Weiyu Huang

hwy23@mails.tsinghua.edu.cn | +86 13533198865

EDUCATION

Tsinghua University, Dept. of Computer Science and Technology **Beijing, China**
Ph.D. in Computer Science 2023.09 – 2028.06
- Advisor: Prof. Jianfei Chen
- Honors: Comprehensive Merit Scholarship (Top 5%), Tsinghua University, 2025

Tsinghua University, Dept. of Mathematical Sciences **Beijing, China**
B.S. in Applied Mathematics; Minor in Computer Science 2019.09 – 2023.06
- Academic Performance: GPA: 3.72/4.0, Rank: 23/110
- Honors: Comprehensive Merit Scholarship (Top 5%), Tsinghua University, 2023

RESEARCH FOCUS

Research focuses on **Efficient Machine Learning**, with interests in model pruning, weight and activation quantization, and optimization techniques for accelerating inference and training in large language models.

WORK EXPERIENCE

Ant Group **Beijing, China**
Research Intern 2025.10 – Present
- **Structured Pruning Framework:** Developed a deterministic, mask-only optimization framework that learns loss-driven sparse structures for LLMs, enabling lightweight pruning for state-of-the-art models.
- **Novel Pruning Design:** Formulated pruning as a deterministic optimization problem and extended the mask-value range during continuous optimization, achieving significant improvements in both perplexity and zero-shot task performance.
- **Large-Scale Experiments:** Scaled pruning to the LLaMA and Qwen model families with up to 30B parameters, achieving low performance degradation at 20–60% sparsity across downstream benchmarks and demonstrating wall-clock speedups with vLLM.
- **Outcome:** First-authored a paper accepted to ICML 2026 and led the experimental design and manuscript preparation.

Shengshu AI **Beijing, China**
Research Intern 2024.11 – 2025.07
- **Semi-Structured Sparse Training:** Researched differentiable semi-structured sparsity-aware training for LLMs and designed a novel sparsity-aware optimizer for N:M sparsity that jointly optimizes weights and sparse patterns.
- **Compression Pipeline:** Implemented end-to-end pruning pipelines across multiple model families, including LLaMA, OPT, and GPT, achieving near-lossless results on downstream suites covering reasoning, understanding, and generation.
- **Practicality Analysis:** Analyzed model performance under various training token budgets and derived a universal scaling law; conducted further experiments on fine-tuning, integration with quantization, and wall-clock inference speedup.
- **Outcome:** First-authored a paper under submission to TPAMI and led the algorithm design and manuscript preparation.

Tsinghua University **Beijing, China**
Teaching Assistant 2023.09 – 2025.06
- Served as a TA for graduate courses: Numerical Analysis (Fall 2023–Fall 2024) and Statistical Machine Learning (Spring 2025).
- Graded homework assignments and provided detailed feedback for **300+ students**; held office hours to answer questions.

PUBLICATIONS

- **Deterministic Differentiable Structured Pruning for Large Language Models** [arXiv] [GitHub]
  Weiyu Huang, et al., Jianfei Chen#, International Conference on Machine Learning (ICML), 2026
- **Pruning Large Language Models with Semi-structured Adaptive Sparse Training** [arXiv] [GitHub]
  Weiyu Huang, et al., Jianfei Chen#, AAAI Conference on Artificial Intelligence (AAAI), 2025
- **CAST: Continuous and Differentiable Semi-Structured Sparsity-Aware Training for Large Language Models**
  Weiyu Huang, et al., Jianfei Chen#, under submission to TPAMI
- **Accelerating Transformer Pre-training with 2:4 Sparsity** [arXiv] [GitHub]
  Yuezhou Hu, ***Weiyu Huang***, et al., Jianfei Chen#, International Conference on Machine Learning (ICML), 2024

COMPETITIONS

- **First Prize**, Chinese Physics Competition for Undergraduates, 2020
- **Silver Medal**, 35th Chinese Physics Olympiad for High School Students, 2018

SKILLS

Programming: MATLAB, C++, C, Python, PyTorch, R

Languages: Mandarin Chinese (native); English (fluent; TOEFL: 107/120; CET-6: 634; GRE: 331/340)

Interest: Football, Badminton, Running