

LINGZHI YUAN

✉ yuan_lz@zju.edu.cn · ☎ (+86) 181-9103-8843 · in lingzhiyuan1 · 🌐 lingzhiyxp

🎓 EDUCATION

Zhejiang University (ZJU), Hangzhou, China

Sep. 2021 – Jun. 2025

B.S. in Automation, expected June 2025

- **GPA:** 3.96/4 90.2/100 | **Rank:** 1/56
- **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)

🔍 RESEARCH EXPERIENCE

Protect T2I Models from Generating Different Kinds of Unsafe Images

Jun. 2024 - Present

Project Leader Advisor: Prof. [Bo Li](#), Associate Professor at University of Chicago/UIUC

Design a new framework to mitigate different kinds of unsafe content generation by T2I models while preserving the benign generation ability of models

- Conduct comprehensive surveys on current safety protection techniques for T2I models and identified key challenges.
- Design and implement an advanced framework capable of addressing multiple categories of unsafe content generation.
- Evaluate and benchmark our approach against existing baseline methods to assess its effectiveness.

Evaluating the Trustworthiness and Safety of Multimodal Models

Mar. 2024 – Oct. 2024

Research Assistant Advisor: Prof. [Bo Li](#), Associate Professor at University of Chicago/UIUC

Conduct an in-depth evaluation of the safety and trustworthiness of MMFM from various perspectives

- Focused on adversarial robustness, employing red-team algorithms such as MMP, GCG, and Attack-VLM on surrogate models to build a comprehensive dataset.
- Assessed advanced targeted models using the curated dataset to benchmark performance.
- 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 unsafe content generation by text-to-image models in a text-agnostic manner

- Assessed generated images using metrics such as CLIP Score, FID, and Nudenet, along with other relevant evaluation tools.
- Tuned hyperparameters to enhance the model's defense against inappropriate content generation.
- Performed data post-processing and visualization to analyze and present findings effectively.

Characterize the Vulnerability of Image Sensors Under EMI

Apr. 2023 – Sept. 2024

Deputy Leader Student Research Training program, Advisor: Prof. [Xiaoyu Ji](#), Professor at Zhejiang University

Identify a new class of vulnerabilities of image sensors and explain the underlying principle

- 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, leveraging experimental data and expertise in signal analysis and image processing.
- Evaluated the attack's effectiveness through case studies on various CV models.

SKILLS

- **Programming Languages:** solid expertise in Python, Matlab, C++, and various algorithms
- **Tools:** PyTorch, Linux/Unix, Transformers, Diffusers, OpenCV
- **Language Proficiency:** TOEFL 99 (Reading: 26 Listening: 25 Speaking: 20 Writing: 28)

HONORS AND AWARDS

<i>Infineon Power Semiconductor Scholarship</i>	Sep. 2023
<i>Zhejiang University Second Class Scholarship</i>	Sep. 2022, Sep. 2023
<i>Nitori International Scholarship</i>	Sep. 2022

CAMPUS ACTIVITIES

Zhejiang University Student SmartLink Club , Hangzhou, China	Sep. 2023 – Jun. 2024
<i>Club Leader</i>	

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