

# Xiang Zheng

---

|                            |  |  |
|----------------------------|--|--|
| CONTACT INFORMATION        |  | Email: xiang.zheng@cityu.edu.hk<br>Webpage: <a href="https://x-zheng16.github.io">https://x-zheng16.github.io</a><br>GitHub: <a href="https://github.com/x-zheng16">https://github.com/x-zheng16</a><br>Google Scholar: 692 citations, h-index 9 |
| RESEARCH INTERESTS         | Reinforcement Learning, Trustworthy AI, Agentic AI, Embodied AI  |  |
| EDUCATION                  | <b>City University of Hong Kong</b> , Hong Kong SAR, China<br><i>Ph.D. in Computer Science</i><br>Supervisor: Prof. Cong Wang<br>Fields: Reinforcement Learning, Trustworthy AI<br>GPA: 4.00/4.00  | <b>2020-2024</b>   |
|                            | <b>Tsinghua University</b> , Beijing, China<br><i>M.S. in Control Science and Engineering</i><br>Supervisor: Prof. Tao Zhang<br>Fields: Reinforcement Learning, Intelligent Robotics<br>GPA: 3.55/4.00   | <b>2016-2019</b>   |
|                            | <b>Beihang University</b> , Beijing, China<br><i>B.S. in Automation, B.S. in Mathematics (Dual Degree)</i><br>Selected into the Shen Yuan Honors College (1/200)<br>GPA: 3.74/4.00 (222.5 credits, 92 courses)   | <b>2012-2016</b>   |
| RESEARCH & WORK EXPERIENCE | <b>City University of Hong Kong</b> , Hong Kong SAR, China<br><i>Postdoctoral Fellow</i><br>Supervisor: Prof. Cong Wang<br>Topic: Reinforcement learning for trustworthy & embodied AI.<br>- Reinforced defense for VLMs, accepted by ICLR'25.<br>- Curiosity-driven auditing for LLMs, accepted by AAAI'25.   | <b>2024-present</b>  |
|                            | <b>City University of Hong Kong</b> , Hong Kong SAR, China<br><i>Graduate Research Assistant</i><br>Supervisor: Prof. Cong Wang<br>Topic: Efficient exploration strategies for reinforcement learning<br>- Designed constrained intrinsic motivation for unsupervised reinforcement learning. This work is accepted by IJCAI 2024.<br>- Developed intrinsically motivated adversarial policy against robotic RL agents. This work is accepted by DSN 2024. | <b>2020-2024</b>   |

**Xi'an Jiaotong University**, Xi'an, China **2019**

*Visiting Researcher*

Supervised by Prof. Chao Shen & Dr. Xingjun Ma

Topic: Adversarial machine learning

**National Institute of Informatics**, Tokyo, Japan **2018**

*Research Intern*

Supervised by Prof. Tetsunari Inamura

Topic: Intelligent robotics

**Tsinghua University**, Beijing, China **2016-2019**

*Graduate Research Assistant*

Supervised by Prof. Tao Zhang

Topic: Intelligent space robot

**The University of New South Wales**, Sydney, Australia **2016**

*Research Intern*

Supervised by Prof. Elias Aboutanios

Topic: Spacecraft de-orbiting control

**PUBLICATION**      **Journal**

- J-1. Xingjun Ma, ..., **Xiang Zheng**, et al., “Safety at Scale: A Comprehensive Survey of Large Model and Agent Safety,” *Foundations and Trends® in Privacy and Security*, vol. 8, pp. 254-469, 2026.
- J-2. Yuxue Cao, Shengjie Wang, **Xiang Zheng**, Wenke Ma, Xinru Xie, Lei Liu, “Reinforcement Learning With Prior Policy Guidance for Motion Planning of Dual-Arm Free-Floating Space Robot,” *Aerospace Science and Technology*, vol. 136, pp. 108098, 2023.  
**Impact Factor 5.6, JCR Rank Q1 (3/34).**
- J-3. Shengjie Wang, Yuxue Cao, **Xiang Zheng**, Tao Zhang, “A Learning System for Motion Planning of Free-Float Dual-Arm Space Manipulator Towards Non-cooperative Object,” *Aerospace Science and Technology*, vol. 131, pp. 107980, 2022.  
**Impact Factor 5.6, JCR Rank Q1 (3/34).**
- J-4. Shengjie Wang, Yuxue Cao, **Xiang Zheng**, Tao Zhang, “Collision-Free Trajectory Planning for a 6-DoF Free-Floating Space Robot via Hierarchical Decoupling Optimization,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 4953–4960, 2022.  
**Impact Factor 5.2, JCR Rank Q2 (10/30).**

**Conference (\* for the corresponding author.)**

- C-1. Yunhan Zhao, **Xiang Zheng\***, Lin Luo, Yige Li, Xingjun Ma\*, Yu-Gang Jiang, “BlueSuffix: Reinforced Blue Teaming for Vision-Language Models Against Jailbreak Attacks,” in **ICLR**, 2025.
- C-2. **Xiang Zheng**, Longxiang Wang, Yi Liu, Xingjun Ma, Chao Shen, Cong Wang\*, “CALM: Curiosity-Driven Auditing for Large Language Models,” in **AAAI**, 2025.
- C-3. **Xiang Zheng**, Xingjun Ma, Chao Shen, and Cong Wang\*, “Constrained Intrinsic Motivation for Reinforcement Learning,” in **IJCAI**, 2024.  
**Acceptance Rate 791/5651=14.00%**
- C-4. **Xiang Zheng**, Xingjun Ma, Shengjie Wang, Xinyu Wang, Chao Shen, and Cong Wang\*, “Toward Evaluating Robustness of Reinforcement Learning With Adversarial Policy,” in **DSN**, 2024.  
**Acceptance Rate 42/203=20.69%**
- C-5. Shengjie Wang, Fengbo Lan, **Xiang Zheng**, Yuxue Cao, Oluwatosin Oseni, Haotian Xu, Tao Zhang, Yang Gao, “A Policy Optimization Method Towards Optimal-Time Stability,” in **CoRL**, 2023.
- C-6. Shengjie Wang, **Xiang Zheng**, Yuxue Cao, Tao Zhang, “A Multi-target Trajectory Planning of a 6-DoF Free-Floating Space Robot via Reinforcement Learning,” in **IROS**, 2021.
- C-7. Shengjie Wang, Yuxue Cao, **Xiang Zheng**, Tao Zhang, “An End-to-End Trajectory Planning Strategy for Free-Floating Space Robots,” in **CCC**, 2021.
- C-8. Shihao Zhao, Xingjun Ma, **Xiang Zheng**, James Bailey, Jingjing Chen, Yu-Gang Jiang, “Clean-Label Backdoor Attacks on Video Recognition Models,” in **CVPR**, 2020.  
**Acceptance Rate 1470/6656 = 22.09%**
- C-9. **Xiang Zheng**, Ziwei Wang, Tao Zhang, “Robust Finite-Time Attitude Tracking Control for Nonlinear Quadrotor With Uncertainties and Delays,” in **ROBIO**, 2017.

### arXiv preprint & Submission

1. **Xiang Zheng**, Xingjun Ma, Wei-Bin Lee, Cong Wang\*, “RedRL: A Light-Weight Benchmark for Reinforcement Learning-Based Red Teaming,” arXiv preprint arXiv:2506.04302, submitted to **Frontiers of Computer Science**.
2. Zilong Wang, **Xiang Zheng\***, Xiaosen Wang, Bo Wang, Xingjun Ma\*, Yu-Gang Jiang, “GenBreak: Red Teaming Text-to-Image Generators Using Large Language Models,” arXiv preprint arXiv:2506.10047, submitted to **CVPR’26**.
3. Xutao Mao, Liangjie Zhao, Liutao, **Xiang Zheng**, Hongying Zan, Cong Wang, “STARE: Step-Wise Temporal Credit Assignment and Red-Teaming

- Engine for Multi-Modal Toxicity Attack”, submitted to **CVPR’26**.
4. Jiayu Li, Yunhan Zhao, **Xiang Zheng**, Zonghuan Xu, Yige Li, Xingjun Ma, Yu-Gang Jiang, “AttackVLA: Benchmarking Adversarial and Backdoor Attacks on Vision-Language-Action Models”, arXiv preprint arXiv:2511.12149, submitted to **CVPR’26**.
  5. Jiale Ding, **Xiang Zheng\***, Yutao Wu, Cong Wang, Wei-Bin Lee, Ling Pan, Xingjun Ma\*, Yu-Gang Jiang, “RedTopic: Toward Topic-Diverse Red Teaming of Large Language Models,” arXiv preprint arXiv:2507.00026, submitted to **ICLR’26**.
  6. Sirui He, Chujie Chen, **Xiang Zheng**, Zhihang Liu, Cong Wang, “Refining Specs For LLM-Based RTL Agile Design,”, submitted to **ICLR’26**.
  7. Longxiang Wang, **Xiang Zheng**, Xuhao Zhang, Yao Zhang, Ye Wu, Cong Wang, “RL<sup>2</sup>eak: Reinforcement Learning Enhanced Prompt Leakage Attack in Multi-tenant Large Language Model Services,” submitted to **ICLR’26**
  8. Yutao Wu, Xiao Liu, Yinghui Li, Yifeng Gao, Yifan Ding, Jiale Ding, **Xiang Zheng**, Xingjun Ma, “ADMIT: Few-shot Knowledge Poisoning Attacks on RAG-based Fact Checking,” arXiv:2510.13842, submitted to **ICLR’26**.
  9. Yunhan Zhao, **Xiang Zheng**, Xingjun Ma, “Defense-to-Attack: Bypassing Weak Defenses Enables Stronger Jailbreaks in Vision-Language Models,” arXiv preprint arXiv:2509.12724, submitted to **ICASSP’26**.
  10. Ruofan Wang, **Xiang Zheng**, Xiaosen Wang, Cong Wang, Xingjun Ma, “RedDiffuser: Red Teaming Vision-Language Models for Toxic Continuation via Reinforced Stable Diffusion,” arXiv preprint arXiv:2503.06223, submitted to **ICDE’26**.
  11. Zonghuan Xu, **Xiang Zheng\***, Xingjun Ma, Yu-Gang Jiang, “TabVLA: Targeted Backdoor Attacks on Vision-Language-Action Models,” arXiv preprint arXiv:2510.10932, submitted to **ICRA’26**.

## PROJECTS

## Participant

1. **Foxconn-CityUHK Joint Research Centre:** Intrinsically Motivated Reinforcement Fine-Tuning for Red Teaming Black-Box LLMs  
2025 → 2026, 500,000 HKD
2. **CRF:** Enabling Metadata-Private and Accountable Networks at Scale  
No. 8730083, 2023 → ..., 7,520,000 HKD
3. **NSFC:** Towards Secure and Privacy-enhanced Machine Learning as a Service  
No. 9054034, 2022 → ..., 1,000,000 CNY
4. **RFS:** Building Privacy-Assured and Scalable Encrypted Databases With Secure Enclave  
No. 9062004, 2022 → ..., 5,155,380 HKD

5. **CAST**: Learning-Based \*\*\* \*\*\* Control Technology for \*\*\* Targets  
2017 → 2019, 1,800,000 CNY
6. \*\*\*\*\*: Research on Deep Learning Algorithms for Intelligent \*\*\*  
2018 → 2019, 600,000 CNY

|                    |  |                  |
|--------------------|--|------------------|
| INVITED<br>TALKS   | <b>Xi'an</b> , CCF YOCSEF<br>Towards Robust Embodied AI: Skill Discovery and Red Teaming   | <b>2025</b>      |
|                    | <b>Shenzhen</b> , Southern University of Science and Technology<br>Reinforcement Learning-Based Adversarial Safety Evaluation<br>and Defense Enhancement for Large Language Models | <b>2025</b>      |
|                    | <b>Xi'an</b> , CCF YOCSEF<br>Reinforcement Fine-Tuning for Large Model Red-/Blue-Teaming   | <b>2025</b>      |
|                    | <b>Xi'an</b> , Northwestern Polytechnical University<br>Curiosity-Driven Auditing for LLMs   | <b>2024</b>      |
|                    | <b>Xi'an</b> , Northwestern Polytechnical University<br>Efficient Intrinsically Motivated Adversarial Policy Learning  | <b>2024</b>      |
|                    | <b>Gold Coast</b> , Griffith University<br>Intrinsically Motivated Adversarial Policy  | <b>2024</b>      |
|                    | <b>Melbourne Connect</b> , The University of Melbourne<br>Towards Efficient Evasion Attacks Against RL   | <b>2024</b>      |
| AWARDS &<br>HONORS | <b>IJCAI Travel Grant</b><br>IJCAI Organization and AIJ Division<br>Amount: USD 500  | <b>2024</b>      |
|                    | <b>International Conference Grant</b><br>City University of Hong Kong<br>Amount: HKD 10,000  | <b>2024</b>      |
|                    | <b>DSN Student Travel Grant</b><br>DSN Student Travel Awards Committee<br>Amount: USD 1,500  | <b>2024</b>      |
|                    | <b>Research Activities Fund</b><br>City University of Hong Kong<br>Amount: HKD 96,000  | <b>2022-2023</b> |

|   |                   |
|---|-------------------|
| <b>Institutional Research Tuition Grant</b>                                   | <b>2020-2024</b>  |
| City University of Hong Kong  |                   |
| Amount: HKD 168,384   |                   |
| <b>CityU Presidential PhD Scholarship (2/73)</b>                              | <b>2020-2024</b>  |
| City University of Hong Kong  |                   |
| Amount: HKD 1,353,120   |                   |
| <b>NII MOU Research Activities Grant (1/106)</b>                              | <b>2018</b>       |
| National Institute of Informatics, Japan                                      |                   |
| Amount: JPY 68,400  |                   |
| <b>CSC Scholarship</b>  | <b>2016</b>       |
| China Scholarship Council   |                   |
| Amount: AUD 6,400   |                   |
| <b>Stars of Advanced Engineering (2/50)</b>                                   | <b>2015</b>       |
| Shen Yuan Honors College, Beihang University                                  |                   |
| The highest honor at Shen Yuan Honors College (2/50)                          |                   |
| <b>TEACHING &amp; EXPERIENCE</b>  |                   |
| <b>Teaching Assistant</b> , City University of Hong Kong                      | <b>Semester B</b> |
| CS5293, Topics on Information Security  | <b>2023/24</b>    |
| <b>Teaching Assistant</b> , City University of Hong Kong                      | <b>Semester A</b> |
| CS4394, Information Security and Management                                   | <b>2023/24</b>    |
| <b>Teaching Assistant</b> , City University of Hong Kong                      | <b>Semester B</b> |
| CS4293, Topics in Cybersecurity   | <b>2021/22</b>    |
| CS5293, Topics on Information Security  |                   |
| <b>Teaching Assistant</b> , City University of Hong Kong                      | <b>Semester A</b> |
| CS4394, Information Security and Management                                   | <b>2021/22</b>    |
| CS5294, Information Security Technology Management                            |                   |
| <b>Teaching Assistant</b> , City University of Hong Kong                      | <b>Semester B</b> |
| CS4293, Topics in Cybersecurity   | <b>2020/21</b>    |
| CS6290, Privacy-enhancing Technologies  |                   |
| <b>Teaching Assistant</b> , City University of Hong Kong                      | <b>Semester A</b> |
| CS2310 & CS2311, Computer Programming   | <b>2020/21</b>    |
| <b>MENTORED STUDENTS</b>  |                   |
| <b>Ph.D. Students</b>   |                   |
| Ruofan Wang (Fudan, VLM+Safety), Zilong Wang (Fudan, Diffusion Model+Safety), |                   |

Longxiang Wang (CityUHK, Agent+RL), Sirui He (CityUHK, LLM), Xutao Mao (CityUHK, LLM+Safety),  
Yutao Wu (Deakin, Agent+Safety)

### **Master Students**

Xiao Li (Fudan, RL), Yunhan Zhao (Fudan, LLM+RL), Shihao Zhao (Fudan, Safety),  
Zixing Chen (Fudan, VLA+Safety), Jiayu Li (Fudan, VLA),  
Chenghao Yao (CityUHK, LLM+RL), Shuxuan Lye (CityUHK, GUI Agent+Safety), Xiao Hu (CityUHK, AI Referee),  
Shengjie Wang (Tsinghua, RL), Yuxue Cao (CAS, RL), Siyuan Zhang (NWPU, LLM)

### **Undergraduate Students**

Jiale Ding (Fudan, LLM+RL), Zonghuan Xu (Fudan, LLM+VLA)

PROFESSIONAL **Program Committee Member**  
ACTIVITY ICML 2026, CVPR 2026, ICLR 2026, AAAI 2026, ICRA 2026, MM 2025, ICLR 2025, AAAI 2025

### **Journal Reviewer**

TDSC, TSC, TC

### **External Conference Reviewer**

NeurIPS 2025, ICNP 2025, ESORICS 2022, AsiaCCS 2022, RAID 2021

### **External Journal Reviewer**

IoT-J

MISC. Language: Python, C/C++  
Framework: PyTorch  
Simulator: MuJoCo  
Platform: Tianshou, Stable Baselines, CleanRL, TRL, etc.