Mengyue Huang

(Formerly Mengyue Xi)

Email: ximy@alumni.sysu.edu.cn | Website | LinkedIn | Google Scholar

Research Interests

AI-assisted Computer Architecture, with a focus on GPU performance analysis, memory systems, and software-hardware co-design.

Technical Skills

Programming: C/C++, Python, Verilog
Systems: CUDA/HIP, CUDA/ROCm ecosystem, LLVM/Clang
AI/ML: PyTorch, LLMs, machine learning

Education

Sun Yat-sen University (SYSU), Master of Computer Science and Technology — Sep. 2023 – June 2026
- GPA: 94.92/100.00, Rank: 2/69, Advised by Prof. Xianwei Zhang
- Master's Thesis: GPU Multi-Level Cache Management with Compiler Optimization and Large Language Models (based on [2] and [3]).

Chongqing University, Bachelor of Computer Science and Technology — Sep. 2019 – June 2023
- GPA: 3.82/4.00, Rank: 16/219
- Bachelor's Thesis: *Reconfigurable (CGRA) Architecture Optimization with MLIR*, implemented polyhedral-model-based loop interchange in LLVM/MLIR, achieving 1.14–1.16× speedup by identifying optimal loop structures.

Research Experience

Performance Architect Intern, **Compute Architecture Group, NVIDIA** — Feb. 2025 – Feb. 2026

Developed **DeepSim**, an AI-assisted GPU performance modeling framework; first-author paper accepted by **SC 2026**.

- Reformulated performance projection as **multi-task learning** over **unit-level speed-of-light (SOL) ratios**, so that DeepSim both projects performance across NVIDIA GPU generations and attributes predicted slowdown to the specific hardware units responsible — unlike black-box regressors that give a number without a reason.
- Curated **290K profiled samples from 284 kernels** across NVIDIA Ampere, Ada, and Hopper, and trained DeepSim to project onto **held-out Blackwell (B200)** at **12.29% MAPE** (**14.78%** on unseen kernels) — accurate projection onto an architecture absent from training, the setting that matters for pre-silicon design-space exploration.
- Built the complete experimental infrastructure, including GPU workload collection, automated profiling, feature extraction, model training, and cross-architecture evaluation.
- Performed GPU workload characterization and performance analysis across HPC and deep-learning workloads to identify architectural bottlenecks and quantify optimization opportunities for future architectures.
- Designed and implemented targeted CUDA microbenchmarks to characterize instruction throughput (IPC) and the memory subsystem (cache hierarchy latency/bandwidth), providing the ground-truth measurements used to calibrate and sanity-check the model.

Graduate Research Assistant, **arcSYS Lab**, Sun Yat-sen University — Sep. 2023 – Present

Authored three works on **GPU software-layer resource management** (two as first author), supported by the General Program of the National Natural Science Foundation of China.

- **Mpache** (ASP-DAC 2025): built an **LLVM**-based compiler pass that chooses, per load instruction, which level of the GPU cache hierarchy to bypass, explicitly modeling L1/L2 interaction rather than treating each level independently — **1.15**× speedup over the default cache policy.

- **CacheC** (Euro-Par 2025): used **LLMs** to reason about inter-kernel cache interference and emit per-kernel cache-management decisions for concurrently executing kernels, yielding a **1.34**× speedup.
- **GoPTX** (DAC 2025, third author): fine-grained kernel fusion that weaves instructions at the **PTX** level to raise instruction-level parallelism, yielding an **11.2%** average speedup.

Publications

Papers [2]–[4] were published under my former name, Mengyue Xi.

[1] DeepSim: SOL-based Explainable GPU Performance Projection with Multi-Task Learning
Mengyue Huang, et al.
Int'l Conf. for High Performance Computing, Networking, Storage and Analysis (**SC**), 2026. *Accepted; to be presented in November 2026.*

[2] CacheC: LLM-based GPU Cache Management to Enhance Kernel Concurrency
Mengyue Xi, et al.
Int'l European Conf. on Parallel and Distributed Computing (**Euro-Par**), 2025. ***[Code]***

[3] Mpache: Interaction Aware Multi-level Cache Bypassing on GPUs
Mengyue Xi, et al.
Asia and South Pacific Design Automation Conf. (**ASP-DAC**), 2025. ***[Code]***

[4] GoPTX: Fine-grained GPU Kernel Fusion by PTX-level Instruction Weaving
Kan Wu, Zejia Lin, Mengyue Xi, et al.
Design Automation Conf. (**DAC**), 2025. ***[Code]***

Selected Projects

Customs Declaration AI Platform — Lab–Industry Collaboration Apr. 2024 – Aug. 2024

- Built an end-to-end prototype combining OCR, retrieval-augmented generation (**RAG**), and open-source LLMs to automate customs declaration document processing.
- Profiled and deployed the serving stack across A100, RTX 3090, and RTX 4090 GPUs, tuning batching and memory footprint to keep accuracy robust under heterogeneous hardware.

Teaching and Academic Service

Teaching Assistant, **Compiler Theory** & **Compiler Construction**, SYSU Feb. 2024 – Aug. 2024

- Developed LLVM-based teaching materials and programming assignments (Flex/Bison and ANTLR) for compiler front-end topics, including lexical analysis, parsing, type checking, and AST construction.

Sub-Reviewer: Int'l Conf. on Networking, Architecture and Storage (NAS), 2024.

Awards and Honors

- **National Scholarship (undergraduate)**, Ministry of Education of China, 2020 — the highest national honor for undergraduate students in China
- **Samsung Scholarship**, Sun Yat-sen University
- **CCF Excellence in Computer Education Teaching Award**, First Prize (team)

Miscellaneous

Class Leader, Computer Science Honors Class (30 students), Chongqing University, 2019–2023 — coordinated class activities and cohort-level administrative affairs.
Member of the English Debate Team, Sun Yat-sen University.
Boxing and swimming enthusiast.