Weili Xu

weilixu2@illinois.edu
weili-0234.github.io
github.com/weili-0234
linkedin.com/in/weili-xu-2a05662a7

306 N Wright St.
Urbana, IL, 61801
Dept. of Electrical & Computer Engineering
University of Illinois Urbana-Champaign

EDUCATION

University of Illinois Urbana-Champaign *Aug. 2023 – Present*
B.S. in Computer Engineering, GPA: 3.97/4.00

Zhejiang University (Dual Degree Program) *Aug. 2023 – Present*
B.S. in Computer Engineering, GPA: 4.20/4.30

Selected Courseworks (all A/A+):
ECE 408 Applied Parallel Programming, ECE 391 Computer Systems Engineering

SKILLS

Programming: Python, C/C++, OpenAI Triton, CUDA C/C++, cuTile, Java, RISC-V Assembly
Frameworks: PyTorch, vLLM, SGLang, SkyRL, DeepSpeed, Flash-Linear-Attention, OpenHands
Tools: Git, Linux/UNIX, slurm, Docker, GDB, PDB, Nsight Systems, Nsight Compute

EXPERIENCE

TileGym (NVIDIA's cuTile Kernel Library) *Dec. 2025 - Present*
Open-Source Code Contributor *Remote*

- Implemented backward-pass kernels for FlashAttention and SwiGLU in cuTile, with numerical correctness test suites and benchmarking harness; achieved >10% higher throughput than optimized Triton baseline.

University of Illinois Urbana-Champaign *Aug. 2025 – Present*
Research Intern in Machine Learning Systems *Urbana, IL*

ThunderAgent: A Fast, Simple, and Robust Program-Aware Agentic Inference System

- Co-first author paper, with 250+ stars on GitHub.
- Proposed a program abstraction that wraps stateless LLM requests into stateful, schedulable programs, enabling lifecycle-aware scheduling of GPU KV caches, tool environments (Docker containers, sandboxes), and cross-node routing for distributed agentic inference and RL rollout across vLLM and SGLang backends.
- Designed a Periodic Thrashing Monitor with shortest-first eviction and exponential time-decay scoring to preemptively detect KV-cache thrashing, and a Global Waiting Queue with load-balanced cross-node restore that eliminates memory imbalance across multi-GPU nodes without sacrificing cache locality.
- Achieved 1.5–3.6× serving throughput and 1.8–3.9× RL rollout throughput on agentic workloads (OpenHands, mini-SWEAgent, ToolOrchestra) on an 8×H100 node. Lifecycle-aware garbage collection yielded 4.2× disk memory savings by automatically reclaiming zombie tool resources upon program termination.

Zhejiang University *Aug. 2023 - Jul. 2025*
Research Intern in Machine Learning *Hangzhou, China*

AuroraLong: Bringing RNNs Back to Efficient Open-Ended Video Understanding (ICCV 2025)

- First author paper that pioneered the use of scalable Linear Attention to unlock hour-long video understanding on edge GPUs with limited VRAM, accelerated inference times by **8.13x** on RTX 3090 GPU.

Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark (ICCVW 2025)

- Developed a benchmark to thoroughly evaluate multimodal reasoning of **90+** video LLMs, earning Oral Presentation and Outstanding Paper Award at Knowledge-Intensive Multimodal Reasoning Workshop.

Zhejiang University-University of Illinois Urbana-Champaign Institute *Aug. 2024 – May 2025*
Undergraduate Teaching Assistant *Hangzhou, China*

- **ECE 220 (Computer Systems & Programming)** *Spring 2025*
- **ECE 120 (Introduction to Computing)** *Fall 2024*