Yicheng Zhang

Tsinghua University, Beijing, China
zhangyic23@mails.tsinghua.edu.cn · github.com/zycccishere

Research Interests

Efficient Model Architecture (Sparse Attention, Quantization), **Machine Learning Systems** (CUDA/Triton Optimization), and **Recursive Self Improvement**.

Education

Tsinghua University **2023.08 – Present**
B.Eng. in Computer Science and Technology *Beijing, China*

- **GPA: 3.96 / 4.0** (Rank: **6 / 173**)

Publications

(* denotes equal contribution; underline indicates myself)

[1] **MLS-Bench: A Holistic and Rigorous Assessment of AI Systems on Building Better AI.**
Bohan Lyu*, Yucheng Yang*, Siqiao Huang*, Jiaru Zhang*, Qixin Xu*, Yicheng Zhang*, et al.
ICML 2026 AI for Math Workshop Honorable Mention Award, 2026.05. [Project Page]

- A comprehensive benchmark evaluating LLM agents on real ML research tasks.
- Led the design and implementation of tasks in the **ML Systems & Efficient ML** section.
- Officially adopted by Kimi-K2.7-Code, Kimi-K3 and Qwen3.8-Max.

[2] **Second-Look Prompting Improves Vision-Language Models.**
Yicheng Zhang, Xun Zhang, Jingyuan Ma, Tianyu Yu, Yuan Yao.
Preprint, 2026. [Project Page]

- Introduced **Second-Look Prompting (SLP)**, a training-free method that repeats the image–question sequence to enable question-aware visual processing.
- Improved overall average accuracy for **all 12 evaluated VLMs** across 12 reasoning and perception benchmarks, with gains of up to 1.7 percentage points.

[3] **A Survey of Efficient Attention Methods: Hardware-efficient, Sparse, Compact, and Linear Attention.**
Jintao Zhang, Rundong Su, Chunyu Liu, Jia Wei, ..., Yicheng Zhang, et al.
Under review at IEEE TPAMI, 2025.11. [Project Page]

- A systematic taxonomy of efficient attention from four complementary perspectives: hardware-efficiency, sparsity, compactness, and linearity.
- Contributed the **Sparse Attention** section.

Honors & Awards

National Scholarship **2024**
Ministry of Education of China
- **Top 0.2%** of undergraduates in China

Tsinghua Student Supercomputing Challenge (THUSCC) **Dec 2025**
Champion (1st Place)
- Ranked **1st out of 48 teams**. Responsible for the **GPU Optimization** track.
- Implemented high-performance attention kernels using **CUDA**.

Comprehensive Scholarship **2025**
Computer Science and Technology, Tsinghua

Leadership & Activities

Class President **2024 – 2025**
Class 34, Department of Computer Science

Student Welfare Representative **2023 – 2024**
Class 34, Department of Computer Science

Student Association for Science and Technology **2024 – 2026**
Member, Intelligent Agents Division

Tsinghua University Drama Troupe **2024 – Present**
Member