

Yike Zhang

📍 Beijing, China ✉ zhangyik21@mails.tsinghua.edu.cn ☎ 152 1078 6258 🌐 zerollin

Education

- BS** **Tsinghua University**, Computer Science and Technology 9/2021 - 6/2025
- GPA: **3.95/4.00**
 - **Coursework:** Linear Algebra(**A**), Data Structures(**A**), Probability and Statistics(**A+**), Artificial Neural Networks(**A**), Database Special Topic Training(**A+**)
 - **Honors:** Outstanding Graduate & Outstanding Bachelor Thesis
- PhD** **Tsinghua University**, Computer Science and Technology 9/2025 - Present
- **Supervisor:** [Jianyong Wang](#)
 - **Research Topic:** Efficient AI/ML

Experience

- Microsoft Research Asia**, Research Intern 12/2024 - 3/2026
- Mentored by [Zhiyuan He](#), [Huiqiang Jiang](#) and [Yuqing Yang](#)
 - Worked on the efficient inference of LLMs and efficient Reinforcement Learning algorithm design
 - Received MSRA "Stars of Tomorrow" award

Publications

LeanK: Learnable K Cache Channel Pruning for Efficient Decoding (EMNLP 2025) [Paper](#), [Code](#)

Yike Zhang, Zhiyuan He, Huiqiang Jiang, Chengruidong Zhang, Yuqing Yang, Jianyong Wang, Lili Qiu

- Explored the static sparsity of K-cache and developed a stabilized training pipeline to obtain compression masks.
- Implemented an efficient TileLang-based decoding kernel for acceleration.

SpinningRL: Efficient RLVR via Theoretical Reward Gain Modeling and Adaptive UCB Sampling (In submission)

Yike Zhang, Zhiyuan He, Bingxiang He, Xufang Luo, Jiahang Xu, Yuqing Yang, Jianyong Wang

- Improve RLVR efficiency through sample selection and scheduling.

Efficient Attention Mechanisms for Large Language Models: A Survey (Cell Patterns 2026) [Paper](#)

Yutao Sun[†], Zhenyu Li[†], **Yike Zhang**[†], Tengyu Pan[†], Bowen Dong[†], Yuyi Guo, Jianyong Wang ([†] denotes co-first author)

- Summarized and categorized efficient attention methods, including linear, sparse, and hybrid architectures.
- Analyzed the co-design of algorithmic innovations and hardware considerations.

FocusLLM: Precise Understanding of Long Context by Dynamic Condensing (ACL 2025) [Paper](#)

Zhenyu Li, **Yike Zhang**, Tengyu Pan, Yutao Sun, Zhichao Duan, Junjie Fang, Jianyong Wang

- Designed a novel framework to enable efficient long context processing of LLMs through memory condensing.

TriangleMix: Accelerating Prefilling via Decoding-time Contribution Sparsity (ACL 2026 Findings) [Paper](#)

Zhiyuan He, **Yike Zhang**, Chengruidong Zhang, Huiqiang Jiang, Yuqing Yang, Lili Qiu

- Identified an unexplored sparsity in the prefilling stage, which could be used as static sparse pattern for acceleration.

VL Norm: Rethink Loss Aggregation in RLVR [Paper](#)

Zhiyuan He, Xufang Luo, **Yike Zhang**, Yuqing Yang, Lili Qiu

- Proposed an unbiased and variance-minimized gradient normalization that stabilizes RLVR training.