He Xiao

+86 18980479369 | hexiaoriver@connect.hku.hk / hxriver@126.com | Hong Kong, China

github | google scholar | personal website

Introduction

I am a second-year Ph.D. candidate in Electrical and Electronic Engineering at the University of Hong Kong, under the supervision of Prof. Ngai Wong and Prof. Can Li. My research centers on developing efficient artificial intelligence systems and AI accelerators, with expertise in in-memory computing, ultra-low-bit quantization, large language models, and hardware–software co-optimization. I am passionate about bridging the gap between advanced AI algorithms and practical, high-performance hardware implementations.

Education

- **The University of Hong Kong** — *Sep 2024 – Aug 2028*
  Ph.D. in Electrical and Electronic Engineering — Hong Kong, China
  ◦ Advisors: Prof. Ngai Wong, Prof. Can Li
- **The University of Western Australia** — *Jul 2019 – Nov 2021*
  M.Sc. in Electrical and Electronic Engineering — Perth, Australia
  ◦ Advisors: Prof. Wen Lei
- **Southwest University** — *Sep 2016 – Jun 2020*
  B.Eng. in Automation — Chongqing, China
  ◦ Advisors: Prof. Xiaofang Hu

Academic Experience

- **The University of Hong Kong, Factulty of Engineering** — *2025 Fall*
  Teacher Assistant, Math 1853 — Hong Kong, China
- **Institute of Microelectronics, Southern University of Science and Technology** — *Aug 2025 – Oct 2025*
  Visiting Scholar, tapeout project collaboration — Shenzhen, China
- **School of Artificial Intelligence, Southwest University** — *Jun 2022 – Jan 2024*
  Research Assistant, team leader — Chongqing, China

Internship Experience

- **PIMIC Inc. (USA), Hong Kong Science and Technology Park** — *Sep 2024 – Aug 2025*
  Research Scientist Intern, Efficient AI & Edge Deployment — Hong Kong, China
  ◦ Led the development of an on-device speech-dialogue chatbot, enabling large-language-model interaction on edge hardware.
  ◦ Conducted model fine-tuning and quantization to achieve real-time inference with 4-bit precision.
  ◦ Initiated preliminary ASIC research for specialized AI accelerators based on project insights.

Selected Publications and Patents

C=Conference, J=Journal, P=Patent

[C.1] **H. Xiao**, R. Yang, Q. Yang, W. Xu, Z. Li, Y. Su, Z. Liu, H. Yang, N. Wong. **PTQTP: Post-Training Quantization to Trit-Planes for Large Language Models**. *Under review*, ICLR 2026.

[C.2] **H. Xiao**, Q. Yang, D. Xie, W. Xu, W. Zhou, H. Liu, Z. Liu, N. Wong. **Exploring Layer-wise Information Effectiveness for Post-Training Quantization in Small Language Models**. *Under review*, AAAI 2026.

[C.3] **H. Xiao**, Y. Zhou, D. Xie, Q. Cheng, X. Hu, Z. Liu, N. Wong. **A Unified Compute-In-Memory Framework for Multisensory Emotion Recognition**. *Accepted as poster*, ASP-DAC 2025.

[C.4] **H. Xiao**, H. Liu, Z. Liu, N. Wong. **Brain-Inspired Quantized Spiking Neural Networks with Efficient Analog Neurons**. *Accepted as poster*, MIND 2025.

[J.1] **H. Xiao**, H. Liu, D. Cheng, W. Xu, J. Xiong, Z. Liu, N. Wong. **HKPIM: A Hybrid Kernel-fused Processing-In-Memory Framework for Efficient Large Language Model Inference**. *IEEE TCAD*, 2025, under review.

[J.2] **H. Xiao**, D. Xie, X. Hu, Y. Zhou, S. Duan. **Brain-Inspired Recognition System Based on Multimodal In-Memory Computing Framework for Edge AI**. *IEEE TCAS-I*, 71(5):2294–2307, 2024.

[J.3] **H. Xiao**, X. Hu, T. Gao, Y. Zhou, S. Duan, Y. Chen. **Efficient Low-bit Neural Network with Memristor-Based Reconfigurable Circuits**. *IEEE TCAS-II*, 71(1):66–70, 2024.

[J.4] **H. Xiao**, Y. Zhou, T. Gao, S. Duan, G. Chen, X. Hu. **Memristor-Based Light-Weight Transformer Circuit Implementation for Speech Recognition**. *IEEE JETCAS*, 13(1):344–356, 2023.

[J.5] **H. Xiao**, H. Sun, T. Zhao, Y. Zhou, X. Hu. **Pure-Attention-Based Multi-Function Memristive Neuromorphic Circuit and System**. *International Journal of Bifurcation and Chaos*, 33(9):2330023, 2023.

[J.6] Wenhao Zhang* and **H. Xiao***, Y. Zhou, S. Duan, X. Hu, **A Global Self-Attention Memristive Neural Network for Image Restoration**, *IEEE TETCI*, 8(3):2613-2624,2024.

[P.1] Y. Zhou, **H. Xiao**, X. Hu, H. Hong, S. Duan. **Speech Recognition System Based on Light-Weight Transformer**. Chinese Patent ZL 2023 1 0065728.1.

[P.2] Y. Zhou, **H. Xiao**, X. Hu, H. Hong, S. Duan. **Memristor-Based Text Sentiment Detection System**. Chinese Patent ZL 2023 1 0091466.6.

[P.3] X. Hu, D. Xie, Y. Zhou, **H. Xiao**. **Image Restoration Method Based on Brain-Inspired Vision Transformer for Edge ADAS Devices**. Chinese Patent ZL 2023 1 1466352.1.

Skills

- **Programming Languages:** Python, C++, Verilog
- **AI/Deep Learning Frameworks:** PyTorch, HuggingFace Transformers, Triton, VLLM
- **EDA Tools:** Cadence, Synopsys
- **Languages:** English (Fluent), Mandarin (Native)

Honors and Awards

- PGS, The University of Hong Kong
- Chongqing Outstanding Individual, ChongQing

Student Experience

- **Chairman** *Jun 2018 - Jun 2019*
  Youth Volunteers Association of Southwest University (University-level Organization)
- **Main Commitee Member** *Jun 2018 - Jun 2019*
  Student Association of Southwest University

Academic and Social Service

- **Reviewer** *Sep 2024 - Present*
  IEEE TCAS-I, IEEE TCAS-II, AAAI, NeurIPS
- **Volunteer** *Sep 2016 - Jun 2019*
  During Bachelor Study at Southwest University
  - Participated in a variety of volunteer service activities, serving groups such as left-behind children, local elderly, and children with special needs. Service organizations included nursing homes, communities, kindergartens, and primary and secondary schools. Assisted in organizing various volunteer publicity activities, with a total service time exceeding 100 hours.