

Chongyu Fan

 fanchon2@msu.edu Website Github Google Scholar Citation 610

Education

Michigan State University, East Lansing, USA 2024.08 - Present
Doctor of Computer Science
Advisor: Prof. Sijia Liu @ OPTML Lab

Huazhong University of Science and Technology, Wuhan, China 2020.09 - 2024.06
Bachelor of Engineering
Outstanding Graduate
GPA: 3.96/4.0

Industry Experience

- **Reinforcement Pre-training (RPT) for Vision-Language Models**
 - Reframed next-token prediction into *verifiable rewards* for reinforcement learning.
 - Improved rollout efficiency by leveraging high-entropy data for exploration.
 - Achieved 8% gains with RPT over pretraining, and retained them during post-training.

ByteDance, Research Scientist Intern, San Jose, USA 2025.05 - 2025.09

Mentor: Jian Du

- **Reinforcement Learning for Multi-Agent Systems**
 - Identified reward imbalance issues in multi-agent reinforcement learning.
 - Proposed a *counterfactual reward* mechanism to fairly evaluate each agent’s contribution.
 - Achieved a 10% performance improvement on math reasoning tasks over state-of-the-art baselines.
 - **Log-to-Leak Attack on Model Context Protocol (MCP)**
 - Discovered security issues in MCP that may lead to leakage of user–agent interaction logs.
 - Developed *Log-to-Leak*, a prompt injection attack framework for MCP.
 - Demonstrated 100% leakage success under real-world agent settings.
 - **Research Outcome:** Paper submitted to **ICLR 2026**.

Research Project Highlights

- **Post-training Knowledge Editing and Alignment for LLMs and Diffusion Models**
My research focuses on developing methods to mitigate the influence of undesired knowledge in foundation models, particularly during the post-training stage of LLMs [[NeurIPS'25a](#)], [[EMNLP'25](#)] and diffusion models [[ICLR'24](#)], [[NeurIPS'24a](#)], [[NeurIPS'24b](#)]. Through these efforts, I aim to enhance the trustworthiness, robustness, and safety of next-generation AI systems.

- **Bi-level and Smooth Optimization for LLMs and Diffusion Models**
My research leverages bi-level optimization to improve both the training and inference of LLMs and diffusion models, enhancing their robustness and interpretability [ECCV'24], and explores smooth optimization to strengthen the safety and reliability of LLMs [ICML'25].

- **Efficient Inference and Training for Reasoning Models**
I investigate optimized test-time computation to enhance reasoning capability while mitigating over-and under-thinking issues [[arXiv'25a](#)], [[NeurIPS'25b](#)], and design dataset condensation techniques to improve the efficiency and scalability of reasoning model training [[arXiv'25b](#)].

Publications

I have published more than ten papers in top-tier machine learning and computer vision venues (*e.g.*, NeurIPS, ICML, ICLR, ECCV, EMNLP), with **five** of them as first author. As of October 23, 2025, my research has garnered **610** citations on Google Scholar.

First-Authored Publications (* indicates equal contribution)

- [ICLR'24] **C. Fan**, J. Liu, Y. Zhang, D. Wei, E. Wong, S. Liu, “*SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation.*” (**Spotlight, acceptance rate 5%; 200+ citations; IBM Pat Goldberg Best Paper Award Finalist**)
- [ECCV'24] **C. Fan**, J. Liu, A. Hero, S. Liu, “*Challenging forgets: Unveiling the worst-case forget sets in machine unlearning.*” (Travel Grant)
- [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.*”
- [EMNLP'25] **C. Fan***, C. Wang*, Y. Zhang, J. Jia, D. Wei, P. Ram, N. Baracaldo, S. Liu, “*Reasoning Model Unlearning: Forgetting Traces, Not Just Answers, While Preserving Reasoning Skills.*”
- [NeurIPS'25a] **C. Fan**, J. Liu, L. Lin, J. Jia, R. Zhang, S. Mei, S. Liu, “*Simplicity Prevails: Rethinking Negative Preference Optimization for LLM Unlearning.*”

Co-Authored Publications

- [NeurIPS'24a] Y. Zhang, **C. Fan**, Y. Zhang, Y. Yao, J. Jia, J. Liu, G. Zhang, G. Liu, R. Komelia, X. Liu, S. Liu, “*UnlearnCanvas: Stylized Image Dataset for Enhanced Machine Unlearning Evaluation in Diffusion Models.*”
- [NeurIPS'24b] 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.*”
- [ICML'25W] J. Lee, Z. Mai, **C. Fan**, W.L. Chao, “*An Empirical Exploration of Continual Unlearning for Image Generation.*”
- [NeurIPS'25b] 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.*” (**Spotlight, acceptance rate 3%**)

Preprint Papers

- [arXiv'25a] **C. Fan**, Y. Zhang, J. Jia, A. Hero, S. Liu, “*CyclicReflex: Improving Large Reasoning Models via Cyclical Reflection Token Scheduling.*”
- [arXiv'25b] J. Jia, H. Reisizadeh, **C. Fan**, N. Baracaldo, M. Hong, S. Liu, “*EPiC: Towards Lossless Speedup for Reasoning Training through Edge-Preserving CoT Condensation.*”
- [arXiv'25c] Y Lang, Y Zhang, **C. Fan**, C Wang, J Jia, S Liu, “*Downgrade to Upgrade: Optimizer Simplification Enhances Robustness in LLM Unlearning.*”
- [arXiv'25d] **C. Fan**, C Wang, Y Huang, S Pal, S Liu, “*LLM Unlearning Under the Microscope: A Full-Stack View on Methods and Metrics.*”

Academic Services

- **Workshop Student Co-Organizer:** New Frontiers in Adversarial ML [NeurIPS'24]
- **Reviewer:** ICLR'25-26, ICML'25, NeurIPS'25, AISTATS'25

Grant/Funding Experience

Cisco Research Award (\$75,000)	2025 – 2026
• Topic: “Vision–Language–Action Model Fine-Tuning through Modularity and Spectral Optimization”	
• PI: Dr. Sijia Liu	
• Role: Co-Proposal Writer	

Honors

Graduate Travel Fellowship, Michigan State University	2024
Graduate Travel Fellowship, Michigan State University	2025
IBM Pat Goldberg Best Paper Award Finalist	2025