HONG HUANG

Github ◇ Google Scholar ◇ Personal Website
Phone: +86-17349764371 ◇ WeChat: Hong4Work ◇ Email: honghuang2000@outlook.com

RESEARCH INTERESTS

Efficient AI, Edge Intelligence, Model Compression & Acceleration, Efficient LLMs

EDUCATION/PROFESSIONAL EXPERIENCE

City University of Hong Kong *Hong Kong, China; Sept. 2024 – Present*
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

Efficient Mobile/Edge LLMs (*Bar Menu*)

- **[Quaff ACL'25] 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. [Link], [Code]
- **[Tequila ICLR'26] Hong Huang**, Decheng Wu, Rui Cen, Guanghua Yu, Zonghang Li, Kai Liu, Jianchen Zhu, Peng Chen, Xue Liu, Dapeng Wu. "Tequila: Trapping-free Ternary Quantization for Large Language Models." The Fourteenth International Conference on Learning Representations (ICLR), 2026. [Link], [Code], [Publicity]
- **[Sherry Preprint] Hong Huang**, Decheng Wu, Qiangqiang Hu, Guanghua Yu, Jinhai Yang, Jianchen Zhu, Xue Liu, Dapeng Wu. "Sherry: Hardware-Efficient 1.25-Bit Ternary Quantization via Fine-grained Sparsification." submitted to ACL 2026. [Link], [Code]

Efficient Federated Learning (*Fed- Series*)

- **[FedRTS NeurIPS'25] Hong Huang**, Hai Yang, Yuan Chen, Jiaxun Ye, Dapeng Wu. "FedRTS: Federated Robust Pruning via Combinatorial Thompson Sampling." The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS), 2025. [Link], [Code]
- **[FedMef CVPR'24] 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. [Link], [Code]
- **[FedTiny ICDCS'23] 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. (Acceptance rate: 18.9%) [Link], [Code]

Efficient ML System (*.cpp Series*)

- **[Prima.cpp ICLR'26]** Zonghang Li, Tao Li, Wenjiao Feng, Rongxing Xiao, Jianshu She, **Hong Huang**, Mohsen Guizani, Hongfang Yu, Qirong Ho, Wei Xiang, Steve Liu "Prima.cpp: Fast 30-70B LLM Inference on Heterogeneous and Low-Resource Home Clusters." The Fourteenth International Conference on Learning Representations (ICLR), 2026. [Link], [Code]

INTERNSHIP EXPERIENCE

Tencent *Shenzhen, China; Aug. 2025 - present*
Research Intern, AI Infra Department *Mentored by Mr. Jianchen Zhu*

- Focusing on ultra-low bit quantization techniques, we achieved innovative breakthroughs in ternary quantization: We first discovered and revealed the ”deadzone trapping” issue in ternary quantization, and proposed the **Tequila** quantization method, introducing dead weights reactivation technology to address the deadzone trapping issue and improve the model's capability.
- We further proposed the **Sherry** quantization method, which uses a 3:4 sparsification strategy to compress the ternary model to a hardware-friendly 1.25-bit, effectively solving the irregularity problem of ternary quantization in hardware deployment and improving inference efficiency.
- Both Tequila and Sherry achieve near-lossless performance on ARC metrics; Tequila was published at **ICLR'26** [Link], Sherry was submitted to **ACL'26** [Link].

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
- Achieved 28.5% memory savings while improving the accuracy by 2%; published in **CVPR'24** [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

- Leading FedPruning Research Group, a group of 15+ junior Ph.D. and M.S. students focused on edge computing and model compression; coordinated research leading to 5 papers accepted/submitted to top-tier conferences and transactions within six months (*e.g.,* NeurIPS'25, TCC with major revision).

AWARD/SCHOLARSHIP/FELLOWSHIP

DAAD AINet Fellowship (Postdoc-NeT-AI)	German Academic Exchange Service 2025
NeurIPS Travel Award	NeurIPS 2025
Research Tuition Scholarship	City University of Hong Kong 2025
Graduate School Fellowship	University of Florida 2021-2023
Zhiyuan Hornor Scholarship	Shanghai Jiao Tong University 2017-2021