Tianchen Zhao

(+86) 176-1027-2031 | suozhang1998@gmail.com | Homepage tianchen-zhao.info | Beijing, China

Education

Tsinghua University Department of Electronic Engineering Ph.D. Student (Advisor: Prof. Yu Wang) *2023.09 – Present*

Beihang University Electronic Engineering B.Eng. & M.Eng. *2016.09 – 2023.06*

Research Interests

Primary Interest: EfficientML algorithm-system co-design for foundation model construction. *Note: a more detailed research framework is available on Homepage*

Multimodal Foundation Models: EfficientML Co-Design & Efficient Sampling

- Sparse and low-bit quantization for visual generative models:
  - **[ECCV'24] MixDQ & [ICLR'25] ViDiT-Q:** Developed improved algorithmic and systems solutions for Diffusion-based visual generation models, enabling lossless INT8 and mixed INT4/8 quantization. Implemented and open-sourced fused kernels, achieving 1.4–1.7x inference speedup.
  - **[DAC'25] PARO & [NeurIPS'25] PAROAttn:** Designed token-reorder sparse/quantized attention optimization for visual generation models with end-to-end algorithm-system co-design and implementation. Achieves training-free, lossless 90% sparsity together with full-attention INT8 quantization, with 3–6x attention speedup.
- Efficient sampling for emerging multimodal models:
  - **[arXiv'26] StreamingVLA:** Addressed multi-stage synchronization bottlenecks in VLA models with action-space flow matching and adaptive early observation, enabling asynchronous execution across stages, with 2.4x execution speedup and 5x reduction in stalling time.
  - **[Ongoing] Streaming Forcing:** For frame-wise autoregressive video generation models (Self_Forcing), exploring an improved flow-matching formulation to optimize control-signal injection for more efficient training and reduced ghosting during long-video extrapolation.

Agentic RL Algorithms and Infrastructure Optimization

- RL post-training algorithms for multi-agent learning and context management:
  - **[Intern] at Miromind:** Designed multi-agent GRPO together with inter-turn context compression for Deep Research Agents. Achieved comparable training quality under substantially reduced maximum context length (32K→8K), significantly improving training efficiency.
- Tile-based megakernel runtime for RL rollout:
  - **[Intern] at Naive.AI:** Targeted low utilization caused by long-tail rollouts in RL training without relying on mainstream scheduling optimizations. Designed a high-efficiency runtime for low-parallelism workloads, following a TileRT-like, tile-based megakernel and static-compilation path to accelerate long-tail rollout inference.

Note: other directions, including 3D point-cloud foundation models, AutoML, and more hardware-oriented work, are omitted here and can be found in the publication list below.

Research Experience

Novauto **Research Intern** *2019.10 – 2020.06*

- Contributed to the LPCVC 2020 competition, focusing on detection model deployment optimization for Google Pixel 4.

Infinigence AI **Research Intern** *2023.06 – 2025.01*

- Worked on efficiency optimization for Diffusion-based visual models:
  - Fast evaluation for text-to-image generation (*FlashEval*, CVPR 2024)
  - Mixed-precision quantization for few-step diffusion models (*MixDQ*, ECCV 2024)
  - Quantization for Diffusion Transformer image and video generation (*ViDiT-Q*, ICLR 2025)
  - Accelerator design for mixed-bitwidth attention quantization (*PARO*, DAC 2025)

ByteDance Seed Vision **Research Intern (TopSeed Program)** *2025.02 – 2025.08*

- Conducted research on efficient visual generative models:
  - Joint algorithm-system optimization for token-reordering-based sparse and quantized attention (*PAROAttention*, NeurIPS 2025)

Miromind **Research Intern** *2025.08 – 2026.01*

- Designed RL post-training algorithms for Deep Research Agents:
  - Explored multi-agent GRPO and context management methods to improve agent quality and efficiency

Naive.AI **Research Intern** *2026.01 – Present*

- Efficiency optimization for RL rollout frameworks in coding agents:
  - Explored inference optimization for small-batch, long-sequence, long-tail rollouts via a megakernel-like tile-based runtime design

Publications

[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]

Academic Service & Honors

- Reviewer: CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML
- Outstanding Reviewer: NeurIPS 2024, NeurIPS 2025, CVPR 2025
- National Scholarship (2025, Tsinghua University)
- 3rd Place, CVPR Workshop Low Power Computer Vision Contest (LPCVC, Detection Track), 2020