# Xin Wang

Fudan University, 220 Handan Rd, Yangpu District, Shanghai
136 2728 3132 • hsinwang22@gmail.com • xinwong.github.io

## Education

2022–Pres. **PhD**, Fudan University, CN
Advisor: Prof. Xingjun Ma and Prof. Yu-Gang Jiang

2018–2021 **MPhil**, Central China Normal University, CN
Advisor: Prof. Qiusha Min

2016 **Exchange Student**, Providence University, TW

2014-2018 **BEng**, Nanyang Institute of Technology, CN

## Research Interests

I am broadly interested in safety and privacy aspects of machine learning with a recent focus on large language models. Most of my past works are in the domain of trustworthy machine learning, particularly adversarial examples and robustness of machine learning algorithms.

## Work Experience

2025 **Research Intern**, *Shanghai AI Lab*, (*Part-time*), Mentor: Jie Li
(1) *Jailbreak Toolbox* (Project Leader): Led the development of an adaptive, evolutionary jailbreak toolbox, scaling it to 25 jailbreak methods with support for text and image-text inputs across single and multi-turns.
(2) *SafeWork-R1* (Core Contributor): Enhanced the trustworthiness of LLMs by developing multiple reward models focused on safety, ethical values, and knowledge. To address the scarcity of trustworthy data, generated over 150k high-quality training samples. Successfully trained a 70B Outcome Reward Model (ORM) and an implicit Preference Reward Model (PRM), dramatically boosting model safety and trustworthiness.

2025 **Research Intern**, *Sea AI Lab*, (*Part-time*), Mentor: Tianyu Pang and Chao Du
Proposed an Imperceptible Jailbreak / Reasoning Energy Attack that uses Unicode Variation Selectors (invisible characters) to append an adversarial suffix, leaving the prompt visually unchanged yet altering its tokenization to bypass safety alignment and evade human review.

2021 **AI Algorithm Engineer**, *iFLYTEK*, (*Full-time*), Mentor: Jianbo Chen
Our team proposed the Baopu platform, as the independently R&D federated learning platform, aiming to provide a secure computing framework to support the federated learning ecosystem. Moreover, our team, YYDS, is among the best-performing teams that participate in the iDash Privacy & Security Workshop.

## Publications (Google Scholar)

[1] **TAPT: Test-Time Adversarial Prompt Tuning for Robust Inference in Vision-Language Models**
*X. Wang*, K. Chen, J. Zhang, *et al.* CVPR 2025

[2] **AdvQDet: Detecting Query-Based Adversarial Attacks with Adversarial Contrastive Prompt Tuning**
*X. Wang*, K. Chen, X. Ma, *et al.* ACM MM 2024

[3] **FreezeVLA: Action-Freezing Attacks against Vision-Language-Action Models**
*X. Wang*, J. Li, Z. Weng, *et al.* arXiv 2025

[4] **SafeWork-R1: Coevolving Safety and Intelligence under the AI-45° Law**
*Shanghai AI Lab* (Core Contributor) Technical Report 2025

[5] **Adversarial Prompt Tuning for Vision-Language Models**
J. Zhang, X. Ma, *X. Wang*, *et al.* ECCV 2024

[6] **SafeVid: Toward Safety-Aligned Video Large Multimodal Models**
Y. Wang, J. Song, Y. Gao, *X. Wang*, *et al.* NeurIPS 2025 D&B Track

[7] **Safety at Scale: A Comprehensive Survey of Large Model Safety**
X. Ma, Y. Gao, Y. Wang, R. Wang, *X. Wang*, *et al.* FnTs® in Privacy and Security 2025

[8] **Argus Inspection: Do Multimodal Large Language Models Possess the Eye of Panoptes?**
Y. Yao, L. Li, J. Song, C. Chen, Z. He, Y. Wang, *X. Wang*, *et al.* ACM MM 2025 Datasets Track

[9] **Lossless Medical Image Compression Based on Anatomical Information and Deep Neural Networks**
Q. Min, *X. Wang*, B. Huang, *et al.* Biomedical Signal Processing and Control 2022

[10] **Web-Based Technology for Remote Viewing of Radiological Images: App Validation**
Q. Min*, *X. Wang**, B. Huang, *et al.* Journal of Medical Internet Research 2020

[11] **Evolve the Method, Not the Prompts: Evolutionary Synthesis of Jailbreak Attacks on LLMs**
Y. Chen*, *X. Wang**, J. Li, *et al.* arXiv 2025

[12] **Simulated Ensemble Attack: Transferable Jailbreaks Across Fine-tuned Vision-Language Models**
R. Wang, *X. Wang*, Y. Yao, *et al.* arXiv 2025

[13] **NAP-Tuning: Neural Augmented Prompt Tuning for Adversarially Robust Vision-Language Models**
J. Zhang, *X. Wang*, X. Ma, *et al.* arXiv 2025

[14] **SafeEvalAgent: Toward Agentic and Self-Evolving Safety Evaluation of LLMs**
Y. Wang, *X. Wang*, X. Ma, *et al.* arXiv 2025

[15] **DarkLLaVA: Scalable Adversarial Attack with Large Language Models**
Y. Sun, *X. Wang*, J. Zhang, *et al.* arXiv 2025

[16] **Adversarial Prompt Distillation for Vision-Language Models**
L. Luo, *X. Wang*, B. Zi, *et al.* arXiv 2024

[17] **$A^2$RM: Adversarial-Augmented Reward Model**
S. Huang, J. Li, *X. Wang*, *et al.* arXiv 2025

[18] **LeakyCLIP: Extracting Training Data from CLIP**
Y. Chen, S. Wang, *X. Wang*, *et al.* arXiv 2025

[19] **Imperceptible Jailbreaking against Large Language Models**
K. Gao, Y. Li, C. Du, *X. Wang*, *et al.* arXiv 2025

[20] **DAVID-XR1: Detecting AI-Generated Videos with Explainable Reasoning**
Y. Gao, Y. Ding, H. Su, J. Li, Y. Zhao, L. Luo, Z. Chen, L. Wang, *X. Wang*, *et al.* arXiv 2025

[21] **Trustworthy Embodied AI: A Comprehensive Survey**
X. Li, X. Zheng, Y. Gao, X. Xia, Y. Wang, R. Wang, *X. Wang*, *et al.* arXiv 2025

## Awards & Honors

2025 First-class Scholarship, Fudan University
2023 Outstanding Student, Fudan University
2021 iDash Privacy & Security Challenge Track 3: Confidential Computing, Ranked 1st
2021 Outstanding Graduate Award, Central China Normal University
2021 Outstanding Master's Thesis Award, Central China Normal University

## Academic Service

Conference Reviewer for ICLR, NeurIPS, ICCV, EMNLP, ACM MM, *et al.*

Journal Reviewer for IJCV, TIP, TCSVT, TNNLS, *et al.*