ZHEYUAN CHEN

+1 2017797262
sephirotheca17@gmail.com

EDUCATION

University of California, Santa Cruz *Sept 2024 – Present*
Ph.D., Computer Science and Engineering

RESEARCH & WORK EXPERIENCE

Microsoft Research Jun 2025 – Sep 2025
Research Intern, RiSE Group *Redmond, WA*

- Developed **SIMT-Step**, a formal operational semantics for GPU warp execution to reason about correctness and portability of performance-critical GPU kernels across architectures.
- Built a Clang AST-based interpreter and fuzzing framework for HLSL kernels to automatically generate warp-execution tests.
- Discovered and reported numerous warp-execution bugs under non-uniform control flow across NVIDIA, Intel, and AMD GPUs, exposing portability issues in kernel execution and compiler behavior.

Mercedes-Benz Research & Development North America Jun 2024 – Dec 2024
Software Engineering Intern *Sunnyvale, CA*

- Contributed to middleware for automated driving systems (ADS), improving communication, data flow, and resource sharing across safety-critical components.
- Developed an automated toolchain to migrate a large Bazel project of CUDA/C++ code to SYCL, reducing manual effort by 90% and improving kernel portability across heterogeneous hardware.

Languages, Systems, and Data (LSD) Lab Jan 2023 – Present
Research Assistant *University of California, Santa Cruz*

- Developed a formal operational semantics for GPU warp execution to provide rigorous foundations for portable GPU and ML kernel reasoning, and verified key behaviors using TLA+.
- Helped launch the **WebGPU** backend in llama.cpp, with primary contributions to the portable **FlashAttention** kernel in **WGSL**.

PUBLICATIONS

Zheyuan Chen, Naomi Rehman, Guido Martínez, and Tyler Sorensen. "SIMT-Step Execution: A Flexible Operational Semantics for GPU Subgroup Behavior." *PLDI 2026* (accepted).

Yanwen Xu, Rithik Sharma, **Zheyuan Chen**, Shaan Mistry, and Tyler Sorensen. "BetterTogether: An Interference-Aware Framework for Fine-grained Software Pipelining on Heterogeneous SoCs." *IISWC 2025*. **Best Paper Award**.

SKILLS

Languages: Rust, C/C++, Python, Go, WGSL, HLSL, GLSL, Metal, SPIR-V, TLA+

Frameworks & Systems: CUDA, WebGPU, SYCL, Vulkan, HIP, OpenCL, LLVM, MLIR