赵天辰

(+86) 176-1027-2031 | suozhang1998@gmail.com | 个人主页 tianchen-zhao.info | 中国北京

教育背景

清华大学 电子工程系 博士研究生（导师：汪玉教授） *2023.09 – 至今*

北京航空航天大学 电子工程专业 学士 & 工学硕士 *2016.09 – 2023.06*

研究方向

面向基座模型构建的 EfficientML 算法与系统协同优化。 注：更具体的研究框架图见 个人主页

面向多模态基模设计: EfficientML 算法系统协同 & 高效采样

- 面向视觉生成模型的稀疏与低比特量化算法：
  - **[ECCV'24] MixDQ & [ICLR'25] ViDiT-Q:** 针对基于 Diffusion 的视觉生成模型，进行算法方案改进，实现了无损的 INT8 与 INT4/8 混合位宽量化，且实现并开源 Fused Kernel，实现 1.4–1.7x 的推理加速。
  - **[DAC'25] PARO & [NeurIPS'25] PAROAttn:** 设计了基于"Token Reorder"的视觉生成模型注意力稀疏-量化优化方案，进行了算法-系统协同设计，完成完整系统实现。可无训练实现完全无损的 90% 稀疏，与全 Attention INT8 量化，可实现 3-6x 的注意力算子加速。
- 面向新兴多模态模型的高效采样方案设计：
  - **[ArXiv'26] StreamingVLA:** 针对 VLA 模型多阶段的等待同步问题，提出基于动作空间 Flow Matching 建模方式与自适应提前观测, 实现了 VLA 的各阶段异步执行，执行速度提速 2.4x，卡顿时间减少 5x。
  - **[ongoing] Streaming Forcing:** 针对目前 Frame-wise AR 的视频生成模型 (Self_Forcing)，提出基于改进 Flow Matching 建模方式，优化控制信号引入方式，实现更高效的训练，并解决长视频外推中的"鬼影"现象。

面向 Agentic RL 的算法与 Infra 优化

- 面向 Multi-Agent 与 Context 管理的 RL Post-training 算法：
  - **[Intern] at Miromind** 针对 Deep Research Agent，设计 Multi-agent GRPO 方案，与多轮间 Context Compression 方案，可在显著压缩最大 Context 长度情况下（32K-8K），获得近似的训练效果，显著提升训练效率。
- 面向 RL Rollout 的 Tile 化 Megakernel Runtime 设计：
  - **[Intern] at Naive.AI** 针对 RL 训练中长尾 Rollout 阻碍高利用率的问题，不通过主流调度优化，旨在通过设计与采用针对低并行度的高效 Runtime 实现，来针对性提速少量长尾 Rollout。采用类似 TileRT，基于 Tile-based Megakernel-like，与静态编译的技术路线，设计高速推理引擎。

注：其他方向，如 *3D* 点云基模、*AutoML* 与更偏硬件工作，未在此列出，见后续"发表论文列表"

实习经历

超星未来 研究实习生 *2019.10 – 2020.06*

- 作为团队成员参与 LPCVC20 竞赛（面向 Google Pixel 4 的 Detection 模型部署优化）

无问芯穹 研究实习生 *2023.06 – 2025.01*

- 围绕基于 Diffusion 的视觉模型效率优化方法开展研究：
  - 文生图模型的快速评测方法（*FlashEval*, CVPR 2024）
  - 少步扩散模型的混合精度量化方法（*MixDQ*, ECCV 2024）
  - 基于 Diffusion Transformer 的图像与视频生成量化方法（*ViDiT-Q*, ICLR 2025）
  - 面向混合位宽注意力量化的加速器设计（*PARO*, DAC 2025）

字节跳动 Seed 视觉 研究实习生 (TopSeed 计划) *2025.02 – 2025.08*

- 开展高效视觉生成模型方向研究
  - 基于令牌重排的注意力稀疏化与量化算法-系统联合优化（*PAROAttention*, NeurIPS 2025）

Miromind 研究实习生 *2025.08 – 2026.1*

- 面向 Deep Research Agent 的 RL post train 算法设计
  - 探索 Multi-agent GRPO 与 Context Management 方法，提升智能体效果与效率

Naive.AI 研究实习生 *2026.01 – Now*

- 面向 Coding Agent 的 RL Rollout 框架效率优化：
  - 探索面向少 Batch，长序列的长尾 Rollout 的推理优化方案（Megakernel-like Tile-based Runtime 设计）

发表论文列表

[1] Yiran Shi, Dongqi Guo, **Tianchen Zhao**, Feng Gao, Liangzhi Shi, Chao Yu, ZhiJian Mo, Qihua Xiao, XiaoShuai Peng, Qingmin Liao, Yu Wang. "StreamingVLA: Streaming Vision-Language-Action Model with Action Flow Matching and Adaptive Early Observation". Project Page [arXiv'26]

[2] Siqi Chen, Ke Hong, **Tianchen Zhao**, Ruiqi Xie, Zhenhua Zhu, Xudong Zhang, Yu Wang. "db-SP: Accelerating Sparse Attention for Visual Generative Models with Dual-Balanced Sequence Parallelism". [MLSys'25]

[3] **Tianchen Zhao**, Ke Hong, Xinhao Yang, Xuefeng Xiao, Huixia Li, Feng Ling, Ruiqi Xie, Siqi Chen, Hongyu Zhu, Yichong Zhang, Yu Wang. "PAROAttention: Pattern-Aware ReOrdering for Efficient Sparse and Quantized Attention in Visual Generation Models". Project Page [NeurIPS'25]

[4] Xinhao Yang*, **Tianchen Zhao***, Hongyi Wang, Wenheng Ma, Shulin Zeng, Zhenhua Zhu, Xuefei Ning, Huazhong Yang, Yu Wang. "PARO: Hardware-Software Co-design with Pattern-aware Reorder-based Attention Quantization in Video Generation Models". [DAC'25]

[5] **Tianchen Zhao**, Tongcheng Fang, Enshu Liu, Wan Rui, Widyadewi Soedarmadji, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, Yu Wang. "ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation". Project Page [ICLR'25]

[6] **Tianchen Zhao***, Xuefei Ning*+, Tongcheng Fang*, Enshu Liu, Guyue Huang, Zinan Lin, Shengen Yan, Guohao Dai, Yu Wang+. "MixDQ: Memory-Efficient Few-Step Text-to-Image Diffusion Models with Metric-Decoupled Mixed Precision Quantization". Project Page [ECCV'24]

[7] Lin Zhao*, **Tianchen Zhao***, Zinan Lin, Xuefei Ning+, Guohao Dai, Huazhong Yang, Yu Wang+. "FlashEval: Towards Fast and Accurate Evaluation of Text-to-image Diffusion Generative Models". Project Page [CVPR'24]

[8] Tongxin Xie, **Tianchen Zhao**, Zhenhua Zhu+, Xuefei Ning, Bing Li, Guohao Dai, Huazhong Yang, Yu Wang+. "DyPIM: Dynamic-inference-enabled Processing-In-Memory Accelerator". [DATE'24]

[9] **Tianchen Zhao**, Xuefei Ning, Ke Hong, Zhongyuan Qiu, Pu Lu, Linfeng Zhang, Yali Zhao, Lipu Zhou, Guohao Dai, Huazhong Yang, Yu Wang. "Ada3D: Exploiting the Spatial Redundancy with Adaptive Inference for Efficient 3D Object Detection". Project Page [ICCV'23]

[10] **Tianchen Zhao**, Niansong Zhang, Xuefei Ning, He Wang, Li Yi, Yu Wang. "CodedVTR: Codebook-based Sparse Voxel Transformer with Geometric Guidance". Project Page [CVPR'22]

[11] Xuefei Ning*, **Tianchen Zhao***, Wenshuo Li, Peng Lei, Yu Wang, Huazhong Yang. "DSA: More Efficient Budgeted Pruning via Differentiable Sparsity Allocation". [ECCV'20]

学术活动与获奖

- 会议审稿人：CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML
- Outstanding Reviewer：NeurIPS 2024, NeurIPS 2025, CVPR 2025
- 国家奖学金（2025，清华大学）
- CVPR Workshop Low Power Computer Vision Contest (LPCVC, Detection Track) 2020, 3rd Place