徐彬琰

香港中文大学　信息工程系
binyxu@ie.cuhk.edu.hk • +86 159-5123-4880 • 个人主页 • 谷歌学术

教育背景

香港中文大学 (CUHK) **2023.9 – 2027.8 (预计)**
博士研究生，信息工程，导师：张克环（正教授） GPA: 3.86

加州大学伯克利分校 (UC Berkeley) **2021.8 – 2021.12**
交换项目，电子工程与计算机科学 (EECS) GPA: 4.0

西安交通大学 (XJTU) **2019.9 – 2023.7**
工学学士，自动化 荣誉毕业生 (钱学森班), GPA: 3.92

第一作者论文 (4 CCF-A)

Binyan Xu *et al.*. From Multi-Agent to Single-Agent: When Is Skill Distillation Beneficial? 审稿中.

Binyan Xu *et al.*. Contextual Agentic Memory is a Memo, Not True Memory. 审稿中.

Binyan Xu *et al.*. Reviewer Scores Are Not Comparable Across Research Areas in ML Peer Review. 审稿中.

Binyan Xu *et al.*. From Internal Diagnosis to External Auditing: A VLM-Driven Paradigm for Data-Free Online Backdoor Defense. In *ICML '26* **(CCF-A)**.

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

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

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

实习经历

AI智能体研究：单智能体与多智能体系统 **2026.1 – 至今**

腾讯，青云计划实习 | LLM Agent，技能蒸馏，多智能体系统，单智能体系统

- 研究将多智能体系统（MAS）蒸馏为单智能体技能（Skill）在何种条件下能带来性能提升。
- 提出 Metric Freedom，一种在构建技能前即可量化技能收益的原则性预测指标。
- 自适应技能在4项任务、11个数据集上以1.4–15倍更低成本达到或超越原始MAS性能。
- 1篇论文投稿至 NeurIPS（审稿中）。

AI辅助平面设计生成 **2022.7 – 2023.6**

微软亚洲研究院 *(MSRA)*，明日之星实习项目 | 扩散模型，大语言模型

- 开发了一套利用自回归视觉-语言模型进行布局生成的极简主义平面设计系统。
- 引入扩散模型（Diffusion Model）增强系统能力，用于生成额外的装饰性图案。
- 在生成质量和约束满足方面均显著优于现有方法。

科研项目

基础大模型辅助的神经网络攻防研究 **2023.9 – 至今**

香港中文大学应用安全研究实验室 | 后门攻防，对抗攻击，视觉语言模型，扩散模型

- 参与 UnivIntruder 研究，一种基于 CLIP 的可迁移攻击，在4个数据集上达到99.4% ASR，Google/百度上达到84%。
- 提出基于熵的毒化数据集分离的 CLIP 引导后门防御方法，可防御各类后门攻击。
- 提出以 VLM 作为外部审计者的无参考数据在线后门防御范式。
- 3篇论文已录用（CCS '25、MM '25、AAAI '26，均为 CCF-A Oral）；1篇投稿 ICML '26（平均分 4.25）。

相关技能

语言能力: 中文（母语），英语（流利）；托福: 104 (R29/L28/S22/W25)，CET-6: 584
编程语言: Python, PyTorch, TensorFlow, C/C++, MATLAB