Lingzhi Yuan

lingzhiy@umd.edu · (+1) 305-216-0051 · Linkedin · Github · Homepage

Education

University of Maryland, College Park, MD, USA — Aug. 2025 – Present

Ph.D in Computer Science, expected May 2030

Zhejiang University, Hangzhou, China — Sep. 2021 – Jun. 2025

B.Eng in Automation, *GPA*: 3.96/4, *Rank*: 1/58
Minor in Intensive Training Honors Program (ITP) at Chu Kochen Honors College (Top 40/5400)

Experience

Furong Huang Lab, University of Maryland, College Park, MD, USA — Aug. 2025 – Present

Research Assistant, focused on Trustworthy ML and LLM Alignment advised by Prof. Furong Huang

Universal Ubiquitous Co., Ltd Hangzhou, China — Mar. 2025 – Jun. 2025

Student Intern, focused on VLM's Instruction Following and Data Augumentation.

Secure Learning Lab, The University of Chicago, Chicago, IL, USA — Mar. 2024 – Dec. 2024

Research Assistant, focused on Trustworthy ML and Multi-modal Safety under the guidance of Prof. Bo Li

USSLAB, Zhejiang University, Hangzhou, China — Mar. 2023 – Mar. 2024

Research Assistant, focused on IoT Security and AI Security under the guidance of Prof. Yanjiao Chen

Nanyang Technological University, Singapore — Aug. 2023 – Sep. 2023

Summer School Participant, Machine Learning and Its Applications program led by Prof. Kezhi Mao

Publications

* Denotes equal contribution

PropmtGuard: Soft Prompt-Guided Unsafe Content Moderation for Text-to-Image Models

Lingzhi Yuan*, Xinfeng Li*, Chejian Xu, Guanhong Tao, Xiaojun Jia, Yihao Huang, Wei Dong, Yang Liu, Bo Li. *Arxiv Pre-print*

MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models

Chejian Xu*, Jiawei Zhang*, Zhaorun Chen*, Chulin Xie*, Mintong Kang*, Zhuowen Yuan*, Zidi Xiong*, Chenhui Zhang, Lingzhi Yuan, Yi Zeng, Peiyang Xu, Chengquan Guo, Andy Zhou, Jeffrey Ziwei Tan, Zhun Wang, Alexander Xiong, Xuandong Zhao, Yu Gai, Francesco Pinto, Yujin Potter, Zhen Xiang, Zinan Lin, Dan Hendrycks, Dawn Song, Bo Li. *International Conference on Learning Representation (ICLR), 2025*

Selected Projects

Soft Prompt-Guided Defense for Text-to-Image Models — Jun. 2024 - Dec. 2024

Project Leader Advisor: Prof. Bo Li, Associate Professor at University of Chicago/UIUC

Developing an effective and efficient framework to mitigate NSFW content generation across different unsafe categories by T2I models.

- Conducted an extensive survey of advanced safety protection techniques for T2I models, identifying critical challenges and areas for improvement.
- Designed and implemented an advanced framework addressing multiple categories of unsafe content generation, achieving state-of-the-art performance in NSFW moderation with high time efficiency.
- Authored and submitted a manuscript for peer review.

Red Teaming and Capability Testing for Speech-to-Speech Models — Sep. 2024 - Dec. 2024

Research Assistant Advisor: Prof. Bo Li, Associate Professor at University of Chicago/UIUC

Conducted a comprehensive evaluation of the state-of-the-art speech-to-speech Audio LLM (GPT-4o-s2s) under various red teaming scenarios.

- Curated multi-choice audio datasets from three perspectives to establish a robust base evaluation framework.
- Designed non-audible prompt injection attacks using various frequency processing techniques and optimization algorithms to craft adversarial examples.
- Assessed GPT-4o-s2s performance against the Qwen2-Audio baseline, and provided a technical report to **OpenAI** detailing identified vulnerabilities and recommending potential improvements.

Comprehensive Assessment of Trustworthiness in Multimodal Models — Mar. 2024 – Oct. 2024

Research Assistant Advisor: Prof. Bo Li , Associate Professor at University of Chicago/UIUC

Conduct a comprehensive evaluation of the safety and trustworthiness of multimodal foundation models, addressing key challenges in robustness and reliability.

- Investigated adversarial robustness by applying cutting-edge red-team algorithms, such as MMP and GCG for Text-to-Image (T2I) models and Attack-VLM for Image-to-Text (I2T) models, on surrogate architectures to develop a robust benchmark dataset.
- Evaluated state-of-the-art multimodal models including Dalle-3, GPT-4o etc. against the curated dataset, providing a detailed analysis of their vulnerability and resilience under adversarial conditions.
- Authored and submitted a manuscript on the findings to **ICLR 2025** for peer review.

Characterize the Vulnerability of Image Sensors Under EMI — Apr. 2023 – Sep. 2024

Deputy Leader Student Research Training program, Advisor: Prof. Xiaoyu Ji, Professor at Zhejiang University

Identified a novel class of vulnerabilities in image sensors and elucidated the underlying principles.

- Assessed attack feasibility and designed experiments using a signal generator and amplifier to emit signals targeting camera sensors on terminal devices.
- Modeled the attack mechanism through Python simulations, integrating experimental data with advanced signal analysis and image processing techniques.
- Evaluated the effectiveness of the attack through case studies on various computer vision models.

Skills

- **Programming Languages**: solid expertise in Python, Matlab, C++, and various algorithms
- **Tools**: PyTorch, Git, Linux/Unix, Transformers, Diffusers, OpenCV
- **Language Proficiency**: TOEFL 105 (Reading: 30 Listening: 27 Speaking: 23 Writing: 25)

Honors and Awards

University of Maryland Graduate School Dean's Fellowship	Aug. 2025
Infineon Power Semiconductor Scholarship	Sep. 2024, Sep. 2023
Zhejiang University Second Class Scholarship	Sep. 2022, Sep. 2023
Nitori International Scholarship	Sep. 2022