

YIHUA ZHANG

zhan1908@msu.edu | (+1) 5179803880 | [Website](#) | [Blogs](#) | [Google Scholar Citation 1805](#)

EDUCATION

Michigan State University, East Lansing, USA 2022 - Present

Ph.D. Candidate in Computer Science and Engineering

- Advisor: [Dr. Sijia Liu](#)
- Ph.D. Committee: [Dr. Xiaoming Liu](#), [Dr. Anil K. Jain](#), [Dr. Kush R. Varshney](#)
- Research Focus: Efficient Machine Learning, Trustworthy Machine Learning, Multimodal Modeling

RWTH Aachen University, Aachen, Germany 2019 - 2021

M.Sc. in Automation

Huazhong University of Science and Technology, Wuhan, China 2015 - 2019

B.Sc. in Automation, [Qiming Honor College of HUST](#)

HONORS

- **IBM PhD Fellowship 2024** (\$40,000, 24 recipients selected worldwide) 2025
- **Fitch H. Beach Award** (\$2,000, highest honor for MSU Ph.D. students) 2025
- **CPAL Rising Star Award** (hosted by Stanford University, 15 recipients selected worldwide) 2025
- **TrustAI Rising Star Award** by NeurIPS 2025 Lock-LLM Workshop 2025
- **MLCommons Rising Star Award** (hosted by NVIDIA, 41 recipients selected worldwide) 2024
- **UAI 2022 Best Paper Runner-up Award** 2022
- CVPR Outstanding Reviewer Award x2 2023 & 2024
- NeurIPS Top Reviewer Award x2 2022 & 2023
- NeurIPS Scholar Award x2 2022 & 2023
- AAAI Travel Grant Award 2023
- ICML Travel Grant Award 2022
- UAI Student Scholarship Award 2022
- National Scholarship (Top 0.2%; highest undergraduate honor in China) x2 2016 & 2017

RESEARCH HIGHLIGHTS

My research advances fundamental optimization methods, including bi-level, second-order, zeroth-order, sharpness-aware optimization, to build trustworthy and efficient AI systems. While application-driven, my work is grounded in algorithmic innovation, aiming to bridge theoretical foundations with real-world deployment. See my [research roadmap](#).

1. LLM ZERO-ORDER FINE-TUNING AT INFERENCE-LEVEL MEMORY COST

I envision a scalable future of LLM tuning that leverages commodity hardware by relying solely on *inference-level memory*. My work [\[ICML'24\]](#) introduces the first comprehensive benchmark and toolbox for zeroth-order LLM optimization, demonstrating that gradient-free tuning can match backprop-based methods when combined with practical tactics and theoretical insights, also highlighted by my [AAAI'24 tutorial](#).

2. MACHINE UNLEARNING AND KNOWLEDGE EDITING ON LLMs AND DIFFUSION MODELS

I am among the early researchers shaping the field of *machine unlearning*, developing methods that span a broad range of model families, including diffusion models [\[ICLR'24\]](#), [\[ECCV'24\]](#), [\[NeurIPS'24\]](#), [\[NeurIPS'24\]](#),

LLMs [NeurIPS'24], [ICML'25], [ICML'25], and MoE-LLMs [ACL'25]. My work introduces principled algorithms, certified evaluation protocols, and new benchmark datasets for trustworthy unlearning, and uncovers a pervasive yet overlooked issue of 'fake unlearning', see my invited [talk](#) at Shanghai AI Lab.

3. BI-LEVEL OPTIMIZATION FOUNDATIONS FOR LARGE-SCALE MACHINE LEARNING PROBLEMS

Many fundamental machine learning tasks—ranging from model pruning and fairness to adversarial training and transfer learning—can be elegantly cast as *bi-level optimization* (BLO) problems. My research shows that with carefully designed solvers, BLO formulations can lead to significant performance gains [ICLR'23], [NeurIPS'22], [ICML'22]. I delivered a tutorial at AAAI'23, providing both theoretical foundations and practical tools to advance the understanding and application of BLO methods in modern ML.

PROFESSIONAL EXPERIENCE

Bytedance Seed - Multimodal Interaction and World Model, San Jose, USA Sep. 2025 - Present
Student Researcher at [Shen Yan](#)'s team, working on the next generation Seed-VLM pretraining.

Project: Optimization on Token-Efficient Multi-Modal Fusion in VLM Pretrain;

- Designing novel optimizers to improve convergence and stability in large-scale multimodal training;
- Improving token efficiency in VLMs to maximize representation power and reduce redundancy;

Meta AI, Full-Time, Sunnyvale, USA May 2025 - Aug. 2025

Project: Prototyping and Scalable Training of Next-Generation Multi-Modal Ads Ranking Foundation Model;

- Prototyping industry-level next-generation ranking foundation model with multi-modality data;
- Efficient training (triton-acceleration), debugging, and monitoring (GPU diagnosis).

Meta AI, Part-Time, Remote Sep. 2024 - May 2025

Paper: ReasonRec: A Reasoning-Augmented Multimodal Agent for Unified Recommendation

- Developed the first multimodal VLM agent with explicit reasoning and uncertainty-aware planning;
- Demonstrated that SFT + augmented data rivals RL in VLM.

Cisco Research, Part-Time & Full-Time, Remote Dec. 2023 - Aug. 2024

Research Scientist Intern at [Ramana Rao Kompella's team](#), Mentor: [Gaowen Liu](#)

Project: Machine Unlearning for Foundation Models (MoE-LLMs, Diffusion Models)

- [Paper 1](#): UnlearnCanvas: Stylized Image Dataset for Enhanced Machine Unlearning Evaluation in Diffusion Models (NeurIPS'24)
- [Paper 2](#): SEUF: Is Unlearning One Expert Enough for Mixture-of-Experts LLMs? (ACL'25 Main)

Amazon AWS AI Lab, Full-Time Seattle, USA May. 2023 - Aug. 2023

Applied Scientist Intern at [Just Walk Out's Team](#), worked with [Tian Lan](#) and [Zhou Ren](#).

Project: In-context learning for Diffusion Models

- Designed novel training algorithms to enable diffusion models to perform in-context adaptation, a capability traditionally limited to autoregressive models
- Pioneered one of the first approaches to **task-generalizable diffusion models**, achieving robust performance on unseen downstream tasks without fine-tuning

PUBLICATIONS

Yihua Zhang has published over 30 papers in top-tier machine learning and computer vision venues (*e.g.*, *NeurIPS*, *ICML*, *ICLR*, *CVPR*, *ICCV*, *ECCV*, *ACL*), including 15 first-author publications. His [Google Scholar](#) citation count tops up to 1805 as of Nov. 3rd, 2025.

FIRST-AUTHORED PUBLICATIONS

(* indicates equal contribution)

- [NeurIPS'25] **Y. Zhang**, C. Wang, Y. Chen, C. Fan, J. Jia, S. Liu, “*The Fragile Truth of Saliency: Improving LLM Input Attribution via Attention Bias Optimization*”, The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025, **Spotlight (3.2% of 21575 submissions)**.
- [NeurIPS'25] **Y. Zhang***, M. Zhang*, J. Jia, Z. Wang, S. Liu, T. Chen, “*One Token Embedding Is Enough to Deadlock Your Reasoning Large Language Model*”, The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
- [ACL'25] **Y. Zhang***, H. Zhuang*, K. Guo, J. Jia, G. Liu, S. Liu, X. Zhang, “*SEUF: Is Unlearning One Expert Enough for Mixture-of-Experts LLMs?*”, The 63rd Annual Meeting of the Association for Computational Linguistics Main Conference, 2025.
- [ICML'25W] **Y. Zhang***, X. Liu, X. Zeng, M. Liang, J. Yang, R. Jin, W.-Y. Chen, Y. Han, B. Long, H. Li, B. Zhang, L. Luo, S. Liu, T. Chen, “*ReasonRec: A Reasoning-Augmented Multimodal Agent for Unified Recommendation*”, Forty-Second International Conference on Machine Learning, 2025.
- [CVPR'25] **Y. Zhang***, H. Wang*, R. Bai, Y. Zhao, S. Liu, Z. Tu, “*Edit Away and My Face Will not Stay: Personal Biometric Defense against Malicious Generative Editing*”, The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025.
- [CPAL'25] **Y. Zhang**, H. Li, Y. Yao, A. Chen, P.-Y. Chen, S. Zhang, M. Wang, S. Liu, “*Visual Prompting Reimagined: The Power of Activation Prompts*”, Conference on Parsimony and Learning, 2025.
- [NeurIPS'24] **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*”, The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024.
- [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, T. Chen, “*Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark*”, The 63rd Annual Meeting of the Association for Computational Linguistics.
- [IEEE SPM'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 2024.
- [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*”, the Thirty-Seventh Annual Conference on Neural Information Processing Systems, 2023.
- [ICCV'23] **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*”, 2023 International Conference on Computer Vision, **Oral (1.7% of 8620 submissions)**.
- [ICLR'23] **Y. Zhang**, P. Sharma, P. Ram, M. Hong, K. Varshney, S. Liu, “*What Is Missing in IRM Training and Evaluation? Challenges and Solutions*”, The Eleventh International Conference on Learning Representations, 2023.
- [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*”, The Thirty-Sixth Annual Conference on Neural Information Processing Systems, 2022.
- [NeurIPS'22] **Y. Zhang***, G. Zhang*, Y. Zhang, S. Liu, S. Chang, “*Fairness Reprogramming*”, The Thirty-Sixth Annual Conference on Neural Information Processing Systems, 2022.

- [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”, The Thirty-Ninth International Conference on Machine Learning, 2022.
- [CVPR'22] **Y. Zhang***, T. Chen*, Z. Zhang*, S. Chang, S. Liu, Z. Wang, “Quarantine: Sparsity Can Uncover the Trojan Attack Trigger for Free”, 2022 Conference on Computer Vision and Pattern Recognition.

Co-AUTHORED PUBLICATIONS

- [EMNLP'25] C. Wang, C. Fan, **Y. Zhang**, Y. Zhang, J. Jia, D. Wei, P. Ram, N. Baracaldo, S. Liu, “Reasoning Model Unlearning: Forgetting Traces, Not Just Answers, While Preserving Reasoning Skills”, The 2025 Conference on Empirical Methods in Natural Language Processing.
- [ICCV'25] Y. Sun, **Y. Zhang**, G. Liu, H. Xie, S. Liu, “Invisible Watermarks, Visible Gains: Steering Machine Unlearning with Bi-Level Watermarking Design”, International Conference on Computer Vision, 2025.
- [ICML'25] C. Fan, J. Jia, **Y. Zhang**, A. Ramakrishna, M. Hong, S. Liu, “Towards LLM Unlearning Resilient to Relearning Attacks: A Sharpness-Aware Minimization Perspective and Beyond”, Forty-Second International Conference on Machine Learning, 2025.
- [ICML'25] C. Wang, Y. Zhang, J. Jia, P. Ram, D. Wei, Y. Yao, S. Pal, N. Baracaldo, S. Liu, “Invariance Makes LLM Unlearning Resilient Even to Unanticipated Downstream Fine-Tuning”, Forty-Second International Conference on Machine Learning, 2025.
- [ICLR'25] H. Li, **Y. Zhang**, S. Zhang, M. Wang, S. Liu, P.-Y. Chen, “When is Task Vector Provably Effective for Model Editing? A Generalization Analysis of Nonlinear Transformers”, The Thirteenth International Conference on Learning Representations, 2025. **Oral, 1.8% of 11,603 submissions.**
- [AAAI'25] C. Jin, T. Huang, **Y. Zhang**, M. Pechenizkiy, S. Liu, S. Liu, T. Chen, “Visual Prompting Upgrades Neural Network Sparsification: A Data-Model Perspective”, The 39th Annual AAAI Conference on Artificial Intelligence, 2025.
- [EMNLP'24] 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”, The 2024 Conference on Empirical Methods in Natural Language Processing.
- [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”, The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024.
- [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”, The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024.
- [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”, European Conference on Computer Vision, 2024.
- [ICLR'24] 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”, The Twelfth International Conference on Learning Representations, 2024. **Spotlight, 5% of 7262.**
- [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”, The Twelfth International Conference on Learning Representations, 2024.

- [IEEE TSP] 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 Network on Group Imbalance”, IEEE Journal of Selected Topics in Signal Processing, 2024.
- [ICML’23] P. Khanduri, I. Tsaknakis, **Y. Zhang**, J. Liu, S. Liu, J. Zhang, Mingyi Hong, “Linearly Constrained Bilevel Optimization: A Smoothed Implicit Gradient Approach”, Fortieth International Conference on Machine Learning, 2023.
- [ICLR’23] B. Hou, J. Jia, **Y. Zhang**, G. Zhang, Y. Zhang, S. Liu, S. Chang, “Textgrad: Advancing Robustness Evaluation in NLP by Gradient-Driven Optimization”, The Eleventh International Conference on Learning Representations, 2023.
- [CVPR’23] A. Chen, Y. Yao, P.-Y. Chen, **Y. Zhang**, S. Liu, “Understanding and Improving Visual Prompting: A Label-Mapping Perspective”, The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023.
- [CVPR’23] H. Zhuang, **Y. Zhang**, S. Liu, “A Pilot Study of Query-Free Adversarial Attack against Stable Diffusion”, The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023.
- [UAI’22] 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”, The 38th Conference on Uncertainty in Artificial Intelligence, 2022. **Best Paper Runner-Up Award.**

COMMUNITY SERVICES

- **Tutorial Speaker:**
 - [IEEE MilCom’25] Robust Machine Unlearning: Securing Foundation Models Against Forgetting Failures
 - [AAAI’24] Zeroth-Order Machine Learning: Fundamental Principles and Emerging Applications in Foundation Models
 - [AAAI’23] Bi-level Optimization in Machine Learning: Foundations and Applications
- **Conference Volunteer:** AAAI’23, ICLR’23
- **Journal Reviewer:** JMLR, IEEE TPAMI, IEEE T-IFS, TMLR
- **Workshop Organizer:** New Frontiers in Adversarial ML [ICML’22], [ICML’23], [NeurIPS’24].

MENTEES

- **Yuhao Sun** (Undergraduate@USTC) — [ICCV’25] May. 2024 - Aug. 2024
- **Hanhui Wang** (Master@USC) — [CVPR’25] May. 2024 - Oct. 2024
- **Chongyu Fan** (Undergraduate@HUST, PhD@MSU) — [ICLR’24 Spotlight] May. 2023 - Aug. 2024
- **Haomin Zhuang** (PhD@Notre Dame) — [CVPRW’23], [ACL’25 Main] Dec. 2022 - Aug. 2024
- **Can Jin** (Undergraduate@USTC, PhD@Rutgers) — [AAAI’25] Aug. 2023 - Dec. 2023
- **Aochuan Chen** (Undergraduate@THU, PhD@HKUST) — [CVPR’23], [ICLR’24] Oct. 2022 - Oct. 2023

GRANT/FUNDING EXPERIENCE

- Center for AI Safety (CAIS)** (\$57,000 of compute credits) “Tamper-Resistant Safeguards” 2025-2027
- **Role:** Co-Proposal Writer (PI: Dr. Sijia Liu)
- Cisco Research Award** (\$75,000): “Towards LifeLong LMM Agents in Embodied AI” 2024-2025
- **Role:** Co-Proposal Writer (PI: Dr. Sijia Liu)
- NAIRR Pilot Resource Awards** (\$20,000): “Enhancing LLM Unlearning across the Lifecycle” 2024-2025
- **Role:** Co-Proposal Writer (PI: Dr. Sijia Liu)