金泽昊

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

教育背景

2026 - 至今 **佐治亚理工学院** 美国, 亚特兰大
计算科学与工程，硕士（GPA: 4.0/4.0）

2021 - 2025 **清华大学** 中国, 北京
力学与航空航天工程，本科（GPA: 3.8/4.0）

工作经历

基础模型研究实习生 | 阶跃星辰 *StepFun* 2026.05 - 至今
- 参与 **Agentic Rubrics 强化学习后训练（RL post-training）**与内部智能体复杂任务评测基准的建设。
- 搭建基于 **Claude Code 与 Codex** 的 rubrics 评测基准自动生成工作流，包括自动生成、质检、自我迭代。
- 在内网部署 **Hermes 与 Openclaw**，将 agent 集成进团队工作流。

基础模型研究实习生 | 美团 *LongCat* 2025.09 - 2025.12
- 部署和后训练**视觉语言模型（VLM）**和**视觉-语言-动作模型（VLA）**，在公司内部平台可用。
- 研究并实现 agentic 策略以提升**多模态大语言模型（MLLM）**在游戏情景下的性能。
- 作为共同第一作者参与了 **SOP-Maze**（ACL Findings 2026）和其他两篇投稿中的大语言模型指令遵循评测论文。

研究经历

通用激活引导：基于流的激活引导用于推理时干预 | 佐治亚理工学院 2026.03 - 2026.05
- 提出 **FLAS（Flow-based Activation Steering）**，用于大语言模型推理时干预，建模多步、随 token 变化的激活轨迹。
- 在激活引导基准测试 AxBench 上以 Gemma-2（2B/9B）实现最优表现（SOTA），是首个在使用的情况下持续超越上下文工程（in-context prompting）的激活引导方法。相较于此前最优方法 HyperSteer，使用 $<$ **4%** 参数量实现 **+67%** 性能。
- 项目主页：flas-ai.github.io，Demo：HuggingFace Space。

稀疏自编码器优化：该稀疏分解时稀疏分解，不该时稠密吸收 | 佐治亚理工学院 2026.03 - 2026.05
- 发现大语言模型残差流中存在一种低秩稠密的**“计算脚手架”**，它对模型计算因果关键，却本质上不适合稀疏分解。
- 提出与标准稀疏自编码器（SAE）并行的 rank-r 线性分支，在稀疏重构前吸收稠密结构，且无需修改 SAE 本身。
- 在 Gemma-2-2B 上将稠密潜在单元（dense latent）数量**最多减少 84%**，并在相同稀疏度下提升稀疏探测性能。

随机注意力：随机路由实现高表达力线性时间注意力机制 | 清华大学 2025.12 - 2026.04
- 提出随机注意力（Stochastic Attention, SA），通过 token 随机化路由实现无参数增强滑动窗口注意力（Sliding Window Attention，SWA）；在 32K 序列长度下相比全注意力实现**最高 28 倍加速**。
- 采用门控 SA+SWA 双路径架构**预训练 360M 参数语言模型**，在 7 个基准上的平均零样本准确率超越此前最优方法。
- 论文：arXiv:2604.00754，under review。

学术出版物

[1] **Zehao Jin***, Ruixuan Deng*, Junran Wang*, Xinjie Shen, Chao Zhang, “Beyond Steering Vector: Flow-based Activation Steering for Inference-Time Intervention”, arXiv: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”, arXiv: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.*

荣誉奖项

英华学者 清华大学（全校 15 名本科生） 2022
星火计划 清华大学（全校 40 名本科生） 2023
综合优秀奖学金 清华大学前 5% 2022, 2024
国际大学生数学建模竞赛 (ICM) 特等奖提名 (Finalist) 全球参赛队伍前 2% 2023

专业技能

语言能力 英语流利（托福 109），中文母语
技术栈 PyTorch, Hugging Face Transformers, Megatron, vLLM, Python, C/C++
研究方向 机制可解释性（MechInterp）、AI 安全对齐（AI Alignment）、大语言模型智能体（LLM Agent）