Jianing Zhu

University of Texas at Austin, Austin, Texas
⌂ Homepage 🎓 Google Scholar ✉ zhujianing9810@gmail.com ☎ +1 412-475-5044

RESEARCH INTERESTS

My research interests lie in **trustworthy machine learning** for building human-aligned machine intelligence, particularly in developing methodologies that improve its `robustness` (e.g., for adversarial examples), `reliability` (e.g., for out-of-distribution data), and `transparency` (e.g., for functionality and traceability), as well as its applications to **construct powerful & responsible AI** for **benefiting social goods**.

PROFESSIONAL EXPERIENCES

Postdoctoral Fellow, UT Austin ECE — **Sept. 2025 – Present**
VITA Group, Advisor: Prof. Zhangyang "Atlas" Wang

Visiting PhD Student, CMU MLD — **Jan. 2025 – June. 2025**
Neuro-Symbolic AI Group, Advisor: Prof. Pradeep Ravikumar

Research Intern, RIKEN AIP — **Dec. 2023 – May. 2024**
Imperfect Info rmation Learning Team, Advisor: Prof. Masashi Sugiyama

EDUCATION

Hong Kong Baptist University (HKBU) — **Sept. 2021 – Jun. 2025**
Ph.D. of TMLR Group, Department of Computer Science — Advisor: Prof. Bo Han

Sichuan University (SCU) — **Sept. 2017 – Jun. 2021**
B.Eng. in CS, College of Computer Science — National Top-Notch UG Program

SELECTED PUBLICATIONS

The full list can refer to [Google Scholar]. As of 09/2025, his works have been cited over 904 times, with h-index = 11, below are his selected publications (* indicates the equal contribution):

ICLR 2025 [link]: Qizhou Wang, Bo Han, Puning Yang, **Jianing Zhu**, Tongliang Liu, Masashi Sugiyama, "Unlearning with Control: Assessing Real-world Utility for Large Language Model Unlearning".
NeurIPS 2024 [link]: Boxuan Zhang*, **Jianing Zhu***, Tongliang Liu, Masashi Sugiyama, "What If the Input is Expanded in OOD Detection?".
NeurIPS 2024 [link]: Zhanke Zhou, Rong Tao, **Jianing Zhu**, Yiwen Luo, Zengmao Wang, Bo Han, "Can Large Language Models Reason Robustly with Noisy Rationales?".
NeurIPS 2024 [link]: Geng Yu, **Jianing Zhu**, Jiangchao Yao, Bo Han, "Self-Calibrated Tuning of Vision-Language Models for Out-of-Distribution Detection".
NeurIPS 2023 [link]: **Jianing Zhu**, Geng Yu, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, Bo Han, "Diversified Outlier Exposure for Out-of-Distribution Detection via Informative Extrapolation".
ICML 2023 [link]: **Jianing Zhu**, Hengzhuang Li, Jiangchao Yao, Tongliang Liu, Jianliang Xu, Bo Han, "Unleashing Mask: Explore the Intrinsic Out-of-Distribution Detection Capability".
ICML 2023 [link]: **Jianing Zhu**, Xiawei Guo, Jiangchao Yao, Chao Du, Li He, Shuo Yuan, Tongliang Liu, Liang Wang, Bo Han, "Exploring Model Dynamics for Accumulative Poisoning Discovery".
ICLR 2023 [link]: **Jianing Zhu**, Jiangchao Yao, Tongliang Liu, Quanming Yao, Jianliang Xu, Bo Han, "Combating Exacerbated Heterogeneity for Robust Models in Federated Learning".
NeurIPS 2022 (Spotlight)[link]: Jianan Zhou*, **Jianing Zhu***, Jingfeng Zhang, Tongliang Liu, Gang Niu, Bo Han, Masashi Sugiyama, "Adversarial Training with Complementary Labels: On the Benefit of Gradually Informative Attacks".
ICLR 2022 [link]: **Jianing Zhu**, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, Hongxia Yang, "Reliable Adversarial Distillation with Unreliable Teachers".
ICLR 2021 (Oral) [link]: Jingfeng. Zhang, **Jianing Zhu**, Gang Niu, Bo Han, Masashi Sugiyama, Mohan Kankanhalli, "Geometry-aware Instance-reweighted Adversarial Training".

PREPRINTS

ICML 2025 Workshop [link]: Jianing Zhu, Zongze Li, Chandler Squires, Qizhou Wang, Bo Han, Pradeep Ravikumar, "On the Fragility of Latent Knowledge: Layer-wise Influence under Unlearning in Large Language Model".
ArXiv 2024 [link]: Jianing Zhu, Bo Han, Jiangchao Yao, Jianliang Xu, Gang Niu, Masashi Sugiyama, "Decoupling the Class Label and the Target Concept in Machine Unlearning".
NeurIPS 2023 Workshop [link]: Xuan Li*, Zhanke Zhou*, **Jianing Zhu***, Jiangchao Yao, Tongliang Liu, Bo Han, "DeepInception: Hypnotize Large Language Model to Be Jailbreaker".
ArXiv 2021 [link]: Jianing Zhu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, Mohan Kankanhalli, Masashi Sugiyama, "Understanding the Interaction of Adversarial Training with Noisy Labels".

HONORS AND AWARDS

- NeurIPS Top Reviewers (Top 8%), NeurIPS 2023
- ICML Best Reviewers (Top 10%), ICML 2021, 2024
- ICLR Notebale Reviewers, ICLR 2025
- Yakun Scholarship Scheme for Mainland Postgraduate Students, HKBU 2024
- Computer Science Department Research Excellence Award, HKBU 2023
- Computer Science Department RPg Performance Award, HKBU 2022-2024
- Excellent Teaching Assistant Performance Award, HKBU 2022-2023
- Nomination of Hong Kong PhD Fellowship Scheme, HKBU 2021
- University Scholarship, SCU 2018-2020
- National Scholarship, Ministry of Education 2018

INVITED TALKS

- How AI Leaks Information — and How It Can Forget @ UT Austin Digital Trust Symposium Nov. 2025
- Towards Trustworthy Machine Learning for Out-of-distribution Data @ SCU COMP Dec. 2023
- Diversified Outlier Exposure for Out-of-distribution Detection @ HKBU Dec. 2023
- Youth PhD Talk for Conference Work Sharing @ AI Time Jun./Nov. 2023

SERVICES

Program Committee & Reviewer:
ICML 2021-2025, NeurIPS 2021-2025, ICLR 2022-2025, ACML 2021-2025, AAAI 2023-2025, IJCAI 2022-2025, AISTATS 2023-2025, TPAMI, JAIR, TMLR, ACM CSUR, TNNLS, MLJ, NN, and so on.

Organization Committee:
Leading Organizer, TMLR Young Scientist Seminar 2023-Present
Workshop Chair Assistant, NeurIPS 2024 2024
- Workshop Proposal Reviewer, NeurIPS 2025
Workshop Assistant, HKBU-COMP & RIKEN-AIP Joint Workshop 2024
Founder Member of Executive Group, RIKEN TrustML Young Scientist Seminars 2022-2023

MENTORING AND TEACHING

Mentoring following students:
Yuanyi Li (CMU Master) → TBD 2024-2025
Zongze Li (HUST Undergrad) → TBD 2024-2025
Jingwei Sun (XJTU Master) → HKBU PhD 2024-2025
Boxuan Zhang (WHU Master) → Rutgers PhD 2024-2025
Xuan Li (UofSouthampton Master) → HKBU PhD 2023-2024
Geng Yu (SJTU Master) → TBD 2023-2024
Hengzhuang Li (HUST Undergrad) → HUST Master 2022-2023

Teaching assistant on following courses @ HKBU:
COMP7240(PG): Recommender Systems, Autumn 2022-2023
COMP7160(PG): Research Methods in Computer Science, Autumn 2022-2023
COMP7250(PG): Machine Learning, Spring 2022
COMP4135(UG): Recommender Systems and Applications, Autumn 2022-2023
COMP3057(UG): Intro to AI and Machine Learning, Autumn 2022