Jiahan Xie
Phone: (607)280-7683
Email: jxie84@ucsc.edu
Website: https://jiahanxie353.github.io/

Education

- **University of California, Santa Cruz** — Santa Cruz, CA
  PhD student in Computer Science — *Expected 2030*
- **Cornell University** — Ithaca, NY
  Master of Science in Computer Science — *Aug 2023 - May 2025*
- **Cornell University** — Ithaca, NY
  Bachelor of Science in Computer Science and Environmental Engineering — ***Magna Cum Laude***

Research Experience

- **Graduate Researcher** — Sep 2025 - Now
  UC Santa Cruz *MASC* Group, led by José Renau
  - Building an AI agent system for hardware design - RTL code generation, verification, debugging, and tapeout; and evaluate the results in Docker environments
- **Graduate Researcher** — Sep 2023 - May 2025
  Cornell *Capra* Group, led by Adrian Sampson
  - Built and optimized an open-source PyTorch-to-FPGA compiler flow via CIRCT, achieving 1.7× performance over AMD Vitis using advanced memory partition techniques
- **Undergraduate Researcher** — Nov 2021 - May 2023
  Cornell *PEESE* Group, led by Fengqi You
  - Built and evaluated an attention-based multi-agent deep reinforcement learning model for smart-city energy management using OpenAI Gym simulation and real-world data.

Publications and Talks

Xie, J., Williams, E., Sampson, A. "From PyTorch to Calyx: An Open-Source Compiler Toolchain for ML Accelerators." Present in 7th C4ML workshop, CGO 2026 and 8th AccML workshop, HiPEAC 2026. Paper Link

"Global Instruction Selection for Scalable Vectors." Students Technical Talk at LLVM Developer Conference 2024.

Xie, J., Ajagekar, A., You, F. "Multi-Agent Attention-Based Deep Reinforcement Learning for Demand Response in Grid-Responsive Buildings." *Applied Energy 2023.* Paper Link

Internship Experience

- **Machine Learning Compiler Engineer** — Jun 2025 - Sep 2025
  Cerebras Systems — *Sunnyvale, CA*
  - Designed and implemented memory analysis algorithms to optimize on-chip/off-chip data movement for ML workloads, including weight and activation loading/eviction. Built profiling and visualization tools using Perfetto to trace data movement behavior back to source code for performance analysis
  - Built a two-pass compile workflow to reduce wafer layout selection optimization process for large MoE models
- **Software Engineer** — Jun 2023 - Aug 2023
  Orenda Power Inc. — *New York City, NY*
  - Built a power grid database with real-time processing and designed a REST API for IoT control operations
- **Machine Learning Research Engineer** — May 2022 - Aug 2022
  Deep Ivy Inc. — *Remote*
  - Developed framework converters and a graph compiler to accelerate ML model transformations

Project Experience

- **Systolic Array Design for Binarized Matrix Multiplication (Dataflow architecture)** Oct 2024 - Nov 2024
  - Designed and optimized an FPGA systolic array for binarized matmul using Allo, achieving a 15.2× speedup
- **Global Instruction Selection for RISC-V Vector Extension (LLVM)** Nov 2023 - Oct 2024
  - Extended GISel for RISC-V vectors, enabling scalable vector support for SAXPY lowering from C to assembly
- **Compiler Development and Optimization for Bril (Advanced Compilers, C++)** Aug 2023 - Dec 2023
  - Built a Bril backend with compiler optimizations (LVN, DCE, LICM), achieving a 10.7% benchmark speedup
- **Pipelined RISC-V Processor with Cache (Computer Architecture, RTL Design)** Sep 2023 - Nov 2023
  - Implemented pipelined RISC-V processors with stalling, bypassing, and instruction/data caches

Coding Languages

- C++, Python, C, JAVA, SystemVerilog