Jie Zhang

Curriculum Vitae

*CFAR and IHPC, A*STAR*
Singapore
☎ *(+65) 87102696*
✉ *zjzacxt@gmail.com*
🖱 *Personal Homepage*

Summary

Research Scientist & Innovation Lead specializing in multimodal generative AI, cultural adaptation in image/video synthesis, and trustworthy AI. Over 40 publications in top AI venues (TPAMI, NeurIPS, ICML, ICLR, CVPR, S&P, USENIX Security, CCS, NDSS). Proven leader of high-performing teams, experienced in large-scale training of diffusion/autoregressive models, and adept at aligning cutting-edge AI research with governance and real-world impact.

Work Experience

08/2024 – **Research Scientist & Innovation Lead**, *A*STAR Centre for Frontier AI Research (CFAR)*, Singapore, work with Dr. Qing Guo and Prof. Ivor Tsang.

Led technical strategy for generative AI safety projects, including cultural adaptation in multimodal synthesis; managed a team of researchers and students; coordinated collaborations with industry (Adobe, NIO) and academia; oversaw large-scale training and evaluation of diffusion-based image/video models.

03/2023 – 07/2024 **Research Fellow**, *Nanyang Technological University*, Singapore, work with Prof. Tianwei Zhang and Prof. Yang Liu.

07/2022 – 02/2023 **Postdoc**, *University of Waterloo*, Canada, remote work with Prof. Florian Kerschbaum.

Education

09/2017 – 06/2022 **PhD of Cyber Science and Technology**, *University of Science and Technology of China*, Hefei, China.

Research Interests

- Trustworthy AI
- Fairness, robustness & watermarking
- Multimodal generative AI (image, video, audio, etc.)
- Agentic AI
- Cultural adaptation in synthesis

Selected Publications (** Equal Contribution | †Corresponding Author*)

⋆ **Jie Zhang**, et al. Robust Model Watermarking for Image Processing Networks via Structure Consistency, *TPAMI, 2024.*

- ⋆ **Jie Zhang**,et al. "Digital Watermarking for Machine Learning Models - Chapter 6: Protecting Image Processing Networks via Model Watermarking", *Springer book, 2023.*
- ⋆ **Jie Zhang**, et al. Poison Ink: Robust and Invisible Backdoor Attack, *IEEE Transactions on Image Processing (TIP), 2022.*
- ⋆ **Jie Zhang**, et al. Deep Model Intellectual Property Protection via Deep Watermarking, *TPAMI, 2021.*
- ⋆ **Jie Zhang**, et al. Passport-aware Normalization for Deep Model Protection, *Advances in Neural Information Processing Systems (NeurIPS), 2020.*
- ⋆ **Jie Zhang**, et al. Model Watermarking for Image Processing Networks, *AAAI Conference on Artificial Intelligence (AAAI), 2020.*
- ⋆ Qidong Huang*, **Jie Zhang***, Wenbo Zhou, Weiming Zhang, Nenghai Yu, Initiative Defense against Facial Manipulation, *AAAI Conference on Artificial Intelligence (AAAI), 2021*.
- ⋆ Yutong Wu, **Jie Zhang**†, et al. Cowpox: Towards the Immunity of VLM-based Multi-Agent Systems, *ICML 25*.
- ⋆ Yutong Wu, **Jie Zhang**†, et al. THEMIS: Regulating Textual Inversion for Personalized Concept Censorship, *the Network and Distributed System Security Symposium (NDSS), 2025*.
- ⋆ Peigui Qi, **Jie Zhang**†, et al. SafeGuider: Robust and Practical Content Safety Control for Text-to-Image Models, *ACM CCS 2025*.
- ⋆ Runyi Hu, **Jie Zhang**†, et al. VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking, *ICLR 2025*.
- ⋆ Boheng Li, **Jie Zhang**†, et al. Towards Reliable Verification of Unauthorized Data Usage in Personalized Text-to-Image Diffusion Models, *IEEE Symposium on Security and Privacy (S&P) 2025*.
- ⋆ Kunsheng Tang, **Jie Zhang**†, et al. GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models, *The ACM Conference on Computer and Communications Security (CCS) 2024*.
- ⋆ Weitao Feng, **Jie Zhang**†, et al. AquaLoRA: Toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA, *ICML 24*
- ⋆ Chang Liu, **Jie Zhang**†, et al. Detecting Voice Cloning Attacks via Timbre Watermarking, *the Network and Distributed System Security Symposium (NDSS), 2024.*

Selected Preprints

- ⋆ Visible Yet Unreadable: A Systematic Blind Spot of Vision–Language Models Across Writing Systems
- ⋆ A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment
- ⋆ Reinforcement learning enhanced llms: A survey

Awards & Honors

2025 Candidates of best paper (One of the Top 15 Papers), ICME, 2025

2024 Distinguished Artifact Award, CCS, 2024

2021 National Scholarship for Doctoral Students, China

2020 Cyberspace Science Scholarship (funded by Academician Xiaomo Wang), China

Selected Grants

2024–2027 **DTC**, *Combatting Prejudice in AI: A Responsible AI Framework for Continual Fairness Testing, Repair, and Transfer*, Co-PI.

2024–2027 **CRPO**, *Secure, Private, and Verified Data Sharing for Large Model Training and Deployment*, Technique Lead.

2023–2026 **AISG Grand Challenge**, *Towards Building Unified AV Scene Representation for Physical AV Adversarial Attacks and Visual Robustness Enhancement*, Co-PI.

2022–2025 **MoE AcRF Tier2**, *A Framework for Intellectual Property Protection of Deep Learning Applications*, Technique Lead.

2021–2024 **NRF, China**, *Research on Basic Theory and Key Technology of Attack and Defense Analysis for Deep Models*, Technique Lead.

Collaborations

Collaborated with MBZUAI, Microsoft Research, Adobe, University of Waterloo, City University of Hong Kong, Wormpex AI Research

Mentorship

Supervised 10+ PhD&Master students across Singapore, Canada, and China, leading to publications in ICLR, S&P, CCS, ICML

Technical Skills

Languages Python, PyTorch

Specialties Diffusion models, autoregressive models, multimodal alignment, cultural adaptation, fairness-aware training, model watermarking

Tools LoRA fine-tuning, mixed precision, visual aesthetics evaluation

Services

- Reviewer for ICML, ICLR, NeurIPS, AAAI, IJCAI, CVPR, ICCV, ECCV, ACL, NAACL, EMNLP, ACM CCS, NDSS, ACM MM, etc.
- Reviewer for TPAMI, IJCV, TIP, TIFS, TDSC, TMM, TCSVT, SPL, etc.
- Organizer for the 4th Workshop on Practical Deep Learning (Practical-DL 2025).

Interests

Sports, Hiking, Traveling, Scuba

Publications Overview (Google Scholar)

Trustworthy AI & GenAI

Vulnerability Evaluation [TIP 2022], [AAAI 2023], [MM 2023], [AAAI 2024], [AAAI 2024], [AAAI 2024], [CCS 2024], [NeurIPS 2024], [Information Fusion 2024], [USENIX Security 2025], [NAACL 2025], [USENIX Security 2025], [TMM 2025], [CVPR 2025], [S&P 2025], [CCS 2025], [USENIX Security 2025]

Proactive Safeguard [AAAI 2021], [MM 2023], [IJCAI 2024], [ICML 2024], [MM 2024], [NDSS 2025], [AAAI 2025], [ICASSP 2025], [TDSC 2025], [TOSEM 2025], [ICML 2025], [ICML 2025], [CCS 2025]

Post-hoc Forensic [AAAI 2020], [NeurIPS 2020], [MM 2020], [TPAMI 2021], [AAAI 2022], [TAI 2023], [Springer Book], [AAAI 2023], [AAAI 2023], [TKDE 2023], [TPAMI 2024], [NDSS 2024], [ICML 2024], [ECCV 2024], [S&P 2025], [TIFS 2025], [ICLR 2025], [ICME 2025], [ICME 2025], [TDSC 2025], [TPAMI 2025]

Others

Affective Computing [MM 2024], [CVPR 2025], [MM 2025]

AI for Science [AI4X 2025], [AI for Science 2025]