

# BINYAN XU

✉ binyxu@ie.cuhk.edu.hk • 📞 +86 159-5123-4880 • 🏠 Homepage • 🎓 Scholar • 🔗 LinkedIn

## Education

<b>The Chinese University of Hong Kong</b> Ph.D. student in Information Engineering	<b>Sep. 2023 – Aug. 2027</b> GPA: 3.86
<b>University of California, Berkeley</b> Exchange Program in EECS	<b>Aug. 2021 – Dec. 2021</b> GPA: 4.0
<b>Xi'an Jiaotong University</b> B.Sc. in Automation	<b>Sep. 2019 – Jul. 2023</b> Honor, GPA: 3.92

## Research Interests

- Deep Neural Networks Security, including Backdoor Attacks & Defenses and Adversarial Attacks & Defenses
- Application of Advanced Technologies (LLM, VLM, Diffusion) in AI Security.
- Security of Embodied AI and Other LLM/VLM Security Issues

## Publications (First Author)

**Binyan Xu et al.**. One Surrogate to Fool Them All: Universal, Transferable, and Targeted Adversarial Attacks with CLIP. In *ACM Conference on Computer and Communications Security (CCS '25)*, 2025, **Oral Presentation (CCF-A)**.

**Binyan Xu et al.**. CLIP-Guided Backdoor Defense through Entropy-Based Poisoned Dataset Separation. In *ACM International Conference on Multimedia (MM '25)*, 2025, **Oral Presentation (CCF-A)**.

**Binyan Xu et al.**. Breaking the Stealth-Potency Trade-off in Clean-Image Backdoors with Generative Trigger Optimization. In *AAAI Conference on Artificial Intelligence (AAAI '26)*, 2026, **Oral Presentation (CCF-A)**.

## Projects

### AI-Aided Ascetic Graphic Design Generation

Microsoft Research Asia (MSRA), Stars of Tomorrow Internship Program

*Open-Ended Generation, Diffusion Model, Large Languages Model* **Jul. 2022 – Jun. 2023**

- Developed an ascetic graphic design system with an autoregressive vision-language model for layout generation.
- Enhanced this graphic design system with diffusion model to generate additional decorative patterns.
- Significantly outperforms previous approaches in both better generation quality and less constraint violation.

### Vision-Language Model Facilitated Neural Network Attacks and Defenses

Laboratory for Applied Security Research, The Chinese University of Hong Kong

*Transferable Adversarial Attack, Backdoor Defense, Vision-Language Model* **Apr. 2024 – Present**

- Involved in UnivIntruder, an attack method targeted vision models without direct access to target models or datasets.
- Achieved 99.4% Attack Success Rate in 4 datasets, 84% on Google and Baidu, and 70% on GPT-4 and Claude-3.5.
- Proposed CLIP-guided backdoor defense using entropy-based data separation, defending all kinds of backdoor attacks.
- 1 paper accepted in ACM CCS 25', 1 in MM 25', both first-authored.

### Stealthy and Adaptive Clean-Image Backdoor Attacks

Laboratory for Applied Security Research, The Chinese University of Hong Kong

*Backdoor Attack, Generative Adversarial Networks* **Sep. 2023 – Present**

- Proposed an InfoGAN-based stealthy backdoor attack, achieving  $\geq 90\%$  ASR with  $\leq 1\%$  Clean Accuracy drop.
- Showcased adaptability to classification, regression, and segmentation tasks with a 0.1% poison rate.
- 1 paper accepted in AAAI in first-authored.

## Selected Awards

- Finalist Award, Mathematical Contest in Modeling
- Outstanding Graduates Scholarship
- Qian Xuesen Program Honorary Graduate
- First-class Funding for Studying Abroad (about ¥100k)

## Skills

**Languages:** Chinese (Native), English (Fluent); TOEFL: 104 (R29/L28/S22/W25), CET-6: 584  
**Programming:** Python, PyTorch, TensorFlow, C/C++, MATLAB