Zhenghao Zhao

ichbill27@gmail.com | +1 (312) 340-1603 | Homepage | Google Scholar | GitHub | LinkedIn

RESEARCH SUMMARY

Data Efficient Foundation Model Training: Developing methods to select, curate, synthesize, and distill training data to improve model capability under limited data and compute. Applications span language, vision, multimodal models, and robotics.
Research Methods: Data selection and curation, synthetic data generation, dataset distillation, continual training, and knowledge distillation.

EDUCATION

University of Illinois Chicago, Ph.D. in Computer Science — August, 2020 - May, 2027 (expected)
Research Assistant, Advisor: Prof. Yan Yan.

Illinois Institute of Technology, Master of Science in Computer Science — August, 2018 - May, 2020

EMPLOYMENT EXPERIENCE

Amazon – Applied Scientist Intern — May, 2026 - August, 2026
Foundation Model Distillation — Santa Cruz, CA

- Developed GRAFT, a continual multi-teacher distillation framework that integrates heterogeneous vision foundation models across 2D, 3D, and vision-language tasks into a single extensible foundation model.
- Introduced teacher-specific readout tokens and an image–text relational loss to transfer heterogeneous representations while preserving previously acquired capabilities.
- Evaluated the unified encoder across image understanding, dense prediction, 3D perception, and vision-language benchmarks, achieving competitive performance of specialized teacher models across tasks with a single encoder.

Argonne National Laboratory – Research Intern — May, 2025 - August, 2025
Large-scale Foundation Model Training — Lemont, IL

- Built and benchmarked FSDP training pipelines for vision and language foundation models on Aurora cluster.
- Scaled FSDP with tensor parallelism to 64 compute nodes and analyzed scaling efficiency.
- Benchmarked FSDP against DeepSpeed ZeRO 3 across large model training workloads.

Argonne National Laboratory – Research Intern — May, 2023 - August, 2023
Distributed Training — Lemont, IL

- Benchmarked distributed training frameworks across multiple deep learning workloads.

SELECTED PUBLICATIONS

* Equal contribution. Full publication list: Google Scholar

[1] **Zhenghao Zhao**, Gaowen Liu, Zhiling Lan, Yan Yan. "Fisher-Guided Submodular Data Selection for Continual Pre-Training of Large Language Models." Preprint, 2026. [Paper]
A Fisher guided submodular data selection framework for LLM continual pretraining that balances knowledge acquisition and retention, achieving 10× token efficiency over replay baselines.

[2] **Zhenghao Zhao**, Chi Zhang, Qingshuang Chen, Yelin Kim. "GRAFT: Growing Agglomerative Foundation Models via Continual Teacher Distillation." Preprint, 2026. [Paper]
A continual multi teacher distillation framework that incrementally integrates 2D, 3D, and vision language capabilities into a single extensible foundation model.

[3] **Zhenghao Zhao**, Yasutomo Kawanishi, Huihuo Zheng, Michael E. Papka, Yan Yan. "Query-based Submodular Data Selection for Efficient Instruction Tuning." Preprint, 2026.
A submodular data selection framework for LLM instruction tuning that balances target relevance and diversity, matching full data performance with only a 0.5% data budget.

[4] **Zhenghao Zhao**, Haoxuan Wang, Junyi Wu, Yuzhang Shang, Gaowen Liu, Yan Yan. "Efficient Multimodal Dataset Distillation via Generative Models." In the Neural Information Processing Systems (**NeurIPS**), 2025. [Paper]
A generative multimodal dataset distillation method enabling scalable dataset distillation with 18× faster processing.

[5] **Zhenghao Zhao**, Haoxuan Wang, Kai Wang, Yuzhang Shang, Yuan Hong, Yan Yan. "Distill Videos into Images." Preprint, 2025. [Paper]
A video dataset distillation framework that distills videos into representative single frames, reducing optimization complexity while achieving up to 5.3% higher accuracy than prior methods.

[6] **Zhenghao Zhao***, Haoxuan Wang*, Yuzhang Shang, Kai Wang, Yan Yan. "Distilling Long-tailed Datasets." In the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2025. [Paper]
A dataset distillation method for long tailed data that mitigates class imbalance bias, matching full dataset performance with a synthetic set only 5% the original data size.

[7] **Zhenghao Zhao**, Yuzhang Shang, Junyi Wu, Yan Yan. "Dataset Quantization with Active Learning-based Adaptive Sampling." In Proceedings of the *European Conference on Computer Vision* (**ECCV**), 2024. [Paper]
An adaptive dataset compression method motivated by class dependent sensitivity to sample quantity, using active learning to optimize sample allocation across classes.

ROBOTICS EXPERIENCE

Data Selection for VLA Fine-Tuning August, 2025 - present

- Set up and validated a Franka robotic arm and collection of expert demonstration trajectories.
- Developing data selection methods for VLA fine-tuning that use target robot or task demonstrations to retrieve relevant trajectories from large robotics datasets, reducing the amount of task specific data collection required.

TECHNICAL SKILLS

Foundation Model Training: PyTorch, Hugging Face Transformers, Diffusers, FSDP, DeepSpeed, DDP, tensor parallelism, TorchTitan

Models & Modalities: LLMs, vision transformers, vision-language models, diffusion models, video models, vision-language-action models

Large-scale Training: Multi-GPU and multi-node distributed training, model parallelism, HPC training

Robotics: Franka robot arm, VLA fine-tuning, robot demonstration collection and evaluation

Programming & Systems: Python, Linux, Git

PROFESSIONAL SERVICE & TALKS

Conference Reviewer / Program Committee: ICLR 2026-2027; CVPR 2025–2026; NeurIPS 2025-2026; ECCV 2024; ICMR 2025–2026; VOCVALC 2023.

Local Co-Chair: ICMR 2025.

Lightning Talk: "Distilling Long-tailed Datasets." Midwest Machine Learning Symposium, June 2025.

Teaching: Teaching Assistant, Illinois Institute of Technology, 2019-2023; Machine Learning, Natural Language Processing, and Probabilistic Graphical Models.