Xinwei Liu

+86-157-9789-6613 | ✉ sinwayliu@gmail.com | 🏠 thinwayliu.github.io | 🎓 Google Scholar

Ph.D. Candidate, University of Chinese Academy of Sciences, Beijing 100049, China

Research Interests

My research interests span trustworthy, robust, and privacy-preserving AI systems. Currently, my work is centered on:

- **Model Safety and Robustness:**
  - Studying adversarial, backdoor, data poisoning, and jailbreaking attacks on vision models, LLMs/VLMs, and diffusion models across both white-box and black-box settings.
  - Designing and mitigating robustness–alignment trade-offs, including how safety fine-tuning, RLHF, and instruction tuning affect model vulnerabilities.
- **Data Security, Privacy, and Copyright Protection:**
  - Developing adversarial machine learning methods for data copyright protection, including invisible watermarks, adversarial examples, and feature-level perturbations that prevent unauthorized training or extraction.
  - Leveraging machine unlearning to support data subject rights: forgetting individual samples, specific datasets, and sensitive or copyrighted concepts from trained models.
  - Studying concept erasure and targeted forgetting to remove harmful or proprietary concepts from generative and multimodal models while preserving utility.
- **Secure and Safe Agents:**
  - Investigating the security of LLM-based agents, including tool-using agents, retrieval-augmented agents, multi-agent systems, and agents built on Model Context Protocol (MCP).
  - Characterizing and defending against prompt injection, tool hijacking, environment manipulation, and cross-agent compromise in agentic workflows.

Education

- **University of Chinese Academy of Sciences** *Sep. 2020 - June 2026*
  Ph.D. Candidate in Computer Application Technology China
  - Advisor: Prof. Xiaochun Cao, currently the Dean of the School of Cyber Science and Technology, SYSU, China
  - GPA: 3.8/4.0
- **Nanchang University** *Sep. 2016 - June 2020*
  B.E. Degree in Information and Computing Science China
  - Grade: 3.7/4.0, Ranking: 1/51

Publications

C=Conference, J=Journal, P=Preprint, S=In Submission, T=Thesis

First Authors:

[P.1] **Xinwei Liu**, Xiaojun Jia, Yuan Xun, Hua Zhang, Xiaochun Cao (2025). **PersGuard: Preventing Malicious Personalization via Backdoor Attacks on Pre-trained Text-to-Image Diffusion Models**. arXiv preprint arXiv:2502.16167.

[C.1] **Xinwei Liu**, Xiaojun Jia, Yuan Xun, Simeng Qin, Xiaochun Cao (2025). **GeoShield: Safeguarding Geolocation Privacy from Vision-Language Models via Adversarial Perturbations**. In *Proceedings of the AAAI Conference on Artificial Intelligence 2026*

[C.2] **Xinwei Liu**, Xiaojun Jia, Yuan Xun, Siyuan Liang, Xiaochun Cao (2024). **Multimodal Unlearnable Examples: Protecting Data against Multimodal Contrastive Learning**. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM) 2024*, pp. 8024–8033.

[C.3] **Xinwei Liu**, Xiaojun Jia, Jindong Gu, Yuan Xun, Siyuan Liang, Xiaochun Cao (2024). **Does Few-shot Learning Suffer from Backdoor Attacks?**. In *Proceedings of the AAAI Conference on Artificial Intelligence 2024*, Vol. 38, No. 18, pp. 19893–19901.

[C.4] **Xinwei Liu**, Jian Liu, Yang Bai, Jindong Gu, Tao Chen, Xiaojun Jia, Xiaochun Cao (2022). **Watermark Vaccine: Adversarial Attacks to Prevent Watermark Removal**. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, Vol. 13674, pp. 1–17.

[J.1] **Xinwei Liu**, Yuchao Tang, Yixuan Yang (2019). **Primal-dual Algorithm to Solve the Constrained Second-order Total Generalized Variational Model for Image Denoising**. *Journal of Electronic Imaging*, Vol. 28, No. 4, 043017. DOI: 10.1117/1.JEI.28.4.043017.

Contributors:

[C.5] Jianbo Chen, **Xinwei Liu***, Siyuan Liang, Xiaojun Jia, Yuan Xun (2023). **Universal Watermark Vaccine: Universal Adversarial Perturbations for Watermark Protection**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2322–2329.

[C.6] Yuan Xun, Xiaojun Jia, **Xinwei Liu**, Hua Zhang (2025). **The Emotional Baby Is Truly Deadly: Does your Multimodal Large Reasoning Model Have Emotional Flattery towards Humans?**. In *Proceedings of the AAAI Conference on Artificial Intelligence 2026*

[C.7] Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, **Xinwei Liu**, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, Xiaochun Cao, Philip Torr (2024). **A Survey on Transferability of Adversarial Examples Across Deep Neural Networks**. *Transactions on Machine Learning Research (TMLR)*, 2024.

[J.2] Yuan Xun, Xiaojun Jia, Jindong Gu, **Xinwei Liu**, Qing Guo, Xiaochun Cao (2024). **Minimalism is King! High-Frequency Energy-based Screening for Data-efficient Backdoor Attacks**. *IEEE Transactions on Information Forensics and Security*, Vol. 19, pp. 4560–4571. DOI: 10.1109/TIFS.2024.3380821.

[J.3] Yuan Xun, Siyuan Liang, Xiaojun Jia, **Xinwei Liu**, Xiaochun Cao (2024). **CleanerCLIP: Fine-grained Counterfactual Semantic Augmentation for Backdoor Defense in Contrastive Learning**. *IEEE Transactions on Information Forensics and Security*

[P.2] Yuan Xun, Siyuan Liang, Xiaojun Jia, **Xinwei Liu**, Xiaochun Cao (2025). **RobustIT: Adapter-Centric and Attack-Agnostic Anti-Backdoor Instruction Tuning**. arXiv preprint arXiv:2506.05401.

Honors and Awards

- **China National Scholarship (Doctoral Level)** *November 2024*
  Ministry of Education of the People's Republic of China
- **China National Scholarship (Undergraduate Level)** *December 2019*
  Ministry of Education of the People's Republic of China
- **Third Prize in 2025 Qiyuan Large Model Adversarial Challenge (Ranked 1st)** *August 2025*
  Qiyuan Laboratory

Invited Talks

- **Unlearnable Examples and Shortcut** *June 2024*
  Invited speaker at JSPS-NSFC Joint Research 1st Workshop, online
- **Adversarial Machine Learning and Data Protection** *December 2024*
  Invited speaker at China Energy Investment, online

Educational and Teaching Activities

Teaching:

- **Artificial Intelligence Security** *Spring 2024*
  Invited Lecturer, Graduate Course at Sun Yat-sen University (Shenzhen)

Mentees:

- **Jianbo Chen**, B.E. of 2021 at Hunan University, Adversarial ML, Pub.: [C.5] *Nov. 2022 - Dec. 2023*

Internship Experience

- **Ant Group** *Mar. 2022 – Jun. 2023*
  Research Algorithm Intern, Machine Intelligence for Security Division China
  - Worked on privacy-preserving computer vision in large-scale security applications, focusing on protecting visual data against unauthorized access and misuse.
  - Led a research project as first author, resulting in one paper accepted to *ECCV 2022*.
- **China Telecom AI Research Institute, Frontier Interdisciplinary Research Center** *Mar. 2025 – Sep. 2025*
  AI Governance Algorithm Intern China
  - Participated in the design and construction of a red-teaming evaluation and attack benchmark for assessing the safety and robustness of large AI models and systems.
  - Co-authored a paper (under review at *Pattern Recognition*) based on the proposed benchmark.
  - Contributed to the open-source benchmark implementation.

Academic Services

- **Journal Reviewer:**
  - IEEE Transactions on Pattern Analysis and Machine Intelligence
  - IEEE Transactions on Dependable and Secure Computing
  - IEEE Transactions on Information Forensics and Security
  - Pattern Recognition
- **Conference Reviewer:** CVPR, ICCV, ECCV, ICLR, ICML, NeurIPS, AAAI, ACM MM.

Skills

- **Coding:** Python, C++, Matlab
- **Languages:** Mandarin Chinese (mother tongue), English (IELTS 7.0, full professional proficiency)