Last Updated on August 21, 2025

Yichi Zhang

Ph.D Candidate @ Tsinghua University

✉ zyc22@mails.tsinghua.edu.cn, tibo_ricky@outlook.com 🔗 zycheiheihei.github.io
📍 FIT Building 1-508, Tsinghua University, Beijing, China, 100084

Education

Tsinghua University, Department of Computer Science and Technology Aug. 2022 – Present
Beijing, China
Ph.D candidate advised by Prof. Jun Zhu
Currently researching into **Trustworthy AI and Multimodal Learning**

Tsinghua University, Department of Computer Science and Technology Aug. 2018 – July 2022
Beijing, China
Bachelor of Engineering, GPA: 3.90/4.00, Ranking: 7/235
Secondary Bachelor of Science in Psychology
NCEE: 709/750, **8th top scorer** of science in Beijing (∼35k students)

Publications

($\ast$ indicates equal contribution)

Published in Conferences

STAIR: Improving Safety Alignment with Introspective Reasoning (***Oral, ∼top 0.9%***)
Yichi Zhang*, Siyuan Zhang*, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, Jun Zhu
International Conference on Machine Learning (ICML), 2025

RealSafe-R1: Safety-Aligned DeepSeek-R1 without Compromising Reasoning Capability
Yichi Zhang, Zihao Zeng, Dongbai Li, Yao Huang, Zhijie Deng, Yinpeng Dong
R2-FM Workshop at International Conference on Machine Learning (ICML), 2025

MultiTrust: A Comprehensive Benchmark Towards Trustworthy Multimodal Large Language Models
Yichi Zhang*, Yao Huang*, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, Jun Zhu
Advances in Neural Information Processing Systems (NeurIPS), 2024

Exploring the Transferability of Visual Prompting for Multimodal Large Language Models (***Highlight, ∼top 2.8%***)
Yichi Zhang, Yinpeng Dong, Siyuan Zhang, Tianzan Min, Hang Su, Jun Zhu
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

Understanding the Robustness of 3D Object Detection With Bird's-Eye-View Representations in Autonomous Driving
Zijian Zhu*, **Yichi Zhang***, Hai Chen, Yinpeng Dong, Shu Zhao, Wenbo Ding, Jiachen Zhong, Shibao Zheng
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023

Exploring the Generalizability of Factual Hallucination Mitigation via Enhancing Precise Knowledge Utilization
Siyuan Zhang, **Yichi Zhang**, Yinpeng Dong, Hang Su
The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), Findings, 2025

Breaking the Ceiling: Exploring the Potential of Jailbreak Attacks through Expanding Strategy Space
Yao Huang, Yitong Sun, Shouwei Ruan, **Yichi Zhang**, Yinpeng Dong, Xingxing Wei
Annual Meeting of the Association for Computational Linguistics (ACL), Findings, 2025

PINNacle: A Comprehensive Benchmark of Physics-Informed Neural Networks for Solving PDEs
Zhongkai Hao, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, **Yichi Zhang**, Songming Liu, Lu Lu, Jun Zhu
Advances in Neural Information Processing Systems (NeurIPS), 2024

Rethinking Model Ensemble in Transfer-based Adversarial Attacks
Huanran Chen, **Yichi Zhang**, Yinpeng Dong, Jun Zhu
International Conference on Learning Representations (ICLR), 2024

Published in Journals

To make yourself invisible with Adversarial Semantic Contours
Yichi Zhang, Zijian Zhu, Hang Su, Jun Zhu, Shibao Zheng, Yuan He, Hui Xue
Computer Vision and Image Understanding (CVIU), 2023

Released as Preprints

Unveiling Trust in Multimodal Large Language Models: Evaluation, Analysis, and Mitigation
Yichi Zhang, Yao Huang, Yifan Wang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, Jun Zhu
ArXiv (Under Review for TPAMI), 2025

Physics-informed machine learning: A survey on problems, methods and applications
Zhongkai Hao, Songming Liu, **Yichi Zhang**, Chengyang Ying, Yao Feng, Hang Su, Jun Zhu
ArXiv, 2022

Competitions

The **1st place** in the Adversarial Robustness track of 2022 International Algorithm Case Competition — Feb. 2023

The **2nd place** in the CVPR 2021 Security AI Challenger Unrestricted Adversarial Attacks on ImageNet — June 2021

The **8th place** in the CIKM 2020 Adversarial Challenge on Object Detection — Sept. 2020

Experience

Qwen Team, Alibaba | *Research Intern* — May. 2025 – Aug. 2025
Agentic System for DeepResearch

University of Sydney | *PhD Visiting Student* — March. 2025 – May. 2025
Trustworthy Machine Learning

RealAI | *Research Intern* — Oct. 2022 – Dec. 2024
Safety and robustness of deep learning models in wide applications

Tencent | *Research Intern* — July 2021 – Sept. 2021
Advertising models and re-ranking models in the recommendation system of Tencent Video Platform

Selected Awards

Tsinghua Outstanding Graduates (**top 2%**) — June 2022

Beijing Outstanding Graduates (**top 5%**) — June 2022

Beijing Merit Students (**Only one student in the department each year**) — Mar. 2022

Tsinghua Overall Excellence Scholarships — Dec. 2019,2020,2021

Service

Organizer

ICML2024 Workshop on Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)
CVPR2025 Workshop on Test-time Scaling for Computer Vision (ViSCALE)

Reviewer

ICML2024, ICLR2025, ICML2025, ACL2025, TPAMI

Teaching

TA in Machine Learning, instructed by Prof. Jun Zhu and Prof. Jie Tang, 2023 Autumn

Skills

Language: TOEFL 115/120, CET4
Programming: Python, C/C++, Java, PyTorch, LaTeX
Hobbies: Basketball, Chorus