

# Kevin He

+1 908-300-2878 | [kevinhe@g.harvard.edu](mailto:kevinhe@g.harvard.edu) | [www.linkedin.com/in/kevin-j-he/](https://www.linkedin.com/in/kevin-j-he/)

## EDUCATION

### Harvard University

*Ph.D. Student, Computer Science*

September 2025 – Present

### University of California, Berkeley

*Bachelor of Arts, Computer Science*

August 2021 – December 2024

*GPA: 3.973/4.0*

## RESEARCH

I am broadly interested in computer architecture and ML systems, with an emphasis on improving the performance and energy efficiency of modern generative AI workloads. My research has focused on hardware–software co-design for efficient machine learning, including designing novel ML accelerator architectures and edge inference serving. I'm also interested in compilers and datacenter-scale optimizations.

### Harvard Architecture, Circuits and Compilers, advised by Prof. David Brooks

*September 2025 – Present*

- Writing custom ISA kernels to enable LLM inference on a novel distributed multi-chiplet transformer accelerator
- Profiling power and performance on end-to-end LLM serving pipelines for resource-constrained edge devices

### Berkeley SLICE Lab, advised by Prof. John Wawrzynek and Prof. Krste Asanović

*February 2023 – January 2025*

- **uArchDB**: Developed an extensible graph-based microarchitecture event logger for debugging open-source processors and accelerators. [Poster](#) / [Slides](#) / [Github](#)
- **GNNs for Circuit Power**: Researched graph neural networks (GNNs) for predicting digital circuit power
- **DiffSampler**: Investigated gradient descent for GPU accelerated circuit SAT sampling for design verification
- **RISC-V Vector Core Tapeout**: Taped out a 2x2mm SoC with a multicore RVV1.0 vector processor with DSP and ML accelerators on Intel's 16nm FinFET process. Wrote efficient RISC-V quantized matmul vector kernels for benchmarking and design space exploration. [Poster](#) / [Slides](#)

## PUBLICATIONS

### High-Throughput SAT Sampling

Arash Ardakani, Minwoo Kang, **Kevin He**, Qijing Huang, John Wawrzynek

Design, Automation and Test in Europe Conference 2025 (*DATE 2025*)

### DEMOTIC: A Differentiable Sampler for Multi-Level Digital Circuits

Arash Ardakani, Minwoo Kang, **Kevin He**, Qijing Huang, Vighnesh Iyer, Suhong Moon, John Wawrzynek

Asia and South Pacific Design Automation Conference 2025 (*ASP-DAC 2025*)

### Late Breaking Results: Differential and Massively Parallel Sampling of SAT Formulas

Arash Ardakani, Minwoo Kang, **Kevin He**, Vighnesh Iyer, Suhong Moon, John Wawrzynek

Design Automation Conference 2024 (*DAC2024*)

## WORK EXPERIENCE

### Apple

Cupertino, CA

*CPU Performance Intern*

*May 2024 – August 2024*

- Annotated major microarchitecture structures and events in Apple's CPU branch predictors for RTL vs. C++ performance model correlation
- Built new Verilog and performance verification infrastructure for mapping branch predict and train events
- Developed Python tools to parse microarchitecture event logs and quickly extract predictor analytics and identify performance anomalies, significantly reducing initial debugging time to identify issues in waveform dumps

### Apple

Austin, TX

*GPU RTL Design Intern*

*May 2023 – August 2023*

- Implemented RTL bug fixes in Apple's GPU design library and analyzed performance & area impacts of fixes
- Designed and formally verified a new, reusable, parameterized, low-area SRAM FIFO module for the IP library

## TEACHING

### EECS151: Intro to Digital Design Teaching Assistant

*January 2024 – May 2024*