Yufei Xue

Ph.D. Candidate @ Hong Kong University of Science and Technology

+86 166-xxxx-xxxx | E-Mail | Google Scholar

SUMMARY

I am currently a Ph.D. candidate at the Hong Kong University of Science and Technology supervised by Prof. Jun Zhang. I received my B.E. degree with honors from Southeast University in 2024. My research interests lie in Efficient LLMs, with a particular emphasis on quantization, distillation, and algorithm–system co-design. Prior to this, I also conducted research on privacy-preserving model serving and wireless communications.

EDUCATION

The Hong Kong University of Science and Technology — **Hong Kong SAR, China**
Ph.D. Candidate in the Department of Electronic and Computer Engineering — *Sep. 2024 – May 2028 (Expected)*

Southeast University — **Nanjing, China**
B.E. in Information Engineering, with Honors (Ranked 1st/246) — *Sep. 2021 – May 2024*
B.E. in Chien-Shiung Wu College; Chien-Shiung Student (Highest Honor) — *Sep. 2020 – May 2021*

EXPERIENCE

Tencent Hunyuan — **Shenzhen, China**
Qingyun Intern @ AI Infra Department; Supervisor: Dr. Guanghua Yu — *Apr. 2026 – Now*

Institute of Artificial Intelligence (TeleAI), China Telecom — **Shanghai, China**
Intern @ AI Flow Group; Supervisor: Dr. Jiawei Shao — *Jun. 2025 – Mar. 2026*

Peking University — **Beijing, China**
Research Assistant @ Institute for Artificial Intelligence; Supervisor: Prof. Meng Li — *Nov. 2023 – May 2024*

PUBLICATIONS

ProQuant: Progressive Quantization-Aware Training for Edge MLLMs
Yufei Xue, Yushi Huang, Jiawei Shao, Pingcheng Dong, Yonghao Tan, Shiyao Li, Kwang-Ting Cheng, Xuelong Li, Jun Zhang
Submitted

VLMQ: Token Saliency-Driven Post-Training Quantization for Vision-Language Models
Yufei Xue, Yushi Huang, Jiawei Shao, Lunjie Zhu, Chi Zhang, Xuelong Li, Jun Zhang
Submitted [Paper]

Flash-VAED: Plug-and-Play VAE Decoders for Efficient Video Generation
Lunjie Zhu, Yushi Huang, Xingtong Ge, Yufei Xue, Zhening Liu, Yumeng Zhang, Zehong Lin, Jun Zhang
Submitted [Paper]

Global Importance Balanced Mixed-Precision Quantization for Vision Language Models
Guilin Li, Yuexiao Ma, Yue Zhang, Xinxiong Wu, Yufei Xue, Jiaqi Zhou, Qingheng Zhang, Yan Zhang, Fei Chao, Xiawu Zheng
Submitted

FLASH: An Efficient Hardware Accelerator Leveraging Approximate and Sparse FFT for Homomorphic Encryption
Tengyu Zhang†, Yufei Xue†, Ling Liang, Zhen Gu, Yuan Wang, Runsheng Wang, Ru Huang, Meng Li
DATE 2025 [Paper]

Channel Estimation for RIS Assisted Millimeter Wave Systems via OMP with Optimization
You You†, Yufei Xue†, Li Zhang, Xiaohu You, Chuan Zhang
IEEE Transactions on Vehicular Technology 2023 [Paper]

† *Co-first authors*

HONORS AND AWARDS

HKUST RedBird PhD Award	*Aug. 2024*
Merit Student of Jiangsu Province (Top 1/246)	*May 2023*
National Scholarship (Top 1%)	*Sep. 2022*
Merit Student Model (Top 1/246)	*Sep. 2022*
Zhishan Student Scholarship	*May 2022*
Annual Chien-Shiung Student (Highest Honor)	*Nov. 2021*
Tong-Ren-Ding Scholarship	*Nov. 2021*
President Scholarship (Top 2/73)	*Sep. 2021*
Merit Student	*Sep. 2021*
Zhishan Student Scholarship	*May 2021*