Haofeng Huang

✉ huanghf22@mails.tsinghua.edu.cn | 🎓 Google Scholar
https://github.com/jason-huang03

EDUCATION

B.E., Tsinghua University — **Beijing, China**
Institute for Interdisciplinary Information Science (Yao Class) — **Sep 2022 - 2026**

- GPA: 3.97, ranked 3rd out of 94.
- Related Coursework: Deep Learning, Natural Language Processing, Machine Learning, Probability and Statistics, Introduction to Computer System, Database System, Calculus, Linear Algebra, High-Performance Computing.

PhD, Tsinghua University — **Beijing, China**
Institute for Interdisciplinary Information Science — **Sep 2026 -**

TECHNICAL PROFICIENCIES AND STRENGTHS

- **Programming Languages:** Familiar with C/C++, CUDA and Python, with a special focus on developing and maintaining deep learning frameworks. Rich experience in Python, Pytorch and CUDA.
- **Software and Tools:** Competent with key tools including Git, LaTeX, and bash.
- **Research and Communication:** Comfortable with reading scientific papers and good at explaining technical concepts and findings in written English.
- **Self-Directed Learning:** Good at quickly and independently learning new technologies and knowledge from open resources.
- **Mathematical Understanding:** Comfortable with mathematical proof and reasoning, familiar with the principles and practices of mathematical problem-solving.

CERTIFICATION & AWARDS

2021 37th CMO (Chinese Mathematical Olympiad) gold medal, National Training Team

2023 Tsinghua University Academic Excellence Award

2024 Tsinghua University Spark Scientific and Technological Innovation Fellowship (top 1% in Tsinghua University)

2024 SenseTime Scholarship (awarded to 25 undergraduate students nationwide)

2024 National Scholarship (top 0.2% of students nationwide)

2025 National Scholarship (top 0.2% of students nationwide)

2025 Young Scientists Fund of the National Natural Science Foundation of China

2025 Yao Award Silver Medal (top 4 of students in Yao Class)

PUBLICATION

1. Peiyuan Zhang*, Yongqi Chen*, **Haofeng Huang***, Will Lin, Zhengzhong Liu, Ion Stoica, Eric P Xing, Hao Zhang. Faster Video Diffusion with Trainable Sparse Attention. In Proceedings of the *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

2. Jintao Zhang*, Jia Wei*, Pengle Zhang, Xiaoming Xu, **Haofeng Huang**, Haoxu Wang, Kai Jiang,Jun Zhu, Jianfei Chen. SageAttention3: Microscaling FP4 Attention for Inference and An Exploration of 8-bit Training. In Proceedings of the *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

3. Jintao Zhang*, **Haofeng Huang***, Pengle Zhang, Jia Wei, Jun Zhu, Jianfei Chen. SageAttention2: Efficient Attention with Thorough Outlier Smoothing and Per-thread INT4 Quantization. In Proceedings of the *International Conference on Machine Learning (ICML)*, 2025.

4. Jintao Zhang*, Chendong Xiang*, **Haofeng Huang***, Jia wei, Haocheng Xi, Jun Zhu, Jianfei Chen . SpargeAttention: Accurate and Training-free Sparse Attention Accelerating Any Model Inference. In Proceedings of the *International Conference on Machine Learning (ICML)*, 2025.

5. Ruyi Xu*, Guangxuan Xiao*, **Haofeng Huang**, Junxian Guo, Song Han. XAttention: Block Sparse Attention with Antidiagonal Scoring. In Proceednigs of the *International Conference on Machine Learning (ICML)*, 2025.

6. Tianyu Fu*, **Haofeng Huang***, Xuefei Ning*, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, Guohao Dai, Huazhong Yang, Yu Wang. MoA: Mixture of Sparse Attention for Automatic Large Language Model Compression. In *Conference on Language Modeling (COLM)*, 2025.

7. Jintao Zhang, Jia Wei, **Haofeng Huang**, Pengle Zhang, Jun Zhu, Jianfei Chen. SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration. In Proceedings of the *International Conference on Learning Representations (ICLR)*, 2025.

8. Tianchen Zhao, Tongcheng Fang, **Haofeng Huang**[†], Rui Wan, Widyadewi Soedarmadji, Enshu Liu, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, Yu Wang. ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation. In Proceedings of the *International Conference on Learning Representations (ICLR)*, 2025.

9. Zhikai Zhang, Yitang Li, **Haofeng Huang**, Li Yi. FreeMotion: MoCap-Free Human Motion Synthesis with Multimodal Large Language Models. In Proceedings of the *European Conference on Computer Vision (ECCV)*, 2024.

10. **Haofeng Huang**, Li Yi. Journey into SPH Simulation: A Comprehensive Framework and Showcase. (Course Project). *Arxiv*, abs/2403.11156.

ACTIVITIES

- Host for the 12th and 13th Anniversasy Celebration of the Institute for Interdisciplinary Information Science.
- TA in Probability and Statistics in Fall 2023.
- TA in Introduction to Computer System in Spring 2024.
- Head of the Academic and Innovation Department of the Yao Class undergraduate joint committee from Summer 2024 to Autumn 2025.
- President of the Science and Technology Association of IIIS since Autumn 2025.

PROJECTS

SageAttention | *Github (2.7k stars)* **Oct 2024 – Jan 2025**

In this project, we implement quantized attention using INT8 and FP8 data types, along with precision-preserving techniques. SageAttention is implemented in CUDA and is optimized for Ampere, Ada, Hopper and Blackwell architecture. Especially, SageAttention matches the speed of FlashAttention3 on Hopper GPUs, while delivering significantly higher accuracy. This project has been adopted by a variety of companies like NVIDIA, Zhipu, Minimax, Shengshu and so on.

AWQ BF16 Support **Dec 2024 – Jan 202r**

In this PR, I add support for BF16 inference pipeline in AWQ, making it possible to serve models like Qwen2.5 72b, which would orignally encounter NaN issue due to overflow. Specifically, I implement fast INT4 to BF16 dequantization and integrate it into the GEMM/GEMV kernels, achieving similar end-to-end latency compared to the original FP16 baseline. This PR has been merged into the main branch.

SPH Fluid Simulation | *Github (170+ stars)* | *arxiv* **Nov 2023 – Jan 2024**

In this project, I implement a simulation framework that integrates various SPH fluid simulation algorithms alongside techniques for rigid-fluid coupling and high-viscosity fluid simulations. It is elected as one of the distinguished course projects of Advanced Computer Graphics, Fall 2023 and has received more than 200k views on social media.

[1] † Kernel Lead