

Zhizhou Sha

+1 512 910 9476 | shazhizhou0@gmail.com | jamessand.github.io

EDUCATION

University of Texas at Austin

Aug. 2025 – Present

Ph.D. in Computer Science (in progress)

Austin, USA

- Advisors: Qiang Liu and Atlas Wang.
- Research: reinforcement learning, numerical optimization, and efficient machine learning.

Tsinghua University

June 2025

B.Eng. in Computer Science and Technology

Beijing, China

- Tsinghua University Comprehensive Scholarship (Top 20%), 2023 and 2024.

RESEARCH EXPERIENCE

Geometry-Aware Reinforcement Learning | Advisor: Atlas Wang

- **The Path Not Taken: RLVR Provably Learns Off the Principals (NeurIPS 2026)** [[Paper](#)]
 - * Analyzed training trajectories and found that weight singular values remain approximately unchanged during reinforcement learning with verifiable rewards (RLVR) in LLMs, in contrast to shrinkage during supervised fine-tuning; deepened understanding of weight-update structure through training dynamics.
- **Isospectral Stiefel Optimizer (ISO)** [[Paper](#)]
 - * Translated the observed geometry into an optimizer that fixes weight singular values and uses the matrix sign mapping to update singular-vector factors on Stiefel manifolds. On math-reasoning tasks, matched AdamW's best accuracy with **up to approximately 70% fewer training steps**; achieved **approximately 10% higher average accuracy** at the same step budget.

Efficient Learning Algorithms | Advisors: Yingyu Liang and Zhenmei Shi

2024 – 2025

- **Almost-linear-time Transformer gradient computation** [[Paper](#)]: constructed Taylor-based low-rank approximations for large-scale attention, reducing theoretical gradient computation from $O(n^2)$ to $O(n^{1+o(1)})$ with provably bounded error, enabling end-to-end almost-linear-time computation across forward and backward passes.
- **Sparse attention acceleration (CPAL 2025)** [[Paper](#)]: used Half-Space Reporting (HSR) to locate active key-value pairs in large-scale attention, reducing cross-attention complexity from $O(mnd)$ to $O(mn^{4/5}d)$.

INTERNSHIP EXPERIENCE

Together AI | Research Intern

May – Aug. 2026

- **Low-rank compression of video-generation KV caches: Budget-PCA**
 - * Addressed large KV-cache storage requirements with a calibration-free, fixed-cost method combining chunk-wise low-rank PCA and channel-wise asymmetric 2-bit residual quantization. Evaluated three video-generation systems: reduced storage from 16 to **at most 2.326 bits per element**, saving **approximately 85% of storage**; outperformed QVG, the state-of-the-art 2-bit video quantization method, on **11 of 18 metrics**, including a **3.48 dB PSNR gain** on LongCat-Video.
- **Model evaluation under memory constraints**
 - * Built an end-to-end GLM-5.2 cache-quantization evaluation pipeline covering reasoning and long-context benchmarks (GPQA, RULER); assessed quantization-induced accuracy loss and reduced 128K cache memory from **10.97 GiB to 2.91–3.08 GiB**.
 - * Analyzed numerical errors across cache components and proposed a hybrid-precision design balancing phase fidelity, tail behavior, and memory usage.

ByteDance | Research Intern

Dec. 2023 – Jan. 2024

- Fine-tuned BERT for content moderation and performed daily incremental learning on streaming text data to adapt to evolving content.

TECHNICAL SKILLS

Programming: Python, C++, CUDA **Machine learning:** PyTorch, Hugging Face, verl, vLLM

Systems: Linux, distributed training (FSDP, DDP, Ray), Slurm

Mathematics and algorithms: Numerical linear algebra (SVD, PCA, low-rank approximation), manifold optimization, computational complexity and approximation error analysis.