Zongmin Zhang

Mobile: +86-17740192857
Email: zzhang498@connect.hkust-gz.edu.cn

Research Interests

- Trustworthy AI, including AI security and safety

Education

- **Hong Kong University of Science and Technology (Guangzhou)** — Guangzhou, China
  Ph.D. in Data Science and Analytics; Supervisor: Prof. Xinlei He — *Aug. 2025 – Present*
- **City University of Hong Kong** — Hong Kong, China
  MSc in Computer Science; GPA: 3.19/4.3 — *Aug. 2022 – Oct. 2023*
- **The University of New South Wales** — Sydney, Australia
  BSc in Computer Science – Database Systems; Credit Award — *Aug. 2018 – Sep. 2021*

Research Experience

- **Ph.D. Student, Data Science and Analytics** — Guangzhou, China
  Hong Kong University of Science and Technology (Guangzhou), Advisor: Prof. Xinlei He — *Aug. 2025 – Present*
  - **Adversarial Attacks on Image Watermarking**: Work on adversarial attacks against image watermarking by adding noise to watermarked images, causing the watermark fail to detect while preserving visual quality.
  - **Backdoor Attack on SAM2**: Work on backdoor attacks against the video segmentation model Segment Anything 2 (SAM2) model by injecting triggers into the training data, causing SAM2 to either fail at precisely segmenting the correct objects or to segment incorrect ones.
  - **Multimodal LLM Jailbreaking**: Involve developing novel jailbreaking strategies for multimodal large language models via optimizing prompt representations, analyzing model vulnerabilities, and safety alignment gaps.
  - **AI-Generated Content on Social Media**: Build data collection and analysis pipelines to quantify the prevalence and impact of AI-generated text on social media, including large-scale measurement, labeling, and detection model evaluation.

Publications

** indicates equal contribution.*

- **Zongmin Zhang***, Zhen Sun*, Yifan Liao, Wenhan Dong, Xinlei He, Xingshuo Han, Shengmin Xu, Xinyi Huang, **"Backdoor Attacks on Prompt-Driven Video Segmentation Foundation Models"**, arxiv preprint, 2025.12.
- Zhen Sun*, **Zongmin Zhang***, Deqi Liang, Han Sun, Yule Liu, Yun Shen, Xiangshan Gao, Yilong Yang, Shuai Liu, Yutao Yue, Xinlei He, **"To Survive, I Must Defec: Jailbreaking LLMs via the Game-Theory Scenarios"**, arxiv preprint, 2025.11.
- Jihui Guo, **Zongmin Zhang**, Zhen Sun, Yuhao Yang, Jinlin Wu, Fu Zhang, Xinlei He, **"6DAttack: Backdoor Attacks in the 6DoF Pose Estimation"**, **AAAI 2026**, Oral.
- Jingyi Zheng, Tianyi Hu, Yule Liu, Zhen Sun, **Zongmin Zhang**, Zifan Peng, Wenhan Dong, Xinlei He, **"CHASM: Unveiling Covert Advertisements on Chinese Social Media"**, **NeurIPS 2025**.
- Ziyi Zhang, Zhen Sun, **Zongmin Zhang**, Jihui Guo, Xinlei He, **"FC-Attack: Jailbreaking Multimodal Large Language Models via Auto-Generated Flowcharts"**, **EMNLP Findings 2025**.
- Zhen Sun*, **Zongmin Zhang***, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, Xinlei He, **"Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media"**, **ACL Main 2025**.
- Han Sun*, Zhen Sun*, **Zongmin Zhang***, Linzhao Jia, Min Zhang, **"SynDec: A Synthesize-then-Decode Approach for Arbitrary Textual Style Transfer via Large Language Models"**, arXiv preprint, 2025.
- **Zongmin Zhang**, Yujie Han, Zhou Zhang, Yule Liu, Jingyi Zheng, Zhen Sun, **"AdSpectorX: A Multimodal Expert Spector for Covert Advertising Detection on Chinese Social Media"**, in *Proceedings of the Third International Workshop on Social and Metaverse Computing, Sensing and Networking* ***SocialMeta 2024, Best Paper Award***.

- Zhen Chen*, **Zongmin Zhang***, Wenwu Guo, Xingjian Luo, Long Bai, Jinlin Wu, Hongliang Ren, Hongbin Liu, **"ASI-Seg: Audio-Driven Surgical Instrument Segmentation with Surgeon Intention Understanding"**, **IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2024**, Oral.

- Xingjian Luo, You Pang, Zhen Chen, Jinlin Wu, **Zongmin Zhang**, Zhen Lei, Hongbin Liu, **"SurgPLAN: Surgical Phase Localization Network"**, **IEEE International Symposium on Biomedical Imaging (ISBI) 2024**.

Work Experience

- **City University of Hong Kong** — Hong Kong, China
  Research Assistant (Advisor: Prof. Antoni B. Chan) — *Jul. 2024 – Jun. 2025*
  - **GPT-based Plagiarism Detection Algorithm**: Involving the development of a plagiarism detection algorithm leveraging GPT-style similarity scores for document- and paragraph-level plagiarism detection in academic texts.
  - **User-facing Tools & Analysis**: Built instructor-facing interfaces and dashboards to visualize similarity scores and plagiarism risk, and carried out experiments to analyze performance under different prompts and thresholds.

- **Centre for Artificial Intelligence and Robotics (CAIR), HKISI, CAS** — Hong Kong, China
  Research Assistant (Advisor: Dr.Jinlin Wu) — *Dec. 2023 – Jun. 2024*
  - **Endoscopic Lesion Segmentation Model**: Developed a medical image segmentation pipeline based on nnU-Net to segment abnormal regions in endoscopic images, and integrated the model with a reporting module that automatically generates draft medical reports from the segmentation results.
  - **Clinical-facing Web Interface**: Built and maintained a web application for clinicians to interact with the medical LLM, with a React frontend and Flask backend, and iterated on new features based on user feedback.

- **City University of Hong Kong** — Hong Kong, China
  Research Assistant (Advisor: Prof. Antoni B. Chan) — *Sep. 2022 – Jun. 2023*
  - **Online Exam Plagiarism Checking System**: Developed a web-based system to check plagiarism among students' online exam submissions, implementing both the backend services in Python (Flask) and the frontend interface in React.

- **Beijing Sankuai Online Technology Co., Ltd. (Meituan)** — Chengdu, China
  System Development Intern — *Feb. 2021 – Aug. 2022*
  - **Financial Reporting System**: Involving the development and maintenance of a Java Spring Boot–based financial reporting system used by business developers to monitor and analyze key operational metrics.
  - **Monitoring Error Cases**: Designed a data monitoring module and reported any error cases that happened on financial reporting functions.