Ziyi Xu

xzy2022@sjtu.edu.cn

Education

Shanghai Jiao Tong University — Shanghai, China
B.S. in Computer Science, ACM Honors Class — *Sept. 2022 – Present*

- GPA: 4.0/4.3 (Rank: 4/30)
- Selected courses:
  - Compiler Design and Implementation: 99/100
  - Computer Architecture: 99/100
  - Operating System: 100/100

Research Experience

Stanford University — Stanford, USA
Research Intern, advised by Prof. Christos Kozyrakis — *Jun. 2025 – Nov. 2025*

Shanghai Jiao Tong University — Shanghai, China
Research Assistant, advised by Prof. Quan Chen — *Jul. 2024 – May 2025*

Research Projects

FailSafe: High-performance Resilient Serving [arxiv]
Submitted to MLSys 2026 — *Jun. 2025 – Nov. 2025*

- A fault-tolerant LLM serving system resilient to GPU failure, with extremely fast recovery from failure and high performance on irregular GPU availability (e.g., **TP7, TP3**).
- Introduce proactive KVCache backup and on-demand weight recovery, achieving **183× faster recovery**.
- Eliminate post-failure GPU resource imbalance with cyclic KVCache placement and hybrid attention mechanisms, achieving up to **2× performance improvement** in various tensor parallelism sizes.

Strata: Hierarchical Context Caching for Long-context LLM Serving [arxiv]
Submitted to ASPLOS 2026 — *Feb. 2025 – May 2025*

- Accelerate long-context LLM serving via hierarchical context caching on top of SGLang.
- Achieve **5× lower TTFT** and **3.8× higher throughput** with custom PCIe transfer CUDA kernels and IO-aware scheduling, significantly reduce GPU stall and underutilization in long context serving.
- Integrated into SGLang and has been deployed **in production**.

Optimizing SLO-oriented LLM Serving with PD-Multiplexing [arxiv]
ASPLOS 2026 (Major Revision) — *Jan. 2025 – Apr. 2025*

- A LLM serving framework that adopts PD-multiplexing paradigm on top of SGLang
- Co-locate prefill and decode stage both **temporally and spatially** on the same GPU, with fine-grained resource control to achieve higher throughput without violating SLO target.
- Improve the performance by **5.1**× over SOTA baselines under the same SLO target.
- Adopted by popular open-source LLM serving engines **SGLang** and **TensorRT-LLM** .

Efficient Function-as-a-Service for Large Language Models [arxiv]
Submitted to OSDI 2026 — *Sept. 2024 – Dec. 2024*

- A FaaS framework for LLM applications that achieves fast startups by tracing fine-grained execution paths.
- Accelerate serverless ML cold start by **2.1**× compared with SOTA solutions.
- Enabled various tracing-guided optimizations **automatically** with **minimal code changes** in user code.

Open-Source Projects

SGLang
A high-performance serving framework for large language models. — *Oct. 2024 –*

- Integrate XGrammar for **10× faster structural generation.** [BLOG]
- Optimize hierarchical caching in long context serving, with up to **80% reduction in TTFT.** [BLOG]
- **Day-0 support** for DeepSeek V3.2 model, with various performance-critical kernel optimizations. [BLOG]

XGrammar
Fast, Flexible and Portable Structured Generation. — *Oct. 2024 –*

- Support C++ static reflection, and implement auto-serialization and structural comparison on top of that.
- Implement thread-safe LRU cache with multi-thread grammar compiler, achieving more than **2× speedup**.

Honors & Awards

Zhiyuan Honorary Scholarship (Top 2%, 2022–2023)

Academic Excellence Scholarship, Second Prize (2023)

Teaching Assistant Experience

Computer Programming (Prof. Huiyu Weng), Fall 2023

Programming Practice (Prof. Yong Yu), Summer 2024

Optimization Methods (Prof. Kuan Yang and Prof. Bo Jiang), Fall 2024

Advanced Compilers (Prof. Yong Yu), Spring 2025

Technical Skills

Programming Languages: C/C++, CUDA, Python, Go, Rust, Verilog, Coq

Tools: Git, Docker, CMake, xmake, LaTeX

Languages: Mandarin (native), English (fluent)

Research Interests

- Building reliable high-performance systems for distributed and large-scale machine learning.
- Exploring compiler-based optimizations, tracing, and static analysis, along with efficient scheduling and execution.
- Developing practical and impactful systems that benefit a broad range of real-world users.