# Wentao Guo

wentaoguo.me | wg0420@princeton.edu

## EDUCATION

**Princeton University**

- Ph.D. in Computer Science, advised by Prof. Tri Dao — 09/2024 - 05/2029

**Cornell University**

- Master of Engineering in Computer Science, GPA: 3.993 — 06/2022 - 12/2023
- B.S. in Computer Science with Honors, Magna Cum Laude, GPA: 3.890 — 09/2018 - 05/2022

## PUBLICATION & MANUSCRIPT

*** denotes equal contribution.**

- **Wentao Guo**, Mayank Mishra, Xinle Cheng, Ion Stoica, Tri Dao. "SonicMoE: Accelerating MoE with IO and Tile-aware Optimizations." *In arXiv preprint 2025.* [paper] [code]
- **Wentao Guo**, Jikai Long, Yimeng Zeng, Zirui Liu, Xinyu Yang, Yide Ran, Jacob R. Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, Beidi Chen, Zhaozhuo Xu. "Zeroth-Order Fine-Tuning of LLMs with Transferable Static Sparsity." *In ICLR'25.* [paper]
- Yuheng Wu, **Wentao Guo**, Zirui Liu, Heng Ji, Zhaozhuo Xu, Denghui Zhang. "How large language models encode theory-of-mind: a study on sparse parameter patterns," *In npj Artificial Intelligence 2025.* [paper]
- A. Feder Cooper*, **Wentao Guo***, Khiem Pham*, Tiancheng Yuan, Charlie F. Ruan, Yucheng Lu, Christopher De Sa. "Coordinating Distributed Example Orders for Provably Accelerated Training." *In NeurIPS'23.* [paper] [code]
- Yucheng Lu, **Wentao Guo**, and Christopher De Sa. "GraB: Finding Provably Better Data Permutations than Random Reshuffling." *In NeurIPS'22.* [paper]
- **Wentao Guo***, Andrew Wang*, Bradon Thymes, Thorsten Joachims. "Ranking with Slot Constraints." *In KDD'24.* [paper] [code]
- Tao Yu*, **Wentao Guo***, Jianan Canal Li*, Tiancheng Yuan*, Christopher De Sa. "MCTensor: A High-Precision Deep Learning Library with Multi-Component Floating-Point." *In the HAET workshop at ICML'22.* [paper] [code]
- Yann Hicke, Abhishek Masand, **Wentao Guo**, Tushaar Gangavarapu. "Assessing the efficacy of large language models in generating accurate teacher responses." *In the BEA workshop at ACL'23.* [paper]

## RESEARCH EXPERIENCE

**Research Assistant**

*Prof. Tri Dao's Lab, Princeton University* — 03/2025 - 12/2025

- Efficient MoE implementation on Hopper GPUs [**arXiv**]
  - Developed an efficient MoE implementation that achieves 1.86x training throughput and reduces the activation memory footprint by 45% compared to a prior SOTA MoE baseline on H100 GPU. Proposed a novel token rounding that eliminates compute wasted on padding by Grouped GEMM kernels and improves 16% training throughput for sparse MoEs while preserving inference quality.

**Research Assistant**

*Prof. Beidi Chen's Lab, Carnegie Mellon University* — 06/2023 - 05/2024

- Memory-efficient LLM fine-tuning [**ICLR'25**]
  - Combined sparse fine-tuning techniques with zeroth-order (ZO) optimization methods to personalize LLM fine-tuning while respecting memory constraints on devices (8 GB).

**Research Assistant**
*Prof. Christopher De Sa's Lab, Cornell University* 06/2021 - 05/2023

- Centralized example ordering for improved optimizer convergence [**NeurIPS'22**]
  - Collaborated to develop the Gradient Balancing (GraB) algorithm that leverages per-example gradients from the prior epoch to determine the example order in the next epoch, with a provably faster convergence rate than the random reshuffling (RR) method.
- Distributed example ordering for improved optimizer convergence [**NeurIPS'23**]
  - Designed the Coordinated Distributed GraB (CD-GraB) algorithm that generalizes the GraB algorithm to the distributed setting without centralized access to all data examples.
- High-precision floating-point computation for learning in hyperbolic space [**HAET@ICML'22**]
  - Developed the MCTensor library that implements high-precision Multiple-Component Format (MCF) algorithms with PyTorch-compatible interfaces, and the HTorch library that integrates hyperbolic space optimization pipelines with MCF algorithms.

**Research Assistant**
*Prof. Thorsten Joachims's Lab, Cornell University* 06/2022 - 02/2023

- Ranking with slot constraints [**KDD'24**]
  - Proposed the MatchRank algorithm that recommends a shortlist of relevant candidates while respecting the set of slot constraints defined by decision-makers.

## TEACHING EXPERIENCE

- **Graduate Teaching Assistant, Princeton**
  - COS 324 Introduction to Machine Learning 08/2025 - 01/2026
- **Graduate Teaching Assistant, Cornell**
  - CS 4787 Principles of Large-Scale Machine Learning Systems 08/2023 - 12/2023
  - CS 4780 Intro to Machine Learning 01/2023 - 05/2023

## ACADEMIC SERVICE

- **Reviewer**: NeurIPS'23, ICLR'24, ICML'24, KDD'24, NeurIPS'24, Journal of DMLR, ACL RR (June 2024, October 2024), ICLR'25, COLM'25, NeurIPS'25, ICLR'26

## HONORS

- Princeton University Graduate Fellowship
- Cornell Engineering Honor Society membership