Yixiao Wang

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

Academic Background

School of Computer Science, Shanghai Jiao Tong University — Software Engineering
GPA: 3.92 / 4.3 (Ranking: 6 / 88)

Performance in Fundamental Science Courses: Circuit Theory (Honors, 99) Linear Algebra (Honors, 97) University Physics (Honors I, 96; Honors II, 97) Mathematical Analysis (Honors, 95)

Performance in Major Courses: Software Engineering Practice (99) Data Structures (97) Internet Application Development Technologies (96) Fundamentals of Computer Systems (I, 93; II, 94)

Research Experience

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

- ELORA is a high-performance cache management system for multi-LoRA LLM serving. By unifying LoRA adapter and KV cache management, it reduces TTFT and TPOT while improving throughput and addressing ineffective cache usage and memory fragmentation in existing systems.
- The system introduces a Dependency-aware Cache Manager and a Performance-driven Cache Swapper, which dynamically manage LoRA and KV caches based on usage dependencies. This enables smarter allocation and utilization of HBM resources, thereby reducing latency and improving throughput.

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

- MOBA is a memory-enhanced planning framework for efficient mobile task automation. It improves long-horizon task execution by addressing limitations of existing mobile agents, including insufficient contextual understanding, weak task decomposition, and limited adaptability in dynamic environments.
- The system leverages multifaceted memory to capture historical interactions, task experiences, and environment-aware information. Using these memories, MOBA refines planning strategies, enabling reliable action selection and improving completion efficiency across mobile automation scenarios.

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

- BuddyMoE is an efficient inference system for memory-constrained Mixture-of-Experts models. It targets the latency bottleneck caused by expert offloading, where prefetch failures require expensive expert transfers from CPU memory to GPU memory and significantly slow down inference.
- The system exploits redundancy among MoE experts by replacing missing experts with semantically similar "buddy" experts resident in GPU memory, reducing CPU-GPU transfer stalls while preserving model accuracy, thereby improving large-scale MoE inference efficiency under limited GPU memory.

Honors & Awards

Outstanding Graduate of Shanghai May 2026
Zhiyuan Honor Scholarship, Shanghai Jiao Tong University 2022–2025, 4 times
Class C Scholarship, Shanghai Jiao Tong University 2023–2025, 3 times

Others

- **Teaching Assistant**: Programming Design and Methodology (Honor, C++), Fall 2024–2025
- **Teaching Assistant**: Probability and Statistics (Honor), Spring 2025–2026