Yichong Zhang

Tsinghua University, Beijing, China | yichong-23@mails.tsinghua.edu.cn

Personal website

Education

Tsinghua University Beijing, China

- Undergraduate in Mathematics and Physics + Electrical Engineering Aug 2023 – Jul 2027 (expected)
- GPA: 3.9/4.0, Rank: 3/52

Harvard College, Harvard University Cambridge, MA, United States

- Visiting Undergraduate Student (VUS) Aug 2025 – May 2026

Research Interests

My research aims to enable agent-driven cross-stack optimization for building efficient and reliable AI systems. My current work spans two complementary directions:

- **Systems and architecture for emerging AI workloads**, spanning GPU kernel optimization, runtime systems, and efficient deployment of embodied AI.
- **Reliable code generation and optimization with coding agents**, with a focus on testing and verification for agent generated code.

Publications

Yichong Zhang, Genghan Zhang, Zishen Wan, Kunle Olukotun, Olivia Hsu, "*Behavioral Test Diversity: Quantifying Test Suite Diversity Through Agentic Code Repair*," Under review at ICLR 2027

Jiayi Qian, **Yichong Zhang**, Hanchen Yang, Chun Tao, Souvik Kundu, Zishen Wan, Tushar Krishna, "*Serving Agentic Workflows with a Physical-Plan Compiler and Adaptive Runtime*," Under review at ASPLOS 2027

Ke Hong, Xiuhong Li, Minxu Liu, Qiuli Mao, Tianqi Wu, Zixiao Huang, Lufang Chen, Zhong Wang, **Yichong Zhang**, Zhenhua Zhu, Guohao Dai, Yu Wang, "*Efficient and Adaptable Overlapping for Computation and Communication via Signaling and Reordering*," accepted by EuroSys 2026

Tianchen Zhao, Ke Hong, Xinhao Yang, Xuefeng Xiao, Huixia Li, Feng Ling, Ruiqi Xie, Siqi Chen, Hongyu Zhu, **Yichong Zhang**, Yu Wang, "*PAROAttention: Pattern-Aware ReOrdering for Efficient Sparse and Quantized Attention in Visual Generation Models*," accepted by NeurIPS 2025

Research Experience

Tests as Specifications for Reliable Code Generation May 2026 – Present

Advisors: Profs. Kunle Olukotun & Olivia Hsu, Department of Computer Science, Stanford University

- Investigated behavioral redundancy in test suites for coding agents, identifying limitations of conventional metrics such as code coverage and mutation score in capturing the diversity of behavioral constraints specified by tests.
- Developed Behavioral Test Diversity (BTD), an agent-conditioned metric that quantifies test-suite diversity by measuring how often code repairs guided by one test also satisfy other tests.
- Applied BTD to guide test-suite pruning and expansion, reducing tests supplied as executable specifications by 61–90% across ProgramBench, vLLM, and ComputeEval while preserving full suite recovery, and achieving 3.2× the behavioral diversity gain of a flat baseline in test suite expansion.

Benchmarking and Characterizing Robotic Manipulation Workloads Feb 2026 – Present

Advisors: Profs. Zishen Wan & Vijay Janapa Reddi, Harvard SEAS, Harvard University

- Developed EmbodiedPerf, a lightweight profiling framework for robotic manipulation benchmarks, enabling measurements of latency, energy consumption, and hardware utilization alongside task success.

- Benchmarked robotic manipulation policies (VLA, WAM, and Code-as-Policy) using EmbodiedPerf across LIBERO, RoboTwin, and BEHAVIOR-1K on H100, RTX 5090, and Jetson Thor and Orin, revealing that smaller model sizes do not necessarily translate into better task-level efficiency and that performance bottlenecks vary across hardware.
- Evaluated different power-management configurations (including power-budget modes and DVFS) on Jetson platforms and asynchronous inference across models, finding that inference-level speedups do not necessarily translate into task-level efficiency gains.
- Deployed and evaluated the profiling framework on a Franka robot arm performing a real-world pick-and-place task, with π0.5 inference running on an RTX 5090 GPU.

Workflow-Aware Serving for Agentic AI Applications Jul 2026 – Sep 2026

Advisor: Prof. Zishen Wan, Department of Computer Science, Columbia University

- Refined the design of Dyserve's physical-plan compiler and adaptive runtime, incorporating pair-specific model–verification profiling, quality-aware plan switching, and staged preparation of pressure-specialized plans.
- Devised evaluation strategies, implemented benchmark experiments, and analyzed accuracy–latency trade-offs and adaptive serving effectiveness across agentic workloads.
- Synthesized key findings to shape the paper's technical narrative, contributed to manuscript writing, and designed and produced its main figures.

Efficiency and Robustness of Vision-Language-Action Models Sep 2025 – Feb 2026

Advisors: Profs. Vijay Janapa Reddi & Zishen Wan, Harvard SEAS, Harvard University

- Evaluated cache reuse and 4- and 8-bit weight-only quantization in VLA models, finding no apparent degradation in task performance or robustness; further characterized aggressive 3- and 2-bit quantization in OpenVLA and OpenVLA-OFT language backbones, observing 0.9% and 16% performance degradation, respectively, on LIBERO.
- Implemented and benchmarked SVD- and RRQR-based low-rank approximations of VLA language backbones, finding that SVD consistently outperformed RRQR across the evaluated ranks.

GPU Systems and Kernel Optimization Jan 2025 – Aug 2025

Advisor: Prof. Yu Wang, Department of Electronic Engineering, Tsinghua University

- Implemented fused CUDA reorder kernels for LayerNorm and RoPE in PAROAttention, a hardware–software co-designed sparse-attention system for video-generation transformers, achieving ~20× speedup over PyTorch; built benchmarking pipelines for kernel-level comparisons against FlashAttention-2, XAttention, and SVG.
- Contributed to a signaling-based computation–communication overlap mechanism for GPU workloads and benchmarked GEMM–collective workloads, demonstrating up to 1.65× speedup over non-overlapped execution on RTX 4090 GPUs.

Extracurricular Activities

"A GreatPerhaps" Career Talk Project Beijing, China Mar 2024 – Jan 2025

Founder

- Founded and led a cross-university team spanning Tsinghua, Peking, Shanghai Jiao Tong, Fudan, and the University of Washington, producing career-focused interviews and content with ~100,000 cumulative views.
- Conducted one-on-one interviews with senior technology executives, including CTOs from Megvii Technology, Pudu Robotics, and TeleAI.
- Project website: https://grephap.org

Skills

Reviewer: ICLR 2027, NeurIPS 2026 Workshop (Verify-Agents Workshop)

Language: Chinese (Native); English (Fluent, with a TOEFL score of 106)

Programming and Software: Python, CUDA, C++, Bash, Git, LaTeX