model-free deep reinforcement learning (DRL) to address navigation control challenges for endoscopes in complex luminal environments. For instance, Ng et al. (2024) proposed a Proximal Policy Optimization (PPO)-based control method for tendon-driven flexible endoscopes, achieving over 90 % navigation success rate with a 3 mm tolerance through simulation training, while demonstrating the transfer capability of pre-trained policies in contact environments. Trovato et al. (2010) implemented early autonomous control for colonoscopy robots using Q-learning algorithms, albeit limited by discrete action space design. Zhang et al. (2022) and Turan et al. (2019) applied PPO and DDPG algorithms to capsule endoscope control, attaining millimeter-level tracking accuracy in gastric coverage scanning and magnetic navigation tasks. Expanding beyond gastrointestinal applications, Li et al. (2023) developed an attention-augmented RL framework (RL-TEE) for ultrasound probe navigation, achieving compliant 3-DOF control in deformable esophageal environments through hybrid CNN-self-attention architectures. However, these methods generally suffer from oversimplified action space designs, exemplified by the C-PPO method (Corsi et al., 2023), which encodes only four discrete actions while neglecting continuous steering motion modeling, resulting in safety risks in practical applications.

2.2. Human intelligence-integrated reinforcement learning

To enhance navigation safety and training efficiency, researchers have explored integrating human intervention (HI) mechanisms into RL frameworks. Wu et al. (2023) developed a human-robot collaborative RL framework that achieves safe navigation in dynamic environments through prioritized experience replay and reward shaping mechanisms during sim-to-real transfer. Pore et al. (2023) and Lazo et al. (2022) employed CNN-based visual servoing systems with ergonomic soft robot designs, improving model generalization through cross-domain data training. Vecerik et al.(Vecerik et al., 2017) innovatively combined behavior cloning (BC) with DDPG, utilizing human demonstrations to address sparse reward problems, showing significant performance improvements over conventional RL algorithms. Nevertheless, current approaches still face limitations: Saunders et al. (2018) require continuous human monitoring for online intervention, incurring high training costs; Wang et al. (2018)'s intervention-aided RL reduces human involvement frequency but lacks a systematic framework for integrating prior knowledge, hindering adaptation to complex geometric constraints in narrow digestive tract lumens.

While autonomous control remains a key objective in endoscopic robotics, medical safety demands necessitate human intervention integration beyond mere post-training supervision. Current research trends embedded human expertise via behavior cloning and guided policy exploration. However, systematically integrating human priors during training while maintaining endoscopic safety constraints remains a critical challenge.

3. Methodology

This study aims to develop a robotized flexible endoscope autonomous navigation system based on deep reinforcement learning (DRL). As described in Section 3.1, we first constructed a reinforcement learning problem model incorporating a hybrid discrete-continuous action space. Section 3.2 elaborates on the triple-enhancement mechanism integrating human intervention-enhancing strategy safety and learning efficiency through Expert-guided Exploration (EEM), Reward-Penalty Adjustment (RPA), and Behavior Cloning Strategy (BCS). Finally, Section 3.3 proposes the HI-PPO integration framework, which combines prioritized experience replay and safety coverage mechanisms to achieve optimized navigation strategies with human intervention.

3.1. DRL problem formulation for RDE agent

Observation Space: The network input of the agent consists of 128 × 128-pixel RGB images, where visual features are extracted

through a CNN encoder. Depth maps are independently generated by a pre-trained Depth Anything network (Yang et al., 2024), exclusively used for computing the navigation target point p_{target}. The target coordinates (x, y) and its relative distance d_{target} to the endoscope are encoded into a 4-dimensional vector (normalized to [-1, 1]), which is concatenated with the 128-dimensional visual features from the CNN, forming a 132-dimensional State representation. This design effectively reduces computational overhead while ensuring the transmission of critical geometric information required for spatial reasoning to the policy network.

Action space: The agent's action space adopts a hybrid discrete-continuous architecture, comprising two independent decision branches for motion control and bending control, achieving 6-DOF manipulation through the Unity physics engine. The motion control branch outputs a 3D continuous action vector $\mathbf{v} \in [-1, 1]^3$, which is directly converted into instantaneous displacement increments $\Delta\mathbf{p} = \mathbf{v} \cdot 10\text{mm/s} \cdot 0.1s$ at the endoscope tip in the virtual environment. This displacement is constrained by depth map-based lumen visibility analysis–when the area of the largest connected region falls below 5% of 128^2 pixels (i.e., $L = 0$), safety mechanisms override speed control. The bending control branch employs a quaternary discrete action set $\{0 : \text{forward bend}, 1 : \text{backward bend}, 2 : \text{left bend}, 3 : \text{right bend}\}$, where each action triggers a 5° bending angle increment. The Unity-built-in joint constraint system ensures bending ranges do not exceed ±90° in each direction. When angles exceed limits, the physics engine automatically executes progressive resetting: $\theta_j(t+1) = \theta_j(t) - \text{sgn}(\theta_j(t)) \cdot 5°/\text{frame}$ until safe posture is restored. Action execution follows a mode-exclusive principle: translation DOFs are locked via an animation State machine during bending to prevent non-physical deformation of virtual soft tissues caused by composite motions.

Reward function: Our primary objective is to make the RDE agent reach the colon terminus while avoiding tissue damage. To achieve this, the agent must maintain the depth map-derived local navigation point p_{target} near the image center $p_{center} = (64, 64)$. We construct a dense reward function based on this principle:

$$r_t = \begin{cases} \left(1 - \frac{\|p_{target} - p_{center}\|_2}{64}\right), & \text{target visible } (L = 1) \\ -0.5, & \text{target lost } (L = 0) \end{cases} \tag{1}$$

where the normalization factor $D_{max} = 64$ corresponds to half the image width. The visibility flag L is determined in real-time through connected-component analysis of the depth map–valid when the largest connected region exceeds 5% of 128^2 pixels. To detect p_{target} in endoscopic images, we implement a depth estimation algorithm based on (Yang et al., 2024) (see Fig. 2), where p_{target} is computed as the centroid of the farthest connected domain from the endoscope in the depth map. A terminal reward of +10 is granted upon successful navigation to the endpoint, while a penalty of -10 is applied if the agent returns to the starting point without human intervention.

3.2. Human intervention techniques

To enhance the safety of RDE navigation, we propose three mechanisms to enable real-time navigation actions to adapt to complex environments. Enhanced Exploration Mechanism (EEM) integrates expert

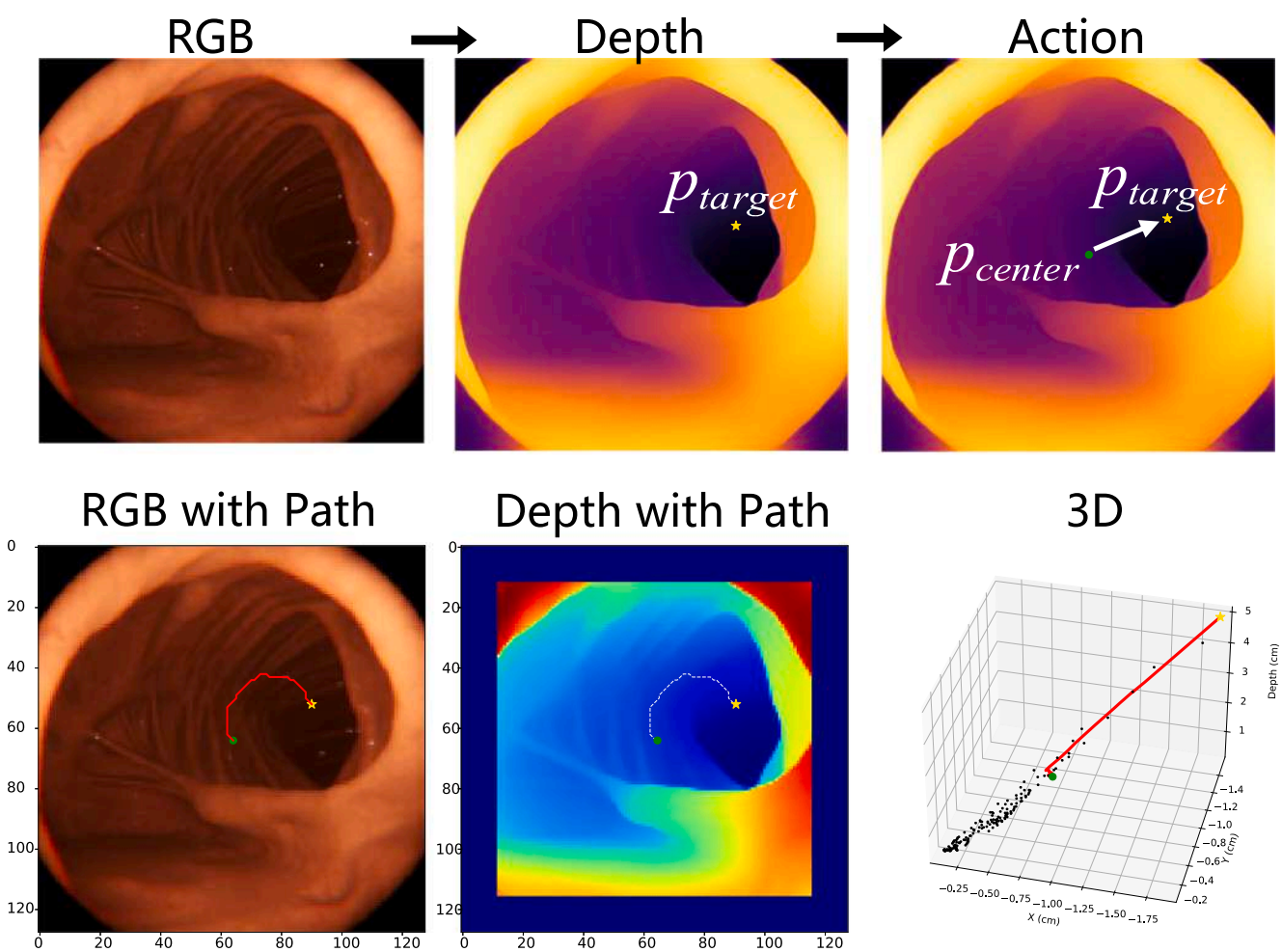


Fig. 2. Detection of endoscopic navigation target point using depth estimation and connected component centroid analysis.

knowledge by improving State exploration efficiency, where human intervention takes control when agents get stuck. The Reward-Penalty Adjustment (RPA) promotes safer policy learning by penalizing unsafe actions through a penalty score during initial interventions. Behavior Cloning Similarity (BCS) improves learning performance by guiding the agent to imitate expert actions, integrating similarity objectives into the actor-critic network for optimal policy learning.

Enhanced exploration mechanism (EEM): While standard PPO expands the agent's exploration space through stochasticity, unsafe exploration in confined colonoscopy scenarios may increase patient injury risks. Specifically, the high-dimensional State space and complex navigation paths in colonic environments require extensive samples due to greedy exploration, leading to prolonged training times and suboptimal trajectories (Corsi et al., 2023).

In contrast, human-guided exploration enhances this process through real-time intervention and expert knowledge, replacing inefficient actions with expert-corrected ones. We propose the Enhanced Exploration Mechanism (EEM) that modifies the behavior policy π^{b} to incorporate human actions during interventions:

$$\pi^{\mathrm{b}}(a_t|s_t) = \begin{cases} \pi_\theta(a_t|s_t), & M_t = 0 \\ \delta(a_t = a_t^H), & M_t = 1 \end{cases} \tag{2}$$

where M_t is a binary intervention indicator ($M_t = 1$ during human intervention; otherwise $M_t = 0$), and a_t^{H} denotes the human action at timestep t. This mechanism ensures direct integration of expert demonstrations into the learning process, improving data quality and exploration efficiency.

Reward-penalty adjustment (RPA): An instantaneous intervention penalty term is added to the base reward r_t:

$$r_t^{\mathrm{new}} = r_t + \lambda \cdot \delta(M_t = 1 \wedge M_{t-1} = 0) \tag{3}$$

where $\lambda = -2.5$ quantifies the intervention penalty magnitude, and the penalty is triggered exclusively when human intervention is initiated ($M_t = 1$ with $M_{t-1} = 0$). This design strategically applies a substantial negative reward at the moment of intervention activation, which serves two synergistic purposes: it incentivizes the agent to proactively avoid behaviors that necessitate human takeover, while maintaining policy optimization stability by restricting penalty imposition to discrete intervention onset events rather than continuous penalization throughout the intervention period.

Behavior cloning strategy (BCS): In addition to reward adjustment, we enhance policy learning by using behavior cloning as an auxiliary objective. Specifically, we add an action similarity term to the standard PPO objective function, enabling the agent to imitate expert behaviors while optimizing the cumulative reward. The BCS-integrated objective function can be denoted as follows:

$$\begin{aligned} \mathcal{J}(\theta) = \mathbb{E}_t\big[\min\big(r_t(\theta)A_t, \mathrm{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)A_t\big)\big] \\ + \alpha \cdot \mathrm{sim}(a_t^\theta, a_t^{\mathrm{H}}) \end{aligned} \tag{4}$$

where $\mathrm{sim}(\cdot)$ is the similarity metric, and $\alpha \in [0,1]$ is the weight coefficient for imitation learning. This design directly aligns the agent's action output with the deterministic actions of human experts on a per-dimension basis, avoiding the incompatibility problem between probability distribution assumptions and deterministic actions (Algorithm 1).

3.3. HI-PPO framework integration

Prioritized experience replay optimization: Experience sampling probabilities are determined by $p(\varsigma_i) \propto \hat{\rho}_i^\alpha$. This mechanism prioritizes experiences with high critic prediction errors and significant human-RL advantage discrepancies during multiple PPO update epochs, thereby enhancing expert experience utilization. Within the PPO framework, experience priority $\hat{\rho}_i$ integrates critic network errors and human action advantages:

$$\hat{\rho}_i = \underbrace{|V_\theta(s_i) - V_{target}|}_{\text{Critic error}} + \beta \cdot \underbrace{\max(0, A_i^H - A_i^{RL})}_{\text{Human-RL advantage gap}} + \epsilon \tag{5}$$

where $V_{target} = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n V_\theta(s_{t+n})$ denotes n-step returns, $\epsilon \in R^+$ is a small normal number to ensure that the value is greater than zero, and A_i^H represents human action advantage computed via actual trajectory returns:

$$A_i^H = \sum_{k=0}^{T-t} \gamma^k r_{t+k}^H - V_\theta(s_t) \tag{6}$$

where A_i^{RL} estimates RL action advantage using Generalized Advantage Estimation (GAE):

$$A_i^{RL} = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t) \tag{7}$$

where β modulates human guidance intensity, and ϵ ensures minimal sampling probability for all experiences.

As a result, the retrieving probability of the tuple ς_i is calculated by the following probability density function p^{new}:

$$p(i)^{new} = \frac{\hat{\rho}_i^\alpha}{\sum_{k=1}^{|D|} \hat{\rho}_k^\alpha} \tag{8}$$

where α is the scaling coefficient and $|D|$ is the size of the experience replay buffer.

Imitation learning enhancement: Building upon PPO's clipping mechanism, this design incorporates a KL divergence constraint to align policy updates with expert behavioral patterns while maximizing expected returns. The policy optimization objective is enhanced through human demonstration guidance. The modified policy loss function from

Algorithm 1 HI-PPO of RDE.

Input: Environment Env, human policy π^H, training epochs E
Output: Optimized policy π_θ

```
Estimate depth map D_t ← DepthAnything(I_t)
Initialize π_θ, V_θ, replay buffer D, priority coefficient β, safety timer τ = 0
for episode = 1 to E do
    Reset Env, obtain initial State s_1, initialize trajectory buffer Q = ∅
    while not terminated do
        if human intervenes then
            a_t ← π^H(s_t), M_t ← 1
        else
            a_t ← π_θ(s_t), M_t ← 0
        Safety override:
        if target lost (L = 0) then
            a_t[vel] ← 0 Freeze insertion speed
            τ ← τ + Δt
            if τ ≥ 5s then
                Extract recent 50 positions from Q, compute smoothed path {q̃_k}
                a_t[vel] ← −0.5 · sat(||q_t − q̃_{t−1}||_2/10mm)
        else
            τ ← 0
        Execute a_t, observe r_t, s_{t+1}, append pre-action position q_t^pre to Q
        Compute modified reward r'_t = r_t − 2.5 · δ(M_t = 1)
        Store transition ζ_t = (s_t, a_t, M_t, r'_t, s_{t+1}) in D
        Sample batch, update priorities via ρ̂_i = |V_θ(s_i) − V_target| + β · max(0, A_i^H − A_i^RL)
    Update V_θ and π_θ
```

Eq. 4 becomes:

$$\mathcal{J}(\theta) = \mathbb{E}_t\left[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)\right] + \lambda \cdot \omega_t \cdot D_{KL}(\pi_\theta(\cdot|s_t)\|\pi^H(\cdot|s_t)) \quad (9)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ denotes the policy probability ratio, A_t represents the Generalized Advantage Estimation (GAE), and D_{KL} quantifies the Kullback-Leibler divergence between current and human policies. The adaptive weight $\omega_t = \frac{\max(0, A_t^H - A_t^{RL})}{\sum A_t^H}$ dynamically scales the KL term, where A_t^H is the human action advantage value. Eq. 9 ensures progressive policy alignment with expert behaviors while maintaining PPO's stable optimization characteristics.

Security mechanism integration: The algorithm interleaves RL and human actions per episode, storing experiences with intervention markers. Prioritized replay elevates critical sample usage by merging value errors and human-RL advantage gaps. At the same time, policy updates blend PPO objectives with behavior cloning loss for efficient human-RL co-learning. A safety protocol operates to mitigate depth calculate failures: upon target loss ($L = 0$), forward velocity is halted ($a_t[vel] \leftarrow 0$) with bending retained for reorientation; if vision remains lost > 5 s, a smooth retraction path is generated from the last 50 trajectory points, guiding withdrawal at ≤ 0.5 mm/s until visibility recovers or safe zones are reached. Path smoothing and velocity saturation ensure tissue protection during the emergency retreats.

4. Experimental verification

This study constructs the HI-PPO framework using triple-enhancement mechanisms (EEM, RPA, and BCS) to realize human-collaborative endoscopic autonomous navigation. The experimental system is developed on a Unity simulation platform. To ensure real-time performance, coroutines are employed to enable asynchronous operations without blocking the main thread during Unity-based reinforcement learning. The main thread, utilizing the ml-agents pipeline, captures Camera Sensor images as observations and processes them through a simple CNN. Depth maps are retrieved via a separate asynchronous thread. Additionally, the target point position is queried every six frames which is viable due to Unity's high frame rate and the minimal short-term positional deviation. To validate the method, comparative experiments include: 1) Navigation performance comparison with DRL methods (SAC, PPO, and C-PPO); 2) Component ablation studies for triple-enhancement mechanisms. Evaluation metrics cover Average Trajectory Error (ATE), Security Score (S), and Trajectory Smoothness (Jerk index), focusing on exploring the effects of exploration efficiency, safety constraint, and expert behavior imitation.

4.1. Experimental platform

We aim to train an RDE reinforcement learning model to control the motion of a digestive endoscopic end-effector via decision commands output by the model. However, direct access to patients for real-world experiments is challenging. Thus, we employ a Unity-based comprehensive simulation platform (Kağan et al., 2021) to advance our research. This environment can simulate realistic tissues and organs. A monocular-camera endoscope is used for training and testing in multi-segment virtual colons. Simulations were conducted on a standard keyboard-equipped computer featuring an Intel Gold 6154 CPU and NVIDIA RTX 3090 GPU. The simulation software is Unity, with algorithms and scripts written in C#. Neural network models were developed using the ML-Agents framework.

4.2. Human guidance implementation

We implemented human guidance to enable seamless human participation and withdrawal during reinforcement learning. Both humans and RL agents share identical perceptual information, including endoscopic images and target distance. Given that humans excel at providing heuristic commands rather than precise numerical inputs, the human action input mode is designed to facilitate heuristic guidance. In the simulation environment, humans issue commands by pressing keyboard keys. Eight commands are available: forward, backward, left, right, left turn, right turn, forward turn, and backward turn. Users indicate primary intent with a single key press (e.g., Left key), while key press duration specifies command intensity. For example, holding the Left key longer generates a larger steering value if the human is dissatisfied with the current turning degree. This converts human heuristic intent into concrete demonstration actions usable by the RL algorithm. When human actions are sent, they override the RL actions at that timestep, as detailed in Table 1.

Table 1
Human intervention commands and their corresponding keys.

Type	Command	Key Set
Translation	Forward	Key_up
	Backward	Key_down
	Left Translation	Key_left
	Right Translation	Key_right
Turning	Upward Turn	B + Key_up
	Downward Turn	B + Key_down
	Left Turn	B + Key_left
	Right Turn	B + Key_right

4.3. Evaluation metrics

RL training was conducted in simulation environments, with overall performance evaluated using the average reward from 5 training runs. The trained RL agent was subsequently tested on the No.3 colon model. The evaluation utilized three metrics: Average Trajectory Error (ATE), Security Coefficient, and Jerk Index, as shown in Table 2.

4.4. Implementation details

Training and evaluation of the proposed framework were conducted on three colon models of increasing complexity, as detailed in Fig. 3. Colon No.1 (Ozyoruk et al., 2021) and No.2 [1] are for training while No.3 (Kağan et al., 2021) is for inference, which are 3D models from CT and Unity-based lighting restoration helps facilitate the transfer of algorithms and results to real-world scenarios.

The primary objective was to evaluate training differences between the proposed HI-PPO and standard PPO (Baseline) methods. The evaluation followed three sequential steps. **Step 1:** Train five instances of both algorithms for two million steps on the No.1 colon model with different random initializations. Select the top 200 policies based on success rates during training (Fig. 5). **Step 2:** Continue training these 200 policies on the No.2 colon model, retaining the best ten models. **Step 3:** Test the final models on the No.3 colon model, with segment-specific evaluation according to difficulty levels (Fig. 4).

5. Results

5.1. Impact of human intervention on training

First, we evaluated and compared the proposed HI-PPO algorithm with the Baseline. A fair comparison was achieved by training both methods five times using the same random seed sequence. Humans intervened only in the early training phase for the human-involved HI-PPO algorithm. The rewards during training are shown in Fig. 5. In the No.1

[1] https://www.turbosquid.com

Table 2
Key Evaluation Metrics.

Metric	Formula	Definition
Average Trajectory Error (ATE)	$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}\lVert \mathbf{p}_i - \hat{\mathbf{p}}_i \rVert_2$	Mean Euclidean distance between actual ($\mathbf{p}_i$) and estimated ($\hat{\mathbf{p}}_i$) 3D points.
Jerk Index	$\text{Jerk} = \frac{d^3x}{dt^3}$	Rate of acceleration change (m/s^3), quantifying motion smoothness. High values indicate potential tissue damage risk.
Security Coefficient	$S = 1 - \left(\alpha \cdot \frac{f(D)}{N} + \beta \cdot \frac{C}{N} + \gamma \cdot \frac{\sum F_i}{C \cdot F_{\text{thr}}}\right)$	Composite safety score considering: distance violations ($D = 5$ mm) (Fig. 4), collision count (C), and force magnitude (F_i) with weights $\alpha = 0.4, \beta = 0.5, \gamma = 0.1$.

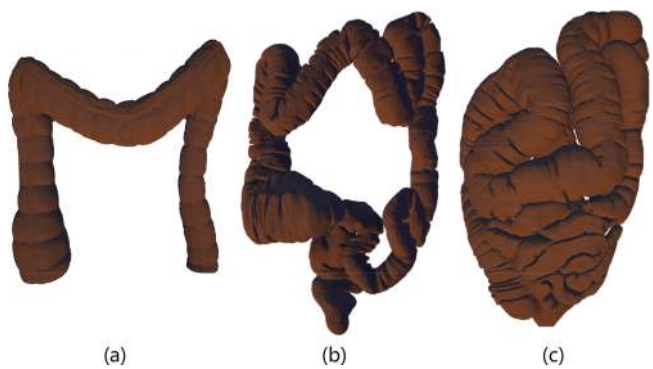


Fig. 3. The colon model used in the experimental phase. (a) Simple colon No.1 for training. (b) Complex colon No.2 for training. (c) Complex colon No.3 for inference.

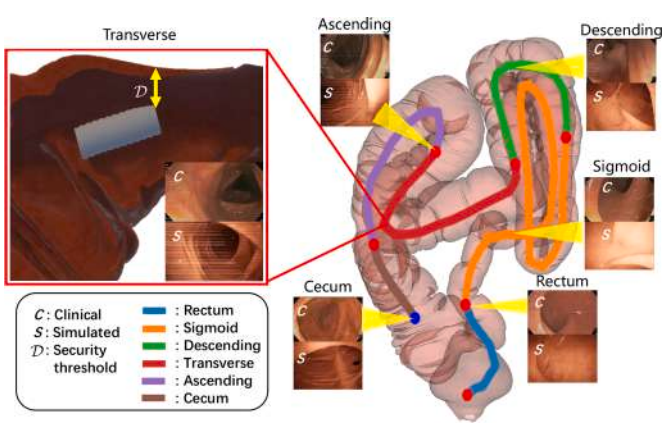


Fig. 4. Experiment configuration and environmental visualization based on different anatomical segments.

colon, the "With HI Intervention" group (orange) saw a significant increase in normalized cumulative reward as training steps progressed, far exceeding the Baseline (blue) in later stages, with minimal fluctuation across 5 training runs. In the No.2 colon, the HI group's rewards fluctuated significantly, still outperforming the Baseline. The red segments in the figure represent the human intervention process. Results show that the Baseline's curve rose slowly due to the lack of an effective exploration mechanism. The HI intervention mechanism aids model learning, leading to more remarkable cumulative reward improvements. These results validate the advantage of HI-PPO in optimizing the learning process and enhancing training performance through intervention in colon models of varying complexities, proving more effective than the baseline method.

5.2. Quantitative comparison and evaluation

Then, we validated the effectiveness of HI-PPO components in the No.3 colon model using a segmented evaluation strategy. For anatomical differences, such as the rectum's complex bending, the ascending colon's entanglement tendency, and the cecum's dilated cavity-phased performance quantification identifies each module's performance bottlenecks under specific anatomical forms, providing an anatomical basis for optimizing the navigation algorithm. We compared our method against three baselines. SAC (Haarnoja et al., 2018), a widely used off-policy reinforcement learning algorithm, maximizes rewards and entropy while promoting diverse exploration. The standard PPO (Schulman et al., 2017) without human intervention served as a baseline, and C-PPO, a PPO with strong constraints, followed (Corsi et al., 2023). All methods underwent parameter-tuning training, with the ground-truth (GT) path as the colon's central line. Compared to the three baselines, our method consistently achieved lower ATE in colon segments of varying lengths and complexities (Table 3). Specifically, HI-PPO reduced the average tracking error (ATE) by 3.7 %–33.2 % versus suboptimal algorithms across all 6 segments. The descending colon's ATE was 9.269 mm, only 3.01 % behind human experts. Standard deviation results show our method's higher stability and minimal error fluctuation. Clinically, with the colon's average diameter at 70 mm (Bassler, 1920), a 10 mm margin is a safe distance in robot-assisted colon surgery (Park & Baik, 2016; Marks et al., 2024). Our method stayed within this range except in the complex sigmoid segment. The sigmoid's safety coefficient $S = 0.8314$ exceeded human experts, verifying the obstacle-avoidance ability of our method in complex bends. The human intervention module boosted cecum security (0.9107) by 5.4 % compared to C-PPO. HI-PPO's 8.436 mm ATE set a record in the ascending colon, though 6.9 % behind the human's 7.85 mm due to path drift from the lengthy tract. In addition,

Table 3
Quantitative comparison across anatomical segments (ATE: mm, Security Coefficient: S).

Colon Segment	SAC		PPO		C-PPO		HI-PPO		Human Expert	
	ATE (Std)	Security	ATE (Std)	Security	ATE (Std)	Security	ATE (Std)	Security	ATE (Std)	Security
Rectum	4.915 (0.977)	0.5531	4.687 (0.590)	0.5867	3.523 (0.856)	0.6478	2.353 (0.629)	0.7749	**1.98 (0.48)**	**0.8121**
Sigmoid	20.12 (13.61)	0.5785	33.38 (17.88)	0.6286	22.15 (12.43)	0.6964	**16.15 (11.37)**	**0.8314**	17.12 (10.82)	0.8303
Descending	20.28 (3.882)	0.6853	17.97 (7.271)	0.7438	13.82 (6.154)	0.7687	9.269 (3.687)	0.8465	**8.99 (3.05)**	**0.8590**
Transverse	15.35 (9.062)	0.7850	9.001 (6.616)	0.8140	8.673 (5.824)	0.8659	**8.348 (5.699)**	0.8970	8.77 (5.41)	**0.9021**
Ascending	9.53 (5.659)	0.7789	15.80 (7.616)	0.8274	12.25 (6.832)	0.8574	8.436 (4.004)	0.9145	**7.85 (3.65)**	**0.9202**
Cecum	12.20 (2.199)	0.7888	3.984 (2.329)	0.7927	3.762 (2.058)	0.8638	3.597 (1.929)	0.9107	**3.49 (1.85)**	**0.9874**
Mean (Approx)	13.73	0.695	14.14	0.732	10.70	0.783	**8.02**	0.862	8.03	**0.885**

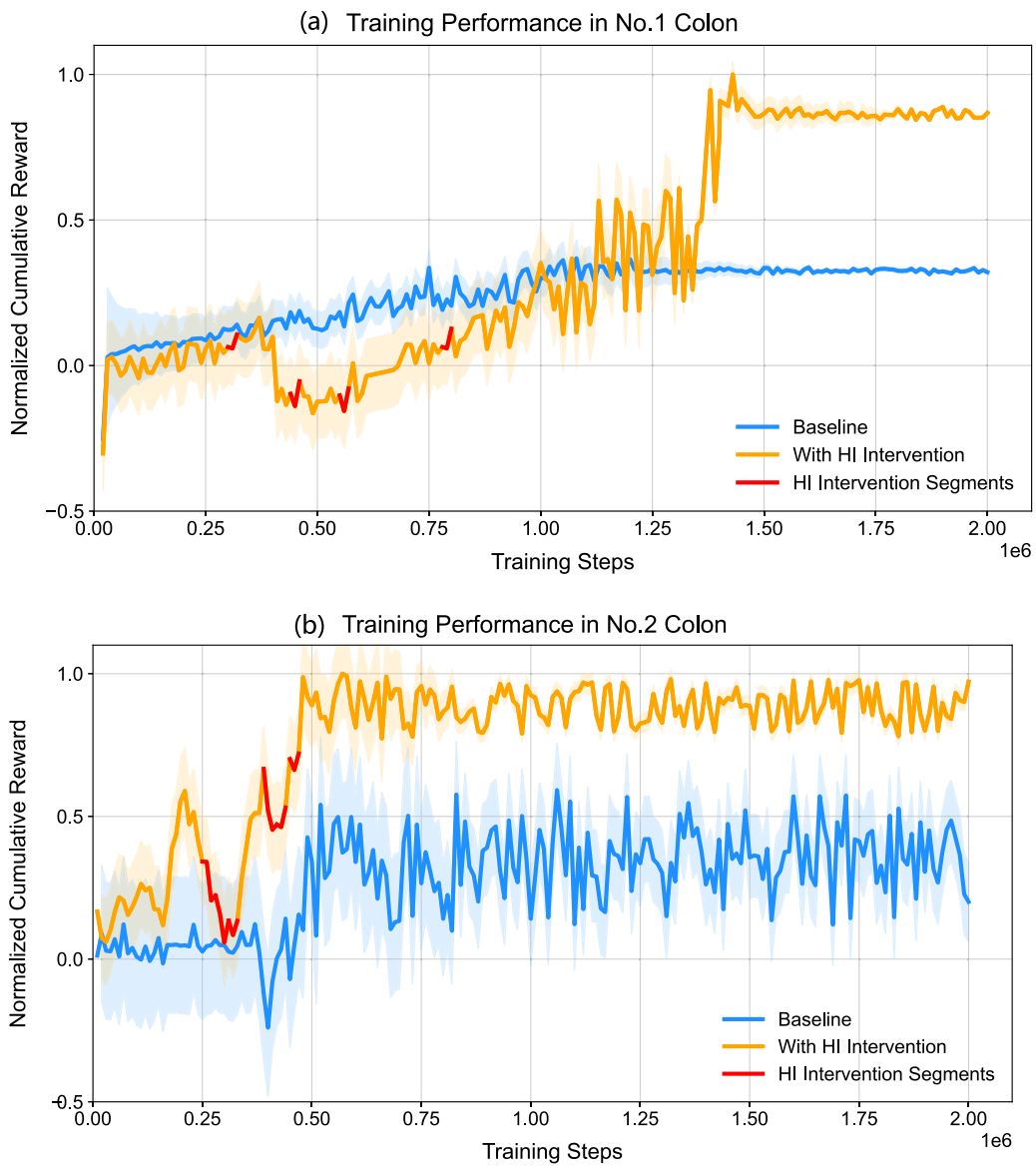


Fig. 5. The learning curve of HI-PPO training on colons of different complexity. Two colon models were used. Cumulative rewards are normalized in the range [-1, 1]. The shaded area represents the range of values obtained over 5 training sessions.

Table 4
Ablation study of HI-PPO across anatomical segments (ATE: mm).

Colon Segment	PPO (Baseline)	+ EEM	+ RPA	+ BCS	EEM + RPA	EEM + BCS	RPA + BCS	HI-PPO (Full)
Rectum	4.687 (0.590)	3.952 (0.621)	4.123 (0.602)	3.845 (0.598)	3.521 (0.615)	3.312 (0.607)	3.674 (0.594)	**2.353 (0.629)**
Sigmoid	33.38 (17.88)	28.15 (15.23)	29.74 (16.02)	27.89 (14.97)	24.57 (13.45)	22.83 (12.69)	25.41 (13.87)	**16.15 (11.37)**
Descending	17.97 (7.271)	14.32 (6.854)	15.08 (7.123)	13.75 (6.215)	12.04 (5.987)	11.29 (5.432)	12.87 (6.102)	**9.269 (3.687)**
Transverse	9.001 (6.616)	8.892 (5.823)	8.945 (6.102)	8.654 (5.632)	8.814 (5.712)	8.645 (5.987)	8.892 (5.614)	**8.348 (5.699)**
Ascending	15.80 (7.616)	12.45 (6.324)	13.27 (6.785)	11.98 (6.012)	10.23 (5.432)	9.854 (5.112)	10.67 (5.687)	**8.436 (4.004)**
Cecum	3.984 (2.329)	3.852 (2.015)	3.867 (2.104)	3.821 (1.987)	3.612 (1.954)	3.625 (1.992)	3.614 (1.945)	**3.597 (1.929)**

HI-PPO outperformed baselines in safety across segments: the rectum's $S = 0.7749$ was 19.6 % higher than C-PPO. While the sigmoid's ATE (16.15 mm) slightly bested human experts (17.12 mm), a standard deviation of 11.37 mm indicates stability needs further improvement. Overall, HI-PPO balances safety and accuracy, laying a technical foundation for autonomous colonoscopy navigation.

5.3. Ablation of human intervention module

As shown in Table 4, the HI-PPO method demonstrates significant advantages in colon navigation tasks. The complete model reduces the average tracking error (ATE) by 49.8 % and 51.6 % compared to baseline PPO in key anatomical segments like the rectum and sigmoid colon. In

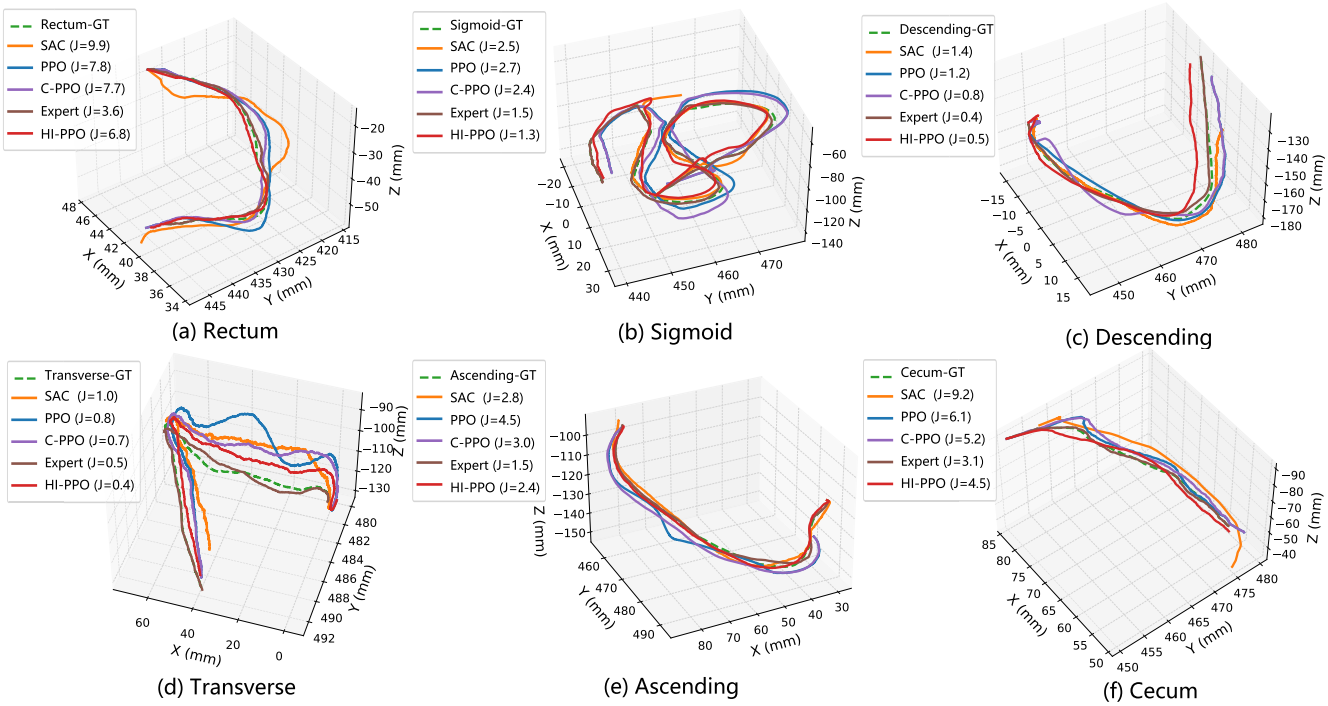


Fig. 6. Visualization of navigation trajectories in different anatomical segments. Jerk index (J, unit: mm/s^3) is overlaid to quantify motion smoothness, where lower values indicate reduced mechanical stress on the intestinal mucosa.

the complex sigmoid colon, the collaboration of the EEM environment-exploration module and RPA path-planning cuts errors from 33.38 mm to 24.57 mm, with standard deviation shrinking by 24.7 %, showcasing these two modules' adaptability to complex anatomy. BCS plays a key role in the descending colon, enabling the HI-PPO model to achieve the 13.75 mm, resulting in a 23.5 % improvement over single modules. The cecum's error of HI-PPO (3.597 mm) needs improvement, likely due to path cumulative effects. A backtracking correction mechanism is planned for improvement.

5.4. Navigation trajectory evaluation

From Fig. 6, HI-PPO provides a more reliable movement strategy, approximating expert operations for colon navigation tasks. Though the ATE in the sigmoid colon segment is prominent due to its length and curvature, our method maintains trajectory consistency with GT. As shown in Fig. 6, compared with SAC (J = 9.9) and PPO (J = 7.8), HI-PPO reduces deviations in complex segments like the rectum (J = 6.8) and transverse colon (J = 0.4), verifying its ability in motion smoothness. In the sigmoid colon, the SAC (J = 2.5) shows apparent deviations in the latter part of turns, deviating from the expected path in the intestinal environment. HI-PPO effectively alleviates the severe oscillation of other RL algorithms by integrating human intervention skills. It is also can be observed that our method can actively adjust the movement direction during sharp turns to reduce acceleration peaks (Fig. 6 (b)). In the descending colon Fig. 6(c) and transverse colon Fig. 6 (d), HI-PPO's J values (0.5 and 0.4) remain low, with trajectories closer to the expert path than algorithms like SAC and PPO. In the ascending colon Fig. 6(e) and cecum Fig. 6(f), the Jerk index of our method is not the lowest compared to the expert. Overall, HI-PPO improves trajectory stability compared with SAC, PPO, and C-PPO, decreasing stress concentration on intestinal mucosa via safety mechanisms. It achieves smoother motion control in most colon segments, validating its effectiveness in navigation trajectory optimization.

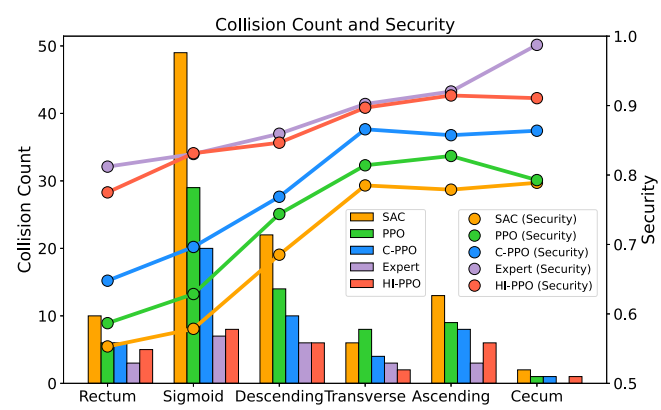


Fig. 7. Collision and security.

5.5. Safety evaluation

Fig. 7 presents collision counts and safety coefficients of different models across colon segments. In the rectum, HI-PPO's collision count (10 collisions) is lower than SAC (10 collisions) and PPO (6 collisions), with a safety coefficient of 0.77, which is higher than SAC's 0.55 and PPO's 0.58. In the sigmoid colon, HI-PPO's collision count (8 collisions) is lower than PPO's 29 collisions, and its safety coefficient exceeds 0.83, notably better than SAC's 0.57. In descending and transverse colon regions, HI-PPO maintains low collision counts and stable safety coefficients of ≥0.84. Compared to experts, HI-PPO's safety coefficients in multiple segments closely match expert levels (e.g., transverse and ascending). Overall, HI-PPO achieves lower collision counts (e.g., rectum and transverse) and higher safety coefficients in most segments than SAC and PPO through human intervention techniques, effectively balancing collision control and safety. This provides a more reliable safety strategy for colon navigation scenarios.

Fig. 8 shows collision force distributions of different methods across colon segments. HI-PPO's collision forces in the rectum concentrate at

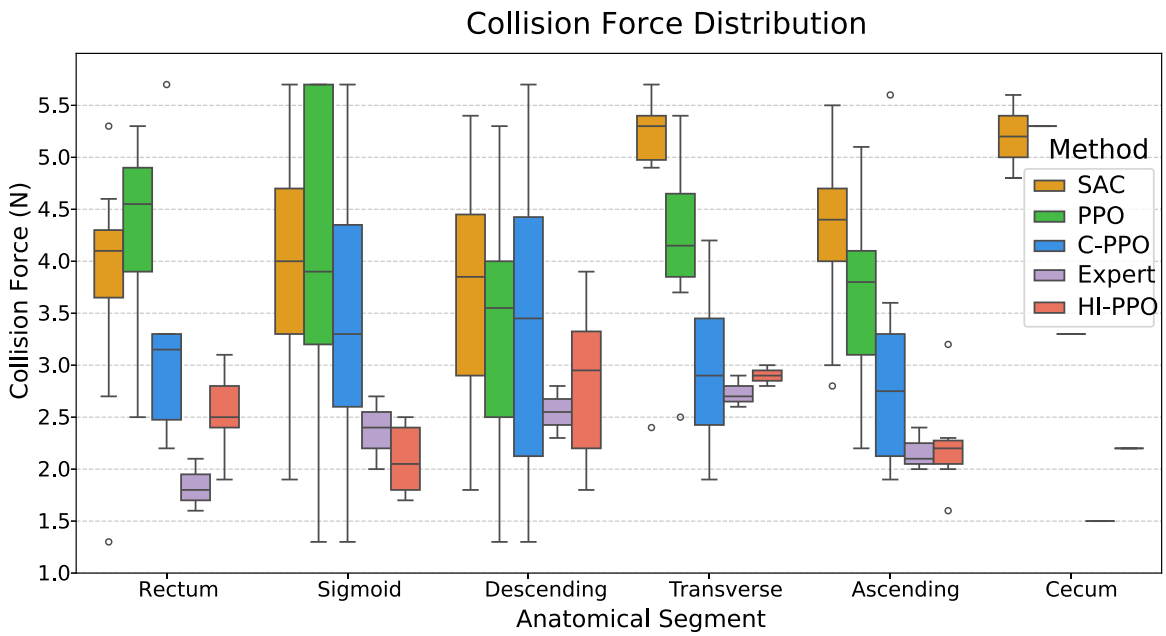


Fig. 8. Collision Force Distribution.

2.0 ~ 2.8 N, lower than SAC (3.5~4.5 N) and PPO (3.5~5.0 N). HI-PPO's force quartile (1.7~2.5 N) is also lower than other methods in the sigmoid colon. In descending and transverse colon regions, HI-PPO's median collision forces remain lowest (e.g., transverse colon 2.7 N vs. C-PPO's 2.9 N). These results benefit from HI-PPO's security mechanism integration detailed in Section 3.3 that regulates operational forces and reduces excessive intestinal wall contact. Compared to SAC and PPO, HI-PPO achieves superior collision force control across all segments, reducing mechanical damage risks to the intestinal mucosa. This validates its safety advantages from a mechanical interaction perspective, offering a more reliable force control solution for colon navigation.

6. Discussion

The proposed HI-PPO framework represents an advancement in robotic digestive endoscopy by integrating human intervention into reinforcement learning to address potential safety challenges in complex anatomical environments. However, several limitations warrant consideration.

First, the sigmoid colon poses a higher challenge to the algorithm's robustness due to its high curvature. When handling sigmoid colon curvature-mutation scenarios, the strategy π_θ of the current algorithm insufficiently captures and utilizes curvature-mutation-related features in the state s_t, and lacks a dynamic adjustment mechanism specifically for curvature mutation, failing to timely adjust the generation logic of the action a_t based on mutation characteristics. Designing a dedicated dynamic adjustment mechanism for curvature mutation to optimize the generation logic of the action a_t may enhance the algorithm's stability in complex sigmoid colon environments.

Second, while the Unity simulation faithfully reconstructs colon morphology, it lacks real-time tissue deformation, which may affect force feedback accuracy. Addressing this requires integrating finite element models of colon mechanics into the simulation framework. Hierarchical Reinforcement Learning (HRL) (Pateria et al., 2021) emerges as a promising solution to better adapt to variable anatomical structures, especially in complex regions like the sigmoid colon. HRL extends RL to intricate long-term navigation tasks and decomposes complex challenges to subtasks via policy hierarchy. At the high-level policy stage, anatomical knowledge is utilized to define subtasks tailored to the sigmoid colon's unique challenges, including navigating sharp curvatures or simple translation. Meanwhile, the low-level policies focus on real-time adjustments, leveraging immediate sensor data such as depth images and contact forces to execute fine-grained actions. This hierarchical division of the decision-making process may enable the algorithm to better cope with the dynamic and variable nature of the sigmoid colon.

Third, in this study, we focused on tip-based safety assessment due to the endoscope tip is the primary actively controlled component in clinical reality, and its motion directly impacts diagnostic accuracy and patient safety. While the shaft is designed to be passive and flexible, ignoring shaft-tissue interactions may underestimate risks in highly curved segments. This limitation may lead to optimistic safety assessments in complex anatomies. For future versions, we plan to: 1) integrate physics-based shaft dynamics using finite element models to simulate tissue friction and bending forces; 2) deploy distributed force sensors along the shaft to provide real-time interaction feedback; and 3) expand the RL state space to include shaft pose parameters. These enhancements will enable the algorithm to optimize both tip and shaft behaviors, improving safety margins in tortuous lumens.

For the depth prediction, while the pre-trained network shows acceptable cross-domain robustness with SSIM differences $\leq 17.5\,\%$ across Unity, physical phantom, and clinical images, this highlights need to enhance generalization for real-world clinical applications. To address the generalization challenges of depth estimation networks in clinical settings, we plan to: 1) use data augmentation via domain-randomization techniques, such as introducing synthetic noise, dynamic illumination variations, and simulations of clinical artifacts; 2) employ transfer learning to fine-tune the pre-trained network on a larger, multi-center clinical dataset to reduce the simulation-reality performance gap; 3) integrate self-supervised learning by pretext tasks like depth completion and motion prediction, leveraging unlabeled clinical data to enhance feature learning efficiency.

For computational time, there is a qualitative difference in execution time between single-threaded and coroutine-based processing methods. For example, single-threaded execution times for different models were approximately 900ms, whereas coroutine-based execution times were within 220ms. For overall inference time across colon segments, our model demonstrated superior performance: in the cecum, it achieved a runtime of ~ 8 seconds. In the sigmoid, our approach again outperformed, completing navigation in approximately 400 seconds. These results show that our model exhibits potential for significantly minimizing

inference latency in endoscopic applications. However, while these findings highlight progress toward efficient navigation, further optimization is needed to address the gap between current performance and the real-time requirements of clinical practice.

In the future, we will conduct ex-vivo model validation experiments, aiming to bridge the simulation-to-reality gap and demonstrate the advanced optimization in balancing motion accuracy and tissue protection. Additionally, exploring hierarchical policy architectures, where high-level planning handles segment transitions and low-level control manages local navigation, could improve adaptability in tortuous pathways. Clinically, the framework allows for integration with commercial endoscopic systems, positioning it as a scalable solution for enhancing procedural safety.

7. Conclusions

This study addresses the challenges of safe and efficient navigation in automated robotic digestive endoscopy (RDE) within the unstructured and narrow confines of the digestive tract. We proposed a human intervention-based PPO framework by incorporating expert knowledge to address the safety and effectiveness of RDE. Experiments have shown that HI-PPO can safely guide RDE compared to existing RL algorithms, indicating its potential for more practical application. Future work will focus on validating the framework in real environments and exploring additional methods to further enhance its safety and practicality.

Credit Author Statement

Min Tan: Conceptualization, Data curation, Methodology, Software, Writing- Original draft preparation, Visualization, Validation; **Yushun Tao:** Data curation, Software, Writing- Original draft preparation, Validation; **Boyun Zheng:** Investigation, Writing- Reviewing and Editing; **Gaosheng Xie:** Data curation, Writing- Reviewing and Editing; **Lijuan Feng:** Validation, Writing- Reviewing and Editing; **Zeyang Xia:** Methodology, Writing- Reviewing and Editing; **Jing Xiong:** Conceptualization, Supervision, Methodology, Writing- Reviewing and Editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We sincerely appreciate the support from the doctors in the Department of Gastroenterology and Hepatology at Shenzhen University General Hospital.

References

Bassler, A. (1920). Diseases of the intestines and lower alimentary tract. FA Davis Company.

Chen, C., Michael, H., & Brenner, H. (2017). The toll of not screening for colorectal cancer. *Expert Review of Gastroenterology & Hepatology*, *11*(1), 1–3. Publisher: Taylor & Francis https://doi.org/10.1080/17474124.2017.1264269

Chen, S. C., & Rex, D. K. (2007). Endoscopist can be more powerful than age and male gender in predicting adenoma detection at colonoscopy. *The American Journal of Gastroenterology*, *102*(4), 856–861. Place: United States. https://doi.org/10.1111/j.1572-0241.2006.01054.x

Ciuti, G., Caliò, R., Camboni, D., Neri, L., Bianchi, F., Arezzo, A., Koulaouzidis, A., Schostek, S., Stoyanov, D., Oddo, C. M., & others (2016). Frontiers of robotic endoscopic capsules: A review. *Journal of Micro-Bio Robotics*, *11*, 1–18. Publisher: Springer.

Corsi, D., Marzari, L., Pore, A., Farinelli, A., Casals, A., Fiorini, P., & Dall'Alba, D. (2023). Constrained reinforcement learning and formal verification for safe colonoscopy navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 10289–10294). https://doi.org/10.1109/IROS55552.2023.10341789

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning* (pp. 1861–1870). PMLR.

Huang, H.-E., Yen, S.-Y., Chu, C.-F., Suk, F.-M., Lien, G.-S., & Liu, C.-W. (2021). Autonomous navigation of a magnetic colonoscope using force sensing and a heuristic search algorithm. *Scientific Reports*, *11*(1), 16491. Publisher: Nature Publishing Group UK London.

Kağan, İ., Celik, I. O., Obeid, A., Gokceler, G. I., Ozyoruk, K. B., Almalioglu, Y., Chen, R. J., Mahmood, F., Gilbert, H., Durr, N. J. et al. (2021). VR-CAPS: A virtual environment for capsule endoscopy. *Medical Image Analysis*, *70*, 101990.

Lazo, J. F., Lai, C.-F., Moccia, S., Rosa, B., Catellani, M., de Mathelin, M., Ferrigno, G., Breedveld, P., Dankelman, J., & De Momi, E. (2022). Autonomous intraluminal navigation of a soft robot using deep-learning-based visual servoing. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 6952–6959). IEEE.

Li, K., Li, A., Xu, Y., Xiong, H., & Meng, M. Q.-H. (2023). Rl-TEE: Autonomous probe guidance for transesophageal echocardiography based on attention-augmented deep reinforcement learning. *IEEE Transactions on Automation Science and Engineering*, *21*(2), 1526–1538. Publisher: IEEE.

Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J., & Montiel, J. M. M. (2017). ORBSLAM-based endoscope tracking and 3D reconstruction. In *Computer-Assisted and Robotic Endoscopy: Third International Workshop, CARE 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 3* (pp. 72–83). Springer.

Manfredi, L. (2021). Endorobots for colonoscopy: Design challenges and available technologies. *Frontiers in Robotics and AI*, *8*, 705454. Publisher: Frontiers Media SA.

Marks, J. H., Keller, D. S., Lagares-Garcia, J. A., Schoonyoung, H. P., Farritor, S. M., Oleynikov, D., & Jobst, M. A. (2024). Safety and efficacy of a novel miniaturized robotic assisted surgery system in colectomy: A prospective, investigational device exemption clinical study using the IDEAL framework. *Diseases of the Colon & Rectum*, (pp. 10–1097).

Martin, J. W., Scaglioni, B., Norton, J. C., Subramanian, V., Arezzo, A., Obstein, K. L., & Valdastri, P. (2020). Enabling the future of colonoscopy with intelligent and autonomous magnetic manipulation. *Nature machine Intelligence*, *2*(10), 595–606. Publisher: Nature Publishing Group UK London.

Ng, C., Gao, H., Ren, T.-A., Lai, J., & Ren, H. (2024). Navigation of tendon-driven flexible robotic endoscope through deep reinforcement learning. In *2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)* (pp. 134–139). https://doi.org/10.1109/ARSO60199.2024.10557794

Ozyoruk, K. B., Gokceler, G. I., Bobrow, T. L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., & others (2021). EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis*, *71*, 102058.

Park, E. J., & Baik, S. H. (2016). Robotic surgery for colon and rectal cancer. *Current Oncology Reports*, *18*, 1–8.

Pateria, S., Subagdja, B., Tan, A.-h., & Quek, C. (2021). Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, *54*(5), 1–35.

Pore, A., Finocchiaro, M., Dall'Alba, D., Hernansanz, A., Ciuti, G., Arezzo, A., Menciassi, A., Casals, A., & Fiorini, P. (2022). Colonoscopy navigation using end-to-end deep visuomotor control: A user study. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 9582–9588). IEEE.

Pore, A., Li, Z., Dall'Alba, D., Hernansanz, A., De Momi, E., Menciassi, A., Gelpi, A. C., Dankelman, J., Fiorini, P., & Vander Poorten, E. (2023). Autonomous navigation for robot-assisted intraluminal and endovascular procedures: A systematic review. *IEEE Transactions on Robotics*, *39*(4), 2529–2548. Publisher: IEEE.

Prendergast, J. M., Formosa, G. A., Heckman, C. R., & Rentschler, M. E. (2018). Autonomous localization, navigation and haustral fold detection for robotic endoscopy. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 783–790). IEEE.

Rex, D. K., Schoenfeld, P. S., Cohen, J., Pike, I. M., Adler, D. G., Fennerty, M. B., Lieb, J. G., Park, W. G., Rizk, M. K., & Sawhney, M. S. (2015). Quality indicators for colonoscopy. *Gastrointestinal Endoscopy*, *81*, 31–53.

Saunders, W., Sastry, G., Stuhlmüller, A., & Evans, O. (2018). Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 2067–2069). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC. https://dl.acm.org/doi/abs/10.5555/3237383.3238074.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Trovato, G., Shikanai, M., Ukawa, G., Kinoshita, J., Murai, N., Lee, J. W., Ishii, H., Takanishi, A., Tanoue, K., Ieiri, S. et al. (2010). Development of a colon endoscope robot that adjusts its locomotion through the use of reinforcement learning. *International Journal of Computer Assisted Radiology and Surgery*, *5*, 317–325.

Turan, M., Almalioglu, Y., Gilbert, H. B., Mahmood, F., Durr, N. J., Araujo, H., Sarı, A. E., Ajay, A., & Sitti, M. (2019). Learning to navigate endoscopic capsule robots. *IEEE Robotics and Automation Letters*, *4*(3), 3075–3082.

Vecerik, M. et al. (2017). Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. arXiv preprint arXiv:1707.08817.

Wang, F. Zhou, B., Chen, K., Fan, T., Zhang, X., Li, J., Tian, H., & Pan, J. (2018). Intervention aided reinforcement learning for safe and practical policy optimization in navigation. In A. Billard, A. Dragan, J. Peters, & J. Morimoto (Eds.), *Proceedings of The 2nd Conference on Robot Learning* (pp. 410–421). PMLR (*vol. 87*).

Wu, J., Zhou, Y., Yang, H., Huang, Z., & Lv, C. (2023). Human-guided reinforcement learning with sim-to-real transfer for autonomous navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 14745–14759.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10371–10381).

Zhang, Y., Bai, L., Liu, L., Ren, H., & Meng, M. Q.-H. (2022). Deep reinforcement learning-based control for stomach coverage scanning of wireless capsule endoscopy. In *2022 IEEE international conference on robotics and biomimetics (ROBIO)* (pp. 01–06). IEEE.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：完颜宇洋
方向：人工智能
单位：中国科学院自动化研究所

题目：Look Before You Leap: A GUI-Critic-R1 Model for Pre-Operative Error Diagnosis in GUI Automation

Look Before You Leap: A GUI-Critic-R1 Model for Pre-Operative Error Diagnosis in GUI Automation

Yuyang Wanyan[1,2,3†*], **Xi Zhang**[3*], **Haiyang Xu**[3‡], **Haowei Liu**[3], **Junyang Wang**[3], **Jiabo Ye**[3], **Yutong Kou**[1], **Ming Yan**[3‡], **Fei Huang**[3], **Xiaoshan Yang**[1,2‡], **Weiming Dong**[1,2], **Changsheng Xu**[1,2]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]Alibaba Group
wanyanyuyang2021@ia.ac,cn, xiaoshan.yang@nlpr.ia.ac.cn,
{shuofeng.xhy, ym119608}@alibaba-inc.com

Abstract

In recent years, Multimodal Large Language Models (MLLMs) have been extensively utilized for multimodal reasoning tasks, including Graphical User Interface (GUI) automation. Unlike general offline multimodal tasks, GUI automation is executed in online interactive environments, necessitating step-by-step decision-making based on the real-time status of the environment. This task has a lower tolerance for decision-making errors at each step, as any mistakes may cumulatively disrupt the process and potentially lead to irreversible outcomes like *deletions* or *payments*. To address these issues, we introduce a **pre-operative critic** mechanism that provides effective feedback prior to the actual execution, by reasoning about the potential outcome and correctness of actions. Specifically, we propose a Suggestion-aware Group Relative Policy Optimization (**S-GRPO**) strategy to construct our pre-operative critic model GUI-Critic-R1, incorporating a novel suggestion reward to enhance the reliability of the model's feedback. Furthermore, we develop a reasoning-bootstrapping based data collection pipeline to create a **GUI-Critic-Train** and a **GUI-Critic-Test**, filling existing gaps in GUI critic data. Static experiments on the GUI-Critic-Test across both mobile and web domains reveal that our GUI-Critic-R1 offers significant advantages in critic accuracy compared to current MLLMs. Dynamic evaluation on GUI automation benchmark further highlights the effectiveness and superiority of our model, as evidenced by improved success rates and operational efficiency. The code is available at https://github.com/X-PLUG/MobileAgent/tree/main/GUI-Critic-R1.

1 Introduction

Recently, Multi-modal Large Language Models (MLLMs), leveraging their remarkable perception and reasoning capabilities, have demonstrated outstanding capabilities in various domains [2, 41]. Among these, GUI automation, as a practical multi-modal application scenario, is emerging as a significant technological revolution in artificial intelligence interactions [38, 16, 50, 15, 44, 26, 27]. To be specific, given an online GUI device and a natural language instruction, it requires the GUI agent driven by MLLMs to generate a series of precise operations similar to how humans do [48, 32], such as click, type, and scroll.

Unlike traditional offline multimodal tasks such as visual question answering [19] and optical character recognition [14], GUI automation task operates within an online interactive environment

*Equal contribution. †Work done during internship at Tongyi Lab, Alibaba Group. ‡Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

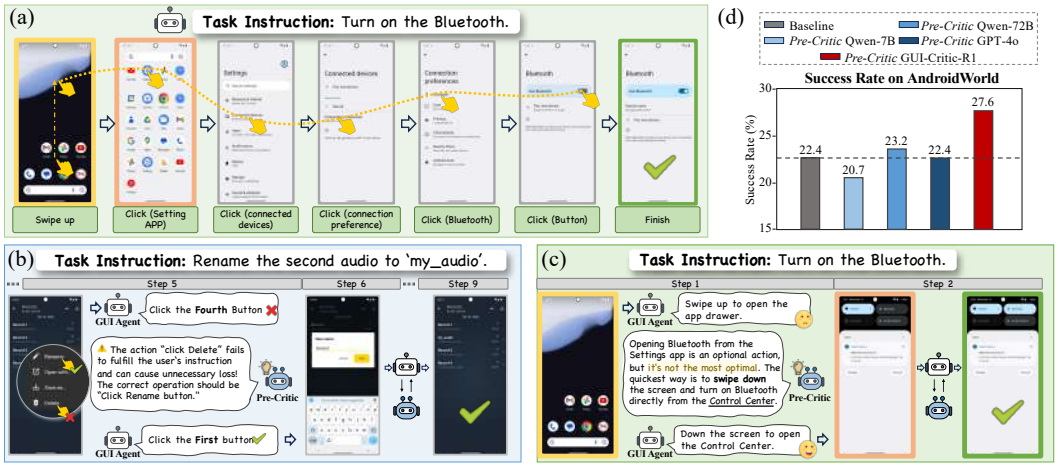


Figure 1: (a) shows an example of GUI automation. Case studies in (b-c) demonstrate how pre-critic prevents erroneous and redundant actions in GUI automation. (d) illustrates the quantitative performance comparison between pre-critic methods and baseline on AndroidWorld [27].

and has some inherent key challenges [32]. To be specific, the agents are required to generate coherent and sequential operations step by step. In this way, an error in one step can have cumulative effects on subsequent operations (*e.g., delete a file shown in Figure 1(b)*), thereby disrupting the entire interactive process. However, constrained by limited reflection capabilities, current MLLMs usually struggle to detect errors independently [34]. Therefore, to ensure the single-step accuracy, it is necessary to provide MLLM-based agents with additional feedbacks that incorporates critic analysis, such as assessments of the action correctness, potential outcomes, and action suggestions. Nevertheless, in the dynamic environment, erroneous operations usually require additional steps to correct (*e.g., refunding after an incorrect payment*), and some dangerous errors may be irreparable (*e.g., deleting files as shown in Figure 1(b)*). Therefore, to prevent these issues from happening, we believe that the critical feedback should be provided to the agent **before the action is actually executed**. Additionally, completing a user instruction on a GUI device typically entails multiple pathways. From an application perspective, the GUI agent is expected to complete instructions with an optimal path that contains the least number of steps. In this way, the prior critical feedback may also prevent the model from selecting a sub-optimal path with more steps as shown in Figure 1(a), thereby enhancing the efficiency of completing instructions.

Considering the above issues, in this paper, we propose a **pre-operative critic** (pre-critic) mechanism for GUI automation. Specifically, before performing an operation in online environments, the pre-critic mechanism first evaluates whether the operation generated by the agent is beneficial to the instruction completion, through comprehending the screenshot and analyzing potential results of the operation. Then, it provides real-time critical feedback to the agent, including the underlying causes of errors and corrective suggestions, to assist in refining its decision-making process. For example, as shown in Figure 1(b), given the instruction *Rename the current audio to 'my_audio'*, the agent initially predicts the action of *click (delete button)* in the 5-th step. Such operation may cause the audio file to be deleted and the task to fail. Fortunately, with the pre-critic mechanism, we can catch the dangerous error before executing, and provide objective feedback to the agent. Besides, Figure 1(c) illustrates a case where the pre-critic identifies a more efficient method to *turn on Bluetooth* (*i.e., enabling Bluetooth via the control center*), thereby reducing the operation steps. We also exhibit some statistic results in Figure 1(d), where the pre-critic helps increase the success rate of baseline from 22.4% to 27.6% in the dynamic GUI automation [27] *. In conclusion, these figures demonstrate that the introduction of pre-critic can effectively alleviate the aforementioned issues of error revision and operational inefficiency in online environments.

An intuitive approach to implementing a pre-critic model is leveraging existing MLLMs. However, closed-source models [13] incur heavy efficiency and cost issues, making them unsuitable for real-time GUI automation. Besides, open-source models [5, 2, 41] struggle as pre-critic models due to inherent

*More detailed experimental results about efficiency can be found in Table 2

limitations in comprehending GUI interface and forecasting interaction outcomes [16, 36, 51]. To this end, we propose a specialized 7B model **GUI-Critic-R1** that harmonizes performance and efficiency for GUI pre-critic. In particular, to enhance the model's GUI reasoning and generalization capabilities, we introduce a Suggestion-aware Group Relative Policy Optimization (**S-GRPO**). In S-GRPO, a suggestion reward is innovatively designed to refine the model's critic reasoning process and ensure it provides reliable suggestions for fixing the error operation. Moreover, since it lacks both training and test datasets for pre-critic in GUI automation, we develop a reasoning-bootstrapping based data collection pipeline to construct a **GUI-Critic-Train** and a **GUI-Critic-Test** dataset. Specifically, GUI-Critic-Train contains 6K high-quality chain-of-thought data about mobile devices for robust training. GUI-Critic-Test includes 1k samples encompassing both mobile and web scenarios, and aims to explicitly evaluate the pre-critic model's diagnostic capabilities when exposed to novel instructions or applications. In experiments, we first compare our GUI-Critic-R1 with advanced MLLMs on the GUI-Critic-Test, and find that our model achieves satisfactory results in the critic accuracy. Then, we also apply the GUI-Critic-R1 to a real-time GUI automation benchmark (*i.e.*, AndroidWorld [27]), the improvements on the success rate further demonstrate the effectiveness of the proposed method.

Our main contributions are summarized as follows:

1. To the best of our knowledge, we are the first to investigate a pre-operative critic mechanism for diagnosing GUI operations. To achieve it, we propose a Suggestion-aware Group Relative Policy Optimization strategy integrating a novel suggestion reward, and develop our pre-critic model GUI-Critic-R1. This model is capable of delivering constructive and insightful feedback to refine the GUI reasoning process.
2. We present a reasoning-bootstrapping based data collection pipeline to construct a GUI-Critic-Train dataset, comprising 6k high-quality chain-of-thought annotations. Additionally, a GUI-Critic-Test dataset is developed to comprehensively evaluate the critic model's performance in both mobile and web domains.
3. Extensive experiments on both GUI-Critic-Test dataset and dynamic GUI automation benchmark validate the efficacy of our GUI-Critic-R1 model in producing reliable judgments and feedback for GUI operations.

2 Related Work

LLM-based GUI Agent. Recently, significant attention has been directed toward Graphical User Interface (GUI) agents for task automation on smart devices designed for the mobile and web environments [50, 15, 36, 49, 56, 45, 10, 52, 20, 30, 39, 40]. Despite these advancements, GUI agents frequently make erroneous decisions. Some research has incorporated reflection modules in GUI automation frameworks to verify operation correctness based on the both current screenshot and the screenshot after execution [**?**, 16, 1]. For example, Mobile-Agent-v2 frameworks [36] employ multi-agent architectures that separate planning, decision-making, and reflection to optimize task tracking, memory, and error correction. But these methods require additional steps to undo actions and pose the risk of irreversible operations, resulting in lower efficiency and accuracy. In this paper, we introduces a pre-critic mechanism for diagnosing potential errors.

Critic Model for LLMs. Large language models (LLMs) do not always generate the best reasoning output when performing challenging inference tasks [34]. To address this limitation, several studies propose self-refinement[21, 31, 24] and self-reflect techniques [29, 9, 35, 47, 55, 6, 7] to reifine the output themselves. However, their effectiveness is largely restricted by their reliance on the inherent capabilities of LLMs, which may hinder the broader application and scalability of these methods [11]. In contrast, several studies [22, 43, 42] explore employing an independent critic model to produce natural language feedback for the evaluation of outputs generated by Large Language Models (LLMs). Critic-V [51] extends this inspiration to the area of VLMs to train a critic vision-language model to locate imperfections in visual content perception and errors in reasoning steps. However, they focus on general offline tasks. Differently, we explore the application of a critic model in more complex scenarios of GUI automation, which presents increased challenges due to its operation in an online environment. In this paper, we perform pre-operative critic to resolve the challenges.

Reinforcement Learning for Reasoning. Recent research has increasingly focused on enhancing the reasoning capabilities of LLMs through Reinforcement Learning (RL), drawing inspiration from

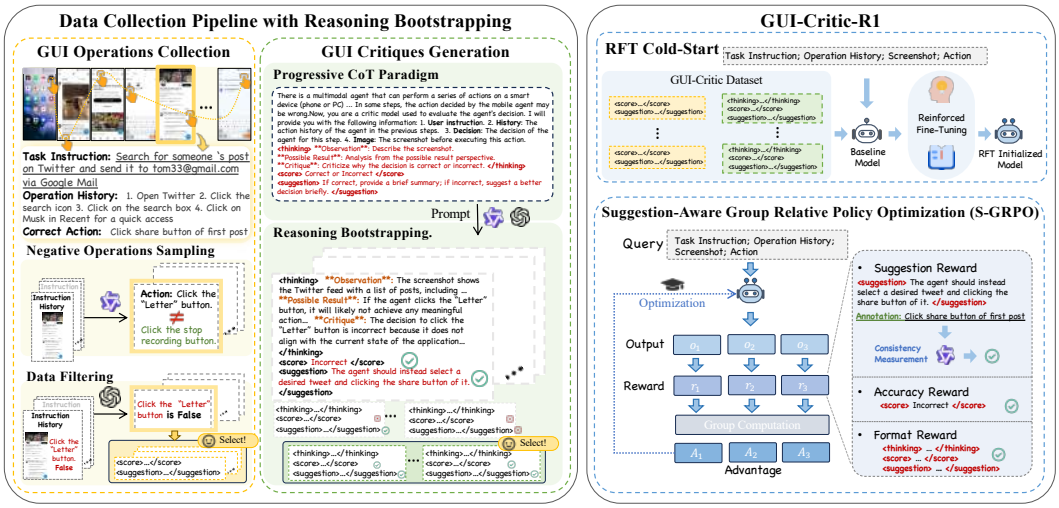


Figure 2: The left shows our reasoning-bootstrapping based data collection pipeline, including the GUI operations collection and GUI critiques generation. Specifically, a progressive CoT paradigm and a reasoning-bootstrapping strategy are employed to ensure the quality of critiques. The right illustrates the training strategy for our GUI-Critic-R1 model. The process begins with a RFT cold-start on the GUI-Critic-Train dataset, and followed by the implementation of our proposed S-GRPO. Besides, a novel suggestion reward is employed to constrain the correctness of suggestions.

models such as DeepSeek-R1 [8] and Kimi-1.5 [33]. These models employ rule-based reward mechanisms to enhance reasoning performance. Several studies attempt to adapt the idea of reinforcement learning with verifiable rewards in multimodal scenarios. For instance, R1-V [53] investigates the application of rule-based RL in geometry problems and object-counting tasks, while Visual-RFT [17] extends this approach to open vocabulary and few-shot detection, reasoning grounding, and fine-grained few-shot classification. Recent studies [12, 25, 23, 53, 46] further generalize the algorithm to address more general tasks such as multimodal mathematical reasoning, decision-making, and planning. In this paper, we extend the reinforcement learning algorithm to train a critic model that diagnoses operations in GUI automation, and propose a S-GRPO strategy with a novel suggestion reward to enhance the model's reasoning capabilities.

3 Method

3.1 Problem Definition of Pre-Operative Critic

The GUI automation task can be formally characterized as a Markov Decision Process: $\mathcal{M} = (\mathcal{E}, \mathcal{A}, \mathcal{P}, \pi_{agent})$. The state of the environment $\epsilon \in \mathcal{E}$ consists of a user instruction, historical interactions, and the current screenshot of the device. The action space $\mathcal{A}$ encompasses all available operations (actions) including *click, long press, type, scroll, home, back*, and *done*. At each step, the MLLM-based agent π_{agent} observes the environment ϵ and selects an action a. Upon executing a, it receives a novel observation ϵ', with the probability given by the state transition function $\mathcal{P}(\epsilon'|\epsilon, a)$. The agent iterates this process until accomplishing the desired instruction or encountering a terminal state.

In this paper, we propose a pre-operative critic model $\pi_{critic}(\epsilon, a)$ to critique the decision-making process of π_{agent}, before the action is actually executed. Specifically, taking state ϵ and action a as inputs, the model produces a correctness score $l \in [0, 1]$ that reflects the rightness of a. Besides, it generates critique c and corrective suggestion s in natural language, with the former explaining the rationale behind the correctness judgment and the latter suggesting a better action if a is incorrect.

3.2 Data Collection Pipeline with Reasoning Bootstrapping

In this section, we detail the data collection pipeline of our proposed GUI-Critic-Train and GUI-Critic-Test datasets. Each sample in the dataset includes the environmental state ϵ of the step, a operation candidate a, as well as annotations for correctness score l and suggestion s. This section

commences by introducing the collection of operations in Section 3.2.1, followed by the presentation of our reasoning bootstrapping method for generating critique c in Section 3.2.2.

3.2.1 GUI Operations Collection

To begin with, we collect successful automation trajectories from publicly accessible datasets, encompassing correct step-level operations across various GUI scenarios. The action in a specific state ϵ is considered as the operation candidate whose correctness score is positive ($l = 1$), and the corresponding action description is considered as the suggestion s.

Negative Operations Sampling. Subsequently, we collect samples with negative score ($l = 0$) based on the states of the above correct operations. To be specific, given these states, open-source MLLM-based agents are first employed to predict a set of operations. Then, these operations are evaluated according to rule-based criteria, and we retain those deemed incorrect. In this way, the obtained negative operations are aligned as closely as possible with the real error distribution in GUI environments, thus ensuring the quality of the dataset. The corrective suggestions for these samples are the same as the corresponding positive samples.

Data Filtering. The data collected using the aforementioned methods may not be entirely reliable. This is because that public datasets could contain erroneous annotations, and the rule-based criteria are not entirely reliable, as they may also erroneously penalize some correct operations. Consequently, further data cleaning is necessary. Specifically, we adopt GPT-4o [13] as a pre-critic model to judge the rightness of the operations in the collected samples, and we retain the samples for which the annotated scores are consistent with the judgment by GPT-4o. Finally, we denote the collected samples as $\mathcal{D}_{c_action} = \{(\epsilon_n, a_n, l_n, s_n)\}_{n=1}^{N}$.

3.2.2 GUI Critiques Generation

The $\mathcal{D}_{c_action}$ collected in the above section still lacks the critique c that elucidates why an action is considered as correct or incorrect. In this section, we resort to Chain-of-Thought (CoT) technique to enabling MLLMs to handle the challenging critique generation.

Progressive CoT Paradigm. We design a progressive CoT paradigm to helps MLLMs to perform deliberate, structured, and analytical thinking about the GUI operations. Specifically, our paradigm contains three parts `<thinking>...</thinking>` `<score>...</score>` `<suggestion>...</suggestion>`, which respectively corresponding to intermediate logical thought t, correctness score l, and suggestion s. The reasoning t contains the content of critique c. In particular, the logical reasoning in the <thinking> field is composed of several essential components:

1. `Observation: Analyze the screenshot's state, ignoring user instructions.`
2. `Possible Result: Speculate the most possible result of the action.`
3. `Critique: Assess correctness of the action with reasoning.`

`Observation` enhances the model's perceptual acuity by demanding a comprehensive understanding of spatial and contextual elements present in GUI. `Possible Result` boosts foresight by prompting the model to predict outcomes from current observations. Finally, `Critique` involves a critical analysis of actions to verify their correctness.

Reasoning Bootstrapping. Following the above format, we leverage current MLLMs to generate CoT reasoning progress. However, performing reverse annotation of CoT conditioned on ground-truth score l and suggestion s is harmful. This is because that the MLLMs may overly depend on pre-known annotations, potentially biasing the thought process toward these outcomes rather than reflecting the actual critique reasoning sequence. Therefore, we propose a reasoning bootstrapping strategy to generate high-quality thoughts without prior knowledge of the ground-truth l and s. Specifically, merely provided with environmental states ϵ and action candidates a, it forces the model to simulate a genuine critic reasoning process for max times as follows:

$$\mathbb{E}_{(\epsilon,a,l,s)\sim\mathcal{D}_{c_action}}(\bar{t}_i, \bar{l}_i, \bar{s}_i)_{i=1}^{\max} = \bar{\pi}(\epsilon, a), \quad \text{Select}\,(\bar{t}_i, \bar{l}_i, \bar{s}_i) \mid (\bar{l}_i = l) \wedge (\bar{s}_i = s), \tag{1}$$

where $\bar{\pi}$ is an excellent MLLM and $\bar{t}$ denotes the generated thought. We select reasoning outputs containing the correctness score and suggestion that match with annotations to compile $\mathcal{D}_{c_cot} = \{(\epsilon_m, a_m, l_m, s_m, t_m)\}_{m=1}^{M}$.

Finally, the GUI-Critic-Train dataset is constructed by aggregating the collected datasets, which can be represented as $\mathcal{D}_c = \mathcal{D}_{c_action} \cup \mathcal{D}_{c_cot}$. For the GUI-Critic-Test dataset that does not require the annotation of reasoning process, we construct it with the pipeline introduced in Section 3.2.1.

3.3 GUI-Critic-R1

Constructing a pre-critic model for GUI automation is not easy since it requires a thorough understanding of GUI knowledge, multimodal processing, and logical reasoning abilities. In this section, to cultivate the deliberative analysis ability of the pre-critic model for complex GUI scenarios, we propose a Suggestion-aware Group Relative Policy Optimization (S-GRPO) strategy for our GUI-Critic-R1. Specifically, we first employ Reinforcement Fine-Tuning Initialization (RFT cold-start) to stabilize the model's reasoning process. Then, the S-GRPO is adopted to further enhance the model's ability for GUI pre-critic.

3.3.1 RFT Cold-Start

We initiate model training with Reinforced Fine-Tuning (RFT) on collected GUI-Critic dataset $\mathcal{D}_c$:

$$\mathcal{L}_{rft} = \mathbb{E}_{(\epsilon,a,l,s)\sim\mathcal{D}_{c_action},\ (\hat{\epsilon},\hat{a},\hat{l},\hat{s},\hat{t})\sim\mathcal{D}_{c_cot}} - \log(\pi_{critic}(l,s|\epsilon,a)) - \log(\pi_{critic}(\hat{l},\hat{s},\hat{t}|\hat{\epsilon},\hat{a})). \quad (2)$$

The cold-start stage distills GUI knowledge from human annotations (*i.e.*, correct operations in $\mathcal{D}_{c_action}$) and distill reasoning experience from existing MLMs (*i.e.*, progressive reasoning processes in $\mathcal{D}_{c_cot}$), thereby equipping the model with foundational capabilities needed for generating operation critiques and valid feedbacks.

3.3.2 Suggestion-aware Group Relative Policy Optimization

After the RFT cold-start, we enable the pre-critic model to self-improve its reasoning ability via online reinforcement learning. Standard GRPO [8] approach samples a group of generated outputs for each input from the policy model, where each output includes CoT thought and an answer. Subsequently, the model is encouraged to favor better answers with a high reward value within the group. Since the pre-critic model is required to output additional valuable critiques and suggestions of the operation, only employing the typical reward that focus on the correctness of format and answer is not enough. Therefore, we introduce a suggestion-aware GRPO method, incorporating a novel suggestion reward specifically designed for GUI pre-critic.

Suggestion Reward. Given the suggestion annotation s', the suggestion reward r_s evaluates the model's constructive outputs as follows:

$$r_s(o) = \mathbb{I}_{similar}(s, s'), \quad (3)$$

where $o = (l, s, c)$ is the output of pre-critic model π_{critic}. $\mathbb{I}_{similar}(\cdot)$ calculates the similarity between the inputs leveraging large language models, and output a binary judgment. Based on Eq. (3), r_s provides a quantitative and objective metric to assess the correctness of suggestions, thereby facilitating more precise evaluations of π_{critic} efficacy. Moreover, this direct feedback enables models to learn from their outputs iteratively, refining their progressive CoT reasoning process to produce more accurate action critique and corrective suggestions.

Format and Accuracy Reward. Additionally, we adopt the rewards introduced in DeepSeek-R1 [8]: format reward $r_f(o)$ and accuracy reward $r_a(o)$. The former evaluates whether the outputs follow the structure described in Section 3.2.2. The latter evaluates whether the generated correctness score align with ground-truth. These rewards imposes constraints solely on the format and final answer, leaving the thought process unregulated, resulting in stimulating the model's intrinsic ability to reason.

Optmization. The final reward function $r(o)$ can be defined as: $r(o) = \lambda_f \cdot r_f(o) + \lambda_s \cdot r_s(o) + (1 - \lambda_f - \lambda_s) \cdot r_a(o)$, where λ_f and λ_s adjust the importance of the format and suggestion rewards. For each input (ϵ, a), we first sample a group of generated outputs $G = \{o_1, o_2, \cdots, o_G\}$ from the frozen policy model $\pi_{\theta_{old}}$, where each output $o = (c, l, s)$ includes critique c, score l and the suggestion s. Then we maximize the following objective and optimizes the critic model π_{critic}:

$$\mathcal{L}_{\text{GRPO}} = \frac{1}{|G|}\sum_{i=1}^{G} \min\Big(\frac{\pi_{critic}(o_i \mid \epsilon, a)}{\pi_{\theta_{\text{old}}}(o_i \mid \epsilon, a)} A_i, \text{clip}\Big(\frac{\pi_{critic}(o_i \mid \epsilon, a)}{\pi_{\theta_{\text{old}}}(o_i \mid \epsilon, a)}, 1-\varepsilon,\ 1+\varepsilon\Big) A_i\Big) - \beta\, D\big(\pi_{critic} \,\|\, \pi_{\text{ref}}\big), \quad (4)$$

where ε and β are the PPO clipping hyperparameters and the coefficient controlling the Kullback–Leibler (KL) penalty, respectively. $D\big(\pi_{critic} \,\|\, \pi_{\text{ref}}\big) = \frac{\pi_{\text{ref}}(o_i|\epsilon,a)}{\pi_{critic}(o_i|\epsilon,a)} - \log\Big(\frac{\pi_{\text{ref}}(o_i|\epsilon,a)}{\pi_{critic}(o_i|\epsilon,a)}\Big) - 1$ is the KL divergence. The relative advantage for the i-th response is computed by normalizing the rewards across the group: $A_i = \frac{r(o_i) - \text{mean}(\{r(o_1),\ldots,r(o_G)\})}{\text{std}(\{r(o_1),\ldots,r(o_G)\})}$.

4 Experiment

4.1 Experiment Settings

Dataset and Benchmark. ***(1) GUI-Critic-Train.*** The data in GUI-Critic-Train were sourced from publicly available GUI operation datasets, including AITZ [54], AMEX [3], Odyssey [18], and AITW [28]. We utilize Qwen2.5-VL-7B [2] to generate negative operations, which were simply deemed incorrect based on rule-based criteria. GPT-4o [13] and Qwen2.5-VL-72B [2] are utilized to generate the Chain of Thought (CoT) processes for $\mathcal{D}_{c_cot}$ outlined in Section 3.2.2. Consequently, GUI-Critic-Train dataset comprises approximately 11k entries, with 6k annotated by high-quality Chain-of-Thought (CoT) processes. ***(2) GUI-Critic-Test.*** To comprehensively evaluate the pre-critic capabilities of MLLMs, we established three main benchmark settings in GUI-Critic-Test: Mobile-Instruction Generalization (**GUI-I**), Mobile-Scenario Generalization (**GUI-S**), and Web-Scenario Generalization (**GUI-W**). Specifically, GUI-I test data are sourced from the AMEX [3], with instructions different from those in the training set. GUI-S test data are drawn from the Odyssey [18], comprising mobile applications that differ from those seen in the training set. Besides, shifting from mobile to web, GUI-Web contains the samples about the web automation, which are randomly sampled from the GUICourse [4]. Each of the settings, GUI-I, GUI-S, and GUI-Web, is manually annotated to ensure label accuracy, resulting in dataset sizes of 656, 114, and 418, respectively. We present critic accuracy and suggestion accuracy as the metrics. The former reflects the ability to assess the correctness of GUI operations, while the latter reflects the suggestion quality by quantifying the similarity between the generated suggestion and the annotation, which is calculated by prompting the Qwen2.5-VL-72B [2].

Implementation Details. We employ Qwen2.5-VL-7B [2] as the backbone. For the RFT cold-start, it is conducted for one epoch on mobile scenarios (GUI-I and GUI-S) and two epochs on web scenarios (GUI-W) via supervised fine-tuning (SFT). Next, we train the model on the $\mathcal{D}_{c_action}$ by S-GRPO for 10 epochs, with a learning rate of $3e^{-6}$ and a batch size of 128. The suggestion and format reward weights (λ_s, λ_f) are both set to 0.1, while the group size is set to 6. In accordance with the methodology outlined in [57], the KL divergence coefficient is set to $1e^{-2}$ by default. All experiments are conducted on eight NVIDIA A100 Tensor Core GPUs.

4.2 Comparision Results

To analyze the performance of our GUI-Critic-R1 comprehensively, we construct experiments from both static and dynamic aspects. Firstly, we evaluate the model's ability to determine operational correctness and suggest the corrective operation on our static GUI-Critic-Test benchmark. Additionally, a crucial application of GUI-Critic lies in its functionality as a pre-critic model within the dynamic GUI automation processes. To this end, we implement online evaluation experiments on AndroidWorld [27] to validate the effectiveness of our approach.

4.2.1 Static Evaluation

Table 1 illustrates the quantitative results of our GUI-Critic-R1 and a variety of baselines on GUI-Critic-Test benchmark. It showcases that GUI-Critic-R1 performs competitively on different scenarios compared to closed- and open-source MLMs in both critic accuracy and suggestion accuracy. For close source MLMs, Claude-3.5 achieves best performance on most settings. Despite the capabilities of close source MLMs, our GUI-Critic-R1 achieves state-of-the-art (SoTA) performance on the GUI-I test dataset. Compared to the base model, GUI-Critic-R1 achieved a significant improvement in critic accuracy from 54.88% to 69.05%. Compared to the excellent GPT-4o, our model also achieves a 3.19% critic accuracy improvement and a 11.89% advantage in suggestion accuracy. When faced with GUI-S, the novel application scenarios containing complex cross-application instructions from Odyssey [18], our model demonstrated commendable generalization capabilities as well. Although

Table 1: Static evaluation performance comparison of closed-source and open-source MLLMs on our GUI-Critic-Test, including three different settings: GUI-I, GUI-S, and GUI-W.

Model	**GUI-I**		**GUI-S**		**GUI-W**	
	Critic Accuracy	Suggestion Accuracy	Critic Accuracy	Suggestion Accuracy	Critic Accuracy	Suggestion Accuracy
Close Source MLMs						
Claude-3.5	67.26	40.71	64.27	46.11	65.55	37.64
GPT-4o	66.01	40.54	62.28	33.33	68.45	28.27
Gemini-2.0-Flash	66.76	42.98	64.91	38.59	62.85	38.78
Open Source MLMs						
Deepseek-VL2-7B [41]	44.36	0.00	43.85	0.00	10.28	0.00
InternVL2.5-8B [5]	53.96	19.35	52.63	14.91	54.67	24.53
InternVL2.5-8B-MPO [5]	51.06	20.42	51.75	19.30	55.37	24.06
Qwen2.0-VL-7B [37]	52.59	21.04	54.42	19.30	53.50	38.78
Qwen2.0-VL-72B [37]	54.27	30.03	53.51	29.80	51.86	21.96
Qwen2.5-VL-7B [2]	54.88	43.14	57.02	37.72	59.11	36.21
Qwen2.5-VL-72B [2]	56.40	49.08	**59.65**	38.79	60.05	38.79
Ours (GUI-Critic-R1)	**69.20**	**52.43**	58.77	**47.37**	**63.08**	**39.48**
Δ (Ours - Qwen2.5-VL-7B)	+14.32	+9.29	+1.75	+9.65	+3.97	+3.27

our model achieves a relatively insignificant improvement in critic accuracy (1.75%), it performs a 9.65% advantage in suggestion accuracy. It demonstrates that our model has a robust understanding of operations even in a new environment. In the web scenarios, where domain differences are more significant, our model exhibits commendable performance, indicating that it not only enhances the ability to determine the operation correctness but also generates effective suggestions. These results validate the superiority and robustness of our proposed S-GRPO.

4.2.2 Dynamic Evaluation

We further evaluate our model on the AndroidWorld [27] benchmark, which provides a live Android emulator and 116 tasks across 20 mobile apps. Specifically, in this platform, the GUI agent can perform operations on an emulated Android phone to attain human instructions, and the results are evaluated automatically. Our pre-critic model can serve as an error diagnosis module in the framework in a plug-and-play manner. For a fair comparison, we conduct both post-operative and pre-operative critic with existing MLLMs as the baseline. The results in Table 2 illustrate that GUI-Critic-R1 achieves the best success rate on the benchmark, verifying the ability of error correction and suggestion of our model. We observe that both post- and pre-critic can improve the performance of the GUI agent, indicating that pre-critic can avoid some potential mistakes, and post-critic can also remedy some accomplished errors. Despite the capabilities of GPT-4o, we find that it can sometimes produce inaccurate critiques and suggestions to mislead the agent when performing as the pre-critic, primarily due to its insufficient common-sense knowledge of the GUI. GPT-4o outperforms the prior-critic on the post-critic because the prior-critic requires the model to predict the potential outcomes of GUI operations, which GPT-4o lacks the common-sense reasoning ability. Furthermore, we evaluated the efficiency of GUI task execution across different approaches. We introduced a metric called Efficiency Advantage Rate (EAR) for fair comparison, which measures the proportion that the 'baseline + critic' model achieves an efficiency advantage (fewer steps) than the baseline, for tasks where they achieve consistent results. The results shown in Table 2 reveal that our model tends to complete tasks in fewer steps, thereby demonstrating its efficiency. In contrast, post-critic methods typically require a greater number of steps.

Table 2: Dynamic evaluation results on the AndroidWorld [27] benchmark.

Model	SR(↑)	EAR(↑)
Baseline	22.4	-
Post-Critic		
GPT-4o [13]	25.0	18.3
Pre-Critic		
Qwen2.5-vl-7B [2]	20.3	21.8
Qwen2.5-vl-72B [2]	23.2	24.4
GPT-4o [13]	22.4	26.1
Ours	27.6	31.8

Table 3: Ablation study on the dataset collection pipeline.

Model	GUI-I		GUI-S	
	Critic Accuracy	Suggestion Accuracy	Critic Accuracy	Suggestion Accuracy
w/o NOS	50.46	1.22	50.30	5.26
w/o DF	67.23	49.54	54.39	42.98
w/o GCG	67.84	51.22	56.14	42.11
Ours	**69.20**	**52.43**	**58.77**	**47.37**

Table 4: Ablation study on training strategies.

RFT	S-GRPO			GUI-I		GUI-S	
	r_f	r_a	r_s	Critic Accuracy	Suggestion Accuracy	Critic Accuracy	Suggestion Accuracy
✓	✗	✗	✗	63.16	45.61	55.26	34.21
✗	✓	✓	✗	67.98	43.44	54.38	39.47
✗	✓	✓	✓	69.05	49.24	56.14	42.10
✓	✓	✓	✗	66.01	47.71	57.89	40.35
✓	✓	✓	✓	**69.20**	**52.43**	**58.77**	**47.37**

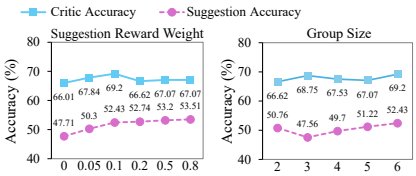


Figure 3: Analysis of suggestion reward weight λ_s and Group Size on the GUI-I setting.

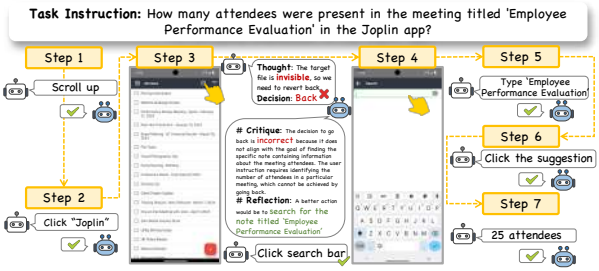


Figure 4: A case of pre-critic on GUI automation.

4.3 Ablation Study

Analysis of Dataset Collection Pipeline. We first conduct ablation studies for the proposed data collection pipeline to verify the contributions of three key components on GUI-I and GUI-S settings. Corresponding to Table 3, we first substitute Negative Operations Sampling (NOS) with a strategy of random decision replacements to acquire negative samples. It underscores that negative samples sampled by random substitution are too naive to undermine the model's capability to detect errors and suggest corrections. Secondly, we substitute the data filtering (DF) phase with a random sample selection approach. Without DF, the quality of training samples may be compromised, potentially degrading the effectiveness of the training process. Finally, we omit the GUI Critique Generation (GCG) process by utilizing only $\mathcal{D}_{c_action}$ in the RFT cold-start stage. The absence of GCG leads to a 1.36% decline in critic accuracy on GUI-I, highlighting its crucial contribution to the model's generalization and cognitive capabilities. These findings collectively emphasize the pivotal importance of our data collection process, establishing it as an integrated mechanism for boosting model training and operational efficiency.

Analysis of Training Strategy. In order to analyze the impact of training strategies on model performance, we conduct ablation experiments focusing on the RFT cold-start and reward components in the GRPO stage. As depicted in Table 4, the results in the first and third lines indicate that the utilization of RFT cold-start provided a robust initial boost to the model's decision-making capabilities, and S-GRPO further increases the ability. Towards the second line, we observe that employing GRPO alone was insufficient to activate the model's GUI-Critic abilities, as the model does not yet possess the basic GUI critic ability, making it difficult to produce high-quality outputs when sampling randomly. In the fourth line, we ablate the suggestion reward, which led to an obvious decrease in suggestion accuracy, indicating this reward is necessary for the model to suggest correct GUI operations. These observations underscore the crucial significance of the RFT cold-start and the proposed suggestion reward in S-GRPO, incrementally enhancing the capability of our GUI-Critic-R1.

Analysis of parameters. Figure 3 illustrates the impact of different suggestion reward weight λ_s or the group size in the R-GRPO phase. Note that when varying λ_s, the weights for r_a and r_f are maintained at a ratio of 4:1, ensuring that the total sum of weights remains equal to 1. By adjusting λ_s, we observe that excessively low values inadequately constrain the suggestion, while excessively high values may compromise accuracy. Consequently, a value of 0.1 was selected as an optimal parameter. Additionally, the size of the group plays a crucial role in balancing performance with resource utilization. Therefore, a compromise was made by selecting a group size of 6.

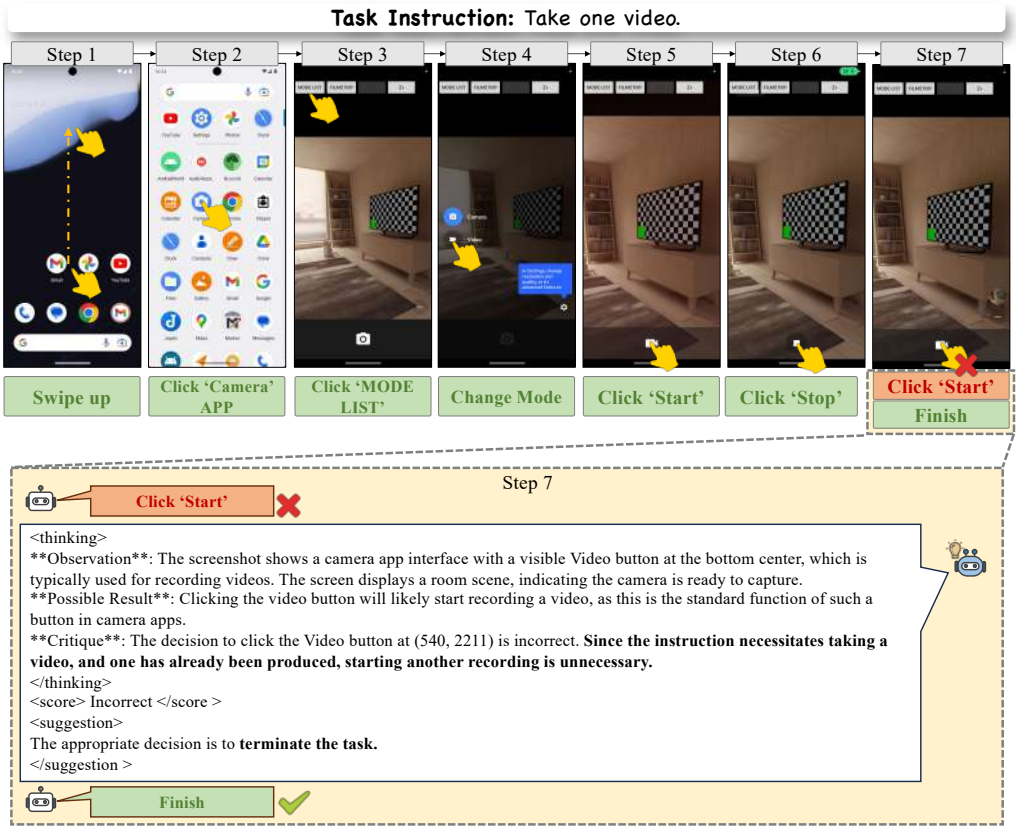


Figure 5: An instance of GUI-Critic-R1 avoiding a redundant action within GUI automation through a pre-operative critique approach.

4.4 Case Study

Figure 4 illustrates an example drawn from the AndroidWorld benchmark [27], demonstrating how our model guides the GUI agent to achieve a correct and effective trajectory. Specifically, the agent is tasked with finding a file in the Joplin app but encounters an interface without the target file visible, prompting it to mistakenly consider rolling back. Our model suggests clicking the search box to locate the target file, thereby assisting the agent in successfully completing the task.

In Figure 5, we depict another example of pre-operative critic for GUI automation. The agent successfully navigates steps 1 through 6, initiating the camera app and completing a video recording. Nonetheless, at step 7, the agent erroneously decides to press the record button again. This action is incorrect, as the instruction clearly stipulates recording only one video. Given that the agent has fulfilled this requirement, the task should be stopped rather than continuing the recording. Our model precisely delineates the state of the current interface and, based on historical operation, identifies the agent's decision as redundant. Consequently, our model advises the agent to terminate the task at this step.

5 Conclusion

In this paper, we introduce a pre-operative critic mechanism aimed at performing error diagnosis in GUI Automation, prior to executing actions within interactive environments. To develop our pre-critic model GUI-Critic-R1, we propose a Suggestion-aware Gradient Policy Optimization (S-GRPO) strategy with a novel suggestion reward. It is designed to ensure effective corrective discrimination for operations and provides reliable feedback for improvement. Futhermore, We develop a reasoning-bootstrapping based data collection pipeline to construct a GUI-Critic-Train and GUI-Critic-Test dataset, addressing existing gaps in GUI critic data. Through extensive static and dynamic experiments, we demonstrate the the efficacy of our approach across diverse GUI scenarios, highlighting its potential to significantly enhance the success rate of automated tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants 62322212, U23A20387), the Beijing Natural Science Foundation (No. L221013), and the CAS Project for Young Scientists in Basic Research (YSBR-116).

References

[1] Saaket Agashe, Kyle Wong, Vincent Tu, Jiachen Yang, Ang Li, and Xin Eric Wang. Agent s2: A compositional generalist-specialist framework for computer use agents. *arXiv preprint arXiv:2504.00906*, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Yuxiang Chai, Siyuan Huang, Yazhe Niu, Han Xiao, Liang Liu, Dingyu Zhang, Peng Gao, Shuai Ren, and Hongsheng Li. Amex: Android multi-annotation expo dataset for mobile gui agents. *arXiv preprint arXiv:2407.17490*, 2024.

[4] Wentong Chen, Junbo Cui, Jinyi Hu, Yujia Qin, Junjie Fang, Yue Zhao, Chongyi Wang, Jun Liu, Guirong Chen, Yupeng Huo, et al. Guicourse: From general vision language models to versatile gui agents. *arXiv preprint arXiv:2406.11317*, 2024.

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[6] Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. Re-rest: Reflection-reinforced self-training for language agents. *arXiv preprint arXiv:2406.01495*, 2024.

[7] Dayuan Fu, Keqing He, Yejie Wang, Wentao Hong, Zhuoma Gongque, Weihao Zeng, Wei Wang, Jingang Wang, Xunliang Cai, and Weiran Xu. Agentrefine: Enhancing agent generalization through refinement tuning. *arXiv preprint arXiv:2501.01702*, 2025.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[9] Priyanshu Gupta, Shashank Kirtania, Ananya Singha, Sumit Gulwani, Arjun Radhakrishna, Sherry Shi, and Gustavo Soares. Metareflection: Learning instructions for language agents using past reflections. *arXiv preprint arXiv:2405.13009*, 2024.

[10] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024.

[11] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

[12] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[14] Noman Islam, Zeeshan Islam, and Nazia Noor. A survey on optical character recognition system. *arXiv preprint arXiv:1710.05703*, 2017.

[15] Yanda Li, Chi Zhang, Wanqi Yang, Bin Fu, Pei Cheng, Xin Chen, Ling Chen, and Yunchao Wei. Appagent v2: Advanced agent for flexible mobile interactions. *arXiv preprint arXiv:2408.11824*, 2024.

[16] Haowei Liu, Xi Zhang, Haiyang Xu, Yuyang Wanyan, Junyang Wang, Ming Yan, Ji Zhang, Chunfeng Yuan, Changsheng Xu, Weiming Hu, et al. Pc-agent: A hierarchical multi-agent collaboration framework for complex task automation on pc. *arXiv preprint arXiv:2502.14282*, 2025.

[17] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[18] Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv preprint arXiv:2406.08451*, 2024.

[19] Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. Robust visual question answering: Datasets, methods, and future challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[20] Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. Coco-agent: A comprehensive cognitive mllm agent for smartphone gui automation. *arXiv preprint arXiv:2402.11941*, 2024.

[21] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[22] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024.

[23] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[24] Jiayi Pan, Yichi Zhang, Nicholas Tomlin, Yifei Zhou, Sergey Levine, and Alane Suhr. Autonomous evaluation and refinement of digital agents. *arXiv preprint arXiv:2404.06474*, 2024.

[25] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[26] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

[27] Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.

[28] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Androidinthewild: A large-scale dataset for android device control. *Advances in Neural Information Processing Systems*, 36:59708–59728, 2023.

[29] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[30] Yunpeng Song, Yiheng Bian, Yongtao Tang, Guiyu Ma, and Zhongmin Cai. Visiontasker: Mobile task automation using vision based ui understanding and llm task planning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–17, 2024.

[31] Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. Adaplanner: Adaptive planning from feedback with language models. *Advances in neural information processing systems*, 36:58202–58245, 2023.

[32] Fei Tang, Haolei Xu, Hang Zhang, Siqi Chen, Xingyu Wu, Yongliang Shen, Wenqi Zhang, Guiyang Hou, Zeqi Tan, Yuchen Yan, et al. A survey on (m) llm-based gui agents. *arXiv preprint arXiv:2504.13865*, 2025.

[33] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[34] Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. *arXiv preprint arXiv:2311.08516*, 2023.

[35] Haoyu Wang, Tao Li, Zhiwei Deng, Dan Roth, and Yang Li. Devil's advocate: Anticipatory reflection for llm agents. *arXiv preprint arXiv:2405.16334*, 2024.

[36] Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *arXiv preprint arXiv:2406.01014*, 2024.

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[38] Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733*, 2025.

[39] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. Autodroid: Llm-powered task automation in android. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 543–557, 2024.

[40] Hao Wen, Hongming Wang, Jiaxuan Liu, and Yuanchun Li. Droidbot-gpt: Gpt-powered ui automation for android. *arXiv preprint arXiv:2304.07061*, 2023.

[41] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.

[42] Yufei Xiang, Yiqun Shen, Yeqin Zhang, and Nguyen Cam-Tu. Retrospex: Language agent meets offline reinforcement learning critic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4650–4666, 2024.

[43] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024.

[44] Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction. *arXiv preprint arXiv:2412.04454*, 2024.

[45] An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, et al. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. *arXiv preprint arXiv:2311.07562*, 2023.

[46] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[47] Siyu Yuan, Zehui Chen, Zhiheng Xi, Junjie Ye, Zhengyin Du, and Jiecao Chen. Agent-r: Training language model agents to reflect via iterative self-training. *arXiv preprint arXiv:2501.11425*, 2025.

[48] Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Guyue Liu, Qingwei Lin, et al. Large language model-brained gui agents: A survey. *arXiv preprint arXiv:2411.18279*, 2024.

[49] Chaoyun Zhang, Liqun Li, Shilin He, Xu Zhang, Bo Qiao, Si Qin, Minghua Ma, Yu Kang, Qingwei Lin, Saravan Rajmohan, et al. Ufo: A ui-focused agent for windows os interaction. *arXiv preprint arXiv:2402.07939*, 2024.

[50] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023.

[51] Di Zhang, Junxian Li, Jingdi Lei, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. *arXiv preprint arXiv:2411.18203*, 2024.

[52] Jiaqi Zhang, Chen Gao, Liyuan Zhang, Yong Li, and Hongzhi Yin. Smartagent: Chain-of-user-thought for embodied personalized agent in cyber world. *arXiv preprint arXiv:2412.07472*, 2024.

[53] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

[54] Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*, 2024.

[55] Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. Agent-pro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint arXiv:2402.17574*, 2024.

[56] Jiani Zheng, Lu Wang, Fangkai Yang, Chaoyun Zhang, Lingrui Mei, Wenjie Yin, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Vem: Environment-free exploration for training gui agent with value environment model. *arXiv preprint arXiv:2502.18906*, 2025.

[57] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

Appendix

A Limitations and Future Works

In this paper, we explore the pre-operative critic mechanism for GUI automation to enhance success rates and decrease the number of operational steps. Future endeavors have the potential to extend our method to lighter models, such as Qwen2.5-VL-3B, to achieve more efficient and improved critic performance. Furthermore, our algorithm is based on single-step GUI visual information and semantic operation history.

Our approach can be enhanced by integrating the trajectory-level critic in future studies, which involves taking a sequence of screenshots as input. It has the potential to provide more comprehensive insights.

B More Dynamic Evaluation

In this paper, we conduct the dynamic evaluation on AndroidWorld [27] and illustrate the experiment results in the main text (Section 4.2.2). In this section, we introduce more details of the implementation of the experiments and more experimental results.

B.1 Supplementary Implementation Detials

We adopt the M3A (Multimodal Autonomous Agent for Android) framework introduced in AndroidWorld [27] as the backbone. In the M3A framework, the decision agent is provided with a list of available action types, guidelines for operating the phone, and a list of UI elements derived from the leaf nodes of the Android accessibility tree. The agent receives the current screenshot along with a Set-of-Mark annotated screenshot, which includes bounding boxes with numeric labels at the top-left corner of each UI element. At this stage, the agent attempts to execute the generated action, referencing specific marks. Prior to action execution, we insert a pre-operative critic to evaluate whether the agent's proposed action is conducive to achieving the instruction. If the action is deemed correct, it proceeds to execution as normal. Conversely, if the action is considered incorrect, we input the critic model's analytical recommendations to the decision agent, prompting it to reassess and formulate a new decision. After action execution, another agent is adopted to deliver a concise summary following the execution of the action. We leverage such semantic summaries to serve as a record of the action history. We follow the Androidworld to configure the UI element detection procedure and the definition of the action space.

B.2 Supplementary Experiments

We extend the dynamic evaluation experiment to incorporate GPT-4o [13] as the decision agent. As evidenced in Table 5, employing GPT-4o [13] as the decision-making agent yields performance benefits over Qwen2.5-VL-72B [2], underscoring GPT-4o's robust perception and reasoning capabilities. The pre-operative critic improves accuracy and exemplifies the efficacy of the baseline in detecting erroneous decisions and offering constructive improvement suggestions. In comparison with other pre-operative models, our GUI-Critic-R1 exhibits superior performance outcomes. When utilizing GPT-4o as the baseline, the advantage of the pre-critic in terms of the Efficient Advantage Rate (EAR) becomes more pronounced. GPT-4o tends to engage in excessive exploration during GUI automation, often opting to attempt further actions rather than terminating early, thereby resulting in a greater number of steps. In contrast, the pre-critic effectively anticipates erroneous or redundant exploration, thereby reducing unnecessary steps and manifesting a superior EAR metric.

Table 5: Supplementary dynamic evaluation results on the AndroidWorld [27] benchmark with GPT-4o [13] as the baseline.

Model	SR(↑)	EAR(↑)
Baselines		
Qwen2.5-VL-72B [2]	22.4	-
GPT-4o [13]	23.9	-
GPT4o+Post-Critic		
GPT-4o [13]	26.4	31.6
GPT4o+Pre-Critic		
Qwen2.5-VL-7B [2]	16.9	32.4
Qwen2.5-VL-72B [2]	25.9	36.8
GPT-4o [13]	24.0	43.1
Ours	**29.4**	**46.2**

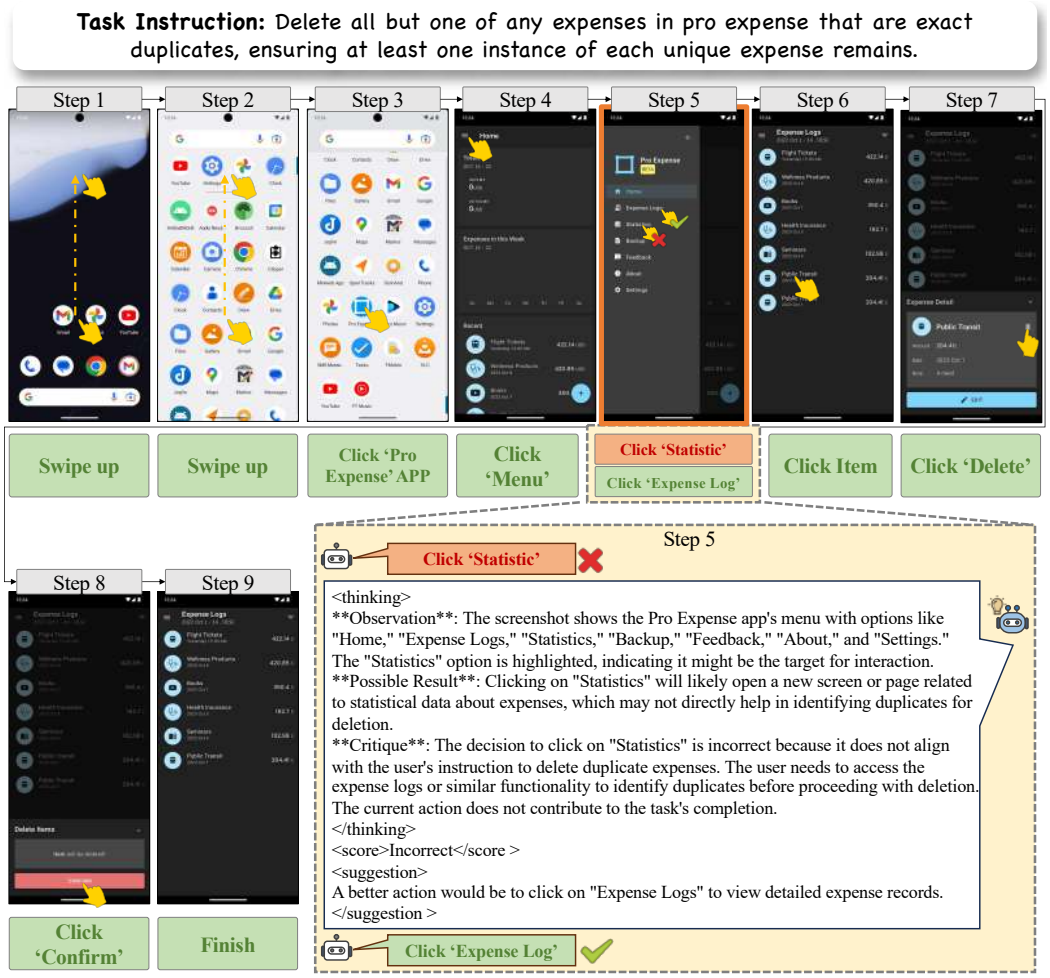


Figure 6: An instance of GUI-Critic-R1 correcting an erroneous decision within GUI automation through a pre-operative critique approach.

B.3 Additional Case Study

Figure 6 illustrates an operational process of GUI automation with pre-operative critic by our GUI-Critic-R1. In this example, the GUI agent initially demonstrates correct behavior by opening the Pro Expense app and accessing the menu. However, at step 5, the agent erroneously decides to click the "statistics" button, which is incorrect. Our GUI-Critic-R1 model successfully identifies this as an incorrect action and analyzes the reason behind the error. The model determines that selecting this button would navigate to a new interface displaying expense statistics, which is irrelevant to the instruction requiring the removal of duplicate expenses. Furthermore, our model provides a suggestion for correction: click on "Expense Logs" to view detailed expense records.

C GUI-Critic Dataset

To construct the GUI-Critic-Train and GUI-Critic-Test datasets, we extract step-level data from publicly available GUI operation datasets, encompassing user instructions, current interface screenshots, text-based operative history, and corresponding correct actions. In the following section, we provide an overview of the datasets employed in our study.

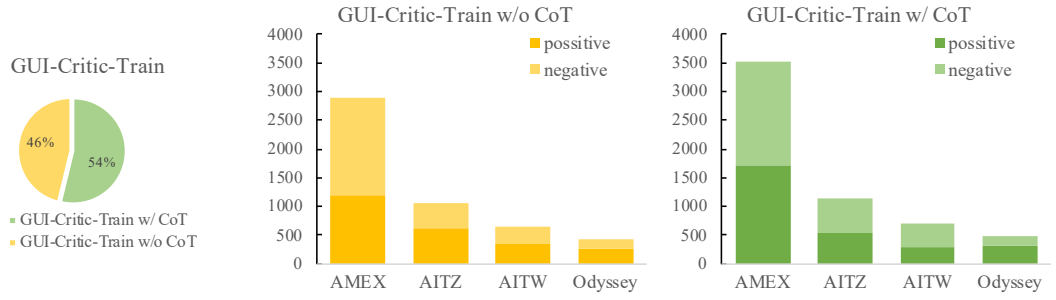


Figure 7: Illustration of the data composition for the GUI-Critic-Train dataset. The left displays the proportion of data without Chain-of-Thought (CoT) annotations versus with CoT. The central part shows the sources of data without CoT annotations, along with the ratio of correct to incorrect actions. The right section depicts the sources of data with CoT annotations, also detailing the ratio of correct to incorrect actions.

C.1 Public Datasets

Firstly, we review the publicly available datasets employed in our research.

Android in the Wild (AITW) [28] dataset comprises human demonstrations of device interactions, encompassing both screen captures and actions, alongside corresponding natural language instructions. It includes a substantial collection of 715,000 episodes covering 30,000 unique instructions, across four versions of the Android operating system (v10–13) and eight different device models ranging from the Pixel 2 XL to Pixel 6, each with distinct screen resolutions. The dataset involves complex multi-step tasks necessitating nuanced semantic understanding of both language and visual context.

Android-In-The-Zoo (AITZ) [54] dataset contains 18,643 screen-action pairs together with chain-of-action-thought annotations, spanning over 70 Android apps, coupled with 4× useful annotations compared with action coordinate labels only. Based on the screen episodes from AITW [28], the authors generate candidate answers for the screen descriptions, action thinkings and next action descriptions. These candidates are further validated and refined by human to guarantee alignment with the screenshots.

GUI Odyssey [18] is a comprehensive dataset for training and evaluating cross-app navigation agents. GUI Odyssey comprises 7735 episodes, meticulously curated from 6 different mobile devices such as Pixel Pro and Tablet. It encompasses 6 types of cross-app navigation tasks spanning from general system tool use to media entertainment, requiring navigating through 201 different apps and 1399 app combos from various fields such as video, music, reading

Android Multi-annotation EXpo (AMEX) [3] is a comprehensive, large-scale dataset designed for generalist mobile GUI-control agents. Their capabilities of completing complex tasks by directly interacting with the graphical user interface (GUI) on mobile devices are trained and evaluated with the proposed dataset. AMEX comprises over 104K high-resolution screenshots from 110 popular mobile applications, which are annotated at multiple levels. AMEX includes three levels of annotations: GUI interactive element grounding, GUI screen and element functionality descriptions, and complex natural language instructions, each averaging 13 steps with stepwise GUI-action chains.

GUICourse [4]. In this paper, the author constructed GUIAct, a GUI navigation dataset in website and android scenarios for enhancing VLMs' knowledge of GUI systems, including 67k single-step and 15k multi-step action instructions.

C.2 GUI-Critic Dataset Components

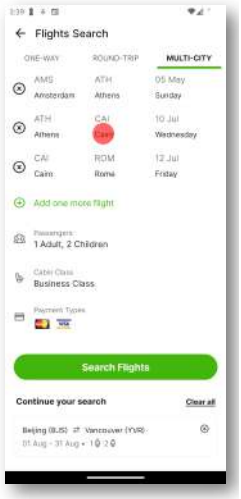
In this subsection, we delineate the origins, constituent parts, and foundational principles of the construction of our GUI-Critic-Train and GUI-Critic-Test datasets.

Instruction: Open Wego. Find multi-city flights from Amsterdam to Athens on May 5, from Athens to Cairo on May 8 and from Cairo back to Amsterdam on May 12. 1 adult. First class. Sort by price and filter only book on Wego. The price must be under US$3000. set as May 12.

History: Step 1: Tap on the 'Wego Flights' app icon to open it. Step 2: Tap on the back arrow to return to the main flight search screen. Step 3: Tap on the 'MULTI-CITY' option to change the search type. Step 4: Tap on the first city to change it to Amsterdam. Step 5: Type 'Amsterdam' into the search field. Step 6: Tap on the 'Amsterdam' option to select it. Step 7: Tap on 'GVA' to change it to 'Athens.' Step 8: Type 'Athens' in the search bar to set it as the arrival city. Step 9: Tap on the first option, 'Athens,' to select it as the destination city. Step 10: Tap on the destination city of the second segment and change Madrid to Cairo. Step 11: Swipe up to navigate to the month of May and locate May 8 on the calendar. Step 12: Tap on May 8, 2024, to select it as the departure date from Athens to Cairo. Step 13: Tap to change the destination from Madrid to Cairo in the second segment. Step 14: Select 'Cairo, Egypt' from the list of popular cities. Step 15: Select 'Cairo, Egypt' as the destination.

Action: {"action": "click", "coordinate": [639, 836]}

Screenshot marked with the candidate action.



Generated CoT process:

<thinking>
Observation: The screenshot is of the "Flights Search" interface within the Wego app. The search type is set to "MULTI-CITY." The current itinerary includes the following segments:
1. Amsterdam (AMS) to Athens (ATH), departing May 5, Sunday.
2. Athens (ATH) to Cairo (CAI), departing May 8, Wednesday.
3. Cairo (CAI) to Rome (ROM), departing July 12, Friday.
The red translucent marker is highlighting the city "Cairo" in the second segment of the itinerary. Below are options to add passengers, change cabin class (currently set to "Business Class"), and payment types.
The history so far involves setting the multi-city search and filling in destinations (Amsterdam, Athens, Cairo) for the first two segments. Based on the user instruction, the third flight segment should have Cairo as the departure city and Amsterdam as the destination, with the departure date set to May 12.
Possible Result: If the agent clicks on "Cairo" as marked, it will likely allow editing Cairo as the destination city for the second flight segment. However, this action is redundant since the city "Cairo" is already correctly assigned as the destination for the second segment. The focus should now shift to modifying the third flight segment to Cairo → Amsterdam with the departure date set as May 12.
This click neither promotes progress toward completing the intended user instruction nor aligns with the detailed requirements of modifying the third segment.
Critique: This decision is incorrect. Based on the history, Cairo is already correctly set as the destination for the second segment. The agent should now focus on correcting the third flight segment, currently set as Cairo → Rome on July 12, to reflect Cairo → Amsterdam on May 12 (per user instructions). Clicking "Cairo" in the second segment achieves no necessary action for advancing the core task.
</thinking>
<score> Incorrect </score >
<suggestion>
Change the third segment of the itinerary: Click on "ROM" and type "Amsterdam."
</suggestion>

Figure 8: A data sample in GUI-Critic-Train dataset.

C.2.1 GUI-Critic-Train

Figure 7 provides a visualization of the data sources utilized in our training set. For the construction of our training dataset, we predominantly selected data from AMEX[3] due to its extensive repository of high-quality GUI operation data. The raw data encompasses decision-making processes enriched with reasoning, which contributes to the reliability of the annotated correct actions. Detailed methodologies for the generation of incorrect decisions and the formulation of Chain-of-Thought (CoT) critiques are elaborated in the main text (Section 3.2). We preserved a higher proportion of samples containing incorrect decisions to facilitate the model's acquisition of enhanced error correction competencies. Figure 8 illustrates a data sample with the generated CoT process in GUI-Critic-Train dataset.

C.2.2 GUI-Critic-Test

Mobile-Instruction Generalization (GUI-I). Data in GUI-I originates from the AMEX dataset [3]. We ensure that the instructions in GUI-I are distinct from those in GUI-Critic-Train. After human annotation, GUI-I retains 656 GUI screenshots, each accompanied by a user instruction, operational history, and a set of candidate actions.

Mobile-Scenario Generalization (GUI-S). We collect 114 test data from the Odyssey dataset [18] for GUI-I. We select the data with applications different from those in GUI-Critic-Train. Specifically, the novel applications are:

1. *TikTok*: A popular social media application for sharing and watching short-form videos, fostering a global community of creators and viewers.
2. *Chaton*: An innovative application powered by intelligent conversation models, designed to enhance user interactions through AI-driven dialogues.

3. *ClevCalc*: A multifunctional calculator tool app that offers a wide range of mathematical operations and unit conversions for everyday use.
4. *Remix*: A versatile audio editing application that allows users to mix, cut, and enhance soundtracks with a variety of editing features.
5. *Redfin Houses*: A comprehensive app for real estate transactions, providing listings, pricing, and market insights for prospective home buyers and sellers.
6. *Tripadvisor*: A travel review platform app where users can rate and explore travel destinations, accommodations, and restaurants worldwide.

Web-Scenario Generalization (GUI-W). It contains 418 samples of web automation, which are randomly sampled from the GUICourse [4]. Although the action space in web environments differs from mobile platforms (e.g., limitations on swipe direction and the incorporation of double-tapping), the underlying operational logic largely remains consistent across both scenarios.

D Agent Prompt

In Table 6, Table 7, Table 8, and Table 9, we present the prompts for pre-critic, post-critic, action similarity evaluation, and adding critiques to the GUI agent decision process. Note that, we conduct minimal modifications to the pre-critic prompt, resulting in the post-critic prompt. These modifications include the removal of predictions concerning possible results and the adjustment of requirements from generating corrective suggestions to remedial suggestions in the `<suggestion>...</suggestion>`. The action similarity evaluation prompt is leveraged in calculating the suggestion reward in S-GRPO and the suggestion accuracy.

System
There is a multimodal agent that can perform a series of actions on a smart device (phone or PC) to automate the completion of user instructions. Possible actions include "click" / "left_click" at (x,y) position, "long press" at (x,y) position, "swipe" from (x1,y1) to (x2,y2), "scroll" down or up, "drag" from (x1,y1) to (x2,y2), "type" (text content), "back", "home", "enter" and so on. User instructions are usually complex and may include several detailed requirements. In some steps, the action decided by the mobile agent may be wrong. Now, you are a critic model used to evaluate the agent's decision. I will provide you with the following information: 1. User instruction. 2. History: The action history of the agent in the previous steps. 3. Decision: The decision of the agent for this step. 4. Image: The screenshot before executing this action. If the action contains positional parameters (such as click and swipe), the interaction area is marked with a translucent red circle or red arrow. Firstly, you need to understand the purpose of the decision. Pay attention to analyzing the interface elements in the screenshot (such as button position, text content, etc.). If there are red marks, focus on the action position. You can take appropriate account of the history information. Then, based on the given information, carefully analyze the decision given by the agent for the current step: 1. Decision Analysis (1). Observation: Observe the screenshot and analyze the state without considering the user's instruction. - Focus on the operable or informative elements related to the operational decision. (2). Possible Result: Speculate the most possible result of executing this decision. - Predicts the screenshot change after the operation. - Whether to promote the progress of core tasks. (3). Critique: Determine whether the decision is correct and explain why. - Focus on historical operations. - Based on the previous analysis and the history, determine if this decision supports the completion of the instruction. - Only perform actions specified in the instructions. - Home button is the correct choice for switching apps. - Both clicking a suggestion and Enter are correct when searching. 2. Based on the above analysis, determine whether this decision is "Correct" or "Incorrect". 3. Reflection: If correct, retell the action; if incorrect, suggest a better action. Propose a one-step action for the current observation, like click, swipe (with direction), type (with information), Home, Back, or Terminate (in 20 words). Assess the current decision's correctness in the following format: <thinking> **Observation**: Describe the screenshot. **Possible Result**: Analysis from the possible result perspective. **Critique**: Criticize why the decision is correct or incorrect. </thinking> <score> Correct or Incorrect </score> <suggestion> If correct, provide a brief summary; if incorrect, suggest a better decision briefly. </suggestion>
User
Below is the information for the current step: 1. User instruction: {User's instruction} 2. History: {Operation History} 3. Decision: {Action} 4. Image is the screenshot of this step. <image>

Table 6: The prompt for the pre-operative critic.

System
There is a multimodal agent that can perform a series of actions on a smart device (phone or PC) to automate the completion of user instructions. Possible actions include "click" / "left_click" at (x,y) position, "long press" at (x,y) position, "swipe" from (x1,y1) to (x2,y2), "scroll" down or up, "drag" from (x1,y1) to (x2,y2), "type" (text content), "back", "home", "enter" and so on. User instructions are usually complex and may include several detailed requirements. In some steps, the action decided by the mobile agent may be wrong. Now, you are a critic model used to evaluate the agent's decision. I will provide you with the following information: 1. User instruction. 2. History: The action history of the agent in the previous steps. 3. Decision: The decision of the agent for this step. 4. Images: The screenshots before and after executing this action. Firstly, you need to understand the purpose of the decision. Pay attention to analyzing the interface elements in the screenshot (such as button position, text content, etc.). If there are red marks, focus on the action position. You can take appropriate account of the history information. Then, based on the given information, carefully analyze the decision given by the agent for the current step: 1. Decision Analysis (1). Observation: Observe the screenshots and analyze the state without considering the user's instruction. - Focus on the operable or informative elements related to the operational decision. (2). Critique: Determine whether the decision is correct and explain why. - Focus on historical operations. - Ensure compliance with each specific requirement in the instructions. - Only perform actions specified in the instructions. - Home button is the correct choice for switching apps. - Both clicking a suggestion and Enter are correct when searching. 2. Based on the above analysis, determine whether this decision is "Correct" or "Incorrect". 3. Reflection: If correct, retell the action; if incorrect, suggest a remedial action on the screen after executing this action. Propose a one-step action for the current observation, like click, swipe (with direction), type (with information), Home, Back, or Terminate (in 20 words). Assess the current decision's correctness in the following format: <thinking> **Observation**: Describe the screenshots. **Critique**: Criticize why the decision is correct or incorrect. </thinking> <score> Correct or Incorrect </score> <suggestion> If correct, provide a brief summary; if incorrect, suggest a remedial decision briefly. </suggestion>
User
Below is the information for the current step: 1. User instruction: {User's instruction} 2. History: {Operation History} 3. Decision: {Action} 4. Images are screenshots before and after executing this action.

Table 7: The prompt for the post-operative critic.

The following two sentences describe the operation that a mobile agent needs to perform on mobile in order to complete a certain user instruction: 1. {Annotated Suggestion} 2. {Generated Suggestion} Determine whether these two sentences describe a similar action? If yes, answer 1, if not 0, no explanation required.

Table 8: The prompt for evaluating the similarity between two action-descriptive sentences.

Note that your last decision may be incorrect! Please make a remedial decision based on the following Critiques, and consider adhering to the Suggestion provided: {Critique}

Table 9: The prompt for integrating critique into the GUI decision-making process.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名： 王梦斐
方向： 人工智能
单位： 中国科学院上海微系统与信息技术研究所

题目： 基于交叉频域特征的光流估计

CFFlow: An Optical Flow Estimation Hinging on Cross-Frequency Attention

Wang Mengfei[1,2]
https://orcid.org/0009-0005-5586-4820
Zhu Dongchen[1,2*]
https://orcid.org/0000-0002-1579-3942
Wang Lei[1,2]
https://orcid.org/0009-0001-6655-188X
Li Jiamao[1,2]
https://orcid.org/0000-0002-7478-4544

[1] Bionic Vision System Laboratory, Science and Technology on Micro-system Laboratory, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China.

[2] University of Chinese Academy of Sciences, Beijing 100049, China.

Abstract

Due to issues in training datasets and motion estimation in optical flow, existing optical flow models overly emphasize surface-level information features, leading to inaccurate optical flow estimations. Through theoretical analysis, we reveal that multi-frame approaches require more attention to interconnections among diverse information than two-frame methods. Consequently, we propose **CFFlow**, which integrates cross-frequency information into optical flow estimation networks. To address challenges posed by large displacements and small object motion in real-world scenarios, we introduce two specialized modules: Large Displacement Attention (LDA) and Small Object Displacement Attention (SODA), designed to handle distinct motion patterns. CFFlow effectively resolves small target mismatches and large displacement challenges, achieving superior optical flow estimation accuracy. On Sintel and KITTI benchmarks, our CFFlow attains an Average Endpoint Error (AEPE) of **0.99** (clean pass) and **1.65** (final pass), with an F1-all error of **4.21%**, ranking first among all three-frame and two-frame methods.

1 Introduction

Optical flow estimation is a fundamental computer vision task that estimates pixel-level correspondence between consecutive frames. As a crucial paradigm of utilizing video temporal continuity, it is widely used in a variety of downstream video-related tasks, such as video interpolation [15], motion recognition [46], object detection [10], video in-painting [42], video understanding and analysis [40].

With the advancement of neural network architectures, many optical flow methods have emerged [4, 25, 28]. However, existing frameworks face two key limitations: (1) **Limited utilization of existing datasets.** Many methods [1, 21, 38] primarily rely on appearance features and visual similarity to match, often overlooking the underlying motion structure within scenes. This limits robustness under complex motion conditions. Although some

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
* Indicates corresponding author.

(a) Input Image (b) Image High-Frequency Feature (c) Image Low-Frequency Feature

Figure 1: **(a)** Input image from the KITTI dataset. **(b)** High-frequency feature preserves fine motion details and highlights fast-moving small objects. **(c)** Low-frequency feature retains global motion structure and captures large displacements more effectively.

work has explored the integration of segmentation information to aid in optical flow estimation [36, 45], our paper focuses on improving optical flow estimation efficiently by leveraging frequency-domain features.

(2) **Overreliance on surface information while overlooking the potential of cross-frequency features.** Many existing methods emphasize pre-trained image features [7, 16], yet optical flow estimation often benefits more from abundant low-level cues [43, 44]. Although frequency-aware features have been applied in dense image prediction and object detection tasks [3, 32], their potential remains underexplored in optical flow. Surface features primarily capture basic visual patterns, whereas cross-frequency features encode rich multi-scale structural information, offering greater promise for motion estimation. This paper aims to bridge that gap by fully leveraging cross-frequency features in flow estimation.

In this paper, we focus on leveraging cross-frequency features to enhance optical flow estimation. Existing methods often ignore multi-band frequency information. **We define cross-frequency features as the integration of motion patterns across frequency bands and gradient variations at image edges**, capturing rich multi-scale motion and structural details. These features enhance localized motion and edge-aware flow estimation, addressing limitations of prior approaches.

To enrich cross-frequency features beyond the limitations of two-frame designs, we adopt a multi-frame framework. We decomposed input images into low-frequency (smooth structural) and high-frequency (detailed) components. As illustrated in Figure 1, low-frequency features help estimate global large displacements, while high-frequency components capture fine details, improving small-object motion estimation and reducing mismatches.

To exploit the distinct roles of cross-frequency components, we design specialized modules for each. The Image Frequency Separation (IFS) Module first decomposes inputs into low- and high-frequency components. The Large Displacement Attention (LDA) Module captures long-range motion in the low-frequency domain, while the Small Object Displacement Attention (SODA) Module focuses on fast-moving small objects in the high-frequency domain. Both modules extract motion cues in their respective domains, contributing jointly to the 4D cost volume.

In summary, the contributions of this paper are as follows:

- We define cross-frequency information in optical flow tasks for the first time and design the IFS module to analyze these components, offering a complementary perspective to traditional spatial surface information.

- To fully exploit low- and high-frequency components, we introduce the LDA and SODA modules, targeting large displacements and small object motions, respectively, enhancing motion estimation across scales.

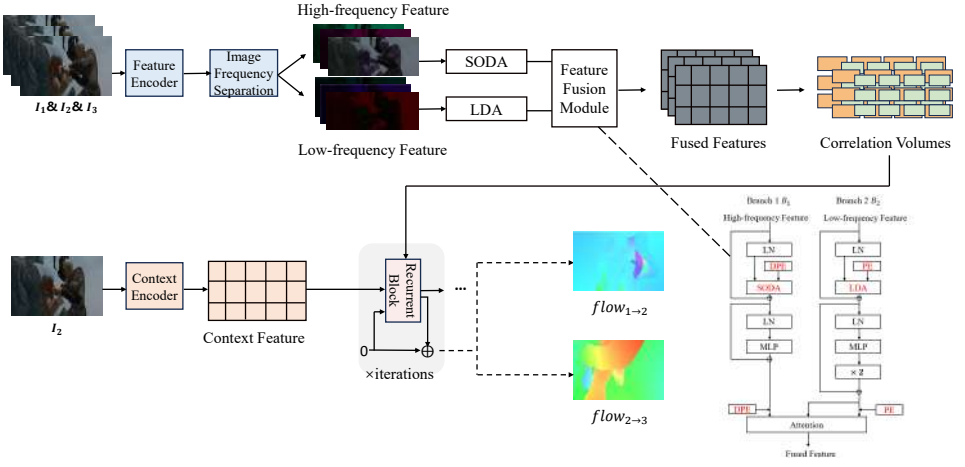


Figure 2: Overview of our LLFlow training pipeline. The IFS module decomposes image features into high- and low-frequency components. Then, SODA models small-object motion from high-frequency features, and LDA captures large displacements from low-frequency features. Finally, the FF module fuses both to produce refined motion estimation.

- Our approach achieves state-of-the-art results on Sintel and KITTI-2015 benchmarks, with AEPE of **0.989** and **1.647** on the clean and final passes, and an F1-all error of **4.21**%, ranking 1st among all three-frame methods.

2 Related Work

2.1 Optical Flow Estimation

Optical flow has long been a core vision task. Classical methods like Lucas–Kanade [22] and Horn–Schunck [11] formulated it as an energy optimization problem with handcrafted priors, but struggled with complex real-world motion. With deep learning and synthetic datasets, modern methods now learn optical flow end-to-end, shifting focus from manual priors to model and data design. PWCFusion [24, 29] fuses frame information via GRU-RCN, while MFCFlow [4] enhances motion features using correlation-based compensation. VideoFlow [25] leverages temporal cues for multi-frame bidirectional estimation, and SplatFlow [35] introduces splatting for differentiable motion alignment. Despite these advances, existing multi-frame approaches often emphasize high-level inter-frame features while overlooking cross-frequency motion correlations.

2.2 Cross-frequency Features

Wavelet transform decomposes signals into low-frequency Approximation Coefficients (AC) for global trends and high-frequency Detail Coefficients (DC) for local variations [19, 34, 41]. Similarly, audio processing integrates low-frequency linguistic structures with high-frequency acoustic cues [18], yet many systems rely on single-band features, ignoring their interplay. This issue extends to optical flow, where high-frequency components capture fine

motion (e.g., edges), and low-frequency ones encode global structure. Pretrained models often overemphasize low-frequency semantics [12], missing fine-grained details. To address this, we propose a frequency-aware framework that fuses high-frequency precision with low-frequency consistency for more accurate optical flow.

3 Method

3.1 Theoretical Analysis

Multi-frame optical flow methods aim to match $N : (I_1, I_2, I_3) \longmapsto (f_1, f_2)$, where I_1, I_2, and I_3 are consecutive frames, and (f_1, f_2) are the optical flows for (I_1, I_2) and (I_2, I_3). A multi-frame optical flow network can be formulated as:

$$(f_1, f_2)^* = N_\theta(I_1, I_2.I_3) = \arg\max_{(f_1, f_2)} p(f_1, f_2 \mid I_1, I_2, I_3), \tag{1}$$

where $(f_1, f_2)^*$ are the optimal estimated flows, N_θ is the network with parameters θ, and $p(f_1, f_2 \mid I_1, I_2, I_3)$ is the posterior distribution of multi-frame optical flow. Using Bayes' theorem, we expand $p(f_1, f_2 \mid I_1, I_2, I_3)$:

$$\begin{aligned} p(f_1, f_2 \mid I_1, I_2, I_3) &= \frac{p(f_1, f_2)p(I_1, I_2, I_3 \mid f_1, f_2)}{p(I_1, I_2, I_3)} \\ &= \frac{p(f_1, f_2)p(I_2 \mid f_1, f_2)p(I_1 \mid I_2, f_1, f_2)p(I_3 \mid I_1, I_2, f_1, f_2)}{p(I_1, I_2, I_3)} \\ &= \frac{p(I_2)p(f_1, f_2 \mid I_2)p(I_1 \mid I_2, f_1)p(I_3 \mid I_1, I_2, f_1, f_2)}{p(I_1, I_2, I_3)} \\ &= \frac{p(I_2)p(f_1 \mid I_2)p(f_2 \mid f_1, I_2)p(I_1 \mid I_2, f_1)p(I_3 \mid I_1, I_2, f_1, f_2)}{p(I_1, I_2, I_3)} \end{aligned} \tag{2}$$

Since $p(f_1, f_2 \mid I_1, I_2, I_3)$ is a probability density of (f_1, f_2), we omit irrelevant terms $p(I_2), p(I_1, I_2, I_3)$ and apply the maximum likelihood estimation to get the following formulation:

$$\begin{aligned} (f_1, f_2)^* = \arg\max_{(f_1, f_2)} [&\underbrace{\{\log(p(f_1 \mid I_2))\}}_{\text{context}} + \underbrace{\{\log(p(f_2 \mid f_1, I_2))\}}_{\text{motion relation}} \\ &+ \underbrace{\{\log(p(I_1 \mid I_2, f_1))\}}_{\text{cost volume 1}} + \underbrace{\{\log(p(I_3 \mid I_1, I_2, f_1, f_2)\}}_{\text{cost volume 2}}]. \end{aligned} \tag{3}$$

Here, according to the definitions of cost volume and context, we get these terms. The motion relation, cost volume 1, and cost volume 2 terms capture high-frequency information, while the context term reflects low-frequency motion data. The context term aids optical flow estimation by enhancing feature understanding.

Based on the cost volume terms, it can be observed that image information can affect the estimation of optical flow. Taking into account the different effects of different image features, we consider introducing frequency information to assist in optical flow estimation.

Some methods use pre-trained visual features for high-level understanding [15], while others focus on cost volume representations to capture long-range dependencies [13, 26]. In multi-frame settings, each cost volume encodes pairwise matching cues, with high-level semantics aiding motion reasoning and low-frequency features providing spatial context. Empirical evidence shows that combining cross-frequency features with attention in patch matching preserves motion continuity effectively.

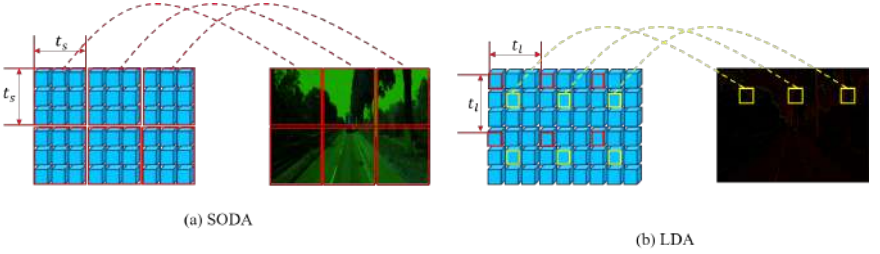


Figure 3: **(a)** SODA module: High-frequency features (blue cubes) are grouped within red boxes. **(b)** LDA module: Low-frequency features with identical border colors indicate grouping for large displacement modeling.

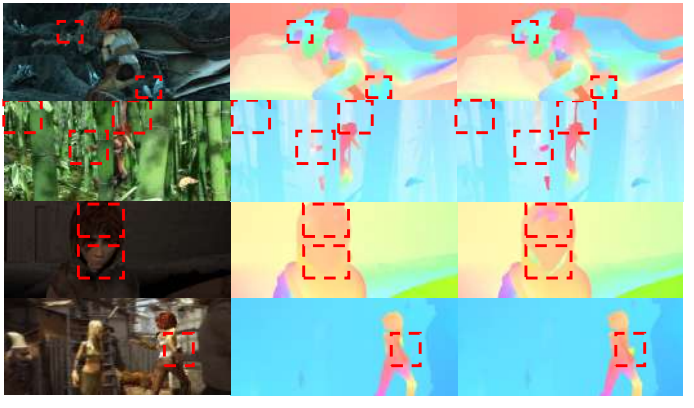


Figure 4: It can be seen from the red box that CFFlow better preserves the small object motion and the large displacement of the objects.

3.2 Overview

As shown in Figure 2, we redesign the optical flow feature extraction process by introducing the Image Frequency Separation Module (IFS), which decomposes extracted features into low- and high-frequency components. Built on this framework, our model CFFlow incorporates two specialized modules: SODA for small object motion and LDA for large displacement, enabling accurate flow estimation across scales. A Feature Fusion Module (FF) is then used to integrate both frequency branches. The following sections detail the design of IFS, SODA, LDA, and FF.

3.3 Image Frequency Separation Module

We introduce image frequency separation to improve robustness in complex scenes. Traditional methods process spatial features as a whole, making them prone to interference. Our IFS addresses this by decomposing features into high- and low-frequency components, handling large and small displacements. These components differ in spatial resolution. Given input features $f \in R^{H\times W}$, the high-frequency f^h and low-frequency f^l are computed as:

$$\begin{aligned} f_0^l &= \text{Downsample}(f) \in R^{H/s\times W/s}, \\ f_0^h &= f - \text{Upsample}(f_0^l) \in R^{H\times W}, \end{aligned} \tag{4}$$

Table 1: **Experiments on Sintel [2] and KITTI [9] datasets.** '*' denotes the multi-frame methods. We use **bold** and _ to highlight the methods that rank 1st and 2nd.

Training Data	Method	Sintel(train)		KITTI-15(train)		Sintel(test)		KITTI-15(test)
		Clean	Final	EPE	F1-all	Clean	Final	F1-all
C+T	FlowNet2 [19]	2.02	3.54	10.08	30.0	3.96	6.02	-
	SKFlow [30]	1.22	2.46	4.27	15.5	-	-	-
	RAFT [33]	1.43	2.71	5.04	17.4	-	-	-
	FlowFormer [13]	0.94	2.33	4.09	14.72	-	-	-
	FlowFormer++ [26]	0.90	2.30	3.93	**14.13**	-	-	-
	CRAFT [28]	1.27	2.79	4.88	17.5	-	-	-
	GMFlow [34]	1.08	2.48	-	-	-	-	-
	GMA [16]	1.30	2.74	4.69	17.1	-	-	-
	Videoflow-BOF [25]	1.03	2.19	3.96	15.33	-	-	-
	Videoflow-MOF [25]	1.18	2.56	**3.89**	14.20	-	-	-
	Ours	**0.89**	**2.11**	3.88	14.24	-	-	-
C+T+S+K(+H)	PWC-Net [29]	(1.71)	(2.34)	(1.50)	(5.3)	3.44	4.58	7.71
	GMFlow [34]	-	-	-	-	1.74	2.90	9.32
	SKFlow* [30]	(0.52)	(0.78)	(0.51)	(0.94)	1.28	2.27	4.48
	RAFT* [33]	(0.75)	(1.21)	(0.63)	(1.5)	1.94	3.18	5.11
	CRAFT [28]	(0.60)	(1.06)	(0.58)	(1.33)	1.45	2.40	4.81
	GMA [16]	(0.62)	(1.07)	(0.57)	(1.2)	1.39	2.48	5.14
	PWC-Fusion* [27]	-	-	-	-	3.43	4.57	7.17
	FlowFormer [13]	(0.48)	(0.74)	(0.53)	(1.11)	1.16	2.09	4.68
	FlowFormer++ [26]	(0.40)	(0.60)	(0.57)	(1.16)	1.07	1.94	4.52
	MatchFlow(R) [6]	(0.51)	(0.81)	(0.59)	(1.3)	1.33	2.64	4.72
	SAMFlow [35]	-	-	-	-	1.00	2.08	4.49
	MatchFlow(G) [6]	(0.49)	(0.78)	(0.55)	(1.1)	1.16	2.37	4.63
	Videoflow-BOF* [25]	(0.37)	(0.54)	(0.52)	**(0.85)**	1.02	1.84	4.44
	SplatFlow* [35]	(0.53)	(0.91)	(0.80)	(2.4)	1.12	2.07	4.61
	Ours	**(0.35)**	**(0.47)**	**(0.45)**	(0.91)	**0.99**	**1.65**	**4.21**

Table 2: **Results on the Sintel Test Set with Different Evaluation Metrics.**

Method	Sintel Test(clean)						Sintel Test(final)					
	All	Matched	Unmatched	s_{0-10}	s_{10-40}	s_{40+}	All	Matched	Unmatched	s_{0-10}	s_{10-40}	s_{40+}
MemFlow [8]	1.064	0.426	6.091	0.253	0.778	5.623	1.914	0.931	9.928	0.430	1.382	10.556
SAMFlow [35]	1.000	0.384	5.966	0.252	0.760	**5.245**	2.080	1.036	10.60	0.515	1.488	11.278
VideoFlow-BOF [25]	1.005	0.389	6.023	0.229	0.695	5.605	1.713	0.812	9.054	**0.387**	1.242	9.422
StreamFlow [31]	1.041	**0.381**	6.416	**0.212**	**0.690**	5.999	1.874	0.824	10.435	0.409	**1.240**	10.674
FlowFormer++ [26]	1.073	0.390	6.635	0.252	0.796	5.810	1.943	0.878	10.627	0.438	1.404	10.712
Ours(CFFlow)	**0.989**	0.396	**5.819**	0.229	0.694	5.470	**1.647**	**0.788**	**8.650**	0.402	1.242	**8.793**

where s is the stride of upsample and downsample. f_0^l and f_0^h represent initial low-frequency features and high-frequency features of f. In order to fully separate high-frequency and low-frequency information, f_0^l and f_0^h will be separated 4 times:

$$\begin{aligned} f_i^l &= F_{ll}(f_{i-1}^l) + \text{Downsample}(F_{hl}(f_{i-1}^h)) \in R^{H/s \times W/s}, \\ f_i^h &= F_{hh}(f_{i-1}^h) + \text{Upsample}(F_{lh}(f_{i-1}^l)) \in R^{H \times W}, \end{aligned} \tag{5}$$

Here, $i = 1,2,3,4$. F_{ll}, F_{hl}, F_{lh}, and F_{hh} denote convolutions between low/high-frequency input and output channels, respectively. The final high- and low-frequency components of input f are given by f_4^h and f_4^l.

3.4 Large Displacement Attention Module

We divide the self-attention mechanism into two parts: Large Displacement Attention Module (LDA) and Small Object Displacement Attention Module (SODA), which handle f_4^l and f_4^h respectively. LDA is designed to handle the low-frequency component f_4^l, primarily to capture long-range attention. Specifically, feature grouping is required. Since most large displacements require long-range

Table 3: **Results of Different Image Frequency Separation Modules.**

Methods	Sintel(train)		KITTI(train)	
	clean	final	F1-epe	F1-all
OctaveConv [5]	0.39	0.54	**0.48**	1.09
Haar Wavelet [27]	0.51	0.49	0.54	**0.89**
OctaveResent [20]	**0.35**	**0.47**	0.45	0.91

attention, it is considered to group the images in intervals, thereby achieving attention to long-distance features.

As shown in Figure 3, for the input feature f_4^l of size $R^{H/s\times W/s}$, sampling and grouping are performed at a fixed interval t_l. Each sampling group G has a size of $R^{H/(st_l)\times W/(st_l)}$.The f_4^l group form G^l is as follows:

$$G^l = [G_1, G_2, ..., G_{k_l}]^T, \tag{6}$$

where k_l is the number of groups, which is related to the t_l value. After grouping, standard self-attention is computed within each group G in LDA.

$$f_{LDA}^l = \text{Self-attention}(G^l). \tag{7}$$

3.5 Small Object Displacement Attention Module

As described in Sec. 3.4, SODA is designed to process the high-frequency feature f_4^h, focusing on short-range attention for small object motion. Since the displacements of small objects are mostly relatively close in distance and relatively small, they may be overlooked. Therefore, attention must be paid to the features of small neighborhoods.

As illustrated in Figure 3, the input $f_4^h \in \mathbb{R}^{H\times W}$ is divided into adjacent $t_s \times t_s$ pixel groups, forming the group structure G^s used by SODA:

$$G^s = [G_1, G_2, ..., G_{k_s}]^T, \tag{8}$$

where k_s is the number of groups, which is related to the t_s value. After grouping, standard self-attention is performed within each SODA group.

$$f_{SODA}^h = \text{Self-attention}(G^s). \tag{9}$$

3.6 Feature Fusion Module

As shown in Figure 2, high-frequency feature f_{SODA}^h and low-frequency feature f_{LDA}^l are processed in separate branches (B_1, B_2), each with Layer Normalization (LN), positional embedding (PE), dynamic PE (DPE), followed by LDA or SODA, another LN, and an MLP (multi-layer perception). The low-frequency branch is upsampled by 2× for alignment, and both branches are fused via attention to produce the final feature f^u.

$$f_{out}^h = B_1(f_{SODA}^h), f_{out}^l = B_2(f_{LDA}^l), \tag{10}$$

where f_{out}^h and f_{out}^l represent the outputs of two branches. Then we focus on the above results to obtain the final fused features through the following formula:

$$f^u = f + Attention(f_{out}^h, f_{out}^l, f), \tag{11}$$

Where f is the input image feature. Specifically, the features f_1, f_2, and f_3 of the input images I_1, I_2, and I_3 are processed through the IFS, SODA, LDA, and FF modules, ultimately resulting in the fused features f_1^u,f_2^u, and f_3^u.

Table 4: **Results of Different Sizes of LDA and SODA**

Module	Group Interval (t_l, t_s)	Things(val)		KITTI(train)	
		clean	final	F1-epe	F1-all
LDA (t_l)	1	1.48	2.11	4.22	15.07
	3	1.31	**1.78**	**3.99**	14.67
	5	1.22	2.03	4.02	**14.28**
	7	**1.19**	1.84	4.10	14.72
SODA (t_s)	1	1.38	2.08	4.17	15.24
	3	1.12	**1.46**	**3.93**	14.47
	5	**1.06**	1.58	4.08	14.53
	7	1.21	1.49	3.99	14.49

4 Experiments

4.1 Implementation Details

Based on some previous works, we utilize the feature encoder and context encoder as ImageNet Twins-SVT. We divide the training process into two stages: the C+T-stage and the C+T+S+K+H-stage. In the C+T stage, we first train for 300k iterations on the FlyingChairs [8] and FlyingThings [23] datasets. Since the FlyingChairs dataset only contains training data pairs, we achieve the multi-frame effect through grouping. In the C+T+S+K+H stage, we fine-tune the model trained in the C+T stage for 120k iterations on the Sintel [2], KITTI-2015 [9], and HD1K datasets [17]. We use the AdamW optimizer and a one-cycle learning rate scheduler. The batch size for all stages is 6. We use AEPE and F1-all as evaluation metrics.

4.2 Comparison with State-of-the-art Methods

4.2.1 Generalization Performance

In Table 1, the "$C+T$" setting reflects the model's cross-dataset generalization ability. Specifically, on the challenging Sintel final pass, CFFlow ranks first and surpasses the VideoFlow series. It is noteworthy that VideoFlow and FlowFormer++ have 13.4% and 52.9% more parameters than CFFlow (13.5M vs. 18.2M vs. 11.9M). Additionally, FlowFormer++ undergoes pre-training using a masked autoencoder strategy on the YouTube-VOS dataset [39]. And for the clean pass and KITTI-2015, our CFFlow achieves performance on par with the state-of-the-art methods.

4.2.2 Dataset-specific Performance

After the "$C+T+S+K(+H)$" stage, we submit to the Sintel and KITTI benchmarks. As shown in Table 1, our model achieves **0.99** and **1.65** AEPE on the Sintel clean and final passes, outperforming most methods, especially SAMFlow and MatchFlow on the final pass. This confirms that cross-frequency features are more effective than standard image feature enhancements. On KITTI-2015, CFFlow achieves a 4.21% F1-all error, surpassing both FlowFormer++ and VideoFlow-BOF, with consistent gains over the latter under the same three-frame setting. Full results are available on the official benchmark websites.

4.3 Performance Analysis for Different Regions

To comprehensively evaluate CFFlow, we present additional metrics in Table 2, including "**Unmatched**" AEPE for pixels that do not match across frames and s_{0-10}, s_{10-40}, s_{40+} for different ranges of flow

Table 5: **The Parameters and Inference of Different Models.**

Model	Para.	Infer.
GMA [16]	5.9M	74ms
FlowFormer [13]	18.2M	149ms
SAMFlow [45]	-	450ms
CRAFT [28]	6.4M	116ms
MatchFlow(G) [7]	15.4M	126ms
CFFlow(Ours)	11.9M	174ms

magnitudes of the ground truth. Compared with five state-of-the-art methods: MemFlow, SAMFlow, VideoFlow, StreamFlow, and FlowFormer++. CFFlow achieves the **best performance in the "Unmatched" region** on both clean and final passes, demonstrating its strength in handling occlusion and mismatches through cross-frequency features (see Figure 4). Moreover, CFFlow remains competitive in large-displacement scenarios caused by fast motion and severe blur, outperforming SAMFlow and MemFlow on the final pass.

4.4 Visualization

In Figure 4, we present examples of visualization results on a clean pass of the Sintel test sets. From left to right, each row represents input images and predicted flows of the multi-frame method MemFlow [6] and our CFFlow. Compared to MemFlow, CFFlow performs better in regions with small object details (third row, fourth row) and areas prone to unmatched issues, such as occlusions (second row). At the same time, CFFlow also demonstrates advantages regarding object integrity (first row). It shows that cross-frequency is valid in large displacement scenarios.

4.5 Ablation Study

4.5.1 Image Frequency Separation Module

We design the Image Frequency Separation Module using Octave Convolution to reduce spatial redundancy and preserve key frequency information. Inspired by YOLOv9's Haar Wavelet DownSample, we adapt it for frequency separation. As shown in Table 3, OctaveConv provides reasonable results but does not fully optimize separation. Haar Wavelet improves Sintel accuracy, but underperforms on KITTI F1-all. In contrast, OctaveResNet, which replaces standard convolutions in ResNet with OctaveConv, achieves the best balance, improving both Sintel (0.35, 0.47) and KITTI F1-all (0.91). These results show that OctaveConv in ResNet improves frequency separation while maintaining important details.

4.5.2 LDA and SODA Module

We analyze the impact of the interval settings (t_l, t_s) on the model performance through optical flow estimation experiments with LDA and SODA under different intervals of the module group (see Table 4). The results demonstrate that the module group interval is one of the key factors affecting the accuracy and generalization capacity of optical flow estimation. For both SODA and LDA, an interval of 3 proves to be an ideal configuration, where the models achieve a better balance in performance across multiple evaluation metrics. However, the performance variations under different intervals reveal differences between the two methods in feature interaction and optical flow information mining, providing valuable references for further optimization of module design in optical flow estimation models.

4.5.3 Parameters and Runtime Analysis

Table 5 shows that our model achieves a good balance between performance and efficiency. With 11.9M parameters, lower than FlowFormer (18.2M) and close to CRAFT (6.4M), it runs at 174 ms, faster than SAMFlow (450 ms) and CRAFT (422 ms), although slightly slower than MatchFlow (126 ms). Despite added processing over FlowFormer (149ms), it maintains competitive efficiency.

5 Conclusion

This paper investigates the integration of cross-frequency features into optical flow estimation. Current multi-frame methods often focus on surface-level cues, overlooking deeper frequency information. To address this, we propose CFFlow, which incorporates cross-frequency features via two modules: LDA for large displacements and SODA for small object motion. Experiments show that CFFlow achieves state-of-the-art performance on Sintel and KITTI, ranking first among three-frame methods. However, fully embedding cross-frequency knowledge remains a challenge for future work.

6 Acknowledgements

This work was supported by National Science and Technology Major Project from Minister of Science and Technology, China (2021ZD0201403), Youth Innovation Promotion Association, Chinese Academy of Sciences (2021233), Shanghai Academic Research Leader (22XD1424500).

References

[1] Wenbo Bao, Xiaoyun Zhang, Li Chen, and Zhiyong Gao. Kalmanflow 2.0: Efficient video optical flow estimation via context-aware kalman filtering. *IEEE Transactions on Image Processing*, 28 (9):4233–4246, 2019.

[2] D Butler, Jonas Wulff, G Stanley, and M Black. Mpi-sintel optical flow benchmark: Supplemental material. In *MPI-IS-TR-006, MPI for Intelligent Systems (2012*. Citeseer, 2012.

[3] Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. Frequency-aware feature fusion for dense image prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[4] Yonghu Chen, Dongchen Zhu, Wenjun Shi, Guanghui Zhang, Tianyu Zhang, Xiaolin Zhang, and Jiamao Li. Mfcflow: A motion feature compensated multi-frame recurrent network for optical flow estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5068–5077, 2023.

[5] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3435–3444, 2019.

[6] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19068–19078, 2024.

[7] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Rethinking optical flow from geometric matching consistent perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1337–1347, 2023.

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[10] Davar Giveki. Human action recognition using an optical flow-gated recurrent neural network. *International Journal of Multimedia Information Retrieval*, 13(3):29, 2024.

[11] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17 (1-3):185–203, 1981.

[12] Ying Jiao Huang, Xia Fei Ma, Rong Rong Xue, and Hao Tong Ma. Cff-net: cross-frequency fusion network for aero-degraded image restoration. In *AOPC 2024: Optical Sensing, Imaging Technology, and Applications*, volume 13496, pages 350–355. SPIE, 2024.

[13] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 668–685. Springer, 2022.

[14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[15] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9000–9008, 2018.

[16] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021.

[17] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016.

[18] Ying-Yee Kong and Louis D Braida. Cross-frequency integration for consonant and vowel identification in bimodal hearing. *Journal of Speech, Language, and Hearing Research*, 54(3):959–980, 2011.

[19] Salim Lahmiri. Wavelet low-and high-frequency components as features for predicting stock prices with backpropagation neural networks. *Journal of King Saud University-Computer and Information Sciences*, 26(2):218–227, 2014.

[20] Pengtao Li, Meihua Wang, Zhun Fan, Han Huang, Guijie Zhu, and Jiafan Zhuang. Our-net: a multi-frequency network with octave max unpooling and octave convolution residual block for pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[21] Yawen Lu, Qifan Wang, Siqi Ma, Tong Geng, Yingjie Victor Chen, Huaijin Chen, and Dongfang Liu. Transflow: Transformer as flow learner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18063–18073, 2023.

[22] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679, 1981.

[23] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[24] Zhile Ren, Orazio Gallo, Deqing Sun, Ming-Hsuan Yang, Erik B Sudderth, and Jan Kautz. A fusion approach for multi-frame optical flow estimation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2077–2086. IEEE, 2019.

[25] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023.

[26] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1610, 2023.

[27] Radomir S Stanković and Bogdan J Falkowski. The haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, 29(1):25–44, 2003.

[28] Xiuchao Sui, Shaohua Li, Xue Geng, Yan Wu, Xinxing Xu, Yong Liu, Rick Goh, and Hongyuan Zhu. Craft: Cross-attentional flow transformer for robust optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17602–17611, 2022.

[29] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[30] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. Skflow: Learning optical flow with super kernels. *Advances in Neural Information Processing Systems*, 35:11313–11326, 2022.

[31] Shangkun Sun, Jiaming Liu, Huaxia Li, Guoqing Liu, Thomas Li, and Wei Gao. Streamflow: streamlined multi-frame optical flow estimation for video sequences. *Advances in neural information processing systems*, 37:9205–9228, 2024.

[32] Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, pages 343–360. Springer, 2024.

[33] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[34] BCM van Wijk, A Jha, W Penny, and V Litvak. Parametric estimation of cross-frequency coupling. *Journal of neuroscience methods*, 243:94–102, 2015.

[35] Bo Wang, Yifan Zhang, Jian Li, Yang Yu, Zhenping Sun, Li Liu, and Dewen Hu. Splatflow: Learning multi-frame optical flow via splatting. *International Journal of Computer Vision*, 132 (8):3023–3045, 2024.

[36] Mengfei Wang, Wenjun Shi, Dongchen Zhu, Lei Wang, and Jiamao Li. Memoflow: Modifying explicit motion of inconsistency in optical flow. In *International Conference on Pattern Recognition*, pages 219–234. Springer, 2025.

[37] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.

[38] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[39] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[40] Mingliang Zhai, Xuezhi Xiang, Ning Lv, and Xiangdong Kong. Optical flow and scene flow estimation: A survey. *Pattern Recognition*, 114:107861, 2021.

[41] Dengsheng Zhang. Wavelet transform. In *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pages 35–44. Springer, 2019.

[42] Kaidong Zhang, Jialun Peng, Jingjing Fu, and Dong Liu. Exploiting optical flow guidance for transformer-based video inpainting. *IEEE transactions on pattern analysis and machine intelligence*, 46(7):4977–4992, 2024.

[43] Zhiyong Zhang, Aniket Gupta, Huaizu Jiang, and Hanumant Singh. Neuflow v2: High-efficiency optical flow estimation on edge devices. *arXiv preprint arXiv:2408.10161*, 2024.

[44] Zhiyong Zhang, Huaizu Jiang, and Hanumant Singh. Neuflow: Real-time, high-accuracy optical flow estimation on robots using edge devices. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5055. IEEE, 2024.

[45] Shili Zhou, Ruian He, Weimin Tan, and Bo Yan. Samflow: Eliminating any fragmentation in optical flow with segment anything model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7695–7703, 2024.

[46] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 熊兵

方向： 人工智能

单位： 中国科学院深圳先进技术研究院

题目： USIGAN: Unbalanced Self-Information Feature Transport for Weakly Paired Image IHC Virtual Staining

USIGAN: Unbalanced Self-Information Feature Transport for Weakly Paired Image IHC Virtual Staining

Yue Peng†,Bing Xiong†,Fuqiang Chen,Deboch Eybo Abera, RanRan Zhang, Wanming Hu,Jing Cai,Wenjian Qin*

***Abstract*—Immunohistochemical (IHC) virtual staining is a task that generates virtual IHC images from H&E images while maintaining pathological semantic consistency with adjacent slices. This task aims to achieve cross-domain mapping between morphological structures and staining patterns through generative models, providing an efficient and cost-effective solution for pathological analysis. However, under weakly paired conditions, spatial heterogeneity between adjacent slices presents significant challenges. This can lead to inaccurate one-to-many mappings and generate results that are inconsistent with the pathological semantics of adjacent slices. To address this issue, we propose a novel unbalanced self-information feature transport for IHC virtual staining, named USIGAN, which extracts global morphological semantics without relying on positional correspondence. By removing weakly paired terms in the joint marginal distribution, we effectively mitigate the impact of weak pairing on joint distributions, thereby significantly improving the content consistency and pathological semantic consistency of the generated results. Moreover, we design the Unbalanced Optimal Transport Consistency Mining (UOT-CTM) mechanism and the Pathology Self-Correspondence Mining (PC-SCM) mechanism to construct correlation matrices between H&E and generated IHC in image-level and real IHC and generated IHC image sets in intra-group level. Experiments conducted on two publicly available datasets demonstrate that our method achieves superior performance across multiple clinically significant metrics, such as IoD and Pearson-R correlation, demonstrating better clinical relevance. The code is available at: https://github.com/MIXAILAB/USIGAN**

***Index Terms*—Histopathology, Virtual Stain, Unbalanced Optimal Transport, Self Information Mining**

I. Introduction

IMMUNOHISTOCHEMISTRY(IHC) virtual staining is a task that aims to generate virtual IHC images consistent with the pathological semantics of adjacent reference IHC-stained slices while preserving the content consistency of H&E images. However, spatial heterogeneity between adjacent slices makes it challenging to directly establish a mapping between H&E and IHC images. In previous studies, most researchers adopted contrastive learning approaches, which maximize mutual information in the corresponding regions to establish cross-domain correlations between the morphological information in H&E images and the staining styles in IHC images. However, this approach still fails to address the issue of spatial heterogeneity, which can lead to misalignment of morphological information from H&E images to incorrect IHC staining patterns, affecting the extraction and accurate mapping of key information in generative models [37, 22, 47]. Therefore, we focus on a previously overlooked concept in information theory—self-information, which quantifies the information content of a single event. Specifically, self-information reflects the rarity of a pixel. For example, when rare structures or patterns appear in pathological images, the corresponding regions exhibit higher self-information, indicating that they contain more valuable diagnostic information and biological significance. In summary, self-information quantifies the model's prioritization of information needs for critical pathological features.

Our goal is to transform self-information from a previously passive learning paradigm into an active focus, further uncovering the process by which generative models establish cross-domain correlations in virtual staining. Self-information represents critical regions in the mapping process, and constructing a correlation matrix allows different feature regions to be assigned higher weights, which is the essence of self-information mining. As illustrated in Figure 1, IHC virtual staining faces two main challenges: (1) spatial heterogeneity and (2) erroneous pathological semantic correlations caused by incorrect one-to-many mappings between morphological features and staining patterns. Self-information mining addresses these challenges by prioritizing key information. We leverage two types of priors in virtual staining to incorporate corresponding self-information and establish correlations: 1) Explicit Prior: Pathological semantic consistency between adjacent slices. By evaluating staining intensity through clinical optical density metrics, we construct a correlation matrix between generated IHC and real IHC image groups to mine pathological semantic self-information. 2) Implicit Prior: Morphological structures and cellular atypia in H&E images correspond to different disease subtypes and their respective IHC staining patterns [23]. Through optimal transport theory, we optimize the expectation of self-information, explicitly introducing self-information to strengthen critical regions and penalizing the effects of erroneous one-to-many mappings.

In previous studies on IHC virtual staining, two paradigms: fully supervised and unsupervised—have been primarily used to passively construct cross-domain correlations. Pix2Pix and its variants attempt to establish cross-domain correlations at

† Yue Peng and Bing Xiong contribute equally to this work.
∗ Wenjian Qin are corresponding author to this work.
B. Xiong, Y. Peng, F.Chen, Deboch. Abera, R. Zhang are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Shenzhen, 518055, China. University of Chinese Academy of Sciences, Beijing, 100049, China. Email: wj.qin@siat.ac.cn.
W. Hu are with Department of Pathology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China,Guangdong Provincial Clinical Research Center for Cancer, Guangzhou 510060, China
J.Cai are with Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong 999077, China

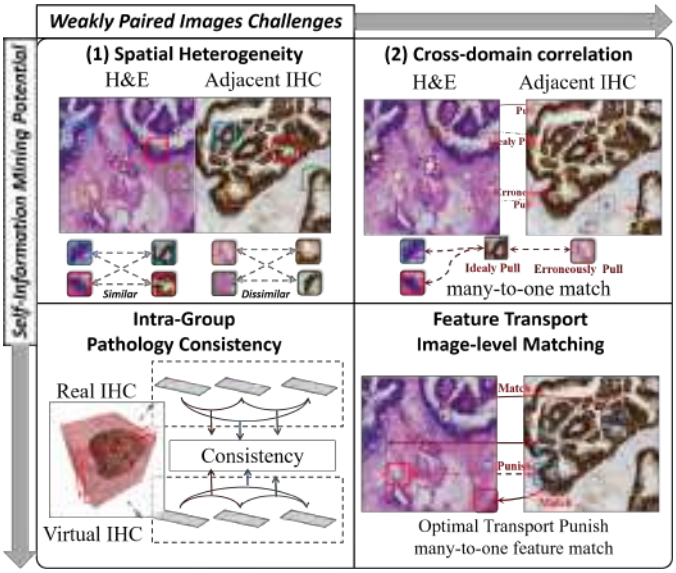


Fig. 1. Spatial heterogeneity between adjacent slices poses a major challenge in virtual staining. This heterogeneity manifests as misalignment of tissue structures and pathological semantics across slices, making it difficult to directly establish cross-domain mappings between morphology and staining styles.

the pixel level by leveraging weakly paired IHC data [14, 34, 11, 41]. However, these methods struggle to effectively learn mapping relationships between tissue regions. Cycle consistency-based methods implicitly construct mappings between morphology and staining styles by allowing two models to map between two domains [59]. However, the assumption of reciprocal mappings between staining domains is often violated for weakly paired data, resulting in low pathological semantic consistency in the generated results. To address these issues, some researchers have adopted contrastive learning approaches that maximize mutual information between patches to explicitly construct mapping relationships [14, 34, 11, 41]. While this non-pixel-level approach alleviates the impact of weakly paired data to some extent, its performance remains constrained by the degree of pairing. These methods overlook the critical role of self-information in pathological images. Passively summarizing mapping relationships from image collections is inherently challenging, whereas self-information, by emphasizing features with high diagnostic value, can guide the model to correctly focus on key regions, thereby improving the performance and consistency of virtual staining results.

In this paper, we propose a novel method for mining self-information at both the image and group levels by eliminating weakly paired terms in the joint marginal distribution, effectively mitigating their impact and significantly enhancing the content and pathological semantic consistency of generated results. Specifically, for image-level self-information mining, we introduce Unbalanced Optimal Transport Consistency Mining (UOT-CTM), a method that preserves geometric structures of high self-information regions to reduce transport costs during optimal transport minimization. Although previous attempts using optimal transport explicitly constrained transport plans from H&E to weakly paired IHC and from generated IHC to weakly paired IHC in feature space—thus aligning morphological and staining style features—these approaches suffer from three primary limitations: 1) Neglecting incompatibility in weakly paired data, leading to forced, low-similarity feature matching. 2) Ignoring inherent semantic differences between morphological and staining-style imitation tasks, resulting in undesirable trade-offs. 3) Overlooking semantic inconsistencies and distributional differences by directly constraining joint marginal distributions involving weakly paired terms. To address these issues, UOT-CTM employs a cyclic consistent transport strategy to explicitly constrain transport plans along paths from H&E to generated IHC, effectively capturing image-level self-information. Additionally, we propose a marginally relaxed unbalanced optimal transport strategy to alleviate the one-to-many mapping issue and construct more accurate correlation matrices, transforming the constraint objective from directly matching $P(x, y)$ and $P(y, z)$ distributions to matching $P(x, z)$ and $P(x, y) \cdot P(y, z)$, thus eliminating weakly paired terms. For group-level self-information mining, we design the Pathology-Consistent Self-Correspondence Mining (PC-SCM) mechanism, which leverages optical density—rare but indicative of high self-information—as a "self-information anchor," establishing correlation matrices between generated and real IHC based on optical density. This ensures prioritization of high-optical-density regions, maintains consistent correlation matrices within batches, avoids negative impacts from directly using weakly paired IHC as learning templates, and effectively captures pathological semantic information. Our contributions can be concluded as follows:

- We introduce the concept of self-information of the image in histopathology. By maximizing this self-information, we achieve a marked boost in cross-domain correlation accuracy, delivering virtual IHC results that are superior in both content and pathological semantic consistency.
- To mine image-level self-information we employ unbalanced optimal transport, which mitigates the adverse effects of flawed one-to-many mappings through feature-matching transport.
- We utilize optical density to directly anchor pathological semantic information, guiding intra-batch self-information mining. This approach focuses on abnormal regions represented by high optical density, thereby indirectly leveraging self-information and mitigating the impact of spatial heterogeneity.

II. Related Work

A. Image-to-Image Translation

The task of image translation aims to map images from a source domain to a target domain while preserving their original content [51, 17, 13, 42]. Pix2Pix used a patch-based discriminator on paired data to alleviate the tendency of previous generative models toward style averaging [14], establishing itself as a cornerstone method in supervised image translation. Subsequent supervised methods introduced various extensions, such as cross-layer connections for high-resolution generation and perceptual constraints [41]. However, due to its strict requirement for pixel-level paired data, Pix2Pix is difficult to

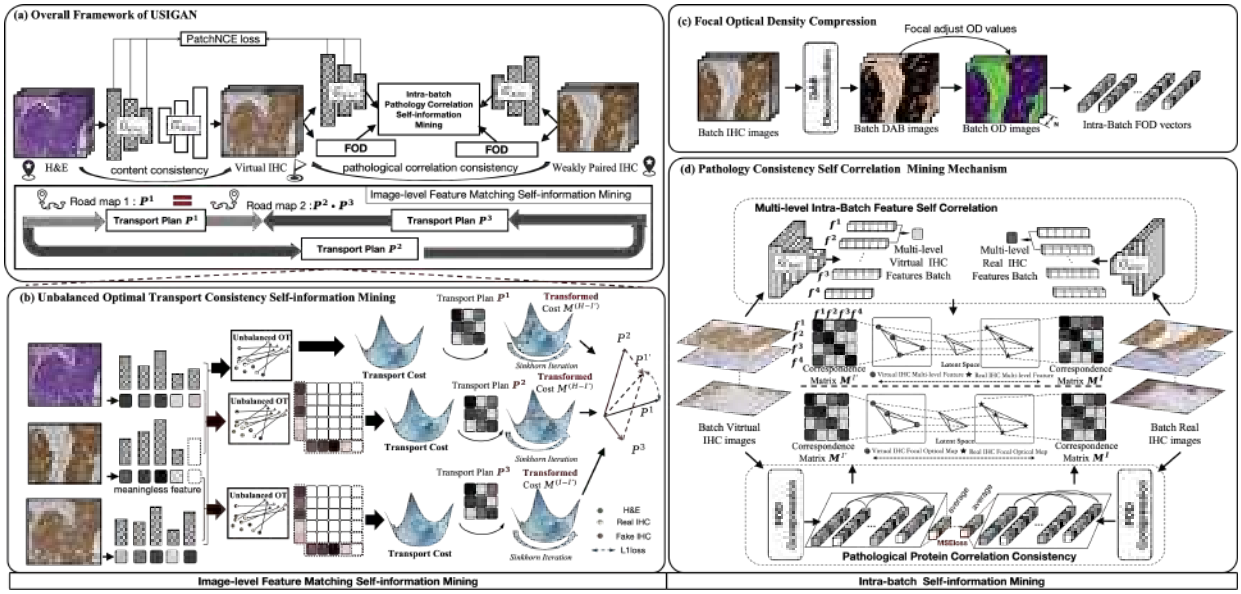


Fig. 2. We leverage weakly paired IHC as an intermediate bridge to ensure global consistency between H&E and IHC while mining image-level self-information through a transport consistency framework. Simultaneously, we utilize focal optical density and DAB deconvolution to extract optical density feature vectors to guide intra-batch self-information mining. Multi-scale features within the batch are used as auxiliary features to fully explore intra-batch self-information.

generalize to broader applications. CycleGAN [59] introduced a framework based on the bidirectional mapping hypothesis and cycle consistency loss, which became the foundation for subsequent methods . To address the overly strict bidirectional mapping assumption of cycle consistency, CUT [7], based on contrastive learning, constructs cross-domain correlations using positive and negative patch samples. Contrastive learning-based image translation methods grounded in optimal transport theory have also garnered significant attention. Techniques such as weighted positive and negative samples [51], hard negative sample generation [43], and unbalanced optimal transport have been extensively studied and achieved notable progress [50, 53].

Although existing image translation methods show great potential in pathological image stain transfer, the lower coupling between content and style in pathological images, along with the higher demand for content consistency [47], makes it challenging for these methods to accurately generate virtual images that maintain both pathological features and content consistency.

B. Immunohistochemical Virtual Staining

Immunohistochemical virtual staining aims to generate staining results for specific protein markers from source-stained images while preserving image content. This technique is gaining increasing attention in digital pathology. Zhang et al. proposed an IHC multi-staining mapping method that measures the accuracy of specific protein markers based on mean intensity optical density (MIOD), introducing a novel evaluation metric for assessing the molecular response accuracy of virtual staining results [55]. Liu et al. proposed a method that annotates positive regions in breast tissue H&E images to achieve high pathological correlation with adjacent slice labels, further demonstrating that morphological information can effectively distinguish positive regions [27]. Chen et al., on the other hand, sought to reduce the demand for labeled positive signal data by using DAB deconvolution for destaining, segmenting positive regions based on optical density thresholds, and employing an additional feature extractor to refine pseudo-labels [3], thereby achieving high preservation of pathological semantics. However, this method heavily relies on the quality of the feature extractor, which limits its practicality. It is evident that the morphological information in H&E images plays a crucial role in maintaining pathological relevance.

At present, the ability of optimal transport to penalize many-to-one mappings has been applied by researchers in weakly paired or unpaired image translation tasks [51, 52]. Guan et al. proposed leveraging optimal transport to compute the transport cost between H&E and virtual IHC, as well as between weakly paired IHC and virtual IHC, imposing an L1 constraint in the feature space [10, 9]. This approach encourages virtual IHC to achieve a balance between content consistency and pathological semantics across H&E and weakly paired IHC. However, due to its disregard of the classical OT assumptions and failure to fully eliminate the influence of weakly paired terms, the method produces suboptimal results.

III. Method

A. Image-level Self-information Mining

Self-information serves as a measure of the information content for specific events. In pathological images, regions exhibiting the most "abnormal" characteristics are associated with high self-information. These abnormal morphological regions often correspond to areas with high expression of molecular markers. On weakly paired data, self-information

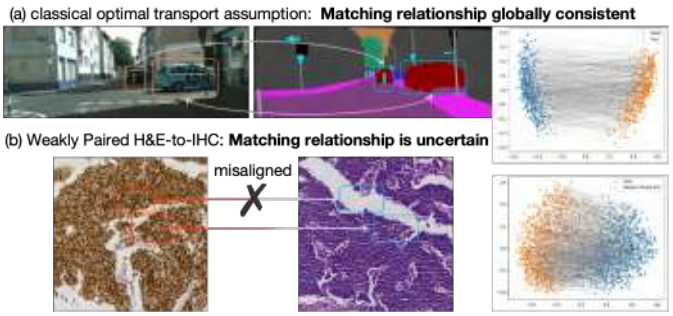


Fig. 3. Strongly paired data (Cityspace dataset) exhibit a clear global matching relationship, while weakly paired data (MIST ER) demonstrate varying matching relationships influenced by spatial heterogeneity. In features reduced by PCA, the classical optimal transport approximates a one-to-one permutation structure, which is visualized through unified open source tools POT [5]. The orange points in the distribution of Weakly Paired IHC may contain multiple subgroups, but the matching patterns of these subgroups do not exhibit global consistency.

provides the model with an intuitive focus on key regions. Unlike previous studies that establish cross-domain consistency by maximizing the mutual information of patches at the same spatial location, self-information reflects intra-image relationships, where abnormal regions consistently exhibit high self-information. More importantly, when cross-domain correlations are established by maximizing mutual information at the same spatial location, patch-level contrastive learning can partially mitigate the impact of weakly paired data. However, under weak pairing conditions, morphological-style patches with different pathological representations may be incorrectly minimized. This many-to-one mapping relationship poses significant challenges for establishing accurate cross-domain correlations.

To eliminate the weakly paired terms in the optimization objective, we aim to utilize self-information, which reflects intra-image relationships, as a bridge to capture feature relationships in weakly paired images, as shown in Figure 2. Optimal transport, as a method to measure the minimal cost of moving between two distributions under a given cost function, effectively penalizes many-to-one mappings. Self-information is incorporated into the dual objective of optimal transport through entropy regularization by influencing the marginal constraints and cost structure. Specifically, entropy regularization introduces an additional term into the optimal transport problem, which encourages smoother and more distributed transport plans. Self-information, as a measure of the information content within the image, is embedded into this regularization term. The Sinkhorn-Knopp algorithm leverages this entropy-regularized formulation to compute the optimal transport plan, where self-information directly impacts the gradients and the dual variables, thereby shaping the alignment process.

As shown in Fig. 3, classical optimal transport relies on the assumption of globally consistent mapping relationships, which is difficult to satisfy in weakly paired data. To construct an accurate transport cost and penalize suboptimal mappings, we propose unbalanced optimal transport to relax the strict marginal constraints. Instead of a "hard" mass conservation constraint, it introduces a "soft" penalty with a divergence measure. The detailed steps are outlined in the pseudo-algorithm 1.

Algorithm 1 Relaxed Sinkhorn Algorithm for Unbalanced Optimal Transport (UOT)

Input: Source features $\mathbf{q} \in \mathbb{R}^{n\times I\times D}$, target features $\mathbf{k} \in \mathbb{R}^{n\times O\times D}$

Parameters: Entropic regularization $\varepsilon > 0$, relaxation $\tau > 0$, iterations max_iter $\in \mathbb{N}$

Output: Transport plan $\mathbf{K} \in \mathbb{R}^{(n)\times I\times O}$ (batched)

Stage 1: Sinkhorn stage (UOT)
Compute batched similarity (einsum):

$$\mathbf{C}[b,i,o] \leftarrow \sum_{d=1}^{D} \mathbf{q}[b,i,d]\ \mathbf{k}[b,o,d]$$

$\mathbf{K} \leftarrow \mathbf{C}$
Mask self-matches: $\mathbf{K}[b,i,i] \leftarrow -10 \quad \forall\, b$
Entropic scaling: $\mathbf{K} \leftarrow \exp\big(\mathbf{K}/\varepsilon\big)$
Set $(n, I, O) \leftarrow \text{shape}(\mathbf{K})$; initialize

$$\mathbf{u} \leftarrow \mathbf{1}_{n\times I}, \qquad \mathbf{v} \leftarrow \mathbf{1}_{n\times O}$$

Initialize dual variables $(\mathbf{u}, \mathbf{v})$
$i \leftarrow 0$
while $i <$ max_iter **do**
  $\mathbf{P} \leftarrow \mathbf{K}$
  $\mathbf{u} \leftarrow \mathbf{u} + \frac{\log(\mathbf{1}_{n\times I}) - \log(\text{bmm}(\mathbf{P}, \mathbf{v}[:,:,\text{None}])[:,:,0])}{\tau}$
  $\mathbf{P} \leftarrow \mathbf{K}$
  $\mathbf{v} \leftarrow \mathbf{v} + \frac{\log(\mathbf{1}_{n\times O}) - \log(\text{bmm}(\mathbf{u}[:,\text{None},:], \mathbf{P})[:,0,:])}{\tau}$
  $i \leftarrow i + 1$
end while
Stage 2: Final transport plan
$\mathbf{K} \leftarrow \mathbf{u}[:,:,\text{None}] \odot \mathbf{K} \odot \mathbf{v}[:,\text{None},:]$
return $\mathbf{K}$

We define the transport plan $T^{(H-I)}$ between H&E images and weakly paired IHC images, capturing distributional differences such as texture structure, and the transport plan $T^{I-I'}$ between generated and weakly paired IHC images, reflecting molecular expression imitation differences; we assume these two transport plans should be as similar as possible in the feature space. By minimizing the cost of the transport plan, regions with high self-information receive priority, facilitating the accurate representation of distributional relationships and semantic mappings. Specifically, given that H&E images X and weakly paired IHC images Y share morphological and structural features, we define the minimal transport cost f_1 from x_{ij} to y_{ij}; similarly, as weakly paired IHC Y and generated IHC Z share staining type and texture features, we define f_2 from y_{ij} to z_{ij}. Consequently, an indirect transport plan $f_{\text{indirect}} = f_1 \cdot f_2$ can represent the sequential feature mapping through these two steps, while the direct transport plan f_{direct} between H&E and generated IHC images is computationally simpler due to their shared features and adherence to classical optimal transport assumptions.

$$\mathcal{L}_{tcyc} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\mathbf{T}^{(H-I)}\cdot\mathbf{T}^{(I-I')} - f_{direct}\right| \quad (1)$$

where f_{direct} represent $\mathbf{T}^{(H-\hat{I})}$, $\hat{I}$ is virtual IHC image. Theoretically, by combining transport plans, the joint probability distribution $P(H,\hat{I})$ between H&E images and generated IHC replaces the joint probability distributions $P(H,I)$ between H&E and weakly paired IHC, and $P(I,I')$ between weakly paired IHC and generated IHC, thereby eliminating the influence of weakly paired data. The joint marginal distributions can be formulated as:

$$P(H,I') = \sum_{I} P(H,I)\cdot P(I,I')$$

The all transport plans are computed based on the Unbalanced Optimal Transport (UOT) method.

B. Intra-batch Self-information Mining

Optical density is an indirect indicator for assessing antibody concentration in IHC-stained images, and it serves as an important basis for clinically evaluating disease subtypes in IHC images. Optical density is highly correlated with abnormal regions in the image, where high optical density values, representing high antibody concentrations, often reflect abnormal biological expressions. Therefore, as a critical metric for evaluating pathological semantics, optical density can act as a self-information anchor to directly guide the generative model in focusing on key regions. To effectively model the pathological correlation among batch-level IHC images, first, we introduce optical density, which is proportional to the concentration of the stain [3]. The amount of stain is the factor determining the OD at a wavelength as per the LambertBeerlaw [38]. It can be formulated as:

$$OD_C = -log_{10}(I_C/I_{0,C}) = A * c_C \quad (2)$$

where I_0,C and I_C denote the light intensity entering and passing the specimen. We use traditional color deconvolution [35]for stain separation. Then we specifically select DAB stain's OD values to generate the RGB image (IHC DAB). Due to the high optical density (OD) values in positive regions, which are typically much smaller in area compared to negative regions, the distribution becomes unbalanced. To amplify the influence of positive regions in self-information mining, we adjust Equation 2 to assign grayscale values to positive signals using a focal calibration map, referred to as Focal Optical Density (FOD).

$$O_C = (-log_{10}(I_C/I_{0,C}))^{\alpha} \quad (3)$$

where O is the FOD with tunable focusing parameter $\alpha > 1$. We convert virtual IHC and real IHC to O^F and O^R, which use ReLU to activate and flatten as vector f_i^F and f_i^R. Then we calculate the optical density relational function between any two real IHC feature within the batch to obtain correlation matrix $\mathbf{M}$, which can be formulated as :

$$\mathbf{M} = \begin{bmatrix} \cos(\mathbf{f}_1,\mathbf{f}_1) & \cos(\mathbf{f}_1,\mathbf{f}_2) & \cdots & \cos(\mathbf{f}_1,\mathbf{f}_n) \\ \cos(\mathbf{f}_2,\mathbf{f}_1) & \cos(\mathbf{f}_2,\mathbf{f}_2) & \cdots & \cos(\mathbf{f}_2,\mathbf{f}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\mathbf{f}_n,\mathbf{f}_1) & \cos(\mathbf{f}_n,\mathbf{f}_2) & \cdots & \cos(\mathbf{f}_n,\mathbf{f}_n) \end{bmatrix}$$

where $cos(\cdot,\cdot)$ denotes the Cosine similarity. $\mathbf{M}^R$ and $\mathbf{M}^F$ denotes real IHC and fake IHC optical density self-correlation map. Then, we propose $\mathcal{L}_{odc}$, which maximize the preservation of pathological correlation among batch samples. It can be formulated as:

$$\mathcal{L}_{odc} = ||M^R - M^F|| + \frac{1}{N^2}||\sum_{i,j} O^R(i,j) - \sum_{i,j} O^F(i,j)||_2^2 \quad (4)$$

We computed the cosine similarity of multi-scale features between real IHC and fake IHC within the generator to construct the self-correlation matrices $M^{F'}$ and $M^{R'}$. These matrices act as a shorter gradient backpropagation path, thereby influencing the encoder. $\mathcal{L}_{\text{cc}}$, as a shorter gradient backpropagation path, influences the encoder.

$$\mathcal{L}_{cc} = ||M^{R'} - M^{F'}|| \quad (5)$$

C. Loss functions

The loss function term in the USI-GAN include adversial loss $\mathcal{L}_{adv}$, PatchNCE loss $\mathcal{L}_{patchnce}$, transport consistency loss $\mathcal{L}_{tcyc}$, correlation consistency loss $\mathcal{L}_{cc}$, integrated optical density correlation consistency loss $\mathcal{L}_{odc}$. Adversial loss can be formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_{y\in Y}\log D(Y) + \mathbb{E}_{x\in X}\log(1 - D(X)) \quad (6)$$

where $\mathbb{E}$ denotes the expectation, X and Y denote the source and target domain feature. G represent Generator and D represent Discriminator. PatchNCE loss $\mathcal{L}_{nce}$ establishes cross-domain correlations by maximizing the mutual information between the input and output, which can be expressed as:

$$\mathcal{L}_{nce}(\mathbf{v},\mathbf{v}^+,\mathbf{v}^-) = -\log\left[\frac{\exp(\mathbf{v}\cdot\mathbf{v}^+/\tau)}{\exp(\mathbf{v}\cdot\mathbf{v}^+/\tau) + \sum_{n=1}^{N}\exp(\mathbf{v}\cdot\mathbf{v}^-/\tau)}\right] \quad (7)$$

where $\mathbf{v}$, $\mathbf{v}^+$ and $\mathbf{v}^-$ are the embeddings of the anchor,positive and negative samples, respectively, τ is a temperature hyper-parameter. The PatchNCE can be formulated as:

$$\mathcal{L}_{patchnce}(X) = \mathbb{E}_{x\in X}\sum_{l=1}^{L}\sum_{s=1}^{S_l}\mathcal{L}_{nce}(\mathrm{z},\mathrm{z}^+,\mathrm{z}^-) \quad (8)$$

where L represent the layer seleted from the multi-layer of encoder feature. S_l is the number of spatial location in each layer, z represent the anchor embedding from output image. the positive z^+ is the embedding of the corresponding patch from the input image, while the negatives z^- are embeddings of the non-corresponding ones.

The total loss for the USI-GAN is shown as follows:

$$\mathcal{L}_{total} = \lambda_{tcyc}\mathcal{L}_{tcyc} + \lambda_{cc}\mathcal{L}_{cc} + \mathcal{L}_{odc} + \mathcal{L}_{nce} + \mathcal{L}_{adv} \quad (9)$$

IV. Experiments And Results

A. Datasets

1) MIST Dataset: In the MIST dataset [19], there are aligned H&E-IHC patches for four different IHC stains (Ki67, ER, PR and HER2) critical to breast cancer diagnosis. All patches are of size 1,024 × 1,024 and non-overlapping.

2) IHC4BC Dataset: Similar to the MIST dataset, there are aligned H&E-IHC patches for four different IHC stains (Ki67, PR, HER2, ER) in the IHC4BC dataset [1]. The IHC4BC dataset contains more patches than the MIST dataset, containing approximately 90,000 H&E-IHC pairs. The patch size of these pairs is 1,000 × 1,000.

TABLE I
Details of MIST and IHC4BC Public Benckmark Datasets. Paired Rate represent Pearson-R of H&E and Weakly Paired IHC.

Dataset	Staining Type	No.WSI	No.Pairs (train)	No.Pairs (test)	Paired Rate
MIST	Ki67	56	4,361	1,000	8.50%
	ER	56	4,153	1,000	11.82%
	PR	56	4,139	1,000	11.46%
	HER2	64	4,642	1,000	9.11%
IHC4BC	Ki67	60	17,745	1,000	7.37%
	ER	59	26,395	1,000	16.62%
	PR	60	20,071	1,000	14.09%
	HER2	52	16,995	1,000	16.45%

The H&E-IHC image pairs in MIST and IHC4BC dataset are registered and achieve overall structual consistency. We employed the Pearson-R correlation coefficient to measure the grayscale correlation between pairs of H&E and IHC images. Each patch was resized to 1024×1024 and cropped to 512×512 during training.

B. Experimental Settings

Our method is implemented with Python based on PyTorch on a computer with Intel(R) Core(TM) i5-10400 CPU, 48 GB RAM, and a NVidia RTX A6000 GPU. All experiments were conducted with an image resolution of 1024×1024, and we trained our method with random 512×512 crops. We used 5-layer PatchGAN as the discriminator and ResNet-6Blocks as the generator for our method. For contrastive learning setting, we keep the same with CUT [31], e.g. 256 negative samples, temperature parameter $\tau = 0.07$ for PatchNCE loss, and batch size as 2. The USIGAN was trained for 80 epochs for MIST (about 2 days) and 20 epochs for IHC4BC dataset (about 3 days). During training, we used the Adam optimizer with a linear decay scheduler and an initial learning rate of 2×10^{-4}. The hyperparameters in Eq.9 were set as: $\lambda_{tcyc} = 10,000$ and $\lambda_{cc} = 10$.

C. Evaluation Metrics

In this study, we employed three dimensions of metrics to evaluate our work. Pixel-level metrics include Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). Pathology-related metrics consist of the IoD and Pearson correlation coefficient. Perceptual metrics include the Fréchet Inception Distance (FID) [12], Deep Image Structure and Texture Similarity (DISTS), and PHV$_{avg}$, which is described in [26].

For pathological evaluation, we utilized the Integrated Optical Density (IoD), following the clinical guidelines described in [15], to quantify the immunohistochemistry (IHC) images. Additionally, we calculated the Pearson correlation coefficient between the fluorescence intensities of real IHC datasets and the generated IHC images as a measure of pathological fidelity, as outlined in [3, 55]. Futhermore, we observed that cross-staining registration in pathological imaging often assumes that grayscale representations of images with different stains share similar distributions and content [45]. Therefore, we computed the Pearson correlation coefficient of pixel values between the grayscale versions of the virtually stained results and the H&E images, serving as a standard for content consistency. We use R_{avg} represent virtual IHC clinical value.

$$R_c = \frac{\sum_{i=1}^{N}(I_1^{(i)} - \bar{I}_1)(I_2^{(i)} - \bar{I}_2)}{\sqrt{\sum_{i=1}^{N}(I_1^{(i)} - \bar{I}_1)^2}\sqrt{\sum_{i=1}^{N}(I_2^{(i)} - \bar{I}_{2)^2}}} \tag{10}$$

$$R_p = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(D_i - \bar{D})(O_i - \bar{O})}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(D_i - \bar{D})^2 \cdot \frac{1}{N-1}\sum_{i=1}^{N}(O_i - \bar{O})^2}} \tag{11}$$

$$R_{avg} = (R_c + R_p)/2 * 100\% \tag{12}$$

where R_c and R_p represents content preserving and pathological preserving Pearson correlation, respectively. N is total numbers of test dataset. $I_1^{(i)}$ and $I_2^{(i)}$ are the grayscale values of the i-th pixel in each image. $\bar{I}_1$ and $\bar{I}_2$ are the mean grayscale values of the two images. D_i and O_i represents i-th image integrated density.

D. Comparison with Competitive Methods

We compare our method with baselines and thirteen state-of-art Image-to-Image translation approches. It can be divided into the following 2 types: image translation SOTA and virtual staining SOTA methods.

1) Results on MIST: Table II presents the quantitative experimental results on the MIST dataset. Our USIGAN demonstrates significant advantages in both clinical value and image quality. We evaluated our method using three categories of metrics: pixel-level, feature-level, and fluorescence intensity. Considering the instability of GAN models and the observation that our method performs better with batch sizes greater than 1 compared to other methods, we also compared the results for a batch size of 1 and results aligned with the batch size used in our method. The results indicate that, compared to supervised learning methods, a batch size of 1 provides noticeable advantages. However, due to the influence of weak pairing, the generated results exhibit lower content consistency. The IoD metric evaluates the difference in total fluorescence intensity between the test set and the adjacent real slices. Ideally, combining IoD with R_{avg} can provide insights into the similarity of virtual IHC staining to its real pathological diagnostic significance. As shown in Figure 4, although supervised learning methods like Pix2Pix achieve high pathological semantic relevance, virtual IHC generated by these methods often suffers from noticeable artifacts and

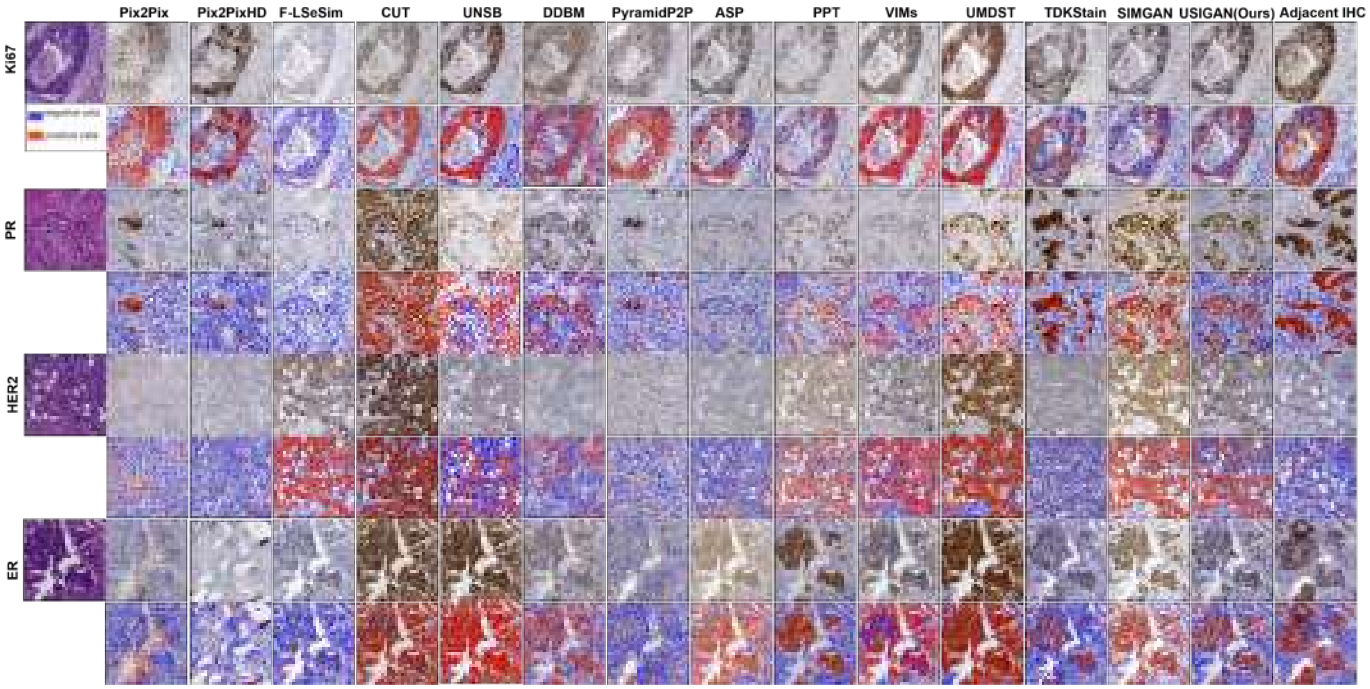


Fig. 4. Selected representative methods exhibit varying performances in virtual IHC staining results visualization on the MIST dataset. The quantitative comparison on different sate-of-art methods. Cell Segmentation and classification is performed using DeepLIIF [6] as follow by [32]

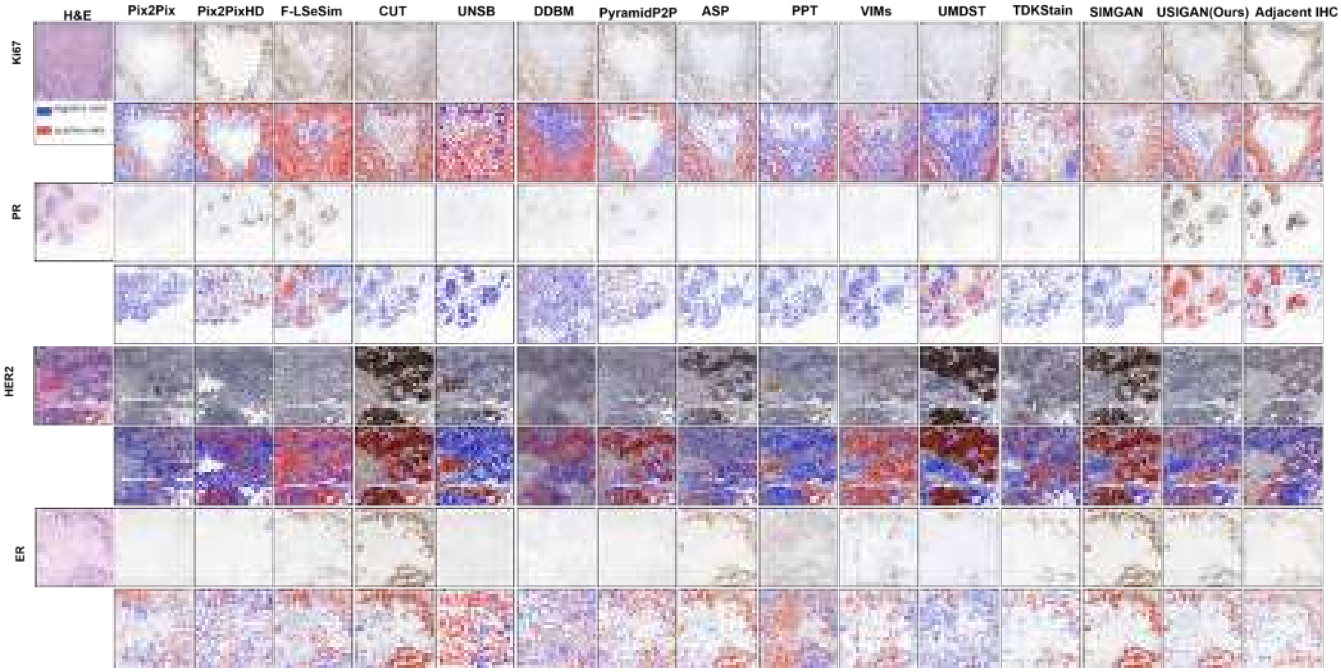


Fig. 5. Selected representative methods exhibit varying performances in virtual IHC staining results visualization on the IHC4BC dataset. The quantitative comparison on different sate-of-art methods. Cell Segmentation and classification is performed using DeepLIIF [6] as follow by [32]

distortions, which limit their practical applicability. Therefore, we believe that R_{avg} provides a more comprehensive evaluation of their clinical value. On the MIST dataset, our method outperforms existing approaches in perceptual and pathological semantic metrics, demonstrating superior performance and clinical relevance.

2) Results on IHC4BC: Table III presents the quantitative experimental results on the IHC4BC dataset. It is important to note that in the PR subset of the IHC4BC dataset, the pathological images exhibit low contrast in non-positive regions, appearing almost transparent. This makes it easier to distinguish positive regions based on optical density (OD) values. Due to this characteristic, the significant contrast in

TABLE II
COMPARISON OF VIRTUAL STAINING PERFORMANCE ON **MIST** DATASET. THE BEST SCORES ARE IN **BOLD** AND SECOND SCORES ARE IN UNDERLINE, RESPECTIVELY. THE METRICA/METRICB INDICATES THE (BATCH SIZE SAME WITH USI-GAN/BATCH SIZE SET AS 1 FOR GAN-BASED METHOD) RESULTS.

Method	HER2							ER						
	Quality (reference)		Pathology		Perception			Quality (reference)		Pathology		Perception		
	SSIM	PSNR	$R_{avg}\uparrow$	IoD$^{\times 10^7}$	FID↓	DISTS↓	PHV$_{avg}$ ↓	SSIM	PSNR	$R_{avg}\uparrow$	IoD$^{\times 10^7}$	FID↓	DISTS↓	PHV$_{avg}$ ↓
GAN-based Image-to-Image Translation SOTA														
Pix2Pix [14]	0.2022 / 0.2224	14.8224 / **15.244**	37.44 / 40.29	-5.6092 / -4.4196	107.40 / 73.32	0.2999 / 0.2912	0.5164 / 0.5260	0.2329 / **0.2535**	14.8243 / **15.4573**	42.39 / 44.97	-6.5789 / -5.646	87.29 /77.33	0.3393 / 0.3236	0.5565 /0.5889
Pix2PixHD [40]	**0.2485** / 0.2093	14.6117 / 14.5924	25.54 / 40.15	-3.6936 / -3.8803	87.51 / 68.48	0.3156 / 0.2899	0.5665 / 0.4752	0.2457 / 0.2419	14.2915 / 14.7984	30.04 / 41.48	-6.7141 / -5.3618	97.68 / 68.61	0.3630 / 0.3129	0.5643 / 0.5072
CycleGAN [59]	0.1805 / 0.2008	13.1507 / 13.6588	58.77 / 60.36	-1.0736 / -0.1638	54.65 / 53.97	0.26681 / 0.2629	0.5053 / 0.5053	0.2191 / 0.2050	13.4436 / 13.1596	63.57 / 63.32	-1.1561 / **-0.8555**	37.44 / 39.99	0.2665 / 0.2655	0.4992 / 0.4976
F-LseSim [57]	0.1813 / 0.1665	13.2261 / 12.5903	58.27 / 61/67	-3.1648 / 1.1324	51.10 / 50.70	0.2685 / 0.2810	0.5335 / 0.5431	0.2132 / 0.1802	13.9336 / 13.2137	65.38 / 54.76	-9.1023 / -4.4095	143.57 / 84.70	0.2903 / 0.2869	0.5451 / 0.5759
CUT [31]	0.1453/ 0.1698	11.6753 / 12.4523	57.78 / 56.90	10.9532 / 5.435	69.55 / 52.55	0.3000 / 0.2845	0.5712 / 0.5405	0.1999 / 0.1954	13.2221 / 13.5516	63.44 / 60.52	2.1214 / 6.0702	59.95 / 79.75	0.2942 / 0.3006	0.5474 / 0.5674
MoNCE [51]	0.1525 / 0.164	11.9814 / 12.082	56.44 / 54.86	7.2754 / 6.4068	60.65 / 61.58	0.2959 / 0.2893	0.5652 / 0.5429	0.1886 / 0.1818	12.6758 / 12.6003	65.86 / 68.88	6.1803 / 2.7424	59.99 / 67.79	0.2870 / 0.2778	0.5430 / 0.5541
QSAttn [13]	0.1492 / 0.157	11.8613 / 11.8377	55.67 / 56.00	8.1264 / 9.0991	65.11 / 63.99	0.2925 / 0.2921	0.5632 / 0.5531	0.1808 / 0.2034	12.4868 / 13.026	65.79 / 58.89	6.944 / 5.0241	71.65 / 83.37	0.2957 / 0.3004	0.5701 / 0.5644
NEGCUT [44]	0.1534 / 0.1234	12.2013 / 10.6738	58.69 / 67.36	2.6799 / -2.3618	53.39 / 53.52	0.2956 / 0.2554	0.5181 / 0.5139	0.1874 / 0.1725	12.6041 / 11.9798	65.57 / 69.33	6.5302 / 4.7934	54.24 / 63.41	0.2830 / 0.3019	0.5482 / 0.5854
SRC [16]	- / 0.1639	- / 12.4392	- / 44.46	- / 0.9507	- / 56.43	- / 0.3053	- / 0.5719	- / 0.1741	- / 12.1321	- / 66.87	- / 5.3383	- / 55.46	- / 0.2930	- / 0.5615
DECENT [46]	0.1719 / 0.1651	12.4824 / 12.4744	50.34 / 54.38	2.1423 / 4.6545	58.23 / 62.36	0.2987 / 0.2994	0.5673 / 0.5701	0.2044 / 0.1961	13.3857 / 13.076	65.97 / 61.81	2.7814 / 3.4981	68.77 / 61.81	0.2881 / 0.2863	0.5558 / 0.5546
EnCo [2]	0.1604 / 0.1558	12.1925 / 12.1163	52.54 / 51.84	4.7079 / 7.279	84.43 / 72.53	0.2975 / 0.3050	0.5734 / 0.5736	0.1739 / 0.1788	12.6444 / 13.0233	59.72 / 63.51	1.4499 / 2.1424	67.58 / 61.33	0.2912 / 0.2856	0.5707 / 0.5567
Diffusion-based Image-to-Image Translation SOTA														
UNSB [18]	0.128 / -	13.2194 / -	58.36 / -	-1.0456 / -	48.77 / -	0.2695 / -	0.5833 / -	0.1572 / -	13.3287 / -	62.78 / -	-9.7522 / -	34.49 / -	0.2644 / -	0.5603 / -
DDBM [58]	0.1891 / -	14.6195 / -	14.87 / -	-7.3041 / -	188.51 / -	0.3442 / -	0.5901 / -	0.2094 / -	14.5504 / -	28.92 / -	-7.0279 / -	180.32 / -	0.3383 / -	0.6001 / -
ControlNet [54]	0.1545 / -	8.6484 / -	-1.74 / -	9.6394 / -	122.23 / -	0.3846 / -	0.6744 / -	0.145 / -	8.376 / -	5.036 / -	24.6374 / -	154.64 / -	0.3445 / -	0.6351 / -
VIMs[4]	0.1595 / -	13.3339 / -	28.75 / -	-2.5329 / -	116.25 / -	0.3084 / -	0.6257 / -	0.191 / -	13.7069 / -	44.52 / -	-3.9112 / -	111.54 / -	0.3014 / -	0.6144 / -
Virtual Staining SOTA														
PyramidP2P [26]	0.1944 / 0.208	14.5908 / 15.1779	38.60 / 41.73	-5.8272 / -3.7689	103.07 / 78.89	0.2764 / 0.2814	0.5078 / 0.5185	0.2236 / 0.2346	14.7975 / 15.2438	43.73 / 46.54	-5.5017 / -5.4274	91.71 / 74.46	0.3066 / 0.2989	0.5432 / 0.5158
PPT [56]	0.2008 / 0.2096	14.3922 / 14.6015	55.98 / 38.99	-4.6784 / -3.4225	54.38 / 62.04	0.2884 / 0.3133	0.6194 / 0.5425	0.2086 / 0.1517	13.9607 / 12.5299	58.13	-4.522	50.44 / 77.14	0.3124 / 0.2959	0.5311 / 0.5781
UMDST [24]	- / 0.2185	- / 12.6884	- / 49.46	- / **-0.0378**	- / 70.41	- / 0.3124	- / 0.6045	- / 0.253	- / 13.6299	- / 63.22	- / 3.6211	- / 44.52	- / 0.3123	- / 0.5165
ASP [19]	0.2025 / 0.1632	14.6295 / 12.9524	76.31 / 57.15	-4.7957 / 0.91334	47.86 / 63.53	0.2792 / 0.2844	0.5020 / 0.5516	0.2143 / 0.1954	14.1445 / 13.5516	74.96 / 72.74	-3.1462 / -2.3544	41.40 / 55.68	0.2646 / 0.2716	0.5071 / 0.5213
TDKStain [19]	0.1957 / 0.2025	14.5065 / 14.4854	44.41 / 45.71	-1.8509 / -1.8614	58.64 / 65.99	0.2465 / 0.2531	0.4718 / 0.4759	0.213 / 0.2205	14.392 / 14.6813	49.66 / 45.88	-2.6782 / -4.4095	48.93 / 73.01	0.2398 / 0.2860	0.4755 / 0.4995
SIMGAN [10]	0.1864	13.6776	73.73	-1.4421	39.66	0.2608	0.5093	0.2165	13.9746	/ 78.27	-1.4809	34.61	0.2532	0.4977
USI-GAN(ours)	0.1871	14.0592	**83.35**	-1.6788	**37.76**	**0.2342**	**0.4708**	0.2019	13.7622	**85.07**	**-0.5109**	**33.06**	**0.2338**	**0.4694**

Method	Ki67							PR						
	Quality (reference)		Pathology		Perception			Quality (reference)		Pathology		Perception		
	SSIM	PSNR	$R_{avg}\uparrow$	IoD$^{\times 10^7}$	FID↓	DISTS↓	PHV$_{avg}$ ↓	SSIM	PSNR	$R_{avg}\uparrow$	IoD$^{\times 10^7}$	FID↓	DISTS↓	PHV$_{avg}$ ↓
GAN-based Image-to-Image Translation SOTA														
Pix2Pix [14]	0.2397 / 0.2766	14.6725 / **15.8554**	37.14 / 41.02	-2.9183 / -3.5921	97.35 / 87.33	0.2667 / 0.3296	0.4885 / 0.5693	0.2321 / 0.2936	14.8758 / **15.8001**	43.27 / 48.29	-6.5099 / 0.8295	81.88 / 93.51	0.3016 / 0.3534	0.5181 / 0.6001
Pix2PixHD [40]	0.2483 / 0.2435	14.8874/ 14.4045	23.51 / 38.70	-6.2699 / -0.5209	369.76 / 65.69	0.4193 / 0.2729	0.6437 / 0.4799	**0.3658** / 0.2763	13.3915 / 15.0074	40.26 / 41.90	-6.5009 / -5.7457	197.80 / 91.64	0.5374 / 0.3420	0.7480 / 0.5600
CycleGAN [59]	0.2294 / 0.2317	13.9732 / 14.0034	58.08 / 57.70	-0.41 / -0.9368	33.97 / 33.54	0.2433 / 0.2444	0.4747 / 0.4786	0.2132 / 0.2139	13.3125 / 13.3928	66.23 / 64.00	-0.4074 / -0.4639	40.21 / 40.38	0.2656 / 0.2666	0.4945 / 0.4976
F-LseSim [57]	0.2349 / 0.191	13.9824 / 12.7051	48.06 / 57.51	-3.8521 / 7.3051	66.63 / 67.47	0.2643 / 0.2810	0.5553 / 0.5399	0.1961 / 0.1848	13.174 / 13.344	63.74 / 70.69	-10.355 / -3.6976	103.67 / 58.28	0.2910 / 0.2827	0.5726 / 0.5551
CUT [31]	0.2174 / 0.2011	13.7533 / 12.7791	64.87 /65.47	3.9548 / 7.7662	54.42 / 59.37	0.2665 / 0.2756	0.5267 / 0.5418	0.205 / 0.2012	12.9842 / 12.7852	57.38 / 53.53	4.2059 / 4.2655	59.05 / 60.76	0.2966 / 0.2958	0.5585 / 0.5582
MoNCE [51]	0.2188/ 0.2158	13.3704 / 13.1576	66.26 / 61.89	4.2888 / 4.1535	55.54 / 56.72	0.2669 / 0.2696	0.5446 / 0.5467	0.2366 / 0.2216	14.2132 / 13.5972	69.28 / 67.89	-1.0985 / 1.9259	48.56 / 42.50	0.2806 / 0.2778	0.5322 / 0.5271
QSAttn [13]	0.1985 / 0.2122	13.437 / 13.4907	60.18 / 66.02	3.4735 / 4.4797	49.45 / 58.86	0.2712 / 0.2749	0.5468 / 0.5413	0.1984 / 0.2000	12.6366 / 12.3894	60.09 / 59.93	7.0856 / 7.5200	69.78 / 73.58	0.2968 / 0.3054	0.5633 / 0.5684
NEGCUT [44]	0.1924 / 0.2158	12.4626 / 13.2313	45.37 / 63.51	7.0443 / 5.1103	67.22 / 59.17	0.2672 / 0.2801	0.5399 / 0.5650	0.2086 / 0.2065	13.2525 / 12.9201	69.26 / 67.38	1.5486 / 3.4305	53.659 / 54.48	0.2847 / 0.2889	0.5514 / 0.5574
SRC [16]	- / 0.2286	- / 13.7734	- / 63.60	- / 1.582	- / 61.81	- / 0.2671	- / 0.5359	- / 0.1984	- / 12.6936	- / 72.75	- / 3.4875	- / 60.31	- / 0.2814	- / 0.5449
DECENT [46]	0.2368 / 0.2223	14.1949 / 14.1141	62.47 / 62.61	1.1197 / 0.9440	54.27 / 60.58	0.2646 / 0.2708	0.5325 / -0.5473	0.2072 / 0.2098	13.3525 / 13.462	68.75 / 66.16	1.7774 / 0.8075	56.39 / 61.50	0.2821 / 0.2905	0.5542 / 0.5585
EnCo [2]	0.1954 /0.2302	13.871 / 14.0677	60.09 / 59.62	1.106 / 1.5699	55.56 / 67.47	0.2654 / 0.2602	0.5472 / 0.5399	0.1871 / 0.1927	12.471 / 12.7945	65.71 / 64.66	3.3781 / 3.2650	81.47 / 77.99	0.2900 / 0.2924	0.5574 / 0.5646
Diffusion-based Image-to-Image Translation SOTA														
UNSB [18]	0.1765 / -	14.052 / -	60.30 / -	-8.0121 / -	39.29 / -	0.2527 / -	0.5684 / -	0.1659 / -	13.6937 / -	61.68 / -	-10.8229 / -	34.90 / -	0.2793 / -	0.5847
DDBM [58]	0.2085 / -	14.0552 / -	34.81 / -	2.2744 / -	150.88 / -	0.2949 / -	0.5773 / -	0.2025 / -	13.3431 / -	41.17 / -	-6.6537 / -	163.58 / -	0.3094 / -	0.5879
ControlNet [54]	0.2744 / -	10.2626 / -	13.48 / -	16.4868 / -	282.97 / -	0.3831 / -	0.6683 / -	0.1382 / -	9.0184 / -	5.944 / -	8.3654 / -	134.27 / -	0.3670 / -	0.6537 / -
VIMs[4]	0.1975 / -	14.1509 / -	34.66 / -	-1.354 / -	97.20 / -	0.2986 / -	0.6107 / -	0.1941 / -	13.9057 / -	47.37 / -	-4.3369 / -	129.18 / -	0.3090 / -	0.6188 / -
Virtual Staining SOTA														
PyramidP2P [26]	0.2823 / **0.2846**	15.2563 / 15.7009	35.47 / 41.47	-2.2201 / -2.9104	111.51 / 78.72	0.3203 / 0.3017	0.5766 / 0.5494	0.2540 / 0.2455	15.2238 / 15.4207	46.38 / 49.32	-5.5563 / -4.5925	92.63 / 76.69	0.3053 / 0.3042	0.5465 / 0.5432
PPT [56]	0.2663 / 0.2516	14.9242 / 15.5841	51.97 / 52.35	-2.5072 / -3.8480	48.64 / 119.88	0.2788 / 0.3391	0.5241 / 0.6085	0.2256 / 0.182	14.3989 / 13.3685	63.20 / 55.76	-6.2238 / -1.6585	50.31 / 65.43	0.2932 / 0.2966	0.5232 / 0.5595
UMDST [24]	- / 0.2256	- / 12.3384	- / 18.94	- / -0.787	- / 64.31	- / 0.3562	- / 0.6478	- / 0.2518	- / 13.6979	- / 67.73	- / -2.5224	- / 58.40	- / 0.3040	- / 0.5987
ASP [19]	0.2163 / 0.2123	14.525 / 14.1183	71.27 / 66.29	-1.8027 / 0.3118	44.89 / 74.11	0.2464 / 0.2561	0.4922 / 0.5268	0.2178 / 0.1848	14.3087 / 13.344	74.34 / 70.69	-5.19 / -3.6976	41.59 / 58.28	0.2554 / 0.2827	0.4924 / 0.5551
TDKStain [33]	0.2422 / 0.2525	15.0263 / 15.0208	42.58 / 42.26	**-0.1888** / -2.7421	61.98 / 64.80	0.2398 / 0.2638	0.4892 / 0.4900	0.2313 / 0.2279	14.7935 / 14.6793	48.38 / 47.76	-3.2448 / -3.8214	57.90 / 64.18	0.2601 / 0.2656	0.4929 / 0.4873
SIMGAN [10]	0.2322	14.2495	73.17	-0.6221	28.52	0.2483	0.5022	0.2071	13.8286	85.16	**-0.2687**	35.87	0.2521	0.4995
USI-GAN(ours)	0.2317	14.4027	**77.75**	-0.2680	**27.36**	**0.2351**	**0.4702**	0.2163	14.2176	**87.20**	-2.0699	**34.64**	**0.2339**	**0.4560**

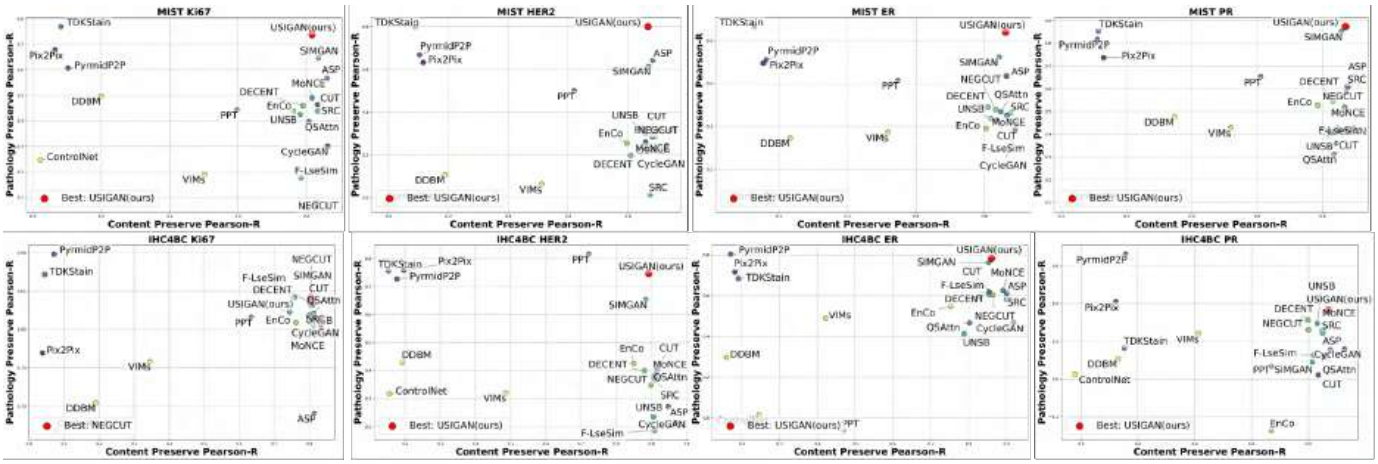


Fig. 6. The visualization results of pathological semantic consistency and content preservation across different methods demonstrate that our approach exhibits outstanding performance in both aspects.

positive regions of the PR subset makes global features, such as boundary definition and contrast intensity, more critical during evaluation.

TABLE III
COMPARISON OF VIRTUAL STAINING PERFORMANCE ON **IHC4BC** DATASET. THE BEST SCORES ARE IN **BOLD** AND SECOND SCORES ARE IN UNDERLINE, RESPECTIVELY. THE METRICA/METRICB INDICATES THE (BATCH SIZE SAME WITH USI-GAN/BATCH SIZE SET AS 1 FOR GAN-BASED METHOD) RESULTS.

Method	HER2							ER						
	Quality (reference)		Pathology		Perception			Quality (reference)		Pathology		Perception		
	SSIM	PSNR	$R_{avg}\uparrow$	IoD$^{\times 10^7}$	FID↓	DISTS↓	PHV$_{avg}$ ↓	SSIM	PSNR	$R_{avg}\uparrow$	IoD$^{\times 10^7}$↓	FID↓	DISTS↓	PHV$_{avg}$ ↓
GAN-based Image-to-Image Translation SOTA														
Pix2Pix [14]	0.1248 / **0.1619**	11.37 / **12.5642**	47.68 / 46.80	-7.1422 / -9.711	93.70 / 173.59	0.2855 / 0.4515	0.5242 / 0.6890	0.4628 / 0.4695	20.5881 / 20.3643	44.96 / 39.67	-1.9269 / -2.0841	102.54 / 149.93	0.2924 / 0.2938	0.4537 / 0.4744
Pix2PixHD [40]	0.143 / 0.1473	11.2992 / 10.9458	42.34 / 38.46	-5.9182 / -10.4474	108.60 / 148.99	0.3754 / 0.3667	0.5946 / 0.6052	0.502 / 0.4757	**20.8655** / 20.2985	37.87 / 41.67	-2.1063 / -1.0089	76.88 / 44.79	0.3384 / 0.2428	0.5061 / 0.4265
CycleGAN [59]	0.128 / 0.1269	10.9497 / 10.8503	59.13 / 61.80	-9.5066 / -7.5383	53.60 / 49.80	0.2452 / 0.2629	0.4687 / 0.4716	0.484 / 0.4615	19.6816 / 18.5512	69.39 / 73.18	-0.4863 / 0.9928	39.04 / 39.94	0.2276 / 0.2654	0.3859 /0.3888
F-LseSim [57]	0.1275 / 0.1243	10.8599 / 10.7016	54.81 / 58.56	-13.4688 / 0.2620	90.63 / 71.41	0.2647 / 0.2568	0.5324 / 0.5237	0.4391 / 0.4175	18.3421 / 18.1839	73.00 / 76.88	0.1354 / 1.0490	61.17 / 56.97	0.2667 / 0.2739	0.4683 / 0.4870
CUT [31]	0.1214 / 0.1195	10.6364 / 10.338	66.00 / 63.32	-2.7794 / 3.0142	53.14 / 51.73	0.2579 / 0.2629	0.5186 / 0.5364	0.413 / 0.4396	17.9916 / 18.6876	73.63 / 64.06	1.7072 / 0.2348	39.27 / 38.29	0.2442 / 0.2763	0.4589 / 0.4575
MoNCE [51]	0.1175 / 0.1211	10.3727 / 10.6194	64.73 / 68.16	-1.3287 / -1.7506	60.06 / 53.57	0.2556 / 0.2529	0.5167 / 0.5160	0.4286 / 0.4242	17.9131 / 18.2314	75.81 / 68.88	1.6675 / 1.7383	46.64 / 44.81	0.2619 / 0.2622	0.4612 / 0.4695
QSAttn [13]	0.1228 / 0.1194	10.8112 / 10.6096	64.11 / 63.19	-2.5190 / -2.0231	55.65 / 57.57	0.2584 / 0.2659	0.5182 / 0.5299	0.442 / 0.4644	18.5162 / 18.7556	63.48 / 67.34	0.1457 / 0.4662	39.35 / 40.83	0.2800 / 0.2659	0.4637 / 0.4472
NEGCUT [44]	0.1204 / 0.1234	10.603 / 10.6738	62.32 / 68.16	-3.2282 / -2.3618	57.79 / 55.51	0.2648 / 0.2554	0.5288 / 0.5139	0.4056 / 0.4042	17.5892 / 16.8981	73.37 / 68.55	1.6634 / 2.8059	48.91 / 54.91	0.2700 / 0.2842	0.4986 / 0.5023
SRC [16]	- / 0.1229	- / 10.7725	- / 63.87	- / -2.6529	- / 54.74	- / 0.2630	- / 0.5318	- / 0.4341	- / 18.3454	- / 74.09	- / 2.0176	- / 40.91	- / 0.2568	- / 0.4595
DECENT [46]	0.1202 / 0.1246	10.9916 / 10.8228	63.99 / 66.66	-6.319 / -2.6494	57.66 / 56.55	0.2497 / 0.2704	0.4986 / 0.5407	0.4227 / 0.4063	18.211 / 17.7322	71.16 / 72.69	-0.8704 / 1.6861	46.79 / 51.41	0.2651 / 0.2728	0.4711 / 0.4966
EnCo [2]	0.1173 / 0.119	10.4794 / 10.7763	63.76 / 65.19	-6.299 / -4.9355	71.14 / 59.60	0.2770 / 0.2689	0.5540 / 0.5556	0.4063 / 0.433	18.2537 / 18.5431	65.07 / 69.80	0.2648 / 0.2673	57.16 / 43.80	0.2719 / 0.2673	0.4946 / 0.4758
Diffusion-based Image-to-Image Translation SOTA														
UNSB [18]	0.1087 / -	11.1355 / -	56.99 / -	-17.3059 / -	49.04 / -	0.2568 / -	0.5809 / -	0.4020 / -	19.4709 / -	60.03 / -	-3.2319 / -	55.30 / -	0.2617 / -	0.5003 / -
DDBM [58]	0.1067 / -	11.6096 / -	31.14 / -	-11.335 / -	198.794 / -	0.3950 / -	0.6627 / -	0.4529 / -	20.3453 / -	22.93 / -	-3.0707 / -	279.76 / -	0.3612 / -	0.5676 / -
ControlNet [54]	0.0983 / -	7.8967 / -	23.63 / -	15.6894 / -	140.56 / -	0.3718 / -	0.6902 / -	0.2639 / -	7.4072 / -	13.03 / -	39.5472 / -	232.41 / -	0.4038 / -	0.7062 / -
VIMs[4]	0.1272 / -	11.4642 / -	40.34 / -	-12.5726 / -	81.40 / -	0.3084 / -	0.6164 / -	0.4239 / -	17.7305 / -	45.63 / -	-1.4041 / -	89.95 / -	0.3014 / -	0.5134
Virtual Staining SOTA														
PyramidP2P [26]	0.1162 / 0.1297	11.1824 / 11.5591	45.24 / 47.26	-8.0986 / -7.8924	92.52 / 74.99	0.2776 / 0.2869	0.5095 / 0.5204	0.4603 / 0.4462	20.675 / 20.6914	48.78 / 47.93	-1.6351 / -1.6706	73.05 / 63.48	0.2725 / 0.2681	0.4364 / 0.4297
PPT [56]	0.1339 / 0.1309	11.3801 / 11.3499	68.78 / 50.50	-8.8844 / -13.0090	56.95 / 61.96	0.2597 / 0.2992	0.5298 / 0.5469	0.3476 / 0.3268	16.5275 / 15.7775	20.36 / 30.08	0.8693 / 1.6747	70.91 / 196.25	0.3037 / 0.3399	0.5376 / 0.6001
UMDST [24]	- / 0.1479	- / 11.1785	- / 65.40	- / -4.1051	- / 63.10	- / 0.3065	- / 0.6403	- / **0.5418**	- / 20.5546	- / 54.60	- / -1.5868	- / 62.68	- / 0.3360	- / 0.5611
ASP [19]	0.1202 / 0.1219	10.9916 / 11.1032	61.00 / 66.95	-9.9207 / -12.4642	66.75 / 69.59	0.2618 / 0.2582	0.5105 / 0.4978	0.4688 / 0.3929	19.5896 / 16.6227	75.66 / 76.03	-0.2174 / 2.6554	39.36 / 66.36	0.2197 / 0.2785	**0.3751** / 0.4975
TDKStain [33]	0.1275 / 0.1216	10.9003 / 11.1563	45.43 / 44.62	-6.645 / 1.1292	66.39 / 92.79	0.2604 / 0.2512	0.5082 / 0.5040	0.4737 / 0.4788	19.5718 / 20.4984	43.83 / 43.63	-0.8196 / -1.2804	56.46 / 59.66	0.2517 / 0.2559	0.4344 / 0.4304
SIMGAN [10]	0.1200	10.7020	76.90	**-0.1793**	58.32	0.2522	0.5196	0.4177	17.7036	80.79	2.0767	37.93	0.2357	0.4217
USI-GAN(ours)	0.1267	11.1342	**81.98**	-7.3421	**48.54**	**0.2345**	**0.4611**	0.452	19.4773	**82.25**	**0.0869**	37.01	**0.2139**	0.3761
Method	Ki67							PR						
	Quality (reference)		Pathology		Perception			Quality (reference)		Pathology		Perception		
	SSIM	PSNR	$R_{avg}\uparrow$	IoD$^{\times 10^7}$↓	FID↓	DISTS↓	PHV$_{avg}$ ↓	SSIM	PSNR	$R_{avg}\uparrow$	IoD$^{\times 10^7}$	FID↓	DISTS↓	PHV$_{avg}$ ↓
GAN-based Image-to-Image Translation SOTA														
Pix2Pix [14]	0.4097 / 0.4309	20.3587 / **20.8746**	40.44 / 40.13	-0.8362 / 1.0083	102.77 / 93.23	0.2837 / 0.2973	0.4887 / 0.5263	0.5038 / **0.6096**	23.083 / 24.9063	26.61 / 26.05	-0.475 / -0.8543	215.53 / 154.69	0.3072 / 0.3622	0.4837 / 0.5701
Pix2PixHD [40]	**0.4279** / 0.4155	20.3387 / 19.8854	44.45 / 48.25	1.3103 / 1.8992	54.70 / 37.54	0.2586 / 0.2328	0.4830 / 0.4253	0.5664 / 0.5368	24.1064 / 23.4964	31.14 / 46.56	-1.1365 / -0.9227	71.70 / 51.58	0.3114 / 0.2570	0.4821 / 0.4182
CycleGAN [59]	0.39 / 0.3911	19.2125 / 19.1552	81.94 / 82.35	0.0964 / 2.1075	39.47 / 25.74	0.2073 / 0.2444	0.3848 / **0.3807**	0.5546 / 0.565	23.2907 / 23.4863	54.22 / 49.74	-0.5593 / -0.8913	43.98 / 37.25	**0.2331** / 0.2666	0.3647 / **0.3625**
F-LseSim [57]	0.3856 / 0.3725	19.1239 / 19.1106	80.94 / 58.86	0.4046 / 2.377	37.69 / 32.89	0.2499 / 0.2538	0.4821 / 0.4762	0.4976 / 0.5269	20.4933 / 22.6819	47.18 / 43.12	0.4286 / -0.6384	68.77 / 50.43	0.3011 / 0.27788	0.4649 / 0.4450
CUT [31]	0.3511 / 0.3738	19.0078 / 19.5456	82.11 / 81.82	0.4491 / 2.2578	39.05 / 32.64	0.2424 / 0.2250	0.4841 / 0.4543	0.5452 / 0.5151	23.3178 / 22.3619	42.81 / 45.85	-0.7435 / -0.2942	42.29 / 36.19	0.2646 / 0.2661	0.4095 / 0.4286
MoNCE [51]	0.3489 / 0.3571	18.5064 / 18.819	80.84 / 81.16	1.2873 / 2.4970	47.04 / 33.94	0.2495 / 0.2403	0.4955 / 0.4671	0.4845 / 0.4784	18.9298 / 21.3643	55.40 / 55.56	1.9725 / 0.6368	72.49 / 38.38	0.3174 / 0.2704	0.4805 / 0.4602
QSAttn [13]	0.384 / 0.355	19.8617 / 19.1989	81.69 / 81.86	-0.2352 / 2.5348	36.89 / 44.19	0.2311 / 0.2353	0.4658 / 0.4711	0.5449 / 0.5168	23.3068 / 22.571	48.79 / 57.81	-0.3421 / **-0.0485**	37.01 / 35.89	0.2540 / 0.2501	0.4119 / 0.4146
NEGCUT [44]	0.3528 / 0.3663	18.8025 / 19.3018	81.45 / 77.93	0.3715 / 2.1969	41.86 / 33.51	0.2410 / 0.2386	0.4829 / 0.4808	0.5095 / 0.4667	22.037 / 20.348	53.04 / 41.10	0.3156 / 0.7382	39.94 / 37.75	0.2612 / 0.2957	0.4466 / 0.47
SRC [16]	- / 0.3489	- / 18.6581	- / 81.47	- / 0.697	- / 37.36	- / 0.2369	- / 0.4797	- / 0.4989	- / 21.1762	- / 54.68	- / 0.8948	- / 37.46	- / 0.2750	- / 0.4454
DECENT [46]	0.3718 / 0.3602	19.2342 / 19.2451	80.75 / 78.65	0.1254 / 2.1868	30.70 / 33.08	0.2422 / 0.2442	0.4860 / 0.4942	0.5466 / 0.5104	23.2836 / 22.0259	55.48 / 54.28	-0.7105 / -0.4417	40.26 / 40.42	0.2450 / 0.2659	0.3791 / 0.4529
EnCo [2]	0.3917 / 0.3751	19.4617 / 19.158	78.53 / 79.05	2.1865 / 2.4694	38.648 / 38.17	0.2341 / 0.2377	0.46315 / 0.4704	0.5038 / 0.5113	20.8565 / 22.5336	19.83 / 43.42	-0.3983 / -0.4417	66.36 / 53.41	0.3330 / 0.2643	0.5234 / 0.4420
Diffusion-based Image-to-Image Translation SOTA														
UNSB [18]	0.3481 / -	20.5063 / -	78.34 / -	-1.9469 / -	31.81 / -	0.2109 / -	0.4611 / -	0.511 / -	23.7755 / -	56.31 / -	-1.7314 / -	37.01 / -	0.2442 / -	0.4502 / -
DDBM [58]	0.3567 / -	18.8877 / -	44.71 / -	2.7059 / -	109.45 / -	0.3013 / -	0.5659 / -	0.5043 / -	22.293 / -	11.78 / -	-1.0088 / -	234.86 / -	0.3171 / -	0.5499 / -
ControlNet [54]	0.2035 / -	8.7388 / -	37.32 / -	10.4722 / -	246.05 / -	0.4117 / -	0.7179 / -	0.2442 / -	8.2439 / -	0.093 / -	21.7282 / -	260.62 / -	0.4449 / -	0.7148 / -
VIMs[4]	0.4015 / -	20.8459 / -	55.17 / -	0.5972 / -	80.61 / -	0.2986 / -	0.5371 / -	0.573 / -	24.4844 / -	32.63 / -	-1.5848 / -	109.73 / -	0.3090 / -	0.4629 / -
Virtual Staining SOTA														
PyramidP2P [26]	0.39 / 0.4042	20.2169 / 20.3703	48.48 / 47.02	-0.7704 / 1.3461	57.24 / 55.18	0.2365 / 0.2581	0.4263 / 0.4666	0.5258 / 0.5225	23.9905 / 24.2178	41.01 / 35.67	-1.1295 / -1.1742	62.83 / 95.60	0.2576 / 0.2782	0.3951 / 0.4161
PPT [56]	0.3911 / 0.2825	19.5607 / 16.7756	72.49 / 62.48	0.91 / 2.7694	54.91 / 102.98	0.2211 / 0.2556	0.4539 / 0.5454	0.529 / 0.2968	23.3723 / 18.4435	36.83 / 16.55	-1.1287 / 0.2280	68.88 / 322.23	0.2676 / 0.3442	0.4240 / 0.6314
UMDST [24]	- / 0.4734	- / 20.1886	- / 43.21	- / 1.0701	- / 56.41	- / 0.2898	- / 0.5723	- / 0.5863	- / 22.5572	- / 48.92	- / -0.0899	- / 44.76	- / 0.3014	- / 0.5216
ASP [19]	0.3743 / 0.3893	19.1065 / 19.3626	75.22 / 74.95	0.2612 / 1.8991	35.86 / 34.31	0.2334 / 0.2233	0.4475 / 0.4347	0.5762 / 0.5654	24.4079 / 24.2786	51.61 / 53.16	-1.2598 / -1.2196	67.65 / 63.55	0.2522 / 0.2449	0.4137 / 0.3957
TDKStain [33]	0.4138 / 0.4006	20.0043 / 19.4149	45.86 / 46.03	-0.6335 / -8.5305	57.74 / 43.57	0.2411 / 0.2225	0.4185 / 0.4194	0.4998 / 0.5416	23.3719 / 23.3434	15.82 / 16.48	-1.1883 / -0.9537	107.13 /63.78	0.2590 / 0.2590	0.4329 / 0.4135
SIMGAN [10]	0.3408	18.8144	81.80	0.1395	38.37	0.2397	0.4934	0.5247	21.999	45.23	-0.3719	40.05	0.2718	0.4396
USI-GAN(ours)	0.3773	19.209	**82.23**	**0.009**	**23.74**	**0.2071**	0.4064	0.2163	14.2176	**61.80**	3.2447	**34.63**	0.3499	0.4559

Our method, which binds optical density (OD) to pathological semantic information, is highly sensitive to positive regions and tends to classify cells as positive. While this is advantageous in datasets with abundant positive regions, it poses a challenge in datasets with fewer abnormal regions. As shown in Figure 6, apart from supervised methods, the pathological semantic relevance of all methods is significantly lower on this dataset compared to others. Nevertheless, our method continues to demonstrate outstanding performance.

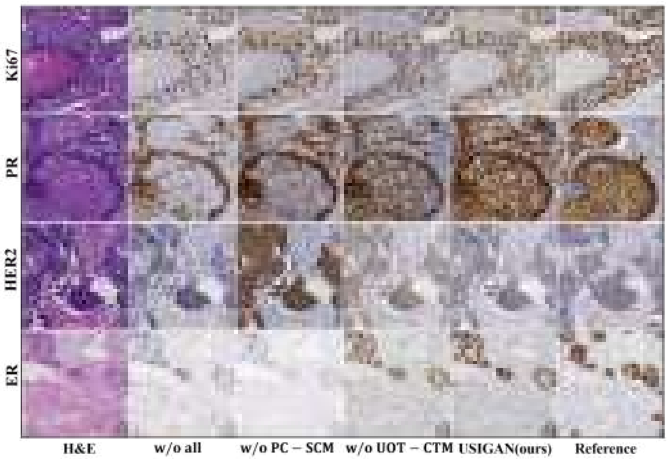


Fig. 7. Visualize the impact of UOT-CTM and PC-SCM on USIGAN on MIST dataset benckmark.

E. Ablation Study and Analysis

To investigate the effectiveness of the cyclic transfer strategies UOT-CTM and PC-SCM, we measured content consistency using the Pearson-R coefficient R_c, pathological semantic consistency using R_p, and overall clinical value using R_{avg}. As shown in Table V and Table VI, incorporating the UOT-CTM strategy represented by $\mathcal{L}_{tcyc}$ significantly improved content consistency, while the improvement in pathological semantic consistency was less pronounced. On the other hand, the PC-SCM mechanism substantially enhanced pathological semantic consistency but performed poorly in maintaining content consistency. By combining both modules, our method

TABLE IV
COMPARISON OF STAIN TRANSFER PERFORMANCE FOR DIFFERENT METHODS ON THE MIST BENCHMARK DATASETS. THE BEST AND SECOND-BEST SCORES ARE IN **BOLD** AND UNDERLINE, RESPECTIVELY. IOD,FID AND PHV$_{avg}$ (THE LOWER IS BETTER.)

Methods	Ki67						ER					
	IoD↓	R_c(%)↑	R_p(%)↑	R_{avg}(%)↑	FID↓	PHV$_{avg}$↓	IoD↓	R_c(%)↑	R_p(%)↑	R_{avg}(%)↑	FID↓	PHV$_{avg}$↓
Baseline	0.4895	77.86	36.34	57.10	27.45	0.5023	0.8598	80.71	42.18	61.44	35.27	0.5103
w/o UOT-CTM	0.2595	80.75	71.81	76.28	29.40	0.5024	1.8404	83.65	81.91	**82.78**	35.35	0.4662
w/o PC-SCM	**0.1915**	**82.70**	44.68	63.69	30.59	0.4977	**0.4649**	**86.35**	46.31	66.33	**32.84**	0.4991
USI-GAN (ours)	0.2680	81.63	**73.88**	**77.75**	**27.36**	**0.4702**	0.5109	86.06	**84.07**	**85.07**	33.06	0.4694

Methods	PR						HER2					
	IoD↓	R_c(%)↑	R_p(%)↑	R_{avg}(%)↑	FID↓	PHV$_{avg}$↓	IoD↓	R_c(%)↑	R_p(%)↑	R_{avg}(%)↑	FID↓	PHV$_{avg}$↓
Baseline	2.4654	82.06	25.24	53.65	37.97	0.5220	1.7549	85.32	32.80	59.06	42.14	0.5082
w/o UOT-CTM	1.6829	83.53	86.73	85.13	43.40	0.4932	**0.7819**	84.32	**82.94**	**83.63**	**37.39**	0.4728
w/o PC-SCM	**0.4736**	87.19	48.90	68.05	34.26	0.5150	1.5598	85.81	21.39	53.60	42.09	0.5247
USI-GAN (ours)	2.0700	**87.03**	**87.37**	**87.20**	34.64	**0.456**	1.6788	**86.63**	80.08	83.35	37.76	**0.4708**

TABLE V
THE ABLATION EXPERIMENTS ON MIST ER FOR ALL MODULES ARE CONDUCTED, WHERE $\mathcal{L}_{\text{TCYC}}$ REPRESENTS THE UOT-CTM MODULE, AND THE REMAINING COLUMNS REPRESENT THE PC-SCM MODULE.

$\mathcal{L}_{odc}$	$\mathcal{L}_{cc}$	$\mathcal{L}_{tcyc}$	R_c(%) ↑	R_p(%) ↑	R_{avg}(%) ↑	FID ↓	PHV$_{avg}$ ↓
–	–	–	80.71	42.18	61.44	35.27	0.5103
✓	–	–	84.06	81.54	82.80	35.16	0.4776
–	✓	–	82.63	75.55	79.09	**32.67**	0.4908
✓	✓	-	83.65	81.91	81.01	35.35	**0.4662**
–	–	✓	**86.35**	46.30	66.33	32.84	0.4991
–	✓	✓	85.88	70.85	78.36	33.05	0.4846
✓	–	✓	85.62	76.58	81.10	33.25	0.5049
✓	✓	✓	86.06	**84.07**	**85.07**	33.06	0.4694

TABLE VI
THE QUANTITATIVE COMPARISON RESULTS OF THE EFFECTS OF UOT AND UOT-CTM ON USIGAN. $\mathcal{L}_1$ W (U)OT REPRESENT CONSTRAIN TRANSPORT STRATEGY BETWEEN $T^{(H-I)}$ AND $T^{(I-I)}$, AND (U)OT REPRESENTS THE THEORY FOR COMPUTING THE COST MATRIX OF FEATURE MATCHING.

Strategy	R$_c$(%) ↑	R$_p$(%) ↑	IoD↓	FID↓	PHV$_{avg}$ ↓
baseline	80.71	42.18	0.8590	35.27	0.5103
$\mathcal{L}_1$ w OT	86.19	37.51	2.5027	33.64	0.5243
$\mathcal{L}_1$ w UOT	85.69	43.61	2.2763	33.03	0.5042
$\mathcal{L}_{tcyc}$(OT)	86.25	40.51	1.4186	33.23	0.5084
$\mathcal{L}_{tcyc}$(UOT)	**86.35**	**46.30**	**0.4640**	**32.84**	**0.4991**

TABLE VII
THE QUANTITATIVE RESULTS OF THE INFLUENCE OF DIFFERENT BATCH SIZE ON THE PERFORMANCE OF THE PC-SCM MECHANISM.

Batch size	R_c(%)	R_p(%)	IoD	FID	PHV$_{avg}$
1	84.58	23.61	1.9810	47.95	0.5337
2	84.19	57.95	0.8635	49.56	0.5136
4	85.23	57.11	2.1414	49.18	0.5302
8	86.14	57.90	2.6316	57.81	0.5245
16	82.79	61.70	4.0556	64.54	0.5480

leverages the advantages of each, achieving a balance between content and pathological semantic consistency compared to the baseline.

Additionally, to explore the effect of optical density anchoring on pathological semantics, we analyzed rows 2, 3, and 4 in Table V. The multi-level feature consistency loss $\mathcal{L}_{cc}$ improves

Fig. 8. Visualize the impact of UOT and transport consistency loss compared to the $\mathcal{L}_1$ loss that directly constrains both transport matrix.

TABLE VIII
THE QUANTITATIVE RESULTS OF THE DIFFERENT RELAX PARAMETER τ ON THE PERFORMANCE OF THE UOT-CTM MECHANISM ON MIST ER.

τ	R_c(%)	R_p(%)	IoD	FID	PHV$_{avg}$
0.0001	86.63	52.81	0.4639	57.39	0.5642
0.001	86.35	46.31	0.4649	32.84	0.4991
0.01	87.00	48.41	0.5813	59.18	0.5772
0.1	86.12	42.69	0.5827	58.66	0.5735

pathological representation consistency by leveraging features across multiple levels. However, due to the inconsistency of low-level features such as morphology and texture in weakly paired data, $\mathcal{L}_{cc}$ reintroduces the influence of weak pairing errors. In contrast, the optical density consistency loss $\mathcal{L}_{odc}$ directly anchors pathological semantics, showing a more significant improvement in pathological semantic consistency R_p compared to $\mathcal{L}_{cc}$, but at the cost of slightly lower image quality as measured by FID. Therefore, we adopted $\mathcal{L}_{cc}$ as an auxiliary term, combining it with $\mathcal{L}_{odc}$ in the PC-SCM strategy. This combination led to significant improvements in perceptual metrics, as illustrated by the visual results in

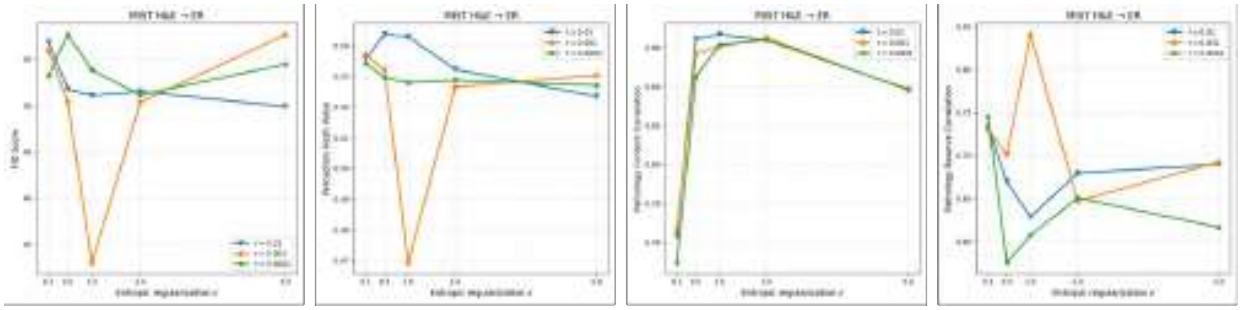


Fig. 9. Ablation study for the unbalanced optimal transport hyperparameters: entropic regularization ϵ and relaxation coefficient τ in the UOT-CTM module when calculating transport costs for MIST ER staining.

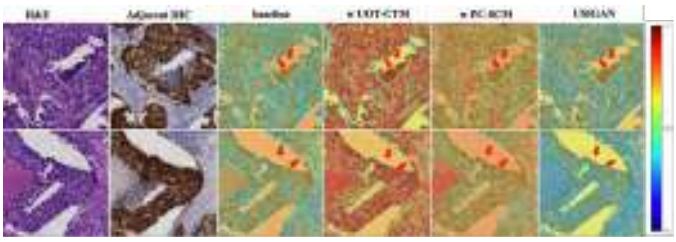


Fig. 10. The ablation study visualizations of model attention region changes demonstrate that models with self-information mining gradually enhance the contrast between attention to key foreground regions and the background.

Figure 7.

To further discuss the effectiveness of eliminating the impact of weakly paired data and the limitations of classical optimal transport (OT) in such scenarios, we analyzed the differences between directly constraining the transport matrices $T^{(H-I)}$ and $T^{(H-\hat{I})}$ using $\mathcal{L}_1$, without eliminating weakly paired terms, and employing the cyclic transfer strategy $\mathcal{L}_{tcyc}$. Additionally, we compared the differences in transport cost computation between unbalanced optimal transport (UOT) and classical OT. As shown in Figure 8 and Table VI, applying OT to compute transport costs without eliminating weakly paired terms resulted in a decrease in pathological semantic consistency R_p. Under cyclic transfer, using OT for transport cost computation also led to a lower R_p compared to the baseline, though it performed better than the strategy without eliminating weakly paired terms. When using UOT to compute transport costs, the cyclic transfer strategy achieved optimal results across all metrics. This demonstrates that the cyclic transfer loss $\mathcal{L}_{tcyc}$ can better learn both content and style information in the transport matrices, avoiding a trade-off between the two. Furthermore, UOT provides a more accurate transport cost matrix, enhancing the overall performance.

1) Sensitivity to batch size on PC-SCM: PC-SCM utilizes intra-batch correlation to guide the generated results in maintaining pathological consistency across batches. Therefore, the batch size influences the effectiveness of PC-SCM. To explore the impact of batch size, we set the image reading size to 256 and randomly crop it to 128 for easier requirement of memory. As shown in Table VII The Optical Density vectors, compressed through Focal Optical Density , exhibit increasing instability as the batch size grows. This phenomenon arises due to the weaker direct correlation between morphological features and the intensity values of optical density. As a result, relatively satisfactory results can be achieved at smaller batch sizes, where the impact of this instability is minimized. Overall, a batch size of 4 achieves the best performance by balancing the trade-off between morphological consistency and self-information mining. However, due to the higher computational cost associated with larger batch sizes, we adopt a batch size of 2 as our training configuration. This choice effectively reduces training overhead while maintaining a satisfactory level of performance.

2) Sensitivity to τ/ϵ on UOT-CTM: The UOT-CTM relies on Sinkhorn-Knopp entropy regularization and KL divergence to achieve optimal transport and relaxed transport boundaries. The relaxation coefficient τ and the entropy regularization parameter ϵ directly affect the quality of the transport cost matrix. The parameter ϵ adds a regularization term that impacts the smoothness of the dual variables in optimal transport. When ϵ is small, the transport plan is closer to the original cost matrix and is more sparse; however, when ϵ is large, the transport plan becomes smoother but may risk losing fine details. The parameter τ determines whether the target mass in the transport plan strictly matches the source mass. In weakly paired datasets with varying matching rates, the optimal τ can vary. When τ is too small, unbalanced optimal transport approaches traditional optimal transport, while excessively large τ may lead to numerical instability. Table VIII and Figure 9 illustrate the results of virtual staining with different τ and τ/ϵ combinations for the ER staining type on the MIST dataset. When ϵ is too large and τ is too small, the transport plan becomes overly sparse and numerically unstable. Conversely, when ϵ is too small and τ is too large, the transport plan is smooth but loses too much detail, resulting in poorer content consistency but good pathological semantic consistency. For all datasets, we selected $\tau = 0.001$ and $\epsilon = 1$ as the default settings. However, we believe that more optimal τ/ϵ combinations may exist for datasets with different matching rates.

F. Subjetive Evaluation

We invited three pathologists to conduct a double-blind subjective evaluation. For each staining type, we randomly selected 10 images for each method, resulting in a total of 400 images from the MIST and IHC4BC datasets. These images were presented in randomized order, showing the H&E image

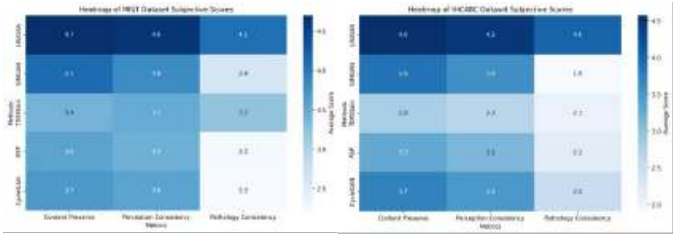


Fig. 11. The heatmap below represents the subjective evaluation scores of pathologists for Immunohistochemistry (IHC) images. These scores were assigned based on their assessment of diagnostic clinical value.

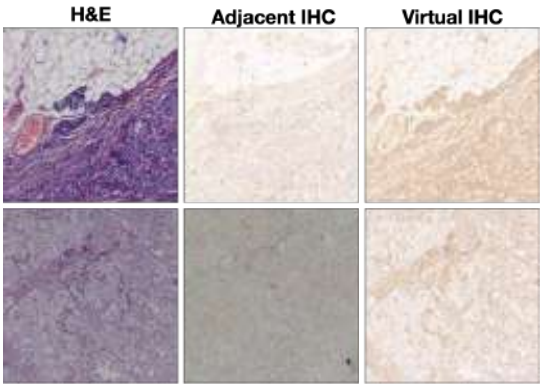


Fig. 12. The heatmap below represents the subjective evaluation scores of pathologists for Immunohistochemistry (IHC) images. These scores were assigned based on their assessment of diagnostic clinical value.

alongside the corresponding adjacent IHC and virtual IHC for scoring. Scores were assigned on a scale of 0 to 5 across three dimensions: (1) structural fidelity, (2) visual quality, and (3) diagnostic usability.

As shown in Figure 11, we summarized the mean subjective scores of our method and the state-of-the-art (SOTA) virtual staining methods across these three dimensions in a heatmap. Scores of 0–2 were classified as unsatisfactory, while scores of 3–5 were classified as satisfactory. We calculated Fleiss' Kappa to assess the inter-rater consistency among the pathologists. In the MIST dataset, the Fleiss' Kappa for structural fidelity is 0.656, for visual quality is 0.628, and for diagnostic usability is 0.818. In the IHC4BC dataset, the Fleiss' Kappa values are 0.849 for structural fidelity, 0.847 for visual quality, and 0.930 for diagnostic usability, indicating a high level of consistency among the evaluators' subjective assessments.

V. Disscussion

A. SSIM/PSNR is a good metric for Virtual Staining?

SSIM and PSNR are traditional image quality evaluation metrics that primarily measure pixel-level similarity and reconstruction quality. However, they face several limitations when applied to pathological images. The variability in brightness and contrast caused by imaging conditions, such as aperture size and light intensity, can lead to significant fluctuations, making strict brightness consistency across images difficult to achieve. As SSIM and PSNR rely on such consistency, their evaluation results can often be misleading in pathological scenarios. Furthermore, these metrics neglect critical pathological features like nuclear morphology and staining intensity, which are essential for accurate diagnosis, and therefore fail to capture high-level pathological semantics. They also struggle with the spatial heterogeneity inherent in weakly paired datasets, being overly sensitive to positional mismatches, which are often irrelevant to diagnostic outcomes. Finally, SSIM and PSNR focus solely on pixel-level similarity, disregarding the tissue structure and pathological semantics required to evaluate the contextual and diagnostic relevance of pathological images. As shown in Figure 12, the widely-used public dataset BCI[26] exhibits noticeable brightness variations. Consequently, when brightness fluctuates, the SSIM metric, as well as pixel-level metrics like PSNR, may exhibit significant changes. However, such changes do not necessarily indicate errors in generating the staining intensity of positive regions; instead, they merely reflect variations in brightness.

Currently, most studies focus on using SSIM/PSNR to evaluate virtual staining results, while largely overlooking the unavoidable brightness variations commonly encountered in clinical settings. This is fundamentally different from the assessment of natural images. As [29] illustrates, slight image rotation results in pixel-level distance changes. We attempted to adjust the brightness of the generated results on the BCI dataset based on the adjacent slices. Before global brightness adjustment, the evaluation metrics were as follows: SSIM = 0.3874 and PSNR = 18.8571. After applying global brightness adjustment, the metrics improved significantly to SSIM = 0.39717 and PSNR = 22.1215. However, despite the noticeable improvement in these metrics, the diagnostic significance of the generated images remained unchanged. Pixel-level metrics struggle to evaluate virtual staining results due to the imperfect registration of weakly paired data. Inspired by [8], we propose Path-FID, which replaces the InceptionV3 backbone of traditional FID with CONCH [28], a model pretrained on billion-scale pathological data. Path-FID captures multi-level pathological semantics and tissue structures, making it robust to varying pairing rates. As shown in Table IX and Table X, Path-FID demonstrates strong consistency in evaluating content and pathological semantics, validating its suitability for virtual staining tasks.

B. Downstream Analysis

Following [39], we categorized the IHC images into different levels based on optical density thresholds, which were initially set and later adjusted by pathologists. Specifically, the HER2 dataset was divided into four levels: 0–500, 500–2000, 2000–5000, and 5000+. For ER and PR staining, the data was divided into two levels: 0–1000 and 1000+, while Ki67 was divided into 0–2000 and 2000+. These levels were used to train a ViT classification model as a downstream task for validation.

Although TDKStain achieved the best performance in grading results, as shown in Table XI and Table XII, its poor content consistency significantly limits its potential for practical applications. In contrast, the grading performance of our method across multiple datasets highlights its advantage in maintaining pathological semantic consistency, which is critical for clinical applicability.

TABLE IX
THE QUANTITATIVE COMPARISON OF PATHOLOGICAL RELEVANCE METRICS AND PATH-FID ON THE MIST DATASET.

Method	Ki67					ER					HER2					PR				
	R_p	R_c	R_{avg}	IoD	P-FID	R_p	R_c	R_{avg}	IoD	P-FID	R_p	R_c	R_{avg}	IoD	P-FID	R_p	R_c	R_{avg}	IoD	P-FID
PyramidP2P	60.69	10.25	35.47	-2.2201	291.7394	71.17	16.29	43.73	-5.5017	188.5554	66.87	10.32	38.59	-5.8272	165.1759	81.72	11.04	0.578	-5.5563	198.0107
ASP	56.63	85.91	71.27	-1.8027	41.2966	63.53	**86.40**	74.96	-3.1462	36.8592	64.40	**88.22**	76.31	-4.7957	51.87	61.03	**87.65**	74.34	-5.1900	42.1095
PPT	55.28	59.66	51.97	-2.5072	101.3484	61.47	54.79	58.13	-4.5220	135.7866	50.02	61.93	55.98	-4.6784	88.6038	65.30	61.10	63.20	-6.2238	97.3034
TDKStain	**76.97**	8.20	42.58	**-0.1888**	68.1517	**86.54**	12.78	49.66	-2.6782	68.1517	79.98	8.84	44.41	-1.8509	84.7010	85.43	11.32	48.38	-3.2448	96.0378
UMDST	37.61	0.25	18.93	-0.787	96.7640	46.19	82.80	64.00	-2.9506	111.9784	15.98	82.93	49.46	**0.1598**	92.8557	52.16	83.30	67.73	-2.5224	177.1798
SIMGAN	64.68	**81.67**	73.17	-0.6221	19.3284	72.28	84.26	78.26	-1.4809	22.6989	61.48	85.98	73.37	-1.4421	23.7088	85.16	85.76	85.55	**-0.2687**	26.9999
Ours	73.87	81.63	**77.75**	-0.268	**17.5884**	84.07	86.06	**85.07**	**-0.5109**	**21.0212**	**80.08**	86.63	**83.35**	-1.6788	**22.5829**	**87.37**	87.03	**87.20**	-2.070	**25.0436**

TABLE X
THE QUANTITATIVE COMPARISON OF PATHOLOGICAL RELEVANCE METRICS AND PATH-FID ON THE IHC4BC DATASET.

Method	Ki67					ER					HER2					PR				
	R_p	R_c	R_{avg}	IoD	P-FID	R_p	R_c	R_{avg}	IoD	P-FID	R_p	R_c	R_{avg}	IoD	P-FID	R_p	R_c	R_{avg}	IoD	P-FID
PyramidP2P	**89.84**	7.12	48.48	-0.7704	77.6012	80.45	17.11	48.78	-1.6351	73.9984	72.79	17.68	45.24	-8.0986	89.7197	**66.28**	15.74	41.01	-1.1295	71.9115
ASP	69.11	**81.43**	75.22	0.2612	57.6038	75.66	**90.04**	75.65	-0.2174	26.8832	27.24	**94.76**	61.00	-9.9207	74.0689	15.59	87.64	51.62	-1.2598	61.4121
PPT	81.60	63.38	72.49	0.9100	58.8053	-6.11	46.85	20.37	0.8693	102.7235	65.39	72.17	68.78	-8.8844	58.1593	6.66	67.01	36.83	-1.1287	99.7219
TDKStain	81.17	4.54	45.86	-0.6335	86.0319	68.50	19.16	43.83	-0.8196	66.3562	75.58	15.28	45.43	-6.645	74.1557	16.46	15.19	15.82	-1.1883	108.7895
UMDST	67.45	18.97	43.21	1.0701	143.4910	26.62	82.58	54.60	-1.5868	164.6007	48.78	82.01	65.40	-4.1051	149.1963	15.53	82.31	48.92	**-0.0899**	176.2453
SIMGAN	83.19	80.41	81.80	0.1395	52.3105	76.31	85.27	80.79	2.0767	31.2988	65.38	88.42	76.90	-0.1793	83.5756	9.10	81.37	45.23	-0.3719	**50.4651**
Ours	84.14	80.33	**82.23**	**0.009**	**16.4687**	**82.25**	86.03	**78.47**	**0.0870**	**21.4066**	**75.76**	89.21	**81.98**	-7.7342	**35.8315**	36.57	87.03	**61.80**	3.2447	130.6920

TABLE XI
THE ACCURACY OF VIRTUAL STAINING STATE-OF-THE-ART (SOTA) METHODS FOR BREAST CANCER GRADING ON THE MIST DATASET.

Method	Ki67		ER		HER2		PR	
	Recall(↑)	F1-Score (↑)	Recall(↑)	F1-Score (↑)	Recall(↑)	F1-Score	Recall	F1-Score
PyramidP2P	0.842	0.7735	0.695	0.6633	0.587	0.5820	0.705	0.7044
ASP	0.759	0.7767	0.623	0.62405	0.378	0.3961	0.637	0.6339
PPT	0.8	0.8057	0.5460	0.5607	0.45	0.4320	0.642	0.5155
TDKStain	0.6400	0.6812	**0.8050**	**0.8087**	**0.7285**	**0.6520**	**0.7790**	**0.7799**
UMDST	0.4091	0.4091	0.5840	0.5962	0.4200	0.5336	0.5770	0.5733
SIMGAN	0.8280	0.8247	0.6780	0.6775	0.6168	0.5000	0.7050	0.7019
Ours	**0.8510**	**0.8404**	0.7240	0.7290	0.6827	0.6080	0.7310	0.7319

TABLE XII
THE ACCURACY OF VIRTUAL STAINING STATE-OF-THE-ART (SOTA) METHODS FOR BREAST CANCER GRADING ON THE IHC4BC DATASET.

Method	Ki67		ER		HER2		PR	
	Recall(↑)	F1-Score (↑)	Recall(↑)	F1-Score (↑)	Recall(↑)	F1-Score	Recall	F1-Score
PyramidP2P	0.686	0.6527	0.497	0.4679	0.505	0.4545	0.64	0.6068
ASP	0.799	0.7975	0.547	0.4958	0.256	0.2369	0.559	0.4989
PPT	0.71	0.6815	0.505	0.4892	0.6080	0.378	0.642	0.5155
TDKStain	0.774	0.7635	0.759	0.7507	0.471	0.4887	**0.759**	**0.7507**
UMDST	0.7890	0.7825	0.573	0.5701	0.412	0.4218	0.599	0.6055
SIMGAN	0.7650	0.7506	0.6960	0.6513	0.6448	**0.5110**	0.6010	0.5982
Ours	**0.8110**	**0.8109**	**0.8100**	**0.8076**	**0.6504**	0.4490	0.3877	0.4450

TABLE XIII
EVALUATE THE EFFICIENCY OF MODEL ON THE MIST ER DATASET.

Method	R_p	Parameters(M)	Inference time
CycleGAN	46.94	11.383	0.0253s
PyramidP2P	71.17	11.383	0.0103s
UMDST	45.18	10.204	0.0892s
ASP	63.53	7.838	0.0114s
TDKStain	86.54	45.593	0.0198s
SIMGAN	72.28	7.838	0.0119s
Ours	84.07	7.838	0.0115s

C. *Efficient deployment improvements*

Due to the limited computational resources in hospitals and the billion-pixel nature of whole-slide imaging (WSI) data, the clinical deployment of virtual staining faces certain challenges and adjustments. As shown in Table XIII, our method achieves comparable parameter efficiency and processing speed to lightweight models. On whole-slide images, our method processes a slide at 40x magnification in approximately 2.51 minutes and at 20x magnification in about 37.7 seconds. However, these results are based on online cloud servers, where computational power is significantly higher than that of deployment environments. Current research has shown that Spiking Neural Networks (SNNs)[25, 48, 36, 30, 49, 21, 20], which operate through discrete spike events in a biologically inspired manner, can significantly reduce the computational resource demands of neural networks. Exploring ways to convert existing models into SNNs holds significant clinical implications.

VI. CONCLUSION

This paper proposes USIGAN, a novel IHC virtual staining method that completely eliminates weakly paired terms. By employing cyclic optimal transport and intra-batch optical density consistency, USIGAN ensures both content and pathological semantic consistency in virtual IHC staining. However, it remains a one-to-one mapping virtual staining model. In future work, we plan to extend the characteristics of optimal transport to multi-domain virtual staining, enabling the model to learn differential features between various staining types within a shared framework.

VII. ACKOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (No. 62271475), Ministry of Science and Technology's key research and development

program (2023YFF0723400), Shenzhen-Hong Kong Joint Lab on Intelligence Computational Analysis for Tumor lmaging (E3G111), and the Youth Innovation Promotion Association CAS (2022365).

REFERENCES

[1] Amir Akbarnejad et al. "Predicting ki67, er, pr, and her2 statuses from h&e-stained breast cancer images". In: *arXiv preprint arXiv:2308.01982* (2023).

[2] Xiuding Cai et al. "Rethinking the paradigm of content constraints in unpaired image-to-image translation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. 2. 2024, pp. 891–899.

[3] Fuqiang Chen et al. "Pathological semantics-preserving learning for H&E-to-IHC virtual staining". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 384–394.

[4] Shikha Dubey et al. *VIMs: Virtual Immunohistochemistry Multiplex staining via Text-to-Stain Diffusion Trained on Uniplex Stains*. 2024. arXiv: 2407.19113 [eess.IV]. URL: https://arxiv.org/abs/2407.19113.

[5] Rémi Flamary et al. "Pot: Python optimal transport". In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8.

[6] Parmida Ghahremani et al. "Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification". In: *Nature machine intelligence* 4.4 (2022), pp. 401–412.

[7] Aritra Ghosh and Andrew Lan. "Contrastive learning improves model robustness under label noise". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2703–2708.

[8] Erik Großkopf et al. *HistDiST: Histopathological Diffusion-based Stain Transfer*. 2025. arXiv: 2505.06793 [eess.IV]. URL: https://arxiv.org/abs/2505.06793.

[9] Xianchao Guan et al. "OT-StainNet: Optimal Transport Driven Semantic Matching for Weakly Paired H&E-to-IHC Stain Transfer". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 3. 2025, pp. 3194–3202.

[10] Xianchao Guan et al. "Supervised Information Mining from Weakly Paired Images for Breast IHC Virtual Staining". In: *IEEE Transactions on Medical Imaging* (2025).

[11] K de Haan et al. *Deep learning-based transformation of H&E stained tissues into special stains. Nat Commun. 2021; 12: 4884*.

[12] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017).

[13] Xueqi Hu et al. "Qs-attn: Query-selected attention for contrastive learning in i2i translation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18291–18300.

[14] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[15] Ellen C Jensen. "Quantitative analysis of histological staining and fluorescence using ImageJ". In: *The Anatomical Record* 296.3 (2013), pp. 378–381.

[16] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. "Exploring Patch-Wise Semantic Relation for Contrastive Learning in Image-to-Image Translation Tasks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 18260–18269.

[17] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. "Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 18260–18269.

[18] Beomsu Kim et al. *Unpaired Image-to-Image Translation via Neural Schrödinger Bridge*. 2024. arXiv: 2305.15086 [cs.CV]. URL: https://arxiv.org/abs/2305.15086.

[19] Fangda Li et al. "Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 632–641.

[20] Yaxin Li et al. "Efficient structure slimming for spiking neural networks". In: *IEEE Transactions on Artificial Intelligence* 5.8 (2024), pp. 3823–3831.

[21] Yaxin Li et al. "Towards efficient deep spiking neural networks construction with spiking activity based pruning". In: *arXiv preprint arXiv:2406.01072* (2024).

[22] Jia-Ren Lin et al. "High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers". In: *Nature cancer* 4.7 (2023), pp. 1036–1052.

[23] Xiaoqi Lin. "Characteristic morphology and immunohistochemical patterns of clear cell papillary renal cell tumours may be observed in renal cell carcinomas, a critical pitfall in renal biopsy cytopathology". In: *Cytopathology* 35.4 (2024), pp. 481–487.

[24] Yiyang Lin et al. "Unpaired multi-domain stain transfer for kidney histopathological images". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2. 2022, pp. 1630–1637.

[25] Chang Liu et al. *Efficient ANN-SNN Conversion with Error Compensation Learning*. 2025. arXiv: 2506.01968 [cs.LG]. URL: https://arxiv.org/abs/2506.01968.

[26] Shengjie Liu et al. "Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1815–1824.

[27] Shuting Liu et al. "Unpaired stain transfer using pathology-consistent constrained generative adversarial

networks". In: *IEEE transactions on medical imaging* 40.8 (2021), pp. 1977–1989.

[28] Ming Y Lu et al. "A visual-language foundation model for computational pathology". In: *Nature medicine* 30.3 (2024), pp. 863–874.

[29] Jiabo MA et al. *Generative AI for Misalignment-Resistant Virtual Staining to Accelerate Histopathology Workflows*. 2025. arXiv: 2509.14119 [cs.CV]. URL: https://arxiv.org/abs/2509.14119.

[30] Wei Miao et al. "SpikingYOLOX: Improved YOLOX Object Detection with Fast Fourier Convolution and Spiking Neural Networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 2. 2025, pp. 1465–1473.

[31] Taesung Park et al. "Contrastive learning for unpaired image-to-image translation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer. 2020, pp. 319–345.

[32] Pushpak Pati et al. "Accelerating histopathology workflows with generative AI-based virtually multiplexed tumour profiling". In: *Nature machine intelligence* 6.9 (2024), pp. 1077–1093.

[33] Qiong Peng et al. "Advancing H&E-to-IHC Virtual Staining with Task-Specific Domain Knowledge for HER2 Scoring". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 3–13.

[34] Yair Rivenson et al. "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning". In: *Nature biomedical engineering* 3.6 (2019), pp. 466–477.

[35] Arnout C Ruifrok, Dennis A Johnston, et al. "Quantification of histochemical staining by color deconvolution". In: *Analytical and quantitative cytology and histology* 23.4 (2001), pp. 291–299.

[36] Jiangrong Shen et al. *Improving the Sparse Structure Learning of Spiking Neural Networks from the View of Compression Efficiency*. 2025. arXiv: 2502.13572 [cs.HC]. URL: https://arxiv.org/abs/2502.13572.

[37] Tehrani. "Nonlinear imaging histopathology: a pipeline to correlate gold-standard hematoxylin and eosin staining with modern nonlinear microscopy". In: *IEEE Journal of Selected Topics in Quantum Electronics* 29.4: Biophotonics (2023), pp. 1–8.

[38] Frency Varghese et al. "IHC Profiler: an open source plugin for the quantitative evaluation and automated scoring of immunohistochemistry images of human tissue samples". In: *PloS one* 9.5 (2014), e96801.

[39] Frency Varghese et al. "IHC Profiler: an open source plugin for the quantitative evaluation and automated scoring of immunohistochemistry images of human tissue samples". In: *PloS one* 9.5 (2014), e96801.

[40] Ting-Chun Wang et al. "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[41] Ting-Chun Wang et al. "High-resolution image synthesis and semantic manipulation with conditional gans". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8798–8807.

[42] Weilun Wang et al. "Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14020–14029.

[43] Weilun Wang et al. "Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14020–14029.

[44] Weilun Wang et al. "Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14020–14029.

[45] Marek Wodzinski et al. "RegWSI: Whole slide image registration using combined deep feature-and intensity-based methods: Winner of the ACROBAT 2023 challenge". In: *Computer methods and programs in biomedicine* 250 (2024), p. 108187.

[46] Shaoan Xie, Qirong Ho, and Kun Zhang. "Unsupervised image-to-image translation with density changing regularization". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 28545–28558.

[47] Bing Xiong et al. "Unpaired Multi-Domain Histopathology Virtual Staining using Dual Path Prompted Inversion". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 8. 2025, pp. 8780–8787.

[48] Qi Xu et al. *Hybrid Spiking Vision Transformer for Object Detection with Event Cameras*. 2025. arXiv: 2505.07715 [cs.CV]. URL: https://arxiv.org/abs/2505.07715.

[49] Qi Xu et al. "Reversing structural pattern learning with biologically inspired knowledge distillation for spiking neural networks". In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 3431–3439.

[50] Fangneng Zhan et al. "Marginal contrastive correspondence for guided image generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10663–10672.

[51] Fangneng Zhan et al. "Modulated contrast for versatile image synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18280–18290.

[52] Fangneng Zhan et al. *Unbalanced Feature Transport for Exemplar-based Image Translation*. 2021. arXiv: 2106.10482 [cs.CV]. URL: https://arxiv.org/abs/2106.10482.

[53] Fangneng Zhan et al. "Unbalanced feature transport for exemplar-based image translation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 15028–15038.

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 3836–3847.

[55] Ranran Zhang et al. "MVFStain: multiple virtual functional stain histopathology images generation based on specific domain mapping". In: *Medical Image Analysis* 80 (2022), p. 102520.

[56] Wei Zhang et al. "High-Resolution Medical Image Translation via Patch Alignment-Based Bidirectional Contrastive Learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 178–188.

[57] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. *The Spatially-Correlative Loss for Various Image Translation Tasks*. 2021. arXiv: 2104.00854 [cs.CV]. URL: https://arxiv.org/abs/2104.00854.

[58] Linqi Zhou et al. "Denoising Diffusion Bridge Models". In: *The Twelfth International Conference on Learning Representations*.

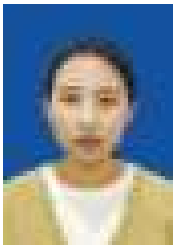
[59] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

Yue Peng is currently pursuing a PhD at the University of Chinese Academy of Sciences and the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. Her research interests include computational pathology, medical image segmentation, and medical image generation.

Bing Xiong recived the B.S. degree in the school of Wuhan University of Technology,Hubei,China,in 2023. He is currently pursuing the M.S degree with ShenZhen Institues of Advanced Technology,University of the Chinese Academy of Sciences. His research intersts include computational pathology and image generation.

Fuqiang Chen recived the B.S. degree in the school of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China, in 2022, the M.S. degree in the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences and University of Chinese Academy of Sciences, in 2025. His research interests include computational pathology.

Deboch Eybo Abera received the B.S. degree in Electronics Information Engineering from the University of Electronic Science and Technology of China in 2021, and the M.S. degree in Information and Communication Engineering from the same university in 2023. He is currently pursuing a Ph.D. degree in Pattern Recognition and Intelligent Systems at the Shenzhen Institute of Advanced Technology, Shenzhen, China. His research interests include digital pathology, multi-modality microscopy image processing etc.



RanRan Zhang recived the B.S. degree in the school of Physics and Physical Engineering,Qufu Normal University,Qufu, China, in 2017, the Ph.D. degree in the school of Information Science and Engineering, Shandong University, Qingdao, China,in 2023.She is currently working as a postdoctoral fellow at the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. Her research interests include computational pathology and deep learning.

Wanming Hu received the B.M. degree in Clinical Medicine from CSMU, Changsha, China (2010) and the M.M. and M.D. degrees in Pathology from Xiangya School of Medicine, CSU, Changsha, China (2013) and Southern Medical University, Guangzhou, China (2020). He is currently an associate Chief Physician, currently serves as Young Editorial Board Member of Chinese Journal of General Surgery, with primary research focusing on central nervous system tumors and head/neck tumors (pathological diagnosis, molecular subtyping, immune microenvironment, AI pathology, etc.).

Jing Cai recived B.S degree in Lanzhou University,Lanzhou,China, and M.S. degree in University of Georgia,USA. and the Ph.D. degree from University of Virginia,USA. he is currently the head professor of Department of Health Technology & Informatics, HK PolyU.His research interst in Novel medical imaging and image preprocessing techniques.

Wenjian Qin is currently the associate professor of the Institute of Biomedical and Health Engineering at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. Dr. Qin obtained his Ph.D. in Pattern Recognition and Intelligent System from the University of Chinese Academy of Sciences. Dr. Qin has been engaged in the research of multi-modality medical imaging with machine learning and computer vision, which explores the new computing theories and methods of the learning-based algorithm in clinical diagnosis and treatment.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：闫雪
方向：人工智能
单位：中国科学院自动化研究所

题目：大语言模型引导的高效的强化学习

EFFICIENT REINFORCEMENT LEARNING WITH LARGE LANGUAGE MODEL PRIORS

Xue Yan [1 2] **, Yan Song** [3]**, Xidong Feng** [3]**, Mengyue Yang** [4]
Haifeng Zhang[† 1 2]**, Haitham Bou Ammar** [3 5]**, Jun Wang**[† 3]

[1]Institute of Automation, Chinese Academy of Science, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]AI Centre, Department of Computer Science, University College London, London, UK
[4]University of Bristol, UK [5]Huawei Technologies, London, UK

ABSTRACT

In sequential decision-making tasks, methods like reinforcement learning (RL) and heuristic search have made notable advances in specific cases. However, they often require extensive exploration and face challenges in generalizing across diverse environments due to their limited grasp of the underlying decision dynamics. In contrast, large language models (LLMs) have recently emerged as powerful general-purpose tools, due to their capacity to maintain vast amounts of domain-specific knowledge. To harness this rich prior knowledge for efficiently solving complex sequential decision-making tasks, we propose treating LLMs as prior action distributions and integrating them into RL frameworks through Bayesian inference methods, making use of variational inference and direct posterior sampling. The proposed approaches facilitate the seamless incorporation of fixed LLM priors into both policy-based and value-based RL frameworks. Our experiments show that incorporating LLM-based action priors significantly reduces exploration and optimization complexity, substantially improving sample efficiency compared to traditional RL techniques, e.g., using LLM priors decreases the number of required samples by over 90% in offline learning scenarios.

1 INTRODUCTION

Many real-world tasks, including robotics (Polydoros & Nalpantidis, 2017; Brunke et al., 2022; Rana et al., 2023), autonomous driving (Naranjo et al., 2005; Song et al., 2022), human-AI dialogue (Li et al., 2019; McTear, 2022), and human-AI gaming (Granter et al., 2017; Yan et al., 2024), involve complex sequential decision-making (SDM) challenges. Traditional approaches to SDM, such as optimal control (Garcia et al., 1989), heuristic search (Świechowski et al., 2023) and reinforcement learning (RL) (Mnih, 2013), have seen substantial success. Notably, AlphaGo (Silver et al., 2016) and AlphaStar (Vinyals et al., 2019), both based on deep reinforcement learning (DRL), have achieved human-level proficiency in the games of Go and StarCraft II, respectively. However, these methods still suffer from high computational complexity, along with poor generalizability and limited applicability across diverse domains (Dulac-Arnold et al., 2015; Cobbe et al., 2019).

Recently, Large Language Models (LLMs) have emerged as effective tools for tackling diverse general-purpose tasks, such as in dialogue systems (Brooks et al., 2023), decision-making (Zhao et al., 2024a), and mathematical reasoning (Imani et al., 2023). Their impressive performance across various domains is largely attributed to the vast amounts of human knowledge compressed during the pre-training phase on extensive corpora (Tucker & Tuckute, 2023). Inspired by how humans make decisions using existing knowledge while learning from new tasks, LLM-based agents are emerging as a promising approach for solving SDM tasks. For instance, researchers leverage human-crafted

[†]Correspondence to : ⟨haifeng.zhang@ia.ac.cn ⟩, ⟨jun.wang@cs.ucl.ac.uk⟩

prompts to guide LLMs in making decisions directly (Zhang et al., 2023; Wang et al., 2023; Ma et al., 2023). While this approach relies heavily on the quality of the prompts and the LLMs' inherent capabilities, another line of research aims to fine-tune LLMs with RL algorithms (RLFT) (Carta et al., 2023; Christianos et al., 2023; Tan et al., 2024; Zhou et al., 2024) to enhance their decision-making precision. A more detailed exploration of related work can be found in Section 5. However, these approaches are resource-intensive, as they require meticulously crafted human prompts and an expensive fine-tuning process for LLMs for each specific task.

Inspired by these, this work stands in the position of combining the advantages of leveraging rich domain knowledge with the strengths of reinforcement learning (RL), but in a different perspective to use fixed LLMs to enhance traditional RL. The key insight is that while it may be challenging for an LLM to generate optimal plans at every state with limited experience, it can still offer suboptimal action proposals or significantly reduce the size of the action space to a manageable scope, thereby alleviating computational burdens. Building on this idea, we treat the LLM as an action prior distribution and incorporate it into Markov Decision Processes (MDPs) solving from a Bayesian inference perspective (Hu et al.). The primary objective is to approximate the posterior action distribution that aligns with task goals through probabilistic inference approaches such as variational inference or direct posterior sampling. Building on these approaches, we derive the incorporation of LLMs into both policy-based and value-based RL frameworks in a simple yet logically sound manner. We conduct experiments on major benchmarks like ALFWorld and Overcooked, demonstrating that the sample efficiency of traditional RL algorithms, such as value-based DQN and policy-based PPO, can largely benefit from integrating LLM action priors.

In summary, our main contributions are three-fold:

1. We present a unified framework for integrating Large Language Models (LLMs) as probabilistic priors into Markov decision-making frameworks. **(Section 2)**
2. We practically implement the framework by leveraging LLMs as the refined action sampler for value-based online RL **(Section 3.1)**, offline RL **(Section 3.2)** or behavior regularizer for policy-based RL **(Section 3.3)**.
3. Extensive experiments on ALFWorld and Overcooked demonstrate that our new framework can significantly boost sample efficiency compared with both pure RL and pure LLM baselines, and also bring in more robust and generalizable value function. **(Section 4)** Source code is available at `https://github.com/yanxue7/RL-LLM-Prior`.

2 FORMULATION

We seek to utilize LLMs to solve complex SDM tasks from a Bayesian inference perspective. Although LLM struggles to directly generate optimal plans in sequential decision-making tasks, it can still provide valuable suggestions for possible sub-optimal actions thanks to its rich prior knowledge Hao et al.; Zhang et al. (2024b). Building on this insight, we treat the LLM as a reliable prior distribution over valid actions and analyze its role in solving MDPs from a Bayesian inference perspective, using traditional probabilistic tools such as variational inference and probabilistic sampling.

Markov Decision Process A Markov Decision Process (MDP) can be described as a tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$. $\mathcal{S}$ is the state space, and $\mathcal{A}$ is the finite action space. We consider the textual action space and state space here. Each state s and each action a is a sentence $s, a \in \mathcal{V}^{\infty}$, where $\mathcal{V}$ is the vocabulary set. $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, mapping the current state s_t to the next state s_{t+1} following the action a_t. $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. γ is the discount factor.

LLM as an action prior It is challenging to directly solve textual decision-making tasks due to the lack of task-specific experience, yet powerful LLMs demonstrate the ability to generate reasonable action proposals Hao et al.; Yao et al. (2024). To maximize the potential of LLMs in decision-making, we treat the LLM as an action sampler, denoted as $p_{\text{LLM}}(a|s_t)$. Regarding the implementation details of sampling from LLM priors, it is challenging for an LLM to output an executable action directly. Therefore, we prompt the LLM with the textual state and admissible

actions, then get a free-form output from the LLM (e.g., a 7B LLM). This output is subsequently mapped to an executable action through a simple rule-based projection, formally described as follows: $p_{\text{LLM}}(a|s_t) = \sum_o p(a|o)\text{LLM}(o|s_t)$, where $p(a|o)$ is a projection from the LLM output o to action a. More detailed explanations of the LLM prior setting can be found in Appendix 9.2.

2.1 Variational Inference For Markov Decision Process

Similar to the *Control as Inference* framework proposed by (Levine, 2018), we introduce a binary optimality variable $\mathcal{O}$ to indicate the quality of a trajectory $\tau = \{s_0, a_0, r_0, s_1, ..., s_n\}$, where $\mathcal{O} = 1$ represents an optimal/successful trajectory, and $\mathcal{O} = 0$ otherwise. Hence, the likelihood, positively related to the reward, can be written as $p(\mathcal{O} = 1|\tau) \propto \exp\left(\sum_t \gamma^t r_t/\alpha\right)$, with α as the temperature parameter. Our goal is to maximize the optimal marginal distribution. To do that, we first apply variational inference:

$$p(\mathcal{O} = 1) \geq \mathbb{E}_{q(\tau)}\left[\log p(\mathcal{O} = 1|\tau)\right] - \text{KL}\left[q(\tau)\|p(\tau)\right] \tag{1}$$

We define the prior $p(\tau)$ as the trajectory distribution, which follows the factorization: $p(\tau) = p(s_0)\prod_t p(a_t|s_t)P(s_{t+1}|s_t, a_t)$. Here, we specify $p_{\text{LLM}}(a|s)$ as the action prior $p(a|s)$. The variational distribution $q(\tau)$ follows a similar factorization: $q(\tau) = p(s_0)\prod_t q(a_t|s_t)P(s_{t+1}|s_t, a_t)$. Substituting these factorizations back into Eq. 1, we obtain the step-wise objective as:

$$\arg\max_\theta \sum_t \mathbb{E}_{\pi_\theta}\left[\gamma^t r_t\right] - \alpha KL\left[\pi_\theta(a_t|s_t)\|p_{LLM}(a_t|s_t)\right], \tag{2}$$

and we learn a language model policy π_θ as the variational distribution $q(a_t|s_t)$, with parameter θ. More detailed derivation can be found in Appendix 8.1.

However, fine-tuning language models in policy-based methods can be challenging in practice. To address this, we propose an alternative approach: **direct posterior inference**, which allows us to sample directly from the optimal posterior distribution $p(\tau|\mathcal{O} = 1)$ for sequential decision-making.

2.2 Direct Posterior Inference for Markov Decision Process

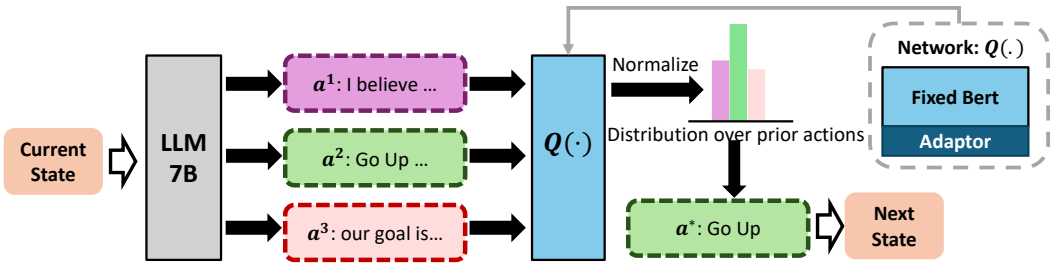


Figure 1: An illustration of the process of approximate sampling from the intractable posterior: $q(a|s, \mathcal{O} = 1) \propto p(\mathcal{O} = 1|a, s_t)p_{\text{LLM}}(a|s_t)$, implemented by reweighting the action prior proposals according to Q-values, which act as likelihood estimates. In our experiments, the Q function adopts BERT to encode textual state-action pairs and output a scalar value through an adaptor network.

Since the optimal posterior distribution over trajectory τ can be factorized into step-wise posterior described as: $p(\tau|\mathcal{O} = 1) \propto p(s_0)\prod_t p(a_t|s_t, \mathcal{O} = 1)P(s_t|s_{t-1}, a_{t-1})$, the learning procedure can be divided into two stages, modeling and inference of step-wise posterior $p(a|s_t, \mathcal{O} = 1)$. The modeling stage will point out and decompose the desirable distribution into more solvable probabilistic terms. The inference stage will give the desirable outputs using sampling-based techniques (Korbak et al., 2022).

Posterior Modeling We first use Bayes' rule to translate the intricate posterior into accessible terms:

$$p(a|s_t, \mathcal{O} = 1) = \frac{p(\mathcal{O} = 1|a, s_t)p(a|s_t)}{p(\mathcal{O} = 1|s_t)} \propto p(\mathcal{O} = 1|a, s_t)p(a|s_t), \tag{3}$$

which is a combination of the likelihood $p(\mathcal{O} = 1|a, s_t)$ and the prior $p(a|s_t)$. Here, we use the LLM as the prior action distribution $p(a|s_t) \leftarrow p_{\text{LLM}}(a|s_t)$. Recall that the optimality likelihood in MDPs

is given by: $p(\mathcal{O}=1|a,s_t) \propto \exp\left(\sum_{i=t}\gamma^{i-t}r_i\right)$, which corresponds to whether the goal state is achieved, similar to the concept of the Q function $Q(s_t,a)$ in RL. In practice, methods such as MCTS or TD-learning can estimate the optimality likelihood $p(\mathcal{O}=1|s_t,a)$.

Action Inference Having described the posterior as a product between LLM prior $p_{LLM}(a|s_t)$ and estimate Q function $Q^\theta(s_t,a)$, we can implement the inference from the posterior distribution $p(a|s_t,\mathcal{O}=1)$ by following these steps:

- Sample actions based on the prior policy p_{LLM}: Obtain k action candidates for the current state s_t, denoted as: $\mathcal{C}^k(s_t)=\{a_1,a_2,\cdots,a_k\}$, where $a_i \sim p_{\text{LLM}}(a|s_t)$.
- Choose the next action from the candidates based on the estimated Q-values: Select an action a^* according to the softmax probability distribution over the estimated Q-values. Specifically, $a^* \sim \text{softmax}(Q^\theta(s_t,a_1)/\alpha, Q^\theta(s_t,a_2)/\alpha,\ldots,Q^\theta(s_t,a_k)/\alpha)$, where α is a hyper-parameter.

This sampling strategy is illustrated in Figure 1, where the action candidates are reweighted based on the Q-values. Below, we establish the connection between variational inference and direct posterior inference.

Proposition 1. *The above action inference strategy—selecting an action from the action priors reweighted based on Q-values—can be described as following a distribution q. As $k\to\infty$, we have:*

$$\lim_{k\to\infty} q(a|s_t) = p_{LLM}(a|s_t)\exp(Q^\theta(s,a)/\alpha)/\mathbb{E}_{a_j\sim p_{LLM}(\cdot|s)}\exp(Q^\theta(s,a_j)/\alpha) \tag{4}$$

The limiting policy corresponds to the policy that optimizes the Q-values with a KL regularizer:

$$\lim_{k\to\infty} q(\cdot|s_t) = \arg\max_\pi \mathbb{E}_{\pi(a|s_t)}[Q^\theta(s_t,a)] - \alpha KL\left(\pi(\cdot|s_t)\|p_{LLM}(\cdot|s_t)\right) \tag{5}$$

Then, the posterior sampling strategy is highly related to the solution of variational inference as shown in Eq. 2. Proof. Please see the appendix 8.2.

3 Practical Implementation

In this section, we introduce the practical implementation of posterior approximation within robust RL frameworks. We describe a policy-based RL method with an LLM prior to realize the variational inference problem. In the meantime, as discussed earlier, optimality likelihood estimation is critical for direct posterior inference and is positively associated with Q-values in the MDP setting. Thereby, we consider Q-value estimation with an LLM prior for value-based RL in both online and offline settings. Finally, we propose three simple yet efficient RL algorithms: one policy-based RL algorithm and two Q-learning variants.

We will first introduce the two Q-learning variants, which perform exploration and optimization within a narrowed yet reliable LLM prior action space, eliminating the need for fine-tuning the language model-based policy. Then, we will present a policy-based PPO variant that incorporates a KL loss with respect to the LLM prior. Detailed descriptions of classic Q-learning algorithms, DQN Mnih (2013) and CQL Kumar et al. (2020), can be found in the Appendix 7.

3.1 Online Value-Based RL with LLM prior

The Q-value estimation is based on DQN (Mnih, 2013), a widely used value-based RL method, under the online setting. Unlike traditional DQN, which performs exploration and Bellman updates over the full action space $\mathcal{A}$, we propose a variant with an LLM prior, referred to as DQN-Prior. This variant limits these processes to the prior action space $\mathcal{C}^k \subseteq \mathcal{A}$, which is sampled from p_{LLM}. This prior action space represents a reduced and more rational subspace where clearly irrelevant actions may be filtered out in advance. Specifically, DQN-Prior alternates between the following steps:

Exploration Roll out new episodes using the posterior inference strategy as described in Sec. 2.2, specifically:

- For a state s_t, sample the action prior space $\mathcal{C}^k(s_t)=\{a_1,a_2,\cdots,a_k\}, a_i\sim p_{\text{LLM}}(a|s_t)$.

- Then sample the action $a_t \sim \text{softmax}\left(Q^\theta(s_t, a_1)/\alpha, \cdots, Q^\theta(s_t, a_k)/\alpha\right)$.
- Apply the action a_t to the environment, and obtain the next state $s_{t+1} \sim P(\cdot|s_t, a_t)$.
- Add the tuple to the replay buffer: $\mathcal{D} = \mathcal{D} \bigcup (s_t, a_t, s_{t+1})$.

Q-function Update After expanding the replay buffer, we use TD-learning to update the Q-network. The loss function for the i-th iteration is given as:

$$\mathcal{L}_i(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[(Q^{\theta_i}(s,a) - y_i)^2\right], \tag{6}$$

where $y_i = \mathcal{B}^* Q^{\theta_{i-1}}(s,a) = r(s,a) + \gamma \max_{a' \in \mathcal{C}^k(s')} Q^{\theta_{i-1}}(s', a')$. Compared to traditional DQN, the DQN-Prior performs exploration and applies the Bellman optimal operator within the LLM prior action space, as highlighted in red in Eq. 6.

For most experiments, we use Qwen-1.5 7B Bai et al. (2023) to generate $k = 5$ prior actions. To represent the Q-network, Bert (Devlin et al., 2019) is used to encode state-action pairs and remains frozen during optimization. Then, we introduce an adapter (a 3-layer MLP) on top of Bert's embeddings, as illustrated in Figure 1, and optimize the adapter's weights by minimizing the temporal difference loss. Therefore, we do not propagate any gradients through the Bert to avoid expensive computational overhead.

3.2 Offline Value-Based RL with LLM Prior

This study also explores the feasibility of using gradient-free LLM priors for successful SDM based solely on offline datasets. (Snell et al., 2023) have explored the use of offline RL for language generation tasks based on the SFT LLM from offline datasets. This approach assumes the entire token vocabulary as the action space, where each action corresponds to a single token, resulting in a vast action space with tens of thousands of possible actions. In our work, we consider a discrete action space where each action is a short phrase consisting of several tokens. For the SDM tasks we address, the action space is significantly smaller, with fewer than a hundred possible actions. Additionally, by leveraging an action prior, we can first obtain a tidy sub-action space $\mathcal{C}^k$, thereby narrowing the optimization space.

We apply this action prior within the CQL framework, a widely used offline RL approach, and refer to this variant as CQL-Prior. The loss function for the proposed CQL-Prior is given as:

$$\mathcal{L}_i(\theta) = \beta \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\log \sum_{a' \in \mathcal{C}^k(s)} \exp(Q^{\theta_i}(s,a')) - Q^{\theta_i}(s,a)\right] + \frac{1}{2}\mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[(Q^{\theta_i}(s,a) - y_i)^2\right], \tag{7}$$

where $y_i = r(s,a) + \gamma \max_{a' \sim \mathcal{C}^k(s')} Q^{\theta_{i-1}}(s', a')$ and β is a hyper-parameter. The main difference from traditional CQL is that it restricts the overestimation of Q-values to the tidy action prior space and applies the Bellman optimal operator within this space, as highlighted in red.

3.3 Policy-based RL with LLM Prior

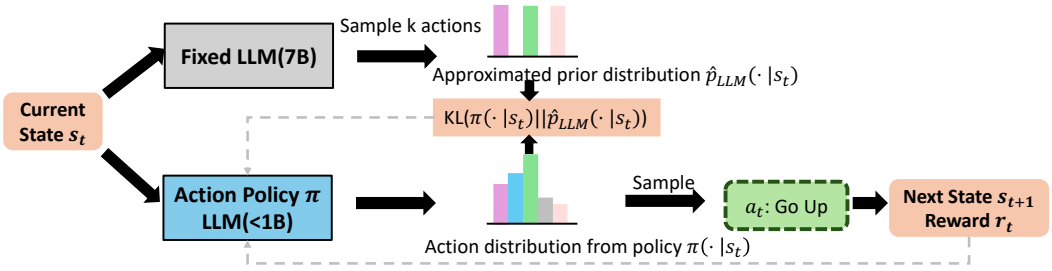


Figure 2: An illustration of the training process of GFlan-Prior, a policy-based RL algorithm with an LLM prior. A fixed LLM serves as a suboptimal prior, while the action policy learns to approximate the posterior action distribution, guided by both environmental rewards and the LLM prior.

We implement variational inference for MDPs with LLMs as action priors using a policy-based RL framework. Specifically, we build upon GFlan (Carta et al., 2023), let the Flan-T5 small model (Rae et al., 2021) (with less than $1B$ parameters) as the action policy for leveraging pre-stored knowledge to solve complex SDMs and fine-tune it via Proximal Policy Optimization (PPO) (Schulman et al., 2017). Our variant, called GFlan-Prior, incorporates a KL regularizer into the PPO loss to optimize Eq. 2. During implementation, we sample $k = 5$ examples from LLM prior distribution $p_{\text{LLM}}(a|s)$ to compute the approximated KL divergence. Formally, for the i-th iteration, the policy π_{θ_i} is learned by optimizing:

$$\arg\max_{\theta} \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{i-1}}} \text{CLIP}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{i-1}}(a_t|s_t)}\right)\hat{A}_t + \eta\mathcal{H}(\pi_\theta(\cdot|s_t)) - \alpha\text{KL}[\pi_\theta(\cdot|s_t)\|\hat{p}_{\text{LLM}}(\cdot|s_t)], \quad (8)$$

where A_t is the estimated advantage, η is the hyperparameter that encourages exploration, and $\hat{p}_{\text{LLM}}(\cdot|s_t)$ is the discrete distribution constructed by k action proposals from the LLM. θ_{i-1} represents the parameters of the action policy learned in the last iteration. The illustration of the training process of GFlan-Prior can be found in Figure 2.

4 EXPERIMENTS

In this section, we demonstrate the effectiveness of incorporating the LLM prior into the RL framework, including both policy-based and value-based algorithms in online and offline settings. LLM priors are used to reduce the exploration and optimization space or regulate the action policy's behavior, thereby resulting in a more efficient RL framework. Further details on experimental settings and more ablation study results can be found in the Appendix 9.

4.1 ENVIRONMENTS

We consider three environments:

ALFWorld (Shridhar et al., 2020) is a popular benchmark for examining LLM-based agents' decision-making ability. This benchmark contains thousands of Textworld (Côté et al., 2019) games with the embodied engine. We focus solely on the text game part, where the action space consists of high-level plans such as "go to a room". The legal actions are finite but large, with the maximum possible admissible action space reaching up to 50, making it challenging to explore from scratch. It contains thousands of tasks, making testing the generalization performance on unseen tasks convenient. We consider two classes of ALFWorld tasks: ALFWorld(pick) and ALFWorld(examine). There are no auxiliary rewards, except for a reward of 1.0 for reaching the final goal.

Overcooked We use the partially observed text overcooked game(Tan et al., 2024); the observation describes the position of visible items, and the agent should take a sequence of actions to deliver a dish. We consider two overcooked tasks: Overcooked(Tomato), deliver a dish of chopped tomato; Overcooked(Salad), deliver a salad containing chopped tomato and lettuce. The maximum possible action space is 8. Besides the reward of 1 for successfully delivering a dish, the textual Overcooked environment from (Tan et al., 2024) also provides dense reward signals. The reward shaping is as follows: 0.2 for correctly chopping an ingredient, 1 terminal reward for successfully delivering the correct dish, -0.1 for delivering any incorrect item, and -0.001 for each time step.

Frozen Lake is a grid world game; the agent should move to the goal position while avoiding falling into holes. There are four admissible actions: up, down, right, and left. The agent will receive a reward of 1 for reaching the final goal and a penalty of -1 for falling into holes. More details of state and action representations can be found in Appendix 9.1.

4.2 BASELINES

We first test the ability of **LLM Prior** to solve SDMs in a zero-shot manner, without deliberately designed prompts. Next, we will introduce the value-based and policy-based RL algorithms:

Valued-Based RL: For the online setting, we use **DQN** and our **DQN-Prior**, which explores and updates the Q-function within the LLM prior space. For the offline setting, we compare the **CQL**, **CQL-prior**, and **Behavioral Cloning(BC)**. Similarly, compared to CQL, CQL-Prior regulates and updates Q-values in the narrowed prior action space. We set $k = 5$ for DQN-Prior and CQL-Prior. The BC utilizes the Flan-T5 small (Rae et al., 2021) as the action policy, and learns the action policy by optimizing: $\arg\min_\pi -\mathbb{E}_{(s,a)\sim\mathcal{D}}[\log \pi(a|s)]$.

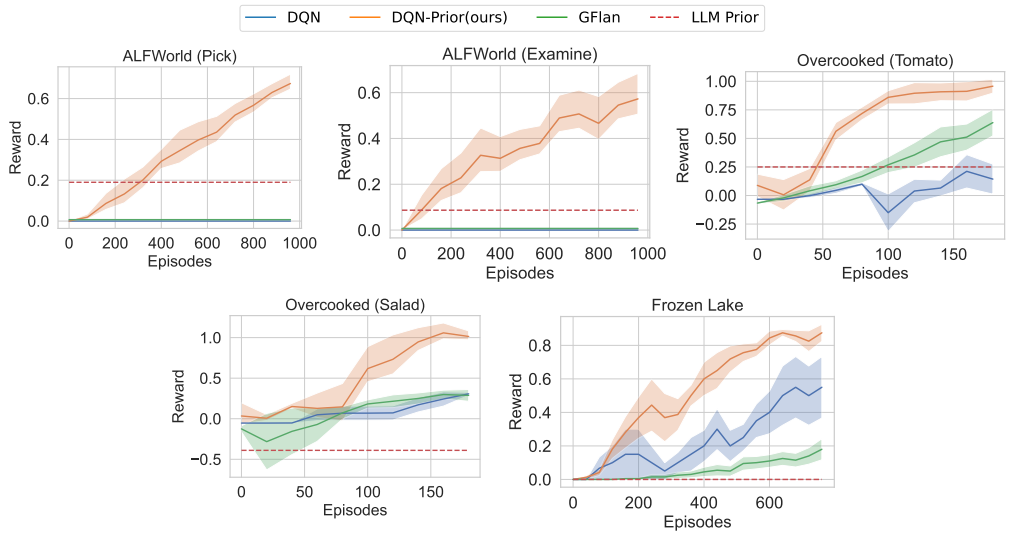


Figure 3: Results of comparison with online baselines. We plot the mean and standard error of the cumulative reward. Periodic evaluation results are plotted for online baselines.

Policy-Based RL: GFlan (Carta et al., 2023) lets the LLM Flan-T5 small as the action policy and finetunes it via PPO (Schulman et al., 2017). **GFlan-Prior** adds a KL constraint between the optimized action policy and LLM prior action on the basis of GFlan, as shown in Eq. 26.

LLM Setting Due to the Qwen-1.5 series (Bai et al., 2023) containing a range of LLMs with different scales, we use Qwen-1.5 for our main experiments. Specifically, we use Qwen-1.5 7B (Bai et al., 2023) as the backbone of the LLM prior for most environments, except for the more complex Overcooked(Salad) task, where Qwen-1.5 14B is used.

4.3 Results Analysis

4.3.1 Incorporating LLM Priors into Value-Based RL

The effectiveness of LLM Prior on online Q-learning As shown in Figure 3, our DQN-Prior outperforms traditional online RL baselines such as DQN and GFlan by exploring and optimizing within a narrowed, more valuable prior action space. In complex ALFWorld tasks, DQN and GFlan struggle to bootstrap due to the large original action space, which can include up to 50 possible actions and the lack of dense reward incentives. In contrast, our DQN-Prior leverages the LLM's domain knowledge to reduce the exploration space to just five suboptimal actions, significantly improving sample efficiency. Additionally, in the Overcooked (Salad) and Frozen Lake tasks, the LLM prior cannot directly solve these tasks by sampling one action per state. However, DQN-Prior still achieves success and demonstrates greater sample efficiency than traditional RL methods, despite the small action space (only 4 or 8 actions). This shows that the LLM prior can provide a suboptimal action proposal space, and conducting RL within this space significantly enhances sample efficiency.

Table 1: Results of offline algorithms. We consider two tasks, ALFWorld (Pick) and Overcooked (Salad), referred to as Pick and Salad for simplicity. To further compare the sample efficiency of the baselines, we use offline datasets of varying sizes for Overcooked (Salad), denoted as Salad (N), each containing approximately N (s, a, s') tuples.

Baseline	Pick	Salad(1000)	Salad(4000)	Salad(8000)	Salad(12000)	Salad(24000)
DQN	0.27	0.14	0.31	0.32	0.32	0.32
CQL	0.03	0.32	0.32	0.32	0.32	**1.33**
DQN-Prior	0.49	0.98	0.81	-	-	-
CQL-Prior	**0.80**	**1.01**	**1.19**	-	-	1.31
BC	0.31	0.57	0.91	0.78	0.53	0.74

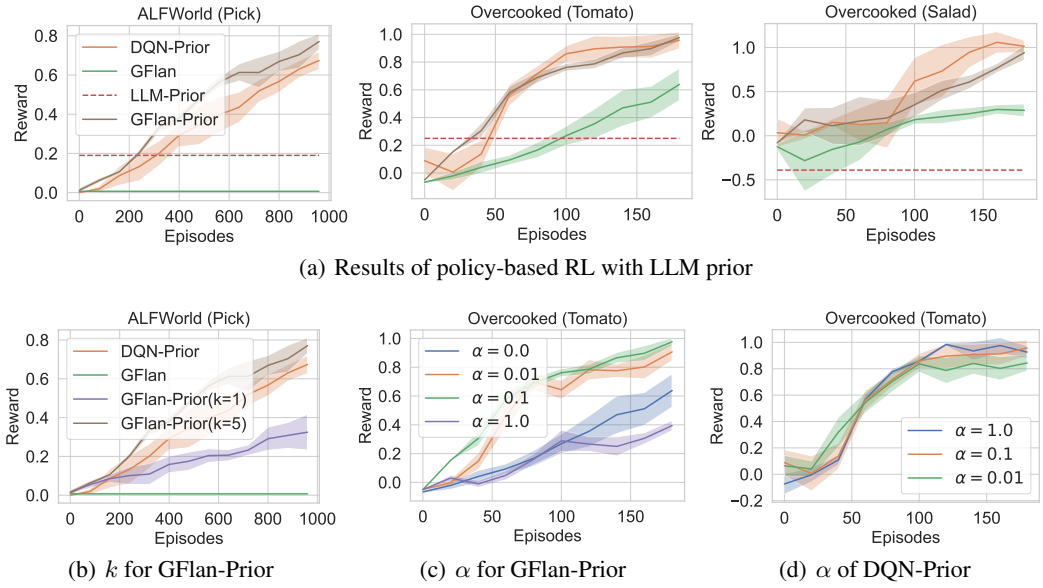
Table 2: Results on the generalization performance of posterior sampling on ALFWorld(Pick). We use Qwen-1.5 7B with $k = 5$ to generate LLM prior action proposals for training the Q function via DQN-Prior and CQL-Prior. This LLM prior configuration is highlighted by (*).

Online (DQN-Prior)								
	Seen Tasks				**Unseen Tasks**			
	$k = 1$	$k = 5$	$k = 10$	$k = 15$	$k = 1$	$k = 5$	$k = 10$	$k = 15$
Qwen-1.5 4B	0	0.35	**0.46**	0.35	0	0.08	0.25	**0.29**
Qwen-1.5 7B	0.19	0.62*	**0.77**	0.65	0.04	0.38	**0.42**	**0.42**
LLaMa-3 8B	0.04	0.62	**0.73**	0.69	0.13	**0.42**	0.38	0.38
Qwen-1.5 14B	0.19	0.65	**0.77**	0.65	0.16	**0.54**	**0.54**	0.38
Qwen-1.5 32B	0.35	0.77	**0.81**	0.73	0.33	**0.5**	**0.5**	0.46
Offline (CQL-Prior)								
	Seen Tasks				**Unseen Tasks**			
	$k = 1$	$k = 5$	$k = 10$	$k = 15$	$k = 1$	$k = 5$	$k = 10$	$k = 15$
Qwen-1.5 4B	0	0.27	0.38	**0.62**	0.0	0.17	0.17	**0.38**
Qwen-1.5 7B	0.19	0.80*	**0.81**	0.77	0.04	0.46	**0.50**	**0.50**
LLaMa-3 8B	0.04	0.85	**0.88**	0.81	0.13	**0.54**	0.50	**0.54**
Qwen-1.5 14B	0.19	0.73	**0.92**	0.73	0.16	0.54	**0.58**	0.54
Qwen-1.5 32B	0.35	0.81	**0.85**	0.69	0.33	**0.63**	0.42	0.50

The effectiveness of LLM Prior on offline Q-learning Table 1 illustrates the performance of baselines trained on offline datasets. A more detailed description of the offline datasets can be found in the Appendix. Table 1 shows that CQL-Prior outperforms all other baselines on the ALFWorld (Pick) and Overcooked (Salad) datasets. Although DQN-Prior lacks constraints on Q-values, it still demonstrates superior performance compared to DQN and CQL, which operate on the full action space. These results indicate that in offline RL, avoiding overestimation of Q-values and performing Bellman updates only within the suboptimal prior action space reduces optimization complexity, thereby improving sample efficiency.

Additionally, we provide further evidence to support this conclusion by testing the sample requirements for CQL and CQL-Prior. We find that traditional CQL, with fewer than 12000 examples, fails to learn how to achieve the final goal, only managing to reach the subgoal of chopping the tomato and lettuce (with a reward of 0.4). In contrast, CQL-Prior, with just around 1000 examples, successfully learns how to achieve the final goal (with a reward of 1.0) with high probability. Therefore, our CQL-Prior reduces the number of required samples by at least 90% compared to CQL on this task.

The generalization ability across unseen tasks and LLMs Table 2 demonstrates the generalization ability of our posterior sampling framework by combining the Q-function, trained with the LLM prior setting of Qwen-1.5 7B and $k = 5$, with other LLMs and varying action proposal numbers (k) during inference. Even though the Q-network is trained with Qwen-1.5 7B and $k = 5$, it can still effectively guide the decision-making process, i.e., action selection, for unseen LLMs of different scales and architectures. It is worth noting that, although Qwen-1.5 4B was originally unable to solve ALFWorld (Pick), it becomes capable of tackling such complex tasks by leveraging the Q-function trained with more powerful LLMs. This value-based RL approach, which utilizes an LLM prior, presents a promising opportunity to train an information-rich Q-network using larger LLMs while deploying only the Q-network and smaller LLMs on the client side. This method enhances reasoning speed and reduces deployment complexity. Furthermore, larger LLMs generally perform better during inference, as their action proposals are of higher quality. $k = 1$ indicates the performance of the LLM on its own, without involving the Q-function. The setting of $k = 5$ was chosen during training for querying efficiency, but it is not necessarily the optimal choice. We observed better test performance with $k = 10$ on seen tasks, indicating that our method can generalize to a larger action space to some extent. In most cases, inference performance increases and then decreases as k increases. This may occur because a larger action space is more likely to include the optimal action, but an excessively large action space can introduce unseen (s, a) pairs during training, leading to out-of-distribution (OOD) problems in both online and offline settings.



(a) Results of policy-based RL with LLM prior

(b) k for GFlan-Prior

(c) α for GFlan-Prior

(d) α of DQN-Prior

Figure 4: (a) Comparison of policy-based RL algorithms, we plot DQN-Prior and LLM-Prior for reference. (b) The ablation study on the number of LLM action proposals k used to approximate the KL divergence. (c) The ablation of the KL coefficient of GFlan-Prior. (d) The ablation of the softmax temperature over Q values of the DQN-Prior

Regarding generalization to unseen tasks, multiple LLM action proposals with $k > 1$, guided by the Q-function, generally outperform the LLM's zero-shot performance with $k = 1$. This demonstrates the generalization capability of our posterior sampling framework, enabled by the powerful LLM as an action prior and the Q-function's ability to compress interactive experiences.

4.3.2 Incorporating LLM Priors into Policy-Based RL

The Figure 4(a), 4(b), 4(c) show the results of policy-based RL algorithms. As shown in Figure 4(a), GFlan-Prior significantly outperforms GFlan and achieves performance comparable to DQN-Prior. Unlike GFlan, GFlan-Prior leverages the prior knowledge embedded in LLMs by aligning with the LLM prior distribution through the incorporation of a KL constraint, as shown in Equation 26. In our main results, we sample $k = 5$ action proposals from the LLM prior to approximate the KL divergence. An ablation study on k can be found in Figure 4(b) and Figure 5 in the Appendix. Similar to the results in Table 2, with $k = 5$, we are more likely to sample the optimal action than with $k = 1$, thereby better supervising the learning of the action policy.

The ablation study on the coefficient of the KL divergence is shown in Figure 4(c). A large α will cause the learned policy π to closely follow the LLM prior proposals, while a small α encourages exploration. In DQN-Prior, the hyperparameter α regulates exploration by acting as the temperature in the softmax over Q, generating a Boltzmann distribution based on Q, as investigated in Figure 4(d). GFlan-Prior appears to be more sensitive to the hyperparameter on controlling exploration than DQN-Prior.

5 Related Work

This study explores the fusion of Bayesian inference and LLMs for decision-making tasks. Herein, we present related works on LLM-Agent for decision-making tasks and how previous research pursues this combination. More discussions on the use of LLMs for value-based inference and reward function generation are provided in Appendix 10.

LLM-driven Agent The idea of LLM-as-Agent has gone viral since the release of powerful large language models. Among all the works, some apply direct prompt engineering techniques

using human-designed prompt and retrieval to harness LLMs for complex interactive tasks, such as ProAgent by (Zhang et al., 2023), Voyager by (Wang et al., 2023) and generative agent by (Park et al., 2023). Many researchers have also tried searching-based methods. (Yao et al., 2023) propose Tree-of-Thought (TOT) using BFS/DFS algorithm to search accurate decision sequences. (Hao et al.) suggests using Monte-Carlo Tree Search to perform a more comprehensive search. These methods often fall into the framework of Proposer-Verifier (Snell et al., 2024) where the language proposes a number of possible sequences and the verifier picks desirable candidates. There are also a few works that relate LLM to conventional reinforcement learning settings and investigate how the superiority in language space helps RL agents' capability in decision-making. For instance, (Brooks et al., 2024) uses LLM as a world model and performs policy iteration by in-context learning. GFlan proposed by (Carta et al., 2023) uses a learnable LLM as a probability likelihood estimator for all possible actions and incorporates it into actor-critic learning frameworks where a value function generates assessment and fine-tuning the LLM correspondingly. (Yan et al., 2023) also resort to an actor-critic learning paradigm but instead of fine-tuning the language model, they train a simple classifier that selects valuable outputs from a fixed language model (e.g. GPT3.5). (Zhang et al., 2024b) use LLM as a behavior regularize and add it to the value estimate. (Wen et al., 2024) leverages entropy-regularized reinforcement learning to train LLM agents for verbal sequential decision-making tasks. Our work also falls into the category of LLM-driven RL agent. However, different from others, we see a large language model as an excellent action proposer and perform Q-learning directly on the reformed action sequence space, thereby achieving better learning efficiency with minimum modification.

LLM in Bayesian Sense The probabilistic nature of transformer-based language models makes them well-suited for a Bayesian inference framework. As highlighted by (Korbak et al., 2022), the Bayesian approach offers a unified perspective on both modeling and inference. In in-context learning (ICL), some researchers argue for the Bayesian properties of ICL ((Jiang, 2023; Wang et al., 2024b; Ye et al., 2024)). Extending beyond ICL, (Yang & Klein, 2021) trains a classifier to approximate the likelihood function, framing conditional text generation as a Bayesian inference problem, with the LLM serving as the prior. (Hu et al.) interprets Chain-of-Thought (CoT) reasoning in natural language processing (NLP) tasks, such as text infilling and sequence continuation, as probabilistic inference problems, using GFlowNet to fine-tune the model. In this context, the language model functions as a unified probabilistic generative process, capable of representing both prior and likelihood distributions. Similarly, (Zhao et al., 2024b) adopts a Bayesian perspective, utilizing Sequential Monte Carlo (SMC) methods to generate undesired outcomes. (Gallego, 2024) frames LLM inference as sampling from a posterior distribution focused on high-reward regions, applying this approach to self-improvement tasks. Inspired by these works, we also formulate our agent framework within a Bayesian setting, exploring the use of LLMs as priors.

6 Conclusion

In this work, we aim to harness the rich domain knowledge and strong reasoning abilities of LLMs to tackle complex decision-making tasks while avoiding the costly processes of prompt design and fine-tuning. Given their extensive prior knowledge but limited task-specific experience, we do not rely on LLMs to directly generate optimal plans. Instead, we simplify their role to producing reliable action proposals. To this end, we treat the powerful LLM as an action prior distribution and, from a Bayesian inference perspective, examine how to integrate LLM priors into solving MDPs. We employ variational inference and direct posterior sampling to achieve this, ultimately proposing both policy-based and value-based RL frameworks incorporating LLM priors. These priors enhance traditional RL frameworks by narrowing the exploration space or guiding the action policy through suboptimal LLM-generated proposals. Experimental results demonstrate a substantial improvement in sample efficiency when incorporating fixed LLM priors into RL frameworks in both online and offline settings. This work focuses specifically on text-based games with finite action spaces. In future research, we plan to explore scenarios with free-form and infinite action spaces and investigate deliberate reasoning approaches to further refine the quality of LLM-generated action proposals.

Acknowledgements

Co-author Haifeng Zhang thanks IPT Project 2024002.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Ethan Brooks, Logan A Walls, Richard Lewis, and Satinder Singh. Large language models can implement policy iteration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Ethan Brooks, Logan Walls, Richard L Lewis, and Satinder Singh. Large language models can implement policy iteration. *Advances in Neural Information Processing Systems*, 36, 2024.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*, pp. 3676–3713. PMLR, 2023.

Filippos Christianos, Georgios Papoudakis, Matthieu Zimmer, Thomas Coste, Zhihao Wu, Jingxuan Chen, Khyati Khandelwal, James Doran, Xidong Feng, Jiacheng Liu, Zheng Xiong, Yicheng Luo, Jianye Hao, Kun Shao, Haitham Bou-Ammar, and Jun Wang. Pangu-agent: A fine-tunable generalist agent with structured reasoning, 2023.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International conference on machine learning*, pp. 1282–1289. PMLR, 2019.

Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pp. 41–75. Springer, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Gabriel Dulac-Arnold, Richard Evans, Hado van Hasselt, Peter Sunehag, Timothy Lillicrap, Jonathan Hunt, Timothy Mann, Theophane Weber, Thomas Degris, and Ben Coppin. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*, 2015.

Fares Fourati, Vaneet Aggarwal, and Mohamed-Slim Alouini. Stochastic q-learning for large discrete action spaces. *arXiv preprint arXiv:2405.10310*, 2024.

Victor Gallego. Distilled self-critique of llms with synthetic data: a bayesian perspective. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Carlos E Garcia, David M Prett, and Manfred Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335–348, 1989.

Scott R Granter, Andrew H Beck, and David J Papke Jr. Alphago, deep learning, and the future of the human microscopist. *Archives of pathology & laboratory medicine*, 141(5):619–621, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pp. 1861–1870. PMLR, 2018.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Edward J Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. In *The Twelfth International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023.

Hui Jiang. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*, 2023.

Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166*, 2023.

Tomasz Korbak, Ethan Perez, and Christopher L Buckley. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Kenneth Li, Samy Jelassi, Hugh Zhang, Sham Kakade, Martin Wattenberg, and David Brandfonbrener. Q-probe: A lightweight approach to reward maximization for language models. *arXiv preprint arXiv:2402.14688*, 2024.

Ziming Li, Julia Kiseleva, and Maarten De Rijke. Dialogue generation: From imitation learning to inverse reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6722–6729, 2019.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Weiyu Ma, Qirui Mi, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *arXiv preprint arXiv:2312.11865*, 2023.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024.

Michael McTear. *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature, 2022.

Volodymyr Mnih. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

José Eugenio Naranjo, Carlos González, Ricardo García, Teresa de Pedro, and Rodolfo E Haber. Power-steering control architecture for automatic driving. *Ieee transactions on intelligent transportation systems*, 6(4):406–415, 2005.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Athanasios S Polydoros and Lazaros Nalpantidis. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, 2017.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Krishan Rana, Ming Xu, Brendan Tidd, Michael Milford, and Niko Sünderhauf. Residual skill policies: Learning an adaptable skill-based action space for reinforcement learning for robotics. In *Conference on Robot Learning*, pp. 2095–2104. PMLR, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016. URL `http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html`.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Charlie Victor Snell, Ilya Kostrikov, Yi Su, Sherry Yang, and Sergey Levine. Offline rl for natural language generation with implicit language q learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Qingyuan Song, Weiping Fu, Wen Wang, Yuan Sun, Denggui Wang, and Jincao Zhou. Quantum decision making in automatic driving. *Scientific reports*, 12(1):11042, 2022.

Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. Monte carlo tree search: A review of recent modifications and applications. *Artificial Intelligence Review*, 56(3): 2497–2562, 2023.

Weihao Tan, Wentao Zhang, Shanqi Liu, Longtao Zheng, Xinrun Wang, and Bo An. True knowledge comes from practice: Aligning large language models with embodied environments via reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Mycal Tucker and Greta Tuckute. Increasing brain-llm alignment via information-theoretic compression. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Audrey Dudzik, Junyoung Chung, Arthur J Tashman, David Horgan, Peter Kroiss, Max Powell, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024a.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Muning Wen, Cheng Deng, Jun Wang, Weinan Zhang, and Ying Wen. Entropy-regularized token-level policy optimization for large language models. *arXiv preprint arXiv:2402.06700*, 2024.

Xue Yan, Yan Song, Xinyu Cui, Filippos Christianos, Haifeng Zhang, David Henry Mguni, and Jun Wang. Ask more, know better: Reinforce-learned prompt questions for decision making with large language models. *arXiv preprint arXiv:2310.18127*, 2023.

Xue Yan, Jiaxian Guo, Xingzhou Lou, Jun Wang, Haifeng Zhang, and Yali Du. An efficient end-to-end training approach for zero-shot human-ai coordination. *Advances in Neural Information Processing Systems*, 36, 2024.

Kevin Yang and Dan Klein. Fudge: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*, 2021.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Naimeng Ye, Hanming Yang, Andrew Siah, and Hongseok Namkoong. Pre-training and in-context learning is bayesian inference a la de finetti. *arXiv preprint arXiv:2408.03307*, 2024.

Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montserrat Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. In *Conference on Robot Learning*, pp. 374–404. PMLR, 2023.

Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. Proagent: Building proactive cooperative ai with large language models. *arXiv preprint arXiv:2308.11339*, 2023.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024a.

Shenao Zhang, Sirui Zheng, Shuqi Ke, Zhihan Liu, Wanxin Jin, Jianbo Yuan, Yingxiang Yang, Hongxia Yang, and Zhaoran Wang. How can llm guide rl? a value-based approach. *arXiv preprint arXiv:2402.16181*, 2024b.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024a.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Baker Grosse. Probabilistic inference in language models via twisted sequential monte carlo. In *International Conference on Machine Learning*, pp. 60704–60748. PMLR, 2024b.

Runlong Zhou, Simon Du, and Beibin Li. Reflect-rl: Two-player online rl fine-tuning for lms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 995–1015, 2024.

7 DESCRIPTION OF CLASSIC Q-LEARNING ALGORITHM

We introduce the classic Q-learning algorithm in both offline and online settings.

Q-learning Q-learning is a classic value-based RL algorithm, which iteratively applies the Bellman optimal operator to train the Q-function, formally written as $\mathcal{B}^*Q(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\max_a' Q(s',a')\right]$. After the optimal Q-function Q^* learned, then the optimal action policy respect to the Q^* is given as: $\pi^*(\cdot|s) = \arg\max_a Q^*(s,a)$. DQN (Mnih, 2013) is a well-known Q-learning algorithm, which can process various classes of the state information such as image and language, which learns a deep neural network $Q^\theta(s,a) \approx Q^*(s,a)$ to approximate $Q^*(s,a)$ and uses TD-learning to iteratively update the Q network, the loss function for the i-the iteration is given as:

$$\mathcal{L}_i(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[(Q^{\theta_i}(s,a) - y_i)^2\right], \tag{9}$$

where $\mathcal{D}$ is the reply buffer, $y_i = \mathcal{B}^*Q^{\theta_{i-1}}(s,a) = r(s,a) + \gamma\max_{a'\in\mathcal{A}} Q^{\theta_{i-1}}(s',a')$. θ_{i-1} is the learned parameters of the last iteration and is frozen during the gradient descent for optimizing the loss function $\mathcal{L}_i$.

Conservative Q-learning CQL (Kumar et al., 2020) framework is designed for offline reinforcement learning, where only offline data is available, but no further online exploration is accessible. Different from traditional Q-learning, CQL introduces an additional regularization term on top of the standard DQN framework or actor-critic framework(such as SAC (Haarnoja et al., 2018)), ensuring that the learned Q-function lower-bounds the actual value. The lower-bounded Q-values help mitigate the overestimation problem commonly encountered with out-of-distribution (OOD) state-action pairs. The CQL loss on the top of value-based DQN is given as:

$$\mathcal{L}_i(\theta) = \beta\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\log\sum_{a'\in\mathcal{A}}\exp(Q^{\theta_i}(s,a')) - Q^{\theta_i}(s,a)\right] + \frac{1}{2}\mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[(Q^{\theta_i}(s,a) - y_i)^2\right], \tag{10}$$

where $y_i = r(s,a) + \gamma\max_{a'} Q^{\theta_{i-1}}(s',a')$ and β is a hyper-parameter.

8 PROOFS

8.1 PROBABILISTIC INFERENCE DERIVATION

In this section, we derive the detailed formulation for probabilistic inference for using LLM as a prior. This includes the policy-based method where we learn a parametrized policy function and the value-based method where we directly perform posterior inference.

Variational Inference Assume that we have the optimality variable $\mathcal{O}$ indicating the quality of a trajectory $\tau = \{s_0, a_0, r_0, s_1, ..., s_n\}, p(\tau) = p(s_0)\prod_t p(a_t|s_t)p(s_{t+1}|s_t,a_t)$. The likelihood function with temperature parameter α can be written as:

$$p(\mathcal{O}=1|\tau) = \exp\left(\sum_t \gamma^t r_t/\alpha\right) \tag{11}$$

Our goal is to maximize the optimal marginal distribution formulated as follows:

$$\begin{aligned} p(\mathcal{O}=1) &= \int p(\mathcal{O}=1,\tau)d\tau \\ &= \int q(\tau)\frac{p(\mathcal{O}=1|\tau)p(\tau)}{q(\tau)}d\tau \end{aligned} \tag{12}$$

where $q(\tau)$ is a parametrized variational distribution. Then, our variational inference objective can be written as:

$$\begin{aligned} \log p(\mathcal{O}=1) &\geq \int q(\tau)\log\left[p(\mathcal{O}=1|\tau)\frac{p(\tau)}{q(\tau)}\right]d\tau \\ &= \mathbb{E}_{q(\tau)}\left[\log p(\mathcal{O}=1|\tau)\right] - \mathbf{KL}\left[q(\tau)|p(\tau)\right] \\ &= \mathbf{ELBO} \end{aligned} \tag{13}$$

In the context of language models, we define prior $p(\tau)$ as trajectory distribution generated through a fixed LLM denoted as:

$$p_{LLM}(\tau) = p(s_0) \prod_t p_{LLM}(a_t|s_t) p(s_{t+1}|s_t, a_t) \tag{14}$$

where $p_{LLM}(a_t|s_t)$ is the probability likelihood of the generated output a_t given input text s_t. The variational distribution $q(\tau)$ can also be decomposed in the same way as:

$$q(\tau) = p(s_0) \prod_t \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t) \tag{15}$$

where the parametrized policy function π_θ can be a learnable language model with parameter θ. Substituting these factorizations back to ELBO we have the step-wise objective function written as:

$$-\mathcal{L} = \sum_t \mathbb{E}_{\pi_\theta} \left[\gamma^t r_t\right] - \alpha \mathbf{KL} \left[\pi_\theta(a_t|s_t) \| p_{LLM}(a_t|s_t)\right] \tag{16}$$

$$\pi^* = \arg\min_\theta \mathcal{L} \tag{17}$$

Direct Posterior Inference Instead of learning a parametrized policy which might involve fine-tuning a language model, we can also try directly sampling from the posterior distribution formulated as:

$$\begin{aligned} p(\tau|\mathcal{O}=1) &= \frac{p(\mathcal{O}=1|\tau)p(\tau)}{p(\mathcal{O}=1)} \\ &\propto p(\mathcal{O}=1|\tau)p(\tau) \\ &= \prod_t p(\mathcal{O}_t = 1|s_t, a_t) p_{LLM}(a_t|s_t) p(s_{t+1}|s_t, a_t) \end{aligned} \tag{18}$$

in which we define $p(\tau)$ as in Equation 14. According to Control-as-Inference framework proposed by Levine (2018), we can formulate posterior inference as Soft Q-Learning where Q-values are updated using a modified soft Bellman equation:

$$\begin{aligned} Q^\pi(s_t, a_t) &= r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p(\cdot|s_t, a_t)} \left[V(s_{t+1})\right] \\ V(s_t) &= \log \int \exp(Q(s_t, a_t)/\alpha) da_t \end{aligned} \tag{19}$$

and the policy can be defined as a softmax function over the Q-values: $\pi(a|s) = \exp(Q^\pi(s,a)/\alpha) / \int_{a'} \exp(Q^\pi(s,a^{'})/\alpha) da^{'}$, ensuring that actions with higher Q-values are more likely to be chosen. In our paper, however, due to the enormously large space of (textual) state-action pair, we also apply sampling approximation methods. Specifically, we leverage similar ideas from Fourati et al. (2024) which sample a random subset of actions from the complete action space and seek the optimal within this subset. In our case, we rely on a fixed LLM to provide such action subspace and can further narrow down the space through repeated sampling (Brown et al. (2024)):

$$\begin{aligned} \pi(a|s) &= \frac{\mathbf{1}_{a \in p_{LLM}(\cdot|s)} \cdot \exp\left(Q^\pi(s,a)/\alpha\right)}{\int_{a'} \mathbf{1}_{a \in p_{LLM}(\cdot|s)} \cdot \exp\left(Q^\pi(s,a')/\alpha\right) da'} \\ &\approx \mathbf{1}_{a \in \{a^i\}} \cdot \frac{1}{N} \frac{\exp(Q^\pi(s,a^i)/\alpha)}{\sum_i \exp(Q^\pi(s,a^i)/\alpha)}, a^i \sim p_{LLM}(a|s) \end{aligned} \tag{20}$$

This means we restrict the Bellman backup to a specific action subspace that can potentially provide a refined set of actions based on context, largely reducing computational complexity and focusing on more relevant actions.

8.2 Proof of Proposition 1

Proposition 1. *The above action inference strategy—selecting an action from the action priors reweighted based on Q-values—can be described as following a distribution q. As $k \to \infty$, we have:*

$$\lim_{k \to \infty} q(a|s_t) = p_{LLM}(a|s_t) \exp(Q^\theta(s,a)/\alpha) / \mathbb{E}_{a_j \sim p_{LLM}(\cdot|s)} \exp(Q^\theta(s,a_j)/\alpha) \tag{4}$$

The limiting policy corresponds to the policy that optimizes the Q-values with a KL regularizer:

$$\lim_{k\to\infty} q(\cdot|s_t) = \arg\max_{\pi} \mathbb{E}_{\pi(a|s_t)}[Q^\theta(s_t,a)] - \alpha KL\left(\pi(\cdot|s_t)\|p_{LLM}(\cdot|s_t)\right) \tag{5}$$

Then, the posterior sampling strategy is highly related to the solution of variational inference as shown in Eq. 2. Proof. Please see the appendix 8.2.

Proof. The proof mainly follows the (Li et al., 2024). Denote the above sampling strategy indeed follows a distribution q, and for each action a, the probability of a is sampled from the distribution q is given by:

$$\begin{aligned} q(a|s_t) &= \mathbb{E}_{\{a_1,\cdots,a_k\}\sim p_{\text{LLM}}(\cdot|s)}\left[\sum_{i=1}^{k}\mathbb{I}(a_i=a)\frac{\exp(Q^\theta(s_t,a_i)/\alpha)}{\sum_{j=1}^{k}\exp(Q^\theta(s_t,a_j)/\alpha)}\right] \\ &= \mathbb{E}_{\{a_1,\cdots,a_k\}\sim p_{\text{LLM}}(\cdot|s)}\left[\frac{\sum_{i=1}^{k}\mathbb{I}(a_i=a)}{\sum_{j=1}^{k}\exp(Q^\theta(s_t,a_j)/\alpha)}\right]\exp(Q^\theta(s_t,a)/\alpha) \\ &= \mathbb{E}_{\{a_1,\cdots,a_k\}\sim p_{\text{LLM}}(\cdot|s_t)}\left[\frac{\frac{1}{k}\sum_{i=1}^{k}\mathbb{I}(a_i=a)}{\frac{1}{k}\sum_{j=1}^{k}\exp(Q^\theta(s_t,a_j)/\alpha)}\right]\exp(Q^\theta(s_t,a)/\alpha) \end{aligned} \tag{21}$$

According to the Law of Large Numbers, when $k\to\infty$, we have:

$$\begin{aligned} \lim_{k\to\infty} q(a|s_t) &= \lim_{k\to\infty}\mathbb{E}_{\{a_1,\cdots,a_k\}\sim p_{\text{LLM}}(\cdot|s_t)}\left[\frac{\frac{1}{k}\sum_{i=1}^{k}\mathbb{I}(a_i=a)}{\frac{1}{k}\sum_{j=1}^{k}\exp(Q^\theta(s_t,a_j)/\alpha)}\right]\exp(Q^\theta(s_t,a)/\alpha) \\ &= \lim_{k\to\infty}\mathbb{E}_{\{a_1,\cdots,a_k\}\sim p_{\text{LLM}}(\cdot|s_t)}\left[\frac{p_{\text{LLM}}(a|s_t)}{\mathbb{E}_{a_j\sim p_{\text{LLM}}(\cdot|s_t)}\exp(Q^\theta(s_t,a_j)/\alpha)}\right]\exp(Q^\theta(s_t,a)/\alpha) \\ &= p_{\text{LLM}}(a|s_t)\frac{\exp(Q^\theta(s,a)/\alpha)}{\mathbb{E}_{a_j\sim p_{\text{LLM}}(\cdot|s)}\exp(Q^\theta(s,a_j)/\alpha)} \end{aligned} \tag{22}$$

Following the proof process from the Appendix A.1. in Rafailov et al. (2023), we have:

$$\arg\max_{\pi}\mathbb{E}_{\pi(a|s_t)}[Q^\theta(s_t,a)] - \alpha\text{KL}\left(\pi(\cdot|s_t)\|p_{\text{LLM}}(\cdot|s_t)\right) = p_{\text{LLM}}(a|s_t)\frac{\exp(Q^\theta(s,a)/\alpha)}{\mathbb{E}_{a_j\sim p_{\text{LLM}}(\cdot|s)}\exp(Q^\theta(s,a_j)/\alpha)} \tag{23}$$

□

Additionally, as shown in Sec. 2, the variational inference approach learns the optimal policy π by maximising:

$$\arg\max_{\pi}\sum_t \mathbb{E}_\pi\left[\gamma^t r_t\right] - \alpha\,\text{KL}(\pi(a|s_t)\|p_{\text{LLM}}(a|s_t)). \tag{24}$$

Introducing the occupancy measure ρ, $\rho(s) = \frac{1}{1-\gamma}\sum_{t=0}^{\infty}[\gamma^t\mathbb{P}(s_t=s|a\sim\pi)]$, we have the following form objective respected to the Q-function

$$\arg\max_{\pi}\mathbb{E}_{\rho(s)}[\mathbb{E}_{a\sim\pi}[Q^\pi(s,a)] - \alpha\text{KL}(\pi(a|s_t)\|p_{\text{LLM}}(a|s_t))] \tag{25}$$

Combining Equations 23 and 25, we find that the variational inference approach and the direct posterior sampling yield similar solutions.

9 Additional Experiment Details

9.1 Environments

We consider three environments: ALFWorld, Overcooked and Frozen Lake.

ALFWorld We consider two classes of ALFWorld tasks: ALFWorld(pick) and ALFWorld(examine). For ALFWorld (Pick), we evaluate the online training baselines on 28 tasks, such as "put the cellphone on the armchair," and test the generalization ability on 26 unseen tasks. For ALFWorld (Examine),

	ALFWorld(Pick)	ALFWolrd(Examine)	Cook(Tomato)	Cook(Salad)	Frozen Lake
Horizon	60	60	30	50	20

Table 3: Maximize horizon of environments.

we use 11 tasks, such as "examine the laptop with the desk lamp." There are no auxiliary rewards, except for a reward of 1.0 for reaching the final goal.

Examples of observations and admissible action descriptions for three environments are shown below:

For ALFWorld(Pick)

Observation: Current observation:You are in the middle of a room. Looking quickly around you, you see a armchair 1, a bed 1, a diningtable 1, a drawer 2, a drawer 1, a garbagecan 1, a sidetable 2, and a sidetable 1.
Your task is to: put some cellphone on armchair..
Admissible actions You are allowed to take the following actions: go to armchair 1, go to bed 1, go to diningtable 1, go to drawer 1, go to drawer 2, go to garbagecan 1, go to sidetable 1, go to sidetable 2, inventory, look.

For Overcooked(Tomato)

Observation: Your task is to serve the dish of a bowl only containing chopped tomato. There is a fixed cutting board in the room. You notice a tomato on the table. Currently you don't have anything in hand.
Admissible actions You are allowed to take the following actions: pick up the tomato, take the bowl, walk to the cutting board, serve nothing, chop nothing.

For Frozen Lake

Observation: You are in a 4x4 grid where each cell is either a frozen cell (F) indicating a safe position, You must avoid to reach a hole (H), and the goal (G) that we aim to reach.
Input: Currently the agent is at position (0, 0), which direction need to take?
Admissible actions: Go Left leads to position (0, 0), start position. Go Right leads to position (0, 1), a frozen cell (F). Go Up leads to position (0, 0), start position. Go Down leads to position (1, 0), a frozen cell (F).
Admissible actions You are allowed to take the following actions: Go Left, Go Right, Go Up, Go Down.

9.2 LLM PRIOR IMPLEMENTATION

There are two ways to sample an action a from the LLM prior $p_{\text{LLM}}(\cdot|s_t)$. First, since the state and action can be described as text, and assuming the action a consists of k tokens, the probability of the LLM generating action a is given by $p(a|s_t) = \prod_{i=1}^{k} p_{\text{LLM}}(a_i|s_t, a_{<i})$. Based on this probability, the first type of LLM prior computes a distribution over actions, denoted as $p_{\text{LLM}}^{\text{dist}}(a|s_t) = \frac{\exp p(a|s_t)}{\sum_{a'} \exp p(a'|s_t)}$. In contrast, the second approach involves sampling a free-form output from the LLM, which is then mapped to an executable action via a simple rule-based projection $\mathcal{P}$. We denote this LLM prior, which relies on mapping the LLM's output, as $p_{\text{LLM}}^{\text{map}}$, where $a \sim p_{\text{LLM}}^{\text{map}}(\cdot|s_t) \leftrightarrow o \sim \text{LLM}(\cdot|s_t), a = \mathcal{P}(o)$.

We observed a performance gap between these two types of LLM action priors. For instance, in the ALFWorld pick & place task, $p_{\text{LLM}}^{\text{map}}$ achieved a winning rate of 19%, while $p_{\text{LLM}}^{\text{dist}}$ failed, with a winning rate of 0%, both using the LLM Qwen1.5-7B. Therefore, in this work, we implement $p_{\text{LLM}}^{\text{map}}$ as the LLM prior for thoughtfully exploiting the power of the LLM in SDM tasks, which we denote as p_{LLM} for simplicity. An Example of LLM's output is given as:

For ALFWorld(Pick) from Qwen-1.5 7B

Input: Current observation:You are in the middle of a room. Looking quickly around you, you see a armchair 1, a bed 1, a diningtable 1, a drawer 2, a drawer 1, a garbagecan 1, a sidetable 2, and a sidetable 1.
Your task is to: put some cellphone on armchair.. You are allowed to take the following actions: go to armchair 1, go to bed 1, go to diningtable 1, go to drawer 1, go to drawer 2, go to garbagecan 1, go to sidetable 1, go to sidetable 2, inventory, look. Please select an action from the admissible actions.
LLM Output to Action: 1: take cd 2 from sidetable 2 , ——> take cd 2 from sidetable 2
2: Open drawer 2 Result: You open door, ——> Open drawer 2
3: look Description: You look around the room., ——> look
4: look > Current observation:You arrive at loc, ——> look
5: open drawer 2 Loc 8: You open drawer 2 Loc 8 ——> open drawer 2 Loc 8
We map the LLM's output to an executable action using a simple rule-based method, extracting the executable actions directly from the LLM's output.

9.3 HYPER-PARAMETERS

Our algorithms are trained on one machine with 2 40G A100. based on Pytorch-GPU 2.1 and cuda12.4. Table 6,7,8,9,10 reports the main hyper-parameters of our algorithms. For all CQL-based algorithms, we set the hyperparameter β for regulating the Q-values, as shown in Eq. 7, as 5.0.

Offline Datasets The composition of offline datasets is shown in Table 11, which reports the number of (s, a, s') examples.

Table 4: The hyperparameters on ALFWorld(Pick)

Baselines	Learning Rate	Epochs	Batch Size	Update Frequency	LLM	α
DQN-Prior	5e-4	4	128	5	/	0.01
CQL-Prior	5e-4	4	128	5	/	0.01
GFlan-Prior	1e-4	16	64	16	Flan-T5 Small	0.01

Table 5: The hyperparameters on ALFWorld(Examine)

Baselines	Learning Rate	Epochs	Batch Size	Update Frequency	LLM	α
DQN-Prior	5e-4	4	128	10	/	0.01
CQL-Prior	5e-4	4	128	10	/	0.01
GFlan-Prior	1e-4	16	64	16	Flan-T5 Small	0.01

Table 6: The hyperparameters on Overcooked(Tomato)

Baselines	Learning Rate	Batch Size	Update Frequency	LLM	α
DQN-Prior	5e-4	128	5	/	0.1
CQL-Prior	5e-4	128	5	/	0.1
GFlan-Prior	1e-4	64	16	Flan-T5 Small	0.1

Table 7: The hyperparameters on Overcooked(Salad)

Baselines	Learning Rate	Batch Size	Update Frequency	LLM	α
DQN-Prior	5e-4	128	5	/	0.1
CQL-Prior	5e-4	128	5	/	0.1
GFlan-Prior	1e-4	64	16	Flan-T5 Small	0.1

Table 8: The hyperparameters on Frozen Lake

Baselines	Learning Rate	Batch Size	Update Frequency	LLM	α
DQN-Prior	5e-4	128	5	/	0.1

Table 9: The hyperparameters on ALFWorld(Pick)

Baselines	Learning Rate	Batch Size	Update Frequency	LLM	α
DQN-Prior	5e-4	128	10	/	0.01
CQL-Prior	5e-4	128	10	/	0.01
GFlan-Prior	1e-4	64	16	Flan-T5 Small	0.01

Table 10: The hyperparameters on ALFWorld(Examine)

Baselines	Learning Rate	Batch Size	Update Frequency	LLM	α
DQN-Prior	5e-4	128	10	/	0.01
CQL-Prior	5e-4	128	10	/	0.01
GFlan-Prior	1e-4	64	16	Flan-T5 Small	0.01

Dataset	Total Examples	Good Examples	Bad Examples
ALFWorld(Pick)	6572	6572	0
Salad(1000)	1186	574	612
Salad(4000)	3892	2209	1683
Salad(8000)	7571	4224	3347
Salad(12000)	11939	4004	7935
Salad(24000)	23869	7191	16678

Table 11: The composition of the offline dataset

9.4 Additional Results

This section presents additional experiments to further demonstrate the effectiveness of our method, including: 1) Our approach, only relying on LLM querying, is more time-efficient compared to the LLM fine-tuning-based approach. 2) Our approach is adaptable to partially observed BabyAI environments. 3) The drawback of frequently querying LLMs for action priors can be mitigated through caching techniques. 4) Inspired by the success of value-based DQN-Prior, conducting policy-based RL exclusively within the prior space is also feasible and significantly improves sample efficiency.

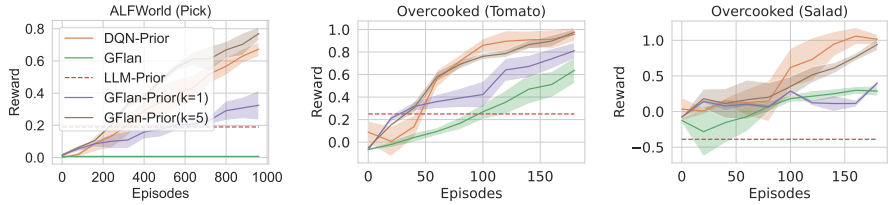


Figure 5: The ablation of the number of action proposals k used for approximating KL divergence between the required action policy and the LLM prior action distribution.

Additional ablation on k for GFLan-Prior As shown in Figure 5, when we sample $k = 5$ action proposals from the LLM to approximate the KL divergence, it performs better than when using $k = 1$. This is because, with more proposals, the optimal action is more likely to be included in the subset, thus better guiding the action policy.

Additional Ablation Study on the Caching Technique for Time Efficiency Most studies on RL fine-tuning for LLMs Tan et al. (2024); Christianos et al. (2023); Wen et al. (2024) face the

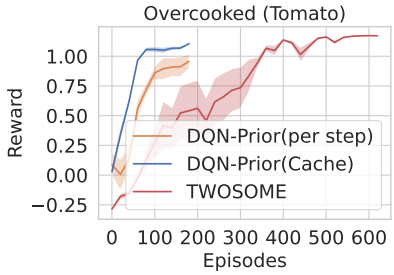


Figure 6: Comparison of DQN-Prior(Cache), DQN-Prior(query LLM per step) and TWOSOME on Overcooked (Tomato).

challenge of querying the LLM for each step. They use large LLMs (up to 7B parameters) as the action policy, and require querying the LLM at each step to collect data for online training. Since the LLM prior is fixed, we can resort to the caching technique within our framework to alleviate the need for querying the LLM at every step. We introduce a variant of the DQN-Prior algorithm, called DQN-Prior (Cache), which uses cached action candidates for previously seen observations and queries the LLM prior, storing its outputs only for new observations. We present the time and computational costs of our DQN-Prior(query LLM each time) and DQN-Prior (Cache), as well as the RLFT baseline, TWOSOME Tan et al. (2024), for learning to converge on Overcooked (Tomato) in Table 12. Results of baselines are shown in Figure 6. TWOSOME fine-tunes the LLM (Qwen-1.5 7B for a fair comparison) using the LoRA technique Hu et al. (2021) and the PPO algorithm. The time-consuming of our DQN-Prior(Cache) is significantly lower than the TWOSOME, while maintaining comparable performance. The caching technique is well-suited to our framework, as the 7B LLM is capable of providing a reliable sub-action space. DQN-Prior (Cache) significantly reduces query time costs but also achieves superior performance compared to DQN-Prior (query LLM per step), as frequent queries introduce more uncertainty.

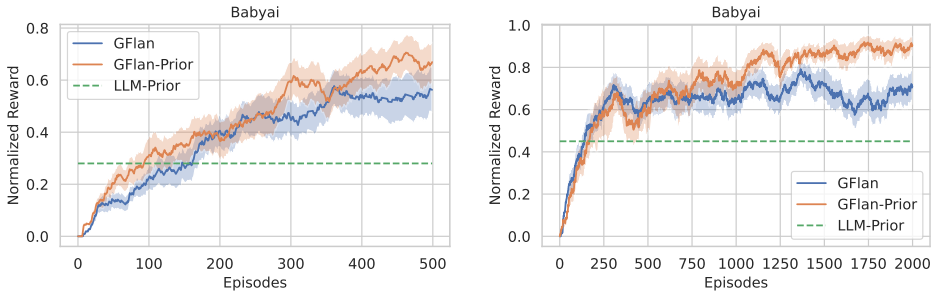
Table 12: The training costs of DQN-Prior (Cache), DQN-Prior (query LLM per step), and TWOSOME on Overcooked (Tomato) are reported for one training seed.

Baselines	Training Time(min)	GPU Usage	Query LLM(Qwen-1.5 7B) Times
DQN-Prior(per step)	31.1	28298 MB	2933
DQN-Prior(Cache)	13.8	28340 MB	434
TWOSOME	60.4	25014 MB	5000

Additional Experiment on BabyAI We compare our policy-based algorithm, GFlan-Prior, with GFlan in the text-based BabyAI environment Carta et al. (2023), which features partial observability and random map initialization for each episode. We focus on the BabyAI task 'Go to the red ball' and set the maximum horizon at 30. The comparison results are presented in Figure 7. Our GFlan-Prior, which integrates both LLM prior knowledge and environmental interaction experience, successfully accommodates the partially observed BabyAI and outperforms both GFlan and the LLM-prior.

Ablation Study on Using a Large LLM for the Q-Network We conducted an experiment on Overcooked(Tomato) to compare the effects of using an LLM as the action prior versus using it to encode state-action pairs. We evaluated three baselines:

- DQN-Prior: Uses Qwen-1.5 7B to generate a sub-action space and BERT to encode (s, a) pairs for Q-value estimation.
- DQN (BERT): Uses BERT to encode (s, a) pairs and performs DQN across the full action space.
- DQN (Qwen-1.5 7B): Uses Qwen-1.5 7B to encode (s, a) pairs and performs DQN across the full action space.



(a) Max Horizon as 30

(b) Max Horizon as 100

Figure 7: The comparison of policy-based algorithms on BabyAI.

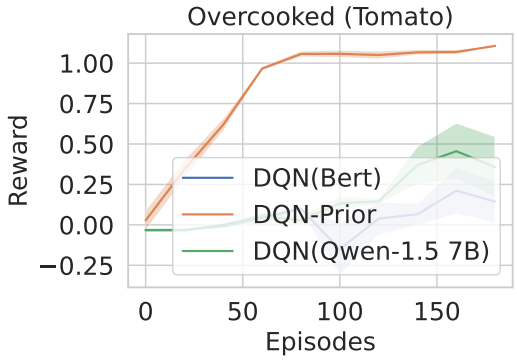


Figure 8: An ablation to verify the effectiveness of using the LLM to provide prior actions versus to encode state-action pairs.

To reduce the overhead of querying the LLM for every state, we store its outputs: For DQN-Prior, we store action proposals. For DQN (Qwen-1.5 7B), we store (s, a) embeddings.

As shown in the Figure 8 and Table 13, DQN-Prior outperforms both DQN (Qwen-1.5 7B) and DQN (BERT). By leveraging the LLM's high-level reasoning, DQN-Prior narrows the exploration space to a reliable action sub-space, improving sample efficiency. While DQN (Qwen-1.5 7B) benefits from the 7B LLM's superior semantic understanding and outperforms DQN (BERT), (s,a) embeddings cannot be directly used for action selection. Instead, a mapping from embeddings to Q-values must be learned from scratch to correctly rank the entire action space, thereby increasing sample complexity.

In addition, DQN-Prior requires less time and CPU memory compared to DQN (Qwen-1.5 7B). The higher memory and time consumption of DQN (Qwen-1.5 7B) are primarily due to its large embedding size (4096 dimensions), which significantly increases the cost of training a mapping (MLP adaptor) from state-action embeddings to Q-values, as well as the memory required to store these embeddings.

Table 13: The training costs of DQN-based baselines: DQN-Prior, DQN(BERT), DQN(Qwen1.5-7B). Both DQN (BERT) and DQN (Qwen-1.5 7B) involve the use of the 7B LLM. To minimize the overhead of querying the LLM for every state, we pre-store its outputs: for DQN-Prior, we store action proposals, while for DQN (Qwen-1.5 7B), we store (s, a) embeddings.

Baselines	Training Time(min)	GPU Usage	CPU usage	Embedding Dimension
DQN-Prior	13.8	28340 MB	18KB	768
DQN(BERT)	16.2	17864MB	/	768
DQN(Qwen-1.5 7B)	74.8	22662MB	22724MB	4096

Applying Direct Posterior Sampling on the Top of PPO Inspired by the success of direct posterior sampling in value-based reinforcement learning (RL), which effectively explores and updates the value function within the LLM prior action space, we recognize that LLMs can provide a narrow and reliable sub-action space. Building on this success, we aim to investigate policy-based RL within this action space. Specifically, we seek to learn a policy that approximates the soft Q-distribution over the LLM prior space:

$$\pi_\theta(a|s) \approx \frac{Q(s,a)}{\sum_{a' \in \mathcal{C}^k(s)} Q(s,a')}$$

In practice, we directly apply Proximal Policy Optimization (PPO) within the prior sub-action space. Formally, for the i-th iteration, the policy π_{θ_i} is learned by optimizing:

$$\arg\max_\theta \mathbb{E}_{\mathcal{C}^k(s_t) \sim p_{LLM}(\cdot|s_t), (s_t,a_t) \sim \pi_{\theta_{i-1}}(\mathcal{C}^k(s_t))} \text{CLIP}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{i-1}}(a_t|s_t)}\right) \hat{A}_t + \eta \mathcal{H}(\pi_\theta(\cdot|s_t)) \quad (26)$$

where A_t is the estimated advantage, η is the hyperparameter that encourages exploration. In the textual environments we considered, we used GFlan Carta et al. (2023) to train PPO with Flan-T5 small as the action policy. For simplicity, we denote the policy-based algorithm as GFlan-Prior(Sub-Action Space) or GFlan-Prior(Sub AS).

The comparison of policy-based RL with LLM prior is shown in Figure 9. GFlan-Prior(KL) indicates the PPO algorithm with the incorporation of LLM prior through the addition of a KL-penalty. These results demonstrate that incorporating LLM prior into PPO training, either by adding a KL-penalty or by providing a sub-optimal sub-action space, can improve the sample efficiency of RL training. Additionally, while both GFlan-Prior(KL) and GFlan-Prior(Sub AS) are able to successfully complete the final task with rewards approaching 1, GFlan-Prior(Sub AS) achieves higher rewards due to its explicit exploration and optimization within the sub-action space, which is more efficient than the implicit use of LLM prior by adding a KL penalty.

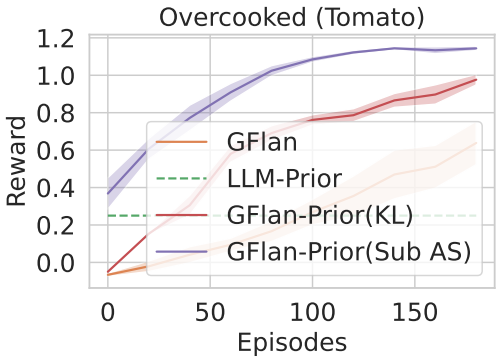


Figure 9: The comparison of policy-based algorithms with LLM prior.

10 Additional Related Works

Value-based Inference of LLM Not limited to the pre-training phase, it has been recently investigated how scaling law applies at inference time for LLM. This is often referred to as searching over an enormously large output space with guidance from value estimation. For example, the naive case is beam search (citation) which selects the top K possible sequence based on cumulative probabilities as opposed to greedy search. The value estimate here is the likelihood of the sentence as predicted by the language model. (Wang et al., 2022) propose CoT-SC, which searches over reasoning paths and selects the most frequent answer. Tree-of-Thought (ToT, (Yao et al., 2023)) adopts depth/breadth-first search. (Hao et al.) introduce Reasoning-via-Planning (RAP) which uses Monte-Carlo Tree Search and value estimate from prompting LLM. TS-LLM designed by (**?**) guides MCTS with a learned value function conditioned on state and a learned Outcome-supervised Reward Model (ORM). (Li et al., 2024) proposes Q-probing to adapt a pre-trained language model to also

maximize a tasks-specific reward function in code generation tasks. Furthermore, recent works investigate Process-supervised Reward Model (PRM, (Lightman et al., 2023)) and apply it in LLM inference, (McAleese et al., 2024) searches over a step-wise critic model in code reviewing tasks. (Wang et al., 2024a) learn a PRM in math reasoning and use it for LLM fine-tuning. (Snell et al., 2024) provides a detailed benchmark of different inference methods. (Zhang et al., 2024a) represents the score as the probability of a single text token (e.g. 'Yes' or 'No') under the context and the prompt. Even though our proposed method can be seen as an example of value-based inference, we focus on a different perspective where we investigate how LLM helps conventional value-based reinforcement learning algorithms in complex environments.

LLM-based Reward Function for RL Beyond serving as high-level planners for SDM tasks, LLMs are also utilized for generating reward signals, which are subsequently used for RL training. For example, (Kwon et al., 2023) leverages LLMs to directly generate reward signals prompted with historical interaction information. Motif (Klissarov et al., 2023) employs LLMs to annotate pairwise preference datasets, which are then used to train a reward function. Given the challenges LLMs face in directly generating low-level actions for robotics tasks with continuous representations, several studies (Yu et al., 2023; Ma et al., 2024) have instead utilized the coding capabilities of GPT-4 to generate human-level reward code. These reward codes are then employed to facilitate RL training.

More discussion can be found in Section 2.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：郑金兰
方向：人工智能
单位：东北师范大学

题目：人工智能基础模型与算法

Neural Networks 174 (2024) 106247

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

A novel predefined-time neurodynamic approach for mixed variational inequality problems and applications

Jinlan Zheng [a], Xingxing Ju [b,c], Naimin Zhang [d], Dongpo Xu [a,*]

[a] Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China
[b] College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
[c] Shaanxi Key Laboratory of Information Communication Network and Security, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China
[d] College of Mathematics and Physics, Wenzhou University, Wenzhou 325035, China

ARTICLE INFO

Keywords:
Mixed variational inequality
Neurodynamic approach
Predefined-time stability
Fixed-time stability
Spare signal recovery
Variational inequality
Game theory

ABSTRACT

In this paper, we propose a novel neurodynamic approach with predefined-time stability that offers a solution to address mixed variational inequality problems. Our approach introduces an adjustable time parameter, thereby enhancing flexibility and applicability compared to conventional fixed-time stability methods. By satisfying certain conditions, the proposed approach is capable of converging to a unique solution within a predefined-time, which sets it apart from fixed-time stability and finite-time stability approaches. Furthermore, our approach can be extended to address a wide range of mathematical optimization problems, including variational inequalities, nonlinear complementarity problems, sparse signal recovery problems, and nash equilibria seeking problems in noncooperative games. We provide numerical simulations to validate the theoretical derivation and showcase the effectiveness and feasibility of our proposed method.

1. Introduction

In various scientific and engineering fields, many problems can be formulated as variational inequality problems (VIPs). Variational inequalities problems are extensively used and highly significant mathematical models that can address a range of problems, including but not limited to composite minimization, saddle point, and nonlinear equations (see, e.g., Facchinei and Pang (2003), Kinderlehrer and Stampacchia (2000), Konnov (2001)). A natural and noteworthy progression of variational inequality problems (VIPs) are mixed variational inequality problems (MVIPs), which have garnered considerable attention and investigation in diverse domains such as optimization, game theory, control theory, and other related disciplines (see, e.g., Cavazzuti et al. (2002), Facchinei and Pang (2003), Kinderlehrer and Stampacchia (2000), Neittaanmäki and Tiba (1988), Scutari, Palomar, Facchinei, and Pang (2010)). The extensive research and application of MVIPs underscore their pivotal role as a significant mathematical model that can effectively tackle complex challenges in various scientific and engineering fields. MVIPs represent a natural and critical extension of VIPs, with their wider scope and applicability making them an indispensable tool for empirical and theoretical analysis in a broad range of applications.

In the realm of variational inequality problems and mixed variational inequality problems, a pivotal area of investigation pertains to the development of efficient methodologies for their resolution. Over the course of recent decades, researchers have devised several effective numerical techniques, including the golden ratio algorithms, inertial extragradient algorithms, splitting extragradient-like algorithms, and proximal algorithms, among others (see, e.g., Ju, Hu, Li, He, and Feng (2021), Malitsky (2020), Parikh et al. (2014), Pham and Trinh (2017), Thong and Van Hieu (2018)). A valuable tool employed in these methods is the proximal operator, which has also found application in addressing numerous other problem domains (Ju, Che, Li, He, & Feng, 2021). Numerical methods are commonly developed within the framework of discrete-time settings, making them well-suited for implementation on digital computers (see, e.g., Malitsky (2020), Parikh et al. (2014), Pham and Trinh (2017), Thong and Van Hieu (2018)). However, their limited efficacy in acquiring real-time online solutions poses a significant challenge. To tackle this issue, a promising avenue of exploration lies in the application of neural-dynamical methods, specifically utilizing recursive neural networks (RNNs), which can be effectively implemented through hardware circuits (see, e.g., Eshaghnezhad, Effati, and Mansoori (2016), Gao and Liao (2016), Hu and Wang (2006), Ju, Che, et al. (2021), Liu and Wang (2015), Vuong and Strodiot (2020)). It is noteworthy to highlight that these investigations predominantly concentrate on elucidating aspects related to asymptotic stability or exponential stability.

* Corresponding author.
E-mail address: xudp100@nenu.edu.cn (D. Xu).

https://doi.org/10.1016/j.neunet.2024.106247
Received 30 December 2023; Received in revised form 20 February 2024; Accepted 15 March 2024
Available online 16 March 2024
0893-6080/© 2024 Elsevier Ltd. All rights reserved.

The concept of finite-time stability, as introduced by Bhat and Bernstein (2000), has been harnessed to enhance the convergence speed of certain recurrent neural networks (RNNs) through the development of RNN approaches exhibiting finite-time (FINt) stability (Di Marco, Forti, Nistri, & Pancioni, 2015; Liu & Wang, 2011). The settling time of these networks is contingent upon the initial conditions and may experience unbounded expansion with increasing distance between the equilibrium point and initial conditions. However, obtaining precise information about the system's initial values proves challenging in practical applications, and estimating the stability time of the system presents a further obstacle. Polyakov introduced fixed-time stability in 2011 as a means of addressing the dependence on initial values in finite-time stability (Polyakov, 2012). This development enables the computation of an upper bound for stable time based on system parameters, irrespective of the initial values. The convergence analysis theory of neural networks has gradually shifted from asymptotic and exponential convergence to finite-time and fixed-time stability, forming a mature theoretical foundation (see e.g., Garg and Baranwal (2020), Garg, Baranwal, Gupta, and Benosman (2023), He, Wen, and Huang (2021), Ju, Hu, et al. (2021), Ju, Li, Han and He (2021), Zheng, Chen, and Ju (2022)). A novel modified proximal dynamical system was proposed to compute the solution of a mixed variational inequality problem (MVIP) within a fixed time in Garg et al. (2023). Additionally, a novel fixed-time converging forward–backward-forward neurodynamic network was investigated to deal with mixed variational inequalities in Ju, Hu, et al. (2021). However, despite the fact that fixed-time stability is superior to classical finite-time stability in a certain sense, it is generally difficult to find a clear relationship, leading to an overestimation of convergence time. The upper bound time is not flexible enough, making it challenging to adjust or predict synchronization time. Consequently, an increasing number of researchers are actively working towards developing less conservative fixed-time stability theorems. Fortunately, Sánchez-Torres proposed the concept of predefined-time stability in 2014 (Sánchez-Torres, Sanchez, & Loukianov, 2014). The stability time varies with the change of user parameters and is independent of the initial values. Users can control the stabilization time of the system by adjusting user parameters, which gives users more autonomy and facilitates users to adjust the time independently. This allows us to arbitrarily set user parameters to meet the user's control requirements for the system. In finite-time/fixed-time analysis, we only need to demonstrate that stability is achieved within a finite/fixed time, without the need to pre-estimate a parameter. In pre-defined time analysis, however, we need to pre-estimate parameters and use them during algorithm execution. The fundamental difference between fixed-time stability analysis and pre-defined time stability analysis lies in the fact that the former focuses on achieving stability within a certain time after running, while the latter involves pre-estimating a time and using it during execution, providing users with greater autonomy and control. There have been some related works on predefined-time research, including (Anguiano-Gijón, Muñoz-Vázquez, Sánchez-Torres, Romero-Galván, & Martínez-Reyes, 2019; Liu et al., 2022). An active control Lyapunov-function design for predefined-time synchronization of chaotic systems was proposed in Anguiano-Gijón et al. (2019). Two novel and general predefined-time stability lemmas were given and applied to the predefined-time synchronization problem in Liu et al. (2022). The use of the concept of predefined-time stability facilitates synchronization and information exchange between different systems. Currently, there are few research results on the combination of predefined-time stability and optimization problems.

To the authors' knowledge, the exploration of neurodynamic approaches utilizing predefined-time stability for addressing mixed variational inequalities has not been previously investigated, which serves as an additional motivation for this paper. Building upon the aforementioned research, this paper introduces a neural network with predefined-time stability (PeNN) to effectively tackle mixed variational inequalities. The notable contributions of this work, in comparison to existing literature, can be summarized as follows:

(i) The novel and more general lemma of predefined-time stability is introduced for the first time. The setting of predefined time T_c is more flexible. The derived lemma provides a more conservative time estimation compared to the time estimation in Chen et al. (2020).

(ii) The theoretical results ensure that the projected neural network can be directly used to solve mixed variational inequality optimization problems. Based on the theory, it is proved that the PeNN can also be applied to solve a wider range of constrained optimization problems and variational inequalities, linear and nonlinear complementary problems, and spare sign recovery problems, nash equilibria (NE) seeking problems in noncooperative games.

(iii) Through theoretical analysis and numerical results, it is shown that the proposed predefined-time convergence method has a faster convergence speed than the methods in Garg et al. (2023), Ju, Che, et al. (2021), Ju, Hu, et al. (2021), Zheng et al. (2022) in variational inequalities, linear and nonlinear complementary problems, spare sign recovery problems.

The present paper is organized as follows. Section 2 provides a comprehensive review of the fundamental definitions and concepts pertinent to the underlying subject matter. In Section 3, we introduce a novel proximal neural network with predefined-time stability (PeNN), for addressing the mixed variational inequality problems. The predefined-time stability of the PeNN is investigated in Section 4, where the upper bound of the convergence time is presented. A detailed account of the application and numerical simulation results of the related optimization problem are presented in Sections 5 and 6, respectively. Finally, the study concludes with a summary of the key findings in Section 7.

Notations: We adopt the following notation conventions throughout this paper: bold uppercase letters represent matrices, bold lowercase letters represent column vectors, standard lowercase letters represent scalars, standard letters represent real-valued functions, and bold letters represent vector-valued functions. Superscript$(\cdot)^\top$ and $|\cdot|$ denote transposed and absolute value, respectively. $\dot{()}$ means taking the derivative of its variable. I is the identity matrix. $\|\cdot\|$ represents the Euclidean 2-norm of the vector, $\|\cdot\|_1$ represents the Euclidean 1-norm of the vector and let $\mathbb{R}$ be the fields of real numbers. $\langle\cdot,\cdot\rangle$ is inner product in Euclidean space in $\mathbb{R}^n$. The definition of a projection operator is $P_\Omega(y) = \arg\min_{z\in\Omega}\|y-z\| := \{x \in \Omega : \|y-x\| \le \|y-z\|, \forall z \in \Omega\}$. The definition of the proximity operator $\mathrm{prox}_{\eta\Phi}(\mathbf{x})$ used in this paper is as follows (Bauschke & Combettes, 2017; Parikh et al., 2014):

$$\mathrm{prox}_{\eta\Phi}(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathbb{R}^n}\left\{\eta\Phi(\mathbf{y}) + \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\right\}. \tag{1}$$

2. Preliminaries

2.1. Blanket definitions

This section provides an overview of the essential definitions and problem formulations. We commence by introducing the concept of a mixed variational inequality problem. Find $\mathbf{x}^* \in \mathbb{R}^n$ such that the following inequality:

$$\langle \mathbf{f}(\mathbf{x}^*), \mathbf{y}-\mathbf{x}^*\rangle + \Phi(\mathbf{y}) - \Phi(\mathbf{x}^*) \ge 0 \quad \forall \mathbf{y} \in \mathbb{R}^n, \tag{2}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is proper convex and lower semi-continuous real-valued function, $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a vector-valued function.

Definition 1. Let $\Phi : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be a proper, lower semi-continuous, convex real-valued function, and $\mathbf{f} : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a vector-valued function, where Ω is a closed convex set. The function $\mathbf{f}(\cdot)$ is regarded as

(i) Strongly monotone, if there exists a constant $\gamma > 0$ such that

$$\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle \ge \gamma\|\mathbf{u}-\mathbf{v}\|^2, \quad \forall \mathbf{u},\mathbf{v} \in \Omega. \tag{3}$$

(ii) Φ-strong pseudomonotone, there exists a constant $\mu > 0$ such that the following inequality holds,

$$\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \Phi(\mathbf{v}) - \Phi(\mathbf{u}) \geq 0$$
$$\Rightarrow \langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \Phi(\mathbf{v}) - \Phi(\mathbf{u}) \geq \mu \|\mathbf{u} - \mathbf{v}\|^2, \quad \forall \mathbf{u}, \mathbf{v} \in \Omega. \tag{4}$$

(iii) the Lipschitz continuous with constant L, if for each pair of points $\mathbf{u}, \mathbf{v} \in \Omega$, such that

$$\|\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})\| \leq L\|\mathbf{u} - \mathbf{v}\|. \tag{5}$$

(iv) Non-expansive, if $\mathbf{f}$ is a Lipschitz continuous function with a constant $L = 1$, i.e.

$$\|\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})\| \leq \|\mathbf{u} - \mathbf{v}\|, \quad \forall \mathbf{u}, \mathbf{v} \in \Omega. \tag{6}$$

2.2. Blanket lemmas

This section lists the definitions and some of the lemmas required for this article.

We consider the following system:

$$\dot{\mathbf{x}}(t) = \mathbf{g}(\mathbf{x}(t)), \tag{7}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$, $t \in (0, \iota]$, $\iota > 0$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ is vector-function($\mathbf{g}(\mathbf{0}) = \mathbf{0}$). The origin is referred to as the equilibrium point of system (7). Without loss of generality, we assume that the origin is the unique equilibrium point of system (7).

Definition 2 (*Bhat & Bernstein, 2000; Polyakov, 2012; Sánchez-Torres, Gómez-Gutiérrez, López, & Loukianov, 2018*). Consider the system (7), and the origin of (7) is called
(i) Finite-time stability. The origin of (7) is deemed globally finite-time stable if it is globally asymptotically stability and if any solution $\mathbf{x}(t)$ of (7) with initial point $\mathbf{x}(0)$ reaches the equilibrium within a finite-time, i.e., $\forall t \geq T(\mathbf{x}(0)) : \mathbf{x}(t) = 0$, where $T : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ represents the settling-time function.
(ii) Fixed-time stability. The origin of (7) is fixed-time stable if it is globally finite-time stable and the settling-time function is bounded, i.e., $\exists T_{\max} > 0 : \forall \mathbf{x}(0) \in \mathbb{R}^n, T(\mathbf{x}(0)) \leq T_{\max}$.
(iii) Predefined time stability. If the parameter $T_c > 0$ can be adjusted to predict the settling time $T(\mathbf{x}(0))$ for achieving fixed-time stability, it implies that the driving response system (7) can attain global predefined time stability, where $T(\mathbf{x}(0)) \leq T_c$.

Lemma 1 (*Chen et al., 2020*). *Suppose $V : \mathbb{R}^n \to \mathbb{R}$ be a continuous, non-negative and unbounded radial function. Satisfying the following inequality:*

$$\dot{V}(\mathbf{x}(t)) \leq -\left(aV^{\alpha}(\mathbf{x}(t)) + bV^{\beta}(\mathbf{x}(t)) + cV(\mathbf{x}(t))\right), \quad \forall \mathbf{x}(t) \in \mathbb{R}^n \backslash \{\mathbf{0}\}, \tag{8}$$

where $a, b, c > 0$, $\alpha \in (0, 1)$, and $\beta > 1$. then, the system (7) achieves convergence within a fixed-time, where $T_{\max} = \frac{1}{c(1-\alpha)} \ln(1 + \frac{c}{a}) + \frac{1}{c(\beta-1)} \ln(1 + \frac{c}{b})$.

Lemma 2. *Consider the system (7). If there exists a positive definite unbounded radial function $V : \mathbb{R}^n \to \mathbb{R}$, and T_c is a user-defined parameter satisfying the following conditions:*
(i) $V(\mathbf{x}(t)) = 0 \Leftrightarrow \dot{\mathbf{x}}(t) = 0$.
(ii) *For any $V(\mathbf{x}(t)) > 0$, there exist $a, b, c, \alpha, \beta, T_c, G_c > 0$, where $\beta > 1$, and $0 < \alpha < 1$, such that the following inequality holds:*

$$\dot{V}(\mathbf{x}(t)) \leq -\frac{G_c}{T_c}\left(aV^{\alpha}(\mathbf{x}(t)) + bV^{\beta}(\mathbf{x}(t)) + cV(\mathbf{x}(t))\right), \tag{9}$$

then the system (7) achieves fixed-time stability within the predefined time T_c, where

$$G_c = \frac{1}{c(1-\alpha)} \ln(1 + \frac{c}{a}) + \frac{1}{c(\beta-1)} \ln(1 + \frac{c}{b}). \tag{10}$$

Proof. The proof can be found in Appendix A.1. □

Remark 1. (i) The parameter G_c in Lemma 2 can be regarded as the maximum upper bound of the stabilization time in fixed-time stability. For example, Lemma 2 can be rewritten as a new fixed-time stability theorem, which degenerates to the form of the Lyapunov function in Chen et al. (2020, Theorem 1, P414): $\dot{V}(\mathbf{x}(t)) \leq -\left(aV^{\alpha}(\mathbf{x}(t)) + bV^{\beta}(\mathbf{x}(t)) + cV(\mathbf{x}(t))\right)$, where $a, b, c > 0$, $\beta > 1, \alpha \in (0, 1)$. Then, the minimal estimate for the stabilization time is given by $T_{\max} = \frac{1}{c(1-\alpha)} \ln(1 + \frac{c}{a}) + \frac{1}{c(\beta-1)} \ln(1 + \frac{c}{b})$.
(ii) For $\forall x > 0$, $\ln(1 + x) < x$, we can observe that $\ln(1 + \frac{c}{a}) < \frac{c}{a}$, and $\ln(1 + \frac{c}{b}) < \frac{c}{b}$. Thus, we obtain $T_{\max} = \frac{1}{c(1-\alpha)} \ln(1 + \frac{c}{a}) + \frac{1}{c(\beta-1)} \ln(1 + \frac{c}{b}) < \frac{1}{a(1-\alpha)} + \frac{1}{b(\beta-1)}$. Therefore, compared with Polyakov (2012, Lemma 1, P2106), Lemma 2 can provide a more accurate estimate for $T_{\max}$.

Lemma 3. *Suppose $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^n$ be a locally Lipschitz continuous vector field, satisfying*

$$\boldsymbol{\phi}(\mathbf{x}^*) = \mathbf{0} \text{ and } \langle \mathbf{x} - \mathbf{x}^*, \boldsymbol{\phi}(\mathbf{x}) \rangle > 0 \tag{11}$$

$\forall \mathbf{x} \in \mathbb{R}^n \backslash \{\mathbf{x}^*\}$, where $\mathbf{x}^* \in \mathbb{R}^n$. Consider the following differential equation:

$$\dot{\mathbf{x}}(t) = -\sigma(\mathbf{x}(t))\boldsymbol{\phi}(\mathbf{x}(t)), \tag{12}$$

where

$$\sigma(\mathbf{x}) := \begin{cases} \rho_1 \frac{1}{\|\boldsymbol{\phi}(\mathbf{x})\|^{1-\alpha_1}} + \rho_2 \frac{1}{\|\boldsymbol{\phi}(\mathbf{x})\|^{1-\alpha_2}} + \rho_3, & if\ \boldsymbol{\phi}(\mathbf{x}) \neq 0, \\ 0, & else, \end{cases} \tag{13}$$

and $\rho_1, \rho_2, \rho_3 > 0$, $\alpha_1 \in (0, 1)$, $\alpha_2 > 1$. Then the right-hand side of (12) is continuous for all $\mathbf{x} \in \mathbb{R}^n$, and for any given initial condition, the solution of (12) exists and is uniquely determined for all $t \geq 0$.

Proof. The proof can be found in Appendix A.2. □

3. The proximal neural network with predefined-time stability

This section presents a proposed proximal neural network for solving mixed variational inequalities. Based on the literature (Chen et al., 2020; Ju, Hu, et al., 2021; Liu et al., 2022), a proximal neural network with predefined-time stability (PeNN) is proposed to solve mixed variational inequalities.

$$\dot{\mathbf{x}} = -\frac{G_c}{T_c}\rho(\mathbf{x})\left(\mathbf{x} - \boldsymbol{\omega}(\mathbf{x}) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}\left(\boldsymbol{\omega}(\mathbf{x})\right)\right), \tag{14}$$

$$\rho(\mathbf{x}) = \begin{cases} \rho_1 \frac{1}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_1}} + \rho_2 \frac{1}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_2}} + \rho_3, & \text{if } \mathbf{x} \in \mathbb{R}^n \backslash \mathrm{Fix}(\boldsymbol{\omega}), \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where G_c is constant, T_c is user-defined parameter, $\boldsymbol{\omega}(\mathbf{x}) = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))$ and $\eta > 0$, $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{x} - \boldsymbol{\omega}(\mathbf{x}) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}\left(\boldsymbol{\omega}(\mathbf{x})\right)$, $\rho_1, \rho_2, \rho_3 > 0$, $\beta_1 \in (0, 1), \beta_2 > 1$ and $\mathrm{Fix}(\boldsymbol{\omega}) := \{\mathbf{x} \in \mathbb{R}^n | \boldsymbol{\omega}(\mathbf{x}) = \mathbf{x}\}$.

We list common assumptions in this article and provides explanations.

Assumption 1. $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is Φ-strong pseudomonotone with coefficient μ, see Definition 1 (ii).

Assumption 2. $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with coefficient $L > 0$, see Definition 1 (iii).

Assumption 3. $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is γ-strong pseudomonotone, see Definition 1 (i).

Remark 2. Assumption 1 guarantees the existence of the solution for MVIPs (2). Assumptions 1 and 2/3 guarantee the uniqueness of the solution for MVIPs (2) in Ju, Hu, et al. (2021). Assumptions 1 and 2/3 guarantee fixed-time convergence. Similar assumptions are presented in relevant literature (see, e.g. Garg et al. (2023), Ju, Che, et al. (2021), Ju, Hu, et al. (2021), Vuong and Strodiot (2020)). At the same time, this Assumption 2 or 3 is relatively weak compared to the strong monotonic condition in Garg et al. (2023).

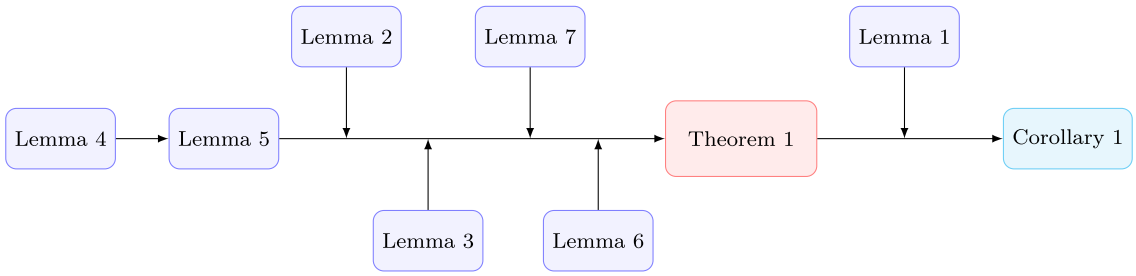


Fig. 1. Relations of the lemmas, theorem and corollary in Section 4.

4. Convergence analysis of the PeNN for MVIPs

In this section, we give the necessary lemmas and relevant conclusions for convergence analysis. Fig. 1 illustrates the interconnection between the theorem, lemmas and corollary presented in this section.

Lemma 4. *Assume $\eta > 0$ and $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a vector-valued function, and* prox *is the operator defined in(1). Then, $\mathbf{x} = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))$ iff $\langle \mathbf{f}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \Phi(\mathbf{y}) - \Phi(\mathbf{x}) \geq 0, \quad \forall \mathbf{y} \in \mathbb{R}^n$.*

Proof. It follows from Bauschke and Combettes (2017, Proposition 12.26, P175), we get

$$\begin{aligned}\mathbf{x} = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x})) &\Leftrightarrow \langle \mathbf{x} - \eta\mathbf{f}(\mathbf{x}) - \mathbf{x}, \mathbf{y} - \mathbf{x}\rangle + \eta\Phi(\mathbf{x}) \leq \eta\Phi(\mathbf{y}) \\ &\Leftrightarrow \eta\langle \mathbf{f}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \eta\Phi(\mathbf{y}) - \eta\Phi(\mathbf{x}) \geq 0 \\ &\Leftrightarrow \langle \mathbf{f}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \Phi(\mathbf{y}) - \Phi(\mathbf{x}) \geq 0, \quad \forall \mathbf{y} \in \mathbb{R}^n. \end{aligned} \tag{16}$$

Therefore, $\mathbf{x} = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))$ iff $\langle \mathbf{f}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \Phi(\mathbf{y}) - \Phi(\mathbf{x}) \geq 0, \quad \forall \mathbf{y} \in \mathbb{R}^n$. □

Lemma 5. *The point $\mathbf{x}$ is an equilibrium point of the PeNN (14), i.e. $\dot{\mathbf{x}} = \mathbf{0}$ iff $\mathbf{x}$ is a solution of MVIPs (2), that is $\mathbf{x} = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))$.*

Proof. Sufficiency: It is obvious that when $\mathbf{x} = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))$, $\mathbf{x} = \boldsymbol{\omega}(\mathbf{x})$, the PeNN (14) reaches a stable point, i.e., $\dot{\mathbf{x}} = 0$. Necessity: Assuming that the point $\mathbf{x}$ is an equilibrium point of the method (14), then

$$\begin{aligned}&-\rho(\mathbf{x})\left(\mathbf{x} - \boldsymbol{\omega}(\mathbf{x}) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}\left(\boldsymbol{\omega}(\mathbf{x})\right)\right) = 0 \\ &\Leftrightarrow \boldsymbol{\psi}(\mathbf{x}) = 0 \quad \text{or} \quad \mathbf{x} \in \mathrm{Fix}(\boldsymbol{\omega}).\end{aligned} \tag{17}$$

This indicates that $\mathbf{x} = \boldsymbol{\omega}(\mathbf{x})$. By Lemma 4, the solution of this system (14) are equivalent to the solution of the mixed variational inequality. □

Lemma 6 (*Ju, Che, et al., 2021*). *Assume that $\mathbf{x}^*$ is a solution of the MVIPs (2). Assumptions 1 and 2 hold, with $\eta L < 1$, then*

$$\text{(i)}\langle \mathbf{x} - \mathbf{x}^*, -\mathbf{x} + \boldsymbol{\omega}(\mathbf{x}) + \eta\mathbf{f}(\mathbf{x}) - \eta\mathbf{f}(\boldsymbol{\omega}(\mathbf{x}))\rangle \leq -\theta\|\mathbf{x} - \mathbf{x}^*\|^2; \tag{18}$$

$$\text{(ii)}\theta\|\mathbf{x} - \mathbf{x}^*\| \leq \|\mathbf{x} - \boldsymbol{\omega}(\mathbf{x}) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}(\boldsymbol{\omega}(\mathbf{x}))\|, \tag{19}$$

where $\boldsymbol{\omega}(\mathbf{x}) = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))$, and $\theta = \frac{(1-\eta L)\eta^2\mu^2}{(1+\eta L+\eta\mu)^2}$.

Lemma 7 (*Ju, Hu, et al., 2021*). *Assumption 2 holds, we have*

$$\|\boldsymbol{\psi}(\mathbf{u}) - \boldsymbol{\psi}(\mathbf{v})\| \leq \xi\|\mathbf{u} - \mathbf{v}\|, \tag{20}$$

where $\xi = 2 + 3\eta L + \eta^2 L^2$. Furthermore, assuming that $\mathbf{x}^*$ is a solution to the MVIPs (2), we have

$$\|\boldsymbol{\psi}(\mathbf{x})\| = \|\mathbf{x} - \boldsymbol{\omega}(\mathbf{x}) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}\left(\boldsymbol{\omega}(\mathbf{x})\right)\| \leq \xi\|\mathbf{x} - \mathbf{x}^*\|, \tag{21}$$

where $\xi = 2 + 3\eta L + \eta^2 L^2$, $\boldsymbol{\omega}(\mathbf{x}) = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))$.

Theorem 1. *Assumptions 1 and 2 hold, $\mathbf{x}^*$ is a solution to the mixed variational inequality (2), with $\eta L < 1$, then the PeNN (14) can achieve the equilibrium point within a predefined time T_c, where T_c is a user-defined parameter, and $G_c = \frac{1}{\rho_3\theta(1-\beta_1)}\ln(1 + \frac{\rho_3\sqrt{2}^{1-\beta_1}\xi^{1-\beta_1}}{\rho_1}) + \frac{1}{\rho_3\theta(\beta_2-1)}\ln(1 + \frac{\rho_3\sqrt{2}^{1-\beta_2}\theta^{1-\beta_2}}{\rho_2})$, $\theta = \frac{(1-\eta L)\eta^2\mu^2}{(1+\eta L+\eta\mu)^2}$ and $\xi = 2 + 3\eta L + \eta^2 L^2$.*

Proof. By Lemma 7 and (18), all the conditions of Lemma 3 are satisfied. Therefore, it can be concluded that the solution to the PeNN (14) exists uniquely. According to Lemma 5, $\mathbf{x}^* \in \mathbb{R}^n$ is an equilibrium point of the PeNN (14). Consider the following Lyapunov function

$$V(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|^2. \tag{22}$$

Differentiating function (22) with respect to time t along with (14) and for all $\forall\mathbf{x}(0) \notin \mathrm{Fix}(\boldsymbol{\omega}) = \{\mathbf{x}^*\}$, we obtain

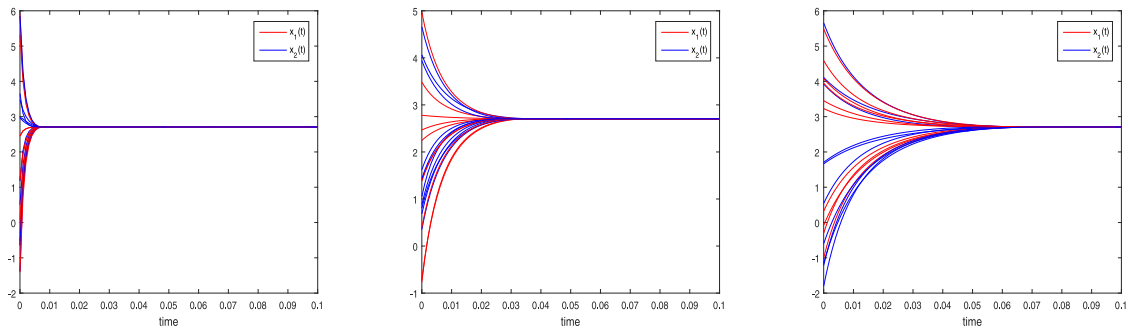
$$\begin{aligned}\dot{V} &= \langle \mathbf{x} - \mathbf{x}^*, \dot{\mathbf{x}}\rangle \\ &\overset{(14)}{=} -\frac{G_c}{T_c}\left\langle \mathbf{x} - \mathbf{x}^*, \rho_1\frac{\boldsymbol{\psi}(\mathbf{x})}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_1}} + \rho_2\frac{\boldsymbol{\psi}(\mathbf{x})}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_2}} + \rho_3\boldsymbol{\psi}(\mathbf{x})\right\rangle \\ &\overset{(18)}{\leq} \frac{G_c}{T_c}(-\rho_1\theta\frac{\|\mathbf{x} - \mathbf{x}^*\|^2}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_1}} - \rho_2\theta\frac{\|\mathbf{x} - \mathbf{x}^*\|^2}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_2}} - \rho_3\theta\|\mathbf{x} - \mathbf{x}^*\|^2) \\ &\overset{(21)}{\leq} \frac{G_c}{T_c}(-\frac{\rho_1\theta}{\xi^{1-\beta_1}}\|\mathbf{x} - \mathbf{x}^*\|^{1+\beta_1} - \rho_2\theta\frac{\|\mathbf{x} - \mathbf{x}^*\|^2}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_2}} - \rho_3\theta\|\mathbf{x} - \mathbf{x}^*\|^2) \\ &\overset{(19)}{\leq} \frac{G_c}{T_c}(-\frac{\rho_1\theta}{\xi^{1-\beta_1}}\|\mathbf{x} - \mathbf{x}^*\|^{1+\beta_1} - \rho_2\theta^{\beta_2}\|\mathbf{x} - \mathbf{x}^*\|^{1+\beta_2} - \rho_3\theta\|\mathbf{x} - \mathbf{x}^*\|^2) \\ &\overset{(22)}{=} \frac{G_c}{T_c}(-\frac{\rho_1\theta}{\xi^{1-\beta_1}}(2V)^{\frac{1+\beta_1}{2}} - \rho_2\theta^{\beta_2}(2V)^{\frac{1+\beta_2}{2}} - 2\rho_3\theta V) \\ &= \frac{G_c}{T_c}(-(\sqrt{2}^{1+\beta_1}\frac{\rho_1\theta}{\xi^{1-\beta_1}}V^{\frac{1+\beta_1}{2}} + \sqrt{2}^{1+\beta_2}\rho_2\theta^{\beta_2}V^{\frac{1+\beta_2}{2}} + 2\rho_3\theta V)) \\ &= \frac{G_c}{T_c}(-\tau_1 V^{\frac{1+\beta_1}{2}} - \tau_2 V^{\frac{1+\beta_2}{2}} - \tau_3 V),\end{aligned} \tag{23}$$

where $\tau_1 = \sqrt{2}^{1+\beta_1}\frac{\rho_1\theta}{\xi^{1-\beta_1}} > 0$, $\tau_2 = \sqrt{2}^{1+\beta_2}\rho_2\theta^{\beta_2} > 0$, $\tau_3 = 2\rho_3\theta$, $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{x} - \boldsymbol{\omega}(\mathbf{x}) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}\left(\boldsymbol{\omega}(\mathbf{x})\right)$. The first inequality is due to (18) in Lemma 6, and the second inequality is obtained from (21) in Lemma 7 and (19) in Lemma 6. Take

$$\begin{aligned}G_c &= \frac{2}{\tau_3(1-\beta_1)}\ln(1 + \frac{\tau_3}{\tau_1}) + \frac{2}{\tau_3(\beta_2-1)}\ln(1 + \frac{\tau_3}{\tau_2}) \\ &= \frac{1}{\rho_3\theta(1-\beta_1)}\ln(1 + \frac{\rho_3\sqrt{2}^{1-\beta_1}\xi^{1-\beta_1}}{\rho_1}) \\ &\quad + \frac{1}{\rho_3\theta(\beta_2-1)}\ln(1 + \frac{\rho_3\sqrt{2}^{1-\beta_2}\theta^{1-\beta_2}}{\rho_2}),\end{aligned} \tag{24}$$

where $\theta = \frac{(1-\eta L)\eta^2\mu^2}{(1+\eta L+\eta\mu)^2}$, $\xi = 2+3\eta L+\eta^2 L^2$. Then, by (23) and Lemma 2, it follows that the equilibrium point of the PeNN (14) is reached within a predefined time T_c. For any $\mathbf{x}(0) \in \mathbb{R}^n$, we have $T(\mathbf{x}(0)) \leq T_{max} = T_c$. □

Remark 3. (i) G_c is a pre-estimated fixed-time estimate that is calculated after setting a parameter. The T_c is a user-defined parameter that allows the user to control the algorithm to achieve stability within



(a) (b) (c)

Fig. 2. Trajectory behavior of $\mathbf{x}(t)$ in PeNN (14) for ten random initial conditions: (a) $T_c = 1$; (b) $T_c = 5$; (c) $T_c = 10$.

a certain time according to their needs. G_c will be automatically calculated after setting other parameters $(\rho_1, \rho_2, \rho_3, \beta_1, \beta_2, \mu, L)$, while T_c is a user-defined parameter.
(ii) The setting of predefined time T_c is more flexible. Users can control the stabilization time of the system by adjusting user parameters, which gives users more autonomy and facilitates users to adjust the time independently.
(iii) In Example 1, from Fig. 2, we can find different T_c values in Fig. 2, which it can control different convergence times, and converge to the unique point $\mathbf{x}^* = (2.707, 2.707)^\top$. This also confirms the theory of predefined time stability. In Examples 1, 2, 3, and 4, it is shown that the proposed predefined-time convergence method has a faster convergence speed than the methods in Garg et al. (2023), Ju, Che, et al. (2021), Ju, Hu, et al. (2021), Zheng et al. (2022).

Corollary 1. *Assumptions 1 and 2 hold, $\mathbf{x}^*$ is a solution to the mixed variational inequality (2), $T_c = G_c$, and $\eta L < 1$, then the PeNN (14) can achieve stability within a fixed time, and the convergence time is $T(\mathbf{x}_0) \le T_{\max} = \frac{1}{\rho_3\theta(1-\beta_1)}\ln(1+\frac{\rho_3\sqrt{2}^{1-\beta_1}\xi^{1-\beta_1}}{\rho_1}) + \frac{1}{\rho_3\theta(\beta_2-1)}\ln(1+\frac{\rho_3\sqrt{2}^{1-\beta_2}\theta^{1-\beta_2}}{\rho_2})$, $\theta = \frac{(1-\eta L)\eta^2\mu^2}{(1+\eta L+\eta\mu)^2}$ and $\xi = 2 + 3\eta L + \eta^2 L^2$.*

Proof. The proof can be obtained by Theorem 1 and Lemma 1. According to Lemma 5, $\mathbf{x}^* \in \mathbb{R}^n$ is an equilibrium point of the PeNN (14). Consider the following Lyapunov function

$$V(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|^2. \tag{25}$$

Differentiating function (25) with respect to time t along with (14) and for all $\forall \mathbf{x}(0) \notin \mathrm{Fix}(\boldsymbol{\omega}) = \{\mathbf{x}^*\}$, we obtain

$$\begin{aligned}
\dot{V} &= \langle \mathbf{x} - \mathbf{x}^*, \dot{\mathbf{x}} \rangle \\
&\overset{(14)}{=} -\left\langle \mathbf{x} - \mathbf{x}^*, \rho_1 \frac{\boldsymbol{\psi}(\mathbf{x})}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_1}} + \rho_2 \frac{\boldsymbol{\psi}(\mathbf{x})}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_2}} + \rho_3 \boldsymbol{\psi}(\mathbf{x}) \right\rangle \\
&\overset{(18)}{\le} -\rho_1\theta \frac{\|\mathbf{x} - \mathbf{x}^*\|^2}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_1}} - \rho_2\theta \frac{\|\mathbf{x} - \mathbf{x}^*\|^2}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_2}} - \rho_3\theta\|\mathbf{x} - \mathbf{x}^*\|^2 \\
&\overset{(21)}{\le} -\frac{\rho_1\theta}{\xi^{1-\beta_1}}\|\mathbf{x} - \mathbf{x}^*\|^{1+\beta_1} - \rho_2\theta \frac{\|\mathbf{x} - \mathbf{x}^*\|^2}{\|\boldsymbol{\psi}(\mathbf{x})\|^{1-\beta_2}} - \rho_3\theta\|\mathbf{x} - \mathbf{x}^*\|^2 \\
&\overset{(19)}{\le} -\frac{\rho_1\theta}{\xi^{1-\beta_1}}\|\mathbf{x} - \mathbf{x}^*\|^{1+\beta_1} - \rho_2\theta^{\beta_2}\|\mathbf{x} - \mathbf{x}^*\|^{1+\beta_2} - \rho_3\theta\|\mathbf{x} - \mathbf{x}^*\|^2 \\
&\overset{(25)}{=} -\frac{\rho_1\theta}{\xi^{1-\beta_1}}(2V)^{\frac{1+\beta_1}{2}} - \rho_2\theta^{\beta_2}(2V)^{\frac{1+\beta_2}{2}} - 2\rho_3\theta V \\
&= -(\sqrt{2}^{1+\beta_1}\frac{\rho_1\theta}{\xi^{1-\beta_1}}V^{\frac{1+\beta_1}{2}} + \sqrt{2}^{1+\beta_2}\rho_2\theta^{\beta_2}V^{\frac{1+\beta_2}{2}} + 2\rho_3\theta V) \\
&= -\tau_1 V^{\frac{1+\beta_1}{2}} - \tau_2 V^{\frac{1+\beta_2}{2}} - \tau_3 V,
\end{aligned} \tag{26}$$

where $\tau_1 = \sqrt{2}^{1+\beta_1}\frac{\rho_1\theta}{\xi^{1-\beta_1}} > 0$, $\tau_2 = \sqrt{2}^{1+\beta_2}\rho_2\theta^{\beta_2} > 0$, $\tau_3 = 2\rho_3\theta$, $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{x} - \boldsymbol{\omega}(\mathbf{x}) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}(\boldsymbol{\omega}(\mathbf{x}))$. Then, by (26) and Lemma 1, it follows that the equilibrium point of the PeNN (14) is reached within a fixed time. That is, for any $\mathbf{x}(0) \in \mathbb{R}^n$, we have $T(\mathbf{x}(0)) \le T_{max} = \frac{1}{\rho_3\theta(1-\beta_1)}\ln(1+\frac{\rho_3\sqrt{2}^{1-\beta_1}\xi^{1-\beta_1}}{\rho_1}) + \frac{1}{\rho_3\theta(\beta_2-1)}\ln(1+\frac{\rho_3\sqrt{2}^{1-\beta_2}\theta^{1-\beta_2}}{\rho_2})$. □

Remark 4. From Theorem 1 and Corollary 1, it can be seen that if we have limited information about the function $\mathbf{f}(\cdot)$, we can make no requirements on T_c and G_c, and the PeNN (14) would degenerate to fixed-time stability. In this case, we can give an upper bound for the time required for the system (14) to reach stability that is independent of the initial conditions, and this bound is more accurate than that given in Ju, Hu, et al. (2021).

5. Applications

In this section, we provide an exposition of the diverse range of applications of the PeNN (2). Specifically, we explore its utility in addressing various problem domains including linear and nonlinear complementarity problems (NCPs), variational inequalities, spare signal recovery as well as NE seeking problems in noncooperative games.

5.1. Variational inequality (VI)

When the function Φ takes the form of an indicator function, the mixed variational inequality problem (2) degenerates into the subsequent variational inequality problem.

$$\langle \mathbf{f}(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \ge 0, \ \forall \mathbf{y} \in \Omega. \tag{27}$$

In this scenario, the proximity operator (1) transforms into the projection operator $P_\Omega(\mathbf{x}) = \arg\min_{\mathbf{y}\in\Omega}\|\mathbf{x} - \mathbf{y}\|$. Consequently, the PeNN (14) reduces into the following form:

$$\dot{\mathbf{x}} = -\frac{G_c}{T_c}\rho(\mathbf{x})\left(\mathbf{x} - P_\Omega(\mathbf{x} - \eta\mathbf{f}(\mathbf{x})) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}\left(P_\Omega(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))\right)\right) \tag{28}$$

$$\rho(\mathbf{x}) = \begin{cases} \rho_1\frac{1}{\|\boldsymbol{\phi}(\mathbf{x})\|^{1-\beta_1}} + \rho_2\frac{1}{\|\boldsymbol{\phi}(\mathbf{x})\|^{1-\beta_2}} + \rho_3, & \text{if } \mathbf{x} \in \mathbb{R}^n \setminus \mathrm{Fix}(\boldsymbol{\omega}), \\ 0, & \text{otherwise} \end{cases}$$

where $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x} - P_\Omega(\mathbf{x} - \eta\mathbf{f}(\mathbf{x})) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}(P_\Omega(\mathbf{x} - \eta\mathbf{f}(\mathbf{x})))$ and $\eta > 0$, $\mathrm{Fix}(\boldsymbol{\omega}) := \{\mathbf{x} \in \mathbb{R}^n | x = P_\Omega(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))\}$, $\rho_1 > 0, \rho_2 > 0, \rho_3 > 0$, $\beta_1 \in (0, 1), \beta_2 > 1$.

5.2. Nonlinear complementarity problem (NCP)

Consider a NCPs: find $\mathbf{x} \in \mathbb{R}^n$ that satisfies the conditions

$$\mathbf{f}(\mathbf{x}) \ge \mathbf{0}, \quad \mathbf{x} \ge \mathbf{0}, \quad \langle \mathbf{x}, \mathbf{f}(\mathbf{x}) \rangle = 0.$$

Lemma 5 states that $\mathbf{x}^*$ is a solution to the NCP iff $\mathbf{x}^*$ is a solution to the following nonlinear projection equation:

$$P_\Omega(\mathbf{x} - \alpha\mathbf{f}(\mathbf{x})) = \mathbf{x}, \quad \forall \alpha > 0.$$

Therefore, the PeNN (14) can effectively be used to solve the nonlinear complementarity problem (NCP). In particular, if the function $\mathbf{f}(\cdot)$ is affine, denoted as $\mathbf{f}(\mathbf{x}) = \mathbf{M}\mathbf{x} + \mathbf{q}$, with $\mathbf{M}$ being an $n \times n$ matrix and $\mathbf{q}$ being an $n \times 1$ vector, then the problem reduces to a linear complementarity problem (LCP). It is worth noting that the solution $\mathbf{x}^*$ satisfies the LCP iff it satisfies the ensuing linear projection

$$P_{\Omega_0}(\mathbf{x} - \alpha(\mathbf{M}\mathbf{x} + \mathbf{q})) = \mathbf{x}, \quad \forall \alpha > 0,$$

where $\Omega_0 = \{\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \geq 0\}$. Therefore, by substituting $\mathbf{f}(\mathbf{x}) = \mathbf{M}\mathbf{x} + \mathbf{q}$ into (28), the LCP problem can be effectively solved.

5.3. Sparse signal recovery (SSR)

Let $\mathbf{x} = [x_1, x_2, \dots, x_n]^\top \in \mathbb{R}^n$ represent an unknown k-sparse signal that requires reconstruction from limited measurements.

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}, \tag{29}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a measurement matrix with $m \ll n$, $\mathbf{y} \in \mathbb{R}^m$ is the observed matrix disturbed by noise $\varepsilon \in \mathbb{R}^m$. The sparse recovery problem (29) can be transformed into the following composite minimization problem (CMP) in Candes and Tao (2009), Garg and Baranwal (2020), Yu, Zheng, and Barbot (2017), that is

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + g(\mathbf{x}) \tag{30}$$

where $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, $g(\mathbf{x}) = \mu\|\mathbf{x}\|_1$, with $\mu > 0$ serving as the balancing parameter. In general, for a sparse vector $\mathbf{x}$ with sparsity, under the constraint isometry property condition, the solution to the SSR problem (29) is shown to be unique and consistent with the solution to CMP (30) in Candes and Tao (2009). Notably, it has been demonstrated in Garg and Baranwal (2020, Lemma 1) that the equivalence between MVIPs (2) and CMP (30) can be established when considering the mappings $\mathbf{f}(\mathbf{x}) = (\mathbf{A}^\top\mathbf{A})(\mathbf{x}) - \mathbf{A}^\top\mathbf{y}$ and $\Phi(\mathbf{x}) = g(\mathbf{x})$. Furthermore, the PeNN (14) in the context of sparse problems is reconfigured into a new form denoted as PeNNssr, expressed as

$$\dot{\mathbf{x}} = -\rho(\mathbf{x})\left(\mathbf{x} - \boldsymbol{\omega}(\mathbf{x}) - \eta\mathbf{f}(\mathbf{x}) + \eta\mathbf{f}(\boldsymbol{\omega}(\mathbf{x}))\right) \tag{31}$$

where $\rho(\mathbf{x})$ is defined by (15), $\mathbf{f}(\mathbf{x}) = (\mathbf{A}^\top\mathbf{A})(\mathbf{x}) - \mathbf{A}^\top\mathbf{y}$, $\boldsymbol{\omega}(\mathbf{x}) = \mathrm{prox}_{\eta\Phi}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x})) = max(|\mathbf{x} - \eta\mathbf{f}(\mathbf{x})| - \eta\mu, 0) \cdot \mathrm{sign}(\mathbf{x} - \eta\mathbf{f}(\mathbf{x}))$, sign is a sign operator. The L-continuity and strong monotonicity of $\mathbf{f}(\mathbf{x})$ have been established in Garg and Baranwal (2020), leading to the straightforward derivation of subsequent results.

Theorem 2. *Let $\mathbf{x}^*$ is the solution of the CMP (30). Assumptions 1 and 2 hold with $\eta\lambda_{\max}(\mathbf{A}^\top\mathbf{A}) < \frac{1}{2}$, then the PeNNssr (31) converges to the solution within a fixed time, and $T(\mathbf{x}(0)) \leq T_{\max} = \frac{1}{\rho_3\theta(1-\beta_1)}\ln(1 + \frac{\rho_3\sqrt{2}^{1-\beta_1}\xi^{1-\beta_1}}{\rho_1}) + \frac{1}{\rho_3\theta(\beta_2-1)}\ln(1 + \frac{\rho_3\sqrt{2}^{1-\beta_2}\theta^{1-\beta_2}}{\rho_2})$, where $\theta = \frac{(1-\eta L)\eta^2\mu^2}{(1+\eta L+\eta\mu)^2}$ and $\xi = 2+3\eta L+\eta^2 L^2$.*

Proof. This proof is similar to the proof of Theorem 1. □

5.4. NE seeking problems in noncooperative games

In this section, we investigate a non-cooperative game problem involving N players. Each player's behavior is denoted by $\mathbf{x}(i)$, and the set of players is defined as $\mathcal{N} = \{1, 2, \dots, N\}$. The strategy combination of all participants is represented by $\mathbf{x} = [x_1, x_2, \dots, x_N]^\top \in \mathbb{R}^N$. The payoff function of the ith player denoted as $\Theta(\mathbf{x}) : \Omega \to \mathbb{R}$, depends on both the behavior of the ith player and the behavior of other participants. Let $\mathbf{x}_{-i} = [x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N]^\top$ represents the set of behaviors of other players. In this section, we consider $\mathbf{x} = (x_i, \mathbf{x}_{-i})$, and $\Theta_i(\mathbf{x})$ is equivalent to $\Theta_i(x_i, \mathbf{x}_{-i})$. We assume the functions $\Theta_i(\mathbf{x})$ are continuously differentiable in $\mathbf{x}$ and convex function, where $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$, and each set $\Omega_i \subset \mathbb{R}$ is closed and convex. All players aim to minimize their own cost function. In this section, we mainly study the Nash equilibrium search algorithm for noncooperative game problems. These types of problems commonly arise in practical scenarios such as energy consumption problems in smart grids (Huang, Zou, & Meng, 2020) and vehicle deployment problems in transportation (Liu & Krstic, 2012). The Nash equilibrium point search for a quadratic noncooperative game problem is defined as follows:

Find the Nash equilibrium point $\mathbf{x}^* = (x_i^*, \mathbf{x}_{-i}^*)$ such that the following inequality holds

$$\Theta_i(x_i^*, \mathbf{x}_{-i}^*) \leq \Theta_i(x_i, \mathbf{x}_{-i}^*), \tag{32}$$

where $x_i \in \Omega_i, \forall i \in \mathcal{N}$, $\mathbf{x}_{-i} = [x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N]^\top$, Ω_i is closed convex set. Let

$$\boldsymbol{\Theta} = [\nabla_1\Theta_1(x_1, \mathbf{x}_{-1}), \nabla_2\Theta_2(x_2, \mathbf{x}_{-2}), \dots, \nabla_N\Theta_N(x_N, \mathbf{x}_{-N})]^\top$$

satisfy Assumptions 1 and 3, where $\nabla_i\Theta_i(x_i, \mathbf{x}_{-i}) = \frac{\partial\Theta_i(x_i,\mathbf{x}_{-i})}{\partial x_i}$ for all $i \in \mathcal{N}$. The proposed PeNN (14) can be reduced to a specific approach called $PeNN_{nesp}$, which aims to find the solution $\mathbf{x}^*$ for the problem (32) from Cavazzuti et al. (2002). The $PeNN_{nesp}$ can be expressed as follows

$$\dot{x}_i = -\frac{G_{ci}}{T_{ci}}\rho(x_i)\left(x_i - \omega(x_i) - \eta\nabla_i\Theta_i(x_i, \mathbf{x}_{-i}) + \eta\nabla_i\Theta_i(\omega(x_i), \omega(\mathbf{x}_{-i}))\right) \tag{33}$$

$$\rho(x_i) = \begin{cases} a_i\frac{1}{\|\phi(x_i)\|^{1-\alpha_i}} + b_i\frac{1}{\|\phi(x_i)\|^{1-\beta_i}} + c_i, & \text{if } x_i \in \mathbb{R}\setminus \mathrm{Fix}(\omega(x_i)), \\ 0, & \text{otherwise} \end{cases}$$

where G_{ci} is constant, T_{ci} is user-defined parameter, $\omega(x_i) = P_{\Omega_i}(x_i - \eta\nabla_i\Theta_i(x_i, \mathbf{x}_{-i}))$ and $\eta > 0$, $\phi(x_i) = x_i - \omega(x_i) - \eta\nabla_i\Theta_i(x_i, \mathbf{x}_{-i}) + \eta\nabla_i\Theta_i(\omega(x_i), \omega(\mathbf{x}_{-i}))$, $a_i, b_i, c_i > 0$, $\alpha_i \in (0, 1), \beta_i > 1$ and $\mathrm{Fix}(\omega(x_i)) := \{x_i \in \mathbb{R} : \omega(x_i) = x_i\}$.

We can get similar results from Theorem 1 under Assumptions 1 and 3.

6. Numerical experiments

This section presents a series of illustrative examples to showcase the efficacy of the PeNN (14). All programming implementations are carried out using Matlab 2015b and executed on a PC desktop equipped with an Intel(R) Core(TM) i7-8550U CPU@1.80 GHz. Let $\rho_1 = \rho_2 = \rho_3 = 100$, $\beta_1 = 0.5$, $\beta_2 = 1.5$, and $\eta = \frac{1}{10L}$ by default in the absence of specific instructions.

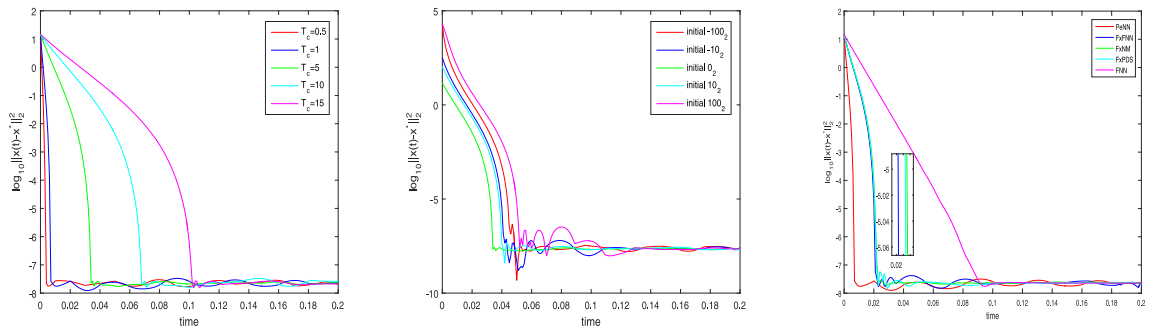
Example 1 (*Hu & Wang, 2006*)**.** Consider a variational inequality problem (27)

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 0.5x_1 - 2x_2 - 10^7 \\ -4x_1 + 0.1x_2^2 - 10^7 \end{bmatrix} \text{ and } \Omega = \{\mathbf{x} \in \mathbb{R}^2 | (x_1 - 2)^2 + (x_2 - 2)^2 \leq 1\}.$$

From Hu and Wang (2006), we know that $\mathbf{f}(\mathbf{x})$ is strongly pseudo-monotone with $\gamma = 11$, and the Lipschitz constant on Ω is $L = 5$. The simulation results converge to $\mathbf{x}^* = (2.707, 2.707)^\top$. In this example, the properties of the function $\mathbf{f}(\cdot)$ are known in more detail. We take $\rho_1 = \rho_2 = \rho_3 = 10$, $\beta_1 = 0.5, \beta_2 = 1.5$, $\eta = 0.02$. We get $G_c = 23.0058$ from Theorem 1 setting different values of T_c. From Fig. 2, we can find different T_c values in Fig. 2, which it can control different convergence times, and converge to the unique point $\mathbf{x}^* = (2.707, 2.707)^\top$. This also confirms the theory of predefined time stability. We take the initial point $\mathbf{x}(0) = \mathbf{0}$, and take different T_c values, and draw convergence responses of the errors in Fig. 3(a). We find that the larger the T_c value, the larger the convergence time, and different T_c values control different convergence time.

We take $T_c = 5$ in Fig. 3(b). We get $G_c = 23.0058$ from Theorem 1 setting different values of the initial point. It shows that the PeNN (14) is independent of the initial point and achieves fixed-time stability within the predefined time $T_c = 5$.

We show the convergence responses of the errors $\log_{10}\|\mathbf{x}(t) - \mathbf{x}^*\|_2^2$ compared with the FxFNN (Ju, Hu, et al., 2021), the FxNM in Zheng et al. (2022), the FxPDS in Garg et al. (2023), the FNN (Ju, Che, et al.,



(a) (b) (c)

Fig. 3. (a) Convergence responses of the errors $\log_{10}\|\mathbf{x}(t)-\mathbf{x}^*\|_2^2$ for the PeNN (14) with initial point $\mathbf{x}(0)=\mathbf{0}$ vs the different values T_c; (b) Convergence responses of the errors $\log_{10}\|\mathbf{x}(t)-\mathbf{x}^*\|_2^2$ for PeNN (14) vs the different initial points; (c) Convergence responses of the errors $\log_{10}\|\mathbf{x}(t)-\mathbf{x}^*\|_2^2$ for PeNN (14) vs the FxFNN (Ju, Hu, et al., 2021), the FxNM in Zheng et al. (2022), the FxPDS in Garg et al. (2023), the FNN (Ju, Che, et al., 2021) with initial point $\mathbf{x}(0)=\mathbf{0}$.

2021) in Fig. 3(c). We take $\rho_1=\rho_2=\rho_3=100$, $\beta_1=0.5, \beta_2=1.5$, $\eta=0.02$, $T_c=1$, and the $G_c=2.3006$ in Fig. 3(c). Fig. 3(c) shows the PeNN (14) is faster in convergence speed than Ju, Hu, et al. (2021), the FxNM in Zheng et al. (2022), the FxPDS in Garg et al. (2023), and the FNN (Ju, Che, et al., 2021). The simulation results demonstrate the correctness of the theory and the efficiency of the methods.

Example 2 (*Eshaghnezhad et al., 2016*). Consider the nonlinear complementarity problem (NCP): Find $\mathbf{x}=(x_1,x_2,\dots,x_n)^\top\in\mathbb{R}^n$ such that

$$\mathbf{x}\in\mathbb{R}^n_+,\quad \mathbf{f}(\mathbf{x})\in\mathbb{R}^n_+,\quad \mathbf{x}^\top\mathbf{f}(\mathbf{x})=0, \tag{34}$$

where $\mathbf{f}(\mathbf{x})=\mathbf{g}(\mathbf{x})+\mathbf{B}\mathbf{x}+\mathbf{c}$, $\mathbf{g}(\mathbf{x})=(\arctan(x_1),\arctan(x_2),\dots,\arctan(x_n))^\top$, $\mathbf{c}=(-\frac{n}{2}+1,\dots,-\frac{n}{2}+n)^\top$ and

$$\mathbf{B}=\begin{bmatrix} 2 & -1 & 0 & \cdots & \cdots & 0\\ -1 & 2 & -1 & 0 & \cdots & \vdots\\ 0 & -1 & 2 & -1 & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & \ddots & 0\\ \vdots & \vdots & \ddots & \ddots & \ddots & -1\\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{bmatrix}_{n\times n}.$$

Clearly, the NCP (34) can be viewed as VI (27), where $\Omega=\mathbb{R}^n_+$. From Eshaghnezhad et al. (2016, Example 4), we know that the Lipschitz constant of the strongly monotone function $\mathbf{f}(\cdot)$ is $L=4.7321$, and the unique solution of NCP (34) is $\mathbf{x}^*=[3.5557,4.4080,3.6080,2.1084,0.7368,0,0,0,0,0]^\top$. We take $n=10$, and set $\lambda=0.8$ and $\beta=0.01$. The PeNN (14) converges to the solution $x^*=[3.5557,4.4080,3.6080,2.1084,0.7368,0,0,0,0,0]^\top$. In this case, since we know little about the properties of the function, we do not set T_c and G_c, and take them separately $\rho_1=\rho_2=\rho_3=100$, $\beta_1=0.2,\beta_2=1.5$, $\eta=0.02$.

Fig. 4(a) shows the trajectory behavior of $\mathbf{x}(t)$ of the PeNN (14) with five random initial points. We show the convergence responses of the errors in different initial conditions in Fig. 4(b). We can observe that the PeNN (14) achieves stability within a fixed-time which is independent of initial conditions. Compared to other fixed-time methods (see, e.g., Garg et al. (2023), Ju, Hu, et al. (2021), Zheng et al. (2022)), this approach exhibits similar convergence time when solving the problem, with minor differences in Fig. 4(c).

Example 3. We conducted an experiment on signal recovery. We generated a sparse signal $\mathbf{x}\in\mathbb{R}^n$ with $n=200$ and $k=20$, where k represents sparsity. The non-zero elements of $\mathbf{x}$ were generated from a normal Gaussian distribution. We used a measurement matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$ with $m=80$. The measurement vector was obtained as $\mathbf{y}=\mathbf{A}\mathbf{x}+\boldsymbol{\epsilon}\in\mathbb{R}^m$, where $\boldsymbol{\epsilon}$ is Gaussian noise with standard deviation $\sigma=0.01$ (Garg et al., 2023). We set the threshold $\mu=\nu\sqrt{2\lg n}$, where $\nu=0.005$, and $\eta=\frac{1}{4\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}$. This experiment is compared with the FxFNNssr (Ju, Hu, et al., 2021), the PNN_{ssr} (Liu & Wang, 2015), and the LCA (Rozell, Johnson, Baraniuk, & Olshausen, 2008). We take $n=200$, $m=80$, $k=20$ in Fig. 5. Fig. 5(a) show the trajectory behavior of $\mathbf{x}(t)$ for the PeNN (14). Figs. 5(b) and (c) show the sparse signal $\mathbf{x}^*$ recovered and the convergence responses of relative errors by the PeNN (14) and other methods (Ju, Hu, et al., 2021; Liu & Wang, 2015; Rozell et al., 2008). We can observe that the PeNN (14) is few faster in convergence speed than other methods (Ju, Hu, et al., 2021; Liu & Wang, 2015; Rozell et al., 2008) from Fig. 5(c). We use different observation matrices m to recover sparse signals and observation matrix of the same dimension to recover different sparse signals in Figs. 6 and 7. The experimental results all show the effectiveness and feasibility of the proposed PeNN (14), and Fig. 8 also shows the comparison of the PeNN (14) corresponding to the error in different situations. The results also showcase the effectiveness and feasibility of the proposed PeNN (14).

Example 4. The objective of the energy consumption game in the context of heating, ventilation, and air conditioning (HVAC) for N players is to find a solution NE $\mathbf{x}^*=(x_i^*,\mathbf{x}_{-i}^*)$ that satisfies the following inequality for all players $i\in\mathcal{N}$ and all potential choices $x_i\in\Omega_i$;

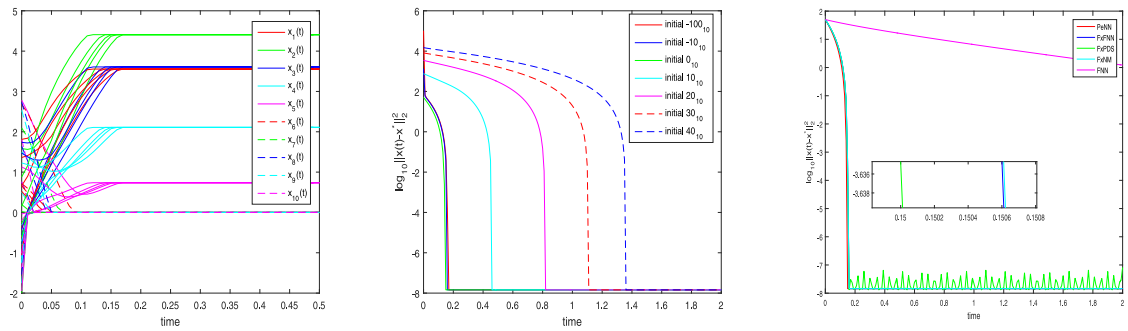
$$\Theta_i(x_i^*,\mathbf{x}_{-i}^*)\le\Theta_i(x_i,\mathbf{x}_{-i}^*),\forall i\in\mathcal{N},\forall x_i\in\Omega_i. \tag{35}$$

Here, the payoff function for the ith player is defined as $\Theta_i(x_i,\mathbf{x}_{-i})=q_i(x_i-p_i)^2+(s\sum_{j=1}^N x_j+r)x_i$, $\forall i\in\mathcal{N}$, where q_i and p_i are constants specific to player i, and $s\sum_{j=1}^N x_j+r$ represents the pricing function. In a way that does not affect generality, we consider a scenario with 6 players ($N=6$). In the simulation, we assume $s=1$ and $r=10$ for all players $i\in\mathcal{N}$, and the feasible range of choices for each player is defined as $\Omega_i=\{x_i|1\le x_i\le 30\}$. Table 1 provides the values of the constants p_i and q_i, the Nash equilibrium solution NE $\mathbf{x}^*$, and the corresponding payoff function $\Theta_i(x_i^*,\mathbf{x}_{-i}^*)$.

To illustrate the application of the proposed method, we consider specific parameter values. In particular, we set $a_i=b_i=c_i=100$, $\alpha_i=0.1,\beta_i=1.1$, $\eta=0.001$ for all players $i\in\mathcal{N}$ in PeNN_{nesp} (33). The transient response of $PeNN_{nesp}$ (33) is presented in Fig. 9(a), demonstrating convergence to the Nash equilibrium solution $\mathbf{x}^*$ from five different random initial points. Fig. 9(b) displays the transient response of the payoff function $\Theta_i(x_i,\mathbf{x}_{-i})$ for the $PeNN_{nesp}$ (33). Furthermore, Fig. 9(c) illustrates the convergence behavior of the errors associated with $PeNN_{nesp}$ (33) for five different initial states.

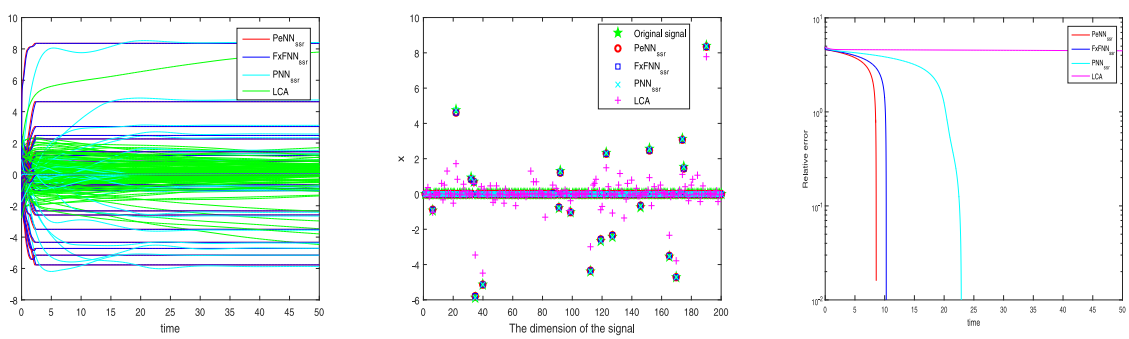
7. Conclusions

The proposed proximal neural network (PeNN) (14) with a predefined-time stability has been shown. The versatility and effectiveness of the PeNN (14) are demonstrated by its ability to handle a wide



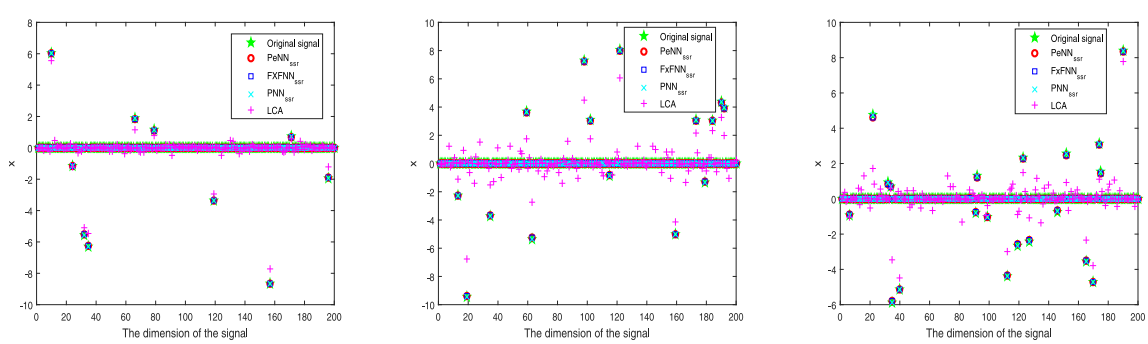
(a) (b) (c)

Fig. 4. (a) Trajectory behavior of $\mathbf{x}(t)$ of the PeNN (14) with five random initial points; (b) Convergence responses of the errors $\log_{10}\|\mathbf{x}(t)-\mathbf{x}^*\|_2^2$ for the PeNN (14) vs the different initial points; (c) Convergence responses of the errors $\log_{10}\|\mathbf{x}(t)-x^*\|_2^2$ for PeNN (14) vs the FxFNN (Ju, Hu, et al., 2021), the FxPDS in Garg et al. (2023), the FxNM in Zheng et al. (2022), the FNN (Ju, Che, et al., 2021) with initial point $\mathbf{x}(0)=\mathbf{0}_{10}$.



(a) (b) (c)

Fig. 5. (a) Trajectory behavior of $\mathbf{x}(t)$ for the PeNN (14); (b) Sparse signal $\mathbf{x}^*$ recovered by the PeNN (14) vs the FxFNN_{ssr} (Ju, Hu, et al., 2021), the PNN_{ssr} (Liu & Wang, 2015) and the LCA (Rozell et al., 2008); (c) Convergence responses of relative errors of the PeNN (14) vs the FxFNN_{ssr} (Ju, Hu, et al., 2021), the PNN_{ssr} (Liu & Wang, 2015) and the LCA (Rozell et al., 2008).



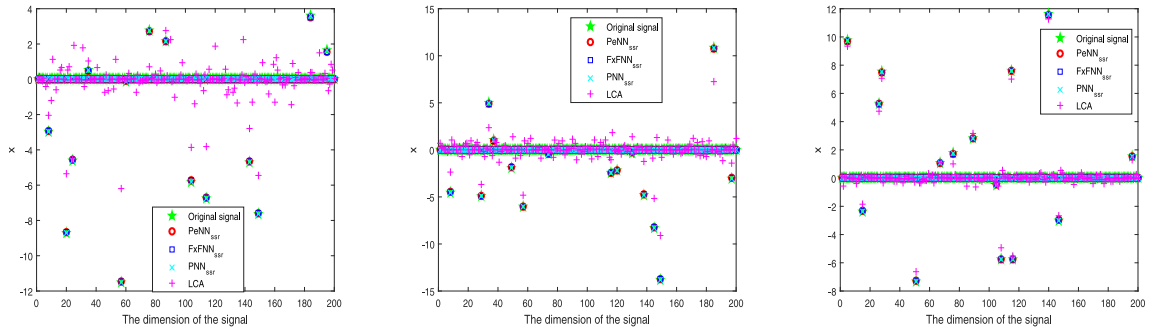
(a) (b) (c)

Fig. 6. (a) $n=200$, $m=80$, $k=10$; (b) $n=200$, $m=80$, $k=15$; (c) $n=200$, $m=80$, $k=20$.

Table 1
Contants q_i, p_i and the value of NE $\mathbf{x}^*$ and the value of the payoff function Θ_i^*.

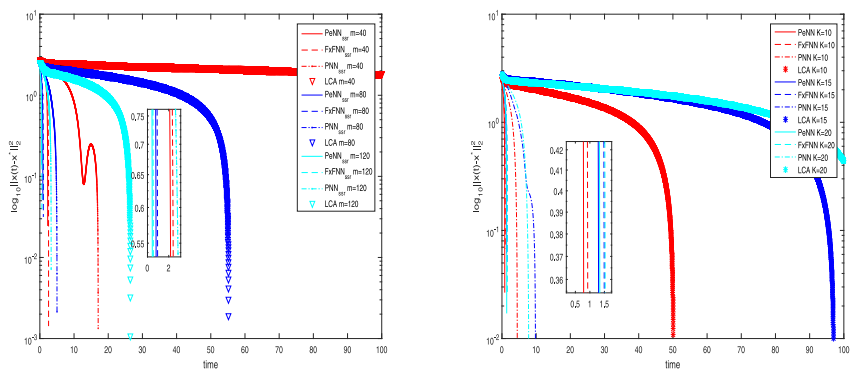
Player i	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$
q_i	3	2	3	4	5	2
p_i	6	9	8	9	3	10
x_i^*	1.4697	2.0576	3.1840	5.1431	1	2.8576
Θ_i^*	99.3594	149.2983	151.4478	191.7408	45.7103	175.5017

range of problems, such as mixed variational inequalities, nonlinear complementarity problems, sparse signal recovery problems, and NE seeking problems in noncooperative games. The numerical simulation results further validate our theoretical derivations and confirm the practical effectiveness of the proposed PeNN (14). This paper provides a different perspective on neurodynamic approaches and sets some examples for the application of the PeNN (14) in solving optimization problems. In the next step, we can consider applications of other math optimization problems within predefined time stability, explore ways to improve convergence speed in continuous-time math optimization, and so on.



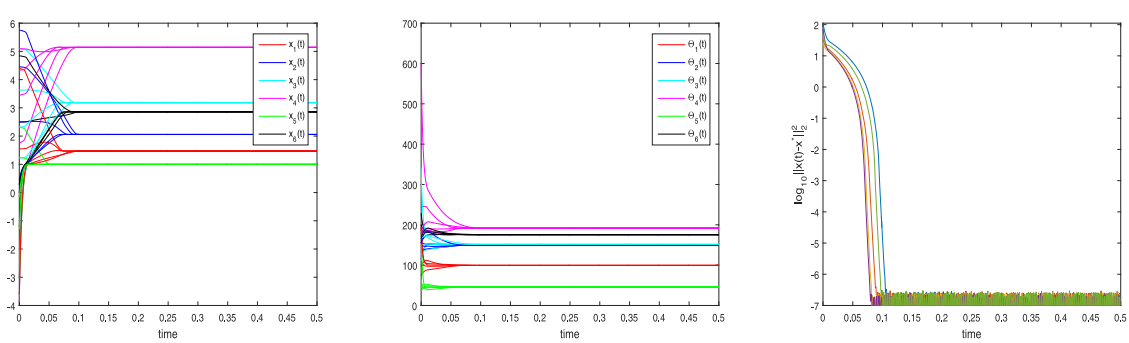
(a) (b) (c)

Fig. 7. (a) $n = 200,\ m = 65,\ k = 15$; (b) $n = 200,\ m = 80,\ k = 15$; (c) $n = 200,\ m = 120,\ k = 15$.



(a) (b)

Fig. 8. (a) Convergence responses of relative errors of the PeNN (14) vs the FxFNN$_{ssr}$ (Ju, Hu, et al., 2021), the PNN$_{ssr}$ (Liu & Wang, 2015) and the LCA (Rozell et al., 2008) in $n = 200,\ k = 5$ vs. different m values; (b) Convergence responses of relative errors of the PeNN (14) vs the FxFNN$_{ssr}$ (Ju, Hu, et al., 2021), the PNN$_{ssr}$ (Liu & Wang, 2015) and the LCA (Rozell et al., 2008) in $n = 200,\ m = 80$ vs. different k values.



(a) (b) (c)

Fig. 9. (a) Trajectory behavior of the players $\mathbf{x}(t)$ for the PeNN (33) in five random initial states; (b) Trajectory behavior of the values of payoff function $\Theta(\mathbf{x}^*)$ for the PeNN (33) in five random initial states; (c) Convergence responses of relative errors of the PeNN (33) in five random initial states.

CRediT authorship contribution statement

Jinlan Zheng: Writing – original draft, Validation, Software, Investigation. **Xingxing Ju:** Writing – review & editing, Methodology, Conceptualization. **Naimin Zhang:** Writing – review & editing, Methodology, Conceptualization. **Dongpo Xu:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No. 62176051), in part by National Key R & D Program of China (No. 2021YFA1003400), in part by the Fundamental Research Funds for Central Universities, China under Grant (No. 2023SCU12009), in part by the Sichuan Province Natural Science Foundation of China under Grant (No. 2023NSFSC1433), in part by the Shaanxi Key Laboratory of Information Communication Network and Security under Grant (No. ICNS202303), in part by the National Funded Postdoctoral Researcher Program of China under Grant (No. GZB20230467), and in part by the China Postdoctoral Science Foundation under Grant (No. 2023M742457).

Appendix. Proofs for the Lemmas in Section 2

A.1. Proof of Lemma 2

Proof. Let $W(s) = V^{1-\alpha}(s)$, then

$$\dot{V}(s) = \frac{1}{1-\alpha}\dot{W}(s)V^{\alpha}(s)$$

It follows from $\dot{V}(\mathbf{x}(t)) \le -\frac{G_c}{T_c}(aV^{\alpha}(\mathbf{x}(t)) + bV^{\beta}(\mathbf{x}(t)) + cV(\mathbf{x}(t)))$,

$$\frac{1}{1-\alpha}\dot{W}(\mathbf{x}(t))V^{\alpha}(\mathbf{x}(t)) \le -\frac{G_c}{T_c}(aV^{\alpha}(\mathbf{x}(t)) + bV^{\beta}(\mathbf{x}(t)) + cV(\mathbf{x}(t))).$$

Since $V(s) = W^{\frac{1}{1-\alpha}}(s)$, we can get

$$\begin{aligned}\dot{W}(\mathbf{x}(t)) &\le (1-\alpha)\frac{G_c}{T_c}\left[-a - bV^{\beta-\alpha}(\mathbf{x}(t)) - cV^{1-\alpha}(\mathbf{x}(t))\right]\\ &\le -a(1-\alpha)\frac{G_c}{T_c}.\end{aligned}$$

Due to $\dot{W}(\mathbf{x}(t)) \le -a(1-\alpha)\frac{G_c}{T_c}$, then there exist a constant $T(\mathbf{x}(0)) = \frac{T_c W(\mathbf{x}(0))}{G_c a(1-\alpha)} > 0$ such that $\lim_{t\to T(\mathbf{x}(0))} W(\mathbf{x}(0)) > 0$ and $W(\mathbf{x}(0)) = 0$ for $\forall t > T(\mathbf{x}(0))$. We can observe that $W(\mathbf{x}(t)) = 0 \Leftrightarrow V(\mathbf{x}(t)) = 0 \Leftrightarrow \mathbf{x}(t) = 0$, We can say that there exists a constant $T(\mathbf{x}(0)) > 0$ such that $\lim_{t\to T(\mathbf{x}(0))}\|\mathbf{x}(t)\| = 0$ for $\forall t > T(\mathbf{x}(0))$. This implies that the origin of system (7) can achieve finite-time stability.

Since the condition (9) holds, then

$$\frac{dt}{dV(\mathbf{x}(t))} \ge \frac{T_c}{G_c}\frac{1}{-aV^{\alpha}(\mathbf{x}(t)) - bV^{\beta}(\mathbf{x}(t)) - cV(\mathbf{x}(t))},$$

and

$$T(\mathbf{x}(0)) \le \frac{T_c}{G_c}\int_{V(\mathbf{x}(0))}^{0}\frac{1}{-as^{\alpha} - bs^{\beta} - cs}ds = \frac{T_c}{G_c}\int_{0}^{V(\mathbf{x}(0))}\frac{1}{as^{\alpha} + bs^{\beta} + cs}ds.$$

Below, we will discuss two different cases.

(i) If $0 \le V(\mathbf{x}(0)) \le 1$,

$$T(\mathbf{x}(0)) \le \frac{T_c}{G_c}\int_0^1\frac{1}{as^{\alpha} + bs^{\beta} + cs}ds \le \frac{T_c}{G_c}\int_0^1\frac{1}{as^{\alpha} + cs}ds.$$

Let $w = s^{1-\alpha}$, we have $dw = (1-\alpha)s^{-\alpha}ds$, and

$$\begin{aligned}\int_0^1\frac{1}{as^{\alpha} + cs}ds &= \frac{1}{1-\alpha}\int_0^1\frac{s^{\alpha}}{as^{\alpha} + cs}dw\\ &= \frac{1}{1-\alpha}\int_0^1\frac{1}{a + cs^{1-\alpha}}dw\\ &= \frac{1}{1-\alpha}\int_0^1\frac{1}{a + cw}dw = \frac{1}{c(1-\alpha)}\ln\left(1 + \frac{c}{a}\right)\end{aligned}$$

Therefore,

$$T(\mathbf{x}(0)) \le \frac{T_c}{G_c}\frac{1}{c(1-\alpha)}\ln\left(1 + \frac{c}{a}\right) \overset{(10)}{\le} T_c.$$

(ii) If $V(\mathbf{x}(0)) > 1$,

$$\begin{aligned}T(\mathbf{x}(0)) &\le \frac{T_c}{G_c}\int_0^1\frac{1}{as^{\alpha} + bs^{\beta} + cs}ds + \frac{T_c}{G_c}\int_1^{V(\mathbf{x}(0))}\frac{1}{as^{\alpha} + bs^{\beta} + cs}ds\\ &\le \frac{T_c}{G_c}\int_0^1\frac{1}{as^{\alpha} + cs}ds + \frac{T_c}{G_c}\int_1^{+\infty}\frac{1}{bs^{\beta} + cs}ds.\end{aligned}$$

Let $z = s^{1-\beta}$, we have $dz = (1-\beta)s^{-\beta}ds$, and

$$\begin{aligned}\int_1^{+\infty}\frac{1}{bs^{\beta} + cs}ds &= \frac{1}{1-\beta}\int_1^0\frac{s^{\beta}}{bs^{\beta} + cs}dz\\ &= \frac{1}{\beta-1}\int_0^1\frac{1}{b + cs^{1-\beta}}dz\\ &= \frac{1}{\beta-1}\int_0^1\frac{1}{b + cz}dz = \frac{1}{c(\beta-1)}\ln\left(1 + \frac{c}{b}\right).\end{aligned}$$

Therefore,

$$T(\mathbf{x}(0)) \le \frac{T_c}{G_c}\left(\frac{1}{c(1-\alpha)}\ln\left(1 + \frac{c}{a}\right) + \frac{1}{c(\beta-1)}\ln\left(1 + \frac{c}{b}\right)\right) \overset{(10)}{=} T_c.$$

The origin of system (7) can achieve fixed-time stability within a predefined time T_c, and

$$G_c = \frac{1}{c(1-\alpha)}\ln\left(1 + \frac{c}{a}\right) + \frac{1}{c(\beta-1)}\ln\left(1 + \frac{c}{b}\right). \quad \square$$

A.2. Proof of Lemma 3

Proof. The proof is similar to that in Garg et al. (2023, Proposition 1,P5031). For completeness, we give the proof. The continuity of the vector field in (12) on $\mathbb{R}^n$ is established by demonstrating its continuity at $\mathbf{x}^* \in \mathbb{R}^n$. This follows from the fact that the function σ is continuous on $\boldsymbol{\phi}(\mathbf{x}) \neq 0$ and the vector field X is locally Lipschitz continuous on $\mathbb{R}^n$. Therefore, it suffices to conclude that the vector field in (12) is continuous for all $\mathbf{x} \in \mathbb{R}^n$. It is important to note that the limit $\lim_{x\to\mathbf{x}^*}\sigma(\mathbf{x}(t))\boldsymbol{\phi}(\mathbf{x}(t)) = 0$ holds true, given that the parameters $\alpha_1 \in (0,1)$ and $\alpha_2 > 1$.

The demonstration of the existence of a solution to (12) for all forward times, given any initial condition, is presented as follows. Leveraging the continuity of the vector field in (12), it can be deduced from Hale (1980, Theorem I.1.1, P14) that for any $\mathbf{x}(0) \in \mathbb{R}^n$, there exists a solution of (12) over an interval $[0, \tau(\mathbf{x}(0))]$, where $\tau(\mathbf{x}(0)) > 0$. Moreover, in accordance with Hale (1980, Theorem I.2.1, P17), any such solution of (12) over the interval $[0, \tau(\mathbf{x}(0)))$ extends to a maximal interval of existence $[0, \tau(\mathbf{x}(0))]$. Next, we introduce the candidate Lyapunov function $V : \mathbb{R}^n \to [0, \infty]$ defined as $V(\mathbf{x}) := \|\mathbf{x} - \mathbf{x}^*\|^2$. Upon considering the time-derivative of the candidate Lyapunov function V along a solution of (12), initiated from $\mathbf{x}(0) \in \mathbb{R}^n$, we obtain

$$\dot{V}(\mathbf{x}(t)) = -\langle \mathbf{x} - \mathbf{x}^*, \sigma(\mathbf{x}(t))\boldsymbol{\phi}(\mathbf{x}(t))\rangle. \tag{A.1}$$

Recalling that $\langle \mathbf{x} - \mathbf{x}^*, \boldsymbol{\phi}(\mathbf{x})\rangle \ge 0$ and $\sigma(\mathbf{x}) \ge 0$ hold true for all $t \in [0, \tau(\mathbf{x}(0)))$, it follows that $V(\mathbf{x}(t)) \le V(\mathbf{x}(0))$ for all $t \in [0, \bar{\tau}(\mathbf{x}(0)))$. Notably, the aforementioned solution lies within the set $L_{\mathbf{x}(0)} := \{\mathbf{y} \in$

$\mathbb{R}^n | \|\mathbf{y}-\mathbf{x}^*\| \le \|\mathbf{x}(0)-\mathbf{x}^*\|\}$. Given that $L_{\mathbf{x}(0)}$ is compact, as per Bhat and Bernstein (2000, Proposition2.1,P752), $\bar{\tau}(\mathbf{x}(0)) = \infty$ and thus the existence of a solution to (12) is established.

This subsequent exposition the uniqueness of solution of (12) across all forward temporal intervals, regardless of the specific initial condition. Let $\mathbf{x}(0) \in \mathbb{R}^n$ be a given initial condition, and let $\mathbf{y}$ represent a solution of (12), with $\mathbf{x}(0) = \mathbf{x}(0)$. The ensuing analysis delves into two distinct cases.

In the first scenario, assume $\mathbf{y}(0) \in \mathbb{R}^n \backslash \{\mathbf{x}^*\}$. We proceed to prove that a solution associated with the vector field in (12) is also a solution linked to the vector field -X by means of a well-suited time reparameterization (Chicone, 2006, Proposition 1.28, P17). Let $T := \inf\{t > 0 | \mathbf{y}(t) = \mathbf{x}^*\}$ and owing to the continuity of $\mathbf{y}$, it ensues that $T > 0$. Now, consider the function $s : [0, T) \to [0, \infty)$ as defined below:

$$s(t) := \int_0^t \sigma(\mathbf{y}(v))dv. \tag{A.2}$$

By virtue of the continuity of the function σ on $\mathbb{R}^n$, the continuity of $\mathbf{y}$ over the interval $[0, T)$, and the continuity of $\sigma(\mathbf{y}(v))$ for any $v \in [0, T)$, it can be deduced that the function s is strictly increasing and continuous. Moreover, it is observed that $\frac{ds}{dt} \neq 0$ holds for all $t \in (0, T)$. Additionally, leveraging the inverse function theorem, it can be ascertained that the function $t := s^{-1}$ not only exists but is also strictly increasing, continuous, and fulfills the condition

$$\frac{dt}{ds}|_{s=s(t)} = \frac{1}{\sigma(\mathbf{y}(t))}, \tag{A.3}$$

for all $t \in (0, T)$. Let $\bar{\mathbf{y}}(t) := -\mathbf{y}(t(s))$ and from the clain rule, it follows that

$$\frac{d\bar{\mathbf{y}}}{ds} = \frac{d\mathbf{y}}{ds}|_{t=t(s)} \frac{dt}{ds}. \tag{A.4}$$

By leveraging Eqs. (A.3) and (A.4), the derivative $\frac{d\bar{\mathbf{y}}}{ds} = -\boldsymbol{\phi}(\bar{\mathbf{x}}(t))$ can be obtained. Consequently, it is observed that a solution corresponding to the vector field in (12) is also a solution for the vector field X, given the time reparameterization provided in (A.2). Additionally, by following similar steps to those in the proof of the first claim and recalling the local Lipschitz continuity of the vector field X on $\mathbb{R}^n$, it can be demonstrated that there exists a unique solution for any given initial condition in correspondence to the vector field X for all forward times, as per Agarwal, Agarwal, and Lakshmikantham (1993, Theorem 3.2.2, P139). Hence, the uniqueness of $\bar{\mathbf{y}}$ can be established, and since the function s is injective and satisfies $s(0) = 0$, it follows that $\mathbf{y}$ is uniquely determined as well.

In the second scenario, considering $\mathbf{y}(0) = \mathbf{x}^*$, we analyze the same candidate Lyapunov function V as introduced in the first claim. By following the steps outlined in the proof of the first claim, it can be demonstrated that the time-derivative of the candidate Lyapunov function V along a solution of (12), commencing from any given initial condition, is consistently nonpositive. By referring to relevant literature, such as Agarwal et al. (1993, Theorem 3.15.1, P160), it can be concluded that $\mathbf{y}$ is uniquely determined. This provides the necessary evidence to establish the validity of the second claim. □

References

Agarwal, R. P., Agarwal, R. P., & Lakshmikantham, V. (1993). *Uniqueness and nonuniqueness criteria for ordinary differential equations*: *vol. 6*, World Scientific.

Anguiano-Gijón, C. A., Muñoz-Vázquez, A. J., Sánchez-Torres, J. D., Romero-Galván, G., & Martínez-Reyes, F. (2019). On predefined-time synchronisation of chaotic systems. *Chaos, Solitons & Fractals*, *122*, 172–178.

Bauschke, H., & Combettes, P. (2017). *Convex analysis and monotone operator theory in hilbert spaces*. Springer.

Bhat, S. P., & Bernstein, D. S. (2000). Finite-time stability of continuous autonomous systems. *SIAM Journal on Control and Optimization*, *38*(3), 751–766.

Candes, E., & Tao, T. (2009). The dantzig selector: Statistical estimation when p is much larger than n. *Quality Control and Applied Statistics*, *54*(1), 83–84.

Cavazzuti, E., Pappalardo, M., Passacantando, M., et al. (2002). Nash equilibria, variational inequalities, and dynamical systems. *Journal of Optimization Theory and Applications*, *114*(3), 491–506.

Chen, C., Li, L., Peng, H., Yang, Y., Mi, L., & Zhao, H. (2020). A new fixed-time stability theorem and its application to the fixed-time synchronization of neural networks. *Neural Networks*, *123*, 412–419.

Chicone, C. (2006). *Ordinary differential equations with applications*: *vol. 34*, Springer Science and Business Media.

Di Marco, M., Forti, M., Nistri, P., & Pancioni, L. (2015). Discontinuous neural networks for finite-time solution of time-dependent linear equations. *IEEE Transactions on Cybernetics*, *46*(11), 2509–2520.

Eshaghnezhad, M., Effati, S., & Mansoori, A. (2016). A neurodynamic model to solve nonlinear pseudo-monotone projection equation and its applications. *IEEE Transactions on Cybernetics*, *47*(10), 3050–3062.

Facchinei, F., & Pang, J.-S. (2003). *Finite-dimensional variational inequalities and complementarity problems*. Springer.

Gao, X., & Liao, L.-Z. (2016). A novel neural network for generally constrained variational inequalities. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(9), 2062–2075.

Garg, K., & Baranwal, M. (2020). CAPPA: Continuous-time accelerated proximal point algorithm for sparse recovery. *IEEE Signal Processing Letters*, *27*, 1760–1764.

Garg, K., Baranwal, M., Gupta, R., & Benosman, M. (2023). Fixed-time stable proximal dynamical system for solving MVIPs. *IEEE Transactions on Automatic Control*, *68*(8), 5029–5036.

Hale, J. (1980). *Ordinary differential equations*. Krieger Publishing Company.

He, X., Wen, H., & Huang, T. (2021). A fixed-time projection neural network for solving ℓ_1 -minimization problem. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 7818–7828.

Hu, X., & Wang, J. (2006). Solving pseudomonotone variational inequalities and pseudoconvex optimization problems using the projection neural network. *IEEE Transactions on Neural Networks*, *17*(6), 1487–1499.

Huang, B., Zou, Y., & Meng, Z. (2020). Distributed-observer-based Nash equilibrium seeking algorithm for quadratic games with nonlinear dynamics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(11), 7260–7268.

Ju, X., Che, H., Li, C., He, X., & Feng, G. (2021). Exponential convergence of a proximal projection neural network for mixed variational inequalities and applications. *Neurocomputing*, *454*, 54–64.

Ju, X., Hu, D., Li, C., He, X., & Feng, G. (2021). A novel fixed-time converging neurodynamic approach to mixed variational inequalities and applications. *IEEE Transactions on Cybernetics*, *52*(12), 12942–12953.

Ju, X., Li, C., Han, X., & He, X. (2021). Neurodynamic network for absolute value equations: A fixed-time convergence technique. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *69*(3), 1807–1811.

Kinderlehrer, D., & Stampacchia, G. (2000). *An introduction to variational inequalities and their applications*. SIAM.

Konnov, I. (2001). *Combined relaxation methods for variational inequalities*: *vol. 495*, Springer Science and Business Media.

Liu, S.-J., & Krstic, M. (2012). *Stochastic averaging and stochastic extremum seeking*. Springer Science and Business Media.

Liu, Q., & Wang, J. (2011). Finite-time convergent recurrent neural network with a hard-limiting activation function for constrained optimization with piecewise-linear objective functions. *IEEE Transactions on Neural Networks*, *22*(4), 601–613.

Liu, Q., & Wang, J. (2015). ℓ_1-Minimization algorithms for sparse signal reconstruction based on a projection neural network. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(3), 698–707.

Liu, A., Zhao, H., Wang, Q., Niu, S., Gao, X., Chen, C., et al. (2022). A new predefined-time stability theorem and its application in the synchronization of memristive complex-valued BAM neural networks. *Neural Networks*, *153*, 152–163.

Malitsky, Y. (2020). Golden ratio algorithms for variational inequalities. *Mathematical Programming*, *184*(1–2), 383–410.

Neittaanmäki, P., & Tiba, D. (1988). A variational inequality approach to constrained control problems for parabolic equations. *Applied Mathematics and Optimization*, *17*, 185–201.

Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, *1*(3), 127–239.

Pham, K. A., & Trinh, N. H. (2017). Splitting extragradient-like algorithms for strongly pseudomonotone equilibrium problems. *Numerical Algorithms*, *76*, 67–91.

Polyakov, A. (2012). Nonlinear feedback design for fixed-time stabilization of linear control systems. *IEEE Transactions on Automatic Control*, *57*(8), 2106–2110.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., & Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, *20*(10), 2526–2563.

Sánchez-Torres, J. D., Gómez-Gutiérrez, D., López, E., & Loukianov, A. G. (2018). A class of predefined-time stable dynamical systems. *IMA Journal of Mathematical Control and Information*, *35*(Supplement_1), i1–i29.

Sánchez-Torres, J. D., Sanchez, E. N., & Loukianov, A. G. (2014). A discontinuous recurrent neural network with predefined time convergence for solution of linear programming. In *2014 IEEE symposium on swarm intelligence* (pp. 1–5). http://dx.doi.org/10.1109/SIS.2014.7011799.

Scutari, G., Palomar, D. P., Facchinei, F., & Pang, J.-S. (2010). Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, *27*(3), 35–49.

Thong, D. V., & Van Hieu, D. (2018). Inertial extragradient algorithms for strongly pseudomonotone variational inequalities. *Journal of Computational and Applied Mathematics*, *341*, 80–98.

Vuong, P. T., & Strodiot, J. J. (2020). A dynamical system for strongly pseudomonotone equilibrium problems. *Journal of Optimization Theory and Applications*, *185*(3), 767–784.

Yu, L., Zheng, G., & Barbot, J.-P. (2017). Dynamical sparse recovery with finite-time convergence. *IEEE Transactions on Signal Processing*, *65*(23), 6146–6157.

Zheng, J., Chen, J., & Ju, X. (2022). Fixed-time stability of projection neurodynamic network for solving pseudomonotone variational inequalities. *Neurocomputing*, *505*, 402–412.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名：周昊龙
方向：人工智能
单位：中国科学院上海微系统与信息技术研究所

题目：基于频率滤波的校正误差鲁棒立体匹配方法

2025 IEEE International Conference on Robotics and Automation (ICRA)
May 19-23, 2025. Atlanta, USA

F^2R^2: Frequency Filtering-Based Rectification Robustness Method for Stereo Matching

Haolong Zhou[†,1,2], Dongchen Zhu[†,1,2], Guanghui Zhang[1], Lei Wang[1,2], Jiamao Li[1,2,*]

***Abstract*— Most stereo matching networks assume that the stereo images are perfectly rectified, ignoring the perturbation of extrinsic parameters due to collisions, mechanical vibrations, and thermal expansion. This leads to poor rectification robustness in real-world stereo systems. That is, even minor rectification errors can lead to failure, making stereo systems unreliable for long-term autonomous operation in complex environments. In this paper, we are the first to propose a frequency filtering-based rectification robustness (F^2R^2) method for stereo matching, which aims to enhance the robustness of existing stereo networks to rectification errors. Specifically, we propose a sensitive frequency filter (SFF) to remove components susceptible to rectification errors within the frequency domain. SFF achieves the filtering through the learning-based adaptive filtering mask (AFM) guided by the spatial-frequency mapping modulation mask (SFM). Moreover, we build the matching feature reconstruction module (MFRM) to recover the features lost during filtering to benefit cost aggregation. Comprehensive experiments on simulated datasets and self-collected data validate that our method can significantly enhance the rectification robustness of stereo matching networks.**

I. INTRODUCTION

Stereo matching recovers scene depth from the stereo pairs, which plays an important role in the fields of robot navigation [1], [2], autonomous driving [3], [4], and 3D reconstruction [5]. The stereo pairs are required to be well-rectified to simplify the dense matching problem into a 1D search [6]. To achieve high-precision calibration, a common approach involves performing offline rectification of fixed stereo cameras [7]. However, the extrinsic parameters of a stereo system can change due to collisions, mechanical vibrations, and thermal expansion after extended operation [8], [9]. As a result, the stereo matching model will experience a substantial drop in accuracy [10]. That is, the sensitivity to rectification errors makes stereo systems unreliable for long-term autonomous operation in complex environments.

To address the issues mentioned above, online self-calibration methods have been proposed to recover stereo extrinsic parameters during operation [11], [12]. However, the corrected images may still have unevenly distributed calibration errors [13]. Besides, this type of method easily fails in scenes with insufficient matched feature points [14], leading to serious consequences for tasks such as obstacle avoidance and navigation. Considering the unreliability of self-calibration, improving the stereo matching network itself to make it less susceptible to decalibration would be a promising solution. From this perspective, Hirschmuller et al. [15] propose a hand-crafted unidirectional spatial filter to adapt to non-ideally rectified stereo images for traditional algorithm in the spatial domain. However, the hand-crafted manner is tedious and relies on professional experience, and spatial domain processing makes it difficult to effectively separate the pattern structure of certain frequency components, which limits the improvement of rectification robustness. Moreover, imprecise filtering can easily cause the loss of geometric details, which has a negative impact on subsequent matching and leads to the degradation of disparity accuracy.

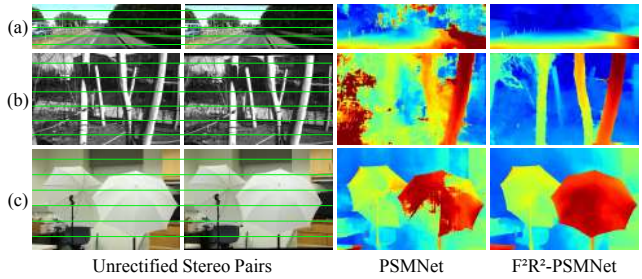


Fig. 1. Evaluation of the robustness of PSMNet with and without our method under rectification errors, using simulated unrectified stereo images from (a) KITTI2015, (b) ETH3D, and (c) Middlebury datasets.

In this paper, based on the principle that structures sensitive to rectification errors correspond to specific frequency domain components [15], we propose a novel Frequency Filtering-based Rectification Robustness (F^2R^2) method for stereo matching networks. Specifically, our method includes Sensitive Frequency Filter (SFF) and the Matching Feature Reconstruction Module (MFRM). Inspired by DFF [16] for domain generalization using deep frequency filtering, we design an Adaptive Filtering Module (AFM) within the SFF to selectively remove sensitive frequency components of matching features through learning the amplitude spectrum distribution of frequency representations. Further, to prevent overfitting to stereo misalignment patterns in the training data, we propose a Spatial-Frequency Mapping modulation mask (SFM). SFM aims to modulate the learning-based mask generated by AFM to handle different magnitudes of rectification errors, leveraging the spatial-frequency mapping relationship. In addition, to mitigate the accuracy degradation

†These authors contributed equally to this work.

*National Science and Technology Major Project from Minister of Science and Technology, China(2021ZD0201403), National Natural Science Foundation of China(62303441), Natural Science Foundation of Shanghai(23ZR1474200), Youth Innovation Promotion Association, Chinese Academy of Sciences(2021233), and Shanghai Academic Research Leader(22XD1424500).

[1]Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China. *Corresponding author: Jiamao Li jmli@mail.sim.ac.cn

[2]University of Chinese Academy of Sciences, Beijing 100049, China.

2025 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3315-4139-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICRA55743.2025.11127707

Authorized licensed use limited to: SHANGHAI INSTITUTE OF MICROSYSTEM AND INFO TECHN. Downloaded on November 14,2025 at 06:41:44 UTC from IEEE Xplore. Restrictions apply.

caused by the loss of geometric details during filtering, we propose the MFRM to reconstruct features lost due to over-filtering. This module fuses the multi-level extracted features and uses a skip connection structure to activate the aggregated cost volume. To our knowledge, this is the first work to consider rectification robustness for stereo matching networks from a frequency domain learning perspective. Notably, the proposed F^2R^2 method is plug-and-play, making it easy to integrate into existing stereo matching networks. Extensive experimental results on public datasets and self-collected data demonstrate that our method significantly enhances the rectification robustness in represented stereo network architectures. A qualitative illustration is shown in Fig. 1. The main contributions of this paper are as follows:

- We devise a Frequency Filtering-based Rectification Robustness (F^2R^2) method for stereo matching by leveraging frequency domain learning.
- We propose a Sensitive Frequency Filter (SFF), which constructs a frequency domain mask to adaptively eliminate sensitive feature components, enabling stereo matching networks to have rectification robustness.
- To ensure disparity accuracy, we build a Matching Feature Reconstruction Module (MFRM) that recovers the geometric details lost during filtering by incorporating multi-level features into the aggregated volume.
- The experimental results on public datasets and self-collected data demonstrate that our model efficiently handles rectification errors.

II. Related work

A. Learning-based Stereo Matching

With the advancement of deep learning, many CNN-based stereo matching methods have been proposed. Mayer et al. [17] extended FlowNet by introducing 1D correlation and 2D convolutions for disparity regression, presenting the first end-to-end stereo matching network. Liang et al. [18] exploited feature constancy to detect matching errors and jointly optimize disparity computation and refinement. AANet [19] designed Intra-Scale and Cross-Scale Aggregation modules for efficient cost aggregation. In contrast to the aforementioned methods using correlation layers for cost volume construction, GC-Net [20] built up 4D cost volume by concatenation and aggregated features with 3D convolutions. PSMNet [21] proposed the stacked hourglass 3D CNN in cost aggregation to capture more contextual information. GwcNet [22] proposed the group-wise correlation to construct the cost volume to obtain a more precise feature similarity measure. Recently, RAFT-Stereo [23] adapted the RAFT [24] architecture to iteratively refine high-resolution disparity estimates using GRU units. Selective-stereo [25] introduced selective recurrent units to reduce the risk of losing critical hidden disparity cue. The above methods only consider the matching of perfectly rectified stereo images.

B. Unrectified Stereo Matching

Recently, a few works have explored stereo matching under imperfect calibration. Yang et al. [26] proposed a coarse-to-fine stereo matching framework, leveraging low-resolution features to mitigate interference from rectification errors. CREStereo [27] introduced an adaptive group local correlation layer to match corresponding pixels that are not located on the same scanline due to rectification errors. Hirschmuller et al. [15] introduced a hand-crafted unidirectional spatial filter for stereo image filtering, retaining only the horizontal gradients for matching cost. However, they are limited to processing in the spatial domain. In this work, we achieve it in the intuitive frequency domain.

C. Frequency Representation Learning in vision tasks

Recently, frequency representation learning has been applied to various vision tasks. Lin et al. [16] introduced Deep Frequency Filtering, enhancing transferable components in the frequency domain to improve the generalization ability of Deep Neural Networks. Li et al. [28] proposed a frequency domain information aggregation strategy in the video super-resolution task to capture long-range spatial-temporal dependencies. Chen et al. [29] introduced frequency analysis into semantic segmentation and proposed the Frequency-Adaptive Dilated Convolution.

III. Method

Our goal is to enhance the robustness of stereo networks $\omega(\cdot)$ to rectification errors, enabling them to obtain accurate disparity from misaligned stereo pairs (I^l,I^r_{mis}). Fig. 2 illustrates the pipeline of our method, which consists of a Sensitive Frequency Filter (SFF) for filtering sensitive feature components and a Matching Feature Reconstruction Module (MFRM) for restoring the over-filtered features. These key components will be elaborated in the following sections.

A. Sensitive Frequency Filter

Motivated by the observation that different frequency components have different sensitivities to rectification errors [15], we develop the SFF from a frequency domain learning perspective to enhance the rectification robustness of stereo networks. The structure of SFF is shown in Fig. 2, comprising the Adaptive Filtering Module (AFM) and the Spatial-Frequency Mapping modulation mask (SFM). Together, the two generate the filtering mask that efficiently removes frequency components sensitive to rectification errors while preserving essential information for matching.

Adaptive Filtering Module. Inspired by DFF [16], the AFM is designed to adaptively remove sensitive components using the learning-based mask in the frequency domain.

AFM takes the extracted misaligned stereo image features $\left(f^l, f^r_{mis}\right) \in R^{C\times H\times W}$ as input. As shown in Fig. 3, we apply the Fast Fourier Transform (FFT) for each channel to obtain the frequency representations $F^l, F^r_{mis} \in R^{2C\times H\times W}$, and then compute the corresponding magnitude A^l, A^r_{mis} and phase Ph^l, Ph^r_{mis}:

$$\begin{aligned} &F(u,v) = FFT\left(f(h,w)\right), \\ &A = \|F(u,v)\| = \sqrt{\left(\mathrm{Re}(u,v)^2 + \mathrm{Im}(u,v)^2\right)}, \\ &Ph = \varphi\left[F(u,v)\right] = \arctan\left[\frac{\mathrm{Im}(u,v)}{\mathrm{Re}(u,v)}\right]. \end{aligned} \tag{1}$$

Authorized licensed use limited to: SHANGHAI INSTITUTE OF MICROSYSTEM AND INFO TECHN. Downloaded on November 14,2025 at 06:41:44 UTC from IEEE Xplore. Restrictions apply.

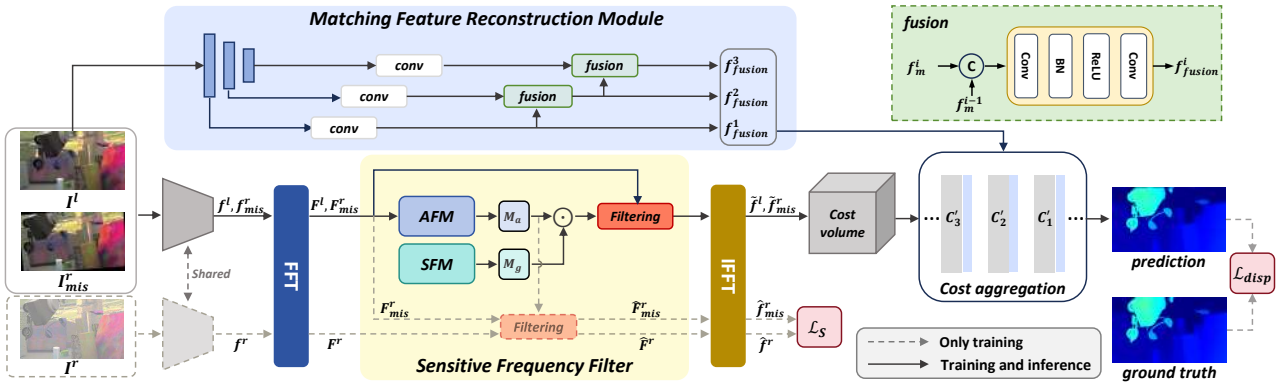


Fig. 2. The overall pipeline of our method. The Sensitive Frequency Filter (SFF) is applied to generate a frequency-domain filtering mask to filter out the frequency components sensitive to rectification errors. In SFF, AFM refers to the Adaptive Filtering Module, and SFM denotes the Spatial-Frequency Mapping modulation mask. Meanwhile, the Matching Feature Reconstruction Module (MFRM) fuses the separately extracted features to enhance the aggregated features.

where $\mathrm{Re}\,(u,v)$ and $\mathrm{Im}\,(u,v)$ represent the real and imaginary part of the frequency representation $F\,(u,v)$.

In terms of signal theory, the distribution of the 2D amplitude spectrum reflects the frequency composition of spatial domain features. Therefore, we develop the Sensitive feature Discrimination Block (SDB) to learn the amplitude spectrum distribution and generate a filtering mask that adaptively filters out sensitive frequency components. The structure of SDB is shown in Fig. 3. First, the given amplitude spectrums A^l, A^r_{mis} are passed through a 1x1 convolution, followed by batch normalization and ReLU activation. The outputs are then concatenated along the channel dimension and passed through three conv stacks to obtain the output feature F_a. After that, we use the activation functions to convert F_a into the range of 0 to 1 to generate the adaptive filtering mask $M_a \in R^{C\times H\times W}$:

$$\begin{aligned} F_a &= SDB\left(A^l, A^r_{mis}\right), \\ M_a &= (sigmoid(ReLU(F_a)) - 0.5) * 2. \end{aligned} \tag{2}$$

We then perform an element-wise multiplication of M_a with the frequency domain representations A^l and A^r_{mis} to filter out the sensitive frequency components. The filtering operation is defined as follows:

$$\hat{A} = M_a \odot A,\ \ A \in \left\{\hat{A}^l, \hat{A}^r_{mis}\right\}, \tag{3}$$

where $\odot$ denotes the Hadamard product, and $\hat{A}$ denotes the filtered magnitude spectrum. Afterward, the filtered magnitude spectrum is combined with its original phase, then fed into the inverse FFT to obtain the filtered spatial domain features $\hat{f}^l, \hat{f}^r_{mis}$. This process can be formulated as follows:

$$\begin{aligned} \hat{F}\,(u,v) &= \hat{A}e^{jPh}, \\ \hat{f}\,(h,w) &= IFFT\left(\hat{F}\,(u,v)\right), \end{aligned} \tag{4}$$

where $Ph \in \{Ph^l, Ph^r_{mis}\}$, $\hat{F}\,(u,v) \in \left\{\hat{F}^l, \hat{F}^r_{mis}\right\}$ and $\hat{A} \in \left\{\hat{A}^l, \hat{A}^r_{mis}\right\}$.

As shown in Fig. 2, during training, the same frequency domain filtering procedure is applied to the perfectly rectified right image feature f^r to generate the filtered feature $\hat{f}^r$. Ideally, the filtering should remove all sensitive components,

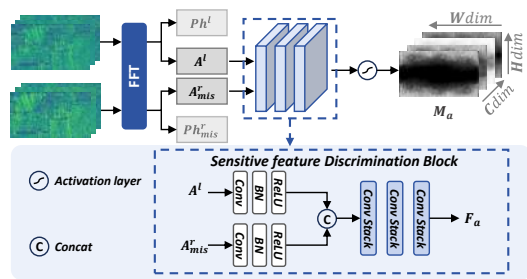


Fig. 3. The structure of Adaptive Filtering Module.

preserving only the features unaffected by rectification errors and relevant to the matching task. In this case, the ideally filtered features $\hat{f}^r$ and $\hat{f}^r_{mis}$ are expected to be consistent. Therefore, we construct a feature consistency loss $\mathcal{L}_s$ to supervise the mask generation process:

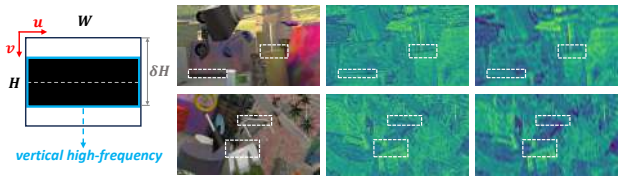
$$\begin{aligned} \mathcal{L}_s &= \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\left(1-\cos\left(\theta_{h,w}\right)\right), \\ \cos\left(\theta_{h,w}\right) &= \frac{\hat{f}^r(h,w)\cdot\hat{f}^r_{mis}(h,w)}{\left\|\hat{f}^r(h,w)\right\|\left\|\hat{f}^r_{mis}(h,w)\right\|}. \end{aligned} \tag{5}$$

In Eq. 5, feature consistency is enforced by calculating the cosine similarity between $\hat{f}^r$ and $\hat{f}^r_{mis}$, ensuring that the M_a effectively suppresses sensitive components.

Spatial-Frequency Mapping modulation mask. Although the AFM can adaptively filter out sensitive frequency components, it is prone to overfitting the stereo misalignment patterns in the training data. To mitigate this issue and enhance the robustness to unknown misalignment patterns, we analyze the relationship between spatial structures and corresponding frequency representations and design a Spatial-Frequency Mapping modulation mask (SFM).

Considering that line scanning principle, the rectification error near the horizontal structure has a greater impact on the disparity accuracy than other regions due to pronounced vertical differences among local features near horizontal structures. From a frequency perspective, these sharp variations in local features along the vertical direction are caused by vertical high-frequency components. Through our derivation, we find that two conjugate symmetric points $P_1 = F\,(u,v)$, $P_2 = F\,(W-u, H-v)$ in the frequency domain correspond to a two-dimensional sine wave feature

Authorized licensed use limited to: SHANGHAI INSTITUTE OF MICROSYSTEM AND INFO TECHN. Downloaded on November 14,2025 at 06:41:44 UTC from IEEE Xplore. Restrictions apply.



(a) (b)

Fig. 4. (a) An illustration of the proposed SFM. (b) Examples that SFM filters out the horizontal structures in left image features f^l and obtains $\widetilde{f}^l$. Left:left image. Middle: left image features before filtering f^l. Right: left image features after filtering $\widetilde{f}^l$.

$f(x,y)$ in the spatial domain:

$$f(x,y) = K\left[\|F(u,v)\|\cos\left(2\pi\left(\frac{ux}{W}+\frac{vy}{H}\right)+\arctan\left(\frac{b}{a}\right)\right)\right], \quad (6)$$

where a, b correspond to the real and imaginary parts of $F(u,v)$, respectively, W, H denote the width and height of the feature map, and K is a constant. The position (u,v) determines the direction $\left(\frac{u}{W}, \frac{v}{H}\right)$ and frequencies $f_x = \frac{u}{W}$, $f_y = \frac{v}{H}$ of the sine wave. Thus, the vertical high-frequency components are mapped to the region in the frequency domain near $v = H/2$, as shown in Fig. 4(a).

Then we design the $M_g \in R^{C\times H\times W}$ in the frequency domain to filter out vertical high-frequency components in each channel, as follows:

$$M_g(u,v) = \begin{cases} 0, & (1-\delta)H < v < \delta H \\ 1, & v \geq \delta H \ or \ v \leq (1-\delta)H \end{cases} \quad \delta \in (0.5, 1), \quad (7)$$

where δ denotes the cutoff frequency . This modulation mask M_g is used to modulate the adaptive filtering mask M_a, producing the final frequency-domain filter mask M:

$$M = M_g \odot M_a. \quad (8)$$

Next, we use the mask M with the same filtering process from Eqs. 3 and 4 to obtain $\widetilde{f}^l$ and $\widetilde{f}^r_{mis}$, which are used to build the matching cost volume. As shown in Fig. 4(b), the horizontal structures in the feature maps are removed by the filtering operation.

Through the modulation of M_g, the network reduces reliance on specific stereo misalignment patterns and more effectively suppresses sensitive frequency components, leading to a more robust feature representation.

B. Matching Feature Reconstruction Module

Through SFF, the frequency components sensitive to rectification errors are removed, greatly enhancing the rectification robustness. However, certain feature components crucial for matching but sensitive to rectification errors are removed during filtering, leading to decreased prediction accuracy. To reconstruct these features, the Matching Feature Reconstruction Module (MFRM) is developed to incorporate the separately extracted features and directly use them to activate the aggregated volumes.

As shown in Fig. 2, MFRM employs a non-sharing feature extraction network to extract multi-level basic features $f^i_m, i = 1,2,3$ from the left image. Features at different levels exhibit a preference for different frequency components [30]. To promote positive interaction between different levels of features and fully exploit the geometric information, we progressively fuse the multi-level features to obtain the fused feature F_{fusion}:

$$\begin{aligned} f^i_{fusion} &= Fusion(f^i_m, f^{i-1}_m), \quad i = 2,3 \\ f^i_{fusion} &= f^i_m, \quad i = 1. \end{aligned} \quad (9)$$

To avoid compromising the robustness, we utilize a skip connection structure that directly employs the unfiltered $F_{fusion} = \left\{f^1_{fusion}, f^2_{fusion}, f^3_{fusion}\right\}$ to guide the 4D cost volume in the aggregation step. Drawing on the guidance strategy from [31], we use F_{fusion} to generate multi-layer activation weights $\sigma_i \in R^{C\times H\times W}$. Then, these weights are replicated along the disparity dimension and used to activate the aggregated cost volumes $C_i{}', i = 1,2,3$, which are used for subsequent disparity computation:

$$\begin{aligned} \sigma_i &= sigmoid(F^i_{fusion}), \\ C_i{}' &= \sigma_i \odot C_i, \quad i = 1,2,3. \end{aligned} \quad (10)$$

This module propagates scene information contained in the fused features, reconstructs features lost during filtering, and adaptively optimizes the feature distribution of cost volume. Finally, a balance between rectification robustness and disparity accuracy is achieved.

C. Training Loss

The training loss of the proposed method is defined as:

$$\mathcal{L} = \mathcal{L}_{disp} + \lambda_S \mathcal{L}_S, \quad (11)$$

where $\mathcal{L}_{disp}$ denotes the smooth L1 loss, and $\mathcal{L}_S$ denotes the feature consistency loss. λ_S denotes the weight of the feature consistency loss.

IV. Experiments

A. Datasets

To quantitatively evaluate the performance of our method, we simulate five unrectified stereo datasets from public datasets, including a synthetic SceneFlow dataset (SF) and four real-world datasets: KITTI2015 (KT15), KITTI2012 (KT12), Middlebury (MB), and ETH3D (ET). Similar to [15], to simulate unrectified stereo images, we apply various magnitudes of vertical translation t and rotation rot around the image center to the right view. These two types perturbations disrupt the 1D search assumption in stereo matching algorithms. In the experiments, $t = a$ and $rot = b$ denote that the right view is perturbed using translation and rotation values uniformly sampled from the ranges $[-a, a]$ and $[-b, b]$, respectively. Note that the unit for translation t is pixels, while the unit for rotation rot is degrees.

B. Implementation Details

We applied our method to three widely used stereo networks with different architectures: PSMNet [21], Gwc-Net [22], and RAFT-Stereo [23]. For Gwc-Net, we use the group-wise correlation version. For the SceneFlow dataset, we train the model for 16 epochs with a batch size of 8 using its cleanpass version. For KITTI 2015 and 2012 datasets, the ground truth disparity labels of the test set are not provided. To facilitate the evaluation of the model

Authorized licensed use limited to: SHANGHAI INSTITUTE OF MICROSYSTEM AND INFO TECHN. Downloaded on November 14,2025 at 06:41:44 UTC from IEEE Xplore. Restrictions apply.

TABLE I

ABLATION STUDY OF EACH KEY COMPONENT ON THE SCENEFLOW AND KITTI2015 DATASETS, USING PSMNET AND GWC-NET AS BASELINES. WE REPORT EPE FOR SCENEFLOW AND EPE/D1 FOR KITTI2015, WITH THE BEST RESULTS IN **BOLD**.

Backbone	Components	$rot = 0$	$rot = 0.5$			$rot = 1.0$			$rot = 1.5$		
		$t = 0$	$t = 5$	$t = 10$	$t = 15$	$t = 5$	$t = 10$	$t = 15$	$t = 5$	$t = 10$	$t = 15$
						SceneFlow					
PSMNet	baseline	0.98	1.53	5.05	9.74	2.88	6.26	9.97	4.60	7.42	10.50
	+AFM	1.01	1.51	3.91	9.59	2.49	5.39	9.93	4.03	6.78	10.68
	+AFM+MFRM	**0.93**	**1.33**	2.01	3.83	1.90	2.77	4.36	2.66	3.50	4.94
	+AFM+SFM	1.24	1.64	1.89	2.59	2.16	2.51	3.47	2.90	3.42	4.70
	+AFM+SFM+MFRM	1.00	1.35	**1.54**	**2.01**	**1.84**	**2.07**	**2.56**	**2.40**	**2.65**	**3.20**
Gwc-Net	baseline	**0.82**	1.33	5.66	14.47	2.79	7.71	14.50	5.17	9.51	15.22
	+AFM	1.02	1.51	3.56	10.61	2.42	5.35	11.25	4.22	7.30	12.36
	+AFM+MFRM	0.91	1.31	1.75	3.01	1.81	2.37	3.54	2.44	3.03	4.11
	+AFM+SFM	1.27	1.68	1.93	2.58	2.15	2.47	3.47	2.70	3.16	4.49
	+AFM+SFM+MFRM	0.95	**1.28**	**1.41**	**1.81**	**1.74**	**1.93**	**2.50**	**2.26**	**2.54**	**3.27**
						KITTI2015					
PSMNet	baseline	**0.71/2.2**	1.06/4.0	4.83/23.9	10.71/44.1	2.24/9.0	5.77/25.6	10.02/41.5	4.47/17.5	7.37/31.6	11.64/45.7
	+AFM	0.76/2.5	0.96/3.5	2.30/13.4	6.63/33.3	1.62/6.7	3.31/16.1	6.45/32.2	3.17/13.4	4.76/23.0	7.85/36.0
	+AFM+MFRM	0.76/2.6	0.89/3.4	1.95/11.2	6.03/30.3	1.37/5.9	2.85/14.0	6.20/29.8	2.54/11.7	3.99/20.2	7.73/33.5
	+AFM+SFM	0.79/2.8	**0.88/3.2**	0.98/4.0	1.47/8.3	1.04/4.1	1.24/5.6	1.96/10.6	1.37/7.4	1.70/9.7	2.96/15.0
	+AFM+SFM+MFRM	0.80/2.9	**0.88/3.2**	**0.94/3.7**	**1.29/6.2**	**1.00/4.0**	**1.16/5.0**	**1.65/8.2**	**1.31/6.6**	**1.53/7.8**	**2.37/12.0**
Gwc-Net	baseline	**0.68/2.1**	**0.86/2.9**	2.93/15.4	9.51/36.9	1.54/6.2	3.97/18.2	9.04/35.8	3.18/13.4	5.44/24.7	9.93/39.0
	+AFM	1.02/5.3	1.38/8.5	2.53/19.1	5.21/35.1	1.88/12.0	3.11/21.7	5.54/34.9	2.90/19.3	4.27/28.5	6.58/38.6
	+AFM+MFRM	0.83/3.2	0.97/4.1	1.42/8.2	3.01/19.6	1.21/6.0	1.81/10.8	3.38/20.5	1.79/10.5	2.53/15.5	4.35/24.2
	+AFM+SFM	0.82/2.9	0.92/3.4	1.06/4.7	1.52/9.7	1.05/4.3	1.24/6.2	1.87/11.9	**1.34**/7.3	**1.55**/9.2	2.52/15.3
	+AFM+SFM+MFRM	0.80/3.0	0.87/3.4	**0.94/3.9**	**1.25/6.4**	**1.02/4.2**	**1.18/5.3**	**1.58/8.3**	**1.34/6.6**	**1.55/8.1**	**2.44/11.5**

TABLE II

COMPARISON BETWEEN FREQUENCY-DOMAIN FILTERING (OURS) AND SPATIAL-DOMAIN FEATURE CONSISTENCY LOSS ON SCENEFLOW. END-POINT ERROR (EPE) IS ADOPTED.

Method		$rot = 0.5$			1.0		
		$t = 5$	10	15	5	10	15
PSMNet	Ours w/o AFM	1.40	3.40	6.40	2.36	4.32	6.67
	Ours	**1.33**	**2.01**	**3.83**	**1.90**	**2.77**	**4.36**
Gwc-Net	Ours w/o AFM	**1.18**	1.83	3.50	**1.74**	2.54	4.01
	Ours	1.31	**1.75**	**3.01**	1.81	**2.37**	**3.54**

under various perturbations, we use 160 stereo pairs from the original training set for training and the rest for testing, then fine-tune the SceneFlow pre-trained model for 300 epochs with the batch size of 8. For the Middlebury and ETH3D datasets, we follow the split settings from [32] to create the training and test sets at quarter resolution. We then fine-tune the SceneFlow pre-trained model for 300 epochs on these two datasets, with the batch sizes of 6 and 5, respectively. For RAFT-Stereo, we train the model for total 200k steps with the batch size of 6 on SceneFlow. Then, fine-tune the pre-trained model for 5K steps on the KITTI datasets and 1K steps on the Middlebury and ETH3D datasets. For different hyper-parameters, the δ in Eq. (7) is set as 0.9. The λ_S in Eq. (11) is set as 20. For all models, during training, we add perturbations based on the size of the stereo images: $rot = 0.5, t = 2$ for the KITTI dataset and $rot = 0.5, t = 3$ for the other data to obtain I^l, I^r_{mis}.

C. Ablation Study

In this section, we conduct detailed ablation studies to evaluate the effectiveness of our method. The evaluation results are listed in Table I.

Adaptive Filtering Module (AFM). As shown in the 1st(baseline) and 2nd of each sub-table in Table I, the AFM is beneficial for improving the robustness to rectification errors. This shows that our AFM is able to adaptively filter out sensitive features under the supervision of the feature consistency loss, forcing the network to rely on more robust geometric features for disparity estimation.

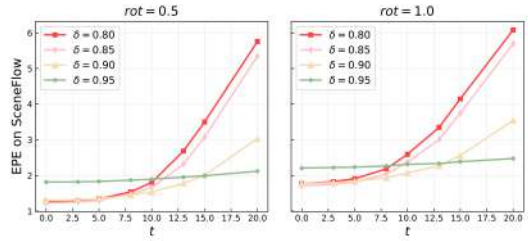


Fig. 5. Ablation study for the cutoff frequency in SFM. End-point error (EPE) is adopted. We select PSMNet as the baseline model and evaluated it on the SceneFlow dataset.

In addition, in order to verify the necessity of frequency-domain-based filtering, we conducted ablated experiments based on our model corresponding to the third-row for simplicity. Table II reports the results. It can be observed that the removal of AFM leads to a decrease in robustness. That is, compared to the model with only the spatial-domain constraint $\mathcal{L}_S$, AFM filtering in the frequency domain more efficiently enhances the rectification robustness, indicating the effectiveness of frequency domain operations.

Spatial-Frequency Mapping modulation mask (SFM). From the 2nd and 4th rows of each sub-table in Table I, the evaluation performance shows further improvement under more severe misalignment conditions than in the training data. This result indicates that our SFM effectively mitigates overfitting tendencies and enhances the ability of the stereo networks to cope with more severe rectification errors.

The results in Fig. 5 suggest that a higher cutoff frequency more completely removes sensitive features, leading to greater robustness against severe rectification errors. However, when the cutoff frequency becomes too high ($\delta = 0.95$), the features used for matching are excessively filtered, leading to a reduction in prediction accuracy. Therefore, we choose $\delta = 0.9$ to balance accuracy under small rectification errors and robustness against large rectification errors.

Matching Feature Reconstruction Module (MFRM). As shown in Table I, the models with the AFM and SFM

Authorized licensed use limited to: SHANGHAI INSTITUTE OF MICROSYSTEM AND INFO TECHN. Downloaded on November 14,2025 at 06:41:44 UTC from IEEE Xplore. Restrictions apply.

TABLE III

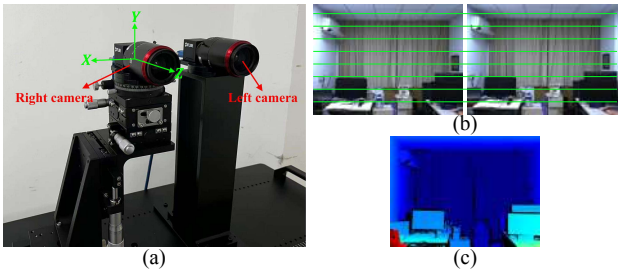
RECTIFICATION ROBUSTNESS WITH AND WITHOUT THE PROPOSED F^2R^2. THE EPE METRIC IS USED.

	F^2R^2	Data	$rot = 0.0$	0.5			1.0		
			$t = 0$	5	10	15	5	10	15
RAFT-Stereo	✘	SF	**0.47**	0.68	1.41	2.60	1.04	1.77	2.95
	✔	SF	0.55	**0.66**	**0.75**	**0.97**	**0.97**	**1.07**	**1.35**
	✘	KT15	**0.65**	0.80	2.13	4.25	1.11	2.34	3.94
	✔	KT15	0.69	**0.73**	**0.79**	**1.22**	**0.82**	**0.96**	**1.61**
	✘	KT12	**0.60**	0.94	4.45	10.42	2.01	4.83	9.85
	✔	KT12	0.68	**0.74**	**0.87**	**1.57**	**0.86**	**1.09**	**2.39**
	✘	MB	**1.10**	1.55	2.87	6.78	2.01	3.57	5.83
	✔	MB	1.34	**1.53**	**1.58**	**1.85**	**1.87**	**1.88**	**2.10**
	✘	ET	**0.30**	**0.41**	1.02	2.16	**0.65**	1.17	1.76
	✔	ET	0.32	0.46	**0.55**	**0.70**	0.68	**0.75**	**0.84**
PSMNet	✘	KT12	**0.69**	1.19	8.18	18.01	3.15	9.06	17.39
	✔	KT12	0.77	**0.92**	**1.05**	**1.66**	**1.19**	**1.36**	**2.39**
	✘	MB	1.44	1.92	4.00	7.42	2.55	4.83	7.17
	✔	MB	**1.42**	**1.69**	**1.83**	**2.18**	**2.14**	**2.23**	**2.59**
	✘	ET	**0.60**	0.89	3.14	7.92	1.55	3.70	5.71
	✔	ET	0.61	**0.85**	**0.98**	**1.82**	**1.22**	**1.49**	**2.39**
Gwc-Net	✘	KT12	**0.65**	**0.92**	3.79	13.21	2.05	5.33	13.89
	✔	KT12	0.92	1.08	**1.25**	**2.73**	**1.36**	**1.81**	**3.70**
	✘	MB	**1.32**	1.81	4.28	11.21	2.52	5.50	10.68
	✔	MB	1.40	**1.67**	**1.77**	**2.23**	**2.09**	**2.14**	**2.47**
	✘	ET	**0.52**	**0.85**	7.15	17.22	1.78	7.46	14.09
	✔	ET	0.65	0.88	**0.97**	**2.03**	**1.22**	**1.37**	**2.55**

exhibits a decline in prediction accuracy on stereo images with minor rectification errors, particularly on the SceneFlow dataset with dense ground truth. This suggests that some critical information beneficial for matching is lost during the filtering process. The 3rd and 5th rows of each sub-table show that MFRM successfully reconstructs the lost features and improves the disparity prediction accuracy. Additionally, the activation in MFRM corrects feature distribution in the cost volume, enhancing network robustness.

D. Universality Analysis

To further evaluate the universality of our method, in addition to PSMNet and Gwc-Net, we apply our method to the represented GRU-based RAFT-Stereo and further conduct experiments on the Middlebury, ETH3D, and KITTI2012 datasets. The results are summarized in Table III. Compared to baseline models, our approach significantly improves prediction accuracy under larger rectification errors while maintaining comparable performance with minor errors or perfect calibration. This shows that our method achieves a trade-off between disparity accuracy and rectification robustness.



(a) (b) (c)

Fig. 6. The data collection system and a data example. (a) our extrinsic parameter-adjustable stereo system. (b) self-collected real unrectified stereo pairs. (c) disparity ground truth of left view.

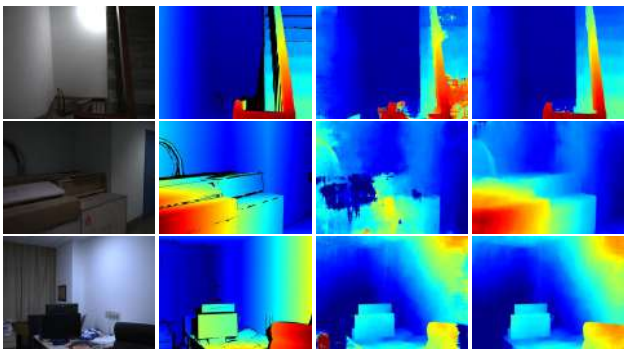
E. Experiments on the self-collected Real-world Data

To validate the performance of our method in real-world stereo systems, we collected an unrectified stereo dataset with real rectification errors. This dataset provides dense disparity ground truth of the left image for quantitative evaluation. To capture the unrectified stereo pairs, we build an extrinsic parameter-adjustable stereo system, as shown in Fig. 6. For stereo pair collection, we fixed the left camera and applied rotational perturbations of $\pm 0.5°$ (x-axis), $\pm 1.5°$ (z-axis), and translational perturbations of $\pm 1cm$ (y-axis) to the right camera. We adopted the same data collection strategy as [33] to obtain the disparity ground truth. We captured 60 stereo pairs with a resolution of 986×730, using 42 pairs for training and the rest for testing.

TABLE IV

QUANTITATIVE EVALUATION ON OUR UNRECTIFIED STEREO DATASET USING THE EPE METRIC, WITH rot_x, rot_z FOR ROTATIONS AROUND THE X AND Z AXES, AND t_y FOR TRANSLATION ALONG THE Y-AXIS. THE TIME DENOTES THE INFERENCE TIME ON SINGLE NVIDIA RTX 3090.

Method	F^2R^2	rot_x	rot_z	t_y	Time(s)
Gwc-Net	✘	2.48	5.16	**0.76**	0.24
	✔	**1.13**	**4.64**	0.84	0.29
PSMNet	✘	3.87	5.73	0.78	0.39
	✔	**0.90**	**4.75**	**0.70**	0.43



Left Image Ground-Truth PSMNet F^2R^2-PSMNet

Fig. 7. Qualitative results on the self-collected test data.

As shown in Table IV, compared to the baseline models, our models effectively improve the disparity prediction accuracy for stereo images with real rectification errors while maintaining a comparable inference time. Compared to rotation, the translation has a smaller impact on matching since the vertical pixel shift decreases as the depth of the corresponding pixels increases, which explains why the translation index is not affected. Additionally, we note that rotation around the z-axis causes additional horizontal shifts, which alters the ground truth disparity and rapidly increases matching errors. Fig. 7 shows the qualitative results. Our method significantly reduces the impact of real rectification errors on matching.

V. CONCLUSIONS

We propose a novel frequency filtering-based rectification robustness (F^2R^2) method for stereo matching. The proposed SFF precisely removes sensitive frequency components, enabling stereo networks to handle rectification errors. Additionally, the proposed MFRM reconstructs matching features by incorporating separately extracted multi-level features, balancing the disparity accuracy and rectification robustness. Through extensive experimental validation, our method shows superior robustness performance in tackling rectification errors.

Authorized licensed use limited to: SHANGHAI INSTITUTE OF MICROSYSTEM AND INFO TECHN. Downloaded on November 14,2025 at 06:41:44 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] J. Biswas and M. Veloso, "Depth camera based localization and navigation for indoor mobile robots," in *RGB-D Workshop at RSS*, vol. 2011, no. 21, 2011.

[2] Y. Bi, C. Li, X. Tong, G. Wang, and H. Sun, "An application of stereo matching algorithm based on transfer learning on robots in multiple scenes," *Scientific Reports*, vol. 13, no. 1, p. 12739, 2023.

[3] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3061–3070.

[4] C. Liu, W. Wei, B. Liang, X. Liu, W. Shang, and J. Li, "Convmlp-mixer based real-time stereo matching network towards autonomous driving," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 2581–2586, 2022.

[5] S. N. Sinha and M. Pollefeys, "Multi-view reconstruction using photo-consistency and exact silhouette constraints: A maximum-flow formulation," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1. IEEE, 2005, pp. 349–356.

[6] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.

[7] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[8] Y. Ling and S. Shen, "High-precision online markerless stereo extrinsic calibration," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1771–1778.

[9] P. Hansen, H. Alismail, P. Rander, and B. Browning, "Online continuous stereo extrinsic parameter estimation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1059–1066.

[10] E. Dexheimer, P. Peluse, J. Chen, J. Pritts, and M. Kaess, "Information-theoretic online multi-camera extrinsic calibration," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4757–4764, 2022.

[11] T. Dang, C. Hoffmann, and C. Stiller, "Continuous stereo self-calibration by camera parameter tracking," *IEEE Transactions on image processing*, vol. 18, no. 7, pp. 1536–1550, 2009.

[12] H. Zhao, Y. Zhang, Q. Chen, and R. Fan, "Dive deeper into rectifying homography for stereo camera online self-calibration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 479–14 485.

[13] T. Dang, C. Hoffmann, and C. Stiller, "Self-calibration for active automotive stereo vision," in *2006 IEEE Intelligent Vehicles Symposium*. IEEE, 2006, pp. 364–369.

[14] J. Wang, D. Scharstein, A. Bapat, K. Blackburn-Matzen, M. Yu, J. Lehman, S. Alsisan, Y. Wang, S. Tsai, J.-M. Frahm *et al.*, "A practical stereo depth system for smart glasses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 498–21 507.

[15] H. Hirschmuller and S. Gehrig, "Stereo matching in the presence of sub-pixel calibration errors," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 437–444.

[16] S. Lin, Z. Zhang, Z. Huang, Y. Lu, C. Lan, P. Chu, Q. You, J. Wang, Z. Liu, A. Parulkar *et al.*, "Deep frequency filtering for domain generalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 797–11 807.

[17] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[18] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2811–2820.

[19] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1959–1968.

[20] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 66–75.

[21] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410–5418.

[22] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3273–3282.

[23] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[24] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[25] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 701–19 710.

[26] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5515–5524.

[27] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 263–16 272.

[28] F. Li, L. Zhang, Z. Liu, J. Lei, and Z. Li, "Multi-frequency representation enhancement with privilege information for video super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 814–12 825.

[29] L. Chen, L. Gu, D. Zheng, and Y. Fu, "Frequency-adaptive dilated convolution for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3414–3425.

[30] Q. Bu, D. Huang, and H. Cui, "Towards building more robust models with frequency bias," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4402–4411.

[31] A. Bangunharcana, J. W. Cho, S. Lee, I. S. Kweon, K.-S. Kim, and S. Kim, "Correlate-and-excite: Real-time stereo matching via guided cost volume excitation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3542–3548.

[32] M. Poggi, D. Pallotti, F. Tosi, and S. Mattoccia, "Guided stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 979–988.

[33] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3260–3269.

Authorized licensed use limited to: SHANGHAI INSTITUTE OF MICROSYSTEM AND INFO TECHN. Downloaded on November 14,2025 at 06:41:44 UTC from IEEE Xplore. Restrictions apply.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：龚帅宇
方向：集成电路
单位：中国科学院微电子研究所

题目：人工反铁磁斯格明子VCMA-MRAM器件及连续型阵列设计

Manipulation of Synthetic Antiferromagnetic Skyrmion by Reduced Voltage via Design of Double Interface Structure

Shuaiyu Gong, Meiyin Yang, Yanru Li, Bowen Shen, Yongliang Li, Peiyue Yu, Yan Cui, and Jun Luo

***Abstract*— Magnetic skyrmions in synthetic antiferromagnet exhibit exceptional stability without stray field and transport bias, which holds significant potential in future memory, logic and artificial intelligent devices. However, achieving voltage control of synthetic antiferromagnetic (SAF) skyrmions poses challenges due to the thick and multi-layer film structure. Here, we design a double interface structure in a magnetic tunnel junction (MTJ) for voltage control of SAF skyrmions. The voltage applied for manipulating SAF skyrmions can be reduced by 50% compared to conventional single interface structure. In addition, the double interface structure allows for more compact device sizes, broader ranges of magnetic property parameters and lower energy consumption. Our studies provide a promising solution for the effective manipulation of antiferromagnetic skyrmion in SAF.**

***Index Terms*— Synthetic antiferromagnetic materials, skyrmions, voltage-controlled magnetic anisotropy.**

I. Introduction

THE magnetic skyrmion is a topologically non-trivial chiral spin texture [1], which has the advantages of high topological stability, small size, low driving current, and high transport speed [2]. Thus, it holds potential across various applications such as logic gates [3], spin nanoscale oscillators [4], artificial synapses [5], artificial neurons [6], reserve pool computing [7], [8], and random number generators [9], [10]. However, ferromagnetic skyrmions are facing critical challenges such as poor stability without external field [11], [12], [13] and transport bias arising from the skyrmion Hall effect [14], [15], [16]. In recent years, synthetic antiferromagnetic (SAF) skyrmions have been proposed, which exhibit good stability and no transport bias due to their antiparallel coupled magnetization structure [17], [18], [19].

Manuscript received 3 March 2024; accepted 20 March 2024. Date of publication 28 March 2024; date of current version 26 April 2024. This work was supported in part by Chinese Academy of Sciences under Grant XDA18000000 and Grant Y201926, in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) under Grant 2020118, and in part by Beijing Nova Program under Grant 20230484358. The review of this letter was arranged by Editor G. Hu. *(Corresponding authors: Jun Luo; Meiyin Yang.)*

Shuaiyu Gong, Meiyin Yang, Bowen Shen, Yongliang Li, Yan Cui, and Jun Luo are with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China, and also with the School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yangmeiyin@ime.ac.cn; luojun@ime.ac.cn).

Yanru Li and Peiyue Yu are with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China.

Color versions of one or more figures in this letter are available at https://doi.org/10.1109/LED.2024.3382686.

Digital Object Identifier 10.1109/LED.2024.3382686

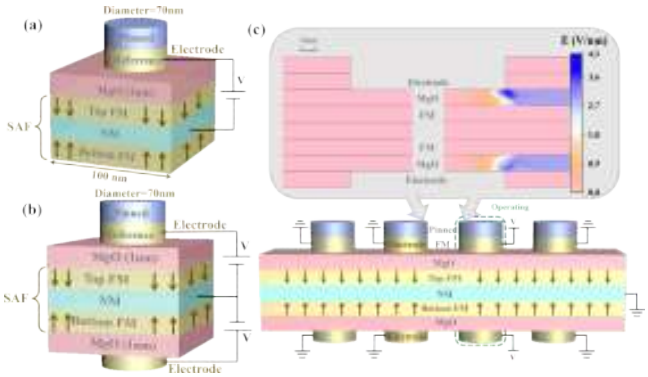


Fig. 1. (a) The schematics of single interface structure (SIS); (b) The schematics of double interface structure (DIS); (c) The schematics of the DIS array and the electric field distribution at the edge of electrode.

The manipulation of SAF skyrmions is essential for its applications in functional devices and various methods had been proposed. For example, the creation of SAF skyrmions can be realized by polarized current [20], external magnetic field [21], [22], laser pulses [23], or deformation of domain wall (DW) [24]. Skyrmions can also be annihilated through external magnetic fields, polarization currents, and deformation at the edges of materials [25]. To save the energy consumption, voltage-controlled magnetic anisotropy (VCMA) is used to manipulate SAF skyrmions via additional ferromagnetic system [26]. However, due to the presence of a highly coupled additional ferromagnetic layer with SAF, it did not completely avoid the problems of ferromagnetic skyrmions in MTJ, such as poor stability under temperature and magnetic field fluctuation. Up to now, direct manipulation of skyrmions in SAF system remains challenging. A significant obstacle is that VCMA is an interface effect, impacting only the interface between the top ferromagnet and MgO (Fig. 1(a)), leaving the bottom ferromagnet unaffected. To achieve high-density skyrmions with ultra-low energy consumption, the generation of SAF skyrmions through VCMA becomes highly desirable.

In this letter, we design a double interface structure (DIS) in a magnetic tunnel junction for manipulating SAF skyrmion directly by voltage. Both top and bottom ferromagnets in SAF can experience VCMA effect for this structure. Compared with the conventional single interface structure (SIS), the DIS significantly reduces the voltage for SAF creation

0741-3106 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Institute of Microelectronics CAS. Downloaded on September 11,2024 at 02:12:51 UTC from IEEE Xplore. Restrictions apply.

and annihilation of skyrmion. Furthermore, the DIS can allow larger range of Dzyaloshinskii-Moriya interaction (DMI) parameter (D) and exchange coupling constant (A_{ex}) for skyrmions manipulation, and also enables smaller device size and lower energy consumption compared with SIS. Our studies provide a promising solution for manipulating antiferromagnetic skyrmion in SAF.

II. Device Structure

The creation and annihilation of skyrmions in SAF structure were conducted through micromagnetic simulations utilizing the object oriented micromagnetic framework (OOMMF). The simulation model comprises a magnetic tunnel junction structure illustrated in Fig. 1(a). The SIS consists of a square free layer of SAF and a MgO barrier layer with dimension of 100 nm × 100 nm. It also includes a circular reference ferromagnetic layer and a pinned layer with varying diameters, which is necessary in MTJ structure. The voltage is applied across the MgO layer, with the reference layer and pinned layer serving as the top electrode for the electric field.

For the DIS shown in Fig. 1(b), an additional MgO layer is introduced beneath the SAF layer, and the bottom electrode is positioned under the MgO layer. Notably, perpendicular SAF can be successfully grown in this structure [27]. The voltage can be applied across the two MgO layers at the same time using the DIS, which can induce VCMA for both ferromagnets (FM) in SAF. To induce the same perpendicular magnetic anisotropy constant (K_u) of the two ferromagnets by voltage, opposite sign of voltage is applied at the top and bottom electrodes for the DIS. The size of calculation unit is 2 nm × 2 nm × 1 nm. The saturation magnetization (M) of both ferromagnetic layers in SAF is 1×10^6 A·m^{-1}. The A_{ex}, the K_u and the D are variable parameters in our research. The antiferromagnetic exchange coupling constant (J_{ex}) between the bottom FM layer and the top FM layer is fixed at -2×10^{-4} J·m^{-2}.

Figure 1(c) illustrates the configuration of the DIS arrays within a continuous SAF, allowing for a more condensed intensity distribution compared to DIS arrays in isolated SAF. A voltage of 3 V was applied to one device while the remaining devices were grounded. The electric field distribution for this arrangement was computed utilizing Gaussian law. The outcomes indicate that the applied voltage on a single device does not impact the adjacent devices, affirming the feasibility of the DIS array configuration.

III. Results and Discussion

Fig. 2 illustrates the creation and annihilation process of SAF skyrmions, which can be realized by reducing and raising the value of K_u, respectively. Similar with the ferromagnetic system [28], the increase of perpendicular magnetic anisotropy energy (E_{PMA}) is less than the decrease of DMI energy (E_{DMI}) due to the reduced K_u (Fig. 2(a)), resulting the formation of SAF skyrmions. When increasing the K_u, the reduction of E_{PMA} is much larger than the increment of E_{DMI} in single domain, which erases the SAF skyrmion (Fig. 2(b)). The K_u can be varied by VCMA effect which makes the control of SAF skyrmions by voltage possible [29].

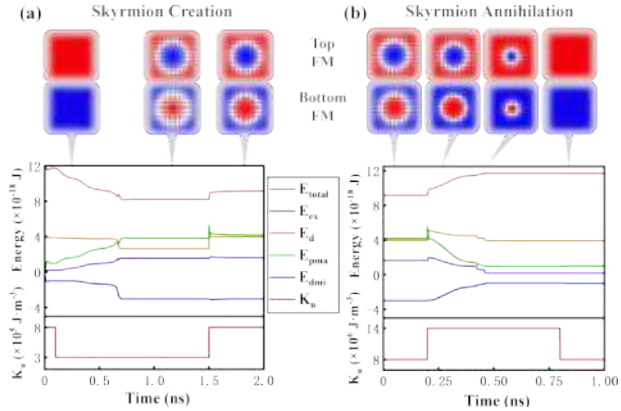


Fig. 2. (a) The energy and K_u vs. time in the creation process of SAF skyrmions. The inset is the spin texture of top and bottom ferromagnetic layer corresponding to the particular time; (b) The energy and K_u vs. time in the annihilation process of SAF skyrmions.

Although the control of skyrmion can be realized by varying K_u, the value for creating and annihilating skyrmions in SIS is very extreme. The SAF skyrmions can be created by K_u below 0.3×10^6 J·m^{-3}, highlighted by the green polygon in Fig. 3(a). The erasing of SAF skyrmions requires K_u higher than 1.4×10^6 J·m^{-3} as shown in Fig. 3(c). To achieve such values of K_u, a very large voltage [30] is required to applied across the MgO, which will destroy the MgO barrier [31]. After adding other MgO/FM interface for the DIS, the critical values of K_u in both FM layers for creating and annihilating skyrmions are 0.55×10^6 J·m^{-3} and 1.05×10^6J·m^{-3} (shown in Fig. 3(b) and (d)), respectively, which are much closer to original K_u (0.8×10^6 J·m^{-3}). The smaller K_u variation required for skyrmion manipulation in the DIS is attributed to the simultaneous flip of spin in the top and bottom FM layers by VCMA (Fig. 3(g)). However, in the SIS, the spin of top FM requires additional energy to flip due to the strong antiferromagnetic exchange coupling from bottom FM, leading to a larger variation of K_u.

The comparation of K_u and corresponding applied voltage value is shown in Fig. 3(e). The corresponding electric field, which determines the applied voltage, is calculated from the VCMA coefficient and the variation of K_u (We take the VCMA coefficient to be 100 fJ/(V·m) [32]). The voltage applied to the DIS is 50% smaller compared to the SIS, which is comparable to the reported voltage values for manipulating skyrmions in ferromagnetic film [28]. Considering the successful experimental manipulation of skyrmions in ferromagnetic film, it suggests that the DIS also makes experimental implementation for feasible.

Moreover, the diameter of the electrodes plays a crucial role in the creation of skyrmions. For the SIS, the optimal diameter range is between 60 nm and 90 nm. Below 60 nm, the electric area at the MgO/FM interface becomes too small to significantly impact the spin textures. When with electrode diameters exceeding 90 nm, spin flips occur primarily at the boundary of the SAF after reducing K_u, making it challenging to create skyrmion in the central region. Interestingly, during the annihilation process, skyrmions do not entirely disappear but instead expand when the electrode diameter is smaller than 60 nm (Fig. 3(c) and (d)). In this case, the region influenced

Authorized licensed use limited to: Institute of Microelectronics CAS. Downloaded on September 11,2024 at 02:12:51 UTC from IEEE Xplore. Restrictions apply.

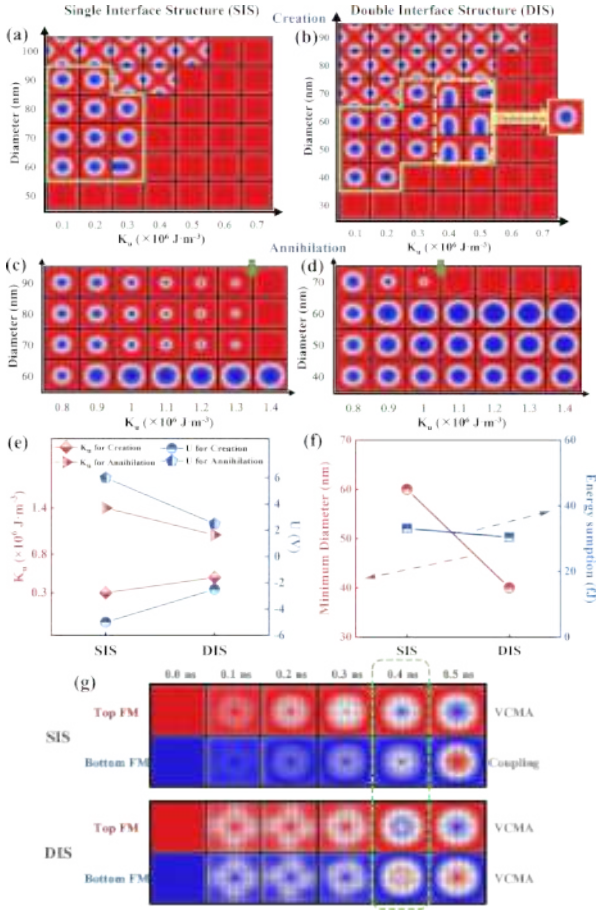


Fig. 3. The spin textures at different electrode diameters when (a) decreasing K_U in the SIS; (b) decreasing K_U in the DIS. Some magnetic domains are touching the boundary of SAF layer. These can be slightly optimized by adjusting material parameters in appropriate range, such as increasing the dimension of SAF by 10%, as shown in the yellow arrow; (c) increasing K_U in the SIS; (d) increasing K_U in the DIS; (e) The critical values of K_U and voltages for creating and annihilating skyrmions in the SIS/DIS; (f) Comparison of the minimum electrode diameters for creating skyrmion and energy consumption in an operation cycle between SIS and DIS. The energy consumption is estimated using the resistivity of MgO (~1 nm) to be 300 Ω·cm [33] without VCMA effect. (g) Spin textures of the top and bottom FM layers during skyrmion creation in the SIS/DIS.

by VCMA is smaller than the overall area of the skyrmion. Only the perpendicular magnetic anisotropy (PMA) at the central area increases, leading to the growing of skyrmion bubble. Comparing the electrode diameters required for creating skyrmions in both structures (Fig. 3(f)), it is evident that the DIS can accommodate smaller electrodes, such as 40 nm.

The energy consumption of the SIS and the DIS for one cycle of skyrmion creation and annihilation are 33 fJ and 30 fJ as plotted in Fig. 3(f). Although there are two voltage channels and requires double operation time for the DIS, the amplitude of applied voltage is about half of that in the SIS, resulting in the lower energy consumption. In addition, when considering the decrease of resistance caused by VCMA [34], the resistance of the SIS at 5V drops 60% more than that of DIS at 2.5V, which results in a higher energy consumption. Therefore, the energy consumption of the DIS is more advantageous compared with SIS. Moreover, the breakdown of MgO in the SIS due to high operating voltage is greatly avoided in the DIS.

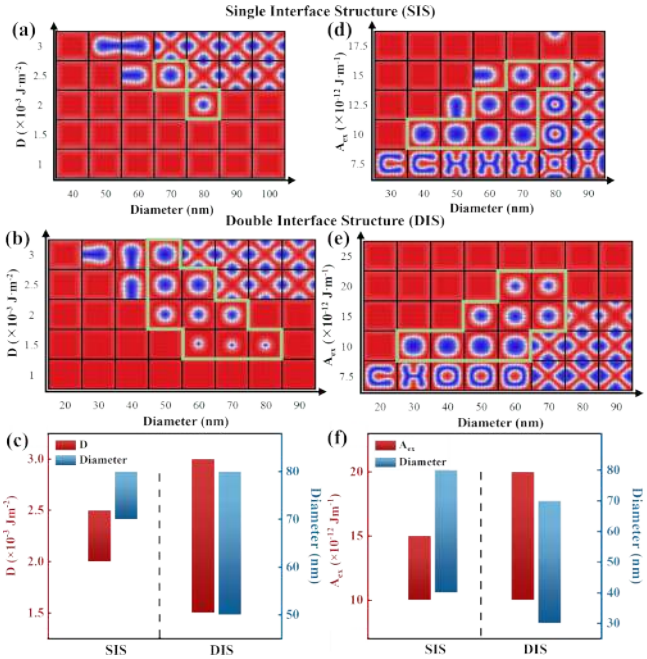


Fig. 4. The spin texture in the creation of skyrmions at different electrode diameters and DMI parameters when (a) in the SIS; (b) in the DIS; (c) Comparison of allowable DMI parameter range and electrode size in both structures; The spin texture in the creation of skyrmions at different electrode diameters and A_{ex} constants when (d) in the SIS; (e) in the DIS; (f) Comparison of allowable A_{ex} constant range and electrode size in both structures.

In addition to the PMA, the DMI and exchange interaction are crucial effects influencing the creation and annihilation of skyrmions. For both the SIS and the DIS, various values of the D were selected, alongside different electrode sizes. The resulting spin textures are illustrated in Fig. 4(a) and (b). In the case of the SIS, a DMI range of 2.0×10^{-3} J·m^{-2} to 2.5×10^{-3} J·m^{-2}, coupled with electrode diameters between 70 nm and 80 nm, is conducive to creating SAF skyrmion (refer to Fig. 4(a)). The DIS widens the DMI range to 1.5×10^{-3} J·m^{-2} -3.0×10^{-3} J·m^{-2} and simultaneously reduces the minimum diameter of the electrode to 50 nm. The Fig. 4(c) provides a comparison of the ranges of D required for skyrmions creation and the electrode diameters for both structures. Similar trends are observed for the A_{ex}, as depicted in Fig. 4(d)-(f). In contrast to the SIS, which necessitates A_{ex} values in the range of 1.0 to 1.5×10^{-11} J·m^{-1}, the DIS extends the maximum A_{ex} to 2.0×10^{-11} J·m^{-1} while also significantly reducing the diameter of the electrodes. The broader magnetic parameter ranges facilitate the experimental implementation.

IV. Conclusion

In summary, we have designed the DIS for the manipulation of SAF skyrmions. By increasing another interface for VCMA, the voltage for manipulating skyrmions is 50% smaller than that of SIS. Furthermore, the DIS features smaller electrode diameters, lower energy consumption and provides greater flexibility in device design, including wider values of DMI parameter and A_{ex} constant. These findings present a promising and efficient solution for the manipulation of skyrmions in SAF, paving the way for advancements in spintronic applications and magnetic storage technologies.

Authorized licensed use limited to: Institute of Microelectronics CAS. Downloaded on September 11,2024 at 02:12:51 UTC from IEEE Xplore. Restrictions apply.

References

[1] A. Abanov and V. L. Pokrovsky, "Skyrmion in a real magnetic film," *Phys. Rev. B, Condens. Matter*, vol. 58, no. 14, pp. R8889–R8892, Oct. 1998, doi: 10.1103/physrevb.58.r8889.

[2] S. Li, W. Kang, X. Zhang, T. Nie, Y. Zhou, K. L. Wang, and W. Zhao, "Magnetic skyrmions for unconventional computing," *Mater. Horizons*, vol. 8, no. 3, pp. 854–868, Mar. 2021, doi: 10.1039/d0mh01603a.

[3] Y. Zhou and M. Ezawa, "A reversible conversion between a skyrmion and a domain-wall pair in a junction geometry," *Nature Commun.*, vol. 5, no. 1, p. 4652, Aug. 2014, doi: 10.1038/ncomms5652.

[4] S. Zhang, J. Wang, Q. Zheng, Q. Zhu, X. Liu, S. Chen, C. Jin, Q. Liu, C. Jia, and D. Xue, "Current-induced magnetic skyrmions oscillator," *New J. Phys.*, vol. 17, no. 2, Feb. 2015, Art. no. 023061, doi: 10.1088/1367-2630/17/2/023061.

[5] K. M. Song, J.-S. Jeong, B. Pan, X. Zhang, J. Xia, S. Cha, T.-E. Park, K. Kim, S. Finizio, J. Raabe, J. Chang, Y. Zhou, W. Zhao, W. Kang, H. Ju, and S. Woo, "Skyrmion-based artificial synapses for neuromorphic computing," *Nature Electron.*, vol. 3, no. 3, pp. 148–155, Mar. 2020, doi: 10.1038/s41928-020-0385-0.

[6] S. Li, W. Kang, Y. Huang, X. Zhang, Y. Zhou, and W. Zhao, "Magnetic skyrmion-based artificial neuron device," *Nanotechnology*, vol. 28, no. 31, Aug. 2017, Art. no. 31LT01, doi: 10.1088/1361-6528/aa7af5.

[7] D. Prychynenko, M. Sitte, K. Litzius, B. Krüger, G. Bourianoff, M. Kläui, J. Sinova, and K. Everschor-Sitte, "Magnetic skyrmion as a nonlinear resistive element: A potential building block for reservoir computing," *Phys. Rev. Appl.*, vol. 9, no. 1, Jan. 2018, Art. no. 014034.

[8] G. Bourianoff, D. Pinna, M. Sitte, and K. Everschor-Sitte, "Potential implementation of reservoir computing models based on magnetic skyrmions," *AIP Adv.*, vol. 8, no. 5, May 2018, Art. no. 055602, doi: 10.1063/1.5006918.

[9] Y. Yao, X. Chen, W. Kang, Y. Zhang, and W. Zhao, "Thermal Brownian motion of skyrmion for true random number generation," *IEEE Trans. Electron Devices*, vol. 67, no. 6, pp. 2553–2558, Jun. 2020, doi: 10.1109/TED.2020.2989420.

[10] D. Pinna, F. Abreu Araujo, J.-V. Kim, V. Cros, D. Querlioz, P. Bessiere, J. Droulez, and J. Grollier, "Skyrmion gas manipulation for probabilistic computing," *Phys. Rev. Appl.*, vol. 9, no. 6, Jun. 2018, Art. no. 064018, doi: 10.1103/physrevapplied.9.064018.

[11] Y. Tokunaga, X. Z. Yu, J. S. White, H. M. Rønnow, D. Morikawa, Y. Taguchi, and Y. Tokura, "A new class of chiral materials hosting magnetic skyrmions beyond room temperature," *Nature Commun.*, vol. 6, no. 1, p. 7638, Jul. 2015, doi: 10.1038/ncomms8638.

[12] H. Du, R. Che, L. Kong, X. Zhao, C. Jin, C. Wang, J. Yang, W. Ning, R. Li, C. Jin, X. Chen, J. Zang, Y. Zhang, and M. Tian, "Edge-mediated Skyrmion chain and its collective dynamics in a confined geometry," *Nature Commun.*, vol. 6, no. 1, p. 8504, Oct. 2015, doi: 10.1038/ncomms9504.

[13] X. Z. Yu, N. Kanazawa, Y. Onose, K. Kimoto, W. Z. Zhang, S. Ishiwata, Y. Matsui, and Y. Tokura, "Near room-temperature formation of a skyrmion crystal in thin-films of the helimagnet FeGe," *Nature Mater.*, vol. 10, no. 2, pp. 106–109, Feb. 2011, doi: 10.1038/nmat2916.

[14] N. Nagaosa and Y. Tokura, "Topological properties and dynamics of magnetic skyrmions," *Nature Nanotechnol.*, vol. 8, no. 12, pp. 899–911, Dec. 2013, doi: 10.1038/nnano.2013.243.

[15] A. Fert, V. Cros, and J. Sampaio, "Skyrmions on the track," *Nature Nanotechnol.*, vol. 8, no. 3, pp. 152–156, Mar. 2013, doi: 10.1038/nnano.2013.29.

[16] J. Iwasaki, M. Mochizuki, and N. Nagaosa, "Current-induced skyrmion dynamics in constricted geometries," *Nature Nanotechnol.*, vol. 8, no. 10, pp. 742–747, Oct. 2013, doi: 10.1038/nnano.2013.176.

[17] J. Barker and O. A. Tretiakov, "Static and dynamical properties of antiferromagnetic skyrmions in the presence of applied current and temperature," *Phys. Rev. Lett.*, vol. 116, no. 14, Apr. 2016, Art. no. 147203, doi: 10.1103/physrevlett.116.147203.

[18] X. Zhang, Y. Zhou, and M. Ezawa, "Antiferromagnetic skyrmion: Stability, creation and manipulation," *Sci. Rep.*, vol. 6, no. 1, p. 24795, Apr. 2016, doi: 10.1038/srep24795.

[19] C. Jin, C. Song, J. Wang, and Q. Liu, "Dynamics of antiferromagnetic skyrmion driven by the spin Hall effect," *Appl. Phys. Lett.*, vol. 109, no. 18, Oct. 2016, Art. no. 182404, doi: 10.1063/1.4967006.

[20] X. Zhang, Y. Zhou, and M. Ezawa, "Magnetic bilayer-skyrmions without skyrmion Hall effect," *Nature Commun.*, vol. 7, no. 1, 2016, Art. no. 10293, doi: 10.1038/ncomms10293.

[21] W. Legrand, D. Maccariello, F. Ajejas, S. Collin, A. Vecchiola, K. Bouzehouane, N. Reyren, V. Cros, and A. Fert, "Room-temperature stabilization of antiferromagnetic skyrmions in synthetic antiferromagnets," *Nature Mater.*, vol. 19, no. 1, pp. 34–42, Jan. 2020, doi: 10.1038/s41563-019-0468-3.

[22] T. Dohi, S. DuttaGupta, S. Fukami, and H. Ohno, "Formation and current-induced motion of synthetic antiferromagnetic skyrmion bubbles," *Nature Commun.*, vol. 10, no. 1, p. 5153, Nov. 2019, doi: 10.1038/s41467-019-13182-6.

[23] R. Juge, N. Sisodia, J. U. Larrañaga, Q. Zhang, V. T. Pham, K. G. Rana, B. Sarpi, N. Mille, S. Stanescu, R. Belkhou, M.-A. Mawass, N. Novakovic-Marinkovic, F. Kronast, M. Weigand, J. Gräfe, S. Wintz, S. Finizio, J. Raabe, L. Aballe, M. Foerster, M. Belmeguenai, L. D. Buda-Prejbeanu, J. Pelloux-Prayer, J. M. Shaw, H. T. Nembach, L. Ranno, G. Gaudin, and O. Boulle, "Skyrmions in synthetic antiferromagnets and their nucleation via electrical current and ultra-fast laser illumination," *Nature Commun.*, vol. 13, no. 1, p. 4807, Aug. 2022, doi: 10.1038/s41467-022-32525-4.

[24] A. Dash, B. Ojha, S. Mohanty, A. K. Moharana, and S. Bedanta, "Device geometry dependent deterministic skyrmion generation from a skyrmionium," *Nanotechnology*, vol. 34, no. 18, Feb. 2023, Art. no. 185001, doi: 10.1088/1361-6528/acb714.

[25] W. L. Gan, S. Krishnia, and W. S. Lew, "Efficient in-line skyrmion injection method for synthetic antiferromagnetic systems," *New J. Phys.*, vol. 20, no. 1, Jan. 2018, Art. no. 013029, doi: 10.1088/1367-2630/aaa113.

[26] S. Qiu, J. Liu, Y. Chen, X. Qi, and L. Fang, "Writing skyrmion at a specific position in synthetic antiferromagnetic racetrack by voltage," *J. Magn. Magn. Mater.*, vol. 554, Jul. 2022, Art. no. 169144, doi: 10.1016/j.jmmm.2022.169144.

[27] Y. Liu, T. Yu, Z. Zhu, H. Zhong, K. M. Khamis, and K. Zhu, "High thermal stability in W/MgO/CoFeB/W/CoFeB/W stacks via ultrathin w insertion with perpendicular magnetic anisotropy," *J. Magn. Magn. Mater.*, vol. 410, pp. 123–127, Jul. 2016, doi: 10.1016/j.jmmm.2016.02.099.

[28] D. Bhattacharya, M. M. Al-Rashid, and J. Atulasimha, "Voltage controlled core reversal of fixed magnetic skyrmions without a magnetic field," *Sci. Rep.*, vol. 6, no. 1, Aug. 2016, Art. no. 31272, doi: 10.1038/srep31272.

[29] B. Rana and Y. Otani, "Towards magnonic devices based on voltage-controlled magnetic anisotropy," *Commun. Phys.*, vol. 2, no. 1, p. 90, Aug. 2019, doi: 10.1038/s42005-019-0189-6.

[30] T. J. Peterson, A. Hurben, W. Jiang, D. Zhang, B. Zink, Y.-C. Chen, Y. Fan, T. Low, and J.-P. Wang, "Enhancement of voltage controlled magnetic anisotropy (VCMA) through electron depletion," *J. Appl. Phys.*, vol. 131, no. 15, Apr. 2022, Art. no. 153904, doi: 10.1063/5.0086768.

[31] D. Bhattacharya, S. A. Razavi, H. Wu, B. Dai, K. L. Wang, and J. Atulasimha, "Creation and annihilation of non-volatile fixed magnetic skyrmions using voltage control of magnetic anisotropy," *Nature Electron.*, vol. 3, no. 9, pp. 539–545, Jun. 2020, doi: 10.1038/s41928-020-0432-x.

[32] X. Li, K. Fitzell, D. Wu, C. T. Karaba, A. Buditama, G. Yu, K. L. Wong, N. Altieri, C. Grezes, N. Kioussis, S. Tolbert, Z. Zhang, J. P. Chang, P. K. Amiri, and K. L. Wang, "Enhancement of voltage-controlled magnetic anisotropy through precise control of mg insertion thickness at CoFeB|MgO interface," *Appl. Phys. Lett.*, vol. 110, no. 5, Jan. 2017, Art. no. 052401, doi: 10.1063/1.4975160.

[33] L. Zhao, M. Yang, J. Gao, T. Yang, Y. Cui, J. Xu, J. Li, Q. Xiang, W. Li, F. Luo, L. Ye, and J. Luo, "Enhancement of magnetic and electric transport performance of perpendicular spin-orbit torque magnetic tunnel junction by Stop-on-MgO etching process," *IEEE Electron Device Lett.*, vol. 44, no. 3, pp. 408–411, Mar. 2023, doi: 10.1109/LED.2023.3236835.

[34] L. Wang, W. Kang, F. Ebrahimi, X. Li, Y. Huang, C. Zhao, K. L. Wang, and W. Zhao, "Voltage-controlled magnetic tunnel junctions for processing-in-memory implementation," *IEEE Electron Device Lett.*, vol. 39, no. 3, pp. 440–443, Mar. 2018, doi: 10.1109/LED.2018.2791510.

Authorized licensed use limited to: Institute of Microelectronics CAS. Downloaded on September 11,2024 at 02:12:51 UTC from IEEE Xplore. Restrictions apply.

Manipulation of synthetic antiferromagnetic skyrmions array in continuous film by VCMA

°Shuaiyu Gong [1,2,3], Meiyin Yang [1,2,3], Yanru Li [1,2,], and Jun Luo [1,2,3]

[1] Key Laboratory of Fabrication Technologies for Integrated Circuits, Chinese Academy of Sciences, Beijing 100029, China
[2] Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China
[3]School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China
E-mail: luojun@ime.ac.cn
yangmeiyin@ime.ac.cn

Abstract

In this paper, we investigate synthetic antiferromagnetic (SAF) skyrmions array in a continuously film (C-SAF array). The SAF skymions array can be written and erased individually by voltage-controlled magnetic anisotropy (VCMA). In addition, the bit density of SAF array in continuous SAF film is at least 20% higher compared with that of isolated device array (I-SAF array). The write, erase time and energy consumption of the C-SAF array also reduces by 75% compared with I-SAF array. Furthermore, C-SAF array can also be used as excellent information transport carriers to realize logic operations and unconventional computing.

1. Introduction

The magnetic skyrmion is a topologically non-trivial chiral spin texture with the advantages of high topological stability, small size, low driving current, and high transport speed [1]. Thus, it holds potential across various applications such as memory device, artificial neurons, reserve pool computing, logic gates. Compared with ferromagnetic skyrmion, synthetic antiferromagnetic (SAF) skyrmions have been proposed, which exhibit good stability and no transport bias due to their antiparallel coupled magnetization structure.

The SAF skyrmions can be manipulated by polarized current, external magnetic field, laser pulses. To save the energy consumption, VCMA is used to manipulate SAF skyrmions. Considering that the high-voltage-breakdown of MgO, it is difficult to effectively control the RKKY coupling SAF with a single MgO/CoFeB interface (Fig. 1a). Thus, a structure of double MgO/CoFeB interface VCMA device [2] (Fig. 1b) were selected in our study. As an important solution to improve device density, the array scheme of these skyrmions VCMA devices is critical but lacks discussion.

In this work, we designed a continuous SAF skyrmions array manipulated by VCMA. We have demonstrated that all the C-SAF array can efficiently achieve the generation and annihilation of skyrmions with higher speed and lower energy consumption than isolated DIS device. We also verified the stability of writing and erasing skyrmions by VCMA. Finally, we show a potential application of our continuous SAF array in logic operation, which can realize the directional transport and capture of skyrmions.

2. Device model

The schematics of C-SAF is shown in Fig. 1c. A 3×4 array of DIS devices with the interval of 85 nm is integrated into a continuous 300 nm × 200 nm SAF. The generation and annihilation of skyrmions are realized by VCMA through the top and bottom MgO/CoFeB interface, which can influence the magnetic anisotropy of the nearby thin film. Fig. 2 illustrates the effect of VCMA on manipulating the generation (+voltage) and annihilation (-voltage) of skyrmions. We simulate this voltage regulation effect by varying the perpendicular magnetic anisotropy energy constant K_u utilizing the object oriented micromagnetic framework (OOMMF). The value of main parameters is shown in the Table 1.

3. Results and Discussion

Fig. 3 illustrates the generation and annihilation process of 3×4 SAF skyrmions, which can be realized by reducing and raising the value of K_u, respectively. Similar with the single skyrmion device, the increase of perpendicular magnetic anisotropy energy (E_{PMA}) in our C-SAF array is less than the decrease of DMI energy (E_{DMI}) and demagnetization energy (E_d) due to the reduced K_u, resulting the generation of SAF skyrmions. The required value of K_u for generating and annihilating skyrmions in C-SAF array and I-SAF device is similar (Fig. 4a). However, the write-erase time of C-SAF array is only one-fifth that of I-SAF array (Fig. 4b), which is attributed to the avoidance of the symmetric boundary in continuous SAF. The symmetric boundary in I-SAF has a repulsive force on the SAF at the center, leading to a negative effect on the reverse of skyrmion and reduced its operation speed. In addition, the single device in C-SAF array also has 20% smaller size and 80% lower energy consumption than I-SAF device (Fig. 4c).

In order to verify the stability of independent write-erase devices in a C-SAF skyrmions array, we performed a series of write-erase operations in a 3×4 array (Fig. 5). The results show that although the operation may cause the weak displacement of the adjacent skyrmions, it does not affect the generation and annihilation of skyrmions through VCMA.

Skyrmions are also stable information carriers. We demonstrate a VCMA-assisted computing application in

Fig. 6, which shows that the array in C-SAF has a good ability to transport and capture skyrmions, and provides a promising potential for the application of skyrmions in unconventional computing.

References

[1] LI S, KANG W, ZHANG X, et al. Materials horizons, 8(3): 854-68, 2021.

[2] GONG S, YANG M, LI Y, et al. IEEE Electron Device Letters, 45(5): 821-4, 2024.

Table 1. Main parameters and values

Symbol	Parameter	Value
M	Magnetization	Bottom: 1×10^6 A/m Top: 1×10^6 A/m
K_u	Perpendicular anisotropy constant	9×10^5 J/m^3
D	Dzyaloshinskii-Moriya interaction parameter	2.3×10^3 J/m^2
α	Damping constant	0.1
J_{ex}	Interlay coupling constant	-2×10^{-4} J/m^2
J_c	Current density	5×10^{11}A/m^2
A	Exchange constant	1.5×10^{-11} J/m

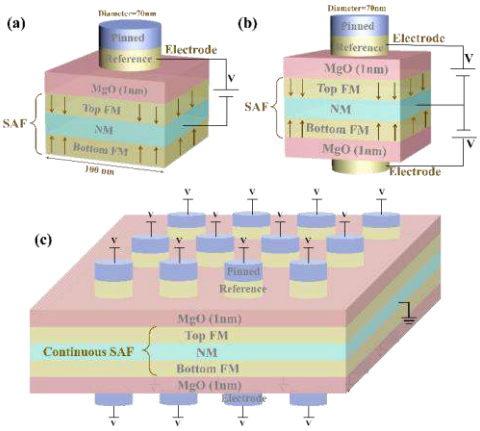


Fig. 1 (a) The single-interface VCMA structure. (b) The double-interface VCMA structure. (c) The structure of VCMA device array for continuous SAF.

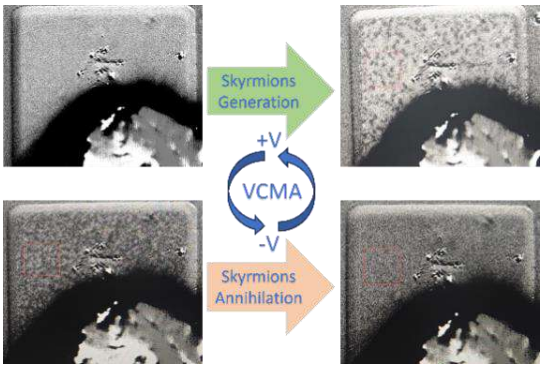


Fig. 2 The use of VCMA to generate or annihilate skyrmions in magnetic layer.

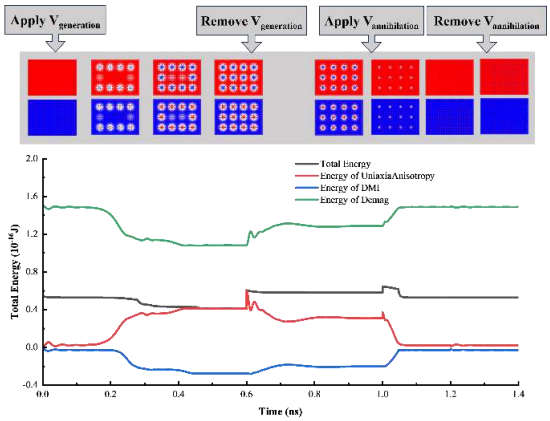


Fig. 3 The energy of C-SAF vs. time in the generation and annihilation process of SAF skyrmions. The inset is the spin texture of top and bottom ferromagnetic layer corresponding to the particular time.

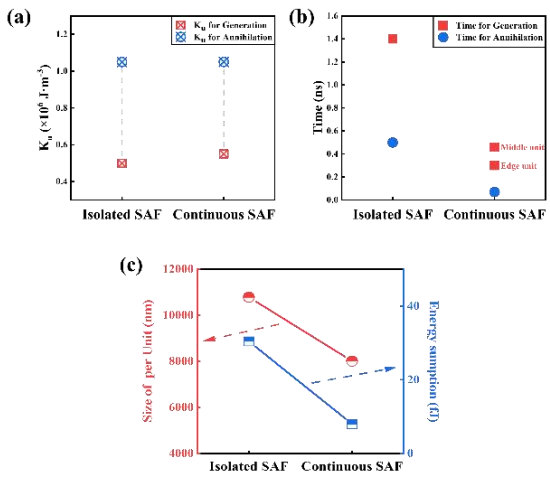


Fig. 4 The performance comparison between continuous and isolated DIS device :(a) required K_u; (b) writing and erasing time; (c) device size and energy consumption.

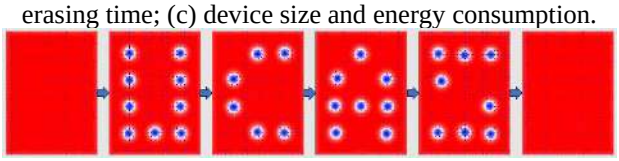


Fig. 5 The writing and erasing effect of DIS array.

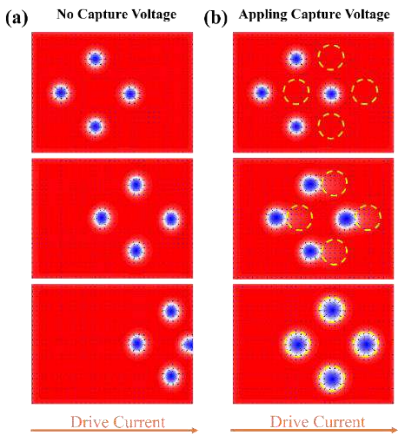


Fig. 6 Demonstration of the effect of current-driven and VCMA trapping of skyrmions.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名： 李维娜

方向： 集成电路

单位： 中国科学院上海微系统与信息技术研究所

题目： 基于极紫外光刻的卤代金属有机框架材料纳米级无光刻胶图案化

First Demonstration of Atomic-Interlayer-Tuning Driven by First Principles Calculations and Atomic Layer Deposition towards High Thermal Stable BEOL IGZO-FETs with SS=62mV/dec, PBTI < 7mV@ 3MV/cm and 353K

Xufan Li[1,2#], Chen Gu[1,2#], Tong Yu[1#], Yue Zhao[1,2], Chuanke Chen[1,2], Chunyu Zhang[1,2], Ziheng Bai[1*], Xinlv Duan[1], Weiwei Li[1], Tianhao Liao[1,2], Shundong Hu[1,2], Nannan You[1], Jiayi Wang[1], Rui Chen[1,2], Zhicheng Wu[1,2], Nianduan Lu[1,2], Jiawei Wang[1], Guanhua Yang[1], Lingfei Wang[1,2*], Di Geng[1,2], Ling Li[1,2*], and Ming Liu[1,2]

[1]State Key Laboratory of Fabrication Technologies for Integrated Circuits, IMECAS, Beijing, 100029, China

[2]University of Chinese Academy of Sciences, Beijing, 100049, China

*Email: baiziheng@ime.ac.cn; wanglingfei@ime.ac.cn; lingli@ime.ac.cn, #equally contributed

Abstract

In this work, first principles calculations, combining AIMD and static calculation are incorporated in the ALD co-designing of IGZO-FETs for the first time, providing in-depth study into the defect kinetics. With emulation of ALD, the optimized IGZO-FETs with outstanding stability up to 500℃ is achieved: **(1)** Sub-0.5nm ALD InO interlayer strengthens the PBTI stability by 51 times @3.5MV/cm, from ΔV_{th} = +174.1mV (5.7V^{-1}) to 3.4mV (294.1V^{-1}). **(2)** PBTI = +6.9 mV; **(3)** 12.1mV @ +3MV/cm and 393K; **(4)** Endurance > 10^{11} cycle with AC $V_{g,bias}$ ~ 2.8V. The first-principles assisted experimental design is practically powerful for oxide FETs with multi-elements and complicated processing methods, and accelerating path-finding of Back-End-of-Line (BEOL) thermal stability enhancement.

Keywords: IGZO, Back-End-of-Line, 2T0C DRAM

Introduction

For current IGZO-technology, gate dielectric (GI) and S/D contact interlayers are crucial for charge transport, device property and reliability. Defect dynamics due to the thermal or electrical stress effects are increasingly complicated, and hindering path-finding of Back-End-of-Line (BEOL) thermal stability enhancement. Thus, with first-principles-calculation and TCAD simulations, bridging the gap between material properties and fabrication conditions, the material-fab co-design is necessarily adopted to mitigate post-annealing degradation and thermal instability (**Fig.1**), contributing to much precise and fine-tuned process, which is significant in BEOL IGZO DRAM cell and array [1,2].

After simulation and experiment verification within 500℃ of different interlayers (None regulation/AlO/InO/GaO), the sub-0.5nm InO interlayer enhances the thermal stability and reliability: sub-0.5nm ALD InO interlayer strengthens the PBTI stability by 51 times, and in 353K, +3MV/cm, PBTI = +6.3mV@1ks, without source/drain (S/D) contact deterioration. This work firstly provides deep insights in defect kinetics and provides first-principles-calculation-experiment co-design for sub-nano interlayer regulation as a powerful tool for BEOL fabrication.

Device Fabrication and First-principles-calculations

Fig. 2a shows the schematic of top-gate atomic-interlayer-tuning (AIT) IGZO FETs. The fabrication flow is in **Fig. 2b**. ITO was deposited as S/D electrode. Contact-interlayer (<0.5nm)/IGZO (5nm)/dielectric-interlayer (<0.5nm) were deposited by ALD. 10nm HfO and 25nm IZO were deposited as GI and top gate. The post-deposition anneal (PDA) (400℃/500℃) was in nitrogen. Fig. 2c shows the cross-sectional TEM and EDS.

Fig.3 shows the modeling workflow of (111/211) bulk ALD-IGZO and the (111) interfaces with or without the InO interlayer. Possible defects in ALD-IGZO are introduced to ideal atomic structures through ab-initio molecular dynamics (AIMD) simulation at 250℃. **Fig.4** shows count of resulting defects of the four atomic structures in **Fig.3** before and after annealing, with/without the InO interlayer. **Fig. 5a** and **Fig. 5b** give the DoS of IGZO without/with migrated oxygen vacancy before/after annealing, and the TCAD results is in **Fig. 5c**.

Reliability Optimization and Experiment Verification

After 400℃ PDA, the FET with InO(<0.5nm) contact-interlayer owns an ideal DIBL (**Fig.6a**), and no obvious Schottky barrier is observed (**Fig.6b**). Transfer curves within 278 to 393K is shown in **Fig.6c** to extract contact barrier (Φ_B) ~80meV of IGZO/InO (<0.5nm)/crystal ITO, which is comparable to ohmic contact [3]. Particularly, the InO interlayer structure shows non-obvious thermal deterioration.

Then, the GI interlayer including "None regulation" (None), AlO, InO and GaO (**Fig.7a**) are studied. The On-current of InO/InO is 1.4μA and SS is 64mV/dec (**Fig.7b, c**). Compared with "None", the sub-0.5nm InO GI interlayer has a lower NBTI ΔV_{th} (**Fig.8a**), decreasing from -9.5mV to -4.6mV@1ks and AlO is -4.3mV. Both the InO and AlO interlayers enhance the NBTI stability 2 times. Notably, the InO GI interlayer is beneficial for PBTI (**Fig.8b**): ΔV_{th}@1ks for +3.5MV/cm is only +3.4mV, much lower than +174.1mV of "None" (increasing stability by 50 times), which is most stable within 10-year prediction (~3×10^8s) in **Fig.8c.** In **Fig.9a** of "InO/InO", the PBTI ΔV_{th} @1ks and 353K for +3MV/cm is 6.9mV, increasing by 20 times compared with the "None", and in 393K (**Fig.9b**), PBTI $|\Delta V_{th}|$ of V_{ov} = 3.5V is 12.1mV. The transfer curves (**Fig.9c**) and PBT/NBTI (**Fig.9d**) after 500℃ PDA are depicted, the SS of "InO/InO" is 62 mV/dec, and the NBTI $|\Delta V_{th}|$ of V_{ov} = -2.5V is 8.4mV. The "InO/InO" and "AlO/InO" are potential candidates for higher thermal budget.

PBTI with 400/500℃ PDA is shown in **Fig.10a** and **Fig.10b**. The InO/InO structure shows better sub-threshold stability than the 'none' one. Statistics on non-negligible defects for interfaces (DoS above 10^{17} cm^{-3}·eV^{-1}) are performed in **Fig.10c**. Combined with **Fig.5**, the DoS with InO interlayer decreases after annealing and alters to deep energy, which is responsible for the high thermal stability in NBTI/PBTI. Finally, the endurance of 2T0C DRAM cell with 400℃ PDA "InO/InO" is carried out in **Fig.11**. Under AC amplitude of ~ +2.8 to -2.8V with 10^{11} cycle, the I_{on}/I_{off} ratio remains > 10^8. For InO/InO structure based 2T0C DRAM, the retention time > 600s are achieved.

Conclusion

AIMD is a partical tool to optimize AIT IGZO FETs for high thermal stability and BTI reliability with 400/500℃ PDA. Therotical prediction is consistent with experiments, sub-0.5nm InO ALD interlayer improves PBTI by 51 times in BEOL thermal budget @3.5MV/cm, from 5.74V^{-1} to 294.1V^{-1}. "InO/InO" AIT-IGZO FETs show great potential in reliable BEOL DRAM cell.

Acknowledgements: This work was supported by National Key Research and Development Program of China (2021YFB3600704) and NSFC (Grant No. 62274178). Postdoctoral Fellowship Program of CPSF (No. GZC20232922)

Reference

[1] Y. Zhao, *et al.*, *IEDM*, 2024. [2] X. Duan, *et al.*, *IEDM*, 2021, pp. 10.15.11-10.15.14. [3] Z. Wu, *et al.*, *IEEE EDL*, vol. 45, no. 3, pp. 408-411, 2024. [4] E. Sarkar, *et al.*, *IEDM*, 2024. [5] K. A. Aabrar, *et al.*, *Symp. on VLSI*, 2024, pp. 1-2. [6] W. Zhao, *et al.*, *IEDM*, 2024. [7] Q. Lin, *et al.*, *IEDM*, 2024. [8] W. Zhao, *et al.*, *IEDM*, 2024. [9] K. Huang, *et al.*, *Symp. on VLSI*, 2022, pp. 296-297. [10] W. Li, *et al.*, *IEDM*, 2024. [11] Z. Zheng, *et al.*, *IEDM*, 2024. [12] Y. Su, *et al.*, *IEDM*, 2024. [13] C. Niu, *et al.*, *IEDM*, 2024.

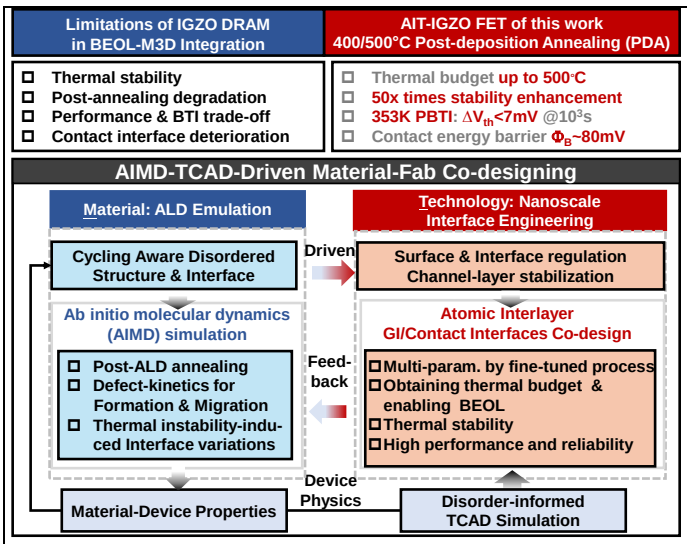


Fig. 1. In terms of limitations of IGZO DRAMs, the AIMD-TCAD method is firstly proposed for defect kinetic description, forming the basis of atomic interface tuning (AIT) in IGZO FETs, exhibiting superior performance (~ 62 mV/dec) and reliability (<7mV PBTI at 353K; immunity to 400/500°C).

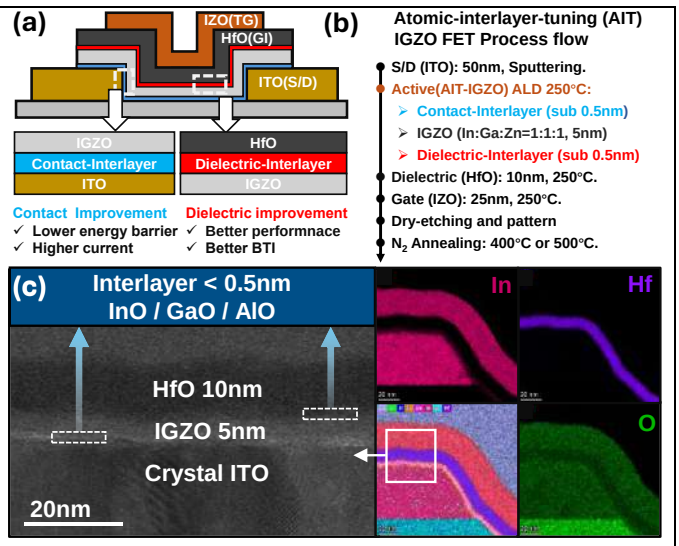


Fig. 2. (a) Structure of top gate AIT-IGZO FET with dielectric and contact interlayers through ALD and (b) relative fabrication flow. (c) TEM and EDS mapping images of AIT-IGZO FET, and the S/D contact is crystal ITO.

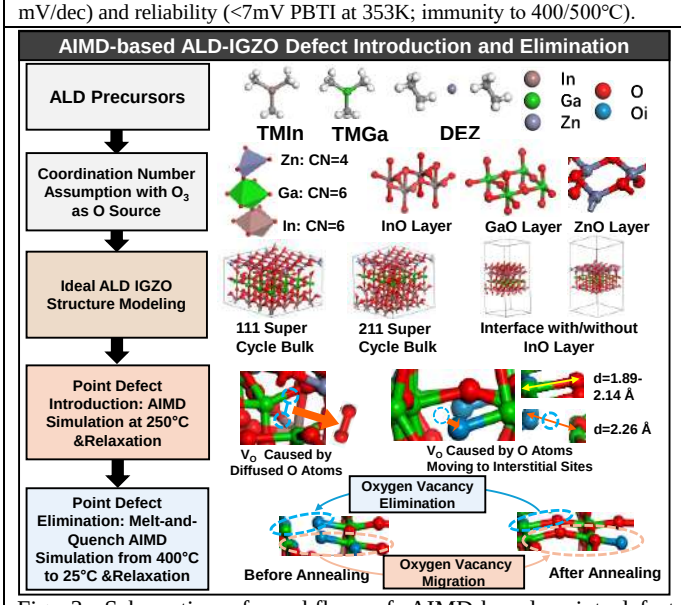


Fig. 3. Schematics of workflow of AIMD-based point defect introduction and elimination for bulk ALD-IGZO and its interface.

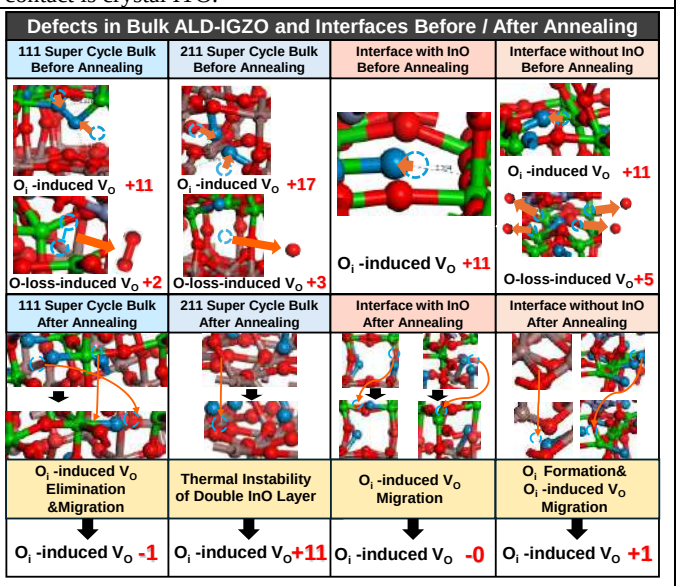


Fig. 4. Counts of Oi- and O-loss-induced Vo defects in bulk ALD-IGZO and interfaces with/without InO before and after annealing.

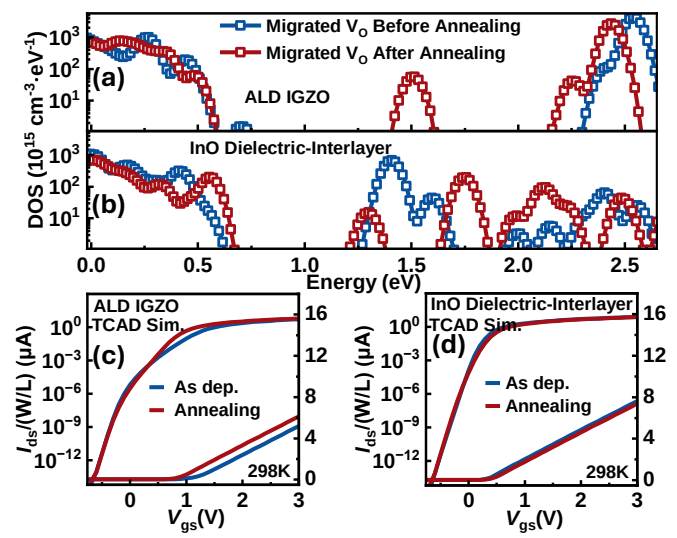


Fig. 5. DoS of (a) IGZO and (b) IGZO with sub-0.5nm InO dielectric interlayer after annealing, and (c,d) corresponding TCAD simulation. The I-V of the FET with InO dielectric-interlayer is more stable after annealing, because of the low DoS.

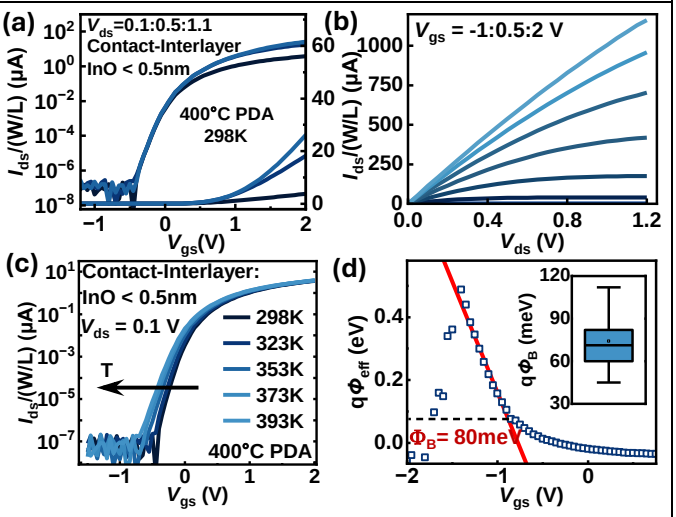


Fig. 6. (a) Transfer and (b) output curves of the AIT-IGZO FET after 400°C PDA with sub-0.5nm InO contact-interlayer. From (c) transfer curves of 298 to 393K, the contact energy barrier (Φ_B) is ~ 80meV, almost the same as as-dep. metal contact as previous study, overcoming metal oxidation in BEOL.

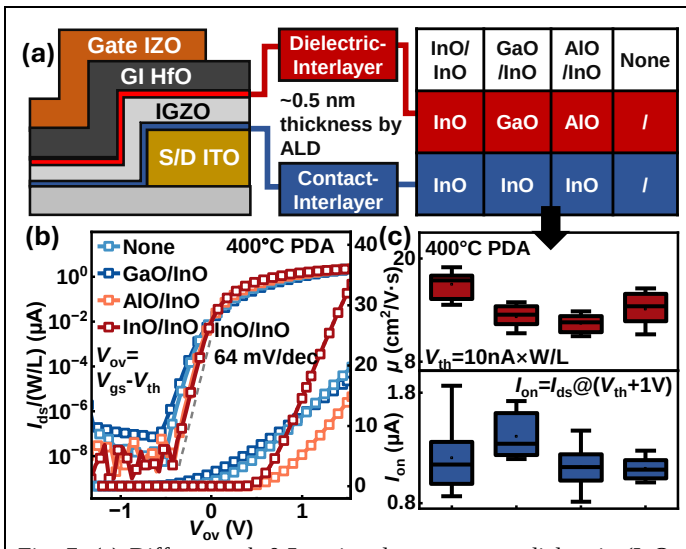


Fig. 7. (a) Different sub-0.5nm interlayers at gate dielectric (InO, GaO, AlO and None) and contact (None and InO). The (b) transfer curves and (c) I_{on} & mobility statistical plots of 400°C PDA FETs, and InO/InO has the best SS of 64mV/dec.

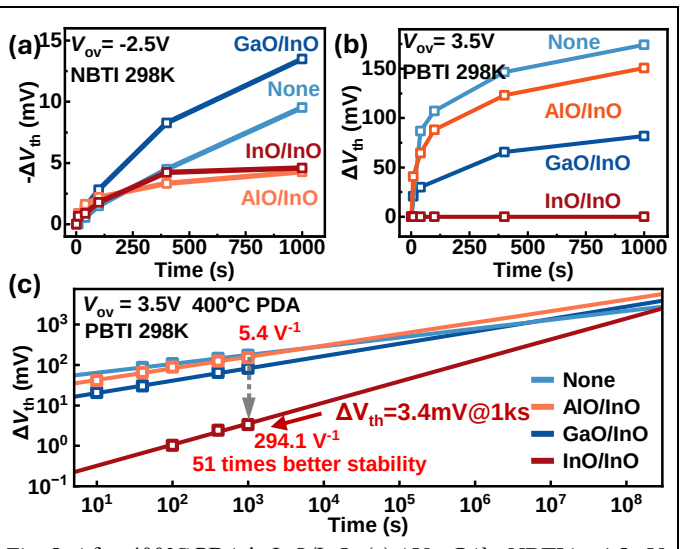


Fig. 8. After 400°C PDA in InO/InO: (a) ΔV_{th} @1ks NBTI is -4.6mV with V_{ov}=-2.5V and (b) ΔV_{th} @1ks = 3.4mV in 3.5MV/cm PBTI, and the PBTI stability of InO/InO is 50 times better than the "None", which is predicted to be the best within 10 years (@3×10^{8}s).

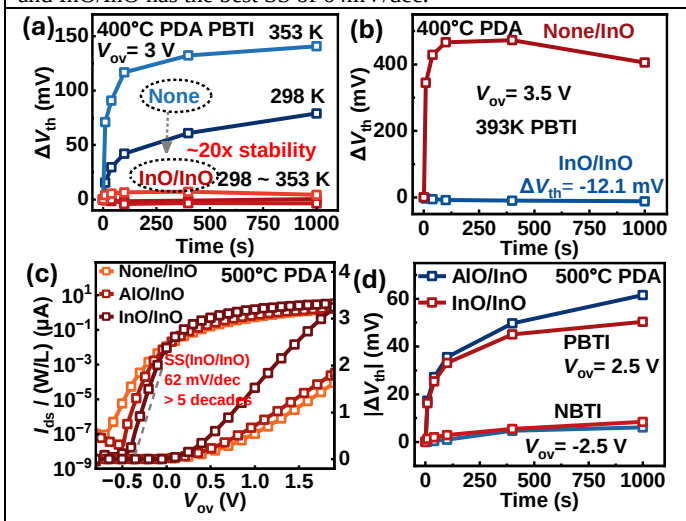


Fig. 9. After 400°C PDA, (a) PBTI stability of InO/InO is much better than "None" interlayer from 298 to 353 K: PBTI @1ks is 6.9mV, (b) and it is 12.1mV in 393K. (c) Both InO/InO and AlO/InO have 500°C thermal budget in PDA. The SS of InO/InO is ~ 62mV/dec, and (d) NBTI @1ks is < 10mV.

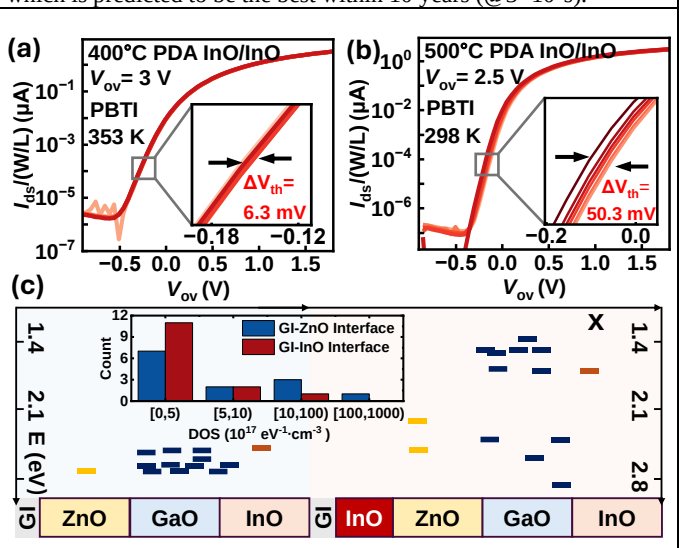


Fig. 10. Both in (a) 400°C and (b) 500°C, PBTI is influenced by trapping-de-trapping. (c) Axial distribution of major defect energy levels for interfaces with/without the InO layer with statistics on number of defects for different DoS ranges.

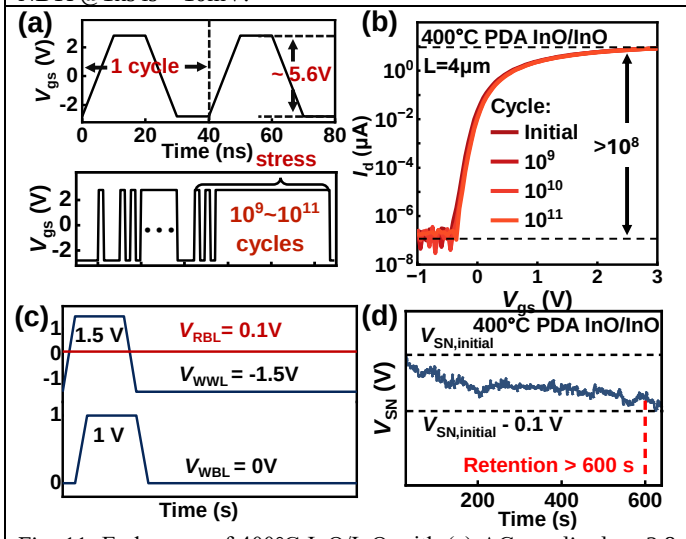
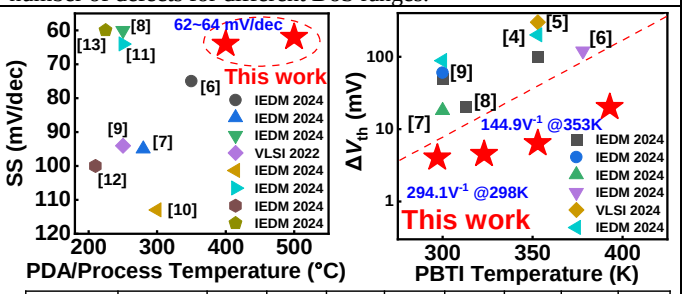


Fig. 11. Endurance of 400°C InO/InO with (a) AC amplitude ~ 2.8 V, and the I_{on}/I_{off} ratio is > 10^8 after 10^{11} cycle. (c) Time sequence operation of 400°C PDA InO/InO 2T0C, and (d) the retention is > 600s at 298K.



	This Work	2024 [4]	2024 [6]	2024 [7]	2024 [8]	2022 [9]	2024 [10]	2024 [11]	2024 [12]
PDA Temp	400°C /500°C	N.A.	350°C	280	300°C	300°C	300°C	N.A	N.A
SS (mv/dec)	62 (500°C) 64 (400°C)	~100	75	95	60	>90	113	>62	>100
PBTI ΔVth (|mV|) and test temp.	3.4 (298K) 3.9 (323K) 6.9 (353K) 12.1 (393K)	88 (298K) 200 (353K)	21 (298K) 100 (378K)	18 (298K)	60 (300K)	20 (313K) 120 (353K)	N.A	N.A	N.A
Contact	ITO	Pd	N.A	Ni	ITO/Ni	TiN	TiN	Pd	N.A
Channel	AIT-IGZO	In_2O_3	IGZO	IWO	IGZO	IGZO	IGZO	IGZO	IGZO
Structure	TG	GAA	GAA	BG	BG	CAA	P-CAA	BG	BG

Fig. 12. Comparison with state-of-the-art oxide FETs.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：李旭帆
方向：集成电路
单位：中国科学院微电子研究所

题目：原子尺度调控的高热稳定性铟镓锌氧化物半导体器件的仿真及工艺实现

ADVANCED SCIENCE Open Access
www.advancedscience.com

Nanoscale Resist-Free Patterning of Halogenated Zeolitic Imidazolate Frameworks by Extreme UV Lithography

Weina Li, Tianlei Ma, Pengyi Tang, Yunhong Luo, Hui Zhang, Jun Zhao, Rob Ameloot, and Min Tu**

Advancements in patterning techniques for metal–organic frameworks (MOFs) are crucial for their integration into microelectronics. However, achieving precise nanoscale control of MOF structures remains challenging. In this work, a resist-free method for patterning MOFs is demonstrated using extreme ultraviolet (EUV) lithography with a resolution of 40 nm. The role of halogen atoms in the linker and the effect of humidity are analyzed through in situ and near-ambient pressure synchrotron X-ray photoelectron spectroscopy. In addition to facilitating the integration of MOFs, the results offer valuable insights for developing the highly sought-after positive-tone EUV photoresists.

1. Introduction

Lithography patterning is a crucial step in integrating new functional materials into electronic applications.[1–4] Metal–organic frameworks (MOFs) have the potential to introduce novel functionalities in microelectronics and optoelectronics, such as low-k dielectrics, chemical sensors, and field-effect transistors.[5–9] MOF patterning methods can be divided into bottom-up and top-down approaches. The bottom-up approach involves the direct growth of MOFs at well-designed locations on a substrate, for instance, via microcontact printing of patterned self-assembled monolayers followed by area-selective MOF growth.[10–15] However, achieving high-resolution (sub-µm) patterns and sharp edges with this technique is challenging, and non-selective growth occurs for thicker MOF films. In contrast, top-down lithographic patterning involves transferring a pattern onto a continuous film, making it more compatible with established industrial fabrication processes.[12–16] Typically, this process uses a photosensitive resist layer as an etch mask, which must be deposited, patterned, and removed without altering the properties of the underlying material. Given this, direct, resist-free lithography of functional materials is highly desirable, as it minimizes process steps and reduces potential contamination sources—especially for MOFs, which are susceptible to pore clogging.[17–26]

W. Li, T. Ma, P. Tang, M. Tu
State Key Laboratory of Transducer Technology
Shanghai Institute of Microsystem and Information Technology
Chinese Academy of Sciences
Shanghai 200050, China
E-mail: min.tu@mail.sim.ac.cn

W. Li, T. Ma, P. Tang, M. Tu
Center of Materials Science and Optoelectronics Engineering
University of Chinese Academy of Sciences
Beijing 100049, China

W. Li, T. Ma, P. Tang, M. Tu
2020 X-Lab
Shanghai Institute of Microsystem and Information Technology
Chinese Academy of Sciences
Shanghai 200050, China

W. Li, P. Tang, M. Tu
School of Graduate Study
University of Chinese Academy of Sciences
Beijing 100049, China

Y. Luo
ShanghaiTech University
School of physical science and technology
Shanghai 201210, China

H. Zhang, J. Zhao
Shanghai Synchrotron Radiation Facility
Shanghai Advanced Research Institute
Chinese Academy of Sciences
Shanghai 201204, China

H. Zhang
National Key Laboratory of Materials for Integrated Circuits
Shanghai Institute of Microsystem and Information Technology
Chinese Academy of Sciences
Shanghai 200050, China

R. Ameloot
Centre for Membrane Separations
Adsorption
Catalysis and Spectroscopy
KU Leuven
Leuven 3001, Belgium
E-mail: rob.ameloot@kuleuven.be

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202415804

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.202415804

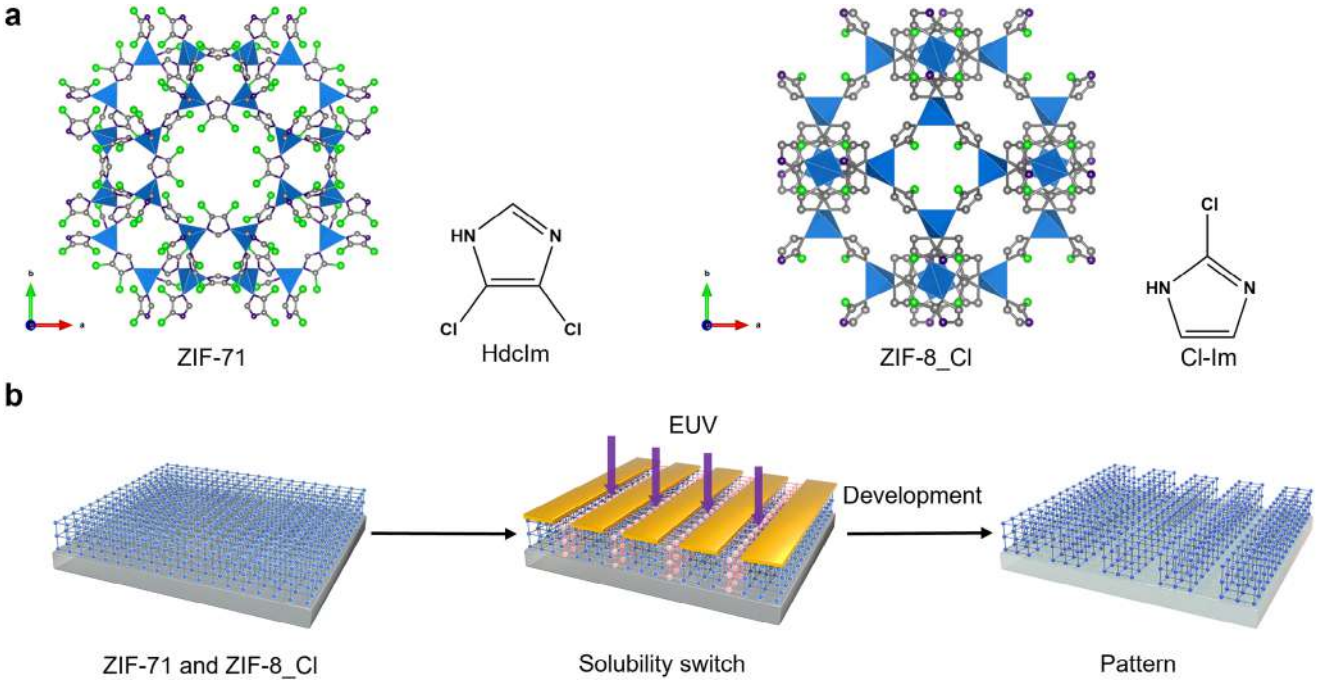


Figure 1. Direct EUV lithographic patterning of halogenated ZIF films. a) Representation of the crystal structures of ZIF-71 and ZIF-8_Cl and the corresponding linkers 4,5-dichloroimidazole (HdcIm) and 2-chloroimidazole (Cl-Im). b) Schematic illustration of the resist-free EUV lithography of ZIF films, which behave as positive-tone resists.

Recent advancements have demonstrated that certain MOFs can be directly patterned with high resolution using electron-beam lithography (EBL).[27–30] Tsapatsis et al. achieved sub-100 nm, resist-free EBL patterning of zeolitic imidazolate framework (ZIF) films, where the exposed ZIFs functioned as either a positive- or negative-tone resist, depending on the exposure dose and solvent used during developing.[27] We have achieved sub-50 nm resolution in direct EBL patterning of halogenated ZIFs, where e-beam irradiation induces structural and chemical changes driven by halogen radicals.[29] While these resist-free approaches enable nanoscale patterning of MOFs, EBL is a maskless, low-throughput method, making it suitable only for niche applications, such as mask fabrication and prototyping.

Extreme UV (EUV) lithography is a mask-based, parallel process compared with EBL, making it well-suited for mass production in high-volume semiconductor fabrication. In EUV lithography, high-energy EUV photons with a wavelength of 13.5 nm and an energy of 92 eV are absorbed by the resist material. This absorption causes electrons to be ejected with energies between 75 and 82 eV, leaving positively charged radicals in the resist layer.[31–34] These EUV-generated electrons and holes trigger chemical reactions, such as bond cleavage and cross-linking, which alter the solubility of the photoresist.[35–38]

The motivation to explore EUV patterning of MOFs is twofold. On the one hand, EUV lithography offers a mask-based and scalable method for integrating the unique properties of MOFs into electronic devices. Since the photoelectrons generated during EUV exposure trigger the solubility change in the resist, MOF materials that can be patterned with EBL may also be suitable for EUV lithography. On the other hand, there is an urgent need to develop novel photoresists tailored for EUV lithography. Conventional organic photoresists, primarily composed of light elements, have limited interaction with high-energy EUV photons, resulting in low exposure sensitivity. Consequently, metal-containing resists are being explored to achieve a higher EUV absorption cross-section.[31,34,39–41] However, these resists are almost exclusively negative-tone, meaning that the pattern is created by retaining the exposed areas after development. This reliance on negative-tone photoresists poses challenges for processes that require positive-tone materials, such as contact-hole patterning in memory devices. Given that some MOFs have demonstrated positive-tone patterning in EBL, investigating these materials as positive EUV photoresists presents an exciting opportunity.

In this study, we report the direct, resist-free patterning of halogenated ZIFs using EUV lithography, achieving a resolution down to 40 nm (**Figure 1**). Additionally, we examined the chemical changes in the halogenated ZIF thin film under EUV irradiation using in situ synchrotron X-ray photoelectron spectroscopy (XPS). We also explored the impact of humidity on the EUV sensitivity of halogenated ZIFs through synchrotron near-ambient XPS.

2. Results and Discussion

Lithographic patterning requires smooth, thin films. To meet these requirements, we use chemical vapor deposition (CVD) to grow halogenated ZIF thin films.[29,42,43] ZIF-71 and ZIF-8_Cl are synthesized by coordinating Zn^{2+} with 4,5-dichloroimidazole (HdcIm) and 2-chloroimidazole (Cl-Im), respectively, forming RHO and SOD topologies.[44,45] Thin films of these MOFs are produced by first depositing ZnO films via atomic layer depo-

 © 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202415804 by Shanghai Institute Of, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

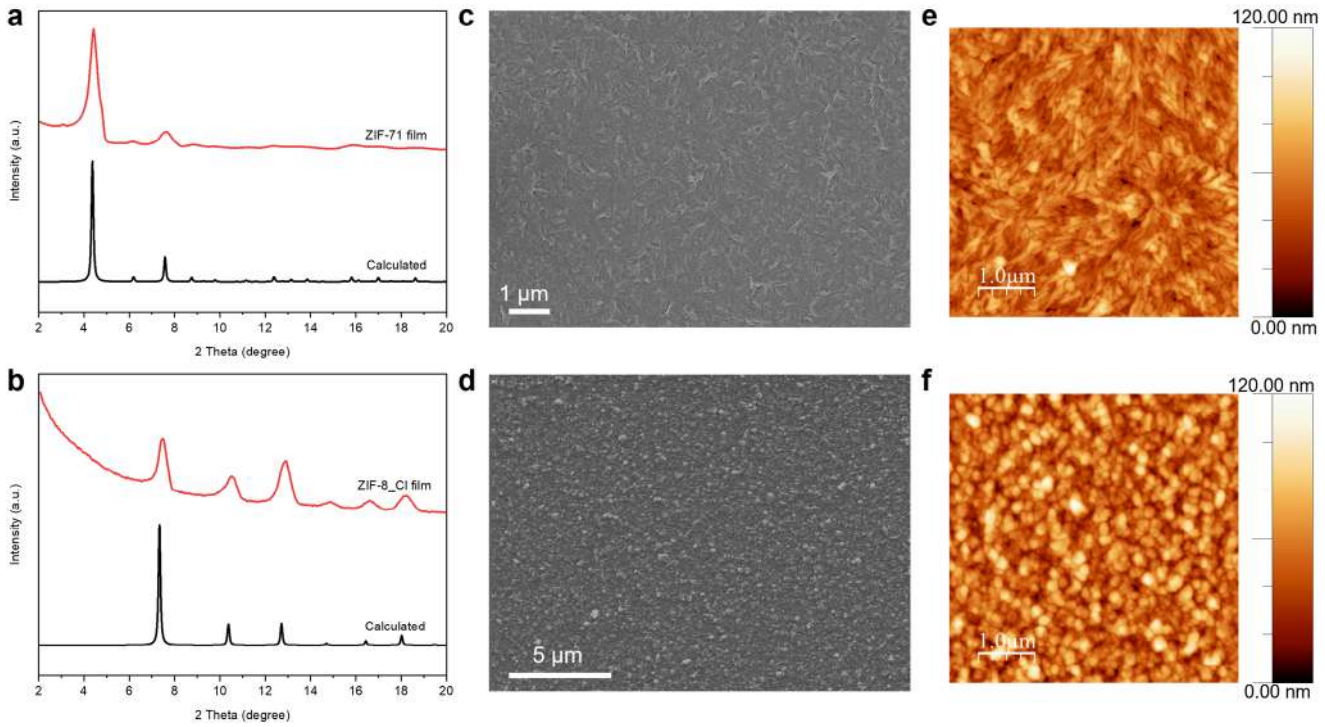


Figure 2. Characterization of halogenated ZIF films. a, b) GIXRD of ZIF-71 and ZIF-8_Cl films. c, d) SEM images of ZIF-71 and ZIF-8_Cl films. e, f) AFM images of ZIF-71 and ZIF-8_Cl films.

sition (ALD) and then converting the precursor films into an atmosphere containing the corresponding linkers and template vapors.[42] The grazing incidence X-ray diffraction (GIXRD) peaks of the resulting ZIF-71 and ZIF-8_Cl films align with the calculated PXRD patterns, confirming the formation of the expected phases (**Figure 2**a,b). Scanning electron microscopy (SEM) and atomic force microscopy (AFM) images show that uniform and homogeneous ZIF-71 and ZIF-8_Cl thin films are successfully deposited (Figure 2c–f). The root mean square roughness values for ZIF-71 and ZIF-8_Cl thin films are ≈12 and 16.7 nm, respectively. ZIF thin films of different thicknesses can be obtained by converting ZnO films with varying initial thicknesses (Figures S1 and S2, Supporting Information). The theoretical expansion factors for ZIF-71 and ZIF-8_Cl films starting from ZnO precursors are 20 and 16.5, respectively.[42] ZIF-71 films with thicknesses of 100 and 60 nm were derived from 5 and 3 nm ZnO layers, respectively, while a ZIF-8_Cl film with a thickness of 80 nm was derived from a 5 nm ZnO layer.

The feasibility of directly patterning halogenated ZIF films using EUV interference lithography was investigated at the Swiss Light Source and the Shanghai Synchrotron Radiation Facility.[15–46] The films were exposed to EUV radiation at varying doses and then developed in dimethylsulfoxide (DMSO). Both ZIF-71 and ZIF-8_Cl films were found to be amenable to direct patterning through EUV lithography (**Figure 3**). The contrast curve for the ZIF-71 film shows a clear step (Figure S3, Supporting Information), with over 90% of the original film being removed at exposure doses above 400 mJ cm^{-2}. For a 50 nm half-pitch (HP) pattern, the optimal EUV dose for a 100 nm thick ZIF-71 film was found to range from 440 to 530 mJ cm^{-2} (Figure S4, Supporting Information). Exposure doses below 440 mJ cm^{-2} led to incomplete development of the exposed regions, while doses above 530 mJ cm^{-2} caused the lines to narrow, increasing the risk of pattern collapse. The optimal exposure dose and resulting pattern quality are strongly dependent on the ZIF film thickness.[47,48] Thinner films require lower exposure doses due to the reduced EUV radiation attenuation,[49] while thicker films are more prone to pattern collapse at high resolution due to the increased aspect ratio of the formed patterns.[47,48] Using an EUV dose of 480 mJ cm^{-2}, a 50 nm pattern with a line-edge roughness (LER) of 5.5 nm was achieved in a 100 nm thick ZIF-71 film (Figure S4, Supporting Information). However, higher resolution patterns (with HPs of 40 and 30 nm) collapsed due to the aspect ratio exceeding 2 (Figures S5 and S6, Supporting Information). To improve pattern quality, a thinner 60 nm ZIF-71 film was deposited (Figure 3; Figures S7–S11, Supporting Information). The thinner film enabled a 50 nm line with a 1.0 nm LER at an exposure dose of 440 mJ cm^{-2} (Figure 3a,b), and a 40 nm line with an 8.5 nm LER at 425 mJ cm^{-2} (Figure 3c,d). Compared to the 100 nm ZIF-71 film, the 60 nm film produced higher-quality patterns with lower LER for 50 nm and smaller features.

We performed in situ synchrotron XPS experiments to investigate the chemical changes in halogenated ZIFs induced by EUV irradiation (**Figure 4**a). After EUV exposure at 92 eV in a vacuum, XPS spectra were collected from the exposed area using 1250 eV photons. This approach ensures that XPS analysis is conducted directly on the EUV-exposed region and prevents air exposure during sample transfer, ensuring that any chemical changes in the films are solely attributed to EUV radiation. The XPS spectra of the Zn 2p region showed no significant changes

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202415804 by Shanghai Institute Of, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

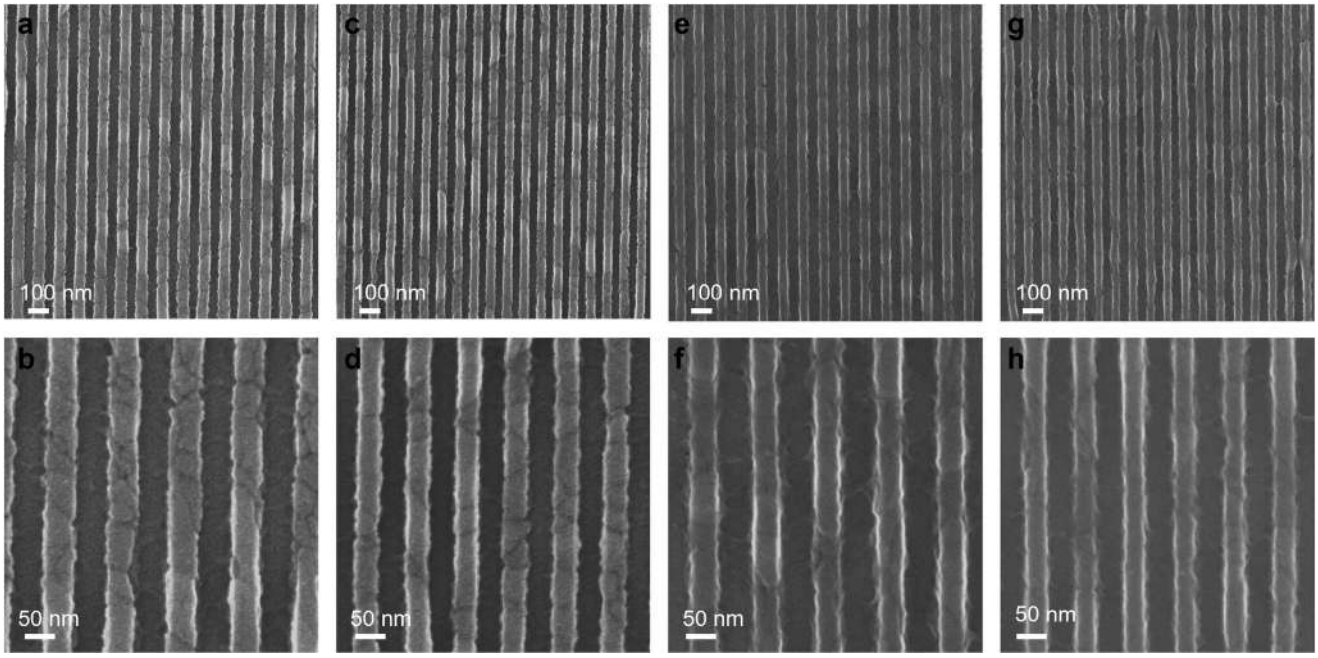


Figure 3. SEM images of ZIF patterns fabricated by EUV lithography. a, b) 50 nm HP pattern from 60 nm ZIF-71 film under a dose of 440 mJ cm^{-2}. c, d) 40 nm HP pattern from 60 nm ZIF-71 film under a dose of 425 mJ cm^{-2}. e, f) 50 nm HP pattern from 80 nm ZIF-8_Cl film under a dose of 400 mJ cm^{-2}. g, h) 40 nm HP pattern from 80 nm ZIF-8_Cl film under a dose of 265 mJ cm^{-2}.

with varying EUV exposure, indicating that Zn^{2+} remained unchanged (Figure S16, Supporting Information).[50] The fitted Cl 2p spectra revealed two components, corresponding to C—Cl (≈201.2 eV) and Zn—Cl (≈198.9 eV) bonds (Figure 4b).[51] The N 1s spectra were fitted with three components, assigned to Zn—N (≈401.6 eV), N—H (≈400.3 eV), and C—N (≈399.7 eV) bonds (Figure 4b).[29,52,53] In the Cl 2p region, EUV exposure led to a decrease in the C—Cl bond signal and a corresponding increase in the Zn—Cl bond signal. The presence of Zn—Cl in the pristine ZIF-71 film was minimal, likely due to beam damage from X-ray exposure during the measurement. In the N 1s XPS spectra, the Zn—N and C—N signals decreased, while the N—H signal increased (Figure 4b,c).

EUV irradiation can cause both primary and secondary damage to the resist material.[32,34,36,54–58] Primary damage occurs when EUV light interacts with the material, leading to bond breaking through photoelectric effects. Secondary damage results from reactions triggered by radiolytic products, such as free radicals generated by energetic electrons. The reduction of the C—Cl bond signal suggests the chlorine radicals are generated by the homolysis of the C—Cl bonds upon the EUV irradiation (Figure 4b,c). These chlorine radicals likely cleave the Zn—N coordination bonds and C—N bonds in the imidazole ring (Figures S17 and S18, Supporting Information).[59–62] Further reactions between the resulting fragments may make the exposed sample soluble in DMSO. While identifying the exact products formed upon EUV irradiation of a thin resist film remains challenging, it is clear the halogen atoms in the linker play a crucial role in the solubility change (Figure S19, Supporting Information).

EUV lithography is typically conducted in a high vacuum to prevent the EUV photons from interacting with the atmosphere, protect sophisticated optical components, and achieve the highest resolution and pattern quality. However, the development process of EUV resists generally occurs under standard cleanroom conditions. As a result, after exposure to EUV light, the latent image in the resist layer is often exposed to humidity. For other types of EUV resists, it has been reported that this exposure to a cleanroom atmosphere during the delay between EUV exposure and pattern development can impact the critical dimensions that can be achieved.[63] We employed synchrotron near-ambient pressure XPS to further explore the effects of humidity on the chemical changes induced by EUV irradiation. Two scenarios were investigated: 1) the ZIF-71film was exposed to EUV radiation in a controlled-humidity atmosphere (**Figure 5**), and 2) the film was exposed EUV under vacuum, followed by exposure to a humid atmosphere. In the first scenario, we found that exposure to EUV in the presence of humidity led to a higher ratio of Zn—Cl to C—Cl bonds compared to exposure in the vacuum (Figure 5). A similar trend was observed in the ratio of the N—H bond to C—N bonds (Figure S20, Supporting Information). This suggests that the presence of water vapor during exposure may immediately trigger hydrolysis of EUV-induced products, resulting in a higher Zn—Cl signal. The second scenario showed similar effects (Figure S21, Supporting Information), even when exposure to humidity was delayed, indicating that humidity influences both primary and secondary types of EUV-induced damage. This finding underscores the importance of controlling the development environment to ensure consistency and reproducibility in EUV lithography.[63]

3. Conclusion

In summary, we demonstrated that halogenated ZIFs (ZIF-71 and ZIF-8_Cl) can be directly patterned by EUV lithography,

 © 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202415804 by Shanghai Institute Of, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

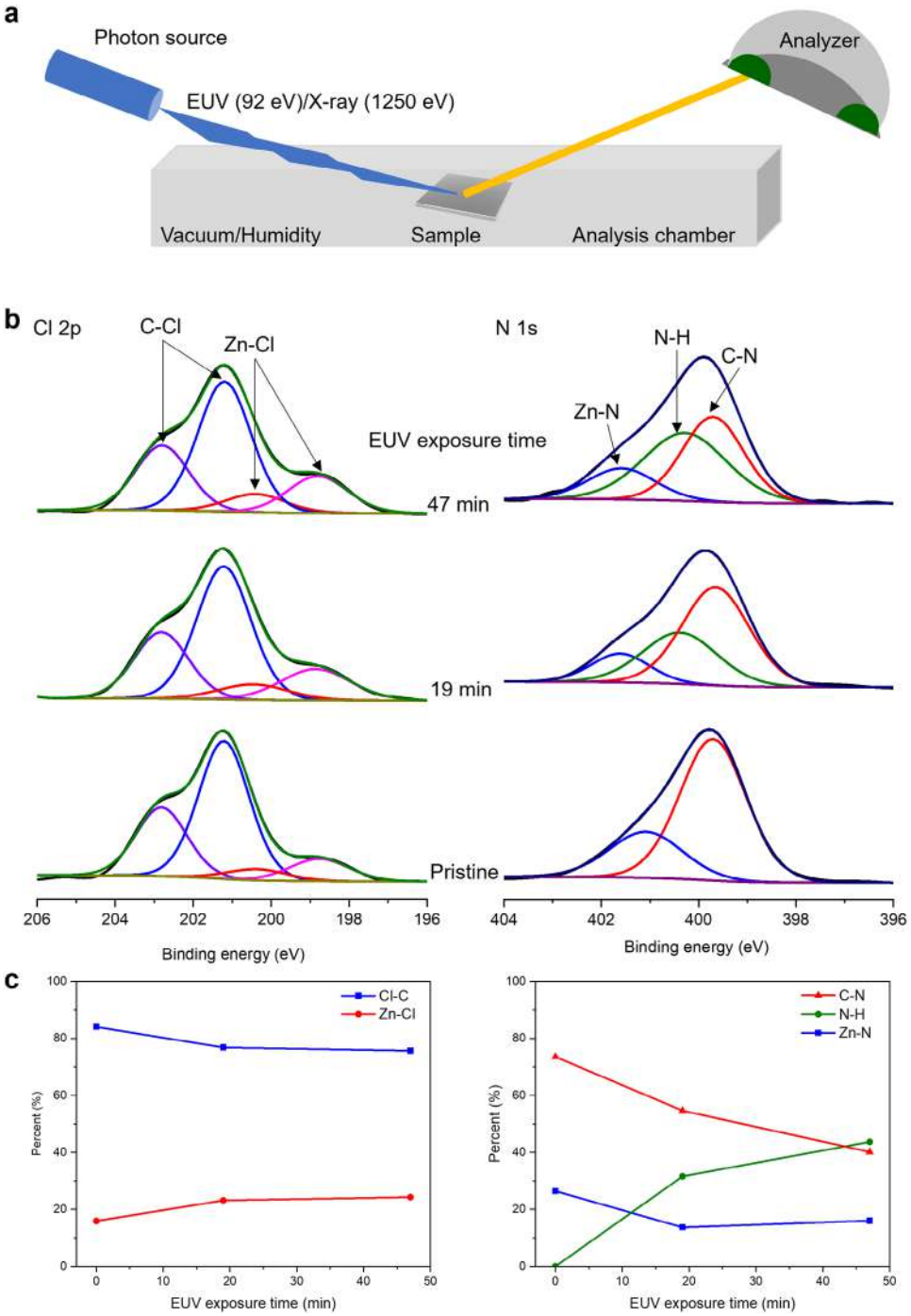


Figure 4. In situ XPS studies of EUV irradiated ZIF-71 film. a) Schematic illustration of the in situ synchrotron XPS measurement. b) XPS spectra of Cl 2p (left) and N 1s (right) as a function of EUV exposure times. c) The proportion of chlorine bonds (C—Cl and Zn—Cl, left) and nitrogen bonds (Zn—N, N—H, and C—N, right) as a function of EUV exposure time.

achieving a half-pitch as small as 40 nm. Further improvements in patterning resolution could be achieved by reducing film thickness and surface roughness. In situ synchrotron XPS studies revealed that EUV exposure leads to the breaking of Zn—N coordination bonds and the ring-opening of the imidazolate linker. Near-ambient pressure XPS analysis indicated that humidity accelerates the dissociation of Zn—N coordination bonds and enhances the reactions between EUV-induced fragments. Our results not only facilitate the integration of MOFs into microelectronics but also provide valuable insights for advancing the development of highly sought-after positive-tone EUV photoresists.

 © 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202415804 by Shanghai Institute Of, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

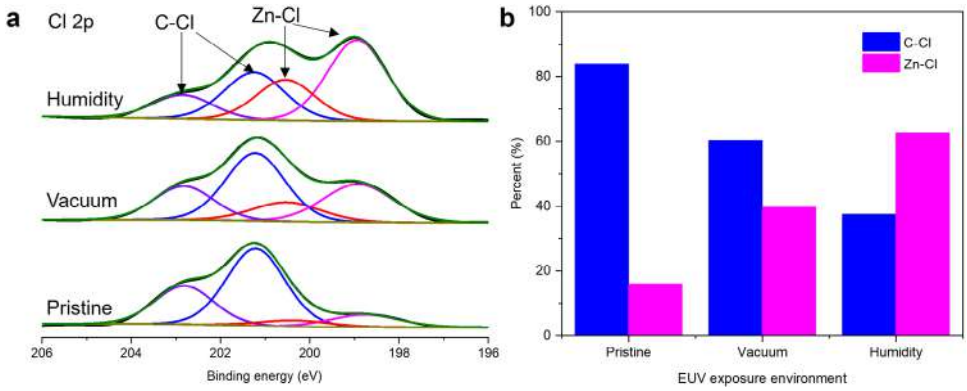


Figure 5. Near-ambient pressure XPS analysis of humidity effects on EUV-irradiated ZIF-71 film. a) XPS spectra of Cl 2p exposed 60 min in different EUV exposure environments (exposure in a vacuum and humidity). b) The proportion of chlorine bonds (C—Cl and Zn—Cl, top) in different EUV exposure environments (exposure in a vacuum and humidity).

4. Experimental Section

Materials: All the reagents were purchased from Sigma-Aldrich and used without further purification. 4,5-dichloroimidazole (HdcIm, 97%), 2-chloroimidazole (Cl-Im, 97%), N, N-dimethylformamide (DMF, 99%), dimethyl sulfoxide (DMSO, 99%), acetone (99%), and diethylzinc (DEZ, 95%).

Deposition of Halogenated ZIF Films: ZIF-71 and ZIF-8_Cl: The deposition of the MOF film by CVD involves two steps: 1) deposition of the ZnO film by ALD; 2) vaporized linker reacted with ZnO film in the presence of a template vapor.[42,50,64] ZnO films were deposited by a Savannah S-200 thermal ALD reactor (Veeco Instruments, Inc.) with deionized water and diethylzinc as precursors. The reactant and purge times were set to 0.015 s and 5 s, respectively. The ALD reactor pressure was ≈0.4 mbar and the reaction temperature was at 120 °C. The growth rate of the ZnO film was ≈1.7 Å per cycle. For ZIF-71 deposition, a ZnO film, 50 mg of HdcIm powder, and 50 µL of DMF were placed separately in a 250 mL Schlenk tube and heated in a convection oven at 110 °C for one day. The obtained ZIF-71 films were activated under a dynamic vacuum. ZIF-8_Cl films were deposited by the same process but with Cl-Im instead of HdcIm.

Characterization: GIXRD measurements of halogenated ZIF films were conducted using the Malvern PANalytical Empyrean diffractometer, which was equipped with a PIXcel3D solid-state detector and a Cu anode (λ = 1.5406 Å) operating at 45 kV and 40 mA. These measurements were taken in reflection geometry with an incident beam angle of 0.02°, covering a 2θ range of 2–20°, with a step size of 0.053° and a counting time of 1000 seconds per step. To control beam divergence, a 1/32° fixed anti-scatter slit was used on the incident beam side. The morphology of the MOF films and patterns were recorded on FEI XL30FEG and Raith E-line instrument. The thin film samples were sputtered with 5 nm platinum, while the pattern samples were measured directly without coating. ImageJ was used to calculate LER based on SEM images.[65,66] AFM images were obtained using a PicoSPM (5100, Agilent Technologies) instrument in intermittent contact mode. The Silicon cantilevers (OLYMPUS, AC160TS-R3) had a resonance frequency of 300 kHz. Data processing was conducted with Gwyddion 2.44.

EUV Exposure and Development: EUV lithography was carried out at the XIL-II Beamline of Swiss Light Source and the Beamline of BL08U1B of the Shanghai Synchrotron Radiation Facility. The pressure in the exposure chamber was ≈10^{-7} mbar. The EUV exposure doses were calculated based on the photo fluxes, exposure times, and areas. After EUV exposure, the MOF film was developed in DMSO for 30 s, rinsed with acetone, and dried using a gentle nitrogen gas stream.

In Situ and Near-Ambient Pressure XPS: The in situ and near-ambient pressure XPS experiments were conducted at the 02B01 beamline of the Shanghai Synchrotron Radiation Facility. Different fresh areas (≈200 × 75 µm^2) of the ZIF-71 film were exposed to varying durations at 92 eV in a vacuum (10^{-9} mbar), followed by XPS measurements at 1250 eV. To investigate the effect of water vapor, the ZIF-71 film was exposed to 92 eV after introducing 1 mbar water vapor, with subsequent XPS measurement at 1250 eV. In another scenario, 1 mbar of water vapor was introduced in the chamber after EUV irradiation, and XPS was measured at the same position. XPS spectra of Cl 2p, N 1s, C 1s, and Zn 2p were acquired. The XPS data were processed by Casa XPS software using the Shirley background and the Gaussisn-Lorentzian function.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2021YFB3200800) and the National Natural Science Foundation of China (No. 22201289). The authors acknowledge the Paul Scherrer Institute, Switzerland, for providing synchrotron radiation beamtime at beamline XIL-II of the Swiss Light Source. This research project has received funding from the EU's H2020 framework program for research and innovation under grant agreement 654360 NFFA-Europe. This work received funding from the Research Foundation Flanders (FWO Vlaanderen) through research projects G85720N and G087422N. This work was supported by the Beamline BL08U1B and 02B01 at the Shanghai Synchrotron Radiation Facility, China.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

EUV lithography, metal–organic frameworks, resist-free patterning, thin films, zeolitic imidazolate frameworks

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202415804 by Shanghai Institute Of, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Received: November 27, 2024
Revised: February 17, 2025
Published online:

[1] J. Jiang, L. Xu, C. Qiu, L.-M. Peng, *Nature* **2023**, *616*, 470.
[2] W. Li, X. Gong, Z. Yu, L. Ma, W. Sun, S. Gao, Ç. Köroğlu, W. Wang, L. Liu, T. Li, H. Ning, D. Fan, Y. Xu, X. Tu, T. Xu, L. Sun, W. Wang, J. Lu, Z. Ni, J. Li, X. Duan, P. Wang, Y. Nie, H. Qiu, Y. Shi, E. Pop, J. Wang, X. Wang, *Nature* **2023**, *613*, 274.
[3] L. Liu, J. Han, L. Xu, J. Zhou, C. Zhao, S. Ding, H. Shi, M. Xiao, L. Ding, Z. Ma, C. Jin, Z. Zhang, L.-M. Peng, *Science* **2020**, *368*, 850.
[4] K. Yi, W. Qin, Y. Huang, Y. Wu, S. Feng, Q. Fang, X. Cao, Y. Deng, C. Zhu, X. Zou, K.-W. Ang, T. Li, X. Wang, J. Lou, K. Lai, Z. Hu, Z. Zhang, Y. Dong, K. Kalantar-Zadeh, Z. Liu, *Nat. Electron.* **2024**, *7*, 1126.
[5] M. D. Allendorf, R. Dong, X. Feng, S. Kaskel, D. Matoga, V. Stavila, *Chem. Rev.* **2020**, *120*, 8581.
[6] I. Stassen, N. Burtch, A. Talin, P. Falcaro, M. Allendorf, R. Ameloot, *Chem. Soc. Rev.* **2017**, *46*, 3185.
[7] W. Li, Z. Zhu, Q. Chen, J. Li, M. Tu, *Cell Rep. Phys. Sci.* **2023**, *4*, 101679.
[8] M. Krishtab, I. Stassen, T. Stassin, A. J. Cruz, O. O. Okudur, S. Armini, C. Wilson, S. De Gendt, R. Ameloot, *Nat. Commun.* **2019**, *10*, 3729.
[9] M. Wang, R. Dong, X. Feng, *Chem. Soc. Rev.* **2021**, *50*, 2764.
[10] R. Ameloot, E. Gobechiya, H. Uji-i, J. A. Martens, J. Hofkens, L. Alaerts, B. F. Sels, D. E. De Vos, *Adv. Mater.* **2010**, *22*, 2685.
[11] J. Zhuang, D. Ceglarek, S. Pethuraj, A. Terfort, *Adv. Funct. Mater.* **2011**, *21*, 1442.
[12] G. Lu, O. K. Farha, W. Zhang, F. Huo, J. T. Hupp, *Adv. Mater.* **2012**, *24*, 3970.
[13] J. Zhuang, J. Friedel, A. Terfort, *Beilstein J. Nanotechnol.* **2012**, *3*, 570.
[14] S. Li, W. Shi, G. Lu, S. Li, S. C. J. Loo, F. Huo, *Adv. Mater.* **2012**, *24*, 5954.
[15] O. Lugier, N. Thakur, L. Wu, M. Vockenhuber, Y. Ekinci, S. Castellanos, *ACS Appl. Mater. Interfaces* **2021**, *13*, 43777.
[16] A. Razmjou, M. Asadnia, O. Ghaebi, H.-C. Yang, M. E. Warkiani, J. Hou, V. Chen, *ACS Appl. Mater. Interfaces* **2017**, *9*, 38076.
[17] Y. Wang, I. Fedin, H. Zhang, D. V. Talapin, *Science* **2017**, *357*, 385.
[18] H. Cho, J. Pan, H. Wu, X. Lan, I. Coropceanu, Y. Wang, W. Cho, E. A. Hill, J. S. Anderson, D. V. Talapin, *Adv. Mater.* **2020**, *32*, 2003805.
[19] J.-A. Pan, H. Wu, A. Gomez, J. C. Ondry, J. Portner, W. Cho, A. Hinkle, D. Wang, D. V. Talapin, *ACS Nano* **2022**, *16*, 16067.
[20] J.-A. Pan, H. Cho, I. Coropceanu, H. Wu, D. V. Talapin, *Acc. Chem. Res.* **2023**, *56*, 2286.
[21] P. Xiao, Z. Zhang, J. Ge, Y. Deng, X. Chen, J.-R. Zhang, Z. Deng, Y. Kambe, D. V. Talapin, Y. Wang, *Nat. Commun.* **2023**, *14*, 49.
[22] H. Wu, Y. Wang, J. Yu, J.-A. Pan, H. Cho, A. Gupta, I. Coropceanu, C. Zhou, J. Park, D. V. Talapin, *J. Am. Chem. Soc.* **2022**, *144*, 10495.
[23] S. Baek, J. S. Son, *Adv. Phys. Res.* **2024**, *3*, 2300069.
[24] M. d. J. Velásquez-Hernández, M. Linares-Moreau, L. A. Brandner, B. Marmiroli, M. Barella, G. P. Acuna, S. D. Zilio, M. F. K. Verstreken, D. E. Kravchenko, O. M. Linder-Patton, J. D. Evans, H. Wiltsche, F. Carraro, H. Wolinski, R. Ameloot, C. Doonan, P. Falcaro, *Adv. Mater.* **2023**, *35*, 2211478.
[25] R. Bo, M. Taheri, H. Chen, J. Bradford, N. Motta, S. Surve, T. Tran-Phu, P. Garg, T. Tsuzuki, P. Falcaro, A. Tricoli, *Adv. Funct. Mater.* **2022**, *32*, 2100351.
[26] M. Linares-Moreau, L. A. Brandner, M. D. J. Velásquez-Hernández, J. Fonseca, Y. Benseghir, J. M. Chin, D. Maspoch, C. Doonan, P. Falcaro, *Adv. Mater.* **2024**, *36*, 2309645.
[27] Y. Miao, M. Tsapatsis, *Chem. Mater.* **2021**, *33*, 754.
[28] Y. Miao, D. T. Lee, M. D. De Mello, M. Ahmad, M. K. Abdel-Rahman, P. M. Eckhert, J. A. Boscoboinik, D. H. Fairbrother, M. Tsapatsis, *Nat. Commun.* **2022**, *13*, 420.
[29] M. Tu, B. Xia, D. E. Kravchenko, M. L. Tietze, A. J. Cruz, I. Stassen, T. Hauffman, J. Teyssandier, S. De Feyter, Z. Wang, R. A. Fischer, B. Marmiroli, H. Amenitsch, A. Torvisco, M. d. J. Velásquez-Hernández, P. Falcaro, R. Ameloot, *Nat. Mater.* **2021**, *20*, 93.
[30] P. Corkery, K. E. Waltz, P. M. Eckhert, M. Ahmad, A. Kraetz, Y. Miao, D. T. Lee, M. K. Abdel-Rahman, Y. Lan, P. Haghi-Ashtiani, A. Stein, J. A. Boscoboinik, M. Tsapatsis, D. H. Fairbrother, *Adv. Func. Mater.* **2024**, *34*, 2311149.
[31] C. Luo, C. Xu, L. Lv, H. Li, X. Huang, W. Liu, *RSC Adv.* **2020**, *10*, 8385.
[32] K. D. Closser, D. F. Ogletree, P. Naulleau, D. Prendergast, *J. Chem. Phys.* **2017**, *146*, 164106.
[33] A. Narasimhan, S. Grzeskowiak, C. Ackerman, T. Flynn, G. Denbeaux, R. L. Brainard, in *Extreme Ultraviolet (EUV) Lithography VIII*, SPIE, Bellingham, WA, **2017**, pp. 195–203.
[34] G. Lim, K. Lee, S. Choi, H. J. Yoon, *Coord. Chem. Rev.* **2023**, *493*, 215307.
[35] J. H. Ma, C. Needham, H. Wang, A. Neureuther, D. Prendergast, P. Naulleau, *ACS Appl. Mater. Interfaces* **2022**, *14*, 5514.
[36] G. Lim, K. Lee, C. Koh, T. Nishi, H. J. Yoon, *ACS Mater. Au* **2024**, *4*, 468.
[37] Y. K. Kang, S. J. Lee, S. Eom, B. G. Kim, C.-C. Hwang, M.-G. Kim, *J. Mater. Chem. C* **2024**, *12*, 15855.
[38] H. Chen, Y. Peng, H. Fu, F. Han, G. Shi, F. Luo, J. Zhao, D. Zhou, P. Chen, X. Peng, *CCS Chem.* **2024**, *6*, 2044.
[39] T. Manouras, P. Argitis, *Nanomaterials* **2020**, *10*, 1593.
[40] B. Cardineau, R. Del Re, M. Marnell, H. Al-Mashat, M. Vockenhuber, Y. Ekinci, C. Sarma, D. A. Freedman, R. L. Brainard, *Microelectron. Eng.* **2014**, *127*, 44.
[41] X. Wang, P. Tao, Q. Wang, R. Zhao, T. Liu, Y. Hu, Z. Hu, Y. Wang, J. Wang, Y. Tang, H. Xu, X. He, *Mater. Today* **2023**, *67*, 299.
[42] M. Tu, D. E. Kravchenko, B. Xia, V. Rubio-Giménez, N. Wauteraerts, R. Verbeke, I. F. J. Vankelecom, T. Stassin, W. Egger, M. Dickmann, H. Amenitsch, R. Ameloot, *Angew. Chem., Int. Ed.* **2021**, *133*, 7631.
[43] P. Su, M. Tu, R. Ameloot, W. Li, *Acc. Chem. Res.* **2022**, *55*, 186.
[44] R. Banerjee, A. Phan, B. Wang, C. Knobler, H. Furukawa, M. O'Keeffe, O. M. Yaghi, *Science* **2008**, *319*, 939.
[45] K. Li, D. H. Olson, J. Seidel, T. J. Emge, H. Gong, H. Zeng, J. Li, *J. Am. Chem. Soc.* **2009**, *131*, 10368.
[46] J. Zhao, S. Yang, C. Xue, L. Wang, Z. Liang, L. Zhang, Y. Wang, Y. Wu, R. Tai, *Int. J. Extreme Manuf.* **2020**, *2*, 012005.
[47] C. Popescu, Y. Vesters, A. McClelland, D. De Simone, G. Dawson, J. Roth, W. Theis, G. Vandenberghe, A. P. G. Robinson, *J. Photopolym. Sci. Technol.* **2018**, *31*, 227.
[48] M. Trikeriotis, W. J. Bae, E. Schwartz, M. Krysak, N. Lafferty, P. Xie, B. Smith, P. A. Zimmerman, C. K. Ober, E. P. Giannelis, in *Advances in Resist Materials and Processing Technology XXVII*, SPIE, Bellingham, WA, **2010**, Vol. *7639*, pp. 120–129.
[49] P. M. Dentinger, G. F. Cardinale, C. C. Henderson, A. Fisher, A. K. Ray-Chaudhuri, in *Emerging Lithographic Technologies IV*, SPIE **2000**, Vol. *3997*, pp. 588–599.
[50] A. J. Cruz, I. Stassen, M. Krishtab, K. Marcoen, T. Stassin, S. Rodríguez-Hermida, J. Teyssandier, S. Pletincx, R. Verbeke, V. Rubio-Giménez, S. Tatay, C. Martí-Gastaldo, J. Meersschaut, P. M. Vereecken, S. De Feyter, T. Hauffman, R. Ameloot, *Chem. Mater.* **2019**, *31*, 9462.
[51] A. W. Taylor, S. Men, C. J. Clarke, P. Licence, *RSC Adv.* **2013**, *3*, 9436.
[52] N. Tiempos-Flores, E. Hernández-Fernández, A. Rico-Barragan, J. Raziel Álvarez, I. Juárez-Ramírez, M. A. Garza-Navarro, J. Rodríguez-Hernández, A. Fonseca-García, D. J. Michaelis, N. E. Davila-Guzman, *Polyhedron* **2022**, *217*, 115736.
[53] S. Japip, K.-S. Liao, Y. Xiao, T.-S. Chung, *J. Membr. Sci.* **2016**, *497*, 248.
[54] J. O. Choi, J. A. Moore, J. C. Corelli, J. P. Silverman, H. Bakhru, *J. Vac. Sci. Technol. B* **1988**, *6*, 2286.

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202415804 by Shanghai Institute Of, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[55] H. Ihee, M. Lorenc, T. K. Kim, Q. Y. Kong, M. Cammarata, J. H. Lee, S. Bratos, M. Wulff, *Science* **2005**, *309*, 1223.
[56] R. N. Widmer, G. I. Lampronti, N. Casati, S. Farsang, T. D. Bennett, S. A. T. Redfern, *Phys. Chem. Chem. Phys.* **2019**, *21*, 12389.
[57] B. A. Naqvi, S. Enomoto, K. Machida, Y. Takata, T. Kozawa, Y. Muroya, S. De Gendt, D. De Simone, *Chem. Mater.* **2024**, *36*, 1459.
[58] Q. Evrard, N. Sadegh, S. Mathew, E. Zuidinga, B. Watts, M. P. Dominguez, A. Giglia, N. Mahne, S. Nannarone, A. Nishimura, T. Goya, T. Sugioka, M. Vockenhuber, Y. Ekinci, A. M. Brouwer, *ACS Appl. Mater. Interfaces* **2024**, *16*, 42947.
[59] M. Lewandowski, D. F. Ollis, *J. Catal.* **2003**, *217*, 38.
[60] R. Yuan, S. Fan, H. Zhou, Z. Ding, S. Lin, Z. Li, Z. Zhang, C. Xu, L. Wu, X. Wang, X. Fu, *Angew. Chem., Int. Ed.* **2013**, *52*, 1035.
[61] P. Sivaguru, Z. Wang, G. Zanoni, X. Bi, *Chem. Soc. Rev.* **2019**, *48*, 2615.
[62] S. Ghosh, H. Yun, P. Kumar, S. Conrad, M. Tsapatsis, K. A. Mkhoyan, *Chem. Mater.* **2021**, *33*, 5681.
[63] S. Castellanos, P. D. Schepper, M. Bhattacharya, J. Doise, J. Wouters, A. K. Narasimhan, B. Cardineau, L. McQuade, C. Needham, M. Kocsis, K. Kasahara, S. Meyers, in *Advances in Patterning Materials and Processes XLI*, SPIE, Bellingham, WA, **2024**, *12957*, pp. 23–27.
[64] I. Stassen, M. Styles, G. Grenci, H. V. Gorp, W. Vanderlinden, S. D. Feyter, P. Falcaro, D. D. Vos, P. Vereecken, R. Ameloot, *Nat. Mater.* **2016**, *15*, 304.
[65] I. Mochi, M. Vockenhuber, T. Allenet, Y. Ekinci, in *Photomask Technology*, SPIE, Bellingham, WA, **2020**, pp. 58-67.
[66] M. J. R. Stegmueller, R. J. Grant, P. Schindele, *Surf. Coat. Technol.* **2019**, *375*, 22.

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202415804 by Shanghai Institute Of, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：李岩
方向：集成电路
单位：中国科学院数学与系统科学研究院

题目：特征值问题的并行轨道更新方法的数值分析

SIAM J. Numer. Anal.
Vol. 63, No. 4, pp. 1886–1908
© 2025 Society for Industrial and Applied Mathematics

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

NUMERICAL ANALYSIS OF THE PARALLEL ORBITAL-UPDATING APPROACH FOR EIGENVALUE PROBLEMS*

XIAOYING DAI†, YAN LI†, BIN YANG‡, AND AIHUI ZHOU†

Abstract. The parallel orbital-updating approach is an orbital/eigenfunction iteration based approach for solving eigenvalue problems when many eigenpairs are required. It has been proven to be efficient, for instance, in electronic structure calculations. In this paper, based on the investigation of a quasi-orthogonality, we present the numerical analysis of the parallel orbital-updating approach for linear eigenvalue problems, including convergence and error estimates of the numerical approximations.

Key words. parallel orbital-updating, eigenvalue problem, convergence, quasi-orthogonality

MSC codes. 65F10, 65J05, 65N25, 65N30

DOI. 10.1137/24M1690084

1. Introduction.

Eigenvalue problems are typical models in scientific and engineering computing. For instance, Hartree–Fock-type and Kohn–Sham equations are widely used mathematical models in electronic structure calculations. The eigenvalues and their corresponding eigenfunctions of these equations provide detailed information about the properties of atoms, molecules, and solids, helping to predict chemical reactions, material properties, and physical behaviors (see, e.g., [9, 17, 20, 22]).

In electronic structure calculations of a large system, the approximations of many eigenpairs are required. With discretization and the self-consistent field iteration [21, 22, 26], solving the Hartree–Fock-type equations or the Kohn–Sham equations is then transformed into repeatedly solving some large-scale algebraic eigenvalue problems. It is known that the computational cost of solving such large-scale eigenvalue problems is high. In particular, the solving process often requires large-scale orthogonalizing operations, which demand global summation operations and limit large-scale parallelization. Nowadays, the computational scale is limited for systems with hundreds to thousands of atoms. Since applications demand and supercomputers are available, it is significant to develop scalable and parallelizable numerical methods to solve such eigenvalue problems.

To reduce the computational cost and improve the parallel scalability, a so-called parallel orbital-updating (ParO) approach has been proposed in [10] and developed in [12, 23, 24] for solving eigenvalue problems and their equivalent models resulting from electronic structure calculations. We mention that there are also some other methods for approximating eigenpairs in the literature such as the density matrix based

*Received by the editors September 3, 2024; accepted for publication (in revised form) June 26, 2025; published electronically August 22, 2025.
https://doi.org/10.1137/24M1690084
Funding: This work was supported by the National Key R & D Program of China under grants 2019YFA0709600 and 2019YFA0709601 and the National Natural Science Foundation of China under grants 12021001 and 92270206. B. Yang also acknowledges the support from the Fundamental Research Funds for the Central Universities and the disciplinary funding of Central University of Finance and Economics.
†SKLMS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190 China; and School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China (daixy@lsec.cc.ac.cn, liyan2021@lsec.cc.ac.cn, azhou@lsec.cc.ac.cn).
‡School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 102206 China (binyang@lsec.cc.ac.cn).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

algorithm [27], the subspace iteration algorithm [31], and the projection method based on the root-finding of the analytic function [32]. With the ParO approach, we avoid solving the large-scale eigenvalue problem directly and instead solve some independent large-scale source problems and small-scale projected eigenvalue problems to obtain approximate eigenpairs. Moreover, we see from the numerical experiments in [10, 24] that the stiff matrix corresponding to the small-scale eigenvalue problem is almost diagonal, which may further reduce the computation cost. Because of their independence, these source problems can be solved intrinsically in parallel. For each source problem, the standard parallel strategies can be applied. It then allows a two-level parallelization: one level of parallelization is obtained by partitioning these source problems into different groups of processors, and another level of parallelization is obtained by assigning each source problem to several processors contained in each group. This two-level parallelization demonstrates that the ParO approach has great potential for large-scale calculations. The numerical experiments in [10, 12, 24] show that the ParO approach is efficient, of good scalability of parallelization, and can produce highly accurate approximations. We conclude that the ParO approach is a powerful parallel computing approach for solving eigenvalue problems, in which many eigenpairs are required, and has been integrated into the electronic structure calculation software Quantum ESPRESSO [1]. However, up to now, there has been no mathematical justification for the ParO approach.

The purpose of this paper is to present the numerical analysis of the ParO approach for linear eigenvalue problems. We observe that during the implementation process of the ParO approach, we are able to obtain approximately orthogonal eigenfunctions, which we call quasi-orthogonal eigenfunctions. The ParO approach can be viewed as utilizing the quasi-orthogonal approximations, for which the computational cost is lower, to obtain orthogonal approximations. Our numerical analysis starts from the introduction and investigation of a quasi-orthogonality, which plays a crucial role in our investigations on approximations of eigenvalue problems. We understand that the presence of both single eigenvalues and multiple eigenvalues renders traditional methods for analyzing single eigenvalues no longer applicable. The difficulty for the case of multiple eigenvalues lies in the fact that the traditional measure for the eigenfunction errors is not valid anymore, because the approximate eigenfunctions obtained in iterations may not approximate the same eigenfunction. Instead of focusing on particular eigenfunctions, in our analysis, we employ the eigenspaces and the gap from the eigenspaces to their approximations, which brings additional analyzing complexities and requires sophisticated functional analysis.

Some approaches for constructing source problems in the ParO approach have been proposed in [10]. As a practical example, the shifted-inverse–based ParO algorithm applies the shifted-inverse approach to construct some source problems and solves a small-scale eigenvalue problem in each iteration to update the shifts to speed up the convergence [10, 24]. To analyze the convergence of the algorithm, we first study its simplified version, which fixes the shifts and does not carry out the steps of solving small-scale eigenvalue problems in iterations. Within the framework of the ParO approach, we demonstrate the convergence of numerical solutions produced by the simplified algorithm, which does not require sufficiently accurate initial guesses. Based on the numerical analysis of the simplified version, we then present a more general and informative convergence result of the shifted-inverse–based ParO algorithm than the classical results of the shifted-inverse approach for simple eigenvalues mentioned in, e.g., [3, 25]. To improve the numerical stability, a modified version is proposed in [24], which augments the projected subspace by using the residuals. We also provide a brief outline of the proof for the convergence of this modified algorithm.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The rest of this paper is organized as follows. We recall some existing results of a model problem and introduce the relevant notation in section 2 and provide some elementary analysis for the quasi-orthogonality in section 3. In section 4, we show the convergence and error estimates of the numerical approximations produced by the ParO approach and its practical versions. We give some concluding remarks in section 5. Finally, we present the corresponding detailed proofs in Appendix A.

2. Preliminaries. In this section, we recall some existing results for an eigenvalue problem (including its finite dimensional approximations) that will be used.

Eigenvalue problem. Suppose H is a real separable Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. Consider an eigenvalue problem: find $\lambda \in \mathbb{R}$ and $0 \neq u \in H$ such that

$$a(u,v) = \lambda b(u,v) \quad \forall v \in H, \tag{2.1}$$

where $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are two symmetric bilinear forms over $H \times H$. We assume that

$$a(v,w) \leqslant C_a \|v\|\|w\| \quad \forall v,w \in H$$

and

$$a(v,v) \geqslant c_a \|v\|^2 \quad v \in H$$

with constants $C_a, c_a > 0$. It follows that $a(\cdot,\cdot)$ is an inner product and the induced norm $\|v\|_a = \sqrt{a(v,v)}$ is equivalent to $\|\cdot\|$ on H. We assume that $b(\cdot,\cdot)$ is another inner product of H and $\|\cdot\|_b \equiv \sqrt{b(\cdot,\cdot)}$ is compact with respect to $\|\cdot\|$. For convenience, we shall denote $\|\cdot\|_a$ as $\|\cdot\|$ in this paper.

It is known that (2.1) has a countable sequence of real eigenvalues $0 < \lambda_1 < \lambda_2 < \cdots$ and λ_i has the multiplicity $d_i (i = 1, 2, \ldots)$. The indices of λ_i are $(i,1), \ldots, (i,d_i)$, that is,

$$\lambda_{i-1} < \lambda_i = \lambda_{i1} = \cdots = \lambda_{id_i} < \lambda_{i+1}, \quad i = 1, 2, \ldots,$$

with $\lambda_0 = 0, d_0 = 0$.

For $1 \leqslant j \leqslant d_i$ and $1 \leqslant s \leqslant d_r$, denote $(i,j) < (r,s)$ if $i < r$ or $i = r, j < s$. Let $M(\lambda_i)$ be the eigenspace corresponding to λ_i and $\{u_{ij}\}_{j=1}^{d_i}$ be the orthonormal basis of $M(\lambda_i)$, that is, $M(\lambda_i) = \mathrm{span}\{u_{i1}, \ldots, u_{id_i}\}$ for $i = 1, 2, \ldots$ with $b(u_{ij}, u_{kl}) = \delta_{ik}\delta_{jl}$, where δ_{ik} and δ_{jl} are the Kronecker delta.

We consider to obtain the smallest N clustered eigenvalues of (2.1) and their corresponding eigenfunctions, and we assume that there exists $q \in \mathbb{N}_+$ such that $\sum_{i=1}^{q} d_i = N$.

An example. A typical example of (2.1) is an eigenvalue problem of a partial differential operator over a bounded domain. Let $\Omega \subset \mathbb{R}^d (d \geqslant 1)$ be a bounded domain. We shall use the standard notation for Sobolev spaces $H^1(\Omega)$ with associated norms (see, e.g., [2]). Let $H = H_0^1(\Omega) = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$ and $(\cdot,\cdot)$ be the standard L^2 inner product. Consider the eigenvalue problem: find $\lambda \in \mathbb{R}$ and $u \in H_0^1(\Omega)$ with $\|u\|_{L^2(\Omega)} = 1$ such that

$$-\nabla \cdot (A\nabla u) + cu = \lambda u,$$

where $A : \Omega \to \mathbb{R}^{d\times d}$ is piecewise Lipschitz and symmetric positive definite, and $0 \leqslant c \in L^\infty(\Omega)$. Its associate weak form reads as follows: find $\lambda \in \mathbb{R}$ and $0 \neq u \in H_0^1(\Omega)$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

such that

$$a(u,v) = \lambda b(u,v) \quad \forall v \in H_0^1(\Omega),$$

where

$$a(u,v) = (A\nabla u, \nabla v) + (cu, v), \quad b(u,v) = (u,v).$$

We see that $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy the assumptions above.

Remark 2.1. We mention that the results obtained in this paper are also valid for a more general bilinear form $a(\cdot,\cdot)$ with

$$\|w\|^2_{H_0^1(\Omega)} - C_1^{-1}\|w\|^2_{L^2(\Omega)} \leqslant C_2 a(w,w) \quad \forall w \in H_0^1(\Omega)$$

holding for some constant $C_1, C_2 > 0$ (see, e.g., Remark 2.9 in [13]).

Distance from one subspace to another. To carry out the analysis, we apply the following distance from the nontrivial subspace $U \subset H$ to the subspace $V \subset H$ ([8, 16, 19]):

$$\operatorname{dist}(U,V) := \sup_{u \in U, \|u\|=1} \inf_{v \in V} \|u - v\|.$$

Consistently, for any $u, v \in H$, we define

$$\operatorname{dist}(u,v) := \operatorname{dist}\left(\operatorname{span}\{u\}, \operatorname{span}\{v\}\right).$$

We see that $\operatorname{dist}(u,v)$ is actually the sine of the angle between u and v and is independent of the norms of vectors. Note that $\operatorname{dist}(U,V) = 1$ when $\dim(U) > \dim(V)$. We also observe that for $U, V \subset H$ with $\dim(U) = \dim(V) < \infty$, there holds that

$$\operatorname{dist}(U,V) = \operatorname{dist}(V,U). \tag{2.2}$$

Define $\mathcal{P}_V$ to be the orthogonal projection from H onto $V \subset H$ with respect to the inner product $a(\cdot,\cdot)$.

Finite dimensional approximation. Let V^h be a finite dimensional subspace of H with $\dim(V^h) = N_g \geqslant N$. The standard finite dimensional discretization of (2.1) is defined as follows: find $\lambda^h \in \mathbb{R}$ and $0 \neq u^h \in V^h$ such that

$$a(u^h, v) = \lambda^h b(u^h, v) \quad \forall v \in V^h. \tag{2.3}$$

For convenience, we assume that there exists $p \geqslant q$ such that $N_g = \sum_{i=1}^p d_i$. Then we may order the eigenvalues of (2.3) as follows:

$$0 < \lambda_{11}^h \leqslant \cdots \leqslant \lambda_{1d_1}^h \leqslant \cdots \leqslant \lambda_{pd_p}^h.$$

The assumption is adopted to simplify the notation in our numerical analysis. In fact, our analysis results in this paper hold for all $N_g \geqslant N$, regardless of whether $N_g = \sum_{i=1}^p d_i$ is satisfied. Assume that the corresponding eigenfunctions u_{ij}^h for $(i,j) \leqslant (p, d_p)$ satisfy that $b(u_{ij}^h, u_{kl}^h) = \delta_{ik}\delta_{jl}$. For $i = 1, \ldots, p$, set $M_h(\lambda_i) = \operatorname{span}\{u_{i1}^h, \ldots, u_{id_i}^h\}$.

We obtain from the minimum-maximum principle [5, 8] that

$$\lambda_i \leqslant \lambda_{i1}^h \leqslant \cdots \leqslant \lambda_{id_i}^h, \quad i = 1, 2, \ldots, p.$$

The following conclusion can be found in [14, 18].

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

PROPOSITION 2.2. *For the eigenvalue problem* (2.1) *and its finite dimensional discretization* (2.3)*, there holds that*

$$0 \leqslant \lambda_{ij}^h - \lambda_i \leqslant \lambda_{ij}^h \operatorname{dist}^2\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right) \quad \forall (1,1) \leqslant (i,j) \leqslant (q, d_q).$$

Given the eigenvalue problem (2.1) and its finite dimensional approximation (2.3), the following result is classical and can be found in [5, 8, 18].

PROPOSITION 2.3. *Let* $\{u_{ij}^h\}$ *be the solutions of* (2.3)*. There exists* $\epsilon_* \in (0,1)$ *satisfying that if* $\operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right) \leqslant \epsilon_*$*, there exists* $\hat{u}_{ij} \in M(\lambda_i)$ *such that*

$$\left\|u_{ij}^h - \hat{u}_{ij}\right\| \leqslant C_* \operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right) \quad \forall (1,1) \leqslant (i,j) \leqslant (q, d_q),$$

where C_* *is a constant that is independent of* V^h*.*

Remark 2.4. Based on Theorem 1 and Theorem 2 in [18], the constants in Proposition 2.3 can be chosen as

$$\epsilon_* = \alpha \min_{i=1,\dots,q+1} \sqrt{\frac{\lambda_i - \lambda_{i-1}}{\lambda_i}} \quad \text{and} \quad C_* = \max_{i=1,\dots,q} \sqrt{\frac{2(\lambda_{i+1} - \lambda_{i-1})\lambda_i}{(1-\alpha^2)(\lambda_i - \lambda_{i-1})(\lambda_{i+1} - \lambda_i)}}$$

with $\alpha \in (0,1)$.

3. Quasi-orthogonality. To understand the philosophy behind the ParO approach and carry out the numerical analysis, we introduce a quasi-orthogonality, which plays a crucial role in our investigations on approximations of eigenvalue problems.

DEFINITION 3.1. *Given* $\delta > 0$ *and* $n \geqslant 2$*,* $\{v_j\}_{j=1}^n \subset H$ *with* $\|v_j\| = 1$ *for* $j = 1, \dots, n$ *is said to be* δ*-quasi-orthogonal if there exists* $\{u_j\}_{j=1}^n \subset H$ *satisfying that*

$$a(u_i, u_j) = \delta_{ij},, \quad \|u_j - v_j\| \leqslant \delta, \quad i,j = 1,2,\dots,n. \tag{3.1}$$

The following proposition tells the approximation property of the orthogonalization of quasi-orthogonal vectors. The proof is provided in Appendix A.1.

PROPOSITION 3.2. *If* $\{v_j\}_{j=1}^n \subset H$ *is* δ*-quasi-orthogonal, then there exists an orthonormal basis* $\{\tilde{v}_j\}_{j=1}^n$ *of* $\operatorname{span}\{v_1, \dots, v_n\}$ *such that*

$$\|\tilde{v}_j - v_j\| \leqslant \sqrt{n}\delta, \quad j = 1, \dots, n.$$

The following conclusion, which can be derived directly by a triangle inequality from Proposition 3.2, shows the approximation property between orthonormal bases.

COROLLARY 3.3. *If* $\{v_j\}_{j=1}^n \subset H$ *is* δ*-quasi-orthogonal, i.e., there exists* $\{u_j\}_{j=1}^n \subset H$ *satisfying*

$$a(u_i, u_j) = \delta_{ij}, \quad \|u_j - v_j\| \leqslant \delta, \quad i,j = 1,2,\dots,n,$$

then there exists $\{\tilde{v}_j\}_{j=1}^n \subset \operatorname{span}\{v_1, \dots, v_n\}$ *such that*

$$a(\tilde{v}_i, \tilde{v}_j) = \delta_{ij}, \quad \operatorname{dist}(u_j, \tilde{v}_j) \leqslant \|u_j - \tilde{v}_j\| \leqslant (\sqrt{n}+1)\delta, \quad i,j = 1,\dots,n.$$

Orthogonal basis approximation. We arrive at the following proposition from Corollary 3.3, which tells the approximation property of orthonormal bases by the distance from one subspace to another and will play a crucial role in our analysis.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

PROPOSITION 3.4. *Take* $\varepsilon \in (0,1)$ *and two subspaces* $U, V \subset H$ *with* $\dim(U) = \dim(V) = n$. *If* $\operatorname{dist}(U,V) \leqslant \varepsilon$, *then for any orthonormal basis* $\{u_j\}_{j=1}^n$ *of* U, *there exists an orthonormal basis* $\{w_j\}_{j=1}^n$ *of* V *satisfying*

$$\operatorname{dist}(u_i, w_i) \leqslant (1+\sqrt{n})\sqrt{2-2\sqrt{1-\varepsilon^2}}, \quad i,j = 1,2,\ldots,n.$$

Proof. First, we show that $\mathcal{P}_V|_U$ is an isomorphism from U to V. Indeed, for $\tilde{u}, \hat{u} \in U$ satisfying $\mathcal{P}_V \tilde{u} = \mathcal{P}_V \hat{u}$, we obtain from

$$\begin{aligned} a(\tilde{u} - \mathcal{P}_V \tilde{u}, v) &= 0, \quad v \in V, \\ a(\hat{u} - \mathcal{P}_V \hat{u}, v) &= 0, \quad v \in V \end{aligned}$$

that

$$a(\tilde{u} - \hat{u}, v) = 0, \quad v \in V,$$

and $\tilde{u} = \hat{u}$ due to $\operatorname{dist}(U,V) \leqslant \varepsilon < 1$. Hence, $\mathcal{P}_V|_U$ is an injection and then is isomorphism from U to V since $\dim(V) = \dim(U)$.

Next we set $v_j = \frac{\mathcal{P}_V u_j}{\|\mathcal{P}_V u_j\|}$ for $j = 1,2,\ldots,n$. Since $\mathcal{P}_V$ is an isomorphism, we have that $V = \operatorname{span}(\{v_j\}_{j=1}^n)$ and

$$\begin{aligned} \|u_j - v_j\| &= \sqrt{\|u_j - \mathcal{P}_V u_j\|^2 + \left\| \mathcal{P}_V u_j - \frac{\mathcal{P}_V u_j}{\|\mathcal{P}_V u_j\|} \right\|^2} \\ &\leqslant \sqrt{\operatorname{dist}^2(U,V) + \left(1 - \sqrt{1 - \operatorname{dist}^2(U,V)}\right)^2} \leqslant \sqrt{2 - 2\sqrt{1-\varepsilon^2}}, \end{aligned}$$

i.e., $\{v_j\}_{j=1}^n$ is $\sqrt{2-2\sqrt{1-\varepsilon^2}}$-quasi-orthogonal. Then by Corollary 3.3, we complete the proof. ∎

Dimension-preserving. For $V^h = \bigoplus_{i=1}^p X_i$, we consider subspaces $X, Y \subset V^h \subset H$ with decompositions as follows:

$$X = \bigoplus_{i=1}^q X_i, \quad Y = \sum_{i=1}^q Y_i,$$

where $\dim(X) = N$ and $\dim(X_i) = \dim(Y_i) = d_i$. Obviously, $\dim(Y) \leqslant \sum_{i=1}^q \dim(Y_i) = N$. We have the following conclusion telling when $\dim(Y) = N$ holds and will be used in our analysis. The proof is given in Appendix A.2.

PROPOSITION 3.5. *Take* $\varepsilon \in (0, \min_{1 \leqslant i \leqslant q} \sqrt{\frac{4(1+\sqrt{d_i})^2 N - 1}{4(1+\sqrt{d_i})^4 N^2}})$. *If*

$$\max_{i=1,\ldots,q} \operatorname{dist}(X_i, Y_i) < \varepsilon,$$

then

$$Y = \bigoplus_{i=1}^q Y_i. \tag{3.2}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

An application of quasi-orthogonality. Note that Proposition 2.3 tells only that each orthonormal eigenfunction of (2.3) approximates some eigenfunction of (2.1). However, these exact eigenfunctions of (2.1) may not be orthonormal to each other (see Corollary 2.11 in [11]). In practical applications, the approximate property of the approximate eigenfunctions to the exact orthonormal eigenfunctions is usually required, which is for structure-preserving and preventing the accumulation of errors of approximations.

From solving (2.3), we may obtain orthogonal or quasi-orthogonal (when the algebraic error of the orthogonalization is taken into account) normalized basis $\{u_{ij}^h\}$ of $M_h(\lambda_i)$. With the investigation of the quasi-orthogonality, we are able to estimate the approximation error of the approximate eigenfunctions to the exact orthonormal eigenfunctions.

THEOREM 3.6. *Let $\{u_{ij}^h\}$ be solutions of* (2.3). *If* $\operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right) \leqslant \epsilon_*$, *then there exists an orthonormal basis $\{u_{ij}^o\}$ of $M(\lambda_i)$ with $b(u_{ij}^o, u_{kl}^o) = \delta_{ik}\delta_{jl}$ such that*

$$\operatorname{dist}\left(u_{ij}^o, u_{ij}^h\right) \leqslant C_{**} \operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right), \quad (1,1) \leqslant (i,j) \leqslant (q, d_q),$$

where $C_{**} = \max_{i=1,\ldots,q} \frac{2(\sqrt{d_i}+1)}{\sqrt{\lambda_i}} C_*$. The constants ϵ_* and C_* are defined in Proposition 2.3, *and C_* is independent of V^h.*

Proof. For the approximate eigenpairs $\left\{(\lambda_{ij}^h, u_{ij}^h)\right\}_{(1,1)\leqslant(i,j)\leqslant(q,d_q)}$, we obtain from Proposition 2.3 that there exist $\hat{u}_{ij} \in M(\lambda_i)$ and the constant C_* that is defined in Proposition 2.3 and is independent of V^h such that

$$\left\|u_{ij}^h - \hat{u}_{ij}\right\| \leqslant C_* \operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right), \quad (1,1) \leqslant (i,j) \leqslant (q, d_q),$$

which together with the fact $\|u_{ij}^h\|_a = \sqrt{\lambda_{ij}^h}$ and $\lambda_{ij}^h \geqslant \lambda_i$ yields that

$$\begin{aligned}\left\|\frac{u_{ij}^h}{\|u_{ij}^h\|} - \frac{\hat{u}_{ij}}{\|\hat{u}_{ij}\|}\right\| &= \frac{1}{\sqrt{\lambda_{ij}^h}}\left\|u_{ij}^h - \hat{u}_{ij} + \left(1 - \frac{\|u_{ij}^h\|}{\|\hat{u}_{ij}\|}\right)\hat{u}_{ij}\right\| \\ &\leqslant \frac{1}{\sqrt{\lambda_i}}\left(\left\|u_{ij}^h - \hat{u}_{ij}\right\| + \left|\|\hat{u}_{ij}\| - \|u_{ij}^h\|\right|\right) \leqslant \frac{2}{\sqrt{\lambda_i}} C_* \operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right).\end{aligned}$$

Then $\{\frac{\hat{u}_{ij}}{\|\hat{u}_{ij}\|}\}_{j=1}^{d_i}$ is $\frac{2}{\sqrt{\lambda_i}} C_* \operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right)$-quasi-orthogonal, and it follows from Corollary 3.3 that there exists an orthonormal basis $\{u_{ij}^o\}$ of $M(\lambda_i)$ with $b(u_{ij}^o, u_{ik}^o) = \delta_{jk}$ such that

$$\operatorname{dist}\left(u_{ij}^o, u_{ij}^h\right) \leqslant C_{**} \operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right), \quad j = 1, \ldots, d_i,$$

where $C_{**} = \max_{i=1,\cdots,q} \frac{2(\sqrt{d_i}+1)}{\sqrt{\lambda_i}} C_*$. We complete the proof. ∎

We see that Theorem 3.6 tells that, for the finite dimensional approximation of an eigenvalue problem, there exists a set of orthonormal eigenfunctions whose distance to the orthonormal approximate eigenfunctions is controlled by the distance of subspaces.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

In the next section, we will present the approximation of iterative solutions $\{(\lambda_{ij}^{(n)}, u_{ij}^{(n)})\}$ produced by the ParO approach to the solutions of the discrete problem (2.3) using Theorem 3.6. Then we obtain the approximation errors of iterative solutions to solutions of (2.1) from Proposition 2.2, Theorem 3.6, and the triangle inequalities

$$\left|\lambda_i - \lambda_{ij}^{(n)}\right| \leqslant \left|\lambda_i - \lambda_{ij}^{h}\right| + \left|\lambda_{ij}^{h} - \lambda_{ij}^{(n)}\right|,$$
$$\operatorname{dist}(u_{ij}^{o}, u_{ij}^{(n)}) \leqslant \operatorname{dist}(u_{ij}^{o}, u_{ij}^{h,o}) + \operatorname{dist}(u_{ij}^{h,o}, u_{ij}^{(n)}),$$

for $(1,1) \leqslant (i,j) \leqslant (q, d_q)$, where $\{u_{ij}^{o}\}_{j=1}^{d_i}$ and $\{u_{ij}^{h,o}\}_{j=1}^{d_i}$ with $b(u_{ij}^{o}, u_{kl}^{o}) = b(u_{ij}^{h,o}, u_{kl}^{h,o}) = \delta_{ik}\delta_{jl}$ are orthogonal bases of $M(\lambda_i)$ and $M_h(\lambda_i)$, respectively.

4. Convergence and error estimates. With the quasi-orthogonality, in this section, we are able to obtain the convergence and error estimates of the numerical approximations produced by the ParO approach for clustered eigenvalue problems.

4.1. Algorithm framework. We first recall the framework of the ParO approach for the first N clustered eigenvalues and their corresponding eigenfunctions of (2.1), which is stated as Algorithm 4.1. We mention that Algorithm 4.1 is indeed a modified version of Algorithm 1.1 in [10] (see Figure 1 for its flowchart). We see from Figure 1 clearly that our framework has a feature of two-level parallelization in step 2: one level is obtained by updating N eigenfunctions intrinsically in parallel; the other level is obtained by applying the standard parallel strategies when updating each eigenfunction.

Step 1. We see that there are several ways to provide initial data in step 1 of Algorithm 4.1. We may obtain initial data from

- physical observation or data (see, e.g., [10, 26]). For instance, in electronic structure calculations, we may choose initial data from Gaussian-type orbitals, Slater-type orbitals, atomic orbital based guesses, and so on;
- solving the eigenvalue problem on a coarse grid;
- neural network–based guesses [15, 34].

Since we look for clustered eigenvalues and their corresponding eigenfunctions, we shall consider the approximation of each eigenspace as mentioned in the Introduction. We understand that it is not trivial to obtain the multiplicity d_i of each eigenvalue λ_i. A possible way to approximate the multiplicities is to cluster the initial guesses $\lambda_1^{(0)} \leqslant \lambda_2^{(0)} \leqslant \cdots \leqslant \lambda_N^{(0)}$. By clustering methods such as the Bayesian information criterion and silhouette method (see e.g., [29, 33]), we can get q' clusters with d_i' eigenpairs in the ith cluster ($i = 1, \ldots, q'$), that is,

$$\left\{\left(\lambda_{ij}^{(0)}, u_{ij}^{(0)}\right)\right\}_{i=1,\ldots,q',j=1,\ldots,d_i'} = \left\{\left(\lambda_k^{(0)}, u_k^{(0)}\right)\right\}_{k=1,\ldots,N}. \tag{4.1}$$

Algorithm 4.1 A framework for the ParO approach.

1. Given a finite dimensional subspace V^h and initial data $(\lambda_k^{(0)}, u_k^{(0)}) \in \mathbb{R} \times V^h$ for $k = 1, 2, \ldots, N$, let $n = 0$.
2. For $k = 1, 2, \ldots, N$, update $u_k^{(n)}$ to obtain $u_k^{(n+1/2)}$ in parallel.
3. Let $u_k^{(n+1)} = u_k^{(n+1/2)}$ for $k = 1, \ldots, N$. Or if necessary, project (2.1) onto $U_{n+1} = \operatorname{span}\left\{u_1^{(n+1/2)}, \ldots, u_N^{(n+1/2)}\right\}$ and obtain eigenpairs $(\lambda_k^{(n+1)}, u_k^{(n+1)})$.
4. If they do not converge, let $n = n + 1$ and go to step 2.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

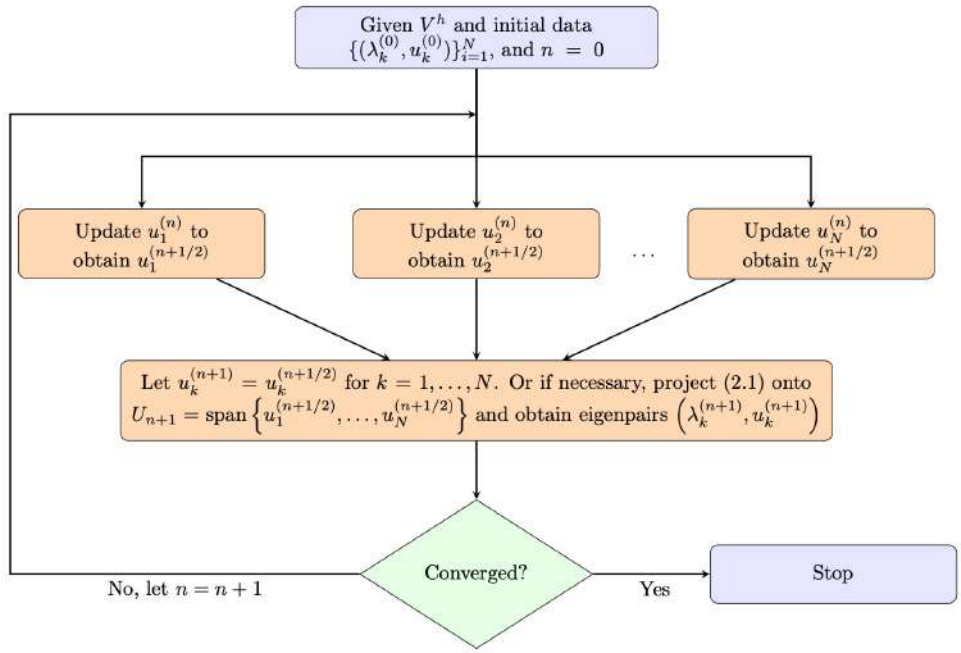


FIG. 1. *Flowchart of Algorithm* 4.1.

In this paper, we assume that $q' = q$ and $d_i' = d_i$ for $i = 1, \ldots, q'$. Such an assumption is likely to hold since the previously mentioned methods may be able to give good initial data. We also want to point out that the numerical experiments in [10, 24] have shown that such an assumption is unnecessary in practical computations.

Define $U_0 = \text{span}\{u_1^{(0)}, u_2^{(0)}, \ldots, u_N^{(0)}\}$ and

$$\lambda_{ij}^{(n+1)} := \lambda_{\sum_{r=0}^{i-1} d_r + j}^{(n+1)}, \quad u_{ij}^{(n+1)} := u_{\sum_{r=0}^{i-1} d_r + j}^{(n+1)}, \quad (1,1) \leqslant (i,j) \leqslant (q, d_q), \quad n \geqslant 0.$$

Set

$$U_n^{(i)} = \text{span}\left\{u_{i1}^{(n)}, \ldots, u_{id_i}^{(n)}\right\}, \quad i = 1, \ldots, q; \tag{4.2}$$

then $U_n = \sum_{i=1}^q U_n^{(i)}$.

Step 2. To update each eigenfunction in step 2 of Algorithm 4.1, as pointed out in [10], we can apply the shifted-inverse approach, Chebyshev filtering, and so on. Define $\mathcal{F}_n : U_n \to U_{n+1}$ as follows:

$$\mathcal{F}_n^{(i)} := \mathcal{F}_n|_{U_n^{(i)}} : u_{ij}^{(n)} = u_{\sum_{r=0}^{i-1} d_r + j}^{(n)} \mapsto u_{\sum_{r=0}^{i-1} d_r + j}^{(n+1/2)} := u_{ij}^{(n+1/2)}, \quad (1,1) \leqslant (i,j) \leqslant (q, d_q).$$

Set

$$U_{n+1/2}^{(i)} := \left\{u_{i1}^{(n+1/2)}, \ldots, u_{id_i}^{(n+1/2)}\right\}, \quad i = 1, \ldots, q.$$

We have $U_{n+1} = \sum_{i=1}^q U_{n+1/2}^{(i)}$.

Step 3. In step 3 of Algorithm 4.1, if we choose to update the approximations, we can solve a small-scale eigenvalue problem as follows: find $(\lambda^{(n+1)}, u^{(n+1)}) \in \mathbb{R} \times U_{n+1}$ satisfying

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$$a\left(u^{(n+1)}, v\right) = \lambda^{(n+1)} b\left(u^{(n+1)}, v\right) \quad \forall v \in U_{n+1} \tag{4.3}$$

to obtain eigenpairs $(\lambda_{ij}^{(n+1)}, u_{ij}^{(n+1)})$ with $b(u_{ij}^{(n+1)}, u_{kl}^{(n+1)}) = \delta_{ik}\delta_{jl}$ for $(1,1) \leqslant (i,j)$, $(k,l) \leqslant (q, d_q)$.

Error estimate. The following theorem shows the approximation errors of eigenpairs when solving a small-scale eigenvalue problem is carried out under the assumption that eigenfunctions are approximated well to a certain in step 2. The proof is given in Appendix A.3.

THEOREM 4.1. *If* $\operatorname{dist}(M_h(\lambda_i), U_{n+1/2}^{(i)}) \ll 1 (i = 1, \ldots, q)$, *then after solving a small-scale eigenvalue problem* (4.3) *in* $U_{n+1} = \sum_{i=1}^{q} U_{n+1/2}^{(i)}$, *there exists an orthonormal basis* $\{u_{ij}^{h,o}\}_{j=1}^{d_i}$ *of* $M_h(\lambda_i)$ *with* $b(u_{ij}^{h,o}, u_{kl}^{h,o}) = \delta_{ik}\delta_{jl}$ *such that for* $(i,j) \leqslant (q, d_q)$,

$$\begin{aligned}
|\lambda_{ij}^{(n+1)} - \lambda_{ij}^{h}| &\leqslant \lambda_{i+1,1}^{h} \operatorname{dist}^2\left(\bigoplus_{i=1}^{q} M_h(\lambda_i), U_{n+1}\right) \\
&\leqslant \lambda_{i+1,1}^{h} \sqrt{N} \max_{1 \leqslant i \leqslant q} \operatorname{dist}^2(M_h(\lambda_i), U_{n+1/2}^{(i)}), \\
\operatorname{dist}(u_{ij}^{h,o}, u_{ij}^{(n+1)}) &\leqslant \tilde{C}_{**} \operatorname{dist}\left(\bigoplus_{i=1}^{q} M_h(\lambda_i), U_{n+1}\right) \\
&\leqslant \tilde{C}_{**} \sqrt{N} \max_{1 \leqslant i \leqslant q} \operatorname{dist}(M_h(\lambda_i), U_{n+1/2}^{(i)}),
\end{aligned}$$

where the constant $\tilde{C}_{**}$ *is independent of* U_{n+1}.

We will see in the next subsection that the requirements $\operatorname{dist}(M_h(\lambda_i), U_{n+1/2}^{(i)}) \ll 1 (i = 1, \ldots, q)$ can be achieved. We understand that the generation and solution of the N-dimensional eigenvalue problem in step 3 of Algorithm 4.1 is computationally expensive. Fortunately, we see from the numerical experiments in [10, 24] that the resulting matrix is almost diagonal: the nondiagonal entries are very small and the computational cost is not very high.

4.2. Shifted-inverse–based ParO algorithm. In this subsection, we shall study the ParO algorithm for solving the clustered eigenvalue problem when the shifted-inverse approach is applied to update each eigenfunction. The shifted-inverse–based ParO algorithm (Algorithm 4.3), which solves a small-scale eigenvalue problem in each iteration to update the shifts, has been proposed in [10, 24]. To show the convergence, we first consider a simplified version (Algorithm 4.2) which fixes the shifts and does not carry out the steps of solving projected eigenvalue problems in iterations. Based on the numerical analysis of the simplified version, we then prove that the approximations produced by Algorithm 4.3 converge rapidly.

Simplified shifted-inverse–based ParO algorithm. We set the shift as any convex combination of $\{\lambda_{ij}^{(0)}\}_{j=1}^{d_i}$ denoted by

$$\bar{\lambda}_i := \mathcal{C}_i\left(\left\{\lambda_{ij}^{(0)}\right\}_{j=1}^{d_i}\right), \quad i = 1, \ldots, q.$$

For instance, we can choose $\mathcal{C}_i(\{\lambda_{ij}^{(0)}\}_{j=1}^{d_i}) = \frac{1}{d_i}\sum_{j=1}^{d_i} \lambda_{ij}^{(0)}$.

In our discussion, we assume that the shifts are always not equal to any eigenvalue of (2.3) in the calculation process. Otherwise, we continue the iterative process on other eigenfunctions while keeping the eigenfunctions unchanged.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Algorithm 4.2 Simplified shifted-inverse–based ParO algorithm.

1. Given a finite dimensional space V^h and $tol > 0$, provide and cluster initial data by (4.1), i.e., $\{(\lambda_{ij}^{(0)}, u_{ij}^{(0)})\}_{(1,1)\leqslant(i,j)\leqslant(q,d_q)} \subset \mathbb{R} \times V^h$. Set $\bar{\lambda}_i = \mathcal{C}_i$ $(\{\lambda_{ij}^{(0)}\}_{j=1}^{d_i})$ and let $n = 0$.
2. For $(1,1) \leqslant (i,j) \leqslant (q,d_q)$, find $u_{ij}^{(n+1/2)} \in V^h$ in parallel by solving
$$a\left(u_{ij}^{(n+1/2)}, v\right) - \bar{\lambda}_i b\left(u_{ij}^{(n+1/2)}, v\right) = \bar{\lambda}_i b\left(u_{ij}^{(n)}, v\right) \quad \forall v \in V^h. \tag{4.4}$$
3. Set $u_{ij}^{(n+1)} = \frac{u_{ij}^{(n+1/2)}}{\|u_{ij}^{(n+1/2)}\|_b}$. If $\frac{\|u_{ij}^{(n+1)} - u_{ij}^{(n)}\|_b}{\|u_{ij}^{(n)}\|_b} > tol$, let $n = n+1$, and go to step 2.

The shifted-inverse approach $\mathcal{F}_n$ is written as follows: for $U_n^{(i)} = \text{span}\{u_{i1}^{(n)}, \ldots, u_{id_i}^{(n)}\}$, $\mathcal{F}_n^{(i)} := \mathcal{F}_n|_{U_n^{(i)}} : U_n^{(i)} \to U_{n+1/2}^{(i)}$ with $u_{ij}^{(n+1/2)} = \mathcal{F}_n^{(i)} u_{ij}^{(n)}$ for $i = 1, \ldots, q, j = 1, 2, \ldots, d_i$ satisfying

$$a(u_{ij}^{(n+1/2)}, v) - \bar{\lambda}_i b(u_{ij}^{(n+1/2)}, v) = \bar{\lambda}_i b(u_{ij}^{(n)}, v) \quad \forall v \in V^h.$$

The simplified shifted-inverse–based ParO algorithm is stated as Algorithm 4.2. Compared with Algorithm 4.1, step 3 is no longer carried out in this simplified version.

If the initial guesses approximate the exact eigenvalues well enough, then the source problems (4.4) will be ill-conditioned. We mention that there are approaches to deal with ill-conditioned systems (see, e.g., [4, 6, 7, 28, 30]). Indeed, it is quite difficult to solve these ill-conditioned systems well, which will be discussed in our other work. Here we assume that such systems can be well solved.

Algorithm 4.2 may be viewed as an extension of the shifted-inverse approach to clustered eigenvalue problems.

The following proposition tells that the approximation to the eigenspace in the iteration process is of dimension-preserving. The proof is given in subsection A.4.

PROPOSITION 4.2. *If $U_{n+1/2}^{(i)}$ is obtained by Algorithm 4.2, then*

$$\dim\left(U_{n+1/2}^{(i)}\right) = \dim\left(U_n^{(i)}\right), \quad i = 1, 2, \ldots, q.$$

With the help of Proposition 4.2, we obtain convergence of eigenspace approximations produced by Algorithm 4.2. The proof is given in Appendix A.5.

THEOREM 4.3. *Assume that*

$$0 < \delta_0 := \max_{(1,1)\leqslant(i,j)\leqslant(q,d_q)} |\lambda_{ij}^h - \bar{\lambda}_i| < \frac{g}{2}, \tag{4.5}$$

where $g := \min_{1\leqslant i<r\leqslant q+1} |\lambda_{id_i}^h - \lambda_{r1}^h|$;

$$\dim\left(U_0^{(i)}\right) = d_i, \quad \text{dist}\left(M_h(\lambda_i), U_0^{(i)}\right) < 1 \quad \forall i = 1, 2, \ldots, q. \tag{4.6}$$

If $U_{n+1/2}^{(i)}$ is produced by Algorithm 4.2, then

$$\text{dist}\left(M_h(\lambda_i), U_{n+1/2}^{(i)}\right) \leqslant \varepsilon_{n+1} := \frac{\delta_0 \varepsilon_n}{\sqrt{(g-\delta_0)^2(1-\varepsilon_n^2) + \delta_0^2 \varepsilon_n^2}}, \quad \lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n} = \frac{\delta_0}{g - \delta_0}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Algorithm 4.3 Shifted-inverse–based ParO algorithm.

1. Given a finite dimensional space V^h and $tol > 0$, provide and cluster initial data by (4.1), i.e., $\left\{(\lambda_{ij}^{(0)}, u_{ij}^{(0)})\right\}_{(1,1)\leqslant(i,j)\leqslant(q,d_q)} \subset \mathbb{R}\times V^h$. Set $\bar{\lambda}_i^{(0)} = \mathcal{C}_i$ $(\{\lambda_{ij}^{(0)}\}_{j=1}^{d_i})$ and let $n = 0$.
2. For $(1,1)\leqslant(i,j)\leqslant(q,d_q)$, find $u_{ij}^{(n+1/2)} \in V^h$ in parallel by solving
$$a\left(u_{ij}^{(n+1/2)}, v\right) - \bar{\lambda}_i^{(n)} b\left(u_{ij}^{(n+1/2)}, v\right) = \bar{\lambda}_i^{(n)} b\left(u_{ij}^{(n)}, v\right) \quad \forall v \in V^h. \tag{4.7}$$
3. Solve an eigenvalue problem: find $(\lambda^{(n+1)}, u^{(n+1)}) \in \mathbb{R}\times U_{n+1} = \mathbb{R}\times \operatorname{span}\{u_{11}^{(n+1/2)}, \ldots, u_{qd_q}^{(n+1/2)}\}$ satisfying
$$a\left(u^{(n+1)}, v\right) = \lambda^{(n+1)} b\left(u^{(n+1)}, v\right) \quad \forall v \in U_{n+1}, \tag{4.8}$$
to obtain eigenpairs $\left\{(\lambda_{ij}^{(n+1)}, u_{ij}^{(n+1)})\right\}$ with $b\left(u_{ij}^{(n+1)}, u_{kl}^{(n+1)}\right) = \delta_{ik}\delta_{jl}$.
4. If $\sum_{i=1}^{q}\sum_{j=1}^{d_i} |\lambda_{ij}^{(n+1)} - \lambda_{ij}^{(n)}| > tol$, set $\bar{\lambda}_i^{(n+1)} = \mathcal{C}_i\left(\{\lambda_{ij}^{(n+1)}\}_{j=1}^{d_i}\right)$, $n = n+1$ and go to step 2.

Theorem 4.3 shows that convergence of Algorithm 4.2 does not require sufficiently accurate initial guesses. (4.5) ensures that the shift is closer to the eigenvalue being approximated. Indeed, (4.5) can be satisfied under the assumption that the finite dimensional discretization (2.3) approximates (2.1) not "too badly." (4.6) guarantees that the dimension of the approximated subspace is preserved. When $N = 1$, (A.11) may be reviewed as the classical shift-inverse convergence result.

Shifted-inverse–based ParO algorithm. We now analyze the convergence of the shifted-inverse–based ParO algorithm proposed in [10, 24], which is stated as Algorithm 4.3.

As mentioned above, we assume that the shifts are always not equal to any eigenvalue of (2.3) in the calculation process. If there are cases where some eigenfunctions are very well approximated, while other eigenfunctions have not yet converged, then we continue the iterative process on the nonconverged eigenfunctions while keeping the well approximated eigenfunctions unchanged.

The following theorem tells the convergence of Algorithm 4.3. The proof is given in subsection A.6

THEOREM 4.4. *Assume that there exists* $0 < \varepsilon_0 \ll 1$ *and an orthonormal basis* $\{u_{ij}^{h,o,0}\}_{j=1}^{d_i}$ *of* $M_h(\lambda_i)(i = 1, \ldots, q)$ *with* $b(u_{ij}^{h,o,0}, u_{kl}^{h,o,0}) = \delta_{ik}\delta_{jl}$ *such that*

$$\operatorname{dist}\left(u_{ij}^{h,o,0}, u_{ij}^{(0)}\right) \leqslant \varepsilon_0, \quad (1,1)\leqslant(i,j)\leqslant(q,d_q), \tag{4.9}$$

$$\zeta_0 := \max_{1\leqslant i\leqslant q}\left|\mathcal{C}_i\left(\{\lambda_{ij}^h\}_{j=1}^{d_i}\right) - \bar{\lambda}_i^{(0)}\right| \ll g, \quad \gamma := \max_{1\leqslant i\leqslant q}\left(\lambda_{id_i}^h - \lambda_{i1}^h\right) \ll g. \tag{4.10}$$

If $\{u_{ij}^{(n+1)}\}$ *are produced by Algorithm* 4.3*, then there exists an orthonormal basis* $\{u_{ij}^{h,o,n+1}\}_{j=1}^{d_i}$ *of* $M_h(\lambda_i)(i = 1, \ldots, q)$ *with* $b(u_{ij}^{h,o,n+1}, u_{kl}^{h,o,n+1}) = \delta_{ik}\delta_{jl}$ *such that*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\operatorname{dist}\left(u_{ij}^{h,o,n+1}, u_{ij}^{(n+1)}\right) \leqslant \varepsilon_{n+1} := \frac{\tilde{C}_{**}\sqrt{DN}\,(\gamma+\zeta_n)\,\varepsilon_n}{\sqrt{(g-\gamma-\zeta_n)^2(1-D\varepsilon_n^2)+D\,(\gamma+\zeta_n)^2\,\varepsilon_n^2}},$$

$$|\lambda_{ij}^h - \lambda_{ij}^{(n+1)}| \leqslant \zeta_{n+1} := \frac{\lambda_{q+1,1}^h}{\tilde{C}_{**}^2}\varepsilon_{n+1}^2, \quad (1,1) \leqslant (i,j) \leqslant (q,d_q),$$

where $D := \max_{1\leqslant i\leqslant q} d_i$ *and the constant* $\tilde{C}_{**}$ *is defined in Theorem* 4.1 *and is independent of* $U_n (n=0,1,2,\ldots)$, *and*

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n} = \frac{\tilde{C}_{**}\sqrt{DN}\gamma}{g-\gamma}. \tag{4.11}$$

Note that $\gamma \ll 1$ when $\operatorname{dist}\left(\bigoplus_{i=1}^q M(\lambda_i), V^h\right) \ll 1$, which together with (4.11) implies that Algorithm 4.3 converges faster when the finite dimensional discretization (2.3) approximates (2.1) better.

If (2.1) is already a discrete eigenvalue problem, then $\gamma = 0$. We obtain from the proof of Theorem 4.4 that

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^3} = \frac{\sqrt{DN}\lambda_{q+1,1}^h}{g\tilde{C}_{**}}, \tag{4.12}$$

which implies that it is a cubic convergence result. Note that the above cubic convergence stems from the convergence of the shifts to some eigenvalues of (2.3), which is exactly what $\gamma = 0$ means. The constant $\tilde{C}_{**}$ in (4.12) comes from the application of Theorem 4.1. Indeed, we can choose a larger $\tilde{C}_{**}$ since (4.12) implies that Algorithm 4.3 converges faster with larger $\tilde{C}_{**}$. However, when $\tilde{C}_{**}$ is larger, more exact initial values are required. For example, if we expect ε_n to decrease toward 0 as $n \to \infty$, a necessary condition,

$$\varepsilon_0 \geqslant \varepsilon_1 = \frac{\tilde{C}_{**}\sqrt{DN}\zeta_0\varepsilon_0}{\sqrt{(g-\zeta_0)^2(1-D\varepsilon_0^2)+D\zeta_0^2\varepsilon_0^2}},$$

implies that ε_0 and ζ_0 are required smaller with larger $\tilde{C}_{**}$. Hence, from Theorem 3.6 and (4.11), we obtain that Algorithm 4.3 applied to a discrete eigenvalue problem converges in cubic rate and goes faster with more exact initial values. The classical result in the discrete case ($N = 1$) is stated as

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^3} \leqslant 1,$$

under the assumption of convergence of the algorithm (see, e.g., [3, 25]). In contrast, Theorem 4.4 is more precise and also for general clustered eigenvalue problems. In addition, (4.11) and (4.12) show what the speed of the convergence depends on.

Modified shifted-inverse–based ParO algorithm. To improve the numerical stability, a modified shifted-inverse–based ParO algorithm is proposed in [24]. We note that step 2 of Algorithm 4.3 can also be written as follows: for $(1,1) \leqslant (i,j) \leqslant (q,d_q)$, find $e_{ij}^{(n+1/2)} \in V^h$ in parallel satisfying

$$a(e_{ij}^{(n+1/2)}, v) - \bar{\lambda}_i^{(n)} b(e_{ij}^{(n+1/2)}, v) = 2\bar{\lambda}_i^{(n)} b(u_{ij}^{(n)}, v) - a(u_{ij}^{(n)}, v) \quad \forall v \in V^h, \tag{4.13}$$

and set $u_{ij}^{(n+1/2)} = u_{ij}^{(n)} + e_{ij}^{(n+1/2)}$. Then instead of solving the N-dimensional projected eigenvalue problem in $U_{n+1} = \operatorname{span}\{u_{11}^{(n+1/2)}, \ldots, u_{1d_1}^{(n+1/2)}, \ldots, u_{qd_q}^{(n+1/2)}\}$, we consider the augmented $2N$-dimensional subspace

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Algorithm 4.4 Modified Shifted-Inverse–Based ParO Algorithm.

1. Given a finite dimensional space V^h and $tol > 0$, provide and cluster initial data by (4.1), i.e., $\left\{(\lambda_{ij}^{(0)}, u_{ij}^{(0)})\right\}_{(1,1)\leqslant(i,j)\leqslant(q,d_q)} \subset \mathbb{R}\times V^h$. Set $\bar{\lambda}_i^{(0)} = \mathcal{C}_i(\{\lambda_{ij}^{(0)}\}_{j=1}^{d_i})$ and let $n=0$.
2. For $(1,1)\leqslant(i,j)\leqslant(q,d_q)$, find $e_{ij}^{(n+1/2)} \in V^h$ in parallel by solving
$$a(e_{ij}^{(n+1/2)}, v) - \bar{\lambda}_i^{(n)} b(e_{ij}^{(n+1/2)}, v) = 2\bar{\lambda}_i^{(n)} b(u_{ij}^{(n)}, v) - a(u_{ij}^{(n)}, v) \quad \forall v \in V^h.$$
3. Solve an eigenvalue problem: find $(\lambda^{(n+1)}, u^{(n+1)}) \in \mathbb{R}\times\tilde{U}_{n+1} = \mathbb{R}\times \operatorname{span}\left\{u_{11}^{(n+1/2)}, \ldots, u_{qd_q}^{(n+1/2)}, e_{11}^{(n+1/2)}, \ldots, e_{qd_q}^{(n+1/2)}\right\}$ satisfying
$$a\left(u^{(n+1)}, v\right) = \lambda^{(n+1)} b\left(u^{(n+1)}, v\right) \quad \forall v \in \tilde{U}_{n+1}, \tag{4.14}$$
to obtain eigenpairs $\left\{\left(\lambda_{ij}^{(n+1)}, u_{ij}^{(n+1)}\right)\right\}$ with $b(u_{ij}^{(n+1)}, u_{kl}^{(n+1)}) = \delta_{ik}\delta_{jl}$.
4. If $\sum_{i=1}^{q}\sum_{j=1}^{d_i} |\lambda_{ij}^{(n+1)} - \lambda_{ij}^{(n)}| > tol$, set $\bar{\lambda}_i^{(n+1)} = \mathcal{C}_i\left(\left\{\lambda_{ij}^{(n+1)}\right\}_{j=1}^{d_i}\right)$, $n = n+1$ and go to step 2.

$$\tilde{U}_{n+1} = \operatorname{span}\left\{u_{11}^{(n+1/2)}, \ldots, u_{1d_1}^{(n+1/2)}, \ldots, u_{qd_q}^{(n+1/2)}, e_{11}^{(n+1/2)}, \ldots, e_{1d_1}^{(n+1/2)}, \ldots, e_{qd_q}^{(n+1/2)}\right\}.$$

For the completeness of this paper, we show the modified shifted-inverse–based ParO algorithm proposed in [24] here, which is stated as Algorithm 4.4.

The convergence of Algorithm 4.4 follows from the similar argument of the proof for Theorem 4.4 together with the fact that

$$\tilde{U}_{n+1} = U_{n+1} \cup U_n, \quad \operatorname{dist}(M_h(\lambda_i), \tilde{U}_{n+1}) \leqslant \operatorname{dist}(M_h(\lambda_i), U_{n+1}).$$

In fact, there have been several numerical experiments for the ParO approach. Numerical experiments in [10] apply finite element discretizations and simulate several typical molecular systems: H_2O (water), $C_9H_8O_4$ (aspirin), $C_5H_9O_2N$ (α amino acid), $C_{20}H_{14}N_4$ (porphyrin), and C_{60} (fullerene). Numerical experiments in [24] apply the plane-wave discretization and simulate several crystalline systems: Si (silicon), MgO (magnesium oxide), and Al (aluminum) with different sizes. These numerical experiments show that the ParO approach is efficient and can produce highly accurate approximations to large-scale systems. Good scalability of parallelization of the ParO approach is also shown in those experiments. It is noteworthy that the ParO approach has been integrated into the electronic structure calculation software Quantum ESPRESSO.

5. Concluding remarks. In this paper, we have provided the numerical analysis of the ParO approach for linear eigenvalue problems based on the investigation of a quasi-orthogonality. Within the framework of the ParO approach, we have demonstrated the convergence of the associated practical algorithms. We point out that numerical experiments in [10, 12, 24] show that the ParO approach is very efficient

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

for electronic structure calculations. Due to the space limitation, we shall address the numerical analysis of the approach for the Kohn–Sham equation in a separate article. It is also our ongoing work to carry out the numerical analysis for the ParO based optimization approach proposed in [12].

Appendix A. Detailed proofs.

A.1. Proof of Proposition 3.2.

Proof. For convenience, we first introduce the following notation. For $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n \subset H$, we set $\mathcal{X} = (x_1, x_2, \ldots, x_n)$, $\mathcal{Y} = (y_1, y_2, \ldots, y_n) \in H^n$ and define

$$\|\mathcal{X}\| = \sqrt{\sum_{i=1}^n \|x_i\|^2}, \quad \text{and} \quad \mathcal{X}^T\mathcal{Y} = \begin{pmatrix} a(x_1, y_1) & \cdots & a(x_1, y_n) \\ \vdots & \ddots & \vdots \\ a(x_n, y_1) & \cdots & a(x_n, y_n) \end{pmatrix}.$$

Let $\mathcal{V} = (v_1, v_2, \ldots, v_n)$; since $\{v_j\}_{j=1}^n \subset H$ is δ-quasi-orthogonal, there exists $\mathcal{U} = (u_1, u_2, \ldots, u_n)$ satisfying that

$$\mathcal{U}^T\mathcal{U} = I, \quad \|\mathcal{U} - \mathcal{V}\| \leqslant \sqrt{\sum_{i=1}^n \|u_i - v_i\|^2} \leqslant \sqrt{n}\delta. \tag{A.1}$$

Let $\operatorname{span}(\mathcal{V}) = \operatorname{span}\{v_1, \ldots, v_n\}$. If $\mathcal{W} \in H^n$ with $\mathcal{W}^T\mathcal{W} = I$ and $\operatorname{span}(\mathcal{W}) = \operatorname{span}(\mathcal{V})$, then there exists the polar decomposition of $\mathcal{W}^T\mathcal{V}$ as follows:

$$\mathcal{W}^T\mathcal{V} = QP, \tag{A.2}$$

where $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $P \in \mathbb{R}^{n\times n}$ is positive definite. Denote $\tilde{\mathcal{V}} = \mathcal{W}Q$, and we see that $\tilde{\mathcal{V}}^T\tilde{\mathcal{V}} = Q^T\mathcal{W}^T\mathcal{W}Q = I$ and

$$\mathcal{V} = \mathcal{W}\mathcal{W}^T\mathcal{V} = \mathcal{W}QP = \tilde{\mathcal{V}}P. \tag{A.3}$$

The decomposition (A.3) is unique since (A.2) implies

$$P = \sqrt{(\mathcal{W}^T\mathcal{V})^T\mathcal{W}^T\mathcal{V}} = \sqrt{\mathcal{V}^T\mathcal{W}^T\mathcal{W}\mathcal{V}} = \sqrt{\mathcal{V}^T\mathcal{V}}.$$

Next, we show that the maximum singular value $\sigma_{\max}(\mathcal{U}^T\tilde{\mathcal{V}}) \leqslant 1$. Consider $\mathcal{Z} \in H^m (m \geqslant n)$ with $\mathcal{Z}^T\mathcal{Z} = I$ and $\operatorname{span}(\mathcal{Z}) = \operatorname{span}(\mathcal{U}) + \operatorname{span}(\tilde{\mathcal{V}})$. We observe that there exist $Q_1, Q_2 \in \mathbb{R}^{m\times n}$ with $Q_1^TQ_1 = Q_2^TQ_2 = I$ satisfying

$$\mathcal{U} = \mathcal{Z}Q_1, \quad \tilde{\mathcal{V}} = \mathcal{Z}Q_2,$$

and

$$\sigma_{\max}(\mathcal{U}^T\tilde{\mathcal{V}}) = \sigma_{\max}((\mathcal{Z}Q_1)^T\mathcal{Z}Q_2) = \sigma_{\max}(Q_1^TQ_2) \leqslant \sigma_{\max}(Q_1)\sigma_{\max}(Q_2) \leqslant 1. \tag{A.4}$$

It follows from (A.3) and (A.4) that

$$\begin{aligned}
\|\mathcal{U} - \mathcal{V}\|^2 &= \|\mathcal{U}\|^2 + \|\mathcal{V}\|^2 - 2\operatorname{tr}(\mathcal{U}^T\mathcal{V}) \geqslant \|\tilde{\mathcal{V}}\|^2 + \|\mathcal{V}\|^2 - 2\sum_{i=1}^n \sigma_i(\mathcal{U}^T\mathcal{V}) \\
&= \|\tilde{\mathcal{V}}\|^2 + \|\mathcal{V}\|^2 - 2\sum_{i=1}^n \sigma_i(\mathcal{U}^T\tilde{\mathcal{V}}\tilde{\mathcal{V}}^T\mathcal{V}) \\
&\geqslant \|\tilde{\mathcal{V}}\|^2 + \|\mathcal{V}\|^2 - 2\sum_{i=1}^n \sigma_{\max}(\mathcal{U}^T\tilde{\mathcal{V}})\sigma_i(P) \\
&\geqslant \|\tilde{\mathcal{V}}\|^2 + \|\mathcal{V}\|^2 - 2\operatorname{tr}(\tilde{\mathcal{V}}^T\mathcal{V}) = \|\tilde{\mathcal{V}} - \mathcal{V}\|^2.
\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Hence, we have

$$\|\tilde{\mathcal{V}} - \mathcal{V}\| \leqslant \|\mathcal{U} - \mathcal{V}\|,$$

which together with (A.1) yields that for $\tilde{\mathcal{V}} = (\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_n)$,

$$\|\tilde{v}_j - v_j\| \leqslant \|\tilde{\mathcal{V}} - \mathcal{V}\| \leqslant \|\mathcal{U} - \mathcal{V}\| \leqslant \sqrt{n}\delta, \quad j = 1, 2, \ldots, n. \qquad \square$$

A.2. Proof of Proposition 3.5.

Proof. Let $\{x_{ij}\}_{j=1}^{d_i}$ be an orthonormal basis of $X_i (i = 1, \ldots, p)$ with $a(x_{ij}, x_{kl}) = \delta_{ik}\delta_{jl}$. We obtain from Proposition 3.4 that there exists an orthonormal basis $\{y_{ij}\}_{j=1}^{d_i}$ of $Y_i (i = 1, \ldots, q)$ satisfying that for $(1, 1) \leqslant (i, j) \leqslant (q, d_q)$,

$$\operatorname{dist}(x_{ij}, y_{ij}) \leqslant \max_{i=1,\ldots,q} (1 + \sqrt{d_i})\sqrt{2 - 2\sqrt{1 - \varepsilon^2}} := \tilde{\varepsilon}. \tag{A.5}$$

Obviously, $\varepsilon < \min_{1 \leqslant i \leqslant q} \sqrt{\frac{4(1+\sqrt{d_i})^2 N - 1}{4(1+\sqrt{d_i})^4 N^2}}$ implies that $\tilde{\varepsilon} < \frac{1}{\sqrt{N}}$. Set $\{\beta_{rt}^{(ij)}\}$ such that $y_{ij} = \sum_{r=1}^{p} \sum_{t=1}^{d_r} \beta_{rt}^{(ij)} x_{rt}$ for $(1, 1) \leqslant (i, j) \leqslant (q, d_q)$, and we have that

$$\left(y_{11}, \ldots, y_{1d_1}, \ldots, y_{q1}, \ldots, y_{qd_q}\right) = \left(x_{11}, \ldots, x_{1d_1}, \ldots, x_{pd_p}\right) B_1,$$

where

$$B_1 = \begin{pmatrix} \beta_{11}^{(11)} & \cdots & \beta_{11}^{(1d_1)} & \cdots & \beta_{11}^{(q1)} & \cdots & \beta_{11}^{(qd_q)} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \beta_{1d_1}^{(11)} & \cdots & \beta_{1d_1}^{(1d_1)} & \cdots & \beta_{1d_1}^{(q1)} & \cdots & \beta_{1d_1}^{(qd_q)} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \beta_{p1}^{(11)} & \cdots & \beta_{p1}^{(1d_1)} & \cdots & \beta_{p1}^{(q1)} & \cdots & \beta_{p1}^{(qd_q)} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \beta_{pd_p}^{(11)} & \cdots & \beta_{pd_p}^{(1d_1)} & \cdots & \beta_{pd_p}^{(q1)} & \cdots & \beta_{pd_p}^{(qd_q)} \end{pmatrix} := \begin{pmatrix} B_2 \\ \star \end{pmatrix},$$

with

$$B_2 = \begin{pmatrix} \beta_{11}^{(11)} & \cdots & \beta_{11}^{(qd_q)} \\ \vdots & \ddots & \vdots \\ \beta_{qd_q}^{(11)} & \cdots & \beta_{qd_q}^{(qd_q)} \end{pmatrix}.$$

We assert that B_2 is strictly diagonally dominant. In fact, we obtain from (A.5) that

$$\tilde{\varepsilon} \geqslant \operatorname{dist}(x_{ij}, y_{ij}) = \operatorname{dist}(y_{ij}, x_{ij}) = \frac{\left\| \sum_{r=1}^{p} \sum_{t=1}^{d_r} \beta_{rt}^{(ij)} x_{rt} - \beta_{ij}^{(ij)} x_{ij} \right\|}{\left\| \sum_{r=1}^{p} \sum_{t=1}^{d_r} \beta_{rt}^{(ij)} x_{rt} \right\|}.$$

Note that

$$\left(\beta_{ij}^{(ij)}\right)^2 \geqslant \left(1 - \tilde{\varepsilon}^2\right) \sum_{r=1}^{p} \sum_{t=1}^{d_r} \left(\beta_{rt}^{(ij)}\right)^2, \quad j = 1, \ldots, d_i$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

implies

$$\left|\beta_{ij}^{(ij)}\right| \geqslant \sqrt{\left(\frac{1}{\tilde{\varepsilon}^2}-1\right)\sum_{(r,t)\neq(i,j)}\left(\beta_{rt}^{(ij)}\right)^2} > \sqrt{(N-1)\sum_{(r,t)\neq(i,j)}\left(\beta_{rt}^{(ij)}\right)^2}$$
$$\geqslant \sum_{(r,t)\neq(i,j)}\left|\beta_{rt}^{(ij)}\right|, \quad (1,1)\leqslant(i,j)\leqslant(q,d_q).$$

It follows from the Gershgorin circle theorem that

$$\left|\lambda-\beta_{ij}^{(ij)}\right| \leqslant \sum_{(r,t)\neq(i,j)}\left|\beta_{rt}^{(ij)}\right|, \quad \forall\lambda\in\sigma(B_2).$$

Consequently, we have

$$|\lambda| \geqslant \left|\beta_{ij}^{(ij)}\right| - \sum_{(r,t)\neq(i,j)}\left|\beta_{rt}^{(ij)}\right| > 0, \quad (1,1)\leqslant(i,j)\leqslant(q,d_q),$$

and $\operatorname{rank}(B_2)=N$. Therefore, $\operatorname{rank}(B_1)=N$, which completes the proof. ∎

A.3. Proof of Theorem 4.1.

Proof. Since $\operatorname{dist}(M_h(\lambda_i),U_{n+1/2}^{(i)}) \ll 1 (i=1,\ldots,q)$, we obtain from Proposition 3.5 that

$$U_{n+1}=\bigoplus_{i=1}^{q}U_{n+1/2}^{(i)}.$$

For $\psi\in\bigoplus_{i=1}^{q}M_h(\lambda_i)$ with $\|\psi\|=1$ and the orthonormal basis $\{v_{ij}^h\}_{j=1}^{d_i}$ of $M_h(\lambda_i)$ with $a(v_{ij}^h,v_{kl}^h)=\delta_{ik}\delta_{jl}$, there exists $\{\alpha_{ij}\}$ satisfying $\sum_{i=1}^{q}\sum_{j=1}^{d_i}\alpha_{ij}^2=1$ and $\psi=\sum_{i=1}^{q}\sum_{j=1}^{d_i}\alpha_{ij}v_{ij}^h$. It holds that

(A.6)

$$\begin{aligned}\operatorname{dist}(\psi,U_{n+1}) &= \left\|\left(\mathrm{I}-\mathcal{P}_{U_{n+1}}\right)\sum_{i=1}^{q}\sum_{j=1}^{d_i}\alpha_{ij}v_{ij}^h\right\| \\ &\leqslant \sum_{i=1}^{q}\sum_{j=1}^{d_i}|\alpha_{ij}|\left\|\left(\mathrm{I}-\mathcal{P}_{U_{n+1}}\right)v_{ij}^h\right\| \leqslant \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{d_i}\left\|\left(\mathrm{I}-\mathcal{P}_{U_{n+1}}\right)v_{ij}^h\right\|^2} \\ &= \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{d_i}\operatorname{dist}^2\left(v_{ij}^h,U_{n+1}\right)} \leqslant \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{d_i}\operatorname{dist}^2\left(M_h(\lambda_i),U_{n+1}\right)} \\ &\leqslant \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{d_i}\operatorname{dist}^2\left(M_h(\lambda_i),U_{n+1/2}^{(i)}\right)} \leqslant \sqrt{N}\max_{1\leqslant i\leqslant q}\operatorname{dist}(M_h(\lambda_i),U_{n+1/2}^{(i)}),\end{aligned}$$

and

$$\operatorname{dist}\left(\bigoplus_{i=1}^{q}M_h(\lambda_i),U_{n+1}\right) \leqslant \sqrt{N}\max_{1\leqslant i\leqslant q}\operatorname{dist}(M_h(\lambda_i),U_{n+1/2}^{(i)}) \ll 1. \tag{A.7}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

It is clear that U_{n+1} is a finite dimensional subspace of V^h. We apply Proposition 2.2 to (4.3) and obtain that $\lambda_{ij}^{(n+1)} \leqslant \lambda_{i+1,1}^h$ due to $\operatorname{dist}(\bigoplus_{i=1}^q M_h(\lambda_i), U_{n+1}) \ll 1$ for $(1,1) \leqslant (i,j) \leqslant (q, d_q)$. Hence, it follows from (A.7), Proposition 2.2, and Theorem 3.6 applied to (4.3) that

$$\begin{aligned}
|\lambda_{ij}^{(n+1)} - \lambda_{ij}^h| &\leqslant \lambda_{i+1,1}^h \operatorname{dist}^2\left(\bigoplus_{i=1}^q M_h(\lambda_i), U_{n+1}\right) \\
&\leqslant \lambda_{i+1,1}^h \sqrt{N} \max_{1\leqslant i\leqslant q} \operatorname{dist}^2(M_h(\lambda_i), U_{n+1/2}^{(i)}), \\
\operatorname{dist}(u_{ij}^{h,o}, u_{ij}^{(n+1)}) &\leqslant \tilde{C}_{**} \operatorname{dist}\left(\bigoplus_{i=1}^q M_h(\lambda_i), U_{n+1}\right) \\
&\leqslant \tilde{C}_{**} \sqrt{N} \max_{1\leqslant i\leqslant q} \operatorname{dist}(M_h(\lambda_i), U_{n+1/2}^{(i)}),
\end{aligned}$$

where the constant $\tilde{C}_{**}$ is independent of U_{n+1} and may depend on V^h (see Proposition 2.3 and Remark 2.4). ∎

A.4. Proof of Proposition 4.2.

Proof. For the n-th iteration, consider the linear operators

$$\mathcal{F}_n^{(i)} : U_n^{(i)} \to U_{n+1/2}^{(i)}, \quad i = 1, 2, \ldots, q,$$

and for $u^{(n)} \in U_n^{(i)}$, $u^{(n+1/2)} = \mathcal{F}_n^{(i)} u^{(n)}$ satisfying

$$a\left(u^{(n+1/2)}, v\right) - \bar{\lambda}_i b\left(u^{(n+1/2)}, v\right) = \bar{\lambda}_i b\left(u^{(n)}, v\right) \quad \forall v \in V^h.$$

We claim that $\mathcal{F}_n^{(i)}$ is an injection. Indeed, $u^{(n+1/2)} = v^{(n+1/2)}$ implies that

$$\begin{aligned}
\bar{\lambda}_i b\left(u^{(n)}, v\right) &= a\left(u^{(n+1/2)}, v\right) - \bar{\lambda}_i b\left(u^{(n+1/2)}, v\right) \\
&= a\left(v^{(n+1/2)}, v\right) - \bar{\lambda}_i b\left(v^{(n+1/2)}, v\right) = \bar{\lambda}_i b\left(v^{(n)}, v\right), \quad v \in V^h,
\end{aligned}$$

and $u^{(n)} = v^{(n)}$.

Since $U_n^{(i)}$ and $U_{n+1/2}^{(i)}$ are finite dimensional, $\mathcal{F}_n^{(i)}$ is indeed an isomorphism, and we arrive at

$$\dim\left(U_{n+1/2}^{(i)}\right) = \dim\left(U_n^{(i)}\right), \quad i = 1, 2, \ldots, q.$$ ∎

A.5. Proof of Theorem 4.3.

Proof. Let us consider $n = 0$ first.

Since $\{u_{ij}^h\}_{j=1}^{d_i}$ is the orthonormal basis of $M_h(\lambda_i)(i = 1, 2, \ldots, p)$ with $b(u_{ij}^h, u_{kl}^h) = \delta_{ik}\delta_{jl}$, for $v_i^{(0)} \in U_0^{(i)}$, there exists $\{\alpha_{rt}^{(i)}\}$ such that

$$v_i^{(0)} = \sum_{r=1}^p \sum_{t=1}^{d_r} \alpha_{rt}^{(i)} u_{rt}^h, \quad i = 1, 2, \ldots, q. \tag{A.8}$$

A simple calculation and the equation

$$a(\mathcal{P}_{M_h(\lambda_i)} v_i^{(0)}, v) = a(v_i^{(0)}, v), \quad v \in M_h(\lambda_i)$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

show that

$$\mathcal{P}_{M_h(\lambda_i)} v_i^{(0)} = \sum_{t=1}^{d_i} \alpha_{it}^{(i)} u_{it}^h.$$

We obtain from (2.2) and (4.6) that there exists $\varepsilon_0 \in (0,1)$ such that

$$\begin{aligned}\varepsilon_0 &\geqslant \operatorname{dist}\left(U_0^{(i)}, M_h(\lambda_i)\right) \geqslant \operatorname{dist}\left(v_i^{(0)}, M_h(\lambda_i)\right)\\ &= \operatorname{dist}\left(v_i^{(0)}, \mathcal{P}_{M_h(\lambda_i)} v_i^{(0)}\right) = \frac{\left\|\sum_{r=1}^p \sum_{t=1}^{d_r} \alpha_{rt}^{(i)} u_{rt}^h - \sum_{t=1}^{d_i} \alpha_{it}^{(i)} u_{it}^h\right\|}{\left\|\sum_{r=1}^p \sum_{t=1}^{d_r} \alpha_{rt}^{(i)} u_{rt}^h\right\|},\end{aligned}$$

which yields

$$\sum_{t=1}^{d_i} \lambda_{it}^h \left(\alpha_{it}^{(i)}\right)^2 \geqslant \left(\frac{1-\varepsilon_0^2}{\varepsilon_0^2}\right) \sum_{1\leqslant r\neq i\leqslant p} \sum_{t=1}^{d_r} \lambda_{rt}^h \left(\alpha_{rt}^{(i)}\right)^2, \quad i=1,2,\ldots,q. \tag{A.9}$$

Let $v_i^{(1/2)} \in V^h$ satisfy

$$a\left(v_i^{(1/2)}, v\right) - \bar{\lambda}_i b\left(v_i^{(1/2)}, v\right) = \bar{\lambda}_i b\left(v_i^{(0)}, v\right) \quad \forall v \in V^h.$$

We may write $v_i^{(1/2)} = \sum_{r=1}^p \sum_{t=1}^{d_r} \beta_{rt}^{(i)} u_{rt}^h$ for $i=1,2,\ldots,q$. Note that (2.3), (4.4), and (A.8) imply

$$\begin{aligned}\bar{\lambda}_i \sum_{r=1}^p \sum_{t=1}^{d_r} \alpha_{rt}^{(i)} b(u_{rt}^h, v) &= \bar{\lambda}_i b(v_i^{(0)}, v) = a\left(v_i^{(1/2)}, v\right) - \bar{\lambda}_i b\left(v_i^{(1/2)}, v\right)\\ &= \sum_{r=1}^p \sum_{t=1}^{d_r} \left(\lambda_{rt}^h - \bar{\lambda}_i\right) \beta_{rt}^{(i)} b(u_{rt}^h, v), \quad v \in V^h.\end{aligned}$$

We have

$$\beta_{rt}^{(i)} = \frac{\bar{\lambda}_i}{\lambda_{rt}^h - \bar{\lambda}_i} \alpha_{rt}^{(i)}, \quad i=1,2,\ldots,q,$$

and hence,

$$v_i^{(1/2)} = \sum_{r=1}^p \sum_{t=1}^{d_r} \frac{\bar{\lambda}_i}{\lambda_{rt}^h - \bar{\lambda}_i} \alpha_{rt}^{(i)} u_{rt}^h, \quad i=1,2,\ldots,q.$$

Note that (4.5) and (A.9) imply that

$$\left|\lambda_{rt}^h - \bar{\lambda}_i\right| \geqslant \left|\lambda_{rt}^h - \lambda_{i1}^h\right| - \left|\lambda_{i1}^h - \bar{\lambda}_i\right| \geqslant g - \delta_0, \quad r \neq i,$$

and

$$\begin{aligned}&\frac{\sum_{1\leqslant r\neq i\leqslant p} \sum_{t=1}^{d_r} \left(\frac{\bar{\lambda}_i}{\lambda_{rt}^h - \bar{\lambda}_i} \alpha_{rt}^{(i)}\right)^2 \lambda_{rt}^h}{\sum_{t=1}^{d_i} \left(\frac{\bar{\lambda}_i}{\lambda_{it}^h - \bar{\lambda}_i} \alpha_{it}^{(i)}\right)^2 \lambda_{it}^h}\\ &\leqslant \left(\frac{\delta_0}{g-\delta_0}\right)^2 \frac{\sum_{1\leqslant r\neq i\leqslant p} \sum_{t=1}^{d_r} \lambda_{rt}^h \left(\alpha_{rt}^{(i)}\right)^2}{\sum_{t=1}^{d_i} \lambda_{it}^h \left(\alpha_{it}^{(i)}\right)^2} \leqslant \left(\frac{\delta_0}{g-\delta_0}\right)^2 \frac{\varepsilon_0^2}{1-\varepsilon_0^2}.\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

We then get that

$$\begin{aligned}\operatorname{dist}\left(v_i^{(1/2)}, M_h(\lambda_i)\right) &= \frac{\left\|v_i^{(1/2)} - \mathcal{P}_{M_h(\lambda_i)} v_i^{(1/2)}\right\|}{\left\|v_i^{(1/2)}\right\|} \\ &= \sqrt{1 - \frac{\sum_{t=1}^{d_i}\left(\frac{\bar{\lambda}_i}{\lambda_{it}^h - \bar{\lambda}_i}\alpha_{it}^{(i)}\right)^2 \lambda_{it}^h}{\sum_{r=1}^{p}\sum_{t=1}^{d_r}\left(\frac{\bar{\lambda}_i}{\lambda_{rt}^h - \bar{\lambda}_i}\alpha_{rt}^{(i)}\right)^2 \lambda_{rt}^h}} \\ &\leqslant \frac{\delta_0\varepsilon_0}{\sqrt{(g-\delta_0)^2(1-\varepsilon_0^2) + \delta_0^2\varepsilon_0^2}} := \varepsilon_1, \quad i = 1, \ldots, q.\end{aligned} \tag{A.10}$$

We see that $v_i^{(1)} = \mathcal{F}_0^{(i)} v_i^{(0)}$, where $\mathcal{F}_0^{(i)}$ is an isomorphism. Then we obtain from (4.5) and (A.10) that $\varepsilon_1 \leqslant \varepsilon_0 < 1$ and

$$\operatorname{dist}\left(M_h(\lambda_i), U_{1/2}^{(i)}\right) \leqslant \varepsilon_1, \quad i = 1, 2, \ldots, q.$$

Similarly, we have

$$\operatorname{dist}\left(M_h(\lambda_i), U_{n+1/2}^{(i)}\right) \leqslant \varepsilon_{n+1} = \frac{\delta_0\varepsilon_n}{\sqrt{(g-\delta_0)^2(1-\varepsilon_n^2) + \delta_0^2\varepsilon_n^2}} \quad \forall n \in \mathbb{N}.$$

Thus, $\varepsilon_{n+1} \leqslant \varepsilon_n \forall n \in \mathbb{N}$ and

$$\begin{aligned}\varepsilon_{n+1} &= \frac{\delta_0\varepsilon_n}{\sqrt{(g-\delta_0)^2(1-\varepsilon_n^2) + \delta_0^2\varepsilon_n^2}} \leqslant \left(\frac{\delta_0}{\sqrt{(g-\delta_0)^2(1-\varepsilon_{n-1}^2) + \delta_0^2\varepsilon_{n-1}^2}}\right)^2 \varepsilon_{n-1} \\ &\leqslant \cdots \leqslant \left(\frac{\delta_0}{\sqrt{(g-\delta_0)^2(1-\varepsilon_0^2) + \delta_0^2\varepsilon_0^2}}\right)^n \varepsilon_0,\end{aligned}$$

which indicates that ε_{n+1} decreases toward 0 as $n \to \infty$. Moreover, there holds that

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n} = \frac{\delta_0}{g - \delta_0}. \tag{A.11}$$

∎

A.6. Proof of Theorem 4.4.

Proof. Let us start by $n = 0$.

Since $\{u_{ij}^{h,o,0}\}_{i=1}^{d_i}$ is an orthonormal basis of $M_h(\lambda_i)$, for $\varphi \in M_h(\lambda_i)$ with $\|\varphi\| = 1$, there exists $\{\alpha_j\}$ satisfying $\varphi = \sum_{j=1}^{d_i} \alpha_j u_{ij}^{h,o,0}$ and $\sum_{j=1}^{d_i} \alpha_j^2 = 1$. Note that

$$\begin{aligned}\operatorname{dist}\left(\varphi, U_0^{(i)}\right) &= \left\|\left(\mathrm{I} - \mathcal{P}_{U_0^{(i)}}\right)\varphi\right\| \leqslant \sum_{j=1}^{d_i} |\alpha_j| \left\|\left(\mathrm{I} - \mathcal{P}_{U_0^{(i)}}\right) u_{ij}^{h,o,0}\right\| \\ &= \sum_{j=1}^{d_i} |\alpha_j| \operatorname{dist}\left(u_{ij}^{h,o,0}, U_0^{(i)}\right) \leqslant \sum_{j=1}^{d_i} |\alpha_j| \operatorname{dist}\left(u_{ij}^{h,o,0}, u_{ij}^{(0)}\right) \leqslant \sqrt{d_i}\varepsilon_0.\end{aligned}$$

We arrive at

$$\operatorname{dist}\left(M_h(\lambda_i), U_0^{(i)}\right) \leqslant \sqrt{d_i}\varepsilon_0 \leqslant \sqrt{D}\varepsilon_0, \quad i = 1, \ldots, q. \tag{A.12}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

For $(1,1) \leqslant (i,j) \leqslant (q,d_q)$, we have from (4.10) that

$$\left|\lambda_{ij}^h - \bar{\lambda}_i\right| \leqslant \left|\lambda_{ij}^h - \mathcal{C}_i\left(\{\lambda_{ij}^h\}_{j=1}^{d_i}\right)\right| + \left|\mathcal{C}_i\left(\{\lambda_{ij}^h\}_{j=1}^{d_i}\right) - \bar{\lambda}_i\right| \leqslant \gamma + \zeta_0 < \frac{g}{2}.$$

Consider $U_{1/2}^{(i)} = \operatorname{span}\{u_{i1}^{(1/2)}, \ldots, u_{id_i}^{(1/2)}\}$ for $i = 1, \ldots, q$. In accordance with Theorem 4.3 and (A.12), there holds that

$$\operatorname{dist}\left(M_h(\lambda_i), U_{1/2}^{(i)}\right) \leqslant \frac{\sqrt{D}\,(\gamma+\zeta_0)\,\varepsilon_0}{\sqrt{(g-\gamma-\zeta_0)^2(1-D\varepsilon_0^2) + D\,(\gamma+\zeta_0)^2\,\varepsilon_0^2}}, \quad i = 1, \ldots, q.$$

Since ε_0 is sufficiently small, we obtain from Proposition 3.5 that $U_1 = \bigoplus_{i=1}^q U_{1/2}^{(i)}$, which together with Proposition 4.2 implies that

$$\dim(U_1) = \sum_{i=1}^q \dim(U_{1/2}^{(i)}) = \sum_{i=1}^q d_i = N.$$

Due to (A.7), we have

$$\operatorname{dist}\left(\bigoplus_{i=1}^q M_h(\lambda_i), U_1\right) \leqslant \frac{\sqrt{DN}\,(\gamma+\zeta_0)\,\varepsilon_0}{\sqrt{(g-\gamma-\zeta_0)^2(1-D\varepsilon_0^2) + D\,(\gamma+\zeta_0)^2\,\varepsilon_0^2}} := \xi_1.$$

We apply Theorem 4.1 to (4.8) and obtain that there exists an orthonormal basis $\{u_{ij}^{h,o,1}\}_{j=1}^{d_i}$ of $M_h(\lambda_i)$ with $b(u_{ij}^{h,o,1}, u_{kl}^{h,o,1}) = \delta_{ik}\delta_{jl}$ such that

$$\operatorname{dist}\left(u_{ij}^{h,o,1}, u_{ij}^{(1)}\right) \leqslant \tilde{C}_{**}\xi_1, \quad \lambda_{ij}^{(1)} - \lambda_{ij}^h \leqslant \lambda_{q+1,1}^h \xi_1^2, \quad (1,1) \leqslant (i,j) \leqslant (q,d_q), \tag{A.13}$$

where $\tilde{C}_{**}$ is a constant that is independent of U_1, i.e., independent of the iteration.

Set $\varepsilon_1 = \tilde{C}_{**}\xi_1$ and $\zeta_1 = \lambda_{q+1,1}^h \xi_1^2$; we then have

$$\left|\mathcal{C}_i\left(\{\lambda_{ij}^h\}_{j=1}^{d_i}\right) - \bar{\lambda}_i^{(1)}\right| = \left|\mathcal{C}_i\left(\left\{\lambda_{ij}^h - \lambda_{ij}^{(1)}\right\}_{j=1}^{d_i}\right)\right| \leqslant \zeta_1.$$

We obtain that $\varepsilon_1 \leqslant \varepsilon_0$ and $\zeta_1 \leqslant \zeta_0$ since ε_0, γ and ζ_0 are sufficiently small. Similarly, there hold that

$$\operatorname{dist}\left(\bigoplus_{i=1}^q M_h(\lambda_i), U_{n+1}\right) \leqslant \xi_{n+1} = \frac{\sqrt{DN}\,(\gamma+\zeta_n)\,\varepsilon_n}{\sqrt{(g-\gamma-\zeta_n)^2(1-D\varepsilon_n^2) + D\,(\gamma+\zeta_n)^2\,\varepsilon_n^2}},$$

$$0 \leqslant \lambda_{ij}^{(n+1)} - \lambda_{ij}^h \leqslant \zeta_{n+1} = \lambda_{q+1,1}^h \xi_{n+1}^2, \quad (1,1) \leqslant (i,j) \leqslant (q,d_q).$$

Therefore there exists an orthonormal basis $\{u_{ij}^{h,o,n+1}\}_{j=1}^{d_i}$ of $M_h(\lambda_i)$ such that

$$\operatorname{dist}\left(u_{ij}^{h,o,n+1}, u_{ij}^{(n+1)}\right) \leqslant \varepsilon_{n+1} = \tilde{C}_{**}\xi_{n+1} = \frac{\tilde{C}_{**}\sqrt{DN}\,(\gamma+\zeta_n)\,\varepsilon_n}{\sqrt{(g-\gamma-\zeta_n)^2(1-D\varepsilon_n^2) + D\,(\gamma+\zeta_n)^2\,\varepsilon_n^2}},$$

for $(1,1) \leqslant (i,j) \leqslant (q,d_q)$, where the constant $\tilde{C}_{**}$ is the same as the one in (A.13) due to the independence of iterations.

We see that $\varepsilon_{n+1} \leqslant \varepsilon_n$ and $\zeta_{n+1} \leqslant \zeta_n$, and both ε_n and ζ_n decrease toward 0 as $n \to \infty$. Finally, we arrive at

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n} = \frac{\tilde{C}_{**}\sqrt{DN}\gamma}{g-\gamma}. \qquad \square$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Acknowledgments. The authors would like to thank Prof. Zhaojun Bai, and the anonymous referees for their constructive suggestions for the enhancement of the conclusions in section 3 and the improvement of the presentation.

REFERENCES

[1] *Quantum ESPRESSO*, http://www.quantum-espresso.org/, 2024.
[2] R. A. Adams and J. J. Fournier, *Sobolev Spaces*, Elsevier, New York, 2003.
[3] P. Arbenz, D. Kressner, and D. Zürich, *Lecture notes on solving large scale eigenvalue problems,* D-MATH, ETH Zurich, 2 (2012), 377.
[4] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1996.
[5] I. Babuška and J. E. Osborn, *Finite element-Galerkin approximation of the eigenvalues and eigenvectors of selfadjoint problems*, Math. Comp., 52 (1989), pp. 275–297, https://doi.org/10.1090/S0025-5718-1989-0962210-8.
[6] Z. Bai, J. Yin, and Y. Su, *A shift-splitting preconditioner for non-Hermitian positive definite matrices*, J. Comput. Math., 24 (2006), pp. 539–552, https://www.jstor.org/stable/43693314.
[7] Z. Bai and S. Zhang, *A regularized conjugate gradient method for symmetric positive definite system of linear equations*, J. Comput. Math., 20 (2002), pp. 437–448, https://www.jstor.org/stable/43693008.
[8] F. Chatelin, *Spectral Approximation of Linear Operators*, SIAM, Philadelphia, 2011.
[9] X. Dai, X. Gong, Z. Yang, D. Zhang, and A. Zhou, *Finite volume discretizations for eigenvalue problems with applications to electronic structure calculations*, Multiscale Model. Simul., 9 (2011), pp. 208–240, https://doi.org/10.1137/090757046.
[10] X. Dai, X. Gong, A. Zhou, and J. Zhu, *A Parallel Orbital-Updating Approach for Electronic Structure Calculations*, preprint, arXiv:1405.0260, 2014.
[11] X. Dai, L. He, and A. Zhou, *Convergence and quasi-optimal complexity of adaptive finite element computations for multiple eigenvalues*, IMA J. Numer. Anal., 35 (2015), pp. 1934–1977, https://doi.org/10.1093/imanum/dru059.
[12] X. Dai, Z. Liu, X. Zhang, and A. Zhou, *A parallel orbital-updating based optimization method for electronic structure calculations*, J. Comput. Phys., 445 (2021), 110622, https://doi.org/10.1016/j.jcp.2021.110622.
[13] X. Dai, J. Xu, and A. Zhou, *Convergence and optimal complexity of adaptive finite element eigenvalue computations*, Numer. Math., 110 (2008), pp. 313–355, https://doi.org/10.1007/s00211-008-0169-3.
[14] E. G. D'yakonov, *Optimization in Solving Elliptic Problems*, CRC Press, Boca Raton, FL, 2018.
[15] S. Hazra, U. Patil, and S. Sanvito, *Predicting the one-particle density matrix with machine learning*, J. Chem. Theory Comput., 20 (2024), pp. 4569–4578, https://doi.org/10.1021/acs.jctc.4c00042.
[16] T. Kato, *Perturbation Theory for Linear Operators*, Classics Math. 132, Springer Science & Business Media, New York, 2013.
[17] E. Kaxiras, *Atomic and Electronic Structure of Solids*, Cambridge University Press, London, 2003.
[18] A. Knyazev, *Sharp a priori error estimates of the Rayleigh-Ritz method without assumptions of fixed sign or compactness*, Math. Notes Acad. Sci. USSR, 38 (1985), pp. 998–1002, https://doi.org/10.1007/BF01157020.
[19] A. V. Knyazev and J. E. Osborn, *New a priori FEM error estimates for eigenvalues*, SIAM J. Numer. Anal., 43 (2006), pp. 2647–2667, https://doi.org/10.1137/040613044.
[20] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140 (1965), pp. A1133–A1138, https://doi.org/10.1103/PhysRev.140.A1133.
[21] G. Kresse and J. Furthmüller, *Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set*, Phys. Rev. B, 54 (1996), pp. 11169–11186, https://doi.org/10.1103/PhysRevB.54.11169.
[22] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, London, 2020.
[23] M. J. Oliveira, N. Papior, Y. Pouillon, V. Blum, E. Artacho, D. Caliste, F. Corsetti, S. De Gironcoli, A. M. Elena, A. García, et al., *The CECAM electronic structure library and the modular software development paradigm*, J. Chem. Phys., 153 (2020), 024117, https://doi.org/10.1063/5.0012901.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[24] Y. Pan, X. Dai, S. de Gironcoli, X.-G. Gong, G.-M. Rignanese, and A. Zhou, *A parallel orbital-updating based plane-wave basis method for electronic structure calculations*, J. Comput. Phys., 348 (2017), pp. 482–492, https://doi.org/10.1016/j.jcp.2017.07.033.

[25] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998.

[26] M. C. Payne, M. P. Teter, D. C. Allan, T. Arias, and A. J Joannopoulos, *Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate gradients*, Rev. Mod. Phys., 64 (1992), pp. 1045–1097, https://doi.org/10.1103/RevModPhys.64.1045.

[27] E. Polizzi, *Density-matrix-based algorithm for solving eigenvalue problems*, Phys. Rev. B, 79 (2009), 115112, https://doi.org/10.1103/PhysRevB.79.115112.

[28] J. D. Riley, *Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix*, Math. Tables Other Aids Comput., 110 (1955), pp. 96–101, https://doi.org/10.2307/2002065.

[29] P. J. Rousseeuw, *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, J. Comput. Appl. Math., 20 (1987), pp. 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.

[30] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2003.

[31] Y. Saad, *Analysis of subspace iteration for eigenvalue problems with evolving matrices*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 103–122, https://doi.org/10.1137/141002037.

[32] T. Sakurai and H. Sugiura, *A projection method for generalized eigenvalue problems using numerical integration*, J. Comput. Appl. Math., 159 (2003), pp. 119–128, https://doi.org/10.1016/S0377-0427(03)00565-X.

[33] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat., 6 (1978), pp. 461–464, https://doi.org/10.1214/aos/1176344136.

[34] Z. Xu and Z. Sheng, *Subspace method based on neural networks for solving the partial differential equation*, Comput. Math. Appl., 195 (2025), pp. 109–138, https://doi.org/10.1016/j.camwa.2025.07.001.

Downloaded 08/22/25 to 39.157.44.76 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 廖天浩

方向： 人工智能

单位： 中国科学院大学前沿交叉科学学院

题目： In2O3基晶体管的热稳定性作用机理与结构调控研究

In_2O_3基晶体管的热稳定性作用机理与结构调控研究

廖天浩[1,2,3]，尤楠楠[2,3,*]，王嘉义[2,3]，徐杨[2,3]，张括[2,3]，胡顺东[2,3]，耿玓[2,3]，杨冠华[2,3]，李泠[2,3]，王盛凯[2,3,*]

1. 前沿交叉科学学院，中国科学院大学，北京 10049，中国
2. 集成电路制造技术全国重点实验室，中国科学院，北京 100029，中国
3. 中国科学院微电子研究所，北京 100029，中国

摘要

面向“后摩尔时代”三维集成（3D Stacking）对后道工艺（BEOL）晶体管的迫切需求，传统硅基技术因其高温工艺（>800 ºC）、单管漏电高导致功耗大而无法兼容先进三维阵列。In_2O_3基晶体管凭借其低温制备（<400 ºC）和高迁移率、低漏电等优异电学性能，成为实现 3D DRAM 和存算一体等未来技术的关键候选。然而，其热稳定性差（导致器件可靠性低）是目前阻碍其产业化应用的核心瓶颈。

本研究深入分析发现，In_2O_3的低热稳定性主要源于两大因素的竞争性影响：一是工艺加热导致的界面态缺陷（如氧空位），其俘获电子导致阈值电压正向漂移；二是工艺中不可避免引入的 H 杂质迁移，其作为浅施主提供额外电子，导致阈值电压负向漂移。这两种相反的漂移机制相互竞争、此消彼长，使得器件在不同工艺和电热应力下的行为难以预测和控制，极大压缩了制造工艺窗口。

本研究工作针对这一关键科学问题，拟从创新的结构调控出发，系统研究其作用机理并提出协同解决方案。主要研究内容包括：栅氧结构协同设计：突破传统插层技术以牺牲 EOT 来换取稳定性的局限，通过协同优化栅氧体层（高 k）与插层（高稳定性），实现界面态抑制与强栅控（即低 EOT）的统一。主动排 H 结构：针对内源 H 机理不清的“黑盒”问题，提出一种基于 H 亲和金属（如 Pt/Pd）的新型器件结构，主动萃取并固定沟道中的 H，实现对 H 浓度影响的定量分析，以理清其作用机理。被动阻 H 钝化层：针对 BEOL 后续工艺中的外源 H 污染，设计一种利用多层膜界面电场与应力的新型钝化层，高效阻挡外源 H 扩散。

本研究工作从界面和杂质两个根源上，努力阐明并解耦In_2O_3热不稳定的竞争机制，提出一套完整的高性能、高可靠性器件结构设计方案，扫清其在 3D 堆叠工艺中应用的核心障碍，为下一代高性能存储芯片的实现奠定器件基础。

一．研究介绍

1.1 算力危机与三维集成范式

进入人工智能与万物互联时代，数据处理需求呈指数级增长，但传统硅基芯片面临物理极限与冯·诺依曼架构下“内存墙”与“功耗墙”的双重桎梏。为了突破这一瓶颈，从二维平面向三维垂直堆叠的范式转移已成为必然。单片三维集成技术通过在底层逻辑电路之上直接堆叠存储或计算单元，能够极大缩短互连长度，实现高带宽、低功耗的数据传输。

1.2 关键挑战：热预算与材料选择

三维集成的核心工艺挑战在于“热预算”。为了保护底层精密的硅基电路，顶层器件的制备温度必须严格控制在 400 ºC 以内（BEOL 兼容）。以 In_2O_3 为代表的氧化物半导体凭借其低温制备工艺（ALD/溅射）、高电子迁移率及极低的关态漏电流（fA 级），成为 3D 堆叠器件的首选材料。

1.3 科学问题：热稳定性失效机制

尽管潜力巨大，In_2O_3 器件在三维集成的多重热循环及封装工艺中，常表现出严重的热不稳定性。本研究明确指出，这种不稳定性并非单一因素所致，而是源于两大核心机制的动态博弈：界面态缺陷：In_2O_3 介质界面处的氧空位 O_v 在正偏压（PBS）下俘获电子，导致阈值电压 V_{th} 正向漂移。氢杂质动力学：氢作为浅施主，在热应力（PBTS）下向沟道迁移并提供电子，导致 V_{th} 负向漂移。这两种机制的竞争使得器件在不同工艺条件下表现出不可预测的漂移行为，本研究旨在通过结构设计精确调控这一物理过程。

二．国内外发展现状

2.1 针对界面态调控的研究现状与局限性

栅氧/沟道界面的质量直接决定了晶体管的电学特性和稳定性。在 In_2O_3 基晶体管中，界面态缺陷（尤其是氧空位,O_v）被认为是导致阈值电压（V_{th}）在正偏压下发生正漂的主要原因。目前，针对界面态的调控主要有两大技术路径：

第一类方法是通过化学手段钝化或补偿界面缺陷。元素掺杂：一大研究热点是在沟道或栅氧中引入 F、N 等非金属元素。例如，Fan 等人于 2018 年在 Materials 上证明 F 元素可以有效取代氧空位，并形成更强的 In-F 键，从而减少缺陷态密度，提升器件稳定性。气氛退火：采用高压氧气、N_2O 等离子体或 O_2 退火等方式，被用于在工艺过程中修复氧空位，如 Fang 等人 2025 年于 IEEE TED 中利用 O_2 改善界面质量。这类方法虽然在一定程度上有效，但也带来了新的工艺挑战。首先，掺杂工艺（如离子注入或 ALD 原位掺杂）会显著复杂化 BEOL 工艺流程，并可能引入新的杂质或损伤，导致迁移率下降。其次，高温或高能的退火过程与 BEOL 的低温热预算（<400°C）相悖，限制了其在 3D 堆叠中的应用。

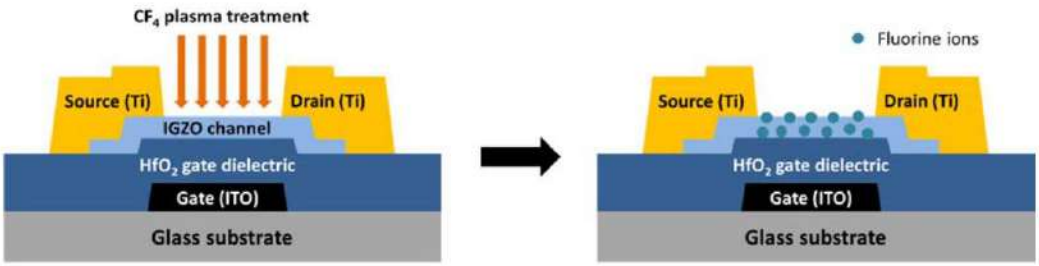


图 2.1 F 掺杂 IGZO TFT 机理的示意图

第二类方法是在 In_2O_3 沟道和主栅氧（通常为 SiO_2）之间，插入一层薄的、高稳定性的介电层，如 Al_2O_3、HfO_2、La_2O_3、GaO 等，如 Shiah 等人于 2021 年于 nature electronics 上利用 IGO 作为 ITZO 插层提升偏压可靠性。其主要思路是利用这些高稳定性材料与 In_2O_3 之间形成的更优异、更稳定的界面，来“隔离”沟道免受主栅氧缺陷的影响。这种“三明治”结构的栅氧栈虽然能改善界面，但其核心代价是牺牲了器件性能。由于插层的引入，整个栅氧层的等效氧化层厚度（EOT）不可避免地增加。根据晶体管缩比定律，EOT 的增加会直接削弱栅极对沟道的控制能力，导致亚阈值摆幅（SS）变差、栅电容降低，并最终导致器件的开态电流和驱动能力下降。这与 3D 堆叠追求高性能的初衷相矛盾。

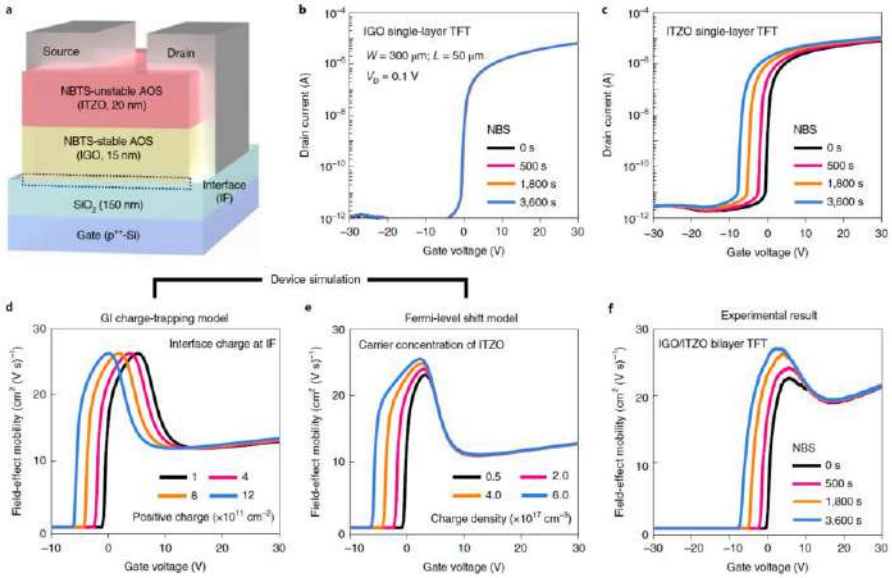


图 2.2 引入高稳定插层 IGO 的器件示意图以及对性能的影响

2.2 针对界面态调控的研究现状与局限性

H 杂质被认为是导致 In_2O_3 基晶体管 V_{th} 在正偏压热应力（PBTS）下发生负漂的最主要原因。H 在 In_2O_3 中通常表现为浅施主，提供额外电子。在 3D 堆叠工艺中，H 的来源广泛（如 ALD/PECVD 的前驱物 H_2O、等离子体中的 H、封装材料等），因此 H 的调控是热稳定性研究的重中之重。

第一类方法主要是针对内源 H，内源 H 是指在沟道和栅氧薄膜沉积过程中，不可避免地引入到材料内部的 H 杂质。机理分析：大量研究集中于分析内源 H 对器件性能的影响，如 Kong 等人于 2022 年在 IEDM 上对 H 机制进行了详细剖析。学术界普遍认同 H 在 In_2O_3 中扮演施主角色，但在加热和偏压下，H 的激活、迁移和与氧空位（O_v）的相互作用机制（例如形成 O_v-H 复合体）极其复杂。工艺调控：部分研究尝试通过优化沉积条件（如使用“无水”前驱物）或增加特定退火步骤来减少薄膜中的初始 H 含量。尽管研究众多，但目前学界对于 H 的具体物性（是以 H^+、H^0 还是 H^-形态存在）、在材料中的浓度阈值、以及其在能带结构中的确切位置对于 In_2O_3 半导体稳定性的具体影响，尚未形成清晰透彻的统一认识。由于缺乏主动、可控地调控器件内部 H 含量的实验手段，导致 H 的作用机理研究仍处于“黑盒”阶段，难以指导工艺进行针对性优化。

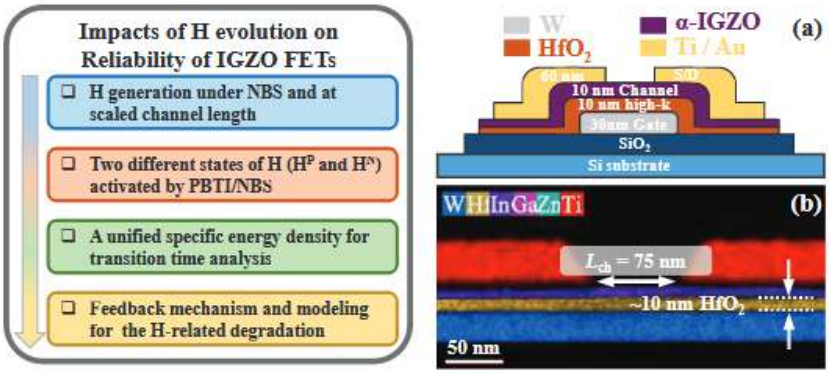


图 2.3 内源 H 对于 IGZO 器件影响作用的机理

第二类是针对外源 H，外源 H 是指在器件制造完成后，在后续的多层堆叠和封装工艺中，从外部环境（尤其是 PECVD 沉积 SiN_x 或 SiO_2 钝化层时）扩散进入器件有源区的 H 杂质，如 Wang 等人于 2020 年于 IEEE Access 上对 ITZO 器件进行了 SAM/Al_2O_3 钝化。钝化层阻隔：为杜绝外源 H 的扩散，目前主流的方法是在器件有源区顶部沉积一层致密的钝化层，如高密度 SiN_x、Al_2O_3 或多层膜结构。工艺优化：另有研究通过优化 PECVD 工艺参数（如降低 H 源气体比例）来减少钝化层本身的 H 含量，试图降低 H 向下的扩散梯度。在复杂的 3D 堆叠工艺中，器件需要经受多次、长时间的 BEOL 热循环（如~400°C 退火）。目前暂时还没有一种被广泛验证的、能够高效且持久阻挡 H 扩散的钝化层方案。单一的钝化层（如 SiN_x）在高温下仍难以完全阻止 H 的渗透，导致器件在完成多层堆叠后，其稳定性相比初始状态仍会发生显著退化。

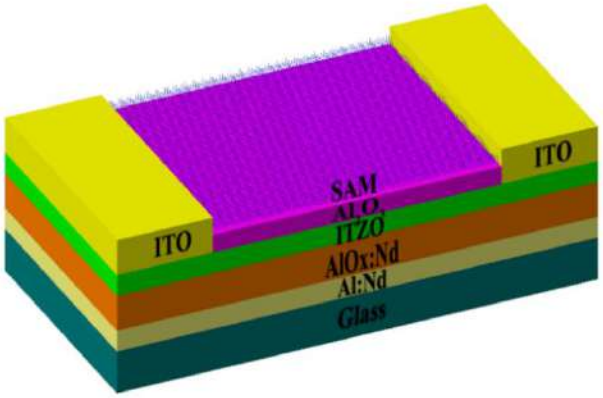


图 2.4 沉积双层钝化 SAM/Al_2O_3 的 ITZO 器件示意图

三．研究进展

本研究围绕“界面态调控”、“内源氢管理”与“外源氢阻隔”三个维度展开，取得了以下阶段性进展与理论突破：

3.1 界面态调控：高 k 复合栅氧栈的协同设计

栅氧体层（Bulk）优选：对比研究不同高 k 介电材料，如 HfO_2,ZrO_2,Al_2O_3 等作为栅氧体层的热稳定性、介电常数及漏电特性。筛选出能承受 BEOL 工艺（~400 ^{o}C）高热预算，并能提供低 EOT、高 k 值的最优体层材料。栅氧插层优化：在最优体层的基础上，研究不同薄层插层（如 Al_2O_3,La_2O_3 等）对 In_2O_3/栅氧界面的调控机制。重点分析插层对界面缺陷态密度（D_{it}）、氧空位（O_v）的钝

化效果，以及其对热过程中元素（如 In, O）互扩散的抑制能力。“体层-插层”协同设计：通过对体层和插层的协同优化，设计一种兼具高热稳定性、高 k 值和低界面缺陷的“复合栅氧栈”结构，打破器件可靠性与高性能（低 EOT）之间的传统制约关系。

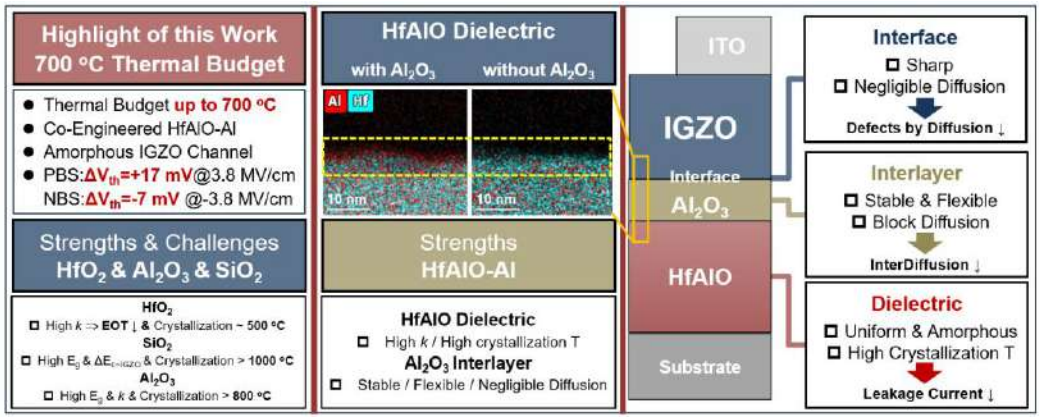


图 3.1 栅氧协同优化示意图

3.2 内源 H 调控：金属主动排 H 结构机制

提出并设计一种基于 H 亲和金属（如 Pt, Pd 等）的“主动排 H”器件结构。利用这类金属对 H 的强吸附和催化特性，在特定工艺节点（如退火）下，主动将 In_2O_3 沟道及附近介电层中的 H 杂质“泵出”并固定。H 物性与浓度表征：通过二次离子质谱（SIMS）、热脱附谱（TDS）等手段，定量对比分析排 H 前后器件内部 H 的浓度、空间分布及化学形态。H 作用机理分析：结合器件电学特性（如 V_{th} 漂移、源漏电阻变化）与材料表征结果，深入解析 H 杂质的浓度、电性（施主特性）及其在能带中的具体位置变化，阐明内源 H 对器件性能和稳定性的物理化学作用机制，并解耦其与界面态的竞争关系。

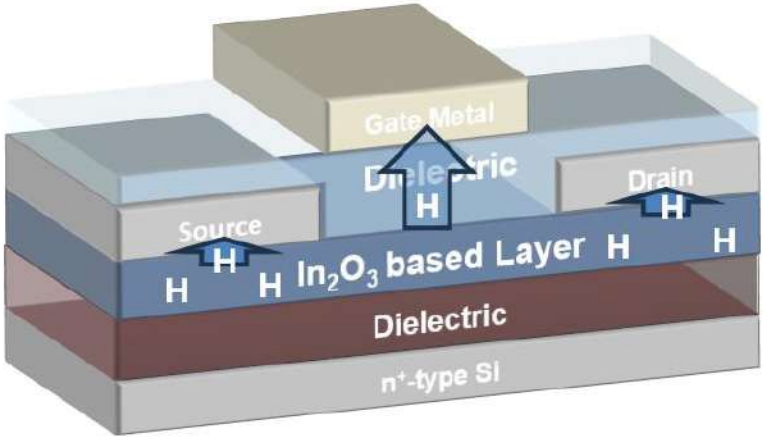


图 3.2 内源排 H 机理示意图

3.3 外源 H 阻隔：界面张力与静电势垒阻 H 结构

跳出传统单一材料钝化的思路，设计一种基于多层膜界面特性的新型“被动阻 H”钝化层结构。阻 H 机理探索：重点研究利用不同材料（如 Al_2O_3/SiN_x）界面处形成的局域电场（利用界面电荷）对 H+等带电离子的排斥力，以及利用高密度界面和晶格失配产生的界面应力场对 H 原子的物理阻挡效应。阻 H 效能验证：通过模拟重度 H 污染的工艺环境（如 H_2 等离子体处理或高温 H_2/N_2 退火），验证所设计的“被动阻 H 层”在严苛 BEOL 热循环下对外源 H 扩散的抑制效率。

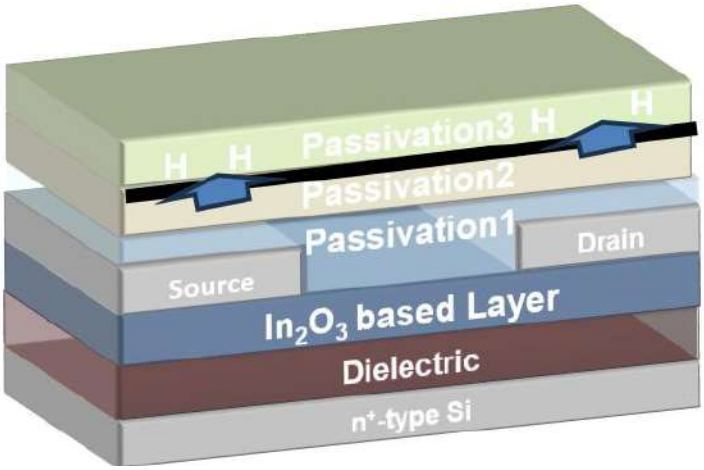


图 3.3 外源阻 H 机理示意图

四．研究结论与阶段性成果总结

本研究针对后摩尔时代单片三维集成中 In_2O_3 基晶体管的热稳定性瓶颈，系统阐明了失效物理机制并验证了结构调控方案的有效性：机制层面：确认了 In_2O_3 器件的热不稳定性本质上是界面态电子俘获（正漂机制）与氢杂质施主掺杂（负漂机制）的竞争平衡。任何单一维度的优化都可能打破平衡导致新的失效，必须进行协同调控。结构层面：确立了体层/插层的复合栅氧结构，在不牺牲 EOT 的前提下显著提升了界面热稳定性。验证了 Pd 金属层的主动吸氢效应，为解决低温工艺下难去除深层氢杂质提供了创新路径，同步实现了高可靠性与低接触电阻。提出了阻 H 钝化方案，有效解决了 3D 集成后续工艺中的外源氢污染问题。

在阶段性成果方面，本人目前已参与发表国际会议论文 6 篇，其中第一作者 4 篇，在投 IEEE EDL 期刊 1 篇返修中，以学生第一发明人申请专利 4 项，在工艺基础、器件制备以及表征分析能力方面已经有了初步的进展。

参考文献

[1] Q. Kong et al., "New Insights into the Impact of Hydrogen Evolution on the Reliability of IGZO FETs: Experiment and Modeling," IEDM, 30.2.1-30.2.4, 2022, doi: 10.1109/IEDM45625.2022.10019394.

[2] C. Fan et al., “Electrical Performance and Reliability Improvement of Amorphous-Indium-Gallium-Zinc-Oxide Thin-Film Transistors with HfO2 Gate Dielectrics by CF4 Plasma Treatment,” Materials, 1996-1944, 2018, doi: 10.3390/ma11050824.

[3] Y. Shiah et al., “Mobility-stability trade-off in oxide thin-film transistors,” Nature Electronics, 800-807, VOL 4, 2021, doi: 10.1038/s41928-021-00671-0.

[4] W. Zhong et al., “Self-Assembled Monolayers SAMs Al2O3 Double Layer Passivated InSnZnO Thin-Film Transistor,” IEEE Access, Vol. 8, 101834-101839, 2020, doi: 10.1109/ACCESS.2020.2997915.

[5] T. Kim, C. Choi, J, Hur, D. Ha, B. Kuh, Y. Kim, M. Cho, S. Kim, J. Jeong, "Progress, Challenges, and Opportunities in Oxide Semiconductor Devices: A Key Building Block for Applications Ranging from Display Backplanes to 3D Integrated Semiconductor Chips," Advanced Materials, vol. 35, no. 43, pp. 2204663, Oct. 2023, doi: 10.1002/adma.202204663.

[6] C. Ye, J. Li, P. Hong, J. Zhao, X. Miao, X. Li, "High-Performance Atomic-Layer-Deposited Dual-Gate InGaO Thin-Film Transistors," Nano Letters, vol. 25, no. 21, pp. 8541-8546, May 2025, doi: 10.1021/acs.nanolett.5c01059.

[7] Z. Zhang, Z. Lin, A. Charnas, H. Dou, Z. Shang, J. Zhang, M. Si, H. Wang, M. A. Alam, P. D. Ye, " Reliability of Atomic-Layer-Deposited Gate-All-Around In2O3 Nano-Ribbon Transistors with Ultra-High Drain Currents," 2022 IEEE International Electron Devices Meeting (IEDM), 30.3.1-30.3.4, Dec. 2022, doi: 10.1109/IEDM45625.2022.10019494.

[8] K. Chen, X. Duan, G. Yang, C. Lu, D. Geng, L. Li, M. Liu, " Inter-Layer Dielectric Engineering for Monolithic Stacking 4F2 -2T0C DRAM with Channel-All-Around (CAA) IGZO FET to Achieve Good Reliability (>104 s Bias Stress, >1012 Cycles Endurance)," 2022 IEEE International Electron Devices Meeting (IEDM), 26.5.1-26.5.4, Dec. 2022, doi: 10.1109/IEDM45625.2022.10019502.

[9] M. Mativenga, F. Haque, J. Um, , A. Siddik, " Impact of Source-to-Gate and Drain-to-Gate Overlap Lengths on Performance of Inverted Staggered a-IGZO TFTs With an Etch Stopper," IEEE Transactions on Electron Devices, vol. 67, no. 8, pp. 3152-3156, Aug. 2020, doi: 10.1109/TED.2020.3004115.

[10] Y. Tai, L, Chou, H. Chiu, B. Chen, " Three-Transistor AMOLED Pixel Circuit With Threshold Voltage Compensation Function Using Dual-Gate IGZO TFT," IEEE Electron Device Letters, vol. 33, no. 3, pp. 393-395, Mar. 2012, doi: 10.1109/LED.2011.2179282.

[11] M. Alip, A. Abliz, D. Wan, " Design of High-Performance Dual-Channel-Layered InGaZnO Thin-Film Transistors With Different Indium Contents," IEEE Transactions on Electron Devices, vol. 72, no. 4, pp. 1802-1808, Apr. 2025, doi: 10.1109/TED.2025.3538524.

[12] P. Nurmamat and A. Abliz, " Rational Design of Different Ga Content Bilayer InGaZnO Thin-Film Transistors With Al_2O_3/HfO_2 Passivation Layer," IEEE Transactions on Electron Devices, vol. 71, no. 5, pp. 3032-3038, May 2024, doi: 10.1109/TED.2024.3383429.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 卢宇鹏

方向： 集成电路

单位： 中国科学院大学前沿交叉科学学院

题目： 抑制IGZO可变电容提升2T0C DRAM稳定性

This article has been accepted for publication in IEEE Electron Device Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LED.2025.3616974

Improved Variability of 2T0C DRAM via Specific Post-Treatment Suppressing Variable IGZO Capacitance

Yupeng Lu, Gangping Yan, Yanyu Yang, Jie Luo, Yunjiao Bao, Longge Mao, Xueli Ma, Jinjuan Xiang, Gaobo Xu, Guilei Wang, Chao Zhao, and Huaxiang Yin

***Abstract*—The voltage drops on the storage node (V_{SN}) determined by the write transistor during data writing and keeping become the major challenge of stable operations in 2T0C DRAM even after optimizing the property uniformity in the read transistor. In this work, we unveil the key role of variable IGZO MOS capacitances on the memory stability based on a 2T0C DRAM scheme with 10-ks retention. By the specific oxygen compensated hydrogen self-adaptive-doping method, the total V_{SN} drop and variation decreases by 22% and 75%, respectively. Benefitted by the uniform distribution of donor defect and interface state in the write transistor, we demonstrate a record-low V_{SN} drop variation <20 mV and ultra-stable read characteristics of 2T0C DRAM. These results pave the way for the large-scale integration of stable multibit DRAM arrays in future.**

***Index Terms*— Storage node voltage, 2T0C DRAM, a-IGZO thin-film transistor (TFT), memory stability, multibit operation.**

I. Introduction

The growing demand for high-density, energy-efficient memory solutions in artificial intelligence (AI) and edge computing has intensified research into next-generation 3D dynamic random-access memory (DRAM). The emerging n-type indium-gallium-zinc-oxide (IGZO) capacitor-less 2T0C-DRAM has become a promising alternative due to its superior retention, 3D integration capabilities [1] [2] [3] [4] [5], and especially multibit operations [6] [7] [8]. However, the charge representing the information, which is stored in the storage node (SN), is capacitively coupled to both the word line (WL) and bit line (BL). During write and read operations, the storage charge is easily redistributed, and results in significant drop on the SN voltage ($V_{SN\text{-}drop}$) along with inferior memory stability. Previous work has focused on V_{SN} pre-charge by additional setup of read WL and BL voltage (V_{RWL},V_{RBL}) to mitigate the V_{SN} drop [9] [10] [11] [12]. In addition, the p-type read transistor (CNT or TeO_x channel) is adopted due to the opposite voltage condition of write and read operations recently [13] [14], so that the V_{SN} drop can be compensated. Nevertheless, the existed efforts are designed based on constant capacitance effect, and the coupling between IGZO channel and dielectric has not been considered. Hence, the variations of V_{SN} drop have not yet been revealed, which can cause drastic data crosstalk and multibit failure. It is urgent to have a deeper understanding and feasible solutions for V_{SN} drop fluctuations.

This work is supported by National Key R&D Program of China (Grant No. 2022YFB3606900), BJSAMT Project(SAMT-ZK-KT-22030102), Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA0330302), and NSFC program (62304247). (Corresponding authors: Gangping Yan, Gaobo Xu, and Huaxiang Yin)

Yupeng Lu, Longge Mao, Gaobo Xu and Huaxiang Yin with the School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing 101408, China. Key Laboratory of Fabrication Technologies for Integrated Circuits, Chinese Academy of Sciences, Beijing 100029, and with the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China. (e-mail: xugaobo@ime.ac.cn; yinhuaxiang@ime.ac.cn).

Yanyu Yang, Jie Luo, Yunjiao Bao with the Key Laboratory of Fabrication Technologies for Integrated Circuits, Chinese Academy of Sciences, Beijing 100029, and with the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China.

Gangping Yan, Xueli Ma, Jinjuan Xiang, Guilei Wang, Chao Zhao, and are with the Beijing Superstring Academy of Memory Technology, Beijing, 100176, China (e-mail: gangping.yan@bjsamt.org.cn).

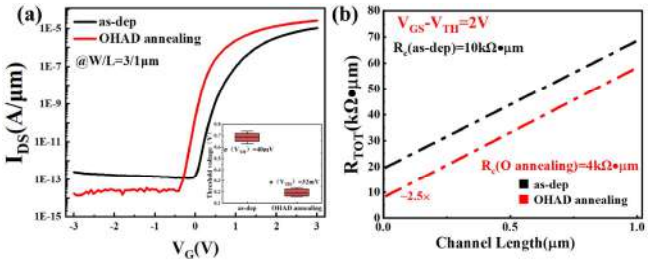


Fig. 1. (a) Transfer characteristics of a IGZO transistor with L_{CH} of 1μm and the variation of V_{TH}. (b) R_C of the IGZO transistor before and after OHAD calculated using the transfer length method (TLM).

In this work, the crucial effect of variable IGZO MOS capacitance on the V_{SN} drop and its variation in IGZO 2T0C DRAM is studied, including two processes of the C-V slope affected by interface states and the total capacitance modulated by IGZO donor defects. Based on the specific attainment of uniform defect distribution, the ultra-stable read characteristics and record-low V_{SN} drop variation <20 mV with a retention of 10ks are experimentally demonstrated in IGZO-based 2T0C DRAM cells. These results establish the foundation for future large-scale multi-bit DRAM arrays with enhanced stability.

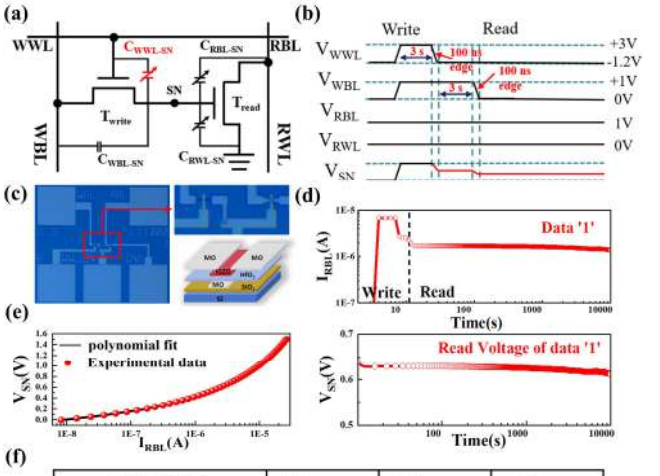
II. Experiment

The 2T0C DRAM is fabricated utilizing back-gate IGZO transistors. The general fabrication process for the back-gate transistor is as follows. 60-nm Mo was deposited as gate metal and source/drain (S/D) contacts. 15-nm HfO_2 and 20-nm IGZO were deposited as the GI and channel, respectively. The write and read transistors of the DRAM were designed with different dimensions: a width-to-length ratio (W/L) of 0.5 μm / 1 μm for the write transistor and 3 μm / 1 μm for the read transistor. The gate-to-source/drain overlap length was set to 0.5 μm, which is related to the parasitic capacitors and consistent with our previous work [15]. Finally, we setup 4 sets of oxygen-compensated hydrogen self-adaptive-doping (OHAD)

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on November 19,2025 at 07:12:16 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Electron Device Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LED.2025.3616974

conditions of 30mins (ST) and 50mins (NT) at 250°C, 30mins (ST) and 50mins (NT) at 300°C. **Fig. 1(a)** illustrates the transfer characteristics of the back-gate IGZO transistor, and improved V_{TH} distribution is obtained. The R_C calculated using the transfer length method (TLM), as shown in **Fig. 1(b)**. The reduction in off-state current and contact resistance indicates that the source-drain region is more heavily doped while the defects in channel region are well compensated after OHAD treatment [15] [16]. **Fig. 2(c)** shows the schematic structure of the IGZO transistor and a photo image of a 2T0C DRAM. All electrical tests were done with a semiconductor analyzer



Capacitor	C_{WWL-SN}	C_{RWL-SN}	C_{RBL-SN}
Area(μm×μm)	0.5×0.5	0.5×3	0.5×3

Fig. 2. (a) 2T0C DRAM circuit diagram considering the capacitive coupling of variable IGZO MOS capacitors. (b) Timing diagram for 2T0C DRAM during write and read operations. (c) A photo image of a 2T0C DRAM and the schematic diagram of a IGZO transistor. (d) Read current and corresponding storage node voltage as a function of time for the DRAM cell. (e) Polynomial fitting of V_{SN} vs. I_{RBL}.

B1500A at room temperature.

III. Results and Discussion

Fig. 2(a) presents the circuit diagram of a 2T0C DRAM cell including variable IGZO MOS capacitors. **Fig. 2(b)** exhibits the timing diagram of the memory cell during write and read operation. The write hold time is 6s for WBL and 3s for WWL. After a 100ns falling edge WBL pulse, the DRAM cell initiates the read operation. V_{RBL} and V_{RWL} is held at 1V and 0V, respectively, to monitor the I_{RBL} and V_{SN}. It is noted that the V_{SN} changes after write word line (WWL) and write bit line (WBL) voltage transition (ΔV_{SN-WWL} and ΔV_{SN-WBL}) during write operations. The rationale for sequentially lowering the potentials of the WWL and the WBL was to better differentiate the impact of their respective coupling capacitances. The polynomial fitting curve of I_{RBL} and V_{SN} of the read transistor in the DRAM cell is depicted in **Fig. 2(e)**, where a large read current memory window >10^4 is obtained. **Fig. 2(d)** shows the read current and corresponding V_{SN} as a function of time. The area information about related capacitors is presented in **Fig. 2(f)** , so that the detailed capacitance changes can be calculated by normalized capacitance value**.** Under the standard of ΔV_{SN} = 0.1V in the read operation, an ultra-long 10-ks retention time is achieved.

The ΔV_{SN-WWL} and relevant variations of nine random memory cells under each different annealing condition are shown in **Fig. 3(a)**. We recorded the voltage at the SN node after the occurrence of WWL and WBL potential falling. The difference between them and the write voltage was calculated as the V_{SN} drop. The total ΔV_{SN-WWL} decreases by 22% as the thermal treatment time and temperature increase, while the optimal variation on with 75% improvement is observed under the NT-250°C condition. To investigate the mechanism of these changes, the normalized C_{WWL-SN} characteristics of the 2T0C DRAM cell is depicted in **Fig. 3(b)**, which dominates the ΔV_{SN-WWL} based on traditional constant capacitive coupling effects [4] [12]. To simulate the real operating state of our DRAM cell, the V_{DS} is set to 1V during capacitance testing. However, as the dielectric capacitance usually does not change after low-temperature annealing [17] [19] [20], these varying C-V properties originate from modifications to the IGZO channel and behave as a typical n-type gate-to-source MOS capacitor. According to the voltage division rule and actual 2T0C DRAM circuit, the capacitive coupling model considering novel IGZO variable capacitance effect on V_{SN} can be derived:

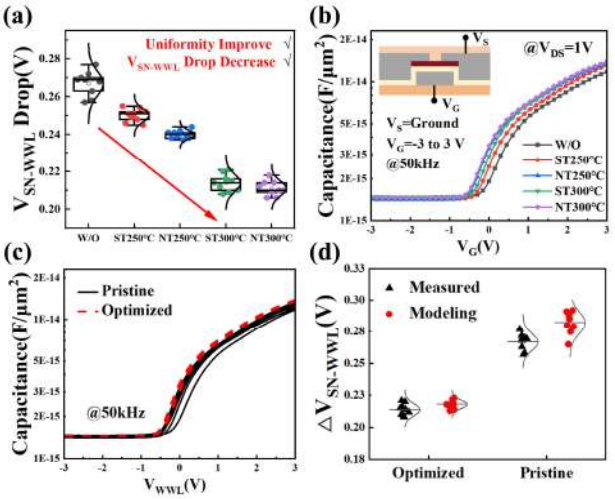


Fig. 3. (a) V_{SN} drop under varying OHAD conditions after WWL voltage transition. (b) Normalized C_{WWL-SN} under different OHAD treatments. (c) Variation of normalized C-V in DRAM cell before and after optimization. (d) Proposed variable IGZO capacitor model, which matches well with measured V_{SN} drop variation.

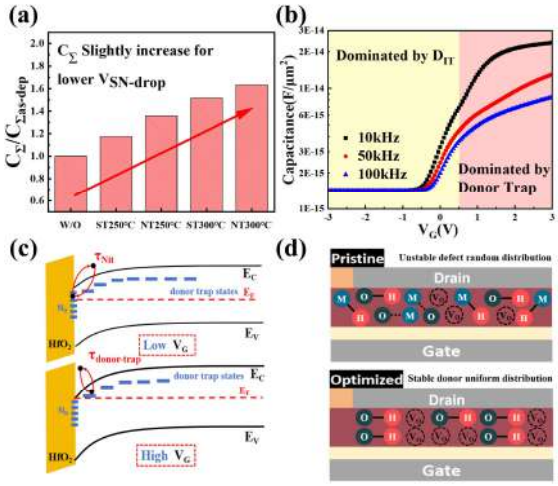


Fig. 4. (a) Total parasitic capacitance under different OHAD treatments (b) Normalized C_{GS} under different frequencies. (c) Energy-band transport diagram of IGZO MOS capacitor with different voltage. (d) Diagram of the optimization for the gate-drain capacitor in an IGZO transistor.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on November 19,2025 at 07:12:16 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Electron Device Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LED.2025.3616974

$$\Delta V_{SN-WWL} = \frac{\int_{\Delta V_{WWL}} C_{WWL-SN} dV + \int_{\Delta V_{RL}} C_{RL-SN} dV}{C_{\Sigma}} \quad (1)$$

Where C_{Σ} is the total capacitance between the SN and other nodes. The measured value for untreated cells is about 10 fF. ΔV_{WWL} are the voltage change at WWL. RL stands for RWL or RBL. Given the prevailing read/write operation schemes [4] [5] [6] [11] [12], it is imperative that the WWL potential be lowered prior to that of the WBL to guarantee the complete cut-off of the write transistor. During WWL transition, V_{RBL} and V_{RWL} are unchanged, and the related impact factors on ΔV_{SN-WWL} are zero. In this case, ΔV_{SN-WWL} solely depends on C_{WWL-SN}, which is dominated by the IGZO variable capacitor. The variation of normalized C-V curve in nine DRAM cells before and after optimization is illustrated in **Fig. 3(c)**, and the frequency of 50 kHz is used. It is found that the uniformity of the C-V curve is highly consistent with the ΔV_{SN} variation. Referring to the proposed variable IGZO capacitor model (**Eqn. (1)**), the estimated ΔV_{SN} variation matches well with measured results (**Fig. 3(d)**). Testing of the recently reported mobile charge sharing mechanism [6] revealed no significant impact on these changes (data not shown). This behavior is likely attributable to the scaling of device dimensions down to the 1 μm or sub-μm regime, causing coupling capacitance to become the dominant factor in memory performance.

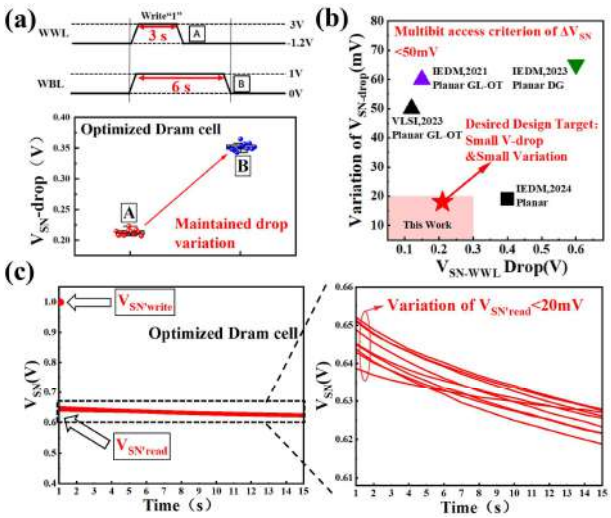


Fig. 5. (a) Maintained V_{SN} drop variation after WWL and WBL transitions. (b) Benchmark of V_{SN-WWL} drop and variation for recent IGZO DRAM, the record-low drop variation is demonstrated. (c) V_{SN}-time characteristics for the 2T0C DRAM cells.

Fig. 4(a) shows the relationship between C_{Σ} and optimization conditions. Since several IGZO variable capacitances increase with annealing time and temperature, C_{Σ} enlarges more than C_{WWL-SN}, thus causing the total ΔV_{SN-WWL} decreases (**Fig. 3(a)**). In IGZO capacitance, AC gate voltage causes the electrons (ΔQ) which have been captured by traps and subsequently emitted from traps to be effectively collected by the drain and source terminals, resulting in a larger capacitance. Generally, the C-V characteristics of an IGZO capacitor can be divided into two processes [21], including both the periods prior to and following the attainment of the threshold voltage by the gate voltage (**Fig. 4(b)**). The energy band diagrams of these two processes in **Fig. 4(c)** shows that the Fermi level shift triggers a transition in dominant charge-storage defects: interface defects prevail at low gate voltages, while donor trap defects dominate once the Fermi level exceeds the interface defect energy levels under high gate voltages. Consequently, the variation of C-V characteristics is determined by the synergistic effect of these two defect types. The large deviation of pristine C-V curves in **Fig. 3(c)** is due to the random distribution of unstable defects (M-H, weak M-O, etc.), as depicted in **Fig. 4(d)**. Due to the specific post-treatment, defects in channel region are well compensated due to the oxygen introduction and hydrogen out-diffusion, while more donor-like hydrogens are generated in S/D region by broken M-H bonds with S/D metal capping [15] [16]. In this case, the V_{TH} variation and the on-current of devices can be optimized, as shown in **Fig. 1**. The improvement in V_{TH} variation, and uniform distribution of stable donor in S/D region both help to improve the variation of V_{SN} drops. On the other hand, the interface states are also well compensated by such thermal treatment. As a result, the ultra-stable capacitance and C-V curve slope are realized, yielding outstanding ΔV_{SN-WWL} variations during write operations (**Fig. 3(a)**).

During normal 2T0C DRAM operations, the ΔV_{SN-WBL} plays an important role as well (**Fig. 2(b)**). The V_{SN} drop triggered by the WBL-SN node is the difference between A and B. As illustrated in **Fig. 5(a)**, it is observed that even after the WBL voltage transition, the V_{SN} drop variation is nearly unchanged, similar to the previously reported work [4], which means that the effect of C_{WBL-SN} is similar to a constant capacitance. Intriguingly, the effective C_{WBL-SN} calculated from the measured V_{SN-WBL} drop in this work, which is coincidentally similar to the C_{WWL-SN}. **Fig. 5(b)** illustrates the benchmark of ΔV_{SN-WWL} and corresponding variation among recent IGZO DRAM, where the best-in-class memory stability is demonstrated, attributed to the enhancement of defect distributions. The typical evolution of V_{SN} over time is shown in **Fig. 5(c)**. The ultra-stable read characteristics indicate that the targeted processing of IGZO variable capacitance effect is imperative to achieve stable readout stability and multi-bit operation in 2T0C DRAM arrays.

IV. Conclusion

In summary, the variable IGZO capacitance impact in capacitive coupling effect of 2T0C DRAM cell is investigated. It is found that the variation of V_{SN} drop is determined by the typical n-type gate-to-source MOS capacitor. By using OHAD treatment to improve the uniformity of IGZO capacitance, the total V_{SN} drop and its variation decreases by 22% and 75%, respectively. Through characterizing the interdependence between variable capacitance variations and V_{SN} drop fluctuations, a novel model is proposed and the corresponding optimization method is suggested. Therefore, a record-low V_{SN} drop variation <20 mV and outstanding readout stability with a retention of 10-ks are demonstrated among 2T0C DRAM. This work provides guidance for the optimization of stable readout and multi-bit operation in IGZO-based 3D DRAM.

References

[1] Yan S Z, Cong Z R, Lu N D, et al. Recent progress in InGaZnO FETs for high-density 2T0C DRAM applications. Sci China Inf Sci, 2023, 66(10): 200404, https://doi.org/10.1007/s11432-023-3802-8.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on November 19,2025 at 07:12:16 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Electron Device Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LED.2025.3616974

[2] K. Nomura, H. Ohta, A. Takagi, T. Kamiya, M. Hirano, and H. Hosono, "Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors," (in English), *Nature,* Article vol. 432, no. 7016, pp. 488-492, Nov 2004 doi: 10.1038/nature03090.

[3] A. Belmonte et al., "Capacitor-less, Long-Retention (>400s) DRAM Cell Paving the Way towards Low-Power and High-Density Monolithic 3D DRAM," 2020 IEEE International Electron Devices Meeting (IEDM), 2020, pp. 28.2.1-28.2.4, doi: 10.1109/IEDM13553.2020.9371900.

[4] A. Belmonte et al., "Tailoring IGZO-TFT architecture for capacitorless DRAM, demonstrating > 103s retention, >1011 cycles endurance and Lg scalability down to 14nm," 2021 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2021, pp. 10.6.1-10.6.4, doi: 10.1109/IEDM19574.2021.9720596.

[5] S. Zhu et al., "First Demonstration of Deeply Scaled 2T0C DRAM With Record Data Retention and Fast Write Speed," in IEEE Electron Device Letters, vol. 46, no. 3, pp. 405-408, March 2025, doi: 10.1109/LED.2025.3531420.

[6] L. Zheng, Z. Wang, Z. Lin and M. Si, "First Demonstration on the Transient Writing Characteristics of Multi-bit ALD IGZO 2T0C DRAM by Fast I-V Measurement," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024, pp. 1-4, doi: 10.1109/IEDM50854.2024.10873579.

[7] Hu Q , Gu C , Li Q ,et al.True Nonvolatile High-Speed DRAM Cells Using Tailored Ultrathin IGZO[J].Advanced Materials, 2023, 35(20):8.DOI:10.1002/adma.202210554.

[8] K. Chen et al., "Improved Multi-bit Statistics of Novel Dual-gate IGZO 2T0C DRAM with In-cell VTH Compensation and ΔVSN/ΔVDATA Boosting Technique," 2023 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2023, pp. 1-4, doi: 10.1109/IEDM45741.2023.10413770.

[9] W. Lu et al., "First Demonstration of Dual-Gate IGZO 2T0C DRAM with Novel Read Operation, One Bit Line in Single Cell, ION=1500 μA/μm@VDS=1V and Retention Time>300s," 2022 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2022, pp. 26.4.1-26.4.4, doi: 10.1109/IEDM45625.2022.10019488.,

[10] K. Toprasertpong et al., "Co-designed Capacitive Coupling-Immune Sensing Scheme for Indium-Tin-Oxide (ITO) 2T Gain Cell Operating at Positive Voltage Below 2 V," 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Kyoto, Japan, 2023, p. oi:10.23919/VLSITechnologyandCir57934.2023.10185433.

[11] O. Phadke, S. Gopal Kirtania, D. Chakraborty, S. Datta and S. Yu, "Suppressed Capacitive Coupling in 2 Transistor Gain Cell With Oxide Channel and Split Gate," in IEEE Transactions on Electron Devices, vol. 71, no. 11, pp. 6749-6755, Nov. 2024, doi: 10.1109/TED.2024.3463628.

[12] L. Zheng, Z. Wang, Z. Lin and M. Si, "The Impact of Parasitic Capacitance on the Memory Characteristics of 2T0C DRAM and New Writing Strategy," in IEEE Electron Device Letters, vol. 44, no. 8, pp. 1284-1287, Aug. 2023, doi: 10.1109/LED.2023.3287942.

[13] Y. Su et al., "Complementary Oxide Semiconductor-Based 2T0C DRAM Macro with CFET Peripherals Using TeOx-PFET/IGZO-NFET for 3D Memory Integration," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024, pp. 1-4, doi: 10.1109/IEDM50854.2024.10873505.

[14] M. Shi et al., "Counteractive Coupling IGZO/CNT Hybrid 2T0C DRAM Accelerating RRAM-based Computing-In-Memory via Monolithic 3D Integration for Edge AI," 2023 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2023, pp. 1-4, doi: 10.1109/IEDM45741.2023.10413876.

[15] G. Yan et al., "First Demonstration of True 4-bit Memory with Record High Multibit Retention >103s and Read Window >105 by Hydrogen Self-Adaptive-Doping for IGZO DRAM Arrays," 2023 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2023, pp. 1-4, doi: 10.1109/IEDM45741.2023.10413762.

[16] Y. Yang et al., "Unveiling the Unique Region Dependence of Bias Stability in Sub-μm IGZO TFTs Using Floating Channel Effect for 3D DRAM," in IEEE Electron Device Letters, vol. 46, no. 2, pp. 290-293, Feb. 2025, doi: 10.1109/LED.2024.3515091.

[17] T. -M. Wang, C. -H. Chang and J. -G. Hwu, "Enhancement of Temperature Sensitivity for Metal–Oxide–Semiconductor (MOS) Tunneling Temperature Sensors by Utilizing Hafnium Oxide (HfO2) Film Added on Silicon Dioxide (SiO2)," in *IEEE Sensors Journal*, vol. 6, no. 6, pp. 1468-1472, Dec. 2006, doi: 10.1109/JSEN.2006.884424.

[18] K. -M. Yu et al., "Low-Temperature Fabrication of High Quality Gate Insulator in Metal–Oxide–Semiconductor Capacitor Using Laser Annealing," in IEEE Electron Device Letters, vol. 40, no. 2, pp. 167-170, Feb. 2019, doi: 10.1109/LED.2018.2889346.

[19] Moldovan O , Castro-Carranza A , Estrada M ,et al.A Complete Charge-Based Capacitance Model for IGZO TFTs[J].IEEE Electron Device Letters, 2019, 40(5):730-733.DOI:10.1109/LED.2019.2905272.

[20] Bhoolokam A , Nag M , Chasin A ,et al.Analysis of frequency dispersion in amorphous In–Ga–Zn–O thin-film transistors[J].Journal of Information Display, 2015, 16(1):31-36.DOI:10.1080/15980316.2014.991769.

[21] Tsao Y C , Chang T C , Chen H M ,et al.Abnormal hump in capacitance–voltage measurements induced by ultraviolet light in a-IGZO thin-film transistors[J].AppliedPhysicsLetters,2017,110(2):023501.DOI:10.1063/1.4973856.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on November 19,2025 at 07:12:16 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名：秦善天
方向：集成电路
单位：中国科学院计算技术研究所

题目：面向边缘高通量应用的数据流架构软硬协同调度优化研究

PANDA: Adaptive Prefetching and Decentralized Scheduling for Dataflow Architectures

SHANTIAN QIN, ZHIHUA FAN, WENMING LI, ZHEN WANG, XUEJUN AN, XIAOCHUN YE, and DONGRUI FAN, State Key Laboratory of Processors, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China and School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China

Dataflow architectures are considered promising architecture, offering a commendable balance of performance, efficiency, and flexibility. Abundant prior works have been proposed to improve the performance of dataflow architectures. Nevertheless, these solutions can be further improved due to the lack of efficient data prefetching and flexible task scheduling. In this article, we propose a novel dataflow architecture with adaptive prefetching and decentralized scheduling (PANDA). First, we present an application-adaptive data prefetching method and on-chip memory microarchitecture designed to overlap memory access latency. Second, we introduce a decentralized dataflow scheduling approach and processing element (PE) microarchitecture aimed at improving hardware utilization. Experimental results show that in a wide range of real-world applications, PANDA attains up to 2.53× performance improvement and 1.79× energy efficiency improvement over the state-of-the-art dataflow architectures.

CCS Concepts: • **Computer systems organization → Data flow architectures**; **Multicore architectures;**

Additional Key Words and Phrases: Prefetching, decentralized dynamic scheduling, reconfigurable on-chip memory architecture

ACM Reference Format:
Shantian Qin, Zhihua Fan, Wenming li, Zhen Wang, Xuejun An, Xiaochun Ye, and Dongrui Fan. 2025. PANDA: Adaptive Prefetching and Decentralized Scheduling for Dataflow Architectures. *ACM Trans. Arch. Code Optim.* 22, 2, Article 62 (June 2025), 27 pages. https://doi.org/10.1145/3721288

1 Introduction

Advances in integrated circuit technology have propelled progress in conventional processors over the past decades. However, this approach is losing effectiveness with the slowdown of Moore's Law [35] and the termination of Dennard Scaling [15]. Fortunately, dataflow architectures exhibit great promise owing to their inherent flexibility and immense potential for high computational parallelism [12, 14].

This paper is sponsored by National Key R&D Program of China (Grant No.2023YFB4503500), Beijing Nova Program (No.20220484054, No.20230484420), Beijing Natural Science Foundation and Changping Innovation Joint Fund (L234078), CAS Project for Youth Innovation Promotion Association.
Authors' Contact Information: Shantian Qin, Zhihua Fan (Corresponding author), Wenming Li (Corresponding author), Zhen Wang, Xuejun An, Xiaochun Ye, and Dongrui Fan, State Key Laboratory of Processors, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China and School of Computer Science and Technology and University of Chinese Academy of Sciences, Beijing, China; e-mails: qinshantian23s@ict.ac.cn, fanzhihua@ict.ac.cn, liwenming@ict.ac.cn, wangzhen21s@ict.ac.cn, axj@ict.ac.cn, yexiaochun@ict.ac.cn, fandr@ict.ac.cn.

This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).
ACM 1544-3973/2025/06-ART62
https://doi.org/10.1145/3721288

Dataflow architectures represent a category of architectures that harness the immense potential of high computational parallelism through direct intercommunication among an array of **processing elements (PEs)** [17, 19]. A dataflow program is delineated by a **dataflow graph (DFG)**, which is a graph that depicts operations as nodes and data dependencies as edges. The fundamental principle of the dataflow execution model is that any DFG node can be executed as soon as all its required operands are available [16]. The dataflow architecture primarily comprises a PE array, an on-chip network, an instruction buffer, a data buffer, and a centralized controller. The PE array is composed of numerous PEs interconnected by the on-chip network. It typically operates as an accelerator alongside a host core, collaboratively executing programs. The instruction buffer stores instructions and configuration information, while the data buffer stores the data. The centralized controller manages the scheduling of configuration information, instructions, and data. It also serves as the interface for interactions between the PE array and the host core, facilitating application iteration, initialization, and teardown.

Abundant prior works have been proposed to improve the performance of dataflow architectures. Some works focus on overlapping the memory access latency, employing techniques like ping-pong data buffering [8, 33, 48, 60] to enhance throughput by decoupling and overlapping memory access and execution. Others aim to improve the computational resource utilization, including adopting dataflow-driven scheduling [7, 13, 21, 22, 37, 39] to extract parallelism through dynamic task issuing on hardware. Nevertheless, these solutions can be further improved due to: *(1) Lack of Efficient Data Prefetching*. From the memory access perspective, they commonly rely on **scratchpad memory (SPM)** as on-chip memory, prefetching all data before program execution to overlap memory access latency. However, this approach is constrained by inefficient prefetching of certain data and the finite capacity of on-chip memory. *(2) Lack of Flexible Task Scheduling*. From the computation perspective, they typically adopt pure static scheduling or leverage software and hardware co-design to achieve dataflow-driven scheduling. However, pure static scheduling suffers from insufficient parallelism due to over-serialization and workload imbalance. Moreover, while dataflow-driven scheduling offers greater flexibility than pure static scheduling, it remains hindered by centralized control, resulting in inefficient communication and inflexible hardware.

By conducting a comprehensive review of existing research on data prefetching and task scheduling for dataflow architectures, we aim to explore novel insights and tackle the aforementioned challenges by proposing an innovative dataflow architecture that leverages these insights.

- We find that some data exhibit optimal suitability for prefetching before execution (prefetchable). Conversely, some data are best accessed via cache during execution (non-prefetchable). Driven by this insight, we introduce an *Application-adaptive Data Prefetching Method*.
- We observe that each PE in dataflow architectures possesses the capability to independently schedule tasks, thereby reducing the requirement for a centralized controller. Inspired by this insight, we propose a *Decentralized Dataflow-driven Dynamic Scheduling Method*.

Building upon the previously discussed insights, we propose a novel dataflow architecture with adaptive **p**refetching **an**d **d**ecentr**a**lized scheduling (**PANDA**) as a solution that aptly aligns with our goal.

Our contributions are as follows:

- We introduce an application-adaptive data prefetching method: (1) It describes memory access interfaces, supporting the prefetch/postback and load/store for prefetchable and non-prefetchable data. It fosters the decoupling of the prefetchable data access from the application execution. (2) It constructs an on-chip memory microarchitecture tailored for multi-core accelerators based on the ISA. It leverages reconfigurable physical storage shared by SPM and cache to reduce on-chip area overhead. (3) It presents a mechanism for

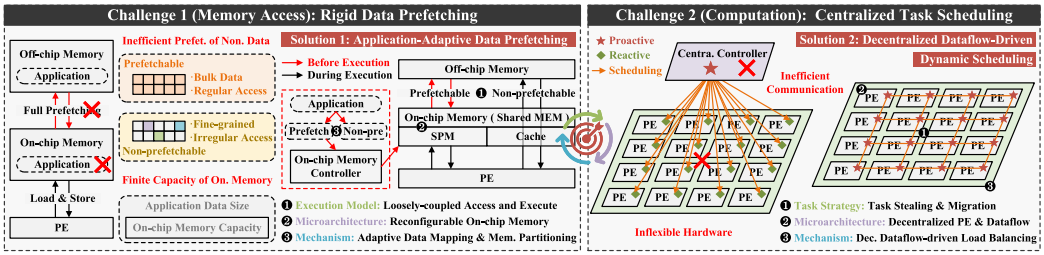


Fig. 1. Traditional dataflow architecture, challenges, and solutions into PANDA design.

application-adaptive data prefetching and on-chip memory partitioning. It facilitates the organized separation of data into prefetchable and non-prefetchable categories.

- We propose a decentralized dataflow-driven dynamic scheduling method: (1) It describes task stealing and migration strategies based on PE state. It facilitates better workload balancing through on-demand real-time task scheduling for busy and idle PEs. (2) It constructs a decentralized PE microarchitecture. It enhances hardware flexibility by empowering PEs to autonomously and proactively schedule their internal tasks. (3) It presents a mechanism for decentralized load balancing. It effectively integrates the aforementioned strategies with hardware microarchitecture.
- We implement all modules of PANDA using Verilog. Moreover, we conduct a comprehensive evaluation that includes the acceleration effect of each innovation feature. Compared to the state-of-the-art dataflow architectures, PANDA outperforms REVEL, Plasticine, DFU, and MTDE by geomean 2.53×, 1.90×, 1.38×, and 1.19× in a wide range of real-world applications, including scientific computing, artificial intelligence, digital signal processing, and graph processing algorithms.

2 Background and Related Works

In this section, we delve into related works concerning data prefetching and task scheduling for dataflow architectures to identify challenges for performance improvement and explore novel insights. Following this analysis, we introduce our proposed architectural overview, focusing on two primary innovations: application-adaptive data prefetching and decentralized dataflow-driven dynamic scheduling.

2.1 Data Prefetching: Overlap Memory Access Latency

The Memory Wall [57] remains a significant barrier to system performance improvement. This challenge is particularly accentuated in multi-core accelerators due to the increased demand for memory bandwidth and the necessity for data sharing among multiple cores on the same chip [6]. As the number of cores increases, this issue intensifies further, imparting a detrimental impact on the overall performance and efficiency of multi-core systems [11, 45]. Additionally, efficient on-chip memory solutions can help bridge the speed gap between processors and memory by reducing memory access latency and optimizing data transfer. Consequently, designing efficient on-chip memory solutions for multi-core accelerators is essential to mitigating the impact of memory wall and enhancing their overall performance. In the realm of on-chip memory solutions, cache and **scratchpad memory (SPM)** stand out.

Cache has been widely used in modern systems as an effective strategy for mitigating the impact of the memory wall [3, 23, 45, 59], which is *globally addressed*. However, cache has several

weaknesses: *(1) Indirect Addressing:* Cache loads and stores specify addresses that hardware must translate to determine the physical location of the accessed data; *(2) Implicit Data Orchestration:* The load request initiators do not directly govern the cache hierarchy's determinations regarding the retention or removal of response data at different storage hierarchy levels [46].

An alternative to the cache is the SPM [8, 10, 17, 18, 20, 26, 37, 41]. SPM is *directly addressed* so it does not have overhead from tag comparisons, which provides significant savings: 34% area and 40% power[4]. SPM also provides *explicit data orchestration* (data in SPM are explicitly allocated by software) and support for rapid and simple *software prefetching of bulk data* as it is managed in software, either by the programmer or through compiler support [42]. These features make SPM appealing, especially for real-time systems [27]. However, SPM also has some inefficiencies: *(1) No Global Addressing:* SPM uses a separate address space disjoint from the global address space, with no hardware mapping between the two; *(2) No Irregular Access Pattern:* The order of accesses to SPM must be known statically upfront and cannot be conditional. The long setup time of **direct memory access (DMA)** leads to the fact that for *fine-grained* data access, *hardware on-demand load* of cache during execution is more efficient than software prefetching of SPM.

Based on the above analysis, cache and SPM each have distinct strengths and weaknesses, excelling only in specific scenarios and lacking broad applicability across various scenarios. There are two main options for combining the advantages of SPM and cache: integrating cache into the on-chip memory architecture alongside SPM, and designing a new on-chip memory architecture that merges the strengths of both SPM and cache. To fully leverage the benefits of cache and SPM, the on-chip memory architecture that integrates both cache and SPM has been proposed and extensively embraced within the industry [2, 38]. However, this configuration significantly amplifies the on-chip area overhead. In addition, the pursuit of an on-chip memory architecture that combines the merits of SPM and cache has garnered significant research attention. There have numerous prior works focused on enhancing on-chip memory architecture by amalgamating the benefits of cache and SPM. ROMA [44] conducted a comprehensive review and comparison of related works, focusing on *Globally Addressed SPM* (e.g., Stash [31]), *Reconfigurable Cache* (e.g., Jenga [50] and Hotpads [51]), and *Decoupled Access Execute (DAE) Architectures* (e.g., DeSC [25] and PDAE [9]). However, none of them fully encompasses all of the benefits illustrated in Table 1.

Challenge1: Traditional dataflow architectures commonly rely on SPM as on-chip memory, prefetching all data to SPM before program execution to overlap memory access latency. However, this approach encounters several limitations, as illustrated in Figure 1 (left). First, not all data can be effectively prefetched, especially small-scale data scattered in off-chip memory. Second, the finite capacity of on-chip memory imposes constraints, particularly for applications with large data sizes, where prefetching all data is impractical. For kernels like matrix multiplication, large data can be split into smaller chunks that fit within on-chip memory for iterative execution. However, for kernels with irregular memory access, such as **Breadth First Search (BFS)**, data access is dynamic and unpredictable. These workloads often involve numerous random data accesses across diverse memory regions, making it difficult to split the data into smaller chunks for iterative execution.

By analyzing the characteristics of numerous real-world applications, we have identified a specific data storage and memory access pattern that exhibits regularity. This pattern involves data being stored and accessed in the form of distinct blocks, each characterized by uniform block size and an identical stride length separating consecutive blocks. For this type of data, provided that the start address, block size, stride length, and the number of blocks are known, it becomes feasible to compute the addresses of all data elements. In previous studies, this pattern is defined as *stream* [37, 52–55]. By prefetching this type of data into on-chip memory before program execution, we

Table 1. Comparison of Cache, SPM, and PANDA

Features	Cache	SPM	PANDA
Global Addressing	✓	✗	✓ (Non-pre.)
Irregular Access Pattern	✓	✗	✓ (Non-pre.)
Hardware Fine-grained Load	✓	✗	✓ (Non-pre.)
Software Bulk Prefetching	✗	✓	✓ (Prefetch.)
Direct Addressing	✗	✓	✓ (Prefetch.)
Explicit Data Orchestration	✗	✓	✓ (Prefetch.)
Reconfigurable Physical Storage	✗	✗	✓

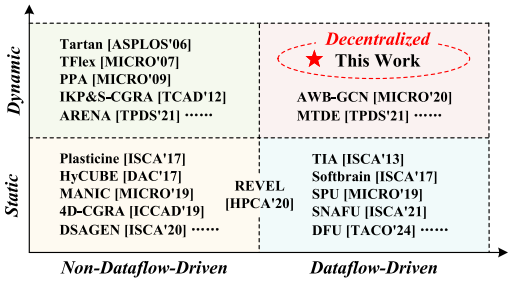


Fig. 2. Scheduling approaches of current research.

can reduce the number of requests and transfers in the memory system, thereby diminishing memory access latency and increasing memory access efficiency.

We observe that the proportion of prefetchable and non-prefetchable data exhibits variability across diverse applications. First, prefetchable data are prevalent and constitute a substantial portion in some applications, such as General Matrix Multiplication, Stencil, Rectified Linear Unit Activation, and Secure Hash Algorithm 256. SPM with DMA supports prefetch before execution to access prefetchable data more efficiently than cache. This highlights the advantage of burst transfer for bulk data. Additionally, non-prefetchable data are also not negligible and account for a large proportion of some applications, particularly in applications like Lagrange Interpolation, BFS, and **PageRank (PR)**. Cache supports load/store during execution to access non-prefetchable data more efficiently than SPM. This is primarily because, in cases of fine-grained data transfers, the burst length is often minimal or even there is no burst transfer. In such scenarios, the total DMA initial time overhead resulting from prefetch becomes very large.

To adeptly accommodate this diversity, we introduce an application-adaptive data prefetching method and design a novel reconfigurable on-chip memory architecture tailored for dataflow architectures, amalgamating the logical organization of SPM and cache while sharing physical storage. The reconfigurable on-chip memory architecture provides support for both prefetching before execution and load/store during execution, ensuring efficient memory access to both prefetchable and non-prefetchable data, as depicted in Figure 4. It is engineered to deliver the benefits of both SPM and cache without incurring excessive area overhead, as shown in Table 1. It is essential to recognize that the data size of prefetchable candidates in certain applications exceeds the storage capacity of on-chip memory. For kernels like matrix multiplication, large data is split into smaller chunks that fit within on-chip memory for iterative execution. However, for kernels like BFS with irregular memory access, splitting the data into smaller chunks for iterative execution is challenging. In such scenarios, proactive endeavors are undertaken to prefetch as many prefetchable candidates as possible from off-chip to on-chip memory before execution. Prefetchable candidates that cannot be prefetched and remain in off-chip memory are then reclassified as non-prefetchable.

In addition, there are key differences that distinguish PANDA prefetching from out-of-order memory accesses to hide memory access latency. In an out-of-order core, memory access latency can be hidden by reordering instructions within the instruction window, but each load instruction still requires tag-checking in caches and frequent fine-grained off-chip memory access. In contrast, adaptive prefetching in PANDA directly links bulk prefetchable data with the Prefetch instruction, fully leveraging the rapid bulk prefetching capabilities of DMA. Furthermore, the on-chip SPM functions as software-exposed stream buffers, eliminating the overhead of dynamic prediction and tag-checking in caches. Additionally, prefetchable data access is decoupled from the execution pipeline. This decoupling is particularly beneficial in mitigating the negative impact of long-latency memory accesses, without requiring a large instruction window. As a result, PANDA

prefetching reduces the instruction pressure on the execution pipeline and enables more efficient memory access.

2.2 Task Scheduling: Improve Computational Utilization

In dataflow architectures, where computational resources are not only abundant but also spatially distributed, improving resource utilization is paramount for enhancing performance and energy efficiency [19]. The choice of task scheduling method plays a pivotal role in this improvement, as it determines the spatial-temporal allocation of workload to hardware architecture [33]. In recent years, numerous studies have focused on enhancing the computational resource utilization of dataflow architectures through optimized task scheduling methods. In this section, we will delve into these works in detail.

Scheduling in dataflow architectures presents greater challenges compared to traditional processors, as it involves both task issuing and task mapping (allocating tasks to spatial resources). Task issuing determines the exploitable parallelism, while task mapping influences resource utilization [7]. In dataflow architectures utilizing pure *static scheduling* [24, 28, 29, 43, 55], the issuing and mapping of tasks are handled statically by the compiler. While these architectures benefit from low hardware overhead, they often suffer from insufficient parallelism due to *over-serialization* and *workload imbalance*. On the one hand, the compiler must create conservative schedules to accommodate the worst-case scenario, leading to over-serialized execution. This is particularly problematic for applications with irregular computation patterns, such as tree and graph algorithms. On the other hand, for large-scale dataflow architectures [5, 47], tasks are typically generated by the compiler and statically mapped to tiles, which are subarrays of interconnected PEs. In such scenarios, achieving workload balance becomes critical for maintaining high throughput. However, achieving balanced execution becomes challenging due to inter-task and intra-task computations variance [40]. Moreover, workloads often change with iteration in irregular applications, rendering pure static scheduling impractical for balanced execution.

For *over-serialization*, some efforts have effectively leveraged the intrinsic characteristics of dataflow architectures by employing a hardware-software co-design approach to achieve *dataflow-driven scheduling* [7, 13, 19, 21, 22, 37, 39]. Instead of relying on the compiler, the dataflow architecture of dataflow-driven scheduling extracts parallelism by dynamic task issuing on hardware. This approach determines when to execute a specific operation based entirely on the control of dataflow, rather than relying on static planning and prediction in advance. It is worth mentioning that REVEL [56] comprises simple PEs in a systolic form (non-dataflow-driven) and dataflow PEs capable of performing complex computations (dataflow-driven), achieving the hybrid static-dynamic task issuing. However, it heavily relies on the compiler's heuristic search for the globally optimal timing decision due to the consideration of balancing between dataflow PEs and systolic PEs, making it significantly less flexible than true dataflow-driven scheduling.

For *workload imbalance*, certain works adjust the allocation of resources and workloads through dynamic mapping during execution to achieve *dynamic scheduling* [1, 7, 21, 30, 34, 40, 49]. There are mainly two kinds of dynamic mapping methods: *(1) Software-driven:* This approach relies on the software compiler. Examples include heuristic virtualization [34], which periodically searches for an optimized mapping to maximize a predefined function, and the bitstream transformation mechanism [40], which allows the shrinking and growing of resources for individual tasks. A common limitation of these approaches is that they lack a detection mechanism to monitor task execution. They statically estimate the workloads of each task, leading to suboptimal mapping when the workload varies and is unpredictable due to dynamic computation variance. Additionally, the algorithmic complexity and sensitivity to function forms and parameters make these approaches less favorable for hardware designs; *(2) Hardware-driven:* This approach is based on real-time hardware

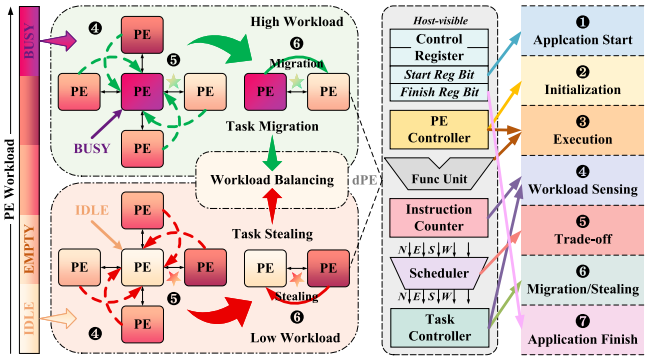


Fig. 3. General workflow of workload-based task migration & stealing and components of decentralized PE.

execution. An example is MTDE [7], which implements an elastic task scheduling scheme. This approach dynamically reschedules tasks at runtime using the tagged-token dataflow paradigm to support dynamic task-level parallelism. Tasks are also dynamically resized during execution through duplication, combination, and substitution operators, ensuring balanced multi-task execution. These hardware-driven methods [7, 21] are more adaptive to applications because decisions are based on real-time workload variations, and the logic is straightforward for hardware designs.

The difference between static and dynamic scheduling lies in whether task mapping is statically predetermined before execution and fixed during execution or can be dynamically adjusted during execution, while the distinction between non-dataflow-driven and dataflow-driven scheduling pertains to whether task issuing is statically pre-planned and compiled in advance or dynamically driven by dataflow during execution. Therefore, existing scheduling schemes in dataflow architectures are classified into four types in Figure 2.

In traditional dataflow architectures, communication between the host core and dataflow PE array, comprised of multiple PEs, is typically facilitated by a centralized controller, including the stream dispatcher & engine in Softbrain [37], the control core in Fifer [36], the stream control unit in REVEL [56], the micro-controller in DFU [19], and the task scheduling & control unit in MTDE [7]. These controllers play several key roles in traditional dataflow architectures. They manage the scheduling of configuration information, instructions, and data through a centralized control mechanism. They also act as the hub for interactions between the PE array and the host core, facilitating application iteration, initialization, and teardown.

Challenge2: While dataflow-driven dynamic scheduling offers greater flexibility than pure static scheduling, it still suffers from centralized control, as shown in Figure 1 (right). First, centralized control necessitates frequent interactions between the centralized controller and PEs, resulting in significant communication and routing overheads. This also leads to increased latency during execution. These issues become increasingly problematic as the scale of PE arrays continues to expand. Second, PEs under centralized control cannot autonomously schedule internal tasks. This absence of autonomous and proactive task scheduling limits hardware flexibility, impeding the achievement of asynchronous inter-PE communication and on-demand real-time scheduling. Achieving workload balancing becomes more arduous in this scenario as well.

We observe that PEs in dataflow architectures are generally sophisticated, and equipped with internal control units and interconnected data paths between PEs. Consequently, each PE possesses the potential to schedule internal tasks independently. However, PEs currently lack the ability to autonomously and proactively schedule tasks, instead relying on reactive scheduling under centralized control.

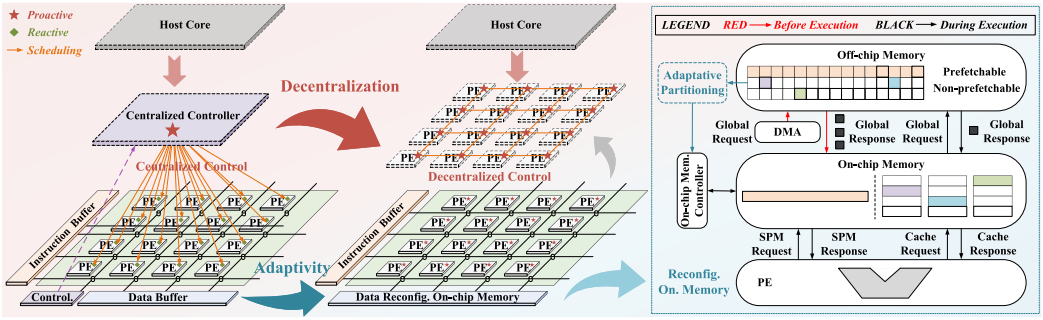


Fig. 4. PANDA overview: Comparison of traditional dataflow architecture and PANDA architecture.

Driven by this insight, we realize that the centralized controller can be replaced by the PE capable of proactive task scheduling to achieve decentralized dataflow-driven dynamic scheduling. To fully harness the advantages of decentralized scheduling, we introduce strategies for dynamic task migration and stealing based on PE workload. To support the decentralized scheduling method, we present a decentralized PE microarchitecture, as illustrated in Figure 3. This decentralized PE enables direct interaction with the host core and autonomous, proactive scheduling of configuration information, instructions, and data.

Through a comprehensive review of existing research on data prefetching and task scheduling for dataflow architectures, we identify two challenges for performance improvement from memory access and computation perspectives. Additionally, we uncover two novel insights to address these challenges and propose an innovative dataflow architecture, PANDA, which leverages these insights, as illustrated in Figure 4. PANDA features two primary innovations: application-adaptive data prefetching and decentralized dataflow-driven dynamic scheduling. We will describe these innovations in detail in the following sections.

3 Adaptive Prefetching

In this section, we first describe an ISA and its programmability, which provides support for prefetch/postback before/after execution and load/store during execution. By extending the memory access ISA, we not only enhance the abstraction level and reduce hardware design complexity but also optimize the execution model by decoupling prefetchable data access from computation.

Subsequently, we construct a reconfigurable on-chip memory microarchitecture with the two primary design principles: We avoid excessive additional area overhead increase, especially for the SRAM, and we have both SPM and Cache corresponding to the two types of memory access commands in the ISA. Here we describe our reconfigurable on-chip memory microarchitecture implementation.

Finally, we present a mechanism for application-adaptive data prefetching and on-chip memory partitioning in applications. It effectively categorizes the data necessitating access during application execution into the two aforementioned categories: prefetchable and non-prefetchable.

3.1 Execution Model of Loosely-coupled Access and Execute

The PANDA ISA abstracts data movement before/after execution and during execution with prefetch/postback and load/store. Table 2 lists the operations provided by PANDA ISA. We decouple prefetchable data access from execution by extending the access ISA with the introduction of Prefetching and Postback instructions. This optimization enhances the execution model and effectively reduces memory access latency.

Table 2. PANDA Instruction Extension

Command Names	Parameters	Description
PANDA_Prefetch	Source Off-chip Memory Address, Step Length, Access Size, Steps, Destination On-chip Memory Address	Data transfer from off-chip memory to on-chip memory before execution
PANDA_Postback	Source On-chip Memory Address, Step Length, Access Size, Steps, Destination Off-chip Memory Address	Data transfer from on-chip memory to off-chip memory after execution
PANDA_Load	Source Address, Destination Register Identifier, Access Selection Bit	Load prefetchable data from on-chip memory or Load non-prefetchable data from off-chip memory
PANDA_Store	Source Register Identifier, Destination Address, Access Selection Bit	Store prefetchable data to on-chip memory or Store non-prefetchable data to off-chip memory

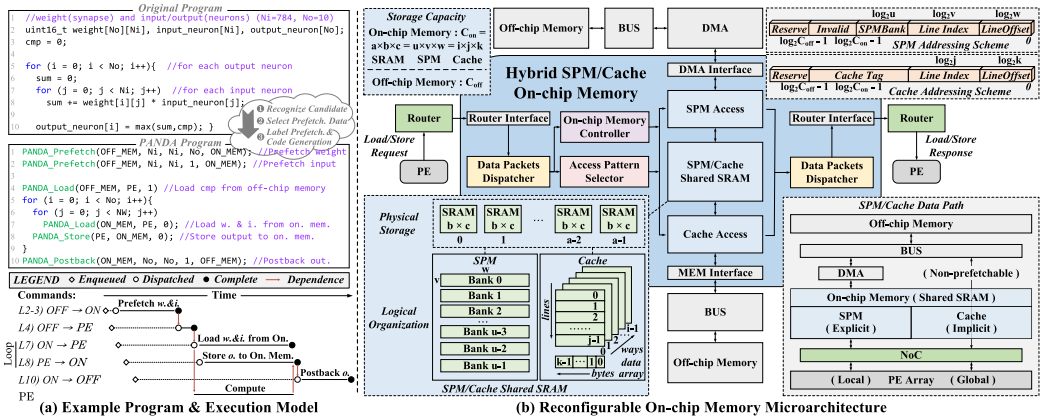


Fig. 5. PANDA memory: (a) Example program & execution model. (b) Reconfigurable microarchitecture.

Prefetch and Postback. Prefetch and *Postback* adopt the two-dimensional affine access, defined by access size (size of lowest level access), step length (size between consecutive accesses), and the number of steps. In addition, the *Prefetch* command specifies the source off-chip memory address and destination on-chip memory address, while the *Postback* command specifies the source on-chip memory address and destination off-chip memory address.

Load and Store. PANDA supports accessing individual elements and loading them into PE registers or storing them into on-chip memory. *Load* or *Store* commands specify a register target and source or destination address.

Example Program. Figure 5(a) illustrates an example neural network classifier, which primarily involves a dense matrix-vector multiplication of synapses (weights) and input neurons. The figure shows the original program, PANDA version, and execution behavior. The transformation to a PANDA version pulls the loading of the weight and input arrays out of the loop using *Prefetch* command (line 2-3) and local *Load* command (line 7). This not only reduces the overhead of requiring long instruction windows in the execution pipeline, but also decreases the number of inefficient and fine-grained data access requests to off-chip memory, thus improving overall memory access efficiency.

In addition, we implement compiler support to identify the prefetchable data and transform the original program to PANDA program. Our implementation uses LLVM [32]. There are three phases: recognizing candidates for prefetchable data, selecting prefetchable data, and code generation & labeling prefetchable data.

Recognizing Candidates for Prefetchable Data. The compiler first analyzes all static memory access instructions within loops, treating each instruction as a candidate for prefetchable data. By

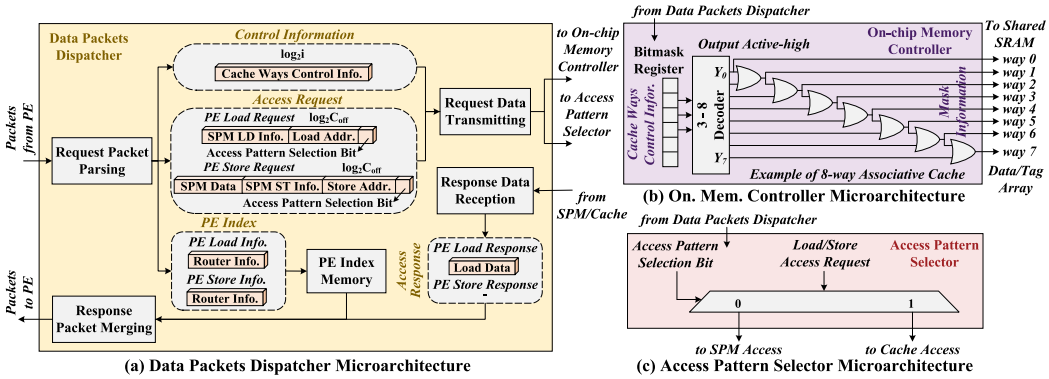


Fig. 6. Dispatcher, controller, and selectore microarchitectures.

evaluating the address, size, and stride of these instructions, it identifies those that conform to the two-dimensional affine access pattern as final prefetchable candidates. Furthermore, the compiler coalesces prefetchable candidates that share the same induction variable and have a small offset between their elements to optimize memory access.

Selecting Prefetchable Data. Due to the limited capacity of on-chip memory, it is necessary to determine whether the total size of the candidates qualified for prefetching exceeds the available on-chip memory capacity. If the total data size exceeds the on-chip memory capacity, an approach that combines memorization search with knapsack dynamic programming techniques will be employed to prefetch as much data as possible. Further details will be described in Section 3.3.

Code Generation & Labeling Prefetchable Data. During this phase, the compiler first generates the prefetching configuration for the selected prefetchable data. The configuration specifies parameters such as address, size, and stride. Then, based on the selected prefetchable data and their associated parameters, the compiler generates *Prefetch* commands and labels these prefetchable data by setting the access pattern selection bit of the *Load* commands to 0. This ensures that during execution, these data can be directly accessed from the on-chip SPM.

3.2 Reconfigurable On-Chip Memory Microarchitecture

At a high level, we combine a data packets dispatcher to assist the interconnection of on-chip memory and **processing element (PE)**, an on-chip memory controller to configure the physical storage, an access pattern selector to allocate the access request to SPM or cache, the SPM with DMA to support prefetch/postback for prefetchable data, the cache to support load/store for non-prefetchable data, and reconfigurable physical storage shared by SPM and cache (see Figure 5).

3.2.1 Data and Packets Dispatch. PEs are typically interconnected through the **network on chip (NoC)** to facilitate information exchange between PEs and on-chip memory through packets. However, the internal units of the reconfigurable on-chip memory communicate primarily through data rather than packets. The data packets dispatcher serves the critical role of converting between data and packets, requiring the parsing and merging of packets from/to PEs, as well as data exchange with the other internal units. The data packets dispatcher fulfills the external interaction of packets with PEs, as well as the internal interaction of data with the controller, selector, SPM, and cache.

The design of the data packets dispatcher is depicted in Figure 6(a). Initially, request packets from PEs are parsed into data. This unit transmits data, including access request information and

control signals, to the selector and controller. It stores PE index information in its internal memory. Furthermore, it receives data associated with access responses from other units. Finally, it merges the stored and received data into response packets destined for PEs.

3.2.2 On-Chip Memory Control & Access Pattern Selection. The on-chip memory controller plays a vital role in managing the SRAM that contains prefetchable data. It incorporates a bitmask register that holds the software-generated on-chip memory configuration information. This bitmask is used to configure specific cache ways as SPM, excluding them from cache lookup and replacement. This mechanism ensures that the prefetchable data accessed through SPM will not be overwritten by non-prefetchable data accessed through the cache.

Request packets from the PE carry SPM/Cache access pattern selection information (the lowest bit of the access request). The data packets dispatcher, upon parsing the access request, conveys the access pattern selection bit to the access pattern selector. If this bit is 0, the access pattern selector sends the load/store access request to the SPM; otherwise, it sends the request to the cache.

The controller flexibly and dynamically allocates the on-chip storage resources based on the prefetchable data size of the application, and the selector routes the access request from the PE to the SPM or cache. They are critical to achieving high generality with low area overhead. By partitioning the on-chip SRAM into SPM and cache based on the prefetchable data size of the application to access the prefetchable and non-prefetchable data, efficient resource utilization is achieved.

3.2.3 SPM and Cache Logical Organization. Our design incorporates both SPM and cache, each with distinct logical organizations, while sharing the same physical storage. SPM contains u banks, each of which is comprised of v lines with w bytes width. SPM interacts with off-chip memory through DMA and with PE array through NoC, where the former can be realized by *Prefetch* and *Postback* commands before/after execution, and the latter can be realized by *Load* and *Store* commands during execution. Cache is organized in i-way set-associative and each way consists of j lines with k bytes width. The interaction between cache and off-chip memory is facilitated by BUS, while its communication with PE array is enabled through NoC. The above two can be realized by *Load* and *Store* commands during execution.

Data Path. Our design provides two distinct data paths, SPM with DMA and cache, to facilitate the access of prefetchable and non-prefetchable data, respectively. Before application execution, DMA in the SPM path efficiently prefetches all prefetchable data from off-chip memory to on-chip memory. Additionally, the SRAMs containing the prefetchable data are configured as SPM and are excluded from cache lookup and replacement by the on-chip memory controller. During execution, PE accesses the prefetchable data directly from on-chip memory through SPM path, while PE accesses the non-prefetchable data from off-chip memory through cache path. This comprehensive setup ensures efficient access to both categories of data.

Coherence amd Consistency. In our approach, the compiler identifies and labels prefetchable data prior to program execution. These labeled data are then prefetched into on-chip memory. The on-chip SRAM storing prefetchable data belongs to the SPM access space, and the SPM capacity is determined by the size of the prefetchable data. The remaining on-chip SRAM is used for cache access and handles non-prefetchable data during execution. To ensure separation between SPM and cache, we utilize a cache way masking mechanism. This mechanism employs a bitmask register that stores software configuration information to control which cache ways are active during the lookup and replacement phases. This allows software to designate certain cache ways as SPM, excluding them from cache lookup and replacement. Consequently, the SPM is explicitly managed by software, while the cache remains implicitly managed by hardware and requires a hardware coherence protocol. PANDA adopts the MESI protocol for cache coherence.

3.2.4 SPM and Cache Physical Storage. The total storage capacity of SPM is $u \times v \times w$ bytes, and the total storage capacity of cache is $i \times j \times k$ bytes. In our design, SPM and cache adopt a unified physical storage, which is called shared SRAM. This shared SRAM consists of a individual SRAM units, each with a depth of b and a width of c bytes. Its total storage capacity is $a \times b \times c$ bytes. Note that the total storage capacity of SPM, cache, and shared SRAM must be equal.

Addressing Scheme. Since SPM and cache use the same physical storage, we need to address them separately based on SPM and cache logical organization. SPM addressing primarily involves banks, line index, and offset (u, v, w), while cache addressing mainly focuses on cache line index and offset (j, k). Assuming that off-chip memory storage capacity is C_{off} bytes and on-chip memory storage capacity is C_{on} bytes. Cache supports off-chip memory global addressing and uses $log_2 C_{off}$ bits valid address for addressing. In contrast, SPM supports only on-chip memory local addressing, and its valid addressing bit width is limited to $log_2 C_{on}$. Additionally, we reserve higher bits of the addressing for potential optimizations and specific design features. For instance, if the on-chip memory adopts a double-buffering technique, an extra bit might be needed to distinguish between the two buffers.

3.3 Adaptive Data Mapping & On-chip Memory Partitioning

In this section, we describe a mechanism for adaptive data mapping and on-chip memory partitioning before application execution. It outlines the process of selecting prefetchable data and generating the necessary on-chip memory configuration information for the on-chip memory controller, thus completing the overall workflow.

The initial step involves recognizing candidates for prefetchable data mapped to the SPM path, with the cumulative size of the candidates being calculated. During this process, the compiler evaluates the address, size, and stride of each memory access, recognizing those that conform to the two-dimensional affine access pattern as prefetchable candidates. Simultaneously, the remaining data are classified as non-prefetchable and mapped to the cache path.

To safeguard the integrity of prefetchable data from modification by non-prefetchable data during execution, the cache ways containing prefetchable data are configured as SPM and excluded from cache lookup and replacement. However, since non-prefetchable data must be accessed via the cache, it is essential to ensure that at least one cache way remains active during the cache lookup phase and possesses proactive global addressing capability. In scenarios where the total size of prefetchable data is excessively large, we employ an approach that combines memorization search with knapsack dynamic programming ideas. Initially, the compiler explores all potential scenarios through dynamic programming and subsequently traces back to identify the scenario that maximizes the prefetchable data size. Additionally, it backtracks to label all prefetchable data candidates within this optimal scenario.

Ultimately, by evaluating the size of prefetchable data, the compiler determines the number of cache ways that must be configured as SPM and generates the cache ways control information for the on-chip memory controller.

4 Decentralized Scheduling

In this section, we first introduce dynamic task migration and stealing strategies based on PE state to better achieve workload balancing through on-demand real-time task scheduling for busy and idle PEs.

Subsequently, we construct a decentralized PE microarchitecture with the following goal: empowering PEs to autonomously and proactively schedule their internal tasks to improve hardware

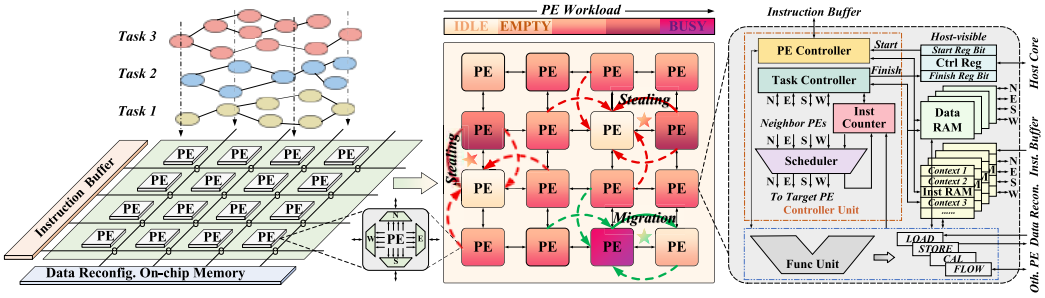


Fig. 7. Workload-based decentralized dataflow-driven task migration/stealing strategies and decentralized PE microarchitecture.

flexibility. Here we describe our decentralized PE microarchitecture implementation and the overall procedure flow based on decentralized PE microarchitecture.

Finally, we propose a decentralized load balancing mechanism. It effectively integrates the aforementioned strategies with hardware microarchitecture in the context of multiple tasks to enhance computational resource utilization.

4.1 Workload-Based Task Stealing & Migration

Task Model. Multiple tasks are assigned to the PE array, with each task being executed collectively by one or more PEs. Each task is represented as a **dataflow graph (DFG)**, as illustrated in Figure 7 (left). Consequently, a single PE may handle segments of various DFGs, and all segments of a DFG assigned to a specific PE are termed as a context. In addition, we use the application-based multi-DFG program execution model [19], where an application consists of multiple sequentially executing tasks. Each task is a DFG comprising multiple computational nodes and directed edges, and it may be executed iteratively multiple times.

First, the various workload states of the PE are delineated, as shown in Figure 7 (middle). Specifically, we define three key states: *IDLE*, *EMPTY*, and *BUSY*. A PE transitions to the *IDLE* state once it has executed all its assigned contexts. Similarly, a PE transitions to the *EMPTY* state upon executing the final context, signifying that there are no remaining contexts. Conversely, a PE enters the *BUSY* state when the number of remaining contexts surpasses the busy threshold. This threshold is dynamically adjusted based on the number of *EMPTY* and *IDLE* PEs present on the current array.

Second, novel workload balancing strategies tailored for the decentralized dataflow architecture are introduced. The objective is to optimize workload balancing by ensuring that more PEs are in a balanced state rather than being in a high-workload (*BUSY*) or low-workload (*IDLE*) state. To achieve this, two workload balancing strategies are proposed: task migration for *BUSY* PEs and task stealing for *IDLE* PEs. It facilitates better workload balancing through on-demand real-time task scheduling for *BUSY* and *IDLE* PEs.

To better support decentralized dataflow-driven task stealing and migration, a decentralized PE microarchitecture is constructed, as illustrated in Figure 7 (right). Detailed elaboration on this microarchitecture will follow in the next section.

4.2 Decentralized PE Microarchitecture

We divide the execution process of a multi-DFG application into seven stages: Application Start, Initialization, Execution, Workload Sensing, Trade-off, Stealing/Migration, and Application Finish. To facilitate the execution of these stages, we design the decentralized PE microarchitecture. This

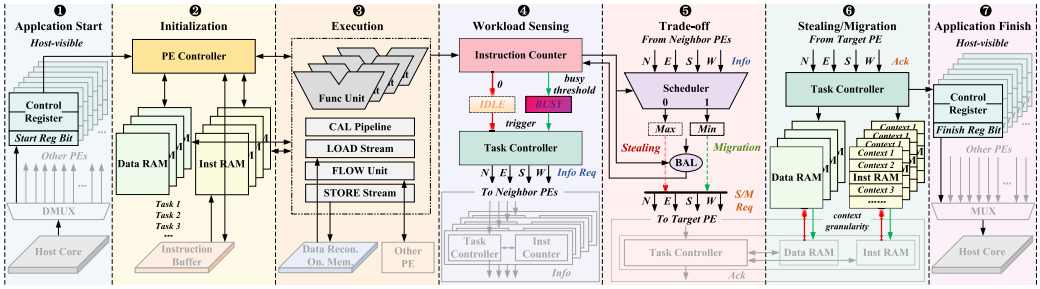


Fig. 8. Overall procedure flow (seven stages) based on decentralized PE microarchitecture.

microarchitecture is primarily composed of three modules: Controller Unit, Memory Unit, and Function Unit. Next, we provide a detailed introduction to each of these modules. Lastly, we elucidate the overall procedure flow (seven stages) based on decentralized PE, as depicted in Figure 8.

Controller Unit. This unit comprises four components, including a PE controller to initialize configuration information and instructions, a task controller to manage workload sensing and stealing/migration, an instruction counter to record the remaining contexts and instructions of the PE, and a scheduler to make workload trade-off decisions.

Memory Unit. This unit encompasses a host-visible control register for the PE to interact with the host core, a data RAM to store data, and an instruction RAM to store instructions.

Function Unit. This unit includes a calculation pipeline to generate the data path for arithmetic operations and logical operations, a load/store stream to transfer data from/to the data reconfigurable on-chip memory, and a flow unit to dispatch data to downstream PE, decoupling the execution of a DFG node into up to four consecutive phases: Load, Calculating, Flow and Store.

Overall Procedure Flow (Seven Stages):

❶ *Application Start.* The host core triggers the start of the application by setting the start register bit of the control register of each PE to 1, indicating the start of execution. This signal prompts the PEs to begin their respective tasks.

❷ *Initialization.* Upon receiving the start signal from the host core, each PE retrieves necessary instructions and configuration information from the instruction buffer, storing the instructions in the PE internal instruction RAM and configuring the data RAM and function units under the control of the PE controller.

❸ *Execution.* PEs commence executing the application program, which involves loading/storing data from/to the data reconfigurable on-chip memory, dispatching data to downstream PE, and dynamically updating the number of remaining contexts and instructions in the instruction counter.

❹ *Workload Sensing.* When the number of remaining contexts and instructions reaches 0 or surpasses the busy threshold, the PE transitions to the *IDLE* or *BUSY* state and initiates workload sensing. It proactively sends a request for workload information to neighbor PEs through the task controller. Upon receiving the signal, the neighbor PEs transmit the remaining workload information to the *IDLE*/*BUSY* PE.

❺ *Tradeoff:* Following the reception of workload information from neighbor PEs, the scheduler within the *IDLE*/*BUSY* PE compares these information. It selects the PE with the maximum/minimum remaining workload as the target for task stealing/migration and tradeoffs based on the remaining workloads of both the *IDLE*/*BUSY* PE and the target PE. Subsequently, the scheduler determines the number of workloads to be stolen/migrated and sends the task stealing/migration request signal accordingly.

❻ *Stealing/Migration:* Once the target PE receives the task stealing/migration request signal, its task controller responds with an acknowledgment signal to the *IDLE*/*BUSY* PE. Concurrently, it manages the instruction RAM and data RAM to facilitate the transfer of instructions and related data with the *IDLE*/*BUSY* PE. Upon receiving the acknowledgment signal, the *IDLE*/*BUSY* PE coordinates the transfer of instructions and data with the target PE, ensuring seamless task stealing or migration. Note that both task stealing and migration operate at the granularity of a context. Furthermore, to ensure proper execution, task stealing/migration must target inactive contexts, which are contexts within a PE that have not yet begun execution.

❼ *Application Finish.* Upon completing the execution of all tasks assigned to it, each PE sets the finish register bit of its control register to 1. Simultaneously, the host core periodically polls the control registers of all PEs. Upon detecting that the finish register bits of all PEs have been set to 1, the host core terminates the execution of the application.

Synchronization. The dataflow-driven execution model in dataflow architectures naturally reduces load/store pressure and pollution potential in on-chip memory by enabling direct data transfer between PEs. Tasks are activated only when the necessary data from upstream PEs are available, implicitly ensuring write serialization and propagation across PEs. Initially, during task allocation, each PE receives its assigned tasks along with configuration information that records upstream and downstream data-dependent tasks and their associated PEs. This allows the PE to send activation signals and data to downstream PEs, as well as acknowledgment signals to upstream PEs upon receiving the required data. Additionally, in dynamic scheduling scenarios, including task stealing and migration, synchronization and data consistency are critical. These processes involve not only transferring relevant data, instructions, and configuration information from the source PE to the destination PE, but also updating the configuration information of upstream and downstream PEs to reflect the new PE identity of the stolen or migrated task. This ensures the correctness of the dataflow-driven execution and maintains synchronization across PEs.

4.3 Decentralized Dataflow-Driven Workload Balancing

Traditional static mapping faces challenges in achieving efficient workload balancing, which motivates our proposal for decentralized dynamic scheduling. PANDA adopts the *DFG balancing method* [19], which is a heuristic static mapping approach aimed at achieving load balancing through instruction mapping among DFG nodes. However, it is hard to generate an absolutely balanced DFG because: (1) the delay of each node during execution is unpredictable, particularly due to stalls caused by hazards or memory access; (2) allocating an equal number of instructions to each DFG node is costly and constrained by the applications itself, which often leads to non-convergence. To address these limitations, we propose decentralized dynamic scheduling to complement the static mapping. It's worth noting that the DFG balancing method can produce a relatively balanced load distribution across the PEs on the array. While potential workload imbalances may occur and are difficult to predict, the likelihood of neighboring PEs in a part of the array being in the same status (either *BUSY* or *IDLE*) is quite rare.

In a scenario where T_n tasks are mapped to an array of P_n PEs, workload balancing in the decentralized dataflow architecture entails two crucial steps: workload balancing trigger and workload balancing implementation.

Initially, PEs in the *imbalance* state trigger the workload balancing process. There are two types of *imbalance* states for PEs: *IDLE* (low-workload state) and *BUSY* (high-workload state). Accordingly, the PE in the *imbalance* state is formulated as follows:

$$\begin{aligned} \xi(imbalance) &= \xi(idle) \cup \xi(busy) \\ \xi(idle) \cap \xi(busy) &= \varnothing, \end{aligned} \tag{1}$$

where the PE in the *IDLE* state is formulated as follows:

$$\xi(idle) \neq \varnothing, \text{ if } C_n = 0 \text{ and } I_n = 0, \tag{2}$$

where the PE in the *BUSY* state is formulated as follows:

$$\xi(busy) \neq \varnothing, \text{ if } C_n > T_b, \tag{3}$$

where C_n represents the number of remaining contexts of the PE, I_n denotes the number of remaining instructions of the context being executed by the PE, T_b signifies the busy threshold, and $\xi(m)$, being non-empty, indicates that the PE is in the m state.

Additionally, the PE in the *EMPTY* state is formulated as follows:

$$\xi(empty) \neq \varnothing, \text{ if } C_n = 0 \text{ and } I_n \neq 0. \tag{4}$$

Whenever a PE transitions to the *IDLE* state, the busy threshold T_b is dynamically updated based on the number of *IDLE* and *EMPTY* PEs ($P_{n,i}$ and $P_{n,e}$) on the current array, as shown in Equation (5). The update depends on the number of *IDLE* and *EMPTY* PEs, as in both scenarios, the number of remaining contexts is 0. The task controller within each PE is responsible for storing and updating the busy threshold. Through this on-demand global threshold adjustment mechanism, we introduce a global hint to the decentralized scheduling, reducing the likelihood of prolonged idle states and mitigating the impact of local minima.

$$T_b = T_n \times \frac{P_{n,i} + P_{n,e}}{P_n} \tag{5}$$

Subsequently, workload balancing is achieved through task stealing for *IDLE* PEs and task migration for *BUSY* PEs.

The number of contexts obtained by the *IDLE* PE through task stealing, denoted as S_n, is given by Equation (6). The principle guiding task stealing for *IDLE* PEs is to alleviate the business of the neighbor PE with maximum remaining workload as much as possible within their capacity.

$$S_n = C_{n,t} - T_b \tag{6}$$

where $C_{n,t}$ represents the number of remaining contexts of the target PE for task stealing (the neighbor PE with maximum remaining workload).

The number of contexts lost by the *BUSY* PE through task migration, denoted as M_n, is given by Equation (7). The principle guiding task migration for *BUSY* PEs is to transition out of the *BUSY* state by migrating tasks to the neighbor PE with minimum remaining workload. Under this principle, if a situation arises where the remaining workload of a single PE is very high, the excessive workload will eventually be shared by the low-workload PEs on the array through continuous task migration. This process ensures more efficient utilization of resources across the array.

$$M_n = C_{n,b} - T_b \tag{7}$$

where $C_{n,b}$ represents the number of remaining contexts of the *BUSY* PE.

5 Evaluation Methodology

Setup. We implement all modules of PANDA using Verilog. We synthesize the design using Synopsys Design Compiler with a commercial 28 nm technology, meeting timing at a 1.25 GHz frequency. We use Synopsys PrimeTime PX for accurate power analysis. Table 3 shows the hardware design parameters, as well as the evaluated area and power breakdown of essential modules in PANDA. Evaluation shows that PANDA has an area footprint of 1.142 mm^2 in a 28 nm process, and consumes a maximum power of 1.217 W at 1.25 GHz.

Table 3. Design Parameters and Area and Power Breakdown (28nm)

Component (Parameter)		Area (mm^2)	Power (mW)
PE	**Function Unit (decoupled execution)**	0.016 (1.40%)	13.8 (1.13%)
	Controller Unit (decentral. control)	0.007 (0.61%)	12.9 (1.06%)
	Memory Unit (1KB inst/data RAM)	0.013 (1.14%)	16.4 (1.35%)
PE Array (4×4 decentralized PEs, SIMD)		0.574 (50.3%)	690.0 (56.7%)
Network on Chip (16 routers, 2D mesh)		0.098 (8.6%)	135.1 (11.1%)
Data Reconfig. On. Mem. (64KB total size)		0.349 (30.5%)	283.6 (23.3%)
Instruction Buffer (16KB total size)		0.121 (10.6%)	108.3 (8.9%)
PANDA (Total)		***1.142***	***1217***

Benchmarks. To evaluate our design, we select several real-world applications from REVEL [56], Plasticine [43], DFU [19], and MTDE [7]. For our experimentation, we deliberately opted for dual scales for each application. In the **small-scale (S)** scenario, the aggregate data size essential for application execution remains below the limits of the on-chip memory capacity. Conversely, the **large-scale (L)** case entails data size that surpasses the available on-chip memory capacity. We also assess the effectiveness of our method on graph applications where the number of edges equals (E) the number of vertices. In such graphs, each vertex typically has few neighbors. Additionally, we classify these applications as follows, based on their memory access and computation patterns. *(1) Regular Applications:* GEMM, STE, CONV, and FIR. These applications exhibit highly regular and fixed access and computation patterns. *(2) Near-Regular Applications:* SHA, ReLU, and SVD. These applications demonstrate relatively regular memory access and computation patterns but contain a small amount of discrete, small-scale parameter accesses and imperfect loops. *(3) Near-Irregular Applications:* FFT, LAG, and CHO. These applications feature complex hopping computations and include some data with a large span of memory access addresses. *(4) Irregular Applications:* BFS and PR. These applications involve imbalanced workloads, large amounts of random memory accesses across diverse memory regions, and redundant computations resulting from the traversal irregularity of edges and the update irregularity of vertex property [58]. Table 4 shows a list of applications and their characteristics.

Comparison Methodology. To quantify the performance of our design, we compare it to four state-of-the-art dataflow designs: REVEL [56], Plasticine [43], DFU [19], and MTDE [7]. REVEL comprises simple PEs in a systolic form and dataflow PEs capable of performing complex computations. It features hybrid static-dynamic task issuing and static task mapping. Plasticine decouples dedicated pattern compute units and memory units. It features static task issuing and static task mapping. DFU is a reconfigurable dataflow architecture for multi-batch data processing. It employs dynamic task issuing and static task mapping. MTDE is a multi-task dataflow engine that supports the elastic task scheduling scheme, featuring dynamic task issuing and dynamic task mapping. For fairness, these designs are extended to achieve similar peak performance and process. The clock frequencies are configured based on the specifications from their original articles.

6 Experimental Results

First, we evaluate the overall performance of PANDA architecture compared to the traditional dataflow architecture. Second, we evaluate the dataflow architecture of adaptive prefetching. In order to verify the optimization effect on the adaptive prefetching, for a fair comparison, we do

Table 4. Profile of Applications

Application	Domain	Data Sizes		Feature
		small	large	
GEMM	Scientific Computing	64×64	128×128	R
Stencil-3d		16×32×4	16×32×25K	R
SHA-256		8K Points	1M Points	NR
Conv-3d	Artificial Intelligence	8×8-32	32×32-32	R
ReLU-3d		8×8-32	32×32-32	NR
SVD		32×32	128×128	NR
FIR	Digital Signal Processing	8K Points	1G Points	R
Cholesky		32×32	128×128	NIR
FFT		1K Points	8K Points	NIR
Lagrange		128 Points	4K Points	NIR
BFS	Graph Processing	|V|=1K |E|=8K	|V|=9K |E|=5M,9K	IR
PageRank		|V|=1K |E|=8K	|V|=9K |E|=5M,9K	IR

In the table, all data types are 32-bit. R/NR/NIR/IR is short for regular, near-regular, near-irregular, and irregular.

not consider the decentralized scheduling. Subsequently, we evaluate the dataflow architecture of decentralized scheduling. Similarly, we do not consider the adaptive prefetching. In addition, we compare PANDA to four state-of-the-art dataflow designs: REVEL, Plasticine, DFU, and MTDE. Finally, we evaluate the overhead of the adaptive prefetching and decentralized scheduling.

6.1 Overall Performance

To evaluate the effectiveness of the methods we proposed, we implement the following four different configurations:

- *Base:* Our baseline is a dataflow architecture with full prefetching (SPM-only) and centralized scheduling.
- *Base + AP:* It only combines the baseline architecture with **A**daptive **P**refetching.
- *Base + DS:* It only combines the baseline architecture with **D**ecentralized **S**cheduling.
- *Base + AP + DS:* It combines the baseline architecture with **A**daptive **P**refetching and **D**ecentralized **S**cheduling.

Figure 9 demonstrates the effectiveness of our proposed methods in terms of performance improvements. **Adaptive prefetching (AP)** achieves an average performance improvement of 1.13× (small-scale) and 1.65× (large-scale) over the baseline, while decentralized scheduling (DS) achieves an average performance improvement of 1.40× (small-scale) and 1.67× (large-scale) over the baseline. The performance benefits of AP for workloads with irregular memory access patterns, such as BFS and PR, stem from the introduction of a cache suitable for handling the dynamic and unpredictable data access typical of such workloads. By combining the adaptive prefetching and decentralized scheduling, the performance of the dataflow architecture can be improved by 1.49× (small-scale) and 2.01× (large-scale). Moreover, for graph applications where the number of edges equals the number of vertices, the adaptive prefetching approach improves performance by 2.48× for BFS and 2.36× for PR, compared to full prefetching (SPM-only) method. Subsequently, we compare the performance gains obtained by each innovation point separately.

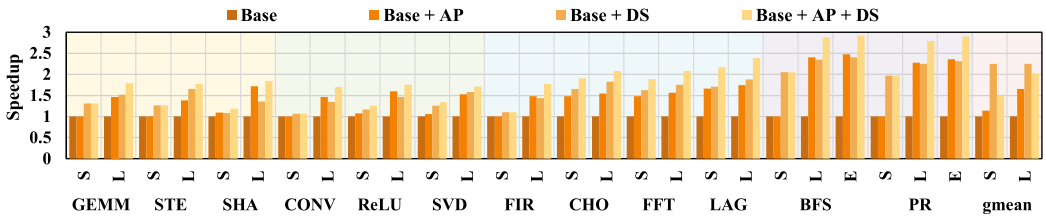


Fig. 9. Overall performance.

6.2 Advantages of Adaptive Prefetching

Our PANDA architecture adopts data reconfigurable on-chip memory to achieve adaptive data prefetching. To evaluate the advantages of adaptive prefetching, we replaced the data reconfigurable on-chip memory with DMA-assisted SPM. We also compared it to a cache-only configuration. Overall, we evaluate the following configurations:

- *SPM:* SPM with DMA support. SPM configuration: 64 KB total size, 16 banks, 128 lines, and 32B line size. DMA configuration: 400 ns initialization time.
- *Cache:* 64KB total size, 8-way set associativity, 128 sets, 64B line size, LRU replacement and write back.
- *PANDA MEM:* 64 KB total size, reconfigurable on-chip memory, SPM and cache logical organizations, SPM and cache shared SRAM.

Figure 10 shows the access latency for our 12 benchmarks with both large and small data sizes using DMA-assisted SPM, cache, and PANDA MEM, normalized to SPM. Compared to SPM with DMA, PANDA MEM effectively mitigates memory access latency, resulting in reductions of 18.2% (small-scale) and 55.1% (large-scale) on average. Additionally, PANDA MEM achieves an average of 47.0% (small-scale) and 16.7% (large-scale) latency improvement over the cache. Subsequently, we analyze these results in more detail.

For small-scale regular applications like small-scale GEMM, STE, CONV, and FIR, where all data are prefetchable, both PANDA MEM and SPM demonstrate equivalent low access latency. In contrast, the cache exhibits high access latency due to its lack of the bulk data prefetching mechanism. Similarly, in cases of small-scale SHA, ReLU, and SVD with a minor proportion of non-prefetchable data, PANDA MEM slightly outperforms SPM in optimizing access latency (by 12.1%, 9.9%, and 8.3%, respectively). Conversely, in scenarios where the proportion of non-prefetchable data is not negligible, PANDA MEM exhibits notable access latency reductions: 44.8% (small-scale FFT), 51.6% (large-scale FFT), 54.5% (small-scale LAG), 55.1% (large-scale LAG), 49.8% (small-scale CHO), 53.8% (large-scale CHO), and 55.1% (large-scale geomean). For irregular applications such as BFS and PR, where tasks are dynamically generated and activated during execution, SPM shows low access latency in small-scale scenarios but high access latency in large-scale scenarios, whereas cache and PANDA MEM show consistently low latency in both scenarios. The performance discrepancy of SPM in irregular applications primarily arises from its lack of the dynamic global addressing capability. In small-scale BFS and PR, where the data size fits within the on-chip memory and all data are prefetchable, SPM demonstrates low access latency. However, in large-scale BFS and PR, where the data size surpasses on-chip memory capacity and the task splitting is more inefficient compared to regular applications like GEMM, SPM shows high access latency. These results highlight the effectiveness of PANDA MEM in reducing access latency compared with SPM, particularly for applications with irregular access patterns. Compared to cache, PANDA MEM benefits from direct addressability to prefetchable data and the ability to efficiently prefetch bulk data. This gives

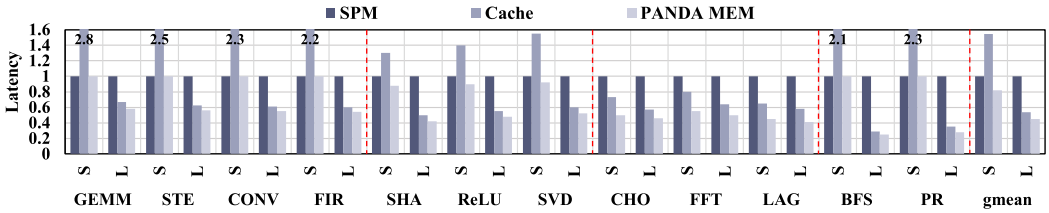


Fig. 10. Access latency comparison for SPM, cache, and PANDA MEM.

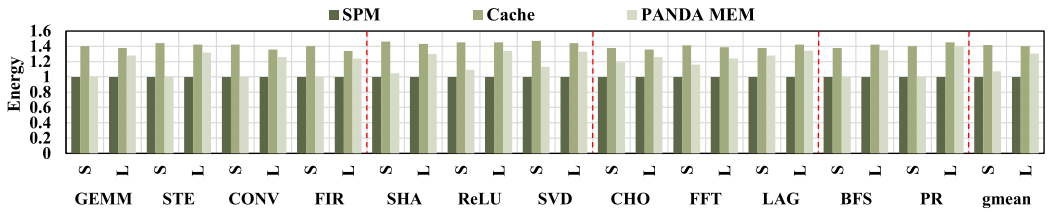


Fig. 11. Access energy comparison for SPM, cache, and PANDA MEM.

PANDA MEM a distinct edge in scenarios with substantial amounts of prefetchable data, leading to decreased access latency in such cases.

Figure 11 shows the access energy for our 12 benchmarks with both large and small data sizes using DMA-assisted SPM, cache, and PANDA MEM, normalized to SPM. We observe that the access energy of PANDA MEM exhibits nuanced behavior. When contrasted with SPM, PANDA MEM consumes more energy, with increases of 7.1% and 30.4% observed for small and large data sizes, respectively. In contrast, when compared to the cache, PANDA MEM exhibits energy advantages, achieving reductions of 32.2% and 7.7% for small and large data sizes, respectively. SPM has less hardware overhead than PANDA MEM and does not need to load non-prefetchable data from off-chip memory during execution. Compared to cache, PANDA MEM greatly reduces the need for tag lookups and minimizes conflict misses when accessing prefetchable data. Hence, PANDA MEM exhibits higher access energy than SPM but lower than cache.

6.3 Advantages of Decentralized Scheduling

Our PANDA architecture adopts decentralized dataflow PEs to achieve decentralized scheduling. To evaluate the advantages of decentralized scheduling, we replaced decentralized dataflow PEs with ordinary dataflow PEs and a centralized controller. Overall, we evaluate the following configurations:

- *Base PE:* 4×4 ordinary dataflow PEs under the control of a centralized controller, centralized scheduling.
- *PANDA PE:* 4×4 decentralized dataflow PEs, no centralized controller, decentralized scheduling.

Figure 12 shows the performance (in marker) and the computational resource utilization (in bar) comparison between the Base PE and PANDA PE, normalized to the centralized dataflow architecture. Compared to the Base PE, PANDA PE achieves an average performance improvement of 1.40× (small-scale) and 1.67× (large-scale). Moreover, PANDA PE achieves 1.23× (small-scale) and 1.34× (large-scale) utilization improvement over the Base PE. Further detailed analysis of these results follows.

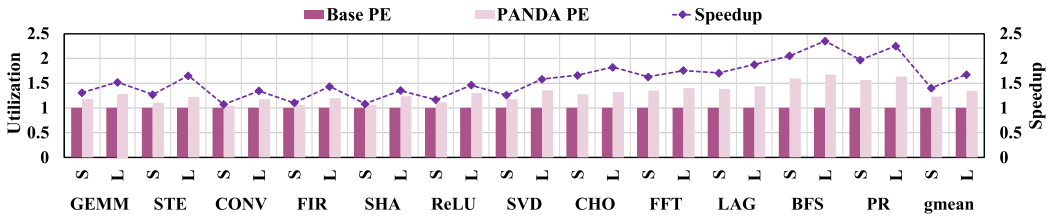


Fig. 12. Speedup (in Marker) and utilization (in Bar) comparisons for base PE and PANDA PE.

For regular and near-regular applications such as GEMM, STE, and SVD, which exhibit regular computational patterns and have limited optimization space for workload balancing in parallel processing, the relatively small performance gains primarily stem from the reduced communication and routing overhead between task iterations enabled by decentralized control. Additionally, as the data size and the number of task iterations increase, the performance improvement becomes more pronounced. For near-irregular applications such as FFT, LAG, and CHO, in addition to reduced communication overhead, performance gains are achieved through dynamic workload adjustment during execution. This adjustment allows for the combination of non-bottleneck tasks, overlapping the high latencies from frequent nested memory access and bottleneck tasks, thereby achieving higher performance. For irregular applications such as BFS and PR, where tasks are dynamically generated and activated during execution, PANDA PE not only maintains the above advantages compared to the Base PE but also enables more efficient dynamic workload adjustments during execution by improving hardware flexibility, resulting in better performance improvements. Similarly, the utilization improvement of PANDA PE compared to the Base PE is mainly attributed to reduced communication and routing overheads due to decentralized control and better workload balancing facilitated by increased hardware flexibility.

6.4 PANDA Outperforms State-of-the-Art Architectures

For comparison, we use four typical dataflow architectures: REVEL [56], Plasticine [43], DFU [19], and MTDE [7]. Table 5 shows the hardware comparisons of our architecture with other architectures.

Figure 13 illustrates the performance (in marker) comparison normalized to REVEL. Our design achieves an average performance improvement of 2.53× over REVEL, 1.90× over Plasticine, 1.38× over DFU, and 1.19× over MTDE. REVEL combines systolic PEs and dataflow PEs for hybrid-dataflow-driven scheduling, which allows better exploitation of application parallelism. However, it heavily relies on the compiler's heuristic search for the globally optimal timing decision due to the consideration of balancing between dataflow PEs and systolic PEs, limiting its performance gains. Plasticine employs non-dataflow-driven scheduling, focusing on a high-width SIMD design and decoupling compute and memory units to enhance performance by overlapping computation execution time and memory access latency. DFU enhances performance through full dataflow-driven scheduling and decouples computation into four pipeline stages, achieving higher levels of parallelism. These optimizations, however, are more effective for regular and near-regular applications, offering limited benefits for near-irregular and irregular applications due to the static scheduling employed by REVEL, Plasticine, and DFU, which lacks dynamic workload adjustment during execution. In contrast, MTDE and our proposed architecture, PANDA, employ dynamic scheduling, significantly improving performance, especially for near-irregular and irregular applications. PANDA stands out by adopting a decentralized design and utilizing highly flexible PEs for

Table 5. Hardware Comparison with State-of-the-art Architectures

Arch.	Scheduling Approach	On-chip Memory	Tech.	Area	Max. Power	Frequency	Peak Performance	Energy Efficiency
REVEL	Static + Hybrid	128KB (Shared SPM)	28 nm	2.709 mm^2	2.282 W	1.25 GHz	320 GFLOPS	25.99~137.29 GFLOPS/W
Plasticine	Static + Non-dataflow.	288KB (9 PMUs)	28 nm	3.760 mm^2	1.933 W	1 GHz	410 GFLOPS	59.25~150.86 GFLOPS/W
DFU	Static + Cen. Dataflow.	512KB(D)+128KB(C)	28 nm	10.28 mm^2	2.040 W	1.25 GHz	320 GFLOPS	88.06~154.25 GFLOPS/W
MTDE	Dynamic + Cen. Dataflow.	256KB (4-bank SPM)	28 nm	2.816 mm^2	2.104 W	0.8 GHz	410 GFLOPS	89.15~187.67 GFLOPS/W
PANDA	Dynamic + Dec. Dataflow.	64KB(D)+16KB(I)	28 nm	1.142 mm^2	1.217 W	1.25 GHz	320 GFLOPS	95.44~234.83 GFLOPS/W

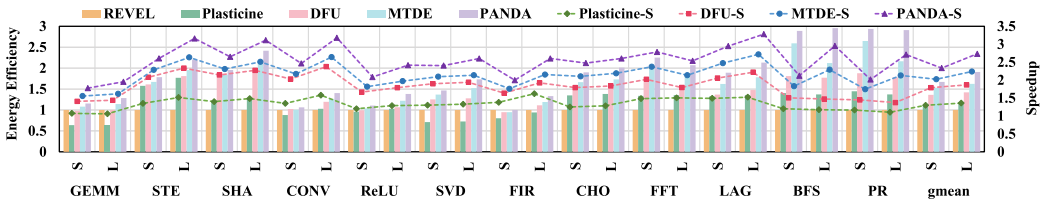


Fig. 13. Speedup (in marker) and energy efficiency (in Bar) comparisons with state-of-the-arts dataflow architectures.

more efficient dynamic scheduling. Additionally, it incorporates an adaptive prefetching approach that reduces memory access latency, further boosting performance compared to MTDE.

Figure 13 shows the energy efficiency (in bar) comparison normalized to REVEL. On average, our design achieves 1.79× energy efficiency improvement over REVEL, 1.60× over Plasticine, 1.28× over DFU, and 1.13× over MTDE. Task mapping significantly influences resource utilization. REVEL, Plasticine, and DFU, with static task mapping, generally exhibit lower utilization compared to MTDE and PANDA, which employ dynamic task mapping. Plasticine, with its high-width SIMD design, achieves high peak performance but demonstrates lower utilization when targeting small-scale applications, thus limiting energy efficiency in these scenarios. REVEL can fully exploit the hardware flexibility of its hybrid systolic-dataflow execution mechanism when applied to well-parallelized applications, achieving high utilization and energy efficiency in these scenarios. DFU, with its reconfigurable interconnections and decoupled execution mechanism, maintains relatively stable utilization and energy efficiency improvements across applications of different data sizes. However, due to the limitations of static mapping and rigid full prefetching, these architectures are less energy efficient for irregular and large-scale applications. MTDE employs dynamic mapping and achieves significant energy efficiency gains, particularly for irregular applications. PANDA also uses dynamic mapping and has a decentralized design. Compared to the centralized control of MTDE, decentralized control can significantly reduce the communication and routing overheads associated with dynamic scheduling and control, further improving energy efficiency.

6.5 Overhead

Finally, we evaluate the hardware overhead of the adaptive prefetching and decentralized scheduling. First is the adaptive prefetching. The hardware overhead of PANDA MEM is shown in Table 6, and PANDA MEM exhibits a total area overhead of 37.4% and 12.9% increase over SPM with DMA and cache of the same size, respectively. Next is the decentralized scheduling. The hardware overhead of PANDA PE is shown in Table 7, and PANDA PE exhibits a total area overhead of 6.5% increase over the Base PE of the same computing fabric.

7 Conclusion

In this article, we describe PANDA, a novel dataflow architecture with application-adaptive data prefetching and decentralized dataflow-driven dynamic scheduling. We implement PANDA in RTL

Table 6. Hardware Comparison of SPM, Cache, and PANDA MEM

	Component	Area (mm^2)	Power (mW)
SPM	**RAM Banks (64KB)**	0.212 (80.5%)	163.9 (81.1%)
	DMA & Control Logic	0.042 (16.5%)	39.6 (19.5%)
	Total	0.254	203.5
Cache	**Data Arrays (64KB)**	0.212 (68.6%)	163.9 (64.2%)
	Tag & Control Logic	0.097 (31.4%)	91.3 (35.8%)
	Total	0.309	255.2
PANDA MEM	**Shared RAM (64KB)**	0.212 (60.7%)	163.9 (57.8%)
	Control Logic	0.137 (39.3%)	119.7 (41.8%)
	Total	0.349	283.6

Table 7. Hardware Comparison of Base PE and PANDA PE

	Component	Area (mm^2)	Power (mW)
Base PE	**PEs (16 ordinary)**	0.478 (88.7%)	486.6 (74.7%)
	Controller	0.061 (11.3%)	165.1 (25.3%)
	Total	0.539	651.7
PANDA PE	**PEs (16 decentralized)**	0.574 (100%)	690.0 (100%)
	Controller (removed)	0 (0%)	0 (0%)
	Total	0.574	690.0

design and demonstrate that in a wide range of real-world applications, including scientific computing, artificial intelligence, digital signal processing, and graph processing algorithms, PANDA attains up to 2.53× performance and 1.79× energy efficiency improvement over the state-of-the-art dataflow architectures. Looking ahead, we will explore ways to efficiently support indirect memory access prefetching within our existing architecture. Our future work will focus on designing a low-cost hardware prefetcher to capture dynamic runtime information and optimizing the PANDA (software-aided) prefetching method through hardware-software co-design.

References

[1] Giovanni Ansaloni, Kazuyuki Tanimura, Laura Pozzi, and Nikil Dutt. 2012. Integrated kernel partitioning and scheduling for coarse-grained reconfigurable arrays. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 31, 12 (2012), 1803–1816. https://doi.org/10.1109/TCAD.2012.2209886

[2] ARM. 2023. Armv8-M Memory Model and Memory Protection User Guide. Retrieved August 7, 2024 from https://developer.arm.com/documentation/107565/0101/

[3] Bahar Asgari, Ramyad Hadidi, and Hyesoon Kim. 2020. ASCELLA: Accelerating sparse computation by enabling stream accesses to memory. In *Proceedings of the 23rd Conference on Design, Automation and Test in Europe (DATE'20)* (Grenoble, France). EDA Consortium, San Jose, CA, 318–321. https://doi.org/10.23919/DATE48585.2020.9116501

[4] Rajeshwari Banakar, Stefan Steinke, Bo-Sik Lee, Mahesh Balakrishnan, and Peter Marwedel. 2002. Scratchpad memory: A design alternative for cache on-chip memory in embedded systems. In *Proceedings of the 10th International Symposium on Hardware/Software Codesign (CODES 2002) (IEEE Cat. No.02TH8627)*. 73–78. https://doi.org/10.1145/774789.774805

[5] Brent Bohnenstiehl, Aaron Stillmaker, Jon J. Pimentel, Timothy Andreas, Bin Liu, Anh T. Tran, Emmanuel Adeagbo, and Bevan M. Baas. 2017. KiloCore: A 32-nm 1000-Processor computational array. *IEEE Journal of Solid-State Circuits* 52, 4 (2017), 891–902. https://doi.org/10.1109/JSSC.2016.2638459

[6] Doug Burger, James R. Goodman, and Alain Kägi. 1996. Memory bandwidth limitations of future microprocessors. In *Proceedings of the 23rd Annual International Symposium on Computer Architecture* (Philadelphia, PA) *(ISCA'96)*. ACM, New York, 78–89. https://doi.org/10.1145/232973.232983

[7] Longlong Chen, Jianfeng Zhu, Yangdong Deng, Zhaoshi Li, Jian Chen, Xiaowei Jiang, Shouyi Yin, Shaojun Wei, and Leibo Liu. 2021. An elastic task scheduling scheme on coarse-grained reconfigurable architectures. *IEEE Transactions on Parallel and Distributed Systems* 32, 12 (2021), 3066–3080. https://doi.org/10.1109/TPDS.2021.3084804

[8] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. 2014. DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. In *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'14)* (Salt Lake City, Utah, USA). ACM, New York, 269–284. https://doi.org/10.1145/2541940.2541967

[9] Tao Chen and G. Edward Suh. 2016. Efficient data supply for hardware accelerators with prefetching and access/execute decoupling. In *Proceedings of the 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-49)* (Taipei, Taiwan). IEEE Press, Article 46, 12 pages. https://doi.org/10.1109/MICRO.2016.7783749

[10] Yu-Hsin Chen, Joel Emer, and Vivienne Sze. 2016. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. In *Proceedings of the 43rd International Symposium on Computer Architecture (ISCA'16)* (Seoul, Republic of Korea). IEEE Press, 367–379. https://doi.org/10.1109/ISCA.2016.40

[11] Hsiang-Yun Cheng, Chung-Hsiang Lin, Jian Li, and Chia-Lin Yang. 2010. Memory latency reduction via thread throttling. In *Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'43)*. IEEE Computer Society, 53–64. https://doi.org/10.1109/MICRO.2010.39

[12] Vidushi Dadu and Tony Nowatzki. 2022. TaskStream: Accelerating task-parallel workloads by recovering program structure. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (Lausanne, Switzerland) *(ASPLOS'22)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3503222.3507706

[13] Vidushi Dadu, Jian Weng, Sihao Liu, and Tony Nowatzki. 2019. Towards general purpose acceleration by exploiting common data-dependence forms. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'52)* (Columbus, OH). ACM, New York, 924–939. https://doi.org/10.1145/3352460.3358276

[14] Jinyi Deng, Xinru Tang, Jiahao Zhang, Yuxuan Li, Linyun Zhang, Boxiao Han, Hongjun He, Fengbin Tu, Leibo Liu, Shaojun Wei, Yang Hu, and Shouyi Yin. 2023. Towards efficient control flow handling in spatial architecture via architecting the control flow plane. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'23)* (Toronto, ON, Canada). ACM, New York, 1395–1408. https://doi.org/10.1145/3613424.3614246

[15] Robert H. Dennard, Fritz H. Gaensslen, Hwa-Nien Yu, V. Leo Rideout, Ernest Bassous, and Andre R. LeBlanc. 1974. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE Journal of Solid-State Circuits* 9, 5 (1974), 256–268. https://doi.org/10.1109/JSSC.1974.1050511

[16] Jack B. Dennis. 1974. First version of a data flow procedure language. In *Programming Symposium*, B. Robinet (Ed.). Springer Berlin, Berlin, 362–376. https://doi.org/10.1007/3-540-06859-7_145

[17] Zhihua Fan, Wenming Li, Shengzhong Tang, Xuejun An, Xiaochun Ye, and Dongrui Fan. 2023. Improving utilization of dataflow architectures through software and hardware co-design. In *: Parallel Processing: Proceedings of the 29th International Conference on Parallel and Distributed Computing (Euro-Par 2023)* (Limassol, Cyprus, August 28 - September 1)). Springer-Verlag, Berlin, 245–259. https://doi.org/10.1007/978-3-031-39698-4_17

[18] Zhihua Fan, Wenming Li, Zhen Wang, Tianyu Liu, Haibin Wu, Yanhuan Liu, Meng Wu, Xinxin Wu, Xiaochun Ye, Dongrui Fan, et al.. 2023. Accelerating convolutional neural networks by exploiting the sparsity of output activation. *IEEE Transactions on Parallel and Distributed Systems* 34, 12 (2023), 3253–3265. https://doi.org/10.1109/TPDS.2023.3324934

[19] Zhihua Fan, Wenming Li, Zhen Wang, Yu Yang, Xiaochun Ye, Dongrui Fan, Ninghui Sun, and Xuejun An. 2024. Improving utilization of dataflow unit for multi-batch processing. *ACM Trans. Archit. Code Optim.* 21, 1, Article 17 (Feb 2024), 26 pages. https://doi.org/10.1145/3637906

[20] Mark Gebhart, Stephen W. Keckler, Brucek Khailany, Ronny Krashinsky, and William J. Dally. 2012. Unifying primary cache, scratch, and register file memories in a throughput processor. In *Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-45)* (Vancouver, B.C., Canada). IEEE Computer Society, 96–106. https://doi.org/10.1109/MICRO.2012.18

[21] Tong Geng, Ang Li, Runbin Shi, Chunshu Wu, Tianqi Wang, Yanfei Li, Pouya Haghi, Antonino Tumeo, Shuai Che, Steve Reinhardt, and Martin C. Herbordt. 2020. AWB-GCN: A graph convolutional network accelerator with runtime workload rebalancing. In *Proceedings of the 2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'20)*. 922–936. https://doi.org/10.1109/MICRO50266.2020.00079

[22] Graham Gobieski, Ahmet Oguz Atli, Kenneth Mai, Brandon Lucia, and Nathan Beckmann. 2021. Snafu: An ultra-low-power, energy-minimal CGRA-generation framework and architecture. In *Proceedings of the 48th Annual International Symposium on Computer Architecture* (Virtual Event, Spain) *(ISCA'21)*. IEEE Press, 1027–1040. https://doi.org/10.1109/ISCA52012.2021.00084

[23] Graham Gobieski, Amolak Nagi, Nathan Serafin, Mehmet Meric Isgenc, Nathan Beckmann, and Brandon Lucia. 2019. MANIC: A vector-dataflow architecture for ultra-low-power embedded systems. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'52)* (Columbus, OH, USA). ACM, New York, 670–684. https://doi.org/10.1145/3352460.3358277

[24] Graham Gobieski, Amolak Nagi, Nathan Serafin, Mehmet Meric Isgenc, Nathan Beckmann, and Brandon Lucia. 2019. MANIC: A vector-dataflow architecture for ultra-low-power embedded systems. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture* (Columbus, OH,) *(MICRO'52)*. ACM, New York, 670–684. https://doi.org/10.1145/3352460.3358277

[25] Tae Jun Ham, Juan L. Aragón, and Margaret Martonosi. 2015. DeSC: Decoupled supply-compute communication management for heterogeneous architectures. In *Proceedings of the 48th International Symposium on Microarchitecture (MICRO-48)* (Waikiki, Hawaii). ACM' New York, 191–203. https://doi.org/10.1145/2830772.2830800

[26] Tae Jun Ham, Lisa Wu, Narayanan Sundaram, Nadathur Satish, and Margaret Martonosi. 2016. Graphicionado: A high-performance and energy-efficient accelerator for graph analytics. In *Proceedings of the 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-49)* (Taipei, Taiwan). IEEE Press, Article 56, 13 pages. https://doi.org/10.1109/MICRO.2016.7783759

[27] Sangyeol Kang and Alexander G. Dean. 2012. Leveraging both data cache and scratchpad memory through synergetic data allocation. In *Proceedings of the 2012 IEEE 18th Real Time and Embedded Technology and Applications Symposium*. 119–128. https://doi.org/10.1109/RTAS.2012.22

[28] Manupa Karunaratne, Aditi Kulkarni Mohite, Tulika Mitra, and Li-Shiuan Peh. 2017. HyCUBE: A CGRA with reconfigurable single-cycle multi-hop interconnect. In *Proceedings of the 54th Annual Design Automation Conference 2017 (DAC'17)* (Austin, TX). ACM, New York, Article 45, 6 pages. https://doi.org/10.1145/3061639.3062262

[29] Manupa Karunaratne, Dhananjaya Wijerathne, Tulika Mitra, and Li-Shiuan Peh. 2019. 4D-CGRA: Introducing branch dimension to spatio-temporal application mapping on CGRAs. In *Proceedings of the 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD'19)*. 1–8. https://doi.org/10.1109/ICCAD45719.2019.8942148

[30] Changkyu Kim, Simha Sethumadhavan, M. S. Govindan, Nitya Ranganathan, Divya Gulati, Doug Burger, and Stephen W. Keckler. 2007. Composable lightweight processors. In *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 40)*. IEEE Computer Society, 381–394. https://doi.org/10.1109/MICRO.2007.41

[31] Rakesh Komuravelli, Matthew D. Sinclair, Johnathan Alsop, Muhammad Huzaifa, Maria Kotsifakou, Prakalp Srivastava, Sarita V. Adve, and Vikram S. Adve. 2015. Stash: Have your scratchpad and cache it too. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture (ISCA'15)* (Portland, Oregon). ACM, New York, 707–719. https://doi.org/10.1145/2749469.2750374

[32] Chris Lattner and Vikram Adve. 2004. LLVM: A compilation framework for lifelong program analysis & transformation. In *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-Directed and Runtime Optimization (CGO'04)* (Palo Alto, CA). IEEE Computer Society, 75. https://doi.org/10.1109/CGO.2004.1281665

[33] Tianyu Liu, Wenming Li, and Zhihua Fan. 2023. DFGC: DFG-aware NoC control based on time stamp prediction for dataflow architecture. In *Proceedings of the 2023 IEEE 41st International Conference on Computer Design (ICCD'23)*. 432–439. https://doi.org/10.1109/ICCD58817.2023.00071

[34] Mahim Mishra, Timothy J. Callahan, Tiberiu Chelcea, Girish Venkataramani, Seth C. Goldstein, and Mihai Budiu. 2006. Tartan: Evaluating spatial computation for whole program execution. In *Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS XII)* (San Jose, CA). ACM, New York, 163–174. https://doi.org/10.1145/1168857.1168878

[35] Gregory E. Moore. 1998. Cramming more components onto integrated circuits. *Proc. IEEE* 86, 1 (1998), 82–85. https://doi.org/10.1109/JPROC.1998.658762

[36] Quan M. Nguyen and Daniel Sanchez. 2021. Fifer: Practical acceleration of irregular applications on reconfigurable architectures. In *Proceedings of the MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'21)* (Virtual Event, Greece). ACM, New York, 1064–1077. https://doi.org/10.1145/3466752.3480048

[37] Tony Nowatzki, Vinay Gangadhar, Newsha Ardalani, and Karthikeyan Sankaralingam. 2017. Stream-dataflow acceleration. In *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA'17)* (Toronto, ON, Canada). ACM, New York, 416–429. https://doi.org/10.1145/3079856.3080255

[38] NVIDIA. 2023. NVIDIA H100 Tensor Core Hopper Whitepaper. Retrieved August 7, 2024 from https://resources.nvidia.com/en-us-tensor-core/

[39] Angshuman Parashar, Michael Pellauer, Michael Adler, Bushra Ahsan, Neal Crago, Daniel Lustig, Vladimir Pavlov, Antonia Zhai, Mohit Gambhir, Aamer Jaleel, et al. 2013. Triggered instructions: A control paradigm for spatially-programmed architectures. In *Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA'13)* (Tel-Aviv, Israel). ACM, New York, 142–153. https://doi.org/10.1145/2485922.2485935

[40] Hyunchul Park, Yongjun Park, and Scott Mahlke. 2009. Polymorphic pipeline array: A flexible multicore accelerator with virtualized execution for mobile multimedia applications. In *Proceedings of the 42nd Annual IEEE/ACM*

International Symposium on Microarchitecture (MICRO 42) (New York, New York). ACM, New York, 370–380. https://doi.org/10.1145/1669112.1669160

[41] Julian Pavon, Ivan Vargas Valdivieso, Adrian Barredo, Joan Marimon, Miquel Moreto, Francesc Moll, Osman Unsal, Mateo Valero, and Adrian Cristal. 2021. VIA: A smart scratchpad for vector units with application to sparse matrix computations. In *Proceedings of the 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA'21)*. 921–934. https://doi.org/10.1109/HPCA51647.2021.00081

[42] Michael Pellauer, Yakun Sophia Shao, Jason Clemons, Neal Crago, Kartik Hegde, Rangharajan Venkatesan, Stephen W. Keckler, Christopher W. Fletcher, and Joel Emer. 2019. Buffets: An efficient and composable storage idiom for explicit decoupled data orchestration. In *Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'19)* (Providence, RI). ACM, New York, 137–151. https://doi.org/10.1145/3297858.3304025

[43] Raghu Prabhakar, Yaqi Zhang, David Koeplinger, Matt Feldman, Tian Zhao, Stefan Hadjis, Ardavan Pedram, Christos Kozyrakis, and Kunle Olukotun. 2017. Plasticine: A reconfigurable architecture for parallel paterns. In *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA'17)* (Toronto, ON, Canada). ACM, New York, 389–402. https://doi.org/10.1145/3079856.3080256

[44] Shantian Qin, Wenming Li, Zhihua Fan, Zhen Wang, Tianyu Liu, Haibin Wu, Kunming Zhang, Xuejun An, Xiaochun Ye, and Dongrui Fan. 2023. ROMA: A reconfigurable on-chip memory architecture for multi-core accelerators. In *Proceedings of the 2023 IEEE International Conference on High Performance Computing (HPCC'23)*. 49–57. https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys60770.2023.00017

[45] Nauman Rafique, Won-Taek Lim, and Mithuna Thottethodi. 2007. Effective management of DRAM bandwidth in multicore processors. In *Proceedings of the 16th International Conference on Parallel Architecture and Compilation Techniques (PACT'07)*. 245–258. https://doi.org/10.1109/PACT.2007.4336216

[46] Ali Sedaghati, Milad Hakimi, Reza Hojabr, and Arrvindh Shriraman. 2022. X-cache: A modular architecture for domain-specific caches. In *Proceedings of the 49th Annual International Symposium on Computer Architecture* (New York, New York) *(ISCA'22)*. ACM, New York, 396–409. https://doi.org/10.1145/3470496.3527380

[47] Steven Swanson, Andrew Schwerin, Martha Mercaldi, Andrew Petersen, Andrew Putnam, Ken Michelson, Mark Oskin, and Susan J. Eggers. 2007. The WaveScalar architecture. *ACM Trans. Comput. Syst.* 25, 2, Article 4 (May 2007), 54 pages. https://doi.org/10.1145/1233307.1233308

[48] Cheng Tan, Nicolas Bohm Agostini, Tong Geng, Chenhao Xie, Jiajia Li, Ang Li, Kevin J. Barker, and Antonino Tumeo. 2022. DRIPS: Dynamic rebalancing of pipelined streaming applications on CGRAs. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA'22)*. 304–316. https://doi.org/10.1109/HPCA53966.2022.00030

[49] Cheng Tan, Chenhao Xie, Tong Geng, Andres Marquez, Antonino Tumeo, Kevin Barker, and Ang Li. 2021. ARENA: Asynchronous reconfigurable accelerator ring to enable data-centric parallel computing. *IEEE Transactions on Parallel and Distributed Systems* 32, 12 (2021), 2880–2892. https://doi.org/10.1109/TPDS.2021.3081074

[50] Po-An Tsai, Nathan Beckmann, and Daniel Sanchez. 2017. Jenga: Software-defined cache hierarchies. In *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA'17)* (Toronto, ON, Canada). ACM, New York, 652–665. https://doi.org/10.1145/3079856.3080214

[51] Po-An Tsai, Yee Ling Gan, and Daniel Sanchez. 2018. Rethinking the memory hierarchy for modern languages. In *Proceedings of the 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-51)* (Fukuoka, Japan). IEEE Press, 203–216. https://doi.org/10.1109/MICRO.2018.00025

[52] Zhengrong Wang, Christopher Liu, Aman Arora, Lizy John, and Tony Nowatzki. 2023. Infinity stream: Portable and programmer-friendly in-/near-memory fusion. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3* (Vancouver, BC, Canada) *(ASPLOS'23)*. Association for Computing Machinery, New York, NY, USA, 359–375. https://doi.org/10.1145/3582016.3582032

[53] Zhengrong Wang and Tony Nowatzki. 2019. Stream-based memory access specialization for general purpose processors. In *Proceedings of the 46th International Symposium on Computer Architecture (ISCA'19)* (Phoenix, AZ) . ACM, New York, 736–749. https://doi.org/10.1145/3307650.3322229

[54] Zhengrong Wang, Jian Weng, Sihao Liu, and Tony Nowatzki. 2022. Near-stream computing: General and transparent near-cache acceleration. In *Proceedings of the 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA'22)*. 331–345. https://doi.org/10.1109/HPCA53966.2022.00032

[55] Jian Weng, Sihao Liu, Vidushi Dadu, Zhengrong Wang, Preyas Shah, and Tony Nowatzki. 2020. DSAGEN: Synthesizing programmable spatial accelerators. In *Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture* (Virtual Event) *(ISCA'20)*. IEEE Press, 268–281. https://doi.org/10.1109/ISCA45697.2020.00032

[56] Jian Weng, Sihao Liu, Zhengrong Wang, Vidushi Dadu, and Tony Nowatzki. 2020. A hybrid systolic-dataflow architecture for inductive matrix algorithms. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA'20)*. 703–716. https://doi.org/10.1109/HPCA47549.2020.00063

[57] Wm. A. Wulf and Sally A. McKee. 1995. Hitting the memory wall: implications of the obvious. *SIGARCH Comput. Archit. News* 23, 1 (Mar 1995), 20–24. https://doi.org/10.1145/216585.216588

[58] Mingyu Yan, Xing Hu, Shuangchen Li, Abanti Basak, Han Li, Xin Ma, Itir Akgun, Yujing Feng, Peng Gu, Lei Deng, et al. 2019. Alleviating irregularity in graph analytics acceleration: A hardware/software co-design approach. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'52)* (Columbus, OH, USA). ACM, New York, 615–628. https://doi.org/10.1145/3352460.3358318

[59] Guowei Zhang, Nithya Attaluri, Joel S. Emer, and Daniel Sanchez. 2021. Gamma: Leveraging Gustavson's algorithm to accelerate sparse matrix multiplication. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'21)* (Virtual, USA). ACM, New York, 687–701. https://doi.org/10.1145/3445814.3446702

[60] Shixuan Zheng, Xianjue Zhang, Leibo Liu, Shaojun Wei, and Shouyi Yin. 2022. Atomic dataflow based graph-level workload orchestration for scalable DNN accelerators. In *Proceedings of the 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA'22).* 475–489. https://doi.org/10.1109/HPCA53966.2022.00042

Received 8 August 2024; revised 22 January 2025; accepted 24 February 2025

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 芮定海

方向： 集成电路

单位： 中国科学院微电子研究所

题目： 基于掩模热-力耦合效应的高精度套刻误差补偿方法

Nanoscale Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Check for updates

Cite this: *Nanoscale Adv.*, 2025, 7, 6563

Method of high-order advanced lithography overlay correction to enhance the manufacturing performance of integrated circuits

Dinghai Rui,[abc] Libin Zhang,[*abc] Yayi Wei[*abc] and Yajuan Su[abc]

As integrated circuit (IC) manufacturing advances toward smaller technology nodes, conventional lithography methods are increasingly challenged by the diffraction-limited resolution, escalating process complexity, and rising costs. Among these challenges, overlays have a particularly pronounced impact on manufacturing quality. To address this issue, this paper proposes a high-order overlay correction model that employs a two-dimensional fifth-order polynomial to accurately fit and characterize the distribution of overlays. The model's effectiveness is validated through finite element simulations. By incorporating an array of piezoelectric actuators, thermally induced deformation control units, and micro-mechanical clamping mechanisms, the model enables precise regulation of complex stress fields and localized temperature variations along the mask boundary, thereby enabling effective compensation of high-order overlay errors. Simulation results demonstrate that the proposed approach reduces the |mean| + 3σ of overlay to below 1 nm. It achieves nearly 100% correction for 1st-order and 2nd-order overlay components, over 80% correction for 3rd-order and 4th-order components, and a correction rate of 68.16% for 5th-order errors. Multiple randomized verification tests indicate average compensation efficiencies of 96.85% in the x-direction and 97.36% in the y-direction, highlighting the model's robustness and consistency. In practical processes, the model successfully reduces actual wafer overlay to |mean| + 3σ values of 4.22 nm and 6.26 nm in the x and y directions, respectively. This study presents an efficient and reliable solution for high-order overlay compensation in advanced lithography, offering significant benefits for enhancing IC manufacturing performance and reliability.

Received 15th July 2025
Accepted 14th August 2025

DOI: 10.1039/d5na00682a

rsc.li/nanoscale-advances

1. Introduction

Lithography is a critical process in integrated circuit (IC) manufacturing, wherein photochemical reactions are employed to transfer predefined patterns from a mask onto a wafer surface.[1–3] Lithographic technology has evolved significantly—from early g-line and i-line light sources to deep ultraviolet lithography (DUVL) utilizing 193 nm excimer laser, and further to extreme ultraviolet lithography (EUVL) at a wavelength of 13.5 nm—each technological leap pushing the limits of manufacturing capabilities.[4–6] As technology nodes continue to shrink, conventional optical lithography faces increasing challenges due to the diffraction-limited resolution, the growing complexity and cost of light source fabrication, and rising process intricacies.[7] To overcome these limitations, a range of emerging lithography techniques have been proposed in recent years, including near-field lithography (NFL), substrate conformal imprint lithography (SCIL),[8–10] surface plasmon lithography (SPL),[11–13] and nanoimprint lithography (NIL).[14–17] These advanced approaches offer promising pathways to surpass the classical diffraction limit, enabling low-cost, large-area, and high-resolution patterning for next-generation semiconductor manufacturing.[18–20]

In IC manufacturing, multiple layers are sequentially fabricated starting from a silicon wafer, and positional deviations between layers—referred to as overlay errors—inevitably arise during the process.[21,22] Overlay accuracy is a critical factor influencing lithographic quality, which in turn directly affects the performance, yield, and reliability of ICs.[1,23] As feature sizes continue to shrink in advanced lithography nodes (7 nm and below), the impact of overlay errors on manufacturing quality becomes increasingly pronounced. According to the 2022 International Roadmap for Devices and Systems (IRDS), for the 3 nm technology node, the critical dimension (CD) control requirement for logic metal layers is 1.8 nm (3σ), while the allowable overlay error is constrained to 2.4 nm (3σ).[23,24] These specifications are expected to become even tighter over the coming decade. Currently, high-order overlay correction techniques such as high order process correction (i-HOPC) are

[a]EDA Center, Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100029, China. E-mail: zhanglibin@ime.ac.cn; weiyayi@ime.ac.cn
[b]School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China
[c]State Key Laboratory of Fabrication Technologies for Integrated Circuits, Beijing 100029, China

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

adopted to mitigate intra-field overlay errors.[25,26] However, state-of-the-art lithography systems are typically limited to supporting precise correction up to the third-order terms.

In advanced multilayer IC structures, the increasing number of process layers gives rise to pronounced challenges such as local structural distortion, high-order nonlinear deformation, and non-uniform strain induced by thermal stress mismatches.[27] Traditional overlay compensation methods based on linear models or low-order polynomial fitting are no longer sufficient to address these complex variations,[28] highlighting the urgent need for a high-precision and high-robustness correction scheme capable of addressing high-order overlay. To this end, the present study proposes an effective method for correcting high-order overlay in advanced lithography processes. This approach aims to enhance overall pattern alignment accuracy, reduce manufacturing defect rates, and meet the stringent requirements of future ultra-precise chip fabrication. The structure of this paper is as follows: Section 1 provides the introduction. Section 2 describes the proposed high-order polynomial-based overlay compensation method. Section 3 presents the simulation validation and discussion. Section 4 concludes the study.

2. Theoretical analysis and modeling of overlays

2.1 The source of overlay error and correction model for high-order overlays

Modern IC manufacturing involves a series of complex processes including wafer growth and preparation, lithography and patterning, as well as etching and deposition,[21,29] as illustrated in Fig. 1(a). This process encompasses a wide range of technologies and requires precise control and meticulous attention at every stage. In particular, advanced lithographic techniques typically involve steps such as photoresist coating, soft baking, exposure, post-exposure baking, development, and overlay metrology. Several key factors critically influence overlay accuracy in lithography, including the gravitational deformation of the mask, thermal effects on the mask, wafer-induced distortions, projection lens distortion in the lithography scanner, and mechanical non-uniformities in the movement of various stages.[30,31] These effects become increasingly significant as semiconductor manufacturing advances toward smaller technology nodes. For instance, the mask may undergo slight bending due to gravity, causing pattern misalignment, while

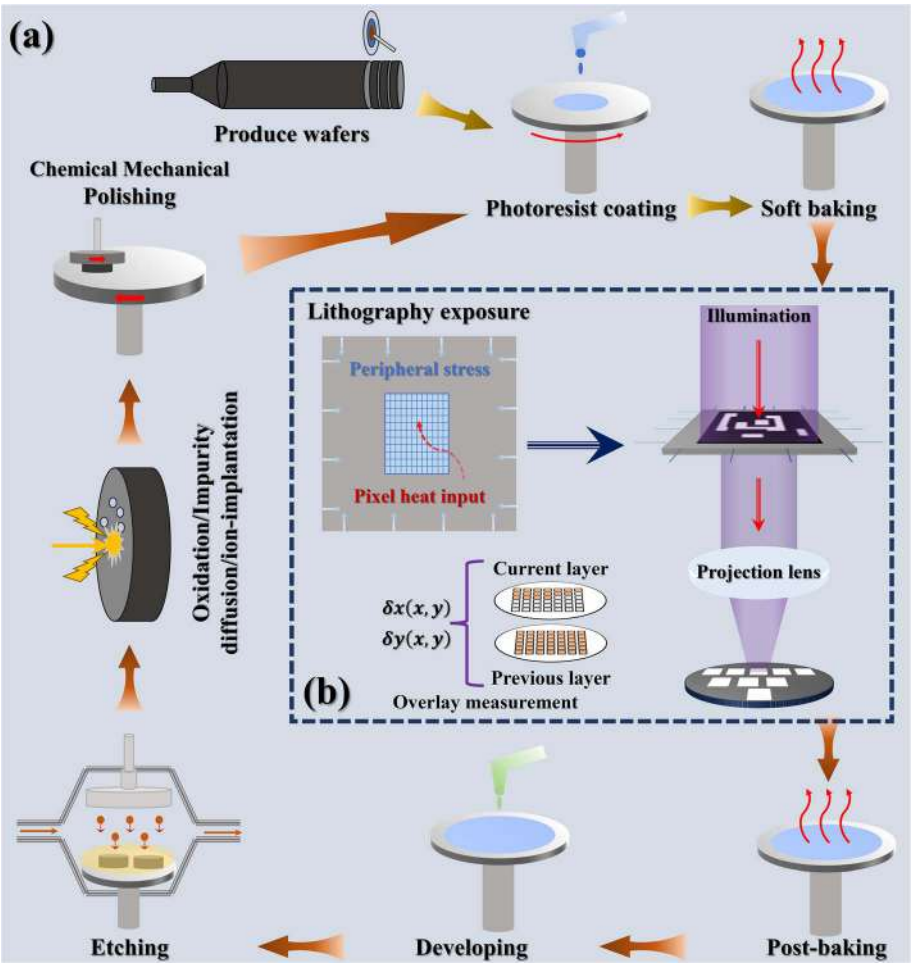


Fig. 1 Schematic diagram of the modern IC manufacturing process and the high-order overlay error correction in the advanced lithography process.[19,28] (a) Modern IC manufacturing process. (b) The high-order overlay correction in the advanced lithography process.

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

absorption of light during exposure leads to localized heating, further affecting pattern dimensions and positions.[32] The combined impact of these effects contributes to increased overlay, which varies dynamically with lithography throughput and exposure dose. Optimizing mask design, chuck architecture, and lithography tool inspection and feedback mechanisms can effectively mitigate these issues. Such improvements enhance the overlay accuracy and reliability of advanced lithography processes, thereby fulfilling the stringent requirements of next-generation semiconductor fabrication at ever-decreasing technology nodes.

From the perspective of a macroscopic lithography system, several feasible approaches for overlay error compensation are available in conventional projection lithography systems, as illustrated in Fig. 1(b). These include optical lens aberration correction, mask pattern adjustments, and wafer stage control. However, in emerging lithography techniques such as SPL, NIL, SCIL, and NFL, the absence of projection optics renders traditional lens-based overlay correction methods ineffective. To address this limitation, a generalized high-order overlay correction model is proposed that can be adapted across various lithography platforms, including the aforementioned non-projection types, while remaining compatible with conventional projection lithography systems. The proposed high-order correction framework relies primarily on three types of precision actuators: an array of piezoelectric actuators for the fine control of complex stress distributions along the mask boundary, thermally induced deformation units that regulate localized temperature fields to induce controlled shape changes, and micromechanical clamping mechanisms that apply directional forces to the mask edges, enabling precise in-plane translation and rotation. Notably, this study implements a compensation strategy consisting of 16 distributed force application units positioned around the mask periphery and 143 sub-millimeter thermal control pixels within a 2 mm × 3 mm (6 mm^2) lithography field region. Future work may explore denser stress distribution configurations and extend the thermal actuation resolution below 1 mm^2, thereby enabling even finer control of overlay.

2.2 The high-order process analysis model

The overlay error data are normalized based on a two-dimensional coordinate system within the exposure field, and all coordinate points are transformed into a standardized unit square domain. Inspired by the third-order correction method employed in ASML's intra-field exposure control, this study adopts a two-dimensional fifth-order polynomial as the basis function to fit and model the distribution of the overlay. The corresponding fifth-order polynomial decomposition model is expressed as follows:

$$\begin{aligned}\delta x(x,y) = {} & k_1 + k_3\cdot x + k_5\cdot y + k_7\cdot x^2 + k_9\cdot xy + k_{11}\cdot y^2 \\ & + k_{13}\cdot x^3 + k_{15}\cdot x^2y + k_{17}\cdot xy^2 + k_{19}\cdot y^3 \\ & + k_{21}\cdot x^4 + k_{23}\cdot x^3y + k_{25}\cdot x^2y^2 + k_{27}\cdot xy^3 \\ & + k_{29}\cdot y^4 + k_{31}\cdot x^5 + k_{33}\cdot x^4y + k_{35}\cdot x^3y^2 \\ & + k_{37}\cdot x^2y^3 + k_{39}\cdot xy^4 + k_{41}\cdot y^5, \end{aligned} \tag{1}$$

Table 1 Lithography simulation related parameters.[19,28]

Parameter	Symbols	Value
Mask thickness (mm)	t	6.35
Density (kg m^{-3})	ρ	2200
Young's modulus (Pa)	E	7.05×10^{10}
Poisson's ratio	ν	0.164
Bulk modulus (Pa)	K	3.497×10^{10}
Shear modulus (Pa)	G	3.0284×10^{10}
Coefficient of thermal expansion (1/°C)	γ	4.6×10^{-7}
Isotropic thermal conductivity (W m^{-1} °C^{-1})	κ	1.38
Exposure absorption efficiency	η	20%
Convection coefficient (W (m^{-2} °C^{-1}))	μ	10

$$\begin{aligned}\delta y\,(x,y) = {} & k_2 + k_4\cdot y + k_6\cdot x + k_8\cdot y^2 + k_{10}\cdot xy \\ & + k_{12}\cdot x^2 + k_{14}\cdot y^3 + k_{16}\cdot y^2x + k_{18}\cdot yx^2 \\ & + k_{20}\cdot x^3 + k_{22}\cdot y^4 + k_{24}\cdot y^3x + k_{26}\cdot x^2y^2 \\ & + k_{28}\cdot yx^3 + k_{30}\cdot x^4 + k_{32}\cdot y^5 + k_{34}\cdot y^4x \\ & + k_{36}\cdot y^3x^2 + k_{38}\cdot y^2x^3 + k_{40}\cdot yx^4 + k_{42}\cdot x^5. \end{aligned} \tag{2}$$

In the equation, $k_j (j = 0, 1, \cdots, 41, 42)$ represents the controllable overlay compensation coefficients implemented *via* various actuators within the lithography system. (x, y) denotes the normalized intra-field coordinates of each exposure field. $\delta x(x, y)$ and $\delta y\,(x, y)$ represent the overlay in the x-direction and y-direction, respectively.

In the thermo-mechanical coupling module of the finite element software, the mathematical and physical models of the lithography system under the parameters listed in Table 1 are typically described using partial differential equations, including the heat equilibrium equation and the structural equilibrium equation. The heat equation governs the evolution of the temperature field, while the structural equation characterizes the resulting displacements and stress distributions caused by thermal variations. For the proposed high-order overlay compensation model, the thermo-mechanical coupling problem is formulated accordingly. Within the elastic deformation range of the mask, the overlay deformation induced by stress in this compensation scheme exhibits linear superposition behavior.[28] The mathematical formulation is as follows:

$$\delta x_{\text{force}} = \varepsilon_x^{\text{force}}\cdot L_x = \frac{L_x}{E}\left(\sum_{i=1}^{n}\alpha_i f_i - \nu\sum_{i=1}^{n}\beta_i f_i\right), \tag{3}$$

$$\delta y_{\text{force}} = \varepsilon_y^{\text{force}}\cdot L_y = \frac{L_y}{E}\left(\sum_{i=1}^{n}\beta_i f_i - \nu\sum_{i=1}^{n}\alpha_i f_i\right). \tag{4}$$

In the equations, α_i and β_i represent the contribution coefficients the n stresses f_i to the total stress in the x-direction and y-direction, respectively. $E = FL/A\cdot \mathrm{d}L$ denotes the Young's modulus, and ν is the Poisson's ratio. δx_{force} and δy_{force} respectively represent the x-direction and y-direction overlay errors for the stress compensation around the mask. Except for the stress variables, all other parameters in eqn (3) and (4) are fixed coefficients, implying that the resulting overlay compensation exhibits a linear relationship with the input stresses. Furthermore, according to thermal deformation theory, within the elastic deformation range, the following can be derived:

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

$$\delta x_{\text{heat}} = \sum_{i=1}^{m} \varepsilon_x{}^i \cdot L_{xi} = \sum_{i=1}^{m} \gamma \cdot \Delta T \cdot L_{xi} = \sum_{i=1}^{m} \gamma \cdot \frac{h_i}{c \cdot \rho \cdot L_{yi} \cdot t}, \quad (5)$$

$$\delta y_{\text{heat}} = \sum_{i=1}^{m} \varepsilon_y{}^i \cdot L_{yi} = \sum_{i=1}^{m} \gamma \cdot \Delta T \cdot L_{yi} = \sum_{i=1}^{m} \gamma \cdot \frac{h_i}{c \cdot \rho \cdot L_{xi} \cdot t}. \quad (6)$$

In the equation, γ denotes the mask thermal expansion coefficient, c is the specific heat capacity of the mask, ρ represents the mask density, and $V = L_x \cdot L_y \cdot t$ represents the volume of m local heat input areas. δx_{heat} and δy_{heat} respectively represent the x-direction and y-direction overlay errors of pixel-level heat compensation within the mask lithography field. Accordingly, within the elastic deformation range of the mask, the proposed high-order overlay correction model can compensate the overlay errors within the lithography field as follows:

$$\widehat{\delta x}(x, y) = \sum_{i=1}^{m} \gamma \cdot \frac{h_i}{c \cdot \rho \cdot L_{yi} \cdot t} + \frac{L_x}{E} \left(\sum_{j=1}^{n} \alpha_j f_j - v \sum_{j=1}^{n} \beta_j f_j \right), \quad (7)$$

$$\widehat{\delta y}(x, y) = \sum_{i=1}^{m} \gamma \cdot \frac{h_i}{c \cdot \rho \cdot L_{xi} \cdot t} + \frac{L_y}{E} \left(\sum_{j=1}^{n} \beta_j f_j - v \sum_{j=1}^{n} \alpha_j f_j \right). \quad (8)$$

Combining eqn (1), (2), (7) and (8), where the parameters of each overlay compensation actuator serve as variables while all other parameters remain fixed coefficients, the high-order compensation model exhibits linear superposability between the generated overlay correction and the applied stress and localized thermal control doses.[28]

2.3 The principle and method of high-order overlay compensation

The main procedure of this high-order overlay error correction model is illustrated in Fig. 2(a) and is fundamentally similar to the conventional overlay correction process used in traditional lithography tools. After completing the exposure and

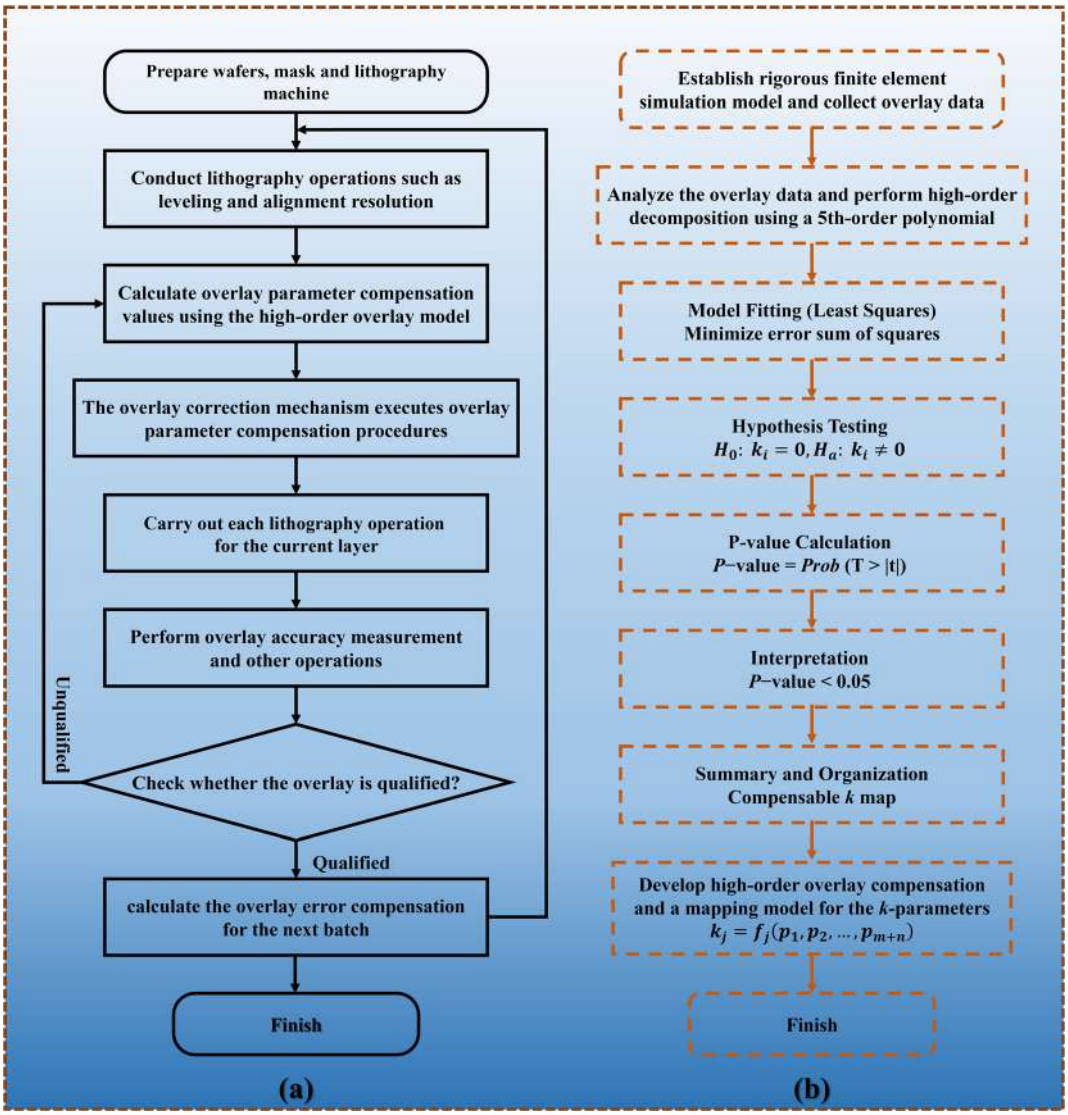


Fig. 2 The overlay compensation flowchart based on the high-order overlay correction model. (a) The flowchart for achieving overlay correction with this model. (b) The flowchart for establishing a high-order overlay compensation model.

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

development of the current layer, overlay metrology is performed on intra-field overlay marks, and the measured overlay distribution is fitted using the high-order polynomial model. The resulting coefficients are then converted into actuator control parameters through the response mapping function, which correspond to distributed piezoelectric forces and localized thermal inputs. These parameters are applied to the mask stage before the next exposure step is initiated, so that the physical and thermal-induced distortions expected during the upcoming exposure are pre-compensated in advance. This predictive correction procedure is repeated for each wafer or lot, enabling continuous adaptation to dynamic process variations. The subsequent steps involve exposure, post-exposure overlay accuracy measurement, and, if necessary, repeated compensation cycles.

The construction process of the fifth-order overlay correction model is illustrated in Fig. 2(b). To avoid destructive experiments, a rigorous finite element simulation is employed to establish the compensation model. The simulated overlay error data are subjected to statistical analysis. By extracting statistical features and ranking the significant weights of the high-order overlay parameters k, the modeling dimensionality can be significantly reduced. This not only simplifies the subsequent development of compensation strategies and the overlay error response function, but also reduces the overall modeling complexity. In this study, a least-squares-based parameter significance test is used to rank the influence of polynomial terms. This enables the identification of dominant terms that have a critical impact on compensation and the extraction of a representative set of compensation control parameters k, thereby improving the efficiency and robustness of the subsequent overlay compensation modeling.

First, a fifth-order polynomial model as defined in eqn (1) and (2) is constructed based on the input overlay distribution data. The estimation of the overlay parameters k is then obtained by minimizing the sum of squared deviations between the measured overlay errors and those predicted by the model.[19,28]

$$\min \sum_{i=1}^{l} \left(\delta x - \widehat{\delta x}(x_i, y_i) \right)^2, \tag{9}$$

$$\min \sum_{i=1}^{l} \left(\delta y - \widehat{\delta y}(x_i, y_i) \right)^2. \tag{10}$$

Next, hypothesis testing is performed to evaluate the statistical significance of each overlay compensation parameter k_j in influencing the overlay, by assessing the corresponding P-values of the hypothesis tests for each parameter k_j.

$$\begin{cases} \text{Original assumption } H_0\text{: } k_j = 0 \rightarrow \text{No significant} \\ \text{Alternative hypothesis } H_a : k_j \neq 0 \rightarrow \text{Significant} \end{cases}. \tag{11}$$

Subsequently, the P-values are calculated based on the test statistics t_j obtained from the t-tests, which evaluate the statistical significance of each parameter:

$$t_j = \frac{\hat{k}_j}{\mathrm{SE}\left(\hat{k}_j\right)} \tag{12}$$

where $\hat{t}$j denotes the estimated value of the parameter k_j, and SE ($\hat{k}$j) represents the standard error of the estimate. Based on the t-distribution and the computed t-value, the corresponding P-value is calculated as:

$$P_j = 2 \times \mathrm{Prob}(T > |t_j|). \tag{13}$$

Then, based on the calculated P-values for the high-order overlay compensation scheme, the k parameters with $P < 0.05$ are selected to simplify the model. Finally, all statistically significant polynomial terms corresponding to the effective overlay compensation parameters from various correction schemes are aggregated and identified as the key polynomial components influencing the actual overlay error. A response mapping model is then established to describe the relationship between the dominant high-order overlay error parameters k_j and the specific control actions of the overlay compensation actuators.

$$k_j = f_j(p_1, p_2, \ldots, p_{n+m}). \tag{14}$$

Here, p_i represents the control parameters related to complex peripheral stresses and localized thermal fine-tuning around the mask, such as the magnitude of electromagnetic actuation forces and the dosage of localized thermal regulation. k_j denotes the response parameters corresponding to the key high-order polynomial terms of overlay error. This functional model can be utilized in a reverse manner to predict the specific actuator control parameters that must be applied by the lithography system in order to achieve a desired overlay error compensation target. In this way, an analytical solution for active compensation control can be obtained. This formulation enables predictive modeling and control of overlay error responses under varying stress and localized thermal modulation conditions.

Furthermore, this paper defines the processing formula for the correction ratio of the mask compensation:

$$R_{\mathrm{corr}} = \frac{\delta_{\mathrm{ori}} - \delta_{\mathrm{res}}}{\delta_{\mathrm{ori}}} \times 100\%, \tag{15}$$

where δ_{ori} and δ_{res} are the |mean| + 3σ value of the original overlay error and the after corrected point-by-point overlay residual, respectively. This metric reflects the percentage reduction of the overlay error variation achieved by the compensation method.

3. Simulation results and discussion

3.1 Compensation effect of each order overlay

Based on the previously described high-order overlay correction model, we conducted simulations to verify the reliability and feasibility of the proposed compensation approach. By employing a fifth-order polynomial decomposition of the overlay error, the simulation results are presented in Fig. 3. This figure illustrates various types of initial random overlay error

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

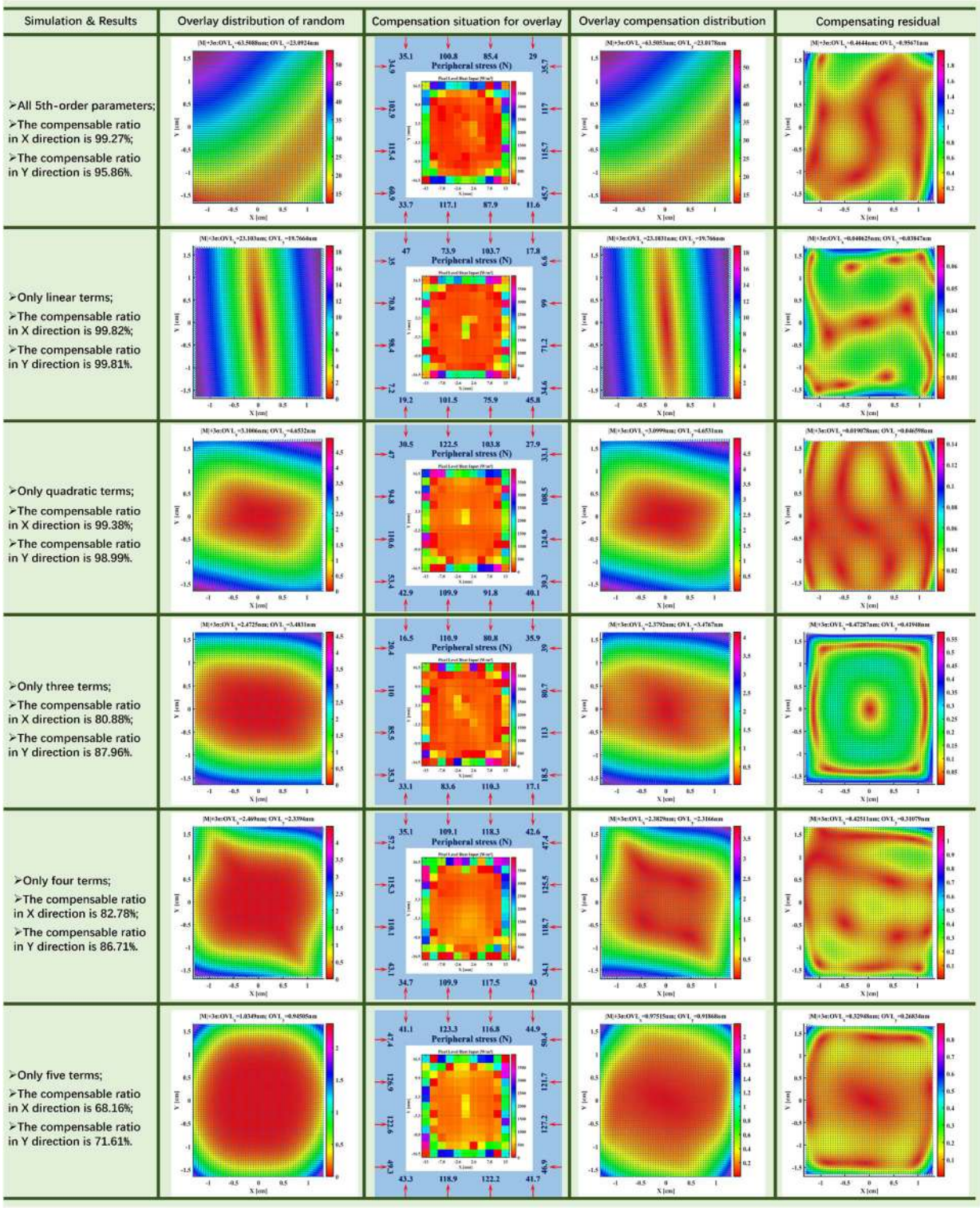


Fig. 3 The compensation effect of the high-order overlay correction method is simulated randomly under different initial overlay error distributions.

distributions, the corresponding localized thermal effects, and complex peripheral stresses around the mask as calculated by the model. The residual overlay errors before and after compensation are evaluated on a point-by-point basis. As shown in the figure, the high-order overlay error correction model effectively reduces the overlay ($|\text{mean}| + 3\sigma$) to below 1 nm. Specifically, for linear and second-order overlay distributions, the model achieves near-complete correction, with residual

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

overlay (|mean| + 3σ) as low as 0.1 nm. In comparison, it is significantly superior to the effect of a certain third-order lithography correction method.[28]

For 3rd-order and 4th-order overlay error distributions, the proposed model is also capable of achieving more than 80% correction rate, with the residual overlay error reduced to approximately 0.5 nm. Furthermore, for 5th-order overlay error distributions, the model still achieves a correction rate of 68.16%, with residual overlay less than 0.4 nm, demonstrating strong reliability. It is worth noting that for conventional DUVL systems, reliable correction of 3rd-order and higher-order components is often unachievable due to the limited degrees of freedom in optical lens-based compensation mechanisms. In summary, the proposed high-order overlay compensation model demonstrates effective correction performance across various types of random overlay distributions.

3.2 Average correction of the overlay

To more comprehensively evaluate the process reliability and stability of the proposed high-order overlay error correction model, we conducted multiple validation tests using fifth-order polynomial-based intra-field overlay correction. The compensation performance for randomly generated 5th-order overlay errors—each consisting of 42 k parameters—is shown in Fig. 4. As indicated by the dashed average lines in the figure, the model achieves an average correction rate of 96.85% in the x-direction and 97.36% in the y-direction across 40 high-order overlay distributions. Additionally, the results from individual compensation instances show that the correction rate consistently remains above 95%, with only a few cases dropping to around 90% due to the presence of stronger high-order components.

3.3 Verification and analysis of overlay compensation in the actual process

To further assess the practical performance of the proposed high-order overlay error correction method, we conducted validation and analysis using overlay error data collected from within the exposure fields of actual fabrication processes. These data were obtained from real wafer manufacturing environments and reflect the true distribution characteristics of overlay errors under complex industrial conditions. This evaluation aims to determine the compensation quality of the high-order correction model in practical applications.

Fig. 5 illustrates the overlay distribution at a specific wafer location in a real fabrication process, along with the corresponding compensation results obtained using the proposed high-order correction model. As shown in Fig. 5(a), the actual overlay distribution from the manufacturing process was fitted using a fifth-order polynomial, with the resulting fitted distribution presented in Fig. 5(b). Notably, Fig. 5(c) shows the residuals between the actual overlay errors and the fifth-order polynomial fit, with the (|mean| + 3σ) residual overlay errors measured at 3.87 nm in the x-direction and 5.43 nm in the y-direction. This indicates that, under a fifth-order compensation scheme, 15.70% and 30.15% of the intrinsic overlay in the x and y directions remain uncompensated. Without incorporating correction mechanisms beyond the fifth order, these residual errors can be considered random and are beyond the compensation capability of current lithography systems. By inputting

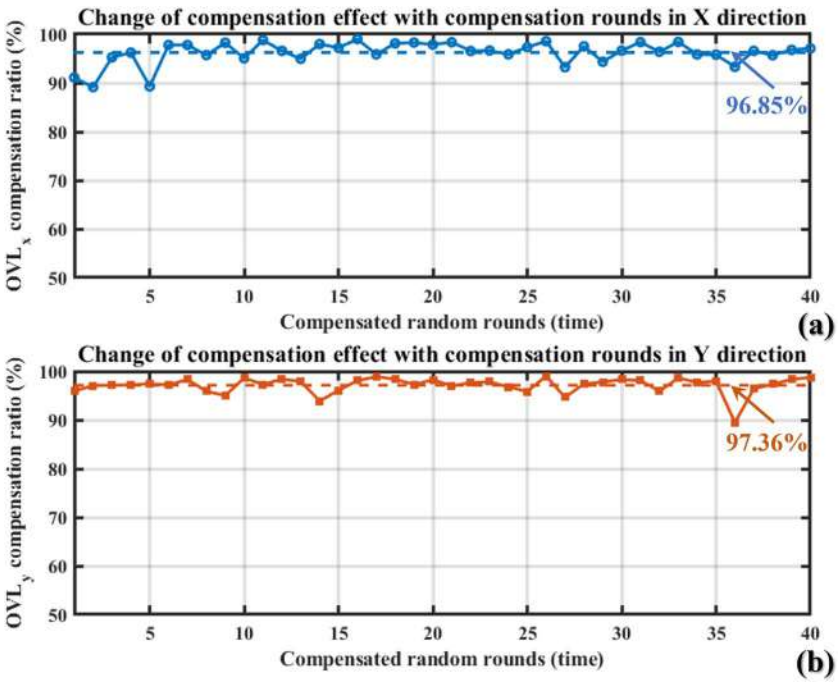


Fig. 4 The average correction effect of the overlay error. (a) The compensation effect in the x-direction after several random overlay compensations. (b) The compensation effect in the y-direction after several random overlay compensations.

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

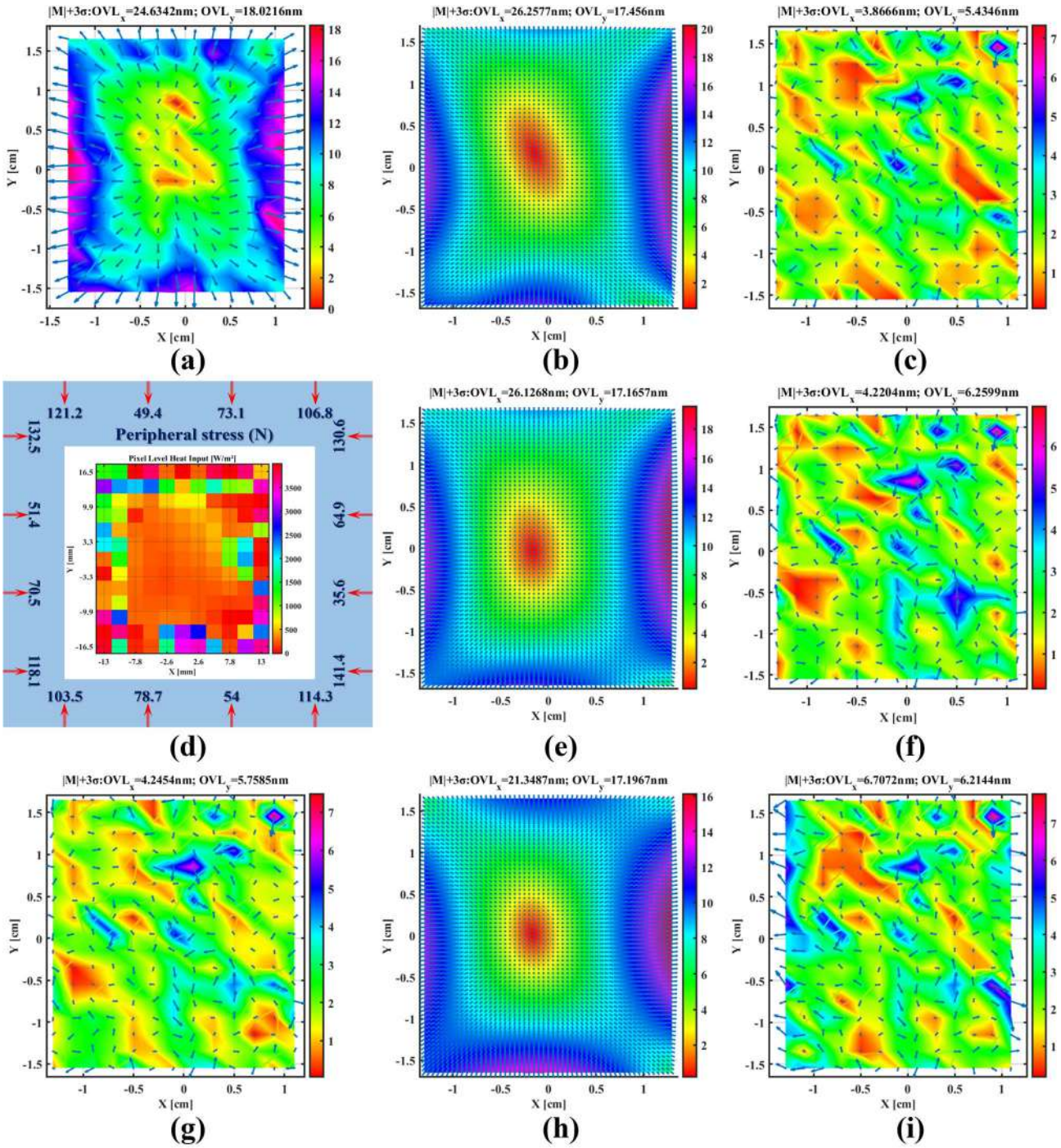


Fig. 5 The distribution of overlay errors in actual processes and the compensation map of this high-order overlay correction model. (a) The overlay distribution within a field of an actual process wafer; (b) the fitting results of the actual overlay errors using a 5th-order polynomial; (c) the residual distribution of overlay errors between the actual process and the 5th-order polynomial fit; (d) application of the correction scheme calculated by the high-order overlay compensation model; (e) the distribution of overlay errors that can be compensated by this model; (f) the compensation residual distribution between the high-order overlay correction model and actual overlay errors; (g) the residual distribution of overlay between the actual process and the 3rd-order polynomial fit; (h) the overlay distribution map using 3rd-order overlay compensation; (i) the overlay residual map using 3rd-order overlay compensation.

this data into the proposed high-order correction model, the resulting spatial distribution of the compensation mechanism is shown in Fig. 5(d), and the corresponding corrected overlay distribution is depicted in Fig. 5(e). Fig. 5(f) presents the point-by-point residual overlay map after applying the high-order mask-based correction model in an actual process environment, demonstrating effective compensation rates of 82.87% in the x-direction and 65.27% in the y-direction. The residuals in

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

the x and y directions are 6.71 nm and 6.21 nm, respectively. As a result, the residual overlay errors were reduced to 4.22 nm and 6.26 nm in the x and y directions, respectively. If a third-order polynomial is employed for compensation, Fig. 5(g) represents the inherent overlay residual distribution. The corresponding overlay distribution after third-order compensation is shown in Fig. 5(h), while Fig. 5(i) illustrates the point-by-point overlay residuals before and after correction. Compared with the third-order overlay compensation model, the fifth-order overlay correction model can effectively reduce the overlay in the x-direction. However, the improvement in the y-direction is not significant. Therefore, in practical processes, the selection of the model should be based on the randomness of overlay data to avoid excessive overlay compensation.

3.4 Analysis of limit correction for overlays based on the k parameter

To provide a more intuitive understanding of the compensation performance of the proposed high-order intra-field overlay control method across different types of overlay distributions, Fig. 6 presents the overlay error compensation results for expansion-induced and shrinkage-induced errors of various polynomial orders. In the figure, the red bars represent the initial overlay errors without any correction (*i.e.*, the randomly generated error distributions), while the blue bars indicate the residual overlay errors after applying the proposed high-order compensation method. It can be observed that, for both expansion and shrinkage overlay effects, the method achieves over 97% correction rate for linear and 2nd-order components. By comparing Fig. 6(a) and (b), we note that the compensation performance for shrinkage-induced errors is comparable to that for expansion-induced errors across all orders. Notably, the compensation results for 4th-order overlay errors outperform those for 3rd-order and 5th-order distributions, with residual overlay errors reduced to approximately 6.11% after compensation.

To further analyze the ultimate performance limits of the proposed overlay compensation method, each of the 42 fifth-order polynomial k-parameters was individually evaluated for compensation performance, as illustrated in Fig. 7. This figure enables a quantitative assessment of the compensation capabilities for different k-parameter distortions, providing insight into the model's ability to address overlay errors arising from both expansion and shrinkage effects. As shown in Fig. 7(a), for expansion-induced overlay errors, the compensation efficiency exceeds 95% for all overlay components associated with 2nd-order or lower k-parameters. However, for high-order components, particularly those corresponding to K15, K17, K19, K20, K28, K33, K38, K40, K41, and K42, residual point-by-point

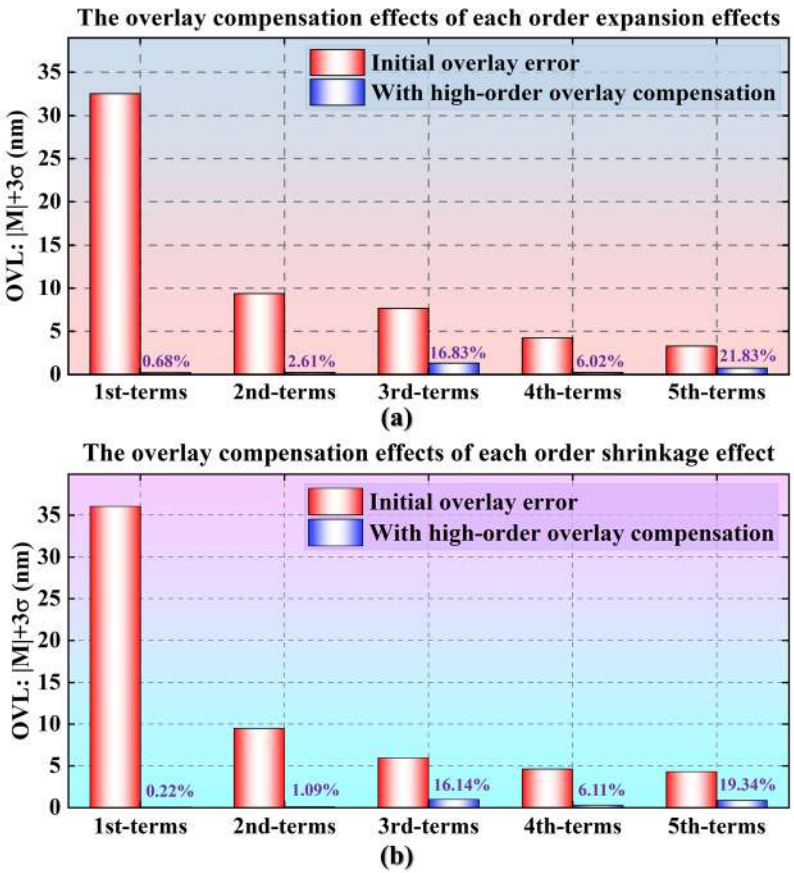


Fig. 6 Diagram of the compensation effects for the overlay of each order. (a) The compensation effect diagram of the overlay for various orders of expansion effects. (b) The compensation effect diagram of the overlay for various orders of shrinkage effects.

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

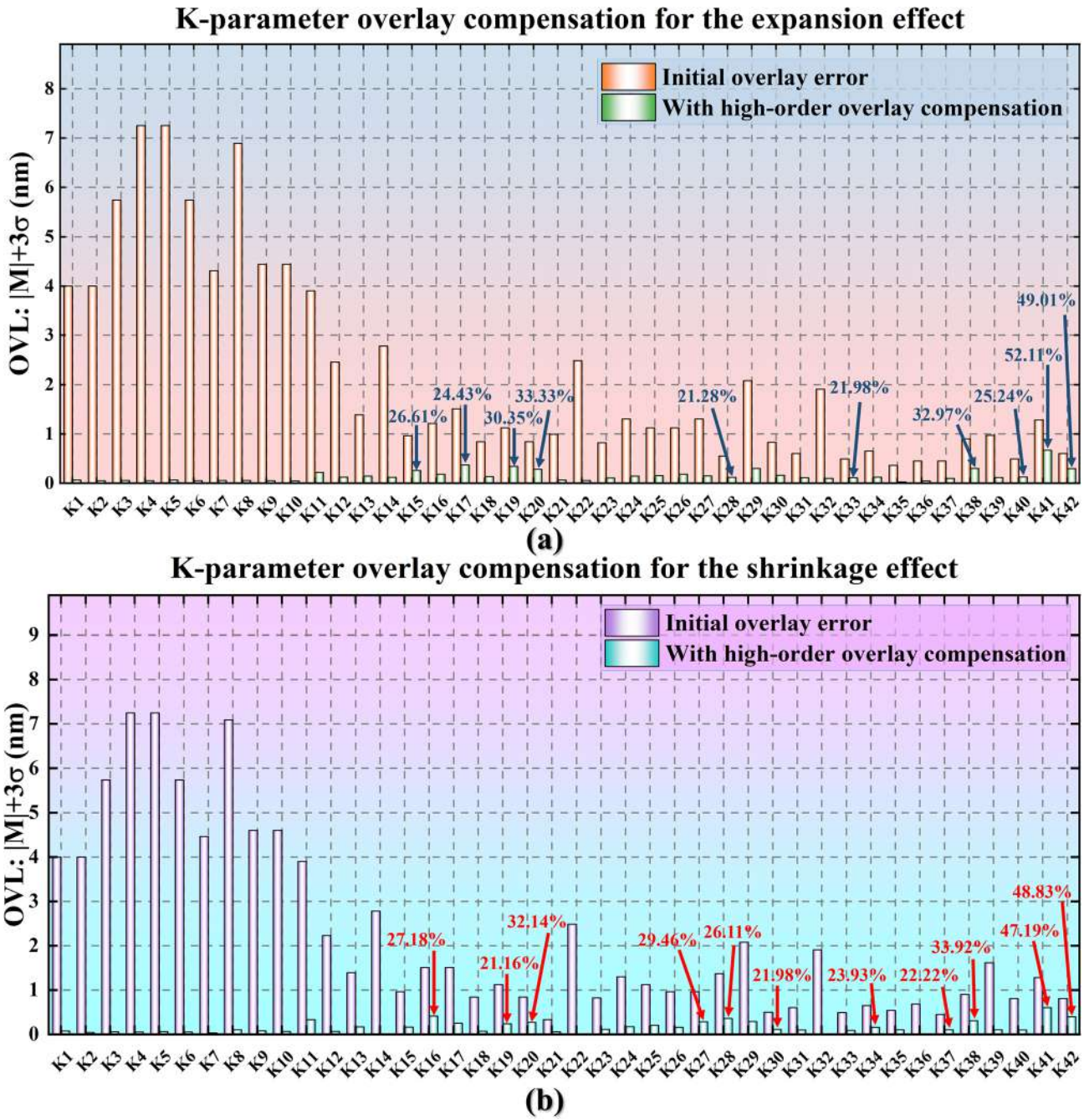


Fig. 7 Correction effect diagram of the high-order correction model under the k-parameter limit case. (a) The overlay compensation effect of the expansion effect under the k-parameter limit. (b) The overlay compensation effect of the shrinkage effect under the k-parameter limit.

overlay errors remain above 20%. Notably, for the expansion effect, the residual overlay errors for K19, K20, K38, K41, and K42 are 30.35%, 33.33%, 32.97%, 52.11%, and 49.01%, respectively, indicating that the proposed high-order compensation method is ineffective in fully correcting these five specific k-components. It is worth noting that similar to commercial lithography systems such as ASML's framework, certain high-order k parameters remain beyond the compensation capability of the current actuator configuration due to physical constraints, including thermal–mechanical coupling nonlinearities and actuator spatial resolution limits, *etc.* Under the present model architecture, the method can stably compensate overlay components up to approximately the 4th–5th polynomial order.

Similarly, the compensation performance for shrinkage-induced overlay distributions is illustrated in Fig. 7(b), which presents the corresponding compensation limits. Using the proposed high-order correction approach, residual overlay exceeding 20% remains for the shrinkage-related components K16, K19, K20, K27, K28, K30, K34, K37, K38, K41, and K42.

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Notably, the model fails to effectively correct K20, K38, K41, and K42, indicating limited compensation capability for these components under shrinkage-dominated overlay conditions.

4. Conclusion

This study addresses the challenge of high-order overlay error correction in advanced lithography by proposing a compensation solution based on high-order polynomial modeling. The aim is to enhance the overall performance and reliability of IC manufacturing. The paper begins by analyzing the origins of overlay errors in modern IC fabrication processes and introduces a high-order overlay correction model, which employs a two-dimensional fifth-order polynomial to fit and characterize the spatial distribution of overlays. The process flow of the proposed correction method is then described, followed by a detailed explanation of the mathematical modeling procedure. Finally, the reliability and feasibility of the high-order overlay correction model are validated through multiple simulations, demonstrating its consistency and robustness.

Simulation results demonstrate that the proposed model effectively reduces overlay errors, bringing the value of $|\text{mean}| + 3\sigma$ below 1 nm. For linear and 2nd-order overlay distributions, the model achieves complete correction, with residual errors less than 0.1 nm. For 3rd-order and 4th-order distributions, more than 80% of the overlay can be corrected, with residuals around 0.5 nm. Furthermore, for 5th-order overlay distributions, the model achieves a correction rate exceeding 68.16%, with residual errors under 0.4 nm. Across multiple randomized simulations, the model shows an average compensation efficiency of 96.85% in the x-direction and 97.36% in the y-direction. In practical manufacturing scenarios, the model effectively compensates for overlays observed on actual wafers, reducing the $|\text{mean}| + 3\sigma$ metric to 4.22 nm and 6.26 nm in the x-direction and y-direction, respectively. In addition, the compensation performance under extreme high-order overlay distributions was investigated. Although certain k-parameter components still exhibit residual overlay errors, the overall correction performance remains significant, demonstrating the model's robustness and applicability in high-precision overlay control.

In conclusion, the high-order overlay correction model proposed in this study offers an effective solution for compensating overlay errors in advanced lithography processes. While some residual errors remain under certain high-order k-parameter components, the model has demonstrated strong adaptability and stability in practical manufacturing environments. Looking forward, the integration of more complex stress modeling, finer localized thermal corrections, and AI-driven real-time compensation systems is expected to further enhance the accuracy and robustness of the approach, thereby supporting the manufacturing demands of future technology nodes with tighter overlay requirements.

Conflicts of interest

The authors declare no conflicts of interest.

Data availability

Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Acknowledgements

This work was supported by Frontier Technology Collaboration Project (QYJS-2023-2900-B), Strategic Priority Research Program of Chinese Academy of Sciences (XDA0330303), University of Chinese Academy of Sciences (118900M032), and Fundamental Research Funds for the Central Universities (E2ET3801).

References

1 J. A. Liddle and G. M. Gallatin, *Nanoscale*, 2011, **3**, 2679.
2 D. Han, T. Ye and Y. Wei, *Nanoscale Adv.*, 2023, **5**, 4424–4434.
3 D. Rui, L. Zhang, Y. Wei, H. Ding and Y. Su, *Opt. Express*, 2025, **33**, 29427.
4 D. Rui, L. Zhang, H. Ding, H. Shen, Y. Wei and Y. Su, *Opt. Express*, 2025, **33**, 17976.
5 I. Giannopoulos, I. Mochi, M. Vockenhuber, Y. Ekinci and D. Kazazis, *Nanoscale*, 2024, **16**, 15533–15543.
6 S.-G. Kim, D.-W. Shin, T. Kim, S. Kim, J. H. Lee, C. G. Lee, C.-W. Yang, S. Lee, S. J. Cho, H. C. Jeon, M. J. Kim, B.-G. Kim and J.-B. Yoo, *Nanoscale*, 2015, **7**, 14608–14611.
7 Y. Luo, L. Liu, W. Zhang, W. Kong, C. Zhao, P. Gao, Z. Zhao, M. Pu, C. Wang and X. Luo, *RSC Adv.*, 2017, **7**, 12366–12373.
8 J. Li, P. Van Nieuwkerk, M. A. Verschuuren, B. Koopmans and R. Lavrijsen, *Appl. Phys. Lett.*, 2022, **121**, 182407.
9 M. A. Verschuuren, M. Megens, Y. Ni, H. Van Sprang and A. Polman, *Adv. Opt. Technol.*, 2017, **6**, 243–264.
10 T. D. G. Nevels, L. J. M. Ruijs, P. Van De Meugheuvel, M. A. Verschuuren, J. G. Rivas and M. Ramezani, *J. Opt.*, 2022, **24**, 094002.
11 W. Kong, L. Liu, C. Wang, M. Pu, P. Gao, K. Liu, Y. Luo, Q. Jin, C. Zhao and X. Luo, *Nanoscale Adv.*, 2022, **4**, 2011–2017.
12 P. Gao, N. Yao, C. Wang, Z. Zhao, Y. Luo, Y. Wang, G. Gao, K. Liu, C. Zhao and X. Luo, *Appl. Phys. Lett.*, 2015, **106**, 093110.
13 P. Gao, M. Pu, X. Ma, X. Li, Y. Guo, C. Wang, Z. Zhao and X. Luo, *Nanoscale*, 2020, **12**, 2415–2421.
14 S. Wang, X. Dou, L. Chen, Y. Fang, A. Wang, H. Shen and Z. Du, *Nanoscale*, 2018, **10**, 11651–11656.
15 T. Ifuku, M. Hiura, Y. Takabayashi, A. Kimura, Y. Suzaki, T. Ito, K. Yamamoto, J. Choi, T. Estrada and D. Resnick, in *Photomask Technology 2022*, ed. B. S. Kasprowicz and T. Liang, SPIE, Monterey, United States, 2022, p. 13.
16 J. Zhang, L. Zhang, L. Han, Z.-W. Tian, Z.-Q. Tian and D. Zhan, *Nanoscale*, 2017, **9**, 7476–7482.
17 S. Y. Chou, C. Keimel and J. Gu, *Nature*, 2002, **417**, 835–837.
18 Z. Alnakhli, Z. Liu, F. AlQatari, H. Cao and X. Li, *Nanoscale Adv.*, 2024, **6**, 2954–2967.
19 D. Rui, L. Zhang, H. Shen, H. Ding, S. Liu, Y. Wei and Y. Su, *J. Vac. Sci. Technol. B*, 2024, **42**, 042601.

Open Access Article. Published on 05 September 2025. Downloaded on 10/7/2025 11:26:30 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

20 D. Rui, L. Zhang, H. Ding, H. Shen, Y. Wei and Y. Su, *J. Opt. Soc. Am. B*, 2024, **41**, 2702.
21 A. J. Den Boef, *Surf. Topogr. Metrol. Prop.*, 2016, **4**, 023001.
22 D. H. Beak, J. P. Choi, T. Park, Y. S. Nam, Y. S. Kang, C.-H. Park, K.-Y. Park, C.-H. Ryu, W. Huang and K.-H. Baik, Lens heating impact analysis and controls for critical device layers by computational method, *Optical Microlithography XXVI*, ed. W. Conley, SPIE, San Jose, California, United States, 2013, p. 86831Q.
23 B. Minghetti, T. Brunner, C. Robinson, C. Ausschnitt, D. Corliss and N. Felix, Overlay characterization and matching of immersion photoclusters, *Optical Microlithography XXIII*, ed. M. V. Dusa and W. Conley, SPIE, San Jose, California, 2010, p. 76400W.
24 IEEE International Roadmap for Devices and Systems, *Lithography and Patterning*, Institute of Electrical and Electronics Engineers, 2024.
25 M. Pike, T. Brunner, B. Morgenfeld, N. Jing and T. Wiltshire, *J. Micro/ Nanolithogr. MEMS MOEMS.*, 2016, **15**, 021406.
26 A. Lam, F. Pasqualini, J. De Caunes and M. Gatefait, Overlay breakdown methodology on Immersion Scanner, *Metrology, Inspection, and Process Control for Microlithography XXIV*, ed. C. J. Raymond, SPIE, San Jose, California, 2010, p. 76383L.
27 W. Niu, W. He, J. Li and Q. Li, *Appl. Therm. Eng.*, 2025, **262**, 125281.
28 D. Rui, L. Zhang, Y. Wei and Y. Su, *Nanotechnology*, 2025, **36**, 205301.
29 A. Hager, L. Güniat, N. Morgan, S. P. Ramanandan, A. Rudra, V. Piazza, A. Fontcuberta I Morral and D. Dede, *Nanotechnology*, 2023, **34**, 445301.
30 Q. Zhang, K. Poolla and C. J. Spanos, Modeling of Mask Thermal Distortion and Its Dependency on Pattern Density, *Photomask and Next-Generation Lithography Mask Technology XII*, ed. M. Komuro, SPIE, Yokohama, Japan, 2005, p. 234.
31 M. Lim, G. Kim, S. Kim, B. Lee, S. Kim, C. Lim, M. Kim and S. Park, Investigation on reticle heating effect induced overlay error, Metrology, *Inspection, and Process Control for Microlithography XXVIII*, ed. J. P. Cain and M. I. Sanchez, SPIE, San Jose, California, USA, 2014, p. 905014.
32 L. Chen, X. Wang, Y. Wei and C. Zhou, *Micromachines*, 2022, **13**, 997.

 © 2025 The Author(s). Published by the Royal Society of Chemistry

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：汤瑞丰
方向：集成电路
单位：中国科学院微电子研究所

题目：Region-Selective Oxygen Vacancy Engineering for Ferroelectric HZO Thin Films Processed at 300 °C

ACS APPLIED MATERIALS & INTERFACES

www.acsami.org

Research Article

Region-Selective Oxygen Vacancy Engineering for Ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ Thin Films Processed at 300 °C

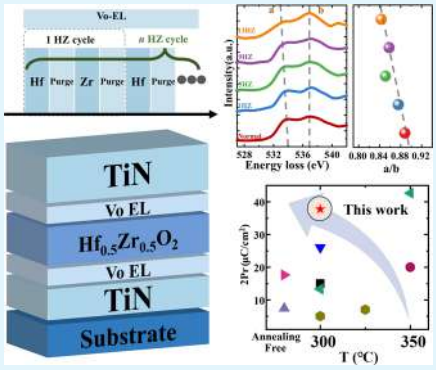
Ruifeng Tang, Yifan Zhang, Yang Yang, Chuanli Zhang, Xun Yu, Yuan Wang, Xiao Long, Pengfei Jiang, Tiancheng Gong, Yan Wang, Wei Wei,* and Qing Luo*

Cite This: https://doi.org/10.1021/acsami.5c08743

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The discovery of ferroelectricity in hafnium oxide (HfO_2) thin films has positioned it as a leading material for next-generation nonvolatile memory. However, the integration of HfO_2-based ferroelectric thin films into back-end-of-line (BEOL) processes remains challenging due to the high thermal budget required to stabilize the ferroelectric orthorhombic (O) phase. In this work, we achieve robust stabilization of the O-phase at a significantly reduced annealing temperature of 300 °C solely through oxygen vacancy engineering. We introduce a region-selective oxygen vacancy engineering strategy to form oxygen vacancy engineering layers (Vo-ELs) within $Hf_{0.5}Zr_{0.5}O_2$ (HZO) thin films during atomic layer deposition (ALD). By delaying the introduction of the oxygen precursor, multiple cycles of metal precursors are deposited before a single oxidation step, creating well-defined Vo-ELs. These Vo-ELs induce a vertical gradient in the oxygen vacancy concentration, as confirmed by electron energy loss spectroscopy (EELS). First-principles calculations further reveal that oxygen vacancies reduce the energy barrier for the tetragonal-to-orthorhombic (T−O) phase transition and enhance the thermodynamic stability of the ferroelectric O-phase. Utilizing this technique, we successfully realize low-temperature (300 °C) fabrication of HZO ferroelectric capacitors, which exhibit a high remanent polarization of 36.4 $\mu C/cm^2$ and outstanding endurance exceeding 10^9 cycles. This work demonstrates the effectiveness of Vo-ELs in enabling low-thermal-budget, high-performance ferroelectric devices compatible with advanced BEOL integration.



KEYWORDS: *ferroelectricity, hafnium oxide, oxygen vacancy, low-thermal budget, BEOL compatibility*

Downloaded via qing luo on September 9, 2025 at 01:45:16 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

■ INTRODUCTION

The discovery of ferroelectricity in hafnium oxide (HfO_2) has positioned this fluorite-structured material as a promising candidate for next-generation nonvolatile memory applications. Such materials exhibit exceptional characteristics, including low power consumption, high speed, miniaturizability, and complementary metal oxide semiconductor (CMOS) compatibility, which impart significant potential to HfO_2-based ferroelectric random access memory (FeRAM) for applications in very large scale integrated (VLSI) circuits.[1−3] The ferroelectricity of HfO_2-based films is primarily attributed to the formation of the orthorhombic (O) phase ($Pca2_1$ space group) after an annealing process. While early studies typically employed annealing temperatures exceeding 400 °C, achieving an optimal balance between polarization and endurance often necessitates even higher thermal budgets.[4,5] However, this high thermal budget poses a significant challenge for integrating ferroelectric functionality into the back-end-of-line (BEOL) processes of standard CMOS integrated circuits, as these processes strictly require a thermal limit below 400 °C.[6,7] Therefore, low-temperature processing of HfO_2-based ferroelectric thin films is essential for ensuring BEOL compatibility.[8] Moreover, it also plays a crucial role in applications such as 3D-stacked nonvolatile memories,[9−11] as well as flexible and wearable electronics.[12−14]

To enable the low-temperature processing of HfO_2-based ferroelectric devices, Zr-doped HfO_2, particularly $Hf_{0.5}Zr_{0.5}O_2$ (HZO), has been widely adopted due to its lower crystallization temperature compared to other doped HfO_2 systems.[15] Many studies have explored low-temperature processing techniques based on HZO films, including alternative annealing methods, diverse substrates, the introduction of interlayers, and other strategies, often by the introduction of additional processing steps. Joh et al.[16] and Kim et al.[17] proposed deep ultraviolet irradiation prior to RTA (DUV-RTA) and high-pressure annealing RTA, reaching a remanent polarization (2Pr) of ∼13.9 $\mu C/cm^2$ after annealing at 350 °C and 2Pr > 20 $\mu C/cm^2$ after annealing at 450 °C, respectively. Lyu et al.[18] demonstrated that growing a ZrO_2

Received: May 1, 2025
Revised: August 8, 2025
Accepted: August 12, 2025

© XXXX American Chemical Society

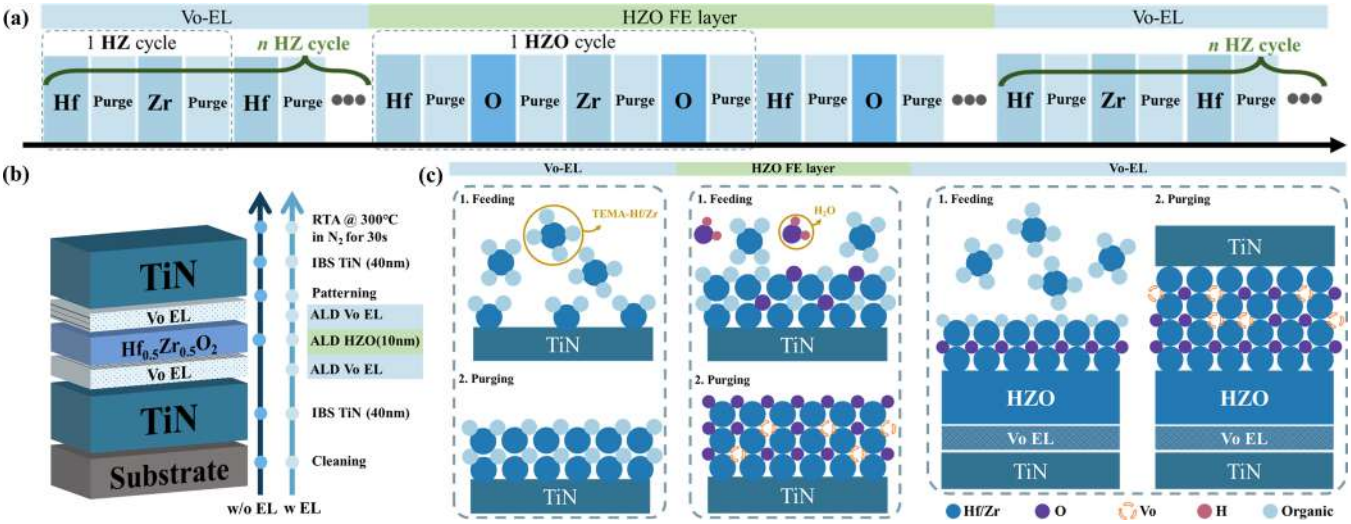


Figure 1. Region-selective oxygen vacancy engineering strategy for precise oxygen vacancy engineering and fabrication flow of ferroelectric HZO capacitors. (a) Schematic illustration of the ALD sequence used to deposit the Vo-ELs and the ferroelectric HZO layer. (b) Process flow for fabricating ferroelectric capacitors incorporating Vo-EL structures. IBS: ion beam sputtering; TiN: titanium nitride. (c) Schematic depiction of precise oxygen vacancy engineering in different layers: (left) bottom Vo-EL, (middle) conventional HZO ferroelectric layer, and (right) top Vo-EL. Orange dashed circles represent randomly distributed oxygen vacancies within the engineered layers.

interfacial layer on a Ge substrate allowed for a 2Pr of 38 μC/cm^2 using RTA at 300 °C. Similarly, Gaddam et al.[19] achieved a 2Pr of ~33.2 μC/cm^2 while lowering the annealing temperature to 350 °C by incorporating a TiO_2/HZO/HfO_2 trilayer structure. While these techniques reduce the process temperature, they either employ annealing methods incompatible with BEOL integration or introduce additional processing steps. Beyond increasing the process complexity, such strategies often make it difficult to isolate the underlying mechanism that enables ferroelectric phase stabilization at low temperatures. Moreover, current approaches lack effective strategies for tuning device properties under such thermal budgets. This indicates that the development of a more straightforward and BEOL-compatible low-temperature process requires further exploration of the intrinsic mechanisms governing ferroelectric phase formation in HZO.

One promising strategy to achieve low thermal budget ferroelectricity is to harness the intrinsic properties of ferroelectric HZO thin films without relying on external techniques.[20,21] Among these, oxygen vacancies (Vo), commonly formed during fabrications, may have potential as intrinsic enablers of low thermal budget ferroelectricity, though this possibility remains unexplored.[22,23] This possibility originates from the role of Vo in stabilizing the ferroelectric orthorhombic phase over nonpolar competing phases, as theoretically validated by density functional theory (DFT) calculations.[24] Nevertheless, the controlled and reproducible introduction of Vo in experimental settings remains challenging, posing a major obstacle to practical implementation. Pujar et al.[25] introduced Vo by uniformly incorporating La dopants into $HfLaO_2$ thin films, achieving this via a spin-coating process followed by multiple long-duration high-temperature annealing steps. However, this approach requires non-BEOL-compatible techniques, including solution-based doping and postdeposition crystallization at 700−800 °C, which limit its practical applicability. Kang et al.[26] successfully introduced and modulated Vo through light ion bombardment, which in turn significantly enhanced the ferroelectricity of HfO_2-based thin films. Nonetheless, the requirement of high-dose ion implantation and high-temperature pulsed laser deposition (PLD) at 700 °C through a sample fabrication process imposes significant limitations on the scalability and BEOL compatibility. Therefore, it remains imperative to develop a controllable and BEOL-compatible strategy for the generation of Vo.

To explore the relationship between the intrinsic Vo defects and low-temperature ferroelectricity, this work developed a region-selective oxygen vacancy engineering strategy within the atomic layer deposition (ALD) process, which is substantially different from conventional cycles, to regulate the Vo concentration in specific regions of HZO films. This method introduces Vo engineering layers (Vo-EL) into the standard HZO films to induce a localized Vo. The ferroelectric films with Vo-EL exhibit crystallization of the O-phase after rapid thermal annealing (RTA) at 300 °C. The 2Pr exhibits a clear dependence on Vo concentration, with 2Pr values up to 36.4 μC/cm^2 and endurance of 10^9. Subsequently, first-principles calculations reveal that Vo can reduce the phase transition barrier from the tetragonal (T) phase to the O-phase, and the barrier height is directly proportional to the Vo concentration. These findings demonstrate that Vo is a key factor in stabilizing the ferroelectricity in HZO films processed at low temperatures. Our proposed method enables the introduction of Vo in specific regions while modulating their concentration, not only achieving highly BEOL-compatible HZO ferroelectric films but also providing an approach to meet the customization requirements of FE devices.

■ RESULTS AND DISCUSSION

Engineering Localized Oxygen Vacancy in HZO Films. A standard process to deposit HZO thin films within ALD involves the alternate introduction of Hf, Zr, and oxygen precursors, with purging steps between each precursor to ensure complete reactions.[27] Specifically, one standard cycle consists of eight steps: Hf precursor feeding, purging, oxygen precursor feeding, purging, Zr precursor feeding, purging, oxygen precursor feeding, and a final purging, which forms a stoichiometric $Hf_{0.5}Zr_{0.5}O_2$. To introduce Vo, we propose a region-selective oxygen vacancy engineering strategy. The

https://doi.org/10.1021/acsami.5c08743
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

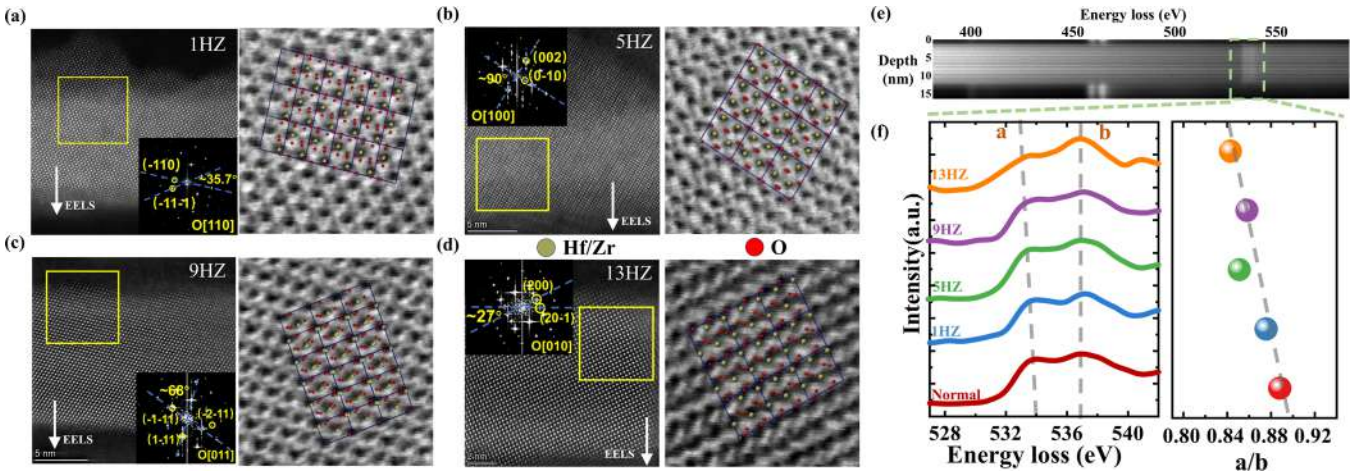


Figure 2. Device characteristics and structural analysis of crystal phases and oxygen vacancies. Characterization results of the four Vo-EL samples: (a–d) correspond to 1HZ, 5HZ, 9HZ, and 13HZ. Left: cross-sectional STEM-HAADF images of the HZO structures with FFT patterns extracted from the regions marked by yellow squares. Arrows indicate the EELS line scan direction used in subsequent measurements. Right: ABF images of the HZO structures with O-phase, overlaid with simulated atomic models along various crystallographic axes. (e) Oxygen K edge EELS spectra for the 9HZ sample, ranging from 330 to 620 eV with a step size of 0.2 eV. (f) (left) Oxygen K edge ELNES spectra and (right) corresponding estimated a/b ratios at the Vo-EL edges across different samples.

strategy modifies the standard cycle to four steps consisting of Hf precursor feeding, purging, Zr precursor feeding, purging, which is named as the 1HZ cycle. After the nHZ cycles, the oxygen precursor is introduced. In this way, we can control the local Vo concentration by adjusting n.

However, one critical question is whether Hf and Zr precursors can accumulate in the absence of an oxygen precursor while being subjected to purging steps. To investigate this, we designed a control experiment using repeated supercycles composed of 9 consecutive Hf–Zr precursor pulses followed by a single oxygen pulse (denoted as 9HZ+1O). A total of 38 such supercycles was performed, nominally equivalent in number to that used for standard 10 nm HZO deposition. If the 9HZ sequence fails to accumulate, the deposition process of supercycles would effectively be equivalent to standard cycles (denoted as 1Hf+1O+1Zr+1O), resulting in an estimated film thickness of approximately 10 nm. Interestingly, the resulting film exhibited a significantly increased thickness of ∼55 nm, corresponding to a growth rate of ∼1.4 nm per supercycle. This confirms that within a certain range of n-values, Hf and Zr precursors can indeed accumulate without immediate oxidation. Furthermore, the 55 nm-thick HZO film exhibited partial crystallization after annealing at 500 °C and achieved 2Pr close to 10 μC/cm^2 (Figure S1).

Based on the proposed strategy, we grow nHZ layers to form Vo-ELs both before and after the formation of the FE layer, as illustrated in Figure 1a. Figure 1b illustrates the fabrication process of metal–ferroelectric–metal (MFM) capacitors with and without Vo-ELs. Atomic-level stoichiometry controls of different layers are illustrated in Figure 1c. Before the growth of the FE layer, n cycles of the Hf (Zr) precursor were introduced (including feeding and purging as well) without the oxygen precursor, constructing an atmosphere lacking oxygen (left).[20] Subsequently, the HZO layer (thickness of 10 nm) is grown following the standard process (middle). Once the oxygen precursor is introduced to the oxygen-lacking environment, the excess Hf (Zr) precursor undergoes incomplete reactions, generating an oxygen-deficient area, i.e., Vo-EL formation. The remaining layer tends to form a stoichiometric HZO layer. After that, the same method is applied to create the upper Vo-EL (right). In summary, this process is employed to grow HZO under an oxygen-deficient environment, temporarily constructing symmetric Vo-ELs at the bottom and the top interfaces near the electrodes. In addition, we also fabricated devices with the Vo-EL embedded at the center of the HZO film (as shown in Figure S2). However, these samples exhibited inferior performance compared to their interfacial counterparts, which is primarily focused on in this work. Furthermore, four device samples with interfacial Vo-ELs of varying Vo concentration (Vo%) are created by controlling the Hf (Zr) precursor cycle numbers n, which are 1, 5, 9, and 13, denoted as 1HZ, 5HZ, 9HZ, and 13HZ, respectively.

Microstructures of HZO Film with Vo-EL. To verify the presence of O-phase grains, aberration-corrected scanning transmission electron microscopy (AC-STEM) and fast Fourier transform (FFT) are applied on HZO capacitors with Vo-EL, which is annealed at 300 °C and woken up by 10^4 cycles. Figure 2a–d shows specific crystal phases and atomic structures of each sample corresponding to 1HZ, 5HZ, 9HZ, and 13HZ. The high-resolution high-angle annular dark field (HAADF) images on the left illustrate the film stack structure of the devices, confirming that the HZO film has a thickness of 10 nm as expected. Distinct crystallized regions are observed in the FE layer, and the FFT results within each STEM image of the yellow-marked region exhibit a perfect match with the simulated diffraction pattern of the O-phase ($Pca2_1$) in different orientations. Correspondingly, the sample without Vo-EL does not exhibit a distinct crystallized region (Figure S3). To further confirm the crystal phase structure of the FE layer in samples annealed at 300 °C, the corresponding annular bright-field (ABF) images (displayed on the right of each figure) are compared with the simulated atomic structure of the O-phase (overlaid on the ABF results). Among these $Pca2_1$-phase grains with different orientations, the [100]-oriented grain in Figure 2b images both the oxygen atoms responsible for polarization switching within the lattice and the oxygen atoms in the spacer layers.[28,29] The intact ferroelectric O-phase grains observed in these samples indicate that Vo-EL is a key factor in triggering the annealing of the annealed O-

https://doi.org/10.1021/acsami.5c08743
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

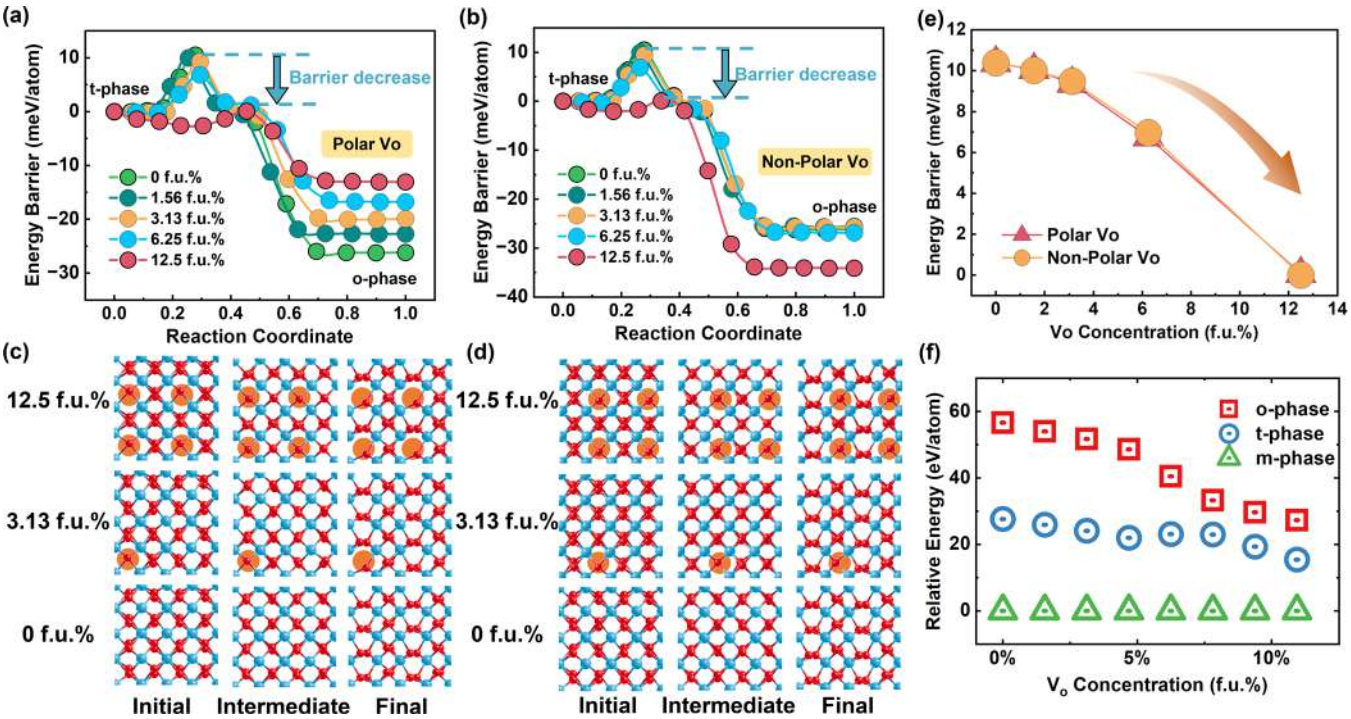


Figure 3. T−O phase transition facilitated by Vo predicted by first-principles calculations. (a) and (b) show the energy evolution pathways of the T−O phase transition calculated by the vc-NEB method, for polar and nonpolar Vo configurations, respectively. (c) and (d) illustrate the corresponding atomic structure evolution during the phase transition, with selected key frames highlighted. (e) Energy barrier of the T−O phase transition as a function of Vo concentration, comparing polar and nonpolar Vo cases. (f) Relative energies of the $P4_2/nmc$ (T-phase) and $Pca2_1$ (O-phase) with respect to the $P2_1/c$ (M-phase), as a function of Vo concentration.

phase grain at 300 °C, which is directly related to the distribution of Vo in the film.

The distribution of Vo inside the HZO films is studied by electron energy loss spectroscopy (EELS) spectrum imaging analysis of the oxygen K edge for each sample, including the device without Vo-ELs. The EELS spectrum, presented in Figure 2e, scans from the top interface layer through the whole FE layer to the bottom layer (as the arrows in Figure 2a-d). The averaged oxygen K spectra are extracted from the 2 nm region around Vo-EL for the five HZO thin films (the left of Figure 2f). The oxygen K energy-loss near edge structure (ELNES) of Vo-EL exhibits a pronounced double-peak feature, labeled as peak *a* (∼534 eV) and peak *b* (∼538 eV),[26] with an energy separation (ΔE_{b-a}) of approximately 4 eV. This characteristic arises from the hybridization of the oxygen 2p orbital with the d_{e_g} and $d_{t_{2g}}$ orbitals of Hf (or Zr), respectively.[30] The characteristic peak splitting of the oxygen K edge ELNES is commonly observed in HfO_2 polymorphs with minor variations in energy shifts and peak broadening.[30] Notably, the sharpness and relative intensity of peak *a* compared to those of peak *b* are influenced by the crystallinity and Vo concentration in HfO_2 films. Poor crystallinity leads to a broadening of the double-peak structure and a reduction in the relative intensity of peak *a*.[31,32] Since EELS are conducted on the same O-phase in all four Vo-EL samples, the most probable cause of the observed changes in the oxygen K edge ELNES is the formation of Vo. As depicted, the significant decrease of the *a*/*b* ratio from 0.88 to 0.84 in the oxygen K edges suggests an increase in Vo defects as the cycle numbers *n* increased from 1 to 13. Furthermore, to establish a comparison, we selected regions of 2 nm within the body of each sample and performed oxygen K edge spectra analysis in the same (Figure S4). We obtained corresponding *a*/*b* ratio results, which showed relatively small differences, showing that the Vo defects did not exhibit significant variations across different samples. More importantly, the sample without Vo-ELs, which is poorly crystallized after RTA at 300 °C, displays a much larger *a*/*b* ratio. Previous studies on Vo engineering[26,32−34] have also demonstrated a similar dependence of the *a*/*b* ratio on Vo concentration. This observation indicates a significant reduction of the Vo concentration in Vo-EL samples compared to the conventional sample.

The introduction of Vo-EL may influence two key factors, i.e., the gradient distribution of Vo and the presence of carbon impurities. However, the EELS analysis of the carbon element shows no significant variation across the entire film (Figure S5), ruling out the possibility that carbon impurities induce ferroelectricity. Consequently, STEM-based structure analysis confirms the enhanced O-phase crystallinity in samples incorporating Vo-EL at low annealing temperatures, while EELS spectra further verify the existence of a Vo concentration gradient across samples with different *n* cycles. These findings indicate a strong correlation between the crystallization of the ferroelectric O-phase and the Vo concentration.

Phase Transition Facilitated by Vo. The ferroelectric properties of HZO thin films primarily originate from the orthorhombic phase, which is widely believed to form through the transformation of the tetragonal (T) phase.[35] This phase transition requires overcoming a significant energy barrier, necessitating a high rapid thermal annealing (RTA) temperature.[36] Using the variable-cell Nudged Elastic Band (vc-NEB) method, the energy barriers of the T-O phase transition are calculated under varying Vo concentrations and types, namely

https://doi.org/10.1021/acsami.5c08743
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

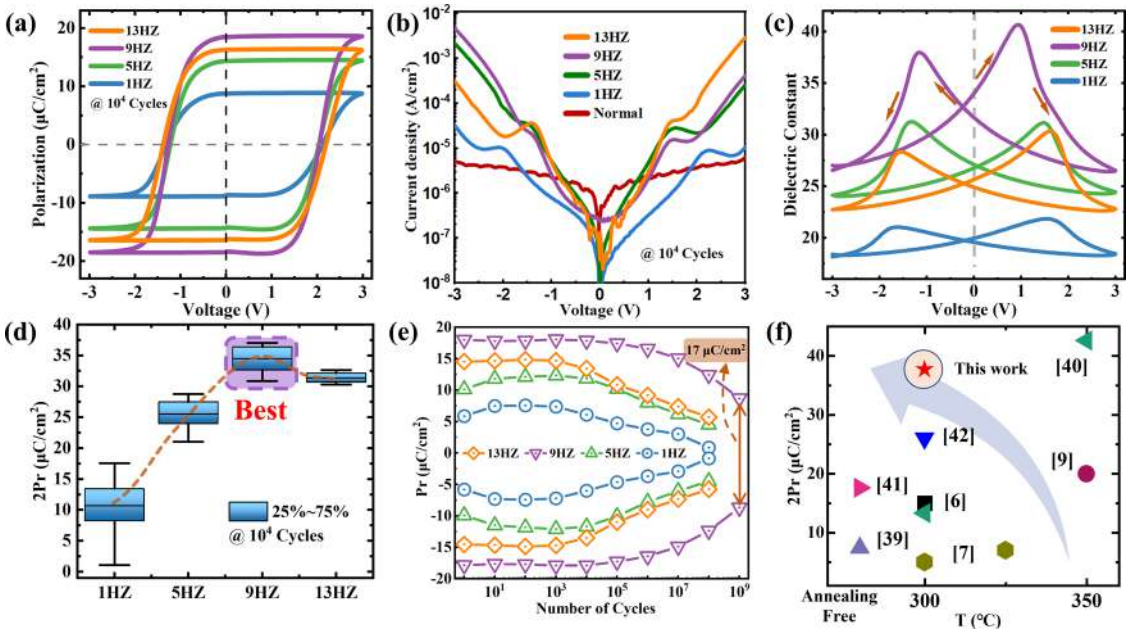


Figure 4. Ferroelectric properties of the Vo-EL capacitors annealed at 300 °C. (a) $P-V$ hysteresis loops measured using the PUND method at 10 kHz. (b) $I-V$ characteristics showing leakage current behavior of the ferroelectric capacitors under a DC voltage sweep. (c) Dielectric constant as a function of voltage for capacitors with different Vo-EL cases. The scan was performed from −3 to +3 V and then back to −3 V using a small-signal amplitude of 50 mV at 10 kHz. (d) Box plot summarizing the 2Pr values extracted from PUND measurements across 20 devices for each fabrication condition. All measurements in (a)−(d) were performed after wake-up cycling using 10^4 trapezoidal pulses at 100 kHz. (e) Endurance characteristics of the various samples, demonstrating the evolution of 2Pr during electric cycling. The endurance tests were conducted under identical conditions using bipolar triangular pulses (±3 V and 1 MHz). (f) Comparison of 2Pr versus annealing temperature among related works.

polar Vo (vacancy of polar oxygen) and nonpolar Vo (vacancy of nonpolar oxygen), as illustrated in Figure 3a,b.

The models with Vo concentrations of 0/1.56/3.13/6.25/12.5 f.u.% are constructed by removing 0/1/2/3/4 oxygen atoms within the unit cell of the O-phase with the $Pca2_1$ space group. With the Vo concentration increasing, the energy barrier decreases. Especially for the Vo concentration of 12.5 f.u.%, the energy barrier drops sharply and vanishes to almost zero, indicating that the T−O phase transition may occur spontaneously. To illustrate the specific movement of oxygen atoms as well as Vo during the phase transition, the initial T-phase, the transition phase (highest-energy state during the evolution process), and the final O-phase with Vo marked by the orange solid circles are shown in Figure 3c (polar Vo) and Figure 3d (nonpolar Vo). During the T−O transition without Vo, all oxygens in the T-phase are initially 4-coordinated. Upon transformation, these 4-coordinated oxygens evolve into a mix of 3-coordinated and 4-coordinated oxygens in the O-phase, respectively. While for the T−O transition with Vo, the transition still proceeds in an ordered manner, suggesting that the Vo-rich O-phase can support a substantial amount of defects while maintaining the polar structure.[37] The variations of the phase transition barrier as a function of Vo concentration are plotted in Figure 3e, revealing a negative correlation between the transition barrier and Vo concentration. Furthermore, we calculated the relative energy differences between the T and the O phases with respect to the M phase as the Vo concentration increased (Figure 3f). The results reveal that as the Vo concentration rises, the energies of the T and O phases progressively converge toward that of the M phase. This reduction in the energy of the T and O phases may promote their crystallization under low-temperature processing conditions. Moreover, the decreasing energy difference between the T and O phases further facilitates the T-to-O phase transition.[37]

Electrical Characterization of Vo-EL Capacities. To validate the applicability of the proposed method in low-temperature processed ferroelectric memory devices, electrical characterizations are performed on HZO ferroelectric capacitors with Vo-EL, which are annealed at 300 °C. The polarization-voltage ($P-V$) hysteresis loops of HZO thin films with Vo-EL are measured by the positive up negative down (PUND) method to evaluate the ferroelectric properties (Figure 4a). All the samples with different Vo-EL exhibit distinct ferroelectric hysteresis characteristics and considerable remanent polarization (2Pr). Notably, the 1HZ sample achieves a 2Pr of 18 $\mu C/cm^2$, whereas the standard sample without Vo-EL does not exhibit any ferroelectric behavior under a 300 °C annealing temperature. This indicates that the introduction of Vo-EL is highly effective in inducing ferroelectricity. Furthermore, the 9HZ sample achieves the highest $2P_r$ of 36.4 $\mu C/cm^2$, demonstrating the strong impact on the polarization enhancement of Vo-EL. Figure 4b, c shows the operating current and dielectric response, respectively. Compared with the standard sample, the Vo-EL samples exhibit higher leakage currents at 3 V, with a general trend of increasing leakage as the number of HZ cycles increases. As for the dielectric constant, the 9HZ sample exhibits the highest value, while the 1HZ sample shows the lowest, following the same trend as the remanent polarization. Furthermore, we measured the 2Pr values of 20 individual devices in each sample, as shown in Figure 4d. The results show that the distribution of 2Pr becomes progressively narrower with an increasing HZ cycle number, indicating improved device-to-device variations at higher HZ values. Specifically, the 13HZ sample exhibits the most concentrated distribution, while the

https://doi.org/10.1021/acsami.5c08743
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

1HZ sample shows the widest spread. Finally, the endurance characteristics of each Vo-EL sample are presented in Figure 4e. All samples exhibit varying degrees of fatigue, likely due to the presence of high Vo concentrations and their domain pinning effects.[22,33,38] Notably, the 9HZ sample, which initially shows a higher 2Pr, maintains a 2Pr of 17 $\mu C/cm^2$ even after 10^9 cycles, demonstrating superior fatigue resistance. Figure 4f presents a benchmark comparison of recent studies on low-temperature annealing processes. In this plot, the upper-left region represents a more favorable trade-off, where both low annealing temperature and high ferroelectric performance are achieved.[39−42] Notably, this work not only demonstrates competitive 2Pr values under a significantly reduced annealing temperature of 300 °C, but also achieves low-temperature ferroelectricity simply by modifying the oxygen precursor injection point, without introducing additional process complexity (Table S1). Above all, these electrical results indicate that different Vo-EL samples enable tunable ferroelectric performance, offering adaptability for diverse application scenarios.

■ CONCLUSIONS

This study presents a region-selective oxygen vacancy engineering strategy for Vo engineering and demonstrates its effectiveness in stabilizing the ferroelectric O phase in HZO thin films processed at a reduced thermal budget of 300 °C. By precise control of the oxygen precursor during ALD to form Vo-EL, localized oxygen vacancies were introduced, resulting in a vertical gradient distribution of Vo, as confirmed by atomic-resolution STEM and EELS analysis. First-principles calculations further reveal that Vo incorporation significantly lowers the energy barrier for the T−O phase transition and improves the thermodynamic stability of the O-phase. Utilizing this approach, the fabricated HZO ferroelectric capacitors exhibited a high 2Pr of 36.4 $\mu C/cm^2$ and outstanding endurance exceeding 10^9 cycles. These results highlight the strong potential of Vo engineering for realizing low-temperature, high-performance ferroelectric devices compatible with BEOL integration, offering a promising pathway for future 3D-stacked nonvolatile memories and flexible electronic applications.

■ METHOD

Device Fabrication. Metal-ferroelectric-metal (MFM) capacitors were fabricated, as illustrated in Figure 1d. The process began with standard cleaning of (100)-oriented p-type silicon wafers with a resistivity of less than 0.005 Ω·cm. A 40 nm thick TiN bottom electrode was then deposited via ion beam sputtering. Subsequently, a 10 nm thick HZO ferroelectric layer was deposited by using ALD at 280 °C using $Hf[N(C_2H_5)CH_3]_4$ (TEMA-Hf) and $Zr[N(C_2H_5)CH_3]_4$ (TEMA-Zr) as metal precursors. H_2O is used as an oxygen source. Following the HZO deposition, a 40 nm thick TiN was patterned as the top electrode via photolithography and sputtering. Each capacitor had a device area of 100 × 100 μm^2. Postmetallization annealing (PMA) was carried out in an N_2 ambient at 300 °C for 30 s using rapid thermal annealing (RTA) to induce crystallization of the HZO layer. The Vo-EL are inserted between the HZO thin film and the electrodes, and their formation is described in detail in the main text.

Electrical Characterization. Electrical measurements were carried out using an Agilent B1500 Semiconductor Parameter Analyzer and a Radiant Workstation Ferroelectric Tester to evaluate the ferroelectric properties of the MFM capacitors. The current density−voltage ($I-V$) and endurance characteristics were measured. Polarization−voltage ($P-V$) behavior was characterized using a PUND waveform with an amplitude of ±3 V and a frequency of 10 kHz. Capacitance−voltage ($C-V$) measurements, corresponding to dielectric constant-voltage tests, were conducted by using an AC probing frequency of 10 kHz with a 50 mV amplitude. For endurance testing, $P-V$ measurements were periodically recorded while the devices were subjected to endurance cycling using a trapezoidal waveform with a 3 V amplitude at 1 MHz.

DFT Calculations. In order to accurately simulate the V_O in the DFT calculation, $HfO_{2-\delta}$ was constructed by removing oxygen atoms from the super cells of HfO_2. The T- and polar O-phases of the 12-atom unit cell (4 hafnium and 8 oxygen) are used to simulate the T-O phase transition in the absence of V_O, as shown in Figure S6. The detailed lattice constants for each V_O concentration configuration are provided in Tables S2 and S3 of the Supporting Information. First-principles calculations are performed in the DS-PAW software using the projector-augmented wave (PAW) method[43,44] and the Perdew−Burke−Ernzerhof exchange-correlation energy functional.[45] All configurations of V_O concentrations are optimized using the conjugate gradient (CG) algorithm for lattice constants and atomic positions until the forces on each atom are less than 0.01 eV/Å. The VCNEB method[46] is used to calculate the phase transition barriers from T to O at different V_O concentrations until the forces on each atom are converged to become smaller than 0.02 eV/Å.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsami.5c08743.

FE layer formed through supercycles mentioned before by ALD, fabrication process, performance of devices with Vo-EL embedded at the center of the HZO film, cross-sectional STEM-HAADF images of normal HZO structure without Vo-EL, oxygen K-ELNES, and estimated a/b ratios within the body of FE layers of various samples, carbon K EELS stripe of various samples, atomic structure of the T phase and O phase, benchmark of recent studies that focus on low thermal budget HZO, and structural data for different polar and nonpolar Vo concentration configurations (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Wei Wei − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*; Email: weiw@ime.ac.cn

Qing Luo − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*; orcid.org/0000-0002-3419-4400; Email: luoqing@ime.ac.cn

Authors

Ruifeng Tang − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of*

https://doi.org/10.1021/acsami.5c08743
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China; School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100029, China

Yifan Zhang − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China; School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100029, China*

Yang Yang − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*

Chuanli Zhang − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China; School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100029, China*

Xun Yu − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China; School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100029, China*

Yuan Wang − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*

Xiao Long − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*

Pengfei Jiang − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*

Tiancheng Gong − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*

Yan Wang − *State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsami.5c08743

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDA0330000; the MOST of China (2022YFB3608400); the National Natural Science Foundation of China (NSFC) (62425407, 62304246, 62404246); the China Postdoctoral Science Foundation (BX20230405); and the postgraduate research opportunities program of HZWTECH (HZWTECH-PROP).

■ REFERENCES

(1) Böscke, T. S.; Müller, J.; Bräuhaus, D.; Schröder, U.; Böttger, U. Ferroelectricity in Hafnium Oxide Thin Films. *Appl. Phys. Lett.* **2011**, *99* (10), 102903.

(2) Khosla, R.; Sharma, S. K. Integration of Ferroelectric Materials: An Ultimate Solution for Next-Generation Computing and Storage Devices. *ACS Appl. Electron. Mater.* **2021**, *3* (7), 2862−2897.

(3) Song, M.-K.; Kang, J.-H.; Zhang, X.; Ji, W.; Ascoli, A.; Messaris, I.; Demirkol, A. S.; Dong, B.; Aggarwal, S.; Wan, W.; Hong, S.-M.; Cardwell, S. G.; Boybat, I.; Seo, J.; Lee, J.-S.; Lanza, M.; Yeon, H.; Onen, M.; Li, J.; Yildiz, B.; Del Alamo, J. A.; Kim, S.; Choi, S.; Milano, G.; Ricciardi, C.; Alff, L.; Chai, Y.; Wang, Z.; Bhaskaran, H.; Hersam, M. C.; Strukov, D.; Wong, H.-S. P.; Valov, I.; Gao, B.; Wu, H.; Tetzlaff, R.; Sebastian, A.; Lu, W.; Chua, L.; Yang, J. J.; Kim, J. Recent Advances and Future Prospects for Memristive Materials, Devices, and Systems. *ACS Nano* **2023**, *17* (13), 11994−12039.

(4) Kim, J.-H.; Onaya, T.; Park, H. R.; Jung, Y. C.; Le, D. N.; Lee, M.; Hernandez-Arriaga, H.; Zhang, Y.; Tsai, E. H. R.; Nam, C.-Y.; Nabatame, T.; Kim, S. J.; Kim, J. Toward Low-Thermal-Budget Hafnia-Based Ferroelectrics *via* Atomic Layer Deposition. *ACS Appl. Electron. Mater.* **2023**, *5* (9), 4726−4745.

(5) Liang, Y.-K.; Huang, Y.-S.; Peng, L.-C.; Yang, T.-Y.; Young, B.-F.; Lu, C.-C.; Yeong, S. H.; Lin, Y.-M.; Su, C.-J.; Chang, E.-Y.; Lin, C.-H. Effects of Annealing Temperature on TiN/$Hf_{0.5}Zr_{0.5}O_2$/TiN Ferroelectric Capacitor Prepared by *In-Situ* Like Consecutive Atomic Layer Deposition. *IEEE Trans. Nanotechnol.* **2022**, *21*, 328−331.

(6) Kim, S. J.; Mohan, J.; Summerfelt, S.; Kim, J. Ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ Thin Films: A Review of Recent Advances. *Jom* **2019**, *71*, 246−255.

(7) Tahara, K.; Toprasertpong, K.; Hikosaka, Y.; Nakamura, K.; Saito, H.; Takenaka, M.; Takagi, S. Strategy Toward HZO BEOL-FeRAM with Low-Voltage Operation (≤ 1.2 V), Low Process Temperature, and High Endurance by Thickness Scaling. In *2021 Symposium on VLSI Technology*; IEEE: Kyoto, Japan, 2021; pp 1−2.

(8) Francois, T.; Pellissier, C.; Slesazeck, S.; Havel, V.; Richter, C.; Makosiej, A.; Giraud, B.; Breyer, E. T.; Materano, M.; Chiquet, P.; Bocquet, M.; Grenouillet, L.; Nowak, E.; Schroeder, U.; Gaillard, F.; Coignus, J.; Blaise, P.; Carabasse, C.; Vaxelaire, N.; Magis, T.; Aussenac, F.; Loup, V. Demonstration of BEOL-Compatible Ferroelectric Hf0.5 Zr0.5 O2 Scaled FeRAM Co-Integrated with 130nm CMOS for Embedded NVM Applications. In *2019 IEEE International Electron Devices Meeting (IEDM)*; IEEE: San Francisco, CA, 2019; pp 15.7.1−15.7.4.

(9) Dutta, S.; Ye, H.; Chakraborty, W.; Luo, Y.-C.; Jose, M. S.; Grisafe, B.; Khanna, A.; Lightcap, I.; Shinde, S.; Yu, S.; Datta, S. Monolithic 3D Integration of High Endurance Multi-Bit Ferroelectric FET for Accelerating Compute-In-Memory. In *2020 IEEE International Electron Devices Meeting (IEDM)*; IEEE: San Francisco, CA, 2020; pp 36.4.1−36.4.4.

(10) Lim, E.; Park, Y.; Youn, C.; Kim, S. 3D Vertical Ferroelectric Capacitors with Excellent Scalability. *Nano Lett.* **2025**, *25* (6), 2166−2172.

(11) Liao, J.; Shi, W.; Yang, J.; Zhang, S.; Yang, W.; Wang, B.; Yan, F.; Xu, H.; Xie, R.; Wang, Y.; Zhou, Y.; Liao, M. Sub-5 Nm Al-Doped HfO2 Ferroelectric Thin Films Compatible with 3D NAND Process. *J. Alloys Compd.* **2024**, *1007*, No. 176327.

(12) Chen, Y.; Yang, Y.; Yuan, P.; Jiang, P.; Xu, Y.; Lv, S.; Ding, Y.; Dang, Z.; Gong, T.; Wang, Y.; Luo, Q. Flexible $Hf_xZr_{1-x}O_2$ Thin Films on Polyimide for Energy Storage With High Flexibility. *IEEE Electron Device Lett.* **2022**, *43* (6), 930−933.

(13) Zhou, X.; Sun, H.; Li, J.; Du, X.; Wang, H.; Luo, Z.; Wang, Z.; Lin, Y.; Shen, S.; Yin, Y.; Li, X. A Flexible $Hf_{0.5}Zr_{0.5}O_2$ Thin Film with Highly Robust Ferroelectricity. *J. Materiomics* **2024**, *10* (1), 210−217.

(14) Kim, H. J.; An, Y.; Jung, Y. C.; Mohan, J.; Yoo, J. G.; Kim, Y. I.; Hernandez-Arriaga, H.; Kim, H. S.; Kim, J.; Kim, S. J. Low-Thermal-Budget Fluorite-Structure Ferroelectrics for Future Electronic Device Applications. *Phys. Status Solidi RRL* **2021**, *15* (5), 2100028.

(15) Park, M. H.; Lee, Y. H.; Hwang, C. S. Understanding Ferroelectric Phase Formation in Doped HfO_2 Thin Films Based on Classical Nucleation Theory. *Nanoscale* **2019**, *11* (41), 19477−19487.

(16) Joh, H.; Anoop, G.; Lee, W.-J.; Das, D.; Lee, J. Y.; Kim, T. Y.; Kim, H.; Seol, W.; Yeom, J.; Jeon, S.; Hong, S.; Yoon, M.-H.; Jo, J. Y. Low-Temperature Growth of Ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ Thin Films Assisted by Deep Ultraviolet Light Irradiation. *ACS Appl. Electron. Mater.* **2021**, *3* (3), 1244−1251.

(17) Kim, T.; Jeon, S. Pulse Switching Study on the HfZrO Ferroelectric Films With High Pressure Annealing. *IEEE Trans. Electron Devices* **2018**, *65* (5), 1771−1773.
(18) Lyu, S.; Wei, W.; Yang, Y.; Zhao, S.; Wang, Y.; Long, X.; Jiang, P.; Gong, T.; Wang, Y.; Luo, Q. Achieving High-Endurance Ferroelectricity in $Hf_{0.5}Zr_{0.5}O_2$ Thin Films on Ge Substrate Through ZrO_2 Interfacial Layer Induced Low-Temperature Annealing. *IEEE Electron Device Lett.* **2024**, *45* (3), 348−351.
(19) Gaddam, V.; Das, D.; Jung, T.; Jeon, S. Ferroelectricity Enhancement in $Hf_{0.5}Zr_{0.5}O_2$ Based Tri-Layer Capacitors at Low-Temperature (350 °C) Annealing Process. *IEEE Electron Device Lett.* **2021**, *42* (6), 812−815.
(20) Gao, T.; Xiao, X.; Dong, Z.; Lu, X.; Mao, L.; Wang, J.; Liu, Y.; Hu, Q.; Xu, J. Application of Defect Engineering *via* ALD in Supercapacitors. *Batteries* **2024**, *10* (12), 438.
(21) Pal, A.; Narasimhan, V. K.; Weeks, S.; Littau, K.; Pramanik, D.; Chiang, T. Enhancing Ferroelectricity in Dopant-Free Hafnium Oxide. *Appl. Phys. Lett.* **2017**, *110* (2), No. 022903.
(22) Kang, S.; Jang, W.-S.; Morozovska, A. N.; Kwon, O.; Jin, Y.; Kim, Y.-H.; Bae, H.; Wang, C.; Yang, S.-H.; Belianinov, A.; Randolph, S.; Eliseev, E. A.; Collins, L.; Park, Y.; Jo, S.; Jung, M.-H.; Go, K.-J.; Cho, H. W.; Choi, S.-Y.; Jang, J. H.; Kim, S.; Jeong, H. Y.; Lee, J.; Ovchinnikova, O. S.; Heo, J.; Kalinin, S. V.; Kim, Y.-M.; Kim, Y. Highly Enhanced Ferroelectricity in HfO_2-Based Ferroelectric Thin Film by Light Ion Bombardment. *Science* **2022**, *376* (6594), 731−738.
(23) Pujar, P.; Cho, H.; Kim, Y.-H.; Zagni, N.; Oh, J.; Lee, E.; Gandla, S.; Nukala, P.; Kim, Y.-M.; Alam, M. A.; Kim, S. An Aqueous Route to Oxygen-Deficient Wake-Up-Free La-Doped HfO_2 Ferroelectrics for Negative Capacitance Field Effect Transistors. *ACS Nano* **2023**, *17* (19), 19076−19086.
(24) He, R.; Wu, H.; Liu, S.; Liu, H.; Zhong, Z. Ferroelectric Structural Transition in Hafnium Oxide Induced by Charged Oxygen Vacancies. *Phys. Rev. B* **2021**, *104* (18), No. L180102.
(25) Grimley, E. D.; Schenk, T.; Sang, X.; Pešić, M.; Schroeder, U.; Mikolajick, T.; LeBeau, J. M. Structural Changes Underlying Field-Cycling Phenomena in Ferroelectric HfO2 Thin Films. *Adv. Electron. Mater.* **2016**, *2* (9), No. 1600173.
(26) Pešić, M.; Fengler, F. P. G.; Larcher, L.; Padovani, A.; Schenk, T.; Grimley, E. D.; Sang, X.; LeBeau, J. M.; Slesazeck, S.; Schroeder, U.; Mikolajick, T. Physical Mechanisms behind the Field-Cycling Behavior of HfO2 -Based Ferroelectric Capacitors. *Adv. Funct Materials* **2016**, *26* (25), 4601−4612.
(27) Silva, J. P. B.; Alcala, R.; Avci, U. E.; Barrett, N.; Bégon-Lours, L.; Borg, M.; Byun, S.; Chang, S.-C.; Cheong, S.-W.; Choe, D.-H.; Coignus, J.; Deshpande, V.; Dimoulas, A.; Dubourdieu, C.; Fina, I.; Funakubo, H.; Grenouillet, L.; Gruverman, A.; Heo, J.; Hoffmann, M.; Hsain, H. A.; Huang, F.-T.; Hwang, C. S.; Íñiguez, J.; Jones, J. L.; Karpov, I. V.; Kersch, A.; Kwon, T.; Lancaster, S.; Lederer, M.; Lee, Y.; Lomenzo, P. D.; Martin, L. W.; Martin, S.; Migita, S.; Mikolajick, T.; Noheda, B.; Park, M. H.; Rabe, K. M.; Salahuddin, S.; Sánchez, F.; Seidel, K.; Shimizu, T.; Shiraishi, T.; Slesazeck, S.; Toriumi, A.; Uchida, H.; Vilquin, B.; Xu, X.; Ye, K. H.; Schroeder, U. Roadmap on Ferroelectric Hafnia- and Zirconia-Based Materials and Devices. *APL Mater.* **2023**, *11* (8), No. 089201.
(28) Lee, H.-J.; Lee, M.; Lee, K.; Jo, J.; Yang, H.; Kim, Y.; Chae, S. C.; Waghmare, U.; Lee, J. H. Scale-Free Ferroelectricity Induced by Flat Phonon Bands in HfO_2. *Science* **2020**, *369* (6509), 1343−1347.
(29) Wei, W.; Zhao, G.; Zhan, X.; Zhang, W.; Sang, P.; Wang, Q.; Tai, L.; Luo, Q.; Li, Y.; Li, C.; Chen, J. Switching Pathway-Dependent Strain-Effects on the Ferroelectric Properties and Structural Deformations in Orthorhombic HfO2. *J. Appl. Phys.* **2022**, *131* (15), 154101.
(30) Mizoguchi, T.; Saitoh, M.; Ikuhara, Y. First-Principles Calculation of Oxygen K-Electron Energy Loss near Edge Structure of HfO_2. *J. Phys.: Condens. Matter* **2009**, *21* (10), 104212.
(31) Wilk, G. D.; Muller, D. A. Correlation of Annealing Effects on Local Electronic Structure and Macroscopic Electrical Properties for HfO2 Deposited by Atomic Layer Deposition. *Appl. Phys. Lett.* **2003**, *83* (19), 3984−3986.
(32) Kim, H.; Dae, K. S.; Oh, Y.; Lee, S.; Lee, Y.; Ahn, S.; Jang, J. H.; Ahn, J. A Simple Strategy to Realize Super Stable Ferroelectric Capacitor *via* Interface Engineering. *Adv. Mater. Interfaces* **2022**, *9* (15), No. 2102528.
(33) Shao, M.; Liu, H.; He, R.; Li, X.; Wu, L.; Ma, J.; Ye, C.; Hu, X.; Zhao, R.; Zhong, Z.; Yu, Y.; Wan, C.; Yang, Y.; Nan, C.-W.; Bai, X.; Ren, T.-L.; Renshaw Wang, X. Programmable Ferroelectricity in $Hf_{0.5}Zr_{0.5}O_2$ Enabled by Oxygen Defect Engineering. *Nano Lett.* **2024**, *24* (4), 1231−1237.
(34) Choi, Y.; Shin, J.; Min, J.; Moon, S.; Chu, D.; Han, D.; Shin, C. Oxygen Reservoir Effect of Tungsten Trioxide Electrode on Endurance Performance of Ferroelectric Capacitors for FeRAM Applications. *Sci. Rep* **2024**, *14* (1), 28912.
(35) Schroeder, U.; Yurchuk, E.; Müller, J.; Martin, D.; Schenk, T.; Polakowski, P.; Adelmann, C.; Popovici, M. I.; Kalinin, S. V.; Mikolajick, T. Impact of Different Dopants on the Switching Properties of Ferroelectric Hafniumoxide. *Jpn. J. Appl. Phys.* **2014**, *53* (8S1), No. 08LE02.
(36) Park, M. H.; Lee, Y. H.; Mikolajick, T.; Schroeder, U.; Hwang, C. S. Thermodynamic and Kinetic Origins of Ferroelectricity in Fluorite Structure Oxides. *Adv. Electron. Mater.* **2019**, *5* (3), No. 1800522.
(37) Ma, L.-Y.; Liu, S. Structural Polymorphism Kinetics Promoted by Charged Oxygen Vacancies in HfO 2. *Phys. Rev. Lett.* **2023**, *130* (9), No. 096801.
(38) Jiang, P.; Luo, Q.; Xu, X.; Gong, T.; Yuan, P.; Wang, Y.; Gao, Z.; Wei, W.; Tai, L.; Lv, H. Wake-Up Effect in HfO2 -Based Ferroelectric Films. *Adv. Electron. Mater.* **2021**, *7* (1), No. 2000728.
(39) Qi, Y.; Xu, X.; Krylov, I.; Eizenberg, M. Ferroelectricity of As-Deposited HZO Fabricated by Plasma-Enhanced Atomic Layer Deposition at 300 °C by Inserting TiO2 Interlayers. *Appl. Phys. Lett.* **2021**, *118* (3), No. 032906.
(40) Tai, L.; Wei, W.; Jiang, P.; Sang, P.; Li, X.; Zhao, G.; Dou, X.; Zhan, X.; Luo, Q.; Wu, J.; Chen, J. Toward Low-Thermal-Budget Processing in Ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ Thin Films by Ozone Interface Oxidation. *IEEE Electron Device Lett.* **2023**, *44* (12), 1959−1962.
(41) Dai, S.; Chai, J.; Duan, J.; Xiang, J.; Han, K.; Wang, Y.; Xu, H.; Zhang, J.; Wang, X.; Wang, W. Investigation of $Hf_{0.5}Zr_{0.5}O_2$ Ferroelectric Films at Low Thermal Budget (300°C). *IEEE Trans. Electron Devices* **2024**, 5150−5155.
(42) Kim, S. J.; Jung, Y. C.; Mohan, J.; Kim, H. J.; Rho, S. M.; Kim, M. S.; Yoo, J. G.; Park, H. R.; Hernandez-Arriaga, H.; Kim, J.-H.; Kim, H. T.; Choi, D. H.; Jung, J.; Hwang, S. M.; Sejoon Kim, H.; Kim, H. J.; Kim, J. Low-Thermal-Budget (300 °C) Ferroelectric TiN/$Hf_{0.5}Zr_{0.5}O_2$/TiN Capacitors Realized Using High-Pressure Annealing. *Appl. Phys. Lett.* **2021**, *119* (24), 242901.
(43) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54* (16), 11169−11186.
(44) Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50* (24), 17953.
(45) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865−3868.
(46) Sheppard, D.; Xiao, P.; Chemelewski, W.; Johnson, D. D.; Henkelman, G. A Generalized Solid-State Nudged Elastic Band Method. *J. Chem. Phys.* **2012**, *136* (7), No. 074103.

https://doi.org/10.1021/acsami.5c08743
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 张程

方向： 集成电路

单位： 中国科学院上海微系统与信息技术研究所

题目： 一种SiC MOSFET的三维电热耦合模型

一种 SiC MOSFET 的三维电热耦合模型

张程[1,2]，陆纹钰[1,2]，周学通[1,2]，程新红[1,2*]，郑理[1,2]

1 集成电路材料国家重点实验室 中国科学院上海微系统与信息技术研究所
2 材料科学与光电子工程中心 中国科学院大学，北京 100049，中国
xh_cheng@mail.sim.ac.cn

摘要

SiC MOSFET 在短路事件中准确分析结温分布面临挑战。现有的确定该分布的方法一直存在分辨率低、收敛问题和精度有限等问题。本文提出了一种用于 SiC MOSFET 芯片在短路期间的快速电热耦合模型。通过考虑温度对载流子迁移率和阈值电压的影响，我们的模型显著扩展了 I_{DS} (V_{DS}, V_{GS}, T) 数据驱动紧凑模型的工作温度范围。通过 TCAD 仿真计算了与温度无关的电场分布 E(x,y,z)。利用这些结果，我们建立了一个功率分布函数 P(x,y,z,T)，作为 COMSOL 中的焦耳热源。这种方法使我们能够快速计算芯片温度分布和瞬态短路电流。验证结果表明，我们的方法将计算时间缩短至 20 分钟以内，在各种总线电压下实现 85%以上的电流预测精度，并准确识别了与实验观察到的短路栅极失效位置一致的最大热应力区域。

关键词：短路，电热耦合，宽温区，结温

I. 引言

SiC MOSFET 具有高开关速度、低导通损耗和在高温下的优异性能[1, 2]。SiC MOSFET 的高功率密度有助于器件小型化和成本降低。然而，这也显著降低了它们的短路能力[3, 4]。为了阐明短路期间的失效机制，分析结温的瞬态行为和空间分布至关重要[5, 6]。但是，由于结温可能在几微秒内急剧上升，传统的测量技术——如热电偶、红外成像或基于热敏电参数的方法——无法准确捕捉快速的温度变化[7]。因此，一个可靠的电热分析模型对于评估短路期间的结温至关重要。

传统的有限元法将泊松方程与热传导方程耦合，需要很长的计算时间并且经常面临收敛问题。传统的 Cauer RC 热模型既不能准确捕捉结温的瞬态行为，也不能捕捉其空间分布[8]。虽然 3D RC 热网络可以简化温度分布的分析[9]，但随着温度节点数量的增加，参数识别的复杂性显著增加[10]。为了快速评估 SiC MOSFET 芯片在短路期间的热行为，我们提出了一种新颖的电热耦合模型，如图 1 所示。

II. 快速电热耦合模型

图 1 展示了我们提出的快速电热耦合模型的工作流程，该模型可以有效地计算短路事件期间的芯片结温。在短路的初始阶段，当结温快速上升但仍保持近乎均匀且环路杂散电感显著时，使用传统的 Cauer RC 热模型确定初始电流和温度分布。接下来，我们将我们的宽温区电流表达式与通过 TCAD 计算的与温度无关的电场分布相结合。这种集成基于关系式 P(x,y,z,T) = j(x,y,z,T)E(x,y,z,T)产生一个功率分布函数，然后在 COMSOL 有限元仿真中用作焦耳热源。该仿真提供了空间温度分布及其瞬态演化。此外，紧凑电热模型能够同时计算瞬态电流响应。

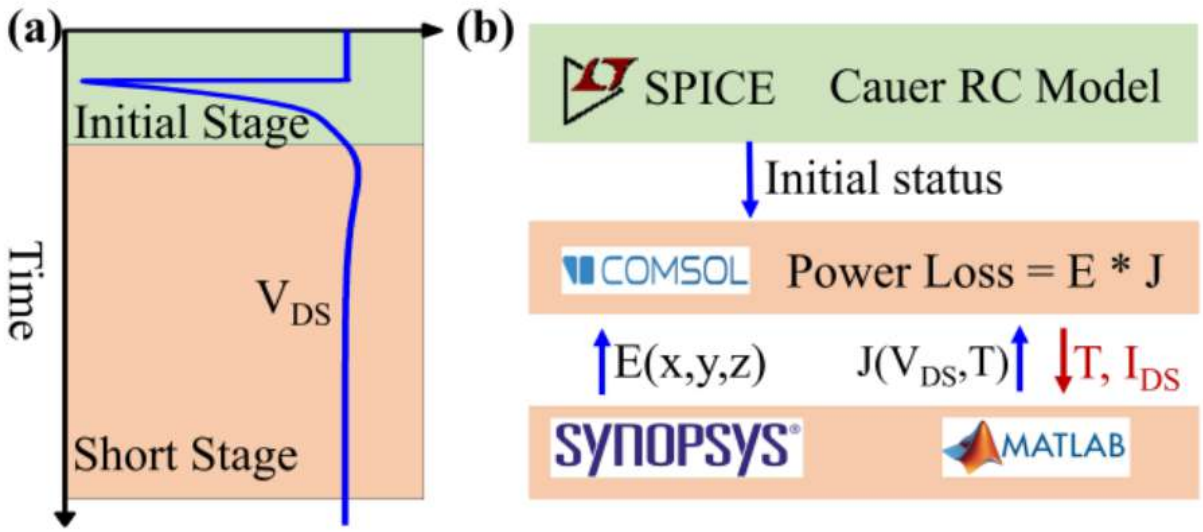


图 1：用于高效求解短路期间芯片结温的快速电热耦合工作流程

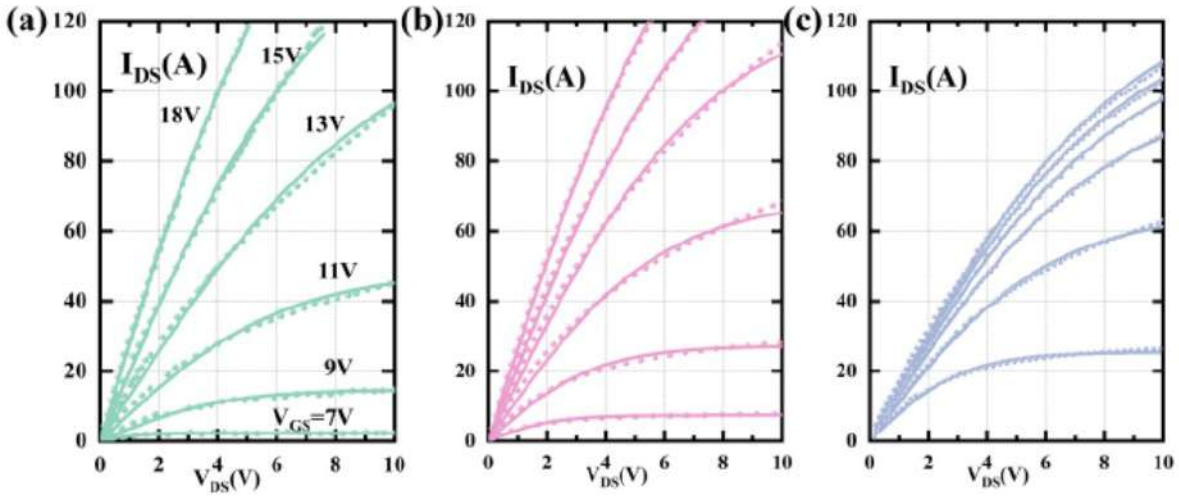


图 2：基于所提出的紧凑模型在(a)-40° C，(b)25° C，和(c)175° C 温度下的测量值和模拟的 I_{DS}-V_{DS} 曲线

在短路期间，SiC MOSFET 经历高温和高压。基于数据拟合的传统 SPICE 模型无法准确捕捉这些极端条件下的短路电流。因此，我们开发了一个基于物理的电热耦合模型，该模型包含阈值电压和迁移率的温度校正，以准确模拟短路电流行为。表 1 列出了这些表达式中使用的参数的物理意义和拟合值。

表 1：拟合的参数

参数	描述
α， β， $E_g(0)$	SiC 带隙参数：4，1300，和 3.26 eV
N_A， Φ_B	沟道浓度，费米势
Φ_{ms}， D_{it}， E_i， Q_f	功函数差，表面态，本征费米能级，固定电荷
g_m， μ， b， c， C_{ox}	跨导，迁移率及其温度参数：b = 1.5，c = 1，栅氧化层

	电容
λ	沟道长度调制参数：0.0011
V_{ds}，V_{gs}	漏源电压，栅源电压
n	三极管区参数：$-0.005V_{gs}+0.165$
m	三极管区参数：-0.102
a	三极管区参数：$-0.1717\ V_{gs}+3.5755$

图 2 比较了在-40° C、25° C 和 175° C 下的测量数据与模拟的 I_{DS}-V_{DS} 曲线。仿真与测量之间的良好一致性证实了这种电热耦合模型在宽温度范围内是可靠的。图 3 显示了短路事件期间 $I_{DS}(V_{DS},T,V_{GS}=18V)$ 的特性，这是我们快速电热耦合模型的基础。利用这些结果，可以使用该电热耦合模型获得电流密度 J(x,y,z,T(t))，并且无需借助复杂的多物理场耦合即可计算芯片功率。

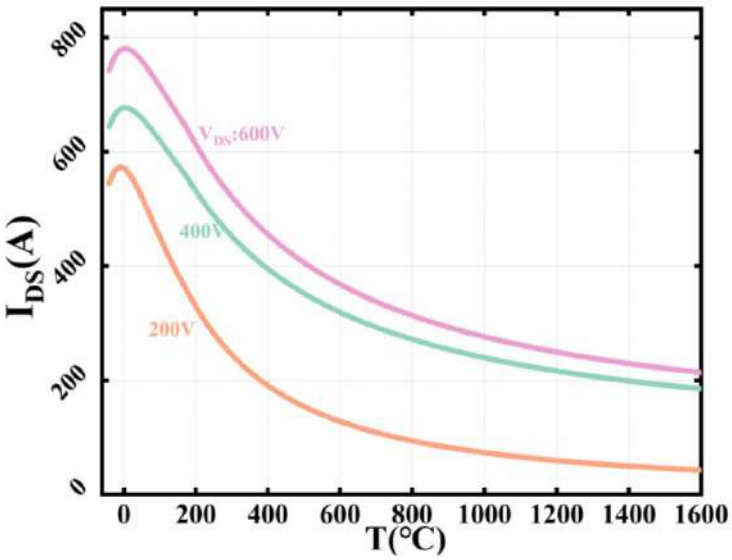


图 3：基于该电热耦合模型的 I_{DS} 在不同 V_{DS} 条件下随温度的变化

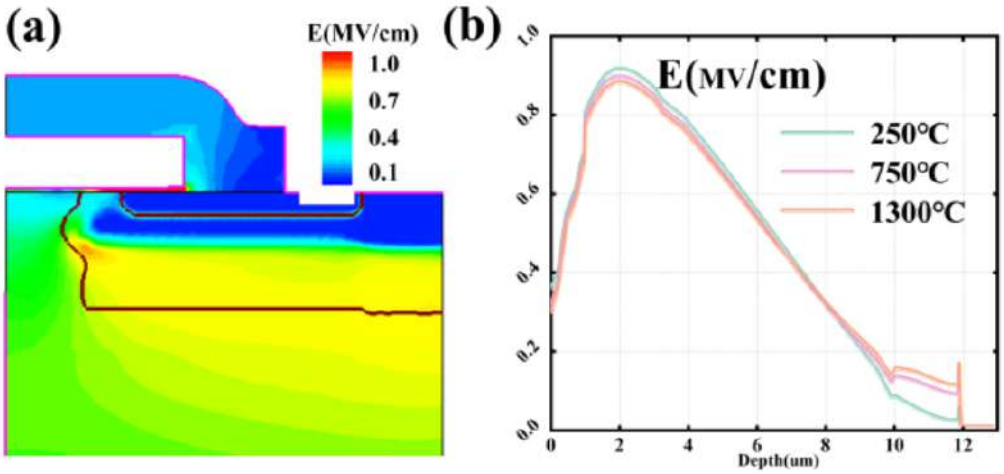


图 4：(a) 短路期间 SiC MOSFET 元胞在 V_{DS}=600V 时的电场分布；(b) 不同温度下的纵向电场分布

图 4(a)显示了在 V_{DS}=600V 时 SiC MOSFET 元胞中的电场分布，这对于计算功率分布至关重要。由于 SiC 中本征载流子浓度低，需要极高的温度（高于 2000 K）才能使热激发载流子与掺杂水平相匹配。如图 4(b)所示，仿真结果表明，在短路事件期间，电场 E(x,y,z,T)可以近似为与温度无关，即 E(x,y,z)。这个假设简化了 TCAD 中的电场计算。

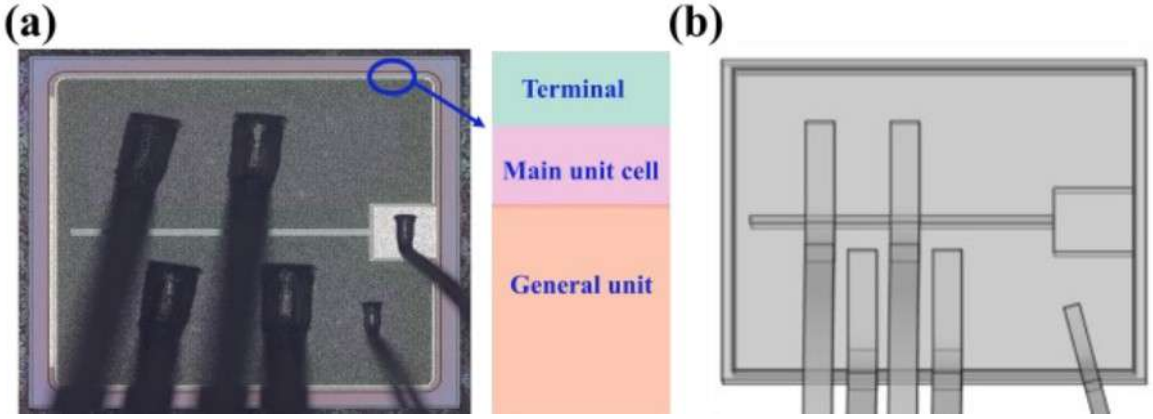


图 5：(a) SiC 裸芯片的实际尺寸图；(b) COMSOL 中建模的芯片结构图

图 5 描绘了 SiC 裸芯片的实际尺寸，而图 5 显示了 COMSOL 中使用的芯片结构图。在短路期间，尽管终端区域不产生热量，但相邻元胞的快速温升导致显著的热应力。由于靠近终端区域的元胞只有一半大小，芯片被分为三个区域：终端区域、主结元胞区域和普通元胞区域。功率分布分别针对每个区域计算，最终的温度分布在 COMSOL 中求解。

III. 快速电热耦合模型的评估效果

使用我们提出的模型，我们模拟了自主研发的 SiC MOSFET 器件（具有明确工艺参数）在不同总线电压下短路事件期间的结温和电流瞬态。然后将这些模拟值与实验测量值进行比较，以验证我们模型的鲁棒性。

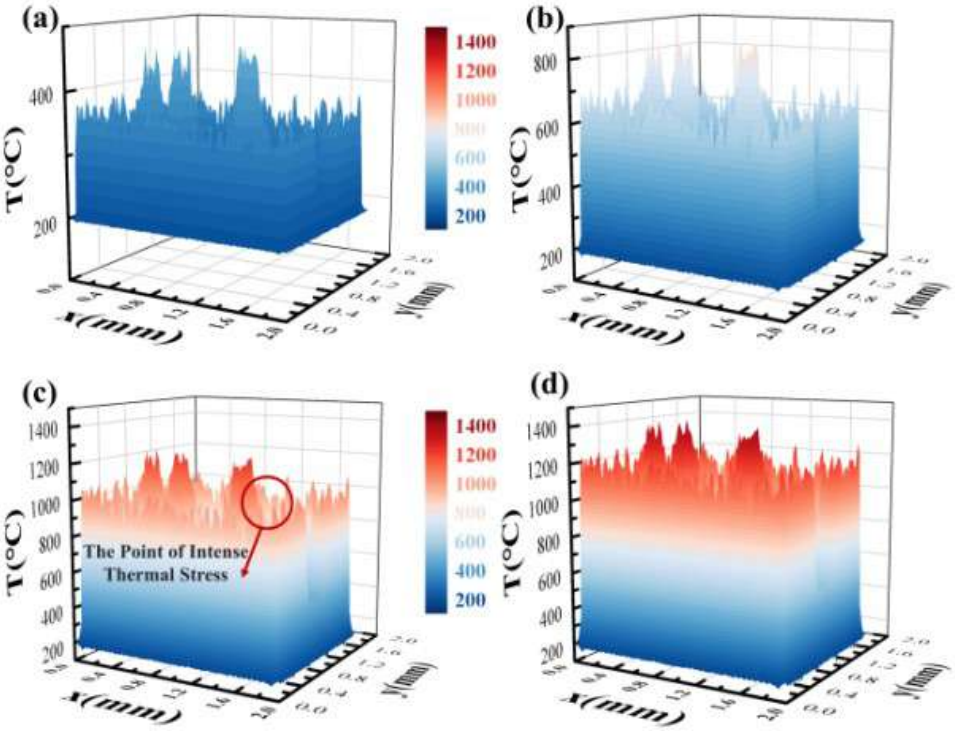


图 6. 在 V_{DS} = 600V 时的瞬态结温分布 (a)t= 0.8us. (b) 1.4us. (c)2.4us. (d)3.2us

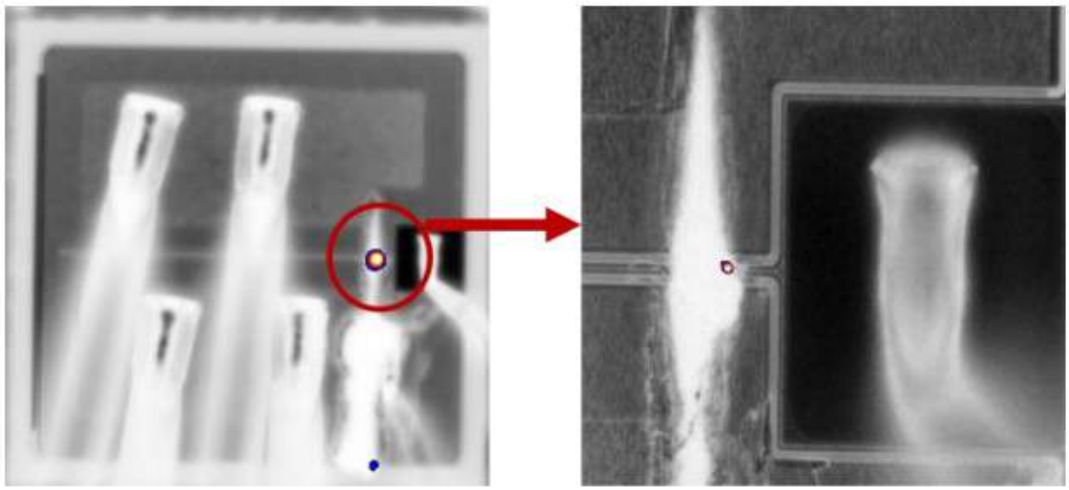


图 7. 栅极热应力失效，标明了失效点

图 6 显示了在 V_{DS}=600V 时，不同时间点（0.8 us，1.4 us，2.4 us，和 3.2 us）芯片表面的瞬态结温分布。结果表明，由于芯片中心和键合线附近的热传导和电流集中的共同影响，出现了几个热点。在某些区域，峰值温度超过平均值 200° C 以上。这一结果突显了在短路期间需要进行详细的空间分析以进行准确的温度评估，这与假设均匀温度和功率密度的传统模型不同。图 7 显示了当 V_{DS}=600V 时的栅极热应力失效，失效位置与我们模型预测的区域相对应。根据图 8，临界栅极失效温度（与最大结温相关）在 1350° C 到 1400° C 之间。在总线电压为 200V 时，即使在稳态下，芯片的最大结温也保持在此临界值以下。在 400V 时，短路耐受时间约为6.5 us，而在 600V 时约为 3.3 us——两者都会导致栅极失效。

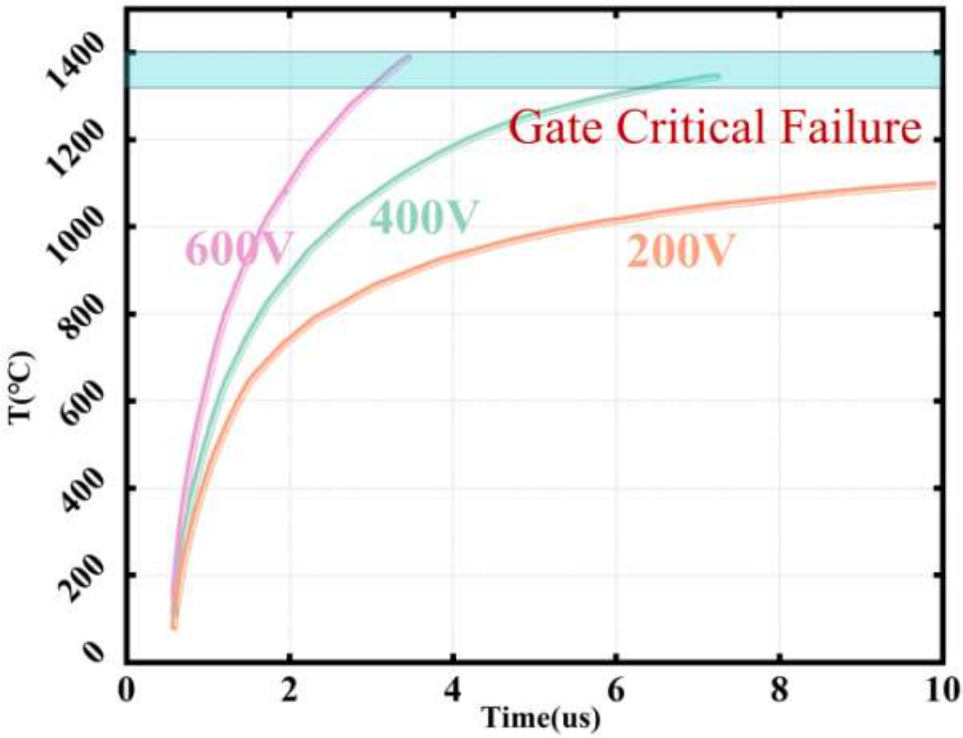


图 8. 最大结温的瞬态变化

图 9 比较了测量的短路电流波形与我们模型预测的波形。显示的电流瞬态对应于失效前的最终短路持续时间。预测电流与测量电流的紧密匹配，以及短路耐受时间的正确估计，证实了我们方法的高精度和鲁棒性。

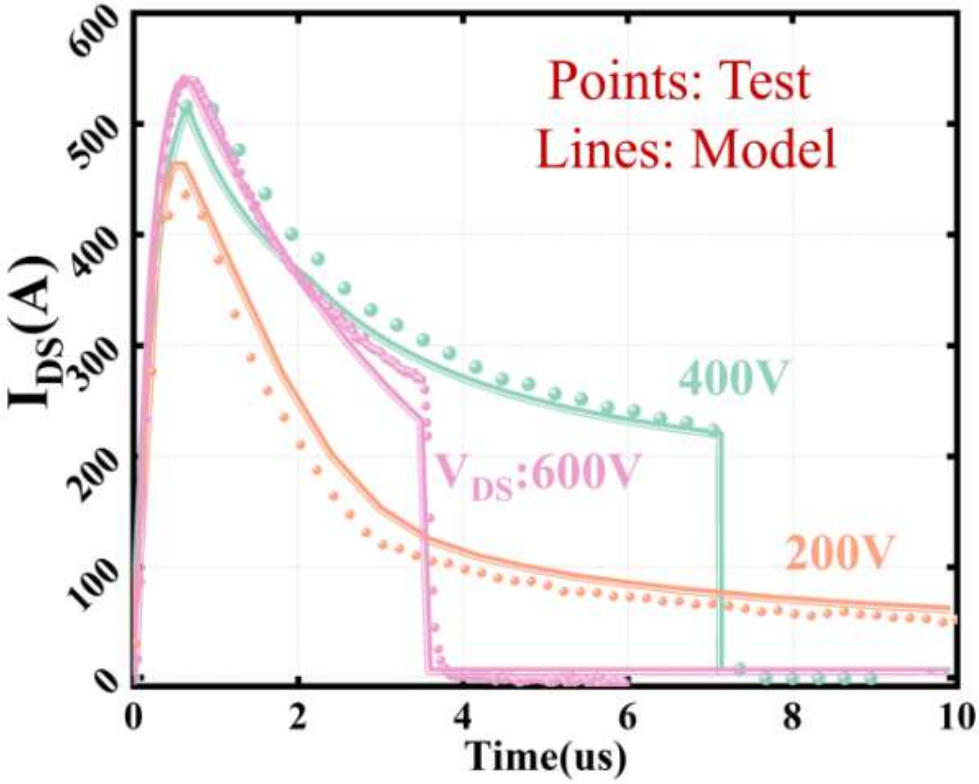


图 9 测量和预测的电流瞬态及耐受时间的比较

在建立验证之后，我们的模型捕捉局部热现象的能力使得能够对 SiC MOSFET 可靠性优化获得关键见解。非均匀的温度分布表明，基于平均温度的传统电热降额标准显著高估了器件的坚固性。特别是，时间相关的热点演化与图 6-8 所示的栅极退化机制相关，其中超过 1350° C 的热积累在多晶硅/栅氧化层界面引起界面分层。这种时空分析为实现自适应保护策略提供了物理基础。

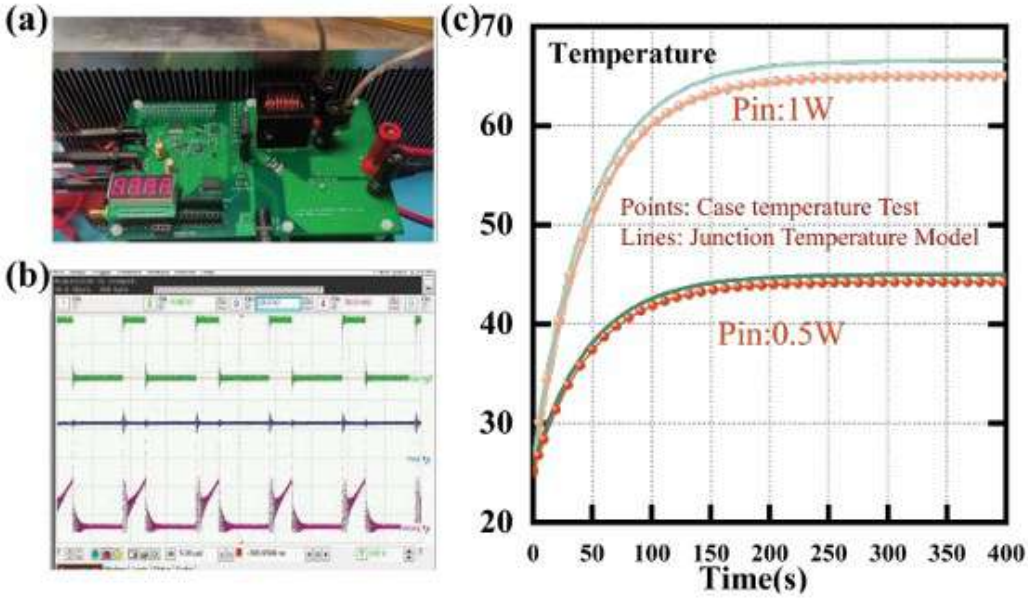


图 10：模型在降压电路中的应用 (a) 测试平台；(b) 测量的波形 V_{GS}，V_{out}，I_{DS}；(c) 不同负载下的温度预测

图 10 展示了我们的模型在降压转换器电路中的应用。我们模型的结温预测与热电偶测量结果非常吻合。具体来说，计算出的稳态芯片到外壳热阻约为 1.5 K/W，与数据手册值一致。这表明我们的模型也适用于分析常规工作条件。

表 2 总结了我们提出的方法与其他传统电热耦合方法的性能比较。我们的快速电热耦合模型在求解时间、预测精度和收敛性方面与传统方法相比显示出明显的优势。此外，与需要在 SPICE 和 FEM 软件之间进行迭代耦合脚本编写的方法相比，该方法简化了代码复杂性。

表 II：不同电热耦合模型的性能比较

方法	求解时间	精度	收敛性	代码简洁性
COMSOL	>10h	<50%	★★★	★★★★
Sentaurus	>48h	<50%	★★	★★
SPICE 和 FEM 迭代	<30min	>80%	★★★	★
本文提出的方法	<20min	>85%	★★★★	★★

VI. 结论

本文提出的快速电热耦合模型有效地模拟了 SiC MOSFET 的短路电流和瞬态结温。它与静态特性数据表现出极好的一致性，并在预测温度和电流瞬态方面表现出高精度。计算时间少于 20 分钟且具有鲁棒的收敛特性，我们的方法优于传统方法。

此外，它能准确预测各种总线电压下的短路电流和耐受时间。总体而言，该模型为分析 SiC MOSFET 芯片在短路事件期间的热行为提供了一种高效可靠的解决方案。

致谢

本工作得到了国家重点研发计划（项目编号：2022YFB3604300, 2022YFB3604301, 2022YFB3604303）、国家自然科学基金（项目编号：11705263）、上海市科学技术委员会（项目编号：23511102602）、中国科学院青年创新促进会和集成电路材料国家重点实验室自主部署项目（编号：SKLJC-Z2024-C02）的资助。

参考文献

1. Kimoto T, Fundamentals of Silicon Carbide Technology: Growth, Characterization, Devices and Applications. 2014.
2. Baliga. Advanced power MOSFET concepts. Springer US, 2010.
3. T. F. Chang, "High Accuracy SPICE Model of 3rd Quadrant Behavior on Both Planar and Trench SiC Power MOSFET, " ICSCRM 2023.
4. M. Mud Holkar, S. Ahmed, M. N. Ericson, S. S. Frank, C. L. Britton and H. A. Mantooth, "Datasheet Driven Silicon Carbide Power MOSFET Model," in IEEE Transactions on Power Electronics, vol. 29, no. 5, pp. 2220-2228, May 2014.
5. H. Wu et al., "Investigation on Short-Circuit Characteristics and Failure Modes of SiC Planar MOSFETs with Linear and Hexagonal Topologies," in IEEE Journal of Emerging and Selected Topics in Power Electronics, doi: 10.1109/JESTPE.2025.3527366.

6. N. Wang, J. Zhang and F. Deng, "Short-Circuit Dynamic Model of SiC MOSFET Considering Failure Modes," in IEEE Transactions on Power Electronics, vol. 39, no. 11, pp. 14656-14669, Nov. 2024, doi: 10.1109/TPEL.2024.3425663.
7. Sadhik M H M , Pou J and Prasanth S. Comparison of IGBT junction temperature measurement and estimation methods-a review[J].IEEE, 2017.DOI:10.1109/ACEPT.2017.8168600.
8. Weihua s , Li R and Zheng Z. Short-circuit Evaluation and Temperature-dependent Model of SiC MOSFET[J]. Proceedings of the Chinese Society of Electrical Engineering. 2018,38(7):2121-2131.DOI:10.13334/.02588013.pesec.170-307.
9. M. Ma et al., "A Three-Dimensional Boundary-Dependent Compact Thermal Network Model for IGBT Modules in New Energy Vehicles," in IEEE Transactions on Industrial Electronics, vol. 68, no. 6, pp. 5248-5258, June 2021, doi: 10.1109/TIE.2020.2991926.
10. A. S. Bahman, K. Ma, P. Ghimire, F. Iannuzzo and F. Blandberg, "A 3-D-Lumped Thermal Network Model for Long-Term Load Profiles Analysis in High-Power IGBT Modules," in IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 4, no. 3, pp. 1050-1063, Sept. 2016, doi: 10.1109/JESTPE.2016.2531631.
11. X. Lu, L. Wang, Q. Yang, F. Yang, Y. Gan and H. Zhang, "Investigation and Comparison of Temperature-Sensitive Electrical Parameters of SiC MOSFET at Extremely High Temperatures," in IEEE Transactions on Power Electronics, vol. 38, no. 8, pp. 9660-9672, Aug. 2023, doi: 10.1109/TPEL.2023.3267472.
12. J. Xin, M. Du, Z. Ouyang and K. Wei, "Online Monitoring for Threshold Voltage of SiC MOSFET Considering the Coupling Impact on BTI and Junction Temperature," in IEEE Transactions on Electron Devices, vol. 68, no. 4, pp. 1772-1777, April 2021.
13. N. Push, S. B and JMookken, "Fast and accurate electro-thermal behavioral model of a commercial SiC 1200V power MOSFET,"IEEE Pulsed Power Conference (PPC), Austin, TX, USA, 2015, pp. 1-5.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名： 赵天成

方向： 集成电路

单位： 中国科学院上海微系统与信息技术研究所

题目： 面向极高功率散热的金刚石基氧化镓异质集成材料与器件

面向极高功率散热的金刚石基氧化镓异质集成材料与器件

赵天成[1,#]，郁鑫鑫[2,#]，徐文慧[1,#,*]，何阳[3]，瞿振宇[1]，孙华锐[3]，李忠辉[2]，

周民[1]，游天桂[1]，欧欣[1,*]

（1. 中国科学院上海微系统与信息技术研究所，上海 200050；2. 南京电子器件研究所，南京 210016；3. 哈尔滨工业大学，深圳 518055）

摘要：面对氧化镓（β-Ga_2O_3）器件在大功率工况下严重的自热效应和寿命短等问题，本文创新发展转印技术，在国际上率先实现了 β-Ga_2O_3 阵列化单晶薄膜与 1 英寸金刚石衬底的异质集成，并在此基础上制备了具有高性能的射频（RF）金属氧化物半导体场效应晶体管（MOSFET）器件。制得的 β-Ga_2O_3 单晶薄膜材料的 X 射线衍射（XRD）摇摆曲线半高宽（FWHM）约为 78 arcsec，表面粗糙度为 0.36 nm。金刚石/β-Ga_2O_3 异质界面过渡层厚度小于 2 nm，测得的界面热阻（TBR）仅为 21.7 m^2·K/GW，是目前已报道的 β-Ga_2O_3 单晶薄膜与金刚石衬底异质集成结构中的最低值。受益于金刚石的超高热导率与构筑的高质量异质界面，金刚石基 β-Ga_2O_3 RF MOSFET 器件性能得到了显著提升。器件饱和漏电流高达 810 mA/mm，最大振荡频率 f_{max} 去嵌前达到 61 GHz，为目前已报道的最高值。在相同功率下，金刚石基 β-Ga_2O_3 器件结温相比 β-Ga_2O_3 同质器件降低 250℃，器件热阻（R_{th}）仅为 5.52 K·mm/W，极大提升了 β-Ga_2O_3 射频器件的热性能。以上研究结果充分证明晶圆级金刚石基 β-Ga_2O_3 异质集成材料具有优异的散热能力和器件应用前景，进一步推动高性能 β-Ga_2O_3 器件的发展，并为金刚石基异质集成材料制备提供新范式。

关键词：氧化镓；金刚石；转印；器件；热管理

1 引言

β-Ga_2O_3 是一种具有重要战略意义的电子材料，因其超宽带隙、高击穿场强、低晶圆成本和高饱和漂移速度[1-3]，在先进射频和功率器件领域备受关注。其约翰逊品质因数（用于衡量功率频率能力）约为氮化镓（GaN）材料的三倍。然而，β-Ga_2O_3 在宽/超宽禁带半导体材料中热导率（0.1-0.27 W/m·K）最低，不及硅材料的 1/5，这使得 β-Ga_2O_3 器件在大功率工况下面临严重自热效应和寿命短等可靠性问题，严重限制了其器件性能，使得散热成为其应用中的主要瓶颈之一。

金刚石是自然界已知热导率最高的材料，是大功率、射频器件的理想热沉衬底材料，将金刚石用于异质集成正在成为大功率器件热管理的重要研究方向。此前研究表明，金刚石基 GaN 高电子迁移率晶体管（HEMT）的射频输出功率密度相比碳化硅（SiC）基 GaN 器件提升了 3.6 倍[4]。因此，将 β-Ga_2O_3 异质集成到具有高导热率的金刚石衬底上，可以显著改善器件散热，从而进一步释放 β-Ga_2O_3 材料的射频和功率潜力。然而，β-Ga_2O_3 与金刚石的异质集成存在两方面的问题：1、两种材料的晶格失配程度大[5]，在金刚石衬底上直接外延生长高质量的 β-Ga_2O_3 薄膜因此受到严重阻碍；2、在活化过程中金刚石表面产生的悬挂键会迅速形成 2×1 重构，这使得晶圆的直接键合极其困难[6]。此外，金刚石本身还存在晶圆形貌和表面平整度差等生产问题，通过常规方法实现金刚石基 β-Ga_2O_3 材料异质集成面临巨大挑战。

本文创新发展转印技术，在国际上率先实现了 β-Ga_2O_3 阵列化单晶薄膜与 1 英寸金刚石的异质集成，成功克服了金刚石异质外延和键合难题。制备的 β-Ga_2O_3 阵列化单晶薄膜排列规整，晶体质量高。进一步制得的金刚石基 β-Ga_2O_3 RF MOSFET 器件表现出优异的散热能力，其器件热阻比 β-Ga_2O_3 同质 RF MOSFET 器件降低了一个数量级。

2 材料制备与表征

首先，通过离子束剥离与转移技术[7]制得 β-Ga_2O_3/SiO_2/Si（GaOOI）异质集成晶圆（图 1a）。然后采用

电感耦合等离子体刻蚀（ICP）将 β-Ga_2O_3 薄膜刻蚀成 200×200 μm 的方块阵列图形，单元间隔 200 μm（图 1b）。薄膜的阵列化一方面通过分割薄膜减少了在整片大面积薄膜转移过程中易受应力影响引起的损伤（裂纹或断裂），化整为零，降低转移难度，另一方面为后续的牺牲层释放提供通道。使用氢氟酸（HF）溶液对 GaOOI 晶圆进行湿法刻蚀，去除 SiO_2 中间层，从而实现 β-Ga_2O_3 薄膜的释放（图 1c）。之后用自制的聚乙烯醇（PVA）作为晶圆级转印的印章，将 β-Ga_2O_3 薄膜阵列从 Si 衬底上检索起来（图 1d），再印刷到 1 英寸金刚石衬底上（图 1e）。随后利用 PVA 的水溶性，将样品浸泡在超纯水中，去除 PVA 后留下 β-Ga_2O_3 薄膜于金刚石衬底上，实现两者的异质集成（图 1f）。最后在 5×10^{-4} Pa 的真空条件下以 900℃作 30 min 的退火处理，在避免金刚石氧化和污染的同时排除异质界面处的残留气体，实现 β-Ga_2O_3 与金刚石更好的结合。图 1g 展示了 1 英寸金刚石基 β-Ga_2O_3 阵列化薄膜异质集成材料的实物图，对应的光镜图像如图 1h 所示。

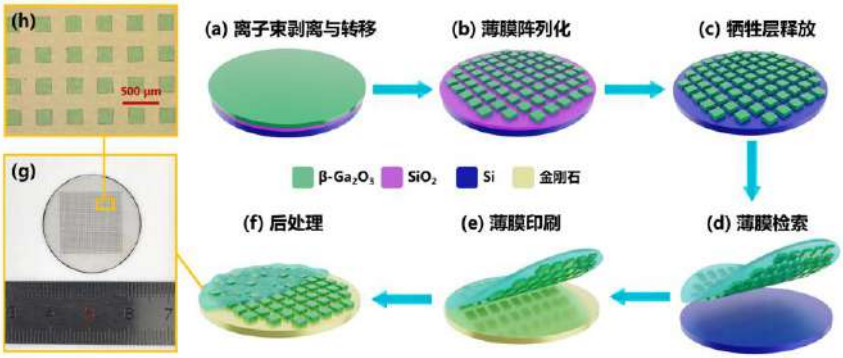


图 1　金刚石基 β-Ga_2O_3 阵列化单晶薄膜制备工艺流程

通过测试 XRD 摇摆曲线的 FWHM 对样品的 β-Ga_2O_3 薄膜晶体质量进行检验，如图 2a 所示，任取三点位置测得结果在 78 arcsec 左右，证明了较好的单晶质量。且三条曲线中均无卫星峰，表明薄膜没有应力且均一性好。采用原子力显微镜（AFM）对转印后 β-Ga_2O_3 薄膜的表面形貌作表征（图 2b），其粗糙度约 0.36 nm，满足器件制备的基本需求。图 2c-f 展示了金刚石基 β-Ga_2O_3 异质结构的透射电子显微镜（TEM）图像，β-Ga_2O_3 薄膜的晶格排列有序，选区电子衍射斑点清晰，表明在整个转印过程中晶体质量没有下降。此外，在 β-Ga_2O_3 和金刚石的异质界面处存在的非晶层厚度小于 2 nm，过渡平滑，有利于界面声子的输运。

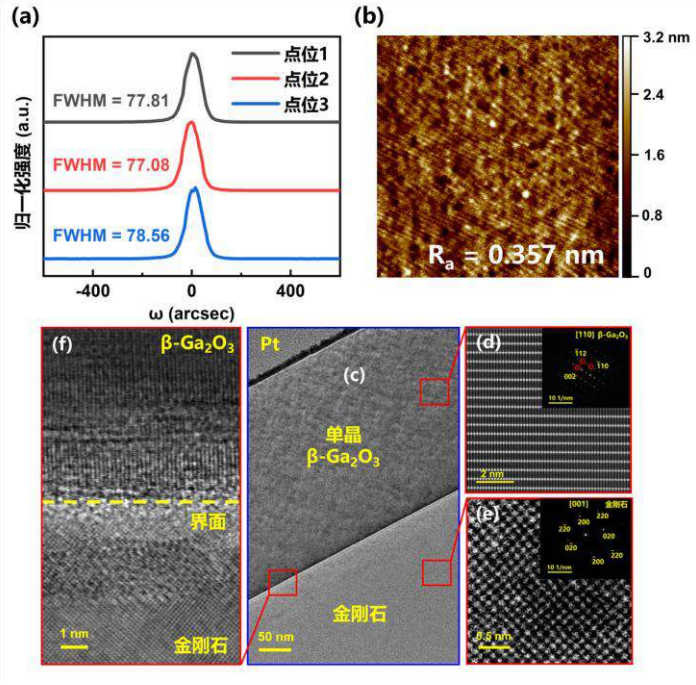


图 2　金刚石基 β-Ga_2O_3 阵列化单晶薄膜材料表征

3 界面热输运性质

对于芯片级热管理，顶部器件层中的热传递受器件结构限制，而衬底中的传热取决于衬底热导率，为固定值。因此，顶层与衬底的界面成为主要瓶颈。通过 TTR 对金刚石基 β-Ga_2O_3 异质界面进行热输运表征（图 3a），β-Ga_2O_3 体材料在 2000 ns 后温度仍有积累，而 β-Ga_2O_3 在 300 ns 内就已实现温度的快速弛豫，突出了其在高速开关和射频器件等方面的应用潜力。图 3b 中的灵敏度曲线显示出了样品热输运对 TBR 的高度敏感性，表明 TBR 显著影响了金刚石基 β-Ga_2O_3 的热性质。从 TTR 曲线提取的金刚石基 β-Ga_2O_3 的 TBR 为 21.7 m^2·K/GW，与机械剥离制备的相同结构（58.8 m^2·K/GW）相比降低了 73%[8]，且远低于分子动力学计算的预测值（83.1 m^2·K/GW）[9]。

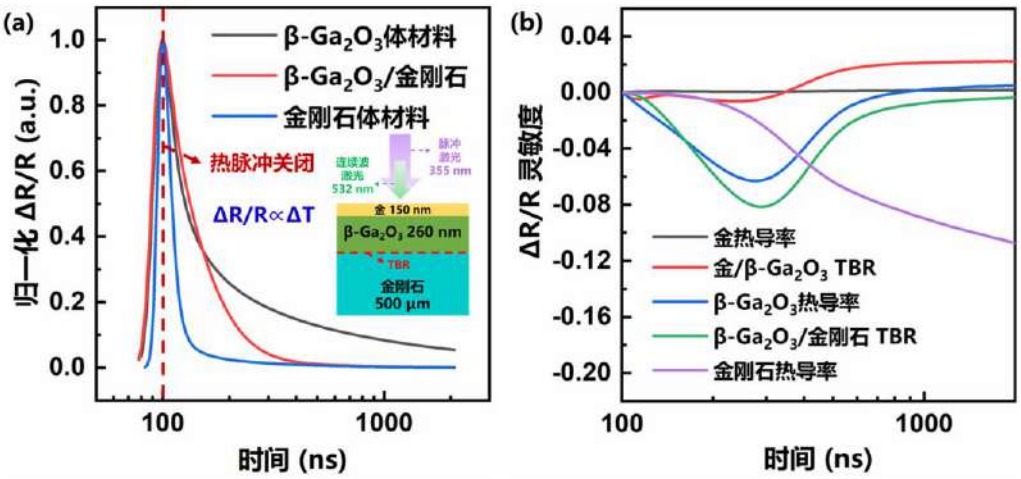


图 3　金刚石基 β-Ga_2O_3 热输运表征

4 器件制备与性能

为了验证通过转印制得的金刚石基 β-Ga_2O_3 材料的器件应用潜力，进一步制备 RF MOSFET，截面示意图如图 4a 所示。图 4b 展示了其有源区的扫描电子显微镜（SEM）图像。首先，采用能量为 15 keV、目标剂量为 3×10^{13} cm^{-2} 的 Si 离子注入，在金刚石基 β-Ga_2O_3 薄膜中形成浅层注入沟道。通过在 N_2 气氛中 950℃下进行快速热退火（RTA），实现 Si 离子的活化。再通过光刻技术定义源极和漏极，并通过电子束蒸发（EBE）沉积 Ti/Au（20/200 nm）金属堆叠。在 N_2 气氛中 470℃下作 3 min 的 RTA 后，实现了 3.2 mΩ 的低欧姆接触电阻（R_c）。然后通过原子层沉积（ALD）在 350℃下沉积了 15 nm 厚的高质量 Al_2O_3 栅介质。最后通过双层抗蚀剂电子束光刻制备 Ni/Au（20/500 nm）T 形栅电极。

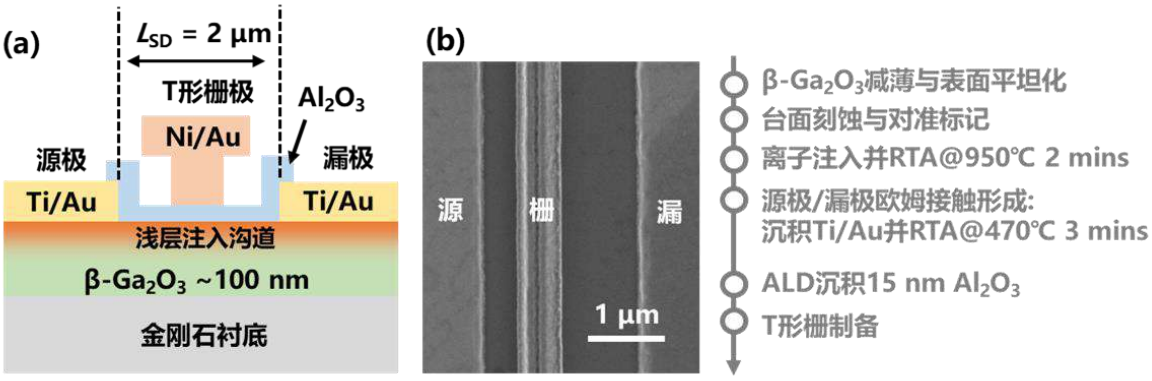


图 4　金刚石基 β-Ga_2O_3 器件制备流程

在图 5a 中，栅长为 0.1 μm 的金刚石基 β-Ga_2O_3 RF MOSFET 实现了 810 mA/mm 的高饱和漏电流（I_{Dsat}），导通电阻（R_{on}）为 12 mΩ（V_{GS}=6 V，V_{DS}=15 V）。从曲线中没有观察到明显自热效应，且 I_{Dsat} 是 β-Ga_2O_3 同质 RF MOSFET（~230 mA）的 3.5 倍[10]，这表现出金刚石衬底卓越的散热效果。图 5b 为在 V_{DS}=15 V 下测得的转移曲线，其阈值电压为-12 V，峰值跨导（g_m）为 50.6 mS/mm。在小信号射频特性方面（图 5c），实现了 34 GHz 的去嵌前外部截止频率 f_T（$|h_{21}|^2$ 下降到 0 dB），这比 β-Ga_2O_3 体器件去嵌前的结果还要大得多。$f_T \cdot L_G$ 积为 3.4 GHz·μm，对应于 2.1×10^6 cm/s 的高载流子饱和速度。同时也实现了高达 61 GHz 的去嵌前 f_{max}（MAG 下降到 0 dB）。图 5d 比较了已报道的 β-Ga_2O_3 RF MOSFET 的 f_{max}-L_G 特性，本文实现的 f_{max} 是最高值，甚至高于去嵌结果。这表明 β-Ga_2O_3 薄膜在转印后没有损坏，且在 β-Ga_2O_3 有源层和金刚石衬底之间构筑了高质量的界面。在此基础上，金刚石基 β-Ga_2O_3 器件的射频性能优于 Ga_2O_3 体器件还归因于金刚石衬底的高导热性和低寄生效应。

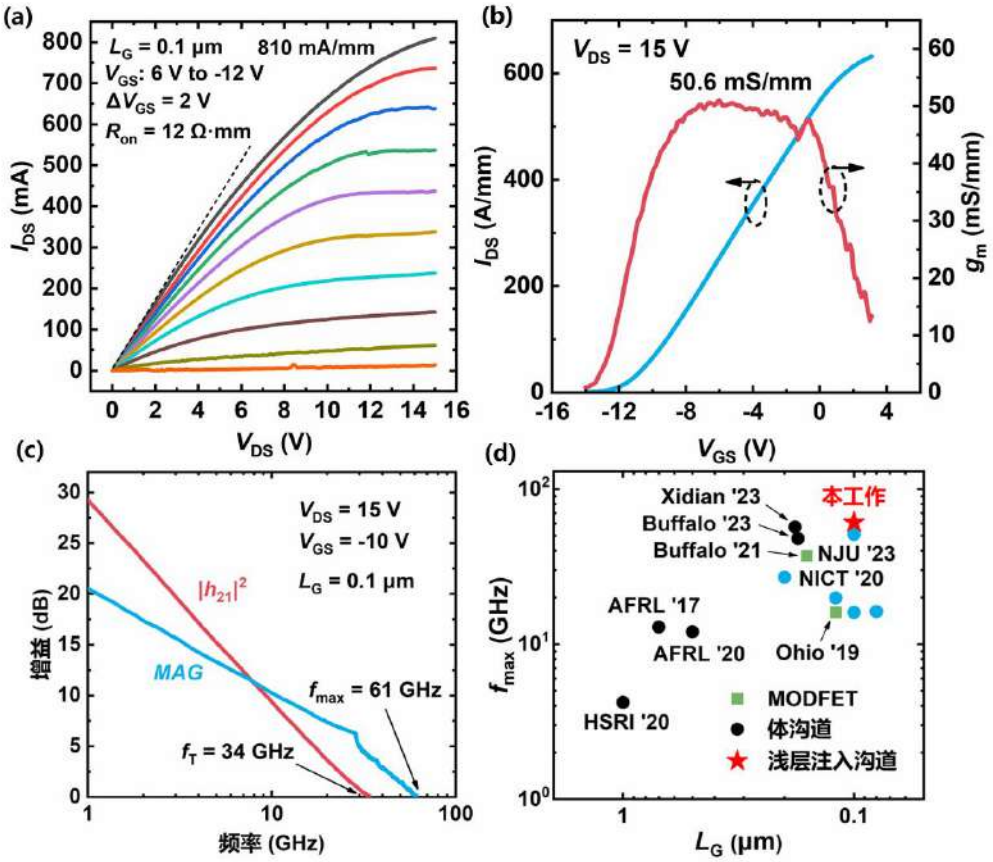


图 5　金刚石基 β-Ga_2O_3 RF 器件性能

图 6a-b 展示了在施加相同的直流功率下 β-Ga_2O_3 RF 器件的红外热成像结果。β-Ga_2O_3 同质 RF 器件结温高达 334.2℃（功率为 0.44 W），而在相近的功率（0.46 W）下金刚石基 β-Ga_2O_3 RF 器件结温大幅降低至 82.9℃。由于金刚石衬底的高效散热，金刚石基 β-Ga_2O_3 RF 器件降低了约 251℃。在图 6c 中，金刚石基 β-Ga_2O_3 RF 器件热阻约为其 β-Ga_2O_3 体器件的十一分之一，表明散热能力显著增强。图 6d 比较了本文金刚石基 β-Ga_2O_3 RF 器件与已报道的其他 β-Ga_2O_3 器件热阻。β-Ga_2O_3 和金刚石的异质集成实现了迄今为止报道的最低器件热阻值，证明了该技术在 β-Ga_2O_3 器件热管理方面的极大潜力。

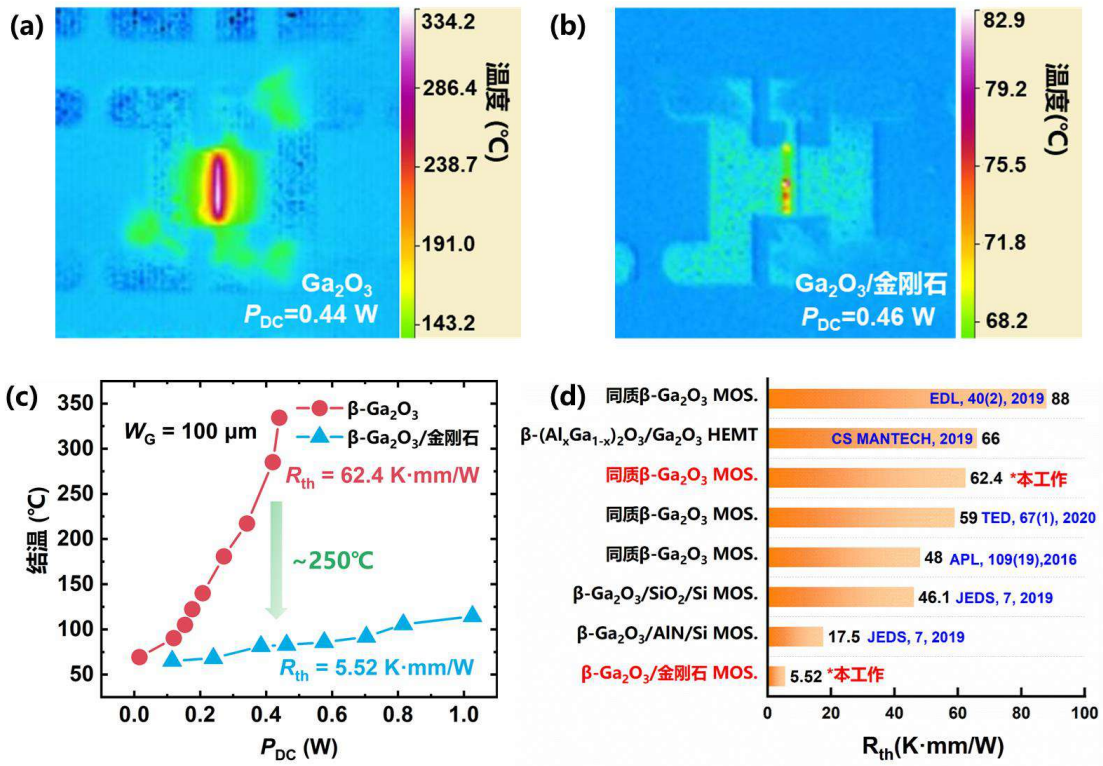


图 6　β-Ga_2O_3 器件热性能对比

5 结论

综上所述，本文通过转印技术实现了 β-Ga_2O_3 阵列化单晶薄膜到 1 英寸金刚石衬底上的异质集成。转印所得的 β-Ga_2O_3 薄膜表现出极高的晶体质量和较低的表面粗糙度。后退火处理优化了 β-Ga_2O_3/金刚石异质界面，从而改善了界面热传输，TBR 实现了目前已报道金刚石基 β-Ga_2O_3 单晶薄膜异质结构中的最低值。金刚石基 β-Ga_2O_3 RF 器件性能显著增强，且散热能力是 β-Ga_2O_3 同质 RF 器件的十倍以上。这些结果证明了金刚石基 β-Ga_2O_3 异质集成材料在器件应用方面的广阔前景。

参考文献：

[1] Hu Z, Nomoto K, Li W, et al. Enhancement-Mode Ga_2O_3 Vertical Transistors With Breakdown Voltage >1 kV[J]. IEEE Electron Device Letters, 2018, 39(6): 869-872.

[2] Higashiwaki M, Sasaki K, Kuramata A, et al. Gallium oxide (Ga_2O_3) metal-semiconductor field-effect transistors on single-crystal β-Ga_2O_3 (010) substrates[J]. Applied Physics Letters, 2012, 100(1): 013504.

[3] Moser N A, Asel T, Liddy K J, et al. Pulsed Power Performance of β-Ga_2O_3 MOSFETs at L-Band[J]. IEEE Electron Device Letters, 2020, 41(7): 989-992.

[4] Chao P C, Chu K, Creamer C, et al. Low-Temperature Bonded GaN-on-Diamond HEMTs With 11 W/mm Output Power at 10 GHz[J]. IEEE Transactions on Electron Devices, 2015, 62(11): 3658-3664.

[5] Cheng Z, Wheeler V D, Bai T, et al. Integration of polycrystalline Ga_2O_3 on diamond for thermal management[J]. Applied Physics Letters, 2020, 116(6): 062105.

[6] Conrad D, Scheerschmidt K, Gösele U. On the possibility of diamond wafer bonding in ultrahigh vacuum[J]. Applied Physics Letters, 2000, 77(1): 49-51.

[7] Xu W, Wang Y, You T, et al. First Demonstration of Waferscale Heterogeneous Integration of Ga_2O_3 MOSFETs on SiC and Si Substrates by Ion-Cutting Process[C]//2019 IEEE International Electron Devices Meeting (IEDM). San Francisco, CA, USA: IEEE, 2019: 12.5.1-12.5.4.

[8] Cheng Z, Yates L, Shi J, et al. Thermal conductance across β-Ga_2O_3-diamond van der Waals heterogeneous interfaces[J]. APL Materials, 2019, 7(3): 031118.
[9] Petkov A, Mishra A, Pomeroy J W, et al. Molecular dynamics study of thermal transport across Ga_2O_3-diamond interfaces[J]. Applied Physics Letters, 2023, 122(3): 031602.
[10] Yu X, Gong H, Zhou J, et al. High-Voltage β-Ga_2O_3 RF MOSFETs With a Shallowly-Implanted 2DEG-Like Channel[J]. IEEE Electron Device Letters, 2023, 44(7): 1060-1063.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 周晓莉

方向： 集成电路

单位： 中国科学院微电子研究所

题目： 高功率密度多结级联VCSEL设计与制备

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

High-Power Low-Divergence Single-Mode Multi-Junction Cascade VCSELs

Xiaoli Zhou, Meng Xun, Guanzhong Pan, Yun Sun, Yinuo Wang, Chao Han, Weichao Wu, Song Wu, Yongming Fei, and Dexin Wu

***Abstract*— Seven-junction cascade vertical-cavity surface-emitting lasers (VCSELs) with a long monolithic cavity have been demonstrated to improve single-mode output power up to 10.8 mW under continuous-wave (CW), room temperature operation. The several active regions in series connected via tunnel junctions are introduced to greatly increase the slope efficiency (SE) up to 4.4 W/A. To avoid significantly increasing the round-trip optical loss, the additionally extended cavity is sandwiched between the active region and n-type distributed Bragg reflectors (DBRs). The cavity spacer is capable of enhancing higher order modes diffraction losses and single-mode optical power. The increased large diffraction loss of higher order modes and the small effective refractive-index difference between the core and cladding regions result in a low-divergence angle of only 16°. Due to the simple fabrication process and the low cost, it is promising for the designed VCSELs to offer alternate solutions for applications requiring high single-mode performances of VCSELs.**

***Index Terms*— Long monolithic cavity, low-divergence, multijunction cascade, vertical-cavity surface-emitting laser (VCSEL).**

I. Introduction

VERTICAL-CAVITY surface-emitting lasers (VCSELs) are characterized by the circular beam, single longitudinal mode operation, low far-field angle, low threshold current, and scalability [1], which have been widely used in many applications, such as laser radar, data communication, and mobile consumption [2]. Single-mode VCSELs are used as indispensable optical sources in some special applications, such as gas detection, laser printing, optical coherence tomography, atomic clocks, and atomic magnetometer [3].

VCSEL is able to work in single longitudinal mode as well as avoid mode hopping due to inherent short cavity length. However, higher order transverse modes usually appear along with the fundamental mode because of the larger transverse dimension of the VCSEL compared with its effective cavity length. The most common method for achieving single-mode operation is to reduce oxide-aperture diameter, but the small oxide-aperture VCSELs exhibit low single-mode output power and high series resistance. There are several methods of increasing the single-mode power of VCSELs, and the key idea is to increase the losses of higher order modes in the cavity; thus, it is feasible to suppress higher order modes emission under large-aperture diameter. One method is to form refractive-index differences by selectively etching the top contact to introduce the shallow surface relief, which may lead to a decrease in the higher order modes. The single-mode optical output power of 850-nm VCSEL reached 6.5 mW at a 6-μm oxide aperture and a 3-μm surface relief diameter [4]. Another method of achieving single-mode VCSELs is to construct a photonic crystal structure. A refractive-index waveguide is fabricated by etching the top distributed Bragg reflectors (DBRs) to form the periodic air holes, resulting in a change in cavity resonance mode, and realizing a single-mode output power of 3.1 mW [5]. Moreover, it is promising to integrate VCSELs with an anti-waveguide structure under large output apertures. Because the refractive index of cladding layers is higher than the core region, the losses of lateral modes, especially the higher order modes, are comparatively large, ultimately achieving a single-mode continuous-wave (CW) power of 7.1 mW at an 8-μm diameter [6]. Although these methods can effectively suppress the higher order modes, complex secondary epitaxial growth and precise fabrication control are both highly required.

Multijunction cascade techniques have been proposed to improve the output characteristics of VCSELs through a combination of several active regions connected via tunnel junctions, in which quantum efficiency is typically >1, while

Manuscript received 1 March 2024; revised 27 June 2024; accepted 1 July 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62374178, Grant 62174178, Grant 62374179, and Grant 62104251; in part by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences (CAS), under Grant ZDBS-LY-JSC031; in part by the Youth Innovation Promotion Association, CAS, under Grant 2022115; in part by Beijing Nova Program under Grant 2022032; and in part by the National Key Basic Research Program under Grant 2022YFB2803000. The review of this article was arranged by Editor J. D. Phillips. *(Corresponding authors: Meng Xun; Guanzhong Pan.)*

Xiaoli Zhou, Meng Xun, Guanzhong Pan, Yun Sun, and Yinuo Wang are with the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China, and also with the School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhouxiaoli@ime.ac.cn; xunmeng@ime.ac.cn; panguanzhong@ime.ac.cn; sunyun@ime.ac.cn; wangyinuo23@mails.ucas.ac.cn).

Chao Han, Weichao Wu, Song Wu, Yongming Fei, and Dexin Wu are with the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China (e-mail: hanchao@ime.ac.cn; wuweichao@ime.ac.cn; wusong@ime.ac.cn; feiyongming@ime.ac.cn; dxwu@ime.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2024.3423801.

Digital Object Identifier 10.1109/TED.2024.3423801

0018-9383 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on July 17,2024 at 01:26:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

the slope efficiency (SE) and power conversion efficiency (PCE) are both drastically increased [7], [8], [9]. However, it is easy for multijunction cascade VCSELs to introduce higher order modes emission. For example, due to the existence of several active regions, it is necessary to incorporate oxide layers to confine injection currents of each active region, and thus, a relatively high SE can be attained. Typically, employing multioxide layers can enhance SE and output power in multijunction cascade VCSELs due to good current confinement. However, multioxide layers are capable of resulting in strong index guiding, which causes more higher order modes emitting [10]. In addition, the low thermal conductivity of the oxide layer also brings an increase in the thermal resistance of the devices. Employing a single-oxide layer can introduce fewer changes to the optical waveguide structure. Transverse spreading of the injection current will get severe in the active regions away from the oxide layer [7]. Therefore, the key point in this article is to increase single-mode optical power by suppressing the higher order modes emission in multijunction cascade VCSELs.

In this letter, a seven-junction cascade VCSEL was designed and fabricated. Shortened spacing among active regions was adopted to reduce the transverse spreading of the injection current. Meanwhile, an extended cavity was introduced to enlarge the effective resonance length to increase the diffraction losses of the higher order modes [10], [11], as the higher order modes suffer more from diffraction than the lowest order mode. Besides, the cavity spacer of homostructure with high thermal conductivity can be as a heat-spreading layer to reduce heat accumulation in the active regions. Eventually, the fabricated VCSEL shows a high single-mode output power of 10.8 mW near 905-nm wavelength. Furthermore, there are only changes in the epitaxial structure of the devices, and the fabrication process is consistent with the traditional VCSELs.

II. Device Structure and Fabrication

The schematic of the device is shown in Fig. 1(a), which was grown by metal–organic chemical vapor deposition (MOCVD) on an n-type GaAs substrate. The top and bottom Bragg reflector stacks consist of p-type $Al_{0.12}Ga_{0.88}As/Al_{0.9}Ga_{0.1}As$ and n-type $Al_{0.12}Ga_{0.88}As/Al_{0.9}Ga_{0.1}As$ Bragg reflectors with 20.5 and 40.5 pairs of quarter-wavelength-thick layers, respectively. Each active region comprises three pairs of 6-nm-thick $In_{0.12}Ga_{0.88}As$ quantum wells separated by 8-nm-thick $Al_{0.3}Ga_{0.7}As$ barriers. Seven active regions are connected by tunnel junctions. Although the multijunction VCSELs exhibit high operation voltage, they can provide a trade-off between voltage and current for driver compatibility. The driver with high voltage and small current can achieve faster rise time and lower cost, which greatly facilitates practical applications. The material of the tunnel junction is GaAs because of its relatively small bandgap energy to increase tunneling probability. The n-type and p-type of tunnel junction are highly doped as 2×10^{19} and 1×10^{20} cm^{-3} for achieving degeneracy. Because the doping concentration of the p-side of the tunnel junction is much higher than that of the n-side, the depletion layer is mainly located on the n-side. Considering the absorption loss, the thickness of the tunnel junction is chosen as 8 nm (p-side) and 12 nm (n-side). To avoid high losses of light absorption, the tunnel junctions are placed in the nodes of the standing wave pattern. Moreover, for better coupling between photons and electrons in the active regions, we place the quantum wells in the antinodes of the standing wave pattern to maximize the optical gain [8].

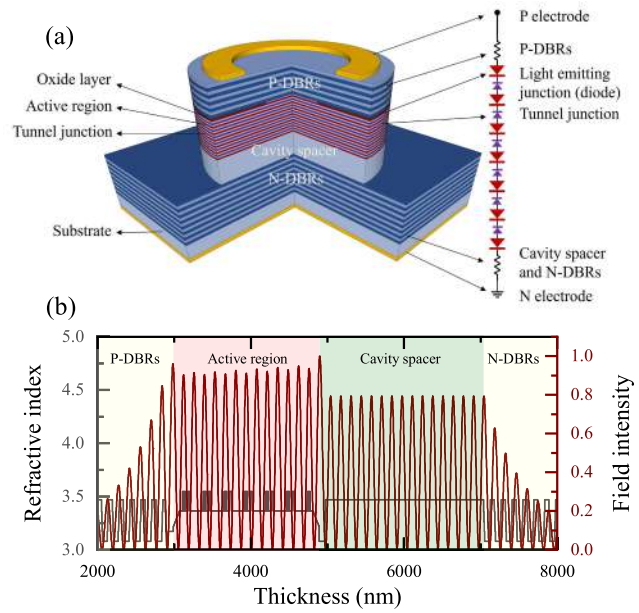


Fig. 1. (a) Schematics and (b) simulated standing wave patterns of the core region of the seven-junction cascade VCSELs.

To increase single-mode output power, a long monolithic cavity was employed to increase the diffraction losses of the higher order modes. It should be noted that the additional extended layer must be made, so that it does not significantly increase the round-trip optical loss. Therefore, the location of the extended layer should be chosen carefully. As the electron mobility is largely higher than the hole mobility, it allows the cavity spacer to be lightly doped in the n-type region to avoid excess light absorption. In the meantime, too much series resistance of the device is not introduced. Consequently, the cavity spacer is incorporated between the active regions and n-type DBRs. To avoid multilongitudinal modes emitting, the length of the cavity should not be too long. The cavity spacer is $Al_{0.12}Ga_{0.88}As$ material of 2-μm thickness and lightly doped by 5×10^{17} cm^{-3}. Current confinement, as well as optical field confinement, is achieved by the wet oxidation of the $Al_{0.98}Ga_{0.02}As$ layer integrated into the node of the first p-type DBR to form insulating AlO_x. In general, the oxide thickness of conventional VCSEL devices is generally 20–30 nm, while the thickness of the oxide layer of our fabricated VCSEL is approximately 15 nm. The thinner oxide layer can reduce the effective refractive-index difference between core and cladding regions, which further reduces the number of higher order modes [16]. Fig. 1(b) shows the simulated standing wave profile of the core region along the growth direction.

The fabrication process is as follows. First, a p-type Ti/Pt/Au ohmic contact electrode was made at the periphery of the light aperture. Next, the high-aluminum layer was exposed by inductively coupled plasma (ICP) to etch to n-type DBRs and subsequently oxidized at 400 °C in a nitrogen/water vapor

Authorized licensed use limited to: ShanghaiTech University. Downloaded on July 17,2024 at 01:26:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

atmosphere. Then, a SiN passivation layer was deposited on the surface of the mesa to make electrical isolation, on which a Ti/Au seed layer was sputtered subsequently. Also, a thick Au pad was electroplated on the periphery of the output optical aperture afterward. The Ti/Au seed layer on the optical aperture was removed by wet etching. Following these steps, the substrate was thinned to be around 150 μm followed by evaporation of AuGeNi/Au ohmic contact electrode on its surface. Finally, rapid thermal annealing was used to form good ohmic contact.

III. Results and Discussion

The performance of the fabricated seven-junction cascade VCSEL with a 3.5-μm oxide aperture was tested under CW conditions. The oxide-aperture diameter of the VCSEL was measured using a high-resolution infrared CCD camera. Due to the different resonance wavelengths of the two regions, the CCD camera can clearly distinguish between the oxidized and nonoxidized regions. Fig. 2 illustrates the light–current–voltage (L–I–V) characteristics in the case-temperature range from 25 °C to 85 °C. It shows that the threshold current is around 1.6 mA at room temperature (25 °C). With the increased ambient temperatures, the threshold currents increased to 1.7 and 1.8 mA at 55 °C and 85 °C, respectively. It is because higher carrier densities are needed to make up for the increased losses owing to the decreased peak material gain. Typically, the optical gain is supposed to increase considerably due to the active regions in series, and thus, the threshold current of multijunction cascade VCSEL is lower than the traditional VCSEL, while the result is not the case. The reason is that the effect of the single-oxide layer may make current transversely spread in the active regions, and consequently, the losses of carriers used for laser emission will increase. Due to partial losses of carriers, the maximum SE is 4.1 W/A at 25 °C. The turn-on voltage is around 9.1 V, and this equates to a voltage drop of 1.3 V per p-n junction, which confirms the excellent property of tunnel junction and no additional voltage drop added. Besides, the maximum output power decreases to 4.73 mW at 85 °C, which is mainly due to the increased internal losses with the temperature. The PCE characteristic is shown in Fig. 2(b). The maximum PCE drops from 21% at 25 °C to 12% at 85 °C.

The spectra of seven-junction cascade VCSELs under different driving currents and ambient temperatures are displayed in Fig. 3. The spectra were measured with a maximum resolution of nominally 0.01 nm. At 25 °C, when the driving currents change from 2 to 3.5 mA, there exists only one peak, indicating a pure single-mode characteristic. As the current continues to increase, side peaks appear in the spectra. Single-mode emission with a side mode suppression ratio (SMSR) of around 40 dB up to a driving current of 4 mA is observed. The maximum single-mode output power is 10.8 mW at the driving current of 4.5 mA, and the SMSR is around 35 dB. It indicates that the seven-junction cascade VCSEL is capable of realizing single-mode emission within the range from the threshold current to 4.5 mA. As the ambient temperature increases to 55 °C and 85 °C, the VCSELs still show good single-mode characteristics till rollover currents. The values of the maximum single-mode power are 7.8 and 4.7 mW at 55 °C and

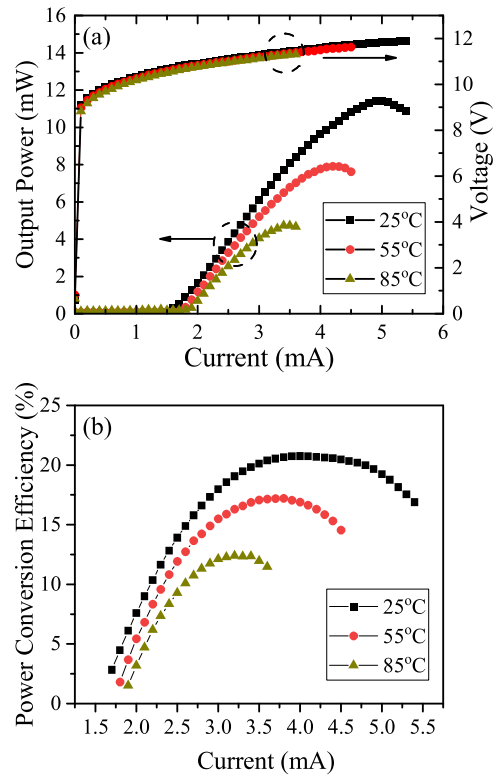


Fig. 2. Measured (a) L–I–V and (b) PCE characteristics curves for the seven-junction cascade VCSELs in the case-temperature range from 25 °C to 85 °C.

85 °C, respectively. The excellent single-mode performance is mainly attributed to the extended cavity structure. The reason is that the diffraction angle of higher order modes is larger than that of the lowest order modes [12]. Thus, the losses of higher order modes may be larger than that of the fundamental mode in the long cavity, which greatly suppresses the higher order modes lasing and effectively improves the defect of being prone to multimode emission in multijunction VCSELs.

The schematic of an experimental setup for far-field measurement is shown in Fig. 4. The VCSEL was placed on a thermoelectric cooler (TEC)-controlled thermochuck. A high-precision current source (Keithley 2601b) was connected to the fabricated VCSEL via a probe. The far-field patterns were measured by a laser beam profiler (Spiricon SP928), in front of which a closed circuit television (CCTV) lens was mounted to reduce the image on the diffusing glass. Fig. 5 presents the far-field profiles and the corresponding divergence angles of the seven-junction cascade VCSEL under different driving currents. The divergence angle is defined as the angle at which the intensity drops to $1/e^2$ of the maximum. Fig. 5(a) shows that the beam sizes become larger with the increase of driving currents. The Gaussian distribution was maintained all the time, showing single-mode operation, which can also be confirmed from the spectra in Fig. 3. It is found that the divergence angle is approximately proportional to the driving current (2.8°/mA), as shown in Fig. 5(b). The increase in divergence angle is because of the thermal lens effect. When the operating current increases, the thermal lens effect can induce a large step of the refractive index and lead to a

Authorized licensed use limited to: ShanghaiTech University. Downloaded on July 17,2024 at 01:26:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

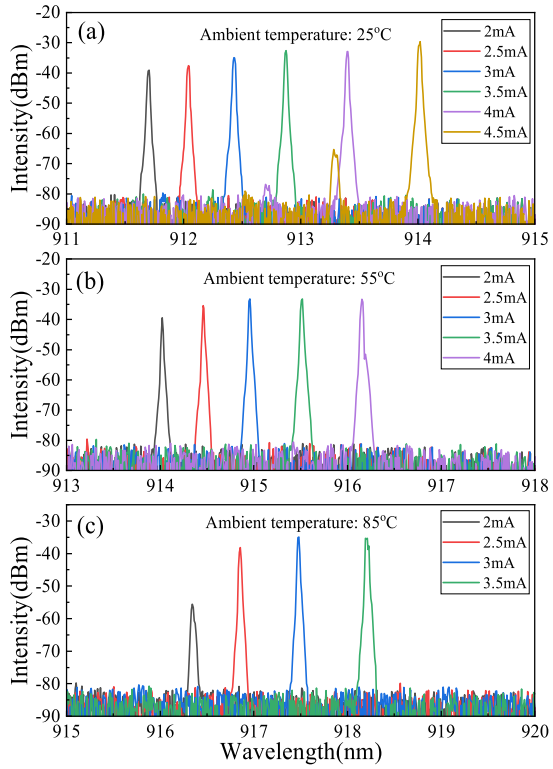


Fig. 3. Spectral characteristics of seven-junction cascade VCSELs with an oxide-aperture diameter of 3.5 μm under different ambient temperatures of (a) 25 °C, (b) 55 °C, and (c) 85 °C.

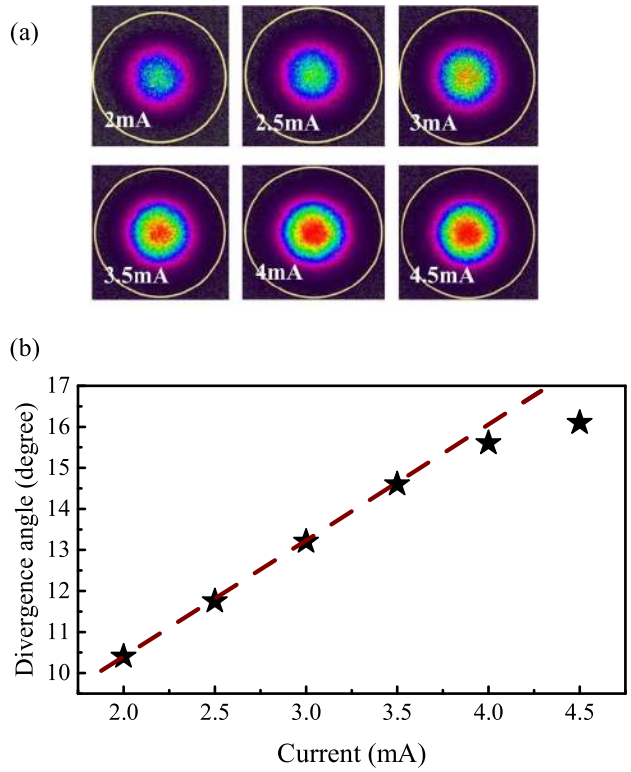


Fig. 5. (a) Far-field profiles and (b) divergence angles of seven-junction cascade VCSELs with an oxide-aperture diameter of 3.5 μm.

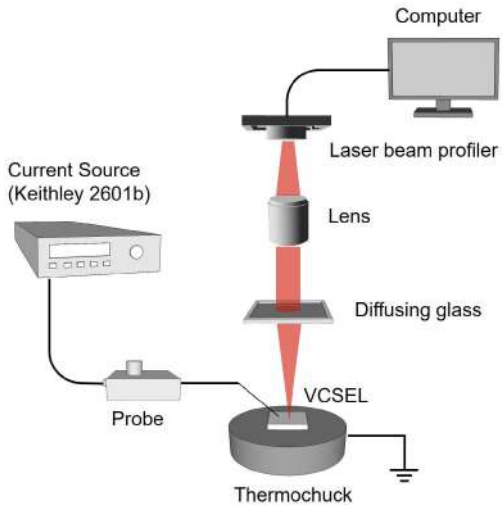


Fig. 4. Schematic of an experimental setup for far-field measurement.

narrowing of beamwidth of the fundamental mode, which results in the far-field divergence angle rapidly increasing with the increase in current [13].

The $1/e^2$ beam divergence angle and the full-width at half-maximum (FWHM) angle increase from 10.4° to 16.1° and from 3.9° to 8.1° over the current range from 2 to 4.5 mA, respectively. The divergence angle of the VCSEL is significantly reduced compared with conventional multijunction VCSELs [14]. According to the equation in [15], the effective refractive-index difference was calculated to be only 0.002, which is smaller than the traditional VCSEL of 0.006–0.01 [16]. The small effective refractive-index difference can also decrease the mode number and make the spectrum pure, which is also beneficial to divergence. Overall, both the large diffraction loss of higher order modes and the small effective refractive-index difference result in low divergence in our designed multijunction VCSELs.

To further investigate the thermal properties of seven-junction cascade VCSELs, the thermal resistance is calculated by combining the emission spectrum with dissipation power using the common experimental method. The specific formula is demonstrated in [17]

$$R_{\text{th}} = \frac{\Delta T}{\Delta P_{\text{diss}}} = \frac{\Delta\lambda/\Delta P_{\text{diss}}}{\Delta\lambda/\Delta T} \tag{1}$$

where $(\Delta\lambda/\Delta P_{\text{diss}})$ is the cavity resonance wavelength versus dissipated power. $\Delta\lambda/\Delta T$ is the spectrum near the threshold current versus the ambient temperature. Fig. 6(a) illustrates that the peak wavelength increases linearly with the ambient temperature, and $\Delta\lambda/\Delta T$ is 0.08 nm/°C, which is larger than the typical temperature drift rate of VCSELs (0.07 nm/°C). The reason is that the refractive index of the extended cavity spacer of $Al_{0.12}Ga_{0.88}As$ with low Al content increases faster with increasing temperature. The red shift of wavelength is mainly because the refractive index of the material changes along with the temperature. Meanwhile, the peak wavelength is proportional to the dissipated power, and $\Delta\lambda/\Delta P_{\text{diss}}$ is 0.11 nm/mW, as shown in Fig. 6(b). The thermal resistance R_{th} of fabricated VCSEL is approximately 1.38 °C/mW, which is slightly lower than the thermal resistance of conventional VCSELs [18], [19], [20].

Authorized licensed use limited to: ShanghaiTech University. Downloaded on July 17,2024 at 01:26:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

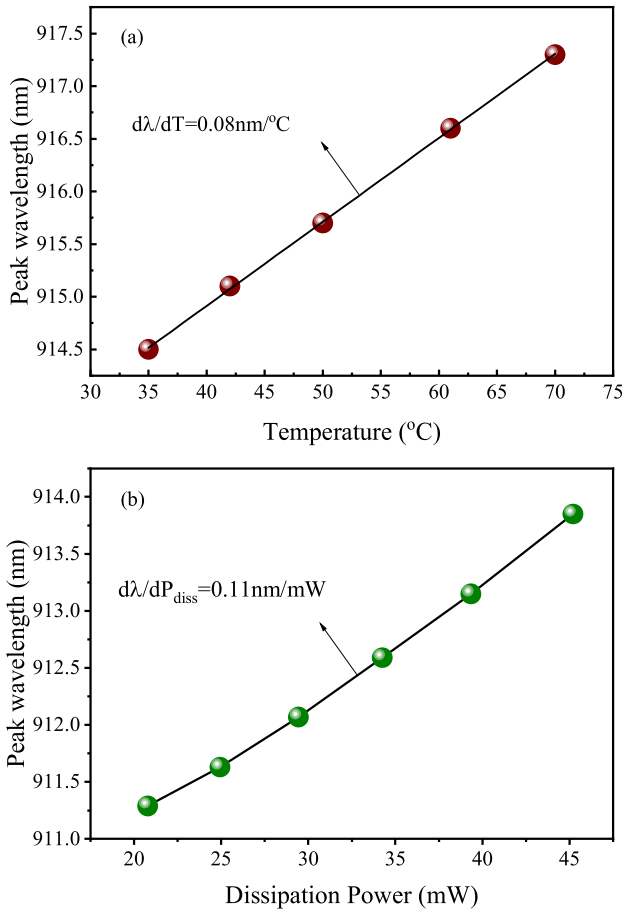


Fig. 6. Measured (a) $\Delta\lambda/\Delta T$ and (b) $\Delta\lambda/\Delta P_{\text{diss}}$ for seven-junction cascade VCSELs with an oxide-aperture diameter of 3.5 μm.

The reason for low thermal resistance may be as follows. Although the oxide layer can provide good confinement for both current and optical fields in conventional VCSELs, the Al_xO_y has a thermal conductivity of 0.7 W/mK [20], which is much lower than the thermal conductivity of GaAs and AlAs of 50 W/mK. Because of the single-oxide layer in our designed seven-junction VCSELs, the active regions away from the oxide layer have better thermal diffusion ability than the active region adjacent to the oxide layer, which reduces the influence of the low thermal conductivity of the oxide layer. In addition, the low thermal resistance is also attributed to additional cavity spacer of homostructure with high thermal conductivity, which facilitates thermal diffusion from active regions to the substrate.

IV. Conclusion

In summary, we have successfully designed and fabricated the single-mode seven-junction cascade VCSEL with one oxide layer. With the extended cavity and shortened active region spacer, it is practical to suppress the higher order modes emission of multijunction cascade VCSELs, resulting in the maximum single-mode output power of 10.8 mW. Although the VCSEL has low SE, it exhibits high single-mode output power and low thermal resistance. In the future, this optimized multijunction VCSEL will be of great significance in the field that requires high single-mode output power.

References

[1] J.-H. Ser, Y.-G. Ju, J.-H. Shin, and Y. H. Lee, "Polarization stabilization of vertical-cavity top-surface-emitting lasers by inscription of fine metal-interlaced gratings," *Appl. Phys. Lett.*, vol. 66, no. 21, pp. 2769–2771, May 1995, doi: 10.1063/1.113469.

[2] H. J. Unold, S. W. Z. Mahmoud, R. Jager, M. Grabherr, R. Michalzik, and K. J. Ebeling, "Large-area single-mode VCSELs and the self-aligned surface relief," *IEEE J. Sel. Topics Quantum Electron.*, vol. 7, no. 2, pp. 386–392, Mar. 2001, doi: 10.1109/2944.954155.

[3] A. Al-Samaneh, M. Bou Sanayeh, S. Renz, D. Wahl, and R. Michalzik, "Polarization control and dynamic properties of VCSELs for MEMS atomic clock applications," *IEEE Photon. Technol. Lett.*, vol. 23, no. 15, pp. 1049–1051, Aug. 12, 2011, doi: 10.1109/LPT.2011.2151853.

[4] A. Haglund, J. S. Gustavsson, J. Vukusic, P. Modh, and A. Larsson, "Single fundamental-mode output power exceeding 6 mW from VCSELs with a shallow surface relief," *IEEE Photon. Technol. Lett.*, vol. 16, no. 2, pp. 368–370, Feb. 2004, doi: 10.1109/LPT.2003.821085.

[5] A. J. Danner, T. S. Kim, and K. D. Choquette, "Single fundamental mode photonic crystal vertical cavity laser with improved output power," *Electron. Lett.*, vol. 41, no. 6, p. 325, Apr. 2005, doi: 10.1049/el:20057841.

[6] D. Zhou and L. J. Mawst, "High-power single-mode antiresonant reflecting optical waveguide-type vertical-cavity surface-emitting lasers," *IEEE J. Quantum Electron.*, vol. 38, no. 12, pp. 1599–1606, Dec. 2002, doi: 10.1109/JQE.2002.805107.

[7] G.-Z. Pan et al., "Multi-junction cascade 905 nm vertical cavity surface emitting lasers with high power density," *Acta Phys. Sinica*, vol. 71, no. 20, pp. 204203-1–204203-7, Oct. 2022, doi: 10.7498/aps.71.20220888.

[8] G. Pan, M. Xun, Z. Zhao, Y. Sun, J. Zhou, and D. Wu, "High slope efficiency bipolar cascade 905 nm vertical cavity surface emitting laser," *IEEE Electron Device Lett.*, vol. 42, no. 9, pp. 1342–1345, Sep. 2021, doi: 10.1109/LED.2021.3098899.

[9] J. F. Seurin et al., "Multi-junction vertical-cavity surface-emitting lasers in the 800–1100 nm wavelength range," *Proc. SPIE*, vol. 11704, Mar. 2021, Art. no. 117040B, doi: 10.1117/12.2578767.

[10] H. J. Unold, S. W. Z. Mahmoud, R. Jager, M. Kicherer, M. C. Riedl, and K. J. Ebeling, "Improving single-mode VCSEL performance by introducing a long monolithic cavity," *IEEE Photon. Technol. Lett.*, vol. 12, no. 8, pp. 939–941, Aug. 2000, doi: 10.1109/68.867967.

[11] S. W. Z. Mahmoud, H. J. Unold, W. Schmid, R. Jäger, R. Michalzik, and K. J. Ebeling, "Analysis of longitudinal mode wave guiding in vertical-cavity surface-emitting lasers with long monolithic cavity," *Appl. Phys. Lett.*, vol. 78, no. 5, pp. 586–588, Jan. 2001, doi: 10.1063/1.1344569.

[12] B. Demeulenaere, R. Baets, and D. De Zutter, "Rigorous electromagnetic study of diffraction loss in VCSEL mirrors," *IEE Proc., Optoelectron.*, vol. 143, no. 4, pp. 221–227, Aug. 1996, doi: 10.1049/ip-opt:19960652.

[13] C.-Y. Peng, K. Tsao, H.-T. Cheng, M. Feng, and C.-H. Wu, "Investigation of the current influence on near-field and far-field beam patterns for an oxide-confined vertical-cavity surface-emitting laser," *Opt. Exp.*, vol. 28, no. 21, pp. 30748–30759, Oct. 2020, doi: 10.1364/OE.397878.

[14] A. Ghods, M. Dummer, G. Xu, and K. Johnson, "Design and fabrication of single-mode multi-junction 905 nm VCSEL with integrated anti-phase mode filter," *J. Lightw. Technol.*, vol. 41, no. 10, pp. 3102–3107, May 15, 2023, doi: 10.1109/JLT.2023.3240123.

[15] G. R. Hadley, "Effective index model for vertical-cavity surface-emitting lasers," *Opt. Lett.*, vol. 20, no. 13, pp. 1483–1485, Jul. 1995, doi: 10.1364/OL.20.001483.

[16] Y. Zhou et al., "Large-aperture single-mode 795 nm VCSEL for chip-scale nuclear magnetic resonance gyroscope with an output power of 4.1 mW at 80 °C," *Opt. Exp.*, vol. 30, no. 6, p. 8991, Mar. 2022, doi: 10.1364/oe.450118.

[17] H. Li, P. Wolf, X. Jia, J. A. Lott, and D. Bimberg, "Thermal analysis of high-bandwidth and energy-efficient 980 nm VCSELs with optimized quantum well gain peak-to-cavity resonance wavelength offset," *Appl. Phys. Lett.*, vol. 111, no. 24, pp. 243508-1–243508-4, Dec. 2017, doi: 10.1063/1.5003288.

[18] M. Xun, G. Pan, Z. Z. Zhao, Y. Sun, J. Zhou, and D. Wu, "190 °C high-temperature operation of 905-nm VCSELs with high performance," *IEEE Trans. Electron Devices*, vol. 68, no. 6, pp. 2829–2834, Jun. 2021, doi: 10.1109/TED.2021.3074902.

[19] Y.-C. Chang and L. A. Coldren, "Efficient, high-data-rate, tapered oxide-aperture vertical-cavity surface-emitting lasers," *IEEE J. Sel. Topics Quantum Electron.*, vol. 15, no. 3, pp. 704–715, May 2009, doi: 10.1109/JSTQE.2008.2010955.

[20] A. Demir, G. Zhao, and D. G. Deppe, "Lithographic lasers with low thermal resistance," *Electron. Lett.*, vol. 46, no. 16, pp. 1147–1149, Aug. 2010, doi: 10.1049/el.2010.8532.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on July 17,2024 at 01:26:53 UTC from IEEE Xplore. Restrictions apply.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 韩英卓

方向： 量子科技

单位： 中国科学院大学物理科学学院

题目： 转角双层石墨烯中赝能隙态的实空间分布研究

Spectral signature of periodic modulation and sliding of pseudogap state in moiré system

Yingzhuo Han[1,2#], Yingbo Wang[1,2#], Yucheng Xue[1#], Jiefei Shi[2], Xiaomeng Wang[2], Kenji Watanabe[3], Takashi Taniguchi[4], Jian Kang[5], Yuhang Jiang[2*], Jinhai Mao[1*]

1. School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.
2. College of Materials Science and Optoelectronic Technology, Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China.
3. Research Center for Electronic and Optical Materials, National Institute for Materials Science, Tsukuba 305-0044, Japan.
4. International Center for Materials Nanoarchitectonics, National Institute for Materials Science, Tsukuba 305-0044, Japan.
5. School of Physical Science and Technology, ShanghaiTech University, Shanghai 200031, China.

Abstract:

The nature of the pseudogap state is widely believed as a key to understanding the pairing mechanism underlying unconventional superconductivity. Over the past two decades, significant efforts have been devoted to searching for spontaneous symmetry breaking or potential order parameters associated with these pseudogap states, aiming to better characterize their properties[1–7]. Recently, pseudogap states have also been realized in moiré systems with extensive gate-tunability, yet their local electronic structure remains largely unexplored[8]. In this study, we report the observation of gate-tunable spontaneous symmetry breaking and sliding behavior of the pseudogap state in magic-angle twisted bilayer graphene (MAtBG) using spectroscopic imaging scanning tunneling microscopy. Our spectroscopy reveals a distinct pseudogap at 4.4 K within the doping range -3 < ν < -2. Spectroscopic imaging highlights a gap size modulation at moiré scale that is sensitive to the filling, indicative of a wave-like fluctuating pseudogap feature. Specifically, the positions of gap size minima (GSM) coincide with regions of the highest local density of states (LDOS) at the filling ν = -2.63, but a unidirectional sliding behavior of GSM is observed for other fillings. In addition, the pseudogap size distribution at certain doping levels also causes a clear nematic order, or an anisotropic gap distribution. Our results have shed light on the complex nature of this pseudogap state, revealing critical insights into the phase diagram of correlated electron systems.

Introduction:

Since the discovery of high-temperature cuprate superconductors, the pseudogap and its coexisting intertwined orders have been intensively studied by various experimental techniques[5–7,9–11]. Elucidating the nature of these phases and their relationship with the pseudogap is critical for understanding the pairing mechanism underlying unconventional superconductivity, such as whether the pseudogap represents a precursor pairing state or a competing phase. Recently, the extraordinary superconducting phenomena have also been observed in moiré systems (such as magic-angle twisted bilayer graphene (MAtBG)[8,12–16], rhombohedral trilayer graphene[17], and twisted bilayer tungsten diselenide[18,19]) and attracted significant attention. Similar to cuprate superconductors, the strong correlation of electrons within the flat bands[20–25] is believed to play a crucial role in forming the superconducting state, complicating the understanding of both the origin of superconductivity and the underlying pairing mechanism. Furthermore, the pseudogap state has also been observed in MAtBG at temperatures above the superconducting transition temperature or under an applied magnetic field[8,26–28]. Nano-scale studies, particularly through local probe measurements, may offer valuable experimental insights into the spatial distribution of the pseudogap state and the presence of potential symmetry-breaking phases.

Here we use a spectra-imaging scanning tunneling microscope (SI-STM) to investigate the pseudogap feature in MAtBG and its evolution in real space at different doping levels. We found that the pseudogap state in MAtBG appears above the superconducting transition temperature. Although it occupies the same filling range as superconductivity, the pseudogap nature can be distinguished using point contact spectroscopy (PCS)[8,16]. The pseudogap size is found to depend on two degrees of freedom: real-space position and doping level. Through extensive dI/dV mapping at various doping levels, we observe modulations in pseudogap size comparable to the moiré superlattice scales and a peculiar phase sliding phenomenon. At one specific doping level (ν = -2.63), the maximum (minimum) gap values are located at positions with the lowest (highest) local density of states (LDOS), leading to a π phase difference between the pseudogap size and LDOS modulation. At other doping levels away from

ν = -2.63, the maximum (minimum) gap values shift away, indicating a pronounced sliding behavior. Additionally, a significant nematic order in the map of pseudogap size is observed, whereas the map of dI/dV intensity retains C_6 rotational symmetry. These experimental observations reveal the complex real-space distribution of pseudogap states at different doping levels, offering deeper insights into the nature of unconventional superconductivity in moiré systems.

Results:

Distinguishing the pseudogap feature

Fig. 1a illustrates the device setup utilized in this study to investigate MAtBG through the STM, where the tip is grounded and bias voltage is applied to the sample. The MAtBG sample is supported by a hexagonal boron nitride (hBN) substrate and placed on a Si/SiO_2 wafer, allowing carrier density tuning via gate voltage V_g applied to the Si substrate. The STM topography in Fig. 1b reveals the typical moiré pattern of MAtBG, with the black dashed line delineating one moiré unit cell. The red, blue, and yellow circles denote three distinct stacking regions: AA, AB, and BA, respectively. According to the measured moiré length $l_M = 13.44$ nm, a twist angle of 1.05°can be extracted, placing our sample within the 'magic angle' regime[29]. In this regime, the band width reaches a minimum, leading to two weakly dispersive flat bands (as shown in the calculated band structure in Fig. 1c), where electron-electron interaction plays a vital role in the electronic ground state. Moreover, each flat band has fourfold degeneracy, corresponding to isospin (spin and valley) degrees of freedom[30–32].

We utilize scanning tunneling spectroscopy (STS) to investigate the evolution of these flat bands under different doping levels. Firstly, as shown in Fig. 1e, each flat band produces a peak in the dI/dV spectrum[22–24,33]. At the charge neutral point (CNP), two peaks are symmetrically positioned on either side of the Fermi level ($V_{Bias} = E_F \equiv 0$ meV). When the flat bands become fully occupied (empty), the respective peaks shift to the left (right) of the E_F, respectively. The substantial difference in the spacing between the two flat bands at empty state (ΔE_{vHs} = 23 meV) and at CNP (ΔE_{vHs} = 49 meV) provides compelling evidence of strong electron correlations[21,34,35]. The filling

factor is defined as ν = -4 when two flat bands are both empty (green curve) and ν = +4 when two flat bands are fully occupied (red curve). Following this criterion, we could determine all the intermediate filling factors ν precisely in Fig. 1d.

By continuously tuning the carrier density via V_g and measuring dI/dV spectroscopy, we observe two flat bands sequentially pass through the E_F (Fig. 1d). At each integer filling, we observed cascade features (indicated by the black dashed line) that are ubiquitously observed in STM studies of MATBG[35–37]. The most striking feature emerges within the doping range of $-3 < \nu < -2$, where a clear gap feature is observed at E_F in both AA and AB stacking regions (green arrows). Previous research has demonstrated the emergence of unconventional superconductivity in this doping range[8,12]. Similar gap features have been visualized in MAtBG sample when superconductivity is suppressed either at temperatures above the transition temperature ($T_c \sim 1.7$ K) or in the presence of a magnetic field[8]. Given that our base temperature is 4.4 K, we attribute this gap to a pseudogap, which is further confirmed by the PCS and temperature-dependent measurements.

To provide a more detailed examination of this pseudogap feature, Fig. 1f shows the zoomed-in gate-dependent dI/dV spectra for both AA and AB stacking regions with finer gate voltage intervals and higher energy resolution. The pseudogap is manifested as a suppression of dI/dV intensity at E_F, further supporting its correlation-induced nature. Fig. 1g presents the dI/dV line cuts as a function of gate voltage ranging from V_g = -12 V to 26 V (left) and the corresponding second derivative of -dI/dV curves (right). This data processing method effectively mitigates the impact of the background intensity, particularly from the nearby flat bands, while enhancing the visibility of the pseudogap features without altering intrinsic gap size[38,39] (see Supplementary Information (SI), Section II for more comparison). The evolution of peaks flanking the pseudogap with the gate voltage can be clearly identified in the right panel (highlighted by the black dashed lines). We further compare the pseudogap features in the AA and AB regions and observe distinct evolution processes in each area. Fig. 1h shows the single dI/dV spectrum at one typical gate voltage in Fig. 1f (V_g = -17 V, corresponding to ν = -2.45), which exhibits a clear gap-size mismatch. This difference suggests a

modulation of the pseudogap size within the moiré scale, which may vary, leading to distinct modulation patterns under different doping conditions (see SI, Section II).

Although both states exhibit gap features, the pseudogap state fundamentally differs from the superconducting state due to its distinct physical mechanism. Consequently, point contact spectroscopy (PCS) measurements in STM experiments are commonly employed to distinguish them[8,40]. In PCS measurements of the SC state, as the tip approaches or even touches the SC surface, electrons from the metallic tip can be reflected as holes, with Cooper pairs propagating into the superconducting sample. This process, known as Andreev reflection, results in an intensity enhancement in the dI/dV spectrum near E_F [41,42]. However, in the pseudogap state, the absence of coherent Cooper pairs precludes Andreev reflection, leaving the gap features unchanged in the dI/dV spectrum near E_F.

We performed PCS measurements on the same sample at both 4.4 K and 0.4 K to compare these two distinct states in Fig. 2. We first cool down the sample to 0.4 K to identify the superconducting gap state. As we gradually increase the tunneling current setpoint to bring the tip closer to the sample surface, the two coherence peaks gradually come close together and eventually merge into one single peak at E_F (Fig. 2a), which is indicative of Andreev reflection. From the individual dI/dV spectra curves obtained with progressively higher setpoint values (Fig. 2b), a clear transition from the normal tunneling regime to the point contact regime can be observed, consistent with the schematic illustration in Fig. 2c. When we apply an increasing out-of-plane magnetic field B, the Andreev reflection peak is gradually suppressed and finally turns into a pseudogap feature beyond the critical value $B_c \sim 50$ mT (Fig. 2d). Previous transport experiments[12] measured this value as tens of mT, which is of the same order of magnitude as our experimental results. Fig. 2e employs Dynes-function to fit the gap in the tunneling spectrum measured at $\nu = -2.6$ using the model quasiparticle DOS for both an s-wave nodeless superconductor (grey curve) and a nodal superconductor (red curve). The fitting yields a superconducting order parameter of $2\Delta' = 2.20$ meV for s-wave superconductor and $2\Delta' = 3.06$ meV for nodal superconductor (see SI, Section III for fitting details). The STS spectra align more closely to a nodal superconductor, such as

higher-angular-momentum pairing (e.g., *p*- or *d*-wave) characterized by an anisotropic gap function. Recent transport measurements in graphene moiré systems have also revealed signatures consistent with nodal superconductivity through superfluid stiffness measurements[43,44]. All the behavior mentioned above is consistent with the previous transport[12,13,15] and probe[8] studies, providing strong evidence for the superconducting state at 0.4 K. However, the PCS measurements conducted at 4.4 K yield distinctly different results. As the tip gets close to the sample, the gap feature at ν = -2.6 persists even at a tunneling current of 150 nA (Fig. 2f and 2g), supporting our hypothesis that this gap corresponds to a pseudogap state. Considering the contrasting results at the two temperature intervals, we can confidently attribute the gap feature detected at 4.4 K to the pseudogap state.

Additionally, we measured the temperature-dependent dI/dV spectrum across the 4.6 K to 9.1 K range (Fig. 2h). As the temperature rises, the two peaks of pseudogap gradually converge, with the gap feature ultimately disappearing at approximately 7 K. Furthermore, gate-dependent dI/dV spectrum at 9 K confirms the lack of pseudogap feature within the doping range of $-3 < \nu < -2$, while the cascade features remain unaffected (Fig. 2i and 2j). The absence of both pseudogap and superconducting gap at higher temperatures further supports their correlation-induced nature, underscoring their similarity to those observed in cuprate superconductors[45].

Imaging spatial gap size modulation

Now we focus on the spatial variation of the pseudogap size (2Δ). We conducted real-space dI/dV mapping at five different doping levels and extracted the distribution of the pseudogap size (Δ-maps). Due to space limitations, results for four doping levels are presented in Fig. 3b-e, with corresponding topography shown in Fig. 3a; the data for ν = -2.40 are provided in SI, Section VII. The black dashed diamonds in each graph delineate the same moiré unit cell within the STM topography, with vertices corresponding to the AA stacking regions. The pseudogap sizes exhibit apparent inhomogeneity at all five doping levels, forming a triangular lattice at the moiré superlattice scale. In this arrangement, purple rounded regions correspond to gap size

minima (GSM), while surrounding orange regions represents larger gap sizes. As mentioned earlier, the AA stacking region is the highest point in terms of topography and is also the area where the LDOS of the flat band is primarily concentrated, whereas the LDOS of the flat band in the AB region is significantly reduced. Given this substantial modulation in the LDOS, the emergence of such real-space modulation is possible[6]. Nevertheless, the GSM does not consistently align with the LDOS configuration across five different fillings. Especially at ν = -2.63, the GSM is located precisely on the AA stacking region, resulting in a significant π phase shift between LDOS and pseudogap size (can be confirmed by the GSM precisely aligning with the vertices of the black rhombus in Fig. 3c). With this doping level as the reference, in the relatively lower hole-doping (ν = -2.45, in Fig. 3b), the GSM shifts upward to the left along one highly symmetric direction relative to the AA stacking region; while in the relatively higher hole-doping (ν = -2.81 and ν = -2.90, in Fig. 3d and 3e), the GSM shifts in precisely the opposite direction (i.e. down-right). This sliding behavior highlights the significant dependence of the spatial distribution of the pseudogap on varying doping levels.

To further quantify this sliding behavior, we extract the pseudogap sizes (2Δ) along the highly symmetric direction where GSM sliding occurs (indicated by the green arrows in each map), and compare them with the corresponding topographic heights (Z), which also reflect the modulation of LDOS. As shown in Fig. 3f-i, both 2Δ and Z exhibit sinusoidal modulation with a periodicity comparable to that of the moiré lattice. Notably, a phase difference is observed between these two quantities. We define δ_{d} as the difference between the position of the GSM and the center of the AA stacking region, and δ_{phase} as $2\pi \times \frac{\delta_{\mathrm{d}}}{L_{\mathrm{moire}}}$ to represent the relative phase shift due to the spatial sliding. The δ_{phase} values as a function of ν are plotted in Fig. 3k. At ν = -2.63, the δ_{phase} is almost zero, indicating a good alignment of the GSM with the AA stacking region. While at ν = -2.40 and ν = -2.45, the δ_{phase} approach $\pi/2$, indicates a noticeable sliding of the GSM away from the AA region in a specific direction. Conversely, at ν = -2.81 and ν = -2.90, the δ_{phase} is approximately $-\pi/2$, signifying sliding of GSM in the opposite

direction. These results highlight the doping-dependent nature of the spatial distribution of the pseudogap size. Based on prior research on superconductivity in MAtBG[8,13], we find that ν = -2.63 is near the "optimal" doped regime, where "optimal" means the superconducting transition temperature is the highest. While ν = -2.40 (ν = -2.90) correspond to the under-doped (over-doped) regimes, respectively. Our results highlight that the pseudogap modulation is not locked to the underlying moiré superlattice, but responses sensitively to the filling of the flat bands, presenting the wave-like fluctuating behavior.

Rotational symmetry-breaking in *Δ*-maps

Besides the δ_{phase} oscillation described above, signatures of rotational symmetry-breaking have also been revealed in two distinct ways. Firstly, the GSM sliding occurs in a unidirectional manner. For instance, at the filling factor of ν = -2.45, when overlaying the STM topographic image with the *Δ*-map, a directional misalignment is observed between the GSM regions (in the *Δ*-map) and the AA regions (in the topography). Rotating the entire dataset by 60° forces the AA regions in the topography to coincide, but the GSM regions near each AA site no longer align with any corresponding features, which constitutes one manifestation of rotational symmetry breaking. Another way of rotational symmetry-breaking can be directly observed from the real-space distribution of the pseudogap at some doping levels. For example, in Fig. 3d and 3e, the pseudogap exhibits a distinct nematic order, with the orientations differing. To further elucidate this characteristic, we applied the self-correlation analysis to the *Δ*-maps (details see SI, Section VIII). This analytical approach yields the correlation coefficients of the image with itself, and has therefore been used to amplify subtle symmetry and periodicity features that may be otherwise difficult to detect[46]. Fig. 4a and 4b show the *Δ*-map taken at ν = -2.63 ("optimal" doped regime) and the corresponding self-correlation image. Each GSM region exhibits an almost circular shape with preserved C_6 rotational symmetry, as clearly reflected by the separated bright circles in Fig. 4b. However, *Δ*-map taken at ν = -2.81 (Fig. 4e) shows contrasting results, where the purple GSM regions take on an ellipse shape, with the interspersed

orange areas displaying a nematic distribution. These features are manifested as one-dimensional bright lines in the self-correlation image (Fig. 4f), indicating a disruption of C_6 rotational symmetry (For the same analysis at more doping levels, see SI, Section X).

A crucial question is whether the rotational symmetry-breaking observed here has intrinsic or extrinsic origins. To address this, we extracted the dI/dV maps at the energy within the pseudogap (E = -1 meV), and performed the self-correlation analysis (as illustrated in Fig. 4c, d for ν = -2.63 and 4g, h for ν = -2.81). Notably, no signature of nematic order was detected in this LDOS channel in either case, suggesting that symmetry breaking is not a global external effect that leads to Fermi surface distortion. This makes the spontaneous symmetry-breaking phase within the pseudogap a more plausible explanation. Moreover, as shown in Fig. 4i, we identified a transition region where the nematic direction undergoes a sharp switch. Although the STM topography of this region (inset in Fig. 4i) reveals a uniformly C_6 symmetric moiré pattern, the two areas separated by the red dashed line exhibit distinct nematic axes, as indicated by the orange and green dashed arrows. Furthermore, we divided the image along the red dashed line and independently performed self-correlation analysis on each section, with the results aligning well with our hypothesis (Fig. 4j and 4k). The presence of such a domain wall not only reinforces the validity of spontaneous symmetry breaking but also suggests that the emergence of the nematic axis is not given by any global factors. Therefore, we can essentially rule out the extrinsic origins (such as strain, a common factor causing symmetry-breaking in two-dimensional materials), and instead focus on other intrinsic physical characteristics related to the origin of the pseudogap.

Notably, our work reveals distinct characteristics of nematic order compared to previous studies in twisted graphene systems[24,47]. While prior works primarily observed nematicity through moiré-scale LDOS stripe patterns or charge density modulations in correlated metallic/insulating states, we identify it via pseudogap size distributions while maintaining rotational symmetry in LDOS, revealing its unique association with superconducting precursor phenomena. This difference stems from our system's sensitivity to doping variations and its special pseudogap ground state, highlighting the

rich diversity of nematic behaviors in flat-band graphene systems.

Discussion

Our work first directly visualized the pseudogap distribution within such all-carbon systems with the flat bands, uncovering several novel findings. Thanks to the convenient carrier density tuning via gate voltage in MAtBG, we observed tunable rotational symmetry breaking—nematic phase drifting and orientation rotation with doping—suggesting nematic bond order rather than atomic site-specific symmetry breaking. This decoupling from the superlattice implies the pseudogap stems from electronic correlations, highlighting twisted 2D systems as ideal platforms for studying pseudogap physics. Remarkably, we confirmed the nematic order could extend over hundreds of nanometers and encompass dozens of moiré unit cells (see SI, Section VIIII). Given the ubiquitous occurrence of nematic ordering in the pseudogap phase of cuprate superconductors[5–7,38,48], this rotational symmetry breaking constitutes an intrinsic universal property of the pseudogap state. Consequently, our results demonstrate that electronic rotational symmetry breaking represents a fundamental order parameter characterizing the pseudogap phase.

The emergence of the nematicity does not cause an apparent translational symmetry breaking, therefore theoretically, it does not necessarily open a gap and thus cause the metal to insulator transition. It is plausible that the measured gap in the pseudogap phase results from another long-range order in the charge, spin, or valley channels. Although this order cannot be directly observed in the STM measurements, it is supported by the periodically modulated rather than uniformly distributed pseudogap observed in the real space. Moreover, the gap measured in the SC state is much larger than $k_B T_c$ ($T_c \sim 2$ K) and has roughly the same magnitude as the one in the non-SC state, suggesting that the gaps found in non-SC and SC states result from the same long-range order. Therefore, two orders, one rooted from the non-SC pseudogap state and much stronger than the pairing one, coexist in the SC states, while the former non-SC order dominates in the STM measured gap. Whether these two orders compete or cooperate still remain unknown. Our work has led to the realization that the pseudogap phase

could be associated with a variety of competing orders, which may coexist with superconductivity and can influence its emergence. Notably, experimental and theoretical studies on the pseudogap in cuprate superconductors also support this "intertwined order" scenario[9], demonstrating the universality of this fundamental physical phenomenon. Such parallel behavior underscores the pseudogap's potential role as a unifying framework for understanding correlated electron phenomena in diverse systems.

Methods:

Sample fabrication. The devices were fabricated using a "tear-and-stack" method[52]. Monolayer graphene (MLG) and few-layer hexagonal boron nitride (hBN) were first exfoliated onto SiO_2/Si substrates and characterized using optical and atomic force microscopes. A polypropylene carbonate (PPC) film and polydimethylsiloxane (PDMS) were used to sequentially pick up the exfoliated flakes—hBN, half of the MLG, and then the remaining half of the MLG. Before picking up the second half of the MLG, the stage was manually rotated by approximately 1° to achieve the desired twisted bilayer graphene (tBG) structure. All pick-up procedures were performed at temperatures between 40-50°C. After assembling the heterostructure, the PPC film was peeled off, flipped, and transferred onto a SiO_2/Si substrate (SiO_2 thickness: 285 nm) at 105°C. The PPC film was then removed by high-vacuum annealing at 350°C. To facilitate scanning tunneling microscopy (STM) measurements, the heterostructures were patterned into gate-tunable devices using e-beam lithography followed by Au/Ti metal evaporation. The devices were then annealed in an ultra-high vacuum chamber at 400°C overnight to remove any residual before the STM measurements.

STM measurements. The STM experiments are performed at 4.4 K with an ultrahigh vacuum better than 1×10^{-10} mbar. The STM tips are calibrated on Cu(111) before use. The topographies are obtained in constant current mode. The dI/dV spectra and maps are collected by standard lock-in technique, with 0.5 mV AC modulation at 711 Hz added to the DC sample bias. The low temperature data were acquired at 0.4 K by ^{3}He evaporation. The magnetic field is provided by the current in the superconducting coil, and the direction of the magnetic field is perpendicular to the sample surface.

STS mapping and Δ-map extraction procedure. We performed real-space STS mapping in the following way: with the feedback loop open, the tip was placed on the desired position in constant current mode. Then, the feedback loop was closed and the bias was swept in constant-height mode. Once the dI/dV curve was obtained, the feedback loop was opened again and the tip moved to the next position in constant current mode. The pseudogap size is extracted from the dI/dV spectra acquired at each spatial point. Specifically, the energy positions of the peaks flanking the gap are identified to determine the gap size. By correlating these gap sizes with their spatial coordinates, we construct the Δ-map.

Data Availability:

Relevant data supporting the key findings of this study are available within the article and the Supplementary Information file. All raw data generated during the current study are available from the corresponding authors upon request.

Acknowledgements:

We appreciate the enlightening discussions with Prof. Kun Jiang, Prof. Fuchun Zhang, and Prof. Jianxin Li.

Author contributions:

J.M. and Y.J. conceived and supervised the experiments. Y.H., J.S., and X.W. were responsible for device fabrication. Y.H. and Y.W. conducted the STM measurements, assisted by Y.X. K.W. and T.T. provided the hBN crystals. All authors participated in discussions of the results and contributed to the manuscript preparation.

Competing interests:

The authors declare no competing interests.

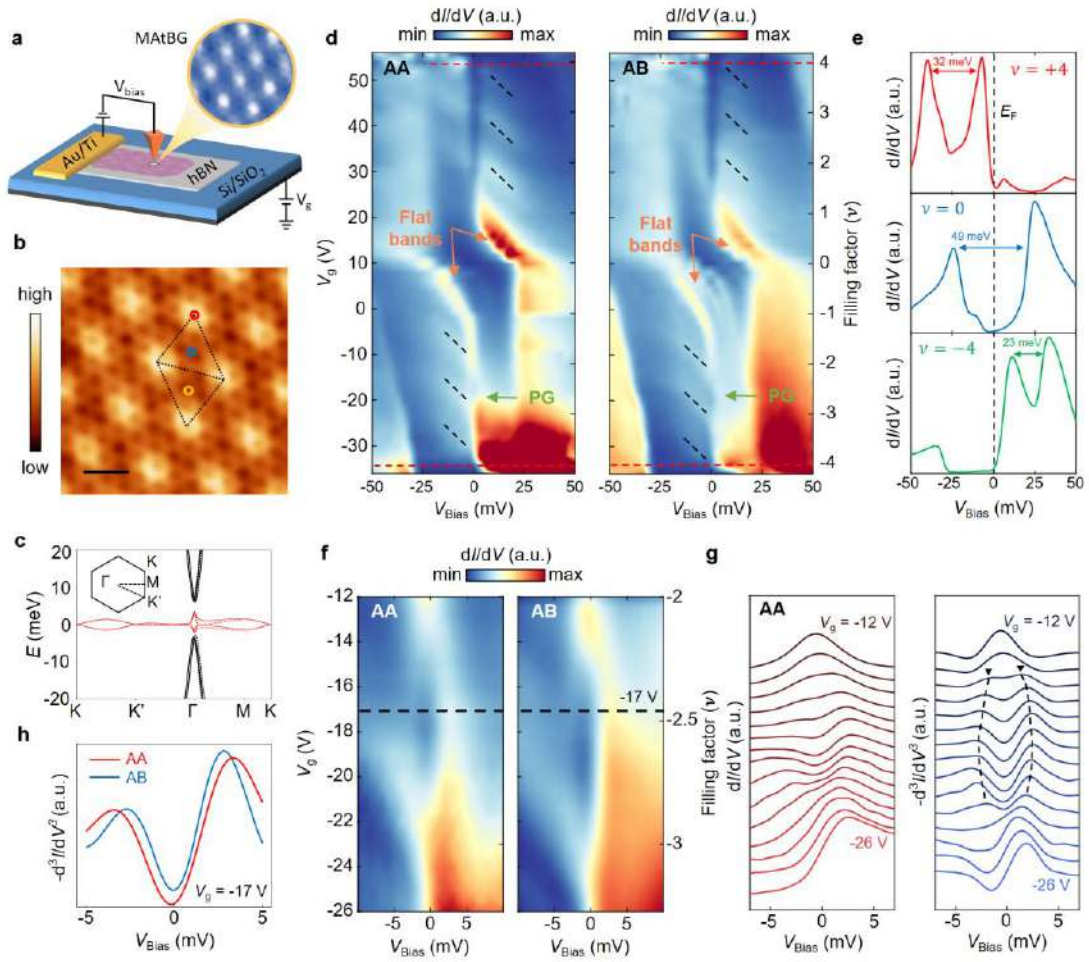


Fig. 1 Electronic structure of MAtBG and pseudogap states at different stacking regions. a, The schematic of the device and STM experimental setup. V_{Bias} is the bias voltage applied between the sample and tip, and V_{g} is the gate voltage used to tune the carrier concentration. **b,** STM topography of the MAtBG (V_{Bias} = -100 mV, I = 0.1 nA; scale bar, 10 nm). The brightest region is the AA stacked region (red), the darkest triangular region is the AB/BA stacked region (blue / yellow), and the domain wall (dashed line) structure connects the AA stacked region. **c,** Band structure calculated using a continuum model for MAtBG (θ = 1.05°), with red color highlighting two flat bands. Solid and dashed lines correspond to the band of **K** and **−K** valley. The inset provides a sketch of its Brillouin zone. **d,** Gate-dependent dI/dV spectra of the AA (left) and AB (right) stacked region (V_{Bias} = -50 mV, I = 0.4 nA). The flat bands / pseudogap (PG) are indicated by orange / green arrows. The black dashed lines label the cascade features and the red dashed lines indicate the filling factors as ±4. **e,** dI/dV spectra taken in the AA regions for the fully occupied band (red curve, ν = +4), charge neutral point (blue curve, ν = 0), and fully empty band (green curve, ν = -4). **f,** High-resolution dI/dV spectra for the AA and AB stacked regions at V_{g} ranging from -12 to -26 V (V_{Bias} = -50 mV, I = 0.4 nA). **g,** Left, dI/dV line cuts for the AA regions at the same V_{g} in **f**. Right, -d^3I/dV^3 curves of the spectra in the left panel. The black dashed lines and arrows indicate the evolution of two peaks. **h,** -d^3I/dV^3 spectra for the AA (red curve) and AB (blue curve) regions at V_{g} = -17 V.

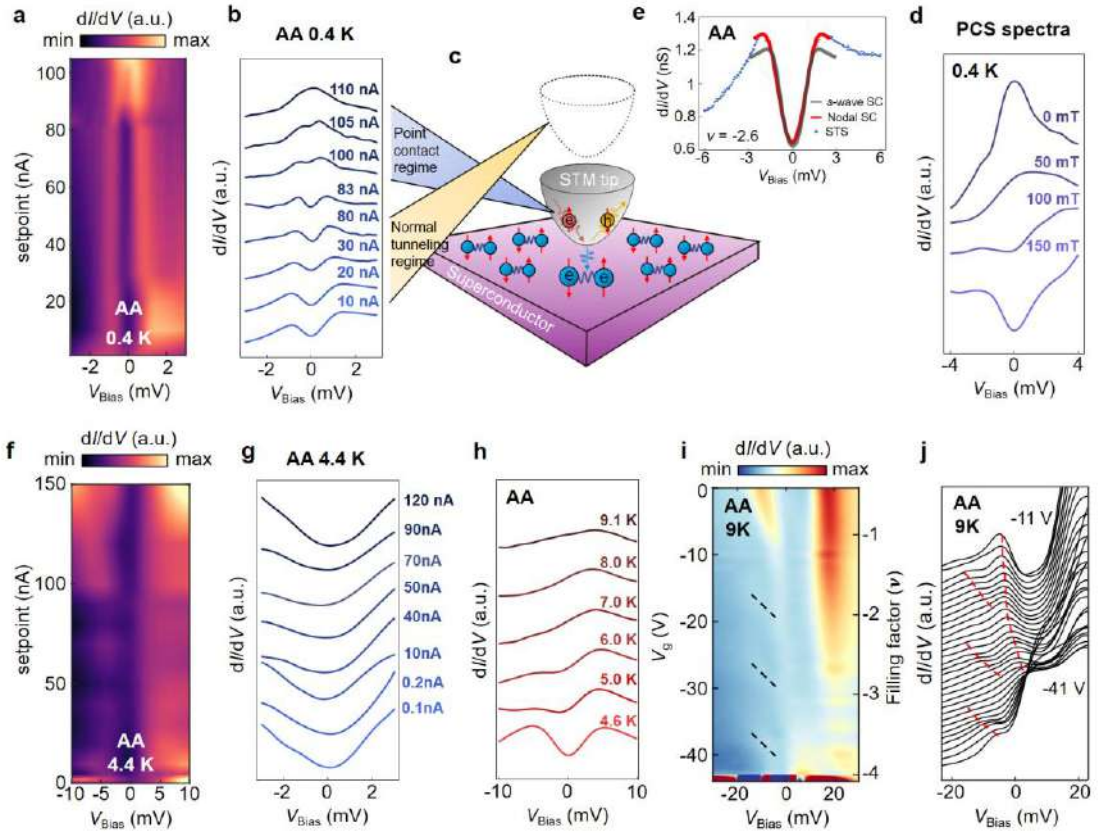


Fig. 2 PCS measurements to identify gap states at ν = -2.6. a, Point contact spectra (PCS) at the superconducting phase (ν = -2.6) for the AA stacked region at 0.4 K. The spectra are normalized for clarity. **b,** Normalized setpoint-dependent dI/dV line cuts ranging from I = 10 nA to 110 nA. When setpoint exceeds 80 nA, the tip transitions from the normal tunneling regime to the point contact regime, with an intensity enhancement occurs near E_F. **c,** A schematic diagram illustrates Andreev reflection occurring at the interface between a metallic STM tip and an SC sample with two regimes. **d,** Magnetic-field dependent PCS dI/dV spectra at the superconducting phase (ν = -2.6). Curves are normalized for clarity. **e,** Dynes-function fits to the experimental tunneling spectrum (blue dots) measured at ν = -2.6 at AA stacked region at 0.4 K using the model quasiparticle DOS for a nodeless *s*-wave superconductor (grey curve) a nodal superconductor (red curve). The STS spectra is clearly more consistent with the nodal superconductor fitting. **f,** PCS at the pseudogap phase (ν = -2.6) for the AA stacked region at 4.4 K. The spectra are normalized for clarity. **g,** Normalized setpoint-dependent dI/dV line cuts ranging from I = 0.1 nA to 120 nA. **h,** Temperature-dependent dI/dV spectra for the pseudogap state (ν = -2.6) of AA stacked region (V_{Bias} = -50 mV, I = 0.4 nA). **i,** Gate-dependent dI/dV spectra taken at 9 K (V_{Bias} = -50 mV, I = 0.4 nA). The pseudogap state is absent between ν = -2 and -3. Cascade features are indicated by the black dashed lines. **j,** Detailed gate-dependent dI/dV intensity plot for the gate voltage range same as **i**. Red dashed lines labeled the cascade feature (isospin flavors) and the dI/dV intensity near E_F.

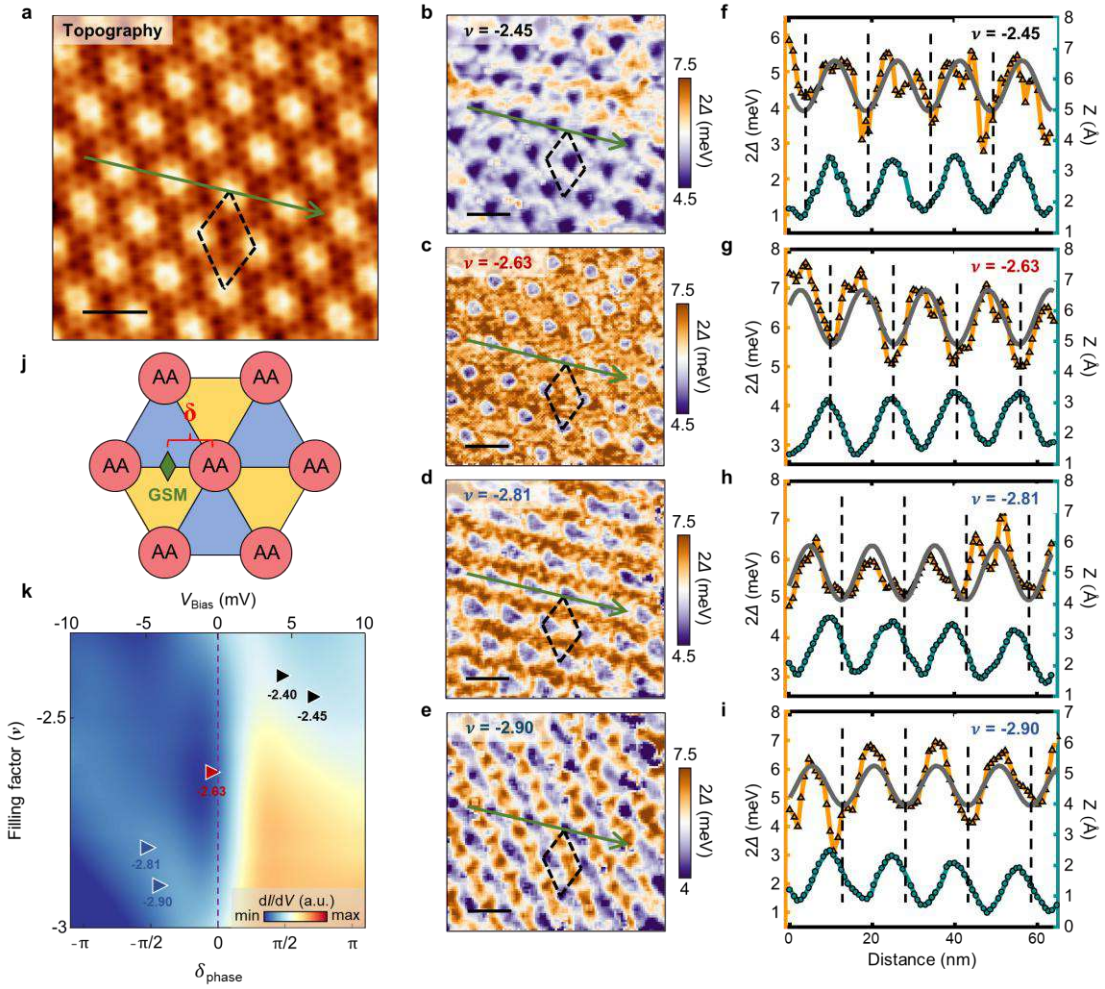


Fig. 3 Spatial variations of the pseudogap. a, The STM topography of 80×80 nm (V_{Bias} = -100 mV, I = 0.1 nA, scale bar, 16 nm). The dashed diamond represents a moiré unit cell, which is also used as a marker to label the same position in **b-e**. The green arrow represents the special highly symmetry direction where the gap size minima (GSM) sliding occurs. **b-e,** Δ-maps extracted from the STS map (V_{Bias} = -50 mV, I = 0.4 nA) taken from the same area as shown in **a** at ν = -2.45 (**b**), ν = -2.63 (**c**), ν = -2.81 (**d**), and ν = -2.90 (**e**). **f-i,** The extracted gap sizes (orange curve) and the topographic heights (turquoise curve) along the green arrow in **b-e** correspondingly. The grey line is fitted curve using a sine function to trace the oscillating gap. The black dashed line indicates the position of the GSM and its corresponding location on the topographic height. **j,** The schematic diagram of the definition of δ_{phase}, which represents the relative position of GSM compared with the center of AA stacking region (i.e. the highest point in topography) within one moiré unit cell. **k,** The extracted δ_{phase} at different ν (marked as solid triangles). The background corresponds to the dI/dV spectra of the AB stacking region at filling factors ν = -2.3 ~ -3. It can be observed that the pseudogap is relatively larger at ν = -2.63 (the "optimal" filling), while it is smaller at the other four filling factors.

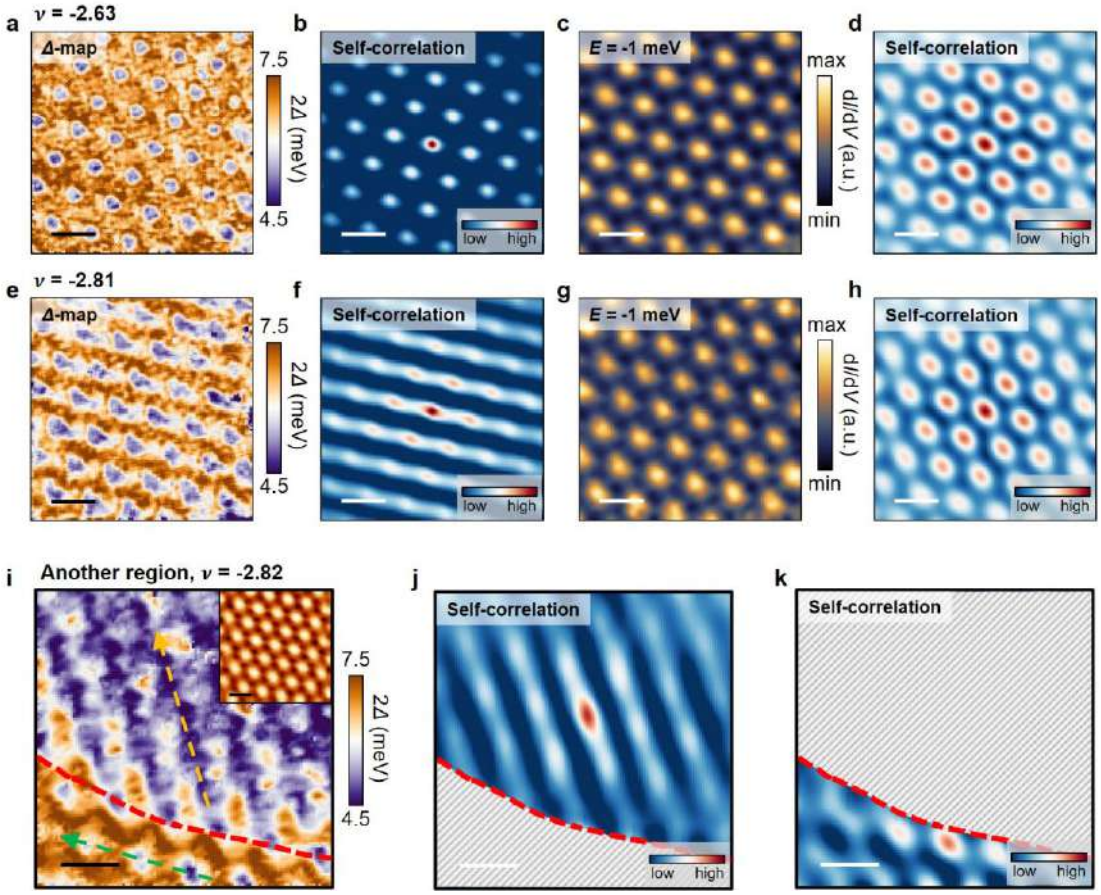


Fig. 4 Visualization of the nematic order in *Δ*-map. a,b, *Δ*-map and the corresponding self-correlation image at ν = -2.63, exhibiting retained C_6 rotational symmetry. **c,d,** d*I*/d*V* intensity map at energy within the gap (*E* = -1 meV) and the corresponding self-correlation image, taken in the same area and doping level as **a**. **e-h,** Same as **a-d**, but taken at ν = -2.81. The *Δ*-map in **e** exhibiting clear nematic order, which is further revealed in the self-correlation image in **f**. **i-k,** *Δ*-map (V_{Bias} = -50 mV, *I* = 0.4 nA) and the corresponding self-correlation image taken at anther region (region B) at ν = -2.82. The red dashed line separate two areas with distinct nematicity axes (pointed by the orange and green dashed arrows in **i**), consistent with the results shown in the self-correlation image in **j** and **k**. The inset in **i** displays the STM topography of the region B. The scale bars in all the images are 16 nm.

Reference

1. Vershinin, M. *et al.* Local Ordering in the Pseudogap State of the High-Tc Superconductor $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Science* **303**, 1995–1998 (2004).
2. Hanaguri, T. *et al.* A 'checkerboard' electronic crystal state in lightly hole-doped $Ca_{2-x}Na_xCuO_2Cl_2$. *Nature* **430**, 1001–1005 (2004).
3. McElroy, K. *et al.* Coincidence of Checkerboard Charge Order and Antinodal State Decoherence in Strongly Underdoped Superconducting $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Phys. Rev. Lett.* **94**, 197005 (2005).
4. Daou, R. *et al.* Broken rotational symmetry in the pseudogap phase of a high-Tc superconductor. *Nature* **463**, 519–522 (2010).
5. Lawler, M. J. *et al.* Intra-unit-cell electronic nematicity of the high-Tc copper-oxide pseudogap states. *Nature* **466**, 347–351 (2010).
6. Kohsaka, Y. *et al.* Visualization of the emergence of the pseudogap state and the evolution to superconductivity in a lightly hole-doped Mott insulator. *Nat. Phys.* **8**, 534–538 (2012).
7. Mukhopadhyay, S. *et al.* Evidence for a vestigial nematic state in the cuprate pseudogap phase. *Proc. Natl. Acad. Sci.* **116**, 13249–13254 (2019).
8. Oh, M. *et al.* Evidence for unconventional superconductivity in twisted bilayer graphene. *Nature* **600**, 240–245 (2021).
9. Fradkin, E., Kivelson, S. A. & Tranquada, J. M. Colloquium: Theory of intertwined orders in high temperature superconductors. *Rev. Mod. Phys.* **87**, 457–482 (2015).
10. Wu, T. *et al.* Magnetic-field-induced charge-stripe order in the high-temperature superconductor $YBa_2Cu_3O_y$. *Nature* **477**, 191–194 (2011).
11. Comin, R. *et al.* Charge Order Driven by Fermi-Arc Instability in $Bi_2Sr_{2-x}La_xCuO_{6+\delta}$. *Science*

343, 390–392 (2014).

12. Cao, Y. *et al.* Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
13. Lu, X. *et al.* Superconductors, orbital magnets and correlated states in magic-angle bilayer graphene. *Nature* **574**, 653–657 (2019).
14. Arora, H. S. *et al.* Superconductivity in metallic twisted bilayer graphene stabilized by WSe2. *Nature* **583**, 379–384 (2020).
15. Yankowitz, M. *et al.* Tuning superconductivity in twisted bilayer graphene. *Science* **363**, 1059–1064 (2019).
16. Kim, H. *et al.* Evidence for unconventional superconductivity in twisted trilayer graphene. *Nature* **606**, 494–500 (2022).
17. Chen, G. *et al.* Signatures of tunable superconductivity in a trilayer graphene moiré superlattice. *Nature* **572**, 215–219 (2019).
18. Xia, Y. *et al.* Superconductivity in twisted bilayer WSe_2. *Nature* 1–6 (2024) doi:10.1038/s41586-024-08116-2.
19. Guo, Y. *et al.* Superconductivity in 5.0° twisted bilayer WSe_2. *Nature* **637**, 839–845 (2025).
20. Kim, K. *et al.* Tunable moiré bands and strong correlations in small-twist-angle bilayer graphene. *Proc. Natl. Acad. Sci.* **114**, 3364–3369 (2017).
21. Xie, Y. *et al.* Spectroscopic signatures of many-body correlations in magic-angle twisted bilayer graphene. *Nature* **572**, 101–105 (2019).
22. Kerelsky, A. *et al.* Maximized electron interactions at the magic angle in twisted bilayer graphene. *Nature* **572**, 95–100 (2019).

23. Choi, Y. *et al.* Electronic correlations in twisted bilayer graphene near the magic angle. *Nat. Phys.* **15**, 1174–1180 (2019).

24. Jiang, Y. *et al.* Charge order and broken rotational symmetry in magic-angle twisted bilayer graphene. *Nature* **573**, 91–95 (2019).

25. Zhang, Z. *et al.* Flat bands in twisted bilayer transition metal dichalcogenides. *Nat. Phys.* **16**, 1093–1096 (2020).

26. Saito, Y., Ge, J., Watanabe, K., Taniguchi, T. & Young, A. F. Independent superconductors and correlated insulators in twisted bilayer graphene. *Nat. Phys.* **16**, 926–930 (2020).

27. Stepanov, P. *et al.* Untying the insulating and superconducting orders in magic-angle graphene. *Nature* **583**, 375–378 (2020).

28. He, M. *et al.* Competing correlated states and abundant orbital magnetism in twisted monolayer-bilayer graphene. *Nat. Commun.* **12**, 4727 (2021).

29. Bistritzer, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl. Acad. Sci.* **108**, 12233–12237 (2011).

30. Song, Y. J. *et al.* High-resolution tunnelling spectroscopy of a graphene quartet. *Nature* **467**, 185–189 (2010).

31. Young, A. F. *et al.* Spin and valley quantum Hall ferromagnetism in graphene. *Nat. Phys.* **8**, 550–556 (2012).

32. Freitag, N. M. *et al.* Large tunable valley splitting in edge-free graphene quantum dots on boron nitride. *Nat. Nanotechnol.* **13**, 392–397 (2018).

33. Li, G. *et al.* Observation of Van Hove singularities in twisted graphene layers. *Nat. Phys.* **6**, 109–113 (2010).

34. Nuckolls, K. P. *et al.* Strongly correlated Chern insulators in magic-angle twisted bilayer graphene. *Nature* **588**, 610–615 (2020).

35. Kang, J., Bernevig, B. A. & Vafek, O. Cascades between Light and Heavy Fermions in the Normal State of Magic-Angle Twisted Bilayer Graphene. *Phys. Rev. Lett.* **127**, 266402 (2021).

36. Wong, D. *et al.* Cascade of electronic transitions in magic-angle twisted bilayer graphene. *Nature* **582**, 198–202 (2020).

37. Zondiner, U. *et al.* Cascade of phase transitions and Dirac revivals in magic-angle graphene. *Nature* **582**, 203–208 (2020).

38. Ye, S. *et al.* The emergence of global phase coherence from local pairing in underdoped cuprates. *Nat. Phys.* **19**, 1301–1307 (2023).

39. Ruan, W. *et al.* Visualization of the periodic modulation of Cooper pairing in a cuprate superconductor. *Nat. Phys.* **14**, 1178–1182 (2018).

40. Kim, H. *et al.* Evidence for unconventional superconductivity in twisted trilayer graphene. *Nature* **606**, 494–500 (2022).

41. Peters, O. *et al.* Resonant Andreev reflections probed by photon-assisted tunnelling at the atomic scale. *Nat. Phys.* **16**, 1222–1226 (2020).

42. Dubouchet, T. *et al.* Collective energy gap of preformed Cooper pairs in disordered superconductors. *Nat. Phys.* **15**, 233–236 (2019).

43. Tanaka, M. *et al.* Superfluid stiffness of magic-angle twisted bilayer graphene. *Nature* **638**, 99–105 (2025).

44. Banerjee, A. *et al.* Superfluid stiffness of twisted trilayer graphene superconductors. *Nature* **638**, 93–98 (2025).

45. Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. *Rep. Prog. Phys.* **62**, 61 (1999).

46. Tsui, Y.-C. *et al.* Direct observation of a magnetic-field-induced Wigner crystal. *Nature* **628**, 287–292 (2024).

47. Rubio-Verdú, C. *et al.* Moiré nematic phase in twisted double bilayer graphene. *Nat. Phys.* **18**, 196–202 (2022).

48. Kohsaka, Y. *et al.* An Intrinsic Bond-Centered Electronic Glass with Unidirectional Domains in Underdoped Cuprates. *Science* **315**, 1380–1385 (2007).

49. Wang, L. *et al.* One-Dimensional Electrical Contact to a Two-Dimensional Material. *Science* **342**, 614–617 (2013).

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 李雅楠

方向： 量子科技

单位： 中国科学院理化技术研究所

题目： mK温区连续制冷的多级绝热去磁制冷机

mK 温区连续制冷的多级绝热去磁制冷机

摘要：随着量子科技、凝聚态物理研究和空间观测等前沿科学的迅速发展，对极低温制冷技术提出了日益严苛的需求。绝热去磁制冷(ADR)作为获取毫开尔文温区的重要技术手段，因其覆盖温区广、本征效率高、无需重力和稀缺资源 ^{3}He 等优势不仅可以满足小型量子系统对极低温环境的需求，同时也是空间探测任务中的主流极低温制冷技术。尽管我国在该技术上已实现多级系统和最低温度的显著突破，但其固有的非连续的制冷模式仍制约着更广泛的实际应用。针对上述问题，本研究设计并搭建了一台三级绝热去磁制冷机，由 GM 型脉管制冷机提供 4K 预冷，高温级选用钆镓石榴石磁热模块，低温级选用铬钾明矾作为顺磁盐工质，级间传热通过自主研制的主动气隙式热开关进行调控。通过对多级 ADR 系统运行时序和级间热耦合特性的不断探索，系统实现了在 500mK 温度下提供 3μW 的连续制冷能力。本文将进一步探讨当前实现 mK 温区连续制冷所面临的关键技术瓶颈，重点解析热开关动态行为、级间热耦合参数以及系统运行策略对整机制冷性能的影响机制。该项研究工作不仅为推动 mK 温区连续制冷系统的进一步发展指明方向，也展示了绝热去磁制冷技术为未来量子科技提供极低温环境的巨大潜力。

关键词：量子科技 极低温 绝热去磁 连续制冷

论文将在其他学术期刊上进行发表

主要成果图：

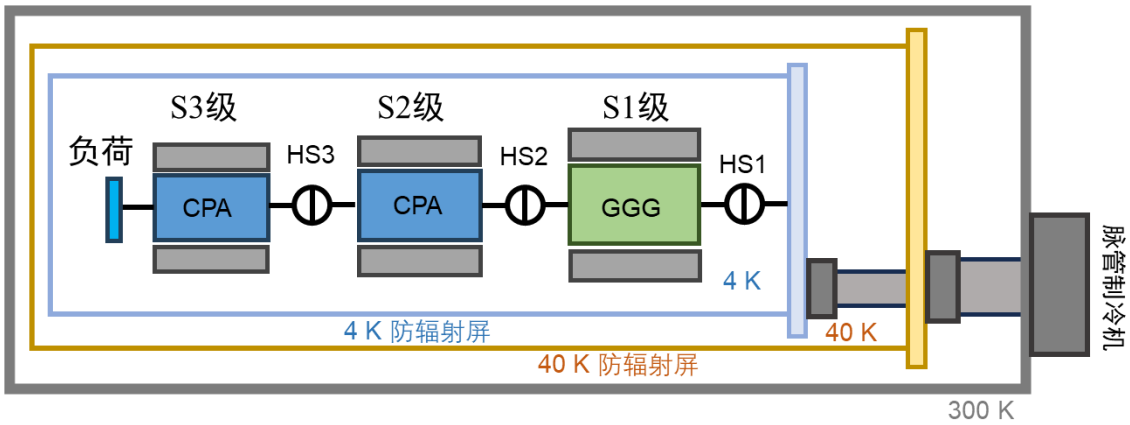


图 1 三级 ADR 系统示意图

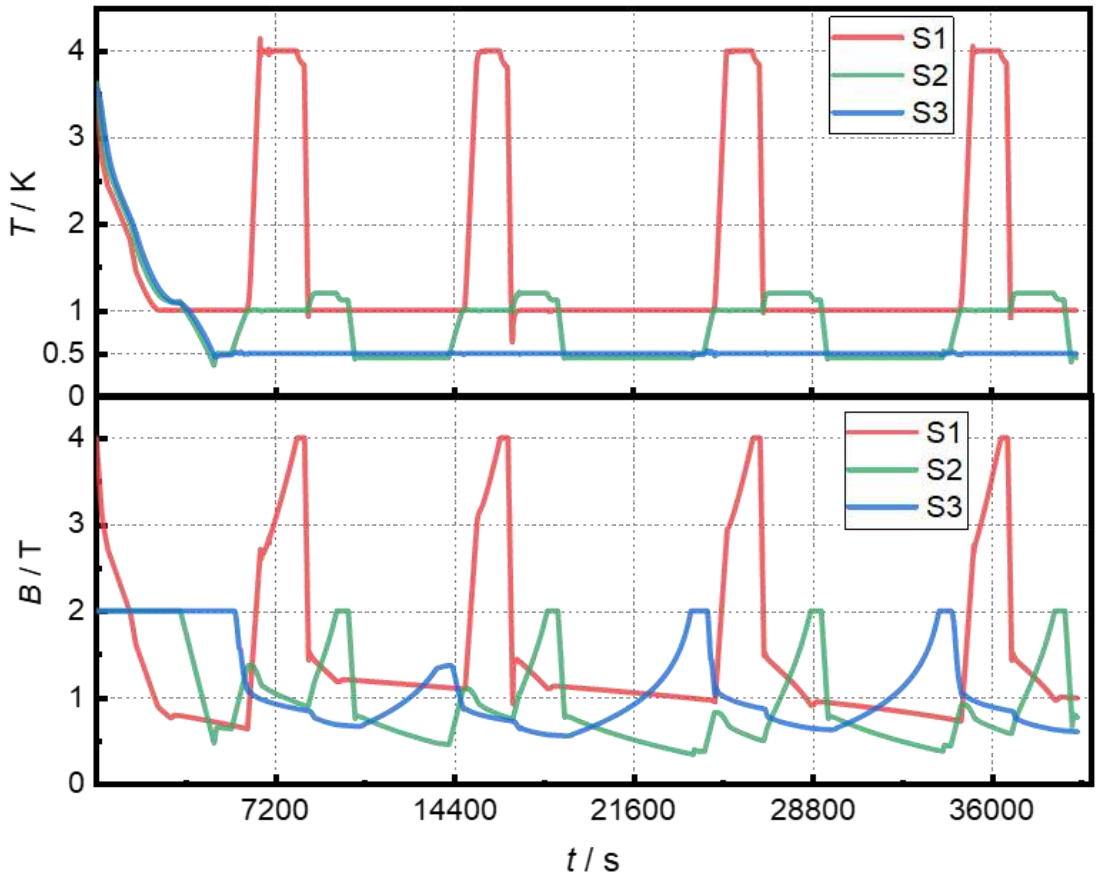


图 2 500mK@3μW 连续制冷

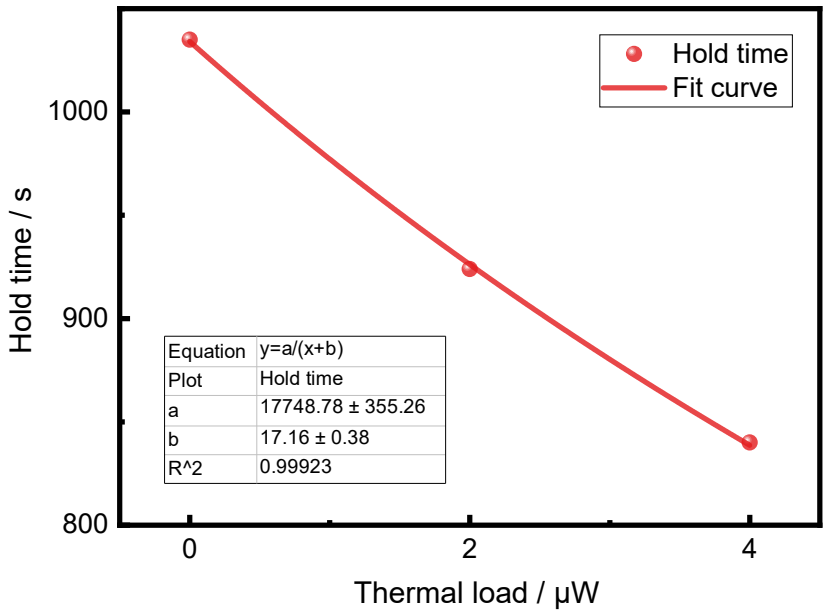


图 3 S3 级漏热（热开关短路 OFF）

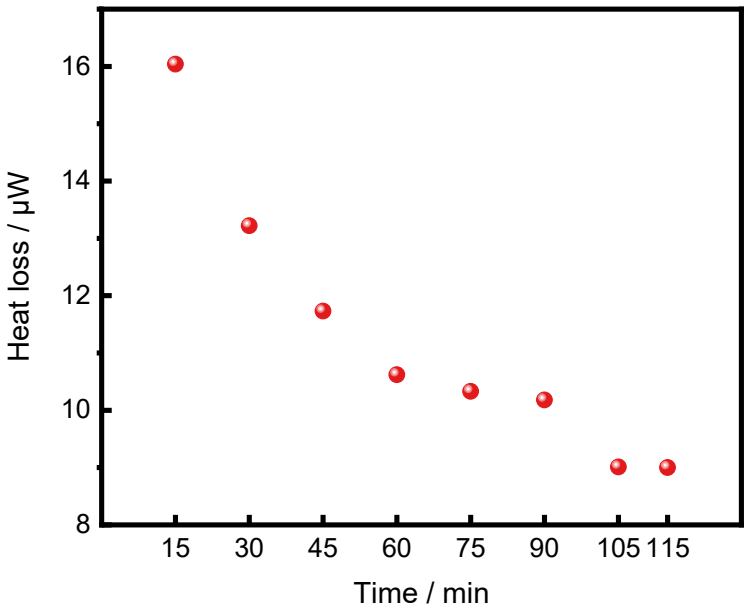


图 4 热开关断开时间测试

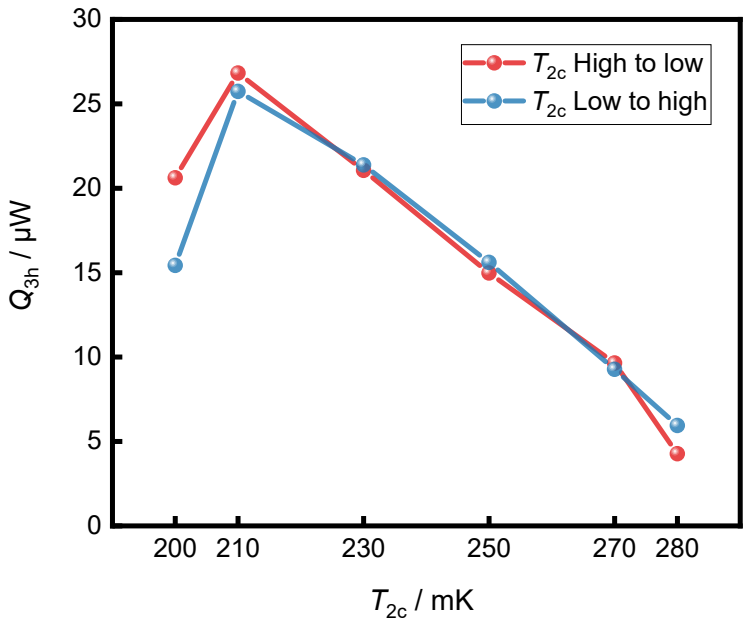


图 5 S2 级不同制冷温度下，S3 级再生热流密度

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名： 王小妮

方向： 量子科技

单位： 中国科学院上海微系统与信息技术研究所

题目： 基于超导纳米线的量子相滑移器件研究

Superconductor Science and Technology

PAPER

Evolution of phase slips with wire width and observation of critical-voltage feature without high-impedance environment in NbN nanowires

To cite this article: Xiaoni Wang *et al* 2024 *Supercond. Sci. Technol.* **37** 045001

View the article online for updates and enhancements.

You may also like

- Enhanced light extraction from organic light-emitting diodes using a quasi-periodic nano-structure
Ju Sung Lee, Yong Sub Shim, Chan Hyuk Park et al.
- Superconductor–insulator transition in nanowires and nanowire arrays
J E Mooij, G Schön, A Shnirman et al.
- 100 kA/10 kV thyristor stack design for the quench protection system in CRAFT
Wei TONG, , Meng XU et al.

This content was downloaded from IP address 180.169.212.98 on 24/02/2024 at 02:37

IOP Publishing Superconductor Science and Technology
Supercond. Sci. Technol. **37** (2024) 045001 (8pp) https://doi.org/10.1088/1361-6668/ad23fe

Evolution of phase slips with wire width and observation of critical-voltage feature without high-impedance environment in NbN nanowires

Xiaoni Wang[1,2], Zulei Xu[1,2], Xiaoyu Liu[1], Fei Xu[1], Yixin Liu[1,2], Wanpeng Gao[1,2], Yu Wu[1], Ming Yang[1,3], Wei Peng[1,2], Zhen Wang[1,2], Gang Mu[1,2,*], Zhi-Rong Lin[1,2,*] and Xiaoming Xie[1,2]

[1] National Key Laboratory of Materials for Integrated Circuits, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, People's Republic of China
[2] University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China
[3] School of Microelectronics, Shanghai University, Shanghai 200444, People's Republic of China

E-mail: mugang@mail.sim.ac.cn and zrlin@mail.sim.ac.cn

Received 18 October 2023, revised 15 January 2024
Accepted for publication 26 January 2024
Published 23 February 2024

Abstract

Quantum phase slip (QPS) is an important effect that occurs in one-dimensional (1D) or quasi-1D superconducting (SC) materials. Due to the extremely high requirements on the uniformity and quality of SC nanowires, it is currently very challenging to obtain repeatable and controllable QPS devices. Here, we report a systematic study of the QPS effect in NbN SC nanowires. High-quality NbN nanowires with different widths are fabricated by combining the processes of photo-lithography, electron beam lithography, and inductively coupled plasma etching. It is found that the electrical transport of the nanowires showed a systematic variation with the wire width, and a clear evolution from single to multiple QPS events with the decrease in wire width is uncovered at low temperatures. Based on these experimental results, the phase diagram reflecting the evolution of the physical states of the nanowires with wire width and temperature is obtained. In addition, we observed the features of the critical voltage, which is a characteristic of the occurrence of coherent QPS (cQPS), in samples with a width of 36 nm in the multiple QPS region. It is worth noting that this cQPS behavior was detected in a non-high-impedance environment.

Keywords: quantum phase slip (QPS), NbN nanowires, phase diagram

1. Introduction

The miniaturization of superconducting (SC) electronic devices and circuits requires a deeper understanding of the transport properties of SC materials at the nanoscale. When the cross-sectional dimensions of a SC nanowire are reduced to near or smaller than the coherence length, it can be treated as a one-dimensional or quasi-1D conducting channel. In such cases, the thermal excitation and quantum tunneling effect will influence the low-temperature physics of SC nanowires [1–4] and cause the fluctuations of the SC order parameters, resulting in the occurrence of phase slip [5–8]. As the temperature decreases, the influence of thermal excitation will gradually be overwhelmed by the quantum tunneling effect, and the quantum phase slip (QPS) effect begins to dominate [9, 10]. This effect occurs with a 2π change in the phase difference of the nanowires over the scale of the SC coherence length,

* Authors to whom any correspondence should be addressed.

© 2024 IOP Publishing Ltd

resulting in the disappearance of the non-dissipative SC state. From another perspective, QPS can result in the localization of Cooper pairs and the tunneling of magnetic flux on both sides of the nanowire [11], which is a macroscopic quantum tunneling phenomenon [12–14] at extremely low temperatures. In the past, researchers have carried out in-depth research on this quantum effect [15–19] and basically clarified the main factors affecting the QPS effect. It has been found that small cross-sectional area of SC nanowires, low critical temperature of superconductivity and high normal-state resistance will help to reduce the energy barrier for QPS, thereby increasing the probability of its occurrence [8].

With the deepening of research, people gradually realized that the QPS effect can serve as an excellent platform for applications in qubit and quantum current standards. In 2005, Mooij *et al* proposed that SC rings with QPS junctions can be used to achieve QPS-based magnetic flux qubit [20]. In the following years, they further showed that the coherent QPS (cQPS) effect, where the QPS rate is strong enough that coherent quantum tunneling can occur between different flux quantum states on both sides of the nanowire, is a dual process of the Cooper-pair tunneling effect in the Josephson junction [21]. Based on this, they predicted the existence of current Shapiro steps in a QPS device similar to the voltage Shapiro steps observed in the Josephson junction, and proposed to verify quantum effects in phase slip by preparing qubit based on SC QPS. In 2012, Tsai's team reported the first cQPS qubit, providing important evidence for verifying the existence of cQPS [22]. In 2022, the current Shapiro steps were observed in NbN nanowires [23], but the flatness of the steps is still not able to meet the requirements of applications in quantum current standards. Currently, there is still a lack of systematic research on the key factors that affect the stability and quality of cQPS. Obviously, the improvement of the current Shapiro step quality based on the cQPS effect, and even the real application of the quantum current standards still needs further efforts by researchers.

In this study, we conducted a systematic investigation of the phase-slips behavior in high-quality NbN SC nanowires, which were fabricated on highly disordered NbN films using a state-of-the-art photo-lithography process. We carefully identified the boundaries between single phase slip and multiple phase slips, as well as those between thermally activated phase slips (TAPS) and QPS. Multiple QPS events were uncovered in the low temperature region in samples with width below 40 nm. The signature of the cQPS effect is observed in this region in terms of a clear critical-voltage feature in the I–V curves without a high-impedance environment.

2. Materials and characterization

2.1. NbN nanowire fabrication

High uniformity of SC nanowires is a prerequisite for investigations of the QPS effect. In order to ensure the homogeneity of the NbN nanowires, NbN film deposition and nanowire pattern transformation were placed as the first and last steps of

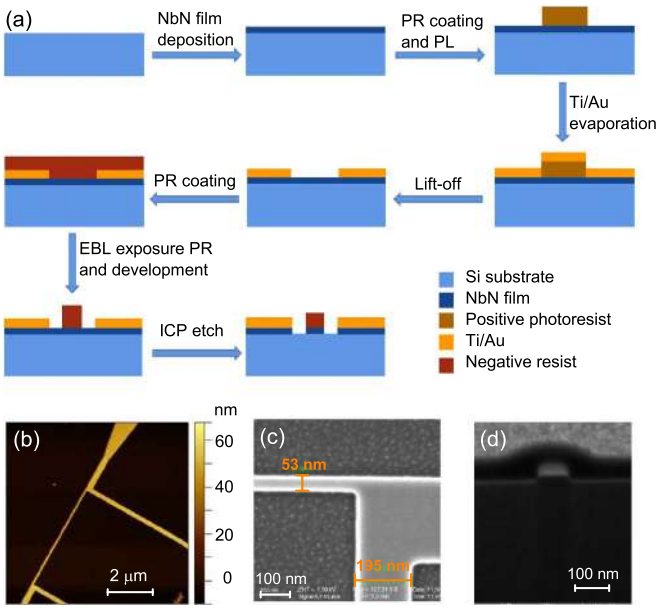


Figure 1. (a) Schematic representation of the fabrication process of NbN superconducting nanowires. (b) AFM image of one typical NbN nanowire with the wire width of 60 nm. (c) SEM picture of one typical NbN nanowire with the wire width of 60 nm. (d) SEM picture of the cross section of one sample with the wire width of 80 nm.

the whole fabrication process. Firstly, we deposited the high-quality NbN films with thickness of 9 nm on Si substrates by dc reactive magnetron sputtering. Details of the film growth and characterization have been reported previously [24]. In order to increase the normal-state resistance of the nanowires and thus obtain a high probability of the occurrence of QPS [8], we use Si substrates which do not match the NbN lattice and a low sputtering temperature (room temperature) to guarantee a high degree of disorder in the NbN SC films. The obtained NbN films show a rather high normal-state sheet resistance of up to 2.5 kΩ.

Secondly, the electrodes used for the transport measurements were patterned using a direct writing photo-lithography (PL) machine (DWL66+) and the metal Ti/Au was evaporated using a composite evaporation system (Ei-5, ULVAC), followed by a standard lift-off process. These structures can also be used as hard masks when defining nanowires in the following steps. The nanowires were then patterned by exposure of a high-resolution negative photoresist (PR) through electron beam lithography (EBL, JEOL-6300FS). Finally, the nanowire pattern was transformed into the NbN film by using the inductively coupled plasma etching machine (ICP, PlasmaPro100, Oxford) with a CF_4 plasma. For samples used for electrical transport measurements, the residual resist after etching was not removed to protect the surface of the nanowire from further oxidation, and all samples were stored in a high-purity Ar environment. The complete fabrication process is shown in figure 1(a).

In order to improve the edge roughness of the nanowires, the thinner inorganic negative-tone hydrogen silsesquioxane

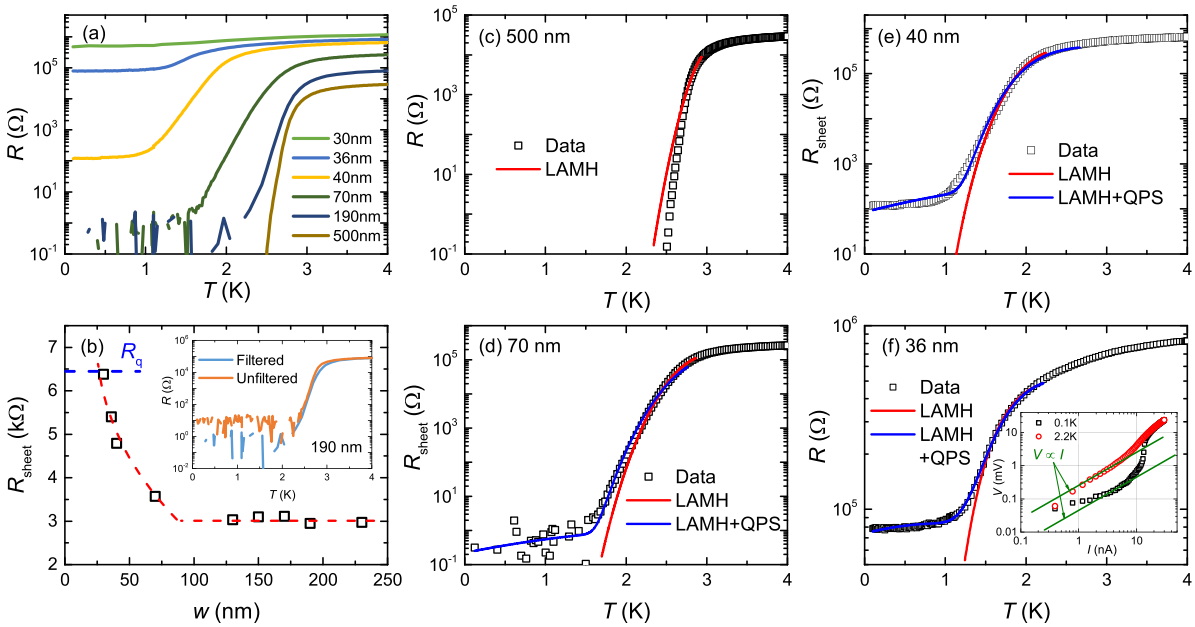


Figure 2. (a) Temperature dependence of resistance of NbN nanowires with different wire widths. (b) The value of normal-state sheet resistance R_{sheet} as a function of w. The red dashed line is a guide for eyes. The inset shows the comparison of the data of one sample with $w = 190$ nm with and without RC filter equipped. (c)–(f) Comparison of the resistance data with the fitting curves based on two different scenarios: pure LAMH (red) and LAMH+QPS (blue). The inset of (f) shows the I–V curves of the sample with $w = 36$ nm in logarithmic coordinates. The olive solid lines represent the linear relation $V \propto I$.

(HSQ) resist (XR-1541.002, Dow Corning) with a higher exposure dose (4700 μC cm^{-2}) was used to improve the adhesion and etching resistance. In addition, a higher development temperature of 50 ^{o}C was adopted to increase the contrast. These measures ensured that the nanowires we fabricated had a fine edge steepness and morphology.

2.2. Characterization

The morphologies and dimensions of the NbN nanowires were measured using an atomic force microscope (AFM, Bruker Dimension Icon) and a scanning electron microscope (SEM, Zeiss Gemini300). Electrical transport measurements were performed with the physical property measurement system (PPMS, Quantum Design) equipped with a dilution refrigerator (DR). The measurements were carried out under the current-biased condition. Current was applied along the longitudinal direction of the NbN nanowires during the transport measurements. The current was 5∼50 nA depending on the sample widths. RC filters were used to filter the external radiation effectively during the electrical transport measurements [25, 26]. As can be seen in the inset of figure 2(b), the use of RC filters can reduce the value of the resistance by an order of magnitude at ultra-low temperatures.

Before carrying out the transport investigations, the morphologies and dimensions of the NbN nanowires were characterized. As an example, we show the AFM and SEM images of one sample with a width of 60 nm in figures 1(b) and (c), respectively. It can be seen that the shape and size of the nanowires are uniform and well-defined. The wire width determined from the SEM picture is 53 nm, which is very close to the designed value of 60 nm. A wider part with a width of 195 nm was used as the electrode for electrical transport measurements. In addition, the cross sections of the nanowires were also checked by SEM, see figure 1(d) as an example for one sample with a width of 80 nm. Very steep sides of the nanowire can be seen. It is worth pointing out that the dimension in the thickness direction includes the thickness of the unremoved photoresist on top of the nanowire, the thickness of the NbN nanowire, and the depth of over-etching of the substrate.

3. Results and discussion

3.1. Resistance data and theoretical simulation

The electrical resistance R under zero field as a function of temperature for the six samples with different widths is shown in figure 2(a). With the decrease in wire width w, the resistance data in both the SC and normal states reveal a continuous evolution. For the samples with $w = 40$ nm, the residual resistance at ultra-low temperatures can be of the order of 100 Ω, which is enhanced to near 10^5Ω for the sample with $w =$ 36 nm. As the width further decreases to 30 nm, the SC transition becomes almost invisible. In figure 2(b), we show the values of the normal-state sheet resistance R_{sheet} as a function of w. In the range with $w > 70$ nm, R_{sheet} remains unchanged with the changing of wire width. In the region of $w < 70$ nm, the sheet resistance reveals a quick upturn with the decrease of w, suggesting that some new effect is likely to occur in this

region. It is notable that the disappearance of the SC transition corresponds precisely to the position where its normal-state sheet resistance is close to the quantum resistance for Cooper pairs $R_q = 6.45$ kΩ (see the blue dashed line in figure 2(b)). This indicates that there is a close relationship between the SC behavior and the normal-state sheet resistance, and when R_{sheet} approaches R_q, the system undergoes a transition from an SC to non-SC state. Such a behavior is consistent with the previous report in MoGe nanowires [12].

It is worth noting that for very narrow samples, the resistance data will be affected by the magnitude of the applied current. For example, the switch current of a sample with $w = 36$ nm is only about 13 nA (as shown in figure 4(c)), and in principle, we should use a very small measurement current. However, as will be elaborated in section 3.3, there is a critical-voltage characteristic caused by cQPS at ultra-low temperatures, which leads to a significant deviation from the linear behavior in the *I*–*V* curves in the small current region, as illustrated in the inset of figure 2(f). In the present study, we adopted a current of 5 nA, which is located in the roughly linear region on the *I*–*V* curves, to measure the sample with $w = 36$ nm.

It is known that the resistive behavior of SC nanowires at finite temperatures below T_c is governed by TAPS. In order to quantitatively analyze the resistance-temperature data, we fitted the data using the Langer–Ambegaokar–McCumber–Halperin (LAMH) model [6, 7, 15], which predicts the resistance behavior due to the TAPS:

$$R_{\text{LAMH}} = \frac{\pi\hbar^2\Omega}{2e^2k_BT}\exp\left(-\Delta F/k_BT\right), \tag{1}$$

where k_B is Boltzmann's constant, Ω is the attempt frequency of the phase slips, and ΔF is an activation energy for a phase slip. It is notable that the values of Ω and ΔF are closely related to the cross-sectional area, sample length L, coherence length ξ, and thermodynamic critical field of the nanowires. The detailed expressions of Ω and ΔF can be seen in [15, 16]. The total resistance R_{total} of the nanowires can be expressed as

$$R_{\text{total}}^{-1} = R_{\text{LAMH}}^{-1} + R_N^{-1}. \tag{2}$$

Here R_N is the normal-state resistance. The fitting results are displayed by the red lines in figures 2(c)–(f). For samples with a width of 500 nm, over a wide temperature range, the experimental data showed a steeper transition than that expected from the LAMH model, indicating that the TAPS is not important in this wide sample. While for the samples with $w = 70$ nm, 40 nm, and 36 nm, the LAMH model can only describe the resistance data near T_c. Meanwhile, a clear tail shape departing from the LAMH tendency emerges in the low temperature region.

In order to describe the contribution to the resistance from the QPS effect, R_{QPS}, the term k_BT in equation (1) was replaced by the energy scale $\hbar/\tau_{\text{GL}}$, which is related to the quantum process [17]. Here τ_{GL} is the characteristic relaxation time in the time-dependent Ginzburg–Landau (GL) theory. In this case, the total resistance can be expressed as

$$R_{\text{total}}^{-1} = \left(R_{\text{LAMH}} + R_{\text{QPS}}\right)^{-1} + R_N^{-1}. \tag{3}$$

During the fitting process, the normal-state resistance R_N, SC critical temperature T_c, and the zero-temperature coherence length ξ_0 are the most important parameters. R_N is fixed to the value determined from the resistance data, while T_c and ξ_0 are treated as fitting parameters. As shown by the blue lines in figures 2(d)–(f), the fitting curves considering both the TAPS and QPS effects simultaneously can give a rather fine description to the experimental data, which provides preliminary evidence for the existence of the QPS effect in the present system. The values of important parameters during the fitting process using equation (3) are listed in table 1.

Table 1. Important parameters during the fitting process using equation (3).

w (nm)	R_N (kΩ)	T_c(K)	L/ξ_0	ξ_0 (nm)
70	306	4.36	826.7	7.26
40	715	4.08	907.8	6.61
36	901	3.98	701.5	8.55

3.2. Phase diagram of phase slip(s)

Current–voltage (*I*–*V*) curves can provide another dimension to detect phase-slip effects. In figure 3(a), we show the *I*–*V* curves of the sample with $w = 40$ nm. We performed the *I*–*V* sweeps 60 times at a fixed temperature $T = 50$ mK. The *I*–*V* curves at low temperature present a steep transition from the zero-resistance to the normal state at a specific switch current point I_{sw} and show an obvious hysteresis behavior. This is a typical feature of the single phase-slip effect, which means that only one phase slip (QPS or TAPS) event can destroy the zero-resistance state at $I = I_{sw}$ [1]. The most notable feature is that the positions of these steep transitions I_{sw} are characterized by irregular fluctuations. This actually reflects the stochastic characteristic of the single phase-slip event [1, 27]. In figure 2(b), we show the histogram of the distribution of the values of I_{sw}.

Next, we study the evolution of the *I*–*V* curves with increasing temperature. As can be seen in figures 4(a)–(f), we show the *I*–*V* curves of three samples with different widths. For the sample with $w = 130$ nm, the *I*–*V* curves at low temperatures present a steep transition, indicating the occurrence of a single phase-slip effect. As the temperature rises above 2.2 K, the transition on the *I*–*V* curves becomes smooth, and the hysteresis feature is inconspicuous (see figures 4(a) and (d)). This indicates that the transition in this temperature region is caused by the collective effect of multiple phase-slip events. It is notable that detecting individual phase-slip event is not possible using the DC measurement setup in the multiple events region, since a single phase-slip event causes a very small voltage change during a very short time [1]. The sample with a width of 40 nm exhibits similar behavior, except that the transition from single phase slip to multiple phase slips occurs at lower

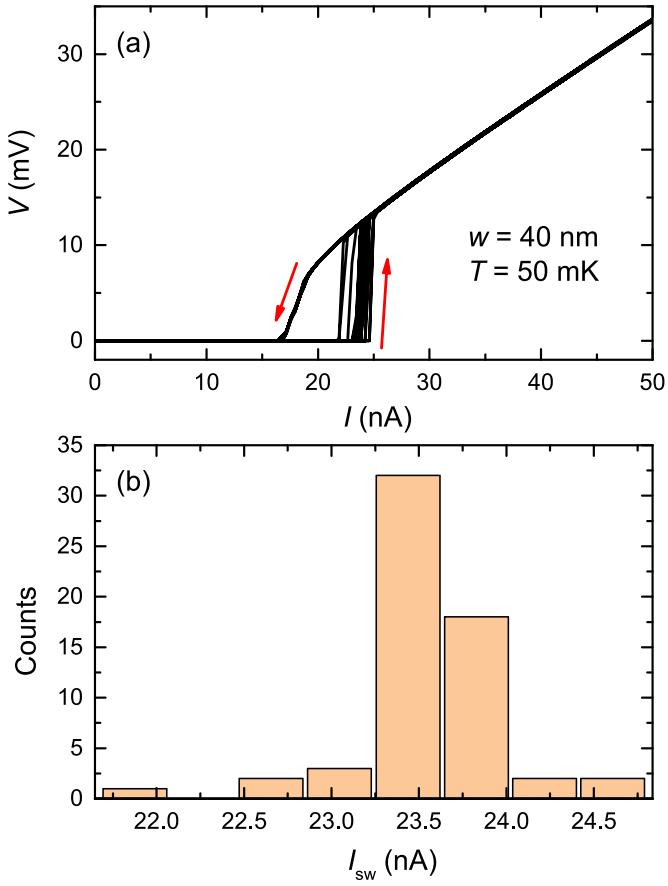


Figure 3. (a) I–V curves for the sample with $w = 40$ nm at 50 mK. The measurements were performed for 60 times at the same condition. The red arrows indicate the directions of sweeping current. (b) The distribution of I_{sw} at 50 mK.

temperatures. For the sample with a width of 36 nm, however, even at temperature as low as 0.1 K, there is no single phase-slip behavior which is characterized by a steep transition in the I–V curves. This evolutionary trend is reasonable since it reflects the fact that the probability of phase slip increases with the decrease in the cross-sectional area of the nanowires [8].

In figures 4(g)–(i), we show the colormaps of the resistance in the T-I space. Here, the value of $\tilde{R}$ is defined as $\tilde{R} = dV/dI$. The existence of a zero-/low-resistance region, normal-state region, and high-resistance transition region (where the transition occurs) can be clearly seen. Compared with these three samples, the difference is that for the wider sample ($w =$ 130 nm), the boundary from the low-resistance region to the normal-state region (the wine region) can extend to the low-current region that is very close to the vertical axis; for the other two narrow samples ($w = 40$ nm and 36 nm), this boundary exists only in a smaller and more limited region. It is also worth noting that, for the sample with $w = 130$ nm, the value of $\tilde{R}$ in the low-temperature region basically returns to the value of R_N on the right side of the boundary. However, the samples with $w = 40$ nm and 36 nm still show significantly higher $\tilde{R}$ values than $\tilde{R}_N$ in the normal-state region at low temperatures. This behavior is particularly obvious in the sample with $w =$ 36 nm.

The temperature dependence of the switch current I_{sw} of four samples is shown in figures 5(a)–(d). It can be seen that, above a certain point of temperature T_q (see the blue arrows), the value of I_{sw} reveals a continuous decrease with the increase in temperature. Below T_q, I_{sw} tends to be saturated with the decrease in temperature. In this temperature range, the fluctuations in the data stem from quantum stochasticity due to the single-phase slip events, as we have described earlier. Actually, T_q basically acts as a demarcation between the TAPS and QPS regions [28]. Although T_q is determined by examining the fluctuation behavior of I_{sw} (δI_{sw}) in [28], comparing figures 2(b) and (c) in this literature, it can be seen that the deviation of δI_{sw}-T data from the TAPS behavior corresponds to the position where I_{sw} becomes saturated. We fitted the evolution of I_{sw} with temperature for two samples with a larger TAPS range using equations based on mean-field GL theory [$I_{sw} \propto (1-t^2)^{3/2}(1+t^2)^{1/2}$, $t = T/T_c$], as shown by the red dashed lines in figures 5(a) and (b). There is an obvious deviation between the experimental data and the fitting curves in the QPS region below T_q. In the inset of figure 5(d), we show the values of T_q as a function of the SC critical temperature $T_{c,50\%}$. Here, $T_{c,50\%}$ is determined using the criterion of $50\%R_N$. It can be seen that T_q follows a roughly linear dependence with $T_{c,50\%}$, being consistent with the previous report and the theory of macroscopic quantum tunneling [27].

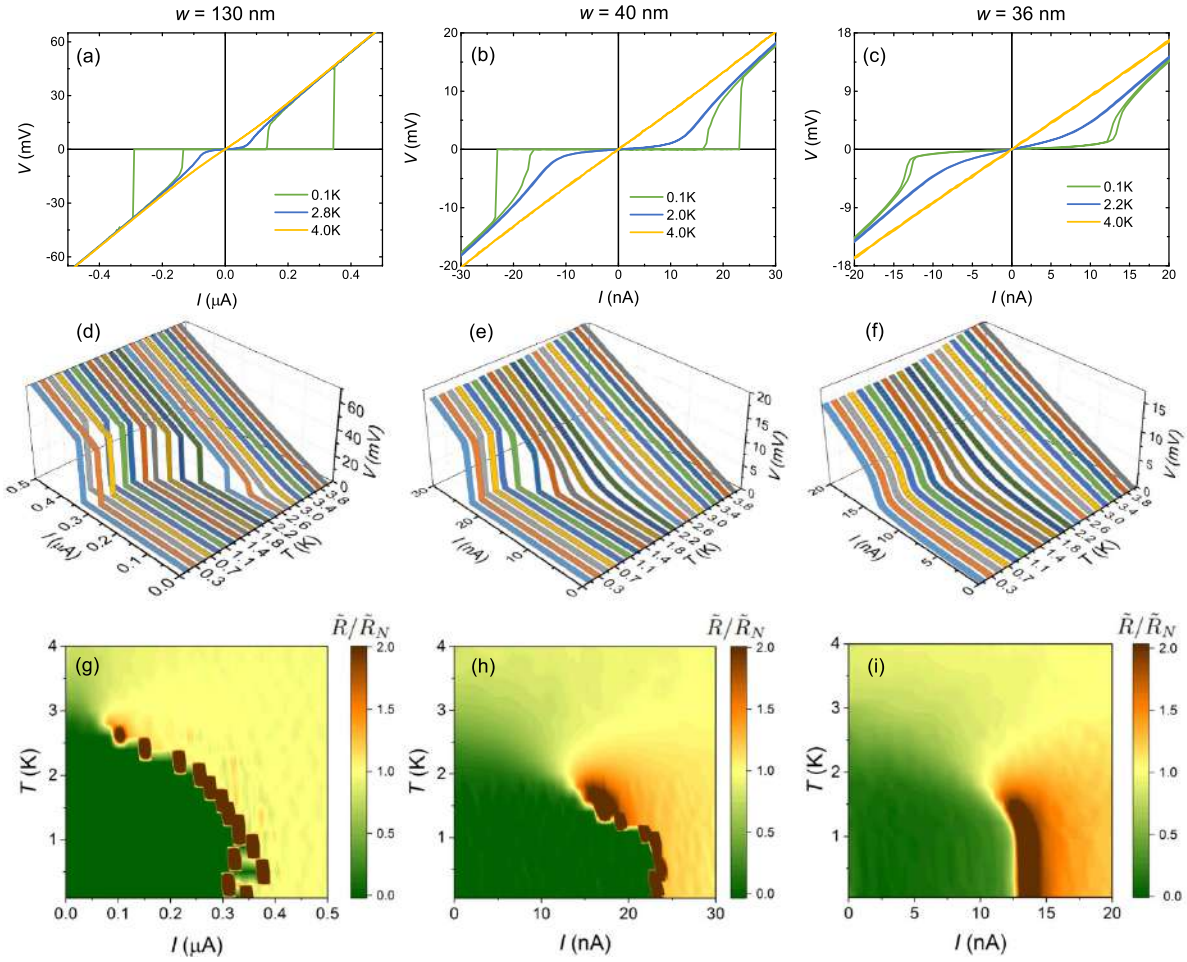


Figure 4. (a)–(c) I–V curves at selected temperatures for the three samples with the widths of 130 nm, 40 nm, and 36 nm, respectively. (d)–(f) 3D bar plot of the I–V data for the three samples. Only the data in the direction of increasing current is shown here. (g)–(i) Colormaps of resistance normalized by $\tilde{R}$ in the T-I space.

Based on the above experimental results, we obtain the phase diagram of the evolution of the phase-slip behavior with temperature and the width of nanowires, as shown in figure 6. As the temperature decreases, the system evolves from the normal state to the regions of multiple TAPS events, single TAPS events, and single QPS events. It is worth noting that as the width of the nanowire decreases, there is an evolution from single QPS to multiple QPS in the region with $w < 40$ nm at low temperatures. As we have stated, this reflects the increase in probability of phase slips with the decrease in the cross-sectional area of the nanowires.

3.3. Coherent QPS in ultrathin nanowires

As we have mentioned, it was proposed that nanowires with cQPS can be regarded as a charge-flux dual device to the famous Josephson junction, which has important potential application prospects in the quantum current standard [20, 21]. The feature of critical voltage V_c, below which the nanowire reveals a zero-conductance state, is a characteristic of cQPS [29]. We observed such a critical-voltage feature in two samples with a width of 36 nm. In figures 7(a) and (c), we show the I–V curves of these two samples at 50 mK in a wide current range. The two samples undergo a transition from a low-resistance region to the normal state at approximately 12 nA and 10 nA respectively, which is caused by the occurrence of multiple QPS events. Enlarging the gentle part of the I–V curve in the low current region, a clear critical-voltage feature can be seen, see figures 7(b) and (d). As indicated by the red arrows, the values of V_c are determined to be 100 μV and 400 μV for samples NW-1 and NW-2 respectively.

Theoretically, the value of V_c can be estimated from the phase-slip tunneling energy E_s [30]:

$$E_s = a\frac{L}{\xi}k_B T_c \frac{R_q}{R_\xi}\exp\left(-b\frac{R_q}{R_\xi}\right), \tag{4}$$

$$V_c = 2\pi E_s/2e, \tag{5}$$

where a and b are constants of order one, R_ξ is the resistance of the nanowire over one coherence length. It is worth pointing out that the exact values of a and b are not yet known. Here, we adopt a typical value of $a = 0.26$ [30]. Using equations (4) and (5), the values of b are calculated to be 1.40 and 1.28 for

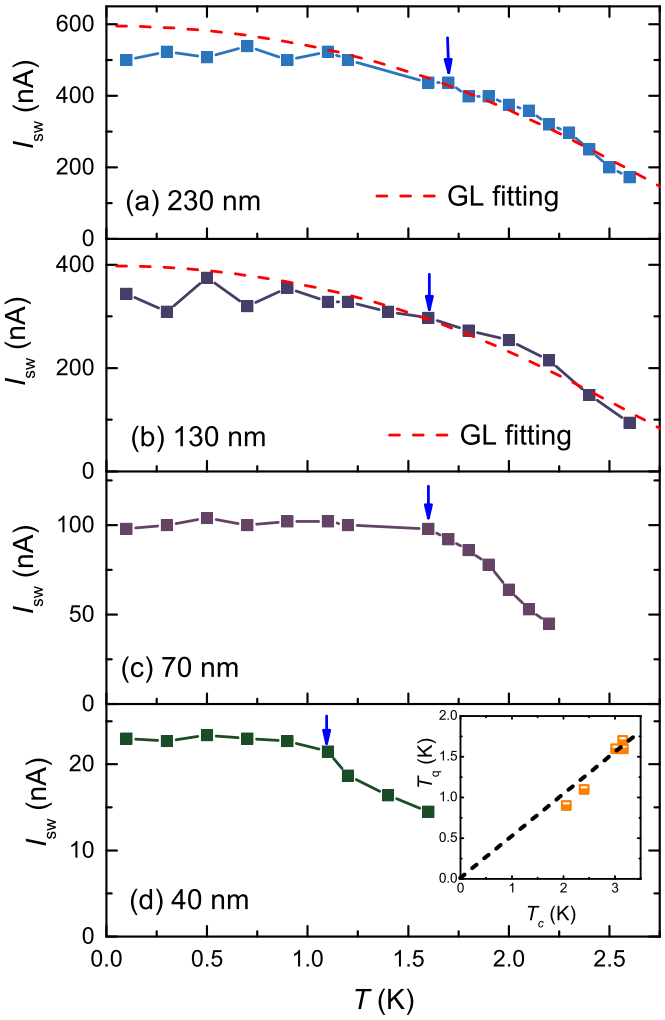


Figure 5. The data of switching current I_{sw} as a function of temperature for the nanowires with width of 230 nm (a), 130 nm (b), 70 nm (c), and 40 nm (d). The red dashed lines in (a) and (b) represent the fitting curves based on the mean-field Ginzburg–Landau (GL) theory. The inset of (d) shows the values of T_q as a function of $T_{c,50\%}$.

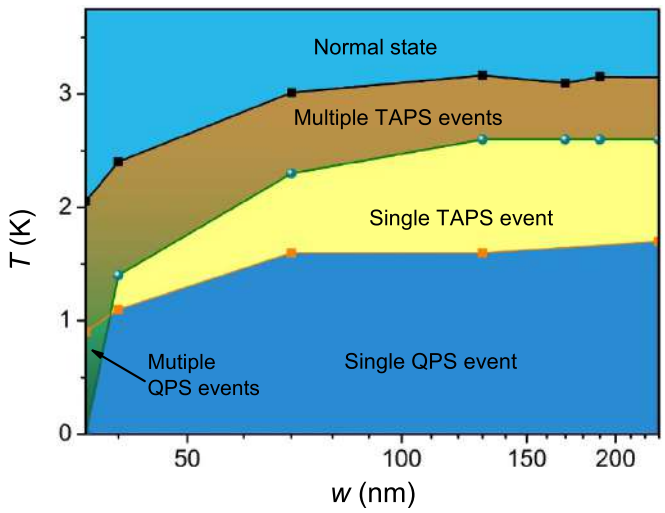


Figure 6. w-T phase diagram for the NbN nanowires. The boundary between normal-state and multiple TAPS events region is determined by $50\%R_N$ in the R-T data.

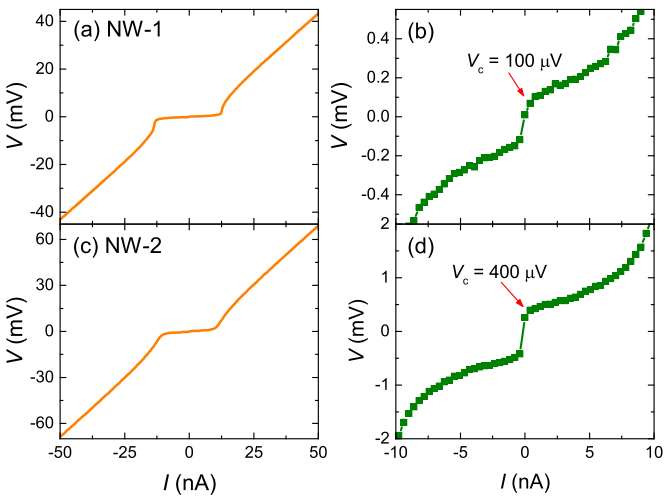


Figure 7. Observation of critical voltage V_c in two samples with $w = 36$ nm. (a), (c) I–V curves of samples NW-1 and NW-2 measured at 50 mK in a wide current range. (b), (d) The enlarged view of the same data in the small current range. The features of V_c can be seen in both samples.

samples NW-1 and NW-2 respectively. The order of magnitude of the obtained b values and the consistency between the two samples indicate the reasonability of our experimental results. The detailed parameters for the two samples are summarized in table 2.

Another noteworthy issue is the impedance environment in the measurement circuits. Due to the fact that charge and phase are dual variables, in order to ensure that the phase has significant uncertainty and achieves strong phase fluctuations, the charge needs to be a classical variable [21]. In experiments, large resistors are generally connected in series in the measurement circuit to reduce the charge fluctuations. Moreover, the introduction of large inductances can ensure a large enough charge relaxation time, so that the current remains stable during the time of measurement. According to previous reports, the critical-voltage behavior was usually observed in the high-impedance environment [23, 31–33]. For example, Arutyunov *et al* pointed out clearly [32] that the so-called Coulomb blockade, accompanied by a critical-voltage characteristic, was observed only after the introduction of high-impedance elements into the measurement circuit, but was not observed even in the thinnest nanowires under low-impedance conditions [34, 35]. It is worth noting that the critical-voltage feature is observed without involving the high-impedance environment in the present work. One possible explanation is that the large resistance values of the 36 nm-wide nanowires themselves in the QPS region may somehow play a role that is needed for high-impedance environments. This benefits from our successful regulation of the NbN thin films, which allows their normal-state resistivity to reach a very high level. The detailed physical mechanism deserves further in-depth investigations in the future.

Table 2. Important parameters of the two samples revealing coherent QPS.

Name	w (nm)	L (μm)	R_ξ (Ω)	V_c (μV)	b –
NW-1	36	6	910.5	100	1.40
NW-2	36	8	961.3	400	1.28

4. Conclusions

We have prepared high quality NbN SC nanowires with different widths and systematically studied the evolution of TAPS and QPS with the dimension of nanowires and temperature. In samples with widths above 40 nm, the system evolves from multiple TAPS to single TAPS and then to single QPS as the temperature decreases. At the low temperature limit, the system changes from a single QPS to multiple QPS scenario with the reduction of the dimension of the nanowires. More importantly, we observed the features of critical voltage, which is the characteristic of the coherent QPS effect, without introducing a high impedance circuit environment.

Data availability statement

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 92065116), the Shanghai Technology Innovation Action Plan Integrated Circuit Technology Support Program (No. 22DZ1100200), and the Key-Area Research and Development Program of Guangdong Province, China (No. 2020B0303030002). The authors would like to thank all the staff at the Superconducting Electronics Facility (SELF) for their assistance.

ORCID iDs

Wei Peng https://orcid.org/0000-0002-9888-1891
Gang Mu https://orcid.org/0000-0001-5676-4702

References

[1] Bezryadin A 2013 *Superconductivity in Nanowires* (Wiley)
[2] Golubev A D Z D S 2019 *Dissipative Quantum Mechanics of Nanostructures: Electron Transport, Fluctuations and Interactions* (Jenny Stanford Publishing)
[3] Varlamov A I L A A 2005 *Theory of Fluctuations in Superconductors* (Clarendon)
[4] Zaikin A G S A D 2022 *Phys. Usp.* **65** 883
[5] Little W A 1967 *Phys. Rev.* **156** 396
[6] Langer J S L V and Ambegaokar V 1967 *Phys. Rev.* **164** 498
[7] McCumber D E M B I and Halperin B I 1970 *Phys. Rev.* B **1** 1054
[8] Arutyunov K, Golubev D and Zaikin A 2008 *Phys. Rep.* **464** 1
[9] Zaikin A D, Golubev D S, van Otterlo A and Zimányi G T 1997 *Phys. Rev. Lett.* **78** 1552
[10] Golubev D S G A D and Zaikin A D 2001 *Phys. Rev.* B **64** 014504
[11] Arutyunov K Y, Lehtinen J S, Radkevich A, Semenov A G and Zaikin A D 2021 *Commun. Phys.* **4** 146
[12] Bezryadin A, Lau C N and Tinkham M 2000 *Nature* **404** 971
[13] Altomare F, Chang A M, Melloch M R, Hong Y and Tu C W 2006 *Phys. Rev. Lett.* **97** 017001
[14] Sahu M, Bae M -H, Rogachev A, Pekker D, Wei T-C, Shah N, Goldbart P M and Bezryadin A 2009 *Nat. Phys.* **5** 503
[15] Giordano N 1988 *Phys. Rev. Lett.* **61** 2137
[16] Sharifi F, Herzog A V and Dynes R C 1993 *Phys. Rev. Lett.* **71** 428
[17] Lau C N, Markovic N, Bockrath M, Bezryadin A and Tinkham M 2001 *Phys. Rev. Lett.* **87** 217003
[18] Rogachev A, Bollinger A T and Bezryadin A 2005 *Phys. Rev. Lett.* **94** 017004
[19] Ning W *et al* 2015 *Nano Lett.* **15** 869
[20] Mooij J E Harmans C J P M 2005 *New J. Phys.* **7** 219
[21] Mooij J E Nazarov Y V 2006 *Nat. Phys.* **2** 169
[22] Astafiev O V, Ioffe L B, Kafanov S, Pashkin Y A, Arutyunov K Y, Shahar D, Cohen O and Tsai J S 2012 *Nature* **484** 355
[23] Shaikhaidarov R S, Kim K H, Dunstan J W, Antonov I V, Linzen S, Ziegler M, Golubev D S, Antonov V N, Ilichev E V and Astafiev O V 2022 *Nature* **608** 45
[24] Wang X *et al* 2023 *Physica* C **606** 1354223
[25] Wang T, Yu A, Mao Y, Fang Y, Liu Y, Zhang H, Guo W, Li W, Peng W and Di Z 2023 *Supercond. Sci. Technol.* **36** 125015
[26] Xing Y *et al* 2021 *Nano Lett.* **21** 7486
[27] Aref T, Levchenko A, Vakaryuk V and Bezryadin A 2012 *Phys. Rev.* B **86** 024507
[28] Li P, Wu P M, Bomze Y, Borzenets I V, Finkelstein G and Chang A M 2011 *Phys. Rev. Lett.* **107** 137004
[29] Constantino N G N, Anwar M S, Kennedy O W, Dang M, Warburton P A and Fenton J C 2018 *Nanomaterials* **8** 442
[30] Mooij J E, Schön G, Shnirman A, Fuse T, Harmans C J P M, Rotzinger H and Verbruggen A H 2015 *New J. Phys.* **17** 033006
[31] Webster C H, Fenton J C, Hongisto T T, Giblin S P, Zorin A B and Warburton P A 2013 *Phys. Rev.* B **87** 144510
[32] Arutyunov K Y, Lehtinen J S and Rantala T 2016 *J. Supercond. Nov. Magn.* **29** 569
[33] Hongisto T T H A B and Zorin A B 2012 *Phys. Rev. Lett.* **108** 097001
[34] Lehtinen J S, Sajavaara T, Arutyunov K Y, Presnjakov M Y and Vasiliev A L 2012 *Phys. Rev.* B **85** 094508
[35] Lehtinen J S and Arutyunov K Y 2012 *Supercond. Sci. Technol.* **25** 124007

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名： 王一诺

方向： 量子科技

单位： 中国科学院大学前沿交叉科学学院

题目： 面向量子精密传感的单模单偏振VCSEL器件研究

High Single-Mode Power 852-nm Two-Junction Cascade VCSELs With Stable Polarization for Quantum Sensing

Yinuo Wang, Meng Xun, *Member, IEEE*, Guanzhong Pan, *Member, IEEE*, Yun Sun, Xiaoli Zhou, Jiaye Zhan, Bingxin Yao, Yongming Fei, Runze Zhang, Xing Li, and Dexin Wu

***Abstract*— We designed and fabricated an 852-nm two-junction cascade vertical cavity surface emitting laser (VCSEL). A high single-mode output power of 6.1 mW is obtained under continuous-wave operation by introducing a stepped oxide structure and an extended cavity. Stable single polarization is achieved by introducing elliptical oxide apertures resulting in a maximum orthogonal polarization suppression ratio (OPSR) of 25.7 dB. The high single-mode output power is fully optimized through epitaxial structure design, which remains compatible with conventional VCSEL fabrication processes without additional steps. This device shows potential for integration in Cs-based quantum sensors.**

***Index Terms*— Vertical cavity surface emitting lasers, single-mode VCSELs, 852 nm, high single-mode power, stable polarization, quantum sensing.**

I. Introduction

WITH the development of global navigation satellite systems, biomagnetic signal detection and other applications, quantum sensors are evolving towards miniaturization and low power-consumption [1], [2]. Single-mode VCSELs have emerged as the preferred light source for small and low power-consumption chip-scale atomic clocks, quantum gyroscopes, and quantum magnetometers by the virtue of low power consumption, circular beam profile, low cost, and easy integration [3], [4], [5], [6]. Quantum sensing systems rely on a high alkali metal atomic polarization rate for a superior signal-to-noise ratio. The output power and stable polarization of the light source is critical for achieving the high polarization rate [4]. This requires the VCSELs to achieve high single-mode output power and stable polarization. Conventional VCSELs are typically multiple transverse modes due to their relatively large lateral dimensions. Several strategies have been proposed to control optical mode [7], [8], [9], [10], [11], [12], including photonic crystal structures, surface relief, lithographic VCSELs, and so on. Although these methods improve mode properties in larger aperture devices, they often suffer from drawbacks such as complex fabrication processes, high accuracy requirements, and elevated production costs.

Recently, multi-junction cascade technology has been reported to improve output power and power conversion efficiency (PCE) [13], [14]. A maximum efficiency of 74% in 15-junction VCSELs has been reported at room temperature under nanosecond driving current [13]. Multi-junction VCSELs typically utilize multiple oxide apertures to confine injection current. However, these oxide layers introduce strong refractive index waveguide that increase the number of higher order modes and thereby degrade single mode operation. Additional measures must be taken to restrict higher-order modes, such as surface relief on the topmost of multi-junction VCSELs [9], [14]. Our group has also demonstrated achieving high single-mode power in multi-junction VCSELs by combining single-oxide-layer and extended cavity. However, the lack of effective current confinement for multi-active-region injection currents by the single oxide layer results in excessive degradation of power conversion efficiency [15].

In this letter, an 852-nm two-junction VCSEL is designed and fabricated. Stepped oxide layers are specifically engineered to mitigate strong optical confinement. An extended cavity structure is incorporated to suppress higher-order modes. To achieve polarization control, elliptical oxide apertures are implemented to induce anisotropic lateral confinement. The fabricated device with a 3 μm $\times$ 5 μm oxide aperture achieves a maximum single-mode output power of 6.1 mW and a maximum OPSR of 25.7 dB. The proposed technique maintains fabrication simplicity while proactively facilitating mass production with scaled-up volumes.

II. Device Design and Fabrication

The structural schematic of the 852-nm VCSEL device is shown in Fig. 1(a). The epitaxial layers are sequentially grown on a 6-inch GaAs substrate by metal-organic chemical vapor deposition (MOCVD). The bottom and top mirrors consist of 34 pairs of n-doped and 17 pairs of p-doped $Al_{0.12}Ga_{0.88}As/Al_{0.9}Ga_{0.1}As$ distributed bragg reflectors, respectively. The two-junction cascade VCSEL incorporates two active regions connected by a heavily doped

Received 15 July 2025; accepted 22 July 2025. Date of publication 28 July 2025; date of current version 29 September 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62374178, Grant 62174178, and Grant 62374179; in part by the Youth Innovation Promotion Association Chinese Academy of Sciences (CAS) under Grant 2022115; in part by Beijing Nova Program under Grant 2022032 and Grant 20240484582; and in part by the National Key Basic Research Program under Grant 2022YFB2803000. The review of this letter was arranged by Editor T.-Y. Seong. *(Corresponding authors: Meng Xun; Guanzhong Pan.)*

Yinuo Wang and Xiaoli Zhou are with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China, and also with the School of Advanced Interdisciplinary Science, University of Chinese Academy of Sciences, Beijing 100049, China.

Meng Xun, Guanzhong Pan, Yun Sun, Jiaye Zhan, Bingxin Yao, Yongming Fei, Runze Zhang, Xing Li, and Dexin Wu are with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China (e-mail: xunmeng@ime.ac.cn).

Digital Object Identifier 10.1109/LED.2025.3593255

1558-0563 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
Authorized licensed use limited to: Shenzhen University of Advanced Technology. Downloaded on September 29,2025 at 08:58:47 UTC from IEEE Xplore. Restrictions apply.
See https://www.ieee.org/publications/rights/index.html for more information.

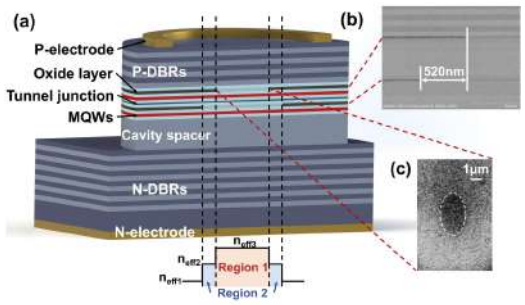


Fig. 1. (a) Schematic view of the two-junction cascade VCSEL. (b) The SEM image of the stepped oxide layers. (c) The infrared microscope image of the elliptical oxide aperture.

GaAs-based tunnel junction. The n-type and p-type doping concentrations are 2×10^{19} cm^{-3} and 1×10^{20} cm^{-3}, respectively. And a 1.5-μm-thick $Al_{0.12}Ga_{0.88}As$ spacer layer is introduced to increase the diffraction loss of the higher-order modes [14]. And this layer is lightly doped to minimize absorption. The two oxide apertures are designed with different aluminum compositions to have distinct oxidation rates, enabling the formation of a stepped oxide structure through a single-step wet oxidation process. The stepped oxide layers generate a graded effective refractive index profile.

A scanning electron microscope (SEM) image of the stepped oxide aperture is shown in Fig. 1(b). Due to the smaller size of the top oxide aperture compared to the bottom one, the top oxide layer dominates the optical confinement effect. The laser radiation is primarily confined to the region 1, where the refractive index contrast between the core and adjacent cladding remains unchanged, closely resembling that of a single oxide layer in the single junction VCSEL. In region 2, minor carrier lateral spreading generates weak laser emission that excites low-intensity higher-order modes. Such modes experience substantial diffraction loss in long-cavity structures and are consequently suppressed. Stable polarization is achieved by elliptical oxide apertures. Although surface grating methods can achieve high polarization suppression ratio [12], they typically require precise e-beam lithography and etching, thereby increasing process complexity/cost and degrading output optical power. In contrast, our device utilizes elliptical oxide apertures to stabilize polarization, while maintaining high compatibility with standard VCSEL fabrication processes. An infrared microscope image of the elliptical oxide aperture is shown in Fig. 1(c).

The device is manufactured on a 6-inch epitaxial wafer. This design requires no additional process steps and its fabrication process is identical to conventional oxide VCSELs. Details of the fabrication process can be found in our previous letter [15]. The size tolerance between upper/lower oxide layers across 6-inch epitaxial wafers shows outstanding repeatability, which results from uniform epitaxial growth and fabrication process. Devices are fabricated with two oxide aperture sizes (upper oxide layer) of 3 μm $\times$ 5 μm and 4 μm $\times$ 6 μm, respectively.

III. Results and Discussion

Fig. 2 shows the L–I–V and optical spectrum characteristics under continuous-wave (CW) operation of the two-junction cascade VCSELs. The 3 μm$\times$5 μm device exhibits a threshold current of 0.8 mA, a slope efficiency (SE) of 2.13 W/A and

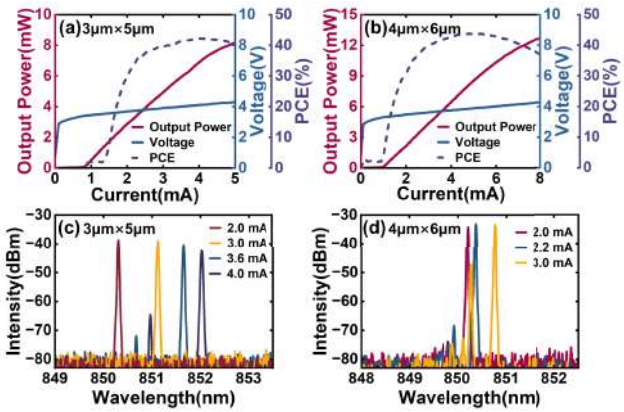


Fig. 2. L-I-V and PCE characteristics of (a) the 3 μm $\times$ 5 μm device and (b) the 4 μm $\times$ 6 μm device. Optical spectrum characteristics of (c) the 3 μm $\times$ 5 μm device and (d) the 4 μm $\times$ 6 μm device.

a maximum PCE of 42.2%. The 4 μm$\times$ 6 μm device shows a slightly higher threshold current of 1 mA, a SE of 2.1 W/A and a maximum PCE of 43.8% under the same operating conditions. For the 3 μm $\times$ 5 μm VCSEL, single-mode emission is maintained up to 3.6 mA. Only a single peak is observed in the spectrum below 3.6 mA. The maximum side-mode suppression ratio (SMSR) is achieved 40 dB at 2 mA. Although a minor side peak emerges at 3.6 mA, the SMSR maintains 31.5 dB, indicating robust single-mode dominance. And the device exhibits a maximum single-mode output power of 6.1mW. The SMSR gradually decreases as the current increased beyond 3.6 mA. The SMSR drops to 24.2 dB when the current is up to 4 mA, which indicates the device is under multimode operation. The 4 μm $\times$ 6 μm device exhibits a maximum SMSR of 39 dB at 2 mA. This device achieves a maximum single-mode output power of 2.59 mW at 2.2 mA with an SMSR of 35 dB. The multiple peaks appear in the spectrum at 3 mA, which confirms the onset of multimode lasing. Due to the reduction in single-mode range, the single-mode output power does not increase with larger sizes.

The superior single-mode performance of the device is primarily attributed to the stepped oxide aperture and the extended cavity. Owing to their larger divergence angles relative to the fundamental mode, higher-order modes experience significantly enhanced diffraction losses in extended-cavity configurations [16]. The stepped oxide aperture is specifically engineered to create a gradual effective index profile, forming a stepped index waveguide structure. Compared to conventional single-oxide designs, the stepped oxide aperture provides stronger electrical confinement without significantly increasing the optical index contrast. This design mitigates the proliferation of the higher-order mode, enabling a balance between output power and mode discrimination.

Fig. 3 shows the far-field distributions and divergence angles of the 3 μm $\times$ 5 μm device and the 4 μm $\times$ 6 μm device under different currents. Details of the setup for far-field measurement can be found in our previous letter [15]. Notably, the far-field patterns of the VCSELs with the elliptical oxide apertures exhibit circular symmetry. Up to a drive current of 3.6 mA, the far-field beam of the 3 μm $\times$ 5 μm device maintains a Gaussian intensity distribution, which confirms single fundamental mode emission. As the current increased

Authorized licensed use limited to: Shenzhen University of Advanced Technology. Downloaded on September 28,2025 at 08:58:47 UTC from IEEE Xplore. Restrictions apply.

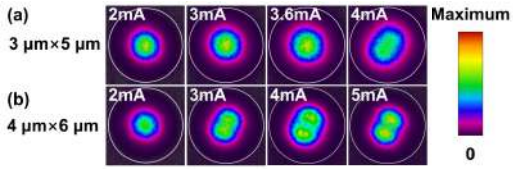


Fig. 3. Far-field beam profiles at different currents for (a) the 3 μm $\times$ 5 μm device and (b) the 4 μm $\times$ 6 μm device.

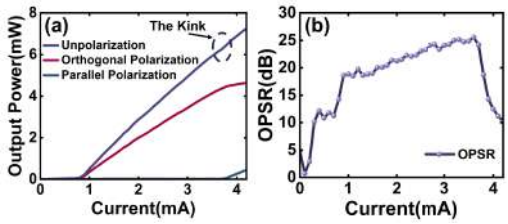


Fig. 4. (a) Polarization-resolved L-I characteristics and (b) OPSR versus injection current for the 3 μm $\times$ 5 μm device.

to 4 mA, the far-field pattern transitions from circular to elliptical, confirming multimode operation. The divergence angle is defined as the angle at which the beam intensity falls to $1/e^2$ of the peak value. Within the regime of single-mode operation, the divergence angle remains within the range of 10.4° to 14.7°. The 4 μm $\times$ 6 μm device exhibits non-Gaussian elliptical emission at 3 mA, confirming transverse multimode operation. As the current increases, the far-field patterns take on an elliptical shape. These observations are consistent with the spectrum characteristics shown in Fig. 2.

Fig. 4(a) shows the polarization-resolved L-I characteristics for the 3 μm $\times$ 5 μm device. Polarization-dependent L-I curves are measured using the following procedure: A linear polarizer is positioned between the VCSEL emitting surface and an optical power meter. The polarizer is continuously rotated to identify the angle corresponding to minimum optical power (indicating the parallel polarization), after which it is precisely rotated by 90° to align with the orthogonal polarization axis. Photometric intensity readings at both orientations are then synchronously recorded with the injection current sweeping. Fig. 4(b) illustrates the OPSR as a function of current for the same device. The OPSR is defined as OPSR $= 10 \times \log (P_{orth}/P_{par})$ [17]. As shown in Fig. 4(a), the polarization remains stable throughout the single-mode operation range. The OPSR exceeds 20 dB between 1.6 mA and 3.7 mA, reaching a maximum of 25.7 dB at 3.6 mA. When the current increases to 3.8 mA, the power in the parallel polarization increases while the orthogonal component decreases, directly correlating to the kink observes in the unpolarized L-I characteristic. These results demonstrate that the elliptical oxide aperture can provide effective and stable polarization control by suppressing polarization mode competition. The excellent polarization behavior is attributed to the anisotropic current confinement induced by the elliptical aperture which introduces asymmetric carrier injection. This asymmetry affects the carrier distribution and absorption in the active region, resulting in anisotropic optical gain and polarization selectivity [18], [19], [20]. From Fig. 4(b), it can be observed that above the threshold current, the OPSR slowly increases with the injection current. This may be attributed to that the short axis direction, due to its higher current density,

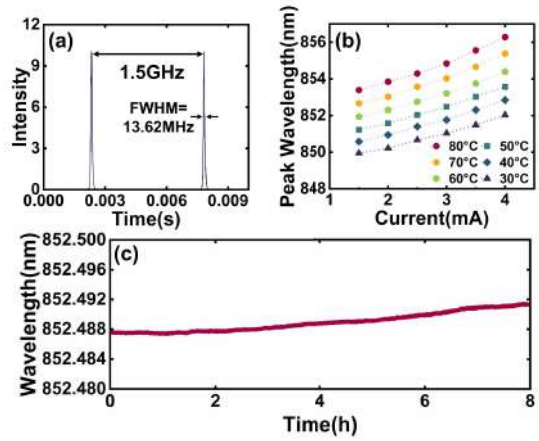


Fig. 5. (a) Spectral linewidth of the device. (b) Variation of the peak wavelength with current under different ambient temperature and (c) the wavelength stability for the 3 μm $\times$ 5 μm device.

experiences a more severe spatial hole burning effect, thereby weakening its gain capability.

Fig. 5(a) shows the spectral linewidth of the device with 3 μm $\times$ 5 μm oxide aperture measured by a scanning Fabry-Pérot interferometer (Thorlabs SA-200) under 3 mA at 25°C. A narrow linewidth of 13.62MHz is reached resulting from the elongated effective cavity length. For thermal behavior analysis, VCSELs were TO-46 packaged incorporating thermoelectric coolers (TECs) for precise temperature control. Fig. 5(b) shows the variation of the central wavelength with current under different ambient temperatures. The peak wavelength exhibits a redshift with increasing injection current, and notably, the redshift rate gradually rises at higher currents. The redshift rate exhibits stronger current dependence at elevated temperatures because of more severe heat accumulation at high temperatures. Fig. 5(c) shows the wavelength stability of the 3 μm $\times$ 5 μm device under 3mA at 50 °C. The peak wavelength is recorded via a wavelength meter (HighFinesse WS7). The result shows that the wavelength shift is about 0.004 nm during 8 hours, maintaining good wavelength stability. The devices undergo accelerated aging testing at a constant bias of 3mA at 100°C. No fail (i. e., no power drop of ≥ -1dB) are observed during 2300 hours of accelerated aging.

IV. Conclusion

In this letter, a two-junction cascade 852-nm VCSEL is demonstrated. Mode discrimination is achieved through a combination of an extended cavity structure and stepped oxide layers. The 3 μm $\times$ 5 μm device achieved a maximum single-mode output power of 6.1 mW and a maximum OPSR of 25.7 dB. The demonstrated 852-nm VCSEL featuring high-power, single-mode, and polarization-stable emission shows strong potential for integration into compact and low power-consumption Cs-based quantum sensors. All performance enhancements are realized through epitaxial design without modifications to the standard VCSEL fabrication process. This approach offers a scalable and fabrication-compatible solution for high-performance VCSELs. Both the stepped oxide and elliptical aperture present good reproducibility across single wafers. However, precise dimensional control for small oxide layer maybe a potential challenge for future high-volume manufacturing.

Authorized licensed use limited to: Shenzhen University of Advanced Technology. Downloaded on September 28,2025 at 08:58:47 UTC from IEEE Xplore. Restrictions apply.

References

[1] E. Batori, N. Almat, C. Affolderbach, and G. Mileti, "GNSS-grade space atomic frequency standards: Current status and ongoing developments," *Adv. Space Res.*, vol. 68, no. 12, pp. 4723–4733, Dec. 2021, doi: 10.1016/j.asr.2020.09.012.

[2] E. Boto, N. Holmes, J. Leggett, G. Roberts, V. Shah, S. S. Meyer, L. D. Muñoz, K. J. Mullinger, T. M. Tierney, S. Bestmann, G. R. Barnes, R. Bowtell, and M. J. Brookes, "Moving magnetoencephalography towards real-world applications with a wearable system," *Nature*, vol. 555, no. 7698, pp. 657–661, Mar. 2018, doi: 10.1038/nature26147.

[3] G. Pan, M. Xun, X. Zhou, Y. Sun, Y. Dong, and D. Wu, "Harnessing the capabilities of VCSELs: Unlocking the potential for advanced integrated photonic devices and systems," *Light, Sci. Appl.*, vol. 13, no. 1, p. 229, Sep. 2024, doi: 10.1038/s41377-024-01561-8.

[4] N. A. Maleev, S. A. Blokhin, M. A. Bobrov, A. G. Kuz'menkov, M. M. Kulagina, and V. M. Ustinov, "Laser source for a compact nuclear magnetic resonance gyroscope," *Gyroscopy Navigat.*, vol. 9, no. 3, pp. 177–182, Jul. 2018, doi: 10.1134/s2075108718030057.

[5] J. Park, H.-G. Hong, T. Y. Kwon, and J.-K. Lee, "Flexible hybrid approach for a 3D integrated physics package of chip-scale atomic clocks," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6839–6846, Mar. 2021, doi: 10.1109/JSEN.2020.3042703.

[6] S. Zhang, Y. Zhou, Z. Chai, Q. Liu, K. Yin, and B. Han, "Geometric-phase-lens collimated vertical-cavity surface-emitting laser turned on Rb D_1 line for miniature atomic magnetometers," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 4000307, doi: 10.1109/TIM.2022.3232658.

[7] Y. Xie, C. Xu, Q. Kan, C. Wang, H. Chen, and G. Shen, "High power single mode output low threshold current photonic crystal vertical cavity surface emitting lasers," in *Proc. OFC/NFOEC*, Feb. 2012, pp. 1–3, doi: 10.1109/LPT.2011.2179796.

[8] D. G. Deppe, J. Leshin, J. Leshin, L. Eifert, F. Tucker, and T. Hillyer, "Transverse mode confinement in lithographic VCSELs," *Electron. Lett.*, vol. 53, no. 24, pp. 1598–1600, Nov. 2017, doi: 10.1049/el.2017.2780.

[9] A. Ghods, M. Dummer, G. Xu, and K. Johnson, "Design and fabrication of single-mode multi-junction 905 nm VCSEL with integrated anti-phase mode filter," *J. Lightw. Technol.*, vol. 41, no. 10, pp. 3102–3107, May 15, 2023, doi: 10.1109/JLT.2023.3240123.

[10] Y. Zhou, Y. Jia, X. Zhang, J. Zhang, Z. Liu, Y. Ning, and L. Wang, "Large-aperture single-mode 795 nm VCSEL for chip-scale nuclear magnetic resonance gyroscope with an output power of 4.1 mW at 80°C," *Opt. Exp.*, vol. 30, no. 6, p. 8991, Mar. 2022, doi: 10.1364/oe.450118.

[11] Y. Wang, C. Li, Y. Zhang, J. Li, and X. Wei, "Mode optimization and polarization control in VCSELs with sub-wavelength gratings," *IEEE Photon. J.*, vol. 17, no. 4, pp. 1–6, Aug. 2025, doi: 10.1109/JPHOT.2025.3578660.

[12] L.-S. Fu, N. Cui, F. Zhang, J.-S. Wang, A. Qiu, B.-L. Guan, and A.-Q. Liu, "Subwavelength-grating coupled-cavity resonance VCSELs with ultra-narrow linewidth," *APL Photon.*, vol. 10, no. 5, May 2025, Art. no. 056112, doi: 10.1063/5.0254855.

[13] Y. Xiao, J. Wang, H. Liu, P. Miao, Y. Gou, Z. Zhang, G. Deng, and S. Zhou, "Multi-junction cascaded vertical-cavity surface-emitting laser with a high power conversion efficiency of 74%," *Light, Sci. Appl.*, vol. 13, no. 1, p. 60, Feb. 2024, doi: 10.1038/s41377-024-01403-7.

[14] Y. Xiao, P. Miao, J. Wang, H. Liu, Y. Gou, Z. Zhang, B. Wang, W. Liu, Q. Wang, G. Deng, and S. Zhou, "Twenty-milliwatt, high-power, high-efficiency, single-mode, multi-junction vertical-cavity surface-emitting lasers using surface microstructures," *Photon. Res.*, vol. 12, no. 9, p. 1899, Sep. 2024, doi: 10.1364/prj.521098.

[15] X. Zhou, M. Xun, G. Pan, Y. Sun, Y. Wang, C. Han, W. Wu, S. Wu, Y. Fei, and D. Wu, "High-power low-divergence single-mode multi-junction cascade VCSELs," *IEEE Trans. Electron Devices*, vol. 71, no. 9, pp. 5472–5476, Sep. 2024, doi: 10.1109/TED.2024.3423801.

[16] M. Xun, G. Pan, Z. Z. Zhao, Y. Sun, W. Wu, W. Jiang, R. Zhang, and D. Wu, "High single fundamental-mode output power from 795 nm VCSELs with a long monolithic cavity," *IEEE Electron Device Lett.*, vol. 44, no. 7, pp. 1144–1147, Jul. 2023, doi: 10.1109/LED.2023.3277837.

[17] P. Qiu, H. Zhou, T. Wang, Q. Wang, R. Zhang, and Q. Kan, "Polarization-stable single-mode 795 nm grating-coupled surface-emitting laser for quantum sensing," *Opt. Exp.*, vol. 31, no. 25, p. 42562, Dec. 2023, doi: 10.1364/oe.507348.

[18] G. Verschaffelt, W. van der Vleuten, M. Creusen, E. Smalbrugge, T. G. van de Roer, F. Karouta, R. C. Strijbos, J. Danckaert, I. Veretennicoff, B. Ryvkin, H. Thienpont, and G. A. Acket, "Polarization stabilization in vertical-cavity surface-emitting lasers through asymmetric current injection," *IEEE Photon. Technol. Lett.*, vol. 12, no. 8, pp. 945–947, Aug. 2000, doi: 10.1109/68.867969.

[19] A. Haglund, J. S. Gustavsson, J. Bengtsson, P. Jedrasik, and A. Larsson, "Design and evaluation of fundamental-mode and polarization-stabilized VCSELs with a subwavelength surface grating," *IEEE J. Quantum Electron.*, vol. 42, no. 3, pp. 231–240, Mar. 2006, doi: 10.1109/JQE.2005.863703.

[20] J. Zhang, W. Zheng, and A. Liu, "Polarized vertical-cavity surface-emitting laser arrays with asymmetrical structures for three-dimensional sensing with polarization imaging," *IEEE J. Quantum Electron.*, vol. 59, no. 5, pp. 1–8, Oct. 2023, doi: 10.1109/JQE.2023.3295399.

Authorized licensed use limited to: Shenzhen University of Advanced Technology. Downloaded on September 28,2025 at 08:58:47 UTC from IEEE Xplore. Restrictions apply.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：杨洁

方向：量子科技

单位：国家纳米科学中心

题目：利用电荷中性点处的拓扑平带增强催化活性

Enhanced catalytic activity by topological flat bands at the charge neutrality point

Yang Song,[1,*] Wen-Han Dong,[2,3,*] Jie Yang,[1] Ziyuan Liu,[1] Mengxi Liu,[4] Jia Zhu,[1] Shixuan Du,[3] Feng Liu,[5,†] and Lizhi Zhang[1,‡]

[1]*Laboratory of Theoretical and Computational Nanoscience, National Center for Nanoscience and Technology, Beijing 100190, China*
[2]*State Key Laboratory of Low Dimensional Quantum Physics and Department of Physics, Tsinghua University, Beijing 100084, China*
[3]*Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*
[4]*Key Laboratory of Standardization and Measurement for Nanotechnology, National Center for Nanoscience and Technology, Beijing 100190, China*
[5]*Department of Materials Science and Engineering, University of Utah, Salt Lake City, Utah 84112, United States*

(Received 16 May 2024; revised 12 November 2024; accepted 8 September 2025; published 1 October 2025)

The topological flat band (TFB) at the charge neutrality point (CNP) has attracted much recent interest for its physical properties in realizing exotic quantum states, while its implication on materials' chemical properties remains largely unknown. Here, we predict the TFB@CNP in two-dimensional covalent and metal organic frameworks (COFs and MOFs) can enhance catalysis. Using density-functional theory and tight-binding calculations, we show that in a class of COFs and MOFs underlined by a diatomic kagome lattice model, TFB@CNP can be designed by tuning the relative intra- and interkagome lattice hopping in correlation with electron filling. Remarkably, using hydrogen evolution reaction as an example, we illustrate that the catalytic activity of abundant active sites in COFs and MOFs can be significantly enhanced when they are involved in forming the TFB@CNP. Our findings open a promising direction that integrates topological physics with reaction chemistry in the new platform of high-performance TFB-enhanced catalysis.

DOI: 10.1103/56zz-7kwp

I. INTRODUCTION

The topological flat band (TFB) represents a unique class of delocalized electronic states beyond the highly localized d or f orbital states [1]. Its extremely narrow bandwidth gives rise to a wide range of intriguing strongly correlated physical phenomena [2–16], such as unconventional superconductivity [9,10] and excitonic Bose-Einstein condensation [16]. However, the implication of TFB on materials' chemical properties remains largely unexplored. In this work, we predict that the presence of TFB at the charge neutrality point (CNP) can significantly enhance a material's existing catalytic activity. The TFB@CNP contributes an extremely high electron density of states (DOS) around the Fermi level to facilitate the charge transfer process for chemical adsorption, which bears some similarity with high DOS of localized d orbitals, but the physical origin and working principle of TFB catalysis have some notable differences from the d-band catalysis [17]. The TFB is formed by destructive interference coded in the lattice and/or orbital symmetry of the whole crystal lattice [1,18] rather than the localized atomic orbitals. Therefore, it may not offer a specific catalytic site, but instead enhance the activity of existing catalytic sites that are involved in forming the TFB. In particular, the TFB may act as an electron reservoir during chemical reaction, changing electron occupation with a nearly fixed Fermi level, while a d-band center will shift position with different electron filling.

Recently, the topological electronic states have been proposed for catalysis [19–22], but the field of topological catalysis is still in its infancy. We note that physically the topological surface state (TSS) originates from bulk Bloch wave functions terminated at the surface via bulk-boundary correspondence. Thus the TSS is robust against surface adsorption adding an "impurity." Conversely, surface adsorption might not react catalytically with a TSS as it does with a trivial local d-orbital state in surface, but only modify slightly the TSS distribution in and around bulk band gap [23,24]. Also, the TSS typically consists of Dirac bands with a vanishing DOS at the Fermi level, unfavorable for catalysis. In these regards, our proposed TFB-enhanced catalysis differs significantly from the previously proposed topological catalysis based on the TSS [19–22].

In general, there are two categories of TFBs. One we call the unit-cell TFB [25,26], hosted in crystalline lattices with a few atoms per unit cell [27–29] and formed by destructive interference with an arbitrary value of nearest-neighbor (NN) hopping t [30]. The other one is the supercell TFB [8], formed by flattened Dirac bands with reduced hopping t and tuned interlayer interaction [31]. As an advantage, the supercell TFBs naturally exist at the CNP to be easily accessible, while the unit-cell TFBs are usually far away from the Fermi level. Partly for this reason, interesting properties of the supercell TFBs have been widely shown recently, including an enhancement of electrochemical reaction in the twisted bilayer graphene (TBG) [32], while experimental demonstration of

*These authors contributed equally to this work.
†Contact author: ftiger.liu@utah.edu
‡Contact author: zhanglz@nanoctr.cn

 ©2025 American Physical Society

the unit-cell TFBs [33] and their properties remains largely elusive. However, of interest here, the unit-cell TFB materials can have some advantages; e.g., they can possess much higher density of catalytic sites than the supercell ones, whose active sites exist only at the AA stacking regions in a very large moiré pattern (e.g., $\sim$170 nm^2 for the 1.1° TBG).

Covalent and metal organic frameworks (COFs and MOFs) have already been widely used as catalysts due to their large accessible area (porous structure), abundant active sites [34,35], and potentially high carrier mobilities [36–38]. Also, COFs and MOFs are well-known unit-cell materials to host TFBs because of their diverse sublattice symmetries [5,29,39–53], such as the well-known kagome lattices [5,45,46]. It is desirable to achieve the TFB@CNP because the conductive COFs and MOFs are usually endowed with a better catalytic activity [54,55]. But, as mentioned above, the unit-cell TFBs are usually located far away from the Fermi level, with only a few exceptions [13,56]. Therefore, a prerequisite condition is to design and discover the TFB@CNP in two-dimensional (2D) organic frameworks.

Here, we propose an effective approach to realize TFB@CNP focusing COFs and MOFs with an underlying diatomic kagome lattice (DKL). Based on density-functional theory (DFT) and tight-binding (TB) calculations, we predict that both a family of 2D COFs (3R-X, X = F, Cl, Br; 3R = aza[3]radialene) and MOFs (HAT-Sn; HAT = 1,4,5,8,9,12-hexaazatriphenylene) can host the TFB@CNP to further enhance the catalytic activity of their active sites that are involved in TFB formation. The prototypes of these COFs and MOFs have already been experimentally synthesized [57,58]. We show that in both families, TFBs can be designed exactly at the CNP by tuning the relative strength of intra- and interkagome lattice hopping in correlation with electron filling. Remarkably, by calculating the free energy diagram of the hydrogen evolution reaction (HER), we show that 3R-X (X = F, Cl, Br) and HAT-Sn possess significantly improved HER activity in the presence of TFB@CNP.

II. COMPUTATION METHODS

DFT calculations were performed in the Vienna *ab initio* simulation package (VASP) [59,60] adopting the projector-augmented wave (PAW) method and the Perdew-Burke-Ernzerhof (PBE) [61] functional. The Grimme-D3 method [62] was used to describe the van der Waals interaction between 3R-Cl COF layer and the metal substrate [63]. Ag (111) substrates were modeled by three layered slabs with two bottom layers fixed and the top layer relaxed. The vacuum layers for all structural models were larger than 15 Å. The atoms were fully relaxed until the force on each atom was less than 0.01 eV Å^{-1} in the monolayer COF or MOF systems and less than 0.02 eV Å^{-1} in the COF/Ag(111) systems. The energy cutoff of the plane-wave basis was set at 600 eV. The Brillouin zone was sampled by a Gamma-centered k mesh [64] of 3×3×1. The lattice constants of 3R-X (X = H, F, Cl, Br) are 15.12, 15.11, 15.13, and 15.11 Å, respectively. The lattice constants of HAT-Sn and HAT-Pb are 15.39 and 15.60 Å, respectively. The *ab initio* molecular dynamics (AIMD) simulations were performed under the canonical (*NVT*) ensemble. The scalar relativistic effect was considered in the PAW pseudopotential for all calculations. The fully relativistic effect (i.e., spin-orbit coupling) was checked to have a negligible influence on the Gibbs free energies. Therefore, these extra calculation results are not included.

The standard hydrogen electrode (SHE) model was used for the proton-electron transfer steps of the HER process [65], in which the applied electric potential is 0 V, pH is 0, temperature is 298.15 K, and the pressure of H_2 is 1 bar. Under these standard conditions, the hydrogen evolution is at equilibrium: $H^+ + e^- \leftrightarrow \frac{1}{2}H_2$.

Since the active sites of the studied COFs and MOFs are spatially separated from each other, we adopted the Volmer-Heyrovský mechanism for HER [66]. In the first step, a Volmer reaction, the hydrogen cation (H^+, i.e., proton) from the solution adsorbs on the COF or MOF surface at the chemically active sites, and transforms into H^*, $H^+ + e^- + * \rightarrow H^*$, where * denotes the active site that binds with H. In the second step, a Heyrovský reaction, another H^+ comes by the adsorbed H atom (H^*), pulling away from the surface and releasing as H_2, $H^+ + e^- + H^* \rightarrow H_2$.

The change of Gibbs free energy (ΔG_{H^*}) for the Volmer reaction is generally calculated using gas-phase approximation as [65]

$$\Delta G_{H^*} \approx \Delta E_H + \Delta E_{ZPE} - T\Delta S_H. \tag{1}$$

Note that the thermal correction derived from the integral of heat capacity is not included in Eq. (1), owing to its minimal contribution at room temperature. Here, ΔE_H is the adsorption energy of hydrogen on the substrate. It can be obtained by the following formula:

$$\Delta E_H = \frac{1}{n}\left[E(H^*) - E(^*) - \frac{n}{2}E(H_2)\right], \tag{2}$$

where $E(H^*)$, $E(^*)$, and $E(H_2)$ are the total energies of the adsorption system, the 2D COFs and MOFs, and the hydrogen molecule, respectively. n is the number of hydrogen atoms. ΔE_{ZPE} is the difference of the zero-point energy (ZPE) of the system between adsorbed and gaseous hydrogen, which can be obtained by the following formula:

$$\Delta E_{ZPE} = \frac{1}{n}\left[E_{ZPE}(H^*) - E_{ZPE}(^*) - \frac{n}{2}E_{ZPE}(H_2)\right], \tag{3}$$

where $E_{ZPE}(H^*)$, $E_{ZPE}(^*)$, and $E_{ZPE}(H_2)$ are the ZPE values of the adsorption system, the 2D COFs and MOFs, and the hydrogen molecule, respectively. n is the number of hydrogen atoms.

In particular, the contributions of all atoms were taken into account in the ZPE calculations that

$$E_{ZPE} = \Sigma_i^{3N}\frac{h\nu_i}{2}, \tag{4}$$

where h is the Planck constant, ν_i is the vibrational frequency of harmonic phonons calculated by the finite displacement method, and N is the total number of atoms in a system. The temperature T is set to 298.15 K. ΔS_H is the corresponding entropy change, as given by

$$\Delta S_H = -\tfrac{1}{2}S^0_{H_2}, \tag{5}$$

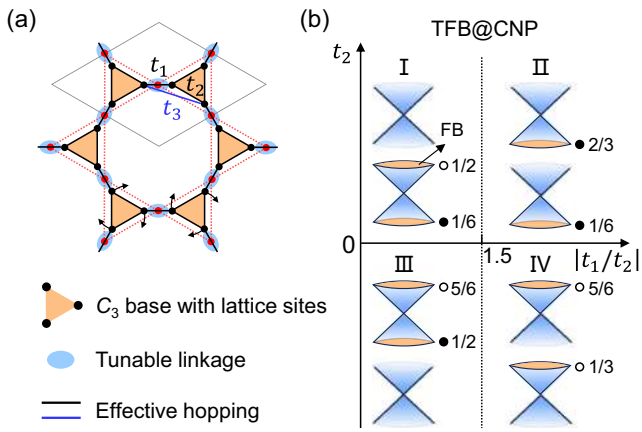


FIG. 1. Designing TFB@CNP in DKL using COFs and MOFs. (a) Schematics of DKL. Red dots are kagome lattice sites; black dots are DKL sites. The orange triangles and blue ovals denote C_3 bases and linkages which constitute COFs and MOFs. The arrows indicate rotation of diatomic dumbbell or equivalently rotation of triangle bases. The first NN inter- and intrakagome hopping are t_1 and t_2, respectively. The second NN interkagome hopping is t_3. (b) Phase diagram of DKL under different hopping conditions. Four regions are divided according to the band structures. The blue hourglasses with orange bottoms denote Dirac bands and connected FBs. The solid (open) dots denote the electronlike (holelike) TFB@CNP under corresponding filling factors.

Considering the small vibrational entropy of adsorbed hydrogen, $S^0_{H_2}$, the entropy of hydrogen under standard conditions is 130.680 J/(K mol), so that $T\Delta S_H \approx 0.202$ eV. The calculation results of ΔE_H and ΔE_{ZPE} are summarized in Table I.

III. RESULTS AND DISCUSSION

A. Theoretical design of TFB@CNP

We first discuss an inverse design process of 2D lattices with TFB@CNP based on a TB lattice Hamiltonian in correlation with electron band filling in the DKL model, as illustrated in Fig. 1(a). The red dots and dashed lines represent the kagome lattice. One can split each kagome lattice site (red dot) symmetrically into two atoms (black dots), like a diatomic dumbbell connected by solid lines, to form a DKL, with two subkagome lattices. Let t_1 and t_3 represent the first and second NN interkagome lattice hopping, respectively, and t_2 the intrakagome hopping [Fig. 1(a)]. It has been shown that [51] at the chiral limit, the DKL is equivalent to a bipartite double cover of a line graph in graph theorem by setting $t_2 = 0$; then chiral symmetry leads to the formation of two sets of kagome bands of opposite chirality, each having a FB sitting above or below two Dirac bands. Importantly, the chiral symmetry guarantees the two TFBs (so-called yin-yang flat bands) to be at or around the CNP [52]. However, in most real materials, chiral symmetry is broken because $t_2 \neq 0$. Therefore, our focus here is to find conditions to have one TFB@CNP for achiral DKL models.

A very rich phase diagram of TB band structures of DKL has been previously mapped out in the parameter spaces of t_1, t_2, and t_3 [53]. Since we are interested in COFs and MOFs with very large ligand groups, t_3 at a large distance can be neglected. Then the six-band TB Hamiltonian of a DKL is written as

$$H = t_1 \sum_{\langle i,j\rangle} c_i^\dagger c_j + t_2 \sum_{\langle\langle i,j\rangle\rangle} c_i^\dagger c_j + \varepsilon \sum_i c_i^\dagger c_i, \tag{6}$$

where $c_i^\dagger$ (c_i) is the creation (annihilation) operator at lattice site i and ε is the on-site energy. As illustrated in Fig. 1(b), there are four typical types of achiral DKL band structures, I, II, III, and IV, depending on the ratio of $|t_1/t_2|$ and the sign of t_2. All four types host two TFBs located at different energies, and the electron filling required to achieve TFB@CNP can be easily determined [see Fig. 1(b)].

Interestingly, the band structures remain the same if one rotates the diatomic dumbbell about the kagome lattice sites [see black arrows in Fig. 1(a)], as the rotated DKL Hamiltonian is a unitary transformation of the original one [46] [see Fig. 2]. This gives extra flexibility in material realization of TFB@CNP, as shown below. We choose the case that t_1, $t_2 > 0$ as an example to explain the underlying reason for the equivalence of the DKL under the C_{3z}-symmetric rotations. We begin with the TB model of DKL. One can consider the following 6×6 Hamiltonian as the matrix representation for Eq. (6), where $\vec{v}_m$ ($m = 1, 2, 3, 4, 5, 6$) are nearest position vectors, as depicted in Fig. 2(a). There are two sets of triangles in a DKL, forming a compound lattice.

$$H^{\mathrm{DKL}}(\vec{k}) = \begin{bmatrix} 0 & t_1 e^{i\vec{k}\cdot\vec{v}_5} & t_1 e^{i\vec{k}\cdot\vec{v}_4} & t_2 e^{i\vec{k}\cdot\vec{v}_1} & 0 & 0 \\ t_1 e^{-i\vec{k}\cdot\vec{v}_5} & 0 & t_1 e^{i\vec{k}\cdot\vec{v}_6} & 0 & t_2 e^{i\vec{k}\cdot\vec{v}_2} & 0 \\ t_1 e^{-i\vec{k}\cdot\vec{v}_4} & t_1 e^{-i\vec{k}\cdot\vec{v}_5} & 0 & 0 & 0 & t_2 e^{i\vec{k}\cdot\vec{v}_3} \\ t_2 e^{-i\vec{k}\cdot\vec{v}_1} & 0 & 0 & 0 & t_1 e^{-i\vec{k}\cdot\vec{v}_4} & t_1 e^{-i\vec{k}\cdot\vec{v}_5} \\ 0 & t_2 e^{-i\vec{k}\cdot\vec{v}_2} & 0 & t_1 e^{i\vec{k}\cdot\vec{v}_4} & 0 & t_1 e^{i\vec{k}\cdot\vec{v}_6} \\ 0 & 0 & t_2 e^{-i\vec{k}\cdot\vec{v}_3} & t_1 e^{i\vec{k}\cdot\vec{v}_5} & t_1 e^{-i\vec{k}\cdot\vec{v}_6} & 0 \end{bmatrix}. \tag{7}$$

By solving the Hamiltonian in Eq. (7) with $t_1/t_2 = 1.8$ and the on-site energy $\varepsilon = -0.8$ eV, we obtain the band structure shown in Fig. 2(d). We find the varying of t_1/t_2 can lead to 2 + 4 to 3 + 3 band transitions, which are summarized in Fig. 1(b). Such a transition means the TFBs in DKL are highly tunable, which offers more chances for achieving TFB@CNP

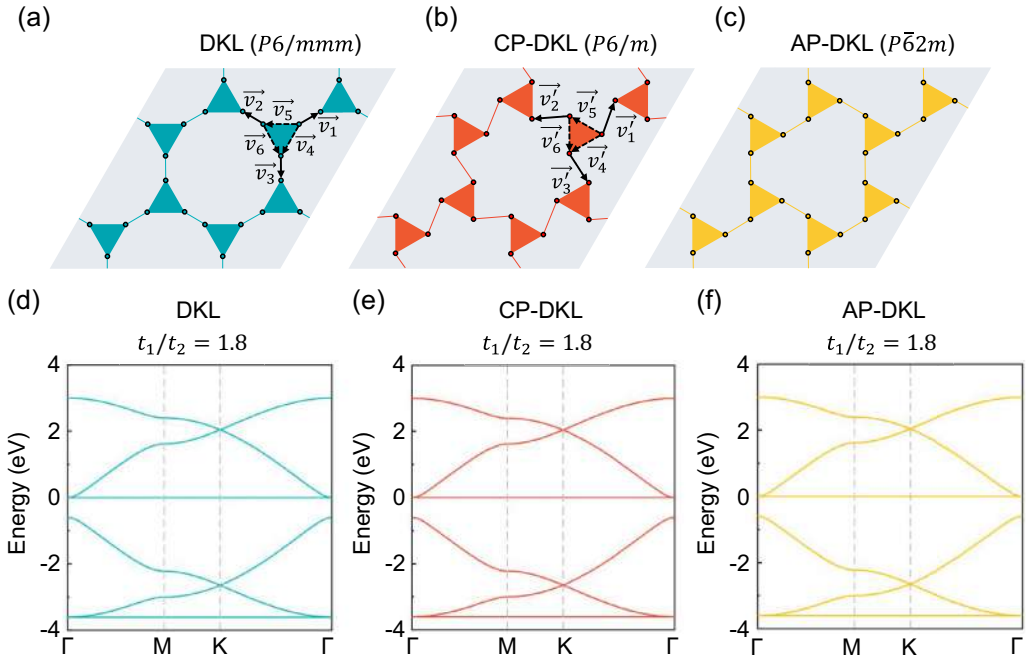


FIG. 2. Equivalence of the DKL under the C_{3z}-symmetric rotations. (a)–(c) Schematics of the (normal) DKL, cophase (CP), and antiphase (AP) rotated DKLs, respectively. (d)–(f) Band structures of the tight-binding Hamiltonian with $t_1/t_2 = 1.8$ for DKL, CP DKL, and AP DKL, respectively. The on-site energy ε is set to -0.8 eV.

than the "simple" lattices, such as kagome lattice. Notice there is also a degree of freedom to rotate the triangles in kagomelike structures. The known example is the rotation transformation from a kagome lattice to a CT lattice [47]. We have demonstrated such a transformation in superatomic zirconium dichlorides [67]. To be more general, we consider the possible triangle rotations in DKLs. As shown in Figs. 2(b) and 2(c), there are two ways to rotate the two triangles, i.e., cophase (CP) rotation and antiphase (AP) rotation. Here, we mainly focus on the CP DKL, and a similar conclusion can be drawn for the AP DKL. For the CP DKL, up to the nearest hopping, $H^{\mathrm{CP}}(\vec{k})$ takes a similar form as $H^{\mathrm{DKL}}(\vec{k})$, while the only difference is to replace $\vec{v}_m$ with $\vec{v}'_m$ ($m = 1, 2, 3, 4, 5, 6$). We prove that $H^{\mathrm{DKL}}(\vec{k})$ can convert into $H^{\mathrm{CP}}(\vec{k})$ by a unitary transformation $UU^{\dagger} = I$, i.e., $H^{\mathrm{CP}}(\vec{k}) = UH^{\mathrm{DKL}}(\vec{k})U^{-1}$, and U takes the form

$$U = \mathrm{diag}[1,\ e^{i\vec{k}\cdot(\vec{v}_5-\vec{v}'_5)},\ e^{i\vec{k}\cdot(\vec{v}_4-\vec{v}'_4)},\ e^{i\vec{k}\cdot(\vec{v}_1-\vec{v}'_1)},\ e^{i\vec{k}\cdot(\vec{v}_2+\vec{v}_5-\vec{v}'_2-\vec{v}'_5)},\ e^{i\vec{k}\cdot(\vec{v}_3+\vec{v}_4-\vec{v}'_3-\vec{v}'_4)}]. \tag{8}$$

It means that the rotation is equal to a unitary transformation between $H^{\mathrm{CP}}(\vec{k})$ and $H^{\mathrm{DKL}}(\vec{k})$, so that $H^{\mathrm{CP}}(\vec{k})$ and $H^{\mathrm{DKL}}(\vec{k})$ share the same band eigenvalues, as shown in Figs. 2(d) and 2(e). For the AP DKL, we also obtained the same band eigenvalues [see Fig. 2(f)]. In this sense, the CP DKL (AP DKL) is equivalent to the pristine one.

B. Material prediction of TFB@CNP

We next predict materials with TFB@CNP in 2D COFs and MOFs underlined by the DKL model. Based on the above analyses, one may approach this in two ways. For given electron filling, one can tune the ratio of t_1/t_2 by changing the ligand groups and/or linkage atoms between the groups. For example, at 2/3 filling, if one increases the $t_1/t_2 > 3/2$, a TFB@CNP can be created when the band structure changes from type I to type II, as shown in the top panel of Fig. 1(b). Alternatively, for a given type of band structure, one tunes the electron filling; for example, the required electron filling for type I is 1/6 or 1/2 to have a TFB@CNP. Here we adopt the first approach, because, on one hand, the tunable linkages are easier to implement in experiments; on the other hand, a direct comparison between the catalytic behaviors of a TFB at and not at CNP can be made to highlight the importance of TFB@CNP for enhancing catalysis.

We first show DFT calculations for COFs with TFB@CNP. As depicted in Fig. 3(a), the 3R-X (X = F, Cl, Br) family has the benzene halides of C_2 symmetry sitting at the kagome sites, which are linked by aza[3]radialene (3R) of C_3 symmetry, forming a CP DKL with a $P6/m$ symmetry and a small counterclockwise rotation of two organic groups [see Fig. 2(b)]. The prototype 3R-Cl COF has recently been synthesized via [1 + 1 + 1] cycloaddition of the precursor on silver surface [57], and our simulated STM image agrees well with the experimental STM images [Fig. 4]. We also confirm the dynamical and thermodynamic stability of 3R-X COFs (see Figs. S1 and S2 in the Supplemental Material (SM) [68]).

Figure 3(b) shows the DFT band structure of 3R-Cl, with one FB located exactly at the CNP. Using the irreducible representations of this FB, parity eigenvalues of $+1$ (Γ point) and -1 (M points) are revealed, indicating a nontrivial $\mathbb{Z}_2 = 1$ band topology [69]. The charge density of this TFB is found to be distributed over the whole lattice including both N and

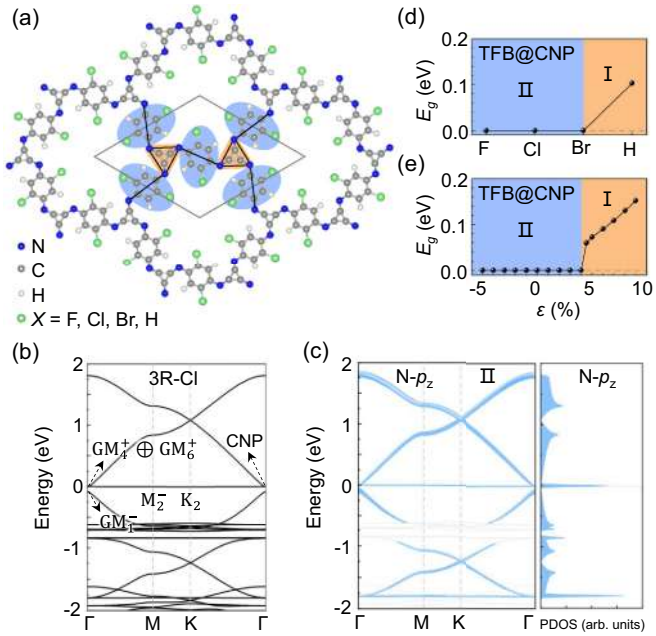


FIG. 3. TFBs in the 3R-X (X = F, Cl, Br, H) COFs. (a) Structure of the 3R-X COFs with a rotated DKL (black lines). Here, the 3R groups serve as the C_3 bases (shaded in orange), while the benzene halides are tunable linkages (shaded in blue). (b) Band structure of 3R-Cl. The irreducible representations of high-symmetry points are labeled, where the sign of superscript denotes parity. (c) Projected bands and PDOS for the p_z orbitals of N atoms. (d) Band gaps of 3R-X COFs. (e) Band gaps of 3R-Cl under biaxial strain (ε). The positive (negative) ε corresponds to tensile (compressive) strain.

C atoms (see SM Fig. S3(a) [68]), consistent with previous works [27]. The projected bands and projected DOS (PDOS) in Fig. 3(c) and SM Fig. S4 [68] confirm that electronic states near the Fermi level can be captured by a simple, effective six-band DKL model, having effective lattice sites near N atoms [see Fig. 3(a)]. By electron counting, we reveal the 3R-Cl COF has 2/3 filled DKL bands, corresponding to case II in Fig. 1(b). Furthermore, one may apply the common strategy of element substitution and strain engineering [29,41] to tune the hopping ratio of t_1/t_2. We found that 3R-X (X = F, Cl, Br) have type II DKL bands with exactly TFB@CNP, while 3R-H is a semiconductor of the type I bands [see Fig. 3(d)].

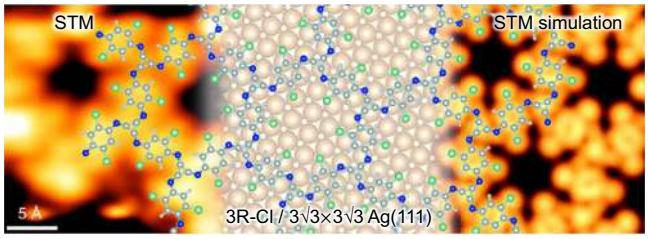


FIG. 4. 3R-Cl synthesis on Ag(111) substrate. Merged image for the experimental STM image, atomic structure, and simulated STM image of 3R-Cl on Ag(111) substrate. The STM image is adapted from Ref. [57]. The energy range of the STM simulation is from −1 eV to the Fermi level (0 eV). The unit cell is 1×1 3R-Cl on $3\sqrt{3}\times3\sqrt{3}$ Ag(111).

Since the electronegativities of the halogen atoms are stronger than that of the hydrogen atoms, the benzene halides are more likely to attract the lone pair electrons of N atoms than benzenes, so that 3R-X (X = F, Cl, Br) have a larger t_1/t_2 ratio than 3R-H. This is supported by the significant different TFB charge distribution of 3R-Cl from that of 3R-H (see SM Fig. S3 [68]). Thus, by modifying the linkages with element substitution, a transition of DKL bands is achieved in 3R-X COFs. Applying biaxial strain shows another way of tuning t_1/t_2 through mostly bond rotation. In Fig. 3(e), one can see the transition of DKL bands of 3R-Cl from type II to type I, which occurs under a tensile strain larger than 4%.

We next predict the MOF with TFB@CNP. We consider the HAT-M (M = Sn, Pb) MOFs, whose prototypes have been recently synthesized by on-surface reaction [58]. As shown in Fig. 5(a), the metal atom sitting on the kagome site acts as a linkage to bond HAT molecules having C_3 symmetry, forming a pristine DKL without rotation [see Fig. 2(a)]. The band structure of HAT-Sn corresponds to type III at 1/2 filling [Fig. 1(b)], featured with the TFB@CNP [Fig. 5(b)]. The projected band structure and PDOS of HAT-Sn show that the six bands near the Fermi level are contributed by p_z orbitals of both Sn atoms and HAT molecules [Fig. 5(c)]. In particular, the TFB at the Fermi level is composed of p_z orbitals of Sn, C, and N atoms, which will participate in the (quick) charge transfer process of HER. Thus, the charge density from the TFB@CNP is globally distributed on the whole framework [Fig. 5(d)]. These features agree well with a DKL model, whose effective sites are shown by the black dots. The black circles approximately mark the charges contributing the effective sites. Differently, HAT-Pb has type IV DKL bands [Fig. 5(e)]. A DKL band transition from type III to type IV is induced by changing the linkage metal atom. For HAT-Pb, the larger Pb atom increases the distance between HAT molecules, leading to a smaller $|t_2|$ compared with HAT-Sn. We also confirm the thermodynamic stability of HAT-M MOFs on a suitable substrate at low temperature (see SM Fig. S5 [68]).

C. Catalytic enhancement by TFB@CNP

Given the material realization of TFB@CNP, we next explore the catalytic activity of both COFs and MOFs, using HER as an example. For HER, it has been shown that the experimentally measured exchange current in solution can be correlated with the calculated H adsorption free energy (ΔG_{H^*}) in the gas phase, which follows a volcano-shaped curve with the peak positions corresponding to the highest activity [65]. In general, the peak position of the highest activity corresponds with $\Delta G_{H^*} \approx 0$ eV. Therefore, the catalyst with the smallest $|\Delta G_{H^*}|$ exhibits the highest activity, which indicates an optimal hydrogen adsorption strength—neither too weak for effective H binding nor too strong for rapid H_2 release [70]. For example, Fig. 6(a) shows the two-step Volmer-Heyrovský mechanism of the HER process [66]. In the first step of a Volmer reaction, the hydrogen cation (H^+, i.e., proton) from solution adsorbs on the 3R-X (X = F, Cl, Br) COF surface at the chemically active sites, which are N atomic sites, and transforms into H^*, $H^+ + e^- + * \rightarrow H^*$, where * denotes the active N site that binds with H. In the

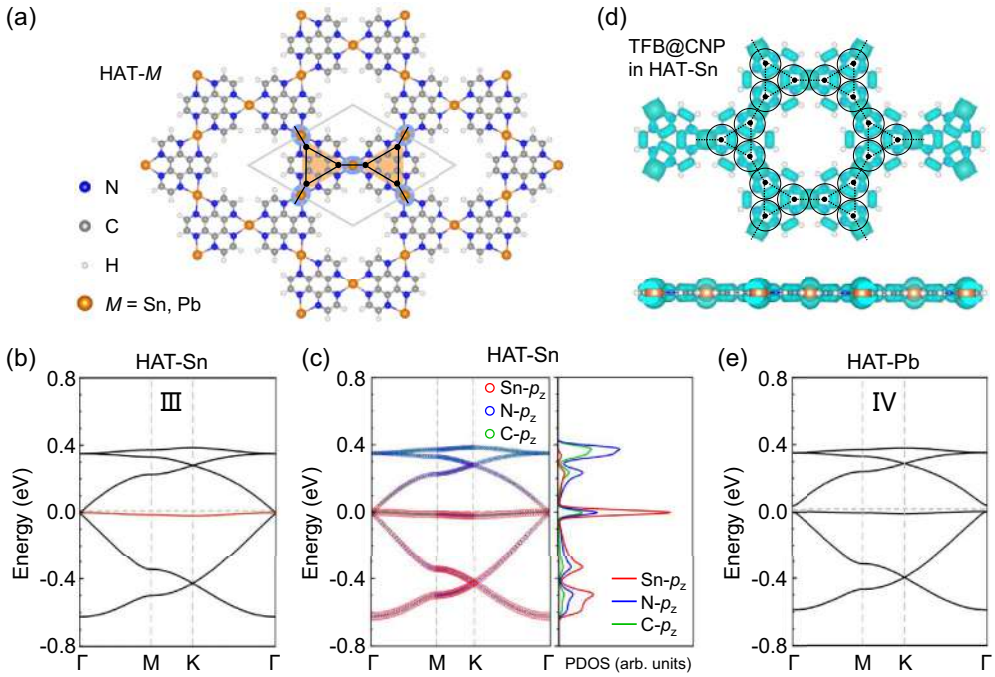


FIG. 5. TFBs in the HAT-M (M = Sn, Pb) MOFs. (a) Structure of the HAT-M with a DKL. The black lines outline a DKL. The orange labeled HAT molecules are C_3 bases, whose effective sites locate at the black dots. The blue-shaded metal atoms are tunable linkages. (b) Band structures of HAT-Sn. The red line highlights the TFB@CNP. (c) Projected bands and PDOS of HAT-Sn. (d) Global charge density distribution of the TFB@CNP of HAT-Sn. (e) Band structures of HAT-Pb.

second step of a Heyrovský reaction, another H^+ comes by the adsorbed H atoms (H^*), pulling away from the surface and releasing as H_2, i.e., $H^+ + e^- + H^* \rightarrow H_2$. Therefore, for simplicity, we have calculated the change of Gibbs free energy ΔG_{H^*} of the first step in the gas phase, as one key indicator for HER activity. We found that ΔG_{H^*} is $\sim$0.06−0.12 eV when the H^* coverage (μ) is 1.0 for 3R-X (X = F, Cl, Br) [see Fig. 6(b)], which is comparable with the activity of Pt (0.08 eV) [71] and Os−N_3S_1 (0.09 eV) [72], as well as other topological catalysts, such as TiB_2 (0.08 eV) [19], TiSi (0.04 eV) [20], PtGa (0.131 eV) [21], and H4CdPc-COF(F) (−0.016 eV) [73]. The rather low absolute value of ΔG_{H^*} (<0.1 eV) indicates high catalytic efficiency for HER. The detailed computational methods of ΔG_{H^*} can be found in the Computation Methods section. ΔG_{H^*} for different μ are evaluated and listed in Table I. When μ is 1.0, the absolute value of ΔG_{H^*} is optimal for HER.

Therefore, the combination of abundant active N sites with TFB@CNP in 3R-X (X = F, Cl, Br) COFs makes them promising for high-performance catalysis. We also test the HER catalytic activity of 3R-Cl on Ag(111) surface (SM Fig. S6 [68]), whose ΔG_{H^*} is −0.003 eV, indicating still an excellent HER performance. Similarly, TFB@CNP in HAT-Sn MOF is found effective for catalyzing HER. Here Sn atoms are found to be active sites for H^* in adsorption in HAT-Sn [Fig. 6(c)]. When the H^* coverage $\mu = 0.33$, the optimal ΔG_{H^*} is 0.08 eV.

D. Comparative studies to confirm the role of TFB@CNP

To verify the role of TFB@CNP in the enhanced catalysis, we perform computational experiments by moving the TFB away from the CNP, using atom substitution or strain engineering. Specifically, we have calculated the ΔG_{H^*} for HER in the semiconducting cases without TFB@CNP, e.g., the 3R-H COF [Fig. 3(d)], the strained 3R-Cl COF [Fig. 3(e)], and the HAT-Pb MOF [Fig. 5(e)]. The calculated ΔG_{H^*} for 3R-H is 0.23 eV [marked by the black line in Fig. 6(b)], significantly larger than that for 3R-Cl (0.07 eV), while the ΔG_{H^*} for 3R-Cl under biaxial strain is 0.09 eV for ε= 5% and 0.15 eV for ε= 10%, respectively, which are noticeably

TABLE I. Parameters used in the calculations of the change of Gibbs free energies (ΔG_{H^*}). Hydrogen adsorption energies ΔE_H, the difference of the zero-point energy between adsorbed and gaseous hydrogen ΔE_{ZPE}, and ΔG_{H^*} with different H^* coverages μ.

Systems	ΔE_H (eV)	ΔE_{ZPE} (eV)	ΔG_{H^*} (eV)
3R-Cl ($\mu = 0.17$)	−1.194	0.199	−0.79
3R-Cl ($\mu = 0.33$)	−1.155	0.191	−0.76
3R-Cl ($\mu = 0.50$)	−1.005	0.189	−0.61
3R-Cl ($\mu = 0.67$)	−0.957	0.181	−0.57
3R-Cl ($\mu = 0.83$)	−0.549	0.164	−0.18
3R-Cl ($\mu = 1.00$)	−0.285	0.154	0.07
3R-Cl ($\mu = 1.00$) (strain: 5%)	−0.282	0.173	0.09
3R-Cl ($\mu = 1.00$) (strain: 10%)	−0.246	0.190	0.15
3R-F ($\mu = 1.00$)	−0.236	0.153	0.12
3R-Br ($\mu = 1.00$)	−0.294	0.150	0.06
3R-H ($\mu = 1.00$)	−0.123	0.151	0.23
Sn-HAT ($\mu = 0.33$)	−0.190	0.069	0.08
Sn-HAT ($\mu = 0.67$)	−0.118	0.063	0.15
Sn-HAT ($\mu = 1.00$)	−0.076	0.057	0.18
Pb-HAT ($\mu = 0.33$)	0.091	0.040	0.33

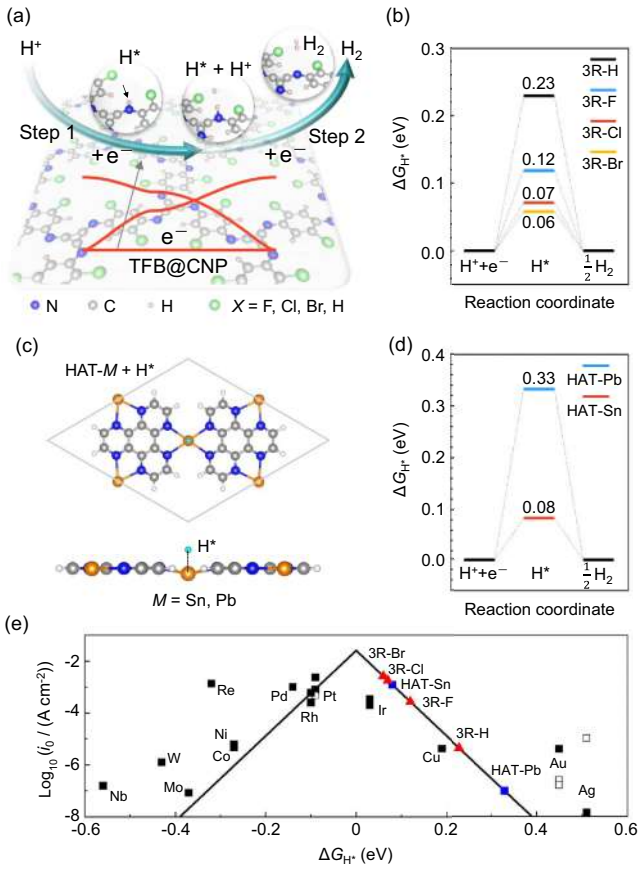


FIG. 6. HER activity of 3R-X COFs and HAT-M MOFs. (a) Schematic of the HER process of 3R-X, in which the N atoms are active sites. (b) The calculated free energy diagrams of HER for 3R-X with H^* coverage $\mu = 1$. (c) Top and side views of the hydrogen adsorption on HAT-M. (d) Calculated free energy diagram of HER for HAT-M with H^* coverage $\mu = 0.33$. (e) Volcano plot of $\log_{10}(i_0)$ vs calculated hydrogen adsorption free energy (ΔG_{H^*}) for HER catalysts. The $\log_{10}(i_0)$ data of the metal surfaces (black squares) are from experiments [65], while those of the COFs and MOFs (colorful shapes) are given by the volcano curve.

increased over the intrinsic 3R-Cl. Similarly for MOFs, ΔG_{H^*} for the semiconducting HAT-Pb is 0.33 eV [see Fig. 6(d) and Table I], much larger than that of HAT-Sn. These comparisons evidently confirm the significant role of TFB@CNP in enhancing the HER activity.

Furthermore, we compared the catalytic activity of our COF and MOF candidates hosting TFB@CNP with some known high-performance catalysts, such as Pt. In experiments, a catalyst for HER is usually measured by the exchange current density, i_0. Interestingly, the experimentally measured exchange current i_0 has been examined as a function of the theoretically calculated H adsorption free energy (ΔG_{H^*}), which follows a volcano-shaped curve with the peak position occupied metal Pt as the best catalyst [70]. Similarly, we have added a plot of the volcano curve to compare the relative catalytic activities of our COFs and MOFs with previous catalysts in Fig. 6(e). The measured exchange currents (i_0) of the plot are adapted from Table I in Ref. [65], which summarizes the data from the original Refs. [74–79]. Remarkably, both 3R-X (X = F, Cl, Br) and HAT-Sn locate near the top of the volcano curve, indicating their high catalytic activity, comparable with noble metals, such as Pt. In contrast, 3R-H and HAT-Pb, without TFB@CNP, locate near the bottom of the volcano curve.

We note that COFs and MOFs are well known to have abundant catalytically active sites due to reaction chemistry, regardless of TFB. On the other hand, the TFB is formed by the lattice symmetry of all lattice sites, which may or may not be catalytically active. For example, in 3R-X COFs, due to atomic rotation, the effective DKL lattice sites are between N and C_1 atoms [Fig. 3(a)], and C_1 contributes more to TFB density than N (see SM Fig. S4 [68]). But only N are active because H are attracted to their lone pair electrons, as shown from the calculated DOS with H adsorbed on N (see Fig. S7(a) [68]), in comparison with H adsorbed on C_1(see Fig. S7(b) [68]). In HAT-M MOFs, metal atoms are active because of strong M-H bond formation. Interestingly, what we have shown is that if the active sites, e.g., N in 3R-X COFs and M in HAT-M MOFs, are involved in forming the TFB, then the TFB@CNP can make them even more active. This is because the TFB@CNP contributes large DOS near the Fermi level [see Figs. 3(c) and 5(c)], which enhances not only the adsorption of the adsorbates on the active sites, but also the charge transfer rate during reactions [17]. Finally, in comparison with previous 2D COF and MOF catalysts, such as C_3N_4@graphene [80], our proposed COFs and MOFs, e.g., 3R-Cl, exhibit intrinsic high HER activity, requiring no external interface engineering.

IV. CONCLUSIONS

Taking DKL as a model system, we propose a generic design strategy for realizing TFB@CNP in 2D organic frameworks. Based on DFT and TB calculations, we predict that the TFBs in 3R-X COFs and HAT-M MOFs, underlined by the DKL, can be tuned by varying linkages or applying biaxial strain. We show that the TFB@CNP can further enhance the HER activity on the abandunt active sites in COFs and MOFs. Our findings open a promising direction for high-performance catalysis, and offer a broad prospective for discovering and engineering TFB@CNP in organic frameworks for chemical applications.

Note added. During the review process, we noted another study [81] suggesting FB may enhance catalytic performance.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFA1204103), and the National Natural Science Foundation of China (Grants No. 22372047 and No. 12404201). W.-H.D. was supported by the Postdoctoral Fellowship Program of CPSF (Grant No. GZC20231363). F.L. acknowledges financial support from the DOE-BES (Grant No. DE-FG02-04ER46148).

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

[1] Z. Liu, F. Liu, and Y.-S. Wu, Exotic electronic states in the world of flat bands: From theory to material, Chin. Phys. B **23**, 077308 (2014).

[2] E. Tang, J.-W. Mei, and X.-G. Wen, High-temperature fractional quantum Hall states, Phys. Rev. Lett. **106**, 236802 (2011).

[3] H. Liu, G. Sethi, D. N. Sheng, Y. Zhou, J.-T. Sun, S. Meng, and F. Liu, High-temperature fractional quantum Hall state in the floquet kagome flat band, Phys. Rev. B **105**, L161108 (2022).

[4] K. Sun, Z. Gu, H. Katsura, and S. D. Sarma, Nearly flatbands with nontrivial topology, Phys. Rev. Lett. **106**, 236803 (2011).

[5] T. Neupert, L. Santos, C. Chamon, and C. Mudry, Fractional quantum Hall states at zero magnetic field, Phys. Rev. Lett. **106**, 236804 (2011).

[6] H. Park, J. Cai, E. Anderson, Y. Zhang, J. Zhu, X. Liu, C. Wang, W. Holtzmann, C. Hu, Z. Liu *et al.*, Observation of fractionally quantized anomalous Hall effect, Nature (London) **622**, 74 (2023).

[7] C. Repellin, The twisted material that splits the electron, Nature (London) **622**, 63 (2023).

[8] Y. Cao, V. Fatemi, A. Demir, S. Fang, S. L. Tomarken, J. Y. Luo, J. D. Sanchez-Yamagishi, K. Watanabe, T. Taniguchi, E. Kaxiras *et al.*, Correlated insulator behaviour at half-filling in magic-angle graphene superlattices, Nature (London) **556**, 80 (2018).

[9] Y. Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, and P. Jarillo-Herrero, Unconventional superconductivity in magic-angle graphene superlattices, Nature (London) **556**, 43 (2018).

[10] M. Oh, K. P. Nuckolls, D. Wong, R. L. Lee, X. Liu, K. Watanabe, T. Taniguchi, and A. Yazdani, Evidence for unconventional superconductivity in twisted bilayer graphene, Nature (London) **600**, 240 (2021).

[11] C. Wu, D. Bergman, L. Balents, and S. D. Sarma, Flat bands and Wigner crystallization in the honeycomb optical lattice, Phys. Rev. Lett. **99**, 070401 (2007).

[12] B. Jaworowski, A. D. Güçlü, P. Kaczmarkiewicz, M. Kupczyński, P. Potasz, and A. Wójs, Wigner crystallization in topological flat bands, New J. Phys. **20**, 063023 (2018).

[13] Y. Chen, S. Xu, Y. Xie, C. Zhong, C. Wu, and S. B. Zhang, Ferromagnetism and Wigner crystallization in kagome graphene and related structures, Phys. Rev. B **98**, 035135 (2018).

[14] D. Jérome, T. M. Rice, and W. Kohn, Excitonic insulator, Phys. Rev. **158**, 462 (1967).

[15] G. Sethi, Y. Zhou, L. Zhu, L. Yang, and F. Liu, Flat-band-enabled triplet excitonic insulator in a diatomic kagome lattice, Phys. Rev. Lett. **126**, 196403 (2021).

[16] G. Sethi, M. Cuma, and F. Liu, Excitonic condensate in flat valence and conduction bands of opposite chirality, Phys. Rev. Lett. **130**, 186401 (2023).

[17] B. Hammer and J. K. Norskov, Why gold is the noblest of all the metals, Nature (London) **376**, 238 (1995).

[18] H. Liu, G. Sethi, S. Meng, and F. Liu, Orbital design of flat bands in non-line-graph lattices via line-graph wave functions, Phys. Rev. B **105**, 085128 (2022).

[19] X. Zhang, L. Wang, M. Li, W. Meng, Y. Liu, X. Dai, G. Liu, Y. Gu, J. Liu, and L. Kou, Topological surface state: Universal catalytic descriptor in topological catalysis, Mater. Today **67**, 23 (2023).

[20] J. Li, H. Ma, Q. Xie, S. Feng, S. Ullah, R. Li, J. Dong, D. Li, Y. Li, and X.-Q. Chen, Topological quantum catalyst: Dirac nodal line states and a potential electrocatalyst of hydrogen evolution in the TiSi family, Sci. China Mater. **61**, 23 (2018).

[21] Q. Yang, G. Li, K. Manna, F. Fan, C. Felser, and Y. Sun, Topological engineering of Pt-group-metal-based chiral crystals toward high-efficiency hydrogen evolution catalysts, Adv. Mater. **32**, 1908518 (2020).

[22] X. Kong, Z. Liu, Z. Geng, A. Zhang, Z. Guo, S. Cui, C. Xia, S. Tan, S. Zhou, Z. Wang *et al.*, Experimental demonstration of topological catalysis for CO_2 electroreduction, J. Am. Chem. Soc. **146**, 6536 (2024).

[23] Z. F. Wang, L. Chen, and F. Liu, Tuning topological edge states of Bi(111) bilayer film by edge adsorption, Nano Lett. **14**, 2879 (2014).

[24] Z. F. Wang, H. Zhang, D. Liu, C. Liu, C. Tang, C. Song, Y. Zhong, J. Peng, F. Li, C. Nie *et al.*, Topological edge states in a high-temperature superconductor $FeSe/SrTiO_3(001)$ film, Nat. Mater. **15**, 968 (2016).

[25] A. Mielke, Ferromagnetic ground states for the Hubbard model on line graphs, J. Phys. A: Math. Gen. **24**, L73 (1991).

[26] A. Mielke, Ferromagnetism in the Hubbard model on line graphs and further considerations, J. Phys. A: Math. Gen. **24**, 3311 (1991).

[27] N. Regnault, Y. Xu, M.-R. Li, D.-S. Ma, M. Jovanovic, A. Yazdani, S. S. P. Parkin, C. Felser, L. M. Schoop, N. P. Ong *et al.*, Catalogue of flat-band stoichiometric materials, Nature (London) **603**, 824 (2022).

[28] A. Bhattacharya, I. Timokhin, R. Chatterjee, Q. Yang, and A. Mishchenko, Deep learning approach to genome of two-dimensional materials with flat electronic bands, npj Comput. Mater. **9**, 101 (2023).

[29] H. Liu, S. Meng, and F. Liu, Screening two-dimensional materials with topological flat bands, Phys. Rev. Mater. **5**, 084203 (2021).

[30] G. Sethi, B. Xia, D. Kim, H. Liu, X. Li, and F. Liu, Graph theorem for chiral exact flat bands at charge neutrality, Phys. Rev. B **109**, 035140 (2024).

[31] R. Bistritzer and A. H. MacDonald, Moiré bands in twisted double-layer graphene, Proc. Natl. Acad. Sci. USA **108**, 12233 (2011).

[32] Y. Yu, K. Zhang, H. Parks, M. Babar, S. Carr, I. M. Craig, M. Van Winkle, A. Lyssenko, T. Taniguchi, K. Watanabe *et al.*, Tunable angle-dependent electrochemistry at twisted bilayer graphene with moiré flat bands, Nat. Chem. **14**, 267 (2022).

[33] M. Pan, X. Zhang, Y. Zhou, P. Wang, Q. Bian, H. Liu, X. Wang, X. Li, A. Chen, X. Lei *et al.*, Growth of mesoscale ordered two-dimensional hydrogen-bond organic framework with the observation of flat band, Phys. Rev. Lett. **130**, 036203 (2023).

[34] A. Dhakshinamoorthy, Z. Li, and H. Garcia, Catalysis and photocatalysis by metal organic frameworks, Chem. Soc. Rev. **47**, 8134 (2018).

[35] Y. Shi, X. Zhang, H. Liu, J. Han, Z. Yang, L. Gu, and Z. Tang, Metalation of catechol-functionalized defective covalent organic frameworks for lewis acid catalysis, Small **16**, 2001998 (2020).

[36] M. Wang, M. Wang, H.-H. Lin, M. Ballabio, H. Zhong, M. Bonn, S. Zhou, T. Heine, E. Cánovas, R. Dong *et al.*, High-mobility semiconducting two-dimensional conjugated covalent organic frameworks with *p*-type doping, J. Am. Chem. Soc. **142**, 21622 (2020).

[37] R. Dong, P. Han, H. Arora, M. Ballabio, M. Karakus, Z. Zhang, C. Shekhar, P. Adler, P. S. Petkov, A. Erbe *et al.*, High-mobility band-like charge transport in a semiconducting two-dimensional metal–organic framework, Nat. Mater. **17**, 1027 (2018).

[38] S. Wan, F. Gándara, A. Asano, H. Furukawa, A. Saeki, S. K. Dey, L. Liao, M. W. Ambrogio, Y. Y. Botros, X. Duan *et al.*, Covalent organic frameworks with high charge carrier mobility, Chem. Mater. **23**, 4094 (2011).

[39] W. Jiang, X. Ni, and F. Liu, Exotic topological bands and quantum states in metal–organic and covalent–organic frameworks, Acc. Chem. Res. **54**, 416 (2021).

[40] E. H. Lieb, Two theorems on the Hubbard model, Phys. Rev. Lett. **62**, 1201 (1989).

[41] W. Jiang, S. Zhang, Z. Wang, F. Liu, and T. Low, Topological band engineering of lieb lattice in phthalocyanine-based metal–organic frameworks, Nano Lett. **20**, 1959 (2020).

[42] B. Pal, Nontrivial topological flat bands in a diamond-octagon lattice geometry, Phys. Rev. B **98**, 245116 (2018).

[43] M. Ezawa, Higher-order topological insulators and semimetals on the breathing kagome and pyrochlore lattices, Phys. Rev. Lett. **120**, 026801 (2018).

[44] K. Essafi, L. D. C. Jaubert, and M. Udagawa, Flat bands and dirac cones in breathing lattices, J. Phys.: Condens. Matter **29**, 315802 (2017).

[45] A. Mielke, Exact ground states for the hubbard model on the kagome lattice, J. Phys. A: Math. Gen. **25**, 4335 (1992).

[46] X. Ni, Y. Zhou, G. Sethi, and F. Liu, π-Orbital Yin–Yang kagome bands in anilato-based metal–organic frameworks, Phys. Chem. Chem. Phys. **22**, 25827 (2020).

[47] S. Zhang, M. Kang, H. Huang, W. Jiang, X. Ni, L. Kang, S. Zhang, H. Xu, Z. Liu, and F. Liu, Kagome bands disguised in a coloring-triangle lattice, Phys. Rev. B **99**, 100404(R) (2019).

[48] Y. Gao, Y.-Y. Zhang, J.-T. Sun, L. Zhang, S. Zhang, and S. Du, Quantum anomalous Hall effect in two-dimensional cudicyanobenzene coloring-triangle lattice, Nano Res. **13**, 1571 (2020).

[49] M. Chen and S. Wan, The quantum anomalous Hall effect on a star lattice with spin–orbit coupling and an exchange field, J. Phys.: Condens. Matter **24**, 325502 (2012).

[50] Y. Zhou, G. Sethi, C. Zhang, X. Ni, and F. Liu, Giant intrinsic circular dichroism of enantiomorphic flat chern bands and flatband devices, Phys. Rev. B **102**, 125115 (2020).

[51] T. Hu, T. Zhang, H. Mu, and Z. Wang, Intrinsic second-order topological insulator in two-dimensional covalent organic frameworks, J. Phys. Chem. Lett. **13**, 10905 (2022).

[52] Y. Zhou, G. Sethi, H. Liu, Z. Wang, and F. Liu, Excited quantum anomalous and spin Hall effect: Dissociation of flat-bands-enabled excitonic insulator state, Nanotechnology **33**, 415001 (2022).

[53] T. Hu, W. Zhong, T. Zhang, W. Wang, and Z. F. Wang, Identifying topological corner states in two-dimensional metal-organic frameworks, Nat. Commun. **14**, 7092 (2023).

[54] M. Ko, L. Mendecki, and K. A. Mirica, Conductive two-dimensional metal–organic frameworks as multifunctional materials, Chem. Commun. **54**, 7873 (2018).

[55] G. Bian, J. Yin, and J. Zhu, Recent advances on conductive 2D covalent organic frameworks, Small **17**, 2006043 (2021).

[56] J.-Y. You, B. Gu, and G. Su, Flat band and hole-induced ferromagnetism in a novel carbon monolayer, Sci. Rep. **9**, 20116 (2019).

[57] D.-Y. Li, Y. Wang, X.-Y. Hou, Y.-T. Ren, L.-X. Kang, F.-H. Xue, Y.-C. Zhu, J.-W. Liu, M. Liu, X.-Q. Shi *et al.*, On-surface synthesis of [3]radialenes via [1+1+1] cycloaddition, Angew. Chem. Int. Ed. **61**, e202117714 (2022).

[58] C.-K. Lyu, Y.-F. Gao, Z.-A. Gao, S.-Y. Mo, M.-Q. Hua, E. Li, S.-Q. Fu, J.-Y. Chen, P.-N. Liu, L. Huang *et al.*, Synthesis of single-layer two-dimensional metal–organic frameworks $M_3(HAT)_2$ (M = Ni, Fe, Co, HAT = 1,4,5,8,9,12-Hexaazatriphenylene) using an on-surface reaction, Angew. Chem. Int. Ed. **61**, e202204528 (2022).

[59] G. Kresse and J. Hafner, *Ab initio* molecular dynamics for liquid metals, Phys. Rev. B **47**, 558 (1993).

[60] G. Kresse and J. Furthmüller, Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. **6**, 15 (1996).

[61] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. **77**, 3865 (1996).

[62] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 Elements H-Pu, J. Chem. Phys. **132**, 154104 (2010).

[63] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, van der waals density functional for general geometries, Phys. Rev. Lett. **92**, 246401 (2004).

[64] H. J. Monkhorst and J. D. Pack, Special points for brillouin-zone integrations, Phys. Rev. B **13**, 5188 (1976).

[65] J. K. Nørskov, T. Bligaard, A. Logadottir, J. R. Kitchin, J. G. Chen, S. Pandelov, and U. Stimming, Trends in the exchange current for hydrogen evolution, J. Electrochem. Soc. **152**, J23 (2005).

[66] C. G. Morales-Guio, L.-A. Stern, and X. Hu, Nanostructured hydrotreating catalysts for electrochemical hydrogen evolution, Chem. Soc. Rev. **43**, 6555 (2014).

[67] Y. Song, W.-H. Dong, K.-R. Hao, S. Du, and L. Zhang, Designing topological and correlated 2D magnetic states via superatomic lattice constructions of zirconium dichloride, Nano Res. **16**, 13509 (2023).

[68] See Supplemental Material at http://link.aps.org/supplemental/10.1103/56zz-7kwp for phonon spectrum, *ab initio* molecular dynamics simulations, and electronic properties of 3R-*X* COFs and HAT-*M* MOFs.

[69] L. Fu and C. L. Kane, Topological insulators with inversion symmetry, Phys. Rev. B **76**, 045302 (2007).

[70] Y. Zheng, Y. Jiao, Y. Zhu, L. H. Li, Y. Han, Y. Chen, A. Du, M. Jaroniec, and S. Z. Qiao, Hydrogen evolution by a metal-free electrocatalyst, Nat. Commun. **5**, 3783 (2014).

[71] B. Hinnemann, P. G. Moses, J. Bonde, K. P. Jørgensen, J. H. Nielsen, S. Horch, I. Chorkendorff, and J. K. Nørskov, Biomimetic hydrogen evolution: MoS_2 nanoparticles as catalyst for hydrogen evolution, J. Am. Chem. Soc. **127**, 5308 (2005).

[72] D. Cao, H. Xu, H. Li, C. Feng, J. Zeng, and D. Cheng, Volcano-type relationship between oxidation states and catalytic activity of single-atom catalysts towards hydrogen evolution, Nat. Commun. **13**, 5843 (2022).

[73] T. Zhang, Z. Jiang, and A. M. Rappe, Hydrogenation of covalent organic framework induces conjugated π bonds and electronic

topological transition to enhance hydrogen evolution catalysis, J. Am. Chem. Soc. **146**, 15488 (2024).

[74] S. Trasatti, Work function, electronegativity, and electrochemical behaviour of metals: III. Electrolytic hydrogen evolution in acid solutions, J. Electroanal. Chem. Interfacial Electrochem. **39**, 163 (1972).

[75] J. Perez, E. R. Gonzalez, and H. M. Villullas, Hydrogen evolution reaction on gold single-crystal electrodes in acid solutions, J. Phys. Chem. B **102**, 10931 (1998).

[76] J. O'M. Bockris and A. K. N. Reddy, *Modern Electrochemistry* (Plenum Press, New York, 1970), Vol. 2.

[77] D. Eberhardt, E. Santos, and W. Schmickler, Hydrogen evolution on silver single crystal electrodes—first results, J. Electroanal. Chem. **461**, 76 (1999).

[78] O. Savadogo and D. L. Piron, New hydrogen cathodes in acid medium: Case of nickel electrodeposited with Heteropolyacids (HPAs), Int. J. Hydrogen Energy **15**, 715 (1990).

[79] N. M. Markovic and P. N. Ross, in *Interfacial Electrochemistry*, edited by A. Wieckowski (Marcel Dekker, New York, 1999).

[80] X. Tan, H. A. Tahini, and S. C. Smith, *p*-Doped graphene/graphitic carbon nitride hybrid electrocatalysts: Unraveling charge transfer mechanisms for enhanced hydrogen evolution reaction performance, ACS Catal. **6**, 7071 (2016).

[81] Y. Huang, Y. Chen, S. Zhang, Z. Zhang, and P. Cui, High-density natural active sites for efficient nitrogen reduction on kagome surfaces promoted by flat bands, npj Comput. Mater. **11**, 154 (2025).

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：张钧翔
方向：量子科技
单位：中国科学院物理研究所

题目：开放量子系统亚稳态在固态体系中的观测

nature communications

Article

https://doi.org/10.1038/s41467-025-64772-6

Observation of metastability in open quantum dynamics of a solid-state system

Received: 2 February 2025

Accepted: 25 September 2025

Published online: 06 November 2025

Check for updates

Jun-Xiang Zhang[1,2,6], Yuan-De Jin[3,4,6], Chu-Dan Qiu[3,4], Wen-Long Ma[3,4] & Gang-Qin Liu[1,2,5]

Metastability is a ubiquitous phenomenon in nonequilibrium physics and classical stochastic dynamics. It arises when the system dynamics settles in long-lived states before eventually decaying to true equilibria. Remarkably, it has been predicted that quantum metastability can also occur in continuous- and discrete-time open quantum dynamics. However, the direct experimental observation of metastability in open quantum systems has remained elusive. Here, we experimentally observe metastability in the discrete-time evolution of a single nuclear spin in diamond, realized by sequential Ramsey interferometry measurements (RIMs) of a nearby nitrogen-vacancy electron spin. We demonstrate that the metastable polarization of the nuclear spin emerges at a metastable region of measurement times determined by the spectral structure of the quantum channel induced by the RIM. Such metastable polarization enables high-fidelity single-shot readout of the nuclear spin and the observation of an ultralong spin relaxation time of more than 10 s at room temperature. Our results represent a concrete step towards uncovering nonequilibrium physics in open quantum dynamics, which is practically relevant for the utilization of metastable information for various quantum information processing tasks.

Metastability, similar to prethermalization in nonequilibrium physics[1], arises when the system relaxes into long-lived states before subsequently decaying to true stationarity much slowly. The concept of metastability is widely used, for example, in the description of the long-lived atomic state of individual atoms[2], the magnetic state of single nanoparticle[3], and the metastable structure and phase of condensed matter systems[4,5]. In these systems, metastability is mainly due to the barriers separating local minima in the energy levels of the systems.

In classical stochastic dynamics, metastability is a manifestation of a separation of time scales that arises from the splitting in the spectrum of the dynamical generator. Recently, quantum metastability theory has been extended to Markovian open quantum dynamics with a similar spectral structure in the generator[6–8], where the manifold of metastable states is argued to be composed of disjoint states, decoherence-free subspaces, and noiseless subsystems. Such metastability theory is physically interesting in its own right and stimulates the prediction of metastability phenomena in various quantum models[9–17]. It is also of practical importance to utilize the preserved information of quantum processes for quantum information processing[18,19], especially in the presence of control imperfections or environmental noise.

For the continuous-time dynamics described by the Lindblad master equation, metastability can be observed from the distinct two-step decay of temporal correlations[6]. However, experimental observations remain scarce, mainly due to the difficulty in measuring temporal second- or higher-order correlations with good resolution[20,21].

[1]Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing, China. [2]School of Physical Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]State Key Laboratory of Semiconductor Physics and Chip Technologies, Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China. [4]Center of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing, China. [5]CAS Center of Excellence in Topological Quantum Computation, Beijing, China. [6]These authors contributed equally: Jun-Xiang Zhang, Yuan-De Jin. e-mail: wenlongma@semi.ac.cn; gqliu@iphy.ac.cn

https://doi.org/10.1038/s41467-025-64772-6

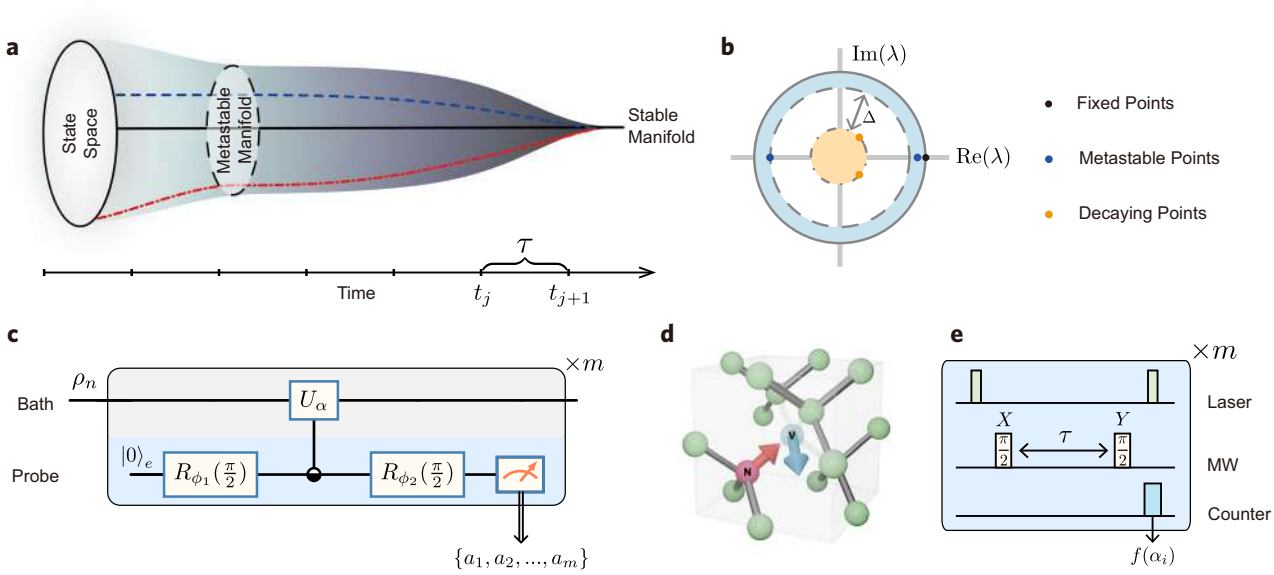


Fig. 1 | Metastability in discrete-time open quantum dynamics and its observation in diamond. a Schematic of metastability in open quantum dynamics. The states outside the metastable manifold (MM, red dot-dashed line) exhibit a two-step relaxation, i.e., first quickly relaxing into the MM and then slowly decaying into the stationary manifold (SM). The discrete-time axis implies that the open quantum dynamics is described by sequential quantum channels. **b** Spectrum of a quantum channel, with all eigenvalues located within the unit circle (solid circle) of the complex plane. Metastable points are located in the region with $|\lambda| \approx 1$ (labeled with blue ring in the spectrum). The metastable region of measurement times is determined by the spectral gap Δ between the decaying region (labeled in orange) and the metastable region in the spectrum. **c** Quantum circuit of sequential RIMs, where $R_\phi(\vartheta)$ is the rotation operator of the probe qubit, U_α is a unitary operator of the bath conditioned on the probe state $|\alpha\rangle_e$ ($\alpha = 0, 1$), and $\{a_1, \ldots, a_m\}$ ($a_i = 0, 1$) is the sequence of measurement outcomes. **d** The experimental system, with the electron spin of a nitrogen-vacancy (NV) center in diamond serving as the probe qubit, and the host ^{14}N nuclear spin as the quantum bath system. **e** Pulse sequences to implement sequential RIMs in experiments. Laser pulses are used to polarize and read out the NV spin state, and resonant microwave (MW) pulses are used to manipulate the probe spin state.

Alternatively, certain signatures of metastability can be observed without measuring the temporal correlations (e.g., through the high-accuracy time-resolved heterodyne detection in the superconducting cavities)[22]; however, such approaches are often not applicable to other experimental settings.

A recent advance in quantum metastability theory is the generalization of the setting from continuous-time to discrete-time open quantum dynamics[23,24]. Such a framework describes the more general scenario in which each discrete evolution is induced by a quantum channel (or a completely positive and trace-preserving quantum map)[25], which may not be generated by continuous-time master equations[26,27]. With this framework, metastable open quantum dynamics of a quantum system can be directly observed by only sequentially measuring an ancilla qubit, instead of continuously measuring the temporal correlations. Remarkably, metastability has been predicted in the commonly used Ramsey interferometry measurements (RIMs)[28–30], which is easy to realize in various quantum platforms. By repeating the RIM on a probe qubit, a nearby bath system can be metastably polarized, which is manifested in statistics of the probe measurement outcomes associated with different quantum trajectories of the bath system.

Here, we report an observation of the metastable open dynamics in diamond based on sequential RIMs, using a nitrogen (^{14}N) nuclear spin as the bath system and the nitrogen-vacancy (NV) electron spin as the probe. Due to the hyperfine coupling between the bath and the probe during free evolution, each round of RIM solely applied on the probe spin induces a quantum channel on the nuclear spin, and sequential RIMs induce discrete-time open quantum dynamics on the nuclear spin (Fig. 1a). By recording the statistical results of the sequential probe measurement, we can continuously monitor the state evolution of the target spin. We directly observe the metastable polarization when varying the number of repetitions m, i.e., the bath spin is first steered to metastable (polarized) states for a finite range of m, and eventually relaxes towards the true stable (maximally mixed) state as m continues to increase.

Results

Metastability in sequential RIMs

We illustrate the principle of metastability by considering a typical class of sequential quantum channels on a quantum bath system, with the channel being induced by a probe qubit under a RIM sequence[23]. The probe qubit is coupled to the bath through the general pure-dephasing coupling

$$H = \sigma_e^z \otimes B_n + \mathbb{I}_e \otimes C_n, \tag{1}$$

where the subscripts e and n refer to the probe and bath, respectively, σ_e^i is the Pauli-i operator of the qubit ($i = x, y, z$) with $\sigma_e^z = |0\rangle_e\langle 0| - |1\rangle_e\langle 1|$, $\mathbb{I}_e$ is the identity operator of the qubit, and $B_n(C_n)$ is the interaction (free) operator of the bath. We denote the probe qubit rotation along an axis in the equatorial plane as $R_\phi(\vartheta) = e^{-i(\cos\phi\sigma_e^x + \sin\phi\sigma_e^y)\vartheta/2}$, with ϕ denoting the rotation axis and ϑ the rotation angle.

For a single RIM, the probe qubit is first initialized to state $|0\rangle_e$, and then prepared in a superposition state $|\psi\rangle_e = (|0\rangle_e - ie^{i\phi_1}|1\rangle_e)/\sqrt{2}$ by a rotation $R_{\phi_1}(\frac{\pi}{2})$. After the composite system evolving with the pure-dephasing coupling [Eq. (1)] for time τ, the probe undergoes another rotation $R_{\phi_2}(\frac{\pi}{2})$, and is finally projectively measured in the basis of σ_e^z with the measurement outcome $a \in \{0, 1\}$.

The RIM sequence of the probe spin induces a quantum channel Φ on the bath,

$$\Phi(\rho_n) = \mathrm{Tr}_e\left[U(\rho_e \otimes \rho_n)U^\dagger\right] = \sum_{a=0,1} M_a \rho_n M_a^\dagger, \tag{2}$$

where $\rho_e = |\psi\rangle_e\langle\psi|$ (ρ_n) is the initial state of the probe (bath) spin, $U = e^{-iH\tau}$ is the propagator for the composite system. By partially tracing over the probe spin, we obtain the Kraus operator $M_a = [U_0 - (-1)^a e^{i\Delta\phi} U_1]/2$, where $U_\alpha = e^{-i[(-1)^\alpha B_n + C_n]\tau}$ is a unitary operator of the bath spin conditioned on the probe spin state $|\alpha\rangle_e$ ($\alpha = 0, 1$), and $\Delta\phi = \phi_1 - \phi_2$ is the phase difference between the rotation axes of two $\pi/2$ pulses in a RIM. With the measurement outcome a, the bath state is steered to $M_a\rho_n M_a^\dagger/p(a)$ with probability $p(a) = \mathrm{Tr}(M_a\rho_n M_a^\dagger)$. Note that the Kraus operators satisfy $\sum_a M_a^\dagger M_a = \mathbb{I}_n$ with $\mathbb{I}_n$ being the identity operator on the bath, which ensures that $\sum_a p(a) = 1$.

The behaviors of repetitive quantum channels can be clearly revealed by the spectral decomposition of a single channel (see "Methods"). A quantum channel has at least one fixed point (Fig. 1b), which is the state that remains unchanged after the channel[31,32]. It has been proved that the fixed points of the channel in Eq. (2) depend on the commutativity of B_n and C_n (see "Methods")[23]. If B_n commutes with C_n, i.e., $[B_n, C_n] = 0$, the fixed points include the simultaneous eigenstates of B_n and C_n, then sequential RIMs of the probe spin will polarize the bath system to one of the eigenstates, also constituting a quantum non-demolition (QND) measurement on the bath spin[33–38]. However, if $[B_n, C_n] \neq 0$, the fixed points are the maximally mixed state in the whole space (or subspace) of the bath, then sequential RIMs cannot effectively measure the bath system but only perturb it.

Quantum metastability emerges when $[B_n, C_n] \neq 0$, but C_n is a small perturbation of B_n, then the bath will be polarized for a finite range of repetitions of RIMs. We suppose that the channel has r fixed points and $q - r$ ($q > r$) metastable points, which collectively span a $(q - 1)$-dimensional metastable manifold (MM). The region of repetition numbers for metastable polarization of the bath spin can be estimated as $(1 - |\lambda_{q+1}|)^{-1} \ll m \ll (1 - |\lambda_q|)^{-1}$ when λ_q is very close to 1, where λ_q is the eigenvalue of metastable point with the smallest eigenvalue. As m increases and enters the metastable region, the bath state evolves into the MM. Beyond this region, the system further relaxes into the stationary manifold (SM), which is spanned by the fixed points. This process, during which the system first reaches the MM and then relaxes to the SM, is referred to as a two-step relaxation (see Fig. 1a).

The metastability behaviors in sequential RIMs can be directly monitored by the measurement statistics of the probe spin[23]. To see this, we decompose the average dynamics of sequential quantum channels into stochastic trajectories,

$$\Phi^m(\rho_n) = \sum_{a_1,\ldots,a_m=0}^{1} \mathcal{M}_{a_m}\cdots\mathcal{M}_{a_1}(\rho_n), \tag{3}$$

where $\mathcal{M}_{a_i}(\cdot) = M_{a_i}(\cdot)M_{a_i}^\dagger$ with $a_i \in \{0, 1\}$ being the measurement outcome of the ith RIM. For a trajectory with the sequence of measurement outcomes $\{a_1, \ldots, a_m\}$, the bath is steered to $\rho_n' = \mathcal{M}_{a_m}\cdots\mathcal{M}_{a_1}(\rho_n)/p(a_1,\ldots,a_m)$ with probability $p(a_1,\ldots,a_m) = \mathrm{Tr}[\mathcal{M}_{a_m}\cdots\mathcal{M}_{a_1}(\rho_n)]$. With the number of outcome 0(1) in $\{a_1, \ldots, a_m\}$ denoted as $m_0(m_1)$ satisfying $m_0 + m_1 = m$, we can define the measurement polarization $X = (m_0 - m_1)/(2m)$[23,24,39], which can be used to classify different classes of stochastic trajectories that the bath system undergoes. The measurement distribution of X can show several peaks, with each peak corresponding to the quantum trajectories that lead to a fixed point (or metastable state) of the channel (see "Methods"). So in the metastable region, the distribution exhibits up to q peaks, corresponding to the metastable states. However, as m surpasses this region, the number of peaks gradually reduces to r, reflecting the final stationary states of the bath.

Experimental implementation in the NV system

We demonstrate this protocol experimentally on an NV center and a nearby ^{14}N nuclear spin in a high-purity diamond (Fig. 1d). Nuclear spins around NV centers have great potential for building quantum networks[40], storing quantum information[38], performing quantum simulations[41–43] and sensing inertial parameters[44]. Access to the nuclear spins is usually based on their hyperfine interaction with the electron spin and the optical interface of the NV center. It is therefore an interesting and fundamental problem to monitor and understand the dynamics of these nuclear spins under sequential measurements acting on the central electron spin.

Specifically, an NV electron spin is a spin-1 system with a zero-field splitting $D = 2.87$ GHz between its $|0\rangle_e$ ($m_s = 0$) and $|\pm1\rangle_e$ ($m_s = \pm1$) state. In our experiment, a moderate external magnetic field of $B = 108.4 \pm 0.2$ G is applied along the NV symmetry axis (z axis) to lift the degeneracy of the $|\pm1\rangle_e$ states and we work in the $\{|0\rangle_e, |-1\rangle_e\}$ subspace. The Hamiltonian in Eq. (1) becomes (in the rotating frame)

$$H = A_{zz}S_zI_z + QI_z^2 - \gamma_n\mathbf{B}\cdot\mathbf{I} + \tilde{H}_1, \tag{4}$$

where $\mathbf{S} = (S_x, S_y, S_z)$ is the NV electron spin-1 operator, $\mathbf{I} = (I_x, I_y, I_z)$ is the ^{14}N nuclear spin-1 operator, $A_{zz} = -2.16$ MHz ($A_\perp = -2.63$ MHz) is the zz-component (transverse component) of the hyperfine tensor, $Q = -4.95$ MHz is the quadrupolar splitting, $\gamma_n = 1.071$ kHz/G ($\gamma_e = 2.802$ MHz/G) being the gyromagnetic ratio of ^{14}N nuclear spin (electron spin), and $\tilde{H}_1 = \sum_\alpha |\alpha\rangle_e\langle\alpha| \otimes H_n^\alpha$ is a second-order perturbation term with $H_n^\alpha \approx \frac{\gamma_e(2-3|\alpha|)}{2D}[-\gamma_e(B_x^2 + B_y^2) + 2A_\perp(B_xI_x + B_yI_y)]$. To introduce a non-commutativity between $B_n = A_{zz}I_z$ and $C_n = QI_z^2 - \gamma_n\mathbf{B}\cdot\mathbf{I} + \tilde{H}_1$, the external magnetic field is slightly misaligned with the NV axis, i.e., there is a small tilt angle θ between the magnetic field and NV axis. Under this circumstance, the perturbation terms $-\gamma_n(B_xI_x + B_yI_y)$ and $\tilde{H}_1$ lead to the metastable polarization of ^{14}N nuclear spin.

For each RIM, the resulting quantum state of the NV spin is read out by counting its spin-dependent photon number. As shown in Fig. 1e, a 532-nm laser is used to excite the NV center, and all experiments are performed at room temperature. Due to the low photon emission rate of the NV center and the poor collection efficiency, only 0.05 photons can be detected on average in each measurement, which is not sufficient to distinguish the spin states. Fortunately, the spin relaxation time of the ^{14}N nuclear spin is very long (see "Discussion" below), and many RIMs can be performed before a state jump occurs. Therefore, to obtain a reasonable signal-to-noise ratio (SNR) and to distinguish the nuclear spin states, we simply sum up the counts of the sequential measurements. We expect that the metastable polarization is easier to observe by readout methods with higher SNR, for example, the resonant excitation of the NV electron spin at low temperature[45,46].

Figure 2 presents the metastable dynamics of the target nuclear spin under sequential RIMs acting on the NV electron spin. Before the measurement, the ^{14}N nuclear spin is in the thermal state, which means that its population is evenly distributed among the set of states $\{|0\rangle_n, |\pm1\rangle_n\}$ ($m_I = 0, \pm1$). For a small number of RIMs ($m < 5$ k), the shot noise of the photon counts exceeds the count difference of the measured states, so that no efficient signal can be detected in this range. As the number of RIMs increases ($m > 5$ k), the measurement results begin to concentrate on two regions, which is a clear signature of the metastable polarization of the bath spin, as can be seen in the middle slice of Fig. 2a, b. We classify the trajectories based on the average photon counts with a threshold value of 0.05. If the average photon counts are less than 0.05, it is classified as a *dark* state; otherwise as a *bright* state. The experimentally resolved bright state corresponds to the maximally mixed state of the subspace spanned by $\{|0\rangle_n, |-1\rangle_n\}$, i.e., $\rho_B = (|0\rangle_n\langle0| + |-1\rangle_n\langle-1|)/2$. This bright state, together with the dark state $\rho_D = |1\rangle_n\langle1|$, are the extremely metastable states (EMSs) of the one-dimensional MM spanned by the fixed point and the metastable point with higher eigenvalue, as shown in Fig. 2d. Beyond the metastable region ($m > 600$ k), the two peaks mentioned above gradually disappear and a single peak appears, corresponding to the stable state of the system, i.e., the maximally mixed state in the

https://doi.org/10.1038/s41467-025-64772-6

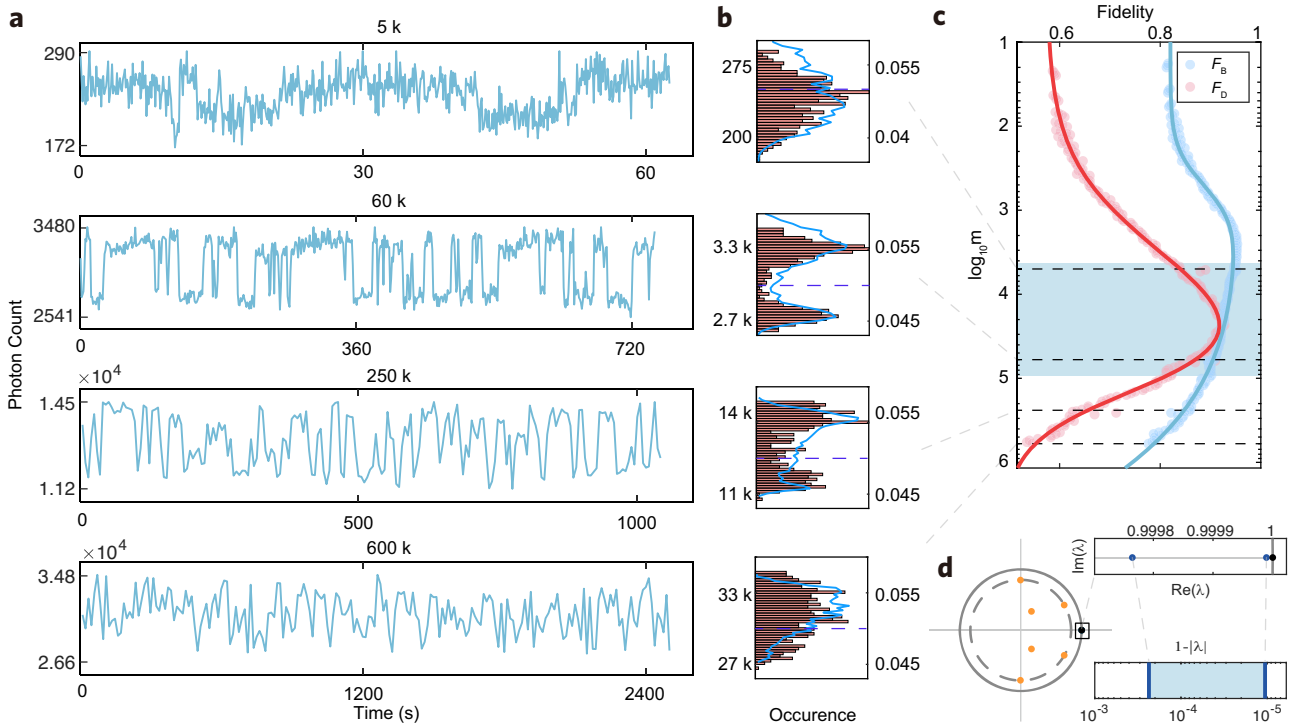


Fig. 2 | Metastable dynamics of a ^{14}N nuclear spin induced by sequential RIMs of a nearby NV electron spin. a, **b** Typical PL time trace of an NV center under sequential RIMs and the corresponding distribution of the measurement results. The total and average photon counts are shown on the left and right axes, respectively. The duration of free evolution, $\tau = 374$ ns. The number of measurement repetition m is 5 k, 60 k, 250 k, 600 k from top to bottom. When $m = 5$ k, it is difficult to distinguish the nuclear spin quantum states. When $m = 60$ k, the jump signal is concentrated in two distinct photon count intervals. As m increases further to 250 k, the overlap between the two distribution peaks becomes more pronounced. When m exceeds the metastable region, the photon counting peaks gradually merge. The blue envelopes in (**b**) indicate the normalized numerical simulation results. **c** Numerical results. The evolution of the nuclear spin state fidelity under sequential RIMs. The initial thermal state of the nuclear spin is polarized to either the dark state $|1\rangle_n\langle 1|$ or the bright state $(|0\rangle_n\langle 0| + |-1\rangle_n\langle -1|)/2$. The solid lines are fits to the simulation results. The fidelity $F_{\mathrm{D,B}}$ quantifies how close the quantum states represented by the trajectories are to the ideal dark or bright states, respectively. **d** The channel spectrum, with one fixed point, two metastable points and six decaying points. The right pane is a magnified view of the spectrum near the metastable points, and the deviations of their eigenvalues from 1 ($1 - |\lambda|$) are shown accordingly. The metastable region and the underlying channel spectrum gap are marked with blue shades in (**c**) and (**d**), respectively. Monte Carlo simulations are conducted with a tilt angle $\theta = 8.8°$ and 3000 samples.

whole space $\mathbb{I}_n/3$. More data and simulations can be found in Figs. S4–S6.

To quantitatively understand the metastable dynamics of the nuclear spin, we perform numerical simulations that take into account the parameters of hyperfine coupling, the strength and orientation of the external magnetic field, and the readout efficiency of the NV electron spin. The metastable states can be characterized by comparing them with the ideal polarized states, i.e., by the fidelity $F_{\mathrm{D(B)}} = F(\rho_{\mathrm{D(B)}}, \rho_n')$, where $F(\rho,\sigma) = \mathrm{Tr}\sqrt{\rho^{1/2}\sigma\rho^{1/2}}$ and ρ_n' is the actual state. As shown in Fig. 2c, the experimentally observed metastable dynamics of the ^{14}N can be well reproduced with a tilt angle of the magnetic field of $\theta = 8.8°$.

Since the ^{14}N nuclear spin is a three-level system, three-state jumps are expected in the PL time trace. The fact that only two stages appear indicates that two of them cannot be distinguished under the current experimental conditions ($B = 108.4$ G and $\theta = 8.8°$). Interestingly, when a smaller angle of the external magnetic field ($\theta < 2°$) is chosen, three-state jumps are observed in the experiment, as shown in Fig. 3a, b. The physical picture behind this is that the external magnetic field can influence the metastability behavior by changing the value of $[B_n, C_n]$. Through numerical simulations we find that under the magnetic field with $\theta = 8.8°$, three peaks emerge at around 5 k measurements and two of them merge into one before 60 k measurements, so that only two-state jumps are observed in our measured region (Fig. 2a). Figure 3c shows the calculated dependence of the two-state jump and three-state jump regions as a function of the orientation of the magnetic field. For a smaller tilt angle of the external magnetic field (e.g., $\theta = 2°$), the metastable region of the three-state jump can survive under a large number of RIM measurements, which allows the experimental observation of the three-peak signals (see Supplementary Note II for more details).

Single-shot readout of a single nuclear spin

The observed metastability provides a simple and efficient way to realize high-fidelity single-shot readout of the nuclear spin. Under a magnetic field tilt angle $\theta = 7° \pm 2°$ and a RIM interval of $\tau = 374$ ns, the quantum jumps of the nuclear spin are clearly observed in the range of measurement numbers from 5 k to 250 k, as shown in Fig. 2a. We then characterize the readout fidelity of the nuclear spin using the method developed in ref. 36,47. By pre-selecting bright-state and dark-state photon count thresholds, we filter quantum jump signals across different measurement repetitions. This results in photon number distributions for bright and dark states with initialization fidelity higher than 99%. We then quantify the state-dependent readout fidelities, as summarized in Fig. 4a and Fig. S9b, c. It is worth noting that both the external magnetic field and the RIM interval affect the occurrence of the metastable states, so as to the optimal readout region. With a repetition number of $m = 80$ k (total measurement time 1.695 s), a readout fidelity of 97% ± 2% with a threshold count value of 2950 is achieved.

Next, we investigate the relaxation mechanism of the ^{14}N nuclear spin under repeated measurements. Previous studies have shown that the flip-flop process in the excited state of the NV center is the dominant factor, and usually a large external magnetic field (>2000 G) is

https://doi.org/10.1038/s41467-025-64772-6

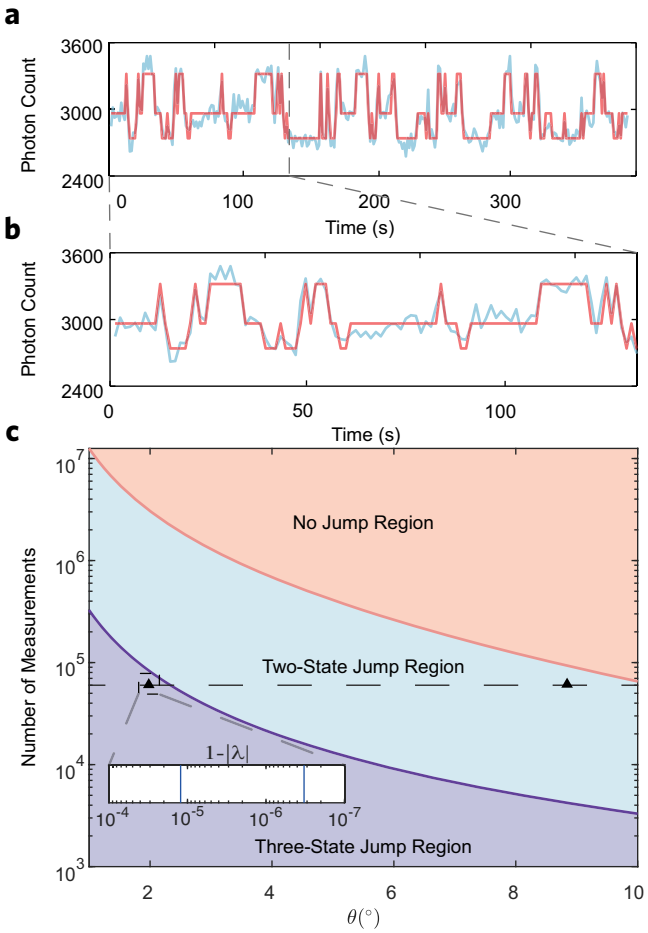


Fig. 3 | Magnetic field dependence of the metastable region. a Three-state jump signal of the ^{14}N nuclear spin. The signal is observed when the magnetic field is close to the NV axis ($\theta = 2°$), with a RIM duration $\tau = 374$ ns and the number of repetitions $m = 60$ k. The red fitting line is extracted from the hidden Markov model. **b** The first 100 points of of three-state jump signal. The three plateau levels correspond to the states $|+1\rangle_n$, $|0\rangle_n$, and $|-1\rangle_n$ of the ^{14}N nuclear spin. **c** Phase diagram for metastability. The metastable region for the number of measurements m depends on the orientation of the magnetic field. The dashed line marks the case of $m = 60$ k, with the two black triangles indicating experimental conditions used. The left point with $\theta = 2°$ gives the three-state jump signals explicitly shown in (**a**) and (**b**). The corresponding metastable region, quantified by the deviations of their eigenvalues from 1, is shown in the inset. The right point with $\theta = 8.8°$ corresponds to the two-state jump signal shown in Fig. 2a.

applied to suppress this term so that single-shot readout of the nuclear spin can be achieved[36,37]. Alternatively, our results show that the nuclear spin relaxation can be largely suppressed by simply shortening the duration of the excitation laser. Moreover, by inserting additional waiting intervals between the sequential RIMs, a longer nuclear spin T_1 is observed, as summarized in Fig. 4b. The longest T_1 measured in our experiment is 15 ± 3.5 s. To our knowledge, this is the first time that a single-shot readout of the ^{14}N nuclear spin has been achieved at room temperature and under a small magnetic field.

Finally, we perform nuclear magnetic resonance (NMR) and Rabi measurements to demonstrate coherent control of the ^{14}N nuclear spin. The nuclear spin state is first prepared by a RIM measurement sequence (either in the bright or dark state), then an RF pulse is applied to flip the nuclear spin, and a second RIM measurement is performed to determine the nuclear spin state. The RF pulse has either a fixed duration (30 μs for NMR) or a fixed frequency (at the resonance frequency of NMR). The sequence is repeated 80 k times, and the probability of nuclear spin flip is recorded. Figure 4c shows the NMR signal at 4973.0 kHz, which corresponds to the ^{14}N nuclear spin being in $|+1\rangle_n$ state. Figure 4d shows the Rabi oscillation of the ^{14}N nuclear spin between the states $|+1\rangle_n$ and $|0\rangle_n$. These results further confirm that the observed quantum jump signal originates from the ^{14}N nuclear spin.

Discussion

In conclusion, we have directly observed metastability in the discrete-time open quantum dynamics of a single nuclear spin in diamond, induced by sequential RIMs of a probe NV electron spin. The observed metastable polarization of the nuclear spin demonstrates that QND measurements of an arbitrary quantum system can be realized in a much broader range of conditions, and go far beyond the QND experiments where the ancilla and the measured system are tuned to maximally entangled states[36–38]. Such schemes can also be easily extended to other physical platforms, including trapped ions, superconducting circuits, and other semiconductor single-spin systems.

The framework for constructing QND measurements can be generalized to implement dynamical-decoupling-based quantum metrology in noisy environments, which enables us to learn the probe-bath and bath-bath coupling strengths[48,49], the magnetic field[46,50,51] and bath spin polarization[52,53]. On the other hand, our results point out a potential issue in constructing quantum networks with NV centers or other single-spin systems, since sequential measurements of an ancilla can cause its nearby memory to equilibrate towards a maximally mixed state, leading to the loss of information about the stored state.

The generality of our protocols also sets a blueprint for future studies of nonequilibrium phenomena in open quantum systems. It will be interesting future work to realize more general channels in both theory and experiments, whose fixed points may include decoherence-free subspaces or noise-free subsystems, and to explore the metastability phenomena for protected coding and logical gates. Such phenomena can be directly relevant to quantum information processing, since the fixed points of a quantum channel with preserved information can become metastable when control errors or environmental noise are taken into account.

Methods

Experimental setup and sample information

The experiments are performed with a [100]-oriented, type IIa single-crystal diamond produced by Element Six, with a natural ^{13}C isotopic abundance of 1.1%. The NV centers used in this study are generated through electron irradiation and subsequently annealed under vacuum conditions. To improve photon collection efficiency, a solid immersion lens (SIL) is etched onto the diamond surface. Under saturation conditions, the photon counting rate of the NV center is about 300 kcps. The diamond was mounted on a custom-built confocal microscope, and a 532-nm laser was used to initialize and read out the NV spin states. The laser pulses are timing-controlled with an acoustic optical modulator (AOM, Gooch & Housego 3350-199) and focused on the diamond sample through an objective (NA = 0.9). The fluorescence emitted from the NV center is collected with the same objective, passed through a dichroic mirror and a 650-nm long-pass filter, then coupled into a multimode fiber and detected with a single photon detector (SPD). The fluorescence photons are counted with an data acquisition card (NI-6343).

To control the NV electron spin and the ^{14}N nuclear spins, we use an arbitrary waveform generator (AWG, Techtronix 5024C) to generate transistor-transistor logic (TTL) signals and low-frequency analog signals. The synchronized TTL signals are used to control the AOM, RF switches and counters. For electron spin manipulation, the microwave (MW) carrier signal from a MW source (Rohde & Schwarz SMIQ03B) is combined with two analog signals using an IQ-mixer (Marki Microwave, IQ1545LMP). An additional analog channel is used to generate the RF signal for nuclear spin manipulation. Both the MW and RF

https://doi.org/10.1038/s41467-025-64772-6

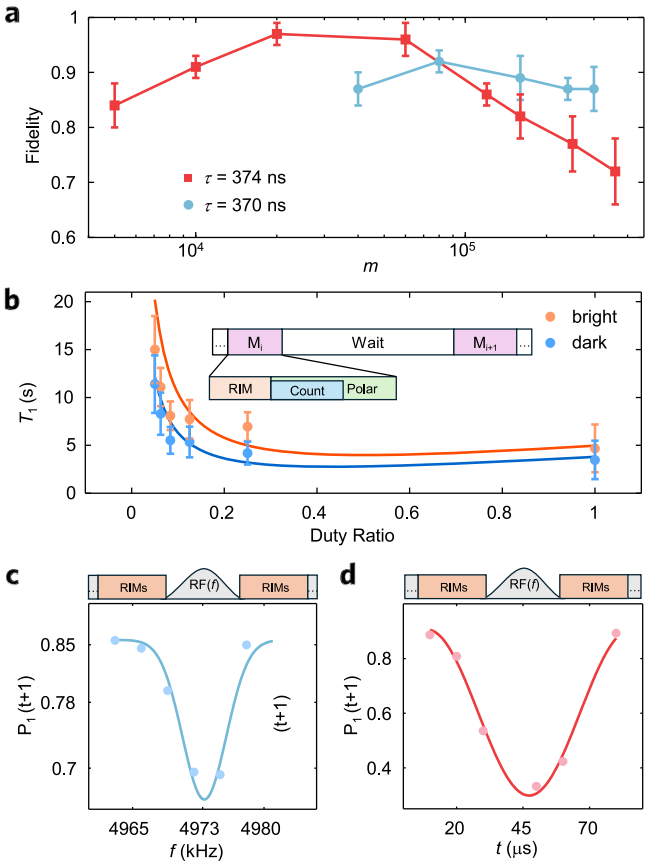


Fig. 4 | Single-shot readout of the nuclear spin state. a Readout fidelity of the ^{14}N nuclear spin as a function of the measurement repetition number. The data points with $\theta = 7° \pm 2°$, $\tau = 374$ ns are shown in red squares, while the data points with $\theta = 7° \pm 2°$, $\tau = 370$ ns are shown in blue circles. **b** ^{14}N nuclear spin relaxation time (bright state in orange and dark state in blue) as a function of the duty cycle of the excitation laser pulses. The inset illustrates the experimental pulse sequence. The duty cycle is defined as the ratio of the measurement time to the waiting time in the sequence. In (**a**) and (**b**), the vertical error bars display the statistical error (s.d.) from the fit. **c** ^{14}N nuclear spin NMR, obtained by varying the radio-frequency (RF) pulse frequency at a fixed pulse duration of 30 μs. **d** ^{14}N nuclear spin Rabi oscillations. A resonant RF pulse of a specific length is applied between every two sequential RIMs to flip the nuclear spin between the $|0\rangle_n$ and $|+1\rangle_n$ states, while the electron is at the $|0\rangle_n$ state.

signals are amplified by amplifiers (Mini Circuits, ZHL-16W-43-S+ and ZHL-32A-S+). The MW and RF signals are combined by a customer-designed duplexer (DC-200 MHz, 1200 MHz-4000 MHz) and transmitted to the NV position via a gold co-planar waveguide deposited on the diamond surface.

As can be seen in the ODMR spectrum of the NV center, there is no splitting due to hyperfine interaction with strongly coupled ^{13}C nuclear spin, and a static magnetic field $B = 108.4 \pm 0.2$ G is applied close to the NV symmetry axis. The transition between the $|0\rangle_e \leftrightarrow |-1\rangle_e$ NV spin states is addressed via resonant microwave pulses (at the $m_I = 0$ state of the ^{14}N nuclear spin). To compensate for the temperature drift of the experimental system, we re-measured the ODMR spectra and adjusted the microwaves accordingly before each of the sequential RIMs. Figure S8b shows the free-induction decay (FID) signal of the measured NV center, which yields $T_{2e}^{*} = 1.15\,\mu$s.

For the ^{14}N nuclear spin, the measurement strength oscillates as a function of the RIM duration τ. Considering the nuclear spin relaxation and the NV electron spin dephasing time, $\tau = 374$ ns is chosen to perform the experiments. Under these circumstances, the dynamics induced by sequential RIMs precedes the ^{14}N nuclear spin T_1 process. For each RIM, an optical pulse of 220 ns is used to read out the population of the NV electron spin. After the first measurement, the nuclear spin randomly collapses to $|0\rangle_n$ or $|\pm 1\rangle_n$, and in the following measurement, it remains at the same state until a quantum jump occurs. Results with more τ value can be found in the Supplementary Information Note III.

Quantum channels and their representations

Quantum channel, also called completely positive and trace-preserving (CPTP) map, describes the most general (closed or open) quantum dynamics that a quantum system can undergo. Each quantum channel has four different representations: the Kraus representation, the Stinespring representation, the natural representation, and the Choi representation. The Kraus representation is the most commonly used one and characterizes a channel by a set of Kraus operators $\{M_a\}_{a=1}^{r}$, which satisfy $\sum_{a=1}^{r} M_a^{\dagger} M_a = \mathbb{I}$, such that $\Phi(\cdot) = \sum_{a=1}^{r} M_a(\cdot) M_a^{\dagger} = \sum_{a=1}^{r} \mathcal{M}_a(\cdot)$. The Stinespring representation is a dilation of a quantum channel, which is realized by coupling the bath system to a probe system, subjecting the composite system to a unitary evolution and then tracing over the probe system, i.e., $\Phi(\rho_n) = \mathrm{Tr}_e[U(\rho_e \otimes \rho_n)U^{\dagger}]$, where $\mathrm{Tr}_e[\cdot]$ denotes the partial trace over the probe qubit.

Since a quantum channel acts on the operator space of the bath system and can be considered as a superoperator, it is more convenient to use its natural representation in the Hilbert-Schmidt (HS) space. In the HS space, an operator is transformed into a vector ($X = \sum_{i,j=1}^{d} x_{ij} |i\rangle\langle j| \leftrightarrow |X\rangle\rangle = \sum_{i,j=1}^{d} x_{ij} |ij\rangle\rangle$), so that each superoperator is transformed into a single matrix ($X(\cdot)Y \to X \otimes Y^T$). Thus, the natural representation of Φ is $\hat{\Phi} = \sum_{a=1}^{r} \hat{\mathcal{M}}_a$ with $\hat{\mathcal{M}}_a = M_a \otimes M_a^{*}$. Note that we add hats on the operators of the HS space. In the natural representation, we can spectrally decompose the quantum channel as

$$\hat{\Phi} = \sum_i \lambda_i |R_i\rangle\rangle\langle\langle L_i|, \tag{5}$$

where λ_i is the ith eigenvalue with the corresponding right (left) eigenvector $|R_i\rangle\rangle$ ($|L_i\rangle\rangle$), satisfying $\hat{\Phi}|R_i\rangle\rangle = \lambda_i |R_i\rangle\rangle$, $\hat{\Phi}^{\dagger}|L_i\rangle\rangle = \lambda_i^{*}|L_i\rangle\rangle$, and the biorthonormalization condition $\langle\langle L_i|R_j\rangle\rangle = \mathrm{Tr}(L_i^{\dagger} R_j) = \delta_{ij}$ with δ_{ij} being the Kronecker delta. The eigenvalues $\{\lambda_i\}$ of a quantum channel are all located within a unit disk of the complex plane. The eigenspaces with $\lambda = 1$ are called fixed points denoted by $|\rho_{\mathrm{fix}}^{i}\rangle\rangle$, those with $\lambda = e^{i\varphi}$ and $\varphi \neq 0$ are rotating points, and those with $|\lambda| < 1$ are decaying points that decay fast under repetitive channels since $\hat{\Phi}^{m}|R_j\rangle\rangle = \lambda_j^{m}|R_j\rangle\rangle \to 0$ for $|\lambda_j| < 1$.

Metastability in sequential quantum channels

Of particular interest are the decaying points with eigenvalue $|\lambda_i| \approx 1$, which are called metastable points Quantum metastability emerges when there are metastable points in the channel spectrum[23]. Although the information of these metastable points will be finally lost, it can be well preserved within a certain range of m. Specifically, for the channel with r fixed points and $q - r$ metastable fixed points, the contributions of the metastable points cannot be neglected when $m \ll \mu' = 1/|\ln|\lambda_q||$, while the contribution of the other decaying points decays fast as m grows and can be omitted when $m \gg \mu'' = 1/|\ln|\lambda_{q+1}||$. So μ' and μ'' delimit a metastable region:

$$\frac{1}{|\ln|\lambda_{q+1}||} \ll m \ll \frac{1}{|\ln|\lambda_q||}. \tag{6}$$

When the eigenvalue of metastable points is very close to 1, and the decaying points relax fast, which is suitable for our case, the

metastable region can be estimated as

$$m \ll \frac{1}{|\ln|\lambda_q||} \approx \frac{1}{1-|\lambda_q|} \tag{7}$$

In the metastable region, $|\lambda_i|^m \approx 1$ for $i \le q$ and $|\lambda_j|^m \approx 0$ for $j > q$, then we have

$$\hat{\Phi}^m|\rho\rangle\rangle \simeq \sum_{i=1}^{r} c_i|\rho_{\text{fix}}^i\rangle\rangle + \sum_{j=r+1}^{q} \tilde{c}_j|R_j\rangle\rangle, \tag{8}$$

for initial state $|\rho\rangle\rangle = \sum_i c_i|R_i\rangle\rangle$ with $\tilde{c}_j = c_j e^{im\varphi_j}$. We note that the metastable points, as well as decaying points, are trace-zero, so any metastable state in metastable manifold is a superposition of both fixed points (trace-one) and metastable points, denoted by $\{c_1, \ldots, c_r, \tilde{c}_{r+1}, \ldots, \tilde{c}_q\}$ with $\sum_{i=1}^r c_r = 1$, corresponding to a point in the $(q-1)$ dimensional HS subspace.

We can transform from the HS subspace to the metastable manifold (MM), which contains the metastable states as a convex combination of q extreme metastable states (EMSs)

$$|\rho_{\text{MS}}\rangle\rangle = \sum_{\nu=1}^{q} p_\nu|\rho_{\text{EMS}}^\nu\rangle\rangle \tag{9}$$

with $p_\nu = \langle\langle P_\nu|\rho_{\text{MS}}\rangle\rangle$ satisfying $\sum_\nu p_\nu = 1$. Here, P_ν is the projector to $|\rho_{\text{EMS}}^\nu\rangle\rangle$, satisfying $\langle\langle P_\nu|\rho_{\text{EMS}}^\mu\rangle\rangle = \delta_{\nu\mu}$ and $\sum_\nu P_\nu = \mathbb{I}$.

For the case of $r = 1$ and $q = 2$, i.e., the channel has one fixed point and one metastable point, forming a one-dimensional MM with the EMSs[23]

$$|\rho_{\text{EMS}}^{1,2}\rangle\rangle = |\rho_{\text{fix}}\rangle\rangle + c_2^{M,m}|R_2\rangle\rangle/h, \tag{10}$$

where c_2^M (c_2^m) is the maximal (minimal) eigenvalue of L_2, and $h = \sqrt{\langle\langle L_2|L_2\rangle\rangle\langle\langle R_2|R_2\rangle\rangle}$ is a normalization coefficient.

Metastability in sequential RIMs

The dynamics of the bath system after a RIM sequence of the probe is described by a quantum channel, which can be represented in the Stinespring representation as

$$\Phi(\rho_n) = \text{Tr}_e[U(\rho_e \otimes \rho_n)U^\dagger] \tag{11}$$

where $\rho_e = R_{\phi_1}(\frac{\pi}{2})|0\rangle_e\langle 0|R_{\phi_1}^\dagger(\frac{\pi}{2})$. For the Hamiltonian

$$H = \sigma_e^z \otimes B_n + \mathbb{I}_e \otimes C_n, \tag{12}$$

the unitary evolution can be decomposed as $U = e^{-iH\tau} = \sum_{\alpha=0,1}|\alpha\rangle_e\langle\alpha| \otimes U_\alpha$ with $U_\alpha = e^{-i[(-1)^\alpha B_n + C_n]\tau}$. After partially tracing over the probe, the channel can be transformed to the Kraus representation as

$$\Phi(\rho_n) = \sum_{a=0,1} \mathcal{M}_a(\rho_n) = \sum_{a=0,1} M_a\rho_n M_a^\dagger, \tag{13}$$

where $\mathcal{M}_a(\cdot) = M_a(\cdot)M_a^\dagger$ is a superoperator with the Kraus operator $M_a = [U_0 - (-1)^a e^{i\Delta\phi}U_1]/2$ and $\Delta\phi = \phi_1 - \phi_2$ is the phase difference between the rotation axes of $R_{\phi_1}(\frac{\pi}{2})$ and $R_{\phi_2}(\frac{\pi}{2})$. Then, the natural representation of the channel for the RIM case is

$$\hat{\Phi} = \hat{\mathcal{M}}_0 + \hat{\mathcal{M}}_1 = (\hat{\mathcal{U}}_0 + \hat{\mathcal{U}}_1)/2, \tag{14}$$

where $\hat{\mathcal{M}}_a = M_a \otimes M_a^*$ and $\hat{\mathcal{U}}_\alpha = U_\alpha \otimes U_\alpha^*$.

It has been proved that the spectrum of the quantum channel induced by RIM is determined by the commutativity of B_n and C_n[23]. If $[B_n, C_n] = 0$, then B_n and C_n can be diagonalized simultaneously

$$B_n = \sum_{k=1}^{d} b_k|k\rangle\langle k|, \quad C_n = \sum_{k=1}^{d} h_k|k\rangle\langle k|. \tag{15}$$

The fixed points are spanned by all the rank-1 projectors $\{|k\rangle\langle k|\}_{k=1}^d$. While if $[B_n, C_n] \neq 0$, B_n and C_n can only be block diagonalized with a unitary transformation,

$$B_n = W\left(\bigoplus_{j=1}^r B_{n,j}\right)W^\dagger, \quad C_n = W\left(\bigoplus_{j=1}^r C_{n,j}\right)W^\dagger \tag{16}$$

where r ($<d$) is the number of blocks. Such a block diagonalization partitions the Hilbert space of the bath system into a direct sum of r subspaces $\mathcal{H} = \bigoplus_{j=1}^r \mathcal{H}_j$. Then it can be proved that the fixed points in this case are spanned by all the projectors $\{\Pi_j\}_{j=1}^r$ to the blocks.

Moreover, it has also been demonstrated that when C_n is a perturbation of B_n, i.e., $\|C_n\|/\|B_n\|$ is small with $\|\cdot\|$ denoting the operator norm, the EMSs are approximately the eigenstates of B_n[23].

Model for the NV system

The full Hamiltonian for the system containing the NV electron and the ^{14}N nuclear spin is

$$H = DS_z^2 + \gamma_e\mathbf{B}\cdot\mathbf{S} + \mathbf{S}\cdot\mathbb{A}\cdot\mathbf{I} + QI_z^2 + \gamma_n\mathbf{B}\cdot\mathbf{I}, \tag{17}$$

where $\mathbb{A}$ is the hyperfine interaction tensor, containing only the zz-component $A_{zz}S_zI_z$ and the transverse component $A_\perp(S_xI_x + S_yI_y)$ for the ^{14}N nuclear spin. Since the large zero-field splitting of the NV electron spin suppresses the transition between $|m_s = 0\rangle$ and $|m_s = \pm 1\rangle$, the Hamiltonian Eq. (17) can be well approximated as a pure-dephasing form in Eq. (1) under the second-order perturbation[54].

Let $H_0 = DS_z^2 - \gamma_e B_z S_z + S_z A_{zz} I_z + QI_z^2 - \gamma_n\mathbf{B}\cdot\mathbf{I}$ and $H_1 = -\gamma_e(B_xS_x + B_yS_y) + A_\perp(S_xI_x + S_yI_y)$, then H_1 is a perturbation on H_0. The second-order perturbation gives $\tilde{H}_1 = \sum_\alpha|\alpha\rangle_e\langle\alpha| \otimes H_n^\alpha$, in which

$$H_n^\alpha = P_\alpha^e H_1 \frac{1}{E_\alpha - (\mathbb{I}_e - P_\alpha^e)H_0(\mathbb{I}_e - P_\alpha^e)} H_1 P_\alpha^e, \tag{18}$$

with $E_\alpha = D\alpha^2 - \gamma_e B_z\alpha$. Specifically, we have

$$H_n^\alpha \approx \frac{\gamma_e(2-3|\alpha|)}{2D}\left[-\gamma_e\left(B_x^2 + B_y^2\right) + 2A_\perp\left(B_xI_x + B_yI_y\right)\right]. \tag{19}$$

In the rotating frame associated with $H_R = DS_z^2 + \gamma_e B_z S_z$, the Hamiltonian becomes

$$\begin{aligned} H &= e^{iH_R t}(H - H_R)e^{-iH_R t} \\ &= A_{zz}S_zI_z + QI_z^2 - \gamma_n\mathbf{B}\cdot\mathbf{I} + \tilde{H}_1, \end{aligned} \tag{20}$$

which is time-independent since $[H_R, \tilde{H}_1] = 0$ under the second-order perturbation. The above equation can be cast in the form of Eq. (12) with $B_n = A_{zz}I_z$ and $C_n = QI_z^2 - \gamma_n\mathbf{B}\cdot\mathbf{I} + \tilde{H}_1$.

When the magnetic field is perfectly aligned along the NV axis, $[B_n, C_n] = 0$, the fixed points of the channel are $P_\beta^n = |\beta\rangle_n\langle\beta|$, with $\beta = 0, \pm 1$. But when the magnetic field deviates slightly from the axis, $[B_n, C_n] \neq 0$, the channel has only one fixed point, which is the maximum mixed state $\mathbb{I}_n/3 = (P_1^n + P_0^n + P_{-1}^n)/3$, and two metastable points. The convex combination of the fixed point and the two metastable points spans a two-dimensional MM holding three EMSs, which are approximately P_β^n. We note that the eigenvalues λ_2 and λ_3 of two metastable points with eigenvalues are usually different ($|\lambda_2| > |\lambda_3|$), which makes it difficult to directly observe the three-peak feature in experiments.

As m is so large that the contribution of one metastable point can be neglected ($|\lambda_3|^m \to 0$), the remaining metastable point, together with

the fixed point, span a one-dimensional MM with two EMSs, which can be calculated by Eq. (10). For Fig. 2, numerical calculations show that the two observed EMSs are $\rho_D = P_1^n$ and $\rho_B = (P_0^n + P_{-1}^n)/2$.

Model for the optical readout process

In this section, we model the optical readout process of NV centers at room temperature. In the readout process, we use a green laser to excite the NV electron and collect the emitted photons. The NV electron spin in state $|\pm 1\rangle_e$ has a greater chance to relax to the singlet state (via inter-system crossing), resulting in a lower photon number in the detection window (650–800 nm). This provides a convenient and efficient mechanism to distinguish the NV spin states by the "bright-dark" contrast, corresponding to the $|0\rangle_e$ and $|\pm 1\rangle_e$ states, respectively.

The optical readout process corresponds to a non-selective projective measurement of the NV electron spin, and can be modeled as a series of Kraus operators $\{K_n\}_{n=0}^{\infty}$. If n photons are collected by the counter, the Kraus operator K_n acting on the NV electron spin is[55]

$$K_n = \sum_{\alpha=0,1} \sqrt{p(n|\alpha)} P_\alpha^e, \tag{21}$$

where P_α^e is projector on subspace α and collected photon numbers obey Poisson distribution $p(n|\alpha) = \frac{1}{n!} e^{-n_\alpha} n_\alpha^n$ with n_α being the average photon number obtained for state $|\alpha\rangle_n$. If n_α is small (about 0.05 in our experiments), truncating to K_0 and K_1 is a good approximation with $p(1|\alpha) = n_\alpha$ and $p(0|\alpha) = 1 - p(1|\alpha)$.

Then, the Kraus operator on the ^{14}N bath spin when obtaining n photons can be derived as

$$\begin{aligned} W_n(\rho) &= \mathrm{Tr}_e\{K_n \mathcal{P}[R_{\phi_2} U R_{\phi_1}(|0\rangle_e\langle 0| \otimes \rho) R_{\phi_1}^\dagger U^\dagger R_{\phi_2}^\dagger] K_n^\dagger\} \\ &= \sum_{\alpha=0,1} p(n|\alpha) M_\alpha \rho M_\alpha^\dagger, \end{aligned} \tag{22}$$

where $\mathcal{P}(\cdot) = \sum_{\alpha=0,1} P_\alpha^e(\cdot) P_\alpha^e$ is the channel for non-selective projective measurement of the electron spin, and $M_\alpha = [U_0 - (-1)^\alpha e^{i\Delta\phi} U_1]$ is the Kraus operator introduced in Eq. (2) for the nuclear spin. Note that

$$\hat{\Phi}' = \sum_n \hat{W}_n = \sum_{\alpha=0,1} \sum_n p(n|\alpha) \hat{M}_\alpha = \hat{\Phi}, \tag{23}$$

then the channel with the set of Kraus operators $\{W_n\}_{n=0}^{\infty}$ is the same as that with $\{M_\alpha\}_{\alpha=0,1}$, so the analysis of the channel decomposition in the section above also applies here.

Measurement statistics of sequential RIMs

The fixed points (or metastable points) can be observed by the measurement statistics of sequential RIMs. A single RIM on the probe spin has an outcome $a \in \{0, 1\}$, and for sequential RIMs we obtain a sequence of binary numbers $\{a_1, \cdots, a_m\}$. This sequence of measurement outcomes defines a quantum trajectory. We then look for some statistical observable that can characterize different quantum trajectories and their corresponding fixed points.

In practice, we often focus on the average of the m measurement results and its expectation, i.e., $\bar{a} = \frac{1}{m}\sum_{n=1}^m a_n$, which in our case coincides with the measurement frequency $f_1 = m_0/m$ and can be related to the measurement polarization by $X = 1/2 - f_1$. The expectation of f_1 can be written as

$$\langle f_1 \rangle = \sum_{a_1} \cdots \sum_{a_m} f_1 p(a_1, \cdots, a_m|\rho), \tag{24}$$

where the probability to get a specific sequence of measurement results $\{a_1, \cdots, a_m\}$ is given by

$$p(a_1, \cdots, a_m|\rho) = \langle\langle \mathbb{I}|\hat{\mathcal{M}}_{a_m} \cdots \hat{\mathcal{M}}_{a_1}|\rho\rangle\rangle. \tag{25}$$

It can be shown that in the asymptotic limit, i.e., when the number of repetitions is large enough $m \to \infty$, $\langle f_1 \rangle$ is only determined by fixed points[24],

$$\lim_{m\to\infty} \langle f_1 \rangle = \sum_{j=1}^{J} c_j \langle f_{1j} \rangle_*, \tag{26}$$

where we assume there are J fixed points with

$$\langle f_{1j} \rangle_* = \langle\langle \mathbb{I}|\hat{\mathcal{M}}_1|\rho_{\mathrm{fix}}^j\rangle\rangle = p(1|\rho_{\mathrm{fix}}^j), \tag{27}$$

with $c_j = \mathrm{Tr}(\rho_{\mathrm{fix}}^j \rho)$ being the probability of obtaining the j-th fixed point given the initial state ρ. Note that the above equation indicates that the expectation of the measurement average coincides with the probability of obtaining result $a = 1$ for the fixed point ρ_{fix}^j in a single RIM,

$$p(a|\rho) = \mathrm{Tr}(M_a \rho M_a^\dagger) = \langle\langle \mathbb{I}|\hat{\mathcal{M}}_a|\rho\rangle\rangle. \tag{28}$$

In other words, the distribution of the measured frequency f_1 can have J peaks, centering around the set of expectations $\{\langle f_{1j}\rangle_*\}_{j=1}^J$. Each peak includes all the trajectories with f_1 close to $\langle f_{1j}\rangle_*$[34].

Taking into account the optical readout process, the average photon number corresponding to the jth fixed (metastable) point is

$$\begin{aligned} \langle n_j \rangle_* &= \sum\nolimits_n \langle\langle \mathbb{I}|n\hat{\mathcal{V}}_n|\rho_{\mathrm{fix}}^j\rangle\rangle \\ &= \sum\nolimits_{n,\alpha} n p(n|\alpha) \langle f_{\alpha j}\rangle_* \end{aligned} \tag{29}$$

where $\hat{\mathcal{V}}_n = W_n \otimes W_n^* = \sum_{\alpha=0,1} p(n|\alpha) \hat{\mathcal{M}}_\alpha$ is the superoperator of W_n in the HS space.

Data availability

The data generated in this study are provided in the Source Data file. Source data are provided with this paper.

Code availability

The codes used for the current study are available on request from the corresponding author.

References

1. Mori, T., Ikeda, T. N., Kaminishi, E. & Ueda, M. Thermalization and prethermalization in isolated quantum systems: a theoretical overview. *J. Phys. B* **51**, 112001 (2018).
2. Hodgman, S. S. et al. Metastable helium: a new determination of the longest atomic excited-state lifetime. *Phys. Rev. Lett.* **103**, 053002 (2009).
3. Balan, A. et al. Direct observation of magnetic metastability in individual iron nanoparticles. *Phys. Rev. Lett.* **112**, 107201 (2014).
4. Karube, K. et al. Robust metastable skyrmions and their triangular–square lattice structural transition in a high-temperature chiral magnet. *Nat. Mater.* **15**, 1237 (2016).
5. Han, T.-R. T. et al. Exploration of metastability and hidden phases in correlated electron crystals visualized by femtosecond optical doping and electron crystallography. *Sci. Adv.* **1**, e1400173 (2015).
6. Macieszczak, K., Guţă, M., Lesanovsky, I. & Garrahan, J. P. Towards a theory of metastability in open quantum dynamics. *Phys. Rev. Lett.* **116**, 240404 (2016).
7. Macieszczak, K., Rose, D. C., Lesanovsky, I. & Garrahan, J. P. Theory of classical metastability in open quantum systems. *Phys. Rev. Res.* **3**, 033047 (2021).

8. Brown, C. A., Macieszczak, K. & Jack, R. L. Unraveling metastable Markovian open quantum systems. *Phys. Rev. A* **109**, 022244 (2024).
9. Rose, D. C., Macieszczak, K., Lesanovsky, I. & Garrahan, J. P. Metastability in an open quantum Ising model. *Phys. Rev. E* **94**, 052132 (2016).
10. Le Boité, A., Hwang, M.-J. & Plenio, M. B. Metastability in the driven-dissipative Rabi model. *Phys. Rev. A* **95**, 023829 (2017).
11. Gong, Z., Hamazaki, R. & Ueda, M. Discrete time-crystalline order in cavity and circuit QED systems. *Phys. Rev. Lett.* **120**, 040404 (2018).
12. Gambetta, F. M., Carollo, F., Marcuzzi, M., Garrahan, J. P. & Lesanovsky, I. Discrete time crystals in the absence of manifest symmetries or disorder in open quantum systems. *Phys. Rev. Lett.* **122**, 015701 (2019).
13. Landa, H., Schiró, M. & Misguich, G. Multistability of driven-dissipative quantum spins. *Phys. Rev. Lett.* **124**, 043601 (2020).
14. Flynn, V. P., Cobanera, E. & Viola, L. Topology by dissipation: Majorana bosons in metastable quadratic Markovian dynamics. *Phys. Rev. Lett.* **127**, 245701 (2021).
15. Defenu, N. Metastability and discrete spectrum of long-range systems. *Proc. Natl. Acad. Sci. USA* **118**, e2101785118 (2021).
16. Jäger, S. B., Schmit, T., Morigi, G., Holland, M. J. & Betzholz, R. Lindblad master equations for quantum systems coupled to dissipative bosonic modes. *Phys. Rev. Lett.* **129**, 063601 (2022).
17. Cabot, A., Carollo, F. & Lesanovsky, I. Metastable discrete time-crystal resonances in a dissipative central spin system. *Phys. Rev. B* **106**, 134311 (2022).
18. Blume-Kohout, R., Ng, H. K., Poulin, D. & Viola, L. Characterizing the structure of preserved information in quantum processes. *Phys. Rev. Lett.* **100**, 030501 (2008).
19. Blume-Kohout, R., Ng, H. K., Poulin, D. & Viola, L. Information-preserving structures: a general framework for quantum zero-error information. *Phys. Rev. A* **82**, 062306 (2010).
20. Fitzpatrick, M., Sundaresan, N. M., Li, A. C. Y., Koch, J. & Houck, A. A. Observation of a dissipative phase transition in a one-dimensional circuit QED lattice. *Phys. Rev. X* **7**, 011016 (2017).
21. Letscher, F., Thomas, O., Niederprüm, T., Fleischhauer, M. & Ott, H. Bistability versus metastability in driven dissipative Rydberg gases. *Phys. Rev. X* **7**, 021020 (2017).
22. Beaulieu, G. et al. Observation of first-and second-order dissipative phase transitions in a two-photon driven Kerr resonator. *Nat. Commun.* **16**, 1954 (2025).
23. Jin, Y.-D., Qiu, C.-D. & Ma, W.-L. Theory of metastability in discrete-time open quantum dynamics. *Phys. Rev. A* **109**, 042204 (2024).
24. Qiu, C.-D., Jin, Y.-D., Zhang, J.-X., Liu, G.-Q. & Ma, W.-L. How coherence measurements of a qubit steer its quantum environment. *Phys. Rev. B* **110**, 024311 (2024).
25. Watrous, J.*The Theory of Quantum Information* (Cambridge University Press, 2018).
26. Gudder, S. Quantum Markov chains. *J. Math. Phys.* **49**, 072105 (2008).
27. Amato, D., Facchi, P. & Konderak, A. Asymptotics of quantum channels. *J. Phys. A* **56**, 265304 (2023).
28. Ramsey, N. F. A molecular beam resonance method with separated oscillating fields. *Phys. Rev.* **78**, 695 (1950).
29. Lee, H., Kok, P. & Dowling, J. P. A quantum Rosetta stone for interferometry. *J. Mod. Opt.* **49**, 2325 (2002).
30. Degen, C. L., Reinhard, F. & Cappellaro, P. Quantum sensing. *Rev. Mod. Phys.* **89**, 035002 (2017).
31. Arias, A., Gheondea, A. & Gudder, S. Fixed points of quantum operations. *J. Math. Phys.* **43**, 5872 (2002).
32. Albert, V. V. Asymptotics of quantum channels: conserved quantities, an adiabatic limit, and matrix product states. *Quantum* **3**, 151 (2019).
33. Bauer, M. & Bernard, D. Convergence of repeated quantum nondemolition measurements and wave-function collapse. *Phys. Rev. A* **84**, 044103 (2011).
34. Ma, W.-L., Li, S.-S. & Liu, R.-B. Sequential generalized measurements: asymptotics, typicality, and emergent projective measurements. *Phys. Rev. A* **107**, 012217 (2023).
35. Wang, P., Yang, W. & Liu, R. Using weak measurements to synthesize projective measurement of nonconserved observables of weakly coupled nuclear spins. *Phys. Rev. Appl.* **19**, 054037 (2023).
36. Dréau, A., Spinicelli, P., Maze, J. R., Roch, J.-F. & Jacques, V. Single-shot readout of multiple nuclear spin qubits in diamond under ambient conditions. *Phys. Rev. Lett.* **110**, 060502 (2013).
37. Neumann, P. et al. Single-shot readout of a single nuclear spin. *Science* **329**, 542 (2010).
38. Maurer, P. C. et al. Room-temperature quantum bit memory exceeding one second. *Science* **336**, 1283 (2012).
39. Ma, W.-L., Wang, P., Leong, W.-H. & Liu, R.-B. Phase transitions in sequential weak measurements. *Phys. Rev. A* **98**, 012117 (2018).
40. Bernien, H. et al. Heralded entanglement between solid-state qubits separated by three metres. *Nature* **497**, 86 (2013).
41. Cai, J. et al. A large-scale quantum simulator on a diamond surface at room temperature. *Nat. Phys.* **9**, 168 (2013).
42. Abobeih, M. H. et al. Fault-tolerant operation of a logical qubit in a diamond quantum processor. *Nature* **606**, 884 (2022).
43. Randall, J. et al. Many-body–localized discrete time crystal with a programmable spin-based quantum simulator. *Science* **374**, 1474 (2021).
44. Soshenko, V. V. et al. Nuclear spin gyroscope based on the nitrogen vacancy center in diamond. *Phys. Rev. Lett.* **126**, 197702 (2021).
45. Blok, M. et al. Manipulating a qubit through the backaction of sequential partial measurements and real-time feedback. *Nat. Phys.* **10**, 189 (2014).
46. Bonato, C. et al. Optimized quantum sensing with a single electron spin using real-time adaptive measurements. *Nat. Nano.* **11**, 247 (2016).
47. Burrell, A. H., Szwer, D. J., Webster, S. C. & Lucas, D. M. Scalable simultaneous multiqubit readout with 99.99% single-shot fidelity. *Phys. Rev. A* **81**, 040302 (2010).
48. Pfender, M. et al. High-resolution spectroscopy of single nuclear spins via sequential weak measurements. *Nat. Commun.* **10**, 594 (2019).
49. Taminiau, T. et al. Detection and control of individual nuclear spins using a weakly coupled electron spin. *Phys. Rev. Lett.* **109**, 137602 (2012).
50. Taylor, J. M. et al. High-sensitivity diamond magnetometer with nanoscale resolution. *Nat. Phys.* **4**, 810 (2008).
51. Sadzak, N. et al. A Hahn-Ramsey scheme for dynamical decoupling of single solid-state qubits. *Front. Photonics* **3**, 932944 (2022).
52. Wang, Y.-X. & Clerk, A. A. Intrinsic and induced quantum quenches for enhancing qubit-based quantum noise spectroscopy. *Nat. Commun.* **12**, 6528 (2021).
53. Jerger, P. C. et al. Detecting spin-bath polarization with quantum quench phase shifts of single spins in diamond. *PRX Quantum* **4**, 040315 (2023).
54. Childress, L. et al. Coherent dynamics of coupled electron and nuclear spin qubits in diamond. *Science* **314**, 281 (2006).
55. Meinel, J. et al. Quantum nonlinear spectroscopy of single nuclear spins. *Nat. Commun.* **13**, 5318 (2022).

Acknowledgements

This work is supported by the Innovation Program for Quantum Science and Technology (Grants Nos. 2023ZD0300600, 2021ZD0302300), the National Key Research and Development Program of China (Grants No. 2023YFA1608900), the National Natural Science Foundation of China (Grant Nos. 12022509, 11934018, T2121001), and the Chinese Academy of Sciences (Grant Nos. YSBR90, YSBR-100, E0SEBB11, E27RBB11).

Author contributions

W.L.M. and G.Q.L. conceived the project. J.X.Z. conducted the experiments with the supervision of G.Q.L. and Y.D.J. C.D.Q. performed theoretical analysis and simulations with the supervision of W.L.M. All authors contributed to the writing of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64772-6.

Correspondence and requests for materials should be addressed to Wen-Long Ma or Gang-Qin Liu.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：郭沛林

方向：生物制造

单位：中国科学院过程工程研究所

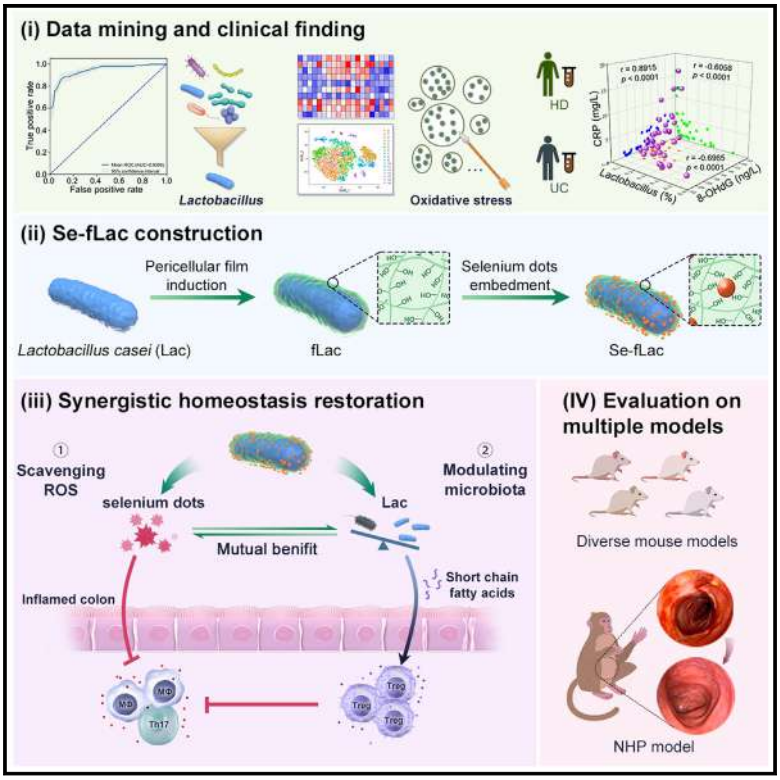
题目：数据驱动工程化益生菌创制及其用于溃疡性结肠炎治疗的研究

Article

Cell Host & Microbe

Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis

Graphical abstract



Highlights

- Data mining revealed alterations in *Lactobacillus* and oxidative stress in UC patients
- Engineered *L. casei* to contain selenium dots embedded in the pericellular film
- Upon oral delivery, Se-fLac synergistically scavenged ROS and modulated the microbiota
- Therapeutic efficacy of Se-fLac was demonstrated in mouse models and NHPs

Authors

Peilin Guo, Wenjing Wang, Qian Xiang, ..., Yimin Cui, Guanghui Ma, Wei Wei

Correspondence

cui.pharm@pkufh.com (Y.C.), ghma@ipe.ac.cn (G.M.), weiwei@ipe.ac.cn (W.W.)

In brief

Guo et al. engineer *Lactobacillus casei* to contain embedded selenium dots (Se-fLac) for synergistic restoration of intestinal redox and microbiota homeostasis upon oral administration. Using mouse and NHP models of colitis, the authors demonstrate that Se-fLac effectively relieves inflammation and reduces colonic damage.

Guo et al., 2024, Cell Host & Microbe *32*, 1–17
September 11, 2024 © 2024 Elsevier Inc. All rights reserved, including those for text and data mining, AI training, and similar technologies.
https://doi.org/10.1016/j.chom.2024.07.028

CellPress

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

Cell Host & Microbe

CellPress

Article

Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis

Peilin Guo,[1,2,3,13] Wenjing Wang,[1,2,13] Qian Xiang,[4] Chao Pan,[5] Yefeng Qiu,[6] Tingting Li,[7] Dongfang Wang,[8] Jian Ouyang,[7] Rongrong Jia,[9] Min Shi,[9] Yugang Wang,[9] Junxia Li,[10] Jiale Zou,[11] Yuan Zhong,[12] Jiawei Zhao,[1,2] Diwei Zheng,[1,2] Yimin Cui,[4,*] Guanghui Ma,[1,2,3,*] and Wei Wei[1,2,3,14,*]

[1]State Key Laboratory of Biochemical Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing 100190, P.R. China
[2]Key Laboratory of Biopharmaceutical Preparation and Delivery, Chinese Academy of Sciences, Beijing 100190, P.R. China
[3]School of Chemical Engineering, University of Chinese Academy of Sciences, Beijing 100049, P.R. China
[4]Institute of Clinical Pharmacology, Peking University First Hospital, Beijing 100191, P.R. China
[5]State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Biotechnology, Beijing 100071, P.R. China
[6]Laboratory Animal Center of the Academy of Military Medical Sciences, Beijing 100071, P.R. China
[7]Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, P.R. China
[8]Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing 100871, P.R. China
[9]Department of Gastroenterology, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, P.R. China
[10]Department of Gastroenterology, Peking University First Hospital, Beijing 100034, P.R. China
[11]Department of Gastroenterology, The Second Medical Centre, Chinese PLA General Hospital, Beijing 100853, P.R. China
[12]State Key Laboratory of Integrated Optoelectronics, College of Electronic Science and Engineering, Jilin University, Changchun 130015, P.R. China
[13]These authors contributed equally
[14]Lead contact
*Correspondence: cui.pharm@pkufh.com (Y.C.), ghma@ipe.ac.cn (G.M.), weiwei@ipe.ac.cn (W.W.)
https://doi.org/10.1016/j.chom.2024.07.028

SUMMARY

Probiotics are potential treatments for ulcerative colitis (UC), but their efficacy is frequently compromised by gastrointestinal conditions that limit adhesion and activity. Here, we use machine learning and bioinformatics to confirm that patients with UC have decreased prevalence of *Lactobacillus* genus and increased oxidative stress, which correlate with inflammation severity. Accordingly, we developed a probiotic-based therapeutic that synergistically restores intestinal redox and microbiota homeostasis. *Lactobacillus casei* (Lac) were induced to form a pericellular film, providing a polysaccharide network for spatially confined crystallization of ultrasmall but highly active selenium dots (Se-Lac). Upon oral administration, the selenium dot-embedded pericellular film efficiently enhanced gastric acid resistance and intestinal mucoadhesion of Lac cells. At the lesion site, the selenium dots scavenged reactive oxygen species, while Lac modulated the gut microbiota. In multiple mouse models and non-human primates, this therapeutic effectively relieved inflammation and reduced colonic damage, thus showing promise as a UC treatment.

INTRODUCTION

Ulcerative colitis (UC) is an inflammatory bowel disease (IBD) characterized by long-lasting inflammation, and the clinical manifestations include diarrhea, mucus or blood in the feces, with an increased risk of colorectal cancer in chronic patients.[1] In clinic, aminosalicylate therapy is used for patients with mild to moderate UC, but ~30% of these patients either show no response at all or will gradually lose clinical efficacy over time.[2,3] Although steroids can alleviate symptoms in patients with moderate to severe UC, they are not a long-term solution due to the risks of side effects.[4] Upon abovementioned unsatisfactory therapeutic outcomes, ~15% of UC patients (within 20 years of being diagnosed) have to undergo colectomy to remove a large section of their colon.[5] Consequently, new alternatives for the efficient treatment of UC are constantly being sought.

Given that an imbalanced composition of gut microbiota with decreased purportedly beneficial bacteria and increased deleterious bacteria has been implicated in UC, many scientists have shifted their attention to bacteriotherapy for restoring gut microbiota homeostasis.[6–9] Although several probiotic strains (e.g., *Bifidobacterium longum*, *Lactobacillus rhamnosus*) have been

All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

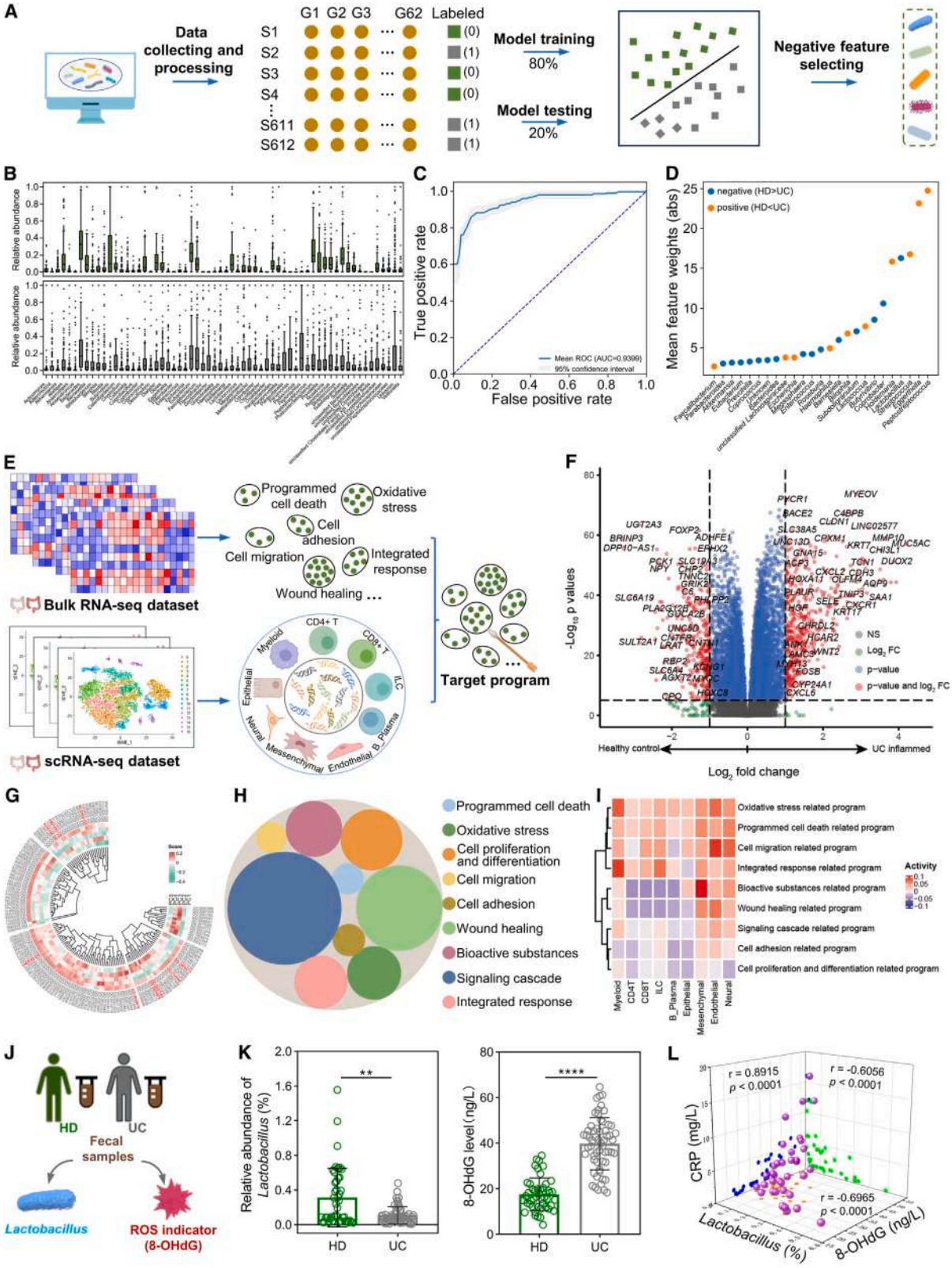


Figure 1. Data mining for revealing *Lactobacillus*-ROS comorbidity in UC cohort

(A) Schematic illustration of machine learning (ML)-driven analyses of gut microbiota for identifying the distinguishing bacterial genera between healthy donors and UC patients. The data included 62 bacterial genera (from G1 to G62) and 612 samples with green (healthy donors, labeled 0) and gray labels (UC patients,

(legend continued on next page)

approved as oral supplements for UC treatment in clinical trials,[10,11] it has not reached the agreement that probiotics have superior performance compared with placebo.[12] One well-known reason could be the harsh physiological environment in the gastrointestinal tract, wherein gastric acid and pepsin in the stomach have an inactivating effect; additionally, the continuous flow in the intestine compromises mucoadhesion, together leading to the supplemented probiotics with a very low amount and viability.[13,14] Moreover, the complicated pathological environment of inflamed tissue also implies the presence of other pathogenic factors.[15]

Upon bioinformatics analyses of publicly available RNA sequencing (RNA-seq) datasets, we herein revealed oxidative stress as a pivotal pathogenic factor for UC intervention. As crucial signaling molecules in oxidative stress, excessive reactive oxygen species (ROS) can trigger proinflammatory cascades, and reducing ROS levels has shown the capacity to restore redox homeostasis and prevent exacerbation of many inflammatory diseases.[16–19] Upon using machine learning (ML) to identify the feature weights of decreased genera and further taking inflammation severity into consideration, we discovered correlations between either two aspects of *Lactobacillus* abundance, 8-hydroxy-2-deoxyguanosine (8-OHdG, an ROS indicator) levels, and C-reactive protein (CRP) values in UC patients. Thus, these findings indicated that, beyond delivering adequate viable *Lactobacillus*, endowing *Lactobacillus* with the function of scavenging ROS can be a further improvement of UC treatment.[20,21]

Considering that selenium is a nutritional trace element conferring antioxidant effects, we proposed that the codelivery of selenium and *Lactobacillus* could fulfill the synergistic restoration of intestinal redox-microbiota homeostasis.[22–24] Given that biofilm had been reported with the capacity of improving bacterial resistance and adhesion,[25,26] we further envisioned that growing selenium-enriched pericellular film surrounding the bacterial cell wall could improve the codelivery efficacy to the intestine.[27–29] Pursuing these issues, we induced *Lactobacillus casei* (a well-known probiotic, denoted as Lac) with pericellular film, wherein the polysaccharide network facilitated spatially confined crystallization of ultrasmall but highly active selenium dots.[30] Upon oral administration, this Lac-based therapeutic prominently scavenged ROS and had the tendency to modulate gut microbiota, and these two aspects were further observed with a potentially mutual benefit. Consequently, we demonstrated potent therapeutic efficacies in classical UC mouse models and, most notably, a rarely reported UC non-human primate (NHP) model that further supported the clinical utility of our therapeutic.

RESULTS

Data mining and clinical verification

To rationally select bacteria for modulating the gut microbiota in UC, we herein employed an ML model to identify the distinguishing bacterial genera between healthy donors (HDs) and UC patients.[31] Initially, we collected and processed sequencing data, including bacterial abundances (a total of 62 bacterial genera, equivalent to 62 features) and phenotypes (HDs labeled 0 and UC patients labeled 1), from the GMrepo metagenomics database.[32] Prioritizing enhanced model interpretability over complexity, we selected L1-regularized logistic regression model and assessed its classification performance. Selecting negative features from model, we could identify the bacteria genera with decreased abundance in UC patients compared to HDs (Figure 1A).

After training and testing the data shown in Figure 1B, our model demonstrated good classification performance, as evidenced by a mean area under the receiver operating curve (AUC) value of 0.9388 ± 0.0035 (Figure 1C). To identify the contributions of diverse bacterial genera, we further interpreted feature importance by ranking the absolute values of mean feature weights (Figure 1D). Focusing on the bacterial genera with negative feature weights, *Lactobacillus* genus had the maximum absolute value of mean feature weight, which was also validated with other datasets (Figures S1A–S1D). Beyond echoing previous study,[33] above data mining provided the rationality to consider *Lactobacillus* genus as a suitable candidate for supplement, which might have the tendency to restore microbiota homeostasis for UC treatment.

Shifting to the inflamed tissue, we herein conducted bioinformatics analyses on publicly available sequencing datasets to pinpoint the pivotal pathogenic factor for intervention (Figure 1E).[34,35] Initially, differentially expressed genes (DEGs) were extracted by comparing the bulk RNA-seq expression profiles of tissues (GSE193677) between UC inflamed and healthy control

labeled 1). After successful modeling, negative features were selected to identify the bacterial genera with decreased abundance in UC patients compared to healthy donors.

(B) Summary of bacterial genera in collected data of healthy donors (green box, n = 313) and UC patients (gray box, n = 299).

(C) Mean areas under the receiver operating curve (AUC) of the L1-regularized logistic regression model for the classification of healthy donors and UC patients.

(D) Top 25 absolute values of mean feature weights in the L1-regularized logistic regression model. Genera that had higher abundance in healthy donors or UC patients are shown in blue or yellow, respectively.

(E) Schematic illustration of pinpointing the pivotal pathogenic factor in colon tissues by conducting systematic bioinformatics analyses on publicly available sequencing datasets.

(F) Volcano plot showing DEGs extracted by comparing the bulk RNA-seq expression profiles of tissues between UC inflamed and healthy control.

(G) Gene set variation analysis showing five distinct modules with varied pathway activity patterns among UC patients.

(H) Schematic illustration of nine annotated meta-programs from the selected pathways in (G).

(I) Heatmap showing the meta-program score across all colon cell types in single-cell RNA-seq datasets from UC patients.

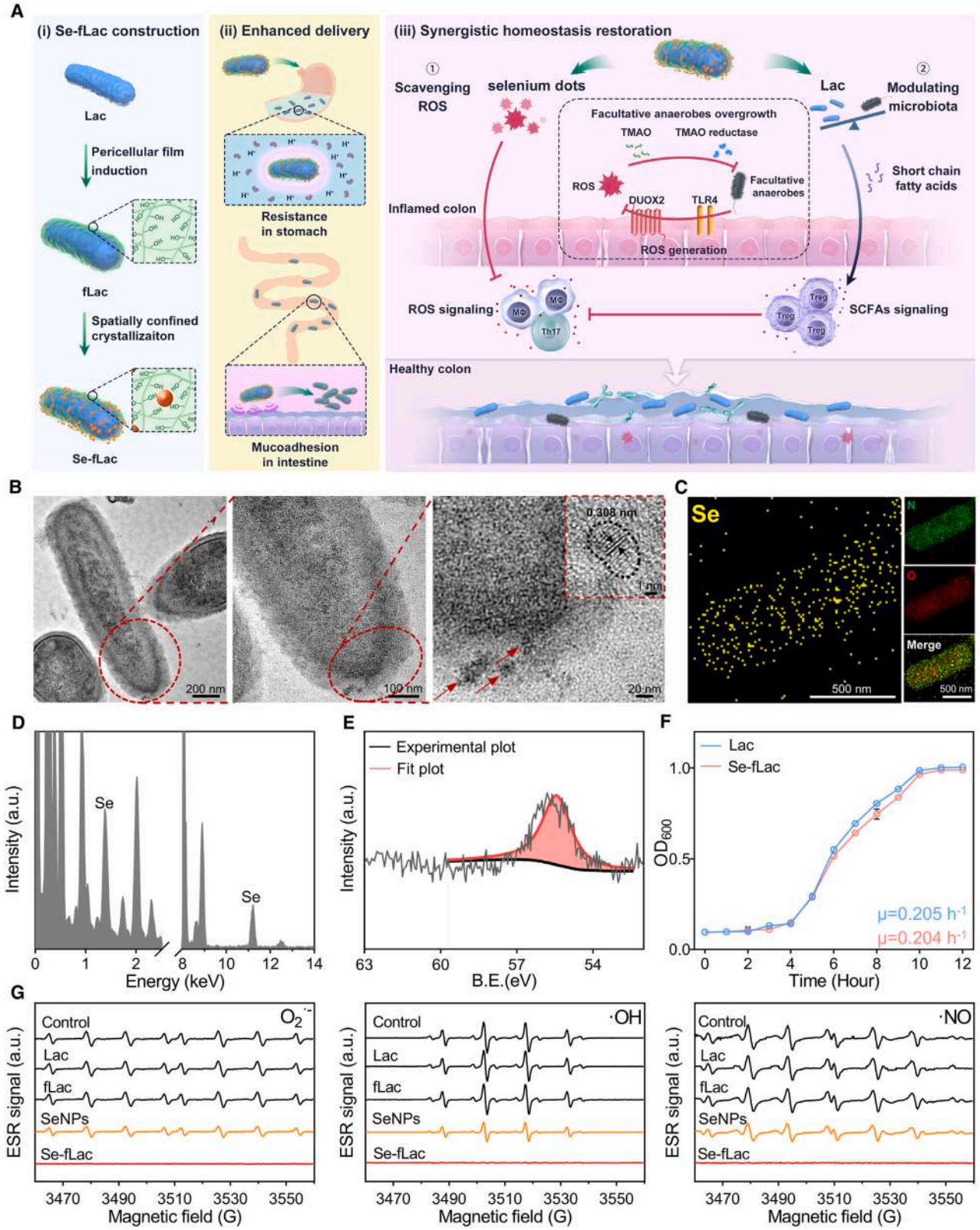
(J) Schematic illustration of fecal sample collection and subsequent measurements in healthy donors (HDs) and UC patients.

(K) Relative abundance of the *Lactobacillus* genus and 8-OHdG level in fecal samples, respectively (HDs, n = 46; UC patients, n = 57).

(L) Correlation analysis examining various relationships among *Lactobacillus* abundance, 8-OHdG levels, and CRP values in UC patients (n = 36).

Data in (K) are presented as mean ± SD and were compared via two-tailed unpaired Student's t test. Data in (L) were compared via correlation analysis. Genes in (F) with an adjusted $p < 0.01$ are statistically significant. $^{**}p < 0.01$, $^{****}p < 0.0001$.

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028



(legend on next page)

(Figure 1F), with those UC-upregulated genes used to perform the GO over-representation analysis. We evaluated the activity of these diverse functional pathways in UC patients through gene set variation analysis (Figure 1G). Varied pathway activity patterns were noted among different patients, leading to five distinct modules, two featuring well-characterized cell-cycle and inflammation-related pathways. Upon annotating the remaining pathways into nine meta-programs (Figure 1H), we were also curious whether they were engaged by all major cell compartments or only certain of them. To address this issue, we inspected the meta-program score using single-cell RNA-seq datasets from UC patients.[35] Strikingly, we noted high activity of the oxidative stress program across all colon cell types (Figure 1I), which was also observed in the scores of Human Microbiome Project, part 2 (HMP2)-derived meta-program (Figures S1E and S1F). Although previous studies have shown oxidative stress associated with UC,[36] we herein provide additional evidence to support it as a pivotal pathogenic factor for targeted intervention, which could be efficiently achieved by scavenging excessive ROS in the colon.

For verification, we collected and analyzed fecal samples from a cohort of HDs (n = 46) and UC patients (n = 57) (Figure 1J; Table S1). As shown in Figure 1K, the abundance of *Lactobacillus* in UC patients was 4.4-fold lower than that in HDs, and the 8-OHdG level in UC patients was 3.7-fold higher than that in HDs. Further considering inflammation severity, we discovered correlations between either two aspects of *Lactobacillus* abundance, 8-OHdG levels, and CRP values in UC patients (Figures 1L and S1G, n = 36). These results from patient samples not only corroborated previous reports on the aberrant decrease in *Lactobacillus* and increase in ROS,[33,36] but also revealed the close association of these two aspects in UC exacerbation.

Rational design, preparation, and characterization of Se-fLac

Above results further promoted us to develop a UC therapeutic to efficiently supplement both *Lactobacillus* and ROS scavenger to the intestine. To accomplish this, *Lactobacillus casei* was initially selected as a microbiota modulator, and induced its pericellular film formation (fLac) via nutrient deficiency culture. We further proposed the idea of functionalizing fLac with ultrasmall but highly active selenium dots (Se-fLac), which could be formed by hydroxyl ligand-mediated spatially confined crystallization in the polysaccharide network of pericellular film (Figure 2Ai). Upon oral administration of resulting Se-fLac, the composite armor of pericellular film and selenium dots enhanced resistance in the stomach, and the pericellular film subsequently facilitated intestinal mucoadhesion (Figure 2Aii). After reaching the colon, ROS scavenging could lower the proportion of inflammatory cells and regulate inflammatory cytokines, and modulating the gut microbiota could increase the proportion of purportedly beneficial bacteria, followed by the elevated production of short-chain fatty acids (SCFAs) to increase the proportion of regulatory T (Treg) cells (Figure 2Aiii). We also envisioned that lower ROS levels and lower proportions of facultative anaerobes might be mutually beneficial, thus magnifying the anti-inflammatory performance for UC therapy.

According to our design, providing Lac cells with the pericellular film is a prerequisite for constructing Se-fLac. Inspired by previous studies demonstrating the elevated production of biofilms under minimal salts glycerol glutamate (MSgg) medium culture, we envisioned that the growth dynamics of biofilms could be tuned via optimization of culture time and MSgg medium.[25] This enabled biofilm formation mainly surrounded the Lac cells (Figure S2A), and the resulting pericellular film thus served as the substrate for selenium crystallization via reduction of Na_2SeO_3.[37] As the transmission electron microscopy (TEM) images shown in Figure 2B, the morphology of the modified fLac cells did not obviously differ from the unmodified fLac cells (Figure S2A), yet the film layer was populated by ~5 nm dots, with a regular interplanar spacing of 0.308 nm for the (221) crystal plane. Further element mapping, energy dispersive spectroscopy (EDS) analysis, X-ray diffraction (XRD), and X-ray photoelectron spectroscopy (XPS) data together demonstrated these dots were enriched with zero-valent selenium (Figures 2C–2E and S2B).[38] This unique *in situ* growth of ultrasmall selenium dots could be attributed to the pericellular film, wherein the three-dimensional framework of polysaccharides and the interaction between hydroxyl ligands and selenium atoms together induced the spatially confined crystallization of selenium (Figures S2C–S2E). After optimization and calculation, the loading content was up to 250 ng selenium dots per 1×10^8 colony-forming unit (CFU) Lac cells (Figures S2F and S2G). Notably, the abovementioned pericellular film and selenium dots had almost no effect on Lac cell viability (Figure 2F).

Figure 2. Preparation and characterizations of Se-fLac

(A) Schematic illustration of the preparation, enhanced oral delivery, and synergistic homeostasis restoration of Se-fLac.

(Ai) *Lactobacillus casei* (Lac) cell was induced to form pericellular film (fLac), and *in situ* growth of selenium dots in pericellular film (Se-fLac) was achieved via hydroxyl ligand-mediated spatially confined crystallization.

(Aii) Facing the harsh gastric environment, the resistance of Lac cells was enhanced by the composite armor of pericellular film and selenium dots. Upon reaching the intestine, the mucoadhesion of Lac cells was improved by the pericellular film.

(Aiii) Synergistic homeostasis restoration: ① selenium dots could scavenge ROS and downregulate ROS signaling, thus decreasing the proportion of inflammatory cells and regulating cytokine levels. ② Supplemented *Lactobacillus* had the tendency of modulating microbiota, followed by the elevated production of SCFAs to increase the proportion of Treg cells. Notably, lower ROS levels and lower abundance of facultative anaerobes were potentially mutual benefit.

(B) Representative TEM images of Se-fLac with gradually increased magnifications. The inserted image shows selenium dots with a (221) crystal plane exhibiting 0.308 nm interplanar spacing.

(C) Representative elemental mapping images of Se-fLac.

(D) EDS elemental analysis of Se-fLac.

(E) XPS spectrum of Se3d in Se-fLac.

(F) Growth curves of Lac cells and Se-fLac cells under normal conditions.

(G) Representative ESR spectra showing the distinct scavenging performances of Lac, fLac, SeNPs, and Se-fLac for $O_2^{\cdot -}$, ·OH, and ·NO.

Data in (F) are presented as mean ± SD from 3 biologically independent experiments.

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

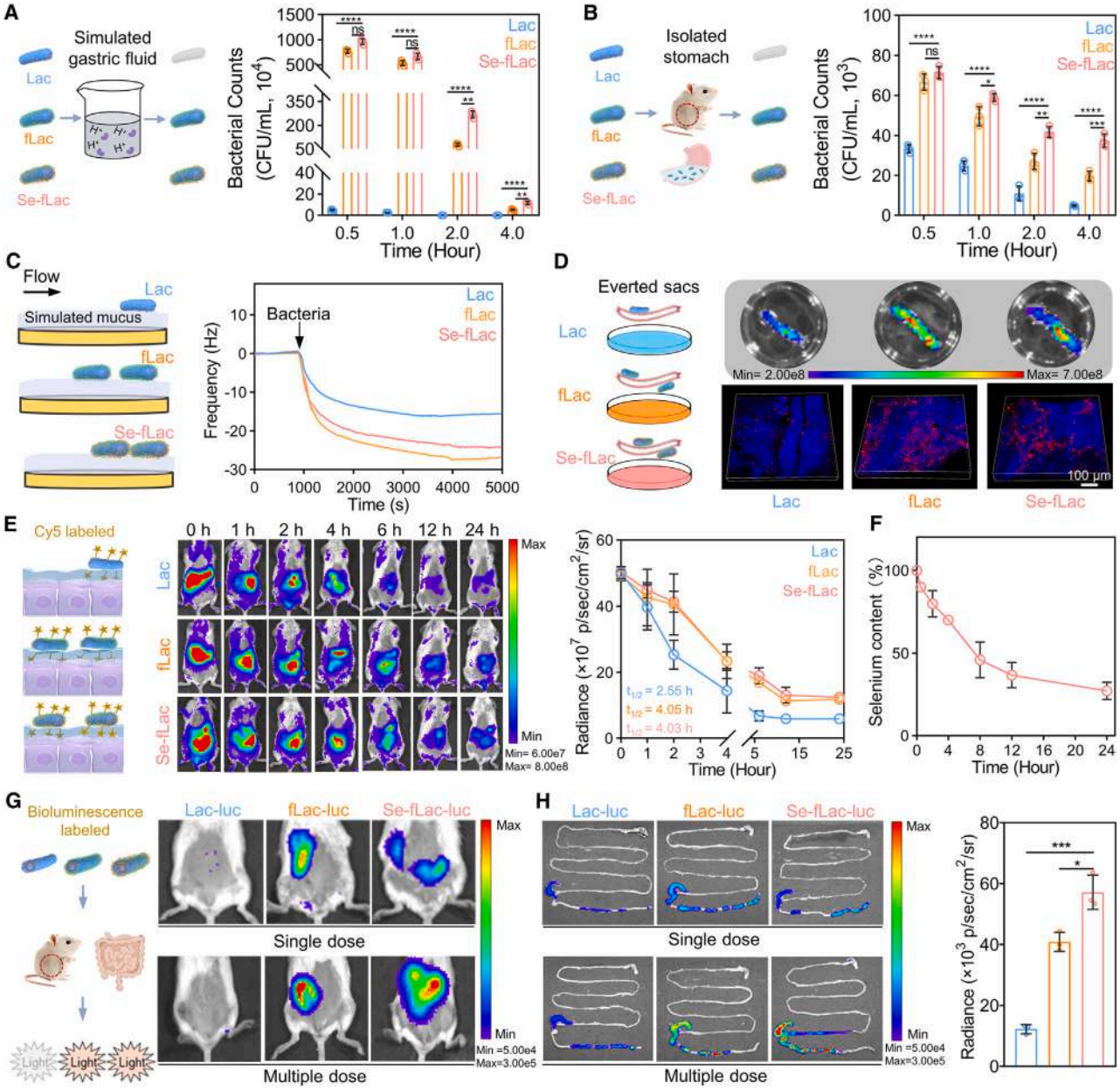


Figure 3. Enhanced resistance and mucoadhesion of Se-fLac in the gastrointestinal tract

(A) *In vitro* resistance of bacteria against SGF. The amounts of survived Lac, fLac, and Se-fLac cells were quantified by bacterial plate counting.

(B) *Ex vivo* resistance of bacteria against gastric environment. The amounts of survived Lac, fLac, and Se-fLac cells in the excised entire stomach were quantified by bacterial plate counting.

(C) *In vitro* mucoadhesion of bacteria evaluated by using QCM. Lac, fLac, and Se-fLac suspensions were flowed into the chamber to evaluate the adhesion between bacteria and mucin 2-modified chips. Representative QCM curves showing changes in the frequency of each group are displayed in the right panel. ΔF represents the value of final change in the frequency at the equilibrium stage.

(D) *Ex vivo* evaluation of Cy5-labeled bacteria adhering to everted sac of colon. Representative fluorescence images of whole sac and 3D reconstructed CLSM images with high magnification are displayed in the right panel.

(E) *In vivo* evaluation of bacterial retention after oral administration. Representative fluorescence images of BALB/c mice and quantitative analysis of fluorescence intensities in the abdomen region are displayed in the right panel.

(F) *In vivo* retention of selenium. BALB/c mice were orally administrated with 1×10^8 CFU Se-fLac, and their intestines were sampled at indicated time points for ICP-MS measurement.

(G) *In vivo* viability of bacteria (transfected with luciferase plasmid) after oral administration. Representative bioluminescence images of BALB/c mice are displayed in the right panel. Mice were orally administrated with 1×10^8 CFU Lac-luc, fLac-luc, or Se-fLac-luc cells for single or multiple doses (once a day for 7 days).

(legend continued on next page)

With the Se-fLac in hand, we used electron spin resonance (ESR) spectroscopy to evaluate its ability to scavenge ROS, including superoxide anions ($O_2^{\cdot -}$), hydroxyl radicals (·OH), and nitroxide radical (·NO). As shown in Figure 2G, neither Lac nor fLac cells showed a sign of ROS reduction. In sharp contrast, all three ESR spectra in the Se-fLac group closely approximated the baseline. Moreover, we also examined conventional selenium nanoparticles (SeNPs, ~60 nm) as the counterpart and observed only a slight decrease in the ESR signal. Quantitatively, Se-fLac exhibited approximately 3-fold higher scavenging abilities for three types of ROS than SeNPs (Figure S2H). Such a high activity of Se-fLac could be attributed to the ultrasmall size of selenium dots, which led to substantial increases in the surface-to-volume ratio and the fraction of surface atoms. Consequently, high surface energy and insufficient atomic coordination facilitate the electron transfer process between selenium dots and ROS, which was verified by using time-resolved transient absorption spectroscopy (Figures S2I–S2K).[39] Notably, high ROS scavenging activity largely remained even after 1-week incubation of Se-fLac in PBS (Figure S2L), which suggested the potential for persistent activity in the intestine.

Enhanced resistance, mucoadhesion, and viability of Se-fLac after oral administration

For efficient oral delivery, Se-fLac cells should first progress through the stomach featuring the harsh gastric environment and should stay longer in the intestinal tract.[40] Thus, we initially compared the survival of Lac, fLac, and Se-fLac cells after incubation with simulated gastric fluid (SGF). Even after 4 h, Se-fLac cells well preserved their structure and exhibited a significantly higher survival rate (2.2-fold) than fLac cells (Figures 3A and S2M–S2O). Beyond the SGF testing, we conducted *ex vivo* experiments by sampling above three types of Lac cells from the stomachs of mice, which were excised and kept at 37°C. At all examined time points, the survival rate increased in the sequence of Lac, fLac, and Se-fLac (Figure 3B). These results thus indicated that the pericellular film enhanced the resistance of Lac cells in the stomach and the embedded selenium dots further strengthened this effect.

To investigate the intestinal mucoadhesion capacity, we conducted a quartz crystal microbalance (QCM) analysis (Figures 3C and S2P). Briefly, Lac, fLac, and Se-fLac suspensions were flowed into the chamber, wherein the chips were precoated with mucin 2. At the equilibrium stage, the adhesion mass (indicated by ΔF) of modified Lac groups (fLac and Se-fLac) was higher than that of pristine Lac group. Moreover, three types of Lac cells (Cy5 labeled) were incubated with everted sacs of mouse colon (Figure 3D), and Cy5 intensity in the fLac and Se-fLac groups was 1.9- and 1.8-fold higher than that in the Lac group, respectively (Figure S2Q). In addition, 3D reconstructed confocal laser scanning microscopy (CLSM) images showed more fLac cells and Se-fLac cells adhered to the surface of intestinal wall (Figure 3D).

For further verification, we treated the mice with Cy5-labeled Lac, fLac, and Se-fLac via single oral administration. Quantification of fluorescence intensity in the abdominal region showed that the half-life retention time in the fLac group ($t_{1/2}$ = 4.05 h) and Se-fLac group ($t_{1/2}$ = 4.03 h) was significantly longer than that in the Lac group ($t_{1/2}$ = 2.55 h) (Figure 3E), with the residual fluorescence intensities of Lac, fLac, and Se-fLac at 24 h decreasing to 11.8%, 21.2%, and 21.5%, respectively. Such a prolonged retention in fLac and Se-fLac groups could be attributed to the improved adherence to the mucosal layer (Figure S2R). In the term of selenium dots, inductive coupled plasma mass spectrometer (ICP-MS) analysis indicated that 27.1% was retained in the intestine 24 h post-administration (Figure 3F). Note that the fluorescence intensity of bacterial cells and amount of selenium dots in the Se-fLac group exhibited a similar attenuation profile in the intestine, thus supporting stable embedment of selenium dots in the pericellular film even after progressing through the gastrointestinal tract.

The enhanced resistance and mucoadhesion promoted us to investigate their viability in the intestine. Briefly, Lac cells were transfected with a luciferase plasmid, and the resulting Lac-luc and modified Lac-luc (fLac-luc and Se-fLac-luc) were orally administered with a single dose. Compared to almost lack of signal in the Lac-luc group, we observed bright bioluminescence signals in the abdomens of mice receiving fLac-luc and Se-fLac-luc treatments (Figure 3G). Moreover, multiple doses of oral administration led to a substantial elevation in bioluminescence intensity for both fLac-luc and Se-fLac-luc groups (Figure 3G), indicating that the accumulation of viable cells in the intestine could be increased by consecutive oral administration. Above mice were also sacrificed and their intestines were immediately excised for imaging. The Lac-luc group showed weak bioluminescence signal throughout the intestine, whereas the bioluminescence signals mainly distributed in the colon region in the fLac-luc and Se-fLac-luc groups, with intensities 3.3- and 4.7-fold higher than that in the Lac-luc group, respectively (Figure 3H). In addition, we investigated the distribution of selenium dots throughout the intestine. The colon showed a majority of selenium accumulation (~71.36%, Figure S2S), thus paving the way for selenium dots to concentrate their ROS scavenging performance at the UC lesion.

Potent therapeutic efficacies of Se-fLac against UC

Encouraged by the above efficient oral delivery, we investigated the *in vivo* therapeutic efficacies using a dextran sulfate sodium (DSS)-induced UC model with a transgenic interferon-gamma (IFN-γ) bioluminescence reporter mouse line and a BALB/c mouse line (Figure 4A).[41] The model mice were divided into five groups, which were orally administered PBS, Lac, fLac, SeNPs + fLac, or Se-fLac (daily administration for 1 week, 1 × 10^8 CFU/day) (Figures S3A–S3D). As *in vivo* imaging data shown in Figure 4B, the PBS group exhibited strong and time-dependent bioluminescence signals in their abdomens, and such

(H) *Ex vivo* viability of bacteria in the intestine after oral administration. Representative bioluminescence images of excised intestines in (G) and quantitative analysis of bioluminescence signal in colon region after multiple doses of oral administration are displayed in the right panel.

Data in (A), (B), (E), (F), and (H) are presented as mean ± SD from 3 biologically independent experiments. Data in (A), (B), and (H) were compared via one-way ANOVA, and the statistical significance was calculated by comparing Se-fLac group with the other groups. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$, ns: non-significance.

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

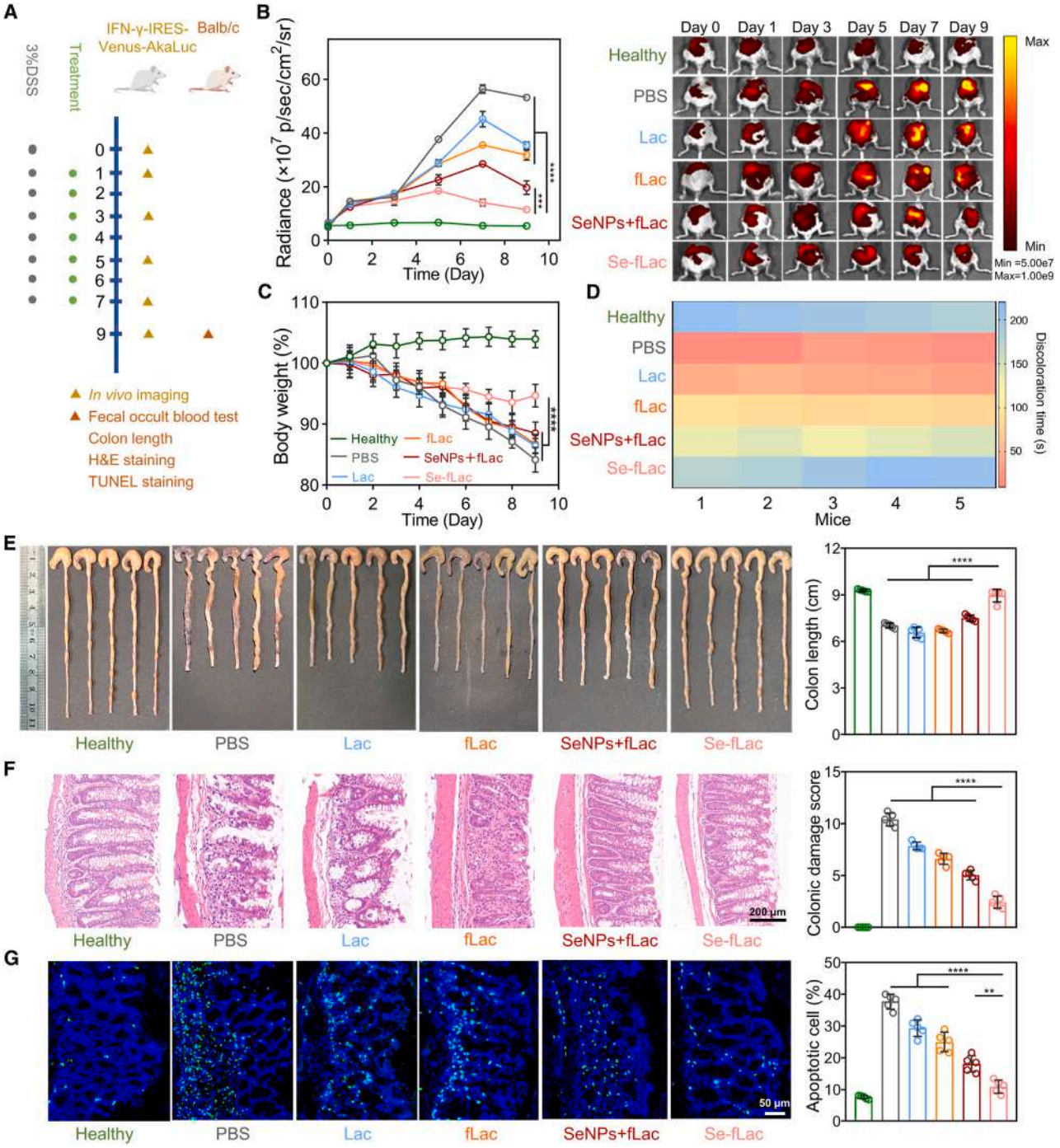


Figure 4. Therapeutic efficacies of Se-fLac in DSS-induced colitis mouse model

(A) Experimental design for establishment, treatment, and evaluation in a DSS-induced colitis mouse model. From days 0 to 7, IFN-γ reporter mice and BALB/c mice were given 3% DSS for model induction, while non-model control animals ("healthy") were administered with plain water. From days 1 to 7, model mice were orally administered with 1×10^8 CFU Lac, fLac, SeNPs + fLac, or Se-fLac (with selenium at an equivalent mass). Indices were detected at the indicated time points.

(B) Quantitative analysis of bioluminescence signal in IFN-γ reporter mice at indicated time points. Representative bioluminescence images of IFN-γ reporter mice are displayed in the right panel.

(C) Daily body weight changes.

(D) Discoloration time of the fecal occult blood test.

(E) Representative photographs of excised colon tissues and the corresponding colon length measured via Image J software.

(legend continued on next page)

remarkably elevated IFN-γ production potentially mirrored the features of exacerbated inflammation in UC. Compared to that in the group treated with pristine Lac, IFN-γ production was suppressed in the fLac group, which might be attributed to the pericellular film enhancing the resistance, mucoadhesion, and viability. Once fLac was mixed with SeNPs (SeNPs + fLac) for ROS scavenging, a further suppressed IFN-γ production was observed, and similar effect was reported previously for anti-inflammatory treatment.[42,43] Given that the selenium dots had higher ROS scavenging activity over SeNPs and the embedment in the pericellular film of Lac cells facilitated the selenium dots to accumulate in the colon, Se-fLac treatment resulted in an even greater reduction in the IFN-γ level than SeNPs + fLac treatment, with the bioluminescence intensity approaching that of the healthy group.

In BALB/c model mice, Se-fLac treatment significantly attenuated body weight loss, while other groups showed >10% reductions in body weight compared with the healthy group (Figure 4C). On day 9, we conducted a fecal occult blood test and found that all mice in the PBS group exhibited discoloration times shorter than 50 s, whereas the mean discoloration time in the Se-fLac group was up to 180 s (Figure 4D). At this time point, the colon length of the Se-fLac group was significantly longer than that of the other treatment groups (Figure 4E). At the histological level, Se-fLac treatment protected the colonic epithelium against pathological damage, and the features of intact epithelia, regular fingerlike crypt structure, and few inflammatory cells together led to the lowest colonic damage score and the mostly reduced level of apoptosis (Figures 4F and 4G).

To further validate the therapeutic efficacy of Se-fLac, we compared it with 5-aminosalicylic acid (5-ASA) treatment, VSL#3 supplements, and their combination (Figures S3E–S3H). Se-fLac significantly outperformed these three counterparts in all therapeutic indices. Considering the difficulty of treating severe UC, we also compared the efficacies of Se-fLac with steroids (dexamethasone [DEX]) in the setting of delayed therapy (Figures S4A–S4E). Although these two treatments had comparable effects in the relieved inflammation, Se-fLac showed more superiorities, especially in protecting against body weight loss, colon shortening, and pathological damage. Given that UC is characterized by recurring episodes of inflammation in the colon, we also established another two chronic colitis models and again confirmed the excellent performance of our Se-fLac (Figures S4F–S4L), thus showing the therapeutic breadth of Se-fLac against UC.

Note that we also conducted a series of biosafety evaluations (Figure S5). Both short-term and long-term studies showed no signs of abnormalities in the histological sections of main organs and no significant difference in hematological parameters or serum biochemical indices, indicating good biosafety of oral delivering Se-fLac. It also deserved notion that even multiple doses of Se-fLac did not significantly increase the serum selenium concentration. After conversion for human use, the equivalent dose of Se-fLac was ~50 μg Se/day (below maximum supplementation of 400 μg Se/day), further excluding the potential side effects (such as type 2 diabetes) associated with selenium overconsumption.[22]

Effective amelioration of inflammatory environment in the colon

To explore the underlying mechanism of Se-fLac's potent therapeutic efficacies, we first focused on the local inflammatory environment in mice on day 9. As the ELISA data shown in Figures 5A and S6A, the levels of proinflammatory cytokines (interleukin [IL]-1β, IL-6, and tumor necrosis factor alpha [TNF-α]) decreased in the sequence of PBS, Lac, fLac, SeNPs + fLac, and Se-fLac groups, while the levels of anti-inflammatory cytokines (IL-10 and transforming growth factor β [TGF-β]) followed a reversed order. Further flow cytometry analysis showed that the accumulation of inflammatory cells (T helper 17 [Th17] cells and macrophages) also decreased after Se-fLac treatment (Figures 5B, S6B, and S6C). Meanwhile, the proportion of Treg cells in the Se-fLac group was 1.8-fold higher than that in the PBS group (Figures 5C and S6D). To investigate the molecular basis of such relieved inflammation, we performed RNA-seq on the colon tissues and found Se-fLac treatment with downregulated gene sets of IBD and IL-17 signaling pathway (Figure 5D). Moreover, Se-fLac treatment significantly downregulated genes associated with inflammation (e.g., *Cxcl2*, *Cd163*, and *Il6*), and genes limiting inflammatory response (e.g., *Cbs* and *Dlk1*) were simultaneously upregulated (Figure 5E).

We speculated that above effective amelioration of inflammatory environment in Se-fLac group could be attributed to the efficient codelivery of both Lac cells and selenium dots. In the aspect of the selenium dots, two-photon microscopic imaging of dihydroethidium (DHE)-stained colon indicated a much lower ROS level in the Se-fLac group (Figures 5F and S6E). Similarly, the ROS indicator 8-OHdG also showed a 2.2-fold decrease compared to that in the PBS group, while the levels of superoxide dismutase (SOD), glutathione (GSH), and glutathione peroxidase 4 (GPX4) showed varying degrees of increase compared to those in the PBS group (Figure 5G). Given that ROS can serve as second messengers to trigger proinflammatory cascades, we further observed strong correlations between either two aspects of ROS levels, proinflammatory cytokine levels, and inflammatory cell proportions (Figures 5H and S6F), thus highlighting the importance of reducing ROS levels for the restoration of redox homeostasis.

Shifting to the Lac cell component, we focused our attention on the SCFAs often produced by purportedly beneficial bacteria, which have been demonstrated with the capacity of regulating the size of the colonic Treg pool.[44,45] Analysis of fecal samples showed that the Se-fLac group exhibited higher levels of SCFAs (e.g., acetate, propionate, and butyrate) than that from other treatment groups (Figure 5I). Upon comparing the Treg cells proportions in wild-type (WT) and G protein-coupled receptor 43 (GPR43)-knockout mice, we further verified that GPR43

(F) Representative hematoxylin and eosin (H&E) staining images of colon sections and the colonic damage scores.
(G) Representative terminal deoxynucleotidyl transferase dUTP nick-end labeling (TUNEL) staining images of colon sections and quantification of apoptotic cells.
Data in (B), (C), (E), (F), and (G) are presented as mean ± SD from 5 biologically independent samples. These data were compared by one-way ANOVA, and the statistical significance was calculated by comparing Se-fLac group with PBS and other treatment groups. $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$.

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

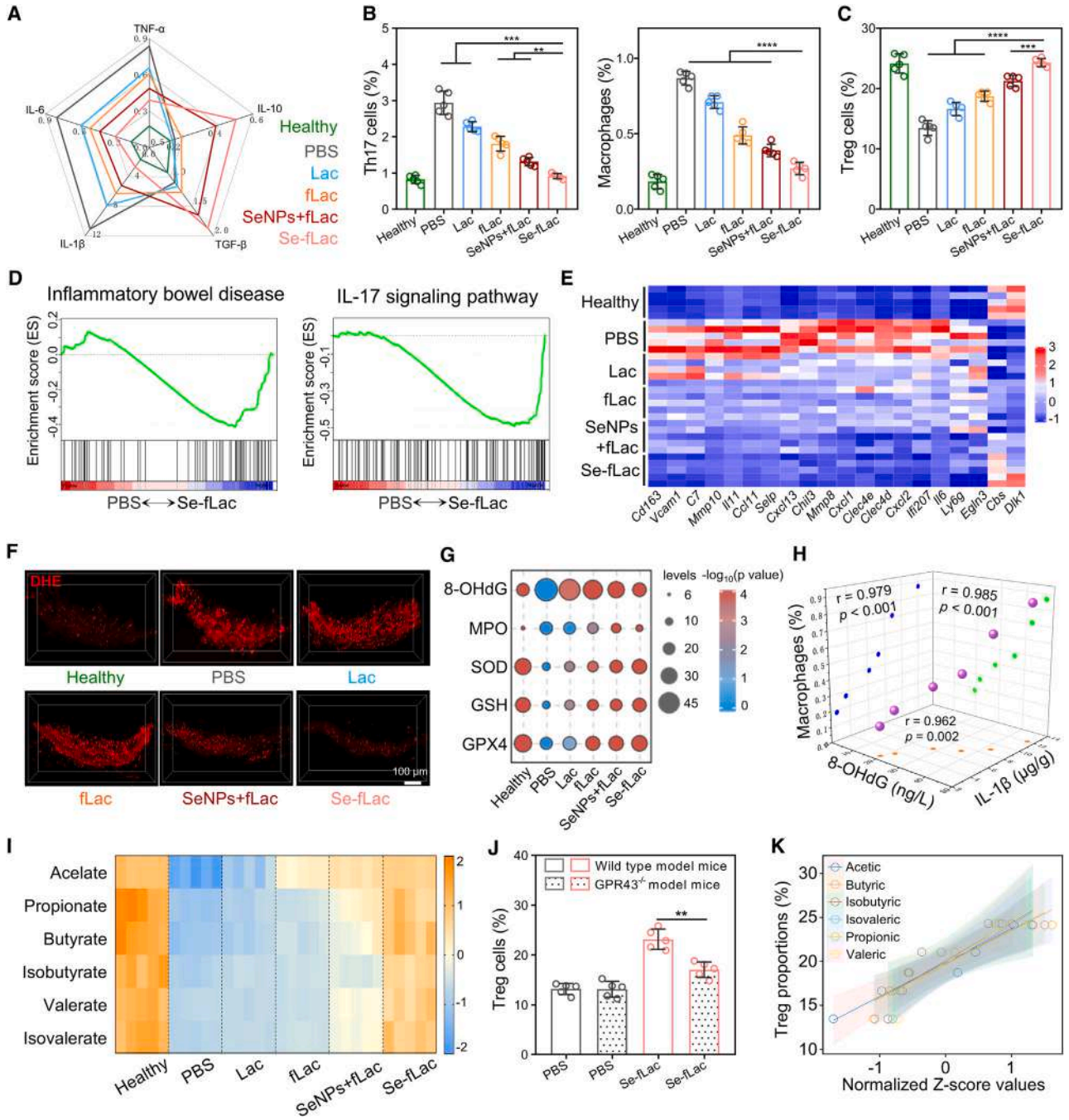


Figure 5. Amelioration of inflammatory environment after Se-fLac treatment

(A) ELISA-based measurements of IL-6, IL-1β, TGF-β, IL-10, and TNF-α levels in colon tissue lysates.

(B) Quantitative flow cytometry analysis of Th17 cells ($CD4^+IL\text{-}17A^+$ cells, gated from $CD3^+$ cells) and macrophages ($CD11b^+$ $F4/80^+$ cells, gated from $CD45^+$ cells) in the colonic lamina propria (LP).

(C) Quantitative flow cytometry analysis of Treg cells ($CD25^+$ $Foxp3^+$ cells, gated from $CD3^+$ $CD4^+$ cells) in the LP.

(D) Gene set enrichment analysis (GSEA) of gene sets with expression profiles that differed between PBS group and Se-fLac group.

(E) Hierarchical clustering heatmap showing DEGs associated with inflammation.

(F) *In vivo* two-photon images of ROS content in the DHE-stained colon.

(G) Detection of redox homeostasis-related indices in collected colon tissues and fecal samples.

(H) Correlation analysis examining various relationships among the ROS levels, IL-1β levels, and macrophage proportions in colon tissues.

(I) Heatmap showing the normalized *Z* scores for SCFAs.

(J) Utilization of GPR43 knockout mice to determine the increase in the Treg cells proportions induced by SCFAs.

(K) Correlation analysis examining the correlations between SCFA levels and Treg cells proportions.

(legend continued on next page)

was indeed the receptor for these SCFAs exerting their function of elevating Tregs proportion (Figure 5J).[46] In addition, strong correlations were observed between SCFA levels and Treg cells proportions (Figure 5K), thus indicating the important role of SCFAs in Treg cell-mediated amelioration of inflammatory environment.

Modulation of gut microbiota and potential mutual action with ROS reduction

In addition to amelioration of the inflammatory environment in the colon, we also realized the attribution of modulating gut microbiota for the therapeutic performance of Se-fLac (Figures S7A–S7C). Analysis of 16S ribosomal RNA (rRNA) gene sequencing data revealed that Se-fLac treatment most significantly improved bacterial richness (observed operational taxonomic unit [OTU] richness) and diversity (Shannon and Chao diversity indices) (Figure 6A). Principal coordinate analysis (PCoA) of microbiota composition showed that the Se-fLac group was more similar to the healthy group compared to the other treatment groups (Figure 6B). A similar trend was also reflected in analysis of the relative proportions of genus-level taxa (Figure 6C), and Se-fLac group showed increased relative abundance of purportedly beneficial bacteria (e.g., *Lactobacillus*; Figures 6D and S7D) and decreased relative abundance of deleterious taxa (e.g., *Enterobacteriaceae*; Figures 6E and S7E) compared to PBS group, supporting the tendency to restore microbiota homeostasis.

Upon initial examination, it may seem that supplementation with Lac cells would be the primary factor in modulating the gut microbiota. However, our results demonstrated that Se-fLac group had a greater tendency toward microbiota modulation than did the fLac group. This intriguing finding suggested that ROS scavenging by selenium dots might also benefit microbiota modulation. Considering that ROS could transform into electron acceptor products in the intestinal lumen and recalling that the deleterious taxa in Figure 6E are mostly facultative anaerobes,[47,48] we herein envisioned that the reduced ROS content might participate in mitigating electron acceptor-mediated anaerobic respiration and further inhibit the overgrowth of facultative anaerobes (Figure 6F). Validating this, Se-fLac group showed lower trimethylamine *N*-oxide (TMAO) levels and *TorA* (encoding a TMAO reductase) expression than fLac group (Figures 6G and 6H).

Given that most facultative anaerobes are gram-negative and that lipopolysaccharide (LPS) can induce oxidative stress,[18,49] we also envisioned that the above-inhibited overgrowth of facultative anaerobes might also suppress ROS production (Figure 6F). Therefore, we next examined the expression of *Tlr4* (a known receptor for LPS) and *Duox2* (a known *Tlr4* target gene encoding a nicotinamide adenine dinucleotide phosphate oxidase enzyme for producing hydrogen peroxide). Mirroring the trends in the abundance of facultative anaerobic *Enterobacteriaceae* in the gut microbiota, the Se-fLac treatment group had the lowest levels of both *Tlr4* and *Duox2* (Figure 6I). The correlation analysis further indicated the mutual and beneficial relationship between lower ROS levels and reduced facultative anaerobic *Enterobacteriaceae* abundance in the Se-fLac group (Figure S7F), and such a synergistic restoration of redox-microbiota homeostasis plausibly explained the potent therapeutic efficacies.

Excellent therapeutic efficacies in a UC NHP model

To increase the clinical relevance of Se-fLac, we explored the therapeutic efficacies of Se-fLac in a UC NHP model. Although the establishment of the UC NHP model has been reported by a few studies,[50–52] the doses and durations of DSS oral administration varied widely among these studies. To establish a standardized and reproducible UC NHP model, we optimized the essential protocols: intragastrical administration was performed on cynomolgus monkeys (4–4.5 years old, 4.5 ± 0.5 kg) with 50 mL of 2.5% DSS/day for 10 days (Figure 7A). After treatment with Se-fLac, no sign of blood in fecal samples were found during the observation period (from days 6 to 10), with the discoloration time in the fecal occult blood test much similar to that of the healthy group (Figure 7B). In contrast, discoloration time gradually decreased in the PBS group, indicating progressive severe bleeding. On day 11, fecal samples were collected for further analysis. Compared with PBS group, Se-fLac group showed a half level of 8-OHdG (Figure 7C), the increased abundance of purportedly beneficial bacteria (e.g., *Ruminococcaceae*), and the decreased abundance of deleterious taxa (e.g., *Prevotellaceae* and *Lachnospiraceae*) (Figure 7D). The restoration of both redox and microbiota homeostasis by Se-fLac treatment further led to substantially relieved inflammation, which could be verified by multiple indices (including calprotectin, CRP, diamine oxidase, neutrophils, and platelets) showing no significant difference to those in the healthy group (Figure 7E).

Further focusing on the lesion site, we conducted several histological analyses. The B-mode abdominal ultrasonography in Figure 7F showed that the PBS group had a relatively thicker mucosal layer (1.96 mm) due to colonic mucosa edema (Figure S7G). In contrast, Se-fLac treatment reduced the thickness of the mucosal layer to 1.16 mm, with no significant difference to that of healthy group. Furthermore, colonoscopy revealed evident ulcers and bleeding in the mucosal layer of the PBS-treated UC model monkey, whereas neither the Se-fLac-treated monkey nor the healthy monkey exhibited these pathological characteristics (Figure 7G). Hematoxylin and eosin (H&E) staining and Alcian blue-periodic acid Schiff (Ab-PAS) staining also revealed that the biopsies from Se-fLac-treated monkey had normal histological features, such as intact epithelial layer, negligible inflammatory cell infiltration, and normal secretion of mucins (Figures 7H and 7I). In addition to the above potent therapeutic outcomes, Se-fLac was also demonstrated to be safe for treating these monkeys (Figures S7H–S7J), further strengthening the promise as a modality for UC therapy.

Data in (B), (C), and (J) are presented as mean ± SD from 5 biologically independent samples. Data in (I) was measured by fecal samples from 5 independent mice in each group. Data in (B), (C), and (G) were compared via one-way ANOVA. Data in (J) were compared via two-tailed unpaired Student's t test. Data in (H) and (K) were compared via correlation analysis. The statistical significance in (B) and (C) was calculated by comparing Se-fLac group with PBS and other treatment groups, and that in (G) was calculated by comparing PBS group with the other groups. $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$.

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

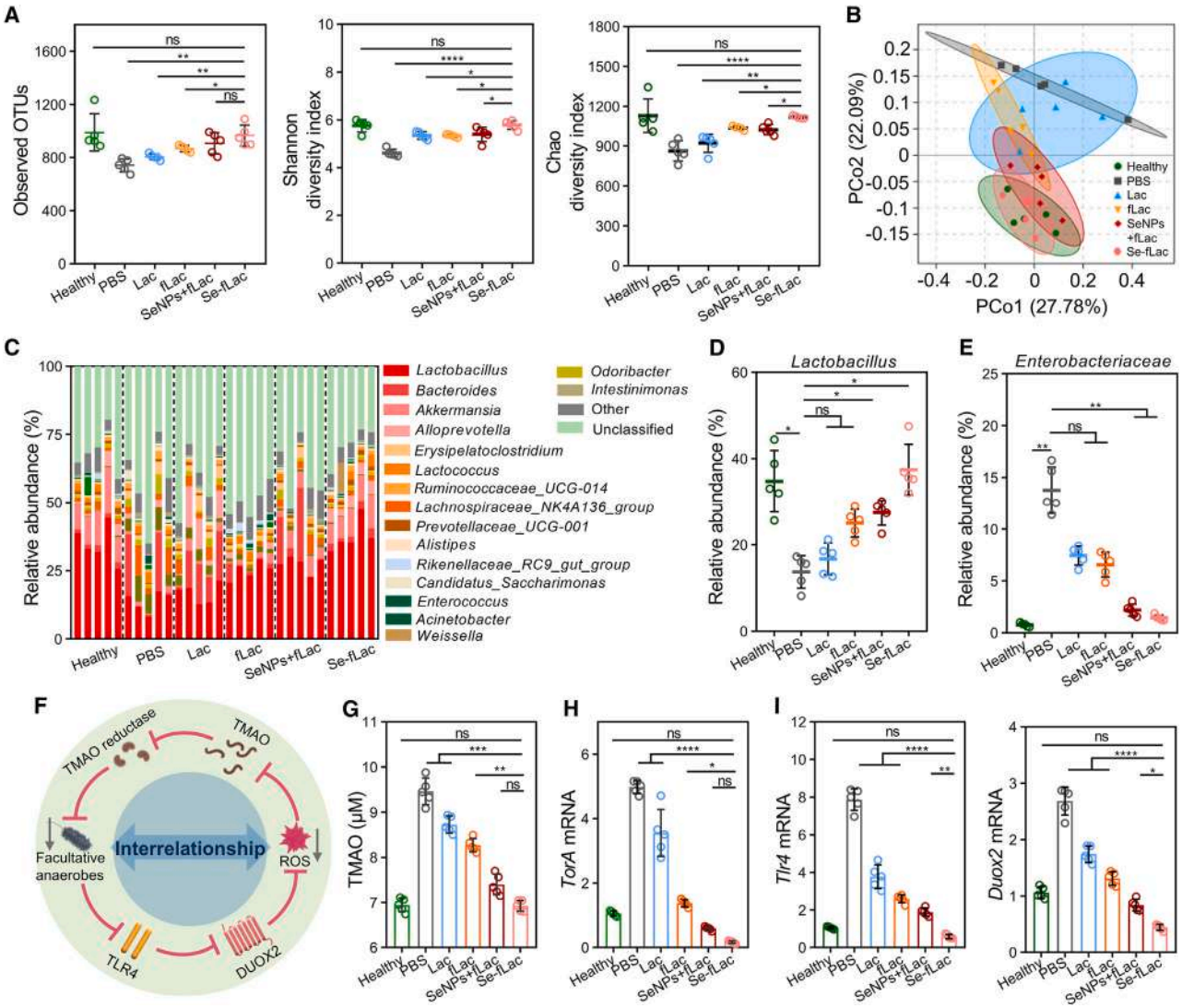


Figure 6. Modulation of gut microbiota and the potential interrelationship with ROS reduction

(A) Estimation of microbial community observed OTUs and α-diversity indices (Shannon and Chao indices) of gut microbiota.

(B) PCoA at the phylum level illustrating the gut microbiota β-diversity of the different groups.

(C) Relative abundance of gut microbiota at genus level showing as a percentage of total sequences.

(D) Relative abundance of *Lactobacillus* at genus level.

(E) Relative abundance of *Enterobacteriaceae* at family level.

(F) Schematic illustration of the potentially mutual benefit of Se-fLac treatment on reducing both ROS and facultative anaerobes.

(G and H) (G) ELISA-based measurement of serum TMAO levels and (H) *TorA* expression in the gut microbiota, measured by ELISA and real-time quantitative reverse transcription polymerase chain reaction (RT-qPCR).

(I) Gene expression of *Tlr4* and *Duox2* in colon tissues determined by RT-qPCR.

Data in (A), (D), (E), and (H) are presented as mean ± SD, and were measured by fecal samples from 5 independent mice in each group. Data in (G) and (I) are presented as mean ± SD from 5 biologically independent samples. Data in (A), (G), (H), and (I) were compared via one-way ANOVA. Data in (D) and (E) were compared via ALDEx2. The statistical significance in (A), (G), (H), and (I) was calculated by comparing Se-fLac group with the other groups. The statistical significance in (D) and (E) was calculated by comparing PBS group with the other groups. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$, ns: non-significance.

DISCUSSION

For the rational design of a therapeutic for UC treatment, we herein conducted data mining of publicly available datasets. Although the selection of ROS for targeted intervention and *Lactobacillus* genus for supplement had been reported for UC treatment, we herein offer the rationalities. On the one hand, our bioinformatics analyses unsealed and validated the high activity of oxidative stress program from a comprehensive perspective across all cell types in the inflamed colon. On the other hand, we employed and validated an L1-regularized logistic regression model to identify the most negative feature weight of *Lactobacillus* genera for UC.

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

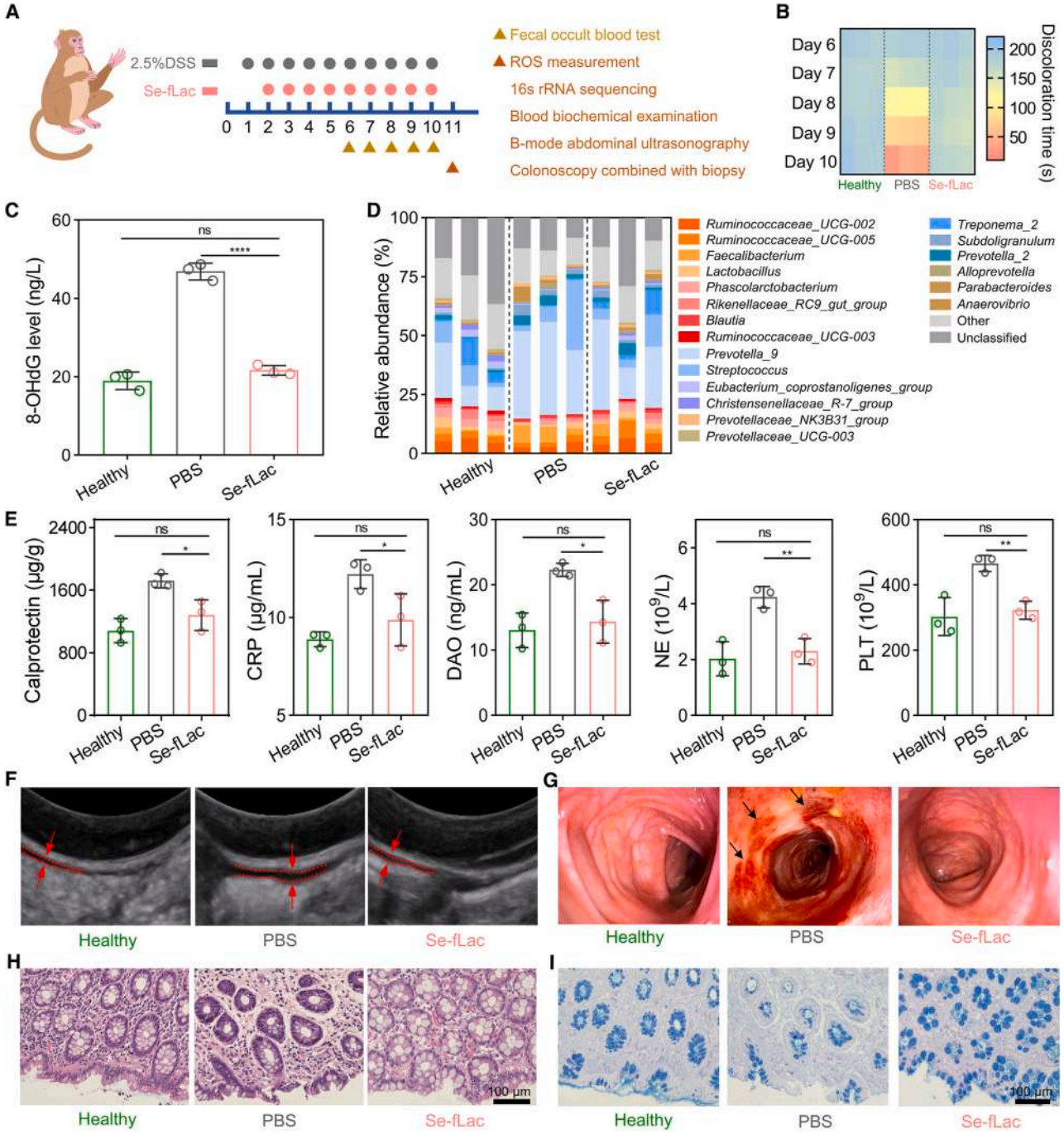


Figure 7. Therapeutic efficacies of Se-fLac in a UC NHP model

(A) Experimental design for establishment, treatment, and evaluation in a UC NHP model. From days 1 to 10, cynomolgus monkeys were provided with water (healthy group) or 2.5% DSS (PBS group) for 10 days. From days 2 to 10, monkeys in the Se-fLac group were orally administered with 1×10^9 CFU Se-fLac. Indices were detected at the indicated time points.

(B) Discoloration time of the fecal occult blood test.

(C) ELISA-based measurement of 8-OHdG level in fecal samples.

(D) Relative abundance of gut microbiota at genus level showing as a percentage of total sequences.

(E) Multiple indices detected in blood and fecal samples. The indices included calprotein, CRP, diamine oxidase (DAO), neutrophils (NE), and platelet (PLT) counts.

(F) Representative images of B-mode abdominal ultrasonography for mucosal layer thickness measurement.

(legend continued on next page)

Although the UC mouse models have been widely utilized to examine therapeutic efficacies, these mouse models might fail to well recapitulate the anatomic structures and physiological functions of the gastrointestinal tract, as well as pathological features of UC in humans. In contrast, there are substantial similarities between NHP and humans, including but not limited to the gastric juice components, gastric emptying time, colon length, and gut microbiota.[53–55] This promoted us to establish an NHP model to bridge mouse- and human-based investigations of Se-fLac, and the potent therapeutic efficacies highlighted the strong translational potential of Se-fLac to develop clinically relevant UC therapeutic.

In addition to the examination of an NHP model, clinically used therapeutics and supplements (5-ASA, steroids, VSL#3, and 5-ASA + VSL#3) were also utilized for comparison. Beyond emphasizing the superior therapeutic efficacies of Se-fLac, several issues bear more considerations. In terms of clinical indications, 5-ASA has been recommended for treating patients with mild to moderate UC, while steroids can be applied to patients with moderate to severe UC.[1] However, Se-fLac exhibited potent therapeutic efficacies in two UC models with distinct disease severities, thus shedding a light on a wide indication for clinical UC treatment. Moreover, for those patients showing no response to 5-ASA treatment or suffering from steroid-associated complications, Se-fLac can also be considered a potential alternative therapeutic.

Besides anti-inflammatory and probiotic agents, natural antioxidants (such as vitamin C and coenzyme Q) have also been tested in a few trials.[56,57] Unfortunately, the curative effects were not impressive for UC, which might be mainly attributed to the instability and poor accumulation of these antioxidants in the colon. Alternatively, researchers have considered the possibility of delivering antioxidative enzymes to the inflamed colon.[21] Given that ROS include multiple types of free radicals and that each antioxidative enzyme scavenges a specific type of free radical, using single antioxidative enzyme was insufficient to remarkably reduce the ROS level. Keeping this in mind, we herein developed a facile and mild approach to achieve *in situ* growth of selenium dots in the pericellular film. These selenium dots not only had good resistance in the gastrointestinal tract but also hitch-hiked the fLac cells to accumulate in the colon. At the lesion site, these selenium dots reacted with multiple types of free radicals, thus ensuring the efficient performance of scavenging ROS.

For guaranteeing and improving the performance of ML approaches, substantially larger and well-characterized datasets are important. Beyond validating the classification performance and contribution of specific bacterial genera with L1-regularized logistic regression model, future efforts can also be made for further validation in newly reported datasets. Another limitation is that the multiple models were all chemically induced. However, we envision that our Se-fLac may also exert potent therapeutic efficacies in T cell adoptive transfer model via synergistic restoration of redox-microbiota homeostasis, since this model has been utilized to demonstrate the therapeutic benefits of the administration of either butyrate-producing bacteria or natural antioxidants.[58,59] Note that growing selenium dots-embedded pericellular film is applicable to diverse bacteria. Upon identifying the purportedly beneficial bacteria with decreased abundance for individual patients, our technology can also be employed to develop personalized therapeutic, thus empowering the precise bacteriotherapy for UC treatment.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Mouse models
  - NHP model
  - Human samples
- METHOD DETAILS
  - Model training and evaluation
  - Bioinformatics analyses of publicly available datasets
  - Preparation and characterization of Se-fLac
  - Resistance assessment *in vitro* and *ex vivo*
  - Evaluation of mucoadhesion
  - *In vivo* viability after oral administration
  - DSS-induced colitis model
  - Therapeutic efficacy of Se-fLac against multiple colitis models
  - Biosafety assessment *in vivo*
  - Measurement of inflammatory environment in the colon
  - Two-photon fluorescence microscopy imaging of ROS in the colon
  - RNA-sequencing analysis
  - Microbiota analysis
  - DSS-induced UC NHP model
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2024.07.028.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant nos. T2225021, U2001224, 21821005, 32101120, and 82373950) and Beijing Natural Science Foundation (JQ21027).

AUTHOR CONTRIBUTIONS

W. Wei, G.M., and Y.C. conceived and designed the study. T.L. and J.O. conducted logistic regression model analysis. D.W., P.G., and W. Wei conducted bioinformatics analyses. P.G., Q.X., J.L., Y.C., R.J., M.S., Y.W., and W. Wei analyzed the clinical samples. P.G. and W. Wang constructed the therapeutic and performed *in vitro* experiments. P.G., W. Wang, and Y.Z. performed and analyzed the time-resolved transient absorption spectroscopy. P.G. and W. Wang performed *in vivo* experiments on multiple UC mouse models. D.Z., J. Zou, and J. Zhao discussed the results and gave advice. P.G., W. Wang, J.

(G) Representative colonscopic images showing ulcers and bleeding in the mucosa layer.
(H and I) (H) Representative H&E staining and (I) Alcian blue-periodic acid Schiff (Ab-PAS) staining images of colon biopsies.
Data in (C) and (E) are presented as mean ± SD from 3 biologically independent samples. These data were compared via one-way ANOVA and the statistical significance was calculated by comparing Se-fLac group with the other groups. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{****}p < 0.0001$, ns: non-significance.

Zou, C.P., and Y.Q. performed UC NHP experiments and analyzed the data. W. Wei and P.G. wrote the manuscript. W. Wei and G.M. further revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 12, 2023
Revised: May 16, 2024
Accepted: July 31, 2024
Published: August 27, 2024

REFERENCES

1. Ungaro, R., Mehandru, S., Allen, P.B., Peyrin-Biroulet, L., and Colombel, J.F. (2017). Ulcerative colitis. Lancet *389*, 1756–1770. https://doi.org/10.1016/S0140-6736(16)32126-2.
2. Bernstein, C.N., Fried, M., Krabshuis, J.H., Cohen, H., Eliakim, R., Fedail, S., Gearry, R., Goh, K.L., Hamid, S., Khan, A.G., et al. (2010). World Gastroenterology Organization Practice Guidelines for the diagnosis and management of IBD in 2010. Inflamm. Bowel Dis. *16*, 112–124. https://doi.org/10.1002/ibd.21048.
3. Marti-Aguado, D., Ballester, M.P., Tosca, J., Bosca-Watts, M.M., Navarro, P., Anton, R., Pascual, I., Mora, F., and Minguez, M. (2019). Long-term follow-up of patients treated with aminosalicylates for ulcerative colitis: predictive factors of response: an observational case-control study. U. Eur. Gastroenterol. J. *7*, 1042–1050. https://doi.org/10.1177/2050640619854277.
4. Kobayashi, T., Siegmund, B., Le Berre, C., Wei, S.C., Ferrante, M., Shen, B., Bernstein, C.N., Danese, S., Peyrin-Biroulet, L., and Hibi, T. (2020). Ulcerative colitis. Nat. Rev. Dis. Primers *6*, 74. https://doi.org/10.1038/s41572-020-0205-x.
5. Monstad, I.L., Solberg, I.C., Cvancarova, M., Hovde, O., Henriksen, M., Huppertz-Hauss, G., Gunther, E., Moum, B.A., Stray, N., Vatn, M., et al. (2021). Outcome of ulcerative colitis 20 years after diagnosis in a prospective population-based inception cohort from South-Eastern Norway, the IBSEN study. J. Crohns Colitis *15*, 969–979. https://doi.org/10.1093/ecco-jcc/jjaa232.
6. Chu, H., Khosravi, A., Kusumawardhani, I.P., Kwon, A.H., Vasconcelos, A.C., Cunha, L.D., Mayer, A.E., Shen, Y., Wu, W.L., Kambal, A., et al. (2016). Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. Science *352*, 1116–1120. https://doi.org/10.1126/science.aad9948.
7. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662. https://doi.org/10.1038/s41586-019-1237-9.
8. Plichta, D.R., Graham, D.B., Subramanian, S., and Xavier, R.J. (2019). Therapeutic opportunities in inflammatory bowel disease: mechanistic dissection of host-microbiome relationships. Cell *178*, 1041–1056. https://doi.org/10.1016/j.cell.2019.07.045.
9. Cao, F., Jin, L., Gao, Y., Ding, Y., Wen, H., Qian, Z., Zhang, C., Hong, L., Yang, H., Zhang, J., et al. (2023). Artificial-enzymes-armed Bifidobacterium longum probiotics for alleviating intestinal inflammation and microbiota dysbiosis. Nat. Nanotechnol. *18*, 617–627. https://doi.org/10.1038/s41565-023-01346-x.
10. Shen, Z.H., Zhu, C.X., Quan, Y.S., Yang, Z.Y., Wu, S., Luo, W.W., Tan, B., and Wang, X.Y. (2018). Relationship between intestinal microbiota and ulcerative colitis: mechanisms and clinical application of probiotics and fecal microbiota transplantation. World J. Gastroenterol. *24*, 5–14. https://doi.org/10.3748/wjg.v24.i1.5.
11. Koretz, R.L. (2018). Probiotics in gastroenterology: how pro is the evidence in adults? Am. J. Gastroenterol. *113*, 1125–1136. https://doi.org/10.1038/s41395-018-0138-0.
12. Suez, J., Zmora, N., Segal, E., and Elinav, E. (2019). The pros, cons, and many unknowns of probiotics. Nat. Med. *25*, 716–729. https://doi.org/10.1038/s41591-019-0439-x.
13. Lin, S., Mukherjee, S., Li, J., Hou, W., Pan, C., and Liu, J. (2021). Mucosal immunity-mediated modulation of the gut microbiome by oral delivery of probiotics into Peyer's patches. Sci. Adv. *7*, eabf0677. https://doi.org/10.1126/sciadv.abf0677.
14. Pan, J., Gong, G., Wang, Q., Shang, J., He, Y., Catania, C., Birnbaum, D., Li, Y., Jia, Z., Zhang, Y., et al. (2022). A single-cell nanocoating of probiotics for enhanced amelioration of antibiotic-associated diarrhea. Nat. Commun. *13*, 2117. https://doi.org/10.1038/s41467-022-29672-z.
15. Xavier, R.J., and Podolsky, D.K. (2007). Unravelling the pathogenesis of inflammatory bowel disease. Nature *448*, 427–434. https://doi.org/10.1038/nature06005.
16. Sies, H., and Jones, D.P. (2020). Reactive oxygen species (ROS) as pleiotropic physiological signalling agents. Nat. Rev. Mol. Cell Biol. *21*, 363–383. https://doi.org/10.1038/s41580-020-0230-3.
17. Dickinson, B.C., and Chang, C.J. (2011). Chemistry and biology of reactive oxygen species in signaling or stress responses. Nat. Chem. Biol. *7*, 504–511. https://doi.org/10.1038/nchembio.607.
18. Campbell, E.L., and Colgan, S.P. (2019). Control and dysregulation of redox signalling in the gastrointestinal tract. Nat. Rev. Gastroenterol. Hepatol. *16*, 106–120. https://doi.org/10.1038/s41575-018-0079-5.
19. Lee, Y., Sugihara, K., Gillilland, M.G., 3rd, Jon, S., Kamada, N., and Moon, J.J. (2020). Hyaluronic acid-bilirubin nanomedicine for targeted modulation of dysregulated intestinal barrier, microbiome and immune responses in colitis. Nat. Mater. *19*, 118–126. https://doi.org/10.1038/s41563-019-0462-9.
20. Liu, J., Wang, Y., Heelan, W.J., Chen, Y., Li, Z., and Hu, Q. (2022). Mucoadhesive probiotic backpacks with ROS nanoscavengers enhance the bacteriotherapy for inflammatory bowel diseases. Sci. Adv. *8*, eabp8798. https://doi.org/10.1126/sciadv.abp8798.
21. Zhou, J., Li, M., Chen, Q., Li, X., Chen, L., Dong, Z., Zhu, W., Yang, Y., Liu, Z., and Chen, Q. (2022). Programmable probiotics modulate inflammation and gut microbiota for inflammatory bowel disease treatment after effective oral delivery. Nat. Commun. *13*, 3432. https://doi.org/10.1038/s41467-022-31171-0.
22. Rayman, M.P. (2012). Selenium and human health. Lancet *379*, 1256–1268. https://doi.org/10.1016/S0140-6736(11)61452-9.
23. Huang, Z., Rose, A.H., and Hoffmann, P.R. (2012). The role of selenium in inflammation and immunity: from molecular mechanisms to therapeutic opportunities. Antioxid. Redox Signal. *16*, 705–743. https://doi.org/10.1089/ars.2011.4145.
24. Huang, Q., Liu, Z., Yang, Y., Yang, Y., Huang, T., Hong, Y., Zhang, J., Chen, Q., Zhao, T., Xiao, Z., et al. (2023). Selenium nanodots (SENDs) as antioxidants and antioxidant-prodrugs to rescue islet β cells in type 2 diabetes mellitus by restoring mitophagy and alleviating endoplasmic reticulum stress. Adv. Sci. (Weinh) *10*, e2300880. https://doi.org/10.1002/advs.202300880.
25. Wang, X., Cao, Z., Zhang, M., Meng, L., Ming, Z., and Liu, J. (2020). Bioinspired oral delivery of gut microbiota by self-coating with biofilms. Sci. Adv. *6*, eabb1952. https://doi.org/10.1126/sciadv.abb1952.
26. Al-Hadidi, A., Navarro, J., Goodman, S.D., Bailey, M.T., and Besner, G.E. (2021). Lactobacillus reuteri in Its biofilm State Improves Protection from Experimental Necrotizing Enterocolitis. Nutrients *13*, 918. https://doi.org/10.3390/nu13030918.
27. Salas-Jara, M.J., Ilabaca, A., Vega, M., and García, A. (2016). Biofilm forming Lactobacillus: new challenges for the development of probiotics. Microorganisms *4*, 35. https://doi.org/10.3390/microorganisms4030035.
28. Yan, J., and Bassler, B.L. (2019). Surviving as a community: antibiotic tolerance and persistence in bacterial biofilms. Cell Host Microbe *26*, 15–21. https://doi.org/10.1016/j.chom.2019.06.002.

29. Geng, Z., Wang, X., Wu, F., Cao, Z., and Liu, J. (2023). Biointerface mineralization generates ultraresistant gut microbes as oral biotherapeutics. Sci. Adv. *9*, eade0997. https://doi.org/10.1126/sciadv.ade0997.

30. Song, X., Xiong, Z., Kong, L., Wang, G., and Ai, L. (2018). Relationship between Putative eps Genes and Production of exopolysaccharide in Lactobacillus casei LC2W. Front. Microbiol. *9*, 1882. https://doi.org/10.3389/fmicb.2018.01882.

31. Topçuoğlu, B.D., Lesniak, N.A., Ruffin, M.T.t., Wiens, J., and Schloss, P.D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. mBio *11*, e00434-20. https://doi.org/10.1128/mBio.00434-20.

32. Dai, D., Zhu, J., Sun, C., Li, M., Liu, J., Wu, S., Ning, K., He, L.J., Zhao, X.M., and Chen, W.H. (2022). GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. Nucleic Acids Res. *50*, D777–D784. https://doi.org/10.1093/nar/gkab1019.

33. Ganji-Arjenaki, M., and Rafieian-Kopaei, M. (2018). Probiotics are a good choice in remission of inflammatory bowel diseases: A meta analysis and systematic review. J. Cell. Physiol. *233*, 2091–2103. https://doi.org/10.1002/jcp.25911.

34. Argmann, C., Hou, R., Ungaro, R.C., Irizar, H., Al-Taie, Z., Huang, R., Kosoy, R., Venkat, S., Song, W.M., Di'Narzo, A.F., et al. (2023). Biopsy and blood-based molecular biomarker of inflammation in IBD. Gut *72*, 1271–1287. https://doi.org/10.1136/gutjnl-2021-326451.

35. Nie, H., Lin, P., Zhang, Y., Wan, Y., Li, J., Yin, C., and Zhang, L. (2023). Single-cell meta-analysis of inflammatory bowel disease with scIBD. Nat. Comput. Sci. *3*, 522–531. https://doi.org/10.1038/s43588-023-00464-9.

36. D'Incà, R., Cardin, R., Benazzato, L., Angriman, I., Martines, D., and Sturniolo, G.C. (2004). Oxidative DNA damage in the mucosa of ulcerative colitis increases with disease duration and dysplasia. Inflam. Bowel Dis. *10*, 23–27. https://doi.org/10.1097/00054725-200401000-00003.

37. Wang, S., Chen, Y., Han, S., Liu, Y., Gao, J., Huang, Y., Sun, W., Wang, J., Wang, C., and Zhao, J. (2022). Selenium nanoparticles alleviate ischemia reperfusion injury-induced acute kidney injury by modulating GPx-1/NLRP3/Caspase-1 pathway. Theranostics *12*, 3882–3895. https://doi.org/10.7150/thno.70830.

38. Zhang, C., Wang, H., Yang, X., Fu, Z., Ji, X., Shi, Y., Zhong, J., Hu, W., Ye, Y., Wang, Z., et al. (2022). Oral zero-valent-molybdenum nanodots for inflammatory bowel disease therapy. Sci. Adv. *8*, eabp9882. https://doi.org/10.1126/sciadv.abp9882.

39. Wolff, C.M., Frischmann, P.D., Schulze, M., Bohn, B.J., Wein, R., Livadas, P., Carlson, M.T., Jäckel, F., Feldmann, J., Würthner, F., et al. (2018). All-in-one visible-light-driven water splitting by combining nanoparticulate and molecular co-catalysts on CdS nanorods. Nat. Energy *3*, 862–869. https://doi.org/10.1038/s41560-018-0229-6.

40. Schubert, M.L. (2017). Physiologic, pathophysiologic, and pharmacologic regulation of gastric acid secretion. Curr. Opin. Gastroenterol. *33*, 430–438. https://doi.org/10.1097/MOG.0000000000000392.

41. Wirtz, S., Popp, V., Kindermann, M., Gerlach, K., Weigmann, B., Fichtner-Feigl, S., and Neurath, M.F. (2017). Chemically induced mouse models of acute and chronic intestinal inflammation. Nat. Protoc. *12*, 1295–1309. https://doi.org/10.1038/nprot.2017.044.

42. Esfandiari, Z., Vakili, B., Ahangarzadeh, S., Esfahani, S.N., and Shoaei, P. (2024). Impact of selenium nanoparticle-enriched lactobacilli feeding against Escherichia coli O157:H7 infection of BALB/c mice. Probiotics Antimicrob. Proteins *16*, 784–795. https://doi.org/10.1007/s12602-023-10081-7.

43. Qiao, L., Chen, Y., Song, X., Dou, X., and Xu, C. (2022). Selenium nanoparticles-enriched Lactobacillus casei ATCC 393 prevents cognitive dysfunction in mice through modulating microbiota-gut-brain axis. Int. J. Nanomed. *17*, 4807–4827. https://doi.org/10.2147/IJN.S374024.

44. Caruso, R., Lo, B.C., and Núñez, G. (2020). Host-microbiota interactions in inflammatory bowel disease. Nat. Rev. Immunol. *20*, 411–426. https://doi.org/10.1038/s41577-019-0268-7.

45. He, J., Chu, Y., Li, J., Meng, Q., Liu, Y., Jin, J., Wang, Y., Wang, J., Huang, B., Shi, L., et al. (2022). Intestinal butyrate-metabolizing species contribute to autoantibody production and bone erosion in rheumatoid arthritis. Sci. Adv. *8*, eabm1511. https://doi.org/10.1126/sciadv.abm1511.

46. Han, K., Nam, J., Xu, J., Sun, X., Huang, X., Animasahun, O., Achreja, A., Jeon, J.H., Pursley, B., Kamada, N., et al. (2021). Generation of systemic antitumour immunity via the in situ modulation of the gut microbiome by an orally administered inulin gel. Nat. Biomed. Eng. *5*, 1377–1388. https://doi.org/10.1038/s41551-021-00749-2.

47. Ni, J., Wu, G.D., Albenberg, L., and Tomov, V.T. (2017). Gut microbiota and IBD: causation or correlation? Nat. Rev. Gastroenterol. Hepatol. *14*, 573–584. https://doi.org/10.1038/nrgastro.2017.88.

48. Yoo, W., Zieba, J.K., Foegeding, N.J., Torres, T.P., Shelton, C.D., Shealy, N.G., Byndloss, A.J., Cevallos, S.A., Gertz, E., Tiffany, C.R., et al. (2021). High-fat diet-induced colonocyte dysfunction escalates microbiota-derived trimethylamine N-oxide. Science *373*, 813–818. https://doi.org/10.1126/science.aba3683.

49. Burgueño, J.F., Fritsch, J., González, E.E., Landau, K.S., Santander, A.M., Fernández, I., Hazime, H., Davies, J.M., Santaolalla, R., Phillips, M.C., et al. (2021). Epithelial TLR4 signaling activates DUOX2 to induce microbiota-driven tumorigenesis. Gastroenterology *160*, 797–808.e6. https://doi.org/10.1053/j.gastro.2020.10.031.

50. Hao, X.P., Lucero, C.M., Turkbey, B., Bernardo, M.L., Morcock, D.R., Deleage, C., Trubey, C.M., Smedley, J., Klatt, N.R., Giavedoni, L.D., et al. (2015). Experimental colitis in SIV-uninfected rhesus macaques recapitulates important features of pathogenic SIV infection. Nat. Commun. *6*, 8020. https://doi.org/10.1038/ncomms9020.

51. McQueen, P., Busman-Sahay, K., Rieder, F., Noël-Romas, L., McCorrister, S., Westmacott, G., Estes, J.D., and Burgener, A. (2019). Intestinal proteomic analysis of a novel non-human primate model of experimental colitis reveals signatures of mitochondrial and metabolic dysfunction. Mucosal Immunol. *12*, 1327–1335. https://doi.org/10.1038/s41385-019-0200-2.

52. Jiminez, J.A., Uwiera, T.C., Douglas Inglis, G., and Uwiera, R.R. (2015). Animal models to study acute and chronic intestinal inflammation in mammals. Gut Pathog. *7*, 29. https://doi.org/10.1186/s13099-015-0076-y.

53. Willmann, S., Edginton, A.N., and Dressman, J.B. (2007). Development and validation of a physiology-based model for the prediction of oral absorption in monkeys. Pharm. Res. *24*, 1275–1282. https://doi.org/10.1007/s11095-007-9247-y.

54. Clayton, J.B., Gomez, A., Amato, K., Knights, D., Travis, D.A., Blekhman, R., Knight, R., Leigh, S., Stumpf, R., Wolf, T., et al. (2018). The gut microbiome of nonhuman primates: lessons in ecology and evolution. Am. J. Primatol. *80*, e22867. https://doi.org/10.1002/ajp.22867.

55. Ikegami, K., Tagawa, K., Narisawa, S., and Osawa, T. (2003). Suitability of the cynomolgus monkey as an animal model for drug absorption studies of oral dosage forms from the viewpoint of gastrointestinal physiology. Biol. Pharm. Bull. *26*, 1442–1447. https://doi.org/10.1248/bpb.26.1442.

56. Halliwell, B. (2024). Understanding mechanisms of antioxidant action in health and disease. Nat. Rev. Mol. Cell Biol. *25*, 13–33. https://doi.org/10.1038/s41580-023-00645-4.

57. Bourgonje, A.R., Feelisch, M., Faber, K.N., Pasch, A., Dijkstra, G., and van Goor, H. (2020). Oxidative stress and redox-modulating therapeutics in inflammatory bowel disease. Trends Mol. Med. *26*, 1034–1046. https://doi.org/10.1016/j.molmed.2020.06.006.

58. Huang, L.J., Mao, X.T., Li, Y.Y., Liu, D.D., Fan, K.Q., Liu, R.B., Wu, T.T., Wang, H.L., Zhang, Y., Yang, B., et al. (2021). Multiomics analyses reveal a critical role of selenium in controlling T cell differentiation in Crohn's disease. Immunity *54*, 1728–1744.e7. https://doi.org/10.1016/j.immuni.2021.07.004.

59. Furusawa, Y., Obata, Y., Fukuda, S., Endo, T.A., Nakato, G., Takahashi, D., Nakanishi, Y., Uetake, C., Kato, K., Kato, T., et al. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. Nature *504*, 446–450. https://doi.org/10.1038/nature12721.

60. Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier,

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

E., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. *32*, 822–828. https://doi.org/10.1038/nbt.2939.

61. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics Oxf. Engl. *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

62. Zhong, Y., Zhang, J., Li, T., Xu, W., Yao, Q., Lu, M., Bai, X., Wu, Z., Xie, J., and Zhang, Y. (2023). Suppression of kernel vibrations by layer-by-layer ligand engineering boosts photoluminescence efficiency of gold nanoclusters. Nat. Commun. *14*, 658. https://doi.org/10.1038/s41467-023-36387-2.

63. Hsu, Y.L., Chen, C.C., Lin, Y.T., Wu, W.K., Chang, L.C., Lai, C.H., Wu, M.S., and Kuo, C.H. (2019). Evaluation and optimization of sample handling methods for quantification of short-chain fatty acids in human fecal samples by GC-MS. J. Proteome Res. *18*, 1948–1957. https://doi.org/10.1021/acs.jproteome.8b00536.

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-Mouse CD3	Biolegend	Cat# 100206; RRID: AB_312663
Anti-Mouse CD4	Biolegend	Cat# 100405; RRID: AB_312690
Anti-Mouse CD25	Biolegend	Cat# 101915; RRID: AB_2616761
Anti-Mouse/Rat Foxp3	ThermoFisher	Cat# 77-5775-40; RRID: AB_469981
Anti-Mouse IL-17A	ThermoFisher	Cat# 17-7177-81; RRID: AB_763580
Anti-Mouse CD45	ThermoFisher	Cat# 35-0451-82; RRID: AB_469718
Anti-Mouse CD11b	ThermoFisher	Cat# 17-0112-82; RRID: AB_ 469343
Anti-Mouse F4/80	ThermoFisher	Cat# 11-4801-82; RRID: AB_ 2637191
Ghost Dye	Tonbo	Cat# 13-0683; RRID: N/A
Bacterial and virus strains		
Lactobacillus casei	Lab stock	CICC20280
Biological samples		
Human fecal samples	This study	N/A
Chemicals, peptides, and recombinant proteins		
Sodium Selenite	Sigma-Aldrich	Cat# S5261
Ascorbic Acid	Sigma-Aldrich	Cat# A4544
S-nitroso-N-acetylpenicillamine	Sigma-Aldrich	Cat# N3398
5,5-Dimethyl-1-pyrroline-N-oxide (DMPO)	Santa Cruz Biotechnology	Cat# sc-202587
5-(Diethoxyphosphoryl)-5-methyl-1-pyrroline-N-oxide (DEPMPO)	Santa Cruz Biotechnology	Cat# sc-202132
D-Luciferin potassium	Merck	Cat# 50227
Cy5 NHS ester	Xi'an ruixi Biological Technology	Cat# R-FR-004
Dextran sodium sulfate	Shanghai Yuanye Bio-Technology	Cat# S14049
Dihydroethidium	Shanghai Yuanye Bio-Technology	Cat# S40454
RPMI 1640, with L-Alanyl-L-Glutamine	VivaCell	Cat# C3010-0500
Fetal Bovine Serum	Gibco	Cat# 10099141C
Liberase TM	Roche	Cat# 5401127001
DNase I	Roche	Cat# 10104159001
Cell Stimulation Cocktail (500X)	ThermoFisher	Cat# 00-4970-93
Brefeldin A (1000X Solution)	ThermoFisher	Cat# 00-4506-51
EDTA	ThermoFisher	Cat# AM9261
Percoll	Beijing Solarbio Science & Technology Co., Ltd.	Cat# P8370
Aka Lumine n-Hydrochloride	Wako	Cat# 012-26706
Trizol	ThermoFisher	Cat# 5596026CN
Critical commercial assays		
One step terminal deoxynucleotidyl transferase-mediated dUTP nick-end labeling (TUNEL) apoptosis assay kit	Beyotime Biotechnology	Cat# C1088
BeyoFast™ SYBR Green One-Step qRT-PCR Kit	Beyotime Biotechnology	Cat# D7268M
RNAeasy™ Animal RNA Isolation Kit with Spin Column	Beyotime Biotechnology	Cat# R0027
Mouse IL-10 ELISA KIT	Beijing Solarbio Science & Technology Co., Ltd.	Cat# SEKM-0010

(*Continued on next page*)

Please cite this article in press as: Guo et al., Engineered probiotic ameliorates ulcerative colitis by restoring gut microbiota and redox homeostasis, Cell Host & Microbe (2024), https://doi.org/10.1016/j.chom.2024.07.028

CellPress

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Mouse TGF-β ELISA KIT	Beijing Solarbio Science & Technology Co., Ltd.	Cat# SEKM-0035
Mouse IL-1β ELISA KIT	Beijing Solarbio Science & Technology Co., Ltd.	Cat# SEKM-0002
Mouse IL-6 ELISA KIT	Beijing Solarbio Science & Technology Co., Ltd.	Cat# SEKM-0007
Mouse TNF-α ELISA KIT	Beijing Solarbio Science & Technology Co., Ltd.	Cat# SEKM-0034
Micro Hydroxyl Free Radical Scavenging Capacity Assay Kit	Beijing Solarbio Science & Technology Co., Ltd.	Cat# BC1325
Superoxide Anion Activity Content Assay Kit	Beijing Solarbio Science & Technology Co., Ltd.	Cat# BC1295
Reduced Glutathione (GSH) Content Assay Kit	Beijing Solarbio Science & Technology Co., Ltd.	Cat# BC1175
Superoxide Dismutase (SOD) Activity Assay Kit	Beijing Solarbio Science & Technology Co., Ltd.	Cat# BC0175
DPPH Free Radical Scavenging Capacity Assay Kit	Beijing Solarbio Science & Technology Co., Ltd.	Cat# BC4755
Foxp3/Transcription Factor Staining Buffer Set	ThermoFisher	Cat# 00-5523-00
Intracellular Fixation & Permeabilization Buffer Set	ThermoFisher	Cat# 88-8824-00
Mouse TMAO (Trimethylamine-N-oxide) ELISA Kit	biorbyt	Cat# orb1736553
8-Hydroxy-desoxyguanosine (8-OHdG) ELISA Kit	Shanghai Yuanmu Biotechnology	Cat# YM-S0328; Cat# YM-S2952
Fecal Occult Blood test Kit	BASO	Cat# BA-2020B
HiPure Stool DNA Kit	Magen Biotechnology	Cat# D3141
AxyPrep DNA Gel Extraction Kit	Axygen Biosciences	Cat# AP-GX-4
Deposited data		
Public 16S sequencing data	Nielsen et al.[60]	SRA: PRJEB1220, ERR209670, PRJNA232056, PRJNA240346, PRJNA368966, PRJNA388210, PRJNA389280, PRJNA400072, PRJNA450340, PRJNA237362; HMP2 dataset (https://ibdmdb.org/results)
Public Bluk RNA-sequencing data	Argmann et al.[34]	GEO: GSE193677; HMP2 (https://ibdmdb.org/results)
Public Single-cell RNA-sequencing data	Nie et al.[35]	Meta-Analysis
Bulk RNA-sequencing data	This study	SRA: PRJNA1112090
Experimental models: Organisms/strains		
Mouse: BALB/c	Vital River Laboratories	N/A
Mouse: C57BL/6	Vital River Laboratories	N/A
Mouse: *Ffar2*-KO C57BL/6	GemPharmatech	N/A
Mouse: IFN-γ-IRES-Venus-AkaLuc C57BL/6	Shanghai Model Organisms Center	N/A
Oligonucleotides		
Tlr4 F	CTCACAACTTCAGTGGCTGGATTTA	N/A
Tlr4 R	CTCACAACTTCAGTGGCTGGATTTA	N/A
Duox2 F	TCCAGAAGGCGCTGAACAG	N/A
Duox2 R	GCGACCAAAGTGGGTGATG	N/A
TorA F	GCAAGTRGGTTGGAACTGTG	N/A
TorA R	CTCTTGACGYTGGATACACC	N/A
βActin F	TGACAGGATGCAGAAGGAGA	N/A
βActin R	CGCTCAGGAGGAGCAATG	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
GraphPad Prism 9.0.0	GraphPad Software Inc.	https://www.graphpad.com/
Image J	ImageJ Software Inc.	https://imagej.nih.gov/ij/index.html
R 4.1.2	R Development Core Team	https://www.r-project.org/
Python 3.7.1	Python Software Foundation	https://www.python.org/
Flowjo	FlowJo, LLC	https://www.flowjo.com/
Silva database	Max Planck Institute	https://www.arb-silva.de/
fastp	Chen et al.[61]	https://github.com/OpenGene/fastp
QIIME2	QIIME2 development team	https://qiime2.org/
Other		
Code	This study	https://github.com/WeiFlora/IntestinalFlora

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Wei Wei (weiwei@ipe.ac.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyses existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. Accessions to raw data generated in this article are listed in the key resources table.
- Our analysis code has been uploaded to the GitHub (https://github.com/WeiFlora/IntestinalFlora).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mouse models

After purchasing from the corresponding companies, 6-8 weeks old female BALB/c, C57BL/6, Ffar2-K, and IFN-γ-IRES-Venus-Aka-Luc mice were kept in specific pathogen-free environment and were housed in individually ventilated isolator cages with free access to food and water (23±2°C, with 55 ± 5% humidity and under a 12-12 h light-dark cycle). The animal experiments were approved by the Institutional Animal Care and Use Committees at the Institute of Process Engineering, Chinese Academy of Sciences (approval ID: IPEAECA2020081). This study was performed in strict accordance with the Regulations for the Care and Use of Laboratory Animals and Guideline for Ethical Review of Animal (China, GB/T 35892-2018).

NHP model

A total of 9 cynomolgus monkeys (4 to 4.5 years old, 4.5 ± 0.5 kg, male) were purchased from the Animal Center of the Academy of Military Medical Sciences and were individually housed in cages with free access to food and water (23±2°C, with 55 ± 5% humidity and under a 12-12 h light-dark cycle). The NHP experiments were approved and conducted in accordance with the institutional guidelines of the Academy of Military Medical Sciences Institutional Animal Care and Use Committee (approval ID: LACUC-DWZX-2021-043). This study was performed in strict accordance with the Regulations for the Care and Use of Laboratory Animals and Guideline for Ethical Review of Animal (China, GB/T 35892-2018).

Human samples

Fecal samples of healthy donors and UC patients were obtained from the Department of Gastroenterology in Tongren Hospital (Shanghai, China) and Peking University First Hospital (Beijing, China). The studies (CDJQ-2022 and 2022-064-01) were approved by the ethical review board at Tongren Hospital and Peking University First Hospital. The age and gender information of healthy donors and UC patients are included in Table S1. In our analysis, sex, gender, or both have no influence on the results of the study. All human subjects are yellow race, East Asian ancestry and the ethnicity is Han Chinese. The socioeconomic status of the study participants was not recorded. After collection, fecal samples were snap frozen in dry ice, and stored at -80 °C until analysis. C-reactive protein (CRP) value was obtained from the hospital medical records. The relative abundance of *Lactobacillus* genus of fecal samples

was measured by 16S rRNA sequencing. 8-hydroxy-2-deoxyguanosine (8-OHdG) levels were measured via ELISA according to the manufacturer's instructions.

METHOD DETAILS

Model training and evaluation

The microbiota data were obtained from the GMrepo metagenomics database (PRJEB1220, ERR209670, PRJNA232056, PRJNA240346, PRJNA368966, PRJNA388210, PRJNA389280, PRJNA400072, and PRJNA450340, Table S2). Downloading the data of healthy donors and ulcerative colitis patients, the samples with less than 20 bacterial genera were excluded. Considering the non-normal distribution of obtained data and our aim to reflect and rank the contribution of each bacterial genus, we herein utilized L1-regularized logistic regression model. For each data splitting, the abundance of test data was scaled to 0-1 according to maximum and minimum value of each bacterium abundance in train set. The missing values were imputed with 0. The logistic regression model was trained using the scikit-learn package (v.1.0.2) in python (v.3.7.1). For hyperparameters tuning, we used 5 cross-validated grid search to find the best hyperparameter combinations. The regularization strength here was 0.1. The area under the receiver operating curve (AUC) value was used as model evaluation metric. For model interpretation, a larger absolute value of feature weight indicated a greater importance in classification of that feature. As the label of healthy donors was 0 and UC patients was 1, so a large negative coefficient was more likely to result in a classification as 1 when the abundance of the relevant feature deceases, which means lower abundance of that genus plays a significant role in leading to the ending of UC. For validation, we additionally obtained the Gevers dataset (PRJNA237362) and the HMP2 dataset (https://ibdmdb.org/results) with their metadata. Samples with fewer than 20 bacterial genera were excluded. For each data splitting, the abundance of test data was scaled to 0-1 according to maximum and minimum value of each bacterium abundance in train set. The logistic regression model was trained as mentioned before. The AUC value was used to estimate the classification performance. For model interpretation, absolute values of feature weights among negative features were evaluated as previous.

Bioinformatics analyses of publicly available datasets

We first downloaded the bulk RNA-seq data from the GEO database with the accession number GSE193677 (Table S3). This dataset contains 461 control participants and 293 samples of inflamed tissues. The differential expression analysis between the inflamed tissue samples and control participants was performed using the "rankSumTestWithCorrelation" function from the "limma" package (v3.56.2) of R. The derived p-values were adjusted by the Bonferroni method. We selected genes displaying higher expression in inflamed tissues (with a $\log_2$-transformed fold change from controls to inflamed tissues greater than 0.2 and the adjusted p-value less than 0.01) for subsequent gene set analyses. Employing the "enrichGO" function from the "clusterProfiler" package (v4.8.3), we carried out over-representation analysis of Gene Ontology (GO) terms based on these genes, with the parameters set as "ont='BP', pAdjustMethod='BH', pvalueCutoff=0.01, qvalueCutoff = 0.05". The obtained GO terms were input to the "simplify" function to remove redundancy, yielding a selection of 250 non-redundant ones. We next evaluated the variation of these gene sets enrichment across UC inflamed tissue samples through the "gsva" function from the "GSVA" package (v1.48.3) with the parameters "min.sz=5, max.sz=500", obtaining a new matrix of gene set activities (gene set × samples). The function "rcorr" from "Hmisc" (v5.1-1) was used to map this matrix to a correlation matrix among different gene sets, which was then input to the "princomp" function, with overall 5 components retained for further analyses. Clustering the loadings of different gene sets on these 5 components obtained 5 gene set modules, with two of them featuring cell-cycle and inflammation-related pathways. The remaining pathways from other modules were classified according to their related types of interventions and then organized into nine meta-programs, including the oxidative stress-related program, the programmed cell death-related program, the cell proliferation and differentiation-related program, the cell migration-related program, the cell adhesion-related program, the wound healing-related program, the bioactive substances-related program, the signaling cascade-related program, and the integrated response-related program.

Next, we retrieved a scRNA-seq dataset of UC patients from a recent integrative study. This dataset contains 117 samples of inflamed tissues with well-annotated major cellular components, including epithelial cells, mesenchymal cells, endothelial cells, neural cells, B_Plasma cells, CD8 T cells, myeloid cells, CD4 T cells, and innate lymphoid cells. We used the "AddModuleScore" function from the "Seurat" package (v4.3.0) to evaluate the activity of the aforementioned nine meta-programs, with default parameters, and for each meta-program, we summarized its activity in each major cellular component as the median value over its activities across all cells from this component.

For validation, we downloaded the bulk RNA-seq data of the HMP2 project (https://ibdmdb.org/results). This dataset contains 74 samples of UC patients and 50 samples of nonIBD controls. We first obtained the log-normalized data with the R package "DESeq2" and then performed the differential expression analysis between the expression profiles of UC patients and nonIBD controls following the above pipeline. The correlation analysis between the fold changes of genes in the dataset1 (GSE193677) and dataset2 (HMP2) revealed highly consistent patterns. We then used the overlapping DEGs to perform the GO enrichment analysis, with the enriched pathways annotated into ten meta-programs including the ossification-related program, the oxidative stress-related program, the programmed cell death-related program, the cell proliferation and differentiation-related program, the cell migration-related program, the cell adhesion-related program, the wound healing-related program, the bioactive substances-related program, the signaling cascade-related program, and the integrated response-related program. Finally, we evaluated the activity of the ten

meta-programs in the scRNA-seq dataset as aforementioned. For each meta-program, we summarized its activity in each major cellular component as the median value over its activities across all cells from this component.

Preparation and characterization of Se-fLac

Lactobacillus casei (CICC20280) were first cultured in 100 mL of MRS medium overnight at 37°C to obtain Lac cells. Then, 100 μL of seed culture medium was resuspended and grown in 4 mL of MSgg medium overnight at 37°C to generate fLac cells. Following a traditional approach for preparing selenium nanoparticles, we incubated fLac cells in a Na_2SeO_3 solution (0.3 mM) in saline at 37°C for 2 h, after which selenium dots formation was induced by ascorbic acid reduction (12 mM) in saline at 37°C for 2 h. The samples for electron microscopy were prepared by directly dropping the aqueous solution on ultra-thin carbon film. The samples were dried completely in air before observation. samples were observed by JEM-1400 Flash, operated at 120 kV. TEM Mapping and EDS analysis were performed using a JEM-2100F microscope with an accelerating voltage of 200 kV. For XPS analysis, Se-fLac was lyophilized and then operated on an ESCALAB 250Xi instrument (150 W, 1253.6 eV) (Thermo Fisher Scientific). We also prepared the Se-fLac cells with half and double loading content of selenium dots. To assess whether the selenium dots-embedded pericellular film on the surface of Lac would affect the growth of Lac, the growth curves of Lac and Se-fLac were measured. The Lac and Se-fLac were diluted and seeded into a 96-well plate with an OD600 value of ~0.15 and incubated at 37°C. The OD600 values were monitored every 1 h for 12 hours in total by a microplate reader. To assess the content and valence of selenium on the Se-fLac cells after experiencing simulated gastric fluid (SGF), Se-fLac cells were resuspended into 1 mL of SGF and further incubated at 37°C with gentle shaking for 4 hours. On the one hand, the solution was lyophilized for XPS analysis. On the other hand, the Se-fLac cells were washed with PBS, then the solution was dropped on ultra-thin carbon film for elemental mapping and digested for ICP-MS measurement. 50 μL samples were digested with 200 μL aqua regia (HNO_3 + HCl) and diluted in distilled water. Digestion was performed at room temperature for 2 hours and measurements were performed by using ICP-MS (iCAP Qc, Thermo Fisher Scientific).

For XRD analysis, Se-fLac was lyophilized and then operated on Empyrean high-resolution powder X-ray diffractometer (PANalytical B.V.). For FT-IR analysis, selenium dots-embedded pericellular film (harvested from ultrasonically treated Se-fLac) were lyophilized and then operated on Nicolet iS50 (Thermo Fisher Scientific). For ^{1}H-NMR analysis, the representative saccharide in pericellular film, *N*-Acetyl-D-Glucosamine (GlcNAc), was selected and mixed with $SeO3^{2-}$, which were operated on Bruker AVANCEIII 600MHz. Superoxide anion, hydroxyl radical and nitrogen free radical scavenging activity was determined from the ESR spectrum (Bruker E500 spectrometer) using DEPMPO, DMPO and DEPMPO as a spin-trapping agent, respectively. Superoxide anions and hydroxyl radicals were produced by the Fenton reaction. Nitrogen-free radicals were produced by SNAP. Lac, fLac, SeNPs, and Se-fLac were added to each mixture. The change in the relative peak intensity in the ESR spectra of DEPMPO-$O2^{\bullet-}$, DMPO- •OH and DEPMPO- •NO adducts indicated a scavenging effect. For quantitative analysis, superoxide anion and hydroxyl radical scavenging activities were measured using a micro superoxide anion kit and micro hydroxyl free radical scavenging capacity kit, respectively, according to the manufacturer's instructions. SeNPs and different concentrations of Se-fLac were added to each kit and fluorescent intensities were measured. Nitrogen free radical scavenging activity was measured using DPPH. DPPH was diluted to 20 μM, and SeNPs and different concentrations of Se-fLac were added. The mixture was incubated at 37°C for 30 min and absorbance was measured at 520 nm. To investigate the mechanism of the much superior ROS scavenging property of Se-fLac, femtosecond-TA spectroscopy was determined as previously described.[62] Briefly, we selected DPPH + SeNPs and DPPH + selenium dots as samples. Solution samples in 1 mm path length cuvettes were excited by the tunable output of the optical parametric amplifier (pump). 320 nm excitation was employed, and the wavelength range was cropped from 350 nm to 600 nm to wipe out interferences from spurious signals. Chirp correction was then carried out to calibrate the initial time.

Resistance assessment *in vitro* and *ex vivo*

SGF was prepared by dissolving NaCl (3.2 g) and pepsin (0.2 g) in 100 mL deionized water, and the pH was adjusted to 2.0 with a dilute HCl solution. SGF was filtered through a 0.22 μm membrane before use. Lac, fLac and Se-fLac cells (1×10^7 CFU) were resuspended into 1 mL of SGF and further incubated at 37°C with gentle shaking. At indicated time points, 200 μL solution was withdrawn from the medium and centrifuged at 5000 rpm for 5 min and resuspended in 200 μL of PBS. Bacterial counts were performed by culturing the samples with suitable dilution on MRS agar plates at 37°C for bacterial plate counting. For *ex vivo* measurement, each mouse was orally administrated with 1×10^7 CFU of Lac, fLac and Se-fLac cells. Their stomachs were excised for culture with indicated time periods at 37°C, then sample collection was conducted on entire stomach specimens for subsequent dilution and bacterial plate counting.

Evaluation of mucoadhesion

In vitro mucoadhesion measurement was performed by QCM. Mucin 2 solution was spun on the surface of an Au QCM crystal at a speed of 3000 rpm for 10 s. Equal amount of Lac, fLac and Se-fLac cells were flowed into the QCM chamber with same speed, and the Δf was measured according to the manufacturer's advice. *Ex vivo* mucoadhesion measurement was performed by everted sac of colon. The everted sacs of colon were prepared and incubated with equal amount of Cy5-NHS labeled Lac, fLac and Se-fLac cells. The amount of Lac, fLac and Se-fLac cells adhered to the everted sac of colon were monitored through fluorescence signals via IVIS Spectrum Imaging System (PerkinElmer) and Confocal laser scanning microscope (CLSM, Nikon). *In vivo* retention and mucoadhesion measurements were performed by fluorescence imaging. Briefly, BALB/c mice (female, 6 to 8 weeks old) were randomly assigned to three groups (n = 3 biologically independent samples in each group), and Cy5-NHS labeled Lac, fLac, or Se-fLac cells

(1×10^7 CFU) solutions were orally administered, respectively. At indicated time points, fluorescence signals in the mice were detected by IVIS Spectrum Imaging System (PerkinElmer). Furthermore, another batch of mice received same amount of Cy5-NHS labeled Lac, fLac, or Se-fLac cells were euthanized at the same time points. The colon tissues were isolated and embedded in tissue freezing medium for frozen sections. Upon further staining mucus with FITC-WGA in the sections of the colon tissues, the amount of Lac, fLac and Se-fLac cells accumulated in the mucus layer were monitored through fluorescence signals via CLSM. To observe the retention of selenium dots *in vivo*, another three more mice received same amount of Se-fLac were euthanized at the same time points, and the gastrointestinal tracts were isolated and digested for the measurement of selenium content via ICP-MS.

In vivo viability after oral administration

To investigate the *in vivo* viability of Lac, fLac and Se-fLac in the intestine, the Lac cells were initially electrotransformed with the pNZ8148-luc plasmid. Then, the fLac and Se-fLac cells were prepared as the former method described. The Lac-luc, fLac-luc and Se-fLac-luc cells (1×10^8 CFU) were orally administered to mice for single and multiple doses (seven days, once a day). The bioluminescence signals of Lac, fLac and Se-fLac cells were monitored at 4 h after the last oral administration. Furthermore, the multiple-administrated mice were euthanized, and the gastrointestinal tracts were isolated for imaging.

DSS-induced colitis model

A dose escalation study was performed firstly and six-week-old female BALB/c mice ($n = 5$ biologically independent samples in each group) fed with normal diet were accumulated for one week. From days 0 to 7, mice were given 3% DSS for model induction, while non-model control animals ("healthy") were administered with plain water. From days 1 to 7, model mice were orally administered with Se-fLac (different doses: 10^6, 10^7, 10^8 and 10^9 CFU) once a day. Daily body weight of each group was recorded. At day 9, the mice were euthanized and their colons were excised. The colon length was measured via Image J. Briefly, the two sides of each colon were outlined in the original image. Then, the scale was set by ruler for 10 cm in each original image. The two sides were measured with freehand lines in software for calculating the average as the individual data point in correspondingly quantitative analysis.

Then we evaluated the therapeutic efficacies of Se-fLac with other counterparts. Six-week-old female BALB/c mice ($n = 5$ biologically independent samples in each group) fed with normal diet were accumulated for one week. From days 0 to 7 mice were given 3% DSS for model induction, while non-model control animals ("healthy") were administered with plain water. From days 1 to 7, mice were orally administered with Lac (1×10^8 CFU), fLac (1×10^8 CFU), SeNPs+fLac (1×10^8 CFU), Se-fLac (1×10^8 CFU), 5-ASA (30 mg/kg), VSL#3 (1×10^8 CFU), 5-ASA+VSL#3, Se incubated Lac (1×10^8 CFU) once a day. At day 9, the mice were euthanized and their colons were excised. After imaging, the colon tissues were used for histological analysis. The histological analysis for evaluating colon damage was performed according to standard procedures for paraffin embedding and H&E staining. Briefly, the colonic tissues were fixed in 4% paraformaldehyde solution, embedded in paraffin, sectioned (4 μm), and stained with H&E. The resulting slides were scanned using a Nikon Intensilight fluorescence microscope. The severity of colonic histological damage was scored in a blinded fashion to prevent observer bias, as previously described.[19] Briefly, colonic damage was assigned scores as follows: 0, normal; 1, hyperproliferation, irregular crypts, and goblet cell loss; 2, mild to moderate crypt loss (10-50%); 3, severe crypt loss (50-90%); 4, complete crypt loss, surface epithelium intact; 5, small- to medium-sized ulcers (<10 crypt widths); 6, large ulcers ($\geq$10 crypt widths). Inflammatory cell infiltration was scored separately for the mucosa (0, normal; 1, mild; 2, modest; 3, severe), submucosa (0, normal; 1, mild to modest; 2, severe) and muscle/serosa (0, normal; 1, moderate to severe). Scores for epithelial damage and inflammatory cell infiltration were summed, resulting in a total scoring range of 0 to 12.

Similarly, six-week-old female IFN-γ reporter mice (IFN-γ-IRES-Venus-AkaLuc, $n = 5$ biologically independent samples in each group) were used to induce colitis model and evaluate the formulations as mentioned in the preceding paragraph. To visualize the production of IFN-γ *in vivo* after different treatments, IFN-γ reporter mice were intraperitoneally injected with 100 μL of 5 mM Aka Lumine n-Hydrochloride (Wako) in Dulbecco's PBS 15 min before acquiring images using the IVIS Spectrum Imaging System (PerkinElmer). Moreover, we evaluated the therapeutic efficacy of Se-fLac with clinical treatments (30 $mg \cdot kg^{-1}$ of 5-ASA, 1×10^8 CFU VSL#3, and 5-ASA+VSL#3) in above mentioned DSS-induced colitis model. The indices were measured as mentioned before.

Therapeutic efficacy of Se-fLac against multiple colitis models

Delayed therapeutic colitis model was induced in the six-week-old female BALB/c mice ($n = 5$ biologically independent samples in each group) and six-week-old female IFN-γ reporter mice (IFN-γ-IRES-Venus-AkaLuc, $n = 5$ biologically independent samples in each group). For model induction, mice were given 3% DSS in drinking water for seven days and orally administered with Se-fLac (1×10^8 CFU) or DEX (1 mg/kg) once a day from days 8 to 11. Daily body weight of each group was recorded. At day 12, IFN-γ reporter mice were intraperitoneally injected with 100 μL of 5 mM Aka Lumine n-Hydrochloride (Wako) in Dulbecco's PBS 15 min before acquiring images using the IVIS Spectrum Imaging System (PerkinElmer). BALB/c mice were euthanized, and the colon length was measured. Colon tissues were excised for histological analysis.

TNBS-induced colitis model was in the female C57BL/6 mice. Briefly, the mice were acclimatized for seven days and were randomly assigned to three groups (Healthy, PBS, Se-fLac, $n = 5$ biologically independent samples in each group). The TNBS-induced mice pre-sensitized by absorption of 1 wt% TNBS solution through the skin at day -7 and kept for seven days. Then, 2 wt% TNBS solution were slowly administered into the colon lumen of mice at day 0. Then, mice were orally administered with Se-fLac (1×10^8 CFU) once a day from days 8 to 11. At day 12, the mice were euthanized, and the colon length was measured. Colon tissues were excised for histological analysis.

For the DSS-induced chronic UC model, BALB/c mice were acclimatized for seven days and were randomly assigned to three groups (Healthy, PBS, Se-fLac, n = 5 biologically independent samples in each group). For model induction, mice were given 3% DSS in drinking water for one week and normal water for one week, and the cycle was repeated three times. During the second, fourth and sixth weeks, mice were orally administered with Se-fLac (1×10^8 CFU) once a day. At day 43, the mice were euthanized, and the colon length was measured. Colon tissues were excised for histological analysis.

For antibiotics pretreated DSS-induced colitis model, BALB/c mice were acclimatized for seven days and were randomly assigned to three groups (Healthy, Se-fLac, Se-fLac [antibiotics pretreatment] n = 5 biologically independent samples in each group). From days -5 to -1, mice in Se-fLac (antibiotics pretreatment) group were pretreated for 5 days with a cocktail of antibiotics (ampicillin, metronidazole, vancomycin and neomycin). From days 0 to 7, mice were given 3% DSS for model induction, while non-model control animals ("healthy") were administered with plain water. From days 1 to 7, model mice were orally administered with Se-fLac (1×10^8 CFU) once a day. At day 9, the mice were euthanized, and the colon length was measured. Colon tissues were excised for histological analysis.

Biosafety assessment *in vivo*

For the short-term biosafety evaluation, BALB/c mice were acclimatized for seven days and were randomly assigned to two groups (Healthy and Se-fLac, n = 5 biologically independent samples in each group). From days 1 to 7, mice were orally administrated with PBS or Se-fLac (1×10^8 CFU/day). At day 8, blood samples were collected for routine blood tests, biochemical analysis (Hitachi Ltd. Hitachi-917, Japan), and selenium content measurement. Major organs, including the heart, liver, spleen, lung, and kidney, were exercised and fixed in 4% paraformaldehyde for histological analysis. For the long-term biosafety evaluation, BALB/c mice were acclimatized for seven days and were randomly assigned to two groups (Healthy and Se-fLac, n = 5 biologically independent samples in each group). Mice were orally administrated with PBS or Se-fLac (1×10^8 CFU/day) for four weeks. After that, blood samples were collected for routine blood tests, biochemical analysis (Hitachi Ltd. Hitachi-917, Japan), and selenium content measurement. Major organs, including the heart, liver, spleen, lung, and kidney, were exercised and fixed in 4% paraformaldehyde for histological analysis.

Measurement of inflammatory environment in the colon

To the same evaluations of therapeutic efficacies of Se-fLac with Lac, fLac, SeNPs+fLac, we performed according to the protocols mentioned before (n = 10 biologically independent samples in each group). At day 9, the blood samples were collected for ELISA-based TMAO measurement. Then the mice were euthanized and their colon tissues were excised. A small portion of collected colon tissues were cut into small pieces and homogenized in RIPA buffer containing 1 mM phenylmethylsulfonyl fluoride. All procedures were conducted on ice. Lysates were then centrifuged at 10,000 × g at 4 °C for 10 min to remove debris and stored at -80 °C until further use. Protein concentration was measured using a bicinchoninic acid assay (Thermo Fisher Scientific). Lysates were thawed on ice and serially diluted (1:100-1:1000) before performing the experiments. Cytokines levels and redox indices were measured via ELISA and commercialized kit according to the manufacturer's instructions. Total RNA was extracted according to the manufacturer's instructions for subsequent RT-qPCR measurement to analyze the expression of different genes and were normalized with that of βActin. A large portion of collected colon tissues were opened longitudinally and Peyer's Patches were removed. Then the tissues were cut into small pieces and incubated in 5 mM EDTA solution twice for 20 min at 37°C with rotation. Lamina propria lymphocytes were isolated by digesting the tissue with Liberase TM and DNase I at 37°C for 30 minutes with rotation. Lymphocytes were further purified using 80:40 Percoll gradient. Before staining, the lymphocytes from lamina propria were stimulated with PMA/Ionomycin for 6 hours in the presence of Brefeldin A. Then single-cell suspensions were subsequently stained with panel 1 (Ghost dye, anti-mouse CD3, anti-mouse CD4, anti-mouse CD25, and anti-mouse Foxp3), panel 2 (Ghost dye, anti-mouse CD3, anti-mouse CD4, and anti-mouse IL-17A), and panel 3 (Ghost dye, anti-mouse CD45, anti-mouse CD11b, and anti-mouse F4/80). After staining, the cells were washed and analyzed using the CytoFLEX LX flow cytometer.

Fecal samples were also collected at day 9 for ELISA-based 8-OHdG measurement, GC-MS-based short chain fatty acids (SCFAs) measurement and the following gut microbiota analysis. Briefly, the fecal samples were suspended in 50 μL of a 15% phosphoric acid solution. Then, 100 μL of a 125 μg/mL isohexanoic acid (internal standard) and 400 μL of ethyl ether were subsequently added, and the mixture was homogenized for 1 min, followed by centrifugation for 10 min at 12000 rpm. The supernatant was transferred into a glass insert prior to analysis. GC/MS analysis was performed by a Trace 1310 gas chromatograph coupled to an ISQ LT mass spectrometer (Thermo Scientific) with a HP-INNOWAX capillary column (Agilent Technologies). The temperatures of the injector, ion source, and transfer line were set at 250 °C, 300 °C, and 250 °C, respectively. The flow rate of helium carrier gas was kept at 1 mL/min. One microliter of the sample was injected in the split mode at a ratio of 1:10. The initial column temperature was 90 °C, ramped to 120 °C at the rate of 10 °C/min, and then finally increased to 250 °C at the rate of 25 °C/min and maintained at this temperature for 2 min. The ionization was carried out in the electron impact (EI) mode at 70 eV. The compounds were quantified in the selected ion monitoring (SIM) mode using the target ion. The contents of SCFAs were determined as previously described.[63]

Two-photon fluorescence microscopy imaging of ROS in the colon

To the same evaluations of therapeutic efficacies of Se-fLac with Lac, fLac, SeNPs+fLac, we performed according to the protocols mentioned before (n = 5 biologically independent samples in each group). At day 9, all the mice were treated with DHE (10 μM) solution with enema for 30 min. Then mice were euthanized, and their colon were excised and completely washed with PBS to remove free DHE. 3D imaging was performed using a two-photon fluorescence microscope (FV1200MPE, Olympus).

RNA-sequencing analysis

To the same evaluations of therapeutic efficacies of Se-fLac with Lac, fLac, SeNPs+fLac, we performed according to the protocols mentioned before (n = 5 biologically independent samples in each group). At day 9, the mice were euthanized and their colon tissues were excised. Total RNA was extracted using Trizol reagent according to the manufacturer's protocol. RNA quality was assessed on an Agilent 2100 Bioanalyzer and checked using RNase free agarose gel electrophoresis. After total RNA was extracted, eukaryotic mRNA was enriched by Oligo(dT) beads. Then the enriched mRNA was fragmented into short fragments using fragmentation buffer and reversely transcribed into cDNA by using NEBNext Ultra RNA Library Prep Kit for Illumina. The purified double-stranded cDNA fragments were end repaired, A base added, and ligated to Illumina sequencing adapters. The ligation reaction was purified with the AMPure XP Beads (1.0X). and polymerase chain reaction (PCR) amplified. The resulting cDNA library was sequenced using Illumina Novaseq6000 by Gene Denovo Biotechnology Co. (Guangzhou, China). To get high quality clean reads, reads were further filtered by fastp (version 0.18.0). The rRNA mapped reads then will be removed. An index of the reference genome was built, and paired-end clean reads were mapped to the reference genome using HISAT2.2.4 and other parameters set as a default. The mapped reads of each sample were assembled by using StringTie v1.3.1 in a reference-based approach. For each transcription region, a FPKM (fragment per kilobase of transcript per million mapped reads) value was calculated to quantify its expression abundance and variations. RNAs differential expression analysis was performed by DESeq2 software between two different groups. The genes/transcripts with the parameter of false discovery rate (FDR) below 0.05 and absolute fold change $\geq$2 was considered differentially expressed genes/transcripts. Gene set enrichment analysis were performed by using software GSEA to identify whether a set of genes in specific KEGG pathways shows significant differences in two groups.

Microbiota analysis

The fecal samples were collected from previous repeated experiment, frozen in liquid nitrogen immediately for 16S rRNA sequencing assay. Microbial DNA was extracted using the HiPure Stool DNA Kits according to manufacturer's protocols. The hypervariable region V3-V4 of the bacterial 16S rRNA gene was amplified with primer pairs 341F (5'-CCTACGGGNGGCWGCAG-3′) and 806R (5'-GGACTACHVGGGTATCTAAT-3′). Amplicons were extracted from 2% agarose gels and purified using the AxyPrep DNA Gel Extraction Kit according to the manufacturer's instructions and quantified using ABI StepOnePlus Real-Time PCR System. Purified amplicons were pooled in equimolar and paired end sequenced on an Illumina platform according to the standard protocols. Raw data containing adapters or low-quality reads would affect the following assembly and analysis. To get high quality clean reads, raw reads were further filtered according to the following rules using FASTP (version 0.18.0): 1) Removing reads containing more than 10% of unknown nucleotides (N); 2) Removing reads containing less than 80% of bases with quality (Q-value) over 20. Then paired end clean reads were merged as raw tags using FLSAH (version 1.2.11) with a minimum overlap of 10 bp and mismatch error rates of 2%. Noisy sequences of raw tags were filtered by QIIME (version 1.9.1) pipeline under specific filtering conditionsto obtain the high-quality clean tags. All chimeric tags were removed using UCHIME algorithm and finally obtained effective tags for further analysis. The effective tags were clustered into operational taxonomic units (OTUs) of $\geq$ 97 % similarity using UPARSE pipeline. The tag sequence with highest abundance was selected as representative sequence within each cluster. The representative OTU sequences were classified into organisms by a naive Bayesian model using RDP classifier (version 2.2) based on SILVA database (version 138.1), with the confidence threshold value of 0.8 to 1. The abundance statistics of each taxonomy was visualized using Krona (version 2.6). Alpha diversity index and Beta diversity analysis were performed in QIIME. For assessing microbiota composition, log-ratio based analysis (ALDEx2) was utilized for calculating median difference in centred log-ratio values between treatment groups and PBS group, and also was utilized for calculating the significance (expected p-value of Wilcoxon rank test) between PBS group and treatment groups.

DSS-induced UC NHP model

Nine cynomolgus monkeys were randomly divided into three groups (n = 3 biologically independent samples in each group) and provided with 50 mL water (Healthy group) or 50 mL water containing 2.5% DSS (PBS and Se-fLac groups) from days 1 to 10. From days 2 to 10, monkeys in the Se-fLac group were administered with 1×10^9 CFU Se-fLac once a day. The fecal traits of the cynomolgus monkeys in each group were observed every day. Discoloration time of the fecal occult blood test was determined by using the Fecal occult blood test kit with Pyramidon method according to the to the manufacturer's instructions. At day 11, the cynomolgus monkeys were anaesthetized for B-mode abdominal ultrasonography, which determined the mucosal layer thickness of monkeys. For calculating the mucosal layer thickness of monkeys, 6 points were selected in the colon from each of nine monkeys. After that, monkeys underwent colonoscopy examination with biopsies. The biopsies were fixed in 4% paraformaldehyde solution, embedded in paraffin, sectioned, and stained with H&E and Ab-PAS. Blood samples were collected for routine blood tests and biochemical analysis (Hitachi Ltd. Hitachi-917, Japan). Fecal samples were collected to perform the 16S rRNA sequencing analysis and ELISA-based 8-OHdG measurement.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was performed by using GraphPad Prism 9.0.0 and Origin 2022. Statistical analysis was performed using two-tailed Student's t test (comparison between two groups) or one-way analysis of variance (ANOVA; comparison among more than

two groups). Log-ratio based analysis (ALDEx2) was utilized for calculating the significance (expected *p*-value of Wilcoxon rank test) between PBS group and treatment groups and in Figure 6. Spearman correlation analysis was performed to test the relationships between *Lactobacillus* abundance, 8-OHdG levels and CRP values in UC patients, the relationships between 8-OHdG level, cytokines levels and inflammatory cell proportions in colitis mice, and the relationships between SCFAs level and Treg proportions. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, ns: non-significance.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：贾坦
方向：生物制造
单位：中国科学院动物研究所

题目：Construction of Linear Heart Tubes via Coaxial Printing

1. Research Background and Scientific Questions

The heart is one of the earliest organs to form and function during human development, and its embryogenesis involves exquisitely precise spatiotemporal regulation. In early embryonic stages, the heart originates from mesodermal cells and undergoes a series of complex morphogenetic events to eventually form a mature four-chambered organ. Among these processes, the formation and looping of the linear heart tube represent foundational events in cardiac morphogenesis and serve as a critical window for investigating the pathogenesis of congenital heart diseases.

This study focuses on two hallmark features of cardiac development: the formation of a three-dimensional tubular structure and the establishment of anteroposterior axial polarity. As the initial morphological structure of the heart, the linear heart tube, with a diameter of only about 500 micrometers, already exhibits clear anteroposterior polarity and differentiation potential. The compartmentalized distribution of the first heart field (FHF) and second heart field (SHF) lays the cellular foundation for subsequent heart tube looping and atrioventricular separation. Underlying this polarity is the crucial regulatory role of concentration gradients of morphogens (e.g., Wnt, BMP) during development.

Based on this developmental biology background, we pose the core scientific question: **How can we reconstruct the microenvironment for polarized heart tube differentiation from embryonic stem cells in vitro?** This question can be broken down into two sub-problems: first, how to construct a 3D tubular structure matching the dimensions of the embryonic heart tube; and second, how to recreate the effect of morphogen concentration gradients in an engineered system. Addressing these issues will not only deepen our understanding of the molecular mechanisms of heart development but also provide a new technological platform for modeling congenital heart diseases and drug screening.

---

2. Research Content and Progress

2.1 Construction of Linear Heart Tubes via Coaxial 3D Bioprinting

2.1.1 Bioink Design and Optimization

Bioink design is central to fabricating linear heart tubes using 3D bioprinting. Through extensive preliminary screening, we identified a composite hydrogel system based on sodium alginate and gelatin. This selection was based on the following considerations: sodium alginate offers excellent ionic crosslinking properties for immediate shape fidelity after printing; gelatin provides RGD motifs essential for cell adhesion, supporting stem cell survival and differentiation; and the combination allows precise tuning of mechanical properties. We characterized the material's temperature sensitivity and crosslinking kinetics. Rheological tests confirmed that the optimized ink exhibits ideal shear-thinning behavior within 15–25°C, ensuring smooth extrusion while maintaining structural stability post-printing. Notably, the introduction of a dual-crosslinking strategy effectively addressed the insufficient structural strength of single-mechanism crosslinking, providing a stable 3D microenvironment for long-term culture.

2.1.2 Printing Process Development

The application of coaxial printing is one of the innovative aspects of this study. Compared to conventional 3D printing, coaxial printing better mimics the bilayered structure of the embryonic heart tube: an outer cell-rich layer simulating the distribution of cardiac progenitor cells, and an inner acellular support layer forming the lumen. By systematically optimizing printing parameters, we successfully fabricated fine tubular structures with a wall thickness of 100 μm and an inner diameter of 400 μm, closely matching the dimensions of the early embryonic heart tube. Post-printing processing was also carefully designed. We developed a custom cutting mold to ensure uniform sample length—a seemingly simple yet critical detail that significantly enhanced experimental reproducibility and data comparability, highlighting the value of engineering principles in biological research.

2.1.3 Regulation of Cellular Behavior

Beyond structural precision, we closely monitored the biological behavior of cells post-printing. Experimental data showed that the optimized printing conditions maintained over 90% cell viability and stable expression of stemness markers, indicating that the process had minimal impact on stem cell properties. For differentiation induction, we adopted a stage-specific regulation strategy. By sequentially adding small molecules and growth factors, we successfully guided the printed structures toward cardiac lineage. Particularly noteworthy was the role of retinoic acid (RA) concentration gradients in inducing heart tube polarity, providing key insights for subsequent morphogen gradient experiments.

2.2 Morphogen Gradient Chip Design

2.2.1 Design Principle and Implementation

The microfluidic chip was designed based on diffusion principles, where precise control of channel geometry enables stable morphogen gradient formation. Compared to conventional microfluidics, our design has two distinctive features: first, it employs static diffusion rather than continuous flow, better mimicking the embryonic microenvironment; second, it uses height difference-driven flow to avoid mechanical disturbance from pump systems. The optimization process balanced practical needs with scientific rigor. By systematically comparing diffusion efficiency across channels of varying lengths and cross-sectional areas, we identified a 5 mm channel length as optimal—ensuring sufficient gradient resolution without unduly prolonging experimental duration.

2.2.2 Gradient Characterization

Validating gradient efficacy was critical for evaluating chip performance. Using dextran standards of different molecular weights, we confirmed that the platform generates stable and reproducible concentration gradients. Notably, the chip performed optimally for small molecules (around 20 kDa), which aligns well with the molecular weight range of common morphogens, laying a technical foundation for subsequent biological experiments.

2.3 Preliminary Integrated Experimental Results

Integrating 3D-printed heart tubes with gradient chips posed the greatest challenge

in this study. Initial results indicated that under RA gradients, the printed heart tubes exhibited region-specific marker expression, suggesting the establishment of axial polarity. Although the data are preliminary, they demonstrate the potential of this integrated platform. Of particular interest was the directional change in calcium signaling—a functional indicator that strongly supports the emergence of polarity. Compared to molecular marker detection alone, electrophysiological assessment more effectively reflects the maturity of cardiac tissue and the establishment of functional polarization.

---

3. Discussion and Outlook

The most significant contribution of this study lies in the innovative application of engineering methodologies to address developmental biology questions. 3D printing technology overcomes the bottleneck of constructing 3D tubular structures in conventional culture systems, while microfluidic chips provide an effective means to replicate embryonic morphogen gradients. The integration of these two approaches creates a novel platform for studying heart development.

From a technical perspective, this research exemplifies interdisciplinary collaboration: breakthroughs in materials science addressed bioink performance requirements, precision mechanical engineering enabled micron-scale printing, and fluid dynamics principles guided gradient chip design. The synergistic effects of such cross-disciplinary integration are difficult to achieve through traditional biological methods alone.

📝1. 研究背景与科学问题 心脏作为人体最早形成并发挥功能的器官之一，其发育过程蕴含着极其精妙的时空调控机制。在胚胎发育早期，心脏起源于中胚层细胞，经过一系列复杂的形态发生事件，最终形成具有完整四腔室结构的成熟心脏。这一过程中，线性心管的形成与扭曲是心脏形态发生的奠基性事件，也是先天性心脏病发病机制研究的关键窗口。 本研究聚焦于心脏发育过程中两个最具代表性的特征：三维管状结构的形成和前后轴向极性的建立。线性心管作为心脏发育的初始形态，其直径仅约 500 微米，却已具备明确的前后极性分化潜能。第一心场（FHF）和第二心场（SHF）的区隔化分布，为后续心管扭曲和房室分隔奠定了细胞基础。这种极性分化的背后，是多种形态发生素（如 Wnt、BMP 等）形成的浓度梯度在发挥关键调控作用。 基于这一发育生物学背景，我们提出核心科学问题：如何在体外重建胚胎干细胞向极性化心管分化的微环境？这一问题可分解为两个关键子问题：其一，如何构建与胚胎心管尺寸匹配的三维管状结构；其二，如何在工程化体系中重现形态发生素的浓度梯度效应。解决这些问题不仅有助于深入理解心脏发育的分子机制，也将为先天性心脏病模型构建和药物筛选提供新的技术平台。 2. 研究内容与进展 2.1 同轴 3D 打印线性心管构建 2.1.1 生物墨水设计与优化 生物墨水的设计是 3D 打印构建线性心管的核心环节。我们通过大量预实验筛选，最终确定了以海藻酸钠和明胶为基础的复合水凝胶体系。这一选择基于以下考量：首先，海藻酸钠具有优异的离子交联特性，可确保打印结构的即时成型；其次，明胶提供细胞粘附所需的 RGD 序列，有利于干细胞存活和分化；再者，二者的复合使用可以实现力学性能的精准调控。在材料表征方面，我们重点考察了温度敏感性和交联动力学。流变学测试表明，优化后的墨水体系在 15-25℃范围内表现出理想的剪切稀化行为，既保证了挤出过程的流畅性，又能维持成型后的结构稳定性。特别值得注意的是，双重交联策略的引入有效解决了单一交联机制的结构强度不足问题，为长期培养提供了稳定的三维微环境。 2.1.2 打印工艺开发 同轴打印技术的应用是本研究的创新点之一。与传统 3D 打印相比，同轴打印能更好地模拟胚胎心管的双层结构特征：外层为细胞富集区，模拟心肌前体细胞的分布模式；内层为无细胞支撑层，形成管腔结构。通过系统优化打印参数，我们成功实现了壁厚 100 微米、内径 400 微米的精细结构打印，这一尺寸与胚胎早期心管高度吻合。打印后处理工艺同样值得关注。 我们开发了专用的管道切割模具，确保实验样本的长度均一性。这一细节设计看似简单，却极大提高了后续实验的可重复性和数据可比性，体现了工程学思维在生物学研究中的价值。 2.1.3 细胞行为调控 在确保结构精度的同时，我们更关注打印后细胞的生物学行为。实验数据表明，优化后的打印条件能维持 90%以上的细胞存活率。且干性标志物表达稳定。说明该技术路线对干细胞特性的影响在可控范围内。 在分化诱导方面，我们采用了阶段特异性调控策略。通过时序添加小分子化合物和生长因子，成功引导打印结构向心肌方向分化。特别值得关注的是视黄酸

（RA）浓度梯度对心管极性的诱导作用，这一发现为后续形态发生素梯度实验提供了重要参考。 2.2 形态发生素梯度芯片设计 2.2.1 设计原理与实现 微流控芯片的设计基于扩散的物理原理，通过精确控制通道几何参数，实现形态发生素梯度的稳定建立。与常规微流控技术相比，我们的设计具有两个显著特点：一是采用静态扩散模式而非持续流动，更贴近胚胎发育的微环境特征；二是引入高度差驱动，避免泵系统带来的机械扰动。 芯片的优化过程充分考虑了实际应用需求。通过系统比较不同通道长度和截面积的扩散效果，我们确定了5mm 通道长度的最优方案。这一长度既能保证足够的梯度分辨率，又不至于过度延长实验周期，体现了实用性与科学性的平衡。 2.2.2 梯度表征 梯度效果的验证是芯片功能评价的关键环节。我们选用系列分子量标准的葡聚糖进行模拟测试，证实该平台可形成稳定、可重复的浓度梯度。特别值得注意的是，芯片对小分子量物质（20kDa 左右）的梯度形成最为理想，这与常见形态发生素的分子量范围高度匹配，为后续生物学实验奠定了技术基础。 2.3 整合实验初步成果 将 3D 打印心管与梯度芯片整合是本研究最具挑战性的环节。初步实验结果显示，在 RA 梯度作用下，打印心管确实表现出区域特异性标志物的差异表达，提示轴向极性的建立。虽然目前数据尚属初步，但已展现出这一整合平台的巨大潜力。 特别值得关注的是钙信号传导的方向性变化，这一功能学指标为极性分化提供了有力佐证。相较于单纯的分子标志物检测，电生理功能的评估更能反映心脏组织的成熟度和功能性极化的建立。 3. 讨论与展望 本研究最显著的价值在于将工程学方法创新性地应用于发育生物学问题。3D 打印技术解决了传统培养方法难以构建三维管状结构的瓶颈，而微流控芯片则为重现胚胎微环境中的形态发生素梯度提供了有效手段。二者的有机结合，创造了一个全新的心脏发育研究平台。 从技术路线来看，本研究体现了多学科交叉的特色。材料科学的突破解决了生物墨水的性能需求，机械工程的精密控制实现了微米级结构的打印，流体力学原理指导了梯度芯片的设计，这些学科交叉产生的协同效应是传统生物学方法难以企及的。

我已根据学术英语规范对您的论文初稿进行了全面润色，在保留原意和专业术语的基础上优化了表达逻辑、句式结构和学术措辞。以下是润色后的版本：

---

1. Research Background and Scientific Questions

The heart is among the first functional organs to form during human embryogenesis, a process governed by precise spatiotemporal regulatory mechanisms. During early embryonic development, the heart originates from mesodermal cells and undergoes a series of complex morphological events to form a mature four-chambered structure. Among these, the formation and subsequent looping of the linear heart tube represent foundational milestones in cardiac morphogenesis and provide a critical window into the mechanisms underlying congenital heart diseases.

This study focuses on two hallmark features of cardiac development: the establishment of a three-dimensional tubular architecture and the emergence of

anteroposterior polarity. As the earliest cardiac structure, the linear heart tube—despite its small diameter of approximately 500 micrometers—exhibits clear anteroposterior polarity and differentiation potential. The spatially segregated distribution of the first and second heart fields (FHF and SHF) lays the cellular groundwork for subsequent heart tube looping and chamber separation. Underlying this polarity are concentration gradients of key morphogens, such as Wnt and BMP, which play essential roles in orchestrating these developmental processes.

Against this background, we pose the following central scientific question: **How can we reconstruct, in vitro, the microenvironment supporting the polarization of embryonic stem cell-derived heart tubes?** This question can be divided into two sub-questions: first, how to engineer a three-dimensional tubular structure that mimics the dimensional properties of the embryonic heart tube; and second, how to reconstitute morphogen concentration gradients within a bioengineered system. Addressing these challenges will not only advance our understanding of the molecular basis of cardiac development but also provide a novel technological platform for modeling congenital heart diseases and supporting drug screening.

---

2. Research Content and Progress

2.1 Fabrication of Linear Heart Tubes via Coaxial 3D Bioprinting

2.1.1 Bioink Design and Optimization

Bioink formulation is central to the success of 3D bioprinting for constructing linear heart tube structures. Through extensive preliminary testing, we developed a composite hydrogel system based on sodium alginate and gelatin. This combination was selected for several reasons: sodium alginate exhibits rapid ion-mediated crosslinking, enabling immediate structural fidelity after printing; gelatin contains RGD motifs that support cell adhesion and promote stem cell survival and differentiation; and together, they allow fine-tuning of mechanical properties. We systematically characterized the material's temperature sensitivity and crosslinking kinetics. Rheological analyses confirmed that the optimized bioink exhibits ideal shear-thinning behavior within the 15–25°C range, facilitating smooth extrusion while maintaining structural stability post-printing. Notably, the implementation of a dual-crosslinking strategy effectively mitigated the limited structural integrity associated with single-mechanism crosslinking, thereby providing a stable 3D microenvironment conducive to long-term culture.

2.1.2 Printing Process Development

The adoption of coaxial bioprinting constitutes a key innovation of this study. Unlike conventional 3D printing, coaxial printing enables the replication of the bilayered organization of the embryonic heart tube: a cell-rich outer layer mimicking the distribution of cardiac progenitor cells, and an acellular inner layer forming a patent lumen. Through systematic optimization of printing parameters, we successfully fabricated tubular structures with a wall thickness of 100 μm and an inner diameter of 400 μm, dimensions that closely match those of the early

embryonic heart tube. A custom-designed cutting mold was also developed to ensure uniform sample length—a simple yet critical refinement that significantly improved experimental reproducibility and data comparability, underscoring the value of engineering precision in biological research.

2.1.3 Regulation of Cellular Behavior

In addition to achieving structural precision, we closely monitored the biological responses of cells post-printing. Under the optimized printing conditions, cell viability exceeded 90%, and stem cell markers were stably expressed, indicating that the printing process did not compromise stem cell integrity. To direct cardiac differentiation, we implemented a stage-specific induction protocol involving the timed administration of small molecules and growth factors. Notably, retinoic acid (RA) concentration gradients were found to induce heart tube polarization—a finding that informed subsequent experiments on morphogen-mediated patterning.

2.2 Design of a Morphogen Gradient-Generating Chip

2.2.1 Design Principle and Implementation

The microfluidic chip was designed based on diffusion principles, with channel geometry precisely controlled to establish stable morphogen gradients. Compared to conventional microfluidic systems, our design offers two distinct advantages: it employs a static diffusion mode rather than continuous flow, better mimicking the embryonic microenvironment; and it utilizes height difference-driven flow, eliminating mechanical disturbances associated with pump systems. During optimization, we balanced practical applicability with scientific rigor. By comparing diffusion performance across channels of varying lengths and cross-sectional areas, we identified a 5 mm channel length as optimal, providing sufficient gradient resolution without unduly prolonging experimental duration.

2.2.2 Gradient Characterization

Validation of gradient stability and reproducibility was essential for assessing chip performance. Using fluorescent dextran standards of varying molecular weights, we confirmed that the platform generates consistent and reproducible concentration gradients. The chip performed particularly well for molecules around 20 kDa—a range that encompasses many common morphogens—thus providing a reliable technical foundation for subsequent biological applications.

2.3 Preliminary Integrated Experimental Results

Integrating 3D-bioprinted heart tubes with the gradient chip represented the most challenging phase of this study. Initial results demonstrated that under RA gradients, the printed heart tubes exhibited regional specificity in marker expression, suggesting the establishment of axial polarity. Although these data are preliminary, they highlight the potential of this integrated platform. Particularly compelling was the observed directional propagation of calcium signaling—a functional readout that provides strong evidence for functional polarization. This electrophysiological assessment offers a more physiologically relevant measure of cardiac maturation than molecular marker expression alone.

---

3. Discussion and Outlook

The primary significance of this study lies in its innovative integration of engineering approaches to address fundamental questions in developmental biology. 3D bioprinting overcomes the limitations of conventional culture methods in constructing three-dimensional tubular architectures, while microfluidic technology enables precise replication of embryonic morphogen gradients. Together, they establish a novel and versatile platform for studying heart development.

From a methodological perspective, this work exemplifies the power of interdisciplinary collaboration: advances in materials science enabled the design of functional bioinks; precision engineering allowed the fabrication of microscale structures; and fluid dynamics principles guided the development of gradient-generating chips. The synergistic effects achieved through such cross-disciplinary integration are beyond the reach of traditional biological approaches alone.

1. 研究背景与科学问题

心脏作为人体最早形成并发挥功能的器官之一，其发育过程蕴含着极其精妙的时空调控机制。在胚胎发育早期，心脏起源于中胚层细胞，经过一系列复杂的形态发生事件，最终形成具有完整四腔室结构的成熟心脏。这一过程中，线性心管的形成与扭曲是心脏形态发生的奠基性事件，也是先天性心脏病发病机制研究的关键窗口。

本研究聚焦于心脏发育过程中两个最具代表性的特征：三维管状结构的形成和前后轴向极性的建立。线性心管作为心脏发育的初始形态，其直径仅约 500 微米，却已具备明确的前后极性分化潜能。第一心场（FHF）和第二心场（SHF）的区隔化分布，为后续心管扭曲和房室分隔奠定了细胞基础。这种极性分化的背后，是多种形态发生素（如 Wnt、BMP 等）形成的浓度梯度在发挥关键调控作用。

基于这一发育生物学背景，我们提出核心科学问题：如何在体外重建胚胎干细胞向极性化心管分化的微环境？这一问题可分解为两个关键子问题：其一，如何构建与胚胎心管尺寸匹配的三维管状结构；其二，如何在工程化体系中重现形态发生素的浓度梯度效应。解决这些问题不仅有助于深入理解心脏发育的分子机制，也将为先天性心脏病模型构建和药物筛选提供新的技术平台。

2. 研究内容与进展

2.1 同轴 3D 打印线性心管构建

2.1.1 生物墨水设计与优化

生物墨水的设计是 3D 打印构建线性心管的核心环节。我们通过大量预实验筛选，最终确定了以海藻酸钠和明胶为基础的复合水凝胶体系。这一选择基于以下考量：首先，海藻酸钠具有优异的离子交联特性，可确保打印结构的即时成型；其次，明胶提供细胞粘附所需的 RGD 序列，有利于干细胞存活和分化；再者，二者的复合使用可以实现力学性能的精准调控。在材料表征

方面，我们重点考察了温度敏感性和交联动力学。流变学测试表明，优化后的墨水体系在 15-25℃范围内表现出理想的剪切稀化行为，既保证了挤出过程的流畅性，又能维持成型后的结构稳定性。特别值得注意的是，双重交联策略的引入有效解决了单一交联机制的结构强度不足问题，为长期培养提供了稳定的三维微环境。

2.1.2 打印工艺开发

同轴打印技术的应用是本研究的创新点之一。与传统 3D 打印相比，同轴打印能更好地模拟胚胎心管的双层结构特征：外层为细胞富集区，模拟心肌前体细胞的分布模式；内层为无细胞支撑层，形成管腔结构。通过系统优化打印参数，我们成功实现了壁厚 100 微米、内径 400 微米的精细结构打印，这一尺寸与胚胎早期心管高度吻合。打印后处理工艺同样值得关注。

我们开发了专用的管道切割模具，确保实验样本的长度均一性。这一细节设计看似简单，却极大提高了后续实验的可重复性和数据可比性，体现了工程学思维在生物学研究中的价值。

2.1.3 细胞行为调控

在确保结构精度的同时，我们更关注打印后细胞的生物学行为。实验数据表明，优化后的打印条件能维持 90%以上的细胞存活率。且干性标志物表达稳定。说明该技术路线对干细胞特性的影响在可控范围内。

在分化诱导方面，我们采用了阶段特异性调控策略。通过时序添加小分子化合物和生长因子，成功引导打印结构向心肌方向分化。特别值得关注的是视黄酸（RA）浓度梯度对心管极性的诱导作用，这一发现为后续形态发生素梯度实验提供了重要参考。

2.2 形态发生素梯度芯片设计

2.2.1 设计原理与实现

微流控芯片的设计基于扩散的物理原理，通过精确控制通道几何参数，实现形态发生素梯度的稳定建立。与常规微流控技术相比，我们的设计具有两个显著特点：一是采用静态扩散模式而非持续流动，更贴近胚胎发育的微环境特征；二是引入高度差驱动，避免泵系统带来的机械扰动。

芯片的优化过程充分考虑了实际应用需求。通过系统比较不同通道长度和截面积的扩散效果，我们确定了 5mm 通道长度的最优方案。这一长度既能保证足够的梯度分辨率，又不至于过度延长实验周期，体现了实用性与科学性的平衡。

2.2.2 梯度表征

梯度效果的验证是芯片功能评价的关键环节。我们选用系列分子量标准的葡聚糖进行模拟测试，证实该平台可形成稳定、可重复的浓度梯度。特别值得注意的是，芯片对小分子量物质（20kDa 左右）的梯度形成最为理想，这与常见形态发生素的分子量范围高度匹配，为后续生物学实验奠定了技术基础。

2.3 整合实验初步成果

将 3D 打印心管与梯度芯片整合是本研究最具挑战性的环节。初步实验结果显示，在 RA 梯度作用下，打印心管确实表现出区域特异性标志物的差异表达，提示轴向极性的建立。虽然目前数据尚属初步，但已展现出这一整合平台的巨大潜力。

特别值得关注的是钙信号传导的方向性变化，这一功能学指标为极性分化提供了有力佐证。相较于单纯的分子标志物检测，电生理功能的评估更能反映心脏组织的成熟度和功能性极化的建立。

3. 讨论与展望

本研究最显著的价值在于将工程学方法创新性地应用于发育生物学问题。3D 打印技术解决了传统培养方法难以构建三维管状结构的瓶颈，而微流控芯片则为重现胚胎微环境中的形态发生素梯度提供了有效手段。二者的有机结合，创造了一个全新的心脏发育研究平台。

从技术路线来看，本研究体现了多学科交叉的特色。材料科学的突破解决了生物墨水的性能需求，机械工程的精密控制实现了微米级结构的打印，流体力学原理指导了梯度芯片的设计，这些学科交叉产生的协同效应是传统生物学方法难以企及的。

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：李欣然

方向：生物制造

单位：中国科学院深圳先进技术研究院

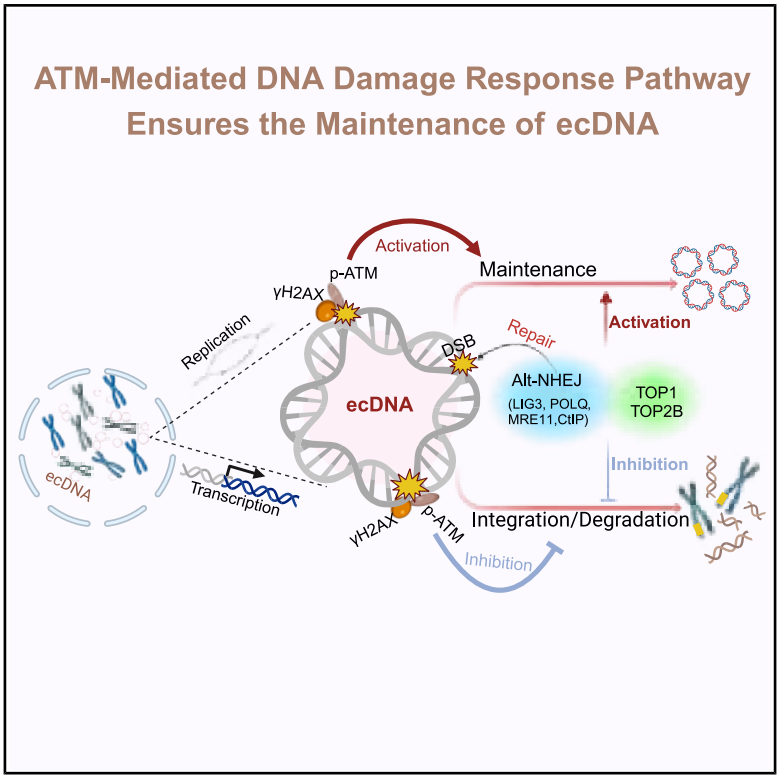
题目：肿瘤中染色体外DNA的复制和维持与DNA损伤间双向调控关系

Article

Cell

Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors

Graphical abstract



Authors

Xing Kang, Xinran Li, Jiaqi Zhou, ..., Zhike Zi, Xiangyu Liu, Haiyun Gan

Correspondence

hy.gan@siat.ac.cn

In brief

This study demonstrates that extrachromosomal DNA (ecDNA) replication induces DNA double-strand breaks and activates the DNA damage response (DDR). The DDR pathways, such as alt-NHEJ, are critical for ecDNA maintenance in tumor cells. Mechanistic insights into ecDNA replication and maintenance unveil a therapeutic approach for treating tumors harboring ecDNA.

Highlights

- ecDNA replication-dependent activation of ATM and DDR ensures ecDNA maintenance
- TOP1 and TOP2B are critical regulators of ecDNA-induced DDR
- Alt-NHEJ pathway promotes ecDNA maintenance
- ATM-mediated DDR is a promising therapeutic target for treating ecDNA+ tumors

Kang et al., 2025, Cell *188*, 1–17
June 26, 2025 © 2025 Elsevier Inc. All rights reserved, including those for text and data mining, AI training, and similar technologies.
https://doi.org/10.1016/j.cell.2025.04.012

CellPress

Please cite this article in press as: Kang et al., Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.012

Cell

CellPress

Article

Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors

Xing Kang,[1,14] Xinran Li,[1,2,14] Jiaqi Zhou,[1,14] Yang Zhang,[1,14] Lingyu Qiu,[1,14] Congcong Tian,[1,14] Zhiwen Deng,[3,14] Xiaoyan Liang,[1] Ziwei Zhang,[1] Songlin Du,[4] Suili Hu,[1,5] Nan Wang,[1] Zhen Yue,[1] Yajing Xu,[1,5] Yuan Gao,[6] Junbiao Dai,[7,8] Zhiquan Wang,[9] Chuanhe Yu,[10] Jinyi Chen,[1] Yuchun Wu,[1,11] Liangming Chen,[1] Yuan Yao,[1] Sitong Yao,[1,5] Xinran Yang,[1] Lixia Yan,[1,2] Qing Wen,[1] Olivia M. Depies,[9] Kuiming Chan,[12] Xiaohuan Liang,[5] Gang Li,[13] Zhike Zi,[1] Xiangyu Liu,[3] and Haiyun Gan[1,15,*]

[1]Guangdong Provincial Key Laboratory of Synthetic Genomics, Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Guangdong Provincial Key Laboratory of Regional Immunity and Diseases, International Cancer Center, Marshall Laboratory of Biomedical Engineering, Shenzhen University Health Science Center, Shenzhen, China
[4]College of Animal Science and Technology, Northwest A&F University, Yangling, Xianyang 712100, Shaanxi, China
[5]College of Veterinary Medicine, South China Agricultural University, Guangzhou, China
[6]Department of Pharmacology, Case Western Reserve University, School of Medicine, Cleveland, OH, USA
[7]CAS Key Laboratory of Quantitative Engineering Biology, Guangdong Provincial Key Laboratory of Synthetic Genomics and Shenzhen Key Laboratory of Synthetic Genomics, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[8]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China
[9]Division of Hematology, Department of Medicine, Mayo Clinic, Rochester, MN, USA
[10]Hormel Institute, University of Minnesota, Austin, MN, USA
[11]School of Basic Medicine, Qingdao University, Qingdao, China
[12]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong, China
[13]Centre of Reproduction, Development and Aging, Institute of Translational Medicine, Faculty of Health Sciences, University of Macau, Macau, China
[14]These authors contributed equally
[15]Lead contact
*Correspondence: hy.gan@siat.ac.cn
https://doi.org/10.1016/j.cell.2025.04.012

SUMMARY

Extrachromosomal DNA (ecDNA) drives the evolution of cancer cells. However, the functional significance of ecDNA and the molecular components involved in its replication and maintenance remain largely unknown. Here, using CRISPR-C technology, we generated ecDNA-carrying (ecDNA+) cell models. By leveraging these models alongside other well-established systems, we demonstrated that ecDNA can replicate and be maintained in ecDNA+ cells. The replication of ecDNA activates the ataxia telangiectasia mutated (ATM)-mediated DNA damage response (DDR) pathway. Topoisomerases, such as TOP1 and TOP2B, play a role in ecDNA replication-induced DNA double-strand breaks (DSBs). A subset of these elevated DSBs persists into the mitotic phase and is primarily repaired by the alternative non-homologous end joining (alt-NHEJ) pathway, which involves POLθ and LIG3. Correspondingly, ecDNA maintenance requires DDR, and inhibiting DDR impairs the circularization of ecDNA. In summary, we demonstrate reciprocal interactions between ecDNA maintenance and DDR, providing new insights into the detection and treatment of ecDNA+ tumors.

INTRODUCTION

Extrachromosomal DNA (ecDNA) is a specific type of acentric, circular DNA exclusively present in tumors. It was identified in neuroblastoma as early as the 1960s and was referred to as "double minutes" for a long time due to its frequent occurrence in pairs.[1,2] Since its initial identification, ecDNA has been detected in most types of human cancer.

Recent studies have revealed the important roles of ecDNA in tumor development, such as promoting the expression of oncogenes, increasing genetic heterogeneity within tumors, facilitating tumor adaptation, and worsening patient prognosis.[1,2]

All rights are reserved, including those for text and data mining, AI training, and similar technologies.

One of the most crucial functions of ecDNA in cancer is the amplification of oncogenes. Unlike chromosomal amplifications, ecDNA can exist in high copy numbers, leading to higher levels of oncogene expression.[1] Key genes, such as *MYC* and *EGFR*, have been identified on ecDNA in various cancers.[1] This amplification confers a growth advantage to cancer cells, as they bypass the normal genomic constraints that regulate oncogene dosage. Because of the lack of centromeres, ecDNA is not subject to the same inheritance rules as chromosomal DNA, leading to uneven distribution during cell division and eventually to increased genetic heterogeneity within tumors.[3] Replication and transcriptional rates are usually higher in cells with ecDNA than in cells without ecDNA,[4] further increasing genetic heterogeneity. The dynamic nature of ecDNA allows tumors to rapidly adjust the expression of genes in response to environmental pressures, such as targeted therapies. For instance, cancer cells may increase the number of ecDNA molecules containing specific drug-resistant oncogenes, leading to the survival of cells despite the presence of inhibitory drugs,[5] allowing for rapid adaptation and evolution. The presence of ecDNA in cancer presents both challenges and opportunities for therapeutic interventions. On the one hand, the highly plastic and adaptable nature of ecDNA-driven tumors makes them difficult to treat with conventional therapies. On the other hand, ecDNA offers a novel biomarker for early diagnosis and could be a target for therapies that either degrade it or inhibit its formation and propagation.

Understanding the molecular mechanisms underlying the basic biological processes related to ecDNA's formation, replication, and maintenance is a growing area of interest. The formation of ecDNA in cancer cells may be closely linked to genomic instability and DNA damage repair mechanisms. However, our current understanding of the mechanisms involved is limited and controversial. For instance, the role of DNA damage repair pathways, such as non-homologous end joining (NHEJ), in ecDNA generation is still under debate. Evidence suggests that NHEJ may be involved, as inhibiting DNA-PKs reduces ecDNA formation.[5] A more recent study has shown that both NHEJ and alternative NHEJ (alt-NHEJ) contribute to the production of extrachromosomal circular DNA.[6] However, knockout of LIG4, an NHEJ-specific DNA ligase, alone or in combination with LIG1, does not significantly impact the generation of extrachromosomal circular DNA.[7]

As ecDNA is a type of circular DNA, its replication introduces additional DNA entanglements and torsional stress. This circular structure makes DNA topoisomerases—essential enzymes that regulate DNA topology during processes such as replication and transcription—attractive targets for therapeutic strategies aimed at tumors containing ecDNA. Topoisomerases resolve DNA entanglements by forming reversible topoisomerase cleavage complexes (TOPCCs).[8] However, these complexes can occasionally become abortive. An abortive topoisomerase 1 covalent cleavage complex (TOP1CC) predominantly generates single-strand breaks (SSBs) but may also result in double-strand breaks (DSBs).[9] By contrast, an abortive TOP2CC typically induces DSBs.[10] The components of the machinery involved in mammalian ecDNA replication are likely different from those involved in linear chromosomal DNA replication. However, current understanding of this process remains limited and incomplete.

ecDNA must be actively maintained during cellular division to sustain its roles in tumor development. Previous studies have revealed that ecDNA remains stable in mitotic cells, indicating autonomous replication.[11,12] Though both the number and structure of ecDNAs are unstable within a cell population,[5] the structure and composition of ecDNAs in most single cells from cultured cell lines do not vary significantly, suggesting stable maintenance of ecDNA within cells.[13] By contrast, episomal DNA other than ecDNA is susceptible to degradation through an unclear mechanism. Despite this knowledge, the molecular mechanism underlying ecDNA maintenance is largely unknown.

Here we report the reciprocal relationship between ecDNA maintenance and the DNA damage response (DDR) in cancer cells. We demonstrate that ecDNA replication and maintenance activate the ataxia telangiectasia mutated (ATM)-mediated DDR pathway. Mechanistically, ecDNA replication leads to the formation and accumulation of abortive TOPCCs, which may trigger DDR. Correspondingly, ecDNA maintenance requires DDR, particularly the alt-NHEJ pathway. Finally, we demonstrate that inhibiting DDR impairs the circularization of ecDNA. Our findings suggest that targeting DDR could be a general approach for treating tumors containing ecDNA.

RESULTS

ecDNA replicates autonomously and is stably maintained

To study the ecDNA replication mechanisms, we generated cell lines containing ecDNA (ecDNA+) using the CRISPR-C method.[14] Two guide RNAs targeting the upstream and downstream regions of *EGFR* were simultaneously delivered into human glioblastoma (GBM) astrocytoma-derived U251 cells and human embryonic kidney-derived 293T cells (Figures 1A and S1A). ecDNA was formed by Cas9 endonuclease cleavage, followed by end-joining-mediated DNA repair. The theoretical end-joining sequence was confirmed by reverse PCR and Sanger sequencing (Figures S1B and S1C). To exclude potential false-positive PCR signals generated from the head-to-tail junction that are derived from tandem duplications, we lysed the cells and removed all linear DNA with endonuclease and then used outward-directed PCR to confirm the existence of circular DNA in these models (Figure S1D). Fluorescence *in situ* hybridization (FISH) performed in metaphase showed that the *EGFR* signal did not colocalize with chromosome 7 (the CEN7 signal) in the generated ecDNA+ cells, while in wild-type cells these two signals colocalized on the same chromosome, confirming displacement of *EGFR* from its original location (Figures 1B and S1E). Additionally, the *EGFR* foci detected outside of chromosome 7 by FISH were significantly higher in ecDNA+ cells (Figures 1C and S1E). The relative *EGFR* copy numbers remained stable for up to 60 days in both cultured U251 and 293T ecDNA+ cells, with a slight increase with longer culture time (Figure S1F). The number of ecDNA per cell in our ecDNA+ cell models cultured in neutral selection conditions[15] was lower than that found in cancer cells in vivo, consistent with previous reports.[16,17] We performed all other experiments after confirming the stability of the relative

Please cite this article in press as: Kang et al., Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.012

CellPress

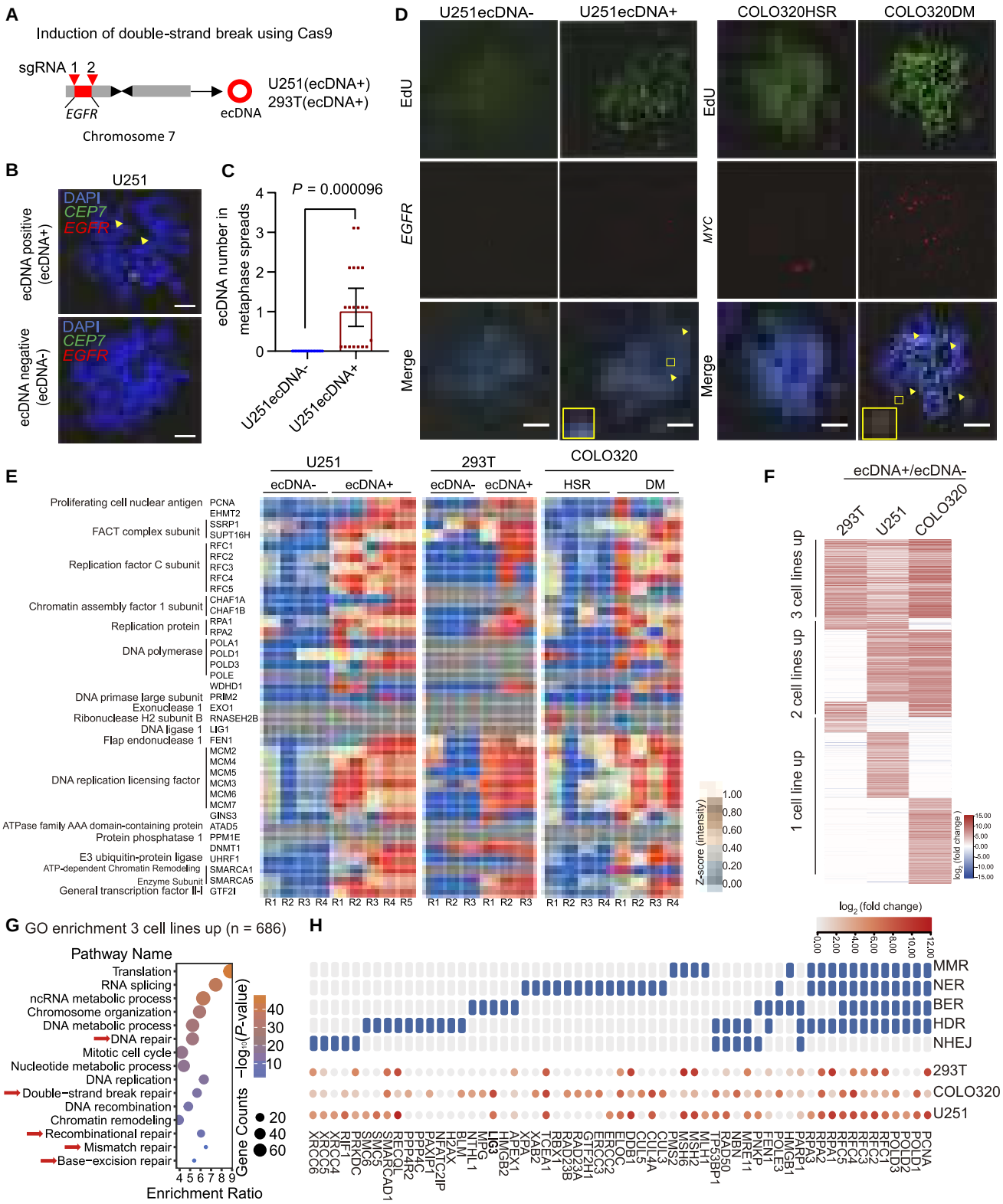


Figure 1. ecDNA replication-related proteins and pathways are enriched in ecDNA+ cells

(A) Schematic illustrating the CRISPR-C strategy used to generate ecDNA+ U251 and 293T cells.

(B) Representative metaphase FISH images for *EGFR* probe in ecDNA+ U251 cells. Scale bar, 10 μm. Yellow triangles indicate *EGFR* foci.

(legend continued on next page)

ecDNA copy number in our models using qPCR (Figures S1G and S1H).

To test whether ecDNA could replicate in these cells, newly synthesized DNA was labeled with 5-ethynyl-2ᵟ-deoxyuridine (EdU). An EdU-click reaction followed by *EGFR*-targeting FISH (EdU-FISH) showed colocalized foci between EdU and FISH signals, indicating replication of ecDNA in U251 cells (Figure 1D). Similar results were observed in a pair of isogenic colorectal cancer cell lines derived from the same patient: COLO320DM (extrachromosomal *MYC* amplification on ecDNA) and COLO320HSR (chromosomal *MYC* amplification on homogeneously staining regions [HSRs]) (Figure 1D).[17] Previously, whole-genome sequencing (WGS) revealed that amplicon copy numbers between the two cell lines are nearly identical.[18] These results demonstrate that we have successfully established two ecDNA+ cell lines (U251ecDNA+ and 293TecDNA+), and ecDNA can replicate and be stably maintained in ecDNA+ cells.

ecDNA maintenance and replication are associated with DDR

To identify key molecules involved in ecDNA replication and maintenance, we employed a method to cross-link proteins associated with newly synthesized DNA labeled by incorporating EdU. In this assay, formaldehyde was used to facilitate the cross-linking, followed by covalent biotin tagging through the EdU-click reaction (Data S1). The identified proteins showed high overlap with previously reported proteins associated with nascent DNA or DNA replication, highlighting the effectiveness of our method (Figure 1E). Notably, several DNA replication-associated proteins, such as MCM2 and POLD, were significantly enriched in ecDNA+ cells, suggesting increased DNA replication in ecDNA+ cells (Figure 1E). Our mass spectrometry results also found a significant number of proteins were specifically enriched in 293TecDNA+, COLO320DM, and U251ecDNA+ cells (Figures 1F and 1G; Data S1). Among these proteins, 686 were enriched in all three ecDNA+ cells, while 1,537 proteins were identified in at least two ecDNA+ cell lines (Table S1; Data S1).

Biological processes associated with DNA replication, and DNA repair were significantly enriched (Figure 1G; Data S1). Specifically, the identified proteins were significantly enriched in DNA repair pathways, including mismatch repair (MMR), base-excision repair (BER), nucleotide excision repair (NER), and DSB repair (Figure 1H). These results implied that DDR is involved in ecDNA replication and maintenance.

ATM-mediated DDR is activated in ecDNA+ cells

The potential involvement of DDR proteins in ecDNA replication led us to evaluate cellular DNA damage levels in ecDNA+ cells. Levels of serine 139-phosphorylated H2AX (γH2AX) were consistently higher in ecDNA+ cells compared with the corresponding control cells using several ecDNA+ cell model pairs (Figures 2A–2C; Data S1). We also observed significantly higher levels of colocalization of γH2AX and *MYC* foci in ecDNA+ cells, indicating active DDR on ecDNA (Figures S2A and S2B). We then used PC3 cells carrying ecDNA with repetitive TetO and endogenously expressing the TetR-EGFP protein[15] [TetO-EGFP PC3 (+)] as an alternative model to study DDR on ecDNA and observed colocalized γH2AX and EGFP foci (Figures 2D and S2C–S2F). Furthermore, greater γH2AX enrichment was found in the *MYC* region in COLO320DM cells (Figure S2G). These data confirm that γH2AX is enriched on ecDNA. The higher levels of γH2AX, a surrogate marker for DSBs, led us to investigate the distribution of DSBs in ecDNA+ cells. We performed END-seq to investigate the distribution of DSBs at nucleotide resolution in an unbiased manner. A high proportion (about 66%) of the peaks we identified in COLO320DM cells were annotated to be localized in circles, and the normalized density around peak centers was higher in COLO320DM cells than in COLO320HSR cells (Figures 2E and 2F). The enrichment ratio in the promoter region was significantly higher than in other DNA elements, including exons and introns (Figure S2H). These results indicate that DSBs may preferentially occur at fragile sites on ecDNA.

The DDR pathway is initiated by DNA damage sensors, transduced by kinases (such as ATM), and then activates downstream effectors. Although total ATM and CHK2 levels were not significantly different between ecDNA+ and ecDNA− cells, both phosphorylated ATM (p-ATM) and phosphorylated CHK2 (p-CHK2) levels were significantly higher in ecDNA+ cells, confirming activation of DDR pathways in ecDNA+ cells (Figure 2C). By contrast, total CHK1 and phosphorylated CHK1 (p-CHK1) were not consistently altered in ecDNA+ cells compared with ecDNA− cells (Data S1). Similar results were observed in an additional pair of isogenic cancer cell lines: HeLa-DM (extrachromosomal *DHFR* amplification on ecDNA) and HeLa-HSR (chromosomal *DHFR* amplification in HSRs),[5] which share a similar copy number in the amplified regions (Data S1). These results reveal

(C) Quantifications of *EGFR* foci numbers in metaphase FISH images of ecDNA+ U251 cells from (B). Mean values ± SEM are presented with *p* value determined using a two-tailed Student's t test.
(D) Representative metaphase FISH images for ecDNA probes showing DNA replication detected by the EdU-click reaction in U251ecDNA−/U251ecDNA+ and COLO320HSR/COLO320DM cells. Scale bar, 10 μm. Yellow triangles indicate colocalized EdU and *EGFR* foci.
(E) Heatmap showing signal intensity differences in replication-associated proteins between ecDNA+ and ecDNA− cell lines.
(F) Heatmap showing proteins enriched in ecDNA+ versus ecDNA− cells among the three cell lines (COLO320, U251, and 293T). Log_2(fold change) of identified protein intensities in ecDNA+/ecDNA− cells are displayed. Proteins identified in all three, two, and only in one cell line are indicated.
(G) Gene Ontology-biological process (GO-BP) analysis using proteins enriched in ecDNA+ cell lines. The *x* axis represents the ratio of enrichment associated with each GO term listed on the left. The 15 statistically significant GO terms are displayed. Red arrows represent the DNA repair pathways.
(H) Illustration showing the distribution of proteins identified by EdU-labeled immunoprecipitation (IP) of ecDNA+ versus ecDNA− cells involved in the MMR, NER, BER, homology-directed repair (HDR), and NHEJ pathways among the three cell lines. The upper heatmap represents which pathway the proteins belong to, and the lower heatmap represents whether the proteins were identified by EdU-labeled IP and the corresponding ecDNA+ versus ecDNA− log_2(fold change) in the three indicated cell lines.
See also Figure S1.

Please cite this article in press as: Kang et al., Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.012

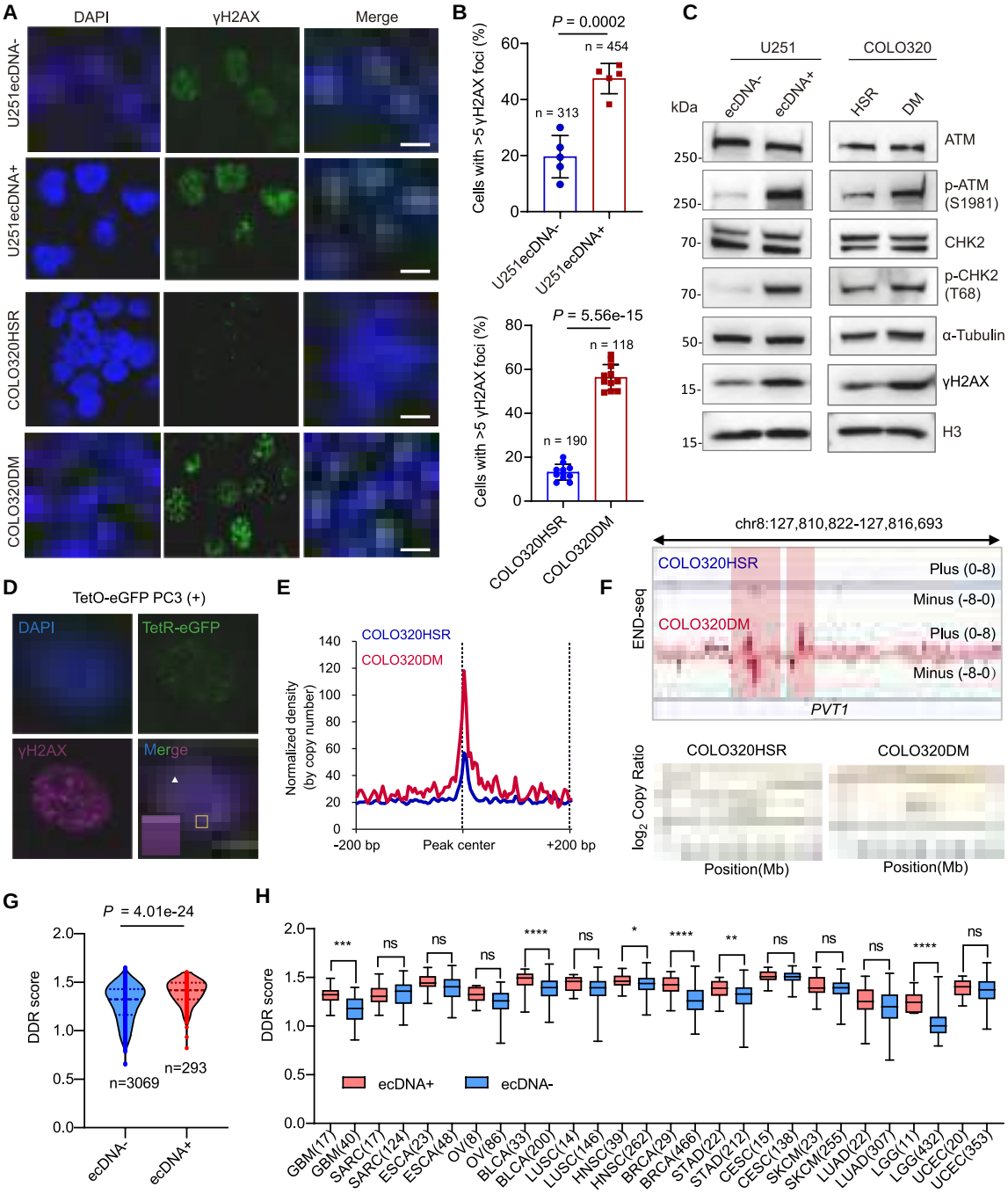


Figure 2. DDR is activated in ecDNA+ cells

(A) Representative immunofluorescence images of γH2AX signals in U251ecDNA−/U251ecDNA+ and COLO320HSR/COLO320DM cell lines. Nucleus is stained with DAPI. Scale bar, 10 μm.

(B) Quantification of percentage of cells with more than five γH2AX foci in U251ecDNA−/U251ecDNA+ and COLO320HSR/COLO320DM cell lines from (A). Mean values ± SD are presented with *p* values determined using a two-tailed Student's t test.

(C) Representative immunoblots of U251 and COLO320 cells were analyzed with the indicated antibodies. ATM, p-ATM, CHK2, and p-CHK2 were normalized to α-tubulin, whereas γH2AX was normalized to histone H3. Molecular weights in kDa (left).

(D) Representative immunofluorescence images showing colocalized ecDNA foci with γH2AX foci in TetO-EGFP PC3 (+) cells. Nucleus is stained with DAPI. Scale bar, 5 μm.

(E) Density of END-seq peaks (±200 bp) at genomic loci with ecDNA amplification (annotated on ecDNA) normalized to corresponding WGS coverage in COLO320HSR (blue) and COLO320DM (red) cells.

(legend continued on next page)

persistent DSBs coupled with ATM-mediated DDR activation in ecDNA+ cells.

We evaluated DDR in clinical samples by analyzing WGS data from the TCGA dataset, which contains 3,362 samples across 14 cancer types. The clinical samples were classified into ecDNA+ and ecDNA− groups, and DDR scores were calculated for each group.[19] ecDNA+ samples exhibited significantly higher DDR scores (Figure 2G). Furthermore, ecDNA+ cancers usually had higher DDR scores, including GBM, lower-grade glioma (LGG), and breast cancer (BRCA) (Figure 2H). These results are highly consistent with what we observed in our ecDNA cell line models.

Topoisomerases are critical regulators of ecDNA-induced DDR

We speculated that the excessive DSBs in ecDNA+ cells primarily originate from internal cellular factors, such as those induced by proteins. From the literature, we have identified several proteins that can directly induce DSBs in DNA, including Spo11, RAG1, RAG2, and type II topoisomerase.[20–25] We checked whether these proteins were identified in our dataset of proteins enriched in ecDNA+ cells (Table S1). Among these proteins, type II topoisomerase TOP2B was consistently identified, and the fold changes in TOP2B levels in ecDNA+ versus ecDNA− cells were significantly different (Table S1). The elevated TOP2B protein levels in ecDNA+ cells were also confirmed by western blot (Figure S3A).

We stably depleted TOP2B and its family member TOP2A in ecDNA+ cells (Figure S3B). Knockdown of TOP2B in ecDNA+ cells significantly reduced the γH2AX signal (Figures 3A–3C and S3C–S3H; Data S2). ecDNA-induced activation of DDR could be partially rescued by TOP2B depletion (Figures 3C and S3G; Data S2). The effects of TOP2A knockdown were not as profound as those of TOP2B knockdown (Figures 3A and 3B). TOP2A is essential for cell division, and its expression pattern is highly cell-cycle dependent.[20] Although we identified TOP2A as a protein enriched in ecDNA+ cells using our cell models, TOP2A levels did not significantly differ between ecDNA+ and ecDNA− cells (Table S1). Meanwhile, knocking down TOP2B in ecDNA− cells led to a significant increase in γH2AX levels, and the trends or extent of changes in other DDR proteins differed from those in ecDNA+ cells (Data S2). These data suggest that the phenotypes of TOP2B relating to γH2AX in ecDNA+ cells may be specific to ecDNA+ cells. Similar results were observed in HeLa-DM and HeLa-HSR (Data S2). We also confirmed interaction between TOP2B and ecDNA using chromatin immunoprecipitation (ChIP)-qPCR (Figure S3H).

The topoisomerase family contains several additional enzymes, including TOP1, TOP3A, and TOP3B.[8] TOP1 was identified in our dataset while the other topoisomerases were not (Table S1). Knockdown of TOP1 in ecDNA+ cells significantly reduced not only the total γH2AX signals but also the γH2AX signals on ecDNA (Figures 3D, 3E, and S3C–S3F; Data S2). ecDNA-induced activation of DDR could be partially rescued by TOP1 depletion (Figure 3F). We demonstrated that the TOP1 phenotype is unique in ecDNA+ cells (Data S2). To further confirm the role of TOP1 in ecDNA-induced DSBs, we performed an *in vivo* proximity ligation assay (PLA) in TetO-EGFP PC3 (+) cells using antibodies targeting TOP1 and TetR. Colocalized signals between PLA and EGFP foci confirmed the presence of TOP1 on ecDNA (Figures 3G and 3H). Treatment with doxycycline, which disrupts the interaction between the TetO sequence and TetR protein, significantly reduced the number of PLA foci per cell (Figures 3G and 3H), confirming the specificity of the PLA signals between TOP1 and TetR.

To assess whether topoisomerase knockdown affects the genome-wide DSB landscape, we performed END-seq using TOP1 or TOP2B knockdown cells. Both TOP1 and TOP2B knockdowns altered γH2AX enrichment pattern (Data S2). The regions enriched in the TOP1 and TOP2B knockdown cells showed different patterns, implying TOP1 and TOP2B preferentially target different sites (Figure 3I).

TOP1 resolves DNA torsional stress by forming TOP1CCs.[20] TOP1CCs are transient structures, but they occasionally become abortive and cannot be readily resolved, generating strand breaks. We observed significantly higher levels of TOP1CCs in ecDNA+ versus ecDNA− cells (Figures 4A–4C; Data S2). Our findings indicate that TOP1CCs could localize on ecDNA (Figures 4D–4F; Data S2). This localization is highly dependent on the presence of TOP1, as knockdown of TOP1 substantially reduces TOP1CC enrichment on ecDNA (Figures 4F–4H). TOP1CC might result in DSBs either during replication or by transcription-associated process.[26] These results reveal that both TOP1 and TOP2B are significant sources of DSBs in ecDNA+ cells, and they are involved in ecDNA replication and the maintenance-mediated DDR pathway.

ecDNA replication and transcription activate DDR

In light of the role of DNA replication and transcription in TOP1-induced DSB formation,[26,27] we investigated the role of ecDNA replication and transcription in ecDNA-induced DSBs and DDR.

To test whether ecDNA replication promoted DSBs and DDR activation, we treated both ecDNA+ and ecDNA− cells with DNA polymerase inhibitor aphidicolin (APH) to suppress DNA replication. After 20 h of APH treatment, neither the level nor distribution of γH2AX signals significantly changed in COLO320HSR cells. By contrast, the percentage of COLO320DM cells with more

(F) Integrative genomics viewer tracks showing a typical region of END-seq data (top) and the corresponding coverage of WGS data (bottom) at ecDNA amplicon region (*PVT1* locus) in COLO320HSR (blue) and COLO320DM (red) cells. Positive and negative strand tracks are displayed. Areas highlighted in red show regions with a higher concentration of reads in COLO320DM versus COLO320HSR cells.

(G) Violin plot showing DDR score calculated by gene set variation analysis (GSVA) of ecDNA+ and ecDNA− samples. *p* value was determined using a two-tailed Mann-Whitney test.

(H) Boxplots showing the distribution of DDR scores calculated by GSVA across 3,362 samples from 14 cancer types presented in the TCGA database. *x* axis indicates cancer type and the corresponding number of samples. *p* values are determined using a two-tailed Mann-Whitney test (* represents p value < 0.05, ** represents p value < 0.01, *** represents p value < 0.001, and **** represents p value < 0.0001; ns, not significant).

See also Figure S2.

Please cite this article in press as: Kang et al., Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.012

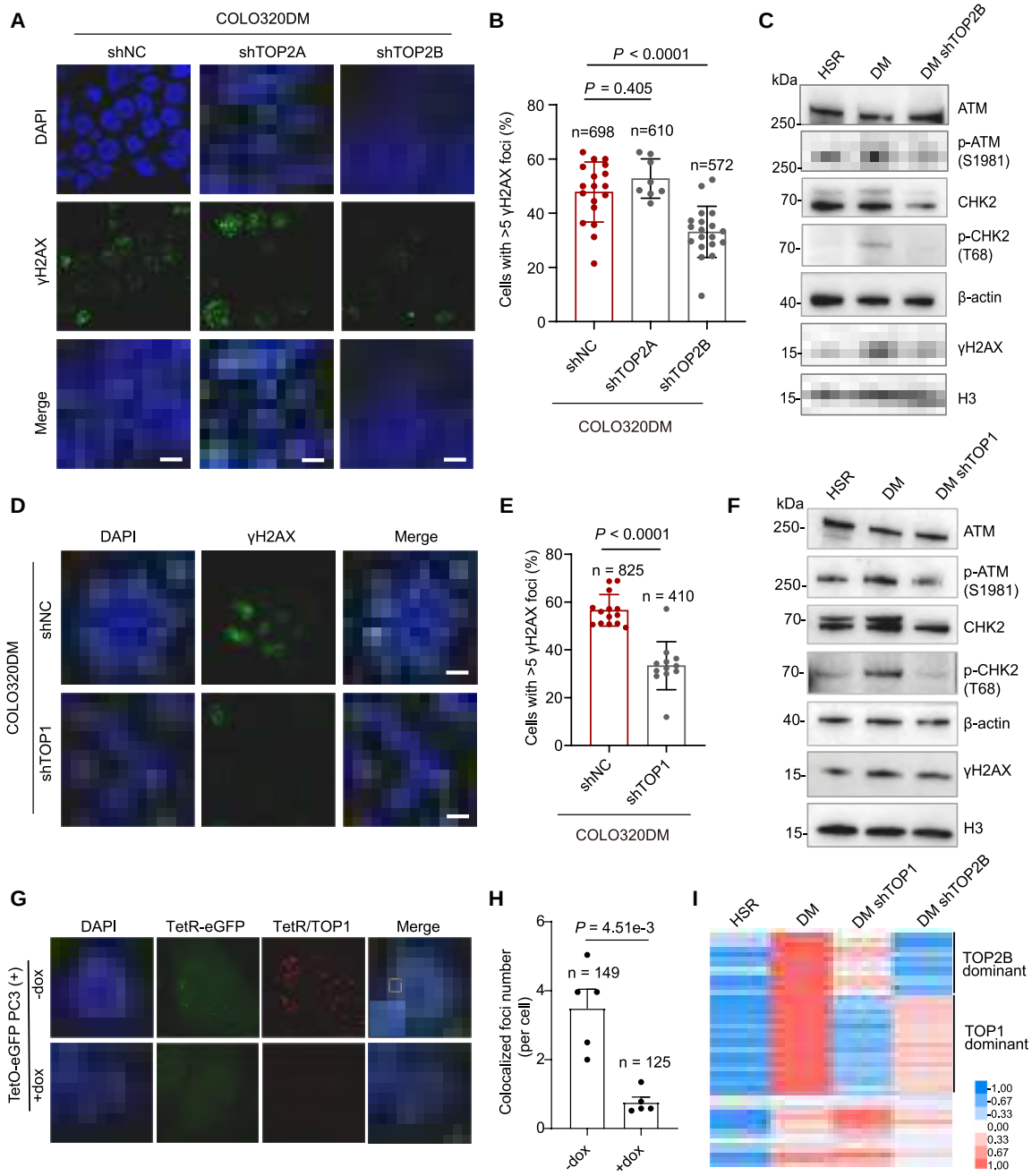


Figure 3. Topoisomerases are critical regulators of ecDNA-induced DDR

(A) Representative immunofluorescence images showing γH2AX signals in COLO320DM cells stably transfected with shRNA targeting TOP2A (shTOP2A), TOP2B (shTOP2B), or scrambled shRNA (shNC). Nucleus is stained with DAPI. Scale bar, 10 μm.

(B) Quantification of percentage of cells containing more than five γH2AX foci in COLO320DM shTOP2A, shTOP2B, and shNC cells from (A). Mean values ± SD are presented with *p* value determined using a one-way ANOVA.

(C) Representative immunoblots of COLO320HSR, COLO320DM, and COLO320DM shTOP2B cells analyzed with indicated antibodies. ATM, p-ATM, CHK2, and p-CHK2 were normalized to β-actin, whereas γH2AX was normalized to histone H3. Molecular weight in kDa (left).

(D) Representative immunofluorescence images showing γH2AX signals in COLO320DM cells stably transfected with shRNA targeting TOP1 (shTOP1) or shNC. Nucleus is stained with DAPI. Scale bar, 10 μm.

(E) Quantification of percentage of cells containing more than five γH2AX foci using COLO320DM shTOP1 and shNC cells from (D). Mean values ± SD are presented with *p* values determined using a two-tailed Student's t test.

(F) Representative immunoblots of COLO320HSR, COLO320DM, and COLO320DM shTOP1 cells were analyzed with indicated antibodies. ATM, p-ATM, CHK2, and p-CHK2 were normalized to β-actin, whereas γH2AX was normalized to histone H3. Molecular weight in kDa (left).

(legend continued on next page)

than five γH2AX foci was significantly lower than in control cells (Figures 5A and 5B). In U251ecDNA− cells, APH treatment significantly increased γH2AX signals, consistent with previous findings that APH increases DNA damage by impairing DNA repair mechanisms.[28] However, in U251ecDNA+ cells, the γH2AX signals were slightly decreased following APH treatment (Figures 5A and 5B), suggesting the DNA damage caused by APH was mitigated by ecDNA replication. To eliminate potential cell-cycle bias, we evaluated γH2AX signals and obtained similar results from APH-treated cells with control cells that displayed a similar cell-cycle pattern (Data S3). Similar results were obtained using several other cell models (Figures S4A–S4F). While APH treatment did not significantly alter the levels of ecDNA in cells, PLA signals between γH2AX and ecDNA were significantly decreased (Figures S4G–S4I). These data suggest that the change in DNA damage levels caused by inhibiting ecDNA replication is more pronounced than that caused by APH treatment.

Additionally, inhibiting DNA replication with APH or by knocking down TOP1 significantly decreases TOP1CC levels in ecDNA+ cells (Figures 4F–4H, 5C, and 5D; Data S3), suggesting that TOP1CC localization levels in ecDNA+ cells are dependent on both DNA replication and TOP1. The conflict between replisomes and TOP1CCs also decreased upon replication inhibition in ecDNA+ cells, as indicated by the reduced PLA signal between PCNA and TOP1CC (Figures 5E and 5F; Data S3).

Transcription activity is significantly higher in ecDNA+ versus ecDNA− cells.[2,18,29] We evaluated the role of transcription in the formation of TOP1CCs by using the transcription inhibitor triptolide (TPL). TOP1CC levels were significantly decreased in TPL-treated ecDNA cells (Data S3), establishing the role of transcription during ecDNA-induced TOP1CC formation. Additionally, inhibiting transcription in ecDNA+ cells significantly decreased DDR, as indicated by γH2AX signals (Data S3). Taken together, these results highlight the essential role of ecDNA replication and transcription in triggering DSBs and DDR activation.

LIG3 is an essential regulator of ecDNA maintenance

Most DSBs must be repaired to maintain genomic stability and cell survival. Excessive DSBs induced by ecDNA raise important questions about how cells regulate their ecDNA levels. Three primary DSB repair pathways involve critical DNA ligases: LIG1, LIG3, and LIG4, which have all been identified in humans. We stably depleted ligases in ecDNA+ cells. Neither LIG1 nor LIG4 knockdown significantly affected ecDNA or HSR levels (Figures 6A and 6B; Data S4). However, transient or stable knockdown of LIG3, which was also identified in our dataset (Table S1), decreased ecDNA levels (Figures 6A and 6B; Data S4). This decrease could be partially rescued by overexpressing LIG3 (Figures S5A–S5C).

Next, we examined whether the ligase activity of LIG3 is crucial for controlling ecDNA levels by overexpressing ligase activity-null LIG3 in LIG3 knockdown ecDNA+ cells or knockdown of LIG3 in LIG3 overexpressed ecDNA+ cells (Figures S5A–S5F). Overexpressing ligase activity-null LIG3 could not restore the ecDNA foci lost via LIG3 knockdown (Figures S5A–S5C). Rescued LIG3 expression could maintain ecDNA levels (Figures S5B, S5C, S5E, and S5F). Knocking down LIG1 and LIG3 simultaneously further decreased ecDNA levels (Figures S5G–S5I). These findings suggest that LIG3 can compensate for LIG1 in maintaining ecDNA, but LIG1 cannot compensate for LIG3 in this process, and also ligase activity of LIG3 is essential for its role in controlling ecDNA levels. This notion was further strengthened by the finding that inhibiting LIG3 increased levels of the temporary linear DNA generated by ecDNA amplification (Figure S5J). Additionally, ecDNA copy number and γH2AX levels significantly decreased with LIG3 knockdown (Figures S5K and S5L; Data S4).

To test whether LIG3 is recruited to ecDNA, we performed the PLA in TetO-EGFP PC3 (+) cells using antibodies targeting LIG3 and TetR (Data S4) and observed clear colocalization between PLA and EGFP foci, confirming the presence of LIG3 on ecDNA (Data S4). In TCGA clinical samples, the mRNA expression level of LIG3 is significantly higher in tumor samples containing ecDNA than in those without ecDNA (Data S4). Within the same tumor types, such as GBM and LGG, the expression of LIG3 is also higher in samples with ecDNA than in those without ecDNA (Data S4). In summary, these findings reveal the importance of LIG3 in ecDNA replication and maintenance.

Alt-NHEJ participates in the maintenance of ecDNA

DSBs can be repaired by several major pathways, including homologous recombination (HR), NHEJ, and alt-NHEJ. The role of alt-NHEJ-specific LIG3 in ecDNA replication and maintenance suggests DDR, particularly alt-NHEJ, might also be critical for these processes.

To test this hypothesis, we depleted key factors associated with different DDR pathways, including LIG4, RAD51, and POLθ using inhibitors (Figures 6C and 6D; Data S4), as well as BRCA1, MRE11, RAD51, and XRCC6, using short hairpin RNAs (shRNAs) (Data S4). We then examined ecDNA levels, as indicated by ecDNA foci numbers and qPCR. In the shRNA knockdown experiment, both BRCA1 and MRE11 knockdown cells exhibited significantly lower ecDNA foci compared with control cells (Data S4). However, only MRE11 knockdown cells showed a significant decrease in ecDNA copy number (Data S4). By contrast, depleting HR and NHEJ components individually, such as RAD51, LIG4, or XRCC6, had no observable impact on ecDNA levels (Figures 6C and 6D; Data S4). We also simultaneously targeted the HR and NHEJ pathways using inhibitors,

(G) *In situ* proximity ligation assay (PLA) detection of interaction between TetR and TOP1 in TetO-EGFP PC3 (+) cells in presence or absence of doxycycline (dox, 10 μM, 96 h). Nucleus is stained with DAPI, and ecDNA foci were visualized with the endogenously expressed TetR-EGFP protein. Scale bar, 5 μm.
(H) Quantification of PLA foci number localized in nuclei of TetO-EGFP PC3 (+) cells from (G). Mean values ± SEM are presented with *p* value determined using a two-tailed Student's t test.
(I) Heatmap showing signal intensity of END-seq peaks at genomic loci with ecDNA amplification in COLO320HSR, COLO320DM, COLO320DM shTOP1, and COLO320DM shTOP2B cells.
See also Figure S3.

Please cite this article in press as: Kang et al., Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.012

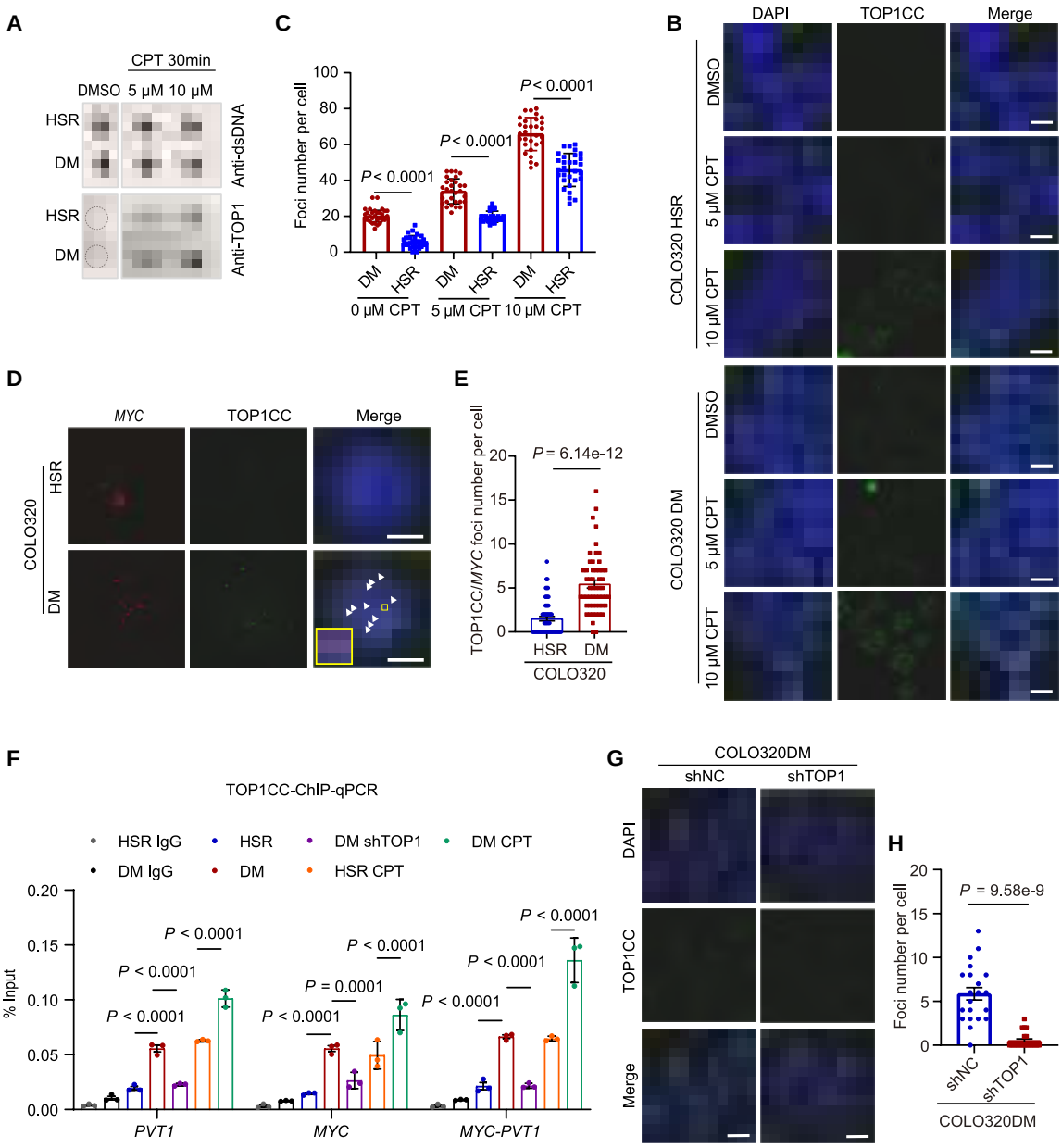


Figure 4. TOP1CCs are induced by and localized on ecDNA

(A) Representative dot blot showing TOP1CC levels in COLO320DM and COLO320HSR cells treated with DMSO or camptothecin (a TOP1 inhibitor, CPT) at indicated concentrations for 30 min. Dashed areas indicate where samples were loaded.

(B) Representative immunofluorescence images showing TOP1CC signals in COLO320DM or COLO320HSR cells treated with DMSO or CPT at indicated concentrations for 30 min. Nucleus is stained with DAPI. Scale bar, 10 μm.

(C) Quantification of TOP1CC foci number per cell in COLO320DM or COLO320HSR cells from (B) treated with DMSO or CPT at indicated concentrations for 30 min. Mean values ± SD are presented with *p* values determined using a two-tailed Student's t test.

(D) Representative images showing immunofluorescence stained with antibody against TOP1CC combined with FISH stained with *MYC* probes in COLO320DM and COLO320HSR cells. Nucleus is stained with DAPI. Scale bar, 5 μm. White triangles indicate merged foci.

(E) Quantification of colocalized TOP1CC and *MYC* foci in COLO320DM and COLO320HSR cells from (D). Mean values ± SEM are presented with *p* value determined using a two-tailed Student's t test.

(F) TOP1CC ChIP-qPCR analysis at representative sites (*PVT1*, *MYC*, and *MYC-PVT1*) that were identified as being localized on ecDNA using END-seq in COLO320DM cells stably transfected with shRNA targeting TOP1 (shTOP1), COLO320DM, or COLO320HSR cells with or without CPT treatment (10 μM, 30 min). Mean values ± SD are shown, with *p* values calculated using a one-way ANOVA.

(legend continued on next page)

and no significant change in ecDNA levels was observed (Figures 6C and 6D; Data S4). This suggests that HR and NHEJ may not be essential for ecDNA maintenance.

Depleting POLθ with an inhibitor significantly reduced ecDNA levels in ecDNA+ cells, suggesting alt-NHEJ participates in ecDNA maintenance (Figures 6C and 6D; Data S4). Knocking down alt-NHEJ components, including CtIP, PLK1, POLθ, FEN1, and XRCC1, significantly decreased ecDNA levels (Figures 6E and 6F; Data S4). These critical factors localized on ecDNA (Figures 6G and 6H; Data S4). Notably, inhibiting upstream alt-NHEJ factors, such as PARP or PLK1, markedly decreased the localization of downstream molecules, such as POLθ and LIG3, on ecDNA (Figures 6G and 6H). Taken together, these results demonstrated that ecDNA maintenance primarily depends on the alt-NHEJ pathway.

Different DNA damage repair pathways exhibit distinct activity patterns at various stages of the cell cycle. For instance, neither HR nor NHEJ is active in mitotic cells, whereas alt-NHEJ has been reported to safeguard genome integrity against DSBs during mitosis.[30,31] We found that mitotic ecDNA+ cells exhibited more γH2AX signals compared with mitotic ecDNA− cells, indicating that DNA damage repair during mitosis may be essential for ecDNA+ cells (Figures S6A–S6D). Inhibiting alt-NHEJ components such as POLθ and PLK1 led to elevated levels of γH2AX in mitotic ecDNA+ cells, as well as increased γH2AX on ecDNA (Figures S6E and S6F). This suggests that alt-NHEJ is critical for DSB repair in mitotic cells (Figures S6A–S6F). Besides, transient knockdown of LIG3 with small interfering RNA (siRNA) led to increased γH2AX signals on ecDNA, as demonstrated by the PLA assay (Figures S6G–S6I).

Recent studies suggested that MDC1, a component of the MDC1-TOPBP1-CIP2A complex that safeguards DSBs during mitosis and a protein identified in our dataset (Table S1), is also important for alt-NHEJ function.[32] Therefore, we investigated whether MDC1 is essential for ecDNA maintenance. Knocking down MDC1 significantly decreased ecDNA levels (Figures S6J–S6M), further substantiating the critical role of alt-NHEJ in ecDNA maintenance.

ATM-mediated DDR is a promising therapeutic target for treating ecDNA+ tumors

Because topoisomerases are sources of DSBs and DSB repair pathways are key factors in the maintenance of ecDNA, we studied the role of topoisomerases in ecDNA maintenance. Notable decreases in ecDNA levels and increases in HSR signals were observed in TOP1- or TOP2B-knockdown or inhibited cells (Figures 7A and 7B; Data S5). Because ecDNA+ cells showed an increase in DDR relative to ecDNA− cells, we investigated whether DDR contributes to the maintenance of ecDNA in cancer cells. ecDNA+ cells were treated with KU55933 (ATMi, an ATM inhibitor) and AZD7762 (CHKi, a checkpoint kinase proteins inhibitor) to suppress the DDR pathway (Figure 7C). Remarkably, most ecDNA foci disappeared and formed HSRs (Figures 7C, 7D, S7A, and S7B; Data S5). Additionally, treatment with ATMi and CHKi significantly reduced the relative ecDNA copy number (Figure 7E; Data S5). By contrast, inhibiting ATR showed no effect on ecDNA maintenance (Data S5).

We further elucidate the relationship between ecDNA replication and the ATM pathway using EdU-immunoprecipitation (IP)-qPCR (Figure S7C). We observed decreased enrichment of EdU at the *MYC-PVT1* fusion breakpoint and *MYC* locus in ATMi-treated cells while incorporating at the genomic *EGFR* locus remained unchanged (Figure S7D). Furthermore, we found ecDNA maintenance requires the proper function of DDR. Inhibiting ATM impaired the circularization of the temporary linear DNA generated by ecDNA replication, as revealed by the elevated levels of ecDNA in linear form (Figure S7E). Collectively, these results suggest that the ATM-mediated DDR pathway is crucial for ecDNA replication and maintenance.

Given that TOP1 and DDR are critical for ecDNA replication and maintenance, we investigated whether ecDNA+ cells were more sensitive to DDR inhibitors than ecDNA− cells. Our drug resistance experiments demonstrated that ecDNA+ cells exhibit increased sensitivity to drugs targeting ATM, CHK2, and TOP1 (Figures 7F and S7F–S7H), suggesting that TOP1 and DDR inhibitors could be employed for treating ecDNA+ tumors (Figure 7G).

DISCUSSION

ecDNA has gained increasing attention in recent years as a potential therapeutic target in various cancers. Its proper function depends heavily on replication and maintenance mechanisms, making these processes promising targets for new therapeutic strategies. However, our current understanding of the underlying mechanisms remains quite limited. This study identified key molecules related to ecDNA replication and maintenance. Our findings show that ecDNA induces ATM-mediated DDR activation in ecDNA+ cells, which depends on ecDNA replication. Mechanistic analysis suggests alt-NHEJ is the major pathway for repairing ecDNA-induced DNA breaks and a critical regulator of ecDNA maintenance. Furthermore, ecDNA maintenance requires properly functioning DDR. Inhibiting DDR impairs the circularization of the temporary linear DNA generated during ecDNA amplification. This reciprocal interaction between ecDNA maintenance and DDR highlights the profound impact of ecDNA on cellular function. These discoveries suggest that targeting the DDR pathway could disadvantage ecDNA+ tumors that rely on ecDNA for growth and survival.

In addition to DDR factors, our work identified topoisomerases, including TOP1 and TOP2B, as critical regulators of ecDNA replication and maintenance. In ecDNA+ cells, replication and transcription activity are significantly higher than in cells without ecDNA,[2,18,29] increasing the likelihood of conflicts between replisomes or transcription complexes and TOPCCs. Elevated collision levels can lead to the formation of abortive TOPCCs and DSBs, resulting in DDR activation in ecDNA+ cells. Consistent

(G) Representative immunofluorescence images showing TOP1CC signals in COLO320DM cells with shTOP1 or scrambled shRNA (shNC). Nucleus is stained with DAPI. Scale bar, 10 μm.

(H) Quantification of TOP1CC foci number in immunofluorescence images using COLO320DM shNC and shTOP1 cells from (G). Mean values ± SEM are shown, with *p* value calculated using a two-tailed Student's t test.

Please cite this article in press as: Kang et al., Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.012

CellPress

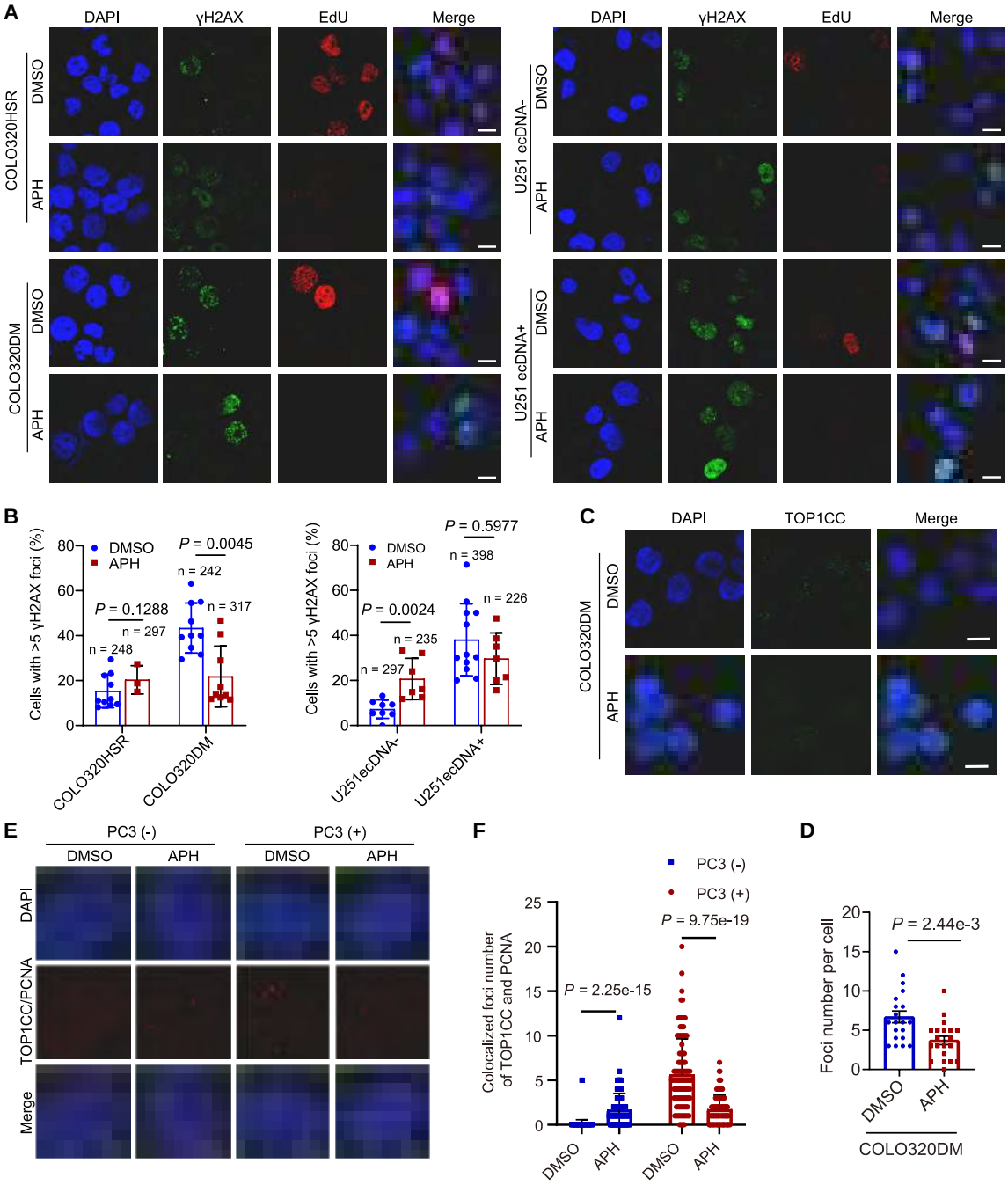


Figure 5. ecDNA replication triggers the DDR pathway

(A) Representative immunofluorescence images of γH2AX signals in COLO320DM/COLO320HSR and U251ecDNA−/U251ecDNA+ cells after treatment with 2 μM aphidicolin (APH) for 20 h combined with 10 μM EdU incorporation. Scale bar, 10 μm.

(B) Quantification of percentage of cells with more than five γH2AX foci in COLO320DM/COLO320HSR and U251ecDNA−/U251ecDNA+ cells after APH treatment from (A). Mean values ± SD are presented with *p* values determined using a two-tailed Student's t test.

(C) Representative immunofluorescence images of TOP1CC signals in COLO320DM cells treated with 2 μM APH (20 h). Scale bar, 10 μm.

(D) Quantification of the TOP1CC foci number in ecDNA+ COLO320DM cells treated with APH from (C). Mean values ± SEM are presented with *p* value determined using a two-tailed Student's t test.

(E) PLA detection of interaction between PCNA and TOP1CC in PC3 (−)/PC3 (+) cells treated with APH. Nucleus is stained with DAPI. Scale bar, 5 μm.

(F) Quantification of PLA signals in nucleus from (E). Mean values ± SEM are presented with *p* values determined using a two-tailed Student's t test.

See also Figure S4.

Please cite this article in press as: Kang et al., Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.012

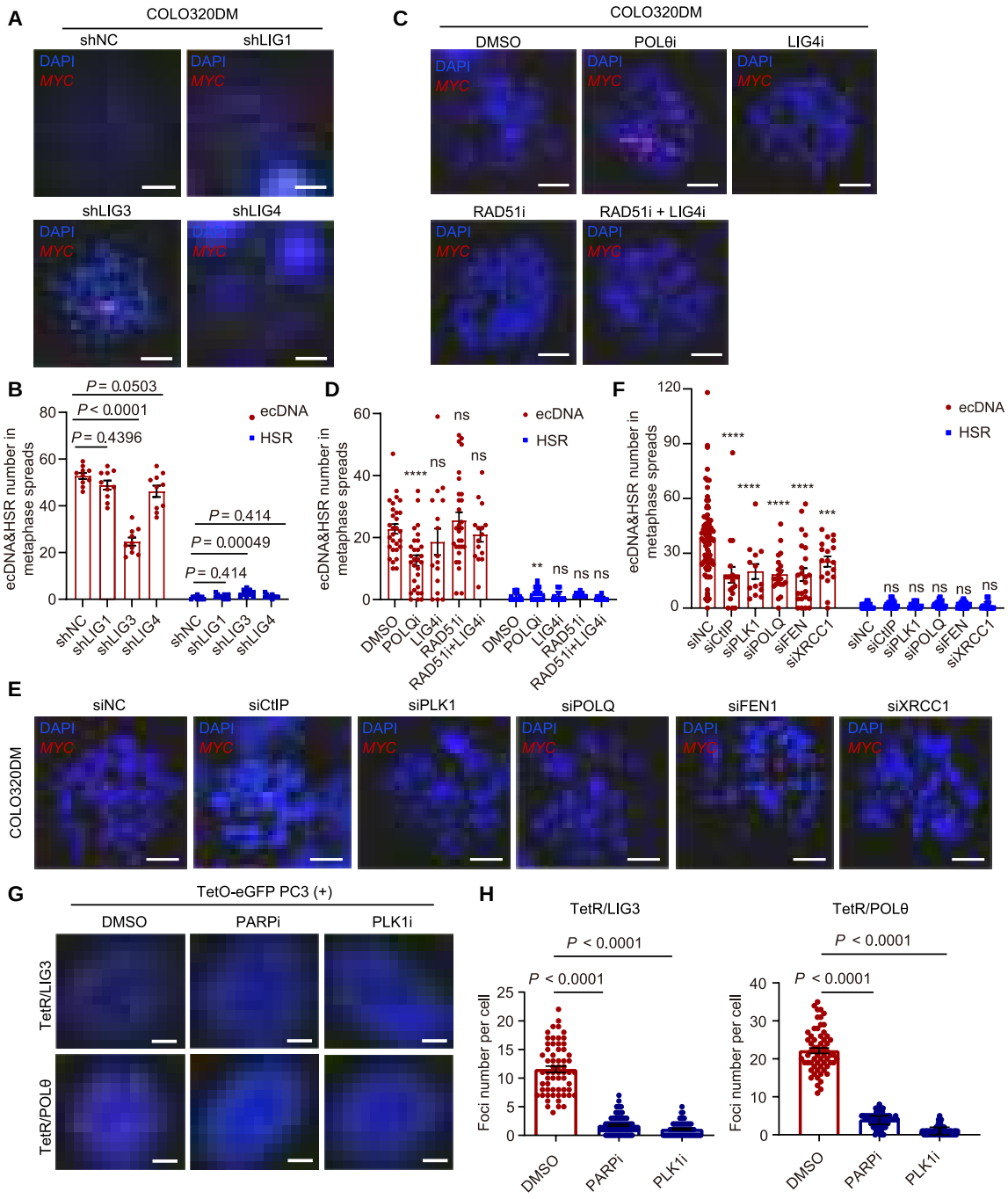


Figure 6. LIG3 and alt-NHEJ are critical for the maintenance of ecDNA

(A) Representative metaphase FISH images for *MYC* probes using COLO320DM cells stably transfected with shRNA targeting *LIG1* (shLIG1), *LIG3* (shLIG3), *LIG4* (shLIG4), or scrambled shRNA (shNC). Scale bar, 10 μm.

(B) Quantification of ecDNA and HSR foci number in metaphase FISH images using COLO320DM shLIG1, COLO320DM shLIG3, COLO320DM shLIG4, or COLO320DM shNC cells from (A). Mean values ± SEM are presented with *p* values determined using a one-way ANOVA.

(C) Representative metaphase FISH images for *MYC* probes using COLO320DM cells treated with 10 μM ART558 (POLθi), 10 μM SCR130 (LIG4i), or 5 μM B02 (RAD51i), or a combination of RAD51i and LIG4i for 24 h. Scale bar, 10 μm.

(D) Quantification of ecDNA and HSR foci number in metaphase FISH images using COLO320DM cells treated with POLθi, LIG4i, and RAD51i from (C). Mean values ± SEM are presented with *p* values determined using a one-way ANOVA (** represents *p* value < 0.01, **** represents *p* value < 0.0001. ns, not significant).

(E) Representative metaphase FISH images for *MYC* probes using COLO320DM cells stably transfected with siRNA targeting *CtIP* (siCtIP), *PLK1* (siPLK1), *POLQ* (siPOLQ), *FEN1* (siFEN1), or *XRCC1* (siXRCC1). Scale bar, 10 μm.

(legend continued on next page)

with this model, inhibiting DNA replication or transcription significantly decreases TOP1CC levels in ecDNA+ cells. Abortive TOPCCs trigger DNA break signaling and DDR.[26,33] Consequently, these breaks are repaired by DNA repair pathways involving DNA ligases such as LIG3. Circular ecDNA introduces additional topological constraints during DNA replication and transcription. When the DNA replication machinery is stalled due to these constraints, topoisomerases are recruited to resolve DNA entanglements and torsional stress that arise. A decrease in the total amount of topoisomerases compromises the resolution of topological constraints, leading to fork collapse and the generation of linear DNA fragments.[34] The linear DNA fragments generated can be integrated into genomic hotspots through DNA repair mechanisms or other processes such as transposition, forming HSRs. In turn, the stalled or collapsed replication fork impairs ecDNA replication, reducing the total amount of ecDNA in cells. Our findings suggest that ecDNA maintenance in cells primarily depends on the alt-NHEJ pathway. How the alt-NHEJ is recruited as the primary repair pathway requires further investigation. In addition, more evidence is needed to confirm the roles of the HR and NHEJ pathways in ecDNA maintenance.

This study sheds new light upon the connection between DDR and the intrinsic mechanisms that have evolved to carry out ecDNA maintenance and replication as well as the molecular mechanisms underlying ecDNA evolution. The extent to which these mechanisms contribute to tumor heterogeneity, tumorigenesis, and drug resistance remains to be determined. Our study strongly suggests that ATM-mediated DDR and the alt-NHEJ pathway are critical for ecDNA maintenance in certain ecDNA+ cancers. Unraveling these pathways could provide new opportunities for advances in diagnosis, prognosis, and the development of a general therapeutic strategy for ecDNA+ tumors.

Limitations of the study

This study has several limitations. First, due to technical limitations, we were unable to directly purify ecDNA-associated proteins. Instead, we employed an alternative global EdU pull-down assay. By comparing data from ecDNA+ and ecDNA− cells, we identified molecules that potentially associate with ecDNA. Additionally, due to technical constraints, we could not measure the kinetics of the replication and degradation of ecDNA. Second, although this study demonstrates that topoisomerases are involved in ecDNA-induced DNA damage, we were unable to fully elucidate the mechanism by which topoisomerases in ecDNA+ cells induced DSBs. Furthermore, we cannot rule out the possibility that ecDNA may induce DSBs through alternative mechanisms. Third, although we demonstrated that the alt-NHEJ pathway is involved in repairing ecDNA-induced DSBs during mitosis, we did not investigate the repair mechanisms of ecDNA-induced DSBs in other phases of the cell cycle. Consequently, we cannot conclusively determine whether HR, NHEJ, or other repair pathways are involved in the repair of ecDNA-induced DSBs. Finally, we showed that alt-NHEJ is essential for maintaining ecDNA. However, we are not sure how alt-NHEJ components are recruited to ecDNA.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Haiyun Gan (hy.gan@siat.ac.cn).

Materials availability

Oligonucleotides, targeting constructs for genome editing, and newly generated cell lines are available upon request and should be directed to the lead contact, Haiyun Gan.

Data and code availability

All MS proteomics data have been deposited with the ProteomeXchange Consortium[35] via the PRIDE Inspector Toolsuite[36] under the accession number PRIDE: PXD060725 and are publicly accessible at https://www.ebi.ac.uk/pride/. Sequencing datasets (WGS and END-seq) have been deposited in the Gene Expression Omnibus (GEO) database under accession ID GEO: GSE272961 and are publicly accessible at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE272961. The DDR core pathways and gene set data are accessible at https://gdc.cancer.gov/about-data/publications/PanCan-DDR-2018.[37] The referenced human proteome is available in UniProtKB/Swiss-Prot and UniProtKB/TrEMBL (version 2021_04). Clinical TCGA data can be accessed via the GDC Data Portal (https://portal.gdc.cancer.gov/). Protein-protein interactions among identified proteins were analyzed using the STRING database (https://string-db.org/cgi/input.pl).

We have made use of publicly available software and tools, as stated in the STAR Methods. No custom algorithms or software are central to the research.

ACKNOWLEDGMENTS

We thank the Paul S. Mischel laboratory for providing TetO-EGFP PC3 (+) cell line, the Don W. Cleveland laboratory for providing HeLa-DM and HeLa-HSR cells, and the Wei Wu laboratory at the Chinese Academy of Sciences for assistance with END-seq. This work was funded by grants from the National Key R&D Program of China (grant no. 2019YFA0903800 to H.G.), the Major Program of the National Natural Science Foundation of China (32090031 to H.G.), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0480000 to H.G.), the General Program of the National Natural Science Foundation of China (32070610 to H.G.), the Guangdong Province Fund for Distinguished Young Scholars (2021B1515020109 to H.G.), the National Natural Science Foundation of China for Young Scholars (32000580 to Q.W., 32100460 to J.Z., 32101178 to Y.Y., 32400470 to C.T., and 32400482 to L.Q.), the Guangdong Basic and Applied Basic Research Foundation (2021A1515110377 to X.K., 2021A1515110483 to N.W., and 2024A1515012730 to C.T.), the Guangdong Provincial Key Laboratory of Synthetic Genomics (2023B1212060054 to H.G.), the Shenzhen Science and Technology Program (JCYJ20230807140602005 to X.Y.), the Shenzhen Medical Research Fund (A2403039 to C.T.), the Shenzhen Key Laboratory of Synthetic Genomics (ZDSYS201802061806209 to H.G.), the China Postdoctoral Science Foundation (2021M703381 to X.K. and 2021M693306 to N.W.), and the SIAT Innovation Program for Excellent Young Researchers (E1G071 to N.W.). The authors thank Kristin Harper, PhD, of Harper

(F) Quantification of ecDNA and HSR foci number in metaphase FISH images using COLO320DM cells depleted with indicated genes from (E). Mean values ± SEM are presented with p values determined using a one-way ANOVA (*** represents p value < 0.001, **** represents p value < 0.0001. ns, not significant).
(G) PLA detection of interaction between TetR and LIG3, as well as TetR and POLθ, in TetO-EGFP PC3 (+) cells treated with 10 μM olaparib (PARPi) or 500 nM volasertib (PLK1i) for 24 h. Nucleus is stained with DAPI, and ecDNA foci were visualized with endogenously expressed TetR-EGFP protein. Scale bar, 5 μm.
(H) Quantification of PLA foci number localized in the nucleus of TetO-EGFP PC3 (+) cells from (G). Mean values ± SEM are presented with p values determined using a one-way ANOVA.
See also Figures S5 and S6.

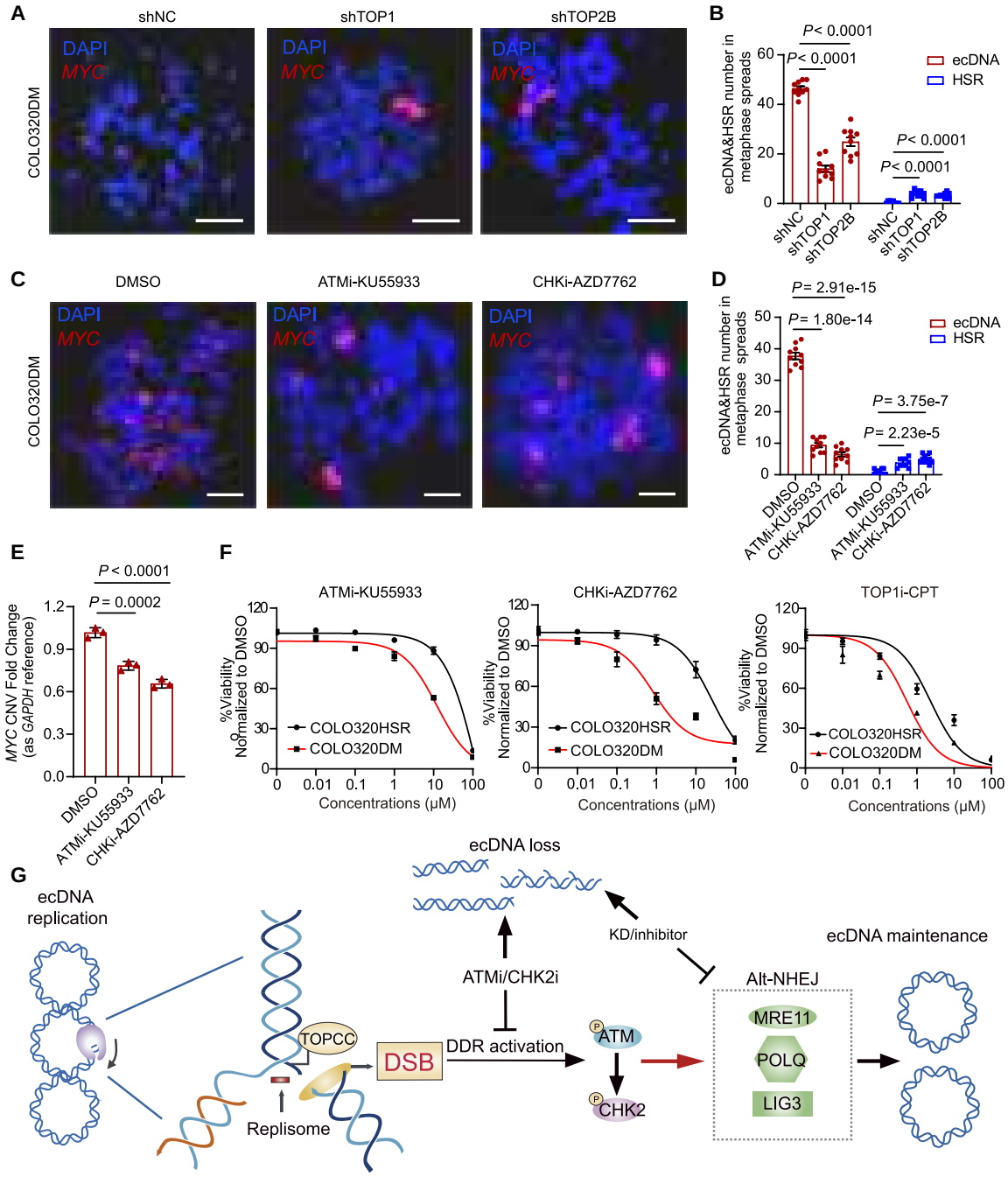


Figure 7. Maintenance of ecDNA associated with topoisomerases and ATM-mediated DDR pathway

(A) Representative metaphase FISH images for *MYC* probes from COLO320DM cells stably transfected with shRNA targeting *TOP1* (shTOP1), *TOP2B* (shTOP2B), or scrambled shRNA (shNC). Scale bar, 10 μm.

(B) Quantification of ecDNA and HSR foci number in FISH images for *MYC* probes from COLO320DM shTOP1, COLO320DM shTOP2B, and COLO320DM shNC cells from (A). Mean values ± SEM are presented with *p* values determined using a one-way ANOVA.

(C) Representative metaphase FISH images for *MYC* probes from COLO320DM cells treated with 5 μM KU55933 (ATMi), 0.2 μM AZD7762 (CHKi), and DMSO (no inhibitor) for 24 h. Scale bar, 10 μm.

(D) Quantification of ecDNA and HSR foci number in FISH images for *MYC* probes from COLO320DM cells treated with ATMi and CHKi from (C). Mean values ± SEM are presented with *p* values determined using a one-way ANOVA.

(E) Quantification of relative *MYC* copy number normalized to *GAPDH* in COLO320DM cells treated with ATMi and CHKi. Mean values ± SD are presented with *p* values determined using a one-way ANOVA.

(legend continued on next page)

Health & Science Communications, LLC, for providing editorial support in accordance with Good Publication Practice guidelines.

AUTHOR CONTRIBUTIONS

H.G., X.K., and Y.Z. conceived and designed the project. X.K., Y.Z., X. Li, J.Z., O.M.D., Y.G., and Z.Y. wrote the manuscript. X.K., Z.D., and X. Li performed and analyzed the metaphase DNA FISH imaging. X.K. and X. Li performed and analyzed the dual FISH-IF imaging. X.K., X. Li, and C.T. tested drug treatment concentration and cytotoxicity. X.K., S.H., and Y.X. designed the sgRNA for Cas9 experiments. X.K., S.H., and Y.X. screened the ecDNA+ single cells. J. Z., N.W., X. Li, and S.H. performed the EdU-labeled biotinylation experiments and data analysis. X.K., S.H., Xiaoyan Liang, Y.Z., and X. Li designed the shRNAs for protein knockdown experiments. X.K., L.Q., S.H., Z. Zhang, and X. Li performed the immunofluorescence experiments. X.K. and L.Q. designed and performed the PLA. Z.D. and X.K. performed the comet assay. X. Li and C. T. performed and analyzed the ChIP-qPCR and CUT&RUN-qPCR experiments. Xiaoyan Liang, J.Z., and Z.D. performed the END-seq assay and data analysis. X. Li, L.Q., and S.D. performed the TOP1CC assay. X.K., J.Z., and C.T. generated the WGS data for COLO320 cells. X.K. and X. Li performed and analyzed the EdU-IP qPCR assay. Xiaoyan Liang and C.T. helped with the FACS experimental design. X.K. and Y.Z. performed the rescue assay and FISH experiment for COLO320 shRNA cells. L.C., Q.W., Y.Y., L.Y., Y.Z., C.T., X.Y., S.Y., K.C., Z.W., C.Y., G.L., Z. Zi, J.D., and X. Liu helped with and provided feedback on the experimental design. J.Z. and Y.W. performed TCGA data analysis. H.G. supervised the study. All authors provided input on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Mammalian cell culture
- METHOD DETAILS
  - Guide RNA design for CRISPR-C
  - ecDNA induction by CRISPR-C
  - Circular ecDNA detection with linear DNA digestion
  - Quantitative PCR (qPCR)
  - Immunofluorescence (IF)
  - Metaphase chromosome spread
  - Metaphase DNA fluorescence *in situ* hybridization (FISH)
  - Dual FISH-immunofluorescence (FISH-IF)
  - Western blotting analysis
  - Detection of EdU incorporation
  - Quantification of nascent RNA production
  - Generation of lentivirus-mediated knockdown cells
  - Overexpression and shRNA-rescued cell lines construction
  - siRNA transfection
  - Drug treatments
  - Synchronization and cell cycle detection
  - Cell cytotoxicity assay
  - TOP1CC detection
  - *In situ* proximity ligation assay (PLA)
  - EdU immunoprecipitation (EdU-IP)-qPCR
  - EdU-labeled cells biotinylation
  - Streptavidin pull-down of biotinylated proteins
  - On-bead digestion and LC-MS/MS
  - LC-MS/MS data acquisition
  - Chromatin immunoprecipitation qPCR (ChIP-qPCR)
  - CUT&RUN-qPCR
  - END-Seq
  - Whole genome sequencing (WGS)
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Label-free MS/MS analysis
  - TCGA clinical data source
  - DNA damage response (DDR) protein analysis
  - Whole genome sequencing (WGS) analysis
  - Amplicon calling and classifying
  - END-seq analysis
  - Colocalization Analysis
  - Statistics and reproducibility

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.04.012.

Received: October 18, 2024
Revised: February 19, 2025
Accepted: April 6, 2025
Published: April 28, 2025

REFERENCES

1. Turner, K.M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., Li, B., Arden, K., Ren, B., Nathanson, D.A., et al. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature *543*, 122–125. https://doi.org/10.1038/nature21356.
2. Kim, H., Nguyen, N.P., Turner, K., Wu, S., Gujar, A.D., Luebeck, J., Liu, J., Deshpande, V., Rajkumar, U., Namburi, S., et al. (2020). Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat. Genet. *52*, 891–897. https://doi.org/10.1038/s41588-020-0678-2.
3. Yan, X., Mischel, P., and Chang, H. (2024). Extrachromosomal DNA in cancer. Nat. Rev. Cancer *24*, 261–273. https://doi.org/10.1038/s41568-024-00669-8.
4. Verhaak, R.G.W., Bafna, V., and Mischel, P.S. (2019). Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat. Rev. Cancer *19*, 283–288. https://doi.org/10.1038/s41568-019-0128-6.
5. Shoshani, O., Brunner, S.F., Yaeger, R., Ly, P., Nechemia-Arbely, Y., Kim, D.H., Fang, R., Castillon, G.A., Yu, M., Li, J.S.Z., et al. (2021). Chromothripsis drives the evolution of gene amplification in cancer. Nature *591*, 137–141. https://doi.org/10.1038/s41586-020-03064-z.
6. Rose, J.C., Belk, J.A., Wong, I.T.L., Luebeck, J., Horn, H.T., Daniel, B., Jones, M.G., Yost, K.E., Hung, K.L., Kolahi, K.S., et al. (2025). Disparate pathways for extrachromosomal DNA biogenesis and genomic DNA repair. Cancer Discov. *15*, 69–82. https://doi.org/10.1158/2159-8290.CD-23-1117.

(F) Dose-response curves showing increased sensitivity to ATMi, CHKi, or camptothecin (a TOP1 inhibitor, CPT) over 48 h in COLO320HSR and COLO320DM cells. Mean values ± SEM were presented.
(G) Schematic illustrating our proposed model of the relationship between ecDNA and DDR. Increased collisions between replisome or transcription complexes and topoisomerase covalent cleavage complexes (TOPCCs) lead to the formation of abortive TOPCCs, resulting in elevated levels of abortive TOPCCs in ecDNA+ cells. A subset of these elevated DSBs persists into the mitotic phase and is primarily repaired by the alt-NHEJ, which involves POLθ and LIG3. Correspondingly, ecDNA maintenance requires DDR, and inhibiting DDR impairs the circularization of ecDNA.
See also Figure S7.

7. Wang, Y., Wang, M., Djekidel, M.N., Chen, H., Liu, D., Alt, F.W., and Zhang, Y. (2021). eccDNAs are apoptotic products with high innate immunostimulatory activity. Nature *599*, 308–314. https://doi.org/10.1038/s41586-021-04009-w.

8. Pommier, Y., Nussenzweig, A., Takeda, S., and Austin, C. (2022). Human topoisomerases and their roles in genome stability and organization. Nat. Rev. Mol. Cell Biol. *23*, 407–427. https://doi.org/10.1038/s41580-022-00452-3.

9. Duardo, R.C., Marinello, J., Russo, M., Morelli, S., Pepe, S., Guerra, F., Gómez-González, B., Aguilera, A., and Capranico, G. (2024). Human DNA topoisomerase I poisoning causes R loop-mediated genome instability attenuated by transcription factor IIS. Sci. Adv. *10*, eadm8196. https://doi.org/10.1126/sciadv.adm8196.

10. Hoa, N.N., Shimizu, T., Zhou, Z.W., Wang, Z.Q., Deshpande, R.A., Paull, T. T., Akter, S., Tsuda, M., Furuta, R., Tsutsui, K., et al. (2016). Mre11 Is Essential for the Removal of Lethal Topoisomerase 2 Covalent Cleavage Complexes. Mol. Cell *64*, 580–592. https://doi.org/10.1016/j.molcel.2016.10.011.

11. van Leen, E., Brückner, L., and Henssen, A.G. (2022). The genomic and spatial mobility of extrachromosomal DNA and its implications for cancer therapy. Nat. Genet. *54*, 107–114. https://doi.org/10.1038/s41588-021-01000-z.

12. Ruiz, J.C., Choi, K.H., von Hoff, D.D., Roninson, I.B., and Wahl, G.M. (1989). Autonomously replicating episomes contain mdr1 genes in a multidrug-resistant human cell line. Mol. Cell. Biol. *9*, 109–115. https://doi.org/10.1128/mcb.9.1.109-115.1989.

13. Chamorro Gonzalez, R., Conrad, T., Stober, M.C., Xu, R., Giurgiu, M., Rodriguez-Fos, E., Kasack, K., Bruckner, L., van Leen, E., Helmsauer, K., et al. (2023). Parallel sequencing of extrachromosomal circular DNAs and transcriptomes in single cancer cells. Nat Genet *55*, 880–890. https://doi.org/10.1038/s41588-023-01386-y.

14. Møller, H.D., Lin, L., Xiang, X., Petersen, T.S., Huang, J., Yang, L., Kjeldsen, E., Jensen, U.B., Zhang, X., Liu, X., et al. (2018). CRISPR-C: circularization of genes and chromosome by CRISPR in human cells. Nucleic Acids Res. *46*, e131. https://doi.org/10.1093/nar/gky767.

15. Lange, J.T., Rose, J.C., Chen, C.Y., Pichugin, Y., Xie, L., Tang, J., Hung, K. L., Yost, K.E., Shi, Q., Erb, M.L., et al. (2022). The evolutionary dynamics of extrachromosomal DNA in human cancers. Nat. Genet. *54*, 1527–1533. https://doi.org/10.1038/s41588-022-01177-x.

16. Levan, G., Mandahl, N., Bengtsson, B.O., and Levan, A. (1977). Experimental elimination and recovery of double minute chromosomes in malignant cell populations. Hereditas *86*, 75–90. https://doi.org/10.1111/j.1601-5223.1977.tb01214.x.

17. Quinn, L.A., Moore, G.E., Morgan, R.T., and Woods, L.K. (1979). Cell lines from human colon carcinoma with unusual cell products, double minutes, and homogeneously staining regions. Cancer Res. *39*, 4914–4924.

18. Tang, J., Weiser, N.E., Wang, G., Chowdhry, S., Curtis, E.J., Zhao, Y., Wong, I.T.L., Marinov, G.K., Li, R., Hanoian, P., et al. (2024). Enhancing transcription-replication conflict targets ecDNA-positive cancers. Nature *635*, 210–218. https://doi.org/10.1038/s41586-024-07802-5.

19. Chen, M., Huang, B., Zhu, L., Wang, Q., Pang, Y., Cheng, M., Lian, H., Liu, M., Zhao, K., Xu, S., et al. (2022). DNA Damage Response Evaluation Provides Novel Insights for Personalized Immunotherapy in Glioma. Front. Immunol. *13*, 875648. https://doi.org/10.3389/fimmu.2022.875648.

20. Uusküla-Reimand, L., and Wilson, M.D. (2022). Untangling the roles of TOP2A and TOP2B in transcription and cancer. Sci. Adv. *8*, eadd4920. https://doi.org/10.1126/sciadv.add4920.

21. Keeney, S., Giroux, C.N., and Kleckner, N. (1997). Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. Cell *88*, 375–384. https://doi.org/10.1016/s0092-8674(00)81876-0.

22. Ru, H., Chambers, M.G., Fu, T.M., Tong, A.B., Liao, M., and Wu, H. (2015). Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures. Cell *163*, 1138–1152. https://doi.org/10.1016/j.cell.2015.10.055.

23. Stewart, J.B., and Chinnery, P.F. (2021). Extreme heterogeneity of human mitochondrial DNA from organelles to populations. Nat. Rev. Genet. *22*, 106–118. https://doi.org/10.1038/s41576-020-00284-x.

24. Krasich, R., and Copeland, W.C. (2017). DNA polymerases in the mitochondria: A critical review of the evidence. Front. Biosci. (Landmark Ed) *22*, 692–709. https://doi.org/10.2741/4510.

25. Lee, J.-H., and Paull, T.T. (2021). Cellular functions of the protein kinase ATM and their relevance to human disease. Nat. Rev. Mol. Cell Biol. *22*, 796–814. https://doi.org/10.1038/s41580-021-00394-2.

26. Rubio-Contreras, D., and Gómez-Herreros, F. (2023). TDP1 suppresses chromosomal translocations and cell death induced by abortive TOP1 activity during gene transcription. Nat. Commun. *14*, 6940. https://doi.org/10.1038/s41467-023-42622-7.

27. Cristini, A., Ricci, G., Britton, S., Salimbeni, S., Huang, S.N., Marinello, J., Calsou, P., Pommier, Y., Favre, G., Capranico, G., et al. (2019). Dual Processing of R-Loops and Topoisomerase I Induces Transcription-Dependent DNA Double-Strand Breaks. Cell Rep. *28*, 3167–3181.e6. https://doi.org/10.1016/j.celrep.2019.08.041.

28. Shaikh, N., Mazzagatti, A., De Angelis, S., Johnson, S.C., Bakker, B., Spierings, D.C.J., Wardenaar, R., Maniati, E., Wang, J., Boemo, M.A., et al. (2022). Replication stress generates distinctive landscapes of DNA copy number alterations and chromosome scale losses. Genome Biol. *23*, 223. https://doi.org/10.1186/s13059-022-02781-0.

29. Wu, S., Turner, K.M., Nguyen, N., Raviram, R., Erb, M., Santini, J., Luebeck, J., Rajkumar, U., Diao, Y., Li, B., et al. (2019). Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature *575*, 699–703. https://doi.org/10.1038/s41586-019-1763-5.

30. Brambati, A., Sacco, O., Porcella, S., Heyza, J., Kareh, M., Schmidt, J.C., and Sfeir, A. (2023). RHINO directs MMEJ to repair DNA breaks in mitosis. Science *381*, 653–660. https://doi.org/10.1126/science.adh3694.

31. Gelot, C., Kovacs, M.T., Miron, S., Mylne, E., Haan, A., Boeffard-Dosierre, L., Ghouil, R., Popova, T., Dingli, F., Loew, D., et al. (2023). Polθ is phosphorylated by PLK1 to repair double-strand breaks in mitosis. Nature *621*, 415–422. https://doi.org/10.1038/s41586-023-06506-6.

32. van Vugt, M.A.T.M., and Tijsterman, M. (2023). POLQ to the rescue for double-strand break repair during mitosis. Nat. Struct. Mol. Biol. *30*, 1828–1830. https://doi.org/10.1038/s41594-023-01168-4.

33. Dang, X., Ogbu, S.C., Zhao, J., Nguyen, L.N.T., Cao, D., Nguyen, L.N., Khanal, S., Schank, M., Thakuri, B.K.C., Wu, X.Y., et al. (2020). Inhibition of topoisomerase IIA (Top2alpha) induces telomeric DNA damage and T cell dysfunction during chronic viral infection. Cell Death Dis. *11*, 196. https://doi.org/10.1038/s41419-020-2395-2.

34. Tuduri, S., Crabbé, L., Conti, C., Tourrière, H., Holtgreve-Grez, H., Jauch, A., Pantesco, V., De Vos, J., Thomas, A., Theillet, C., et al. (2009). Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. Nat. Cell Biol. *11*, 1315–1324. https://doi.org/10.1038/ncb1984.

35. Deutsch, E.W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J.J., Mendoza, L., Kundu, D.J., Wang, S., Bandla, C., Kamatchinathan, S., et al. (2023). The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Res. *51*, D1539–D1548. https://doi.org/10.1093/nar/gkac1040.

36. Perez-Riverol, Y., Xu, Q.W., Wang, R., Uszkoreit, J., Griss, J., Sanchez, A., Reisinger, F., Csordas, A., Ternent, T., Del-Toro, N., et al. (2016). PRIDE Inspector Toolsuite: Moving Toward a Universal Visualization Tool for Proteomics Data Standard Formats and Quality Assessment of ProteomeXchange Datasets. Mol. Cell. Proteomics *15*, 305–317. https://doi.org/10.1074/mcp.O115.050229.

37. Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency

across The Cancer Genome Atlas. Cell Rep. *23*, 239–254.e6. https://doi.org/10.1016/j.celrep.2018.03.076.

38. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308. https://doi.org/10.1038/nprot.2013.143.

39. Stewart, S.A., Dykxhoorn, D.M., Palliser, D., Mizuno, H., Yu, E.Y., An, D.S., Sabatini, D.M., Chen, I.S.Y., Hahn, W.C., Sharp, P.A., et al. (2003). Lentivirus-delivered stable gene silencing by RNAi in primary cells. Rna *9*, 493–501. https://doi.org/10.1261/rna.2192803.

40. Yang, M., Yu, H., Yu, X., Liang, S., Hu, Y., Luo, Y., Izsvák, Z., Sun, C., and Wang, J. (2022). Chemical-induced chromatin remodeling reprograms mouse ESCs to totipotent-like stem cells. Cell Stem Cell *29*, 400–418.e13. https://doi.org/10.1016/j.stem.2022.01.010.

41. Deshpande, V., Luebeck, J., Nguyen, N.D., Bakhtiari, M., Turner, K.M., Schwab, R., Carter, H., Mischel, P.S., and Bafna, V. (2019). Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat. Commun. *10*, 392. https://doi.org/10.1038/s41467-018-08200-y.

42. Luebeck, J., Huang, E., Dameracharla, B., Kim, F., Liefeld, T., Ahuja, R., Prasad, D.S., Prasad, G., Kim, S., Kim, H., et al. (2024). 13. AmpliconSuite: Analyzing focal amplifications in cancer genomes. Cancer Genet. *286–287*, S5. https://doi.org/10.1016/j.cancergen.2024.08.015.

43. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics *26*, 2204–2207. https://doi.org/10.1093/bioinformatics/btq351.

44. Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinformatics *47*, 11.12.1–11.1234. https://doi.org/10.1002/0471250953.bi1112s47.

45. Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics *35*, 421–432. https://doi.org/10.1093/bioinformatics/bty648.

46. Talevich, E., Shain, A.H., Botton, T., and Bastian, B.C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput. Biol. *12*, e1004873. https://doi.org/10.1371/journal.pcbi.1004873.

47. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

48. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

49. Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675. https://doi.org/10.1038/nmeth.2089.

50. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

51. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372. https://doi.org/10.1038/nbt.1511.

52. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

53. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

54. Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., Liu, Y., Feng, J., Chen, H., He, Y., et al. (2023). TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. Mol. Plant *16*, 1733–1742. https://doi.org/10.1016/j.molp.2023.09.010.

55. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

56. Morton, A.R., Dogan-Artun, N., Faber, Z.J., MacLeod, G., Bartels, C.F., Piazza, M.S., Allan, K.C., Mack, S.C., Wang, X., Gimple, R.C., et al. (2019). Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. Cell *179*, 1330–1341.e13. https://doi.org/10.1016/j.cell.2019.10.039.

57. Zhao, B., Liu, P., Fukumoto, T., Nacarelli, T., Fatkhutdinov, N., Wu, S., Lin, J., Aird, K.M., Tang, H.Y., Liu, Q., et al. (2020). Topoisomerase 1 cleavage complex enables pattern recognition and inflammation during senescence. Nat. Commun. *11*, 908. https://doi.org/10.1038/s41467-020-14652-y.

58. Torrecilla, I., Ruggiano, A., Kiianitsa, K., Aljarbou, F., Lascaux, P., Hoslett, G., Song, W., Maizels, N., and Ramadan, K. (2024). Isolation and detection of DNA-protein crosslinks in mammalian cells. Nucleic Acids Res. *52*, 525–547. https://doi.org/10.1093/nar/gkad1178.

59. Sirbu, B.M., Couch, F.B., and Cortez, D. (2012). Monitoring the spatiotemporal dynamics of proteins at replication forks and in assembled chromatin using isolation of proteins on nascent DNA. Nat. Protoc. *7*, 594–605. https://doi.org/10.1038/nprot.2012.010.

60. Li, N., Kuo, C.L., Paniagua, G., van den Elst, H., Verdoes, M., Willems, L.I., van der Linden, W.A., Ruben, M., van Genderen, E., Gubbens, J., et al. (2013). Relative quantification of proteasome activity by activity-based protein profiling and LC-MS/MS. Nat. Protoc. *8*, 1155–1168. https://doi.org/10.1038/nprot.2013.065.

61. Li, X., Zhou, J., Zhao, W., Wen, Q., Wang, W., Peng, H., Gao, Y., Bouchonville, K.J., Offer, S.M., Chan, K., et al. (2022). Defining Proximity Proteome of Histone Modifications by Antibody-mediated Protein A-APEX2 Labeling. Genomics Proteomics Bioinformatics *20*, 87–100. https://doi.org/10.1016/j.gpb.2021.09.003.

62. Fang, D., Gan, H., Lee, J.H., Han, J., Wang, Z., Riester, S.M., Jin, L., Chen, J., Zhou, H., Wang, J., et al. (2016). The histone H3.3K36M mutation reprograms the epigenome of chondroblastomas. Science *352*, 1344–1348. https://doi.org/10.1126/science.aae0065.

63. Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat. Commun. *10*, 1930. https://doi.org/10.1038/s41467-019-09982-5.

64. Tian, C., Zhou, J., Li, X., Gao, Y., Wen, Q., Kang, X., Wang, N., Yao, Y., Jiang, J., Song, G., et al. (2023). Impaired histone inheritance promotes tumor progression. Nat. Commun. *14*, 3429. https://doi.org/10.1038/s41467-023-39185-y.

65. Zhang, X., Smits, A.H., van Tilburg, G.B., Ovaa, H., Huber, W., and Vermeulen, M. (2018). Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat. Protoc. *13*, 530–550. https://doi.org/10.1038/nprot.2017.147.

66. Ding, K., He, Y., Wei, J., Fu, S., Wang, J., Chen, Z., Zhang, H., Qu, Y., Liang, K., Gong, X., et al. (2022). A score of DNA damage repair pathway with the predictive ability for chemotherapy and immunotherapy is strongly associated with immune signaling pathway in pan-cancer. Front. Immunol. *13*, 943090. https://doi.org/10.3389/fimmu.2022.943090.

67. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 7. https://doi.org/10.1186/1471-2105-14-7.

68. Wong, N., John, S., Nussenzweig, A., and Canela, A. (2021). END-seq: An Unbiased, High-Resolution, and Genome-Wide Approach to Map DNA Double-Strand Breaks and Resection in Human Cells. Methods Mol. Biol. *2153*, 9–31. https://doi.org/10.1007/978-1-0716-0644-5_2.

Please cite this article in press as: Kang et al., Extrachromosomal DNA replication and maintenance couple with DNA damage pathway in tumors, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.012

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-mouse-HRP	Beyotime	Cat# A0216; RRID: AB_2860575
anti-rabbit-HRP	Beyotime	Cat# A0208; RRID: AB_2892644
ATM	Abcam	Cat# ab32420; RRID: AB_725574
ATR	ProteinTech	Cat# 19787-1-AP; RRID: AB_10639516
BRCA1	Cell Signaling Technology	Cat# 9010S; RRID: AB_2228244
CHK1	Abcam	Cat# ab32531; RRID: AB_726821
CHK2	Abcam	Cat# ab207446; RRID: AB_3683633
CtIP	ProteinTech	Cat# 12624-1-AP; RRID: AB_2175250
DNA Ligase I	Abcam	Cat# ab177946; RRID: AB_3683634
DNA Ligase III	Santa Cruz	Cat# sc-390922; RRID: AB_3683638
DNA Ligase IV	Abcam	Cat# ab193353; RRID: AB_2801534
Donkey anti-Mouse IgG- Alexa Fluor 488	Invitrogen	Cat# A-21202; RRID: AB_141607
Donkey anti-Mouse IgG- Alexa Fluor 555	Invitrogen	Cat# A-31570; RRID: AB_2536180
Donkey anti-mouse IgG-Alexa Fluor 647	Invitrogen	Cat# A-31571; RRID: AB_162542
Donkey anti-Rabbit IgG-Alexa Fluor 488	Invitrogen	Cat# A-21206; RRID: AB_2535792
Donkey anti-Rabbit IgG-Alexa Fluor 555	Invitrogen	Cat# A-31572; RRID: AB_162543
dsDNA	Abcam	Cat# ab27156; RRID: AB_470907
FEN1	ProteinTech	Cat# 14768-1-AP; RRID: AB_2102780
HA-Tag	Cell Signaling Technology	Cat# 3724S; RRID: AB_1549585
Histone H3	Abcam	Cat# ab1791; RRID: AB_302613
MDC1	NOVUS	Cat# NB100-397; RRID: AB_2143264
MRE11	Cell Signaling Technology	Cat# 4847T; RRID: AB_10693469
PCNA	Abcam	Cat# ab18197; RRID: AB_444313
pH3S10	Abcam	Cat# ab267372; RRID: AB_2934071
phospho-ATM	ROCKLAND	Cat# 200-301-400; RRID: AB_217868
phospho-ATR	Abcam	Cat# ab178407; RRID: AB_2721899
phospho-CHK1	Abcam	Cat# ab47318; RRID: AB_869137
phospho-CHK2	Abcam	Cat# ab85743; RRID: AB_10858765
PLK1	Cell Signaling Technology	Cat# 4513S; RRID: AB_2167409
POLQ	ProteinTech	Cat# 28590-1-AP; RRID: AB_3086067
Rabbit IgG	Abcam	Cat# ab37415; RRID: AB_2631996
RAD51	Abcam	Cat# ab133534; RRID: AB_2722613
TetR	Abcam	Cat# ab302642; RRID: AB_3683635
TOP1	Abcam	Cat# ab109374; RRID: AB_10861978
TOP1CC	Sigma	Cat# MABE1084; RRID: AB_2756354
TOP2A	Cell Signaling Technology	Cat# 12286S; RRID: AB_2797871
TOP2B	Abcam	Cat# ab72334; RRID: AB_1271267
XRCC1	ProteinTech	Cat# 21468-1-AP; RRID: AB_2878865
XRCC6	Santa Cruz	Cat# sc-56130; RRID: AB_794206
α-Tubulin	Beyotime	Cat# AF5012; RRID: AB_3683636
β-actin	Beyotime	Cat# AF0003; RRID: AB_2893353
γH2AX	Cell Signaling Technology	Cat# 9718S; RRID: AB_2118009
γH2AX	Cell Signaling Technology	Cat# 2577S; RRID: AB_2118010

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
γH2AX	Cell Signaling Technology	Cat# 80312S; RRID: AB_2799949
Bacterial and virus strains		
*Trans*5α chemically competent Cell	TransGen Biotech	Cat# CD201-01
PX459	Feng Zhang[38]	addgene:62988; RRID# Addgene_62988
psPAX2	Didier Trono	addgene:12260; RRID# Addgene_12260
pMD2.G	Didier Trono	addgene:12259; RRID# Addgene_12259
pLKO/PuroR vectors	Robert A Weinberg[39]	addgene:8453; RRID# Addgene_8453
PiggyBac vector	Yang et al.[40]	N/A
SB100X	Yang et al.[40]	N/A
Chemicals, peptides, and recombinant proteins		
2× Kapa HiFi HotStart Ready Mix	ROCHE	Cat# 7958935001
4% paraformaldehyde	Beyotime	Cat# P0099-500ml
Analytical column	ThermoFisher	Cat# P/N 164534
Antifade Mounting Medium with DAPI	Beyotime	Cat# P0131
Aphidicolin (DNA polymerase inhibitor)	MCE	Cat# HY-N6733
ART558 (POLθ inhibitor)	MCE	Cat# HY-141520
AZD7762 (CHK inhibitor)	MCE	Cat# HY-10992
B02 (RAD51 Inhibitor)	MCE	Cat# HY-101462
Biotin-TEG-azide	Berry & Associates	Cat# BT1085
Blasticidin	InvivoGen	Cat# ant-bl-05
BSA	Yeason	Cat# 36105ES80
Camptothecin (TOP1 inhibitor)	MCE	Cat# HY-16560
Chloroacetamide (CAA)	Sigma	Cat# 194921
Complete EDTA-free protease inhibitor tablets	Roche	Cat# 4693132001
Concanavalin-coated (ConA) magnetic beads	Bangs Laboratories	Cat# 14794
$CuSO_4$	Sangon Biotech	Cat# A603008-0250
DAPI	Beyotime	Cat# P0131
dA-tailing buffer	NEB	Cat# B6059S
DHFR probe	Empiregenomics	Cat# DHFR-CHR05-20-ORGR
Digitonin	Sigma	Cat# D141-500MG
DMSO	Solarbio	Cat# D8371
DNase-free RNase A	Beyotime	Cat# ST576
DNAzol	Invitrogen	Cat# 10503027
Doxycycline	MCE	Cat# HY-N0565
Dulbecco's modified Eagle's medium (DMEM)	Procell	Cat# PM150210
Dynabeads streptavidin C1	Invitrogen	Cat# 65001
EdU	Beyotime	Cat# ST067
EGFR probe	MetaSystems	Cat# D-6005
Electroporate solutions	Celetrix	Cat# 1201
Endopeptidase	Wako Pure Chemical Industries	Cat# 125-05061
Exonuclease T	NEB	Cat# M0265L
Exonuclease V	NEB	Cat# M0345S
Exonuclease VII	NEB	Cat# M0379L
Glycine	Sangon Biotech	Cat# A610235-0005
ICRF-193 (TOP2 inhibitor)	MCE	Cat# HY-118590
Klenow fragment	NEB	Cat# M0210S
Klenow fragment (3'->5' exo-)	NEB	Cat# M0212L
KU55933 (ATM inhibitor)	MCE	Cat# HY-12016
L-glutamine	Procell	Cat# PB180420

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Lipofectamine 3000	Invitrogen	Cat# L3000075
MNase	NEB	Cat# M0247S
MYC probe	MetaSystems	Cat# D-6008
NEB Buffer 2	NEB	Cat# B7002S
Nocodazole (G2/M synchronizing agent)	MCE	Cat# HY-13520
NU7441 (DNA-PK inhibitor)	MCE	Cat# HY-11006
Olaparib (PARP inhibitor)	MCE	Cat# HY-10162
Opti-MEM	Gibco	Cat# 31985-070
pA-MNase	Vazyme	Cat# S701-01
pG-MNase	Vazyme	Cat# S702-01
PBS	Procell	Cat# PB180327
Penicillin and streptomycin	Procell	Cat# PB180120
Phenol-chloroform-isoamyl alcohol (PCI)	Shanghai Acmec biochemical	Cat# AC13309
PI	Invitrogen	Cat# P1304MP
Polybrene	MCE	Cat# HY-112735
Protein A/G-magnetic beads	MCE	Cat# HY-K0202
Proteinase inhibitor cocktails	Roche	Cat# 4693132001
Proteinase K	TIANGEN	Cat# RT403-02
Proteinase K for END-Seq	Qiagen	Cat# 158146
Puromycin	InvivoGen	Cat# ant-pr-1
Quick Ligase	NEB	Cat# M2200L
Reversed-phase columns	ThermoFisher	Cat# P/N 164535
RIPA buffer	Yeasen	Cat# 20101ES
RNase A	Qiagen	Cat# 158156
RPMI-1640	VivaCell	Cat# C3010-0500
SCR130 (LIG4 inhibitor)	MCE	Cat# HY-139297
Sodium acetate	Solarbio	Cat# A1070
Sodium ascorbate	Jena Bioscience	Cat# CLK-MI005
Streptavidin-sepharose beads	GE Healthcare	Cat# 17-5113-01
SYBR Green PCR Master Mix	TaKaRa	Cat# CM0139
T4 DNA polymerase	NEB	Cat# M0203L
T4 DNA Polynucleotide Kinase	NEB	Cat# M0201L
T4 Ligase Buffer	NEB	Cat# B0202S
THPTA	Sigma	Cat# 762342
Thymidine (DNA synthesis inhibitor)	MCE	Cat# HY-N1150
Triptolide (Transcription inhibitor)	MCE	Cat# HY-32735
Tris (2-carboxyethyl) phosphine (TCEP)	ThermoFisher	Cat# 77720
Triton X-100	Sangon Biotech	Cat# A600198-0500
Trypsin	Promega	Cat# V5280
Ultra-sensitive chemiluminescent substrate	Beyotime	Cat# P0018FM
USER enzyme	NEB	Cat# M5505S
VE821 (ATR inhibitor)	MCE	Cat# HY-14731
Volasertib (PLK1 inhibitor)	MCE	Cat# HY-12137
β-agarase enzyme	NEB	Cat# M0392S
Critical commercial assays		
BCA assay	Beyotime	Cat# P0009
BeyoClick™ EdU Cell Proliferation Kit with Alexa Fluor 488	Beyotime	Cat# C0071S

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BeyoClick™ EdU Cell Proliferation Kit with Alexa Fluor 555	Beyotime	Cat# C0075S
BeyoClick™ EU RNA synthesis kit	Beyotime	Cat# R0301L
CHEF Mammalian Genomic DNA plug kit	Bio-Rad	Cat# 1703591
Click-iT EdU Cell Proliferation Kit for Imaging Kit	Invitrogen	Cat# C10377
DNA Clean & Concentrator™-5	ZYMO	Cat# D4004
Electroporation kits	Celetrix	Cat# 1201
FastPure Cell/Tissue DNA Isolation Mini Kit	Vazyme	Cat# DC102-01
Duolink *in situ* Detection Reagent Red (proximity ligation assay)	Sigma-Aldrich	Cat# DU092008
Qubit assay	Invitrogen	Cat# Q33231
VAHTS® Universal DNA Library Prep Kit	Vazyme	Cat# ND607
Deposited data		
DDR core pathways and gene set	Knijnenburg et al.[37]	https://gdc.cancer.gov/about-data/publications/PanCan-DDR-2018
Human reference proteome	UniProt Database	UniProtKB/Swiss-Prot and UniProtKB/TrEMBL, version 2021_04
Links between identified proteins	STRING database (v11.5)	https://string-db.org/cgi/input.pl
Sequencing datasets (WGS and END-seq)	This paper	GSE272961
TCGA clinical data (Gene expression data)	GDC HUB: HTSeq-FPKM	https://portal.gdc.cancer.gov/
TCGA sample analysis	UCSC Xena	https://xenabrowser.net/
The MS proteomics data	This paper	PXD060725
Experimental models: Cell lines		
293T	ATCC	RRID# CVCL_0063
293TecDNA-	This study	N/A
293TecDNA+	This study	This study
COLO320DM	ATCC	RRID# CVCL_0219
COLO320HSR	ATCC	RRID# CVCL_0220
HeLa-DM (PD29429i)	Don W. Cleveland (Ludwig Cancer Research, University of California at San Diego)[5]	N/A
HeLa-HSR (PD29429u)	Don W. Cleveland (Ludwig Cancer Research, University of California at San Diego)[5]	N/A
HT-29	ATCC	RRID# CVCL_0320
HT-29 MTX	ATCC	RRID# CVCL_6834
PC3 (-)	ATCC	RRID# CVCL_0035
PC3 (+)	Paul S. Mischel (Stanford University)[15]	N/A
TetO-eGFP PC3 (+)	Paul S. Mischel (Stanford University)[15]	N/A
U251	ATCC	RRID# CVCL_0021
U251ecDNA-	This study	N/A
U251ecDNA+	This study	N/A
Oligonucleotides		
Oligonucleotides	This study	Table S2
Illumina Nextera Dual Indexing primers	This study	Vazyme, Cat# N323

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Recombinant DNA		
Plasmid: LIG3-HA (LIG3-OE)	This paper	N/A
Plasmid: LIG3^{K508V}-HA (LIG3 OE-M)	This paper	N/A
Software and algorithms		
AmpliconArchitect (AA)	Deshpande et al.[41]	https://github.com/AmpliconSuite/AmpliconArchitect
AmpliconSuite-pipeline (v.1.3.3)	Luebeck et al.[42]	https://github.com/AmpliconSuite/AmpliconSuite-pipeline
bedGraphToBigWig (v.4)	Kent et al.[43]	https://open.bioqueue.org//home/knowledge/showKnowledge/sig/ucsc-bedgraphtobigwig
Bedtools (v.2.29.2)	Quinlan[44]	https://bedtools.readthedocs.io/en/latest/
Bio-Rad software v.4.1	N/A	https://bio-rad-cfx-manager.software.informer.com/3.1/
BLOCK-iT™ RNAi Designer	N/A	https://rnaidesigner.thermofisher.com/rnaiexpress/setOption.do?designOption=shrna
bowtie2 (v.2.4.2)	Langmead et al.[45]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BWA-MEM (v0.7.17)	Li.	https://bio-bwa.sourceforge.net/
CNVKit (v.0.9.6)	Talevich et al.[46]	https://cnvkit.readthedocs.io/en/stable/
Cytoscape (version 3.9.1)	Shannon et al.[47]	https://cytoscape.org/
Homer suite (v.2.0)	Heinz et al.[48]	http://homer.ucsd.edu/homer/
ImageJ v150	Schneider et al.[49]	https://imagej.net/software/imagej/
MACS2 (v.2.2.9.1)	Zhang et al.[50]	https://github.com/taoliu/MACS/tree/ master/MACS2
MaxQuant (version 1.6.10.43)	Cox and Mann[51]	https://www.maxquant.org/
MetaScape (v3.5)	Zhou et al.[52]	https://metascape.org/gp/index.html#/main/step1
Picard MarkDuplicates (v2.18.29)	"Picard Toolkit." 2019. Broad Institute, GitHub Repository.	http://broadinstitute.github.io/picard/
Primer-BLAST software	N/A	https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi
R (Version 4.03)	R Project	https://www.r-project.org/
Samtools (v.1.7)	Li et al.[53]	http://samtools.sourceforge.net/
Tanon App for Biology Software (Version 1.0.0000)	N/A	https://biotanon.com/
TBtools-II (v1.115)	Chen et al.[54]	https://github.com/CJ-Chen/TBtools-II
Trimmomatic (v.0.39)	Bolger et al.[55]	https://github.com/usadellab/Trimmomatic
Foci Analyzer	N/A	https://github.com/Bioimaging-NKI/Foci-analyzer

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Mammalian cell culture

The human glioblastoma cell line U251, embryonic kidney cell line 293T, prostate cancer cell lines [PC3 (+) with ecDNA and its isogenic counterpart PC3 (-) without ecDNA], TetO-eGFP PC3 (+) (prostate cancer cell line with TetR-eGFP labelled ecDNA), cervical adenocarcinoma cell lines (HeLa-DM with ecDNA and its isogenic counterpart HeLa-HSR), and colorectal adenocarcinoma cell lines (HT-29 MTX with ecDNA and its isogenic counterpart HT-29) were cultured in Dulbecco's modified Eagle's medium (DMEM, Procell,

PM150210) containing 10% fetal bovine serum (FBS, Excell, FSP500), supplemented with 100 U ml^{-1} penicillin, 100 mg ml^{-1} streptomycin, and 2 mM L-glutamine, at 37°C under 5% CO_2. COLO320DM and COLO320HSR cells were maintained in RPMI-1640 (VivaCell, C3010-0500) with the same supplements as above. All cell lines were authenticated by karyotyping and routinely tested to confirm the absence of mycoplasma contamination.

METHOD DETAILS

Guide RNA design for CRISPR-C

sgRNAs were designed to target both the upstream and downstream regions of the *EGFR* gene, which is highly amplified on ecDNA and commonly observed in human glioblastoma cells.[56] We used 1,000 base pairs (bp) of sequence flanking each end of this segment (Chr7: 55,015,000–55,016,000; Chr7: 55,218,000–55,219,000; hg38, UCSC Genome Browser) to design guide RNAs using the CHOPCHOP web tool (https://chopchop.cbu.uib.no). These sequences are listed in Table S2.

ecDNA induction by CRISPR-C

U251 and 293T cells were trypsinized, quenched with DMEM (Glutamax, 10% FBS), and centrifuged at 180 g for 5 min. Cells were washed once with PBS (10 mM sodium phosphate, 0.15 M NaCl, pH 7.4) before resuspension in Opti-MEM (Gibco, 31985-070), supplemented with 5% FBS, and 1% Glutamax. The sgRNA was cloned into a PX459[38] (Addgene, 62988) construct co-expressing wild-type SpCas9 and a PGK-Venus cassette. Then 2.5 μg sgRNAs vectors were co-transfected with electroporate solutions (Solution A and Solution B) (Celetrix, 1201) into U251 and 293T cells (6.5×10^5 per ml) by electroporation according to the manufacturer's instructions. We used a 200 μl pipet tip and electroporated with the Celetrix Transfection System using the following optimized settings: 480 V, 30 ms pulse width. Electroporated cells were dispensed into 1 ml DMEM and seeded into a 12-well plate.

For positive selection, 3 days after electroporation, cells were passaged and changed to a medium containing 0.5 μg ml^{-1} puromycin. The medium was changed every 2–3 days, and cells were passaged when at 70%–80% confluence. Cells were collected after 7 days of puromycin incubation and screened by breakpoint PCR according to the sgRNA cutting sites around the *EGFR* gene. Primers for breakpoint PCR are listed in Table S2. Cells with correct PCR products were isolated by diluting and cultured for 2–3 weeks. Individual clones were expanded for genotyping, and the existence of extrachromosomal *EGFR* (*ecEGFR*) amplicons was verified using Sanger sequencing.

Circular ecDNA detection with linear DNA digestion

Cells were suspended in 0.2 ml lysis buffer (50 mM KCl, 1.5 mM $MgCl_2$, 0.5% NP40 and 0.5% Tween-20, 10 mM Tris, pH 8.5) and incubated with 10 μl proteinase K (19.1 mg/ml, TIANGEN, RT403-02) for 2.5 h at 55°C to remove proteins. The samples were then incubated at 95°C for 10 min to inactivate proteinase K, followed by cooling at room temperature for 15 min. The proteinase K-digested samples were stored at -20°C.

For the exonuclease digestion experiment, where linear DNA products were not added, a maximum of 2 μg total DNA was digested with Exonuclease V (NEB, M0345S) in a 50 μl reaction volume at 37°C for 18 h. The exonuclease was then inactivated at 70°C for 30 min. After inactivation, the digested samples were used as PCR templates. Gel images of the corresponding PCR products across the junction site (*EGFR* junction site) were obtained using DNA templates before (Exo-) and after exonuclease treatment for 18 h (Exo+). DNA isolated from U251ecDNA- cells was used as a negative control. *GAPDH* was used as an indicator of linear DNA.

For the exonuclease digestion experiment, where linear DNA products were added, a maximum of 2 μg total DNA with 200 ng linear products was digested with the Exonuclease V in a 50 μl reaction volume at 37°C for 1 h. The exonuclease was then inactivated at 70°C for 30 min. After inactivation, the digested samples were used as PCR templates. Gel images of PCR products using equal amounts of DNA templates (before exonuclease treatment) isolated from COLO320DM cells treated with ATMi or transiently transfected with siRNA targeting *LIG3* and scramble siRNAs. *EGFR* was used as an indicator of linear DNA.

Quantitative PCR (qPCR)

Genomic DNA was extracted using FastPure Cell/Tissue DNA Isolation Mini Kit (Vazyme, DC102-01) according to the manufacturer's instructions, including a 30 min incubation at 65 °C during the proteinase K digestion step, and DNA was eluted with 50 μl ddH_2O. Amplicons for the *EGFR*, *MYC*, *MYC-PVT1* fusion breakpoint, *DHFR*, and *GAPDH* were designed using Primer-BLAST software (https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi). qPCR was performed using 50 ng or 1.5 μl template DNA and 0.5 μM primers with the SYBR Green PCR Master Mix (TaKaRa, CM0139) in 8-well PCR tubes. Reactions were run and monitored on a qTower 3 or Bio-Rad CFX Manager (Version 3.1), and Ct values were calculated with the qTower or Bio-Rad software v.4.1. Primers for qPCR are listed in Table S2.

Immunofluorescence (IF)

Cells grown on confocal dish or coverslips were fixed with 4% paraformaldehyde for 15 min and permeabilized in 0.3% Triton X-100 for 20 min at room temperature (RT). After fixation and permeabilization, the cells were washed with PBS three times and blocked with

3% BSA in PBS at RT for 1 h. Subsequently, the cells were incubated with primary antibodies [anti-γH2AX antibody (1:200), anti-TOP1CC antibody (1:200),[57,58] and anti-pH3S10 antibody (1:200)] diluted in blocking buffer at 4°C overnight in a humidified cold chamber.

Following overnight incubation, the cells were washed three times with PBS and then incubated with fluorescent secondary antibodies [anti-rabbit or anti-mouse IgG-Alexa Fluor (1:200)] diluted in 1% BSA for 1 h at RT in the dark. After incubation, the cells were washed three times with PBST (PBS containing 0.2% Triton X-100). The samples were then washed with PBS, mounted, and counterstained with 4',6-diamidino-2-phenylindole (DAPI)-containing antifade mountant. Cell images were captured at 60× or 100× resolution on a Nikon confocal microscope. Images were analyzed using ImageJ software (v150).[49]

Metaphase chromosome spread

Cells in metaphase were obtained by nocodazole treatment at 0.5 μg ml^{-1} for 24 h. Cells were collected, washed twice with PBS, resuspended with 75 mM KCl, and incubated at RT for 30 min to 1 h. Samples were then fixed with Carnoy's fixative (3:1 methanol: glacial acetic acid, v/v) for 10 min and washed an additional three times with fixative before being dropped onto humidified glass slides for metaphase cell preparations.

Metaphase DNA fluorescence *in situ* hybridization (FISH)

Fixed samples on slides were equilibrated briefly in 2×SSC buffer, followed by dehydration in ascending ethanol series (70%, 85%, 100%) for approximately 2 min each. Pre-warmed FISH probes were diluted in hybridization buffer (1:1 formamide: 2×SSC buffer, v/v) and added onto a slide, and the coverslip was applied and sealed with rubber cement. The FISH probe and samples were co-denatured on a 75°C hotplate for 5 min, and then hybridization was carried out at 37°C overnight in a dark, humidified chamber. Samples were washed with 0.4×SSC at 50°C, followed by a final wash in 2×SSC with 0.1% Tween-20 at RT (twice, 2 min at a time). After washing, samples were sealed with DAPI-antifade mounting medium, with coverslips covered. Cell images were captured at 100× resolution on a Nikon confocal microscope. FISH foci number was quantified using ImageJ software (v150).[49]

Dual FISH-immunofluorescence (FISH-IF)

For *ecMYC* FISH-IF experiments, cells were swollen, spread, fixed and dehydrated as above. After incubation with the *MYC* probe overnight, samples were refixed with 4% paraformaldehyde for 15 min and permeabilized in 0.3% Triton X-100 for 20 min at RT, then washed with PBS, and blocked with 100 μg ml^{-1} DNase-free RNase A (Beyotime, ST576) in blocking buffer (3% BSA in PBS) for 1 h at 37°C. γH2AX or TOP1CC antibody was then used at 1:200 dilution in blocking solution overnight at 4°C and washed with 0.05% Tween in PBS, followed by incubation with Alexa Fluor 488-conjugated anti-rabbit IgG or anti-mouse IgG for 2 h at RT and three washes in PBS with 0.05% Triton X-100. FISH-IF foci number were quantified as described above.

Western blotting analysis

Cells were harvested by trypsinization, pelleted, washed in PBS, and lysed in RIPA buffer (Yeasen, 20101ES) on ice for 30 min. Protein concentrations of cell lysates were determined by BCA assay (Beyotime, P0009). Next, 30 μg of each sample was mixed with protein loading buffer, boiled, and separated on SDS-PAGE gels. Proteins were then transferred to PVDF membrane (Millipore, ISEQ00010) and blocked with 5% non-fat milk in TBST for 1 h, then washed three times in 1×TBST for approximately 5 min each. The membrane was then incubated with the corresponding primary antibody overnight at 4°C. Following primary antibody incubation, the membrane was washed with TBST and incubated with secondary antibodies (anti-rabbit-HRP or anti-mouse-HRP) for 1 h at RT, followed by additional 1×TBST washes. Bands were visualized using Tanon App for Biology Software (Version 1.0.0000) after incubating the membranes with ultra-sensitive chemiluminescent substrate (Beyotime, P0018FM).

Detection of EdU incorporation

To detect ecDNA replication, cells in 6-well plates were incubated continuously with culture medium containing 10 μM EdU for 20 h. Cells were fixed with 4% paraformaldehyde for 10 min, permeabilized with 0.3% Triton X-100 for 15 min at RT, and washed with PBS. After washing with PBS, EdU incorporation was detected using the BeyoClick™ EdU Cell Proliferation Kit with Alexa Fluor 488 (Beyotime, C0071S) or Alexa Fluor 555 (Beyotime, C0075S) following the manufacturer's instructions. The Click-iT reaction was carried out for 30 min at RT. Then, cells were washed with PBS, followed by DNA staining with DAPI for 5 min at RT, and then they were mounted with Antifade Mounting Medium with DAPI (Beyotime, P0131) and viewed with an A1R confocal microscope (Nikon, Japan).

Quantification of nascent RNA production

Cells grown on 6-well plates were pulsed-labelled with 10 mM EU for 30 min, washed twice with 1×PBS and fixed in 4% formaldehyde for 20 min at RT. After three washes with 1×PBS, the cells were permeabilized with 1×PBS containing 0.3% Triton X-100 for 15 min at RT. EU incorporation was detected with BeyoClick™ EU RNA synthesis kit with Alexa Fluor 488 (Beyotime, R0301L) following the manufacturer's instructions. The Click-iT reaction was carried out for 30 min at RT. Then, cells were washed with PBS, and DNA was stained with DAPI for 5 min at RT, mounted with Antifade Mounting Medium with DAPI (Beyotime, P0131) and viewed with an A1R confocal microscope (Nikon, Japan).

Generation of lentivirus-mediated knockdown cells

Stable knockdown cell lines were generated using lentiviral pLKO/PuroR vectors (Addgene, 8453)[39] after puromycin selection. Knockdown efficiency was confirmed by western blot. For lentivirus production, 5×10^6 293T cells were electro-transfected with the vector of target and appropriate packaging plasmids psPAX2 (Addgene, 12260) and pMD2.G (Addgene, 12259) and seeded into 10-cm dishes containing 10 ml Opti-MEM. The medium was replaced, and supernatants were collected at 24 h and 48 h. Collected supernatants were filtered with 0.45 μm filters. Cells were seeded in 12-well plates at a density of 1×10^5 cells per well. After incubating the cells to reach 50% confluence, we infected them with 40 μl per ml lentiviral particles expressing a shRNA targeting the desired gene sequences, or a control scramble construct expressing puromycin in the presence of polybrene (10 mg ml^{-1}) according to the manufacturer's protocol, and subsequently kept the cells in a 5% CO_2 incubator at 37 °C. The medium was changed 48 h later, and cells were selected with puromycin for at least 1 week. shRNA sequences were designed with BLOCK-iT™ RNAi Designer and are shown in Table S2.

Overexpression and shRNA-rescued cell lines construction

To further monitor ecDNA dynamics related to topoisomerases and LIG3, we constructed overexpression plasmid for LIG3 and TOP1 via homologous recombination of CDS region onto a PiggyBac[40] vector harboring blasticidin S deaminase (BSD). Overexpression plasmids were co-transfected with transposase-expressing plasmid SB100X[40] at a ratio of 2:1 into COLO320DM and corresponding shRNA knock-down cells. One day after transfection, stable transfectants were selected by 10 μg ml^{-1} blasticidin at least 10 days.

siRNA transfection

siRNAs were purchased from GenePharma Company. To transiently knock down gene expression, siRNAs were transfected using Lipofectamine 3000 according to the manufacturer's instructions. Briefly, cells were seeded to 70–90% confluence. Then, 7.5μl Lipofectamine 3000 and 7μl siRNA (20μM) were each diluted with 125 μl Opti-MEM medium. The diluted siRNA and Lipofectamine 3000 were mixed and incubated for 15 min at room temperature. The siRNA-Lipofectamine 3000 complex was then added to the cells. Cells were harvested for western blot, FISH, immunofluorescence, or PLA 48 or 72 hours after transfection. siRNA sequences were shown in Table S2.

Drug treatments

For the FISH experiments, 1×10^6 cells, including from the COLO320DM, COLO320HSR, PC3 (+), PC3 (-), U251ecDNA+, U251ecDNA-, HT-29, HT-29 MTX, HeLa-DM, and HeLa-HSR cell lines, were seeded in 10 cm plates and cultured under standard conditions for 24 h. Cells were then treated for about 24 h or over time with one of the following inhibitors: 10 μM ART558, 0.2 μM AZD7762, 5 μM B02, 5 μM KU55933, 10 μM NU7441, 10 μM olaparib, 10 μM SCR130, 5 μM VE821, 500 nM volasertib, or DMSO. For the living-cell imaging assay, 1×10^5 TetO-eGFP PC3 (+) cells were seeded and cultured 24 h in confocal plates, and then treated with ART558, AZD7762, B02, KU55933, or SCR130 as described above. For the RT-PCR experiment, 1×10^5 COLO320DM and COLO320HSR cells were seeded in 6-well plates and cultured under standard conditions for 24 h. Cells were then treated for about 24 h with the inhibitors, as described above. For the IF assay, 2 μM aphidicolin (APH) and 1 μM triptolide (TPL) (final concentration) were used to treat cells for 20 h and 5 h, respectively.

Synchronization and cell cycle detection

To synchronize cells into the early S phase, cells were seeded to 20–30% confluence. PC3 (-), PC3 (+), U251 and U251 ecDNA cells were treated with 3 mM thymidine for 16 h, released by washing three times with fresh media, and then cultured in fresh media for 9 h, followed by 16 h treatment with 3 mM thymidine. COLO320DM/HSR cells were treated with 3 mM thymidine for 18 h, released by washing three times, and then cultured in fresh media for 12 h, followed by another 18 h treatment with 3 mM thymidine. After three washes, the cells were released into fresh media for 3 h, then treated with 2 μM APH for 0, 5, 9,16 and 20 h, individually. Cells were harvested, washed once and resuspended in 300 μl cold PBS, then 700 μl cold 100% ethanol was added with shaking. Cells were fixed overnight at -20 °C, then spun down at 500 g for 5 min. Cells were resuspended in 1 ml PBS, and incubated for 10 min. After centrifugation and removing the supernatant, cells were resuspended staining buffer [100 mM Tris-HCl pH 7.4; 150 mM NaCl, 1 mM $CaCl_2$, 0.5 mM $MgCl_2$, 0.1% NP40, 0.2% PI (Invitrogen, P1304MP)]. After 15 min incubation at room temperature protected from light, samples were analyzed by flow cytometry (BECKMAN COULTER CytoFLEX S and CytExpert 2.3).

Cell cytotoxicity assay

Cells, including COLO320DM, U251, PC3, and HT-29 cell lines, were seeded into white flat-bottom 96-well plates (Corning) in 100 μl medium at a density of 1×10^4 cells per well. Cells were treated with either vehicle or the inhibitors, including camptothecin (CPT), AZD7762, and KU55933, at a series of concentrations (0.01 μM to 100 μM) for 48 h, respectively. Cell cytotoxicity was determined using Cell Counting Kit-8 according to the manufacturer's protocol.

TOP1CC detection

1×10^6 COLO320DM/HSR cells were treated with camptothecin, or vehicle at indicated concentrations for 30 min. Cells were collected and lysed with 600 μl DNAzol (Invitrogen), followed by precipitation with 300 μl ice-cold absolute ethanol. The lysate

was precipitated by centrifugation at 4,800 g for 15 min at 4 °C. Pellets were washed with 75% ethanol, resuspended in 200 μl 8 mM NaOH, and heated at 65 °C for 15 min. Each sample containing 200 ng DNA was subjected to dot-blot for immunoblotting with antibodies targeting TOP1 overnight at 4 °C. In addition, each sample containing 200 ng DNA was subjected to dot-blot for immunoblotting with an anti-double-stranded DNA (dsDNA) antibody as a loading control. Three washes were performed with TBST, followed by incubation with HRP-conjugated secondary antibodies for 1 h at RT, followed by an additional TBST wash three times. Signals were visualized using Tanon App for Biology Software (Version 1.0.0000) after incubating with ultra-sensitive chemiluminescent substrate (Beyotime, P0018FM).

In situ proximity ligation assay (PLA)

The PLA was performed with a kit following the protocol supplied by Sigma-Aldrich. After being harvested, cells were permeabilized with 0.5% Triton X-100 for 1 min and then fixed with 4% paraformaldehyde for 15 min. Fixed cells were re-permeabilized with 1% Triton X-100 for 5 min. After blocking with Duolink blocking buffer at 37 °C for 60 min, cells were incubated with 1:200 dilutions of primary rabbit antibodies (TOP1, LIG3, TOP1CC, PCNA, POLθ, CtIP, FEN1, and γH2AX) and mouse anti-TetR with 1× Duolink Antibody Diluent for 60 min at 37 °C, or overnight at 4 °C. After washing twice with Duolink *in situ* wash buffer (Buffer A), cells were stained with Duolink *in situ* PLA Probe Anti-Rabbit PLUS [affinity-purified donkey anti-rabbit IgG (H + L)] and Duolink *in situ* PLA Probe Anti-Mouse MINUS [affinity-purified donkey anti-mouse IgG (H + L)] for 60 min at 37 °C. Cells were incubated with the ligation-ligase solution (Duolink *in situ* Detection Reagents Red) for 30 min at 37 °C after washing twice with Buffer A. Then, the cells were incubated with the amplification-polymerase solution (Duolink *in situ* Detection Reagents Red) for 100 min at 37 °C in dark. After washing with Buffer A twice and Duolink *in situ* wash buffer (Buffer B), plates were covered with Duolink *in situ* Mounting Medium with DAPI. Imaging was performed with a Nikon confocal microscope and nuclear PLA foci number were quantified.

EdU immunoprecipitation (EdU-IP)-qPCR

EdU-IP-qPCR was performed in COLO320 cells according to Cortez's protocol[59] with some modifications. Exponentially growing COLO320 cells were treated with EdU and KU55933 to a final concentration of 10 μM for 20 h, rapidly harvested for a total of 1×10^7 cells and washed with ice-cold PBS. Cells were then split into two samples and resuspended with 0.5 ml cold lysis buffer (1% SDS, 0.5 M NaCl, 100 mM Tris-HCl pH 8.0, 2 mM EDTA) supplied with 20 μl proteinase K and 10 μl RNAse A and then incubated for 2 h at 65 °C with shaking. The lysed cell suspension was sonicated for 5 min (5 s on/ 5 s off) using a Scientz-IID at 60 W until the average fragment size was less than 500 bp. After sonication, phenol-chloroform-isoamyl alcohol (PCI) extraction was performed. DNA in the aqueous phase was precipitated using 100% ethanol and 30 mM sodium acetate overnight at -20 °C, followed by centrifugation at 16,000 g for 15 min at 4 °C. The DNA pellet was washed with 75% ethanol and resuspended in EdU click reagents added in the following order to the incubated final concentration: PBS, biotin-TEG-azide (100 μM), sodium ascorbate (10 mM), and $CuSO_4$ (2 mM). Samples were placed on a nutating mixer and incubated for 1 h at RT. After incubation, the DNA pellet was extracted, precipitated as described above, and resuspended in 0.1×TE (10 mM Tris-HCl pH 8, 1 mM EDTA). After reserving an aliquot for input, biotinylated DNA was pulled down using Dynabeads streptavidin C1 (Invitrogen, 65001) following the manufacturer's protocol. After three washes, the beads were resuspended in 30 μl 0.1×TE, and qPCR was performed using the SYBR Green Premix Master Mix (TaKaRa) with a qTower3 or Bio-Rad CFX Manager (Version 3.1). Primers sequences are provided in Table S2.

EdU-labeled cells biotinylation

Cells were treated with EdU (Beyotime, ST067) at 10 μM for 2 h. Then 1×10^8 EdU-labeled cells were collected into 50 ml tubes and crosslinked using freshly prepared formaldehyde at a final concentration of 1% at RT for 15 min. Cells were quenched with 1/10 volume of 1.25 M glycine for 5 min, and centrifuged at 1,000 g for 5 min. The pellet was then washed three times with PBS and centrifuged at 1,000 g for 5 min at 4 °C. Cells were pelleted and resuspended in PBS containing 0.2% Triton X-100. After two washes with PBS, the click reaction was assembled by resuspending cells in PBS and adding the click reagents in the following order (final volume 500 μl): 1×Click-iT® cell buffer (Invitrogen, C10377), 0.5 mM THPTA (Sigma, 762342) with 2 mM $CuSO_4$, Click-iT® cell buffer additive (Invitrogen, C10377), 10 mM sodium ascorbate (Jena Bioscience, CLK-MI005) and 10 μM Biotin-TEG-azide (Berry&associates, 875770-34-6).

Streptavidin pull-down of biotinylated proteins

After biotinylation, EdU-labeled cell pellets were lysed with 500 μl lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$, 1 mM EGTA, 1% SDS, 1% NP-40, 1 mM PMSF, 1mM DTT, 1×Roche Complete EDTA-free protease inhibitor tablets and 0.4% sodium deoxycholate) for 15 min on ice. Lysates were sonicated for 3 min (3 s on/3 s off). Cell extracts were centrifuged at 16,000 g for 10 min at 4 °C. The supernatants were collected, and the SDS in the sample was diluted to 0.2% with 1×cold lysis buffer (without SDS). Streptavidin-Sepharose beads (GE Healthcare, 17-5113-01) were washed twice with 1 ml 1× cold lysis buffer (0.2% SDS), after which 50 μl beads were added to each sample, and incubated with rotation for 4 h at 4 °C. Subsequently, beads were washed twice with the following buffer: 1 ml wash buffer (50 mM Tris- HCl pH 7.5, 1% SDS), 1 ml RIPA lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$, 1 mM EGTA, 0.2% SDS, 1% NP-40, 1mM DTT), 1 ml 8 M urea buffer, 1 ml 30% acetonitrile (ACN) and 1 ml 20 mM ammonium bicarbonate. Finally, the beads were subjected to LC-MS/MS analysis.

On-bead digestion and LC-MS/MS

MS-based proteomic experiments were performed as previously described, with minor modifications.[60] Briefly, after the enrichment and washing steps, we resuspended the beads in 200 μl on-bead digestion buffer (50 mM HEPES pH 8.0, 1 μM $CaCl_2$, 2% ACN), and then added 10 mM Tris (2-carboxyethyl) phosphine (TCEP, ThermoFisher, 77720) and 40 mM chloroacetamide (CAA, Sigma, 194921), and incubated for 30 min at RT. The beads were washed with 1 ml on-bead digestion buffer, resuspended in 100 μl on-bead digestion buffer with endopeptidase (Wako Pure Chemical Industries, 125-05061) and incubated at 37°C for 3 h. Then, on-bead digestion buffer with 0.5 mg trypsin (Promega, V5280) was added for digestion at 37°C for 16 h.

Samples were desalted using StageTips, made of C18 material inserted into 200 μl pipette tips, before LC-MS/MS analysis. To desalt the peptide samples, the C18 material was washed once with 200 μl ACN, once with 200 μl StageTips buffer B [0.1% (v/v) formic acid (FA) in 50% (v/v) ACN/H_2O], and twice with 100 μl StageTips buffer A [0.1% (v/v) FA in H_2O]. Peptide samples were loaded on StageTips and washed twice with 100 μl StageTips buffer A. Finally, peptide samples were eluted with 100 μl StageTips buffer C [0.1% (v/v) FA in 40% (v/v) ACN/H_2O] and 100 μl StageTips buffer B. Solutions were passed through the StageTips via centrifugation at 500 g for 5 min at RT. Elution fractions were collected, and the solution was evaporated from peptide samples in a SpeedVac at 45 °C. Finally, 10 μl StageTips buffer A was added to the samples to perform LC-MS/MS analysis.

LC-MS/MS data acquisition

LC-MS data analysis was performed according to Gan's protocol.[61] All peptide digests were reconstituted in 0.1% (v/v) FA in H_2O and separated on reverse-phase columns [trapping column: particle size = 3 μm, C18, length = 20 mm (ThermoFisher, P/N 164535); analytical column: particle size =2 μm, C18, length =150 mm (ThermoFisher, P/N 164534)] on an Ultimate 3000 RSLCnano system (ThermoFisher) coupled to an Orbitrap Q-Exactive HF system (ThermoFisher). Peptide separation was implemented mainly using a 60-min gradient (buffer A: 0.1% FA in H_2O; buffer B: 0.1% FA in 80% ACN/H_2O) at a flow rate of 300 ml min-1 and analyzed by Orbitrap Q-Exactive HF system in a data-dependent mode. The Orbitrap Q-Exactive HF mass spectrometer was operated in positive ion spray voltage mode (2.1 kV) with ion transfer tube temperature set at 275 °C. Full-scan MS spectra (m/z 350-2000) were acquired in the Orbitrap with a resolution of 60,000. Higher collisional dissociation fragmentation was carried out at a normalized collision energy of 28%. The MS2 automatic gain control target was set to 5E4 with a maximum injection time of 50 ms, and dynamic exclusion was set to 30 s.

Chromatin immunoprecipitation qPCR (ChIP-qPCR)

ChIP-qPCR was performed according to Zhang's protocol.[62] Cultured cells were crosslinked with formaldehyde (final concentration 1%) for 10 min and quenched with glycine (125 mM) for 5 min at RT. Cells were scraped and collected by centrifugation, then washed once with PBS. Pellets were resuspended in cell lysis buffer and incubated on ice for 10 min. Cells were centrifuged, washed with MNase digestion buffer, and resuspended in MNase digestion buffer containing proteinase inhibitor cocktails (Roche, 4693132001). Lysates were incubated with MNase (NEB, M0247S) at 37°C for 20 min with continuous shaking in a thermal mixer (Fisher Scientific). The same volume of sonication buffer was added to stop the digestion. Lysates were sonicated for 3 min (10 sec on/10 sec off) using a Scientz-IID at 60 W and centrifuged at 20,000×g for 10 min at 4°C. Supernatants were collected, and chromatin content was estimated via Qubit assay (Invitrogen, Q33231). Anti-TOP1CC (Sigma, MABE1084), anti-γH2AX (CST, 9718S), anti-TOP2B (Abcam, 72334), or IgG were then added to the chromatin at 1:100 dilution and incubated on a rocker overnight at 4 °C. Next, 30 μl protein A/G-magnetic beads (MCE, HY-K0202) were added for 3 h incubation. The beads were washed twice with the following buffers: ChIP buffer (100 mM Tris-HCl pH 8.0, 20 mM NaCl, 2% TritonX-100, 0.2% sodium deoxycholate), high salt buffer (1 ml ChIP buffer, 72 μl NaCl), LiCl buffer (10 mM Tris-HCl pH 8.0, 0.25 M LiCl, 0.5% NP-40, 1 mM EDTA, 0.5% sodium deoxycholate), and 1×TE buffer (50 mM Tris-HCl, 10 mM EDTA, pH 8.0). Bound chromatin was eluted into elution buffer at 65 °C for 15 min. DNA was treated with RNase A and proteinase K to reverse cross-links and then purified using the DNA Clean & Concentrator™-5 kit (ZYMO, D4004). Primer sequences are provided in Table S2.

CUT&RUN-qPCR

CUT&RUN-qPCR was performed according to both Henikoff's and Gan's protocols.[63,64] Briefly, for each sample, 5×10^5 fresh cells were washed twice with 1 ml wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine, 1×protease inhibitor cocktail) at RT. Concanavalin-coated (ConA) magnetic beads (Bangs Laboratories, 14794) were washed twice with 1 ml binding buffer (20 mM HEPES pH 7.9, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$). After washing, cells were resuspended in 1 ml wash buffer and incubated with 15 μl pre-washed ConA beads at RT for 15 min. Beads (bound to cells) were incubated with primary antibody (1:100) (including anti-γH2AX and anti-IgG antibody) in antibody buffer (wash buffer containing 0.08% Digitonin, 2 mM EDTA and 0.1% BSA) on a rotating platform at 4 °C overnight. The beads were then permeabilized in digitonin-containing buffer (wash buffer, 0.08% digitonin), washed in the digitonin buffer, resuspended in 50 μl digitonin buffer with pA-MNase or pG-MNase and incubating for 1 h at 4 °C. After two washes with ice-cold digitonin buffer to remove unbound pA-MNase or pG-MNase, beads were subjected to digestion (after pA-MNase or pG-MNase activation) through the addition of pA-MNase or pG-MNase digestion buffer (digitonin buffer containing 2 mM $CaCl_2$), followed by incubation on a nutator for 2 h at 4 °C. Solubilized chromatin was then released using CUT&RUN stop buffer (340 mM NaCl, 20 mM EDTA, 4 mM EGTA, 0.08% digitonin, 50 μg ml^{-1} glycogen and 100 μg ml^{-1} RNase A) and incubated at 37 °C for 30 min. DNA purification was carried out using phenol-chloroform-isoamyl alcohol. The DNA products of CUT&RUN were divided,

one aliquot used as "input", and the other used for library PCR amplification and purification, yielding "IP products." Primers for specific regions were provided in Table S2.

END-Seq

Single-cell suspensions of COLO320DM (10 million), COLO320HSR (10 million), and shRNA knockdown and drug-treated (10 million) cells were prepared by washing them twice with cold PBS. Cells were resuspended in cell suspension buffer, embedded in 2% agarose, and transferred into plug molds using the CHEF Mammalian Genomic DNA Plug Kit (Bio-Rad, 1703591). Plugs were allowed to solidify at 4 °C for no more than 30 min, then incubated with Proteinase K solution for 1 h at 50 °C, followed by 7 h at 37 °C. Subsequent to incubation, the plugs underwent consecutive washes in the wash buffer (50 mM EDTA, 10 mM Tris, pH 8.0), followed by two washes in TE buffer (10 mM Tris, pH 8.0, and 1 mM EDTA). Plugs were then treated with RNase A (Qiagen, 158156) for 1 h at 37 °C, washed three times in wash buffer to remove residual RNase A, and stored at 4 °C.

Prepared plugs were washed four times with EB buffer (10 mM Tris, pH 8.0) to remove EDTA. Following the final EB buffer wash, exonuclease digestion was performed. Plugs were incubated in a reaction mix containing exonuclease VII (NEB, M0379L) at 37 °C for 1 h. Subsequently, the Exo VII reaction mix was removed, and the plugs were subjected to an exonuclease T (NEB, M0265L) reaction at 25 °C for another 1 h. Next, plugs were rinsed with EB buffer washes, followed by a dA-Tailing reaction using dA-tailing buffer (NEB, B6059S) containing 30 U Klenow fragment exo-(3'->5') (NEB, M0212L) at 37 °C for 1 h. After the dA-Tailing reaction, the residual buffer was removed, and a Quick Ligation reaction using quick ligase buffer containing END-seq adapter 1 and Quick Ligase (NEB, M2200L) was initiated at 25 °C for 1 h with shaking at 400 rpm and tubes tapped every 20 min. To eliminate any unligated and free END-seq adapter, thereby reducing primer-dimer formation in subsequent PCR reactions, plugs were transferred to a 50 ml tube with 45 ml plug wash buffer and mixed minimally for 8 h or overnight in a thermomixer at 23 °C with intermittent mixing (10 seconds at 450 rpm followed by 15 min at 0 rpm).

Four washes were performed with agitation using TE buffer. Next, each plug was transferred individually with a disposable spatula to an empty 1.5 ml tube. The agarose plugs were melted by incubating the tubes in a water bath or heat block at 70 °C for 2–3 min. The tubes were then immediately transferred to a 43 °C water bath and incubated for 5 min. Subsequently, 1.5 μl β-agarase enzyme (NEB, M0392S) was added to the center of the melted plug in each tube, and the solution was gently pipetted up and down 3–5 times using a P200 wide-bore tip. The tubes were then incubated at 43 °C for 45 min.

For drop dialysis, 15 ml TE buffer was pipetted into a 6-cm Petri dish for each sample, and a 0.1 μm dialysis membrane was placed upon the TE buffer to hydrate for 10 min. The β-agarase digested DNA sample solution was transferred to the center of the hydrated dialysis membrane for 1 h at room temperature. Following dialysis, 1 μl 10% SDS and 4 μl proteinase K (Qiagen, 158146) (20 mg ml^{-1}) were added to the dialyzed DNA samples and incubated for 15 min at 50 °C. Each DNA sample was then sheared to a median size of 175 bp using the following cycle settings on a Covaris S220 series ultrasonicator: duty 10%, intensity 175, cycles/burst 200, 240 seconds, 4–7 °C. Sheared DNA was transferred to a new tube and TE buffer was added to a total volume of 200 μl. After adding 1 μl glycogen and 20 μl sodium acetate (3 M, pH 5.2), the solution was vortexed well, and 500 μl 100% ethanol was added, then the mixture was incubated on dry ice for 15 min. The chilled samples were pelleted by spinning at maximum speed for 20 min at 4 °C in a microcentrifuge. The supernatant was decanted, and the pellet was washed twice with 1 ml 70% ethanol, and dissolved with 70 μl TE buffer by incubating at 50 °C for 5 min. DNA concentration was measured with a NanoDrop spectrometer.

Pre-washed Dynabeads MyOne C1 beads were reconstituted in 70 μl 2× WB buffer (0.2% Tween-20, 10 mM Tris-HCl, 2 M NaCl, 2 mM EDTA) and then added to 70 μl DNA sample. The mixture was incubated in a thermomixer at 700 rpm for 30 min. Following incubation, the beads were washed three times with 1 ml 1× WB buffer, twice with 1 ml EB buffer (10 mM Tris, pH 8.0), and once with 1 ml 1× NEB T4 Ligase Buffer (NEB, B0202S). Beads were then resuspended in 50 μl end repair reaction buffer, which contains 5 μl 10× NEB T4 DNA ligase buffer, 2 μl 10 mM dNTPs (KAPA:KN1009), 2 μl End Repair Enzyme Mix [1:5:5 Klenow fragment (NEB, M0210S):T4 DNA polymerase (NEB, M0203L):T4 DNA Polynucleotide Kinase (NEB, M0201L) and 41 μl nuclease free water], and incubated for 30 min at 24 °C with shaking at 700 rpm in the thermomixer. To remove residual reaction buffer, the beads were washed once with 1 ml 1× WB buffer, twice with 1 ml EB buffer, and once with 1 ml 1× dA-Tailing Reaction Buffer. Then dA-Tailing reaction was initiated by incubating beads with 50 μl dA-tailing Reaction Mix for 30 min at 37 °C with shaking at 700 rpm in the thermomixer. Subsequently, beads were washed once with 1 ml 1× NEB Buffer 2 (NEB, B7002S) in preparation for the ligation of END-Seq Adapter 2. Beads were resuspended in 115 μl ligation mix and incubated at 25 °C with shaking at 700 rpm in the thermomixer for 20 min. The ligation reaction was terminated by adding 12 μl 0.5 M EDTA, and the beads were then washed three times with 1 ml 1× WB and three times with 1 ml of EB to prevent the subsequent reaction from being disturbed by EDTA. The beads were resuspended in 18 μl USER reaction buffer, which contains 8μL EB buffer, 9 μl 2× Kapa HiFi HotStart Ready Mix (ROCHE, 7958935001) and 1 μl of USER enzyme (NEB, M5505S), and incubated for 30 min at 37 °C with shaking at 700 rpm in the thermomixer. Next, library PCR was performed using standard Illumina Nextera Dual Indexing primers (Vazyme, N323). Samples were pooled and sequenced using paired-end (2×150 bp) sequencing on an Illumina Novaseq or HiSeq X Ten platform by Annoroad Gene Technology Co. Ltd.

Whole genome sequencing (WGS)

After complete digestion (2 h) with proteinase K (0.5 mg ml^{-1}) and RNase A (0.2 mg ml^{-1}) in PBS at 65 °C, genomic DNA was extracted from cells with phenol:chloroform:isoamyl alcohol (25:24:1, PH 7.8) and precipitated with ethanol. Genomic DNA was resuspended in TE buffer and sonicated (Covaris M220) in AFA milliTUBEs (Covaris. 520045) to obtain 200–500 bp fragments. The sonication

parameters were: PIP 50, duty factor 10%, cycles 200, processing time 170 s. Then, DNA was purified using the DNA Clean & Concentrator™-5 (ZYMO, D4004) kit. WGS libraries were prepared with the fragmented DNA (1 μg) using the VAHTS® Universal DNA Library Prep Kit for Illumina V3 (Vazyme, ND607). WGS libraries were sequenced for at least 50 G data with PE150 on a NovaSeq 6000.

QUANTIFICATION AND STATISTICAL ANALYSIS

Label-free MS/MS analysis

Raw data were processed with MaxQuant (version 1.6.10.43)[51] and its built-in Andromeda search engine for feature extraction, peptide identification, and protein inference. The human reference proteome from UniProt Database (UniProtKB/Swiss-Prot and UniProtKB/TrEMBL, version 2021_04) combined with manually annotated contaminants were applied to search the peptides and proteins. The false discovery rate (FDR) values were set to 0.01, and the match-between-runs algorithm was enabled. After the search, reverse hits and proteins only identified by one site were removed. Filtered results were exported to the R package DEP (v1.18.0)[65] for quality control and analysis. The plot_pca function was used to do principal component analysis for repeatability estimates, the filter_missval function was used to filter on missing values, normalize_vsn function was used to normalize the minimum-filled data, and the test_diff function was used to conduct differential enrichment analysis of proteins identified in ecDNA+/ecDNA-. Enrichment analysis for GO terms associated with biological processes was performed using MetaScape (v3.5).[52] Significantly enriched GO terms were identified using the Fisher classic algorithm. The interaction network was implemented in Cytoscape (version 3.9.1) using links between identified proteins downloaded from the STRING database (v11.5). All Venn diagrams and heatmaps were generated using the TBtools software (v1.120).[54]

TCGA clinical data source

We extracted tumor sample data from the TCGA database, which were identified by Kim et al. and they divided these samples into two groups: samples containing ecDNA (ecDNA+) and samples without ecDNA (ecDNA-).[2] To further expand the clinical data sample size, we defined ecDNA- samples as those with copy numbers less than 4 for 21 commonly amplified oncogenes (*MDM2*, *MYC*, *EGFR*, *CDK2*, *ERBB2*, *SOX2*, *TERT*, *CCND1*, *E2F3*, *CCNE1*, *CDK6*, *MDM4*, *NEDD9*, *MCL1*, *AKT3*, *BCL2L1*, *ZNF217*, *LRAS*, *PDGFRA*, *AKT1*, *MYCL*).[2] We kept tumor types with more than 4 ecDNA+ and ecDNA- samples, and samples with solid expression data for subsequent analyses. As a result, we obtained ecDNA+ and ecDNA- samples across the following tumor types: glioblastoma (GBM), sarcoma (SARC), esophageal (ESCA), ovarian (OV), bladder (BLCA), lung squamous (LUSC), head and neck (HNSC), breast (BRCA), gastric (STAD), cervical (CESC), skin (SKCM), lung adeno (LUAD), lower grade glioma (LGG), and uterine corpus endometrial (UCEC). All TCGA data were sourced from UCSC Xena (https://xenabrowser.net/). Gene expression data was obtained from the processed RNA-seq data for the corresponding tumor in the GDC HUB: HTSeq-FPKM.

DNA damage response (DDR) protein analysis

DDR core pathways and gene sets were obtained from published research (https://gdc.cancer.gov/about-data/publications/PanCan-DDR-2018).[37] DDR scoring is a method to quantify the tumor signature of DDR pathways and was calculated here based on the method reported previously.[66] We used 20 genes provided (*POLQ*, *EXO1*, *EME1*, *BRIP1*, *FANCA*, *BLM*, *FANCI*, *BRCA2*, *XRCC2*, *RAD51*, *DLX5*, *UBE2T*, *BRCA1*, *FANCB*, *FANCD2*, *FEN1*, *XRCC3*, *BARD1*, *GEN1* and *PRKDC*) as a tumor-related core DDR gene set.[66] Gene Set Variation Analysis (GSVA) analysis was implemented in R package GSVA (v1.38.2)[67] to calculate DDR scores based on this core DDR gene set for each TCGA sample.

Whole genome sequencing (WGS) analysis

Samples underwent WGS library preparation and were then sequenced on an Illumina NovaSeq 6000 at 2×150 bp and 20× coverage. Raw reads in fastq files were trimmed of adapter content with Trimmomatic (v.0.39)[55] and aligned to hg38 using BWA-MEM (v0.7.17) with the default parameters. Only paired aligned reads were retained for subsequent analysis. Duplicated reads were marked using Picard MarkDuplicates (v2.18.29) (http://broadinstitute.github.io/picard/).

Amplicon calling and classifying

AmpliconArchitect (AA)[41] was used to detect the copy-number aware breakpoint graph from WGS data. AA was run on the aligned reads and seed regions identified using the PrepareAA pipeline (v.0.931.0, https://github.com/jluebeck/PrepareAA) with the CNVKit (v.0.9.6).[46] The default parameters were used as described in the documentation (https://github.com/virajbdeshpande/AmpliconArchitect). Then, amplicon classification was implemented using AmpliconClassifier (AC), the tool classifies the AA outputs and predicts types of focal amplifications. To maximize efficiency, we utilize the AmpliconSuite-pipeline (v1.3.3),[42] a multithread-enabled end-to-end wrapper for AA and AC to enable analysis of focal copy number amplifications such as ecDNA or breakage-fusion-bridges (BFBs) from paired-end WGS data, with the default parameters described in the documentation (https://github.com/AmpliconSuite/AmpliconSuite-pipeline), except for "–AA_insert_sdevs 9" for avoiding the influence of poorly-controlled insert size. Following this workflow, we obtained copy number variations (CNVs), structural variations (SVs), classification of amplicons, intervals and segments to build an amplicon graph, which is determined as ecDNA amplified location (in circle loci).

END-seq analysis

In this study, we used END-seq, a highly sensitive next-generation sequencing technique for evaluating DNA DSB levels across the genome.[68] All samples were pooled and sequenced with paired-end (2 × 150 bp) sequencing on an Illumina NovaSeq 6000. All raw data were trimmed with Trimmomatic (v.0.39)[55] and aligned to reference genome hg38 using bowtie2 (v.2.4.2).[45] Only paired aligned reads were retained for subsequent analysis. Uniquely aligned reads were extracted using Samtools (v.1.7),[53] and reads from positive and negative strand were separated. END-seq peaks were detected using MACS2 (v.2.2.9.1)[50] with a *P*-value cutoff of 0.001. Fragments per kilobase million (FPKM) for peak regions were calculated with custom Perl (v5.32.1) scripts, and peaks with FPKM less than 1 were filtered out. Bedtools (v.2.29.2)[44] was used to generate BigWig files and counting matrices in bins around peak summits, and bedGraphToBigWig (v.4)[43] was used for visualization. The Homer suite (v.2.0)[48] was used to annotate peaks.

The relative enrichment ratios of annotated peaks for COLO32DM cells were calculated using the formula:

$$\text{Enrichment ratio} = log_2\left(\frac{\text{obs}}{\text{exp}}\right)$$

where "obs" is the observed total length of annotated peaks of certain types, including promoter, 5' UTR, exon, intron, transcription termination sites (TTS), 3' UTRs and intergenic regions. The "exp" is the corresponding experimental length of peaks of certain types, evaluating using the equation:

$$exp = \frac{\text{experimental length of peaks}}{\text{length of whole genome}} \times \text{length of target peaks}$$

where "target peaks" are the annotated peaks identified using END-seq data.

Signal quantification was implemented mainly using custom Perl (v.5.26.2) scripts. These scripts calculated normalized densities of END-seq signals for peak regions (± 200 bp from peak center) in which ecDNA amplified (in circle loci) and other (not in circle loci), with a scan bin size of 100 bp. The signal of END-seq peak was first normalized using the read depth from global sequencing, then further normalized using the copy number of WGS. The specific locations of ecDNA amplification and the copy number of the corresponding regions were identified using WGS data described above.

Colocalization Analysis

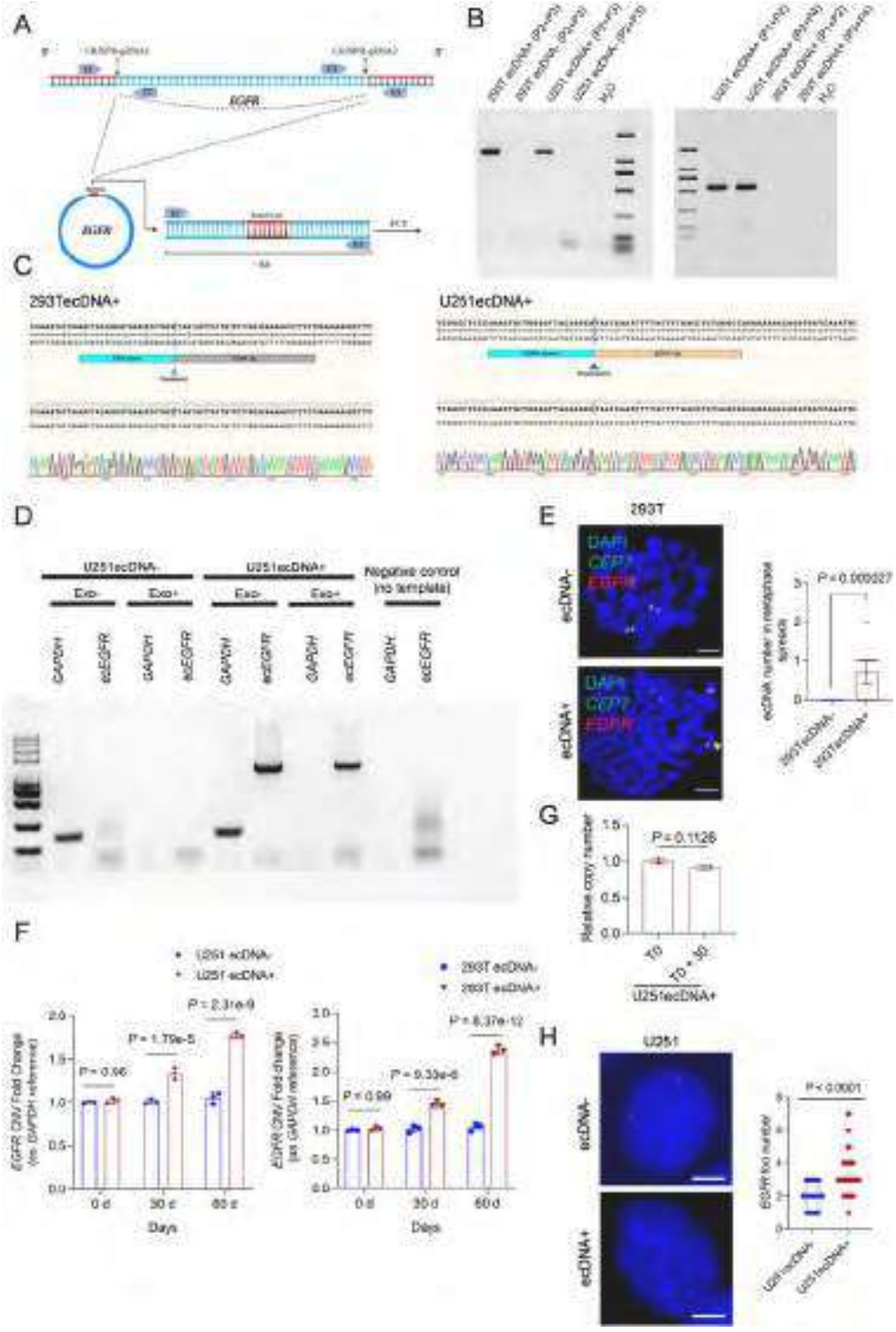
A Foci Analyzer tool (https://github.com/BioImaging-NKI/Foci-analyzer) was used to automatically identify the cell nucleus, γH2AX foci, and ecDNA foci. Parameters, including the nuclear area, the total number of γH2AX and ecDNA foci, the average size of both types of foci, and the overlap area between ecDNA and γH2AX foci, were calculated. To estimate the overlap area between γH2AX and ecDNA foci by chance, we employed a method involving random placement. Fluorescence image profile and co-localization analysis were performed by ImageJ (v150).[49]

Statistics and reproducibility

Student's *t*-test (for two groups) and ANOVA (for three or more groups) were used to test the mean difference. Statistics for box and violin plots were shown with median, upper and lower quartiles. To compare the gene expression levels, and DDR scores for ecDNA+ and ecDNA- TCGA samples, we used the unpaired two-tailed Mann-Whitney test for the *P*-value calculations. Information about the detailed statistical tests used in each experiment is provided in the figure legends. All experiments were performed with three biological replicates unless otherwise stated in the figure legends.

CellPress

Supplemental figures



(legend on next page)

Figure S1. ecDNA replicates autonomously and can be maintained in cultured cells, related to Figure 1

(A) Schematic of CRISPR-C strategy used to generate a single ecDNA in U251 and 293T cells containing *ecEGFR*.

(B) PCR detection of *EGFR* sequence cyclization after cutting with CRISPR-C using agar gel electrophoresis.

(C) Sanger sequencing confirming *EGFR* sequence cyclization with 5^0 UTR connected to 3^0 UTR sequences.

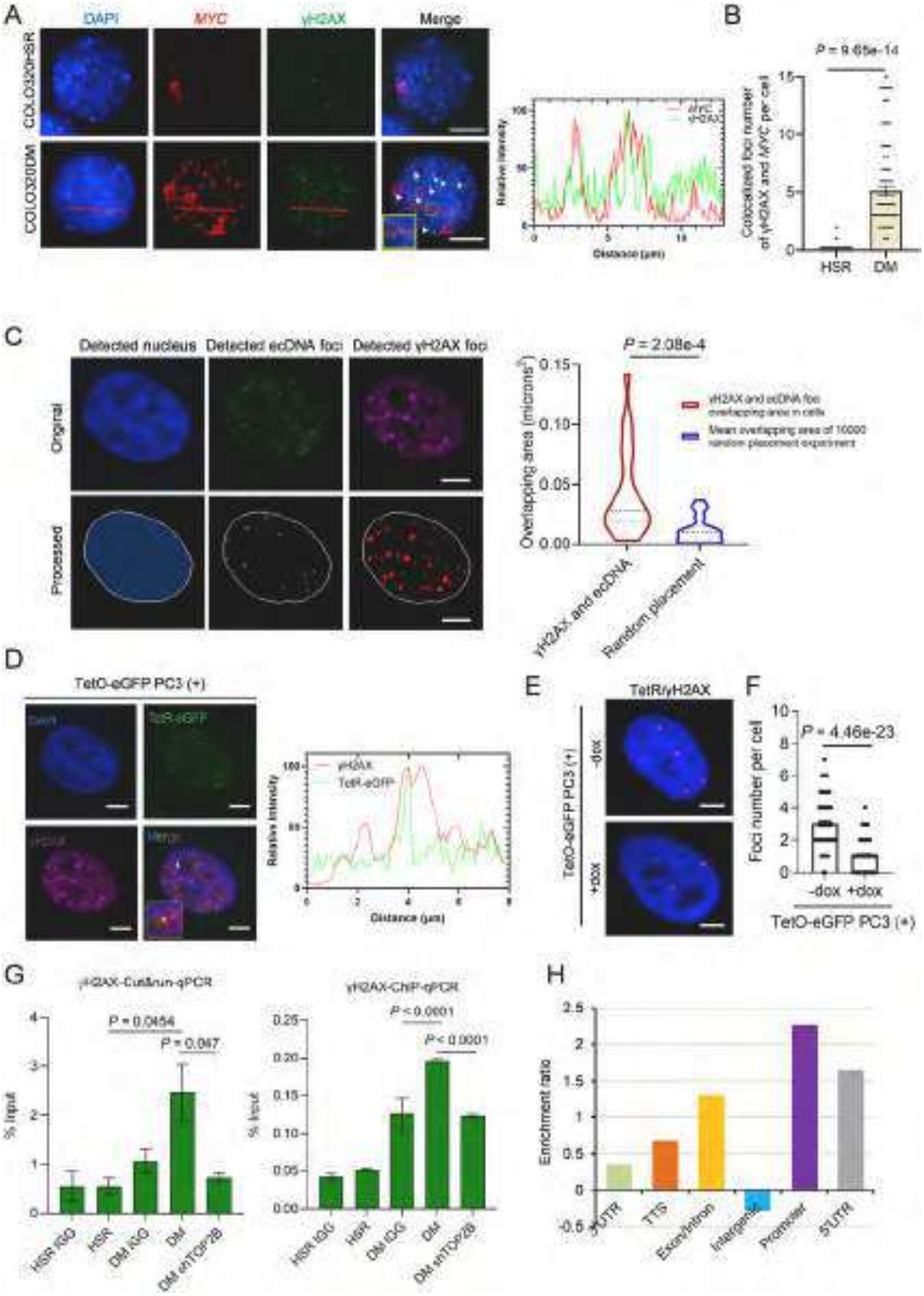
(D) Gel images of corresponding PCR products across junction site (containing *EGFR*) using DNA templates before (Exo−) and after (Exo+) exonuclease treatment for 18 h. DNA isolated from U251ecDNA− cells was used as negative control. *GADPH* was used as indicator of linear DNA.

(E) Representative metaphase FISH images and quantifications for *EGFR* probe in ecDNA+ 293T cells. Chromosomal control was included. Scale bar, 10 μm. Yellow triangle indicates colocalized EdU and *EGFR* foci.

(F) Quantifications of relative *EGFR* copy number normalized to *GAPDH* in 293T and U251 cell lines at three time points. Mean values ± SD are presented with *p* value determined using a two-tailed Student's t test. CNV, copy-number variant.

(G) Quantifications of relative *EGFR* copy number, normalized to *GAPDH*, in U251ecDNA+ cells were performed at a time point T0 and 30 days later (T0 + 30). Mean values ± SD are presented with *p* value determined using a two-tailed Student's t test.

(H) Representative metaphase FISH images and quantifications of *EGFR* foci numbers in metaphase FISH images for *EGFR* probe using ecDNA+ U251 cells. Mean values ± SD are presented with *p* value determined using a two-tailed Student's t test.



(legend on next page)

Figure S2. DSB and γH2AX are enriched on ecDNA, related to Figure 2

(A) Representative images showing immunofluorescence staining with γH2AX antibody and FISH staining with *MYC* probes in COLO320DM and COLO320HSR cells. Nucleus is stained with DAPI. Scale bar, 5 μm. The red lines indicate regions used for fluorescence image profile analysis.

(B) Quantification of colocalized γH2AX and *MYC* foci per cell in COLO320DM and COLO320HSR from (A). Mean values ± SEM are presented with *p* value determined using a two-tailed Student's t test.

(C) Illustration of the nucleus and ecDNA and γH2AX foci detected for Figure 2D. The *p* value for the difference in overlapping area between ecDNA and γH2AX foci observed in cells, compared with what would be expected if colocalization was random (mean overlapping area from 10,000 randomly placed of foci), is shown on the right.

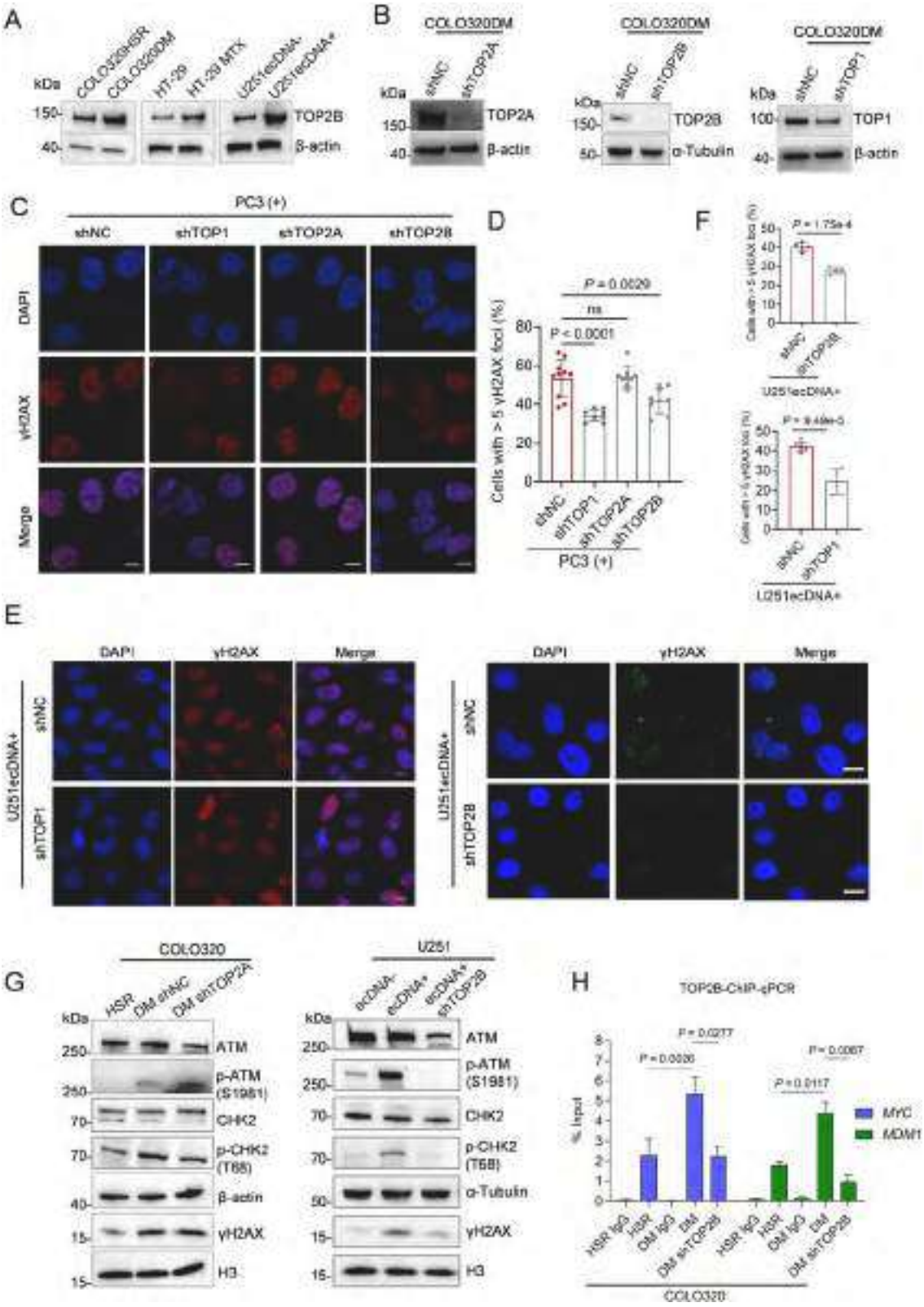
(D) Fluorescence image profile analysis showing colocalized signals in TetO-eGFP PC3 (+) for Figure 2D.

(E) PLA detection of interaction between TetR and γH2AX in TetO-eGFP PC3 (+) cells, with or without doxycycline (dox) treatment. Nucleus stained with DAPI. Scale bar, 5 μm.

(F) Quantification of PLA signals between TetR and γH2AX in nucleus in TetO-eGFP PC3 (+) cells, with or without 10 μM doxycycline (dox) treatment for 96 h, from (E). Mean values ± SEM are presented with *p* value determined using a two-tailed Student's t test.

(G) γH2AX-Cut&run-qPCR and γH2AX-ChIP-qPCR at selected regions in COLO320DM, shTOP2B COLO320DM, and COLO320HSR cells. Mean values ± SD are shown, with *p* values calculated using a two-tailed Student's t test.

(H) Integration analysis showing which types of genomic regions were enriched for END-seq peaks in COLO320DM cells. UTR, untranslated regions; TTS, transcription termination sites.



(legend on next page)

Figure S3. Topoisomerases promote ATM-mediated activation of DDR in ecDNA+ cells, related to Figure 3

(A) Immunoblots showing TOP2B levels in ecDNA− and ecDNA+ cells, with β-actin used as loading control.

(B) Immunoblots showing shRNA knockdown efficiency for TOP2A (shTOP2A), TOP2B (shTOP2B), and TOP1 (shTOP1) in COLO320DM cells, with β-actin or α-tubulin used as loading control.

(C) Representative immunofluorescence images of γH2AX signals in shNC, shTOP1, shTOP2A, and shTOP2B PC3 (+) cells. Nucleus stained with DAPI. Scale bar, 10 μm.

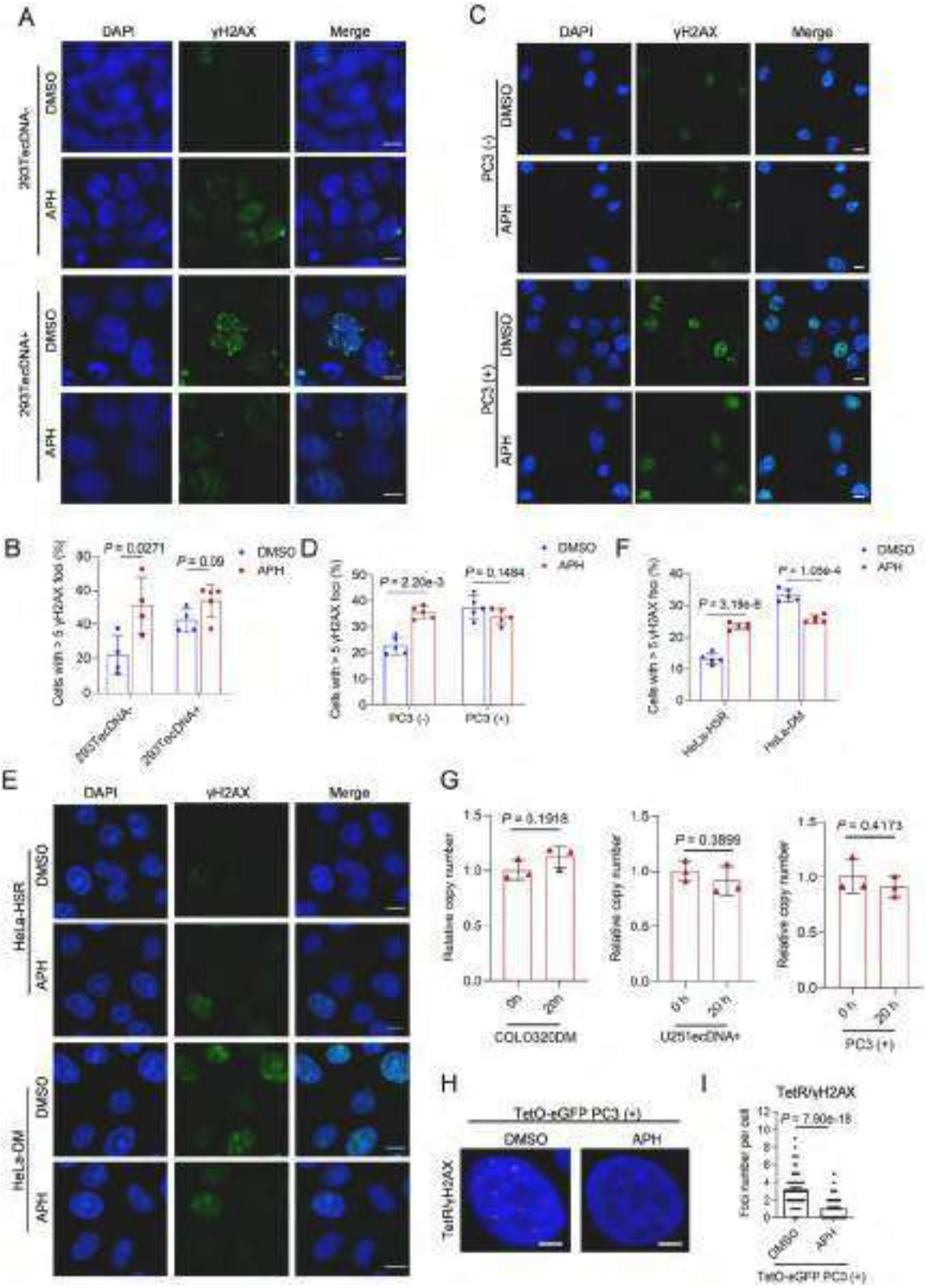
(D) Quantification of percentage of cells with more than five γH2AX foci in shNC, shTOP1, shTOP2A, and shTOP2B PC3 (+) cells from (C). Mean values ± SD presented with *p* values determined using a one-way ANOVA.

(E) Representative immunofluorescence images of γH2AX signals in shNC, shTOP2B, and shTOP1 U251ecDNA+ cells. Nucleus stained with DAPI. Scale bar, 10 μm.

(F) Quantification of percentage of cells with more than five γH2AX foci in shNC, shTOP2B, and shTOP1 U251ecDNA+ cells from (E). Mean values ± SD presented with *p* values determined using a two-tailed Student's t test.

(G) Representative immunoblots of COLO320 cells (COLO320HSR, shNC COLO320DM, and shTOP2A COLO320DM) and U251 cells (U251ecDNA−, U251ecDNA+, and shTOP2B U251ecDNA+) were analyzed with indicated antibodies. Molecular weight (MW) in kDa (left).

(H) Results of ChIP for TOP2B or immunoglobulin G (IgG) followed by qPCR at *MYC* and *MDM1* in COLO320HSR, COLO320DM, and shTOP2B COLO320DM cells. Mean values ± SEM with *p* values determined using a one-way ANOVA.



(legend on next page)

Figure S4. ecDNA replication triggers activation of DDR in ecDNA+ cells, related to Figure 5

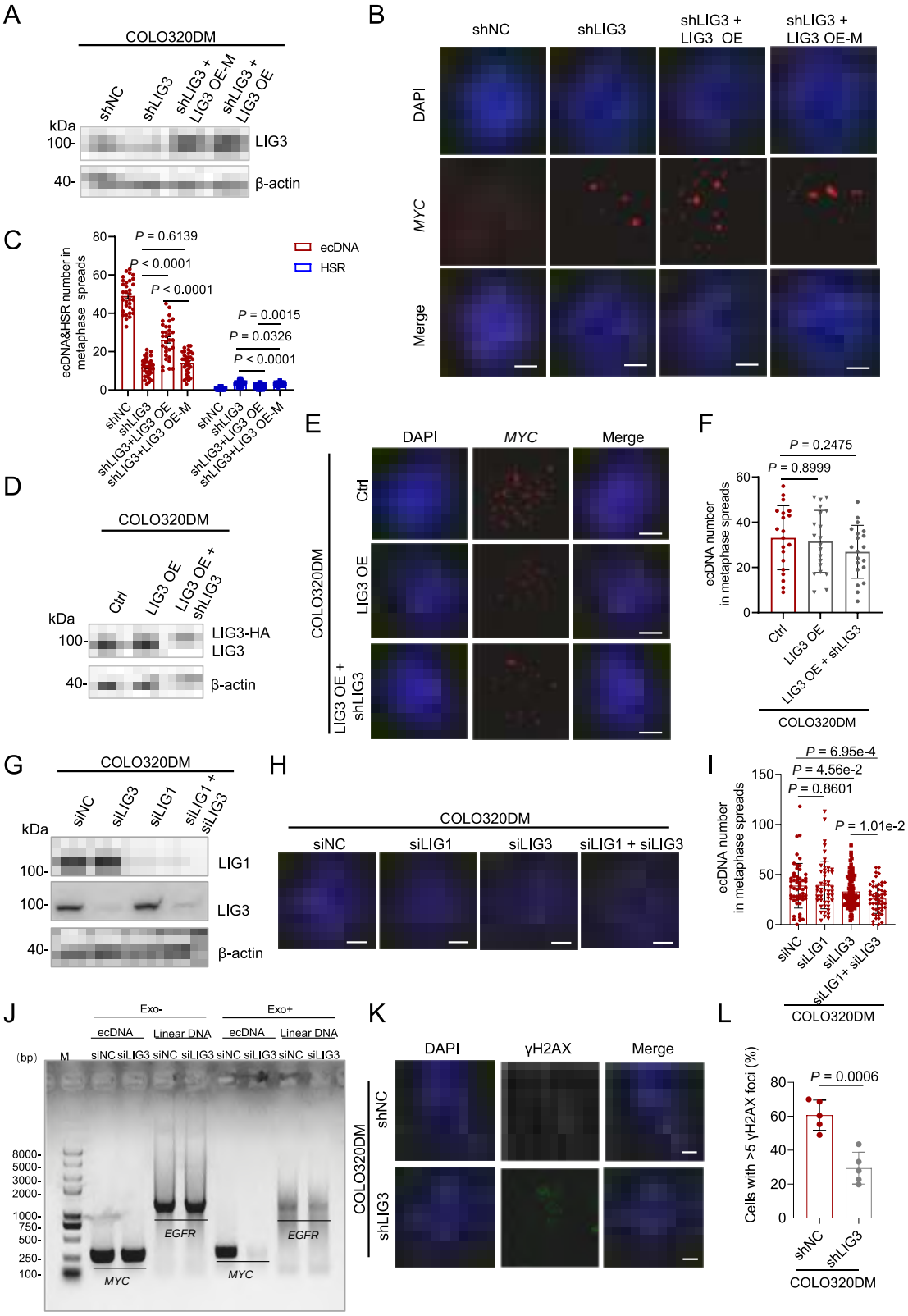
(A, C, and E) Representative immunofluorescence images of γH2AX signals in 293TecDNA−/293TecDNA+ (A), PC3 (−)/PC3 (+) (C), and HeLa-HSR/HeLa-DM cells (E) after treatment with 2 μM APH for 20 h. Nucleus stained with DAPI. Scale bar, 10 μm.

(B, D, and F) Quantification of percentage of cells with more than five γH2AX foci in 293TecDNA−/293TecDNA+ (B), PC3 (−)/PC3 (+) (D), and HeLa-HSR/HeLa-DM cells (F) after treatment with 2 μM APH for 20 h. Mean values ± SD presented with *p* values determined using a two-tailed Student's t test.

(G) Quantifications of relative *MYC* copy numbers, normalized to GAPDH, in ecDNA+ cells before and after treatment with 2 μM APH for 20 h. Mean values ± SD are presented, with *p* values determined using a two-tailed Student's t test.

(H) *In situ* PLA detection of interaction between TetR and γH2AX in APH-treated TetO-eGFP PC3 (+) cells. Nucleus stained with DAPI. Scale bar, 5 μm.

(I) Quantification of PLA foci number localized in the nucleus of APH-treated TetO-eGFP PC3 (+) cells from (H). Mean values ± SEM are presented with *p* value determined using a two-tailed Student's t test.



(legend on next page)

Figure S5. LIG3 is coupled to ecDNA maintenance, related to Figure 6

(A) Immunoblots showing rescue efficiency of LIG3 in COLO320DM cells transfected with HA-tagged full-length LIG3 expression vectors (OE) or ligase activity-null LIG3 (K508V) expression vectors (OE-M), with β-actin used as loading control.

(B) Representative metaphase FISH images of *MYC* probes in LIG3-rescued shLIG3 COLO320DM cells. Scrambled shRNA (shNC) COLO320DM cells were used as controls. Scale bar, 10 μm.

(C) Quantification of ecDNA and HSR foci number per spread in LIG3-rescued shLIG3 COLO320DM cells from (B). shNC COLO320DM cells were used as controls. Mean values ± SEM presented with *p* values determined using a one-way ANOVA.

(D) Immunoblots showing LIG3 OE and LIG3 OE + shRNA LIG3 knockdown pattern in COLO320DM cells, with β-actin used as loading control.

(E) Representative metaphase FISH images for *MYC* probes using COLO320DM cells, LIG3 OE COLO320DM cells, and LIG3 OE + shLIG3 knockdown COLO320DM cells. Scale bar, 10 μm.

(F) Quantification of ecDNA foci number in metaphase FISH images in COLO320DM cells, LIG3 OE COLO320DM cells, and LIG3 OE + shLIG3 knockdown COLO320DM cells from (E). Mean values ± SD are presented with *p* values determined using a one-way ANOVA.

(G) Immunoblots showing siRNA knockdown efficiency for LIG1 (siLIG1), LIG3 (siLIG3), or LIG1 + LIG3 (siLIG1 + siLIG3) in COLO320DM cells, with β-actin used as loading control. Scrambled siRNA (siNC) COLO320DM cells created for use as controls.

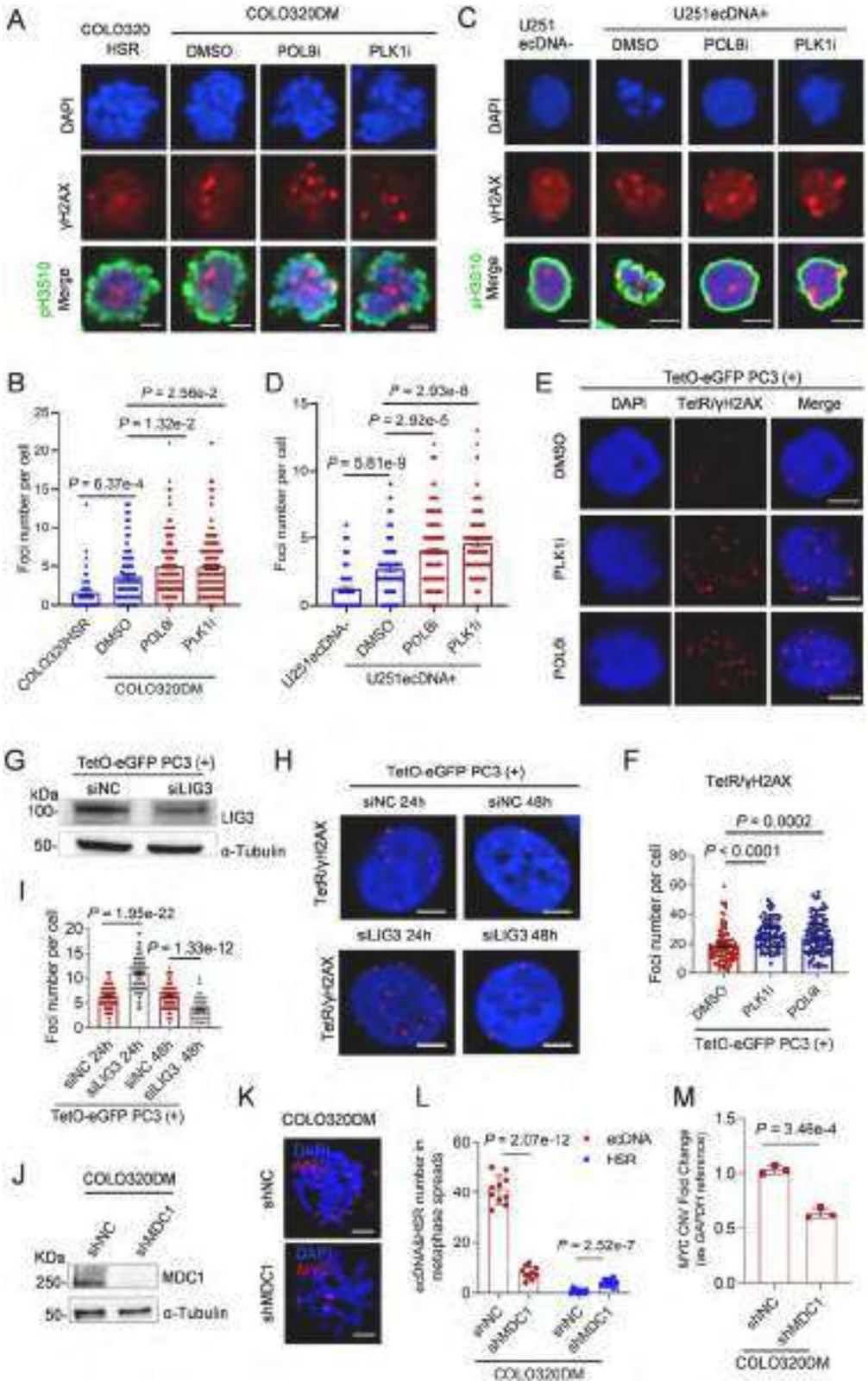
(H) Representative metaphase FISH images of *MYC* probes in LIG1 and LIG3 single or double siRNA knockdown COLO320DM cells. Scale bar, 10 μm.

(I) Quantification of ecDNA foci number in metaphase FISH images in LIG1 and LIG3 single or double siRNA knockdown COLO320DM cells from (H). Mean values ± SD presented with *p* values determined using a two-tailed Student's t test.

(J) Gel images of PCR products using equal amounts of DNA template isolated from COLO320DM cells transfect with siLIG3 and siNC. DNA templates were treated with (Exo+) or without (Exo−) exonuclease to remove linear DNA. *EGFR* used as indicator of linear DNA.

(K) Representative immunofluorescence images showing γH2AX signals in shNC and shLIG3 COLO320DM cells. Nucleus stained with DAPI. Scale bar, 10 μm.

(L) Quantification of percentage of cells with more than five γH2AX foci in shNC and shLIG3 COLO320DM cells from (K). Mean values ± SD presented with *p* value determined using a two-tailed Student's t test.



(legend on next page)

Figure S6. Critical molecules that safeguard genome stability during mitosis are crucial for ecDNA maintenance, related to Figure 6

(A) Representative immunofluorescent images for γH2AX signals in mitotic COLO320HSR and COLO320DM cells treated inhibitors targeting POLθ (POLθi) and PLK1 (PLK1i) for 24 h. pH3S10 used to mark mitotic nuclei. DNA stained with DAPI. Scale bar, 5 μm.

(B) Quantification of γH2AX foci number in mitotic COLO320HSR and COLO320DM cells from (A) treated with POLθi and PLK1i. Mean values ± SD are presented with *p* values determined using a one-way ANOVA.

(C) Representative immunofluorescent images for γH2AX signals in mitotic U251ecDNA− and U251ecDNA+ cells treated with POLθi and PLK1i. pH3S10 used to mark mitotic nuclei. DNA stained with DAPI. Scale bar, 10 μm.

(D) Quantification of γH2AX foci number in mitotic U251ecDNA− and U251ecDNA+ cells from (C) treated with POLθi and PLK1i. Mean values ± SD are presented with *p* values determined using a one-way ANOVA.

(E) PLA detection of interaction between TetR and γH2AX in mitotic TetO-eGFP PC3 (+) cells treated with 10 μM ART558 (a POLθ inhibitor, POLθi) or 500 nM volasertib (a PLK1 inhibitor, PLK1i) for 6 h. Scale bar, 10 μm.

(F) Quantification of PLA foci number for tested interactions between TetR and γH2AX using TetO-eGFP PC3 (+) cells treated with PLK1i and POLθi from (E). Mean values ± SEM presented with *p* value determined using a one-way ANOVA.

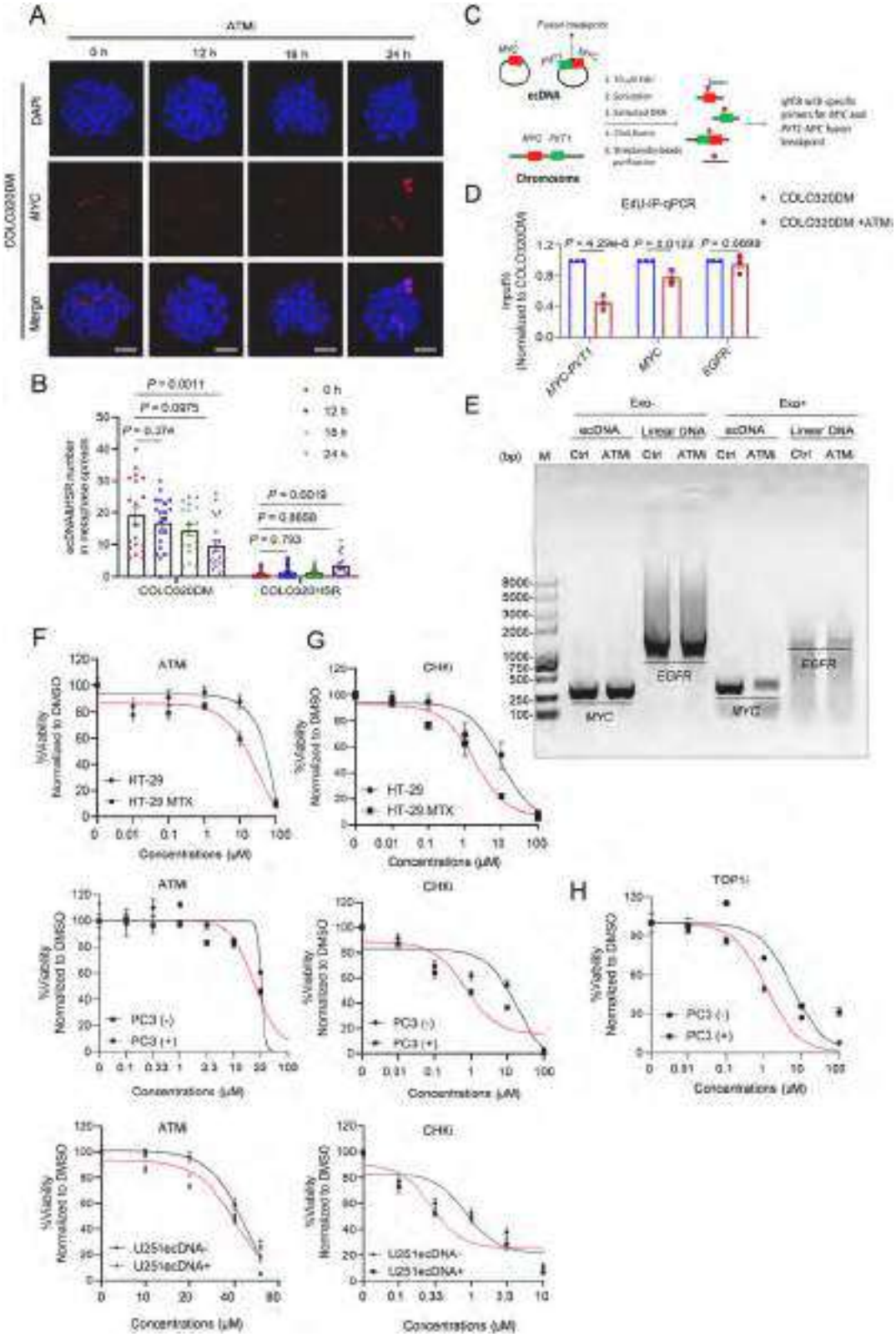
(G) Immunoblot showing the knockdown efficiency of LIG3 (siRNA3) in TetO-eGFP PC3 (+) cells via siRNA for 24 h, α-tubulin used as loading control. Scrambled shRNA (shNC) TetO-eGFP PC3 (+) cells created for use as controls. MW in kDa (left).

(H) *In situ* PLA detection of interaction between TetR and γH2AX in TetO-eGFP PC3 (+) cells with LIG3 knockdown via siRNA (siLIG3) for 24 or 48 h. Scale bar, 5 μm.

(I) Quantification of PLA foci number for tested interactions between TetR and γH2AX using siLIG3 TetO-eGFP PC3 (+) cells from (H). Mean values ± SEM presented with *p* value determined using a two-tailed Student's t test.

(J) Immunoblot showing MDC1 shRNA knockdown efficiency in COLO320DM cells, α-tubulin used as loading control. Scrambled shRNA (shNC) COLO320DM cells created for use as controls.

(K–M) Representative metaphase FISH images of *MYC* (K), quantification of ecDNA and HSR foci number (L), and relative copy number of ecDNA (M) in shNC and shMDC1 COLO320DM cells. Scale bar, 10 μm. Mean values ± SD presented with *p* values determined using a two-tailed Student's t test.



(legend on next page)

Figure S7. ecDNA+ cells are more sensitive to targeted ATM pathway and TOP1 inhibition than ecDNA− cells, related to Figure 7

(A) Representative metaphase FISH images of *MYC* probes in COLO320DM cells treated with the ATMi for various times. Scale bar, 10 μm.

(B) Quantification of ecDNA and HSR foci number in ATMi-treated COLO320DM cells from (C). Mean values ± SEM presented with *p* values determined using a one-way ANOVA.

(C) Schematic depicting EdU-IP-qPCR strategy used to detect ecDNA replication.

(D) Relative ecDNA replication rates assessed by EdU-IP-qPCR in COLO320DM cells treated with an ATM inhibitor (ATMi). *MYC-PVT1* and *MYC* were used as indicators of ecDNA. Chromosomal *EGFR* was used as a negative control. Mean values ± SEM presented with *p* values determined using a two-tailed Student's t test.

(E) Gel images of PCR products using equal amounts of template DNA isolated before exonuclease treatment (Exo−) from cells with and without ATMi. Exonuclease was added (Exo+) to remove linear DNA. *EGFR* was used as an indicator of linear DNA.

(F and G) Dose-response curves showing increased sensitivity to (D) ATMi (KU55933, 48 h) and (E) CHKi (AZD7762, 48 h) in ecDNA+ versus ecDNA− cells. Mean values ± SEM are presented.

(H) Dose-response curve showing increased sensitivity to TOP1i (CPT, 48 h) in ecDNA+ versus ecDNA− cells. Mean values ± SEM are presented.

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生
前沿交叉学科学术论坛论文成果

姓名：刘伟松
方向：生物制造
单位：中国科学院天津工业生物技术研究所

题目：生物电催化二氧化碳还原

nature communications

Article

https://doi.org/10.1038/s41467-024-53946-3

Bioelectrocatalytic carbon dioxide reduction by an engineered formate dehydrogenase from *Thermoanaerobacter kivui*

Received: 17 April 2024

Accepted: 29 October 2024

Published online: 17 November 2024

Check for updates

Weisong Liu[1,2,3], Kuncheng Zhang[1,2,3], Jiang Liu[3], Yuanming Wang[3], Meng Zhang[3,4], Huijuan Cui[1,2,3], Junsong Sun[2,5] & Lingling Zhang[1,2,3]

Electrocatalytic carbon dioxide (CO_2) reduction by CO_2 reductases is a promising approach for biomanufacturing. Among all known biological or chemical catalysts, hydrogen-dependent carbon dioxide reductase from *Thermoanaerobacter kivui* (*Tk*HDCR) possesses the highest activity toward CO_2 reduction. Herein, we engineer *Tk*HDCR to generate an electro-responsive carbon dioxide reductase considering the safety and convenience. To achieve this purpose, a recombinant *Escherichia coli Tk*HDCR overexpression system is established. The formate dehydrogenase is obtained via subunit truncation and rational design, which enables direct electron transfer (DET)-type bioelectrocatalysis with a near-zero overpotential. By applying a constant voltage of −500 mV (*vs.* SHE) to a mediated electrolytic cell, 22.8 ± 1.6 mM formate is synthesized in 16 h with an average production rate of 7.1 ± 0.5 μmol $h^{-1}cm^{-2}$, a Faradaic efficiency of 98.9% and a half-cell energy efficiency of 94.4%. This study provides an enzyme candidate for high efficient CO_2 reduction and opens up a way to develop paradigm for CO_2-based bio-manufacturing.

Carbon dioxide (CO_2) emissions pose significant environmental pressure, such as climate change. The effective reduction of CO_2 to fuels and chemicals is a promising approach for realizing a carbon-neutral economy[1]. Liquid products such as formic acid and methanol are easy to store and transport, and their electrochemical production has been considered an attractive technology, especially in green chemical/biological transformation[2]. The core issue in electrochemical CO_2 reduction lies in the design of electrocatalysts, including chemical electrocatalysts and bioelectrocatalysts. Chemical electrocatalysts are generally based on transition metals (e.g., Cu, Bi, Pb), and significant progress has been made in recent years. For instance, Xia et al. produced nearly 100% pure formic acid at a Pb electrode, which was derived from waste lead–acid batteries[3]. In addition, Sargent et al. reported that multi-carbon products were produced from CO_2 catalyzed by Cu-based catalysts in acidic media[4,5]. Similarly, Yang et al. elucidated the breakthrough potential of Cu-based nanograins as electrocatalysts for generating C2+ products[6]. However, chemical electrocatalysts encounter the challenges of high overpotentials and low atom economy due to the high energy barrier and inevitable side reactions. To address these issues, bioelectrocatalysts provide potential solutions. They can lower the activation energy more significantly than chemical catalysts can, thereby reducing the overpotential. Surely, enzymes are highly specific, ensuring efficiency, yield, and purity in the production of target products[7].

Formate dehydrogenases (FDHs), which catalyze the interconversion between CO_2 and formate, have been intensively studied. Nielsen et al.[8] and Moon et al.[9] comprehensively reviewed FDHs with regard to organism sources, classification, structures, and electron donors.

[1]Key Laboratory of Engineering Biology for Low-carbon Manufacturing, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, 32 west 7th Avenue, Tianjin Airport Economic Area, Tianjin, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]In vitro Synthetic Biology Center, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China. [4]College of Biotechnology, Tianjin University of Science and Technology, 9, 13th Avenue, Tianjin Economic and Technological Development Area, Tianjin, China. [5]Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China. e-mail: zhangll@tib.cas.cn

Notably, a majority of FDHs exhibit quite sluggish kinetics toward CO_2 reduction, with k_{cat} values of 0.01-1 s^{-1}. A few FDHs (e.g., FDHs from *Syntrophobacter fumaroxidans* and *Desulfovibrio vulgaris*) exhibit appreciable catalytic activity toward CO_2 reduction, they are prone to oxidize formate at relatively high catalytic rates[9]. Mo-dependent formylmethanofuran dehydrogenase exhibits preferential directionality for the reductive reaction[10]. However, the CO_2 reduction activity of formylmethanofuran dehydrogenase is 0.24 U mg^{-1}, which is insufficient for practical applications. Hydrogen-dependent carbon dioxide reductases (HDCRs) from *Acetobacter woodii* and *Thermoanaerobacter kivui* exhibit remarkable activity for CO_2 reduction, with k_{cat} values of 28 and 2654 s^{-1}, respectively, making them promising candidates for CO_2 biotransformation[11,12]. Generally, the HDCR unit is a heterohexamer consisting of a formate dehydrogenase subunit, three electron transfer subunits, and two iron-iron hydrogenase subunits. Due to the multisubunit structure and the complicated cofactors, the expression of HDCR is challenging. It was reported that the expression level of *Tk*HDCR was as low as 0.05 mg L^{-1} in its original strain. Therefore, it is desirable to establish efficient overexpression methods for sufficient *Tk*HDCR production in common expression hosts, such as *Escherichia coli*.

Compared to hydrogen as an electron donor, electricity might be a better alternative because of its safety and convenience. Significant progress has been made in electricity-driven FDH catalysis. Reda et al. reported the first example of bioelectrocatalytic formate/CO_2 interconversion based on W-containing *Sf*FDH1 from *S. fumaroxidans*[13]. Reisner et al. studied the reversible bioelectrocatalysis of Mo-containing *Ec*FDH-H from *E. coli* and utilized electrochemical and catalytic inhibition methodologies to comprehensively investigate the interconversion mechanism[14,15]. Li et al. revealed the redox chemistry of the distal Fe-S cluster in *Cl*FDH from *Clostridium ljungdahlii*, demonstrating its significant involvement in mediating electron transfer between the electrode and metal active site of the enzyme[16]. Strategies such as the employment of conductive polymers/electronic mediators and oriented immobilization were employed to improve the electron transfer efficiency. For example, conductive polyaniline (PANi) hydrogel has been explored to fix *Cl*FDH at the electrode, and a formate yield rate of 1.42 μmol h^{-1} cm^{-2} was reached[17]. An electron mediator such as benzylpropylviologen was also used to improve the rate of electron transfer from the electrode to *Ec*FDH[18]. Directed immobilization of enzymes has emerged as a direct approach to improve the efficiency of direct electron transfer. Alvarez-Malmagro et al. designed a covalently immobilized strategy for *Dv*FDH from *D. vulgaris* oriented immobilization on the electrode to improve the direct electron transfer rate[19]. Badiani et al. demonstrated that *Dv*FDH was oriented fixed through electrostatic interaction between the negative spot of the Fe-S cluster and the positively functionalized electrode with $-NHMe^{2+}$group[20]. In addition, Reisner et al. reported that the local environment of bioelectrocatalysis, including CO_2 concentration and pH, was also a key factor affecting the performance of CO_2 reduction[21,22]. They also revealed that the fast CO_2 hydration kinetics could improve enzymatic CO_2 reduction catalysis[23,24].

In this work, we study the heterologous expression of *Tk*HDCR in *E. coli*. and its rational design. The maturation of W-MGD and Fe-Fe active sites involves several chaperone proteins. Through gene identification and codon optimization, plasmids and expression hosts are optimized. To replace the hydrogen supply with electricity, HDCR is truncated to different extents, generating the FdhF_HycB3_HycB4, FdhF_HycB3, and FdhF variants. Further structural adaptation of FdhF_HycB3 is performed to generate the robust variant FdhF_HycB3Δ159-184. In formate bioelectrosynthesis, detailed bioelectrocatalytic kinetics and direct/mediated bioelectrosynthesis of FdhF_HycB3Δ159-184 are studied to understand the catalytic mechanism of FDH catalysis/interfacial electron transfer and guide electro-assisted CO_2 biotransformation (Fig. 1).

Results and discussion

Heterologous expression of *Tk*HDCR in *E. coli*

The challenges associated with the heterologous expression of *Tk*HDCR include the unclear maturation mechanism of active sites and

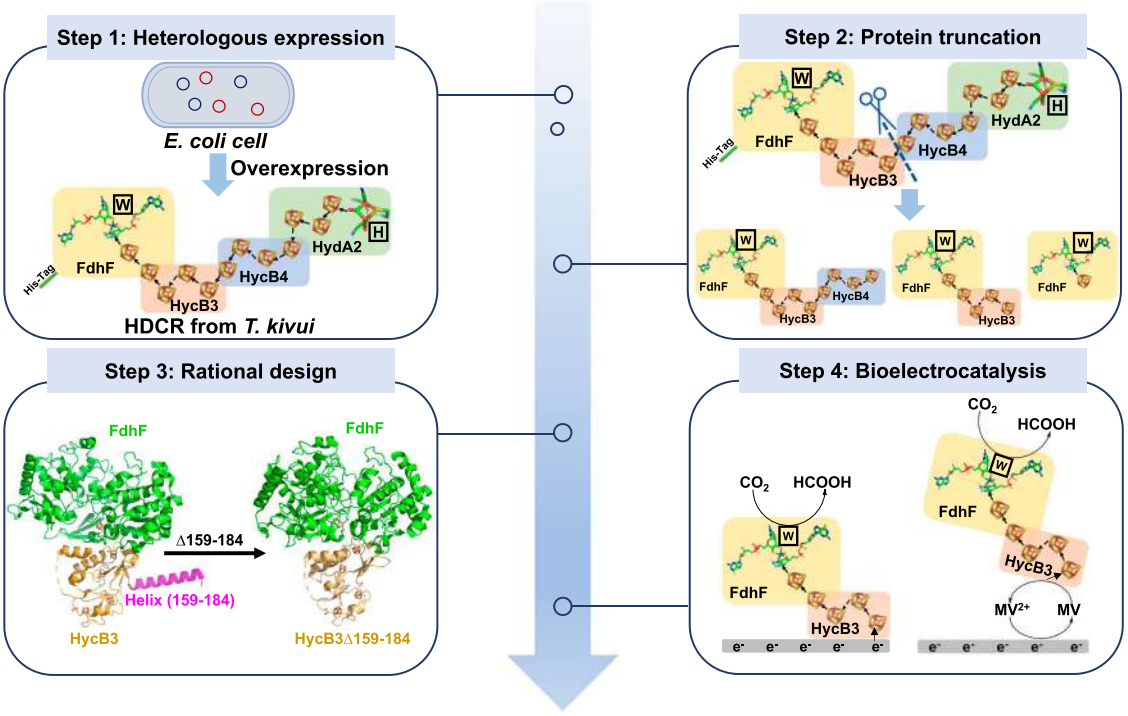


Fig. 1 | Schematic illustration of this study. Step 1: HDCR from *T. kivui* was heterologously expressed in *E. coli*, step 2: Formate dehydrogenase modules were acquired by truncation of HDCR, step 3: The robust variant FdhF_HycB3Δ159-184 was obtained by rational design, step 4: Bioelectrocatalysis systems were constructed based on FdhF_HycB3Δ159-184.

https://doi.org/10.1038/s41467-024-53946-3

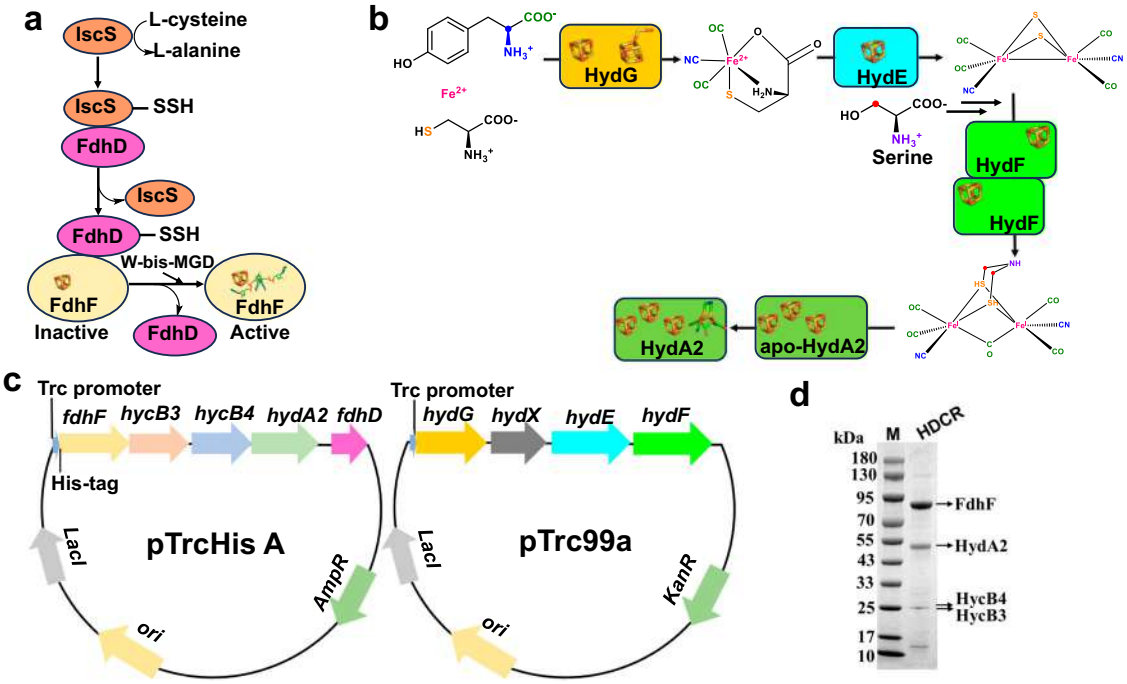


Fig. 2 | Strategy for heterologous expression of HDCR complex in *E. coli* MC1061. a The maturation process of FdhF was assisted by FdhD, IscS: L-cysteine desulfurases[30], (**b**) Biosynthesis of the catalytic H-cluster of [FeFe]-hydrogenase by HydG, HydE and HydF[33], (**c**) HDCR gene fdhF, hycB3, hycB4, hydA2 and maturation factor gene fdhD from *T. kivui* were cloned and inserted into plasmid pTrcHis A, the [FeFe]-hydrogenases mature proteins gene hydG, hydX, hydE, hydF from *Shewanella oneidensis* MR-1were cloned and inserted into plasmid pTrc99a. **d** SDS–PAGE analysis of purified HDCR complex expressed in *E. coli* MC1061; M: standard protein marker. Three independent measurements were repeated. Source data are provided as a Source Data file.

the deficiency of auxiliary factors caused by the homologous disparity between *T. kivui* and *E. coli*. The maturation of W-bis-MGD and di-Fe active centers is crucial for the activation of *Tk*HDCR. According to the cryo-EM structure, W-bis-MGDs are buried deep inside the FdhF subunit[25], suggesting that their insertion is intimately connected to protein folding and that the necessary chaperone is involved[26,27]. Two possible genes, fdhD, and fdhC, were found in the *Tk*HDCR gene cluster, while their corresponding proteins did not appear in the final mature HDCR. It is reasonable to speculate that these proteins might be molecular chaperones involved in HDCR maturation[28]. However, the sequence identity between FdhD of *E. coli* and FdhD from *T. kivui* is only 27.3%, implying that FdhD from *E. coli* may not effectively assist the correct insertion of W-bis-MGD into the FdhF subunit of *T. kivui*, which is different from what is observed in *Aw*HDCR expressed in *E. coli*[29]. Therefore, the roles of the fdhD and fdhC genes from *T. kivui* were investigated experimentally (Supplementary Fig. 1). To investigate the role of FdhD and FdhC in the maturation of FdhF, the fdhD and fdhC genes were codon-optimized and cloned into plasmid pETDuet-1. As shown in Supplementary Fig. 1, the bands at 82.5 kDa were assigned to FdhF, but no activity is detected with respect to fdhF_fdhC, while fdhF_fdhD and fdhF_fdhD_fdhC show nearly equal activity towards formate oxidation (Data was not shown). This suggests that FdhD is necessary for the maturation of FdhF, but FdhC has less influence. As a consequence, FdhD is necessary for the functional integrity of FdhF, and it may catalyze the transfer of sulfur from L-cysteine desulfurase (IscS) to Mo/W-bis-MGD, forming the terminal sulfur ligand Mo/W = S and assisting in the assembly of mature Mo/W-bis-MGD into the active center of FDH (Fig. 2a)[30,31]. The H-clusters of [FeFe]-hydrogenases also require accessory proteins for the maturation of the active center[32]. The chaperones HydE, HydF, and HydG are conserved, and their roles have been well studied (Fig. 2b)[32–34]. [FeFe]-hydrogenases from *C. reinhardtii* and *C. pasteurianum* exhibited high-yield expression in *E. coli* along with the maturation proteins (HydG, HydX, HydE, and HydF) from *Shewanella oneidensis* MR-1[35]. The HydG, HydE, and HydF proteins from *T. kivui* exhibit sequence identities of 47.0%, 36.8%, and 49.1%, respectively, to their counterparts in *Shewanella oneidensis* MR-1. On this basis, the pACYCDuet-hydGXEF plasmid was constructed by cloning the maturase gene cluster, hydG, hydX, hydE, and hydF from *Shewanella oneidensis* MR-1 (Supplementary Fig. 2a).

When pETDuet-*hdcr* and pACYCDuet-*hydGXEF* were both transformed into *E. coli* BL21 (DE3) for *Tk*HDCR overexpression (Supplementary Fig. 2a), only the FdhF subunit was observed (Supplementary Fig. 2b, lane 1). When an extra His-tag was fused at the C-terminus of HydA2, HydA2 can be captured (Supplementary Fig. 2b, lane 2), suggesting unsuccessful expression of the HycB3 and HycB4 subunits or improper assembly of each subunit of the HDCR complex. The expression of sole HycB3 and sole HycB4 formed massive inclusion bodies, which was confirmed to be the reason for the failure of *Tk*HDCR complex assembly (Supplementary Fig. 2c). In these circumstances, the vector pTrcHis was used because the Trc promoter is weaker than the T7 promoter, allowing for mild cofactor maturation. It has been widely employed in the expression of multi-subunit enzymes or insoluble enzyme[36,37]. The plasmid pTrcHisA-*hdcr*, along with its counterpart pTrc99a-*hydGXEF*, was constructed and expressed in *E. coli* MC1061 (Fig. 2c and Supplementary Fig. 3). Remarkably, all four subunits of *Tk*HDCR, FdhF, HydA2, HycB4, and HycB3, were distinctly observed (Fig. 2d), demonstrating successful expression of the *Tk*HDCR complex. Using methyl viologen (MV^{2+}) and reduced methyl viologen (MV) to assess catalytic activity, respectively, the obtained HDCR exhibited a CO_2 reduction activity (MV: CO_2-oxidoreductase activity) of 1318.7 min^{-1} at 30 °C and a formate oxidation activity (formate: MV^{2+}-oxidoreductase activity) of 202.9 min^{-1}. This achievement represents the successful production of active HDCR from *T. kivui* in *E. coli*, revealing a convenient and expeditious method.

Engineering of *Tk*HDCR to acquire FDH module

To obtain the FDH module, HydA2 and its possible electron-transfer counterparts were gradually truncated from the *Tk*HDCR[38], generating

https://doi.org/10.1038/s41467-024-53946-3

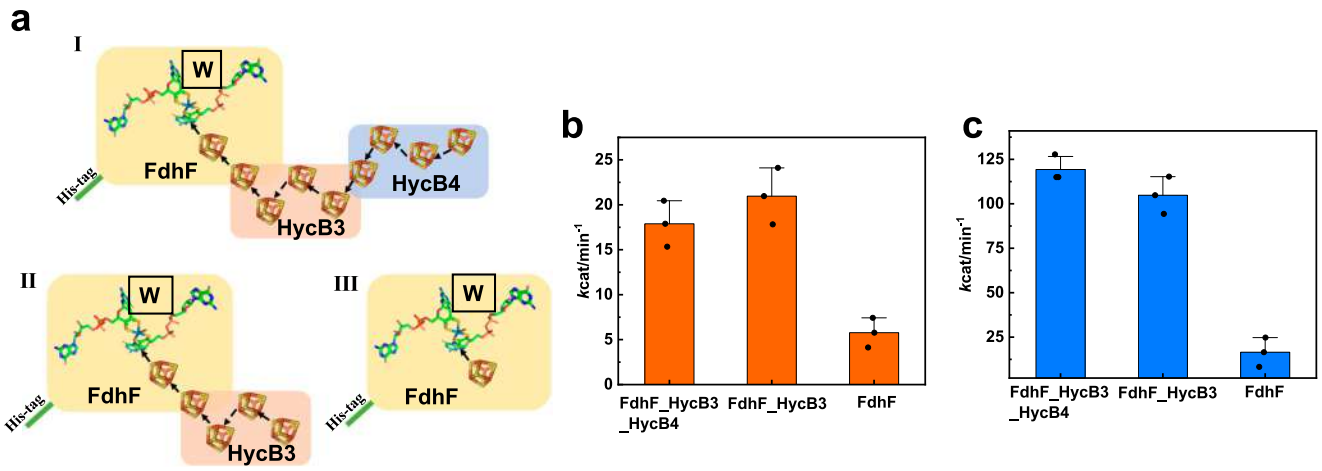


Fig. 3 | Truncation of *Tk*HDCR. a Schematic models of three *Tk*HDCR variants[38], (I) FdhF_HycB3_HycB4, (II) FdhF_HycB3, and (III) FdhF. **b** The turnover frequency of *Tk*HDCR variants for formate oxidation. **c** The turnover frequency of *Tk*HDCR variants for CO_2 reduction. The error bars correspond to the standard deviation of at least three independent measurements, and the center value for the error bars is the average of the three independent measurements. Source data are provided as a Source Data file.

three variants, FdhF_HycB3_HycB4 (HDCRΔHydA2), FdhF_HycB3 (HDCRΔHydA2ΔHydcB4), and FdhF (HDCRΔHydA2ΔHydcB4ΔHydcB3) (Fig. 3a and Supplementary Fig. 4a). The activity determination results indicate that FdhF_HycB3 and FdhF_HycB3_HycB4 exhibits equivalent activity in both the reduction of CO_2 and formate oxidation (Fig. 3b, c). However, FdhF's activity decreases by approximately 80% compared to FdhF_HycB3, suggesting that HycB3 is crucial as an electron transfer protein. The distal [Fe-S] cluster in HycB3 is likely helpful for electron shuffling between electron carriers and the metal active site (Supplementary Fig. 5). When one of its coordinated residues, C83, was substituted by serine (Supplementary Fig. 4b), the catalytic activity completely vanished. Compared to FdhF_HycB3_HycB4, FdhF_HycB3 has a smaller molecular weight, which may facilitate better fixation and transfer of electrons on the electrode. Therefore, FdhF_HycB3 was selected for further study.

It is noted that the activity of FdhF_HycB3 in reducing CO_2 was lower than that of *Tk*HDCR, which may be attributed to self-aggregation and abnormal conformations. In the original *Tk*HDCR, the formation of the nanowire structure relied on the crosslinking mediated by the highly similar C-terminal α-helical structures in HycB3 (sequences G159-K184, purple part in Fig. 4a) and HycB4 (sequences P190-R210)[25]. FdhF_HycB3 may undergo self-aggregation via fragments 159–184, thereby decreasing its specific activity. Dynamic light scattering (DLS) measurements validate this hypothesis with a wide range of size distributions (Supplementary Fig. 6a). In addition, FdhF and HycB3 fail to adopt the correct conformation when HycB3 does not participate in the formation of nanowire. Structural analysis revealed that the FdhF subunit directly interacts with Helix 159–184 to form an incorrect conformation, resulting in an increase in the distance between the two [Fe-S] clusters in HycB3 and FdhF (from 11.3 Å to 30.7 Å) and disrupting the electron hopping between them (Supplementary Fig. 7a, b).

Based on these observations, a variant of FdhF_HycB3 was generated by removing the α-helix (G159-K184), namely, FdhF_HycB3Δ159–184 (Supplementary Fig. 4b and Fig. 4a). Activity determination, DLS and docking analysis of FdhF_HycB3Δ159–184 were also carried out. As a result, the catalytic activities partially improved to 553.7 ± 122.4 min^{-1} for CO_2 reduction and 400.8 ± 85.8 min^{-1} for formate oxidation (Fig. 4b). DLS analysis revealed an enhanced distribution below 50 nm, which approached the size of the monodispersed enzyme molecule (Supplementary Fig. 6b). In the structural modeling of HycB3Δ159-184, a conformation overlaying well with FdhF and HycB3 in the PDB structure was observed (Supplementary Fig. 7c, d), confirming the adverse structural effect of the C-terminal α-helix of HycB3 on the variant FdhF_HycB3. In addition, we have performed elemental analysis (W) of FdhF_HycB3Δ159-184 using ICP-OES, which revealed 13.6 nmol of W was detected when the amount of protein is 16.6 nmol of protein. The above 80% of metal filling in enzyme molecules suggests the reliability of k_{cat} determination and the proper enzyme folding/maturation during heterologous expression in *E. coli*.

After obtaining the beneficial variant FdhF_HycB3Δ159-184, the temperature profile was determined since *Tk*HDCR is a thermostable enzyme with the maximum catalytic activity at 70 °C[12]. As shown in Fig. 4c, the activity of formate oxidation increases with the elevated temperature, reaching 1110.7 ± 57.6 min^{-1} at 70 °C. Meanwhile, the maximum CO_2 reduction activity reaches 1579.5 ± 100.9 min^{-1} at 60 °C and decreases to 774.4 ± 129.7 min^{-1} at 70 °C. The decreased reduction activity at 70 °C might be caused by partial enzyme inactivation. Similar behavior for the temperature-dependent activity of *Dv*FDH was reported[39]. This was further proved by the nonlinear fitting of the temperature-dependent model (Supplementary Fig. 8)[40].

Electrocatalytic characterization of FdhF_HycB3Δ159-184 electrode towards CO_2 reduction

The electrocatalysis of FdhF_HycB3Δ159-184 was characterized by constructing an FdhF_HycB3Δ159-184/carbon nanotube/glassy carbon electrode (FdhF_HycB3Δ159-184/CNT/GC). A cyclic voltammetric (CV) experiment of FdhF_HycB3Δ159-184/CNT/GC electrode was carried out upon adding an equal amount of formate and $NaHCO_3$. As shown in Supplementary Fig. 9, the reduction potential of CO_2 catalyzed by FdhF_HycB3Δ159–184/CNT/GC electrode vary from pH 6.2 to pH 8.0. Similar to that observed previously for *Sf*FDH1[13] and *Ec*FDH-H[14], the initial potential is − 410 ± 2 mV *vs*. SHE at pH 7.0, which is consistent with the formic acid /CO_2 redox potential (− 412 mV *vs*. SHE) at pH 7.0[13]. The zero-current point (Net formate oxidation occurs at all potentials above the intersection, and net CO_2 reduction occurs at all potentials below it) represents the thermodynamic reduction potential for the CO_2/formate interconversion. The near-zero overpotential endows FdhF_HycB3Δ159–184 with superior performance to chemical catalysts for energy-saving direct CO_2 transformation[41]. While the onset of CO_2 reduction positively shifted to − 370 ± 4 mV *vs*. SHE at pH 6.2 and negatively shifted to − 471 ± 3 mV *vs*. SHE at pH 8.0. The decreased catalytic current for CO_2 reduction at pH 8.0 compared to

https://doi.org/10.1038/s41467-024-53946-3

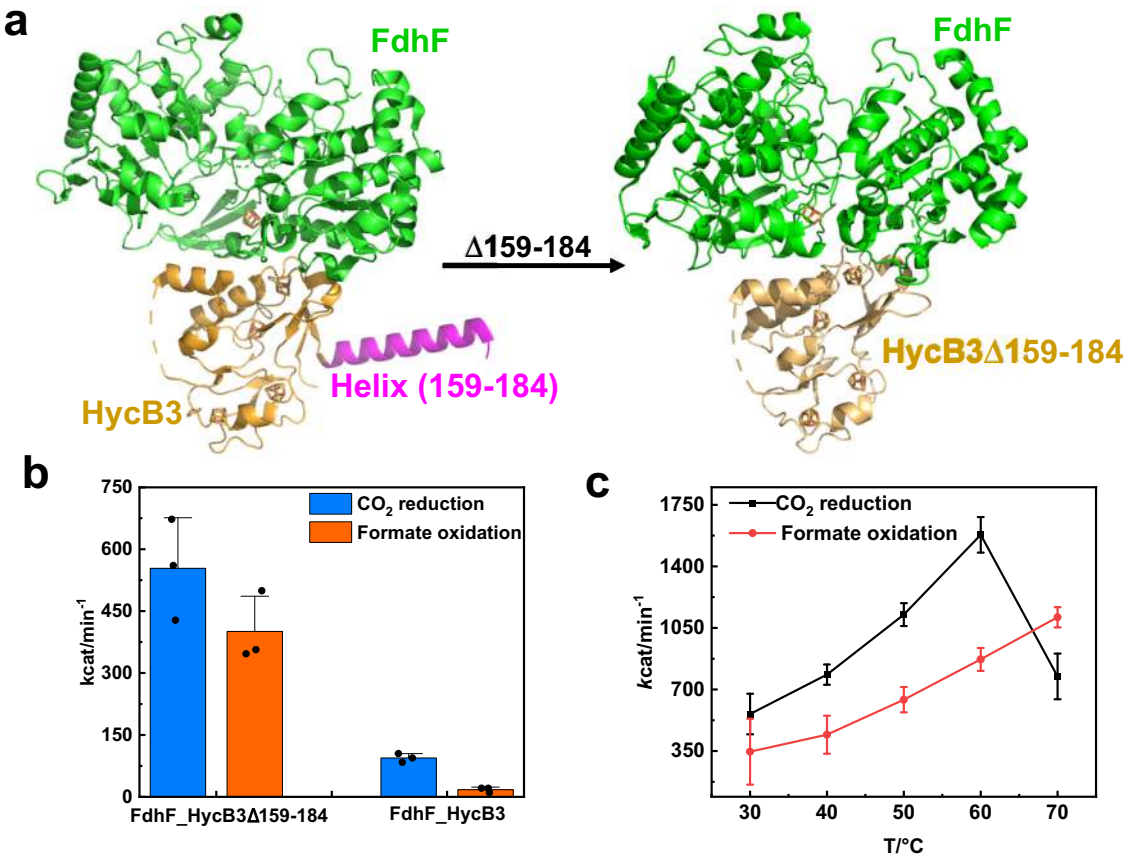


Fig. 4 | Rational design of FdhF_HycB3 for improved catalytic activity. **a** Engineering strategy. **b** The turnover frequency of FdhF_HycB3 and FdhF_HycB3Δ159-184 toward formate oxidation and CO_2 reduction. **c** Temperature profile of FdhF_HycB3 Δ159-184 catalyzing formate oxidation and CO_2 reduction. The error bars correspond to the standard deviation of at least three independent measurements, and the center value for the error bars is the average of the three independent measurements. Source data are provided as a Source Data file.

those at pH 6.2 and 7.0 might be caused by the accelerated CO_2 hydration in alkaline solutions[23]. It has been previously reported that the CO_2 molecule, rather than its hydrated HCO_3^- and CO_3^{2-}species, is the substrate of formate dehydrogenases, so is it in the case of FdhF_HycB3Δ159-184[42]. The FdhF_HycB3Δ159-184/CNT/GC electrode was also tested in a CO_2-saturated electrolyte. The current density (the capacitance current doesn't participate in the calculation of the enzymatic turnover current) for CO_2 reduction reached $270.0 \pm 7.1\,\mu A\ cm^{-2}$ at −550 mV *vs.* SHE (Fig. 5a and Table 1). Considering that the maximum catalytic activity of FdhF_HycB3Δ159-184 was observed at 60 °C, the electrocatalytic activity was also assessed at different temperatures. As shown in Fig. 5b, the catalytic current increases with the elevated temperature (from 30 °C to 60 °C), reaching $600.0 \pm 94.4\,\mu A\,cm^{-2}$ at −550 mV *vs.* SHE at 60 °C. In mediated electron transfer (MET) mode, MV is supplemented in the electrolyte. As shown in Fig. 5c, the catalytic current reaches $800.0 \pm 83.1\,\mu A\,cm^{-2}$ at −550 mV *vs.* SHE, which is 3 times greater than that in DET mode.

Mass transfer and kinetics of electrocatalysts are two main factors dominating electrocatalytic performance. To study the effect of mass transfer, a rotating disk electrode (RDE) was used. Figure 5d displays the LSVs of FdhF_HycB3Δ159-184/CNT/RDE obtained at different rotation rates (0, 100, and 900 rpm). Interestingly, no significant differences in current were observed with increasing rotation rate, suggesting that mass transfer has little influence on FdhF_HycB3Δ159-184 catalyzing the CO_2 reduction reaction and that the intrinsic enzymatic capability is the limiting factor, in accordance with the findings shown in Fig. 5b. To determine the kinetic parameters, a potential of −550 mV *vs.* SHE was applied to FdhF_HycB3Δ159-184/CNT/GC and current were recorded at different concentrations (0–50 mM) of $NaHCO_3$, which was widely used for studying the kinetics of bioelectrodes in the context of CO_2 reduction[10,18] (Fig. 5e), and according to Fig. 5e, the relationship between the current density and the substrate concentration was plotted in Fig. 5f. This result conforms to a typical Michaelis-Menten curve, with a calculated apparent *K*m value of 0.7 ± 0.1 mM CO_2. As the solubility of CO_2 at ambient temperature and pressure is ~33 mM, which is much higher than the concentration needed for maximum catalysis (10 mM), these results explain why rotation fails to affect the electrocatalytic performance. This phenomenon is different from that of the electrocatalytic oxygen reduction reaction, in which mass transfer limitations exist in most cases because of the extremely low solubility of oxygen in aqueous solution (0.5 mM at 25 °C and 101.3 kPa) [43]. To further illustrate the influence of FDH kinetics on electrocatalytic CO_2 reduction, two reported FDHs, FDH from *Cupriavidus necator* and FDH from *Rhodobacter aestuarii*, were tested. *Cn*FDH[44] exhibited a moderate CO_2 reduction activity of $1.0\,U\,mg^{-1}$, and *Ra*FDH[36] exhibited a low CO_2 reduction activity of $0.3\,U\,mg^{-1}$. As expected, slight catalytic currents were generated in both cases (Supplementary Fig. 10), demonstrating that the activity of FDH is still a determinant and that continuous enzyme design/engineering is necessary.

Electrocatalytic synthesis of formate from CO_2 at FdhF_HycB3Δ159-184 electrodes

Before constructing the electricity-driven formate synthesis system, the chemical energy-driven CO_2-formate conversion mode was

https://doi.org/10.1038/s41467-024-53946-3

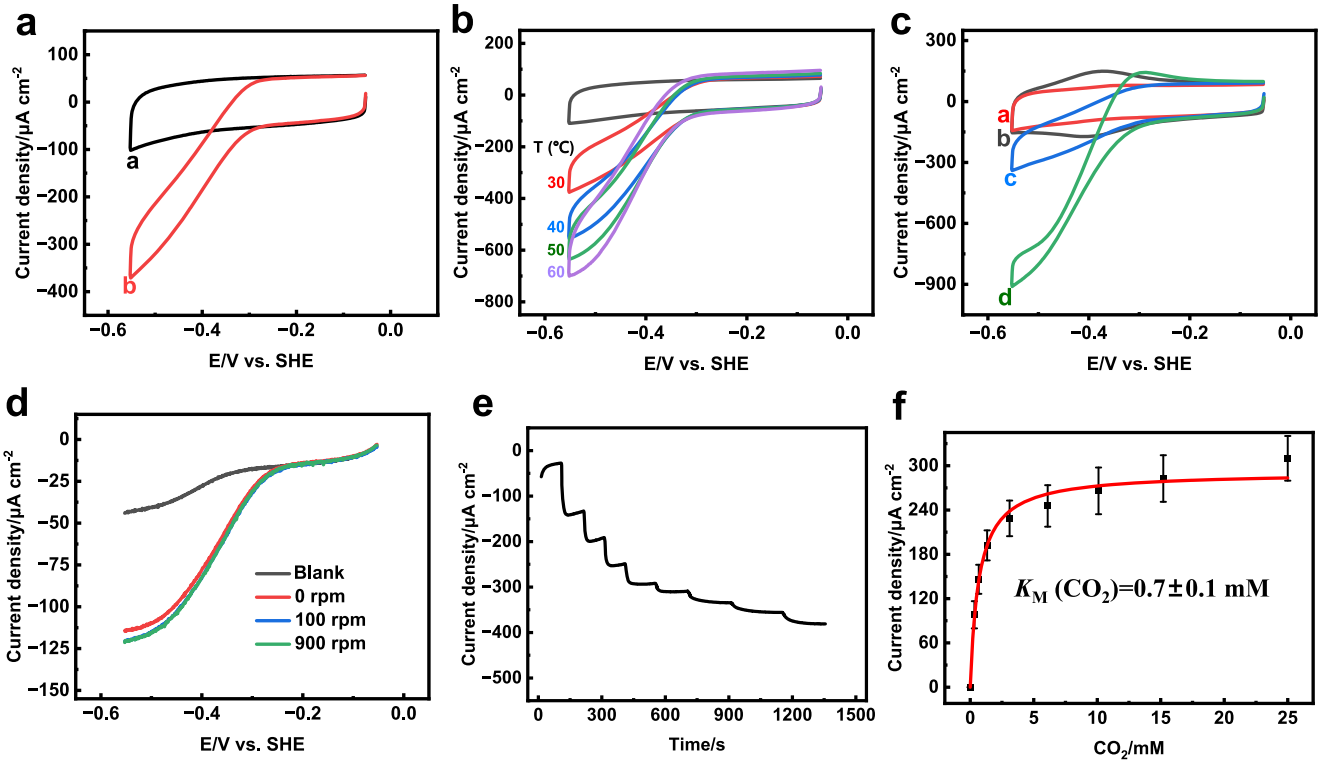


Fig. 5 | Electrocatalytic characterization of FdhF_HycB3Δ159-184/CNT/GC electrodes. a Cyclic voltammogram of FdhF_HycB3Δ159-184 for electrocatalytic CO_2 reduction; line a (black) was performed in 100% N_2, and line (**b**) (red) was performed in an atmosphere of CO_2 by bubbling CO_2 with a flow rate of 20 sccm. **b** Cyclic voltammogram of FdhF_HycB3Δ159-184 for electrocatalytic CO_2 reduction at different temperatures. The flow rate of CO_2 was 20 sccm (**c**) Cyclic voltammogram of FdhF_HycB3Δ159-184 electrocatalytic analysis of direct and indirect electron transport, line a (red): FdhF_HycB3Δ159-184/CNT/GC electrode was evaluated in 100% N_2; line b (black): the electrode was evaluated in the presence of 0.2 mM methyl viologen (MV) and 100% N_2; line c (blue): the electrode was evaluated in 100% CO_2; line d (green): the electrode was evaluated in the presence of 0.2 mM MV and 100% CO_2. **d** Linear sweep voltammetry (LSV) at 0 rpm (red), 100 rpm (blue), and 900 rpm (green) in an atmosphere of CO_2. **e** A potential of −550 mV *vs.* SHE was applied to FdhF_HycB3Δ159-184/CNT/GC for CO_2 reduction in mixed-buffer 0.1 M (Acetate, MES, HEPES, and TAPS), pH 6.0 buffer and $NaHCO_3$ (0–50 mM) was used as the substrate, 25 °C. **f** The catalytic currents and substrate concentrations are plotted. The experiments of (a), (b), (c), and (d) were carried out in 0.1 M HEPES, pH 6.2, buffer containing 0.1 M NaCl at 25 °C (excluding temperature optimization experiments), and at a scan rate of 5 mV s^{-1} (LSV was conducted at a scan rate of 1 mV s^{-1}). HEPES: 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid, CNT: Carbon nanotubes. The error bars correspond to the standard deviation of at least three independent measurements, and the center value for the error bars is the average of the three independent measurements. Source data are provided as a Source Data file.

evaluated. With sodium dithionite (DTH) as the sacrificial reductant (electron donor) and MV as the electron mediator, 1.3 mM formate was detected after 1 h of aqueous catalysis (Supplementary Fig. 11a), validating the effectiveness of the FdhF_HycB3Δ159–184 variant and the feasibility of constructing an electricity-driven formate synthesis system. In DET mode (Fig. 6a and Supplementary Fig. 12a), − 500 mV *vs.* SHE was applied for uninterrupted 14 h of electrolysis, and formate was generated as the sole product, as confirmed by HPLC (Supplementary Fig. 11b) and ^{1}H NMR (Supplementary Fig. 13). The yield of formate increases to 8.0 ± 0.5 mM in 14 h, with an average synthesis rate of 2.9 ± 0.2 μmol h^{-1} cm^{-2}, which is higher than the rate of 1.42 μmol h^{-1} cm^2 catalyzed by conductive hydrogel-mediated *Cl*FDH[17]. At 4 h, the system reached the maximum faradaic efficiency (FE%) and cathodic energy efficiency (CEE%) of 99.1% and 94.6%, respectively, and the synthesis rate of 5.8 μmol $h^{-1}cm^{-2}$ (Fig. 6b). This high CEE is even greater than that of Bi-based catalysts, which have excellent selectivity (57%, 86%)[45,46], because of the high selectivity and near-zero overpotential of enzyme catalysis. In the MV-mediated mode (Fig. 6c and Supplementary Fig. 12b), formate accumulated rapidly in the 6 hours and then became moderate, probably because of the poor electrode stability, until it reached 22.8 ± 1.6 mM at 16 h. The average production rate is as high as 7.1 ± 0.5 μmol $h^{-1}cm^{-2}$. At 6 h, the system reached the maximum FE% and CEE% of 98.9% and 94.4%, respectively, and the synthesis rate of 15.2 ± 0.6 μmol $h^{-1}cm^{-2}$ (Fig. 6d). This rate is lower than the best formate production rate of 55 μmol h^{-1} cm^{-2} reported by Reisner et al.[21], but it is higher than most bioelectrodes used for the electrocatalytic synthesis of formic acid (Table 1). Further optimization of immobilization strategies could improve the formic acid production efficiency of this system.

In the present work, we achieved convenient and efficient heterologous expression of *Tk*HDCR by *E. coli*, grounding enzyme engineering for an electro-responsive carbon dioxide reductase and its massive production. The engineering of *Tk*HDCR by subunit truncation and rational design revealed that the HycB3 subunit is essential for electron shuffling, while its original "arm", helix 159-184, is harmful for enzyme molecule dispersion and proper conformation preservation. Thus, FdhF_HycB3Δ159-184 was generated, with a maximum specific activity of 1579.5 min^{-1}at 60 °C. The electrocatalytic determination of FdhF_HycB3Δ159-184 demonstrated DET-type bioelectrocatalysis. The catalytic current in the MET mode is ~ 3 times greater than that in the DET mode. Since the distal Fe-S cluster of HycB3 facilitates electron transfer between HycB3 and FdhF, immobilizing HycB3 in a specific orientation on the electrode might improve direct electron transfer. Since the *K*m value is calculated as 0.7 ± 0.1 mM CO_2, which is much lower than the solubility of CO_2, mass transfer is not the rate-limiting factor. By applying an applied voltage to the MV-mediated FdhF_HycB3Δ159-184 electrode for CO_2 transformation, a formate concentration of 22.8 ± 1.6 mM and a yield rate of 7.1 ± 0.5 μmol $h^{-1}cm^{-2}$ were observed after 16 hours. An FE of 98.9% and a CEE of 94.4% were attained. This demonstrates the potential of FdhF_HycB3Δ159-184 as a

https://doi.org/10.1038/s41467-024-53946-3

catalyst for environmentally friendly and sustainable formate production. Some fixation strategies may further improve the stability and efficiency of FdhF_HycB3Δ159-184 bioelectrode. Moreover, the generated formate holds promise for conversion into higher-value commodities through enzymatic catalysis.

Methods

Reagents and materials

All the chemicals were of analytical grade or of higher purity and were purchased from Sigma–Aldrich, Macklin (Shanghai, China), and Sinopharm (Beijing, China) unless otherwise noted. Primers and genes were synthesized by Tsingke (Beijing, China), and Primers were dissolved in Milli-Q water to a final concentration of 10 μM. PrimeSTAR Max DNA Polymerase was obtained from Takara Bio (Japan). The methylation-sensitive restriction enzyme (*Dpn*I) was obtained from New England Biolabs. A ClonExpress-II one-step cloning kit was obtained from Vazyme (Nanjing, China). Lysogeny Broth (LB) medium (1 L) containing 10 g of NaCl, 10 g of tryptone and 5 g of yeast extract. For the LB agar plates, 20 g of agar was added. All the media were then autoclaved at 121 °C for 20 min. For the anaerobic culture medium, 100 mM 3-(N-morpholino) propanesulfonic acid (MOPS), 5 g L^{-1} glucose, 25 mM sodium fumarate, and 1 mg L^{-1} resazurin were added to LB media, and the pH was adjusted to 7.4. The anaerobic culture medium was autoclaved at 115 °C for 30 min, after which the oxygen was replaced with 100% N_2. Carbon nanotubes (CNTs) were obtained from XFNANO (Nanjing, China).

Construction of plasmids

The *hdcr* gene was codon optimized and synthesized and subsequently cloned and inserted into pET20b by Tsingke. The polymerase chain reaction (PCR) mixture (50 μL) consisted of 25 μL of PrimeSTAR Max DNA Polymerase, 20 ng of the template (pET20b-*hdcr*), and 500 nM primers P1 and P3, followed by the PCR protocol: 98 °C for 2 min (1 cycle); 98 °C for 20 s; 53 °C for 20 s; 72 °C for 2 min (25 cycles); and 72 °C for 10 min (1 cycle). PCR products were characterized by agarose gel electrophoresis. The Megawhop PCR mixture (50 μL) contained 25 μL of PrimeSTAR Max DNA Polymerase, 20 ng of the pTrcHisA plasmid as the template, and 500 nM primers P2 and P4. The PCR protocol was as follows: 98 °C for 2 min (1 cycle); 98 °C for 30 s, 53 °C for 20 s, and 72 °C for 2 min (25 cycles); and 72 °C for 10 min (1 cycle). All the methylated templates were specifically degraded by *Dpn*I. The PCR products were purified with a universal DNA purification and recovery kit (TIANGEN, China, Beijing). The above two PCR products were linked with a ClonExpress-II one-step cloning kit (37 °C, 30 min). The linked product was subsequently transformed into 100 μL of *E. coli* Top10 competent cells (heated at 42 °C for 45 sec) and recovered at 37 °C for 1 h. The cells were subsequently plated on LB agar plates supplemented with 50 mg L^{-1} ampicillin for growth overnight at 37 °C.

The construction of the plasmid pTrc99a was performed with kanamycin resistance. The kanamycin resistance gene was cloned from the template pET28a using the primers P11 and P12. pTrc99a was generated by removing the original ampicillin resistance gene via the primers P13 and P14, and the pTrc99a-kan$^+$ plasmid harboring the kanamycin resistance gene was generated via the ClonExpress-II one-step cloning kit. The linked product was subsequently transformed into 100 μL of *E. coli* Top10 competent cells, which were subsequently allowed to recover at 37 °C for 1 h. The cells were plated on LB agar plates supplemented with 50 mg L^{-1} kanamycin for growth overnight at 37 °C.

The plasmid pTrc99a-kan$^+$-*hydGXEF* containing the maturation factor genes *hydG*, *hydX*, *hydE*, and *hydF* of FeFe-hydrogenase from *Shewanella oneidensis* MR-1 was constructed using the primers P15, P16, P17 and P18. pTrc99a-kan$^+$-*hydGXEF* and pTrcHisA-*hdcr* were subsequently transformed into *E. coli* MC1061 competent cells. The

Table 1 | Comparison of electrocatalytic CO_2 reduction by Mo/W-containing FDHs

Enzymes	Electrode modification	Electron transfer types	Potential/V vs. SHE	Current density /μA cm^{-2} (mA cm^{-2})	Concentration of formate/mM	Electrode area/cm^2	Production rate/ μmol $h^{-1}cm^{-2}$	Reference
*Sf*FDH-1	Pyrolytic graphite electrode	DET	−0.8	−80	–	–	–	13
*Ec*FDH	Pyrolytic graphite electrode	DET	−0.6	−200	–	–	–	14
*Cl*FDH	Pyrolytic graphite electrode	DET	−0.8	−400	–	–	–	16
*Cl*FDH	Conductive polyaniline (PANi) hydrogel	MET	−0.4	–	–	1.0	1.42	17
*Ec*FDH	Grafted to poly(allylamine) modified carbon electrode	MET	−0.66	−62	–	0.8	0.4	18
*Dv*H-FDH	4-nitrophenyl radical-modified graphite electrode	DET	−0.7	−200	0.004	0.2	0.44	19
*Dv*H-FDH	MesoTiO_2 electrode	DET	−0.6	−100	4.9	0.25	0.82	48
*Dv*H-FDH	IO-TiO_2 I Perovskite I $BiVO_4$	DET	–	(−5)	–	0.19	7	49
*Dv*H-FDH	Mesoporous indium-tin-oxide	DET	−0.8	(−3.6)	–	0.19	55	21
*Dv*H-FDH	$NHMe_2^+$ modified carbon nanotubes	DET	−0.6	−247	1.98	0.07	1.2	20
		MET	−0.6	–	4.18	0.07	2.5	20
FdhF_HycB3Δ159-184	Carbon nanotubes	DET	−0.5	−270	8.0	1.0	2.9	This study
		MET	−0.5	−800	22.8	1.0	7.1	This study

DET: Directed electron transfer. MET: Mediated electron transfer.

https://doi.org/10.1038/s41467-024-53946-3

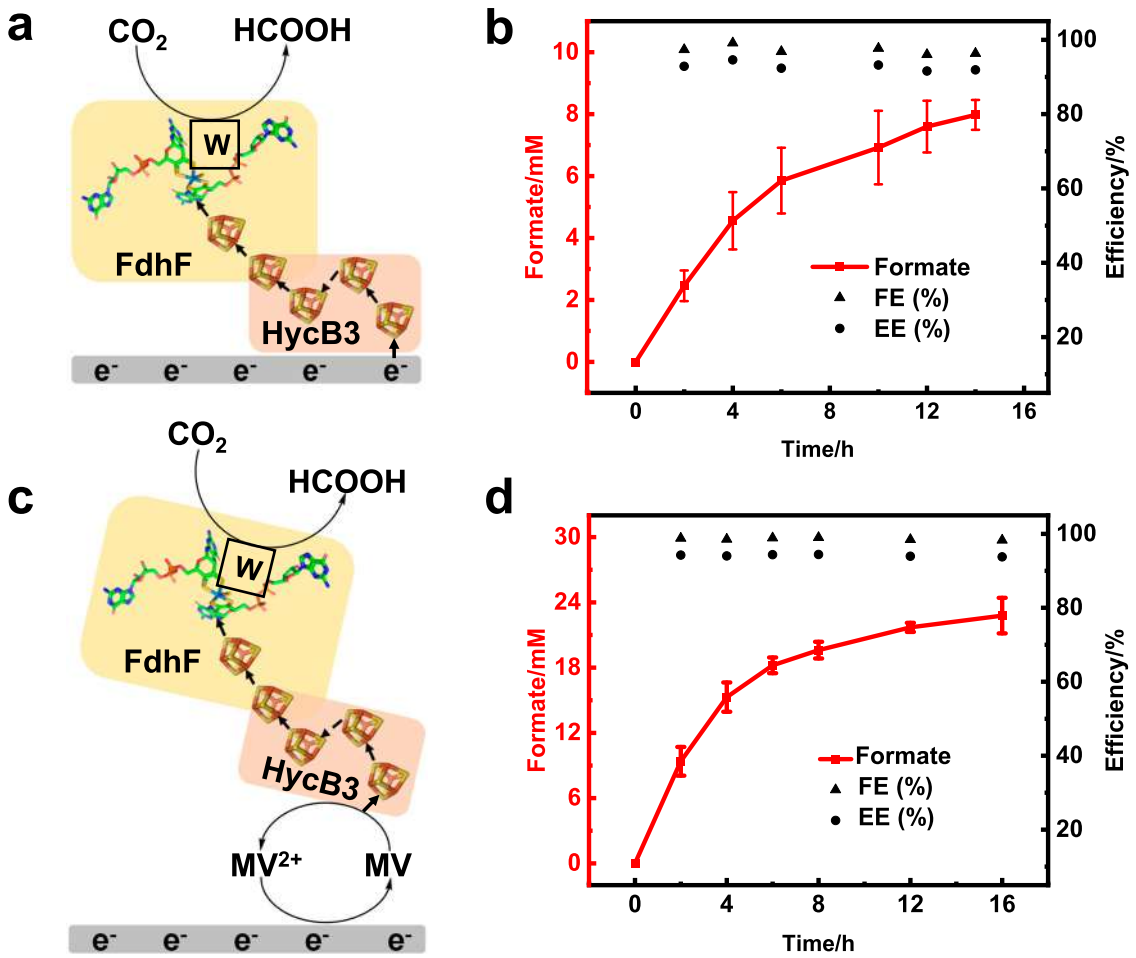


Fig. 6 | Electrosynthesis of formate from CO_2 at FdhF_HycB3Δ159-184 electrodes. a Electrocatalysis via direct electron transfer; (**b**) formate produced by electrocatalysis via direct electron transfer; (**c**) electrocatalysis via mediated electron transfer by adding 0.2 mM MV into the electrolyte; (**d**) formate produced by electrocatalysis via mediated electron transfer. Electrolyte: 0.1 M HEPES/NaOH, 0.1 M NaCl buffer, pH 7.0. A constant voltage of − 500 mV vs. SHE was used. The error bars correspond to the standard deviation of at least three independent measurements, and the center value for the error bars is the average of the three independent measurements. Source data are provided as a Source Data file.

clone was picked from LB agar plates containing kanamycin and ampicillin. *E. coli* MC1061 containing pTrc99a-kan^{+}-*hydGXEF* and pTrcHisA-*hdcr* was used to express HDCR. The pTrcHisA-*fdhF_hycB3_hycB4* (pTrcHisA-*hdcrΔhydA2*), pTrcHisA-*fdhF_hycB3* (pTrcHisA-*hdcrΔhydA2ΔhycB4*), and pTrcHisA-*fdhF* (pTrcHisA-*hdcrΔhydA2ΔhycB4ΔhycB3*) were constructed using primers P5 and P6 P5 and P7, and P5 and P8, respectively. The plasmid was transformed into *E. coli* MC1061 competent cells. The clone was picked from the LB agar plates containing ampicillin. The FdhF_HycB3Δ159-184 variant-expressing plasmid pTrcHisA-*fdhF_hycB3Δ159-184* (pTrcHisA-*hdcrΔhydA2ΔhycB4Δ159-184*) was constructed by using pTrcHisA-*fdhF_hycb3* as a template, and using primer P9 and P10. The variant FdhF_HycB3-C83S expression plasmid was constructed by using primer P38 and P39, and pTrcHisA-*fdhF_hycB3* was used as the template.

The plasmid pETDuet-hdcr was constructed by cloning *fdhF and hycb3* into multiple cloning sites 1 (MCS1) of pETDuet-1 using P19 and P20, P21 and P22. The *hycB4 and hydA2* genes were subsequently cloned and inserted into the MCS2 of pETDuet-1 using P23 and P24, P25, and P26. The maturation factor gene FdhD of FdhF was also cloned and inserted into MCS1 of pETDuet-1 using P20 P27, P28, and P29. In addition, the pACYCDuet-*hydGXEF* plasmid, which contains the maturation factor genes *hydG*, *hydX*, *hydE*, and *hydF* of FeFe-hydrogenase from *Shewanella oneidensis* MR-1, was constructed by cloning the *hydG and hydX* genes into the MCS1 of pACYCDuet-1 using P30 and P31, P32 and P33, and cloning *hydE and hydF* into the MCS2 of pACYCDuet-1 using primers P34 and P35, P36 and P37. Both pETDuet-*hdcr* and pACYCDuet-*hydGXEF* were transformed together into *E. coli* BL21 (DE3) competent cells. The clone was picked from LB agar plates containing ampicillin and chloramphenicol. For primer sequences, please refer to Supplementary Table 1.

Protein expression and purification

For the expression of HDCR in *E. coli* BL21 (DE3), cells were grown anaerobically in anaerobic culture medium[29] supplemented with ampicillin (50 mg L^{-1}), chloramphenicol (30 mg L^{-1}), and 1 mM sodium tungstate. Cultures were grown anaerobically at 37 °C to an OD_{600} of 0.6–0.8, 2 mM $FeSO_4$, and 0.4 g L^{-1} L-cysteine were added. A final concentration of 0.2 mM isopropyl beta-D-1-thiogalactopyranoside (IPTG) was used for the induction of protein production for 20 h under anaerobic conditions at 18 °C and 150 rpm. For the expression of HDCR in *E. coli* MC1061, the conditions were the same as above, except for the change from ampicillin (50 mg L^{-1}) or chloramphenicol (30 g L^{-1}) to ampicillin (50 mg L^{-1}) and kanamycin (50 mg L^{-1}). The expression of the HDCR variants in *E. coli* MC1061 was performed under the same conditions but with only the addition of ampicillin (50 mg L^{-1}).

All purification steps were performed under strictly anaerobic conditions at room temperature in an anaerobic chamber (Vigor Gas Purification Technologies Company, China Suzhou) filled with 96–98%

N_2 and 2–4% H_2, and an oxygen content below 2 ppm. For HDCR purification, cultures were centrifuged at 4 °C for 10 min at 4500 × *g*, resuspended in 30 mL of buffer A (25 mM Tris, 20% [v/v] glycerol, 20 mM $MgSO_4$, pH 7.5, and 1 mg L^{-1} resazurin), centrifuged again and resuspended in buffer A supplemented with 40 mg L^{-1} PMSF, 0.1 g L^{-1} DNaseI, 2 g L^{-1} lysozyme, and 0.5 mM dithiothreitol, and then incubated at 37 °C for 1 h. The cells are completely broken up by freezing and thawing three times with liquid nitrogen. Cell debris was removed by centrifugation at 17500 × *g* and 4 °C for 30 min. The cleared lysate was added to 3 mL of Ni-nitrilotriacetate (Ni-NTA), which was equilibrated with 15 mL of buffer A containing 40 mg L^{-1} PMSF and 0.5 mM dithiothreitol. The resin was subsequently washed with 25 column volumes of buffer A containing 40 mM imidazole, 40 mg L^{-1} PMSF, and 0.5 mM dithiothreitol, and the retained protein was eluted with ~ 5 mL of the same buffer A containing 250 mM imidazole, 40 mg L^{-1} PMSF and 2 mM dithiothreitol. For the purification of the other HDCR variants, buffer B (100 mM Tris, 100 mM NaCl, 10 mM $NaNO_3$, pH 8.0, and 1 mg L^{-1} resazurin) was used, and the purification procedure was the same as above. The protein purifier (AKTA purifier 10) equipped with Superdex 200 (10/300 GL, General Electric Company) was used for further purification. The flow of buffer was 0.9 mL min^{-1}. The protein concentration was measured according to the Bradford method. The proteins were separated on 15% polyacrylamide gels and stained with Coomassie brilliant blue G250.

Enzyme activity assay

Enzyme activity assays were performed in an anaerobic chamber. Buffer C (HEPES/NaOH 100 mM, 20 mM $MgSO_4$, pH 7.0, 2 mM DTT) was used for HDCR, buffer D (HEPES/NaOH 100 mM, pH 7.0, 2 mM DTT) was used for HDCR variants, and DTT was used as the protective reagent for enzymes. For formate oxidation, 20 mM substrate, sodium formate, and 20 mM electron acceptor, MV^{2+}, were added into Buffer C/D, and the total volume was 2 mL. The absorption at 604 nm ($\varepsilon = 13.9\ mM^{-1}\ cm^{-1}$), assigned to the reduced MV^{+}, was monitored by Evolution One (Thermo Fisher Scientific) with a temperature control module after the addition of HDCR and its variants. One unit of oxidation activity was defined as oxidizing 1 μmol of formate (reducing 2 μmol of MV^{2+}) per minute under the assay conditions.

For CO_2 reduction, MV was first reduced by sodium dithionite (DTH) to produce the reduced MV, which is used as the electron donor for CO_2 reduction. Similarly, the absorption decreased at 604 nm ($\varepsilon = 13.9\ mM^{-1}\ cm^{-1}$), and Evolution One (Thermo Fisher Scientific) was monitored with a temperature control module after the addition of HDCR and its variants. One unit of reduction activity was defined as reducing 1 μmol of CO_2 (oxidizing 2 μmol of MV^{+}) per minute under the assay conditions.

For a temperature-dependent experiment, the temperature is controlled by a water bath and temperature-controlling module in a UV-Vis spectrophotometer (Evolution One, Thermo Fisher Scientific). Enzyme aliquots were stored in a water bath at a specific temperature for 10 min, and the pre-mixed reaction buffer without enzyme was incubated inside the spectrophotometer with the temperature module controlled via the software.

Enzyme catalytic CO_2 reduction in solution

The enzyme catalytic reaction was performed in an anaerobic chamber. 20 mM of reduced MV was used as the electron donor, CO_2 was used as the substrate, and 50 μg of FdhF_HycB3Δ159–184 was used as the catalyst in 5 mL of the reaction solution. The control used inactive FdhF_HycB3Δ159-184, which was prepared by boiling for 20 min. After 1.5 h at 30 °C, 100 μL samples were taken, and 5 mM sulfuric acid was added to stop the reaction. The samples were centrifuged at 12000 × *g* for 10 min, and the produced formic acid was detected by high-performance liquid chromatography (HPLC) equipped with a RID–20A detector (Nexera XR, Shimadzu) and an Aminex HPX-87H column (300 × 7.8 mm, Bio-Rad). The mobile phase was 5 mM sulfuric acid, the flow rate was 0.6 mL min^{-1}, and the column temperature was 60 °C.

Electrochemical experiments

Electrochemical measurements were carried out using a potentiostat (CHI660e) and a rotating disk electrode (PINE) in the anaerobic chamber with 100% N_2. A three-electrode system was used. The glassy carbon electrodes (0.07 cm^2) were polished with an alumina slurry (0.05 μm) on a polishing cloth and then sonicated for 1 min in deionized water and ethanol, respectively. A homogenous black CNT was dispersed in water and dimethylformamide (DMF) (ratio of 1:1) to a concentration of 1 g L^{-1}. A 10 μL aliquot of the CNT suspension was drop-cast on a fresh GC electrode, denoted as the CNT/GC electrode. For the bioelectrode, 10 μL of enzyme (1 g L^{-1}) was drop-cast on the CNT/GC electrode and then left to stand at 4 °C for 4 h in an anaerobic chamber. The excess enzyme solution was gently washed off before testing. A platinum wire was used as the counter electrode, and a saturated Ag/AgCl electrode was used as the reference electrode. All applied potentials were converted versus the standard hydrogen electrode (SHE) using the following equation **1** at 25 °C and pH 7.0.

$$ESHE = EAg/AgCl + 197mV \quad (1)$$

Cyclic voltammetry (CV) and linear sweep voltammetry (LSV) analyses were performed at 25 °C in 100 mM HEPES/NaOH and 100 NaCl at pH 7.0. In the pH-dependence experiment, pH was controlled by using a mixed-buffer system [100 mM Acetate, 2-morpholinoethane sulfonic acid (MES), *N*-cyclohexyl-2 aminoethane sulfonic acid (HEPES), and *N*-tris(hydroxymethyl)methyl-3-aminopropanesulfonic acid (TAPS)], and checked following experimentation. The electrode speed was 900 rpm, the scan rate was 5 mV $s^{-1}$for CV, and the scan rate was 1 mV s^{-1} for LSV. The electrolytic cell with a heating layer was used for the temperature optimization experiment. The temperature of the pool is controlled by a flowing water bath, and cyclic voltammetry is performed after reaching the desired temperature, as confirmed by a thermometer test.

Enzyme-electrocatalytic CO_2 reduction to formate

A carbon paper (TORAY, Japan) with an area of 1 cm^2 was used as the supporting enzyme, 100 μL of CNTs (1 g L^{-1}) was added to the carbon paper, and after drying, 95 μL of FdhF_HycB3Δ159-184 solution (1 g L^{-1}) and 5 μL of Nafion (0.02%) were mixed and added to the CNTs to build FdhF_HycB3Δ159-184/carbon nanotube/carbon paper (FdhF_HycB3Δ159-184/CNT/CP) electrode, after which the mixture was allowed to stand at 4 °C for 4 h in an anaerobic chamber. The excess enzyme solution was gently removed. A three-electrode electrochemical cell with FdhF_HycB3Δ159-184/CNT/CP electrode as the working electrode, Ag/AgCl electrode as the reference electrode, and platinum sheet (1 cm^2) as the counter electrode was used. Chronoamperometry was carried out in 5 mL of buffer (100 mM HEPES/NaOH, 100 mM NaCl, pH 7.0) that was continuously fed pure CO_2, and a constant voltage of − 0.5 V vs. SHE was used. For the MET enzyme-electrocatalytic CO_2 reduction to formate, 0.2 mM MV was added to the electrolyte. The formate concentration in the electrolyte was detected by a high-performance liquid chromatography (HPLC) instrument equipped with a RID-20A detector (Nexera XR, Shimadzu) and an Aminex HPX-87H column (300 × 7.8 mm, Bio-Rad). The samples were prepared by adding H_2SO_4 at a final concentration of 5 mM and were centrifuged for 10 min at 12000 × *g*. The mobile phase was 5 mM sulfuric acid, the flow rate was 0.6 mL min^{-1}, and the column temperature was 60 °C. The standard curve for formate was generated by employing known concentrations of formate ranging from 1 to 100 mM. For the ^{1}H-NMR experiment, 400 μL of electrolyte and 200 μL of D_2O were mixed in a clean NMR tube and subsequently

analyzed via one-dimensional ^{1}H liquid NMR spectroscopy (Bruker Advance 600 MHZ). The faraday efficiency (FE) and half-cell cathodic energy efficiency (CEE%) of formate were calculated according to Eqs. 2 and 3[46].

$$FE(\%) = \frac{Q(formate)}{Q(tot)} * 100\% = \frac{c(formate) * V * Z * F}{j * t} \quad (2)$$

where c (formate) is the formate concentration after reaction time t (s), V is the volume of the electrolyte, z is the number of electrons required to reduce CO_2 to formate, and the value is 2, F is the Faraday constant (96,485 C mol^{-1}), and j is the recorded current.

$$CEE(\%) = \frac{1.23 - E(formate)}{1.23 - E(applied)} * FE * 100\% \quad (3)$$

where E (formate) = − 0.42 V *vs.* SHE is the reduction potential of CO_2 reduction to formate at pH 7.0. FE is the faradaic efficiency of formate production. E (applied) = − 0.5 V *vs.* SHE is the applied potential *vs.* SHE.

Dynamic light-scattering assay

A NanoBrook 90plus instrument (Brookhaven Instruments Corporation, USA) was used to analyze the particle size of the proteins. The concentration of all the samples was 0.5 g L^{-1} and the samples were centrifuged for 10 min at 12000 × g and 4 °C. Each measurement was made three times at 25 °C.

ICP-OES analysis

The W element in FdhF_HycB3Δ159-184 was analyzed through ICP-OES (Thermo, ICPOES7200) by Beijing Zhongke Baice Technology Service Co., Ltd.

Bioinformatic methods

All DNA sequences were retrieved from the National Center for Biotechnology Information database, and all protein structures were obtained from the Protein Data Bank. The plasmid map was generated by SnapGene, and the protein structures were visualized using PyMOL. Protein docking analysis was performed via ZDOCK[47].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this Article is available as a Supplementary Information file. Source data are provided in this paper.

References

1. Liu, Z., Wang, K., Chen, Y., Tan, T. & Nielsen, J. Third-generation biorefineries as the means to produce fuels and chemicals from CO_2. *Nat. Catal.* **3**, 274–288 (2020).
2. Li, L., Li, X., Sun, Y. & Xie, Y. Rational design of electrocatalytic carbon dioxide reduction for a zero-carbon network. *Chem. Soc. Rev.* **51**, 1234–1252 (2022).
3. Fang, W. et al. Durable CO_2 conversion in the proton-exchange membrane system. *Nature* **626**, 86–91 (2024).
4. Huang, J. E. et al. CO_2 electrolysis to multicarbon products in strong acid. *Science* **372**, 1074–1078 (2021).
5. Xie, Y. et al. High carbon utilization in CO_2 reduction to multi-carbon products in acidic media. *Nat. Catal.* **5**, 564–570 (2022).
6. Yang, Y. et al. Operando studies reveal active Cu nanograins for CO_2 electroreduction. *Nature* **614**, 262–269 (2023).
7. Zhang, Y.-H. P. J., Zhu, Z., You, C., Zhang, L. & Liu, K. In *vitro* biotransformation (ivBT): definitions, opportunities, and challenges. *Synth. Biol. Eng.* **1**, 10013 (2023).
8. Nielsen, C. F., Lange, L. & Meyer, A. S. Classification and enzyme kinetics of formate dehydrogenases for biomanufacturing via CO_2 utilization. *Biotechnol. Adv.* **37**, 107408 (2019).
9. Moon, M., Park, G. W., Lee, J., Lee, J.-S. & Min, K. Recent progress in formate dehydrogenase (FDH) as a non-photosynthetic CO_2 utilizing enzyme: A short review. *J. CO2 Util.* **42**, 101353 (2020).
10. Sahin, S. et al. Bioelectrocatalytic CO_2 reduction by Mo-dependent formylmethanofuran dehydrogenase. *Angew. Chem. Int. Ed.* **62**, e202311981 (2023).
11. Schuchmann, K. & Muller, V. Direct and reversible hydrogenation of CO_2 to formate by a bacterial carbon dioxide reductase. *Science* **342**, 1382–1385 (2013).
12. Schwarz, F. M., Schuchmann, K. & Muller, V. Hydrogenation of CO_2 at ambient pressure catalyzed by a highly active thermostable biocatalyst. *Biotechnol. Biofuels* **11**, 237 (2018).
13. Reda, T., Plugge, C. M., Abram, N. J. & Hirst, J. Reversible interconversion of carbon dioxide and formate by an electroactive enzyme. *Proc. Natl. Acad. Sci. USA* **105**, 10654–10658 (2008).
14. Bassegoda, A., Madden, C., Wakerley, D. W., Reisner, E. & Hirst, J. Reversible interconversion of CO_2 and formate by a molybdenum-containing formate dehydrogenase. *J. Am. Chem. Soc.* **136**, 15473–15476 (2014).
15. Robinson, W. E., Bassegoda, A., Reisner, E. & Hirst, J. Oxidation-state-dependent binding properties of the active site in a Mo-containing formate dehydrogenase. *J. Am. Chem. Soc.* **139**, 9927–9936 (2017).
16. Li, W. et al. Direct detection of a single [4Fe-4S] cluster in a tungsten-containing enzyme: Electrochemical conversion of CO_2 into formate by formate dehydrogenase. *Carbon Energy* **5**, https://doi.org/10.1002/cey2.304 (2023).
17. Kuk, S. K. et al. NADH-free electroenzymatic reduction of CO_2 by conductive hydrogel-conjugated formate dehydrogenase. *ACS Catal.* **9**, 5584–5589 (2019).
18. Yuan, M. et al. Creating a low-potential redox polymer for efficient electroenzymatic CO_2 reduction. *Angew. Chem. Int. Ed.* **130**, 6692–6696 (2018).
19. Alvarez-Malmagro, J. et al. Bioelectrocatalytic activity of W-formate dehydrogenase covalently immobilized on functionalized gold and graphite electrodes. *ACS Appl Mater. Interfaces* **13**, 11891–11900 (2021).
20. Badiani, V. M. et al. Engineering electro- and photocatalytic carbon materials for CO_2 reduction by formate dehydrogenase. *J. Am. Chem. Soc.* **144**, 14207–14216 (2022).
21. Edwardes Moore, E. et al. Understanding the local chemical environment of bioelectrocatalysis. *Proc. Natl. Acad. Sci. USA* **119**, e2114097119 (2022).
22. Cobb, S. J., Rodríguez-Jiménez, S. & Reisner, E. Connecting biological and synthetic approaches for electrocatalytic CO_2 reduction. *Angew. Chem. Int. Ed.* **63**, e202310547 (2024).
23. Cobb, S. J., Dharani, A. M., Oliveira, A. R., Pereira, I. A. C. & Reisner, E. Carboxysome-inspired electrocatalysis using enzymes for the reduction of CO_2 at low concentrations. *Angew. Chem. Int. Ed.* **62**, e202218782 (2023).
24. Cobb, S. J. et al. Fast CO_2 hydration kinetics impair heterogeneous but improve enzymatic CO_2 reduction catalysis. *Nat. Chem.* **14**, 417–424 (2022).
25. Dietrich, H. M. et al. Membrane-anchored HDCR nanowires drive hydrogen-powered CO_2 fixation. *Nature* **607**, 823–830 (2022).
26. Hartmann, T., Schwanhold, N. & Leimkühler, S. Assembly and catalysis of molybdenum or tungsten-containing formate dehydrogenases from bacteria. *Biochim. Biophys. Acta Proteins Proteom.* **1854**, 1090–1100 (2015).

27. Niks, D. & Hille, R. Molybdenum- and tungsten-containing formate dehydrogenases and formylmethanofuran dehydrogenases: Structure, mechanism, and cofactor insertion. *Protein Sci.* **28**, 111–122 (2018).
28. Schlindwein, C., Giordano, G., Santini, C. L. & Mandrand, M. A. Identification and expression of the Escherichia coli fdhD and fdhE genes, which are involved in the formation of respiratory formate dehydrogenase. *J. Bacteriol.* **172**, 6112–6121 (1990).
29. Leo, F., Schwarz, F. M., Schuchmann, K. & Müller, V. Capture of carbon dioxide and hydrogen by engineered *Escherichia coli*: hydrogen-dependent CO_2 reduction to formate. *Appl. Microbiol. Biotechnol.* **105**, 5861–5872 (2021).
30. Thomé, R. et al. A sulfurtransferase is essential for activity of formate dehydrogenases in *Escherichia coli*. *J. Biol. Chem.* **287**, 4671–4678 (2012).
31. Arnoux, P. et al. Sulphur shuttling across a chaperone during molybdenum cofactor maturation. *Nat. Commun.* **6**, 6148 (2015).
32. Mulder, D. W. et al. Stepwise [FeFe]-hydrogenase H-cluster assembly revealed in the structure of HydAΔEFG. *Nature* **465**, 248–251 (2010).
33. Britt, R. D., Rao, G. & Tao, L. Biosynthesis of the catalytic H-cluster of [FeFe] hydrogenase: the roles of the Fe–S maturase proteins HydE, HydF, and HydG. *Chem. Sci.* **11**, 10313–10323 (2020).
34. Britt, R. D., Rao, G. & Tao, L. Bioassembly of complex iron–sulfur enzymes: hydrogenases and nitrogenases. *Nat. Rev. Chem.* **4**, 542–549 (2020).
35. Tyagi, A. K. et al. High-yield expression of heterologous [FeFe] hydrogenases in *Escherichia coli*. *PLoS ONE* **5**, e15491 (2010).
36. Min, K. et al. Elevated conversion of CO_2 to versatile formate by a newly discovered formate dehydrogenase from *Rhodobacter aestuarii*. *Bioresour. Technol.* **305**, 123155 (2020).
37. Ebrahimpour, A., Rahman, R. N. Z. R. A., Basri, M. & Salleh, A. B. High level expression and characterization of a novel thermostable, organic solvent tolerant, 1,3-regioselective lipase from *Geobacillus* sp. strain ARM. *Bioresour. Technol.* **102**, 6972–6981 (2011).
38. Dietrich, H. M. & Müller, V. Ferredoxin as a physiological electron donor for carbon dioxide fixation to formate in a bacterial carbon dioxide reductase. *ACS Catal.* **13**, 12374–12382 (2023).
39. Cobb, S. J. et al. A photoelectrochemical-thermoelectric device for semi-artificial CO_2 fixation employing full solar spectrum utilization. *Device* 100505, https://doi.org/10.1016/j.device.2024.100505 (2024).
40. Wojcik, M. & Miłek, J. A new method to determine optimum temperature and activation energies for enzymatic reactions. *Bioprocess Biosyst. Eng.* **39**, 1319–1323 (2016).
41. Yang, J. Y., Kerr, T. A., Wang, X. S. & Barlow, J. M. Reducing CO_2 to HCO_2^- at mild potentials: lessons from formate dehydrogenase. *J. Am. Chem. Soc.* **142**, 19438–19445 (2020).
42. Meneghello, M. et al. Formate dehydrogenases reduce CO_2 rather than HCO_3^-: An electrochemical demonstration. *Angew. Chem. Int. Ed.* **60**, 9964–9967 (2021).
43. Parkhill, K. L. & Gulliver, J. S. Indirect measurement of oxygen solubility. *Water Res.* **31**, 2564–2572 (1997).
44. Yu, X., Niks, D., Mulchandani, A. & Hille, R. Efficient reduction of CO_2 by the molybdenum-containing formate dehydrogenase from *Cupriavidus necator* (*Ralstonia eutropha*). *J. Biol. Chem.* **292**, 16872–16879 (2017).
45. Chang, S., Xuan, Y., Duan, J. & Zhang, K. High-performance electroreduction CO_2 to formate at Bi/Nafion interface. *Appl. Catal. B* **306**, 121135 (2022).
46. Li, Y. et al. Enhancing local CO_2 adsorption by L-histidine incorporation for selective formate production over the wide potential window. *Angew. Chem. Int. Ed.* **62**, e202313522 (2023).
47. Zhang, W., Meng, Q. Z., Tang, J. J. & Guo, F. Exploring effectiveness of ab-initio protein-protein docking methods on a novel antibacterial protein complex dataset. *Brief. Bioinform.* **22**, 13 (2021).
48. Miller, M. et al. Interfacing Formate dehydrogenase with metal oxides for the reversible electrocatalysis and solar-driven reduction of carbon dioxide. *Angew. Chem. Int. Ed.* **58**, 4601–4605 (2019).
49. Edwardes Moore, E. et al. A semi-artificial photoelectrochemical tandem leaf with a CO_2 -to-formate efficiency approaching 1%. *Angew. Chem. Int Ed. Engl.* **60**, 26303–26307 (2021).

Acknowledgements

This study was supported by Tianjin Synthetic Biotechnology Innovation Capacity Improvement Projects (TSBICIP-CXRC-024 and TSBICIP-KJGG-007 to L.Z.), CAS Project for Young Scientists in Basic Research (YSBR-072-3 to L.Z.), and Strategic Priority Research Program of the Chinese Academy of Sciences (XDC0120103 to L.Z.).

Author contributions

Experiment, data analysis and writing the initial manuscript were performed by W.L. and K.Z. analyzed data. J.L. provided experimental material. The anaerobic operation was supervised by J.S. Bioelectrocatalytic operation was supervised by Y.W., M.Z., and H.C. Project conception, fund supporting, supervision, and writing the paper were performed by L.Z.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53946-3.

Correspondence and requests for materials should be addressed to Lingling Zhang.

Peer review information *Nature Communications* thanks Samuel Cobb and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：尹若鹏
方向：生物制造
单位：中国科学院微生物研究所

题目：益生菌细胞壁改善酒精性肝病

ADVANCED SCIENCE Open Access
www.advancedscience.com

Postbiotics From *Lactobacillus Johnsonii* Activates Gut Innate Immunity to Mitigate Alcohol-Associated Liver Disease

*Ruopeng Yin, Tao Wang, Jingzu Sun, Huanqin Dai, Yuting Zhang, Ningning Liu, and Hongwei Liu**

Prolonged alcohol consumption disrupts the gut microbiota and the immune system, contributing to the pathogenesis of alcohol-associated liver disease (ALD). Probiotic-postbiotic intervention strategies can effectively relieve ALD by maintaining gut homeostasis. Herein, the efficacy of heat-killed *Lactobacillus johnsonii* (HKLJ) in mitigating alcoholic liver damage is demonstrated in mouse models of ALD. The gut-liver axis is identified as a pivotal pathway for the protective effects of *L. johnsonii* against ALD. Specifically, HKLJ is found to upregulate the expression of intestinal lysozymes, thereby enhancing the production of immunoregulatory substances from gut bacteria, which subsequently activated the Nucleotide-binding oligomerization domain 2 (NOD2)-interleukin (IL-23)-IL-22 innate immune axis. The elevated IL-22 upregulated the antimicrobial peptide synthesis to maintain intestinal homeostasis and moreover activated the Signal transducer and activator of Transcription3 (STAT3) pathway in the liver to facilitate the repair of hepatic injuries. The heat-killed *L. johnsonii* provoked immunity helps correct the gut microbiota dysbiosis, specifically by reversing the reduction of butyrate-producing bacteria (such as *Faecalibaculum rodentium*) and the expansion of opportunistic pathogens (such as *Helicobacter* sp. and *Pichia kudriavzevii*) induced by ethanol. The findings provide novel insights into the gut microbiota-liver axis that may be leveraged to enhance the treatment of ALD.

1. Introduction

Long-term alcohol abuse and the resulting alcohol-associated liver disease (ALD) are recognized as the major contributors to approximately one-quarter of liver disease-related deaths worldwide.[1,2] Excessive alcohol intake impairs the gut barrier function and immune homeostasis in patients with ALD which leads to the release of microbial toxins or the translocation of pathogens from the gut into the liver, exacerbating ALD.[3–5] Inspired by these mechanisms mediated via the gut-liver axis, intervention strategies such as fecal microbiota transplantation (FMT), phage therapy, and the oral supplementation of prebiotics, probiotics, and postbiotics are being used as therapeutic regimens for liver diseases by improving gut barrier function and immune homeostasis.[6–9]

The currently available live microbial formulations have shown marginal or highly variable in vivo efficacies in patients.[10,11] In contrast to live probiotics, inactivated microorganisms and/or their components (called postbiotics) show more advantages

R. Yin, T. Wang, J. Sun, H. Dai, Y. Zhang, H. Liu
State Key Laboratory of Mycology
Institute of Microbiology
Chinese Academy of Sciences
Beijing 100101, China
E-mail: liuhw@im.ac.cn

R. Yin, H. Dai, Y. Zhang, H. Liu
Medical School
University of Chinese Academy of Sciences
Beijing 100049, China

N. Liu
CAS Key Laboratory of Pathogenic Microbiology and Immunology
Institute of Microbiology
Chinese Academy of Sciences
Beijing 100101, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202405781

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.202405781

in good stability, safety, and efficacy repeatability. Accumulating investigations on postbiotics have paved the way to unlocking the substances and mechanisms underlying the beneficial effects of gut microbiota on animals and human beings. For example, peptidoglycan fragments produced by commensal *Enterococcus* have been shown to enhance immunotherapy and host defense,[12,13] polysaccharide A from the cell wall of specific *Bacteroides* modulates anti-inflammatory T cell populations,[14,15] and surface polysaccharides of *Bifidobacterium* attenuate inflammation by enhancing intestinal regulatory T cell production.[16] In studies of ALD, a flagellin from *Roseburia intestinalis* produced an anti-ALD effect by activating the Toll-like receptors (TLR) 5 receptor to restore gut barrier function, and postbiotic nanoparticles derived from *Lactobacillus rhamnosus* GG alleviated ALD symptoms via regulation of intestinal microRNA194-FXR signaling.[17] These results suggest that gut microbial postbiotics are promising for the prevention and treatment of ALD by interacting with host intestinal cells, especially immune cells.

Metagenomic analysis revealed a relatively high abundance of *Lactobacillus* species in the gut of healthy individuals.[18] *Lactobacillus* species and strains have been shown to have diverse benefits in enhancing host immunity through distinctive modes of action.[19] Among them, *Lactobacillus johnsonii*, a commonly detected probiotic in the intestines of healthy hosts possesses important bioactivities in enhancing the efficacy of immunotherapy, improving gut barrier functions, and reducing fatty liver disease.[20–22] Interestingly, inactivated *L. johnsonii* was demonstrated to protect against both sepsis-mediated liver injury and acute alcohol-induced liver injury.[23,24] However, the protective mechanisms of inactivated *L. johnsonii* have yet to be elucidated.

In this work, the anti-ALD effect of gut commensal *L. johnsonii* and its postbiotics was demonstrated in two ALD mouse models. The protective effects of *L. johnsonii* postbiotics were mechanistically linked to the activation of the NOD2-IL-22 signaling pathway in intestinal immune cells and the upregulation of the IL-22-regulated STAT3 pathway in the liver. Moreover, supplementation with *L. johnsonii* postbiotics helped maintain gut microbiota homeostasis under long-term ethanol feeding, suppressing the ethanol-induced overgrowth of opportunistic gut pathogens.

2. Results

2.1. *L. Johnsonii* Protects Against Liver Injury and Maintains Gut Integrity in ALD

Fresh *L. johnsonii* (LJ) was pasteurized to generate heat-killed *L. johnsonii* (HKLJ). Scanning electron microscope (SEM) revealed that HKLJ (right) had broken cell walls according to its micromorphology (**Figure 1A**; Figure S1, Supporting Information). To investigate the protective effect of LJ on ALD injuries, mice fed a normal diet were pretreated with either live LJ (2×10^8 CFUs/day) or HKLJ for two weeks and then fed a Lieber-DeCarli ethanol liquid diet in a chronic manner while being continuously administrated LJ or HKLJ at the same dosage (Figure 1B). Pair-fed mice and ethanol-fed mice were used as negative controls and positive controls, respectively. After six weeks of feeding with ethanol, 33.3% (4 out of 12) of mice died in the ethanol-fed group, and the surviving mice showed a 10.6% reduction in body weight compared to that at the start of chronic feeding. Supplementation of ethanol-fed mice with LJ and HKLJ suppressed body weight loss and protected the mice from death (Figure S2A,B, Supporting Information). Both LJ and HKLJ administration alleviated liver injury, as evidenced by significantly decreased alanine aminotransferase (ALT) and aspartate aminotransferase (AST) levels, respectively, in comparison with those in the ethanol-fed model group (Figure 1C). Importantly, HKLJ more potently reduced AST level than LJ. In addition, supplementation with LJ and HKLJ alleviated the ethanol-induced increase in the liver index and hepatic triglyceride content and recovered hepatic antioxidant potency as reflected by a 38.8% and 38.5% increase in the ratio of reduced glutathione (GSH) to oxidized glutathione (GSSG), respectively (Figure 1D–F). Histopathological analysis revealed that mice fed ethanol exhibited severe hepatic steatosis (indicated by red arrows), inflammatory cell infiltration, lipid deposition, and fibrosis. Notably, both the LJ and HKLJ interventions significantly attenuated these pathological changes (Figure 1G–I). Further qPCR assays showed that supplementation with LJ and HKLJ down-regulated the hepatic gene expression of inflammation-related cytokines including interleukin (IL)-1β, tumor necrosis factor-alpha (TNFα), and inducible nitric oxide synthase (iNOS), as well as lipid synthesis rate-limiting enzymes, such as fatty acid synthetase (FAS) and acetyl-CoA carboxylase 1 (ACC1) (Figure S2C, Supporting Information). HKLJ has the same effect as LJ in lowering the levels of inflammation-related cytokines. Immunofluorescence staining of F4/80-positive hepatic macrophages further supported the reduction in hepatic inflammation in the HKLJ-treated group (Figure S2D, Supporting Information). The protective effect of HKLJ against ALD was also verified in a chronic plus-binge alcohol feeding model (NIAAA model), as shown by significant reductions in the liver index, and levels of ALT and AST (Figure S3, Supporting Information). These findings suggest that pasteurization may enhance the beneficial effects of LJ on ALD injuries.

Alcohol consumption impairs gut barrier functions, which consequently increases circulating endotoxemia and aggravates alcoholic steatohepatitis.[25] In the current work, the LJ or HKLJ treatments significantly reversed the negative effects of ethanol feeding, as indicated by a notable increase in the levels of lipopolysaccharide (LPS) and pro-inflammatory cytokines (TNFα, IL-6, and IL-1β) in the circulation of mice (**Figure 2A–D**). Notably, a 32.3% and 42.5% increase in anti-inflammatory cytokine (IL-10) concentrations was observed in the LJ or HKLJ-treated group, respectively (Figure 2E). As indicated by H&E staining and Alcian blue & periodic acid-Schiff (AB-PAS) staining, supplementation with LJ or HKLJ resulted in a recovery of the ileal villus length, the crypt depth (divided by a dotted line between villi and crypts), and the number of Paneth cells in the intestinal crypts compared with ALD controls (Figure 2F–I). Consistent with the intestinal pathological observation, supplementation with LJ or HKLJ alleviated gut barrier dysfunction, as indicated by tight junctions (*Zo-1*, *Zo-2*, and *Occludin*) and gut defense molecules (*Lysozyme-1*, *Lysozyme-2*, and *Defensin-5*) in the qPCR assay (Figure 2J,K). Taken together, the above data supported that LJ and HKLJ supplementation restored ethanol-disrupted gut barrier function.

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

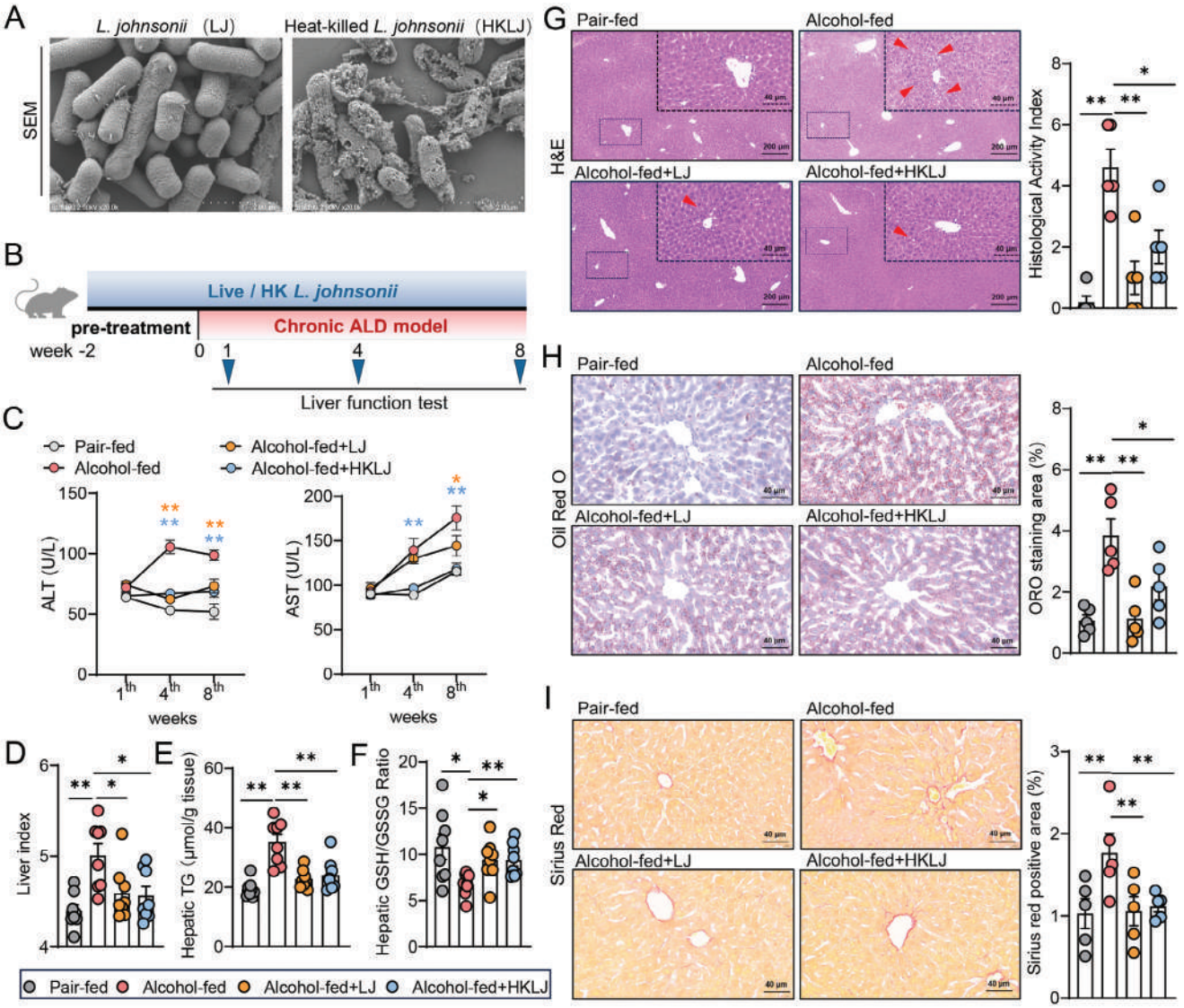


Figure 1. *L. johnsonii* and heat-killed strain attenuate liver injury. A) Representative SEM images of *L. johnsonii* (LJ) and heat-killed strains (HKLJ). B) Schematic diagrams of chronic alcohol feeding-induced ALD mouse model with treatment. C) Serum levels of ALT and AST at weeks 1, 4, and 8 during the intervention. D) Liver-to-body weight ratio at the time of sacrifice. E) Triglyceride levels in liver tissue. F) Ratio of GSH and GSSG contents in liver tissue. G) Representative H&E staining images of liver sections and quantitative analysis of liver histologic scores. H) Staining of mouse liver sections for oil red O and quantitative analysis. I) Staining of mouse liver sections for Sirius red and quantitative analysis. Each dot represents data from one animal (biological replicate). Biological replicate numbers in each group are as follows: (Figure 1C, D, E, F) $n = 8$, (Figure 1G, H, I) $n = 5$. Statistical significance was determined using one-way analysis of variance, and data represent the mean ± s.e.m., unless otherwise indicated. Statistically significant differences are shown with asterisks as follows: $^{*}p$ values < 0.05; $^{**}p$ values < 0.01; ns, no significant differences. SEM: scanning electron microscope; ALT: alanine aminotransferase; AST: aspartate aminotransferase; GSH: reduced glutathione; GSSG: oxidized glutathione; H&E: hematoxylin and eosin.

2.2. Heat-killed *L. Johnsonii* Protects Against Alcohol-Induced Gut Microbiota Dysbiosis

Ethanol-induced impairment of gut integrity is linked to gut microbiota dysbiosis in individuals with ALD.[26] Given that supplementation with the postbiotics (HKLJ) rescued ethanol-disrupted gut barrier functions and promoted the secretion of antimicrobial peptides, we performed 16S ribosomal RNA and ITS gene sequencing of cecum contents to detect changes in the gut microbiome, respectively. HKLJ supplementation significantly increased the bacterial richness according to the Simpson's index compared with that in the ethanol-fed group (**Figure 3A**) and greatly altered the bacterial composition according to the beta diversity determined via PCoA (weighted UniFrac distance) (Figure 3B; Figure S4A, Supporting Information). Linear discriminant analysis of effect size (LEfSe) was used to detect dominant differences in the predominance of bacterial communities in the HKLJ-treated group. At the genus level, LEfSe analysis revealed specific enrichment of the genera *Psychrobacter*, *Pseudomonas*, *Eggerthella*, and *Pseudobutyrivibrio* in the HKLJ-treated group, comparing with the dominance of *Helicobacter* and *Muribaculum* in the alcohol-fed group, which was further confirmed

 © 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

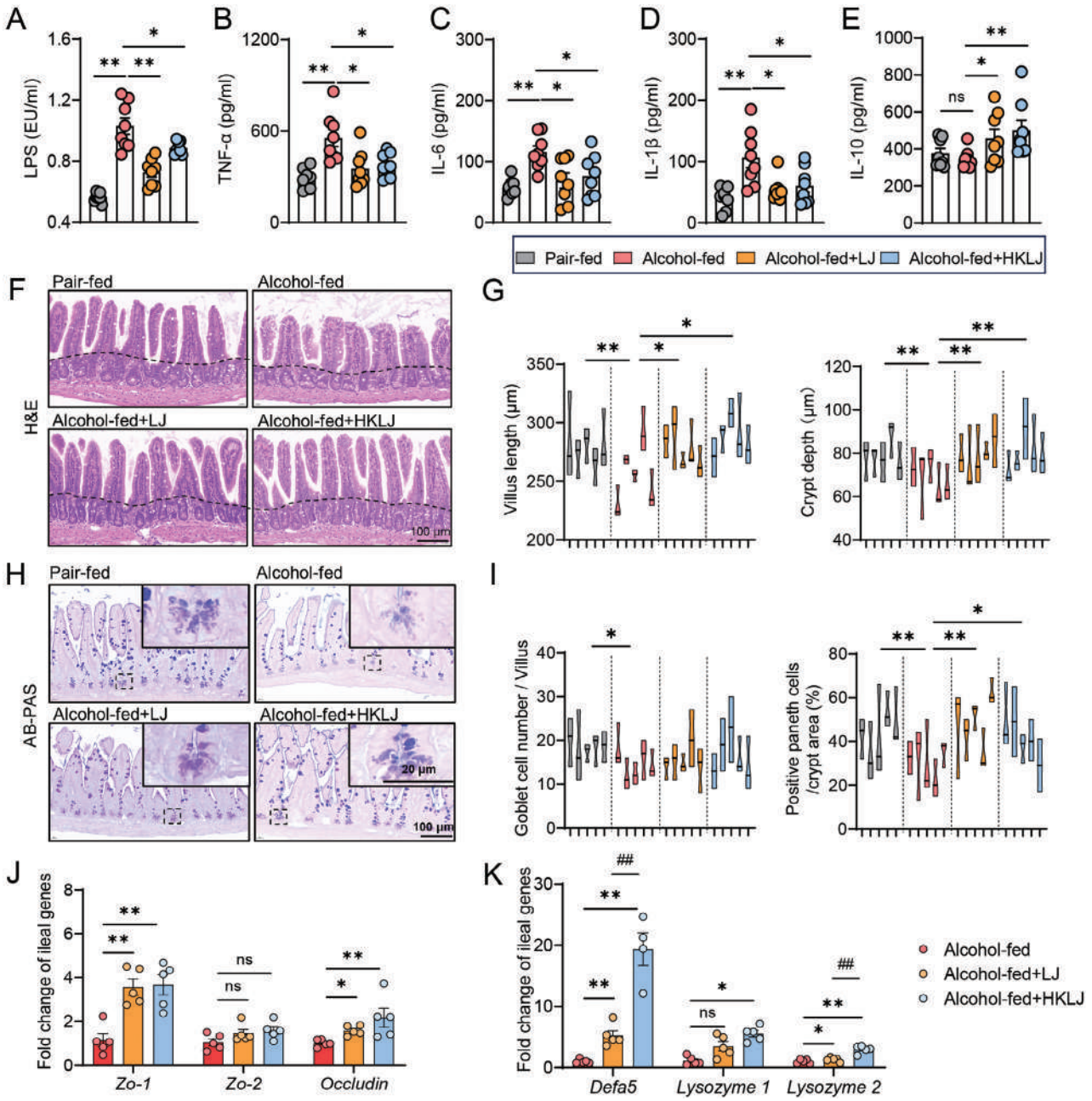


Figure 2. *L. johnsonii* and heat-killed strains maintain intestinal integrity. A) LPS levels in serum. B) Levels of TNF-α in serum. C) Levels of IL-6 in serum. D) Levels of IL-1β in serum. E) Levels of IL-10 in serum. F) Representative images of H&E staining of mouse ileum. G) Measurements of villi and crypts in the mouse ileum. H) Representative images of AB-PAS staining of mouse ileum. I) Measurement of goblet cells and Paneth cells in mouse ileum. J) Ileal mRNA expression levels of *Zo-1*, *Zo-2*, and *Occludin*, normalized to *Gapdh*. K) Ileal mRNA expression levels of *Defa5*, *Lysozyme1*, and *Lysozyme2*, normalized to *Gapdh*. Each violin plot representing one animal, and at least 3 villi/crypts were counted per animal (Figure 2G,I). Each dot represents data from individual biological replicate. Biological replicate numbers in each group are as follows: (Figure 2A–E) $n = 8$, (Figure 2J–K) $n = 4$-5. Statistical significance was determined using one-way analysis of variance, and data represent the mean ± s.e.m., unless otherwise indicated. Statistically significant differences are shown with asterisks as follows: $^{*}p$ values < 0.05; $^{**}p$ values < 0.01; ns, no significant differences. AB-PAS: Alcian blue-periodic acid Schiff; *Gapdh*: glyceraldehyde-3-phosphate dehydrogenase.

by relative abundance analysis of the genera (Figure S4B–D, Supporting Information). At the species level, random forest analysis indicated an increased proportion of *Bifidobacterium pseudolongum*, *Mucispirillum schaedleri*, *Enterorhabdus* sp., *Faecalibaculum rodentium*, and *Flavonifractor plautii* in the HKLJ-supplemented group (Figure 3C,D). Therein, we found that in the gut of mice treated with HKLJ, the abundance of *F. rodentium* was ≈30-fold greater than that in the ALD controls, and *B. pseudolongum*, which was not detected in ALD controls, significantly expanded (Figure 3E).

 © 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

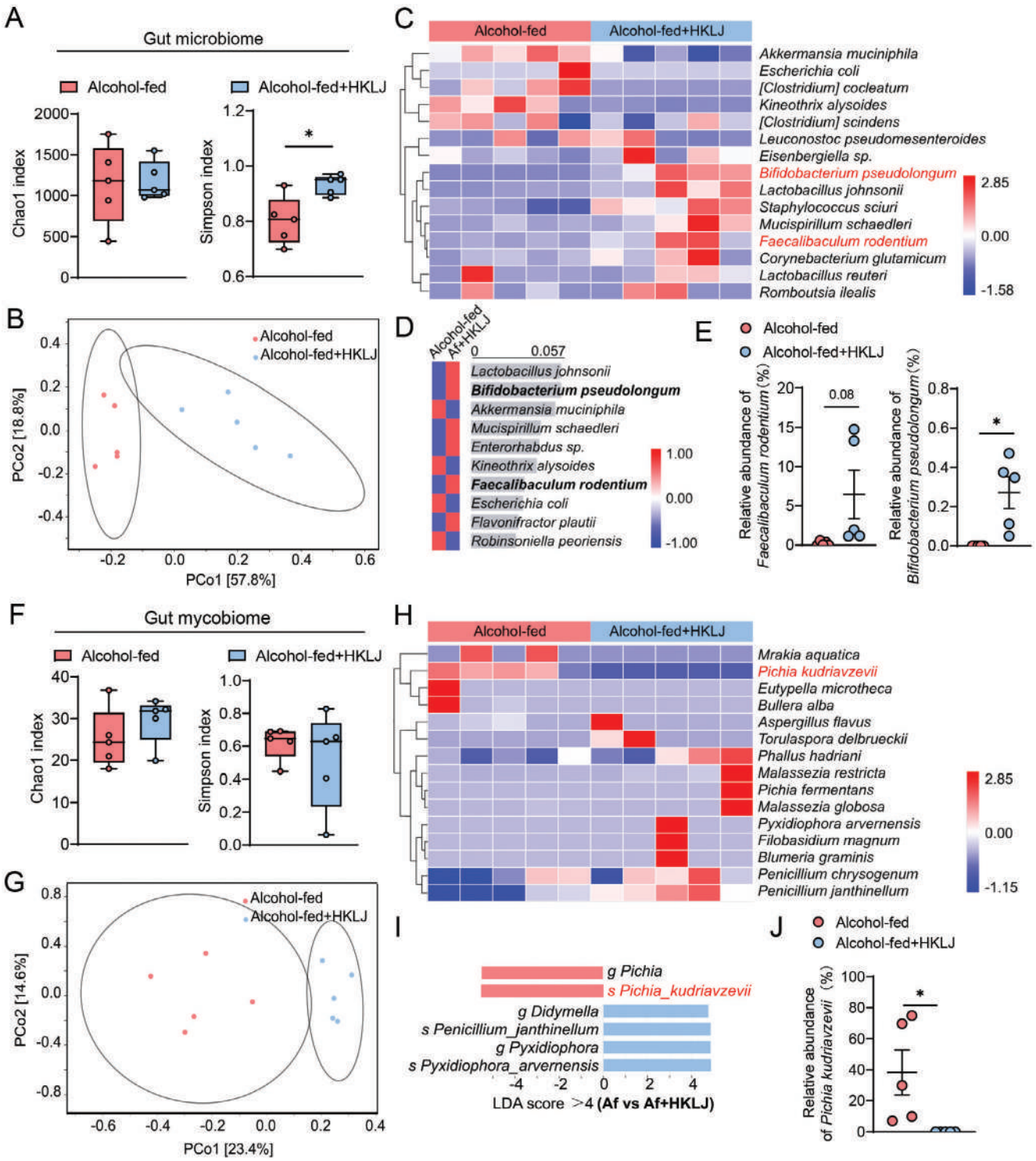


Figure 3. Heat-killed *L. johnsonii* changes gut bacterial and fungal composition. A) Chao1 index and Simpson index (alpha diversity) of the gut bacterial composition. B) PCoA (beta diversity) plot of gut bacterial composition generated using unweighted Unifrac distance. C) Heatmap for the gut bacterial composition in species taxonomic level. D) Based on the Random Forest analysis, the top 10 species in terms of importance scores for differences between groups were ranked with decreasing scores from top to bottom. E) The relative abundance of *Faecalibaculum rodentium* and *Bifidobacterium pseudolongum* in the gut. F) Chao1 index and Simpson index (alpha diversity) of the gut fungal composition. G) PCoA (beta diversity) plot of gut fungal composition generated using Jaccard distance. H) Heatmap for the gut bacterial composition in species taxonomic level. I) LDA score representing the enriched taxon in the gut mycobiome between two groups. Only LDA scores (>4) are shown. Red indicates enriched taxa in the alcohol-fed group; blue indicates enriched taxa in the HKLJ group. J) The relative abundance of *Pichia kudriavzevii* in the gut. Each dot represents data from individual biological replicate. Five biological replicate numbers in each group. Statistical significance was determined using two-tailed Mann-Whitney U test. Data represent

 © 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The expansion of gut fungi characterizes intestinal homeostasis disorders in patients with ALD.[27] Analysis of Chao1 and Simpson's index showed similar alpha diversity in the mycobiome between the HKLJ-treated group and the ethanol-fed group (Figure 3F). Further beta diversity analysis indicated a shift in the mycobiome after HKLJ supplementation (Figure 3G; Figure S5, Supporting Information). It was found that long-term ethanol intake changed the gut mycobiome, especially by enriching the alcohol-tolerant /producing fungus *Pichia kudriavzevii* (Figure 3H–J). Treatment with HKLJ effectively suppressed the alcohol-induced overgrowth of *P. kudriavzevii* in the gut (Figure 3H–J). To further compare the influences of the gut microbiota from the HKLJ-treated ALD mice and the ethanol-fed model mice on alcoholic liver injury, fecal microbiota transplantation (FMT) was performed in NIAAA pseudo germ-free animals (**Figure** 4A). The antibiotics-treated mice receiving faecal microbiota from the HKLJ-treated ALD mice showed much lower levels of ALT, AST, and hepatic triglyceride, and less hepatic fat deposition, as compared to the antibiotics-treated mice receiving faecal microbiota from ethanol-fed model mice (Figure 4B–F).

Significant reduction of butyrate level in the gut was characterized as one of the main features of gut dysbiosis in ALD patients and animals.[28,29] In this study, HKLJ supplementation prevented this decrease (Figure 4G; Figure S6, Supporting Information). We cultured two commensals enriched in the HKLJ-treated group, *F. rodentium*, and *B. pseudolongum*, and assayed their ability to generate SCFAs. As a result, *F. rodentium* was confirmed to be a butyric acid-producer, and *B. pseudolongum* was identified as an acetic acid-producing bacterium (Figure 4H). Intestinal butyrate plays important roles in maintaining hypoxia in the lumen by activating the epithelial PPARγ to control the expansion of opportunistic gut pathogens and retaining gut barrier functions.[30,31] A higher protein level of nuclear PPARγ was detected in the intestinal epithelia of HKLJ-treated mice than in those of ethanol-fed model mice (Figure 4I). Consistent with the changes in PPARγ protein expression, HKLJ supplementation reduced the mRNA levels of intestinal *Nos2* (Figure 4J). In the following assay using the NIAAA model, supplementation with butyrate sodium produced beneficial effects against ALD similar to those of FMT from HKLJ (Figure 4K–O). Butyric acid contributes to the alleviation of ALD injuries by suppressing the growth of opportunistic gut pathogens and maintaining gut barrier integrity. Based on the above analysis, supplementation with HKLJ in ethanol-fed mice is beneficial to correct ethanol-induced gut dysbiosis.

2.3. Enhanced IL-22 Production Contributes to the Benefits of Heat-killed *L. Johnsonii*

To identify the mechanism by which HKLJ alleviates ALD, we performed RNA-seq data analysis of intestinal gene expression profiles (**Figure** 5A). A total of 10276 co-occurring genes were found between the alcohol-fed group (n = 4) and the HKLJ group (n = 4), of which 1177 genes were differentially expressed (Figure 5B). This difference caused changes in the transcriptional profiles (Figure S7A,B). Gene set enrichment analysis (GSEA) revealed the upregulated tight junction-associated pathways in the HKLJ-treated group (Figure S7C, Supporting Information), which further confirmed the reinforcement of gut barrier function. Among the differentially expressed genes (DEGs), 429 genes were up-regulated and 748 genes were down-regulated after treatment (Figure 5C). In the HKLJ-upregulated gene set, two antimicrobial peptide genes (*Reg3g* and *Reg3b*) were noted to be significantly elevated, respectively (Figure 5D; Figure S7D, Supporting Information). Early studies have illustrated that IL-22 is a critical factor for regulating the expression of Reg3 family proteins and maintaining gut integrity.[32] Here, we measured the expression of intestinal IL-22 by immunofluorescence, as well as the levels of IL-22 in the small intestine tissue and circulation by ELISA. Up-regulated expression and higher levels of IL-22 were detected in the HKLJ-treated group (Figure 5E–G). Intestinal IL-22 is mainly produced by group 3 innate lymphoid cells (ILC3s) in the gut.[33] Further flow cytometric analysis of intestinal lamina propria cells showed that the number of IL-22$^+$ ILC3s (gated in RORγT$^+$) was markedly increased after HKLJ supplementation (Figure 5H; Figure S8, Supporting Information). In an early ALD study, IL-22 was reported to alleviate ALD injury via upregulating hepatic signal transducer and activator of transcription (STAT) 3.[34] Herein, the gene expression of hepatic IL-22 receptors (*Il22ra* and *Il10rb*) was up-regulated 1.8 fold and 1.9 fold, respectively (Figure 5I). Compared to that in ethanol-fed mice, an activation of the STAT3 pathway in the HKLJ-treated group was verified by the relatively higher level of phosphorylated STAT3 in the liver (Figure 5J). Correspondingly, the mRNA levels of genes downstream of STAT3 and involved in hepatic cell proliferation and repair were increased (Figure 5K). Overall, HKLJ is capable of upregulating intestinal IL-22 secretion, thus reinforcing gut barrier function and repairing alcohol-induced liver injury.

2.4. Cell Precipitate from *L. Johnsonii* Alleviate ALD Injuries via Activation of NOD2-IL-22 Signaling

In order to gain insight into the effective components of HKLJ, HKLJ was suspended in a buffer solution and separated into obtain water-soluble extracts (WSE) and cell precipitate (CP), as shown in the schematic diagram (**Figure** 6A). In the NIAAA model mice, both HKLJ and CP significantly reduced the liver index, ALT, AST, and hepatic triglycerides, whereas WSE from HKLJ had little effect (Figure 6B–E). Oil red O staining of liver sections also illustrated that CP from HKLJ significantly decreased fat deposition in the livers of ALD (Figure 6F). Consistent with HKLJ, CP from HKLJ promoted IL-22 secretion, whereas WSE did not influence IL-22 in the gut (Figure 6G). Meanwhile, CP from HKLJ markedly upregulated the transcriptional levels of tight junction proteins, gut defensins, lysozyme, and antimicrobial peptides (*Reg3g* and *Reg3b*), similar to those in HKLJ (Figure 6H).

the mean ± s.e.m. For box-and-whiskers plots, the center line represents the median, the bounds of the box represent quartiles and the whiskers represent min to max. Statistically significant differences are shown with asterisks as follows: $^*p$ values < 0.05; $^{**}p$ values < 0.01; ns, no significant differences. PCoA: Principal co-ordinates analysis.

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

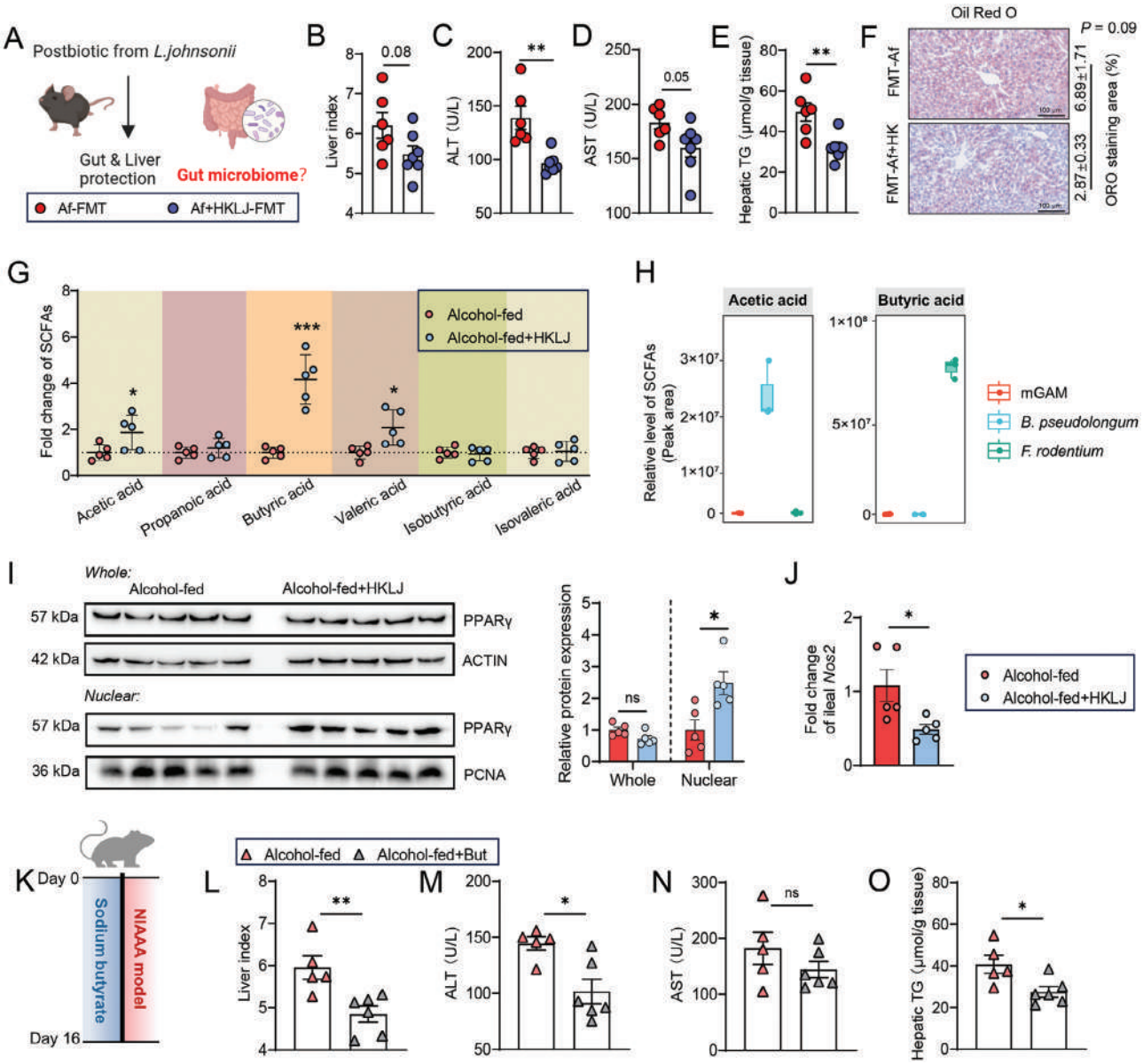


Figure 4. Microbial butyric acid-mediated intestinal PPARγ helps to suppress ALD. A) Schematic diagrams of Figure 4 (created by BioRender.com). B) Liver-to-body weight ratio at the time of sacrifice. C) Serum levels of ALT. D) Serum levels of AST. E) Triglyceride levels in liver tissue. F) Staining of liver sections for oil red O and quantitative analysis. The physiological & biochemical indices and histopathology in (B–F) were measured using serum and liver from fecal microbiota transplant recipient mice from the NIAAA model. G) The relative levels of SCFAs in the fecal samples. H) The levels of acetic acid and butyric acid in the cultured supernatant of *F. rodentium* and *B. pseudolongum*. I) The protein expression of PPARγ in the intestinal epithelium lysate and the cell nucleus lysate by western blot, normalized to β-actin or PCNA (nucleus). J) Ileal mRNA expression levels of *Nos2*, normalized to *Gapdh*. K) Schematic diagrams of NIAAA mouse model with butyrate treatment. L) Liver-to-body weight ratio at the time of sacrifice. M) Serum levels of ALT. N) Serum levels of AST. O) Triglyceride levels in liver tissue. The physiological & biochemical indices in (L–O) were measured using serum and liver from NIAAA model mouse with sodium butyrate treatment. Each dot represents data from individual biological replicate. Biological replicate numbers in each group are as follows: (Figure 4B–E) n = 6-7, (Figure 4F) n = 3-4, (Figure 4G,I,J) n = 5, (Figure 4H) n = 3, (Figure 4L–O) n = 5-6. Statistical significance was determined using unpaired Student's two-tailed t test. Data represent the mean ± s.e.m. For box-and-whiskers plots, the center line represents the median, the bounds of the box represent quartiles and the whiskers represent min to max. Statistically significant differences are shown with asterisks as follows: $^{*}p$ values < 0.05; $^{**}p$ values < 0.01; ns, no significant differences. FMT: Fecal microbiota transplantation.

In light of the change in the lysozyme expression, we deduced that HKLJ and CP enhance the expression and secretion of lysozyme from the host intestine. Lysozymes, mainly present in the intestinal lumen, facilitate the release of antigenic components from bacteria, thereby eliciting an immune response (**Figure** 7A). Subsequently, lysozyme levels in ileal tissues and feces of LJ- and HKLJ-treated ALD mice were assayed, revealing that postbiotics administration significantly increased the level of lysozyme in the intestinal mucosa and lumen (Figure 7B–C). At the interface between the mucosa and the microbiome, a high

 © 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

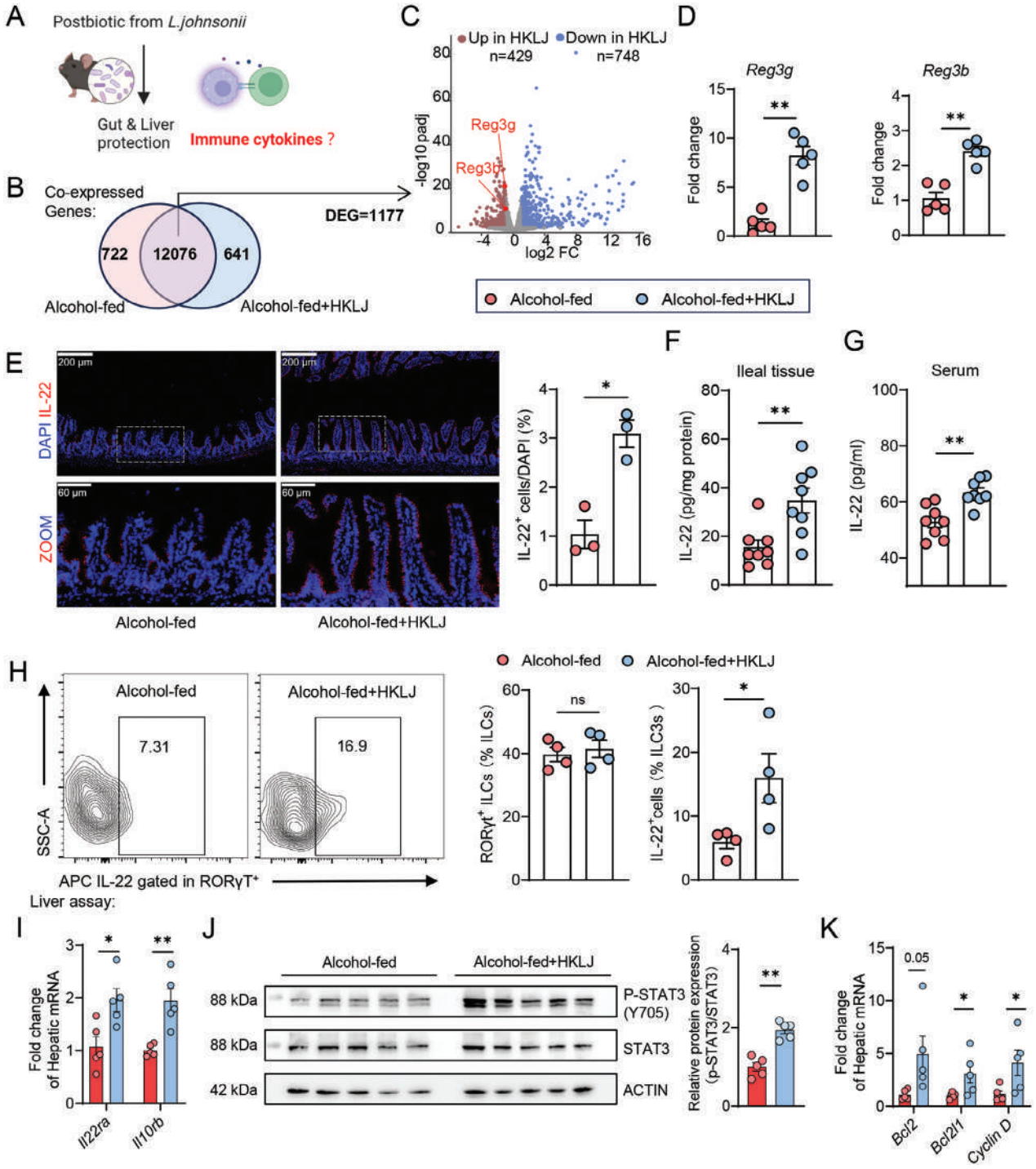


Figure 5. Heat-killed *L. johnsonii* facilitates intestinal secretion of antimicrobial peptides and hepatocyte repair by promoting IL-22 production. A) Schematic diagrams of Figure 5 (created by BioRender.com). B) Intestinal co-expressed genes in ALD mice with administration of vehicle or HKLJ. C) Volcano plots of differentially expressed genes (DEGs) in mouse intestine (blue, downregulated; red, upregulated; change > 2-fold, $p < 0.05$). D) Ileal mRNA expression levels of *Reg3 g* and *Reg3b*, normalized to *Gapdh*. E) Representative immunofluorescence staining of IL-22 and quantitative analysis for IL-22-postive cells in ileums. F) Levels of IL-22 in ileal tissue. G) Levels of IL-22 in serum. H) Representative flow cytometry plots of IL-22-postive ILC3s in the ileal lamina propria. I) Hepatic mRNA expression levels of *Il22ra, Il10rb*, normalized to *Gapdh*. J) Hepatic protein expression of p-STAT3 and STAT3 in the mice by western blot assay, normalized to β-actin. K) Hepatic mRNA expression levels of *Bcl2, Bcl2l1*, and *Cyclin D*, normalized to *Gapdh*. Each dot represents data from individual biological replicate. Biological replicate numbers in each group are as follows: (Figure 5B-5C) $n = 4$, (Figure 5D) $n = 5$, (Figure 5E) $n = 3$, (Figure 5F and 5G) $n = 8$, (Figure 5H) $n = 4$, (Figure 5I-5K) $n = 5$. Statistical significance was determined using unpaired Student's two-tailed *t* test, and data represent the mean ± s.e.m., unless otherwise indicated. Statistically significant differences are shown with asterisks as follows: $^{*}p$ values < 0.05; $^{**}p$ values < 0.01; ns, no significant differences. ILC3s: group 3 innate lymphoid cells.

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

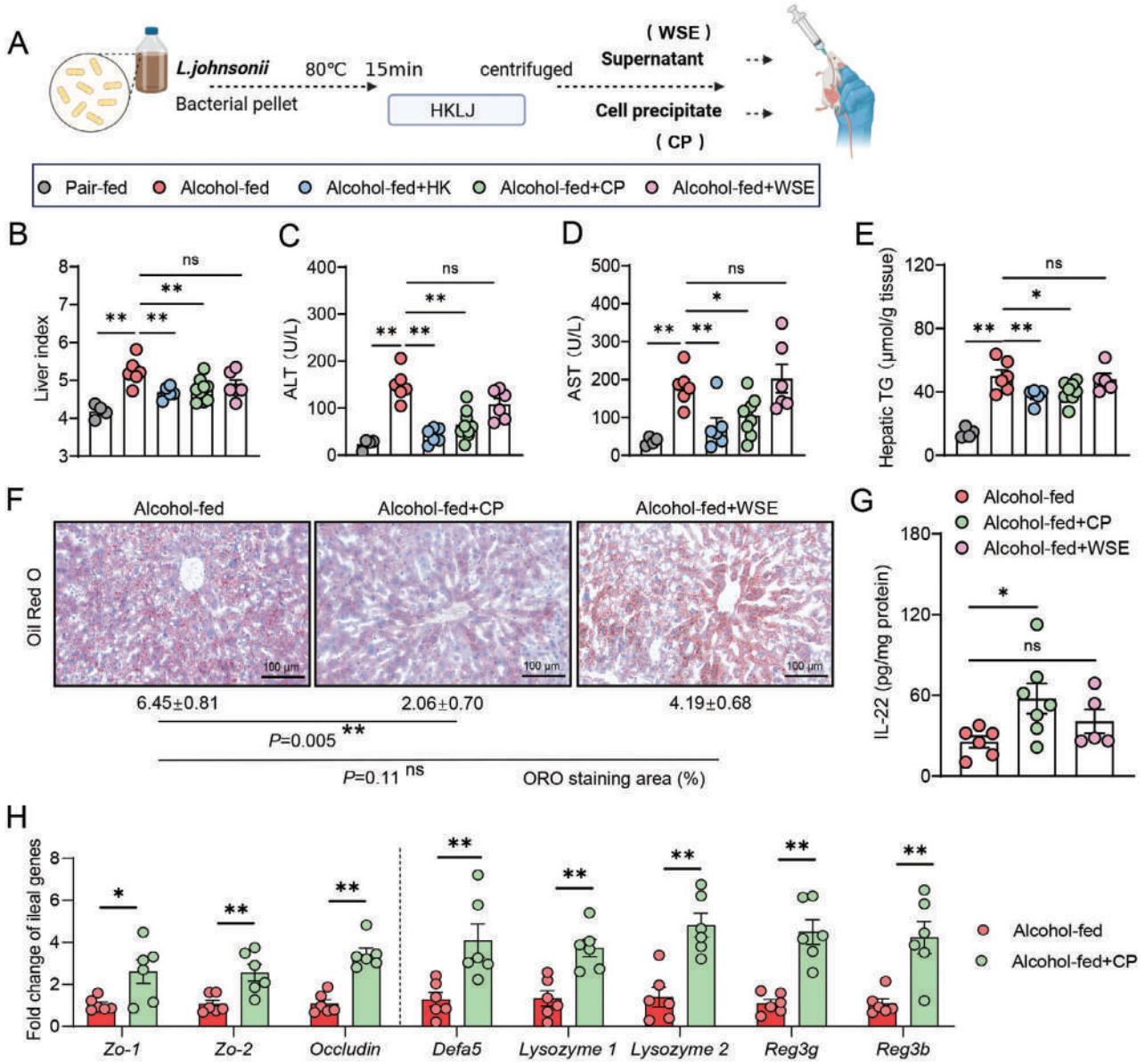


Figure 6. Identification of crude active components in heat-killed *L. johnsonii*. A) Schematic diagrams of crude extraction about *L. johnsonii* (created by BioRender.com). B) Liver-to-body weight ratio at the time of sacrifice. C) Serum levels of ALT. D) Serum levels of AST. E) Triglyceride levels in liver tissue. F) Staining of liver sections for oil red O and quantitative analysis. G) Levels of IL-22 in ileal tissue. H) Ileal mRNA expression levels of *Zo-1*, *Zo-2*, *Occludin*, *Defa5*, *Lysozyme1*, *Lysozyme2*, *Reg3 g* and *Reg3b*, normalized to *Gapdh*. Each dot represents data from individual biological replicate. Biological replicate numbers in each group are as follows: (Figure 6B–E) n = 4-9, (Figure 6F) n = 4-6, (Figure 6G) 5–7, (Figure 6H) n = 6. Statistical significance was determined using one-way analysis of variance (Figure 6B–G) or unpaired Student's two-tailed t test (Figure 6H), and data represent the mean ± s.e.m., unless otherwise indicated. Statistically significant differences are shown with asterisks as follows: $^{*}p$ values < 0.05; $^{**}p$ values < 0.01; ns, no significant differences. CP: Cell precipitate; WSE: Water-soluble extracts.

level of lysozyme is secreted from myeloid dendritic cells stimulated by microbial antigens.[35] In the present study, HKLJ supplementation significantly elevated the proportions of both F4/80^{+} myeloid cells (macrophages) and MHC II^{+} myeloid cells (dendritic cells, DCs), indicating increases in the activation and antigen presentation of myeloid cells (Figure 7D; Figure S9, Supporting Information). Considering the increased expression of intestinal lysozyme by HKLJ and its CP, we treated the precipitates of HKLJ with lysozyme in vitro to give the corresponding cell lysates (HKLJ-Lys). Using bone marrow-derived DCs (BMDCs) as an in vitro model, we assayed the activation of the NOD2/RIP2 pathway and the production of IL-23 (an upstream cytokine of IL-22) by HKLJ-Lys with the known NOD2 ligand MDP as positive control. Treatments at 1 and 10 μg mL^{-1} effectively enhanced both the expression of NOD2 and RIP2 (a NOD2 downstream switch), as well as the production of IL-23 (Figure 7E–G). In the NIAAA experiments, the anti-ALD effects of HKLJ, except for ALT reduction, were largely blocked by the NOD2 inhibitor

 © 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

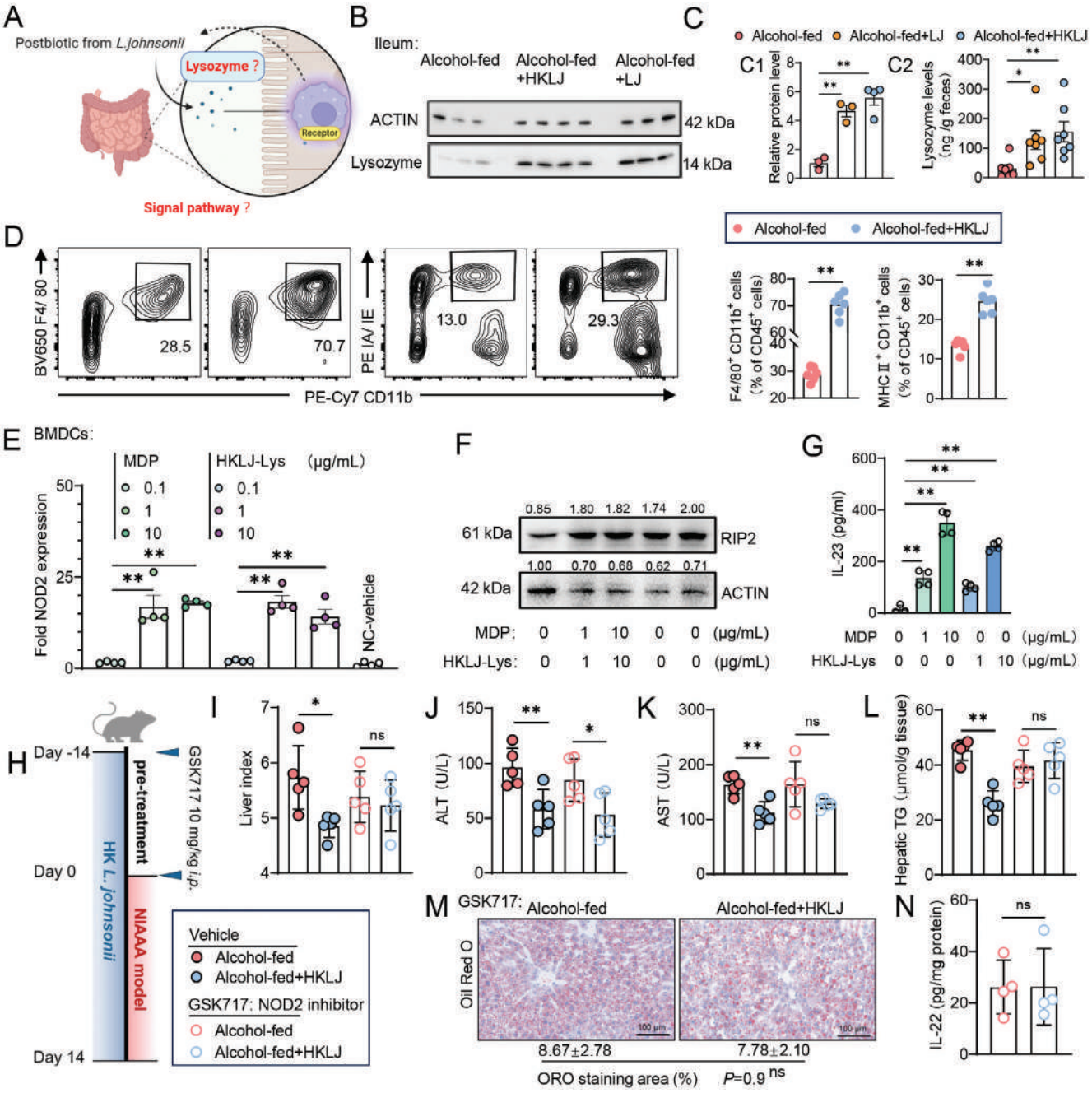


Figure 7. Postbiotics from *L. johnsonii* alleviates ALD via activation of the NOD2-IL-22 relays. A) Schematic diagrams of Figure 7 (created by BioRender.com). B) The western blot assay of ileal lysozyme, normalized to β-actin. C) 1) The protein expression levels of lysozyme in the ileum and 2) levels of lysozyme in the rodent feces. D) Representative flow cytometry plots of F4/80-postive or IA/IE-positive (MHC II-positive) myeloid cells (CD11b-postive gated) in ileum, and quantification analysis. E) The expression of NOD2 activated by HKLJ-Lys. Data are presented as fold change relative to vehicle-treated group (negative control, NC). F) The protein expression of RIP2 in the BMDCs were detected by western blot assay, normalized to β-actin. G) Levels of IL-23 in the culture supernatant of BMDCs were detected by ELISA. H) Schematic diagrams of NIAAA mouse model with GSK717 intervention. I) Liver-to-body weight ratio at the time of sacrifice. J) Serum levels of ALT. K) Serum levels of AST. L) Triglyceride levels in liver tissue. M) Staining of liver sections for oil red O and quantitative analysis. N) Levels of IL-22 in ileal tissue. Each dot represents data from individual biological replicate. Biological replicate numbers in each group are as follows: (Figure 7B, C_1) $n = 3$, (Figure 7C_2) $n = 7$, (Figure 7D) $n = 6$, (Figure 7E-7G) $n = 4$, (Figure 7I-7L) $n = 5$, (Figure 7M and 7N) $n = 4$. Statistical significance was determined using one-way analysis of variance or unpaired two-tailed t test, and data represent the mean ± s.e.m., unless otherwise indicated. Statistically significant differences are shown with asterisks as follows: $^{*}p$ values < 0.05; $^{**}p$ values < 0.01; ns, no significant differences. BMDCs: Bone marrow derived dendritic cells; MDP: Muramyl dipeptide; NOD2: Nucleotide binding Oligomerization Domain 2; RIP2: Receptor-interacting protein kinase 2.

 © 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

GSK717 (Figure 7H–N). Thus, it was evidenced that supplementation with HKLJ induces the expression of intestinal lysozyme, which in turn enhances the release of immunoregulatory components from HKLJ that bind to the NOD2 receptor to induce myeloid cell activation and antigen presentation, triggering the production of IL-22 from ILC3. The aforementioned data illustrate the intricate interactions between the postbiotics of *L. johnsonii* and ALD mice. Further studies on the immunoregulatory components of *L. johnsonii* are needed.

3. Discussion

The discovery and characterization of new postbiotics not only provide insights into the mechanism of action for probiotics but also advance new therapeutic strategies for diseases. In this study, we identified a gut commensal *L. johnsonii*-derived postbiotics that protects against gut microbiota dysbiosis, mucosal damage, and hepatic injury induced by ethanol. Mechanistically, we demonstrated that daily supplementation with heat-killed *L. johnsonii* induced the expression of intestinal lysozymes to increase the production of immunoregulatory substances from gut bacteria that subsequently activated the NOD2-IL-23-IL-22 innate immune axis.

Lysozymes secreted from gut mucosal cells play important roles in liberating NOD1/2 ligands from microbes.[36–38] In addition, the activation of the NOD2/RIP2 pathway by intestinal ligands positively mediates the enterocytic lysozyme sorting.[39,40] Herein, we found that the lysozyme-lysates from HKLJ enhanced the expression of NOD2 and RIP2 in vitro. In this context, we introduce a feedback loop involving lysozyme and postbiotics. Specifically, administration of HKLJ postbiotics induces the upregulated expression of intestinal lysozymes, which in turn increases the production of immunoregulatory ligands from HKLJ. The increased ligands further trigger NOD2-related signaling in dendritic cells to promote cellular lysozyme sorting and downstream immune cytokines production. This loop stimulated by HKLJ plays a crucial role in mediating its anti-ALD effects. In addition to the NOD2 pathway, *L. johnsonii* was reported to activate TLR1/2 in macrophages for relieving colitis,[20] suggesting that other immune signaling pathways be involved in the interaction between this symbiont and host. The immunoregulatory components originated from *L. johnsonii* and their regulatory pathways beyond NOD2 deserve further investigation.

Interleukin-22 (IL-22) plays important roles in regulating tissue regeneration and repair, pathogen clearance, and immune homeostasis. We found that heat-killed *L. johnsonii* supplementation enhanced the level of IL-22 in ethanol-fed mice. In the field of ALD treatment, IL-22 agonists have been demonstrated to be effective in the clinic.[41] We also demonstrated that the lysozyme-lysates of *L. johnsonii* stimulated IL-23 production from the activating DCs, which is the upstream cytokine of IL-22. DCs and ILC3s hook up through IL-23 to maintain the production of intestinal IL-22 in heat-killed *L. johnsonii*-treated ethanol-fed mice, further upregulating the expression of gut tight junction proteins and defense proteins and promoting hepatocyte regeneration via activation of STAT3.

Long-term alcohol consumption greatly disturbs the gut microbiota, leading to the overexpansion of potential pathogens and a reduction in butyrate-producing bacteria.[29,28] Moreover, the ethanol-induced gut microbiota is confirmed to be a disease-driving factor exacerbating experimental ALD symptoms by FMT experiment.[42] Owing to a heat-killed *L. johnsonii*-induced increase in the expression of Reg3 family defense proteins, the ethanol-caused expansion of enteric *Escherichia coli* and *Helicobacter* sp. was effectively suppressed. These pathogens reportedly boost the progression of ALD.[16,43,44] Interestingly, heat-killed *L. johnsonii* increased the abundance of SCFA-producing *Faecalibaculum rodentium* and *Bifidobacterium pseudolongum* in the gut of ethanol-fed mice. In an early study, the SCFA-producing *Faecalibaculum spp.* and *Bifidobacterium spp.* were suggested to be cross-feeding.[45] The *F. rodentium* and *B. pseudolongum* strains have been demonstrated to be beneficial for treating gastrointestinal and hepatic disorders.[46,47] In addition to the expansion of facultative bacteria, the overgrowth of gut fungi was confirmed to accelerate the progression of ALD via enhancing pro-inflammatory signaling through the secreted toxins or β-glucan.[3,27,48] In human cohorts, the abundance of *Pichia* sp. was found to be elevated in alcoholic hepatitis and alcoholic cirrhosis patients.[27] Herein, supplementation of heat-killed *L. johnsonii* significantly restricted the ethanol feeding-induced overgrowth of *Pichia kudriavzevii* (early named *Candida krusei*). It has been demonstrated that gut microbial butyrate triggers PPARγ signaling on the gut epithelium to maintain an intestinal hypoxic environment and control gut microbiota homeostasis,[30] which assists in the amelioration of ALD. Consistent with the observed changes in the gut microbiome, we found that heat-killed *L. johnsonii* increased the level of gut butyrate and activated the epithelial PPARγ signaling pathway. The maintenance of gut hypoxia helps to control the expansion of facultative bacterial pathogens and fungi in the gut of ALD mice.

In summary, postbiotics derived from gut commensal *L. johnsonii* protect against ethanol-disrupted gut barrier functions, restore ethanol-induced gut microbiota dysbiosis, and attenuate hepatic injury by triggering the gut NOD2-mediated innate immune pathway in ALD mice. Increased IL-22 was confirmed to be an effector mediating the aforementioned benefits of *L. johnsonii* postbiotics. These anti-ALD phenotypes conferred by *L. johnsonii* make this symbiont a promising candidate for clinical interventions. However, being a study in the mouse model, a randomized controlled trial with a placebo is needed to assess the safety and efficacy of HKLJ postbiotics in a clinical setting. Altogether, this study in combination with earlier research on probiotics, postbiotics, and prebiotics, supports the development of microbiome-based therapeutics for the treatment of ALD.

4. Experimental Section

Animal Models and Animal Ethics: The research complied with all relevant ethical regulations. The animal experiments were approved by the research ethics committee of the Institute of Microbiology, Chinese Academy of Sciences (permit APIMCAS2023115). Age-matched male C57BL/6JGpt mice (strain no. N000013; aged 8–10 weeks; body weight of 22–25 g) were purchased from GemPharmatech Co., Ltd. (Nanjing, China). Two mice were housed in an independent animal facility under specific-pathogen-free conditions and subjected to the chronic feeding model or chronic plus-binge model (NIAAA model).

For the chronic feeding model,[49] mice were fed with a Lieber-DeCarli liquid diets (BioServ, NO. F1258SP, USA) containing ethanol (alcohol-fed)

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

or isocaloric control liquid diets containing maltodextrin (Aladdin, NO. M305460, China) (pair-fed) for 8 weeks. In detail, mice were fed the Lieber-DeCarli liquid diets by gradually increasing the ethanol concentration to 4.5% (wt/vol) ethanol (24% calories in the diet) in the first two weeks and then gradually increasing the ethanol concentration to 5.6% (wt/vol) ethanol (30% calories in the diet) in the last two weeks.

For the chronic plus-binge alcohol feeding model (NIAAA model),[49] 8-week-old mice were fed the Lieber-DeCarli liquid diets. The content of alcohol in the Lieber-DeCarli liquid diets gradually increased from 0% (wt/vol) to 5% (wt/vol) over 0–5 days and remained at 5% for the subsequent 10 days. On the last day, the mice were gavaged with a single dose (5 g per kilogram of body weight) of ethanol at 9:00 a.m. and euthanized 9 h later. Pair-fed control mice were fed an iso-caloric liquid diet and were euthanized 9 h after intragastric administration of the iso-caloric maltodextrin solution.

Microbe Strains: *Lactobacillus johnsonii* (LJ) (CGMCC No. 27014) which was previously isolated from feces of a healthy human infant was used for this study (APIMCAS2017049). The bacterial strains were plated on De Man, Rogosa, and Sharpe (MRS) agar medium plates (HB0384-5, Hopebio, China) and incubated in an anaerobic atmosphere (15% H_2, 5% CO_2, and 80% N_2) for 48 h at 37 °C. Prior to use, mono-clones were picked and inoculated with fresh sterile MRS broth. The bacterium was anaerobically incubated for 24 h at 37 °C to the logarithmic growth phase (Log-phase).

Faecalibaculum rodentium and *Bifidobacterium pseudolongum* were also isolated from the ileal mucosa of C57BL/6J mice and preserved in the laboratory. These strains were anaerobically incubated in modified Gifu Anaerobic Medium (mGAM) (HB8518, Hopebio, China) for 48 h at 37 °C. Prior to use, mono-clones were picked and inoculated with fresh sterile mGAM broth, anaerobically incubated for 24 h at 37 °C to the Log-phase.

Extraction of L. Johnsonii Cell Components: The *L. johnsonii* was grown as a 50 mL culture in MRS broth to mid-logarithmic phase (1×10^8 CFU mL^{-1}) and then pelleted by centrifugation at 5000 rpm for 20 min at 4 °C. Cell pellets were then rewashed with 50 mL of phosphate buffer saline (PBS), resuspended in 5 mL of PBS, and stored at 4 °C until use. The bacterial pellet of *L. johnsonii* was then resuspended in phosphate buffer saline (PBS) and pasteurized at 85 °C for 15 min. Further centrifugation at 13 000 rpm for 10 min provided the corresponding supernatant and the cell precipitate.

To get the cell components, the cell precipitate of *L. johnsonii* was resuspended in 1 mL of 5% SDS in PBS and then boiled for 25 min. After centrifugation at 20 000 rpm, the pellets were washed with MilliQ water to remove SDS traces. Next, the pellets were sequentially enzymatically treated with 2 mg mL^{-1} of proteinase K (P109033, Aladdin, China) in PBS for 30 min at 37 °C; 200 µg mL^{-1} of trypsin (T105532, Aladdin, China) in PBS for 2 h at 37 °C with shaking; 10 µg mL^{-1} of bi-nuclease (B401536, Aladdin, China) in PBS for 1 h at 37 °C. After boiling for 15 min, the suspension was centrifugated at 20 000 rpm for 10 min. The insoluble fractions were digested with lysozyme (A610308, BBI, China) for 18 h at 37 °C with shaking. The resulting lysates (HKLJ-Lys) were boiled for 15 min and lyophilized.

Histopathological Analysis: Fresh liver left lobe tissue was resected and fixed with paraformaldehyde (4%, pH 7.4). A portion of the tissue was first embedded in paraffin and then cut into 4 µm sections, followed by hematoxylin and eosin (H&E) staining. The remaining part was wrapped in optimal cutting temperature (OCT) compound and cut into 10 µm sections, followed by oil red O and Sirius red staining (Solarbio, G1261 and G1472, China). Quantitative analysis of liver histologic scores was performed according to a combination of the following two standards: degree of inflammation in the confluent area: 0, normal; 1, mild inflammatory infiltrate; 3, moderate inflammatory infiltrate. Steatosis score: 0, normal; 1, a small amount of ballooning degeneration; 3, degeneration involving more than 25% of the liver's left lobe. The fresh ileum was also sectioned and stained following the standard histological procedure described above. All sections were observed by a Zeiss Imager A2-M2 microscope (Carl Zeiss AG, Germany). At least 3 fields of view per section of each intestine were measured and the images were manually delineated and analyzed using Image-Pro Plus 6.0 (from the NIH, USA).

RNA Extraction and RT-qPCR: Total RNA was extracted from the ileal or liver tissues of animals using TRIeasy Total RNA Extraction Reagent (Yeasen, 10606ES60, China) according to the manufacturer's instructions. cDNA was reverse transcribed from 500 ng of total RNA using Hifair II 1st Strand cDNA Synthesis SuperMix (Yeasen, 11123ES60, China). The target primers used are listed in Table S1 (Supporting Information). RT-qPCR was performed using NovoStar SYBR qPCR SuperMix Plus (Novoprotein, E096-01A, China), and amplification was initiated in a CFX Connect Real-Time PCR Detection System (Bio-Rad, 1855201, USA). For each target gene, the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (*gapdh*) was used for internal normalization.

Enzyme-Linked Immunosorbent Assay (ELISA): ELISA kits were used to detect the protein levels of cytokines (TNF-a, IL-6, IL-1β, IL-10, and IL-22), lysozyme, and lipopolysaccharide (LPS) in the rodent serum or feces, as well as the level of IL-23 in the cell culture supernatants, separately. Whole blood was collected from mice using the inner canthus method. Serum was prepared by centrifugation at 3000 rpm for 15 min at room temperature. The 100 mg of feces was collected and homogenized (65 Hz) with 1 mL PBS using zirconia beads. The supernatant was separated from the homogenized suspension by centrifugation at 13 000 rpm for 15 min and placed at 4 °C. Cell culture supernatants were collected by centrifugation at 3000 rpm for 5 min. The experiment was carried out according to the manufacturer's instructions.

Western Blot Analysis: These protein lysates were separated by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and then transferred to polyvinylidene difluoride (PVDF) membranes. After being blocked by 5% skimmed milk (Biofroxx, 1172GR500, Germany), this membrane was incubated with primary antibody at 4 °C overnight. After incubation with a corresponding second antibody, the whole membrane was immersed in a super sensitive ECL luminescence reagent (Meilunbio, MA0186, China). The protein bands in the membrane were detected and quantified using a fully automated chemiluminescence image analysis system (Tanon, 5200, China) and Image J software.

Cecal DNA Extraction and Sequencing: Cecal tissue was obtained at −80 °C, and the cecal contents were transferred to sterile 1.5 mL centrifuge tubes. The bacterial genomic DNA of the contents was isolated using the Stool Genomic DNA Kit (CWBIO, CW2092, China). Genomic DNA purity was detected by a Nano-Photometer spectrophotometer (IMPLEN, USA). The V3-V4 region of the bacterial 16S rRNA genes was amplified by polymerase chain reaction (PCR) using Platinum Hot Start PCR master mix (Invitrogen, 13000012, USA), primers, and genomic DNA (20 ng). After the individual quantification step, the amplicon libraries were pooled in equal amounts, and paired-end 2–250 bp sequencing was performed using the Illumina NovaSeq-PE250 platform at Personal Biotechnology Co., Ltd (Shanghai, China). A similar approach was applied to the mycobiome targeting the internal transcribed spacer (ITS) V1 region of fungal ribosomal genes, as well as an optimized and standardized protocol for the preparation of ITS1 sample libraries.

Alpha diversities were analyzed using the Chao index and the Simpson index calculated with observed amplicon sequence variants (ASVs). Beta diversity was visualized via principal coordinate analysis (PCoA), and assessed using unweighted UniFrac distance or Jaccard distance. Taxonomic analysis was performed using the QIIME 2 feature-classifier trained on the NCBI database. For precise classification, each unique ASV was searched against the NCBI database by the BLAST tool to find the closest class group at the species level according to percent identity. Linear discriminant analysis effect size (LEfSe) was performed with the default parameters to detect diverse taxa between groups.

Quantification of Short-Chain Fatty Acids (SCFAs): Individual mice were placed in an empty sterile facility (without bedding) and removed after 45 min to collect fecal pellets, which were then stored at −80 °C for gas chromatography-mass spectrometry (GC-MS) analysis. For the extraction of SCFAs from feces, 50 mg freeze-dried fecal samples were homogenized in 1 mL of methanol (Sigma, CAS No. 67-56-1, USA) on ice for 10 min. The supernatant enriched with SCFAs was isolated by centrifugation. For the extraction of SCFAs from the medium, bacterial broth cultured to the log phase was treated with an equal amount of ethyl acetate (Sigma, CAS No. 141-78-6, USA). After shaking, the organic

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

phase of the supernatant was obtained by centrifugation. The extractions were aliquoted into 2 mL ampoules after passing through 0.2 µm nylon filter membranes (Agilent, 5191-4341, USA) for GC-MS analysis (Agilent, 7890A/5975C, USA). GC-MS analysis was performed according to a previously described method.[50] GAM broth was used as the negative control.

Isolation of Intestinal Lamina Propria Cells: The protocol for isolating intestinal lamina propria cells has been slightly modified from the previous description.[51] Collected intestinal pieces after removing Peyer's patches were shaken (150 g) for 10 min at 37 °C in Hank's buffer (Gibco, 14175095, USA) containing 5 mM EDTA (Invitrogen, AM9262, USA), 1 mM DTT (Solarbio, D8220, China) to deplete the epithelial layer and mucus. Mucus-depleted intestinal pieces were digested in RPMI-1640 medium (Gibco, C11875500BT, USA) with 5% v/v FBS (Meilun-bio, PWL001, China), 1.0 mg mL^{-1} collagenase II and III (Worthington, WBC-LS004176 and WBC-LS004182, USA), and 0.2 g L^{-1} DNase I (Roche, 10104159001, USA) for 30 min at 37 °C. After centrifugation, the intestinal lamina propria cells were passed through a 70 µm cell filter and washed twice with Hank's buffer for flow cytometry assay.

Flow Cytometry: After 6 h of incubation with RPMI-1640 medium containing cell activation cocktail (Biolegend, 423304, USA), intestinal lamina propria cells were blocked with anti-CD16/32 (BD, 553141, USA) antibody for 5 min, and then incubated on ice with Zombie NIR fixable viability kit (Biolegend, 423106, USA) for 15 min under dark. During the incubation, the cells surface was stained with the antibody mixture for 30 min at 4 °C. Finally, cells were permeabilized and fixed with transcription factor buffer (BD, 562574, USA) and TF perm/wash buffer (BD, 9008102, USA) before staining with the antibody mixture (omitted in DCs protocol). Finally, the stained cells were resuspended in Stain buffer (BD, 554656, USA) and were detected in a Flow Cytometer (BD, LSR-Fortessa, USA). The analysis was completed using FlowJo software (Tree Star Inc., USA).

Cell Culture: C57BL/6J mice were submerged in 75% alcohol for 5 min after euthanasia (CO_2 asphyxiation). In a sterile hood, the cavities of the tibia and femur were cracked open, and the bone marrow was flushed out with PBS and blown into a single-cell suspension. The collected cells were cultured in RPMI-1640 complete medium containing 20 ng mL^{-1} of granulocyte-macrophage colony stimulating factor (GM-CSF) (MCE, HY-P7361, USA) and 1 ng mL^{-1} of IL-4 (Novoprotein, P07750, China).[52] The cells were cultured at 37 °C in a humidified environment with 5% CO_2, and were induced into Bone marrow-derived dendritic cells (BMDCs) after one week. BMDCs were collected after 12 h of drug stimulation and stored at −20 °C for assay.

Statistics: All statistical analyses were conducted using the R package or GraphPad Prism software (version 8.0.2). The data were presented as the mean ± SEM or as individual data points unless otherwise specified. The normality of the probability distribution was confirmed by the Shapiro Wilk test and the variance in the probability distribution was confirmed by the Brown-Forsythe test to determine the appropriate statistical analysis method using SPSS (version R27.0.1.0). Statistical comparisons between only two groups were conducted using unpaired Student's two-tailed *t*-test (normal probability distribution and homogeneity of variance) or the Mann Whiney U test (abnormal probability distribution or heterogeneity of variance). For comparisons of more than two groups, a one-way analysis of variance (ANOVA) test (normal probability distribution) or non-parametric Kruskal Wallis test (abnormal distribution) followed by a post hoc Dunnett's multiple comparison test using Alcohol-fed group data as control was performed. Statistically significant differences are shown with asterisks as follows: $^{*}p$ values < 0.05; $^{**}p$ values < 0.01; ^{ns}p values > 0.05.

Additional detailed methods are provided in the Supporting Information.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was sponsored financially by the National Key Technologies R&D Program (2021YFA0717000). The authors express the gratitude to Dr. Wenzhao Wang for his contributions to gas chromatography-mass spectrometry (GC-MS) analysis, to Dr. Yaxin Zhu for her technical support regarding the anaerobic workstation, and to Dr. Zhao Tong for her assistance with flow cytometry. at the Institute of Microbiology, CAS.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

R.Y. performed most of the experiments, analyzed and interpreted data, and prepared the submitted manuscript. T.W. performed some animal experiments. J.S. isolated and identified all strains. H.D. contributed to the critical discussion of the project. Y.Z. provided technical support for flow cytometry and cell assay. N.L. provided the design of animal models and contributed to manuscript revision. H.L. conceived, designed, and supervised the study and revised the manuscript.

Data Availability Statement

The data that support the findings of this study are available in the Supporting Information of this article.

Keywords

alcoholic-associated liver disease, gut microbiota, innate immunity, *Lactobacillus johnsonii*, postbiotics

Received: May 26, 2024
Revised: October 10, 2024
Published online:

[1] H. Rumgay, K. Shield, H. Charvat, P. Ferrari, B. Sornpaisarn, I. Obot, F. Islami, V. E. P. P. Lemmens, J. Rehm, I. Soerjomataram, *Lancet Oncol.* **2021**, *22*, 1071.
[2] D. Q. Huang, P. Mathurin, H. Cortez-Pinto, R. Loomba, *Nat. Rev. Gastroenterol. Hepatol.* **2023**, *20*, 37.
[3] H. Chu, Y. Duan, S. Lang, L. Jiang, Y. Wang, C. Llorente, J. Liu, S. Mogavero, F. Bosques-Padilla, J. G. Abraldes, V. Vargas, X. M. Tu, L. Yang, X. Hou, B. Hube, P. Stärkel, B. Schnabl, *J. Hepatol.* **2020**, *72*, 391.
[4] G. Szabo, P. Mandrekar, *Exp. Res.* **2009**, *33*, 220.
[5] Y. Duan, H. Chu, K. Brandl, L. Jiang, S. Zeng, N. Meshgin, E. Papachristoforou, J. Argemi, B. G. Mendes, Y. Wang, H. Su, W. Sun, C. Llorente, T. Hendrikx, X. Liu, M. Hosseini, T. Kisseleva, D. A. Brenner, R. Bataller, P. Ramachandran, M. Karin, W. Fu, B. Schnabl, *Nat. Commun.* **2021**, *12*, 7172.
[6] Y. Duan, C. Llorente, S. Lang, K. Brandl, H. Chu, L. Jiang, R. C. White, T. H. Clarke, K. Nguyen, M. Torralba, Y. Shao, J. Liu, A. Hernandez-Morales, L. Lessor, I. R. Rahman, Y. Miyamoto, M. Ly, B. Gao, W. Sun, R. Kiesel, F. Hutmacher, S. Lee, M. Ventura-Cots, F. Bosques-Padilla, E. C. Verna, J. G. Abraldes, R. S. Brown, V. Vargas, J. Altamirano, J. Caballería, et al., *Nature* **2019**, *575*, 505.
[7] G. Ferrere, L. Wrzosek, F. Cailleux, W. Turpin, V. Puchois, M. Spatz, D. Ciocan, D. Rainteau, L. Humbert, C. Hugot, F. Gaudin, M.-L. Noordine, V. Robert, D. Berrebi, M. Thomas, S. Naveau, G. Perlemuter, A.-M. Cassard, *J. Hepatol.* **2017**, *66*, 806.

 © 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[8] C. L. Hsu, B. Schnabl, *Nat. Rev. Microbiol.* **2023**, *21*, 719.
[9] H. Shen, L. Zhou, H. Zhang, Y. Yang, L. Jiang, D. Wu, H. Shu, H. Zhang, L. Xie, K. Zhou, C. Cheng, L. Yang, J. Jiang, S. Wang, Y. Han, J. Zhu, L. Xu, Z. Liu, H. Wang, S. Yin, *Cell Host Microbe* **2024**, *32*, 1331.
[10] F. E. Dailey, E. P. Turse, E. Daglilar, V. Tahan, *Curr. Opin. Pharmacol.* **2019**, *49*, 29.
[11] J. Suez, N. Zmora, E. Segal, E. Elinav, *Nat. Med.* **2019**, *25*, 716.
[12] V. A. Pedicord, A. A. K. Lockhart, K. J. Rangan, J. W. Craig, J. Loschko, A. Rogoz, H. C. Hang, D. Mucida, *Sci. Immunol.* **2016**, *1*, eaai7732.
[13] M. E. Griffin, J. Espinosa, J. L. Becker, J.-D. Luo, T. S. Carroll, J. K. Jha, G. R. Fanger, H. C. Hang, *Science* **2021**, *373*, 1040.
[14] D. Erturk-Hasdemir, S. F. Oh, N. A. Okan, G. Stefanetti, F. S. Gazzaniga, P. H. Seeberger, S. E. Plevy, D. L. Kasper, *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 26157.
[15] C. Ramakrishna, M. Kujawski, H. Chu, L. Li, S. K. Mazmanian, E. M. Cantin, *Nat. Commun.* **2019**, *10*, 2153.
[16] R. Verma, C. Lee, E.-J. Jeun, J. Yi, K. S. Kim, A. Ghosh, S. Byun, C.-G. Lee, H.-J. Kang, G.-C. Kim, C.-D. Jun, G. Jan, C.-H. Suh, J.-Y. Jung, J. Sprent, D. Rudra, C. De Castro, A. Molinaro, C. D. Surh, S.-H. Im, *Sci. Immunol.* **2018**, *3*, eaat6975.
[17] B. Seo, K. Jeon, S. Moon, K. Lee, W.-K. Kim, H. Jeong, K. H. Cha, M. Y. Lim, W. Kang, M.-N. Kweon, J. Sung, W. Kim, J.-H. Park, G. Ko, *Cell Host Microbe* **2020**, *27*, 25.
[18] T. S. Ghosh, J. Arnoux, P. W. O'Toole, *Gut Microbes* **2020**, *12*, 1822729.
[19] Y. Xiao, Q. Zhai, H. Zhang, W. Chen, C. Hill, *Annu. Rev. Food Sci. Technol.* **2021**, *12*, 213.
[20] D.-J.-C. Jia, Q.-W. Wang, Y.-Y. Hu, J.-M. He, Q.-W. Ge, Y.-D. Qi, L.-Y. Chen, Y. Zhang, L.-N. Fan, Y.-F. Lin, Y. Sun, Y. Jiang, L. Wang, Y.-F. Fang, H.-Q. He, X.-E. Pi, W. Liu, S.-J. Chen, L.-J. Wang, *Gut. Microbes.* **2022**, *14*, 2145843.
[21] D. Jia, Q. Wang, Y. Qi, Y. Jiang, J. He, Y. Lin, Y. Sun, J. Xu, W. Chen, L. Fan, R. Yan, W. Zhang, G. Ren, C. Xu, Q. Ge, L. Wang, W. Liu, F. Xu, P. Wu, Y. Wang, S. Chen, L. Wang, *Cell* **2024**, *187*, p1651.
[22] T. Wang, J. Han, H. Dai, J. Sun, J. Ren, W. Wang, S. Qiao, C. Liu, L. Sun, S. Liu, D. Li, S. Wei, H. Liu, *Carbohydr. Polym.* **2022**, *295*, 119862.
[23] X. Fan, C. Mai, L. Zuo, J. Huang, C. Xie, Z. Jiang, R. Li, X. Yao, X. Fan, Q. Wu, P. Yan, L. Liu, J. Chen, Y. Xie, E. L.-H. Leung, *Acta Pharm. Sin. B* **2023**, *13*, 1164.
[24] C.-H. Chuang, C.-C. Tsai, E.-S. Lin, C.-S. Huang, Y.-Y. Lin, C.-C. Lan, C.-C. Huang, *Molecules* **2016**, *21*, 1456.
[25] H. Tilg, P. Mathurin, *Gut.* **2016**, *65*, 728.
[26] J. S. Bajaj, *Nat. Rev. Gastroenterol. Hepatol.* **2019**, *16*, 235.
[27] A.-M. Yang, T. Inamine, K. Hochrath, P. Chen, L. Wang, C. Llorente, S. Bluemel, P. Hartmann, J. Xu, Y. Koyama, T. Kisseleva, M. G. Torralba, K. Moncera, K. Beeri, C.-S. Chen, K. Freese, C. Hellerbrand, S. M. L. Lee, H. M. Hoffman, W. Z. Mehal, G. Garcia-Tsao, E. A. Mutlu, A. Keshavarzian, G. D. Brown, S. B. Ho, R. Bataller, P. Stärkel, D. E. Fouts, B. Schnabl, *J. Clin. Invest.* **2017**, *127*, 2829.
[28] V. B. Dubinkina, A. V. Tyakht, V. Y. Odintsova, K. S. Yarygin, B. A. Kovarsky, A. V. Pavlenko, D. S. Ischenko, A. S. Popenko, D. G. Alexeev, A. Y. Taraskina, R. F. Nasyrova, E. M. Krupitsky, N. V. Shalikiani, I. G. Bakulin, P. L. Shcherbakov, L. O. Skorodumova, A. K. Larin, E. S. Kostryukova, R. A. Abdulkhakov, S. R. Abdulkhakov, S. Y. Malanin, R. K. Ismagilova, T. V. Grigoryeva, E. N. Ilina, V. M. Govorun, *Microbiome* **2017**, *5*, 141.
[29] R. Singhal, H. Donde, S. Ghare, K. Stocke, J. Zhang, M. Vadhanam, S. Reddy, L. Gobejishvili, P. Chilton, S. Joshi-Barve, W. Feng, C. McClain, K. Hoffman, J. Petrosino, M. Vital, S. Barve, *Gut. Microbes.* **2021**, *13*, 1946367.
[30] M. X. Byndloss, E. E. Olsan, F. Rivera-Chávez, C. R. Tiffany, S. A. Cevallos, K. L. Lokken, T. P. Torres, A. J. Byndloss, F. Faber, Y. Gao, Y. Litvak, C. A. Lopez, G. Xu, E. Napoli, C. Giulivi, R. M. Tsolis, A. Revzin, C. B. Lebrilla, A. J. Bäumler, *Science* **2017**, *357*, 570.
[31] H. P. Savage, D. J. Bays, C. R. Tiffany, M. A. F. Gonzalez, E. J. Bejarano, T. P. Carvalho, Z. Luo, H. L. P. Masson, H. Nguyen, R. L. Santos, K. L. Reagan, G. R. Thompson, A. J. Bäumler, *Cell Host Microbe* **2024**, *32*, 1103.
[32] K. K. Jang, T. Heaney, M. London, Y. Ding, G. Putzel, F. Yeung, D. Ercelen, Y.-H. Chen, J. Axelrad, S. Gurunathan, C. Zhou, M. Podkowik, N. Arguelles, A. Srivastava, B. Shopsin, V. J. Torres, A. M. Keestra-Gounder, A. Pironti, M. E. Griffin, H. C. Hang, K. Cadwell, *Cell Host Microbe* **2023**, *31*, 1450.
[33] N. Serafini, A. Jarade, L. Surace, P. Goncalves, O. Sismeiro, H. Varet, R. Legendre, J.-Y. Coppee, O. Disson, S. K. Durum, G. Frankel, J. P. Di Santo, *Science* **2022**, *375*, 859.
[34] S. H. Ki, O. Park, M. Zheng, O. Morales-Ibanez, J. K. Kolls, R. Bataller, B. Gao, *Hepatology* **2010**, *52*, 1291.
[35] H. Lelouard, S. Henri, B. De Bovis, B. Mugnier, A. Chollat-Namy, B. Malissen, S. Méresse, J. Gorvel, *Gastroenterology* **2010**, *138*, 173.
[36] S. A. Ragland, A. K. Criss, *PLoS Pathog.* **2017**, *13*, e1006512.
[37] Y. Wang, E. Goossens, V. Eeckhaut, E. Pérez Calvo, R. Lopez-Ulibarri, I. Eising, M. Klausen, N. Debunne, B. De Spiegeleer, R. Ducatelle, F. Van Immerseel, *Br. J. Nutr.* **2021**, *126*, 641.
[38] Q. Zhang, Y. Pan, B. Zeng, X. Zheng, H. Wang, X. Shen, H. Li, Q. Jiang, J. Zhao, Z.-X. Meng, P. Li, Z. Chen, H. Wei, Z. Liu, *Cell Res.* **2019**, *29*, 516.
[39] Q. Zhang, Y. Pan, R. Yan, B. Zeng, H. Wang, X. Zhang, W. Li, H. Wei, Z. Liu, *Nat. Immunol.* **2015**, *16*, 918.
[40] H. Wang, X. Zhang, Z. Zuo, Q. Zhang, Y. Pan, B. Zeng, W. Li, H. Wei, Z. Liu, *J.I* **2017**, *198*, 3729.
[41] J. P. Arab, T. S. Sehrawat, D. A. Simonetto, V. K. Verma, D. Feng, T. Tang, K. Dreyer, X. Yan, W. L. Daley, A. Sanyal, N. Chalasani, S. Radaeva, L. Yang, H. Vargas, M. Ibacache, B. Gao, G. J. Gores, H. Malhi, P. S. Kamath, V. H. Shah, *Hepatology* **2020**, *72*, 441.
[42] M. Llopis, A. M. Cassard, L. Wrzosek, L. Boschat, A. Bruneau, G. Ferrere, V. Puchois, J. C. Martin, P. Lepage, T. Le Roy, L. Lefèvre, B. Langelier, F. Cailleux, A. M. González-Castro, S. Rabot, F. Gaudin, H. Agostini, S. Prévot, D. Berrebi, D. Ciocan, C. Jousse, S. Naveau, P. Gérard, G. Perlemuter, *Gut.* **2016**, *65*, 830.
[43] M. Qian, J. Liu, D. Zhao, P. Cai, C. Pan, W. Jia, Y. Gao, Y. Zhang, N. Zhang, Y. Zhang, Q. Zhang, D. Wu, C. Shan, M. Zhang, B. Schnabl, S. Yang, X. Shen, L. Wang, *Cellul. Molecul. Gastroenterology and Hepatol.* **2022**, *13*, 233.
[44] T. Hendrikx, S. Lang, D. Rajcic, Y. Wang, S. McArdle, K. Kim, Z. Mikulski, B. Schnabl, *Gut* **2023**, *72*, 1959.
[45] L. De Vuyst, F. Leroy, *Int. J. Food Microbiol.* **2011**, *149*, 73.
[46] E. Zagato, C. Pozzi, A. Bertocchi, T. Schioppa, F. Saccheri, S. Guglietta, B. Fosso, L. Melocchi, G. Nizzoli, J. Troisi, M. Marzano, B. Oresta, I. Spadoni, K. Atarashi, S. Carloni, S. Arioli, G. Fornasa, F. Asnicar, N. Segata, S. Guglielmetti, K. Honda, G. Pesole, W. Vermi, G. Penna, M. Rescigno, *Nat. Microbiol.* **2020**, *5*, 511.
[47] Q. Song, X. Zhang, W. Liu, H. Wei, W. Liang, Y. Zhou, Y. Ding, F. Ji, A. Ho-Kwan Cheung, N. Wong, J. Yu, *J. Hepatol.* **2023**, *79*, 1352.
[48] X. Zhao, X. Hu, J. Han, R. Yin, S. Zhang, H. Liu, *WIREs Mechan. Dis.* **2023**, *15*, e1611.
[49] A. Bertola, S. Mathews, S. H. Ki, H. Wang, B. Gao, *Nat. Protoc.* **2013**, *8*, 627.
[50] S. Qiao, L. Bao, K. Wang, S. Sun, M. Liao, C. Liu, N. Zhou, K. Ma, Y. Zhang, Y. Chen, S.-J. Liu, H. Liu, *Cell Rep.* **2020**, *32*, 108005.
[51] B. Weigmann, I. Tubbe, D. Seidel, A. Nicolaev, C. Becker, M. F. Neurath, *Nat. Protoc.* **2007**, *2*, 2307.
[52] J. Helft, J. Böttcher, P. Chakravarty, S. Zelenay, J. Huotari, B. U. Schraml, D. Goubau, C. Reis e Sousa, *Immunity* **2015**, *42*, 1197.

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/advs.202405781 by CAS-Institute Of Microbiology, Wiley Online Library on [22/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：张鹏鹏
方向：生物制造
单位：中国科学院天津工业生物技术研究所

题目：改造亚胺还原酶催化轴手性化合物的合成

How to cite: *Angew. Chem. Int. Ed.* **2025**, *64*, e202416569
doi.org/10.1002/anie.202416569

Enzyme Catalysis | Hot Paper

Biocatalytic Desymmetrization for the Atroposelective Synthesis of Axially Chiral Biaryls Using an Engineered Imine Reductase

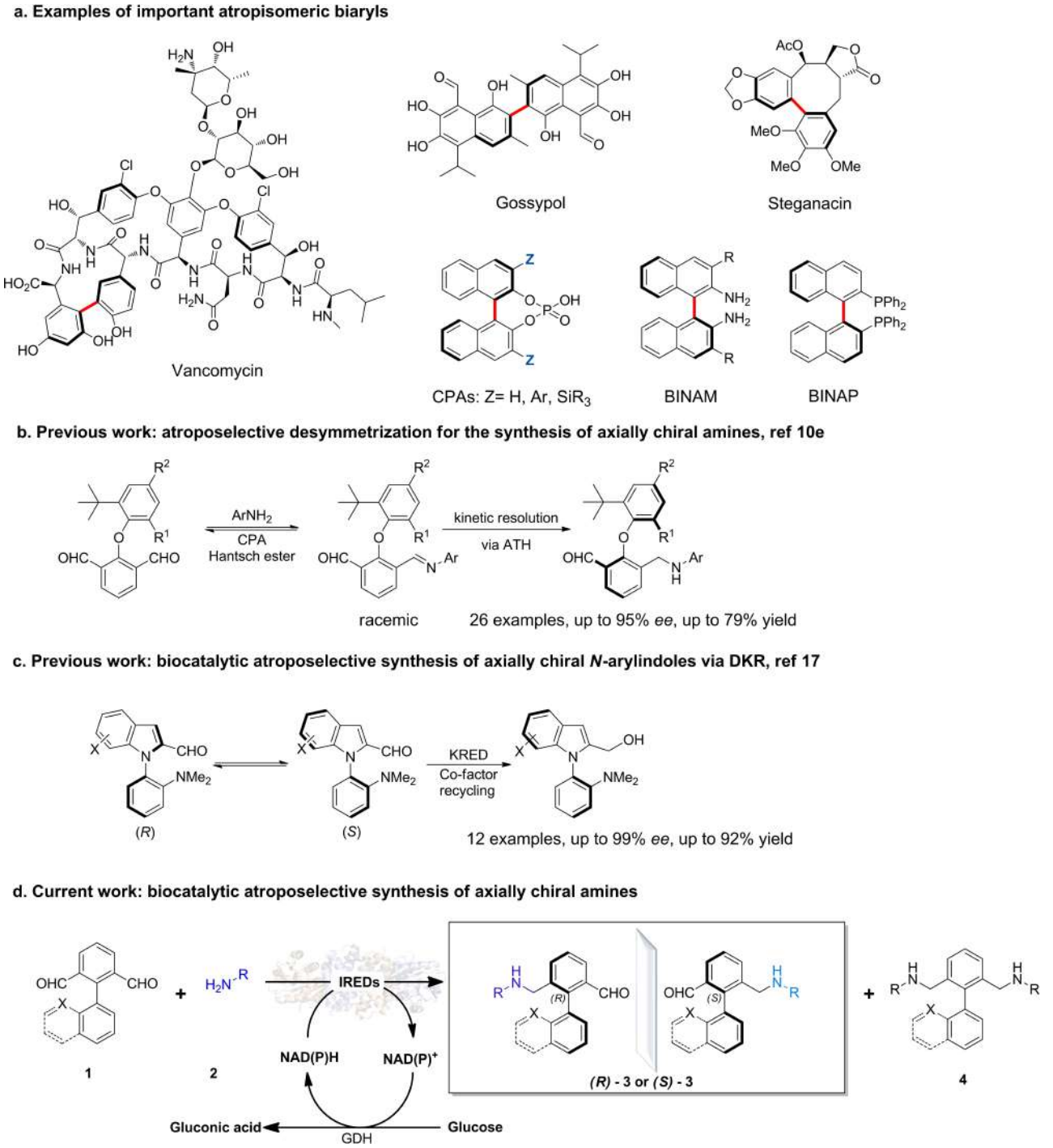
*Pengpeng Zhang[+], Bo Yuan[+], Junkuan Li, Congcong Li, Jiaxin Guo, Bowen Zhang, Ge Qu, Hao Su, Nicholas J. Turner, and Zhoutong Sun**

Abstract: The enzymatic atroposelective synthesis of biaryl compounds is relatively rare, despite considerable attention received by biocatalysis in the academic and industrial sectors. Imine reductases (IREDs) are an important class of enzymes that have been applied in the asymmetric synthesis of chiral amine building blocks. In this study, two IREDs (IR140 and IR189) were identified to catalyze the efficient desymmetrization of biaryls utilizing various amine donors. Further protein engineering enabled the identification of variants (IR189 M8–M9 and IR189 M13–M14) that are able to catalyze the formation of both (*R*) and (*S*) atropisomers in excellent yields and atroposelectivities (24 examples, up to 99 % *ee* and yield). The absolute configuration and rotational barriers were confirmed, and the reactions were readily scaled up to allow isolation of the atropisomeric product in 99 % *ee* and 82 % yield. The optically pure biaryl amines were further derivatized into various synthetically useful atropisomers. To shed light on the molecular recognition mechanisms, molecular dynamics (MD) simulations were performed, offering plausible explanations for the improved atroposelectivity and enzymatic activity. The current strategy expands the scope of the IRED-catalyzed synthesis of axially chiral biaryl amines, contributing significantly to the field of atroposelective biocatalysis.

Biaryl atropisomers play vital roles in natural products,[1] bioactive molecules,[2] molecular machines[3] and functional materials[4] (Scheme 1a), for example, Vancomycin, Gossypol, Steganacin, CPA (chiral phosphoric acids), BINAM (1,1'-binaphthyl-2,2'-diamine), BINAP (1,1′-Binaphthyl-2,2′-diphemyl phosphine2,2′-bis(diphenylphosphino)-1,1′-binaphthyl) are typical examples for useful products bearing axially chiral biaryl motifs.[5] Many excellent studies have been reported in the construction of axially chiral frameworks.[5c,6] For example, a number of methodologies were reported to construct C–C, C–N or N–N axially chiral compounds via asymmetric metal catalysis,[7] N-heterocyclic carbene (NHC) catalysis,[8] thiourea catalysis,[9] or chiral phosphoric acid (CPA) catalysis[10] (Scheme 1b). However, transition metal catalysts require directing groups that limit the substrate scope, heavy metal disposal that increases the cost, and stoichiometric oxidants that are usually toxic and generate byproducts, which also occurs when using other organic catalytic methodologies. In contrast, biocatalysis takes place in mild conditions and room temperature with high selectivity, especially biocatalysts are evolvable for customized substrates and selectivities.[11] Among the broad range of biocatalysts developed to enable the synthesis of a variety of structurally diverse compounds, imine reductases (IREDs) are an important class of enzymes that catalyze the reductive amination of ketones to yield chiral amines.[12] During the past decade, we and others have demonstrated the enormous potential that IREDs possess for the synthesis of chiral amine intermediates at industrially relevant scales.[13] However, until recently, biocatalysis has only been occasionally utilized in the synthesis of axially chiral compounds.[14] For example, enzymatic coupling has been exploited using laccases or P450s for atroposelective synthesis of biaryl compounds.[15] Especially, P450 was engineered to catalyze the coupling reaction for atropisomeric biaryl compounds, albeit in with moderate selectivity.[16] Rodríguez-Salamanca et al. explored the biocatalytic synthesis of axially chiral N-arylindoles with C–N axis using a panel of commercial ketoreductases (KREDs) (Scheme 1c).[17] Lewis et al. reported a DKR of 3-aryl-4(3H)-quinazolinones employed engineered flavin-dependent halogenases (FDHs) catalyzed atroposelective synthesis.[18] Previously, we reported the biocatalytic synthesis of axially chiral biaryl with C–C or C–O axis by a variant of galactose oxidase (GOase $M_{3\text{-}5}$),[19] alcohol dehydrogenases (ADHs) or ketoreductases (KREDs)[20] via desymmetrization, kinetic resolution (KR) and dynamic kinetic resolution (DKR).[21] Additionally, we and others utilized IREDs for the synthesis of various biaryls.[22] Very recently during the preparation of the current manuscript, Hao et al. reported an excellent example of synthesizing

[*] P. Zhang,[+] Dr. B. Yuan,[+] Dr. G. Qu, Dr. H. Su, Prof. Z. Sun
University of Chinese Academy of Sciences, Beijing 100049, P. R. China
E-mail: sunzht@tib.cas.cn
P. Zhang,[+] Dr. B. Yuan,[+] Dr. J. Li, Dr. C. Li, J. Guo, B. Zhang, Dr. G. Qu, Dr. H. Su, Prof. Z. Sun
Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, P. R. China
P. Zhang,[+] Dr. B. Yuan,[+] Dr. J. Li, Dr. C. Li, Dr. G. Qu, Dr. H. Su, Prof. Z. Sun
Key Laboratory of Engineering Biology for Low-Carbon Manufacturing, Tianjin 300308, P. R. China
Prof. N. J. Turner
Department of Chemistry, Manchester Institute of Biotechnology, University of Manchester, Manchester M1 7DN, U. K.
[+] These authors contributed equally to the current work.

© 2024 Wiley-VCH GmbH



Scheme 1. Atroposelective synthesis of biaryls. a) Axially chiral biaryl motifs in bioactive nature products and pharmaceuticals. b) Previous work utilizing chiral phosphoric acids as catalysts. c) Previous work utilizing ketoreductases for the DKR of the biaryls aldehydes. d) Current work investigating desymmetrization to produce axially chiral biaryl amines.

atropisomeric biaryls through dynamic kinetic resolution (DKR) catalyzed by IREDs.[22b] This study is clearly different from the current desymmetrization strategy, though both provide nice complementary routes leading to atropisomers. Herein we present identification and engineering of IREDs to further expand the biocatalytic toolbox to include a panel of diverse range of biaryls biaryl compounds synthesized by desymmetrization of prochiral dialdehydes, yielding both (*R*) and (*S*) atropisomeric amines in high efficiency and atroposelectivity (Scheme 1d).

To synthesize (*R*)- or (*S*)- **3a**, we screened our IRED library with 268 wild-type enzymes from various literature and databases.[23] The phylogenetic tree (Figure S1) was constructed and the IREDs from the library were identified

© 2024 Wiley-VCH GmbH

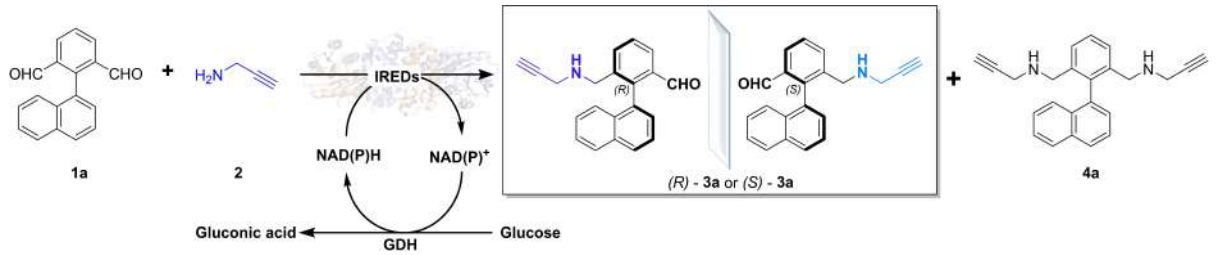
15213773, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202416569 by <Shibboleth>-member@ub.cas.cn, Wiley Online Library on [03/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

from diverse evolutionary origins. Using the pro-chiral biaryl substrate **1a** and the propargylamine **2** as the model substrates, initial results were obtained with a maximum of 99 % yield and enantiomeric excess (*ee*) for (*R*)-**3a** by IR140 from *Aspergillus udagawae.* On the other hand, only one (*S*)-selective IRED was identified from the library, which is IR189 from *Sandaracinus amylolyticus*, affording only 9 % *ee* in 96 % yield (Table 1 and Tables S1–S3). Subsequently, we optimized the dosage of amine donors and co-solvents utilizing the model substrate (Table S4).

For the purpose of obtaining highly atroposelective IREDs for both (*R*)- and (*S*)- biaryls, we initiated the protein engineering campaign on the (*S*)-selective IR189 (Table 2). The structure of IR189 was predicted by AF2,[24] and molecular docking analysis utilizing the imine (*S*)-**10** as a model substrate reveals several key residues (Figure S2). Initially, 14 sites (I118, M119, A120, T121, D169, L173, L176, S235, A236, A238, P239, G240, A242, S243) were selected as hot spots based on proximity to the substrate for formation of saturation libraries individually on each of these residues (Table S5–S6). The best-performing single mutants are IR189 M1 (A236P), M2 (G240A), M3 (S243M), M4 (T121F), M5 (L176F), yielding up to 96 % *ee* and 45 % of (*S*)-**3a**, which is a dramatic improvement comparing to 9 % *ee* when IR189 WT was utilized, though at the cost of enzymatic activity (composition of **3a** lowered from 96 % to 66 %). Further pairwise combinations of the best-performing single mutations led to the identification of the triple mutant IR189 M6 (A236P/G240A/S243M), with considerably improved activity (91 % **3a**) (Table S7). We then added further mutations to M6 to obtain IR189 M7–M9. The final quintuple mutants IR189 M8 (A236P/G240A/S243M/T121G/L176F) and M9 (A236P/G240A/S243M/T121F/L176F) afford 99 % *ee* for **3a**, with 55–94 % yields (Table 2a). Interestingly, with one point mutation on S235 or G240, the atroposelectivity of IR189 is reversed. IR189 M10, M11 and M12 turned out to be the best-performing (*R*)-selective single mutants, yielding up to 99 % *ee* and 52 % yield. Next, we combined the three single mutations pairwise to obtain IR189 M13 (S235M/G240H) and M14 (S235E/G240H), yielding 99 % *ee* and 66–68 % (*R*)-**3a** (Table 2b). As a result, we were able to obtain both (*S*)-selective (M8 and M9) and (*R*)-selective (M13 and M14) mutants for desymmetrization of **1a** utilizing only one enzyme (IR189).

Next, with the best mutants in hand, we further investigated a variety of biaryl dialdehydes (Scheme 2). IR140 performs consistently well to afford (*R*)-**3a–t** in up to 99 % *ee* and 99 % yield. The mutants of IR189 M13 and M14 also perform well and offer nice supplement to IR140. For example, IR189 M13 or M14 outperforms IR 140 to produce **3f**, **3s** and **3t** in higher *ee.* IR189 WT is weakly selective towards the model substrate **1a**, and its selectivity is reversed for substrates **1j–p**, **1r**, and **1t**. However, when highly (*S*)-selective mutants IR189 M8 and M9 were utilized, no reversal of the selectivity was observed. IR189 M8 and M9 offer good compensation, as M8 generally provides higher yields, while M9 typically results in higher atroposelectivities. For example, yields are over 90 % for **3b**, **3d**, **3f**, **3g, 3h, 3k**, **3m** when catalyzed by IR189 M8, whereas *ee* are over 90 % for **3b–e**, **3g**, **3p** by IR189 M9. Diverse electron-withdrawing or electron-donating substitutions on the naphthyl rings or phenyl rings of the biaryl substrates were well tolerated by IR140 WT, IR189 and its mutants. The quinoline substituted **3f–g** also show promising activity with up to 99 % *ee* and yields for both enantiomers. Further increasing the bulkiness of the biaryl motifs to include anthraphenyl building blocks, IR189 and its mutant M8 yield

Table 1: Screening of the IRED library for the model reaction.



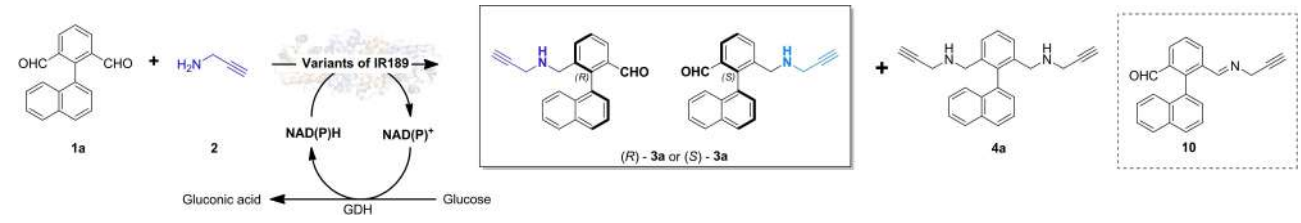
Entry	Enzyme	Source	3 a [%]	*ee* of 3 a [%]	Atroposelectivity
1	IR28	*Amycolatopsis keratiniphila*	75	99	*R*
2	IR42	*Cladophialophora immunda*	60	99	*R*
3	IR138	*Amycolatopsis regifaucium*	93	93	*R*
4	**IR140**	***Aspergillus udagawae***	**99**	**99**	***R***
5	IR165	*Aspergillus nidulans* FGSC A4	71	71	*R*
6	IR173	*Arabidopsis lyrata subsp. lyrata*	62	61	*R*
7	**IR189**	***Sandaracinus amylolyticus***	**96**	**9**	***S***
8	No IRED		ND*	–	–
9	No turnover system		ND*	–	–

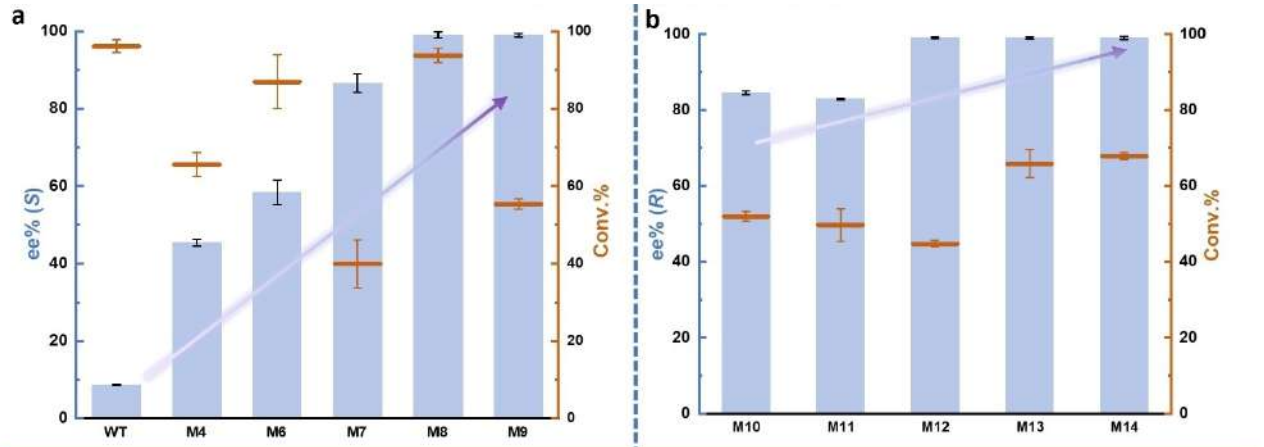
Reaction conditions: **1a** (5 mM), **2** (10 mM), DMSO (5 % v/v) in PBS buffer (50 mM, pH 7.4), IREDs (0.2 g mL^{-1} wet cell), lysozyme (1 mg mL^{-1}), DNase I (6 U mL^{-1}), GDH (1 mg mL^{-1}), NADP^{+} (1 mM), glucose (20 mM), 0.5 mL total volume, 30 °C, 1000 rpm, 12 h. All experiments were performed in duplicate. The HPLC determined yields are referred as 'yields' hereafter; *ee* values were determined by HPLC on a chiral stationary phase.* ND: not detected.

© 2024 Wiley-VCH GmbH

15213773, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202416569 by <Shibboleth>-member@ib.cas.cn, Wiley Online Library on [03/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2: Protein engineering of IR189.





Mutant	Mutations	3a	ee	4a	Sel.
WT	-	96%	9%	ND	(*S*)
M1	A236P	93%	39%	ND	(*S*)
M2	G240A	86%	31%	ND	(*S*)
M3	S243M	75%	22%	ND	(*S*)
M4	T121F	66%	45%	1	(*S*)
M5	L176F	96%	41%	ND	(*S*)
M6	A236P/G240A/S243M	91%	56%	3	(*S*)
M7	A236P/G240A/S243M /T121F	40%	85%	ND	(*S*)
M8	**A236P/G240A/S243 M/T121G/L176F**	**94%**	**99%**	**ND**	**(*S*)**
M9	**A236P/G240A/S243 M/T121F/L176F**	**55%**	**99%**	**ND**	**(*S*)**

Mutant	Mutations	3a	ee	4a	Sel.
M10	S235M	52%	85%	ND	(*R*)
M11	S235E	50%	83%	ND	(*R*)
M12	G240H	45%	99%	ND	(*R*)
M13	**S235M/G240H**	**66%**	**99%**	**ND**	**(*R*)**
M14	**S235E/G240H**	**68%**	**99%**	**ND**	**(*R*)**

Reaction conditions: **1a** (5 mM), **2** (10 mM), DMSO (5% v/v) in PBS buffer (50 mM, pH 7.4), IREDs (0.2 g mL^{-1} wet cell), lysozyme (1 mg mL^{-1}), DNase I (6 U mL^{-1}), GDH (1 mg mL^{-1}), NADP$^+$ (1 mM), glucose (20 mM), 0.5 mL total volume, 30 °C, 1000 rpm, 12 h. All experiments were performed in duplicate. Error bars display mean + s.d. The HPLC determined yields are referred as 'yields' hereafter, ee were determined by HPLC on a chiral stationary phase.* ND: not detected. Sel.: atroposelectivity.

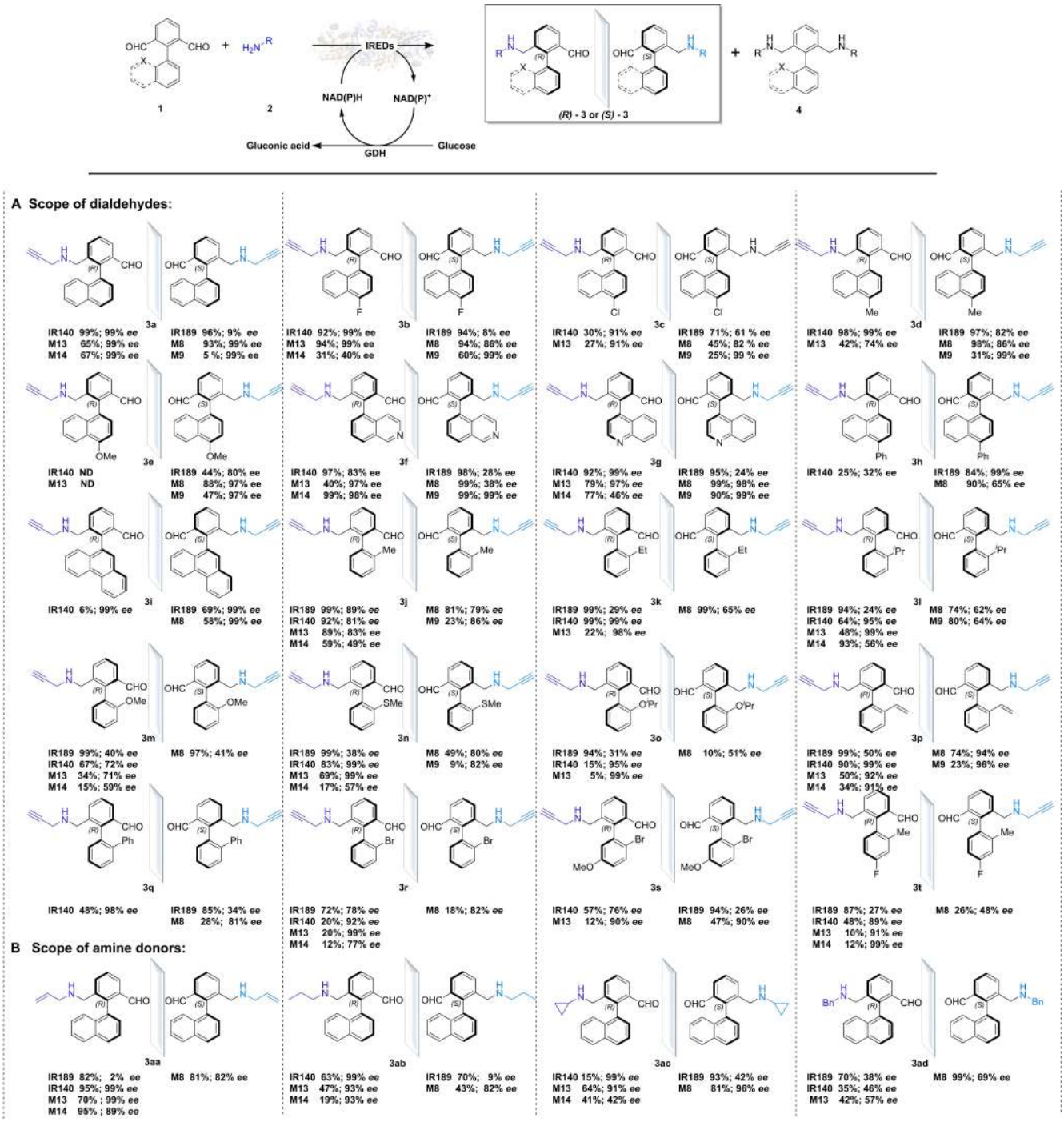
an impressive *ee* (99%) with moderate activity, while IR140 shows diminished activity for **1i**. For the bulky **3h**, IR189 WT affords an impressive 99% *ee* in 84% yield. **3j–t** are products with a smaller phenyl substitution, and the mutants performed well for substrates with either electron-withdrawing or electron-donating substitution groups. For example, *ee* are up to 99% for (*R*)-**3i**, **3k**, **3n**, **3o**, **3p**, **3r**, and **3t**, and *ee* are over 90% for (*S*)-**3p** and **3s**. Interestingly, the thioether product (*R*)-**3n** was obtained in excellent *ee* (99%) and good yield (83%), showing the high potential for the current strategy to be used in compounds bearing hetero-atoms. Finally, several amine donors were applied in the reductive aminations, including allylamine (**2a**), propylamine (**2b**), cyclopropylamine (**2c**) and benzylamine (**2d**). Mutants IR189 M8 and M9 outperform the IR189 WT in all of the screenings, whereas mutants IR189 M13 and M14 mostly improve the yields of the products since IR140 is highly selective for most of the occasions (99% *ee* for **3aa**–**3ad**).

To confirm the absolute conformation of the products, we reduced the aldehyde group on (*R*)-**3a** to an alcohol to obtain **5**, and the crystal structure was obtained via X-ray diffraction (CCDC:2380375, Table S8, Figure S3 and Scheme 3a).[25] Furthermore, the racemization barrier of **3a** was measured to be 32.8 kcal mol^{-1} (Table S9).

Finally, derivatization of (*R*)-**3a** was performed via a series of synthetic transformations to demonstrate the applicability of the current strategy (Scheme 3). We firstly

© 2024 Wiley-VCH GmbH

15213773, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202416569 by <Shibboleth>-member@ib.cas.cn, Wiley Online Library on [03/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



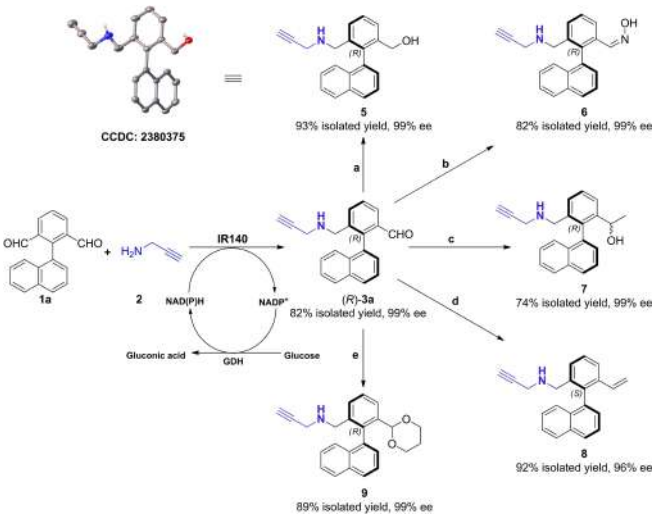
Scheme 2. Products from atroposelective desymmetrization of biaryl dialdehydes. Reaction conditions: **3a–t** (5 mM), **2a–e** (10 mM), DMSO (5% v/v) in PBS buffer (50 mM, pH 7.4), IRED (0.2 g mL^{-1} wet cell), lysozyme (1 mg mL^{-1}), DNase I (6 U mL^{-1}), GDH (1 mg mL^{-1}), NADP$^+$ (1 mM), glucose (20 mM), 0.5 mL total volume, 30 °C, 1000 rpm, 12 h. All experiments were performed in duplicate. Yields and ee were determined by HPLC on a chiral stationary phase.* ND: not detected.

performed a semi-preparative scale (200 mL, 1 mmol substrate) reaction of the IRED-catalyzed desymmetrization of **1a** utilizing IR140. High isolated yields (82 %) and excellent *ee* (99 %) of (*R*)-**3a** were obtained, subsequently, formation of oxime **6** via condensation with hydroxylamine hydrochloride (Scheme 3b), secondary alcohol **7** via Grignard reactions (Scheme 3c), alkene **8** via Wittig reactions (Scheme 3d), and cyclic acetal **9** via acetalizations (Scheme 3e) were achieved with little to no loss of *ee* (96–99 %) for the products in high isolated yields (74–92 %).

To further elucidate the catalytic mechanism underlying the atroposelectivities of the mutants, we conducted MD simulations on the IR189 WT, IR189 M8, and IR189 M13 in complex with the imine intermediate **10** formed from **1a** and

© 2024 Wiley-VCH GmbH

15213773, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202416569 by <Shibboleth>-member@ib.cas.cn, Wiley Online Library on [03/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Scheme 3. Derivatization of (*R*)-**3a**. Semi-preparative reaction conditions: **1a** (5 mM), **2** (10 mM), DMSO (5% v/v) in PBS buffer (50 mM, pH 7.4), IR140 (0.2 g mL^{-1} wet cell), lysozyme (1 mg mL^{-1}), DNase I (6 U mL^{-1}), GDH (1 mg mL^{-1}), NADP$^+$ (1 mM), glucose (20 mM), 200.0 mL total volume, 30 °C, 220 rpm, 12 h. a) $NaBH_4$ (1.2 equiv.), THF/H_2O = 4:1, 0 °C, 2 h. b) NH_2OH•HCl (1.2 equiv), Cs_2CO_3 (1.2 equiv), DMSO/H_2O = 3:1, 100 °C, 12 h. c) MeMgBr (1.2 equiv), dry THF, 0 °C, 2 h. d) $[MePPh_3]^+Br^-$ (1.2 equiv), K_2CO_3 (1.3 equiv), dioxane, 100 °C, 12 h. e) 1,3-propandiol (3.0 equiv), TsOH (1.0 equiv), toluene, 100 °C, 6 h.

2. Analysis of the RMSD fluctuations of the protein backbone atoms over time (Figure 1a) revealed that, after reaching convergence, both mutants exhibited greater conformational changes compared to the WT, with the M8 mutant showing particularly pronounced variations. Further analysis of the representative conformations indicated that mutations in the Loop$_{\beta5\beta6}$ and Helix 12 expanded the substrate-binding pocket (as indicated by the gray shading and the marked distances highlighted in Figure 1b,c), thereby providing adequate space for the substrate. As shown in Figure 1b, the mutation of T121 to G121 on Loop$_{\beta5\beta6}$ resulted in a smaller side chain, thereby reducing steric hindrance. In the Helix 12 region, partial unfolding occurred primarily due to the increased rigidity following the A236P mutation, which provided adequate space for the (*S*)-**3a** conformation. The mutation of G240 to A mitigated the excessive rigidity imposed by P236 on Helix 12. Additionally, F176 and M243 forms stable π–π and S-π non-bonded interactions (D2) with the substrate, respectively, collectively maintaining the enzyme's specific selectivity for imine (*R*)-**10**. For the M13 mutant, the two mutation sites established stable interactions with the naphthalene ring of the substrate, strongly favoring the binding of the (*R*)-enantiomer (Figure 1c). Analysis of the attack distance revealed that, compared to the WT, the two mutants exhibited a more stable attack distance (Figure 1d). On the other hand, the WT enzyme showed reduced selectivity due

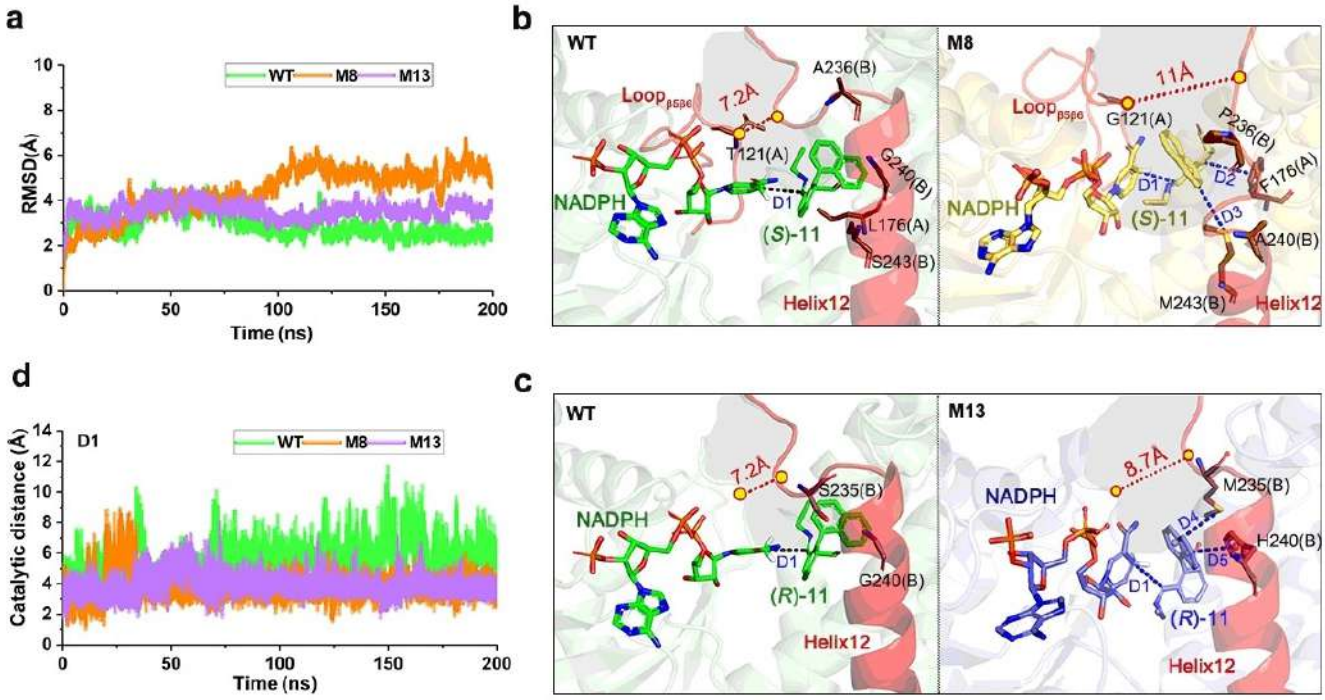


Figure 1. Molecular dynamics (MD) analysis for IR189 and its variants. a) Plot of RMSD values over time for the three simulated systems. b) Differences in interactions between WT and M8 mutant with imine (*S*)-**10** were observed. c) Distinctions in key residue interactions between WT and the M13 mutant with the imine (*R*)-**10** were noted. d) The time evolution of catalytic distances across the three systems was analyzed. Key conformational change regions (Loop$_{\beta5\beta6}$ and Helix12) were represented in red cartoon style. D1 indicates the catalytic distance, while D2, D3, D4, and D5 represent conjugated interactions. The changes in pocket size are highlighted with gray shading.

© 2024 Wiley-VCH GmbH

15213773, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202416569 by <Shibboleth>-member@ib.cas.cn, Wiley Online Library on [03/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to spatial limitations and difficulties in forming stable binding sites.

In summary, we have established the asymmetric synthesis of axially chiral biaryls via IRED-catalyzed reductive amination of aldehydes. IR140 and IR189 were identified and characterized to be utilized in the synthesis of atropisomeric amines in high *ee* and yields. The reactions were readily scaled up and further derivatized to afford novel products. This approach greatly expands the scope of biocatalytic reactions for valuable axially chiral compounds. Future work may include further expansion of the substrate scope to include a variety types of non-biaryl atropisomers as well as improving the atroposelectivity of IREDs to meet the industrialization needs.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2019YFA0905100), the National Natural Science Foundation of China (No. 32171462 and No. 32301277), Tianjin Synthetic Biotechnology Innovation Capacity Improvement Project (No. TSBICIP-CXRC-040), the Natural Science Foundation of Tianjin (No. 21JCJQJC00110).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords: atropisomers · imine reductases · protein engineering · biaryls · desymmetrization

[1] a) G. Bringmann, T. Gulder, T. A. M. Gulder, M. Breuning, *Chem. Rev.* **2011**, *111*, 563–639; b) M. C. Kozlowski, B. J. Morgan, E. C. Linton, *Chem. Soc. Rev.* **2009**, *38*, 3193–3207; c) S. H. Reisberg, Y. Gao, A. S. Walker, E. J. N. Helfrich, P. S. Baran, *Science* **2020**, *367*, 458–463; d) H. Aldemir, R. Richarz, T. A. M. Gulder, *Angew. Chem. Int. Ed.* **2014**, *53*, 8286–8293.

[2] a) J. Clayden, W. J. Moran, P. J. Edwards, S. R. LaPlante, *Angew. Chem. Int. Ed.* **2009**, *48*, 6398–6401; b) S. R. LaPlante, L. D. Fader, K. R. Fandrick, D. R. Fandrick, O. Hucke, R. Kemper, S. P. F. Miller, P. J. Edwards, *J. Med. Chem.* **2011**, *54*, 7005–7022; c) P. W. Glunz, *Bioorg. Med. Chem. Lett.* **2018**, *28*, 53–60.

[3] Y. Zhang, J. P. Calupitan, T. Rojas, R. Tumbleson, G. Erbland, C. Kammerer, T. M. Ajayi, S. Wang, L. A. Curtiss, A. T. Ngo, S. E. Ulloa, G. Rapenne, S. W. Hla, *Nat. Commun.* **2019**, *10*, 3742.

[4] a) L. Pu, *Acc. Chem. Res.* **2012**, *45*, 150–163; b) D. W. Zhang, M. Li, C. F. Chen, *Chem. Soc. Rev.* **2020**, *49*, 1331–1343.

[5] a) D. A. Evans, M. R. Wood, B. W. Trotter, T. I. Richardson, J. C. Barrow, J. L. Katz, *Angew. Chem. Int. Ed.* **1998**, *37*, 2700–2704; b) J. W. Jaroszewski, T. Strøm-Hansen, L. L. Hansen, *Chirality.* **1992**, *4*, 216–221; c) J. A. Carmona, C. Rodriguez-Franco, R. Fernandez, V. Hornillos, J. M. Lassaletta, *Chem. Soc. Rev.* **2021**, *50*, 2968–2983; d) D. L. Boger, S. Miyazaki, S. H. Kim, J. H. Wu, O. Loiseleur, S. L. Castle, *J. Am. Chem. Soc.* **1999**, *121*, 3226–3227.

[6] a) J. Wencel-Delord, A. Panossian, F. R. Leroux, F. Colobert, *Chem. Soc. Rev.* **2015**, *44*, 3418–3430; b) J. K. Cheng, S. H. Xiang, S. Li, L. Ye, B. Tan, *Chem. Rev.* **2021**, *121*, 4805–4902; c) E. Kumarasamy, R. Raghunathan, M. P. Sibi, J. Sivaguru, *Chem. Rev.* **2015**, *115*, 11239–11300; d) G.-J. Mei, W. L. Koay, C.-Y. Guan, Y. Lu, *Chem* **2022**, *8*, 1855–1893; e) X. F. Bai, Y. M. Cui, J. Cao, L. W. Xu, *Acc. Chem. Res.* **2022**, *55*, 2545–2561; f) N. Di Iorio, S. Crotti, G. Bencivenni, *Chem. Rec.* **2019**, *19*, 2095–2104; g) J. M. Lassaletta, *Atropisomerism and Axial Chirality*, Singapore, **2019**, chapter 6, *pp* 1–68; h) H. H. Zhang, T. Z. Li, S. J. Liu, F. Shi, *Angew. Chem. Int. Ed.* **2024**, *63*, e202311053; i) D. Liang, W.-J. Xiao, S. Lakhdar, J.-R. Chen, *Green Synth. Catal.* **2022**, *3*, 212–218.

[7] a) W. Yao, C. J. Lu, L. W. Zhan, Y. Wu, J. Feng, R. R. Liu, *Angew. Chem. Int. Ed.* **2023**, *62*, e202218871; b) T. Li, L. Shi, X. Wang, C. Yang, D. Yang, M. P. Song, J. L. Niu, *Nat. Commun.* **2023**, *14*, 5271; c) X. M. Wang, P. Zhang, Q. Xu, C. Q. Guo, D. B. Zhang, C. J. Lu, R. R. Liu, *J. Am. Chem. Soc.* **2021**, *143*, 15005–15010; d) Q. Xu, H. Zhang, F. B. Ge, X. M. Wang, P. Zhang, C. J. Lu, R. R. Liu, *Org. Lett.* **2022**, *24*, 3138–3143; e) S. Y. Yin, Q. S. J. Zhou, C. X. Liu, Q. Gu, S. L. You, *Angew. Chem. Int. Ed.* **2023**, *62*; f) J. Luo, T. Zhang, L. Wang, G. Liao, Q.-J. Yao, Y.-J. Wu, B.-B. Zhan, Y. Lan, X.-F. Lin, B.-F. Shi, *Angew. Chem. Int. Ed.* **2019**, *58*, 6708–6712; g) J. Zhang, J. Wang, *Angew. Chem. Int. Ed.* **2018**, *57*, 465–469.

[8] a) S. J. Wang, X. Wang, X. Xin, S. Zhang, H. Yang, M. W. Wong, S. Lu, *Nat. Commun.* **2024**, *15*, 518; b) Y. Wu, M. Li, J. Sun, G. Zheng, Q. Zhang, *Angew. Chem. Int. Ed.* **2022**, *61*, e202117340; c) R. Song, Y. Xie, Z. Jin, Y. R. Chi, *Angew. Chem. Int. Ed.* **2021**, *60*, 26026–26037; d) S. Lu, S. B. Poh, Y. Zhao, *Angew. Chem. Int. Ed.* **2014**, *53*, 11041–11045; e) S. Shee, S. Shree Ranganathappa, M. S. Gadhave, R. Gogoi, A. T. Biju, *Angew. Chem. Int. Ed.* **2023**, *62*, e202311709.

[9] a) W. Lin, Q. Zhao, Y. Li, M. Pan, C. Yang, G. H. Yang, X. Li, *Chem. Sci.* **2021**, *13*, 141–148; b) E. S. Munday, M. A. Grove, T. Feoktistova, A. C. Brueckner, D. M. Walden, C. M. Young, A. M. Z. Slawin, A. D. Campbell, P. H. Cheong, A. D. Smith, *Angew. Chem. Int. Ed.* **2020**, *59*, 7897–7905.

[10] a) K. W. Chen, Z. H. Chen, S. Yang, S. F. Wu, Y. C. Zhang, F. Shi, *Angew. Chem. Int. Ed.* **2022**, *61*, e202116829; b) Y. Wei, F. Sun, G. Li, S. Xu, M. Zhang, L. Hong, *Org. Lett.* **2024**, *26*, 2343–2348; c) J. Qin, T. Zhou, T. P. Zhou, L. Tang, H. Zuo, H. Yu, G. Wu, Y. Wu, R. Z. Liao, F. Zhong, *Angew. Chem. Int. Ed.* **2022**, e202205159; d) H. Bao, Y. Chen, X. Yang, *Angew. Chem. Int. Ed.* **2023**, *62*, e202300481; e) L. Dai, Y. Liu, Q. Xu, M. Wang, Q. Zhu, P. Yu, G. Zhong, X. Zeng, *Angew. Chem. Int. Ed.* **2023**, *62*, e202216534.

[11] a) E. L. Bell, W. Finnigan, S. P. France, A. P. Green, M. A. Hayes, L. J. Hepworth, S. L. Lovelock, H. Niikura, S. Osuna, E. Romero, K. S. Ryan, N. J. Turner, S. L. Flitsch, *Nat. Rev. Methods Primers 1.* **2021**, *46*; b) S. Wu, R. Snajdrova, J. C. Moore, K. Baldenius, U. T. Bornscheuer, *Angew. Chem. Int. Ed.* **2021**, *60*, 88–119.

[12] a) B. Yuan, D. Yang, G. Qu, N. J. Turner, Z. Sun, *Chem. Soc. Rev.* **2024**, *53*, 227–262; b) K. Wu, J. Huang, L. Shao, *ChemCatChem* **2022**, *14*, e202200921; c) A. K. Gilio, T. W. Thorpe, N. Turner, G. Grogan, *Chem. Sci.* **2022**, *13*, 4697–4713.

[13] a) M. Schober, C. MacDermaid, A. A. Ollis, S. Chang, D. Khan, J. Hosford, J. Latham, L. A. F. Ihnken, M. J. B. Brown, D. Fuerst, M. J. Sanganee, G.-D. Roiban, *Nat. Catal.* **2019**, *2*, 909–915; b) R. Kumar, M. J. Karmilowicz, D. Burke, M. Burns,

© 2024 Wiley-VCH GmbH

15213773, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202416569 by <Shibboleth>-member@ib.cas.cn, Wiley Online Library on [03/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

L. A. Clark, C. G. Connor, E. Cordi, N. M. Do, K. M. Doyle, S. Hoagland, C. A. Lewis, D. Mangan, C. A. Martinez, E. L. McInturff, K. Meldrum, R. Pearson, J. Steflik, A. Rane, J. Weaver, *Nat. Catal.* **2021**, *4*, 775–782; c) T. W. Thorpe, J. R. Marshall, V. Harawa, R. E. Ruscoe, A. Cuetos, J. D. Finnigan, A. Angelastro, R. S. Heath, F. Parmeggiani, S. J. Charnock, R. M. Howard, R. Kumar, D. S. B. Daniels, G. Grogan, N. J. Turner, *Nature* **2022**, *604*, 86–91; d) B.-B. Li, J. Zhang, F.-F. Chen, Q. Chen, J.-H. Xu, G.-W. Zheng, *Green Synth. Catal.* **2021**, *2*, 345–349; e) P. Wu, D. Luo, Y Wang, X. Shang, B. Wang, X. Deng, J. Yuan, *BioDesign Res.* **2024**, *6*, 0048.

[14] a) B. Skrobo, J. D. Rolfes, J. Deska, *Tetrahedron* **2016**, *72*, 1257–1275; b) O. F. B. Watts, J. Berreur, B. S. L. Collins, J. Clayden, *Acc. Chem. Res.* **2022**, *55*, 3362–3375.

[15] S. Obermaier, W. Thiele, L. Furtges, M. Muller, *Angew. Chem. Int. Ed.* **2019**, *58*, 9125–9128.

[16] L. E. Zetzsche, J. A. Yazarians, S. Chakrabarty, M. E. Hinze, L. A. M. Murray, A. L. Lukowski, L. A. Joyce, A. R. H. Narayan, *Nature* **2022**, *603*, 79–85.

[17] P. Rodríguez-Salamanca, G. de Gonzalo, J. A. Carmona, J. López-Serrano, J. Iglesias-Sigüenza, R. Fernández, J. M. Lassaletta, V. Hornillos, *ACS Catal.* **2023**, *13*, 659–664.

[18] H. M. Snodgrass, D. Mondal, J. C. Lewis, *J. Am. Chem. Soc.* **2022**, *144*, 16676–16682.

[19] B. Yuan, A. Page, C. P. Worrall, F. Escalettes, S. C. Willies, J. J. W. McDouall, N. J. Turner, J. Clayden, *Angew. Chem. Int. Ed.* **2010**, *49*, 7010–7013.

[20] a) S. Staniland, B. Yuan, N. Gimenez-Agullo, T. Marcelli, S. C. Willies, D. M. Grainger, N. J. Turner, J. Clayden, *Chem. Eur. J.* **2014**, *20*, 13084–13088; b) M. Ye, C. Li, D. Xiao, G. Qu, B. Yuan, Z. Sun, *JACS Au* **2024**, *4*, 411–418.

[21] S. Staniland, R. W. Adams, J. J. W. McDouall, I. Maffucci, A. Contini, D. M. Grainger, N. J. Turner, J. Clayden, *Angew. Chem. Int. Ed.* **2016**, *55*, 10755–10759.

[22] a) S. P. France, G. A. Aleku, M. Sharma, J. Mangas-Sanchez, R. M. Howard, J. Steflik, R. Kumar, R. W. Adams, I. Slabu, R. Crook, G. Grogan, T. W. Wallace, N. J. Turner, *Angew. Chem. Int. Ed.* **2017**, *56*, 15589–15593; b) X. Hao, Z. Tian, Z. Yao, T. Zang, S. Song, L. Lin, T. Qiao, L. Huang, H. Fu, *Angew. Chem. Int. Ed.* **2024**. e202410112.

[23] a) J. Zhang, D. Liao, R. Chen, F. Zhu, Y. Ma, L. Gao, G. Qu, C. Cui, Z. Sun, X. Lei, S. S. Gao, *Angew. Chem. Int. Ed.* **2022**, *61*, e202201908; b) S. L. Montgomery, A. Pushpanath, R. S. Heath, J. R. Marshall, U. Klemstein, J. L. Galman, D. Woodlock, S. Bisagni, C. J. Taylor, J. Mangas-Sanchez, J. I. Ramsden, B. Dominguez, N. J. Turner, *Sci. Adv.* **2020**, *6*, eaay9320; c) J. R. Marshall, P. Yao, S. L. Montgomery, J. D. Finnigan, T. W. Thorpe, R. B. Palmer, J. Mangas-Sanchez, R. A. M. Duncan, R. S. Heath, K. M. Graham, D. J. Cook, S. J. Charnock, N. J. Turner, *Nat. Chem.* **2021**, *13*, 140–148.

[24] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589.

[25] Deposition number CCDC 2380375 (for **5**) contains the supplementary crystallographic data for this paper. These data are provided free of charge by the joint Cambridge Crystallographic Data Centre and Fachinformationszentrum Karlsruhe Access Structures service (http://www.ccdc.cam.ac.uk/structures).

Manuscript received: August 29, 2024
Accepted manuscript online: September 13, 2024
Version of record online: October 31, 2024

© 2024 Wiley-VCH GmbH

15213773, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202416569 by <Shibboleth>-member@ib.cas.cn, Wiley Online Library on [03/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

中国科学院大学
University of Chinese Academy of Sciences

前沿交叉科学学院
School of Advanced Interdisciplinary Sciences

中国科学院大学2025年研究生前沿交叉学科学术论坛论文成果

姓名：钟晓彤

方向：生物制造

单位：中国科学院微生物研究所

题目：人工智能助力解锁黏细菌天然产物生物合成新机制

nature communications

Article

https://doi.org/10.1038/s41467-025-63441-y

Genetic engineering of *Sorangium cellulosum* reveals hidden enzymology in myxobacterial natural product biosynthesis

Received: 19 February 2025

Accepted: 18 August 2025

Published online: 27 August 2025

Check for updates

Xiaotong Zhong[1,2,3,5], Shan Liu[1,2,3,5], Bingda Ma[1,2,3], Kaining Gao[1,2,3], Dayong Jiang[1,2,3], Yingshuo Hou[1,2], Huliang Chen[1,2,3], Jiaqi Lv[1,2], James I. Bowen[4], Matthew P. Crump[4], Christine L. Willis[4] ✉ & Luoyi Wang[1,2,3] ✉

Sorangium cellulosum is a cellulolytic myxobacterium that produces a vast array of complex natural products with diverse chemical scaffolds and biological activities. However, biosynthetic investigations of these metabolites have been hindered by the scarcity of genetic manipulation tools available for their producing microorganisms. Here, we develop an efficient electroporation method for transforming foreign DNA into various *Sorangium* strains, enabling effective genetic engineering via homologous recombination. This facilitates delineation of the biosynthetic pathway to ambruticin, unveiling several previously undisclosed steps. Notably, AmbK is identified as the elusive epoxide hydrolase responsible for the formation of the tetrahydropyran ring during post-polyketide synthase (PKS) modification, while the terminal PKS module AmbH is shown to catalyse dual rounds of chain elongation during polyketide assembly. Our findings provide significant insights into the intricate molecular machinery governing myxobacterial natural product biosynthesis and greatly enhance our ability to further engineer *Sorangium* strains to unlock their biosynthetic potentials.

Natural products derived from microbial sources have long been a cornerstone of pharmaceutical drug discovery due to their structural diversity and therapeutic potential[1–3]. Among the myriad of microbial natural product producers, such as fungi and actinomycetes, the Gram-negative myxobacteria have gained increasing recognition for their rich secondary metabolism in recent decades[4–6]. *Sorangium cellulosum*, a cellulolytic member of the myxobacteria, accounts for nearly half of all known metabolites produced by this bacterial group[7]. These include the anticancer epothilones[8], the antifungal soraphens[9], the antibacterial sorangicins[10] and the antiviral ratjadones[11] (Fig. 1a). Strains of *Sorangium* are distinguished by their unusually large genomes, reaching up to 15 Mbp, which encode a wide array of secondary metabolite biosynthetic gene clusters. This genomic richness underscores their substantial potential for natural product discovery, offering a diverse chemical repertoire for biotechnological and pharmaceutical applications[12,13]. Prior studies have suggested that *Sorangium* harbours biosynthetic machinery unique to myxobacteria, featuring rare and unusual enzymology not typically observed in other microbial organisms[14].

Genetic manipulation of *Sorangium* strains is crucial for exploiting their biosynthetic potential and gaining an in-depth understanding of these intricate pathways. However, such efforts have been extremely challenging due to the lack of robust genetic modification tools[15,16]. These organisms are slow-growing and possess a complex social

[1]State Key Laboratory of Microbial Diversity and Innovative Utilization, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [2]Beijing Key Laboratory of Genetic Element Biosourcing & Intelligent Design for Biomanufacturing, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [3]University of Chinese Academy of Sciences, Beijing, China. [4]School of Chemistry, University of Bristol, Bristol, UK. [5]These authors contributed equally: Xiaotong Zhong, Shan Liu. ✉e-mail: chris.willis@bristol.ac.uk; wangluoyi@im.ac.cn

https://doi.org/10.1038/s41467-025-63441-y

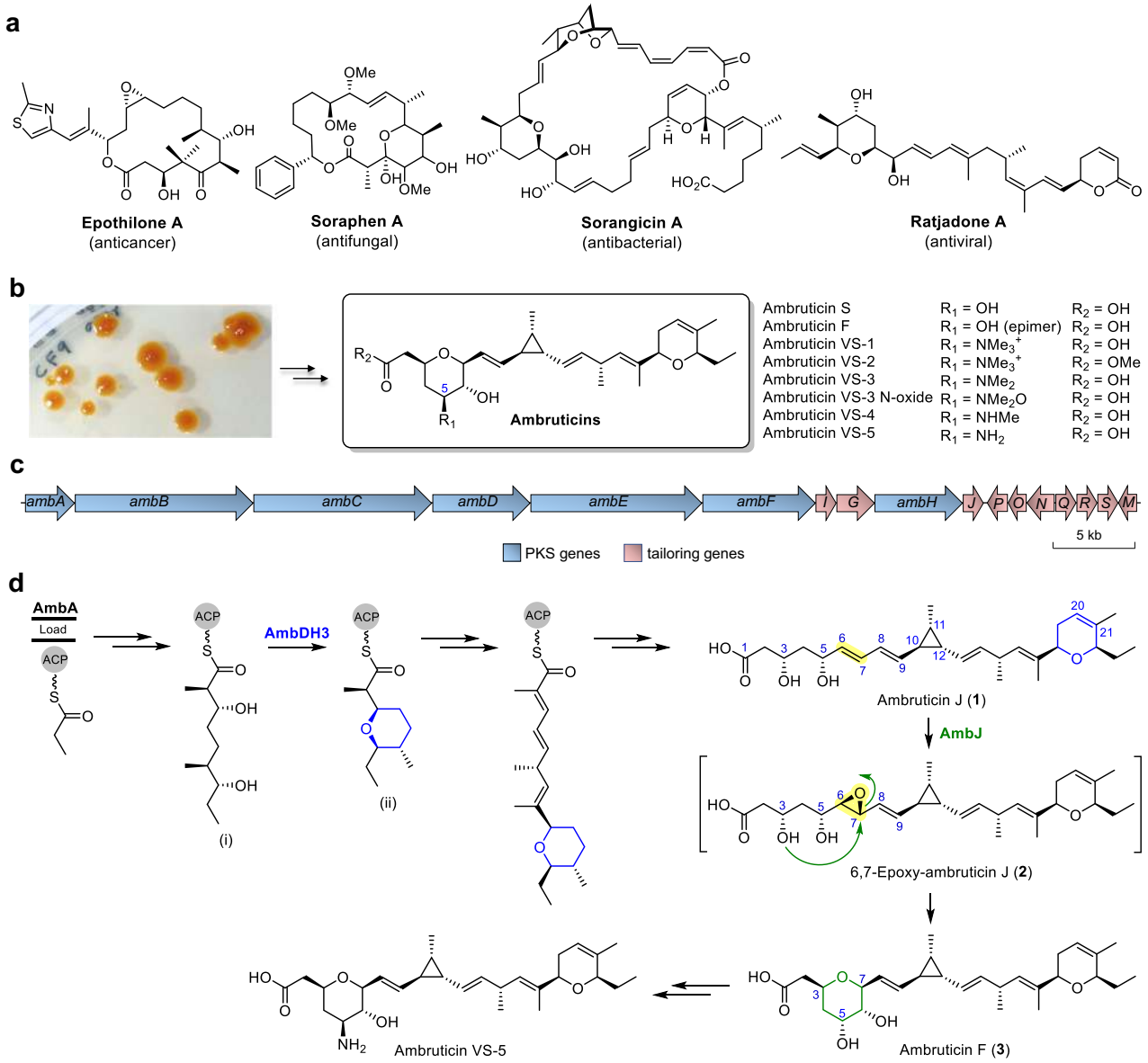


Fig. 1 | Selected bioactive natural products produced by *S. cellulosum* and their biosynthesis. a Examples of natural products from *S. cellulosum* with diverse chemical scaffolds and biological activities. **b** Colonies of *S. cellulosum* So ce10 showing a distinctive red colour on agar plate and the structures of ambruticins. **c** Genetic organisation of the *amb* gene cluster. **d** Previously proposed biosynthetic pathway of ambruticin showing key steps involving tetrahydropyran ring formation, ACP = acyl carrier protein, see Supplementary Fig. 1 for the full pathway.

lifecycle, including fruiting body formation, which complicates their cultivation and genetic manipulation. The extensive GC-rich genome of *S. cellulosum*, containing many repetitive sequences and intrinsic antibiotic resistance genes, poses further difficulties for genomic editing[12,17]. Previously, plasmids that could insert into the middle of the target gene were mainly introduced into *Sorangium* cells through conjugation with a helper *E. coli* strain, and gene disruptions were achieved via a single-crossover recombination event[18,19]. The conjugation approach is time-consuming and has been found to be strain-specific; protocols developed for one particular strain often cannot be applied to another[20]. Electroporation has been reported scarcely and has not seen widespread application[21].

The ambruticins were first reported in the late 1970s and are among the earliest characterised metabolites produced by *Sorangium* strains[22,23]. They have since attracted sustained interest due to their unusual chemical architectures and potent antifungal activities against a variety of fungal pathogens, including *Coccidioides immitis*, *Histoplasma capsulatum*, and *Blastomyces dermatitidis*[24–30]. The ambruticins share several key structural features including a dihydropyran (DHP) and tetrahydropyran (THP) with an unsaturated hydrocarbon linker and an unusual trisubstituted cyclopropyl ring, but differ mainly in the functionality at C-5 (Fig. 1b). In 2006, Reeves and co-workers reported analysis of the ambruticin biosynthetic gene cluster (BGC) in *S. cellulosum* So ce10 and proposed several unusual features of the modular biosynthetic pathway and post-PKS tailoring based on extensive bioinformatic analysis, selected gene disruption experiments and isotopic labelling studies (Fig. 1c)[31]. Interestingly, one of the PKS modules in AmbE was hypothesised to catalyse three iterative chain elongations and downstream processes, including a Favorskii rearrangement and a carbon atom excision event would afford the cyclopropyl intermediate for further modifications (Supplementary Fig. 1)[31]. However, these steps remain speculative and await further experimental validation.

Since 2014, Hahn and colleagues have undertaken a series of biosynthetic investigations on the ambruticins and the closely related jerangolids[32–38]. Through in vitro assays using synthetic substrate surrogates, their studies on key enzymes have elucidated several steps involved in ambruticin biosynthesis. The DHP is formed via an intramolecular oxa-Michael addition catalysed by AmbDH3 of linear polyketide (Fig. 1d, structure i) to give THP (Fig. 1d, structure ii)[32]. AmbDH3 has broad substrate specificity and has been used in the total synthesis of the diarylheptanoid centrolobine[39]. More recent studies indicate that oxidation of the THP to the DHP occurs prior to generation of the C3 to C7 dihydroxylated tetrahydropyran in ambruticin biosynthesis, but the precise timing remains unknown[40].

Using a combination of homology modelling, phylogenetic analysis, and in vitro experiments, Hahn and co-workers reannotated AmbG as a fatty acyl:adenylate ligase (FAAL)−acyl carrier protein (ACP) didomain enzyme with an unusually broad substrate tolerance, providing an important advancement in the understanding of the biosynthesis of the ambruticin middle fragment[37]. Nevertheless, much of the fundamental biosynthetic machinery of ambruticin, including formation of the central cyclopropane ring and generation of the C3-C7 THP, remains elusive. This underscores the need for further exploration utilising advanced genetic engineering techniques to elucidate the remaining complexities involved in its biosynthesis.

Using the ambruticin-producing strain *S. cellulosum* So ce10 as a model system, we here present an efficient approach to genetically manipulate *Sorangium* strains and delineate the intricate molecular machinery of ambruticin biosynthesis. Through targeted genetic engineering of the ambruticin BGC and characterisation of key metabolites from generated mutant strains, combined with feeding studies and whole-cell biotransformation, we have established the roles of the otherwise cryptic AmbK in THP formation and AmbH in PKS module iteration in ambruticin biosynthesis. Our work marks a significant step towards a deeper understanding of the complex chemistry and biology of ambruticin biosynthesis, laying the groundwork for further investigations aimed at elucidating the mechanistic details of unusual enzymology in polyketide assembly and modifications.

Results

Development of the genetic engineering method for *S. cellulosum*

Strains of *S. cellulosum* typically exhibit bright pigmentation attributed to the presence of carotenoids (Fig. 1b). These pigments are usually biosynthesised through a well-conserved biosynthetic pathway (Fig. 2a)[41]. Inactivation or knockout of the *crtB* gene, which encodes phytoene synthase, a key enzyme in carotenoid biosynthesis, has been observed to cause an orange-to-white phenotype change in many microorganisms[42–44]. Our comparative analysis of all available whole-genome sequences of *S. cellulosum* revealed a high level of conservation in their carotenoid biosynthetic gene clusters (Supplementary Fig. 2 and Supplementary Table 1), indicating that *crtB* could potentially serve as a reporter gene for *Sorangium*.

To construct the plasmid for targeted inactivation of *crtB* in *S. cellulosum* So ce10, an antibiotic selection marker must be inserted between the upstream and downstream homology arms of the target site. Due to intrinsic antibiotic resistance genes in their genome, strains of *Sorangium* are resistant to many commonly used antibiotics[12]. Our preliminary screening indicated that *S. cellulosum* So ce10 is sensitive to hygromycin B; therefore, the hygromycin-resistance gene *hyg* was initially chosen as the selection marker. For the plasmid backbone, the pEX18 suicide vector, which has been widely used for Gram-negative bacteria[45,46], was selected. Consequently, the *hyg* marker flanked by upstream and downstream homology arms of ~1500 bp was ligated with the backbone vector to create the knockout plasmid targeting *crtB*, designated as pXZ-crtBKO (Fig. 2b).

We next aimed to deliver the plasmid pXZ-crtBKO into *S. cellulosum* So ce10 by electroporation. Electrocompetent cells prepared in 10% glycerol were mixed with plasmid DNA, electroporated, and immediately transferred to recovery medium. The recovered cells were then plated on hygromycin-containing selection plates. After 5-7 days of incubation, both orange and white colonies were observed, indicating successful transformation (Fig. 2c, i). By homologous recombination, a single-crossover event would integrate the plasmid into the genome without disrupting *crtB*, thus retaining the orange phenotype (Fig. 2bi, ii). In contrast, a double-crossover event would result in the replacement of *crtB* by the hygromycin selection marker, leading to a white phenotype (Fig. 2biii). PCR verification confirmed that the red and white colonies are indeed single-crossover and double-crossover transformants, respectively (Fig. 2cii).

To enhance our workflow efficiency, we extensively screened and optimised electroporation conditions. This included adjusting optical density and the temperature of competent cells, as well as refining electroporation voltage and the number of pulses (Supplementary Table 2). A transformation efficiency of up to 1.2×10^3 colony-forming units (CFU)/μg DNA was achieved when electrocompetent cells were prepared at room temperature[47] using cells at OD_{600} 0.5, with electroporation carried out at a voltage of 15 kV/cm for 3 times (Fig. 2d). In addition, the impact of homology arm length was explored. We determined that a minimum arm length of 1000 bp is necessary for efficient homologous recombination (Fig. 2e). However, an arm length of ~1500 bp is preferred to maximise the occurrence of double-crossover events (Fig. 2f). Similar results were observed when using other plasmid backbones such as pKC1139 and pYH7, which are commonly employed for gene knockout studies in *Streptomyces* strains[48,49]. For sequential genetic manipulation, alternative antibiotic selection markers were investigated. Both the tetracycline-resistance gene and the chloramphenicol-resistance gene were found to be effective markers for So ce10, albeit with lower efficiency compared to the hygromycin marker.

The versatility of our electroporation method was tested on other *Sorangium* strains, including the jerangolid producer So ce307 and the epothilone producer So ce836. In both cases, successful disruption of the *crtB* gene was achieved with comparable efficiency to that of So ce10 (Fig. 2g and Supplementary Fig. 3). Notably, previous attempts to genetically manipulate So ce307 using conjugation methods were unsuccessful[31] while no prior genetic manipulation studies have been reported for So ce836[50–52]. To directly benchmark our electroporation protocol against the traditional conjugation-based method, we investigated *crtB* inactivation in these strains using both approaches in parallel. Despite repeated efforts, the conjugation method failed to yield any transformants in So ce307 or So ce836. In the case of So ce10, only a few single-crossover mutants were obtained, and no double-crossover mutants could be recovered (Supplementary Fig. 4). Additionally, the conjugation process was complicated by substantial background growth of the donor *E. coli* S17.1 strain. In contrast, our electroporation protocol consistently achieved efficient transformation and a reproducible frequency of double-crossover events across strains, highlighting both its technical advancement and broader applicability (Supplementary Table 3).

AmbK is an epoxide hydrolase catalysing THP ring formation

Our next goal was to investigate key steps in the ambruticin biosynthetic pathway in *S. cellulosum* So ce10 using the genetic engineering strategy we had established, beginning with formation of the C3-C7 THP ring. In their biosynthetic studies, Reeves and co-workers disrupted *ambJ*, which encodes a flavin-dependent epoxidase and isolated ambruticin J (**1**) as the putative PKS product lacking the 6,7-epoxide[31]. Thus, it is hypothesised that AmbJ catalyses selective epoxidation of the 6,7-double bond to give 6,7-epoxy-ambruticin J (**2**), followed by epoxide-opening at C-7 through nucleophilic attack by the

https://doi.org/10.1038/s41467-025-63441-y

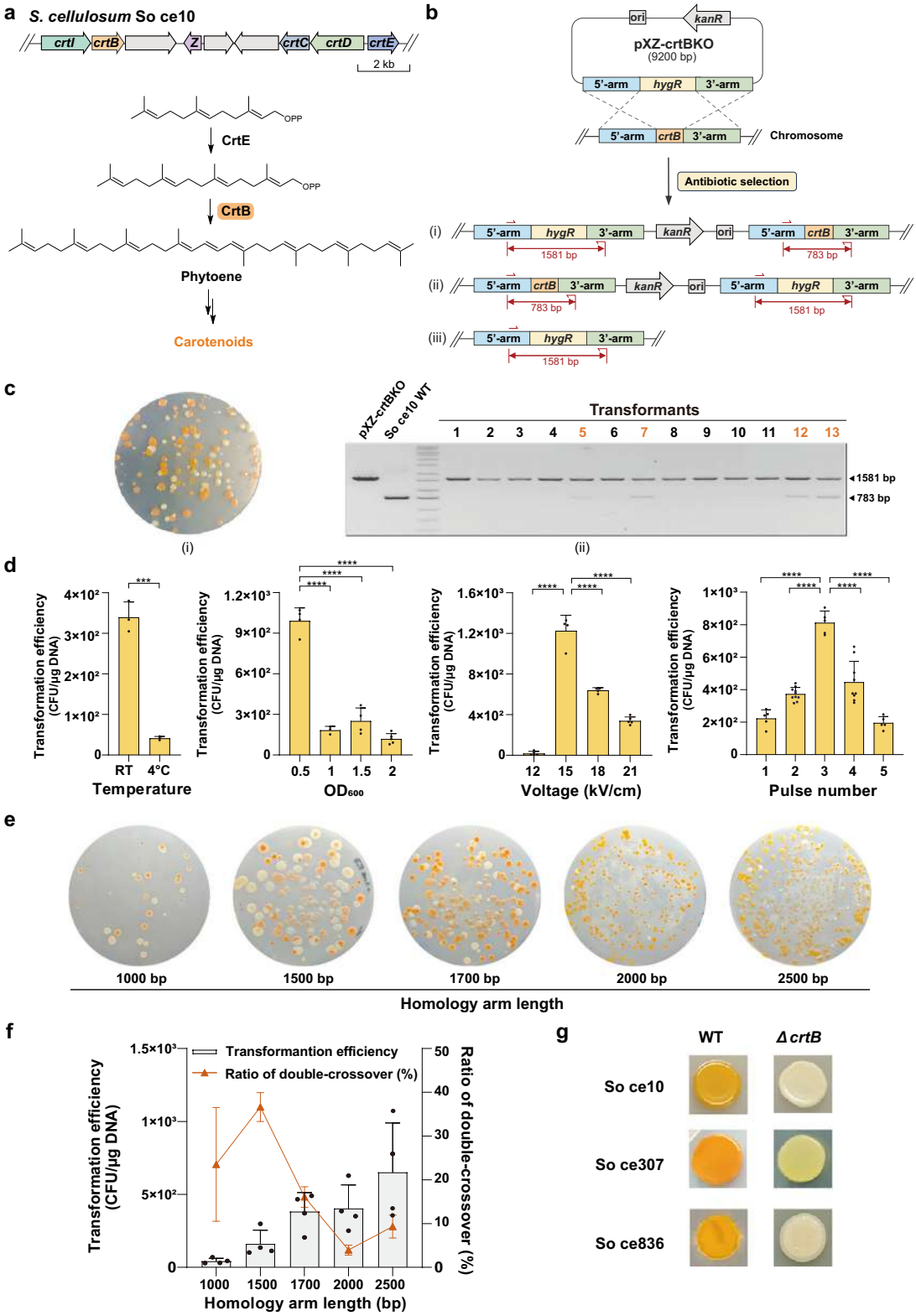


Fig. 2 | Development of the genetic engineering method for *S. cellulosum*. **a** Biosynthesis of carotenoids in *S. cellulosum* So ce10. **b** Construction of the pXZ-crtBKO plasmid and possible outcomes from homologous recombination. **c** Verification of *crtB* disruption by phenotype (i) and genotype (ii). The colour of the numbers above the gel lanes indicates the colony appearance: orange numbers indicate colonies with an orange phenotype, and black numbers indicate colonies with a white phenotype. A total of 42 white colonies from three independent biological replicates ($N = 3$) were validated by genotype, uncropped gels are provided in the Source data file. **d** Optimisation of the electroporation conditions. Error bars represent mean ± SD. $N = 3$ biological replicates for each temperature condition; $N = 4, 3, 4, 5$ replicates for OD_{600} of 0.5, 1, 1.5, 2, respectively; $N = 5, 4, 4, 5$ replicates for Voltage of 12, 15, 18, 21 kV/cm, respectively; $N = 5, 10, 5, 9, 5$ replicates for pulse number of 1, 2, 3, 4, 5, respectively. **e**, **f** Impact of homology arm length on overall transformation efficiency and ratio of double-crossover events. Error bars represent mean ± SD, $N = 4$ biological replicates. **g** Disruption of the *crtB* gene in *S. cellulosum* So ce307 and So ce836. Statistical analysis was performed using two-tailed Student's *t* test or one-way ANOVA followed by Tukey's multiple comparisons test (***$p < 0.001$; ****$p < 0.0001$). Source data are provided as a Source data file.

https://doi.org/10.1038/s41467-025-63441-y

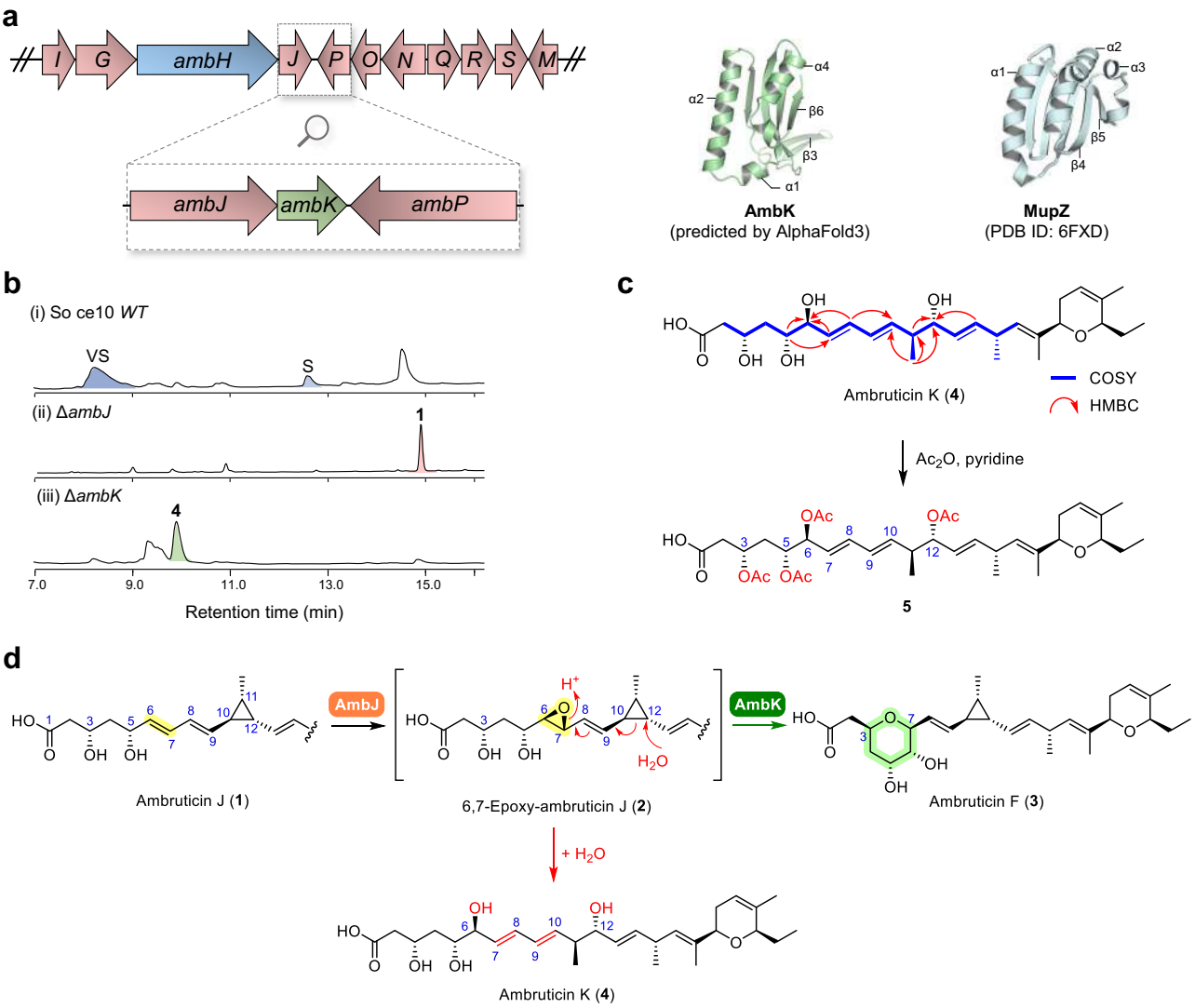


Fig. 3 | AmbK is an epoxide hydrolase catalysing THP ring formation in ambruticin biosynthesis. a Identification of *ambK* from the *amb* BGC. **b** HPLC traces of the So ce10 Δ*ambJ* and Δ*ambK* mutants. **c** Structural elucidation of ambruticin K (**4**). **d** Confirmation of AmbK as the epoxide hydrolase responsible for THP ring formation.

3-hydroxyl group to form the six-membered THP ring and thereby generating ambruticin F (**3**) (Fig. 1d). Similar epoxide formation/epoxide-opening cascades have been identified in the biosynthesis of lasalocid[53,54], mupirocin[46] and xiamenmycin[55], and in each case an additional epoxide hydrolase (Lsd19, MupZ and XimE respectively) is required to favour formation of the typically disfavoured THP ring[56]. In the absence of these epoxide hydrolases, spontaneous cyclisation products with tetrahydrofuran (THF) rings are isolated from the epoxide intermediates (Supplementary Fig. 5).

To date, no epoxide hydrolase has been identified within the *amb* BGC. Hence, it has been speculated that either the epoxidase AmbJ may be multifunctional, or alternatively, the presence of the 8,9-alkene moiety could increase the susceptibility of C-7 of the oxirane to nucleophilic attack[57,58]. Indeed, our recent model studies using truncated surrogates of ambruticin J indicated that both the 5-OH and the 8,9-alkene may influence the intramolecular ring opening, leading to the formation of a THP product[59]. However, details of the biosynthetic mechanism of tetrahydropyran ring formation in the ambruticins remain unknown.

With this in mind, we carefully examined the genomic neighbourhoods of the *ambJ* gene in the BGC to search for a potential epoxide hydrolase. Interestingly, a previously unidentified short region of ca. 500 bp in length immediately downstream of *ambJ* was found (Fig. 3a). Sequence analysis of this region identified two potential overlapping open reading frames (ORFs) in the same orientation, which could be translated into peptide sequences of 136 (ORF1) and 128 (ORF2) amino acids, respectively (Supplementary Fig. 6). BLAST searches of these two sequences suggested that ORF1 might be non-functional as no significant similarities were found, whereas ORF2 appears to encode for a YciI family protein, which we designated as AmbK. Protein structure prediction of AmbK by AlphaFold3[60] indicated a similar $\alpha+\beta$ overall fold to that of Lsd19[54], MupZ[46], and XimE[55], despite low sequence homology (Fig. 3a).

We hypothesised that AmbK might be a candidate for the cryptic epoxide hydrolase implicated in THP formation and so investigated the function of this putative enzyme by first disruption of the corresponding gene in the ambruticin-producing strain *S. cellulosum* So ce10 (Supplementary Fig. 7). Using our developed method, *ambJ* and ambK knock-out mutant strains were successfully generated. As expected, cultures of the resultant Δ*ambJ* mutant produced ambruticin J (**1**, *m/z* 457 $[M\text{-}H]^-$) as the major metabolite (Fig. 3bii)[31]. In contrast, cultures of the Δ*ambK* mutant strain contained none of the usual metabolites (e.g. ambruticin S and the VS series of compounds) found in the wild-type strain (Fig. 1b) but instead yielded a compound with a mass of 492 Da (**4**, *m/z* 491 $[M\text{-}H]^-$) as the major metabolite (Fig. 3biii). ^{1}H and ^{13}C NMR data analysis of compound **4**, compared with those of

ambruticin J (**1**) and F (**3**), showed that a DHP ring and 3,5-dihydroxy acid were present but there were no signals which could be attributed to the cyclopropyl group (Supplementary Fig. 8). Further detailed analysis of its 2D NMR data, including ^{1}H-^{1}H COSY, HSQC and HMBC, suggested that compound **4**, named ambruticin K, bears hydroxyl groups at C-6 and C-12, and features a conjugated 7*E*,9*E*-diene (Supplementary Table 4), as opposed to the 6*E*,8*E*-diene present in ambruticin J (Fig. 3c). Acetylation of ambruticin K (**4**) gave tetraacetate **5** and the ^{1}H-NMR spectrum showed with characteristic downfield shifts of the signals assigned to 3-H, 5-H, 6-H and 12-H in accord with the proposed structure (Fig. 3c and Supplementary Fig. 9). The absolute configuration at C-12 was determined to be *R* by Mosher ester analysis of the acetonide derivative of **4** (Supplementary Methods 1 and 2, Supplementary Fig. 10 and Supplementary Table 5)[61]. We propose that ambruticin K is formed via addition of water to C-12 of the proposed intermediate 6,7-epoxy-ambruticin J (**2**), leading to opening of the cyclopropane ring, and concomitant formation 7,9-diene and opening of the 6,7-epoxide to give the 7,9-dien-6-ol **4** (Fig. 3d). Whilst **2** was not seen in crude extracts, it is well established that in many biosynthetic pathways (e.g. lasalocid[53,54] and mupirocin[46]) and synthetic studies[59] involving epoxide forming/epoxide opening cascades that epoxide intermediates are unstable and not detectable being readily transformed to a more stable product.

Isolation of ambruticin K (**4**) from cultures of the Δ*ambK* mutant of *S. cellulosum* indicated that AmbK is the elusive epoxide hydrolase required for THP ring formation in ambruticin biosynthesis. To verify the function of AmbK, in vivo biotransformations were undertaken using *E. coli* cells overexpressing AmbJ and AmbK. Soluble expression of both proteins was achieved after condition optimisations (Supplementary Fig. 11a). First, ambruticin J (**1**) was incubated with AmbJ-expressing *E. coli* cells, which gave the expected ambruticin K (**4**) (Supplementary Fig. 11b, ii). Pleasingly, incubation of ambruticin J with *E. coli* cells co-expressed with both AmbJ and AmbK gave ambruticin F (**3**) with the characteristic THP ring as confirmed by NMR (Supplementary Fig. 11biii). Furthermore, ambruticin K (**4**) with the co-expressed AmbJ and AmbK simply returned the starting material **4** (Supplementary Fig. 11biv) in accord with the proposal that ambruticin K is not an intermediate in ambruticin biosynthesis but arises through addition of water to the proposed intermediate 6,7-epoxy-ambruticin J (**2**). Overall, these studies confirm that AmbJ catalyses selective epoxidation of ambruticin J (**1**) to produce 6,7-epoxy-ambruticin J (**2**), followed by the epoxide hydrolase AmbK catalysing regioselective 6-membered ring formation to generate the THP ring in ambruticin biosynthesis (Fig. 3d).

AmbH is an iterative module for polyketide backbone assembly

Our next goal was to gain a deeper understanding of assembly of the polyketide backbone. The chain assembly has been proposed to initially follow the canonical type I PKS biosynthetic logic by the loading module and elongation modules 1–6 in AmbA-AmbE (Supplementary Fig. 1)[31]. Module 7 in AmbE then performs three iterative chain elongations, followed by further downstream processes involving AmbF, AmbG, and AmbI to generate the cyclopropane ring-containing backbone. It was postulated that this backbone then undergoes another round of elongation by the terminal PKS module AmbH, and is subsequently hydrolysed by its thioesterase (TE) to release ambruticin J[31]. For many years, these steps lacked detailed, direct experimental evidence to support the proposed assembly process in ambruticin biosynthesis. However, important insights were revealed from studies by Hahn and co-workers using extensive bioinformatic analysis and in vitro enzymatic assays[37]. They reannotated AmbG as a fatty acyl:adenylate ligase (FAAL)–acyl carrier protein (ACP) didomain and showed that it adenylates a wide range of free carboxylic acids with broad structural diversity (Fig. 4a).

Inspired by Hahn's work, we hypothesised that inactivating AmbG may facilitate capture of cryptic intermediates during polyketide chain assembly. With this in mind, the FAAL domain of AmbG was disrupted and the Δ*ambG* mutant strain obtained (Supplementary Fig. 7c). A compound with a mass of 370 Da was isolated from cultures of the Δ*ambG* mutant strain *of S. cellulosum* (Fig. 4b). Structural elucidation by HRMS and extensive NMR analysis confirmed that this product included an unsaturated carboxylic acid with intact cyclopropane and dihydropyran rings. Signals were apparent for 4 double bonds in the chain, including a 2*E*,4*E*-conjugated diene (Supplementary Table 6). The compound, named ambruticin G (**6**) (Fig. 4c), is shorter than the mature chain of ambruticin J, effectively lacking two acetate units in accord with its potential role as a biosynthetic precursor to ambruticin J (Supplementary Fig. 12). Since AmbH is presumed to be the sole PKS module acting after AmbG (Supplementary Fig. 1), we hypothesised that AmbH could extend ambruticin G by two rounds of chain elongation to afford ambruticin J. This hypothesis is supported by the KS-AT-KR-ACP-TE domain architecture of the AmbH module, where recursive chain extension and ketoreductase-catalysed reduction would be expected to yield the C-3 and C-5 hydroxyl groups found in ambruticin J[62].

To test our hypothesis, the entire continuous region of the *amb* BGC upstream of *ambG* (*ambA-ambI*) was knocked out from the Δ*ambJ* mutant to give the Δ*ambA-I*/Δ*ambJ* multi-domain mutant (Supplementary Fig. 7d). As expected, cultures of this strain produced no ambruticin-related metabolites as the main PKS modules were absent. However, AmbG and AmbH remained intact and functional in this strain. Subsequent feeding studies of ambruticin G (**6**) into this Δ*ambA-I*/Δ*ambJ* mutant were conducted, and full conversion to ambruticin J (**1**) was observed after 3 hours of incubation (Fig. 4d). In a parallel experiment, the entire region of *ambA-ambI* was deleted from the *wild-type* strain. Feeding ambruticin G into the resulting Δ*ambA-I* mutant showed conversion to ambruticin S and the VS series compounds, as is found in the *wild-type* strain (Fig. 4e). These results clearly reveal that ambruticin G is a biosynthetic intermediate and highlight AmbH as an iterative module catalysing dual chain elongation of ambruticin G to afford ambruticin J (Fig. 4f).

An updated pathway for ambruticin biosynthesis

Reeves and co-workers proposed that starting from AmbF in ambruticin biosynthesis, unusual steps occur, including a Favorskii rearrangement and a carbon atom excision event, which would deliver a cyclopropyl polyketide intermediate to the terminal PKS module AmbH for a final round of chain elongation, followed by chain release to afford ambruticin J for further post-PKS modifications (Fig. 5a)[31]. Collectively, the results described herein have revealed the critical roles of the otherwise cryptic AmbH and AmbK, providing significant insights into both the assembly of the polyketide backbone and the post-PKS tailoring processes involved in ambruticin biosynthesis in *S. cellulosum* So ce10. These findings have led us to propose a revised biosynthetic pathway for ambruticin (Fig. 5b).

With identification of ambruticin G (**6**) as a biosynthetic intermediate, it is now evident that the terminal module AmbH is iterative and catalyses two rounds of chain elongation (Fig. 5b). Subsequent THP ring formation from ambruticin J (**1**) requires AmbJ, which functions as the epoxidase, and in this work, an additional epoxide hydrolase, AmbK, has been identified. In the absence of AmbK, hydrolysis of epoxide **2** occurs giving ambruticin K (**4**), which is not an intermediate in ambruticin biosynthesis. Ambruticin F (**3**) contains all 3 rings (THP, DHP, and trisubstituted cyclopropane) characteristic of the ambruticins, and further tailoring modifications at C-5 occur to yield the final metabolites (Fig. 5b).

A further important feature to note is that the structure of ambruticin G (**6**) incorporates the DHP ring characteristic of the

https://doi.org/10.1038/s41467-025-63441-y

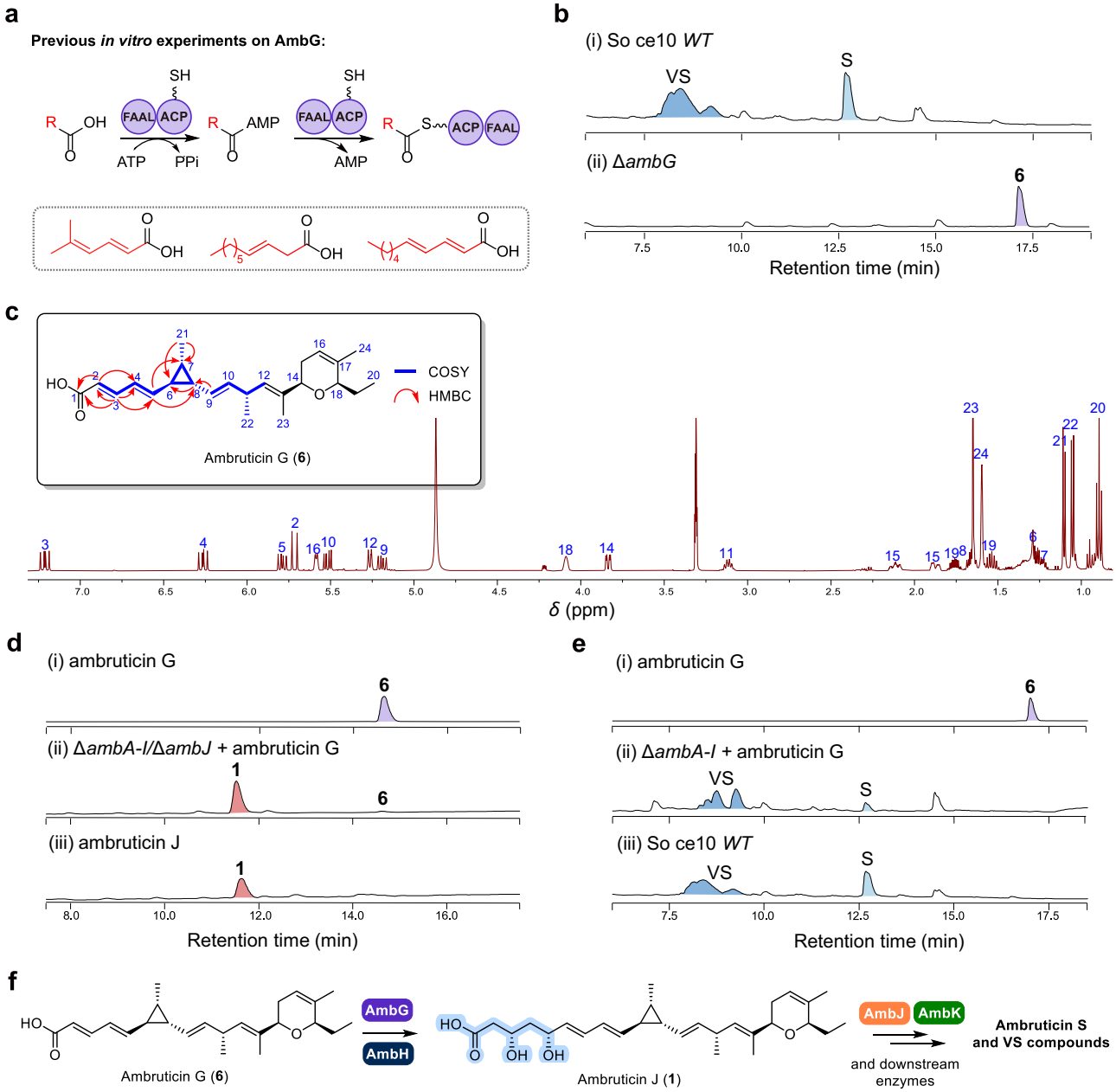


Fig. 4 | AmbH is an iterative module for polyketide backbone assembly in ambruticin biosynthesis. a Previous in vitro studies of AmbG by Hahn and co-workers[37]. **b** HPLC trace of the So ce10 Δ*ambG* mutant. **c** Structural elucidation of ambruticin G (**6**). **d** Feeding study of ambruticin G (**6**) by the Δ*ambA-I*/Δ*ambJ* mutant. **e** Feeding study of ambruticin G (**6**) by the Δ*ambA-I* mutant. **f** The iterative module AmbH converts ambruticin G to ambruticin J, which is further metabolised by AmbJ and AmbK along with downstream enzymes to the final ambruticin VS compounds.

ambruticins. This result is in accord with our recent studies, which indicate that desaturation of the THP to a DHP catalysed by the Rieske oxygenase AmbP occurs prior to formation of the C3-C7 THP ring[40]. The results described herein indicate that the timing of the desaturation step can be as early as the PKS assembly stage.

The mechanism of formation of the cyclopropane ring moiety in ambruticin biosynthesis remains unknown. However, since the truncated metabolite ambruticin G (**6**) contains an intact cyclopropane ring, it is now apparent that the cyclopropane ring must be formed prior to AmbG channelling the free carboxylic acid intermediate back onto the assembly-line for additional PKS processing. The hydrolase domain (H) and pyridoxal cofactor binding motif domain (Px) in the AmbF module, along with the flavin-dependent monooxygenase AmbI, are most likely to play pivotal roles in this process (Fig. 5b). Studies are currently underway in our laboratories to investigate these fascinating biotransformations.

Discussion

The studies reported herein have led to the development of an efficient electroporation method for transforming foreign DNA into various *Sorangium cellulosum* strains, enabling effective genetic engineering via double-crossover homologous recombination within a single step in these genetically intractable myxobacteria. By using different antibiotic selection markers, we also achieved sequential deletions in these organisms. Nevertheless, a potential limitation of our study is the current unavailability of a broader panel of *Sorangium* strains to comprehensively assess the generalisability of our electroporation approach. While our method demonstrates high efficiency in

https://doi.org/10.1038/s41467-025-63441-y

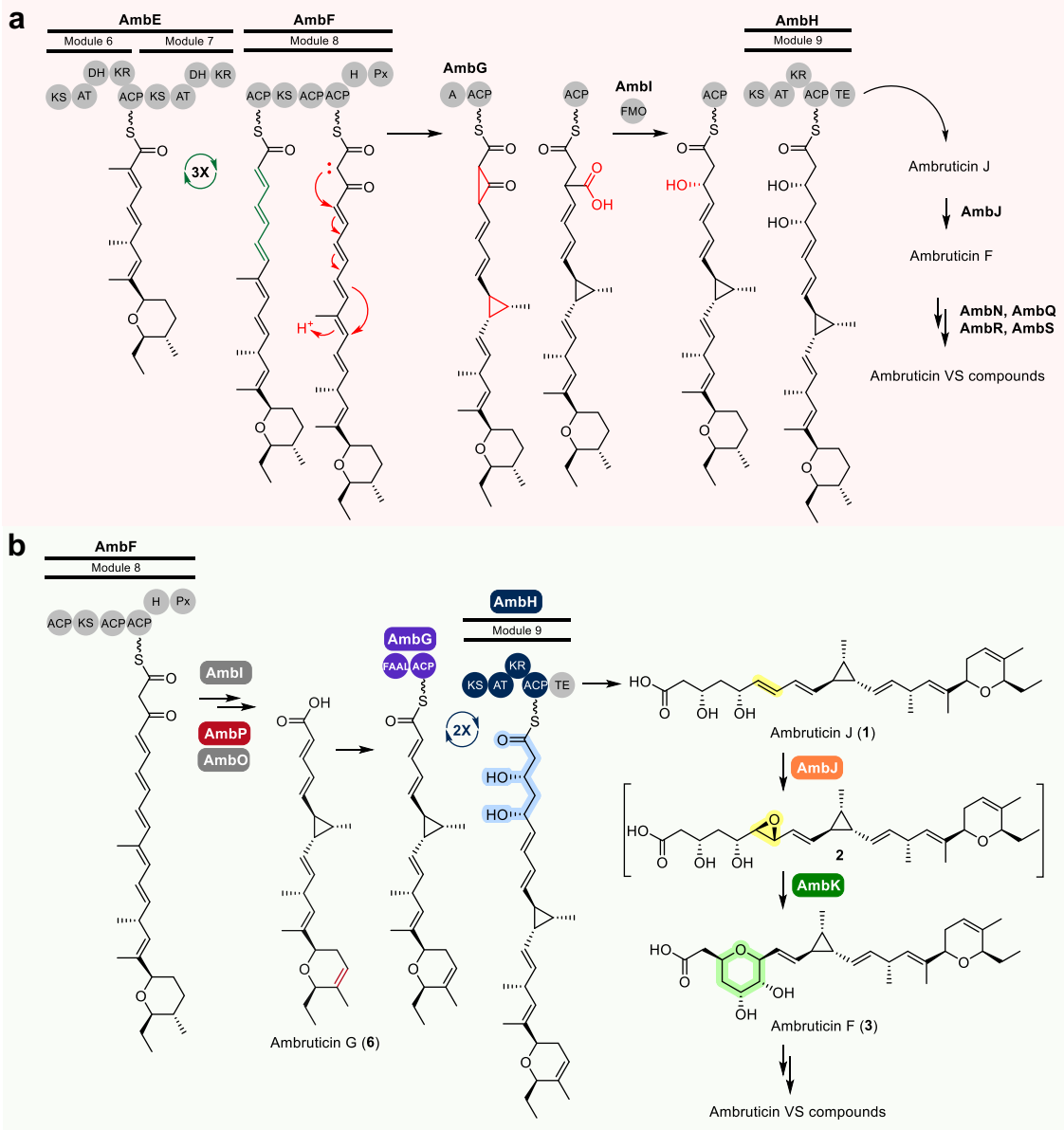


Fig. 5 | Genetic engineering of *S. cellulosum* So ce10 leads to an updated pathway for ambruticin biosynthesis. a Previously proposed biosynthetic pathway of ambruticin[31]. **b** Revised biosynthetic pathway of ambruticin based on this study.

Sorangium strains that grow in suspension, its applicability to strains exhibiting clumped or aggregated growth remains to be fully evaluated. Our preliminary experiments with the recently isolated icumazole-producing strain *Sorangium cellulosum* MBD-1, which displays dense clumping during growth, suggest that transformation is still possible, albeit with reduced efficiency. These observations underscore the influence of morphological characteristics in transformation outcomes and highlight the need for further strain-specific optimisation to extend the applicability of this method to a wider range of *Sorangium* isolates.

This method has been applied to genetic engineering in *S. cellulosum* So ce10, leading to a greater understanding of the biosynthesis of the antifungal polyketides, the ambruticins. Through a combination of targeted gene knockouts, whole-cell biotransformation, and feeding experiments, we identified a truncated metabolite, ambruticin G (**6**), and revealed the role of the iterative PKS module AmbH in polyketide backbone assembly of ambruticin biosynthesis. Furthermore, a previously unknown epoxide hydrolase, AmbK, has been identified and shown to be crucial for THP ring formation during post-PKS modification. Building on these findings, we have proposed an updated pathway for ambruticin biosynthesis (Fig. 5b), which lays the foundation for further investigations aimed at elucidating mechanistic details of its unique enzymology, particularly the cyclopropyl ring formation step.

Despite myxobacteria being renowned for producing a wealth of structurally diverse and biologically active natural products, we are still only beginning to uncover the full extent of their capabilities[6]. In future, using the tools to genetically manipulate *Sorangium* strains reported herein will significantly contribute to unlocking the vast and largely untapped biosynthetic potential of myxobacteria. Such efforts hold immense promise for the creation of novel analogues with altered chemical structures, potentially leading to compounds with enhanced or entirely new biological activities, such as improved antimicrobial, anticancer, or antiviral properties. As we refine these tools and gain deeper insights into the complex enzymology of these biosynthetic pathways, we may also uncover entirely new classes of natural products, further expanding the chemical and biological diversity accessible from myxobacteria.

Methods

General experimental

Strains of *Sorangium cellulosum* were purchased from the German Collection of Microorganisms and Cell Cultures (DSMZ) and maintained on HS agar (0.4% glucose, 0.15% Casein peptone, 0.1% KNO_3, 0.1% $MgSO_4{\cdot}7H_2O$, 0.008% NaFe-EDTA, 0.2% HEPES, 0.00625% K_2HPO_4, 0.0075% $CaCl_2{\cdot}2H_2O$, 1.5% agar, pH 7.2). Oligonucleotides for polymerase chain reactions (PCR) were purchased from Tsingke (China). PCR amplification was performed using Phanta Flash Super-Fidelity DNA Polymerase (Vazyme, China). Plasmids were purified using a HighPure Rapid Mini Plasmid Kit (Biomed, China), and PCR products were purified with the HiPure Gel Pure DNA Mini Kit (Magen, China). Restriction enzymes were purchased from New England Biolabs (Ipswich, MA, USA). DNA ligation was performed using ClonExpress Ultra One Step Cloning Kit (Vazyme, China). Molecular biology experiments were conducted following the manufacturer's instructions. Strains, plasmids, and oligonucleotides used in this study are listed in Supplementary Tables 7 and 8 and Supplementary Data 1.

LC-MS data were obtained on a SHIMADZU LC-MS system comprising a SIL-20AXR autosampler, LC-20ADXR liquid chromatograph, SPD-M40 photodiode array detector, and SHIMADZU LCMS-2010 EV liquid chromatography-mass spectrometry. HPLC-grade H_2O and MeCN were added with 0.1% formic acid as the solvent system. Analytical LC-MS data were obtained using a Phenomenex Kinetex column (C18, 250 × 4.6 mm, 5 μm) at a flow rate of 1 mL min^{-1}. Column chromatographic separations were carried out using Sephadex LH-20 (Cytiva) as packing materials. Preparative HPLC purification was carried out using a SilGreen CH0525010C 18AQ column (C_{18}, 250 × 10.0 mm, 5 μm) at a flow rate of 4.7 mL min^{-1}. High-resolution electrospray ionisation mass spectrometry (HR-ESI-MS) was performed on a Waters ACQUITYI-Class PLUS/Xevo G2-XS Tof instrument. ^1H, ^{13}C NMR and 2D NMR data were collected on a Bruker Avance III 500 MHz or Bruker Ascend 800 MHz NMR spectrometer.

Genetic engineering of *S. cellulosum*

The upstream and downstream fragments of the target gene or region were amplified from genomic DNA via PCR. Knockout plasmids were constructed by inserting the hygromycin, tetracycline, or chloramphenicol selection marker, flanked by the upstream and downstream homology arms, into the KpnI/HindIII-linearised pXZ vector. These plasmids were introduced into *Sorangium* cells via electroporation, and double-crossover transformants were selected on HS agar supplemented with appropriate concentrations of hygromycin, tetracycline, or chloramphenicol. Sequences of the antibiotic resistance genes and the working concentrations of antibiotics are shown in Supplementary Data 2.

For electrocompetent cells preparation, a 1% (v/v) inoculum from mid-exponential phase preculture of *S. cellulosum* strain was inoculated into 100 mL HS broth, cultivated in shake flasks at 30 °C, 200 rpm. Upon reaching an OD_{600} of 0.5 (approximately 30 hours after inoculation), cells were harvested by centrifugation at 4500 × g. The pellet was then washed three times with 10 mL 10% (v/v) glycerol at 15,000 × g. The cell pellet was then resuspended in 10% glycerol to a final volume of 900 μL and aliquoted into 90 μL/ EP tubes for storage at −80 °C.

For electroporation, the electrocompetent cells were thawed on ice, and 500 ng–5 μg of plasmid was mixed with 90 μL of competent cells. The mixture was then transferred into a precooled 1 mm cuvette, and subjected to electroporation using a Bio-Rad MicroPulser, at a voltage of 15 kV/cm for 3 times. Immediately after pulsing, 500 μL of precooled liquid HS medium was added, followed by transferring into 3 mL of liquid HS medium. Cells were allowed to recover at 30 °C and 200 rpm for 2 days. Subsequently, cells were collected by centrifugation and spread on HS agar plates supplemented with appropriate antibiotics. Transformants typically appear after incubation at 30 °C for 5–7 days. Knockout mutants were screened by PCR and further confirmed by sequencing.

For conjugative transformation of plasmid pXZ-crtBKO into *Sorangium cellulosum*, the plasmid was first introduced into chemically competent *E. coli* S17.1 cells via heat shock. The resulting *E. coli* donor strain carrying pXZ-crtBKO was cultured overnight in LB medium containing kanamycin (50 μg/mL) at 37 °C. A 1% v/v inoculum was then transferred to 50 mL of fresh kanamycin-containing LB medium and grown to early exponential phase ($OD_{600} \approx 0.4–0.6$), while the recipient *Sorangium* cultures were grown in 50 mL HS medium at 30 °C until reaching an OD_{600} of 0.5. Both cell types were harvested by centrifugation and washed with LB and S42 medium[31], respectively. The *Sorangium* cells were resuspended and subjected to a 10-min heat shock at 50 °C in pre-warmed S42 medium. Equal volumes (typically 1 mL each) of donor and recipient were mixed, centrifuged, and resuspended in 200 μL of glucose-free S42 medium. The cell mixture was spotted onto an antibiotic-free S42 agar plate and incubated at 30 °C for 48 h. The resulting mating puddle was resuspended in 0.5 mL S42 liquid medium and plated onto S42 agar supplemented with appropriate concentrations of hygromycin and carbenicillin. Plates were incubated at 30 °C for 7–10 days until *Sorangium* transformant colonies emerge.

General fermentation procedure

Wild-type or mutant strains of *S. cellulosum* So ce10 were inoculated on HS agar plates and incubated at 30 °C for 3 days. Colonies from these plates were scraped and used to inoculate 100 mL of liquid HS medium in a 500 mL flask. This seed culture was incubated at 30 °C, 200 rpm for 2–3 days. For production fermentation, 20% of the seed culture was used to inoculate SF1-P medium (0.3% soy peptone, 0.6% fructose, 0.1% $MgSO_4{\cdot}7H_2O$, 0.1% $CaCl_2{\cdot}2H_2O$, 0.008% ferric citrate, and 0.05 M HEPES, pH 7.6; ferric citrate and HEPES were filter sterilised and added after autoclaving). The culture was supplemented with 1 g L^{-1} of fructose every 24 h, starting on the second day, and incubated at 30 °C, 200 rpm for 7 days. The supernatant was extracted three times with ethyl acetate (EtOAc). The combined EtOAc extracts were evaporated *in vacuo* to yield a crude extract, which was then analysed by LC-MS or subjected to further purification.

Isolation and purification of ambruticin J, ambruticin K and ambruticin G

A 2.0 L scale fermentation of the Δ*ambJ* mutant of *S. cellulosum* So ce10 was performed following the general fermentation procedure. The crude extract was initially purified using Sephadex LH-20 column chromatography with MeOH as the eluent to give a crude fraction of ambruticin J. This fraction was further purified by HPLC eluting with a gradient elution of 40% to 95% acetonitrile in water over 20 min, with ambruticin J eluting at ~85% acetonitrile, yielding 1.6 mg of the compound. Similarly, ambruticin K (eluting at ~70% acetonitrile) was isolated from the Δ*ambK* mutant of *S. cellulosum* So ce10, following the same purification protocol, with a yield of 1.4 mg L^{-1}. Ambruticin G (eluting at approximately 90% acetonitrile) was obtained from the Δ*ambG* mutant strain with a yield of 1 mg L^{-1}. For HR-ESI-MS and NMR data of the compounds, see in Supplementary Tables 4 and 6 and Supplementary Figs. 13–38.

Chemical derivatisation of ambruticin K

For acetylation of ambruticin K, **4** was dissolved in 25 μL of pyridine and mixed with 25 μL of acetic anhydride. The reaction was stirred at room temperature overnight and then quenched with water. The reaction mixture was extracted with EtOAc, and the organic layer was dried under vacuum to obtain a crude extract. This extract was then purified using HPLC with a gradient elution of 5% to 95% MeCN in water over 20 min to yield compound **5**.

For preparation of the acetonides derivatives of ambruticin K and their Mosher esters, see Supplementary Methods 1 and 2.

Feeding studies of ambruticin G

A 50 mL culture of the So ce10 Δ*ambA-I* or So ce10 Δ*ambA-I*/Δ*ambJ* strain in SF1-P medium was incubated at 30 °C for 48 h. Subsequently, 0.1 mg of ambruticin G dissolved in 10 µL of MeOH was added to each culture. For the So ce10 Δ*ambA-I* strain, incubation continued at 30 °C for an additional 30 min, while for the So ce10 Δ*ambA-I*/Δ*ambJ* strain, 0.1 mg of ambruticin G was added every 30 min, with a total incubation period of 6 h. Three volumes of EtOAc were used for extraction, and the solvent was removed under vacuum. The dried samples were redissolved in 200 µL MeOH and analysed by LC-MS.

Expression of AmbJ and AmbK in *E. coli*

The *ambJ* and *ambK* genes were amplified by PCR from the genomic DNA of *S. cellulosum* So ce10 using primers listed in Supplementary Data 1. The resulting PCR products were introduced into pET151 and pET28a vectors, respectively. For AmbJ expression, the pET151-ambJ construct and pGro7 (expressing GroES-GroEL) were co-transformed into *E. coli* BL21(DE3) cells. The cells were grown overnight in 5 mL of lysogeny broth (LB) medium supplemented with 50 µg mL^{-1} carbenicillin and 25 µg mL^{-1} chloramphenicol at 37 °C with shaking at 200 rpm. 1 mL of this pre-culture was used to inoculate 100 mL of LB medium containing 50 µg mL^{-1} carbenicillin, 25 µg mL^{-1} chloramphenicol, and 0.5 g L^{-1} L-arabinose. The culture was incubated at 37 °C with shaking at 200 rpm until the OD_{600} reached 0.5-0.6. Protein expression was induced by adding 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and the cells were incubated overnight at 16 °C with shaking at 120 rpm. For AmbK expression, the pET28a-ambK construct was transformed into *E. coli* BL21(DE3). Protein expression followed similar procedure with LB medium supplemented with 50 µg mL^{-1} kanamycin. For co-expression of AmbJ and AmbK, the same procedure used for AmbJ expression was followed, but with the LB medium supplemented with 50 µg mL^{-1} carbenicillin, 50 µg mL^{-1} kanamycin, and 25 µg mL^{-1} chloramphenicol.

Whole-cell biotransformation

An overnight culture (50 mL) of *E. coli* BL21 (DE3) cells overexpressing AmbJ or both AmbJ and AmbK was grown in LB medium. The cells were centrifuged at 4 °C and 2400 × *g* for 5 min, and the pellets were resuspended in 2 mL of 100 mM potassium phosphate buffer (pH 7.2) containing 20 mM glucose. Subsequently, 0.5 mg of ambruticin J or ambruticin K dissolved in 50 µL of MeOH was added. The mixtures were incubated at 30 °C and 200 rpm for 12 h. Reactions were quenched by adding three volumes of ethyl acetate, followed by vortexed and centrifugation. The ethyl acetate layer was dried under vacuum, dissolved in 400 µL methanol, and subjected to LC-MS analysis.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Whole genome sequences of *Sorangium cellulosum* So ce10 and *Sorangium cellulosum* So ce307 have been deposited at NCBI's GenBank under accession number CP162580 and CP162579, respectively. Other data supporting the findings of this study are available in the main text, supplementary information and from corresponding author(s) upon request. Source data are provided with this paper.

References

1. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).
2. Atanasov, A. G. et al. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
3. Hemmerling, F. & Piel, J. Strategies to access biosynthetic novelty in bacterial genomes for drug discovery. *Nat. Rev. Drug Discov.* **21**, 359–378 (2022).
4. Weissman, K. J. & Müller, R. Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.* **27**, 1276 (2010).
5. Saggu, S. K., Nath, A. & Kumar, S. Myxobacteria: biology and bioactive secondary metabolites. *Res. Microbiol.* **174**, 104079 (2023).
6. Wang, C. et al. Recent advances in discovery and biosynthesis of natural products from myxobacteria: an overview from 2017 to 2023. *Nat. Prod. Rep.* **41**, 905–934 (2024).
7. Gerth, K., Perlova, O. & Müller, R. *Sorangium cellulosum*. *Myxobacteria: Multicellularity and Differentiation* (ed. Whitworth, D. E.) 329–348 (ASM Press, 2014).
8. Höfle, G. et al. Epothilone A and B-novel 16-membered macrolides with cytotoxic activity: isolation, crystal structure, and conformation in solution. *Angew. Chem. Int. Ed.* **35**, 1567–1569 (1996).
9. Naini, A., Sasse, F. & Brönstrup, M. The intriguing chemistry and biology of soraphens. *Nat. Prod. Rep.* **36**, 1394–1411 (2019).
10. Lilic, M. et al. The antibiotic sorangicin A inhibits promoter DNA unwinding in a *Mycobacterium tuberculosis* rifampicin-resistant RNA polymerase. *Proc. Natl. Acad. Sci. USA* **117**, 30423–30432 (2020).
11. Fleta-Soriano, E. et al. The myxobacterial metabolite ratjadone A inhibits HIV infection by blocking the Rev/CRM1-mediated nuclear export pathway. *Microb. Cell. Fact.* **13**, 17 (2014).
12. Schneiker, S. et al. Complete genome sequence of the myxobacterium *Sorangium cellulosum*. *Nat. Biotechnol.* **25**, 1281–1289 (2007).
13. Han, K. et al. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. *Sci. Rep.* **3**, 2101 (2013).
14. Wenzel, S. C. & Müller, R. Myxobacterial natural product assembly lines: fascinating examples of curious biochemistry. *Nat. Prod. Rep.* **24**, 1211 (2007).
15. Julien, B. & Rodriguez, E. Genetic engineering of myxobacterial natural product biosynthetic genes, *Manual of Industrial Microbiology and Biotechnology* (eds. Baltz, R. H. et al.) 426–437 (ASM Press, 2014).
16. Yue, X., Sheng, D., Zhuo, L. & Li, Y. Genetic manipulation and tools in myxobacteria for the exploitation of secondary metabolism. *Eng. Microbiol.* **3**, 100075 (2023).
17. Whitworth, D. E., Sydney, N. & Radford, E. J. Myxobacterial genomics and post-genomics: a review of genome biology, genome sequences and related 'omics studies. *Microorganisms* **9**, 2143 (2021).
18. Jaoua, S., Neff, S. & Schupp, T. Transfer of mobilizable plasmids to *Sorangium cellosum* and evidence for their integration into the chromosome. *Plasmid* **28**, 157–165 (1992).
19. Pradella, S. et al. Characterisation, genome size and genetic manipulation of the myxobacterium *Sorangium cellulosum* So ce56. *Arch. Microbiol.* **178**, 484–492 (2002).
20. Kopp, M. et al. Critical variations of conjugational DNA transfer into secondary metabolite multiproducing *Sorangium cellulosum* strains So ce12 and So ce56: development of a mariner-based transposon mutagenesis system. *J. Biotechnol.* **107**, 29–40 (2004).
21. Kopp, M., Irschik, H., Pradella, S. & Müller, R. Production of the tubulin destabilizer disorazol in *Sorangium cellulosum*: biosynthetic machinery and regulatory genes. *ChemBioChem* **6**, 1277–1286 (2005).
22. Ringel, S. M. et al. Ambruticin (W7783), a new antifungal antibiotic. *J. Antibiot.* **30**, 371–375 (1977).
23. Connor, D. T., Greenough, R. C. & Von Strandtmann, M. W-7783, a unique antifungal antibiotic. *J. Org. Chem.* **42**, 3664–3669 (1977).

24. Knauth, P. & Reichenbach, H. On the mechanism of action of the myxobacterial fungicide ambruticin. *J. Antibiot.* **53**, 1182–1190 (2000).
25. Shubitz, L. F. et al. Efficacy of ambruticin analogs in a murine model of coccidioidomycosis. *Antimicrob. Agents Chemother.* **50**, 3467–3469 (2006).
26. Vetcher, L. et al. The antifungal polyketide ambruticin targets the HOG pathway. *Antimicrob. Agents Chemother.* **51**, 3734–3736 (2007).
27. Masschelein, J., Jenner, M. & Challis, G. L. Antibiotics from Gram-negative bacteria: a comprehensive overview and selected biosynthetic highlights. *Nat. Prod. Rep.* **34**, 712–783 (2017).
28. Marcos-Torres, F. J., Volz, C. & Müller, R. An ambruticin-sensing complex modulates *Myxococcus xanthus* development and mediates myxobacterial interspecies communication. *Nat. Commun.* **11**, 5563 (2020).
29. Hahn, F. & Guth, F. M. The ambruticins and jerangolids – chemistry, biology and chemoenzymatic synthesis of potent antifungal drug candidates. *Nat. Prod. Rep.* **37**, 1300–1315 (2020).
30. Ma, S., Mandalapu, D., Wang, S. & Zhang, Q. Biosynthesis of cyclopropane in natural products. *Nat. Prod. Rep.* **39**, 926–945 (2022).
31. Julien, B., Tian, Z., Reid, R. & Reeves, C. D. Analysis of the ambruticin and jerangolid gene clusters of *Sorangium cellulosum* reveals unusual mechanisms of polyketide biosynthesis. *Chem. Biol.* **13**, 1277–1286 (2006).
32. Berkhan, G. & Hahn, F. A dehydratase domain in ambruticin biosynthesis displays additional activity as a pyran-forming cyclase. *Angew. Chem. Int. Ed.* **53**, 14240–14244 (2014).
33. Friedrich, S. et al. Characterisation of the broadly-specific O-methyl-transferase JerF from the late stages of jerangolid biosynthesis. *Molecules* **21**, 1443 (2016).
34. Berkhan, G., Merten, C., Holec, C. & Hahn, F. The interplay between a multifunctional dehydratase domain and a C-methyltransferase effects olefin shift in ambruticin biosynthesis. *Angew. Chem. Int. Ed.* **55**, 13589–13592 (2016).
35. Sung, K. H. et al. Insights into the dual activity of a bifunctional dehydratase-cyclase domain. *Angew. Chem. Int. Ed.* **57**, 343–347 (2018).
36. Lindner, F., Friedrich, S. & Hahn, F. Total synthesis of complex biosynthetic late-stage intermediates and bioconversion by a tailoring enzyme from jerangolid biosynthesis. *J. Org. Chem.* **83**, 14091–14101 (2018).
37. Hemmerling, F., Lebe, K. E., Wunderlich, J. & Hahn, F. An unusual fatty acyl:adenylate ligase (FAAL)-acyl carrier protein (ACP) didomain in ambruticin biosynthesis. *ChemBioChem* **19**, 1006–1011 (2018).
38. Guth, F. M. et al. Rieske oxygenase-catalyzed oxidative late-stage functionalization during complex antifungal polyketide biosynthesis. *ACS Chem. Biol.* **18**, 2450–2456 (2023).
39. Hollmann, T. et al. Biocatalysts from biosynthetic pathways: enabling stereoselective, enzymatic cycloether formation on a gram scale. *ACS Catal.* **10**, 4973–4982 (2020).
40. Bowen, J. I. et al. Combining total synthesis and genetic engineering to probe dihydropyran formation in ambruticin biosynthesis. *Chem. Sci.* **15**, 5319–5326 (2024).
41. Liang, M., Zhu, J. & Jiang, J. Carotenoids biosynthesis and cleavage related genes from bacteria to plants. *Crit. Rev. Food Sci. Nutr.* **58**, 2314–2333 (2018).
42. Du, K. et al. Reprogramming the endogenous type I CRISPR-Cas system for simultaneous gene regulation and editing in *Haloarcula hispanica*. *mLife* **1**, 40–50 (2022).
43. Wang, J., Wei, J., Li, H. & Li, Y. High-efficiency genome editing of an extreme thermophile *Thermus thermophilus* using endogenous type I and type III CRISPR-Cas systems. *mLife* **1**, 412–427 (2022).
44. Wu, Y. et al. Dissecting the arginine and lysine biosynthetic pathways and their relationship in haloarchaeon *Natrinema gari* J7-2 via endogenous CRISPR-Cas system-based genome editing. *Microbiol. Spectr.* **11**, e00288–23 (2023).
45. Hmelo, L. R. et al. Precision-engineering the *Pseudomonas aeruginosa* genome with two-step allelic exchange. *Nat. Protoc.* **10**, 1820–1841 (2015).
46. Wang, L. et al. A Rieske oxygenase/epoxide hydrolase-catalysed reaction cascade creates oxygen heterocycles in mupirocin biosynthesis. *Nat. Catal.* **1**, 968–976 (2018).
47. Tu, Q. et al. Room temperature electrocompetent bacterial cells improve DNA transformation and recombineering efficiency. *Sci. Rep.* **6**, 24648 (2016).
48. Bierman, M. et al. Plasmid cloning vectors for the conjugal transfer of DNA from *Escherichia coli* to *Streptomyces* spp. *Gene* **116**, 43–49 (1992).
49. Sun, Y. et al. Organization of the biosynthetic gene cluster in *Streptomyces* sp. DSM 4137 for the novel neuroprotectant polyketide meridamycin. *Microbiol* **152**, 3507–3515 (2006).
50. Gemperlein, K. et al. Polyunsaturated fatty acid biosynthesis in myxobacteria: different PUFA synthases and their product diversity. *Chem. Sci.* **5**, 1733 (2014).
51. Boldt, J. et al. Bursts in biosynthetic gene cluster transcription are accompanied by surges of natural compound production in the myxobacterium *Sorangium* sp. *Microbiol. Biotechnol.* **16**, 1054–1068 (2023).
52. Xie, F. et al. Insights into the biosynthesis of icumazole, unveiling a distinctive family of crotonyl-CoA carboxylase/reductase. *Cell Rep. Phys. Sci.* **4**, 101394 (2023).
53. Shichijo, Y. et al. Epoxide hydrolase Lsd19 for polyether formation in the biosynthesis of lasalocid A: direct experimental evidence on polyene-polyepoxide hypothesis in polyether biosynthesis. *J. Am. Chem. Soc.* **130**, 12230–12231 (2008).
54. Hotta, K. et al. Enzymatic catalysis of anti-Baldwin ring closure in polyether biosynthesis. *Nature* **483**, 355–358 (2012).
55. He, B. et al. Enzymatic pyran formation involved in xiamenmycin biosynthesis. *ACS Catal.* **9**, 5391–5399 (2019).
56. Bowen, J. I., Wang, L., Crump, M. P. & Willis, C. L. Synthetic and biosynthetic methods for selective cyclisations of 4, 5-epoxy alcohols to tetrahydropyrans. *Org. Biomol. Chem.* **20**, 1150–1175 (2022).
57. Hemmerling, F. & Hahn, F. Biosynthesis of oxygen and nitrogen-containing heterocycles in polyketides. *Beilstein J. Org. Chem.* **12**, 1512–1550 (2016).
58. Trentadue, K., Chang, C. F., Nalin, A. & Taylor, R. E. Enantioselective total synthesis of the putative biosynthetic intermediate ambruticin. *J. Chem. Eur. J.* **27**, 11126–11131 (2021).
59. Bowen, J. I., Wang, L., Crump, M. P. & Willis, C. L. Ambruticins: tetrahydropyran ring formation and total synthesis. *Org. Biomol. Chem.* **19**, 6210–6215 (2021).
60. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature* **630**, 493–500 (2024).
61. Hoye, T. R., Jeffrey, C. S. & Shao, F. Mosher ester analysis for the determination of absolute configuration of stereogenic (chiral) carbinol carbons. *Nat. Protoc.* **2**, 2451–2458 (2007).
62. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed.* **48**, 4688–4716 (2009).

Acknowledgements

This work was supported by the National Key Research and Development Programme of China (2024YFC3407000 to L.W.), the National Natural Science Foundation of China (32270083 to L.W.), the Strategic Priority Research Programme of the Chinese Academy of Sciences (XDB0810000 to L.W.), and the Beijing Municipal Science & Technology Project, China (Z241100007724009 to L.W.). We are grateful to the

EPSRC Bristol Chemical Synthesis Centre for Doctoral Training (EP/L015366/1) for a studentship to J.I.B.

Author contributions

L.W. and C.L.W. conceptualised the project. X.Z., S.L., B.M., K.G., D.J., Y.H., H.C. and J.L. performed experiments. X.Z., S.L., J.I.B., M.P.C., C.L.W. and L.W. analysed data. X.Z., S.L., M.P.C., C.L.W. and L.W. wrote the manuscript with contributions from all the authors. All authors discussed and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-63441-y.

Correspondence and requests for materials should be addressed to Christine L. Willis or Luoyi Wang.

Peer review information *Nature Communications* thanks Rolf Müller, Kenji Watanabe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025