ZEHAO JIN

+1 404-668-1916 | lunamos.thu@gmail.com | lunamos.github.io

EDUCATION

Georgia Institute of Technology, Atlanta, Georgia, US Jan 2026 – Present
Master of Science in Computational Science and Engineering (GPA: 4.0/4.0)

Tsinghua University, Beijing, China Sept 2021 – Jul 2025
Bachelor of Science in Mechanics and Aerospace Engineering (GPA: 3.8/4.0)

PROFESSIONAL EXPERIENCE

StepFun May 2026 – Present
Foundation Model Research Intern | Shanghai, China

- Contributed to **Agentic Rubrics RL post-training** and an internal **agentic evaluation** benchmark.
- Built an automated rubrics-benchmark generation workflow based on **Claude Code** and **Codex**.
- Deployed **Hermes** and **Openclaw** on the internal network to integrate agents into team workflows.

Keeta (Meituan) LongCat Sept 2025 – Dec 2025
Foundation Model Research Intern | Shanghai, China

- Deployed and post-trained **vision-language models (VLM)** and **Vision-Language-Action models (VLA)**.
- Researched and implemented agentic strategies to improve **Multimodal LLM (MLLM)** performance.

RESEARCH HIGHLIGHTS

Beyond Steering Vector: Flow-based Activation Steering for Inference-Time Intervention March 2026 – May 2026
Georgia Institute of Technology | Atlanta, Georgia

- Proposed FLAS (Flow-based Activation Steering), a concept-conditioned velocity field for LLM inference-time intervention that models multi-step, token-varying activation trajectories.
- Achieved **state-of-the-art** zero-shot steering on AxBench with Gemma-2 (2B/9B), becoming the **first method** to consistently outperform in-context prompting and HyperSteer while using under 4% of the parameters.
- Project Page: https://flas-ai.github.io, Demo: https://huggingface.co/spaces/Lunamos/flas-demo.

Decompose Sparsely Where You Should, Absorb Densely Where You Should Not March 2026 – May 2026
Georgia Institute of Technology | Atlanta, Georgia

- Identified a low-rank, dense ***computational scaffold*** in LLM residual streams that is causally critical yet inherently unsuitable for sparse decomposition.
- Proposed a rank-r linear bottleneck parallel to standard **SAEs**, absorbing dense structure before sparse reconstruction without modifying the SAE.
- Reduced dense latent count by up to 84% while improving sparse probing at matched sparsity on Gemma-2-2B.

Stochastic Attention: Randomized Routing for Expressive Linear-Time Attention Dec 2025 – Apr 2026
Tsinghua University | Beijing, China

- Proposed Stochastic Attention (SA), a parameter-free enhancement for sliding-window attention by introducing random token permutations. Achieved up to **28× speedup over full attention** at 32K sequence length.
- **Pretrained 360M-parameter language models** on SlimPajama with a gated SA+SWA dual-path architecture, achieving the best average zero-shot accuracy across 7 benchmarks over previous state-of-the-art methods.
- Paper link: https://arxiv.org/abs/2604.00754, under review.

SELECTED HONORS AND AWARDS

- **Ying-Hua Fellowship**, Tsinghua University (for 15 undergrads) Dec 2022
- **Spark Fellowship**, Tsinghua University (for 40 undergrads) May 2023
- **Overall Excellence Scholarship**, Tsinghua University (for top 5% undergrads) Nov 2022 & 2024
- **Finalist**, Interdisciplinary Contest in Modeling (ICM) (for top 2% teams) May 2023

PUBLICATION HIGHLIGHTS

[1] **Zehao Jin***, Ruixuan Deng*, Junran Wang*, Xinjie Shen, Chao Zhang, "Beyond Steering Vector: Flow-based Activation Steering for Inference-Time Intervention", https://arxiv.org/abs/2605.05892, Under review.

[2] Jiaming Wang*, Zhe Tang*, **Zehao Jin***, Hefei Chen*, Yilin Jin*, Peng Ding*, Xiaoyu Li, Xuezhi Cao, "SOP-Maze: Evaluating Large Language Models on Complicated Business Standard Operating Procedures", https://arxiv.org/abs/2510.08942, ACL Findings 2026.

[3] **Zehao Jin***, Yaoye Zhu, Chen Zhang, Yanan Sui, "Whole-Brain Connectome-Instantiated Model for Whole-Body Movement Control in Drosophila", Cosyne 2026 Poster.

SKILLS

Language: Fluent in English (TOEFL 109) and Mandarin.

Technical: PyTorch, Hugging Face Transformers, Megatron, vLLM, Python, C/C++.

Research Focus: MechInterp, AI Alignment, LLM Agent.