Research Area and Approach

I am Xinwei Liu, a Ph.D. candidate at the University of Chinese Academy of Sciences, with an research interest in creating trustworthy, robust, and privacy-preserving AI systems. The landscape of AI has undergone a paradigm shift, evolving from traditional architectures to modern foundations and, most recently, autonomous agents. While these advancements have allowed AI to permeate every aspect of our digital lives, they have also introduced a dual crisis of reliability and sovereignty. On the one hand, the escalating complexity of these models has exponentially expanded the attack surface, making security risks a critical issue. On the other hand, the intentsive data requirements for training such large-scale models increasingly conflict with the boundaries of user privacy and copyright protection.

The majority of my current research applies adversarial ML techniques as a mechanism for data privacy protection and preventing malicious model exploitation. In addition, I also investigate the security risks of various models, primarily vision models and multimodal models.

Current and Past Research

0.1 Data Privacy Protection

Early in my doctoral studies, the work of Ruiz et al. [1] on using adversarial examples as defensive tools inspired me to explore harnessing adversarial machine learning to counter malicious violations of privacy and copyright. My first project addresses robust visible watermarking [2]. Since DNNs can effectively remove watermarks, we propose a defense mechanism dubbed Watermark Vaccine. We optimize imperceptible perturbations on host images to proactively disrupt blind watermark-removal networks. Building on this, I also supervised a student to extend our approach to host-image–independent universal adversarial watermarks, demonstrating strong generalizability [3].

Similarly, we extend adversarial attack techniques to more advanced VLMs and to the geoprivacy in images. [4]. As VLMs like GPT-4o can accurately infer locations from images, we propose GeoShield, a novel framework for real-world scenarios. Addressing the limitations of existing methods on high-resolution images, GeoShield utilizes feature disentanglement, exposure-element identification, and scale-adaptive enhancement to jointly optimize perturbations, effectively protecting geoprivacy without degradation.

Adversarial attacks also have been proposed to mitigate privacy risks in diffusion model personalization; however, we find these perturbation-based methods impractical and easily bypassed by data transformations. In contrast, we propose PersGuard, a novel framework leveraging model backdoors to prevent unauthorized personalization. [5] PersGuard embeds persistent backdoors into the pre-trained model: fine-tuning on protected data triggers defensive behaviors, while standard inputs yield normal outputs. We formulate the injection as a unified optimization problem incorporating a backdoor retention loss to withstand downstream adaptation.

Beyond adversarial attacks, I explore unlearnable examples to prevent unauthorized data exploitation. For Multimodal Contrastive Learning (MCL), we propose Multi-step Error Minimization [6], a framework that jointly optimizes visual perturbations (via PGD) and text triggers (via HotFlip) to induce models to memorize artificial shortcuts rather than semantic content. Building on this, I am also initiating a new line of work that systematically compares shortcut behaviors in conventional image classification versus MCL.

0.2 Model Security Risk

In addition to leveraging adversarial ML techniques for data privacy protection, I also conduct research on model security from both the attack and defense perspectives.

In earlier work, My colleagues and I conducted a comprehensive survey on transfer-based adversarial attacks in image classification [7], systematically categorizing methods that enable cross-architecture transferability. In this work, I led the analysis of optimization-based algorithms. Building on this foundation, I have recently extended transfer attacks to VLMs, specifically leveraging these techniques to test the robustness of closed-source systems.

My collaborators and I have also devoted effort to explore backdoor threats. A key focus of work has been evaluating the security of Few-Shot Learning (FSL) [8]. While our initial analysis showed that previous attacks struggle in FSL settings due to overfitting and poor stealthiness, we demonstrate that this resistance is fake. To prove this, we propose a attack, which generates triggers that maximize feature separation to prevent overfitting and utilizes imperceptible perturbations to provide stealthiness. In addition, I have contributed to works on backdoor defense [9, 10] and backdoor detection [11]. However, due to the prohibitive scale of modern pre-training datasets, successfully implanting backdoors via data poisoning is becoming increasingly difficult and less realistic. Consequently, this direction could not be the central point of my future research. .

This year, I expanded my focus to LLM and Agent security. During my internship at TeleAI, I co-developed a novel red-teaming benchmark for LLM safety, specifically leading the integration of diverse attack algorithms; this work is currently under review at Pattern Recognition. Concurrently, I am working a study on Vision-Language Retrieval-Augmented Generation (VL-RAG) poisoning. Leveraging the structural reliance of VLRAG systems on external databases, I design black-box poisoning strategies to induce targeted hallucinations, thereby exposing critical vulnerabilities in retrieval-dependent architectures.

Future Research Directions

Based on my current and past research, I outline several potential research directions for the future, keeping in mind the fast-paced development of AI technologies, which may lead to adjustments in my plans. During my postdoc work, my goal is not only to produce more research outputs but also to ensure that these contributions are impactful in community.

Model Security in LLMs and LVLMs I aim to explore emerging security risks in LLMs and LVLMs, such as more effective adversarial attacks or jailbreak techniques. Additionally, I hope to work on areas including safety fine-tuning and model alignment. Due to resource limitations in our research group, my colleagues and I are not yet very familiar with this direction. If possible, I would like to contribute to this area, as I believe it has the potential to provide meaningful contributions to the community.

Agent Security Starting this year, I have explored the literature on agents and agent security, and I believe this is a promising direction. Existing studies on agent attacks often remain relatively naive, while agents can be highly vulnerable. Therefore, I look forward to conducting research on both offensive and defensive strategies related to agent security.

Embodied AI Security I believe embodied intelligence could be a crucial component of future AGI. Recently, several papers have explored adversarial attacks and defenses on VLA models. While this area may not be as actively researched as LLM security at present, I consider its long-term potential to be profoundly significant. Although I have not directly worked on VLA, I have previously engaged with adversarial attacks on point clouds and physical adversarial attacks.

References

[1] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European conference on computer vision*, pages 236–251. Springer, 2020.

[2] Xinwei Liu, Jian Liu, Yang Bai, Jindong Gu, Tao Chen, Xiaojun Jia, and Xiaochun Cao. Watermark vaccine: Adversarial attacks to prevent watermark removal. In *European conference on computer vision*, pages 1–17. Springer Nature Switzerland Cham, 2022.

[3] Jianbo Chen, Xinwei Liu, Siyuan Liang, Xiaojun Jia, and Yuan Xun. Universal watermark vaccine: Universal adversarial perturbations for watermark protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 2322–2329, 2023.

[4] Xinwei Liu, Xiaojun Jia, Yuan Xun, Simeng Qin, and Xiaochun Cao. Geoshield: Safeguarding geolocation privacy from vision-language models via adversarial perturbations. *AAAI*, 2026.

[5] Xinwei Liu, Xiaojun Jia, Yuan Xun, Hua Zhang, and Xiaochun Cao. Persguard: Preventing malicious personalization via backdoor attacks on pre-trained text-to-image diffusion models. *arXiv preprint arXiv:2502.16167*, 2025.

[6] Xinwei Liu, Xiaojun Jia, Yuan Xun, Siyuan Liang, and Xiaochun Cao. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8024–8033, 2024.

[7] Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023.

[8] Xinwei Liu, Xiaojun Jia, Jindong Gu, Yuan Xun, Siyuan Liang, and Xiaochun Cao. Does few-shot learning suffer from backdoor attacks? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19893–19901, 2024.

[9] Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. Cleanerclip: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning. *arXiv preprint arXiv:2409.17601*, 2024.

[10] Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. Ta-cleaner: A fine-grained text alignment backdoor defense strategy for multimodal contrastive learning. *arXiv e-prints*, pages arXiv–2409, 2024.

[11] Yuan Xun, Xiaojun Jia, Jindong Gu, Xinwei Liu, Qing Guo, and Xiaochun Cao. Minimalism is king! high-frequency energy-based screening for data-efficient backdoor attacks. *IEEE Transactions on Information Forensics and Security*, 19:4560–4571, 2024.