Senior Research Scientist in Intuit AI Research Dec. 2024 – Now

RESEARCH FOCUS

Machine Learning: LLM Agent, World Model, Prompt Learning, Model Pruning, Adversarial ML.

Optimization: Black-box Optimization, Zeroth-order Optimization, Bi-level Optimization.

SCHOOL

Ph.D. in CSE, Michigan State University (MSU) Advisor: Sijia Liu Jan. 2021 – May. 2025

Ph.D. student in CSE, Tsinghua/MSU Advisor: Zhichao Cao, Yunhao Liu Aug. 2018 – Dec. 2020

B.S. in Automation, Tsinghua University Advisor: Hong Wang Aug. 2014 – July. 2018

Exchange in CSE, École Polytechnique Fédérale de Lausanne (EPFL) Aug. 2016 – Feb. 2017

WORK

Research Intern in Cisco Research Advisor: Gaowen Liu Feb. 2023 – Jun. 2024

Research Intern in MIT-IBM Watson AI Lab Advisor: Quanfu Fan May. 2021 – Aug. 2021

Research Intern in DiDi AI Lab Advisor: Yashu Liu Nov. 2017 – Feb. 2018

Research Intern in HKUST Advisor: Pan Hui June. 2017 – Sep. 2017

SELECTED PUBLICATION

Google Scholar Citations: 1645

[1] **Y. Yao***, Y. Chen*, Y. Zhang, B. Shen, G. Liu, S. Liu, Safety Mirage: How Spurious Correlations Undermine VLM Safety Fine-Tuning and Can Be Mitigated by Machine Unlearning, *ICLR 2026*.

[2] K. Chen, Z. Lin, Z. Xu, Y. Shen, **Y. Yao**, J. Rimchala, J. Zhang, L. Huang, R2I-Bench: Benchmarking Reasoning-Driven Text-to-Image Generation, *ACL 2025*

[3] **Y. Yao***, J. Liu*, Y. Gong*, X. Liu, Y. Wang, X. Lin, S. Liu, Can Adversarial Examples Be Parsed to Reveal Victim Model Information?, *WACV 2025*.

[4] **Y. Yao***, Z. Pan*, G. Liu, B. Shen, H. Zhao, R. Kompella, S. Liu, From Trojan Horses to Castle Walls: Unveiling Bilateral Backdoor Effects in Diffusion Models, *NeurIPS 2024*.

[5] **Y. Yao**, G. Xiao, V. Asnani, Y. Gong, J. Liu, X. Lin, X. Liu and S. Liu, Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks, *Foundations and Trends in Privacy and Security 2024*.

[6] S. Pal, **Y. Yao**, R. Wang, B. Shen, S. Liu, Backdoor Secrets Unveiled: Identifying Backdoor Data with Optimized Scaled Prediction Consistency, *ICLR 2024*.

[7] J. Jia*, J. Liu*, P. Ram, **Y. Yao**, G Liu, Y Liu, P. Sharma, S. Liu, Model Sparsity Can Simplify Machine Unlearning, *NeurIPS 2023* ***Spotlight***.

[8] A. Chen, **Y. Yao**, P. Chen, Y. Zhang, S. Liu, Understanding and Improving Visual Prompting: A Label-Mapping Perspective, *CVPR 2023*.

[9] **Y. Yao***, Y. Zhang*, P. Ram, P. Zhao, T. Chen, M. Hong, Y. Wang, S. Liu, Advancing Model Pruning via Bi-level Optimization, *NeurIPS 2022*.

[10] A. Chen*, P. Lorenz*, **Y. Yao**, P. Chen, S. Liu, Visual Prompting for Adversarial Robustness, *ICASSP 2022*.

[11] **Y. Yao***, Y. Gong*, Y. Li, Y. Zhang, X. Liu, X. Lin, S. Liu, Reverse Engineering of Imperceptible Adversarial Image Perturbations, *ICLR 2022*.

[12] Y. Zhang, **Y. Yao**, J. Jia, J. Yi, M. Hong, S. Chang, S. Liu, How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective, *ICLR 2022* ***Spotlight***.

[13] **Y. Yao**, Z. Ma, Z. Cao, LoSee: Long-Range Shared Bike Communication System Based On LoRaWAN Protocol, *EWSN 2019* ***Best Poster***.

TALKS

- 11/2024: **Large Vision Language Model.** Invited Talk at Wayne State University.
- 10/2023: **Machine Unlearning via Model Sparsification.** Invited Talk at Cisco.
- 06/2023: **Reverse Engineering of Deceptions.** CVPR 2023.
- 06/2022: **How to Robustify Black-box Models?** Invited Talk at UIUC.

TEACH

- **CSE 480: Database Systems** of MSU, Fall 2020, Fall 2023, Spring 2024, Fall 2024.
- **CSE 801: Big Data Analysis** of MSU, Spring 2023.
- **CSE 232: C++** and **CSE 231 & CSE 102: Python** of MSU, Spring & Fall 2021.
- **CSE 477: Web Applications** of MSU, Spring 2021.
- **CSE 891: Artificial Intelligence and Internet of Things (AIOT)** of MSU, Spring 2020.

HONOR

- **Travel Grant** at NeurIPS 2022, CVPR 2023, NeurIPS 2024.
- **Best Poster Award** at EWSN 2019.
- **Outstanding Undergraduate** of Automation Department, Tsinghua University, 2018.
- **Excellent Academic Scholarship** of Tsinghua University, 2015, 2016, 2017.

SERVICE

- **Chair** of AdvML Frontiers at ICML'22, ICML'23, and NeurIPS'24.
- **Reviewers** of NeurIPS, ICLR, ICML, ACL, CVPR, ACMMM, ICASSP; TPAMI.

PATENT

- Model Assembly with Knowledge Distillation, US18243259
- Domain Adaptation through Model Pruning, US18598148
- Modality-Agnostic Diffusion Prompting, Pending
- Semantic Segmentation using LLM Supervision, Pending
- Bi-Directional LoRA for Machine Unlearning and Retaining Information, Pending
- Generation Model Pruning for Fairness, Pending
- Mixed-Precision Quantization for Machine Unlearning Through Contrastive Learning, Pending
- Optimizing Automated Machine Learning Model Continual Learning Systems, Pending
- Continual Rule Learning for Large Language Models to Reason with Tools, Pending
- Automatic Schema Discovery with Auto-Ingestion and Retrieval Planning, Pending

MENTORSHIP

- Zhuoshi Pan, B.S./M.S. in Automation, Tsinghua University. Poisoning Diffusion, NeurIPS 2024.
- Brian Zhang, Northville High School, MI. Visual Prompt Ensemble, ICASSP 2024.
- Aochuan Chen, B.S. in Automation, Tsinghua University. Robust Visual Prompt, ICASSP 2023.

FUNDING

- Cisco Research Award for "Towards Lifelong LMM Agents in Embodied AI" – Research Gift of $75,000 (Co-lead in proposal development)