


Yichi Zhang

Ph.D Candidate @ Tsinghua University

✉ zyc22@mails.tsinghua.edu.cn, tibo_ricky@outlook.com  [zycheiheiei.github.io](https://github.com/zyc22)
📍 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

(* indicates equal contribution)

PUBLISHED IN CONFERENCES

STAIR: Improving Safety Alignment with Introspective Reasoning (***Spotlight**, ~top 2.6%*)

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

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

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

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

Rethinking Model Ensemble in Transfer-based Adversarial Attacks

Huanran Chen, **Yichi Zhang**, Yinpeng Dong, Jun Zhu

International Conference on Learning Representations (ICLR), 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

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

STAIR: Improving Safety Alignment with Introspective Reasoning

Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, Jun Zhu

ArXiv, 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

WORK EXPERIENCE

RealAI | *Research Intern*

Oct. 2022 – Dec. 2023

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