Zhiteng Li

Feb 19, 2026

CONTACT INFORMATION

School of Computer Science, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, China
Homepage: https://zhitengli.github.io/

ieeezhitengli@gmail.com
Tel: +86-134-140-18498
Google Scholar, Github

RESEARCH INTERESTS

LLM pretraining and scaling; Image/Video generation;
Model compression and inference acceleration (e.g., Quantization); Data synthesis.

EDUCATION

Shanghai Jiao Tong University, Shanghai, China — Sep 2023 – Present

M.S. Student, Computer Science and Technology — Advisor: *Prof.* Yulun Zhang

- GPA: 3.94/4.0 (Top 5%)

Shanghai Jiao Tong University, Shanghai, China — Sep 2019 – Jun 2023

B.S., Computer Science and Technology (IEEE Honor Class)

- GPA: 4.0/4.3 (Rank 1/113)

PROFESSIONAL EXPERIENCE

ByteDance Seed, Beijing, China — May 2025 – Present
Research Intern — Mentors: Huizhuo Yuan, *Prof.* Quanquan Gu
Project: LLM pretraining with MuP & Scaling Law.

Xiaohongshu, Shanghai, China — Jan 2025 – Apr 2025
Research Intern — Mentor: Shuang Sun
Project: Video DiT acceleration (applied our QuantCache to real-world business scenarios).

Sony AI, Remote — Dec 2023 – May 2024
Research Intern — Mentor: Lele Chen
Project: Generative Data Augmentation (GenDataAgent is accepted at **ICLR 2025**).

Amazon Shanghai AI Lab (ASAILab), Shanghai, China — Jun 2022 – Sep 2023
Applied Scientist Intern — Mentors: Mufei Li, Minjie Wang
Project: Contributed to **DGL** & Graph Transformer research.

PUBLICATIONS

5×ICLR, 1×ICML, 1×ICCV, 8×Preprints. ⋆ indicates equal contribution.

1. **[ICLR 2025] ARB-LLM: Alternating Refined Binarizations for Large Language Models**.
   Zhiteng Li⋆, Xianglong Yan⋆, Tianao Zhang, Haotong Qin, Dong Xie, Jiang Tian, Zhongchao Shi, Linghe Kong, Yulun Zhang, and Xiaokang Yang.
2. **[ICLR 2025] GenDataAgent: On-the-fly Dataset Augmentation with Synthetic Data**.
   Zhiteng Li, Lele Chen, Jerone Andrews, Yunhao Ba, Yulun Zhang, and Alice Xiang.
3. **[ICLR 2026] DVD-Quant: Data-free Video Diffusion Transformers Quantization**.
   Zhiteng Li⋆, Hanxuan Li⋆, Junyi Wu, Kai Liu, Haotong Qin, Linghe Kong, Guihai Chen, Yulun Zhang, and Xiaokang Yang.
4. **[ICCV 2025] QuantCache: Adaptive Importance-Guided Quantization with Hierarchical Latent and Layer Caching for Video Generation**.
   Junyi Wu⋆, **Zhiteng Li**⋆, Zheng Hui, Linghe Kong, Yulun Zhang, and Xiaokang Yang.
5. **[ICLR 2026] Quant-dLLM: Post-Training Extreme Low-Bit Quantization for Diffusion Large Language Models**.
   Tianao Zhang⋆, **Zhiteng Li**⋆, Xianglong Yan, Haotong Qin, Yong Guo, and Yulun Zhang.
6. **[ICLR 2026] PT^2-LLM: Post-Training Ternarization for Large Language Models**.
   Xianglong Yan⋆, Chengzhu Bao⋆, **Zhiteng Li**, Tianao Zhang, Kaicheng Yang, Haotong Qin, Ruobing Xie, Xingwu Sun, and Yulun Zhang.

7. **[ICML 2025] BiMaCoSR: Binary One-Step Diffusion Model Leveraging Flexible Matrix Compression for Real Super-Resolution**.
Kai Liu, Kaicheng Yang, Zheng Chen, **Zhiteng Li**, Yong Guo, Wenbo Li, Linghe Kong, and Yulun Zhang.

Preprints

1. **AdaSVD: Adaptive Singular Value Decomposition for Large Language Models**.
Zhiteng Li*, Mingyuan Xia*, Jingyuan Zhang, Zheng Hui, Haotong Qin, Linghe Kong, Yulun Zhang, and Xiaokang Yang.
2. **BinaryHPE: 3D Human Pose and Shape Estimation via Binarization**.
Zhiteng Li, Yulun Zhang, Jing Lin, Haotong Qin, Jinjin Gu, Xin Yuan, Linghe Kong, and Xiaokang Yang.
3. **VEQ: Modality-Adaptive Quantization for MoE Vision-Language Models**.
Guangshuo Qin*, **Zhiteng Li***, Zheng Chen, Weihang Zhang, Linghe Kong, and Yulun Zhang.
4. **ReCalKV: Low-Rank KV Cache Compression via Head Reordering and Offline Calibration**.
Xianglong Yan*, **Zhiteng Li***, Tianao Zhang, Haotong Qin, Linghe Kong, Yulun Zhang, and Xiaokang Yang.
5. **D^{2}Quant: Accurate Low-bit Post-Training Weight Quantization for LLMs**.
Xianglong Yan*, Chengzhu Bao*, **Zhiteng Li**, Tianao Zhang, Haotong Qin, Shaoqiu Zhang, Ruobing Xie, Xingwu Sun, and Yulun Zhang.
6. **Progressive Binarization with Semi-Structured Pruning for LLMs**.
Xianglong Yan, Tianao Zhang, **Zhiteng Li**, Haotong Qin, and Yulun Zhang.
7. **CondiQuant: Condition Number Based Low-Bit Quantization for Image Super-Resolution**.
Kai Liu, Dehui Wang, **Zhiteng Li**, Zheng Chen, Yong Guo, Wenbo Li, Linghe Kong, and Yulun Zhang.
8. **Low-bit model quantization for deep neural networks: A survey**.
Kai Liu, Qian Zheng, Kaiwen Tao, **Zhiteng Li** et al.

Academic Service

Reviewer

- International Conference on Machine Learning (ICML) 2025, 2026
- Advances in Neural Information Processing Systems (NeurIPS) 2025
- International Conference on Learning Representations (ICLR) 2025, 2026
- International Conference on Computer Vision (ICCV) 2025
- Computer Vision and Pattern Recognition (CVPR) 2025, 2026

Honers and Awards

- Yang Yuanqing Education Foundation 2025
- Excellent Graduate of Shanghai Jiao Tong University 2023
- Ruiyuan-Sequoia Talent Development Fund 2022
- National Scholarship for Undergraduate Excellence (Top 0.2% nationwide) 2021
- Excellent Undergraduate Scholarship 2020, 2021, 2022

Skills

- Programming: Python, Pytorch, C/C++, LaTeX