ZIWEI XU

B1-08, COM3, 11 Research Link, Singapore 119391
Email: ziwei.xu@u.nus.edu ◇ Homepage: https://home.ziwei-xu.com/

EXPERIENCE

National University of Singapore *December 2022 - Present*
Research Fellow Supervisor: Prof. Mohan Kankanhalli

National University of Singapore *July 2017 - July 2023*
Doctor of Philosophy in Computer Science Advisor: Prof. Mohan Kankanhalli
Thesis: *Solving Computer Vision Problems under the Compositionality Principle*

Center for Research in Computer Vision, University of Central Florida *September 2021 - February 2022*
Visiting Scholar Advisor: Dr. Yogesh S. Rawat, Prof. Mubarak Shah

University of Science and Technology of China *August 2013 - July 2017*
Bachelor of Engineering (Honours) in Electronic Information Engineering

RESEARCH INTERESTS

Knowledge-Enhanced Machine Learning

- Integration of symbolic knowledge and formal methods with data-driven machine learning models

Safety and Trustworthiness of AI Systems

- Hallucination, value alignment, and adversarial attacks on large language models
- Soundness and robustness of reasoning models

SELECTED WORKS

Conference Papers

- **Z. Xu** and M. Kankanhalli, *Strong Preferences Affect the Robustness of Preference Models and Value Alignment*, International Conference on Learning Representations (ICLR), 2025.
- **Z. Xu**, Y. S. Rawat, Y. Wong, M. Kankanhalli, and M. Shah, *Don't Pour Cereal into Coffee: Differentiable Temporal Logic for Temporal Action Segmentation*, Neural Information Processing Systems (NeurIPS), 2022.
- Y. Wong, S. Fan, Y. Guo, **Z. Xu**, K. Stephen, R. Sheoran, A. Bhamidipati, V. Barsopia, J. Liu, and M. Kankanhalli, *Compute to Tell the Tale: Goal-Driven Narrative Generation*, ACM International Conference on Multimedia (ACM MM), 2022.
- **Z. Xu**, X. Shen, Y. Wong, and M. Kankanhalli, *Unsupervised Motion Representation Learning with Capsule Autoencoders*, Neural Information Processing Systems (NeurIPS), 2021.
- Y. Xie, **Z. Xu**, K. Meel, M. Kankanhalli, and H. Soh, *Embedding Symbolic Knowledge into Deep Networks*, Neural Information Processing Systems (NeurIPS), 2019.

Journal Papers

- G. Wang, Y. Guo, **Z. Xu**, Y. Wong, M. S Kankanhalli, *Semantic-Aware Triplet Loss for Image Classification*, IEEE Transactions on Multimedia, Volume 25, 2022.
- **Z. Xu**, G. Wang, Y. Wong, and M. Kankanhalli, *Relation-aware Compositional Zero-Shot Learning for Attribute-Object Pair Recognition*, IEEE Transactions on Multimedia, Volume 24, 2022.

Preprints

- J. Lu, **Z. Xu**, and M. Kankanhalli, *Reasoning LLMs are Wandering Solution Explorers*, arXiv:2505.20296, 2025.
- **Z. Xu**, U. Sanghi, and M. Kankanhalli, *Bullying the Machine: How Personas Increase LLM Vulnerability*, arXiv:2505.12692, 2025.

- **Z. Xu**, S. Jain, and M. Kankanhalli, *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, arXiv:2401.11817, 2024.

TALKS

- *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, Online Seminar, US Food and Drug Administrative, March 2025.
- *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, Online Talk, Beijing Academy of Artificial Intelligence (Zhiyuan Institute), June 2024.
- *Combining Common Knowledge with Deep Learning Models*, Invited Talk, NUS Centre for Research in Privacy Technologies (N-CRiPT), NUS, May 2023.

AWARDS

- Champion of the ICML 2024 TiFA Workshop MLLM Attack Challenge (Team Member) *July 2024*
- AInet Fellowship, German Academic Exchange Service (DAAD) *April 2024*
- Research Achievement Awards, School of Computing, National University of Singapore *August 2022*
- NeurIPS Student Travel Grant *December 2022*
- Research Achievement Awards, School of Computing, National University of Singapore *December 2020*
- NeurIPS Student Travel Grant *December 2019*
- Outstanding Graduates, University of Science and Technology of China *July 2017*

TEACHING AND MENTORING

Teaching Assistantship

- CS5260 Neural Networks and Deep Learning II, School of Computing, NUS *January 2020*
- CS3244 Machine Learning, School of Computing, NUS *August 2019*
- CS5228 Knowledge Discovery and Data Mining, School of Computing, NUS *January 2019*
- CS5242 Neural Network and Deep Learning, School of Computing, NUS *August 2018*
- CS2010 Data Structures and Algorithms II, School of Computing, NUS *January 2018*

Mentoring

- Udit Sanghi, *Personality Traits can Make LLMs Vulnerable to Adversarial Manipulations*, B.Comp. Dissertation, *April 2025*.

SERVICES

Conference Reviewer

- NeurIPS (since 2023), ICML (since 2024), ICLR (since 2024)
- AAAI (since 2023), IJCAI (since 2023)
- CVPR (since 2025), ICCV (since 2025)
- ACM MM (since 2024)

Journal Reviewer

- IEEE TPAMI, TIP, TMM, TCSVT
- ACM TOMM

Volunteer

- Group Leader, NeurIPS High School Outreach *December 2022*
- Program Committee Member, Computing Research Week *August 2020*