

# Genghan Zhang

zhang677.github.io | ghzhang19@gmail.com

## EDUCATION

### Stanford University

*PhD Student in Computer Science*

- Research Interests: Domain-specific compiler and computer architecture

September 2023 - Present  
*Stanford, USA*

### Tsinghua University

*Bachelor of Engineer in Electronic Information Science and Technology*

- GPA: 3.94/4.00 (Top 3%)

August 2019 - June 2023  
*Beijing, China*

## RESEARCH EXPERIENCE

### Research Assistant

*Department of Computer Science, Stanford University*

- Advisor: Prof. Kunle Olukotun
- Working on using LLM agents to automate high-performance library development.

April 2024 - Present  
*Stanford, CA*

### Research Assistant

*Department of Computer Science, Stanford University*

- Advisor: Prof. Azalia Mirhoseini
- Proposed GPU kernel fusion techniques to accelerate FFN layers for LLM inference by utilizing the sparsity of activation. Accepted by COLM 2024

January 2024 - March 2024  
*Stanford, CA*

### Research Assistant

*Department of Computer Science, Stanford University*

- Advisor: Prof. Fredrik Kjolstad
- Designed an algorithm template and code generation algorithm for *sparse workspace* to solve the sparse scattering problem with a sparse tensor algebra compiler called TACO. Accepted by PLDI 2024.

March 2022 - December 2023  
*Remote*

### Undergraduate Research Assistant

*Nanoscale Integrated Circuits and Systems Lab (NICS), Tsinghua University*

- Advisors: Prof. Yu Wang, Prof. Guohao Dai (SJTU) and Prof. Sitao Huang (UC Irvine)
- Proposed *atomic parallelism*, a new optimization space for sparse-dense hybrid algebra and *segment group*, a new abstraction for sparse compilation theory based on atomic parallelism. Accepted by CCF Transactions on High Performance Computing.

August 2021 - July 2022  
*Beijing, China*

## SELECTED PUBLICATIONS

### • Adaptive Self-improvement LLM Agentic System for ML Library Development

Genghan Zhang, Weixin Liang, Olivia Hsu, Kunle Olukotun.

*International Conference on Machine Learning (ICML), 2025*

### • Compilation of Modular and General Sparse Workspaces

Genghan Zhang, Olivia Hsu, Fredrik Kjolstad.

*Programming Language Design and Implementation (PLDI), 2024*

### • CATS: Context-Aware Thresholding for Sparsity in Large Language Models

Donghyun Lee, Jaeyong Lee, Genghan Zhang, Mo Tiwari, Azalia Mirhoseini.

*Conference on Language Modeling (COLM), 2024*

### • Sgap: Towards Efficient Sparse Tensor Algebra Compilation for GPU

Genghan Zhang, Yuetong Zhao, Yanting Tao, Zhongming Yu, Guohao Dai, Sitao Huang, Yuan Wen, Pavlos Petoumenos, Yu Wang.

*CCF Transactions on High Performance Computing, 2023*

## SERVICE

- Reviewer: ICML 2025, ICLR 2025 (Notable Reviewer), NeurIPS 2024, ICLR 2025 DL4C Workshop, NeurIPS 2024 Sys2-Reasoning Workshop, NeurIPS 2022 GLFrontiers Workshop

- Artifact Evaluation Committee: ASPLOS 2025 summer, PLDI 2025

- Program Committee: LATTE 2025

## WORK EXPERIENCE

### Software Engineer

*NVIDIA*

- Mentor: Andrew Kerr
- Compiler for Tile IR

June 2024 - September 2024  
*Santa Clara, USA*

### Part-time Research Assistant

*Infinigence Tech*

- Mentor: Prof. Xiuhong Li (PKU)
- Assembled an in-house GPU kernel library for LLM inference which demos the company's first-generation product.

May 2023 - July 2023  
*Beijing, China*

### Part-time Research Assistant

*HPC-AI Tech*

- Mentor: Prof. Yang You (NUS)
- Developed novel automatic parallelization techniques for gaint deep learning models.

October 2022 - November 2022  
*Beijing, China*

## HONORS AND AWARDS

### Awards in Tsinghua University

- Academic Excellence Award 2020, 2021, Excellence Award 2022