Han Wang

Graduate Student
Department of Electrical and Computer Engineering
University of Illinois Urbana-Champaign
hanw14@illinois.edu

Education

University of Illinois Urbana-Champaign, *Urbana, IL* — Aug. 2024 – Present
Ph.D. in Electrical and Computer Engineering

Zhejiang University, *Hang Zhou, China* — Sept. 2020 – June 2024
B.Eng. in Electronic Information Engineering — *GPA:3.94/4.00*

Research Interests

Trustworthy Machine Learning, Trustworthy Large Language Model

Work Experiences

University of Illinois Urbana-Champaign, *Urbana, IL* — Aug. 2024 – Present
Research Assistant, Advisor: Prof. Huan Zhang

Publications & Manuscripts

(* = equal contribution)

- Jingnan Zheng*, **Han Wang***, An Zhang, Tai D. Nguyen, Jun Sun, Tat-Seng Chua. ALI-Agent: Assessing LLMs'Alignment with Human Values via Agent-based Evaluation. *Advances in Neural Information Processing Systems (NeurIPS), 2024.* [Paper][Code]
- **Han Wang**, Yixuan Li. Bridging OOD Generalization and Detection: A Graph-Theoretic View. *Advances in Neural Information Processing Systems (NeurIPS), 2024.*
- An Zhang*, **Han Wang***, Xiang Wang, Tat-Seng Chua. Disentangling Masked Autoencoders for Unsupervised Domain Generalization. *In Proceedings of the 18th European Conference on Computer Vision (ECCV), 2024.* [Paper] [Code]
- Mengze Li*, **Han Wang***, Wenqiao Zhang, Jiaxu Miao, Wei Ji, Zhou Zhao, Shengyu Zhang, Fei Wu. Winner: weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.* [Paper]

Research Experiences

An Efficient Defense against VLM Jailbreaks with Feature Steering — June 2024 – Present
Advisor: Prof. Huan Zhang, University of Illinois Urbana-Champaign

- Introduce an efficient adversarial training method to detoxify and safeguard the responses of VLMs by steering the hidden states without further fine-tuning
- Demonstrate the effectiveness of the defense while not harming the utility in several popular VLMs

Agent-based Evaluation of LLMs' Alignment with Human Values — Jan. 2024 – May. 2024
Advisor: Prof. Tat-Seng Chua, National University of Singapore, NExT++ Lab — *NeurIPS 2024*

- Propose an evaluation framework that leverages the autonomous abilities of LLM-powered agents to conduct in-depth and adaptive alignment assessments
- Demonstrate the empirical effectiveness across three aspects of human values (i.e., stereotypes, morality, and legality) in ten popular LLMs

Graph-Theoretic Understanding for OOD Generalization and Detection — May 2023 – Oct. 2023
Advisor: Asst. Prof. Sharon Yixuan Li, University of Wisconsin-Madison — *NeurIPS 2024*

- Propose a novel graph-theoretical framework for understanding both OOD generalization and detection
- Present theoretical insight by analyzing closed-form solutions for the OOD generalization and detection error
- Evaluate the performance through a set of experiments and provide empirical evidence of robustness and alignment with our theoretical analysis

Disentangling MAE for Unsupervised Domain Generalization Oct. 2022 – May 2023
Advisor: Prof. Tat-Seng Chua, National University of Singapore, NExT++ Lab *ECCV 2024*

- Devise a disentangling MAE framework to discover the disentangled representations that faithfully reveal the intrinsic features and superficial variations in an unsupervised manner
- Demonstrate the effectiveness beyond state-of-the-art unsupervised domain generalization methods and domain generalization methods

Weakly-supervised Spatio-temporal Video Grounding Jun. 2022 – Dec. 2022
Advisor: Prof. Fei Wu, Zhejiang University, DCD Lab *CVPR 2023*

- Present a novel perspective of hierarchical video language decomposition and alignment to alleviate spurious correlations brought by limited annotations
- Introduce a framework that encapsulates the structural attention and top-down backtracking for hierarchical understanding, using multi-hierarchy contrastive learning
- Outperform state-of-the-art weakly supervised methods, even surpass some supervised methods

Honors

Outstanding Graduates of Zhejiang University	June 2024
School of Electrical Engineering NR Scholarship, Zhejiang University	Oct. 2023
Zhejiang University Scholarship - Second Prize (Top 8%)	Oct. 2023
Zhejiang Province Government Scholarship	Nov. 2022
Zhejiang University Scholarship - Second Prize (Top 8%)	Oct. 2022
Zhejiang University Scholarship - First Prize (Top 3%)	Oct. 2021

Selected Courses

All GPA 4.0/4.0

- **Mathematics**: Calculus (A), Linear Algebra, Probability and Mathematical Statistics, Partial Differential Equations, Information Theory and Coding
- **CS**: Fundamentals of Data Structures, Object-Oriented Programming, Computer Organization and Design, Computer Network and Communication
- **EE**: Electric Circuit and Electronic Technology, Signal Analysis and Processing, Engineering Electromagnetic Fields and Waves, Power Electronics, Principles of Automatic Control

Skills

Programming Skills: Python, C/C++, Matlab, CUDA, VHDL/Verilog
Language Skills: Chinese (Native), English (Fluent, TOEFL iBT 100/120)