Zhihua Wang Carey

+86 15885844762 · wzh2657751462@gmail.com
Homepage: https://kirito14it.github.io/ · GitHub: https://github.com/Kirito14IT
Class of 2023

Total repository stars: approx. 2.5k | ACM team member (Silver) | National Encouragement Scholarship ×5

Education

Fudan University College of Computer Science and Artificial Intelligence — Sep. 2027–Jul. 2030
Electronic Information - **Computer Technology (Direct-entry Master's)**

Research groups: **FVL** and **TEAI** (Head: **Vice President Yu-Gang Jiang**); — Mar. 2026–Present
Advisor: **Xingjun Ma**
Research interests: LLM Coding Agent Agentic Training TrustWorthy AI
English CET-6/CET-4: 516/522

University of Shanghai for Science and Technology (Undergraduate) — Sep. 2023–Jul. 2027
School of Optical-Electrical and Computer Engineering; **Computer Science and Technology (Academician Innovation Class)**; **Rank: 3/85**

Academic Publications and Intellectual Property

[1] **Zhihua Wang**, Nan Li, ..., Yunhao Feng, Songyang Wu. *ScamWeb: A Multimodal Benchmark for Grounded Understanding of Cyber-enabled Fraud Webpages*. Preprint; AAAI 2027 (CCF-A); Will be accepted(accept*4) **[First Author]**

[2] **Zhihua Wang**, Yunhao Feng, Yifeng Gao, Rui Pan, Baiheng Zhao, Zixing Chen, Xingjun Ma. *MemHarmBench: Evaluating Cross-Session Prompt Injection in Memory-Persistent AI Agents*. Preprint; WWW 2027 (CCF-A) **[First Author]**

[3] **Zhihua Wang**, Yifeng Gao, ..., Xingjun Ma. *Smooth Inliers, Erratic Outliers: Frequency-Aware Mamba for Two-View Correspondence Learning*. Preprint; IEEE TIP (CAS Q1 Top, CCF-A) **[First Author]**

[4] **Zhihua Wang**, Junfan Chen, ..., Guoxiang Tong. *SafeCodeRL: Security-Constrained Multi-Agent Reinforcement Learning for Trustworthy LLM-Generated IoT/CPS Software*. Sensors, 26(11), 3502 Accept (CCF-C) **[First Author]**

Software copyrights: DL-Driven Robust Digital Watermarking Network Application and Code Security Analysis System Based on CodeQL and LLMs. **First author of both.**

Patent application: A related application from the **Key Laboratory of Information Network Security, Ministry of Public Security (Third Research Institute of the Ministry of Public Security)** is under review.

Research Experience

MemHarmBench: Cross-Session LLM Agent Safety Benchmark for Persistent Memory Poisoning — Jul.–Oct. 2026

上海创智学院 Shanghai Innovation Institute

Visiting student advised by **Yu-Gang Jiang** & **Xingjun Ma**, College of Computing and Intelligence, Fudan University
Recommended by Shanghai Innovation Institute | Recommended by MiraclePlus Alpha Sight | **WWW 2027** | **First Author**

- Built an "exposure–trigger" benchmark for cross-session memory poisoning with **288** case environments (**24** attack templates × **12** business scenarios) and **12** target categories.
- On samples retaining malicious rules, cross-session indirect attack success reached **59.38%–76.39%**, exposing decision manipulation despite otherwise normal execution.

Paper: https://arxiv.org/abs/2611.01700 **Code:** https://github.com/Kirito14IT/MemHarmBench **Data:** https://huggingface.co/datasets/kirco567/MemHarmBench

ScamWeb: A Multimodal Benchmark for Fraud Webpage Detection, Fine-Grained Classification and Evidence Grounding — Apr.–Aug. 2026

Research intern advised by **Xingjun Ma** & **Yu-Gang Jiang**, Institute of Trustworthy Embodied AI, Fudan University
AAAI 2027 (accept*4) | **First Author**

- Built a multimodal benchmark of **10,026** webpages spanning **25** fraud categories and **47** languages; proposed training-free ScamDet for detection, fine-grained classification and evidence grounding.
- Achieved binary-classification F1 of **78.7%**; zero-shot evaluation on a **3,324**-page fraud subset yielded **97.3%** recall and **63.9%** fine-grained classification accuracy.

Paper **Code:** https://github.com/Kirito14IT/ScamWeb **Data:** https://huggingface.co/datasets/kirco567/ScamWeb-Dataset
Project Page: https://kirito14it.github.io/ScamWeb/

XSafeClaw: Trustworthy Runtime Safety Governance for Agents ★ Stars 165 — Mar.–Jun. 2026

Featured in Synced

Institute of Trustworthy Embodied AI, Fudan University | Advisor: **Xingjun Ma** | Featured in "Latest Research" on the institute homepage | **Core Contributor**

- Provided session risk assessment, tool interception and human approval for OpenClaw, Hermes and nanobot, supporting safety protection across multiple Agent runtimes.
- Led nanobot integration, unified session identities and tool events, implemented concurrent communication and token and cost accounting across three runtimes, and strengthened safety and integration tests.

Website: https://xsafeclaw.ai **Code:** https://github.com/XSafeAI/XSafeClaw **Coverage:** https://mp.weixin.qq.com/s/1Dp5uTasxnm3ues_hNPlWQ

Smooth Inliers, Erratic Outliers: Frequency-Aware Mamba for Two-View Correspondence Learning — Sep. 2025–Feb. 2026

Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University | Advisor: Yizhang Liu | IEEE TIP (CAS Q1 Top, CCF-A) | **First Author**

- Designed local spectral-geometric attention and frequency-gated bidirectional Mamba to suppress high-frequency outlier noise with near-linear complexity ($O(N \log N)$).
- Reached mAP5° of **73.83** on unseen YFCC100M scenes (SIFT), **6.23** percentage points above MatchMamba, with **2.01M** parameters.

Paper: https://arxiv.org/abs/2606.04493 **Code:** https://github.com/Kirito14IT/SFMambaNet **Project Page:** https://kirito14it.github.io/SFMambaNet/

Internship Experience

Key Laboratory of Information Network Security, Ministry of Public Security (Third Research Institute, Shanghai) AI Research Intern — Apr. 2026–Present

Advisor: Songyang Wu **Project 1:** Independently built a joint webpage image/text and URL analysis pipeline with fraud-evidence grounding; completed **one paper as sole first author**.

Project 2: Designed a **lightweight NLP method for short-text risk topic discovery**, combining character n-grams with a pretrained multilingual MiniLM encoder. Used logistic regression for message risk scoring and conversation ranking, and semantic clustering with extractive topic naming to organize traceable group-chat evidence for local CPU screening.

International Organization (United Nations Volunteers)–UNDP Data Research Intern (Remote) — Jan.–Apr. 2026

Advisor: Johnny **Project:** Worked remotely to clean, organize, analyze and visualize data collected by the team, and produced a research report.

Competition Experience

2025 ACM Shanghai Collegiate Programming Contest	**Silver Medal**	Jul. 2025
16th Lanqiao Cup National Software and Information Technology Professional Talent Competition, National Finals, C/C++ (2025)	**National Second Prize**	Jun. 2025
2025 RAICOM Robot Developer Competition, Programming Track, National Finals	**National First Prize**	Aug. 2025

Skills and Service

Programming and tools: C/C++, Python; PyTorch/TensorFlow, vllm; Transformers, Codex, Claude Code;
NeurIPS 2026 Reviewer