# Xiang Zheng


| | | |
|---|---|---|
| Contact Information | YEUNG-G7352<br>City University of Hong Kong<br>Kowloon, Hong Kong SAR, China | Email: xiang.zheng@cityu.edu.hk<br>Webpage: https://x-zheng16.github.io<br>GitHub: https://github.com/x-zheng16<br>Google Scholar: 692 citations, h-index 9 |


Research Interests: Reinforcement Learning, Trustworthy AI, Agentic AI, Embodied AI

## Education

**City University of Hong Kong**, Hong Kong SAR, China **2020-2024**
*Ph.D. in Computer Science*
Supervisor: Prof. Cong Wang
Fields: Reinforcement Learning, Trustworthy AI
GPA: 4.00/4.00

**Tsinghua University**, Beijing, China **2016-2019**
*M.S. in Control Science and Engineering*
Supervisor: Prof. Tao Zhang
Fields: Reinforcement Learning, Intelligent Robotics
GPA: 3.55/4.00

**Beihang University**, Beijing, China **2012-2016**
*B.S. in Automation, B.S. in Mathematics (Dual Degree)*
Selected into the Shen Yuan Honors College (1/200)
GPA: 3.74/4.00 (222.5 credits, 92 courses)

## Research & Work Experience

**Hong Kong Institute of AI for Science (HKAI-Sci)**, Hong Kong SAR, China **2025-present**
*Research Assistant Professor*
Topic: Reinforcement learning for trustworthy & embodied AI.

**City University of Hong Kong**, Hong Kong SAR, China **2024-2025**
*Postdoctoral Fellow*
Supervisor: Prof. Cong Wang
Topic: Reinforcement learning for trustworthy & embodied AI.
- Reinforced defense for VLMs, accepted by ICLR'25.
- Curiosity-driven auditing for LLMs, accepted by AAAI'25.

**City University of Hong Kong**, Hong Kong SAR, China **2020-2024**
*Graduate Research Assistant*
Supervisor: Prof. Cong Wang
Topic: Efficient exploration strategies for reinforcement learning

- Designed constrained intrinsic motivation for unsupervised reinforcement learning. This work is accepted by IJCAI 2024.
- Developed intrinsically motivated adversarial policy against robotic RL agents. This work is accepted by DSN 2024.

**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$^2$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 Talks

**Xi'an**, CCF YOCSEF **2025**
Towards Robust Embodied AI: Skill Discovery and Red Teaming

**Shenzhen**, Southern University of Science and Technology **2025**
Reinforcement Learning-Based Adversarial Safety Evaluation and Defense Enhancement for Large Language Models

**Xi'an**, CCF YOCSEF **2025**
Reinforcement Fine-Tuning for Large Model Red-/Blue-Teaming

**Xi'an**, Northwestern Polytechnical University **2024**
Curiosity-Driven Auditing for LLMs

**Xi'an**, Northwestern Polytechnical University **2024**
Efficient Intrinsically Motivated Adversarial Policy Learning

**Gold Coast**, Griffith University **2024**
Intrinsically Motivated Adversarial Policy

**Melbourne Connect**, The University of Melbourne **2024**
Towards Efficient Evasion Attacks Against RL

## Awards & Honors

**IJCAI Travel Grant** **2024**
IJCAI Organization and AIJ Division
Amount: USD 500

**International Conference Grant** **2024**
City University of Hong Kong
Amount: HKD 10,000

**DSN Student Travel Grant** **2024**
DSN Student Travel Awards Committee
Amount: USD 1,500

**Research Activities Fund** **2022-2023**

City University of Hong Kong
Amount: HKD 96,000

**Institutional Research Tuition Grant** **2020-2024**
City University of Hong Kong
Amount: HKD 168,384

**CityU Presidential PhD Scholarship (2/73)** **2020-2024**
City University of Hong Kong
Amount: HKD 1,353,120

**NII MOU Research Activities Grant (1/106)** **2018**
National Institute of Informatics, Japan
Amount: JPY 68,400

**CSC Scholarship** **2016**
China Scholarship Council
Amount: AUD 6,400

**Stars of Advanced Engineering (2/50)** **2015**
Shen Yuan Honors College, Beihang University
The highest honor at Shen Yuan Honors College (2/50)

## Teaching & Experience

**Teaching Assistant**, City University of Hong Kong **Semester B**
CS5293, Topics on Information Security **2023/24**

**Teaching Assistant**, City University of Hong Kong **Semester A**
CS4394, Information Security and Management **2023/24**

**Teaching Assistant**, City University of Hong Kong **Semester B**
CS4293, Topics in Cybersecurity **2021/22**
CS5293, Topics on Information Security

**Teaching Assistant**, City University of Hong Kong **Semester A**
CS4394, Information Security and Management **2021/22**
CS5294, Information Security Technology Management

**Teaching Assistant**, City University of Hong Kong **Semester B**
CS4293, Topics in Cybersecurity **2020/21**
CS6290, Privacy-enhancing Technologies

**Teaching Assistant**, City University of Hong Kong **Semester A**
CS2310 & CS2311, Computer Programming **2020/21**

## Mentored Students

**Ph.D. Students**

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 Activity

**Program Committee Member**

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.