Yichi Zhang

Hangzhou, China | yichics02@gmail.com | tel:+86-138-8430-8849 | yichics.github.io

Education

Zhejiang University, B.S. in Information Engineering — Aug. 2021 – Now

- **GPA**: 3.92/4.0
- Minor in Computer Science and Technology.
- **Honor**: Third-class Scholarship, Zhejiang University;

Research Experience

Research Intern, Zhejiang Lab & Zhejiang University — July 2023 – Mar. 2024
Advisor: Prof. Xiaogang Xu — Hangzhou, China

- Focus on **AI-generated content detection**, particularly images from diffusion models and GANs.
- Leveraging the denoising characteristic of diffusion models to amplify high-frequency differences between generated and real images, we designed a novel classification feature that achieves SOTA results on four datasets with excellent generalization performance and 20x faster extraction speed.
- First author of a paper currently under review at IJCV.

Research Intern, NESA Lab, Zhejiang University — July 2023 – Now
Advisor: Prof. Shouling Ji — Hangzhou, China

- Focus on **backdoor attack in horizontal federated learning**.
- Revisiting existing defense mechanisms, we show theoretically and empirically that differences in the distribution of backdoor model updates arise from poisoned data distribution variations, and we establish a theoretical attack boundary for detection-based defenses.
- Second author and main code contributor of a paper currently under review at IEEE S&P 2025.

Research Intern, ALPS Lab, Stony Brook University — July 2024 – Sept. 2024
Advisor: Prof. Ting Wang — New York, United States

- Focus on security issues in LLM applications, particularly **LLM jailbreaking** and **poisoning RAG system**.
- Exploring vulnerabilities in the security alignment mechanism, we fine-tune a jailbreak prompt generator with low perplexity prompts and propose an efficient method for jailbreaking LLMs.
- Analyzing the security vulnerabilities of Graph RAG, we utilize instruction injection attacks on the LLM used for graph construction to manipulate the graph index and achieve knowledge poisoning.

Publications

Diffusion Noise Feature: Accurate and Fast Generated Image Detection

Yichi Zhang, Xiaogang Xu

Revisiting Defense Mechanisms in Federated Learning: Effective and Efficient Backdoor Attack via Trigger Pre-optimization

Xi Chen, Rui Zeng, ***Yichi Zhang***, Chunyi Zhou, Yuwen Pu, Qingming Li, Lu Zhou, Zhe Liu, Shouling Ji

Project

Dynamic 3D Point Cloud Compression Technology Based on Neural Implicit Representation

Undergraduate Graduation Project, Advisor: Prof. Qianqian Yang

Parallel Solving for Power-Ground Network Simulation Based on Machine Learning

Student Research Training Project, Advisor: Prof. Cheng Zhuo