Haocheng Xi

Berkeley AI Research, University of California, Berkeley | xihc@berkeley.edu
haochengxi.github.io

EDUCATION

University of California, Berkeley — Berkeley, CA
Ph.D. in Computer Science, Berkeley AI Research (BAIR) — 09/2024 – Present
Advisor: Prof. Kurt Keutzer

Tsinghua University — Beijing, China
B.Eng. in Computer Science & Technology — 09/2020 – 06/2024
Institute for Interdisciplinary Information Sciences (IIIS)
Yao Class, led by Prof. Andrew C.C. Yao

University of Washington — Seattle, WA
Visiting Student, Paul G. Allen School of Computer Science & Engineering — 02/2023 – 08/2023
Advisor: Prof. Sheng Wang

Beijing No.8 High School — Beijing, China
Experimental class for gifted and talented young, Excellent Graduate — 09/2015 – 07/2020

RESEARCH INTERESTS

My research interests lie in efficient machine learning, such as quantization and sparsity. I aim to push the boundaries of how we can effectively compress and accelerate deep learning models while maintaining their accuracy and robustness.

SELECTED PUBLICATIONS

Efficient Video Generation

Sparse VideoGen: Accelerating Video Diffusion Transformers with Spatial-Temporal Sparsity

Haocheng Xi, Shuo Yang*, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, Song Han*
International Conference on Machine Learning (ICML), 2025. [arxiv] [code] [website] [poster]

Sparse VideoGen 2: Accelerate Video Generation with Sparse Attention via Semantic-Aware Permutation

Shuo Yang, Haocheng Xi*, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, Jianfei Chen, Song Han, Kurt Keutzer, Ion Stoica*
Conference on Neural Information Processing Systems (NeurIPS), 2025. [arxiv] [code] [website]
Selected as **Spotlight Paper** in NeurIPS 2025.

Radial Attention: $O(n \log n)$ Sparse Attention with Energy Decay for Long Video Generation

Xingyang Li, Muyang Li*, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, Song Han*
Conference on Neural Information Processing Systems (NeurIPS), 2025. [arxiv] [code] [website]

StreamDiffusionV2: A Streaming System for Dynamic and Interactive Video Generation

Tianrui Feng, Zhi Li, Shuo Yang, Haocheng Xi, Muyang Li, Xiuyu Li, Lvmin Zhang, Keting Yang, Kelly Peng, Song Han, Maneesh Agrawala, Kurt Keutzer, Akio Kodaira, Chenfeng Xu
Arxiv, 2025. [arxiv] [code] [website]

Efficient Language Model

COAT: Compressing Optimizer states and Activation for Memory-Efficient FP8 Training
Haocheng Xi, Han Cai, Ligeng Zhu, Yao Lu, Kurt Keutzer, Jianfei Chen, Song Han
International Conference on Learning Representations (ICLR), 2025. [arxiv] [code] [website]

Jetfire: Efficient and Accurate Transformer Pretraining with INT8 Data Flow and Per-Block Quantization
Haocheng Xi, Yuxiang Chen, Kang Zhao, Kai Jun Teh, Jianfei Chen, Jun Zhu
Internation Conference on Machine Learning (ICML), 2024. [arxiv] [code] [poster]
Selected as **Spotlight Paper** in ICML 2024.

Training Transformers with 4-bit Integers
Haocheng Xi, Changhao Li, Jianfei Chen, Jun Zhu
Conference on Neural Information Processing Systems (NeurIPS), 2023. [arxiv] [code]

QuantSpec: Self-Speculative Decoding with Hierarchical Quantized KV Cache
Rishabh Tiwari, Haocheng Xi*, Aditya Tomar, Coleman Hooper, Sehoon Kim, Maxwell Horton, Mahyar Najibi, Michael W. Mahoney, Kurt Keutzer, Amir Gholami*
International Conference on Machine Learning (ICML), 2025. [arxiv] [poster]

Oscillation-Reduced MXFP4 Training for Vision Transformers
Yuxiang Chen, Haocheng Xi, Jun Zhu, Jianfei Chen
International Conference on Machine Learning (ICML), 2025. [arxiv]

INTERNSHIP EXPERIENCE

Nvidia Research, Research Intern 03/2024 – 08/2024, 2025/05 – Present
Advisor: Prof. Song Han

COAT: Compressing Optimizer states and Activation for Memory-Efficient FP8 Training

- Introduced COAT, a framework that quantizes optimizer states and activations to FP8 precision, significantly reducing memory usage during large-scale model training.
- Proposed Dynamic Range Expansion for Optimizer states and Mixed-Granularity Activation Quantization, achieving outstanding accuracy and efficiency.
- Achieved a 1.54× reduction in training memory footprint and a 1.43× speedup compared to BF16 training, also doubled the training batch size to utilize GPU better.
- Training loss curve and downstream task performance were consistent with BF16 training, across language models and vision language models.

HONORS

Fellowship of Tsinghua Xuetang Talents Program Among top 300 / 3000 Tsinghua students each year
Athletic Excellence Scholarship In 2022
First Prize of National Senior High School Mathematics Competition In 2019

SKILLS

Language: TOFEL: Total 110 (Reading 29, Listening 29, Speaking 24, Writing 28)
GRE: Quantitive 170, Verbal 158, Writing 4.0
Programming and Software: Python, CUDA, C++, Bash, Git, LaTeX
Deep Learning Package: PyTorch, Transformers, Triton, PEFT, TransformerEngine, VeRL
Conference review: ICML (2025), NeurIPS (2024, 2025), ICLR (2025, 2026)