# Yushi Huang

yh4717023@gmail.com · Harahan Harahan.github.io

## Education

**Hong Kong University of Science and Technology** — Expected 2025.02
*Ph.D. student at ECE, advised by Prof. Jun Zhang* — Hong Kong SAR, China

**Beihang University** — 2020.09 – 2024.06
*Bachelor of Computer Science and Technology* — Beijing, China

- Grade Point Average (GPA): 3.86/4.00
- Weighted Score: 93.2/100

## Research Interest

My research interest is in building efficient vision/language generative models. I am currently working on inference acceleration strategies, *e.g.*, quantization, and pruning in a post-training manner.

## Publications

"*" and "†" denote equal contributions and corresponding authors.

1. **TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models**
   **Yushi Huang***, Ruihao Gong*, Jing Liu, Tianlong Chen, Xianglong Liu†
   Conference on Computer Vision and Pattern Recognition (CVPR), 2024. **(Highlight)**
2. **LLMC: Benchmarking Large Language Model Quantization with a Versatile Compression Toolkit**
   Ruihao Gong*, Yang Yong*, Shiqiao Gu*, **Yushi Huang***, Chengtao Lv, Yunchen Zhang, Dacheng Tao, Xianglong Liu†
   Conference on Empirical Methods in Natural Language Processing: Industry Track (EMNLP Industry Track), 2024.
3. **PTSBench: A Comprehensive Post-Training Sparsity Benchmark Towards Algorithms and Models**
   Zining Wang, Jinyang Guo, Yang Yong, Ruihao Gong, Aishan Liu, **Yushi Huang**, Jiaheng Liu, Xianglong Liu†
   ACM International Conference on Multimedia (ACM MM), 2024.

## Preprint

"*" and "†" denote equal contributions and corresponding authors.

1. **HarmoniCa: Harmonizing Training and Inference for Better Feature Cache in Diffusion Transformer Acceleration**
   **Yushi Huang***, Zining Wang*, Ruihao Gong†, Jing Liu, Xinjie Zhang, Jun Zhang†
   In Submission to International Conference on Learning Representations (ICLR), 2025.
2. **Temporal Feature Matters: A Framework for Diffusion Model Quantization**
   **Yushi Huang**, Ruihao Gong, Xianglong Liu†, Jing Liu, Yuhang Li, Jiwen Lu, Dacheng Tao
   In submission to Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

## Projects

**LLMC: Towards Accurate and Efficient LLM Compression** **(>300 Stars)**
Core Contributors: Yushi Huang, Yang Yong, Shiqiao Gu

- Implement many quantization methods for LLM, like QuaRot, GPTQ, SmoothQuant, OmniQuant...
- Build an end-to-end LLM quantization tool, that supports multiple model architectures, evaluation approaches, and inference backends...
- Provide best practices for quantization on LLM under different conditions.

## Experience

**SenseTime** — 2023.05 – Present
*Research Intern, mentored by Dr. Ruihao Gong* — Beijing, China
Compression and acceleration for vision/language generative models.

## Academic Services

- **Conference Reviewer**: NeurIPS, ICLR

## Skills

- **Programming Languages**: Python, C, Java
- **Scientific Packegs**: Pytorch, Numpy

## Others

- **Languages**:
  - Mandarin Chinese (native)
  - English: 107 (R: 28 L: 29 S: 23 W: 27) in TOEFL iBT TEST