JIAYI QIAN

(+1)404-452-2566 jiayiqian@gatech.edu

EDUCATION

Georgia Institute of Technology, Atlanta, GA *Aug 2023 - May 2025*
Master of Science in Computer Science
College of Computing

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

PUBLICATIONS

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

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

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

RESEARCH EXPERIENCE

Dynamic Quantization Training for Large Language Models (LLMs)

Master Thesis, Georgia Institute of Technology Jul 2024 - Present
Advisor: Prof. Yingyan (Celine) Lin

- Developed a novel dynamic quantization framework for LoRA-based fine-tuning on LLMs.
- Conducted fine-tuning and comprehensive evaluations to verify the effectiveness of proposed framework with various open-source models.
- Performed attention analysis to explore the underlying mechanisms of the proposed approach.

Optimizing Cooperative Embodied AI and Neuro-Symbolic-Probabilistic Systems

Research Intern, Synergy Lab, Georgia Institute of Technology *Aug 2024 - Present*
Advisor: Prof. Tushar Krishna

- Conducted literature reviews and profiling to identify bottlenecks in both Cooperative Embodied AI systems and Neuro-Symbolic-Probabilistic systems.
- Proposed and implemented multiple algorithm-level optimization strategies for Cooperative Embodied AI systems, including multi-step planning and plan-then-comm approaches, to enhance real-time efficiency.
- Contributed to the design of a hierarchical cooperative planning scheme to address scalability challenges in Cooperative Embodied AI systems.
- Replaced closed-source LLMs with open-source alternatives for model-level acceleration and reformulated planning tasks as multiple-choice questions to mitigate performance gaps between closed-source and open-source models.

- Three Submissions under review at **ASPLOS 2025**, **MLSys 2025** and **ISPASS 2025**.

Constructing Any-Shape LLMs for Efficient and Instant Deployment

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

- Conducted profiling of existing pruning methods to evaluate their impact on LLM latency performance across diverse platforms (GPU servers, edge GPUs), using multiple inference frameworks (MLC-LLM, VLLM, TensorRT-LLM, Pytorch)
- Contributed to the design of a novel compression technique tailored for efficient deployment.
- Designed and implemented a latency measurement method to support the latency evaluation of irregular models across the aforementioned frameworks.
- One paper accepted to **NeurIPS 2024**.

Domain-specific Compressed LLMs

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

- Generated domain-specific data with GPT-4 API for model fine-tuning.
- Conducted both LoRA-based and full fine-tuning on Llama models using both self-generated data and open-source datasets to validate the effectiveness of proposed model compression methods.

Trusted Visual Features in Visual SLAM systems

Bachelor's Thesis, Tsinghua University *Oct 2022 - Jun 2023*
Advisor: Prof. Gang Liu

- Integrated deep learning-based feature extraction methods with the ORB-SLAM3 system.
- Developed a SuperPoint-based Visual SLAM system, achieving 20× accuracy improvement over the original ORB-SLAM3 Mono.
- Created a multi-feature fusion Visual SLAM system based on ORB-SLAM3.

WORK EXPERIENCE

Teaching Assistant, Georgia Institute of Technology *Jan 2024 - May 2024*
Course: CSE 4140/6140: Computer Science & Engineering Algorithms
Advisor: Prof. Xiuwei Zhang

- Held office hours and online discussions to assist students.
- Developed exam questions, homework assignments, and projects to assess and enhance student understanding.
- Graded assignments and exams.

SKILLS

Programming Languages: C/C++, Python, Shell, MATLAB, CUDA
Machine Learning Frameworks: PyTorch, TensorFlow, Keras
LLM Inference Frameworks: MLC-LLM, TensorRT-LLM, VLLM