Yihua Zhang

Website Google Scholar GitHub
✉ zhan1908@msu.edu ☎ (517)980-3880

RESEARCH INTERESTS

Foundation Models (LLM/Diffusion Model): Trustworthiness (Machine Unlearning, Alignment, Privacy), Efficiency (Model Sparsification, MoE, Memory-Efficient Fine-Tuning, Parameter-Efficient Fine-Tuning)

Machine Learning: Bi-Level Optimization, Zeroth-Order Optimization, Invariant Risk Minimization

EDUCATION

Michigan State University (MSU) Jan. 2022 - Present
Ph.D. Candidate, Computer Science Advisor: Prof. Sijia Liu

Huazhong University of Science and Technology (HUST) Sep. 2015 - Jun. 2019
B.S.c, Automation Qiming Honor College of HUST
National Scholarship * 2 (Top 0.2%, highest undergraduate honor in China) 2016 & 2017

PROFESSIONAL EXPERIENCE

Meta AI Sep. 2024 - Present
Research Scientist Intern, Supervisor: Dr. Xi Liu
Project: Multi-Agent LLM for recommendataion system.

Cisco Research Dec. 2023 - Aug. 2024
Research Intern, Supervisor: Dr. Gaowen Liu
Project: Machine Unlearning for Foundation Models: LLMs, Diffusion Models, and MoEs.

Amazon AWS AI Lab May. 2023 - Aug. 2023
Applied Scientist Intern, Supervisor: Dr. Zhou Ren, Dr. Tian Lan
Project: In-context learning for vision generative models: design, training, and generalization study.

JD AI Research (JD Explore Academy) Jan. 2021 - Aug. 2021
Research Intern, Supervisor: Dr. Jinfeng Yi
Project: Model robustness, fairness, and explanability co-design.

HONORS

Research Awards

- **MLCommons Rising Star Award** [ML Commons News] 2024
- **UAI 2022 Best Paper Runner-up Award** [Certificate] 2022
- **CVPR Outstanding Reviewer Award * 2** [2023] & [2024]
- **NeurIPS Top Reviewer Award * 2** [2022] & [2023]

Conference Travel Grants

- **NeurIPS Scholar Award * 2** 2022 & 2023
- **AAAI Travel Grant Award** 2023
- **ICML Travel Grant Award** 2022
- **UAI Student Scholarship** 2022

PUBLICATIONS

Yihua Zhang has co-authored over 20 papers in top-tier machine learning and computer vision venues (NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, *etc.*) and published over 10 first-authored papers. Below are his publications: *** indicates an equal contribution, and ‡ denotes the author is his mentee.**

▷ Thrust I. Trustworthy Machine Learning

NeurIPS'24 D&B Track: **Y. Zhang**, C. Fan, Y. Zhang, Y. Yao, J. Jia, G. Zhang, G. Liu, R. Kompella, X. Liu, S. Liu, “UnlearnCanvas: A Stylized Image Dataset to Benchmark Machine Unlearning for Diffusion Models and Beyond”, [PDF], [Code], [Website], [Demo], [Dataset], [Benchmark].

NeurIPS'24: J. Jia, J. Liu, **Y. Zhang**, P. Ram, N. Baracaldo, S. Liu, “WAGLE: Strategic Weight Attribution for Effective and Modular Unlearning in Large Language Models”, [PDF], .

NeurIPS'24: Y. Zhang, X. Chen, J. Jia, **Y. Zhang**, C. Fan, J. Liu, M. Hong, K. Ding, S. Liu, “Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models”, [PDF], [Code].

EMNLP'24 Main: J. Jia, **Y. Zhang**, Y. Zhang, J. Liu, B. Runwal, J. Diffenderfer, B. Kailkhura, S. Liu, "SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning", [PDF], [Code].

ECCV'24: Y. Zhang, J. Jia, X. Chen, A. Chen[‡], **Y. Zhang**, J. Liu, K. Ding, S. Liu, "To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy To Generate Unsafe Images ... For Now", The 18th European Conference on Computer Vision , [PDF], [Code], [Website].

ICCV'23 Oral: **Y. Zhang**, R. Cai, T. Chen, G. Zhang, P.-Y. Chen, H. Zhang, S. Chang, W. Zhang, S. Liu, "Robust Mixture-of-Expert Training for Convolutional Neural Networks", International Conference on Computer Vision 2023, [PDF], [Code].

ICLR'24 Spotlight: C. Fan[‡], J. Liu, **Y. Zhang**, E. Wong, D. Wei, S. Liu, "Salun: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation", 12th International Conference on Learning Representations, [PDF], [Code].

ICLR'23: **Y. Zhang**, P. Sharma, P. Ram, M. Hong, K. R. Varshney, S. Liu, "What Is Missing in IRM Training and Evaluation? Challenges and Solutions", 11th International Conference on Learning Representations, [PDF], [Code].

ICLR'23: B. Hou, **Y. Zhang**, J. Jia, G. Zhang, Y. Zhang, S. Liu, S. Chang, "TextGrad: Advancing Robustness Evaluation in NLP by Gradient-Driven Optimization", 11th International Conference on Learning Representations, [PDF], [Code].

ICML'23: P. Khanduri, I. Tsaknakis, **Y. Zhang**, J. Liu, S. Liu, J. Zhang, M. Hong, "Linearly Constrained Bilevel Optimization: A Smoothed Implicit Gradient Approach", 40th International Conference on Machine Learning , [PDF].

NeurIPS'22: **Y. Zhang**, G. Zhang*, Y. Zhang, W. Fan, Q. Li, S. Liu, S. Chang, "Fairness Reprogramming", 36th Conference on Neural Information Processing Systems, [PDF], [Code], [Website].

UAI'22 Best Paper Runner-Up Award: G. Zhang, S. Lu, **Y. Zhang**, X. Chen, P.-Y. Chen, Q. Fan, L. Martie, M. Hong , S. Liu, "Distributed Adversarial Training to Robustify Deep Neural Networks at Scale", 38th Conference on Uncertainty in Artificial Intelligence, [PDF], [Code], [Award].

ICML'22: **Y. Zhang**, G. Zhang, P. Khanduri, M. Hong, S. Chang, S. Liu, "Fast-BAT: Revisiting and Advancing Fast Adversarial Training through the Lens of Bi-level Optimization", 39th International Conference on Machine Learning, [PDF], [Code], [Talk].

CVPR'22: **Y. Zhang***, T. Chen*, Z. Zhang*, S. Chang , S. Liu , Z. Wang, "Quarantine: Sparsity Can Uncover the Trojan Attack Trigger for Free", 3Computer Vision and Pattern Recognition Conference 2022, [PDF], [Code], [Website].

▷ Thrust II. Efficient Machine Learning

ICML'24: **Y. Zhang**, P. Li, J. Hong, J. Li, Y. Zhang, W. Zheng, P.-Y. Chen, J. Lee, W. Yin, M. Hong, Z. Wang, S. Liu, and T. Chen, "Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark", The Forty-first International Conference on Machine Learning, [PDF], [Code], [Website].

IEEE Signal Process. Mag.'24: **Y. Zhang**, P. Khanduri, I. Tsaknakis, Y. Zhang, M. Hong, S. Liu, "An Introduction to Bi-level Optimization: Foundations and Applications in Signal Processing and Machine Learning", IEEE Signal Processing Magazine, vol. 41, no. 1, pp. 38-59, 2024, [PDF] (Feature Article).

ICLR'24: A. Chen[‡], Y. Zhang, J. Jia, J. Diffenderfer, J. Liu, K. Parasyris, **Y. Zhang**, Z. Zhang, B. Kailkhura, S. Liu, "DeepZero: Scaling up Zeroth-Order Optimization for Deep Model Training", 12th International Conference on Learning Representations, [PDF], [Code].

IEEE J. Sel. Topics Signal Process.'24: H. Li, S. Zhang, **Y. Zhang**, M. Wang, S. Liu, P.-Y. Chen, "How Does Promoting the Minority Fraction Affect Generalization? A Theoretical Study of One-Hidden-Layer Neural Network on Group Imbalance", IEEE Journal of Selected Topics in Signal Processing, 2024, [PDF].

NeurIPS'23: **Y. Zhang**, Y. Zhang, A. Chen[‡], J. Jia, J. Liu, G. Liu, S. Chang, M. Hong, S. Liu, "Selectivity Drives Productivity: Efficient Dataset Pruning for Enhanced Transfer Learning", 37th Conference on Neural Information Processing Systems, [PDF], [Code], [Website].

CVPR'23: A. Chen[‡], Y. Yao, P.-Y. Chen, **Y. Zhang**, S. Liu, "Understanding and Improving Visual Prompting: A Label-Mapping Perspective", 2023 Conference on Computer Vision and Pattern Recognition, [PDF], [Code].

CVPR'23: H. Zhuang[‡], **Y. Zhang**, S. Liu, "A Pilot Study of Query-Free Adversarial Attack against Stable Diffusion", 2023 Conference on Computer Vision and Pattern Recognition, [PDF], [Code].

NeurIPS'22: **Y. Zhang**, Y. Yao, P. Ram, P. Zhao, T. Chen, M. Hong, Y. Wang, S. Liu, "Advancing Model Pruning via Bi-level Optimization", 36th Conference on Neural Information Processing Systems, [PDF], [Code], [Website].

Under Review: **Y. Zhang**, H. Li, Y. Yao, A. Chen, P.-Y. Chen, S. Zhang, M. Wang, S. Liu, “Visual Prompting Reimagined.

TUTORIALS AND INVITED TALKS

- **Tutorial** at AAAI 2024, Topic: Zeroth-Order Machine Learning: Fundamental Principles and Emerging Applications in Foundation Models, [Website] Feb. 2024
- **Tutorial** at AAAI 2023, Topic: Bi-level Optimization in Machine Learning: Foundations and Applications, [Website] Feb. 2023
- **Invited Talk** as Lecture Speaker, Department of Electrical and Computer Engineering, University of Minnesota (UMN) Apr. 2022
- **Invited Talk** at INFORMS Annual Conference, Department of Computer Science Oct. 2022
- **Invited Talk** as Lecture Speaker, Department of Computer Science, UCSB Apr. 2022

SERVICES

Conference Volunteer: AAAI’23, ICLR’23

Conference Reviewer: ICLR’22/23/24, NeurIPS’21/22/23/24, ICML’22/23/24, CVPR’23/24, ICCV’23, ECCV’24, AISTATS’22/23, UAI’22/23

Journal Reviewer: JMLR, IEEE TPAMI, IEEE T-IFS, TMLR

Workshop Student Chair: Workshop Series: `AdvML: New Frontiers in Adversarial Machine Learning` [ICML’22], [ICML’23], [NeurIPS’24].

MENTEES

Zheyu Fan (Undergraduate, HKU)	May. 2024 - Current
Xiaoxuan Qiu (Undergraduate, Tsinghua)	May. 2024 - Current
Yuhao Sun (Undergraduate, USTC)	May. 2024 - Current
Hanhui Wang (Master student, USC)	May. 2024 - Current
Zheng Han (Undergraduate, ZJU)	May. 2024 - Current
Chongyu Fan (Undergraduate, HUST) — [[ICLR’24 Spotlight]]	May. 2023 - Current
Haomin Zhuang (Undergraduate, SCUT) — [[CVPRW’23]]	Dec. 2022 - Current
Aochuan Chen (Undergraduate, Tsinghua University) — [[CVPR’23], [ICLR’24]]	Oct. 2022 - Oct. 2023
Mohammad Jafari (Undergraduate, Sharif University of Technology) — [[ICASSP’24]]	May. 2023 - Oct. 2023

Last updated: October 9, 2024.