Lingzhi Yuan

lingzhiyxp@gmail.com · (+86) 181-9103-8843 · Linkedin · Github · Homepage

Education

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

B.Eng in Automation, expected June 2025

- **GPA**: 3.96/4 90.2/100 | **Rank**: 1/58
- **Research Interests**: Trustworthy ML, Multi-modal Safety, Adversarial Robustness
- **Relevant Courses**: Fundamental of Data Structure (96), Embedded System (96), Probability and Mathematical Statistics (97), Signal Analysis and Processing (93), Computer Vision (100), AI Security (93)
- **Minor**: Intensive Training Honors Program of Innovation and Entrepreneurship (ITP) at Chu Kochen Honors College (Top 40/5400)

University of Chicago, Chicago, U.S. — Mar. 2024 – Present

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

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

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

Publications

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, XiaoFeng Wang, Bo Li
Submitted to CVPR 2025

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
Submitted to ICLR 2025

Research Experience

Soft Prompt-Guided Unsafe Content Moderation for Text-to-Image Models Jun. 2024 - Present

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 to **CVPR 2025** for peer review.

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

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.

Mitigating Inappropriate Content Generation in Text-to-Image Models Nov. 2023 – Mar. 2024

Research Assistant Advisor: Prof. Yanjiao Chen, ZJU100 Professor at Zhejiang University

Develop a framework to mitigate sexually-explicit content generation by T2I models in a text-agnostic manner.

- Assessed generated images using metrics such as CLIP Score, LPIPS Score, FID Score, and Nudenet, along with other relevant evaluation tools.
- Conducted ablation studies on multiple hyperparameters to enhance the model's robustness against inappropriate content generation.
- Related work has been accepted by **CCS 2024**.

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, Linux/Unix, Transformers, Diffusers, OpenCV
- **Language Proficiency**: TOEFL 105 (Reading: 30 Listening: 27 Speaking: 23 Writing: 25)

Honors and Awards

Infineon Power Semiconductor Scholarship Sep. 2024, Sep. 2023
Zhejiang University Second Class Scholarship Sep. 2022, Sep. 2023
Nitori International Scholarship Sep. 2022

Campus Activities

Club Leader

- Orginize the internal training activities for club members, including software and hardware development
- Participate in undertaking the enterprise campus tour activities