Jiayin Qin

(+1)7634646626 University of Minnesota qin00162@umn.edu

Research Interests

Computer Architecture, AI Hardware Accelerators, LLM-Assisted RTL Design, Chiplet-Based Systems, Efficient Machine Learning Systems, Vision-Language-Action Models, Embodied AI

Education

- **Ph.D. Student in Electrical Engineering**, University of Minnesota Sep 2024 – Present
  GPA: 3.96/4.00
  Anticipated Graduation Date: Jun 2029
- **B.E. in Microelectronic Science and Engineering**, Fudan University Sep 2020 – Jun 2024
  GPA: 3.70/4.00

Experience

- **AdaVLA: Progress-Aware Adaptive Inference for Vision-Language-Action Models**
  Graduate Research Assistant, University of Minnesota May 2026 – Present
  Advisor: Prof. Yang (Katie) Zhao
  - Characterizes VLA models across server-edge platforms to identify system bottlenecks.
  - Implements adaptive VLA inference by dynamically adjusting pruning ratios.
  - Designs a lightweight progress estimator for real-time task progress and reward estimation.
- **PipeSpec-VLA: Pipelined Speculative VLA Inference through Server–Edge Co-deployment**
  Graduate Research Assistant, University of Minnesota May 2026 – Present
  Advisor: Prof. Yang (Katie) Zhao
  - Designs a server–edge speculative inference pipeline with staged action-chunk verification to overlap VLM prefill, draft generation, and communication, improving latency and accuracy.
  - Explores top-K draft trajectories to improve action acceptance and reduce inference latency while preserving task success.
- **VLSI II (EE 5324)**
  Teaching Assistant, University of Minnesota Spring 2026
  - Designed a Verilog-based FlashAttention accelerator project covering systolic-array architecture, RTL implementation, synthesis, and power/performance optimization.
  - Provided technical support for VLSI design, synthesis, analysis and layout with VCS, Design Compiler, IC Compiler II, PrimeTime, and Cadence Virtuoso with the TSMC 16 nm PDK.
- **A Flexible and Scalable Chiplet-Based Simulator for Efficient LLM Serving**
  Research Intern, Georgia Institute of Technology Jul 2025 – Sep 2025
  Advisor: Prof. Yingyan (Celine) Lin
  - Develops a flexible chiplet-based simulation framework for evaluating LLM serving systems.

Publications (* = Co-first Authors)

- RTGS: Real-Time 3D Gaussian Splatting SLAM via Multi-Level Redundancy Reduction
  Leshu Li*, **Jiayin Qin***, Jie Peng, Zishen Wan, Huaizhi Qu, Ye Han, Pingqing Zheng, Hongsen Zhang, Yu (Kevin) Cao, Tianlong Chen, Yang (Katie) Zhao
  International Symposium on Microarchitecture (MICRO), 2025, Acceptance rate: 20.8%, [Paper].
  Leads the RTGS hardware architecture design, including units for imbalanced workload scheduling, rendering data reuse, and gradient aggregation to reduce multi-level redundancies.

- Mozart: Modularized and Efficient MoE Training on 3.5D Wafer-Scale Chiplet Architectures
  Shuqing Luo*, Ye Han*, Pingzhi Li*, **Jiayin Qin***, Jie Peng, Yang Katie Zhao, Yu Cao, Tianlong Chen
  Annual Conference on Neural Information Processing Systems (NeurIPS), 2025, Spotlight, Acceptance rate: 3.5%.

  Leads the Mozart hardware architecture design, including 3D logic-on-memory chiplets, hierarchical memory, and a tree-based NoP to reduce compute and communication overhead for MoE training.

- MAHL: Multi-Agent LLM-Guided Hierarchical Chiplet Design with Adaptive Debugging
  Jinwei Tang*, **Jiayin Qin***, Nuo Xu, Pragnya Sudershan Nalla, Yu Cao, Yang (Katie) Zhao, Caiwen Ding
  International Conference on Computer Aided Design (ICCAD), 2025, Acceptance rate: 24.7%, [Paper].

  Leads the structure design of MAHL chiplet generation framework and builds the hierarchical design decomposition engine and hardware design-space exploration engine.

- HiVeGen – Hierarchical LLM-based Verilog Generation for Scalable Chip Design
  Jinwei Tang*, **Jiayin Qin***, Kiran Thorat, Chen Zhu-Tian, Yu Cao, Yang (Katie)Zhao, Caiwen Ding
  International Conference on LLM-Aided Design (ICLAD), 2025, Best Paper Award, Acceptance rate: 34.0%, [Paper].

  Leads the design of the Hierarchy-aware Prompt Generation Engine to enable scalable HDL generation through hierarchical module decomposition.

- CGRA-HD: An Efficient Reconfigurable Accelerator for Hyperdimensional Computing
  Jiayin Qin*, Yuan Dai*, Lingli Wang
  International Conference on Field Programmable Technology (FPT), 2024, [Paper].

  Leads the CGRA-HD architecture design, integrating specialized HDC-oriented PEs with general-purpose PEs to improve performance and hardware efficiency.

- HDLxGraph: Bridging Large Language Models and HDL Repositories via HDL Graph Databases
  Pingqing Zheng, **Jiayin Qin**†, Fuqi Zhang, Niraj Chitla, Zishen Wan, Shang Wu, Yu Cao, Caiwen Ding and Yang Zhao [† = Presenter]
  Asia and South Pacific Design Automation Conference (ASP-DAC), 2026, Acceptance rate: 29.9%, [Paper].

- LACE: Large Language Model Aided Multi-Agent Framework for Agile RISC-V Instruction Extension
  Pingqing Zheng, **Jiayin Qin**†, Fuqi Zhang, Zishen Wan, Shang Wu, Yu Cao, Caiwen Ding and Yang Zhao [† = Presenter]
  International Conference on LLM-Aided Design (ICLAD), 2026, Acceptance rate: 30.9%, [Paper].

Awards

- Outstanding Graduates at Fudan University (2024)

Skills

- **Programming Languages:** C/C++, Python, Verilog, Assembly Language, MATLAB
- **Tools:** SPICE, COMSOL Multiphysics, Multisim, Icarus Verilog, Vivado, Vitis HLS, Synopsys Design Compiler, Synopsys IC Compiler II

Concepts

Hardware Engineer, Robotics, ASIC, SoC, AI Agent, Verification, VLSI