邹 暢

上海交通大学 | 人工智能学院 | 计算机科学与技术

基本信息

男 21岁 电话/微信：13608211202 邮箱：shenyizou@outlook.com

上海交通大学 人工智能学院 博士生 Ph.D. Student. Advised by Prof. Linfeng Zhang 2026-2031(预估)

研究方向： 精准高效的图像/视频生成 Precise and Efficient AIGC

电子科技大学 英才实验学院 本科生 Undergraduate 2022-2026

- **均分 93.35，GPA 3.96/4.00，专业排名 3/26， (从全校范围选拔后的荣誉专业，全员优先推免)**
- **CET-4: 556, CET-6: 520; 连续多年获得优秀学生奖学金**

实习经历

- **腾讯-青云计划 混元大模型方向 多模态模型部 基础模型中心 实习生** 2025.05 至今

作为青云人才计划实习生开展长期实习，研究**下一代图像与视频生成大模型的算法设计**，并在超大规模集群上开展训练。参与**HunyuanImage3.0，HunyuanVideo1.5** 等基础图像与视频生成模型相关研究，并以**核心贡献成员**身份 (**Core Contributors**) 参与 HunyuanVideo 1.5 技术报告，主要负责基础模型的特征缓存加速与模型蒸馏技术等基础研究。实习期间主导并开展以可学习的特征缓存为主的相关研究，提出了**MeanFlow蒸馏的改进方案**，并结合可学习缓存技术，实现了在蒸馏后模型上的进一步加速，成功达到行业领先水准，相关部分算法改进已被成功应用于多款基础模型，并于本人于实习期间**以第一作者身份**发表相关论文于计算机视觉领域顶会 **CVPR 2026。**

Technical Report (Core Contributor): [arXiv:2511.18870]
"**HunyuanVideo-1.5 Technical Report**" **- Tencent Hunyuan Foundation Model Team**

主要学术成果

* Equal contribution

本人已在多个高水平学术会议/期刊上以**第一作者发表论文4篇，一作在投论文2篇，非一作论文发表10篇。**

1. **[ICLR 2025] "Accelerating Diffusion Transformers with Token-wise Feature Caching."** - **Chang Zou*,** Xuyang Liu*, Ting Liu, Siteng Huang and Linfeng Zhang.
2. **[ICCV 2025] "Features of Diffusion Models are Predictable: Accelerating Diffusion Transformers with Taylor Seers."** - Jiacheng Liu*, **Chang Zou***, Yuanhuiyi Lyu, Junjie Chen and Linfeng Zhang.
3. **[ACM MM 2025] "SpeCa: Accelerating Diffusion Models with Speculative Sampling."** Jiacheng Liu*, - **Chang Zou***,Yuanhuiyi Lyu, Fei Ren, Shaobo Wang, Kaixin Li, and Linfeng Zhang.
4. **[CVPR 2026] "DisCa: Accelerating Video Diffusion Transformers with Distillation-Compatible Learnable Feature Caching." - Chang Zou**, Changlin Li, Yang Li, Patrol Li, Jianbing Wu, Xiao He, Songtao Liu, Zhao Zhong, Kailin Huang and Linfeng Zhang.
5. **[ICCV 2025] "Efficient Image Editing via Spatial Locality Caching."** Zexuan Yan, Yue Ma, **Chang Zou**, Wenteng Chen and Linfeng Zhang.
6. **[ACM MM 2025] "Compute only 16 tokens in one timestep: Accelerating Diffusion Transformers with Cluster-Driven Feature Caching."** - Zhixin Zheng*, Xinyu Wang*, **Chang Zou**, Shaobo Wang, Linfeng Zhang.
7. **[NeurIPS 2025] "EfficientVLA: Training-Free Acceleration and Compression for Vision-Language-Action Models"** - Yantai Yang, Yuhao Wang, Zichen Wen, Luo Zhongwei, **Chang Zou**, Zhipeng Zhang, Chuan Wen and Linfeng Zhang.
8. **[AAAI 2026]** "Forecast then Calibrate: Feature Caching as ODE for Efficient Diffusion Transformers." - Shikang Zheng, Liang Feng, Xinyu Wang, Qinming Zhou, Peiliang Cai, **Chang Zou**, Jiacheng Liu, Yuqi Lin, Junjie Chen, Yue Ma and Linfeng Zhang.
9. **[ICLR 2026] "HiCache: A Plug-in Scaled-Hermite Upgrade for Taylor-Style Cache-then-Forecast Diffusion Acceleration."** - Liang Feng*, Shikang Zheng*, Jiacheng Liu, Yuqi Lin, Qinming Zhou, Peiliang Cai, Xinyu Wang, Junjie Chen, **Chang Zou**, Yue Ma and Linfeng Zhang.

邹 暢

上海交通大学 | 人工智能学院 | 计算机科学与技术

主要学术成果-Page2

* Equal contribution

10. **[ICLR 2026] "Z-Cache: Accelerating Diffusion Transformers via Self-Reflection."** -Zegang Cheng, Zhikai Wang, Jiacheng Liu, **Chang Zou**, Junjie Chen, Yue Ma, Zhiyuan Ma and Linfeng Zhang
11. **[CVPR 2026] "LESA: Learnable Stage-Aware Predictors for Diffusion Model Acceleration."** - Peiliang Cai, Jiacheng Liu, Haowen Xu, Xinyu Wang, **Chang Zou** and Linfeng Zhang.
12. **[CVPR 2026] "From Sketch to Fresco: Efficient Diffusion Transformer with Progressive Resolution"** - Shikang Zheng, Guantao Chen, Lixuan He, Jiacheng Liu, Yuqi Lin, **Chang Zou** and Linfeng Zhang.
13. **[CVPR 2026] "Beyond Fixed Formulas: Data-Driven Linear Predictor for Efficient Diffusion Models"** - Zhirong Shen*, Rui Huang*, Jiacheng Liu, **Chang Zou**, Peiliang Cai, Shikang Zheng, zhengyi shi, Liang Feng and Linfeng Zhang.
14. **[(Under Review) TIP (Major Revision)] "Rethinking Token-wise Feature Caching: Accelerating Diffusion Transformers with Dual Feature Caching"** - **Chang Zou***, Shikang Zheng*, Evelyn Zhang, Runlin Guo, Haohang Xu, Conghui He, Xuming Hu and Linfeng Zhang.
15. **[(Under Review) TIP] "Accelerate Diffusion Transformers with Feature Momentum."** - Jiaxin Fang*, **Chang Zou***, Jiacheng Liu, Yuanhuiyi Lyu, Xuming Hu and Linfeng Zhang.
16. **[(Under Review) ICML 2026] "dLLM-Cache: Accelerating Diffusion Large Language Models with Adaptive Caching."** - Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, **Chang Zou,** Qingyan Wei, Shaobo Wang, Yichen Zhu, Linfeng Zhang
17. **[(Under Review) ICML 2026] "Let Features Decide Their Own Solvers: Hybrid Feature Caching for Diffusion Transformers."** -Shikang Zheng, Guantao Chen, Qinming Zhou, Yuqi Lin, Lixuan He, **Chang Zou**, Peiliang Cai, Jiacheng Liu and Linfeng Zhang.
18. **[(Under Review) ICML 2026] "FreqCa: Accelerating Image generation and editing via Frequency-Aware Caching"** - Jiacheng Liu*, Peiliang Cai*, Qinming Zhou, Yuqi Lin, Deyang Kong, Benhao Huang, Yupei Pan, Haowen Xu, **Chang Zou**, Junshu Tang, Shikang Zheng and Linfeng Zhang.
19. **[(Under Review) ECCV 2026] "Accelerating Diffusion Models with Gaussian Process Rectified Cache"** -Zhirong Shen*, Rui Huang*, **Chang Zou**, Shikang Zheng, Jiacheng Liu, Peiliang Cai, zhengyi shi, Zheng Lu, Liang Feng, tuxiaobing, Jinkui Ren, Xiantao Zhang and Linfeng Zhang.
20. **[(Under Review) ECCV 2026] "AViTS: Adaptive Spatiotemporal Token Selection for Efficient Dynamic-Resolution Generation"** - Haoran Qin, Zhengan Yan, Shikang Zheng, Xiaobing Tu, Jiacheng Liu, Yuqi Lin, **Chang Zou**, JinShan Liu, Peiliang Cai, Xiantao Zhang, Jinkui Ren and Linfeng Zhang.

主要科研经历与学术成果

- **上海交通大学 人工智能学院 EPIC-Lab　视觉生成式大模型推理加速　　导师：张林峰**　　2024.06 至今

近年来，随着Diffusion Transformer (DiT)的提出，图像与视频生成的质量取得了极大提升。然而，相应的计算成本也大幅提升，例如OpenAI的视频生成模型OpenSora 生成10分钟视频需要数小时甚至半天，HunyuanVideo生成5s视频需要耗时接近半小时等，导致其在实际应用中难以部署。自2024年6月起，我加入上海交通大学人工智能学院张林峰老师团队，作为**核心成员和大部分项目领导者**, 开展以视频和图像为主的生成式大模型推理加速研究为主题的科研实习。

【研究基础】扩散模型的推理过程中，从初始噪声出发，以多步迭代形式逐渐进行去噪，最终得到干净清晰的生成目标。特征缓存方法洞见了在扩散过程中相邻时间步上特征的高度相似性，通过对已经推理的特征进行缓存并跨时间步进行复用，实现大幅压缩计算量，无需训练即可带来数倍的推理加速。我们的工作从特征缓存方法出发，从包括但不限于更细致的缓存粒度[1,5,6,8]，更优的缓存和推理范式[2,3,7,12,13,14]，更多模态更多任务上的优化[4,9,10]等方向开展研究。

【关键工作1】 [ICLR 2025] “Accelerating Diffusion Transformers with Token-wise Feature Caching.”

TL;DR: 通过细粒度到token级的特征缓存，对图像更重要的部分提供更多计算，实现更快更好的加速。

传统的基于特征缓存的方法往往依赖于Stable Diffusion特殊的U-Net结构，无法直接迁移到DiT模型中，且以往的方法往往以整个特征图为单位进行复用。我们的研究发现，不同深度的层，以及同一层的不同token上缓存引入的误差，可以有数十，百倍的差异，因此有必要**考虑更细粒度的token级的特征缓存-复用方案**（ToCa），并设计了从不同角度出发的4种token选择方法。在文生图（PixArt-alpha，FLUX），文生视频（OpenSora），以及类标签生图（DiT）等模型上开展的充分实验充分证明了我们方法不仅无需训练，还具有极佳的加速效果，例如在OpenSora上能实现2.36倍的无损加速。作为团队在特征缓存方向的首篇工作，也是领域奠基工作之一，ToCa已获超过160 github stars, 并被公众号量子位报道。

【关键工作2】 [ICCV 2025] “Features of Diffusion Models are Predictable: Accelerating Diffusion Transformers with Taylor Seers.”

TL;DR: 通过改进“缓存-复用”范式到“缓存-预测”，引入时序建模到特征缓存，显著降低缓存误差，实现加速。

以往的扩散模型的特征复用策略通过直接将特征在后续的一/多步中直接复用来实现加速，而这样的复用由于新的时间步本质上与旧的时间步并不相同，会引入特征复用导致的误差。在这篇本人作为共同第一作者的工作（TaylorSeer）中，我们创新地提出扩散模型的**特征预测**的概念，并通过前面数步的特征来通过差分方法逼近特征随时间步变化函数的各阶导数，以**使用泰勒展开来预测接下来各步的特征变化**，从而实现大幅减小特征复用导致的误差。在DiT，FLUX, HunyuanVideo 等多种任务的前沿模型上开展的实验结果能充分地证明，TaylorSeer具有远超同期方法的加速性能，例如在同等加速比的DiT模型下，TaylorSeer的质量损失仅是先前SOTA方法的2.98%；保证极高质量生成的前提下，在FLUX.1和HunyuanVideo以及Wan2.1模型上突破性地达到了接近5倍的计算压缩（对比同期的SOTA方法仅有3.5倍）。通过将TaylorSeer与序列并行技术结合，我们成功实现了在5s内生成5s的视频，使高质量实时生成视频成为可能。TaylorSeer 目前已获超过230 github stars, 并受知名公众号量子位约稿；团队也基于此项目与工业界开展长期合作。

【关键工作3】 [ACM MM 2025] “Accelerating Diffusion Models with Speculative Sampling.”

TL;DR: 将自回归的投机采样方法引入到扩散模型，为更难样本分配更多计算，并以缓存方法对简单样本快速推理。

以往的加速方法往往对所有样本采用完全一致的加速方案，这与我们直觉上认识的“样本间有同样有难度差异”的认识不相符。一个很自然的观点是，**更难的样本值得分配更多计算，而简单样本只需要少量计算**即可。在这篇本人为共同第一作者的工作（SpeCa）中，我们提出了一种**可用于扩散模型的投机采样方案**，并以目前表现最好的加速方案TaylorSeer作为快速推理的draft model, 构建了“缓存-预测-检查”的流程。SpeCa 在中间步骤优先使用TaylorSeer进行高效精准的推理，而当TaylorSeer在最后层的预测值和最后层单独计算的值差异过大，则切换回原模型进行完整计算。该方法通过动态地调整计算量在sample间的合理分配，进一步提升了扩散模型的推理效率。保证极高质量生成的前提下，SpeCa分别实现了在FLUX上5.70倍的计算压缩和HunyuanVideo上6.16倍的计算压缩。