Xiaoyu Guo

Email: xiaoyu.guo@wisc.edu **Phone:** +1 (608) 690-9300

RESEARCH INTERESTS

AI Bias, AI Ethics, Natural Language Processing (NLP)

EDUCATION

University of Wisconsin-Madison (UW-Madison) *Sep. 2023 - May. 2025 (Expected)*
M.S. in Information Science Madison, WI
Advisor: Chaowei Xiao

Jinan University / University of Birmingham *Sep. 2019 - Jun. 2023*
JNU BSc / UoB BSc Applied Mathematics in Information Computing (Dual Degree) Guangzhou, China
Graduated with Honors
Advisor: Junwei Duan

PUBLICATIONS

(P: Preprint, C: Conference, W: Workshop, *: Equal contribution)

[C1] JailBreakV-28K: A Benchmark for Assessing the Robustness of MultiModal Large Language Models against Jailbreak Attacks
Weidi Luo*, Siyuan Ma*, Xiaogeng Liu*, **Xiaoyu Guo**, Chaowei Xiao
Conference on Language Modeling (COLM) 2024

RESEARCH EXPERIENCE

Chaowei Xiao Lab Dec. 2023 - Present
Research Assistant, Advised by Chaowei Xiao Madison, WI

- **Benchmark for testing MLLMs' resistance to jailbreak attacks** [C1]
  Contributed to data preprocessing, running experiments, and the initial stages of dataset creation for the evaluation of jailbreak vulnerabilities in multimodal large language models (MLLMs). Assisted in cleaning and preparing the dataset, and supported the implementation of various attack methods used to test the robustness of MLLMs.

Xudong Hu Lab (Remote) Mar. 2024 - Present
Research Assistant, Advised by Xudong Hu Hongkong, China

- **Finetuning Diffusion Models for Fair Representation in Occupational Group Images (In Progress)**
  Focused on extending individual-level debiasing techniques to group image generation in diffusion models, targeting occupational images where demographic attributes (gender, race, age) reflect real-world distributions. Applied a combination of distributional alignment loss (DAL) and direct fine-tuning (DFT) to ensure demographic fairness in both scenarios: images with high-visibility individuals, where key figures are prominently positioned, and regular group images without focal points.

Data Driven Intelligent System Lab Sep. 2019 - Jun. 2023
Research Assistant, Advised by Junwei Duan Guangzhou, China

- **Classification Model for Osteoporosis Images Based on Broad Learning System (BLS)**
  Worked on a project to classify osteoporosis images, enhancing images, extracting features, and integrating CNN with BLS to improve accuracy. Developed the app's core functionalities, including user login, image uploads, and result outputs.

HONORS AND REWARDS

- Jinan University Tier One Scholarship, top 10% Nov. 2021
- Jinan University Outstanding Student Leader Award Nov. 2020, Nov. 2021

TEACHING EXPERIENCE

- MATLAB Programming Design Spring 2021
- Algorithm Design and Analysis Spring 2023

WORK EXPERIENCE

FJ Dynamics Technology Co., Ltd, Shenzhen, China Jun. 2022 –Jan. 2023

Data Analysis Intern

- Automated sales reporting and data filtering using Python with FCRM data.
- Visualized sales trends and used time series analysis to forecast product demand.
- Recommended inventory adjustments to optimize supply and demand balance.

OTHER EXPERIENCE

Jinan University-University of Birmingham Joint Institute, Guangzhou, China Sep. 2020 – Sep. 2021

Head of Student Recruitment Team

- Led a mascot design competition, managed TikTok account posts, and interviewed teachers and alumni.
- Edited videos for social media and collaborated with officials on funding.