Personal Information



Name: Wenju Zhao
Gender: Male
Born: April 2001
Major: Computer Science and Technology
Status: PhD Student
Advisor: Prof. Xiaofei Liao

Education

Huazhong University of Science and Technology
Sep. 2022–Jun. 2027

School of Computer Science and Technology; PhD Student

- **PhD Dissertation**: Accelerating Deep Learning Inference through Model Compression
- **Research interests**: LLM inference acceleration, model compression, quantization, software–hardware co-design, and computer architecture.
- **GPA**: 3.77 / 4 (rank: 1 / 11)

National University of Singapore
Aug. 2026–Apr. 2027

School of Computing; Visiting PhD Student

- **Host advisor**: Prof. Bingsheng He (IEEE Fellow)
- **Research interests**: LLM inference acceleration and model compression.

Wuhan University of Technology
Sep. 2018–Jun. 2022

School of Computer Science and Artificial Intelligence; Undergraduate

- **Major**: Data Science and Big Data Technology
- **GPA**: 4.32 / 5 (rank: 1 / 33)

Research Experience

National Science and Technology Major Project: Next-Generation AI
Hardware-Efficient LLM Inference Engine — **Core Researcher**, 2025–present

- **Contribution 1**: Proposed a new dot-product computation paradigm for quantized LLM inference, achieving a 1.51–1.98× speedup over existing methods (DAC 2026; **first author**, accepted).
- **Contribution 2**: Proposed a new quantization algorithm and a software–hardware co-designed accelerator for LLM inference, achieving a 1.44–2.72× speedup over existing methods (MICRO 2026; **first author**, accepted).
- **Contribution 3**: Developed **MetaKV**, which further quantizes the metadata introduced by KV-cache quantization to reduce memory overhead for long-context LLM inference (**HPCA 2027**, **first author**, under review).

National Key R&D Program
Graph Computing for Complex Scenarios — **Core Researcher**, 2022–present

- **Contribution 1**: Designed a microedge-centric hypergraph neural network accelerator, introducing a microedge abstraction to eliminate redundant computation and memory accesses. Achieved a 10.51× speedup over a GPU-based inference system (HPCA 2025; **first author**, published).
- **Contribution 2**: Developed **HyperPIM**, a redundancy-aware heterogeneous PIM accelerator for hypergraph neural networks (**TCAD 2026**; **co-first author**, accepted).

Email: wjzh@hust.edu.cn Phone: (86) 187-3885-3795 Website: wenjuzhao.github.io

- **Contribution 3**: Conducted a systematic survey of the full-stack graph processing ecosystem, covering domain-specific accelerators, parallel graph computing systems, and distributed techniques, and analyzed their combined roles in accelerating graph applications in scientific computing and AI (JCST; **student first author**; accepted as a **cover article**).

Joint Training Program at the National University of Singapore

Efficient Inference for Embodied AI — **Core Member**, 2026–present

- **Main contribution**: Addressing the computational cost and real-time constraints of deploying vision–language–action (VLA) models on robots. Analyzing computation and memory-access patterns across actions and inference stages, and investigating **action-aware dynamic computation and model compression**.

Honors, Awards and Service

- 2026, 2024, 2020: **National Scholarship** (PhD: 2026, 2024; undergraduate: 2020)
- Jun. 2026: **Best Paper Award, 4th CS Academic Annual Conference, HUST**
- Apr. 2026: **Artifact Evaluation Committee Member, ISCA 2026**
- Aug. 2025: **Invited Talk on Graph Computing Accelerators, GPU and Domain-Specific Accelerator Architectures Session, CCFSys 2025**
- Mar. 2024: **Cover Article, Journal of Computer Science and Technology**

Research Output (Projects, Publications, Patents)

Research Projects

- **2025–present: Hardware-Efficient LLM Inference Engine**
  National Science and Technology Major Project: Next-Generation AI / Core Researcher
- **2025–2026: Low-Bit Quantization Accelerator for LLM Inference**
  Technology Innovation Fund, HUST Innovation Research Institute / Project Lead
- **2022–present: Graph Computing for Complex Scenarios**
  National Key R&D Program / Core Researcher
- **2021–2022: Situational Awareness Based on Whistle Sound Understanding**
  National Undergraduate Innovation Training Program / Project Lead

Publications (Student First Author*, Co-First Author=)

- **[1]** **Wenju Zhao**, Jianhui Yue, Pengcheng Yao, Jiaqi Zhai, Yu Huang, Shunsen Lv, Bing Zhu, Zhongtian Long, Long Zheng, Chencheng Ye, Lihong Gu, Jingling Xue, Hai Jin, Xiaofei Liao. SkewQ: A Skew-Aware Cross-Layer Quantization Accelerator for LLM Inference. International Symposium on Microarchitecture **(MICRO 2026, First Author, Accepted)**
- **[2]** **Wenju Zhao**, Jianhui Yue, Pengcheng Yao, Yujia Cui, Qinggang Wang, Yufei Sun, Jiaqi Zhai, Hai Jin, Xiaofei Liao. FuseDot: A Multiplication-Fused Dot Product Accelerator for Efficient LLM Inference. The Chips to Systems Conference **(DAC 2026, First Author, Accepted)**

- **[3]** **Wenju Zhao**, Pengcheng Yao, Dan Chen, Long Zheng, Xiaofei Liao, Qinggang Wang, Shaobo Ma, Yu Li, Haifeng Liu, Wenjing Xiao, Yufei Sun, Bing Zhu, Hai Jin, Jingling Xue. MeHyper: Accelerating Hypergraph Neural Networks by Exploring Implicit Dataflows. International Symposium on High-Performance Computer Architecture **(HPCA 2025, First Author, Accepted)**
- **[4]** **Zhongtian Long[=], Wenju Zhao[=]**, Mo Zou, Hongrui Guo, Jin Zhao, Qiyuan Niu, Yutao Fu, Shunsen Lv, Zixiao Wang, Yuluo Guo, Yongwei Zhao, Yu Zhang, Hai Jin. HyperPIM: A Redundancy-Aware Heterogeneous PIM Accelerator for Hypergraph Neural Networks. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **(TCAD 2026, Co-First Author, Accepted)**
- **[5]** **Xiaofei Liao, Wenju Zhao***, Hai Jin, Pengcheng Yao, Yu Huang, Qinggang Wang, Jin Zhao, Long Zheng, Yu Zhang, Zhiyuan Shao. Towards High-Performance Graph Processing: From a Hardware/Software Co-Design Perspective. Journal of Computer Science and Technology **(Cover Article, JCST 2024, Student First Author, Accepted)**
- **[6]** **Wenju Zhao**, Bowen Rao, Pengcheng Yao, Jianhui Yue, Shunsen Lv, Zhongtian Long, Jiaqi Zhai, Nanxun Wu, Zhen Zeng, Yu Huang, Bingsheng He, Jingling Xue, Hai Jin, Xiaofei Liao. MetaKV: Metadata-Aware KV-Cache Quantization for Efficient Long-Context LLM Inference via Software-Hardware Co-Design. International Symposium on High-Performance Computer Architecture **(HPCA 2027, First Author, Under Review)**
- **[7]** Zou Qingyun, Bin Gao, **Wenju Zhao**, Weng-Fai Wong, Bingsheng He, and Tulika Mitra. ActTune: Action-Aware Precision and GPU Operating-Point Adaptation for Energy-Efficient Vision-Language-Action Inference. arXiv preprint arXiv:2610.08444v1, 2026
- **[8]** Minnan Pei, Qiwei Dong, Yihan Zhou, Gang Li, Yuchen Zhu, **Wenju Zhao**, Zhongtian Long, Siting Wang, Peisong Wang, and Jian Cheng. From Splats to Silicon: Rethinking Computational Efficiency of 3DGS. arXiv preprint arXiv:2609.06157, 2026
- **[9]** Yufei Sun, Pengcheng Yao, Long Zheng, Yundong Zhang, Yu Huang, Qinggang Wang, Huize Li, **Wenju Zhao**, Bing Zhu, Haifeng Liu, Xiaofei Liao, and Hai Jin. SpecTwin: Accelerating Tree-based Speculative Decoding via Differential Token Quantization. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **(TCAD 2026, Accepted)**
- **[10]** Xinyu Zhu, Qinggang Wang, **Wenju Zhao**, Bing Zhu, Runze Wang, Long Zheng, Xiaofei Liao, Hai Jin, Jingling Xue. RAIL: A Reconfigurable Spatial Accelerator for Linear-Time Sequence Model Inference. International Symposium on Microarchitecture **(MICRO 2026, Accepted)**
- **[11]** Jiaqi Zhai, Xuanhua Shi, **Wenju Zhao**, Kaiyi Huang, Chencheng Ye, Shunsen Lv, Zhongtian Long, Bingsheng He, Hai Jin. ParetoES: Hardware-Accelerated Sparse Embedding Similarity via Pareto-Optimal Pruning. The International Symposium on Computer Architecture **(ISCA 2026, Accepted)**
- **[12]** Zhongtian Long, Yu Zhang, Yutao Fu, Qiyuan Niu, Shunsen Lv, Jiaqi Zhai, **Wenju Zhao**, Yingqi Zhao, Hai Jin. A High-Efficiency 3D-Stacked Accelerator for Deep Graph Neural Network Inference. Scientia Sinica Informationis **(SSI 2026, Accepted)**
- **[13]** Long Zheng, Bing Zhu, Pengcheng Yao, Yuhang Zhou, Chengao Pan, **Wenju Zhao**, Xiaofei Liao, Hai Jin, and Jingling Xue. PRAGA: A Priority-Aware Hardware/Software Co-design for High-Throughput Graph Processing Acceleration. ACM Transactions on Architecture and Code Optimization **(TACO 2025, Accepted)**

Patents

- **[1] Method and System for Sound Semantic Recognition Combining a Dual-Threshold Method and Residual Neural Networks (Granted)**.
  Inventors: Pengfei Duan, Wenju Zhao, Peisheng Zhao, Long Li. Patent No.: ZL 202111536963.X
- **[2] Accelerator and Acceleration Method Based on Hypergraph Neural Networks (Under Substantive Examination)**.
  Inventors: Xiaofei Liao, Wenju Zhao, Pengcheng Yao, Long Zheng, Dan Chen, Qinggang Wang, Haifeng Liu. Application No.: 202510037272.7
- **[3] LLM Inference Accelerator Design Method Based on Cross-Layer Power-Law-Aware Quantization (Under Substantive Examination)**.
  Inventors: Wenju Zhao, Xinpeng Wang, Xiaofei Liao, Pengcheng Yao, Yujia Cui. Application No.: 202511475284.4