

Jiachen Ma

☎ +86-13011707397 | ✉ majiachen@pjlab.org.cn | 📄 Google scholar | 🔗 mjc-ma-01 |

EDUCATION

Fudan University & Shanghai AI Lab (Joint Program) <i>Ph.D. in Computer Science</i>	Shanghai, China <i>Sep. 2025 – Present</i>
Zhejiang University <i>M.S. in Software Engineering</i>	Zhejiang, China <i>Sep. 2022 – Mar. 2025</i>
Hainan University <i>B.S. in Internet of Things Engineering</i>	Hainan, China <i>Sep. 2018 – Jun. 2022</i>

RESEARCH EXPERIENCE

Shanghai Artificial Intelligence Laboratory <i>Research Intern, advised by Prof. Bowen Zhou and Prof. Chao Yang</i>	Jul. 2024 – Present <i>Shanghai, China</i>
• Research Areas: Trustworthy Machine Learning; LLM/VLM Safety; Indirect Jailbreaks; Alignment. • Designed a “Reflection-Enhanced” defense framework and proposed Reflector to internalize step-wise reflection mechanisms, neutralizing sophisticated indirect jailbreaks without retraining. • Proposed ReSAM, a representation-level safety margin alignment method; discovered that safety gradients concentrate in low-rank subspaces, achieving 94.6% safety with only 5 samples.	
Zhejiang University Bigdata Lab <i>Graduate Researcher, advised by Prof. Junbo Zhao</i>	Sep. 2022 – Mar. 2025 <i>Zhejiang, China</i>
• Research Areas: Text-to-Image Safety; Diffusion Models; Adversarial Attacks. • Proposed JPA, the first automated, black-box attack framework targeting text-to-image models (e.g., Stable Diffusion) and online services (DALL-E 2, Midjourney).	

PUBLICATIONS AND PREPRINTS

Reflector: Internalizing Step-wise Reflection against Indirect Jailbreaks

Jiachen Ma, Jiawen Zhang, Xiangtian Li, Chaochao Lu, Chao Yang. *Submitted to ICML 2026.*

ReSaM: Representation-Level Safety Margin Alignment for Vision-Language Models

Jiachen Ma, Jiawen Zhang, Chaochao Lu, Chao Yang. *Submitted to ICML 2026.*

Jailbreaking Prompt Attack: A Controllable Adversarial Attack against Diffusion Models

Jiachen Ma, Yijiang Li, Zhiqing Xiao, Anda Cao, Jie Zhang, Chao Ye, Junbo Zhao. *NAACL 2025.*

CAME: Contrastive Automated Model Evaluation

Ru Peng, Qiuyang Duan, Haobo Wang, Jiachen Ma, Yanbo Jiang, Yongjun Tu, Xiu Jiang, Junbo Zhao. *ICCV 2023.*

HONORS AND AWARDS

2025 Outstanding Graduate Student (Top 5%)	Zhejiang University
2023 Outstanding Member of the Communist Youth League	Zhejiang University
2022 Baogang National Scholarship (Top 1%)	Hainan University
2019 & 2021 First Class Academic Scholarship (Top 1%)	Hainan University
2022 Outstanding Graduate of Hainan Province (Top 5%)	Dept. of Education, Hainan

TECHNICAL SKILLS

Programming: Python, C/C++, JavaScript, HTML/CSS

Frameworks: PyTorch, TensorFlow, Hugging Face

Tools and Technologies: Git, Linux

Areas of Interest: LLM Red Teaming, Multimodal Alignment (VLM), Adversarial Robustness, Agent Safety