王熠笑

yixiao.site | yixiaowang@sjtu.edu.cn | +8613390606996

教育背景

上海交通大学计算机学院—软件工程 GPA：3.92 / 4.3（排名：6 / 88）

基础科学课程成绩：电路理论（荣誉，99）、线性代数（荣誉，97）、大学物理（荣誉 I，96；荣誉 II，97）

专业课程成绩：软件工程实践（99）、数据结构（97）、互联网应用开发技术（96）、计算机系统基础（94）

科研经历

ELORA: Efficient LoRA and KV Cache Management for Multi-LoRA LLM Serving [论文]

- ELORA 是一种面向多 LoRA 大语言模型服务的高性能缓存管理系统，通过统一管理 LoRA 适配器与 KV Cache，降低 TTFT 和 TPOT、提升吞吐量，并缓解缓存利用低效与显存碎片化问题。
- 系统提出了依赖感知的缓存管理器和性能驱动的缓存换入换出机制，能够基于使用依赖关系动态管理 LoRA 与 KV Cache，从而更智能地分配和利用 HBM 资源，降低推理延迟并提升系统吞吐量。

MOBA: Multifaceted Memory-Enhanced Adaptive Planning for Efficient Mobile Task Automation [论文]

- MOBA 是一种面向高效移动任务自动化的记忆增强规划框架，通过改善现有移动智能体在长程任务中的上下文理解、任务分解和动态环境适应能力，提升任务执行的稳定性与完成效果。
- 系统利用多维记忆机制捕获历史交互、任务经验和环境感知信息，并基于这些记忆动态优化规划策略，从而实现更可靠的动作选择，并提升复杂移动自动化场景下的任务完成效率。

BuddyMoE: Exploiting Expert Redundancy to Accelerate Memory-Constrained Mixture-of-Experts Inference [论文]

- BuddyMoE 是一种面向显存受限混合专家模型的高效推理系统，针对专家卸载带来的推理延迟瓶颈，尤其是预取失败时 CPU 到 GPU 的专家传输造成的推理停顿问题进行优化。
- 系统通过挖掘 MoE 专家之间的冗余关系，使用已驻留在 GPU 显存中的语义相似“伙伴专家”替换缺失专家，从而减少传输停顿，在保持模型精度的同时提升有限显存条件下的大规模 MoE 推理效率。

荣誉奖项

上海市优秀毕业生 2026 年 5 月

致远荣誉奖学金，上海交通大学 2022–2025，4 次

C 等奖学金，上海交通大学 2023–2025，3 次

其他经历

- **助教**：程序设计思想与方法（荣誉，C++），2024–2025 秋季学期
- **助教**：概率统计（荣誉），2025–2026 春季学期