Yiwei Chen

chenyiw9@msu.edu | Homepage | LinkedIn | +1 5172046319

Education

Michigan State University *East Lansing, MI, USA*

- **Ph.D. in Computer Science** Aug. 2024 -
  Advisor: Prof. Sijia Liu

Xi'an Jiaotong University *Xi'an, Shaanxi, China*

- **M.S. in Computer Science and Technology.** Ranking: 7/173 Sept. 2021 - Jun. 2024
- **B.Eng. in Automation.** Sept. 2017 - Jun. 2021
  Honor Class in Qian Xuesen Honors College. Ranking: 3/24
- **Special Class for Gifted Young.** Sept. 2015 - Jun. 2017
  A honor program for nationwide selected gifted young students.

Research Interests

Machine Unlearning, Safety Alignment, Vision-Language Models, Adversarial Machine Learning

Publications

(* means Equal Contribution, † means Corresponding)

[1] **Yiwei Chen***, Yuguang Yao*, Yihua Zhang, Bingquan Shen, Gaowen Liu, Sijia Liu†. Safety Mirage: How Spurious Correlations Undermine VLM Safety Fine-tuning. Submitted to *ICCV*, 2025. [Paper]

[2] Zhihao Zhang*, **Yiwei Chen***, Weizhan Zhang†, Caixia Yan, Qinghua Zheng, Qi Wang, Wangdu Chen. Tile Classification Based Viewport Prediction with Multi-modal Fusion Transformer. *ACM Multimedia (ACM-MM)*, 2023. [Paper]

[3] Zhang Lingling†, **Yiwei Chen**, Wu Wenjun, Wei Bifan, Luo Xuan, Chang Xiaojun, Liu Jun. Interpretable Few-Shot Learning with Contrastive Constraint. *Chinese Journal of Computer Research and Development*, 2021. [Paper]

Research Experience

OPTML, Michigan State University *East Lansing, US*

Research Assistant. Advisor: Prof. Sijia Liu Jul. 2024 -

- VLM Unlearning: Conventional supervised safety fine-tuning of VLMs suffers from the "safety mirage" problem due to training data bias, resulting in spurious correlations and over-rejections following one-word attacks. Employing unlearning algorithms on VLMs effectively removes harmful content and addresses these safety issues. Submitted to ICCV 2025 [1].
- Backdoor Attack and Defense in VLMs: Investigate how a backdoored vision encoder can compromise the integrity and functionality of VLMs when deployed for downstream applications, such as object detection. Develop backdoor purification methods for VLMs using in-context learning.

Xi'an Jiaotong University *Xi'an, China*

Research Assistant. Advisor: Prof. Jun Liu & Prof. Qinghua Zhang Jan. 2021 - May. 2024

- Video Understanding: Develop MFTR , a transformer-based model for viewport prediction, achieving state-of-the-art long-term accuracy.
- Few-shot Learning: Develop INT-FSL , an interpretable few-shot learning method leveraging positional attention and contrastive constraints to enhance reasoning and address limited supervision.

Awards & Honors

SCHOLARSHIPS

- Outstanding Master Graduate, Xi'an Jiaotong University (Top 5% of all graduates) Jun. 2024
- Outstanding Bachelor Graduate, Xi'an Jiaotong University (Top 5% of all graduates) Jun. 2021
- Outstanding Freshman Scholarship, Xi'an Jiaotong University (Top 10%) Nov. 2021
- First Prize Scholarship, Xi'an Jiaotong University (Top 10%) Nov. 2016-2019, 2022-2023

COMPETITIONS

- Meritorious Winner in Mathematical Contest in Modelling (top 8% in 25370 teams) Mar. 2019
- The First Prize in Shaanxi of China Undergraduate Mathematical Contest in Modeling Oct. 2018

Academic Service

Conference Reviewer

- The International Joint Conference on Neural Networks (IJCNN 2025)
- The International Conference on Learning Representations (ICLR 2025)
- IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2025)
- ACM International Conference on Multimedia (ACM-MM 2023-2024)
- European Conference on Artificial Intelligence (ECAI 2023)

Language & Skills

Language	Mandarin (Native) English (TOEFL iBT 103, Reading 28 Listening 28 Speaking 25 Writing 22)
Programming	Python, C/C++, Bash, LaTeX, MATLAB, HTML/CSS
Model Frameworks	CLIP, LLMs (e.g., Llama, Vicuna), VLMs (e.g., LLaVA, miniGPT)
Libraries / Softwares	Pytorch, Tensorflow, Deepspeed, Huggingface, OpenCV
Developer Tools	Git, Docker, Vim, VSCode, PyCharm