Yushi Huang

(+86) 15015597343 · yh4717023@gmail.com · Harahan · Harahan.github.io

Education

The Hong Kong University of Science and Technology 2025.02 – Present
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 and language generative models. I am currently working on improving the efficiency of inference and training while maintaining performance and robustness for large-scale models.

Publications

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

1. **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), 2025
2. **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
3. **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)***
4. **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
5. **LLMC+: Benchmarking Vision-Language Model Compression with a Plug-and-play Toolkit**
   Chengtao Lv, Bilang Zhang, Yang Yong, Ruihao Gong†, **Yushi Huang**, Shiqiao Gu, Jiajun Wu, Yumeng Shi, Jinyang Guo, Wenya Wang†
   Association for the Advancement of Artificial Intelligence (AAAI), 2026
6. **SlimInfer: Accelerating Long-Context LLM Inference via Dynamic Token Pruning**
   Lingkun Long, Rubing Yang, **Yushi Huang**, Desheng Hui, Ao Zhou†, Jianlei Yang†
   Association for the Advancement of Artificial Intelligence (AAAI), 2026
7. **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. **MoDES: Accelerating Mixture-of-Experts Multimodal Large Language Models via Dynamic Expert Skipping**
   Yushi Huang, Zining Wang, Zhihang Yuan†, Yifu Ding, Ruihao Gong, Jinyang Guo, Xianglong Liu, Jun Zhang†
   In submission to Conference on Computer Vision and Pattern Recognition (CVPR), 2026
2. **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 International Conference on Learning Representations (ICLR), 2026

3. **LinVideo: A Post-Training Framework towards $\mathcal{O}(n)$ Attention in Efficient Video Generation**
   Yushi Huang, Xingtong Ge, Ruihao Gong†, Chengtao Lv, Jun Zhang†
   In submission to Conference on Computer Vision and Pattern Recognition (CVPR), 2026
4. **Towards Efficient Post-Training Quantization For Large Vision-Language Models Via Token-Wise Redundancy Elimination**
   Yufei Xue, **Yushi Huang**, Jiawei Shao, Lunjie Zhu, Chi Zhang, Xuelong Li†, Jun Zhang†
   In submission to International Conference on Learning Representations (ICLR), 2026
5. **Feed-Forward 3D Gaussian Splatting Compression with Long-Context Modeling**
   Zhening Liu*, Rui Song*, **Yushi Huang**, Yingdong Hu, Xinjie Zhang, Jiawei Shao, Zehong Lin, Jun Zhang†
   In submission to Conference on Computer Vision and Pattern Recognition (CVPR), 2026

Projects

LightCompress: Towards Accurate and Efficient AIGC Model Compression (600+ Stars)

One of the core contributors who:

- Implements many quantization methods for LLM, like QuaRot, GPTQ, SmoothQuant, OmniQuant, *etc*;
- Builds an end-to-end LLM quantization tool that supports multiple model architectures, evaluation approaches, and inference backends;
- Provides best practices for quantization on LLM under different setups.

Experience

SenseTime Research 2023.05 – Now
Research Intern, mentored by Ruihao Gong Beijing, China
Compression and acceleration for vision and language generative models.

Microsoft Research Asia 2024.12 – 2025.02
Research Intern, mentored by Fangyun Wei Beijing, China
Video generation and world models.

Academic Services

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

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