Ziyi Yin

Ph.D. Student
College of Information Sciences and Technology
The Pennsylvania State University
E348 Westgate Building, University Park, PA 16802
zmy5171@psu.edu
Tel +1 (814) 321-4788
https://ericyinyzy.github.io

EDUCATION

The Pennsylvania State University — PA, U.S.
Ph.D. Student in Informatics — Sep. 2022 - Present
Advisor: Dr. Fenglong Ma

Xi'an Jiaotong University — Xi'an, China
Master of Science in Control Science and Engineering — Sep. 2019 - Jun. 2022

Xi'an Jiaotong University — Xi'an, China
Bachelor of Science in Automation (Honors Youth Program) — Sep. 2015 - Jun. 2019

RESEARCH INTERESTS

I am broadly interested in data science and artificial intelligence, with a focus on machine learning. Specifically, my research centers on multimodal learning systems, such as multimodal large language models (MLLMs), with a particular emphasis on their robustness and intrinsic security challenges. My related work has been published at top-tier conferences including ACL'25, AAAI'24, and NeurIPS'23 (see publications below).

Currently, another branch of my research focuses on applying Agentic AI to the field of health informatics. Specifically, my work involves designing a multi-agent LLM framework aimed at enhancing the accuracy and explainability of ICD coding predictions. Besides, I have experience with model distillation, and the corresponding work has been accepted at KDD'25.

PUBLICATIONS

Peer-Reviewed Conferences

ACL25 — **Ziyi Yin**, Muchao Ye, Yuanpu Cao, Aofei Chang, Jiaqi Wang, Han Liu, Jinghui Chen, Ting Wang and Fenglong Ma. *Shadow-Activated Backdoor Attacks on Multimodal Large Language Models.* Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), 2025.

KDD25 — Jiaqi Wang*, **Ziyi Yin***, Quanzeng You, Lingjuan Lyu, and Fenglong Ma. *Collaborative Diagnosis: Empowering Underserved Regions with Asymmetrical Reciprocity-based Cross-silo Federated Learning*, Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2025. (* equal contribution.)

AAAI24 — **Ziyi Yin**, Muchao Ye, Tianrong Zhang, Han Liu, Jinghui Chen, Ting Wang and Fenglong Ma. *VQAttack: Transferable Adversarial Attacks on Visual Question Answering via Pre-trained Models* . The 39-th Annual AAAI Conference on Artificial Intelligence (AAAI), 2024

NeurIPS24 Yuanpu Cao, Tianrong Zhang, Bochuan Cao, **Ziyi Yin**, Lu Lin, Fenglong Ma, Jinghui Chen. *Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization.* The 38-th Annual Conference on Neural Information Processing Systems(NeurIPS), 2024

IJCAI24 Jiaqi Wang, Junyu Luo, Muchao Ye, Xiaochen Wang, Yuan Zhong, Aofei Chang, Guanjie Huang, **Ziyi Yin**, Cao Xiao, Jimeng Sun, and Fenglong Ma. *Recent Advances in Predictive Modeling with Electronic Health Records*, Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024) Survey Track

SDM24 Yuan Zhong, Suhan Cui, Jiaqi Wang, Xiaochen Wang, **Ziyi Yin**, Yaqing Wang, Houping Xiao, Mengdi Huai, Ting Wang and Fenglong Ma. *MedDiffusion: Boosting Health Risk Prediction via Diffusion-based Data Augmentation.* Proceedings of the SIAM International Conference on Data Mining (SDM), 2024

NeurIPS23 **Ziyi Yin**, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang and Fenglong Ma. *VLATTACK: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models.*, Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), 2023

NeurIPS23 Muchao Ye, **Ziyi Yin**, Tianrong Zhang, Tianyu Du, Jinghui Chen, Ting Wang and Fenglong Ma. *UniT: A Unified Look at Certified Robust Training against Text Adversarial Perturbation.* , Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), 2023

EMNLP23 Xiaochen Wang, Junyu Luo, Jiaqi Wang, **Ziyi Yin**, Suhan Cui, Yuan Zhong, Yaqing Wang and Fenglong Ma *Hierarchical Pretraining on Multimodal Electronic Health Records.* , Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. (EMNLP), 2023

IJCAI21 **Ziyi Yin**, Ruijin Liu, Zhiliang Xiong, Zejian Yuan. *Multimodal Transformer Network for Pedestrian Trajectory Prediction.* The 30th International Joint Conference on Artificial Intelligence (IJCAI), 2021.

BMVC21 **Ziyi Yin**, Ruijin Liu, Zhiliang Xiong, Zejian Yuan. *Order-independent Matching with Shape Similarity for Parking Slot Detection.*, The 32nd British Machine Vision Conference (BMVC), 2021.

ICPR21 Ziyang Song, **Ziyi Yin**, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, Shenghao Zhang. *Attention-Oriented Action Recognition for Real-Time Human-Robot Interaction.* The 25th International Conference on Pattern Recognition (ICPR), 2021

ICIP19 **Ziyi Yin**, Ziyang Song, Zejian Yuan. *Learning to Plan Semantic Free-Space Boundary*, The 26th IEEE International Conference on Image Processing (ICIP), 2019

CORE RESEARCH PROJECTS

In Progress Projects

Agentic ICD Coding System (Ongoing Project) 03/2025 - Now

To build a more general, accurate, and explainable automatic ICD coding system, we

- propose a multi-agent LLM framework that first performs information retrieval to identify potential ICD codes, and then leverages collaborative reasoning and critical thinking among multiple LLMs to accurately extract ICD codes from the patient's discharge summary.
- We are currently working on data generation and the initial stage of model fine-tuning.

Defend Jailbreak Attacks on MLLMs (In Submission) 03/2024 - 09/2024
To build a more robust and safe MLLM against jailbreak attacks, we

- Propose an adversarial training algorithm SafeMLLM, which is a two-step training framework for tuning an MLLM.
- Achieve excellent defense performance across six MLLMs and six jailbreak attack methods spanning multiple modalities in both white-box and black-box scenarios.

Safety Risks in MLLMs (ACL'25) 10/2023 - 02/2024
To explore the potential security risks in current MLLMs, we

- Define a new threat model on MLLMs, when an attacker aims to automatically inject malicious content into user's response via a backdoor trigger.
- Propose an attention-regularization strategy to inject malicious behavior which only need a few training samples (90% ASR for <200 image-text pairs).

Asymmetrical Cross-silo Federated Learning (KDD'25) 9/2023 - 11/2023
To use federated learning techniques to mitigate the global issue of geographic health disparities, we

- Propose a novel cross-silo federated learning framework, which aims at using distillation techniques to alleviate geographic health disparities and fortifying the diagnostic capabilities of underserved regions.
- Conduct comprehensive experiments encompassing multi-class and medical image classification tasks as well as 2D and 3D semantic segmentation tasks.

Multimodal Adversarial Attacks on Vision-Language Tasks. (NeurIPS'23, AAAI'24) 09/2022 - 08/2023
To verify adversarial vulnerability on unified multi-modal models, we

- Conduct adversarial attacks on multiple vision language tasks via pertained models, which is a more realistic setting.
- Propose a multi-modal attack strategy to cross-search image and text perturbations from both single-modal and multi-modal levels.

Finished Projects

Parking Slot Detection (BMVC'21) 03/2021 - 09/2021
To detect parking slots in the more general scenarios (variant sizes an shapes), we

- Construct a Large-scale and Remote-view Parking Slot dataset (LRPS).
- Propose a two-level order-independent matching strategy to solve the order induced rotation problem

Pedestrian Trajectory Prediction (IJCAI'21) 09/2020 - 03/2021
To solve the problems of current CNNs or RNNs in compensating the highly dynamic motion information and massive parameters usages, we

- Introduce specific areas of optical flow to compensate the dynamic motion information, and also propose a compact representation method to improve the computational efficiency.
- Propose a Multimodal Transformer Network (MTN) to integrate distinct modalities in a multi-granularity manner.
- Achieves the SOTA performance with 107x fewer parameters on public datasets.

INTERNSHIP EXPERIENCE

Amazon, Inc.
Towards parameter-efficiency on SMoE LLMs Summer 2025
Applied Scientist Intern
Mentor: Zhengyang Wang & Hejie Cui

TEACHING EXPERIENCE

The Pennsylvania State University
DS310: Machine Learning for Data Analytics Fall 2024
Teaching Assistant
Instructor: Dr. Fenglong Ma

Xi'an Jiaotong University
Computer Vision and Pattern Recognition Fall 2021
Teaching Assistant
Instructor: Dr. Zejian Yuan

ACADEMIC SERVICE

- Program Committee Member
  - 2026: AAAI, ICLR, CVPR
  - 2025: ICLR, CVPR, ICCV, ACL, EMNLP, NeurIPS
  - 2024: CVPR, NeurIPS, ICLR, AISTAS, KDD, PAKDD
- Journal Reviewer
  - IEEE Transactions on Emerging Topics in Computational Intelligence
  - Journal of Artificial Intelligence Research

HONORS & AWARDS

2023 Scholar award by NeurIPS'23

2021 First-Class Graduate Student Scholarship by Xi'an Jiaotong University