Kaiwen Zhou

Kaiwen Zhou | kevinz-01.github.io | kzhou35@ucsc.edu

Education

University of California, Santa Cruz — Sep. 2021 – Present
Ph.D. in Computer Science and Engineering — Advisor: Prof. Xin Eric Wang.
Research focus: AI safety, AI agents, embodied AI.

Zhejiang University — Sep. 2017 – June 2021
B.S. in Statistics

Selected Publications

- **SIRAJ: Diverse and Efficient Red-Teaming for LLM Agents via Distilled Structured Reasoning**.
Kaiwen Zhou, Ahmed Elgohary, A S M Iftekhar, Amin Saied.
In submission.
- **Presenting a Paper is an Art: Self-Improvement Aesthetic Agents for Academic Presentations**.
Chengzhi Liu*, Yuzhe Yang*, **Kaiwen Zhou**, Zhen Zhang, Yue Fan, Yannan Xie, Peng Qi, Xin Eric Wang.
In submission.
- **SafeKey: Amplifying Aha-Moment Insights for Safety Reasoning**.
Kaiwen Zhou, Xuandong Zhao, Gaowen Liu, Jayanth Srinivasa, Aosong Feng, Dawn Song, Xin Eric Wang.
EMNLP 2025.
- **The Hidden Risks of Large Reasoning Models: A Safety Assessment of R1**.
Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, Xin Eric Wang.
IJCNLP-AACL 2025.
- **Multimodal Situational Safety**.
Kaiwen Zhou*, Chengzhi Liu*, Xuandong Zhao, Anderson Compalas, Dawn Song, Xin Eric Wang.
ICLR 2025, *NeurIPS Workshop on RBFM 2024 (Oral).*
- **Muffin or Chihuahua? Challenging Large Vision-Language Models with Multipanel VQA**.
Yue Fan, Jing Gu, **Kaiwen Zhou**, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, Xin Eric Wang.
ACL 2024.
- **ViCor: Bridging Visual Understanding and Commonsense Reasoning with Large Language Models**.
Kaiwen Zhou, Kwonjoon Lee, Teruhisa Misu, Xin Eric Wang.
Findings of ACL 2024.
- **Navigation as the Attacker Wishes? Towards Building Byzantine-Robust Embodied Agents under Federated Learning**.
Yunchao Zhang, Zonglin Di, **Kaiwen Zhou**, Cihang Xie, Xin Eric Wang.
NAACL 2024.
- **ESC: Exploration with Soft Commonsense Constraints for Zero-shot Object Navigation**.
Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, Xin Eric Wang.
ICML 2023.
- **JARVIS: A Neuro-Symbolic Commonsense Reasoning Framework for Conversational Embodied Agents**.
Kaizhi Zheng*, **Kaiwen Zhou***, Jing Gu*, Yue Fan*, Jialu Wang*, Zonglin Di, Xuehai He, Xin Eric Wang.
NeSy 2025 (Oral), *SoCal NLP 2022*
- **FedVLN: Privacy-preserving Federated Vision-and-Language Navigation**.
Kaiwen Zhou, Xin Eric Wang.

ECCV 2022.

Selected Research Projects

Diverse and Efficient Red-Teaming for LLM Agents Jun. 2025 – Sep. 2025

Develop a red-teaming framework that generates diverse seed tests and iteratively crafts adversarial attacks using a red-teamer trained via structured reasoning with supervised fine-tuning and reinforcement learning. **Deployed in Microsoft RAI product for agent safety.**

Improving the Safety Alignment of Large Reasoning Models March 2025 – May. 2025

Identify the safety aha-moment of large reasoning models (LRMs), and amplify it for safer LRM with the proposed SafeKey training method, leading to significant safety improvement.

Safety Analysis on Large Reasoning Models Jan. 2025 – Feb. 2025

Identify safety gaps and safety behaviors in open-source reasoning models, including increased harmfulness level in unsafe responses, harmful reasoning outputs, and failure safety thinking when facing adversarial attacks, etc.

Multimodal Situational Safety Apr. 2024 – Sep. 2024

Propose a novel safety problem where the situation in visual input affects the safety of the user's intent in chat and embodied scenarios; benchmark MLLMs and propose multi-agent pipelines to improve situational safety.

Visual Commonsense Reasoning with LLMs and VLMs Mar. 2023 – Sep. 2023

Define VCR as visual commonsense inference or understanding, and propose a workflow maximizing the capability of LLMs and VLMs to solve them.

LLM Commonsense Reasoning for Zero-shot Object Navigation Jun. 2022 – Jan. 2023

Combine commonsense reasoning of pre-trained LLMs and classical embodied navigation via Probabilistic Soft Logic (PSL) to achieve SOTA zero-shot object navigation performance.

Amazon Alexa Prize SimBot Challenge Jan. 2022 – Apr. 2023

Build dialog-based embodied instruction following agent; won first place in the public challenge (phase I) and third place in real-user interaction stage (phase II).

Privacy-preserving Federated Learning for Navigation Agents Sep. 2021 – March 2022

Build a two-stage federated learning framework for vision-and-language navigation agents to preserve users' data privacy while maintaining navigation performance.

Work Experience

Research Intern, Microsoft Responsible AI Mentor: Ahmed Elgohary Jun. 2025 – Sep. 2025

Research Intern, Samsung Research America Mentor: Yilin Shen Jun. 2024 – Sep. 2024

Research Intern, Honda Research Institute Mentor: Kwonjoon Lee Apr. 2023 – Dec. 2023

Research Intern, Samsung Research America Mentor: Yilin Shen Jun. 2022 – Sep. 2022

AI Technical Skills

Post-training, alignment, reinforcement learning, supervised fine-tuning, reasoning, multimodal LLMs.

Miscellaneous

- Dissertation-Year Fellowship, UCSC (2025-2026)
- First place of Alexa Prize SimBot Public Benchmark Challenge.
- Third place of Alexa Prize SimBot Challenge.