

# YUSHI HUANG

✉ [yh4717023@gmail.com](mailto:yh4717023@gmail.com) · 🌐 [Harahan](#) 📄 [Harahan.github.io](#)

## EDUCATION

### Hong Kong University of Science and Technology

Ph.D. student at ECE, advised by Prof. [Jun Zhang](#)

2025.02 – Now

Hong Kong SAR, China

### Beihang University

Bachelor of Computer Science and Technology

2020.09 – 2024.06

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 and training acceleration strategies.

## PUBLICATIONS

“\*” and “†” denote equal contributions and corresponding authors.

### 1. **HarmoniCa: Harmonizing Training and Inference for Better Feature Caching in Diffusion Transformer Acceleration** 📄 🌐

Yushi Huang\*, Zining Wang\*, Ruihao Gong†, Jing Liu, Xinjie Zhang, Jinyang Guo, Xianglong Liu, Jun Zhang†  
International Conference on Machine Learning (ICML), 2025.

### 2. **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)**

### 3. **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.

### 4. **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. **QVGen: Pushing the Limit of Quantized Video Generative Models** 📄

Yushi Huang, Ruihao Gong†, Jing Liu, Yifu Ding, Chengtao Lv, Haotong Qin, Jun Zhang†  
In submission to Conference on Neural Information Processing Systems (NeurIPS), 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  
Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Minor Revision.

## PROJECTS

### LLMC: Towards Accurate and Efficient LLM Compression 🌐 (**>400 Stars**)

One of the Core Contributors who:

- 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

Research Intern, mentored by Dr. [Ruihao Gong](#)

2023.05 – 2024.12 & 2025.02 – Present

Beijing, China

Compression and acceleration for vision/language generative models.

## Microsoft Research Asia

Research Intern, mentored by Dr. [Fangyun Wei](#)

Video generation and world models.

2024.12 – 2025.02

Beijing, China

## ACADEMIC SERVICES

---

- **Conference Reviewer:** NeurIPS, ICLR, ICML, COLM

## SKILLS

---

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

## OTHERS

---

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