Xiwen MIN

alexismxw@gmail.com +1 (201) 600-2218 Pittsburgh, PA, USA

EDUCATION

New York University Sep. 2023 – Dec. 2025 (expected)
Courant Institute of Mathematical Sciences
M.S. in Computer Science
- GPA | 3.96/4.0
- Selected Courses | Operating Systems(A), Big Data and ML Systems(A), Distributed Systems(A-), Algorithms(A)

Shanghai Jiao Tong University Sep. 2019 – Jun. 2023
School of Electronic Information and Electrical Engineering
B.Eng. *with distinction* in Electronic Instrumentation, minor in Computer Science
- GPA | 87.4/100, top 10%
- Selected Courses | Computer System Architecture(A), Computer Network(A), Algorithms for Big Data(A+)

RESEARCH

Efficient Structured Generation System for LLM Reasoning Jun. 2025 - present
InfiniAI Lab @ Carnegie Mellon University, Advisor: Prof. Beidi Chen
- Developed an application-centric serving engine based on SGLang, accelerated the parallel generation of applications with dynamic DAG structures including split/merge of reasoning branches and Monte-Carlo Tree Search (MCTS).
- Designed and implemented abstraction layer to define applications of general dynamic structures, allowing easy user access to optimized scheduler and KV cache management.
- Integrated with a popular RL framework, Slime, to support GRPO with dynamic structure rollout and customized reward, adapted Megatron-LM for specific attention pattern.

Scalable System for 3D Gaussian Splatting (3DGS) Training Jun. 2024 – May 2024
Courant Systems Group @ New York University, Advisor: Prof. Jinyang Li
- Architected a memory-efficient 3DGS training system for large-scale scene reconstruction with a CPU-GPU offloading strategy that utilizes sparsity of 3DGS rendering.
- Developed custom CUDA kernels to improve training throughput, overlapping GPU computation, CPU computation, and CPU-GPU communication with awareness of data locality.
- Enabled training of up to 102 million points (~6B parameters) on a single RTX4090 GPU with 24 GB memory, allowing 3DGS to scale with limited memory while preserving rendering quality.

Real-time Signal Processing and Platform for Wearable Impedance Pneumography Dec. 2022 – May. 2023
EIT Lab @ Shanghai Jiao Tong University, Advisor: Prof. Yixin Ma
- Engineered a multithreaded data acquisition host system using Bluetooth serial communication, implemented a producer-consumer buffer for low-latency signal visualization.
- Designed signal denoising algorithms to isolate respiratory features from cardiac artifacts, validating system accuracy via linear regression calibration and Standard Error of Measurement (SEM) statistical analysis.

PUBLICATION

- Hexu Zhao*, **Xiwen Min***, Xiaoteng Liu, Moonjun Gong, Yiming Li, Ang Li, Saining Xie, Jinyang Li, Aurojit Panda, "CLM: Removing the GPU Memory Barrier for Gaussian Splatting", *To Appear in ASPLOS 2026*
- Hexu Zhao, Xiaoteng Liu, **Xiwen Min**, Jianhao Huang, Youming Deng, Yanfei Li, Ang Li, Jinyang Li, Aurojit Panda, "Scaling Point-based Differentiable Rendering for Large Scale 3D Reconstruction", *Submitted to SOSP 2025*

WORK EXPERIENCE

Graduate Course Assistant | New York University Sep. 2025 – Dec. 2025
CSCI-GA 3033 - Efficient AI Computing: Algorithm and Implementation by Prof. Saiqian Zhang

Software Engineer Intern | Microsoft Jun. 2022 – Sep. 2022
Microsoft 365 Team

PROJECTS

Multiverse: Your Language Models Secretly Decide How to Parallelize and Merge Generation

- Enabled online serving of the SGLang-based Multiverse Engine that dynamically schedules split/merge of reasoning branches for parallel generation, achieving throughput comparable to auto-regressive serving engines.
- Integrated CUDA graph, mixed decode/prefill batch generation, zero-overhead scheduling to reduce overhead of split/merge operations, improving overall throughput.

DynamoW: An Elixir implementation of Dynamo-style KV Store

- Built a highly available, Dynamo-style distributed KV store in Elixir, implementing consistent hashing, replication, and hinted handoff for fault tolerance and eventual consistency.
- Simulated multi-node behavior using Elixir's actor model to test robustness under partial failures.

Mario: A Vison-based Autonomous Pneumatic Pipe Robot

- Developed a photic features detection algorithm using OpenCV in C++ for gesture control in dark environments inside narrow pipes, deployed on embedded platform and integrated with RTSP service for real-time monitoring.
- Awarded National Secondary Prize (top 10) in National Embedded System Design Competition.

HONERS & AWARDS

Suzanne McIntosh Research Fellowship, New York University, 2025
Outstanding Graduate - Class of 2023, Shanghai Jiao Tong University, 2023
Academic Excellence Scholarship, Shanghai Jiao Tong University, 2019, 2020, 2021

MISCELLANEOUS

Skills

Python, C/C++, CUDA | PyTorch | Hugging Face | NVIDIA Nsight Tools

Personal Experiences

Audio Engineer/Jazz and Rock keyboard amateur, built live audio system for 2k+ audience.
Participated in music and musical productions. Familiar with Cubase, Pro Tools, Gig Performer.