

# YUPENG SU

UC Santa Barbara | [yupengsu@ucsb.edu](mailto:yupengsu@ucsb.edu) | [yupengsu.github.io](https://github.com/yupengsu) | [linkedin.com/in/yupeng-su](https://www.linkedin.com/in/yupeng-su)

## Research Interests

---

**Efficient Pre-training/Post-training Optimization for Large Language Models:** focusing on quantization, sparsification, and low-rank decomposition techniques to improve training and inference efficiency

**High-Performance Computing Infrastructure on GPU/FPGA Architectures:** including compiler optimization and GPU/FPGA kernel adaptation for high-performance, energy-efficient inference and practical edge deployment

## Education

---

**University of California, Santa Barbara**, Ph.D. in Computer Science – CA, USA Sept 2025 – present  
 • Advisor: Prof. Zheng Zhang (ECE, UCSB).

**Southern University of Science and Technology**, B.E. in Microelectronics – China Sept 2021 – July 2025  
 • Advisor: Prof. Hao Yu (SME, SUSTech).  
 • CGA/GPA: 3.93/4.0, with 178 units completed.  
 • Awards: Nominee of Top Ten Graduates of SUSTech, Top Ten Graduates of the College of Engineering, SUSTech  
 • Dissertation: *Enhanced Mix-Precision Post-Training Quantization for Large Language Models: Block Awareness and Outlier Identification*.

## Publications

---

**APTQ: Attention-aware post-training mixed-precision quantization for large language models**

Ziyi Guan, Hantao Huang, **Yupeng Su**, Hong Huang, Ngai Wong, Hao Yu  
[10.1145/3649329.3658498](https://arxiv.org/abs/10.1145/3649329.3658498) (IEEE/ACM DAC, 2024)

**LLM-Barber: Block-Aware Rebuilder for Sparsity Mask in One-Shot for Large Language Models**

**Yupeng Su**, Ziyi Guan, Xiaoqun Liu, Tianlai Jin, Dongkuan Wu, Chenfei Chen, Graziano Chesi, Ngai Wong, Hao Yu  
[10.1109/ICCAD66269.2025.11240704](https://arxiv.org/abs/10.1109/ICCAD66269.2025.11240704) (IEEE/ACM ICCAD, 2025)

**EdgeLLM: A Highly Efficient CPU-FPGA Heterogeneous Edge Accelerator for Large Language Models**

Mingqiang Huang, Ao Shen, Kai Li, Haoxiang Peng, Boyu Li, **Yupeng Su**, Hao Yu  
[10.1109/TCSI.2025.3546256](https://arxiv.org/abs/10.1109/TCSI.2025.3546256) (IEEE TCAS I: Regular Papers)

**Quantization Meets Reasoning: Exploring Low-Bit Quantization Degradation for Mathematical Reasoning**

Zhen Li\*, **Yupeng Su\***, Runming Yang, Congkai Xie, Zheng Wang, Zhongwei Xie, Ngai Wong, Hongxia Yang  
[arxiv.org/abs/2501.03035](https://arxiv.org/abs/2501.03035) (Under Review, \*Equal Contribution)

**PTQTP: Post-Training Quantization to Trit-Planes for Large Language Models**

He Xiao, Runming Yang, Qingyao Yang, Wendong Xu, Zhen Li, **Yupeng Su**, Zhengwu Liu, Hongxia Yang, Ngai Wong  
[arxiv.org/abs/2509.16989](https://arxiv.org/abs/2509.16989) (Under Review)

## Experience

---

**Student Research Assistant**, Prof. Hongxia Yang's Lab @ PolyU – Hongkong, China Nov 2024 – May 2025  
 • Investigated the effects of compression on reasoning and explored data-driven strategies for restoration.

**Student Research Assistant**, Next Gen AI (NGai) Lab @ HKU – Hongkong, China Aug 2024 – Feb 2025  
 • Contributed to the lightweight networks for efficient inference under specified computing architectures.

**Summer Intern**, High Performance Integrated Circuit Design Lab @ SUSTech – China June 2024 – Sept 2024  
 • Implement end-to-end pipelines for compressing, optimizing, and deployment for efficient device.  
 • Published [EdgeLLM](#) about the project and results