JIAYI QIAN

(+1)404-452-2566 jiayiqian@gatech.edu https://jiayi-19.github.io/

Research Interests: Efficient Compositional AI Systems; Algorithm–System–Hardware Co-design for Agentic, Embodied, and Neuro-Symbolic AI.

EDUCATION

Georgia Institute of Technology, Atlanta, GA *Aug 2025 - Present*
Ph.D. in Electrical and Computer Engineering, College of Engineering
Advisor: Prof. Tushar Krishna

Georgia Institute of Technology, Atlanta, GA *Aug 2023 - May 2025*
Master of Science in Computer Science, College of Computing
Advisors: Prof. Yingyan (Celine) Lin and Prof. Tushar Krishna

Tsinghua University, Beijing, China *Aug 2019 - Jul 2023*
Bachelor of Science in Electronic Information Science and Technology
Department of Electronic Engineering

PUBLICATIONS (* INDICATES EQUAL CONTRIBUTION)

♦ ***Efficient Compositional AI Systems***

Zishen Wan, Hanchen Yang, **Jiayi Qian**, Ritik Raj, Joongun Park, Chenyu Wang, Arijit Raychowdhury, Tushar Krishna, "Compositional AI Beyond LLMs: System Implications of Neuro-Symbolic-Probabilistic Architectures" (**ASPLOS 2026**)

Zishen Wan, Yuhang Du, Mohamed Ibrahim, **Jiayi Qian**, Jason Jabbour, Yang Zhao, Tushar Krishna, Arijit Raychowdhury, Vijay Janapa Reddi, "ReCA: Integrated Acceleration for Real-Time and Efficient Cooperative Embodied Autonomous Agents" (**ASPLOS 2025**)

Jiayi Qian, Zishen Wan, Souvik Kundu, Tushar Krishna, "DyServe: Dynamic Strategy Generation for Agent Serving" (**SCALE @ ICML 2026**, full paper targeting **ATC 2026**)

Jiayi Qian, Zishen Wan, Zheng Du, Hanchen Yang, Ananda Samajdar, Arijit Raychowdhury, Tushar Krishna, "Fast-AIPS: Integrated Acceleration for Efficient Compositional AI Planning Systems" (**ASPLOS 2027 under review**)

Zishen Wan, **Jiayi Qian**, Yuhang Du, Jason Jabbour, Yilun Du, Yang Zhao, Arijit Raychowdhury, Tushar Krishna, Vijay Janapa Reddi, "Generative AI in Embodied Systems: System-Level Analysis of Performance, Efficiency and Scalability" (**ISPASS 2025**)

Hanchen Yang, **Jiayi Qian**, Zishen Wan, Jingtian Dang, Ziwei Li, Yilun Du, Tushar Krishna, "Towards System-2 AI: Workloads and Characterizations of Energy-Based Models" (**ISPASS 2026**)

♦ ***Efficient AI Algorithms***

Yonggan Fu, Zhongzhi Yu*, Junwei Li*, **Jiayi Qian***, Yongan Zhang, Dachuan Shi, Xiangchi Yuan, Roman Yakunin, Yingyan (Celine) Lin, "AmoebaLLM: Constructing Any-Shape Large Language Models for Efficient and Instant Deployment" (**NeurIPS 2024**)

Zhenbang Du*, Yonggan Fu*, Lifu Wang*, **Jiayi Qian**, Xiao Luo, Yingyan (Celine) Lin, "Fewer Denoising Steps or Cheaper Per-Step Inference: Towards Compute-Optimal Diffusion Model Deployment" (**ICCV 2025**)

♦ ***Computer Architecture***

Zishen Wan, Che-Kai Liu, **Jiayi Qian**, Hanchen Yang, Arijit Raychowdhury, Tushar Krishna, "REASON: Accelerating Probabilistic Logical Reasoning for Neuro-Symbolic AI" (**HPCA 2026**)

EXPERIENCE

Research Intern, Synergy Lab, Georgia Institute of Technology *Jun 2024 - May 2025*
Advisor: Prof. Tushar Krishna

- **Embodied AI**: Profiled in system and operator level (NVIDIA Nsight, Intel VTune); drove algorithm-level improvements (LLM quantization, fine-tuning) and system-level optimizations (planner scheduling, dual-memory hierarchy, hierarchical cooperative planning).
- **Neuro-Symbolic AI**: Reviewed existing workloads; conducted system and operator-level profiling (Nsight, VTune); drove algorithm-level optimization (GEMM-based Deductive Database) and system-level optimizations (flexible LLM mapping, pipelined execution).

Research Intern, EIC Lab, Georgia Institute of Technology *Oct 2023 - Jun 2024*
Advisor: Prof. Yingyan (Celine) Lin

- Benchmarked pruning methods for LLMs to quantify latency across GPU servers and edge GPUs using MLC-LLM, vLLM, TensorRT-LLM.
- Designed and implemented a latency measurement method to support the latency estimation of irregular models across frameworks.
- Contributed to the design of model compression technique tailored for efficient deployment and conducted model fine-tuning.

Teaching Assistant: CSE 4140/6140 (Algorithms), Georgia Institute of Technology *Jan 2024 - May 2024*
Advisor: Prof. Xiuwei Zhang

- Designed programming assignments on heuristic algorithms and authored homework/exam materials.
- Held office hours and discussions; graded assignments and exams.

TECHNICAL SKILLS

Programming Languages: Python, PyTorch, C/C++, Verilog, CUDA, Shell, MATLAB
Frameworks and Tools: vLLM, MLC-LLM, TensorRT-LLM, Vivado, Docker, Nsight, VTune