HONG HUANG

Github ◇ Google Scholar ◇ Personal Website

Phone: +86-17349764371 ◇ WeChat: Hong4Work ◇ Email: honghuang2000@outlook.com

EDUCATION

City University of Hong Kong — *Hong Kong, China; Sept. 2024 – Jun. 2026 (Expected)*
Ph.D. in Computer Science — *Advised by Dr. Dapeng Wu*

City University of Hong Kong — *Hong Kong, China; Sept. 2023 – Aug. 2024*
Reseach Assistant in Computer Science — *Advised by Dr. Dapeng Wu*

University of Florida — *Gainesville, United States; Aug. 2021 – May 2023*
MSc. in Electrical and Computer Engineering — *Advised by Dr. Ruogu Fang and Dr. Dapeng Wu*

Shanghai Jiao Tong University — *Shanghai, China; Aug. 2017 – June 2021*
BE. in Computer Science and Technology — *Advised by Dr. Jian Cao*

SELECTED PUBLICATIONS

Hong Huang, Dapeng Wu "Quaff: Quantized Parameter-Efficient Fine-Tuning under Outlier Spatial Stability Hypothesis." The Annual Meeting of the Association for Computational Linguistics (ACL), 2025. PDF Code

Hong Huang, Hai Yang, Yuan Chen, Jiaxun Ye, Dapeng Wu. "FedRTS: Federated Robust Pruning via Combinatorial Thompson Sampling." The Conference on Neural Information Processing Systems (NeurIPS), 2025. PDF

Hong Huang, Weiming Zhuang, Chen Chen, and Lingjuan Lyu. "FedMef: Towards Memory-efficient Federated Dynamic Pruning." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. PDF

Hong Huang, Lan Zhang, Chaoyue Sun, Ruogu Fang, Xiaoyong Yuan, and Dapeng Wu. "Distributed Pruning Towards Tiny Neural Networks in Federated Learning." IEEE 43rd International Conference on Distributed Computing Systems (ICDCS), 2023. (CCF-B, Acceptance rate: 18.9%) PDF

EXPERIENCE

Tencent — *Shenzhen, China; Aug. 2025 - present*
Research Intern, Machine Learning Platform Department (MLPD) — *Mentored by Mr. Jianchen Zhu*

- Designed a ternary quant algorithm ({-1, 0, +1}) with bias compensation, significantly reducing quant errors.
- Achieved 5.2× inference speedup on LLaMA3.2-3B on Intel i7-13700H while maintaining accuracy compared to the BF16 baseline; submitted to ICLR 2026.

SONY AI — *Tokyo, Japan; Mar. 2023 - Aug. 2023*
Research Intern, Privacy-Preserving Machine Learning (PPML) Team — *Mentored by Dr. Lingjuan Lyu*

- Developed FedMef, a novel memory-efficient federated dynamic pruning framework
- Achieving 28.5% memory savings while improving the accuracy by 2%; published in CVPR 2024 [Link]

Meta — *Menlo Park, United States; Mar. 2022 - Dec. 2022*
Research Assistant, Video Infrastructure Group — *Mentored by Dr. Zhijun Lei*

- Developed TMAP, a CNN-based texture- and motion-aware in-loop filter for AV1
- Achieved reduction of 4.32% BD-rate and 3.79% VMAF; published in JVCIR [Link]

LEADERSHIP

- Admin for the FedPruning Research Group, a group of 10+ junior Ph.D. and M.S. students focused on edge computing and model compression; coordinated research leading to 4 papers accepted/submitted to top-tier conferences and transactions within six months (e.g., NeurIPS 2025, TPDS with major revision).