Yuezhou Hu

yuezhouhu@berkeley.edu | (341)333-8818 | yuezhouhu.github.io

Research Experience

PhD Student, Berkeley Artificial Intelligence Research Lab (BAIR), University of California, Berkeley, Advisor: Prof. Kurt Keutzer — Sep. 2025 – Present

Research Intern, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Advisor: Prof. Tuo Zhao — Jul. 2024 – Aug. 2024

Research Intern, Statistical Artificial Intelligence & Learning Group, Tsinghua University, Advisor: Prof. Jianfei Chen, Prof. Jun Zhu — May 2022 – Jul. 2025

Education

University of California, Berkeley, Computer Science Ph.D Student — Sep. 2025 – Present

- Department of Electrical Engineering and Computer Sciences

Tsinghua University, B.S. in Computer Science — Sep. 2021 – Jul. 2025

- Department of Computer Science and Technology

Research Interests

My research interests primarily focus on **efficient machine learning**, particularly efficient training and inference. Recently, I am interested in:

- Diffusion Language Model; Speculative Decoding; Large Reasoning Model; Dynamic Sparse Training

Publications

3 first/cofirst-author papers, *equal contribution

- **ParallelBench: Understanding the Trade-offs of Parallel Decoding in Diffusion LLMs** [arXiv]
  Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Hooper, ***Yuezhou Hu***, Hyung Il Koo, Nam Ik Cho, Kangwook Lee
  International Conference on Learning Representations (ICLR), 2026
- **AdaSPEC: Selective Knowledge Distillation for Efficient Speculative Decoders** [arXiv] [OpenReview]
  Yuezhou Hu*, Jiaxin Guo*, Xinyu Feng, Tuo Zhao
  Neural Information Processing Systems (NeurIPS), 2025 (**Spotlight**)
- **Pruning Large Language Models with Semi-Structural Adaptive Sparse Training** [arXiv] [Project page]
  Weiyu Huang, ***Yuezhou Hu***, Guohao Jian, Jun Zhu, Jianfei Chen
  AAAI Conference on Artificial Intelligence (AAAI), 2025
- **S-STE: Continuous Pruning Function for Efficient 2:4 Sparse Pre-training** [arXiv] [OpenReview] [Project page]
  Yuezhou Hu, Jun Zhu, Jianfei Chen
  Neural Information Processing Systems (NeurIPS), 2024
- **Accelerating Transformer Pre-training with 2:4 Sparsity** [arXiv] [OpenReview] [PDF] [Project page]
  Yuezhou Hu, Kang Zhao, Weiyu Huang, Jianfei Chen, Jun Zhu
  International Conference on Machine Learning (ICML), 2024

Preprints

- **Residual Context Diffusion Language Models** [arXiv] [Code]
  Yuezhou Hu*, Harman Singh*, Monishwaran Maheswaran*, Haocheng Xi, Coleman Hooper, Jintao Zhang, Aditya Tomar, Michael W. Mahoney, Sewon Min, Mehrdad Farajtabar, Kurt Keutzer, Amir Gholami, Chenfeng Xu
- **Arbitrage: Efficient Reasoning via Advantage-Aware Speculation** [arXiv] [Project page]

Monishwaran Maheswaran[*], Rishabh Tiwari[*], ***Yuezhou Hu***[*], Kerem Dilmen, Coleman Hooper, Haocheng Xi, Nicholas Lee, Mehrdad Farajtabar, Michael Mahoney, Kurt Keutzer, Amir Gholami

- **A Survey of Efficient Attention Methods: Hardware-efficient, Sparse, Compact, and Linear Attention**
  Jintao Zhang, Rundong Su, Chunyu Liu, Jia Wei, Ziteng Wang, Pengle Zhang, Haoxu Wang, Huiqiang Jiang, Haofeng Huang, Chendong Xiang, Haocheng Xi, Shuo Yang, Xingyang Li, ***Yuezhou Hu***, Tianyu Fu, Tianchen Zhao, Yicheng Zhang, Boqun Cao, Youhe Jiang, Chang Chen, Kai Jiang, Huayu Chen, Min Zhao, Xiaoming Xu, Yi Wu, Fan Bao, Jun Zhu, Jianfei Chen
- **CAST: Continuous and Differentiable Semi-Structured Sparsity-Aware Training for Large Language Models** [arXiv]
  Weiyu Huang, ***Yuezhou Hu***, Jun Zhu, Jianfei Chen
- **Identifying Sensitive Weights via Post-quantization Integral** [arXiv]
  Yuezhou Hu, Weiyu Huang, Zichen Liang, Chang Chen, Jintao Zhang, Jun Zhu, Jianfei Chen

Technical Skills

Deep learning programming: Python, Pytorch

GPU Programming: OpenAI Triton, C++, CUDA

Others: Docker, Django, Rust

Language Skills: TOEFL 104 (R:28/L:27/S:23/W:26)

Honors

- Tsinghua Academic Preeminence Scholarship — Fall 2024
- 84 Future Innovation Scholarship — Fall 2024