Peirong **Zheng**

PhD Candidate in Computing · Efficient Vision–Language–Action Systems

Hong Kong SAR

+852 9283 3476 | peirong.zheng@connect.polyu.hk

Education

The Hong Kong Polytechnic University (PolyU) — *Hong Kong SAR*

PhD in Computing — *Jan. 2024 - Present*

- Best Poster Award, COMP 50th Anniversary Research Student Conference, 2024.

Nanyang Technological University — *Singapore*

MSc in Communications Engineering — *Aug. 2022 - Dec. 2023*

University of Electronic Science and Technology of China — *Chengdu, China*

BEng in Communication Engineering | CGPA: 3.65/4.00 | Yingcai Honors Program — *Sep. 2018 - Jul. 2022*

University of Cambridge and University of Oxford — *Cambridge and Oxford, UK*

Visiting Student | Summer Institute — *Aug. 2019*

Research Experience

`AR-VLA`: Conditional Trajectory Reuse for Efficient Autoregressive Vision-Language-Action Inference — *Hong Kong SAR*

Current Research Project — *2026 - Present*

- Developed a training-free conditional trajectory-reuse framework for frozen autoregressive VLA policies, combining per-task dual-view VLA-native memory, threshold-gated top-2 future-action aggregation, and periodic model anchors.
- Designed kinematic descent and grasp-transition guards to truncate unsafe reuse; served **64.1%** of actions from memory and achieved **3.1%** better average success rate on LIBERO tasks with a **2.08**× end-to-end speedup.

Efficient LLM Systems Research — *Hong Kong SAR*

PhD Researcher | Advisor: Prof. Wenchao Xu, currently at HKUST — *Jan. 2024 - Present*

- Achieved up to **3.72**× **decoding speedup** over state-of-the-art methods while maintaining nearly lossless accuracy.
- **`HALO` and `Coda` (INFOCOM 2026; MobiCom Poster 2025):** Developed semantic- and context-aware distributed LLM inference that relaxes synchronization under lossy edge networks.
- Implemented and evaluated LLaMA-series inference on Raspberry Pi clusters; HALO delivered up to **3.41**× **end-to-end speedup** under packet loss, while Coda achieved up to **5.62**× **speedup**.

Technical Skills

Programming	Codex, PyTorch, C++, MATLAB
Systems	OpenVLA, LIBERO, trajectory retrieval, distributed inference, GPU/CPU expert offloading
Research Areas	Vision-language-action models, efficient embodied AI, LLM inference acceleration

Publications

Self-Speculative Decoding for On-device MoE Acceleration — *WWW '26*

Peirong Zheng, Wenchao Xu, and Haozhao Wang — *DOI*

HALO: Semantic-Aware Distributed LLM Inference in Lossy Edge Network — *INFOCOM '26*

Peirong Zheng, Wenchao Xu, Haozhao Wang, Jinyu Chen, and Xuemin Sherman Shen

Poster: Coda: Context-aware Acceleration for Distributed LLM Decoding in Edge — *MobiCom '25*

Peirong Zheng, Wenchao Xu, Haozhao Wang, and Jinyu Chen — *DOI*

Deploying Foundation Model Powered Agent Services: A Survey — *IEEE CST*

Wenchao Xu, Jinyu Chen, **Peirong Zheng**, Xiaoquan Yi, Tianyi Tian, Wenhui Zhu, Quan Wan, Haozhao Wang, Yunfeng Fan, Qinliang Su, and Xuemin Shen — *DOI*

FedNLR: Federated Learning with Neuron-wise Learning Rates — *KDD '24*

Haozhao Wang*, **Peirong Zheng***, Xingshuo Han, Wenchao Xu, Ruixuan Li, and Tianwei Zhang — *DOI*

A Semi-supervised Algorithm for Atrial Fibrillation Attack Prediction — *EMBC '23*

Yuchen Jiang*, **Peirong Zheng***, and Dakun Lai | *Equal contribution — *DOI*