Beichen Huang

beichen8@illinois.edu +1 217-418-0628 beichenhuang.github.io

Education

University of Illinois Urbana-Champaign, USA *Sept. 2025 - Expt. 2027*

- Master of Science in Computer Science. Advisor: Minjia Zhang GPA: *3.8 / 4.0*

McMaster University, Canada *Sept. 2019 - May. 2025*

- Bachelor of Engineering: Mechatronics Engineering Cumulated GPA: *11.0 / 12.0*

Research Papers

Hidden States as Early Signals: Step-level Trace Evaluation and Pruning for Efficient Test-Time Scaling[PDF] ***Preprint***

Zhixiang Liang **Beichen Huang***, Zheng Wang, Minjia Zhang* *UIUC*

- Identified **KV-cache–induced preemption and waiting** as the dominant source of end-to-end inference latency in parallel test-time scaling for LLM reasoning, and designed a **GPU memory–aware test-time scaling system** to accelerate generation.
- Extended the **vLLM scheduler** to monitor **KV cache usage** and trigger **memory-driven trace pruning** upon GPU memory saturation, proactively releasing KV cache and eliminating queuing delays.
- Co-designed algorithm and inference engine, introducing **step-level hidden-state scoring** with a lightweight MLP to monitor reasoning quality, and embedding the signal into scheduler to support real-time, memory-triggered pruning decisions.
- **Reduced 45%–70%** of the end-to-end inference latency while **improving 4%-8%** of reasoning accuracy across multiple reasoning benchmarks, by preventing trace suspension/recomputation and pruning unpromising traces early.

MiLo: Efficient Quantized MoE Inference with Mixture of Low-Rank Compensators [PDF] ***MLSys'25***

Beichen Huang, Yueming Yuan*, Zelei Shao*, Minjia Zhang* *UIUC*

- Conducted a comprehensive analysis and comparison of low-bit quantization techniques, factoring in MoE model structures and weight distributions, leading to **an adaptive and effective quantization strategy** tailored for complex MoE architectures.
- Proposed a novel calibration-free quantization algorithm, with jointly optimized low-rank compensators. Achieved **time-efficient** and nearly **loss-less 3-bit quantization** with minor memory overhead.
- Recovered **87% performance recovery** with **22% compression ratio**, and significantly outperformed prevailing quantization methods across 7 benchmark tests. Implemented a highly efficient 3-bit GeMM CUDA kernel, delivering a **1.3× speedup**.

Multidimensional Fractional Programming for Normalized Cuts [PDF] ***NeurIPS'24***

Yannan Chen, **Beichen Huang***, Licheng Zhao, Kaiming Shen* *The Chinese University of Hong Kong (ShenZhen)*

- Extended **fractional programming theory** by generalizing the quadratic transform to multidimensional ratio objectives, enabling tractable optimization of discrete multi-class normalized cut problems.
- Proposed a multidimensional fractional programming formulation for multi-class normalized cut clustering, enabling direct optimization of discrete 0–1 assignments, and leading to lower optimization objectives.
- Achieved lower NCut objectives across 8 benchmark datasets, and delivered 73% speedup over baseline methods.

Professional Experience

UIUC: Research Intern, Advisor: Minjia Zhang *May 2024 - Dec. 2024*

- Improved efficiency and accuracy of the large-scale model inference through algorithm and system co-design, and designing cost-effective machine learning systems.

The Chinese University of Hong Kong (ShenZhen): Research Intern, Advisor: Kaiming Shen *Sept. 2023 - May 2024*

- Addressed fractional programming problems, and applied the method in solving complex challenges in machine learning.

Work Experience

Teaching Assistant CS105:Intro Computing *UIUC, Aug. 2025 - Dec. 2025*

Teaching Assistant Embedded System *McMaster University, Dec. 2021 - May 2022*

Skills

Programming: Python, CUDA, C, C++ **Tools & Frameworks:** PyTorch, vLLM