# 徐彬琰

binyxu@ie.cuhk.edu.hk • +86 159-5123-4880 • 个人主页 • Semantic Scholar • GitHub • LinkedIn

## 教育背景

**香港中文大学** **2023.9 – 2027.10（预计）**
信息工程博士研究生；导师：张克环教授 *GPA: 3.914*

**加州大学伯克利分校** **2021.8 – 2021.12**
电子工程与计算机科学交换项目 *GPA: 4.0*

**西安交通大学** **2019.9 – 2023.7**
自动化工学学士，钱学森学院，钱学森班 *GPA: 3.92*；排名：前 *5%*

- 通过少年班从初中阶段进入西安交通大学；该选拔项目从 4,000+ 名申请者中录取约 100 人。

## 研究主线与定位

### Agentic LLM 能力构建与安全边界：长程交互、经验学习与行为验证

- 围绕长程开放环境中的 **Agentic LLM 能力构建**，研究工具调用、多轮交互、运行状态管理、GRPO 后训练、轨迹/技能学习与经验巩固。
- 同时研究**模型行为验证与安全边界**，覆盖训练数据完整性、跨模型攻击、生成式后门、VLM 外部审计与自主执行风险。
- 主导 4 篇 CCF-A 已录用、4 篇在投的第一作者工作；当前关注将 tool-use/multi-turn RL、trajectory learning 与 model-agnostic verification 用于网络攻防、威胁分析等 Cyber 场景。

## 代表论文与投稿

*4 篇 CCF-A 类会议已录用、4 篇在投，均为第一作者工作。*

### 一、Agentic LLM 能力构建：状态管理、工具交互与技能学习

**[1] LLM Agents Are Latent Context Managers: Eliciting Self-Management via Proprioception** 在投
Binyan Xu, Haitao Li, Kehuan Zhang.
Agentic RL：显式暴露运行状态，构建 context-tool 动作空间并以 GRPO 学习自适应策略。

**[2] From Multi-Agent to Single-Agent: When Is Skill Distillation Beneficial?** 在投
Binyan Xu, Dong Fang, Haitao Li, Kehuan Zhang.
轨迹与技能学习：判断何时可将多智能体协作及评测闭环内化为单智能体能力。

**[3] Contextual Agentic Memory is a Memo, Not True Memory** 在投
Binyan Xu*, Xilin Dai*, Kehuan Zhang. *Equal contribution.
经验分析：通过消融与统计分析，检验检索式记忆的泛化上限和持续性风险。

### 二、复杂工具执行、外部核验与风险评测

**[4] When Agent Automation Becomes Profitable: Quantify and Insure Autonomous AI Risk** 在投
Binyan Xu, Xilin Dai, Fan Yang, Kehuan Zhang.
轨迹级评测：将 Agent 工具交互轨迹映射为任务损失、控制信号与部署风险。

**[5] From Internal Diagnosis to External Auditing: A VLM-Driven Paradigm for Online Defense** ICML 2026
Binyan Xu, Xilin Dai, Fan Yang, Di Tang, Kehuan Zhang. Poster
外部评测：以独立 VLM 在线核验不可信模型行为，构建模型无关的 verifier。

### 三、基础模型攻防与训练数据完整性

**[6] CLIP-Guided Backdoor Defense through Entropy-Based Poisoned Dataset Separation** ACM MM
Binyan Xu, Fan Yang, Xilin Dai, Di Tang, Kehuan Zhang. Oral Presentation
数据完整性：利用 CLIP 语义信号分离毒化训练样本，并指导模型移除隐藏行为。

**[7] One Surrogate to Fool Them All: Universal and Transferable Adversarial Attacks with CLIP** CCS 2025
Binyan Xu, Xilin Dai, Di Tang, Kehuan Zhang. Oral Presentation
威胁分析：利用公开模型构造跨模型、跨任务、跨真实服务的通用定向攻击。

**[8] Breaking Stealth-Potency Trade-off in Clean-Image Backdoors with Trigger Optimization** AAAI 2026
Binyan Xu, Xilin Dai, Fan Yang, Di Tang, Kehuan Zhang. Oral Presentation
攻防交互：以生成式触发器优化刻画隐蔽攻击，并扩展至回归和分割任务。

## 项目经历

### 长程 Agentic LLM：状态管理、后训练与技能学习 **2026.1 – 2026.7**

腾讯 *IEG*，光子工作室群，*X-Data Research Team*；青云计划实习 全职研究实习生

- 主导 *LLM Agents Are Latent Context Managers* [1]，提出智能体缺失对自身上下文状态的 *proprioception*，设计 VISTA self-managed context 层，将工作记忆表示为 typed/addressable blocks，并暴露 token usage、recency 与 budget 等状态。
- 在 LOCA-Bench 75 个长程工具任务上将 Gemini-3-Flash 成功率从 22.7% 提升至 50.7%，优于 ReAct、SLIM、Active Context Compression 与 Claude Code。
- 构建 context-tool 动作空间及 GRPO 后训练/评测闭环：显式状态接口在 BC+ 同分布/GAIA→BC+ 跨分布达 31.3%/21.3%，超过对应基线 27.3%/17.3%，验证其可协同提升策略能力与运行可靠性。
- 主导技能蒸馏研究 [2]，提出 *Metric Freedom* 判断何时可将多智能体协作/评测结构内化为单 agent 技能；在 4 项任务、11 个数据集上以 1.4–15× 更低成本达到或超过原系统。
- 共同主导 agentic memory 研究 [3]，指出当前“记忆”模块多为上下文扩展而非真正记忆系统；构建评测、消融、统计分析与可复现 artifact，并分析其泛化上限与持续性投毒风险。

### 可控生成式 AI：美学平面设计自动化 **2022.7 – 2023.6**

微软亚洲研究院，*Data, Knowledge and Intelligence*；明日之星实习项目 全职研究实习生

- 共同开发 AI 辅助美学设计系统，结合版式规划、视觉语言理解与扩散式风格/装饰生成，面向真实设计中的可控内容生成。
- 实现可控生成模块，在提升视觉质量的同时降低版式、元素关系与设计约束违背，为后续研究可控生成、约束满足与人机反馈闭环提供应用侧经验。
- 获得微软亚洲研究院 Stars of Tomorrow Internship Program 的 *Award of Excellence*。

### 基础模型全链路安全：数据完整性、外部审计与攻击面分析 **2023.9 – 2026.1**

香港中文大学应用安全研究实验室 博士阶段研究

- 围绕**不完全可信模型的行为核验**，覆盖自主 AI 风险、VLM 外部审计、训练数据完整性、CLIP 迁移攻击与生成式 clean-image 后门 [4–8]。
- 提出以 VLM 作为外部审计者的数据无关在线后门防御方法 [5]，将防御从模型内部诊断转向外部语义审计，适用于不完全信任被部署模型的场景。
- 设计生成式 clean-image 后门 [8]，并在真实 AI 服务和多模态模型上验证跨系统迁移攻击 [7]，刻画开放基础模型能力被滥用时的攻击面。
- 构建 CLIP 引导的毒化数据分离与模型修复流程 [6]；产出 ICML、AAAI、CCS 与 ACM MM 等 CCF-A 论文，形成覆盖数据、模型、部署与运行阶段的安全研究积累。

## 开源与影响力

- 在 github.com/binyxu 开源第一作者工作代码，覆盖 long-horizon agent 评测、VISTA self-managed context、agentic memory、技能蒸馏、VLM 外部审计、后门攻防与对抗迁移。
- Agent-memory 工作发布后引发较广泛讨论，包括 DAIR.AI 在 X 上转发（26K+ 浏览、700+ 点赞/收藏）、Facebook AI 社区、YouTube 解读、小红书转载与新智元报道。

## 荣誉与技能

- **学术服务**: ICML、ICLR、ACM MM、AAAI 正式审稿人；AAAI、ACM CCS、ACM MM 口头报告，ICML poster 展示；Lead TA 指导 40+ 名本科毕业设计学生。
- **荣誉**: 美国大学生数学建模竞赛 Finalist（Top 2%）；入选西安交通大学少年班（4,000+ 名申请者中前 2.5%）；钱学森班荣誉毕业生/优秀毕业生；伯克利交换项目一等资助；微软亚洲研究院 Stars of Tomorrow Award of Excellence。
- **研究方向**: Agentic LLM post-training、tool-use/multi-turn RL、trajectory/skill learning、behavior verification 与 Cyber 场景适配；基础模型安全。
- **编程与工具**: Python（主力）；PyTorch、Hugging Face、vLLM、FSDP、veRL；GRPO 后训练、agent harness、多轮环境交互、tool-use evaluation 与多卡实验。
- **语言**: 中文（母语），英语（流利；TOEFL 104）。