

# Kaiwen Zhou

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

## EDUCATION

---

**University of California, Santa Cruz**  
Ph.D. in Computer Science and Engineering  
Research focus: AI safety, AI agents, embodied AI.

Sep. 2021 – Present  
Advisor: Prof. Xin Eric Wang.

**Zhejiang University**  
B.S. in Statistics

Sep. 2017 – June 2021

## WORK EXPERIENCE

---

**Research Fellow, MATS** Mentor: William Saunders (Anthropic) Jan. 2026 – Present

- Interpreting and monitoring LLM misaligned reasoning process, collaborating with Anthropic's Safety team.

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

- Developed a red-teaming framework for LLM agents that iteratively crafts adversarial attacks.
- Built an effective and efficient red-teamer trained via distilled structured reasoning using SFT and RL.
- **Impact:** Deployed in Microsoft RAI product; a first-author paper (*Findings of EACL 2026*).

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

- Developed prototype LLM-based agents for coding, scientific idea verification, and literature search.

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

- Developed a Novel framework for visual reasoning, maximizing the capability of foundation models.
- Achieved state-of-the-art training-free performance on visual reasoning tasks (*Findings of ACL 2024*).

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

- Combined LLM reasoning with Probabilistic Soft Logic (PSL) for zero-shot object navigation.
- Achieved state-of-the-art performance in zero-shot embodied navigation tasks (*ICML 2023*).

## SELECTED PUBLICATIONS

---

- SafePro: Evaluating the Safety of Professional-Level AI Agents  
**Kaiwen Zhou**, Shreedhar Jangam, Ashwin Nagarajan, Tejas Polu, Suhas Oruganti, ..., Xin Eric Wang.
- [EACL 2026 Findings] SIRAJ: Diverse and Efficient Red-Teaming for LLM Agents via Distilled Structured Reasoning  
**Kaiwen Zhou**, Ahmed Elgohary, A S M Iftekhar, Amin Saied.
- [ICLR 2026] 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.
- [EMNLP 2025] SafeKey: Amplifying Aha-Moment Insights for Safety Reasoning  
**Kaiwen Zhou**, Xuandong Zhao, Gaowen Liu, Jayanth Srinivasa, Aosong Feng, Dawn Song, Xin Eric Wang.
- [IJCNLP-AAACL 2025] The Hidden Risks of Large Reasoning Models: A Safety Assessment of R1  
**Kaiwen Zhou**, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, ..., Dawn Song, Xin Eric Wang.
- [ICLR 2025] Multimodal Situational Safety  
**Kaiwen Zhou\***, Chengzhi Liu\*, Xuandong Zhao, Anderson Compalas, Dawn Song, Xin Eric Wang.
- [ACL 2024] 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 Findings] ViCor: Bridging Visual Understanding and Commonsense Reasoning with Large Language Models

**Kaiwen Zhou**, Kwonjoon Lee, Teruhisa Misu, Xin Eric Wang.

- [NAACL 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.
- [ICML 2023] 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.
- [NeSy 2025 (Oral)] 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.
- [ECCV 2022] FedVLN: Privacy-preserving Federated Vision-and-Language Navigation  
**Kaiwen Zhou**, Xin Eric Wang.

## SELECTED RESEARCH PROJECTS

---

**AGI Safety: Safety Evaluation for Professional-Level AI Agents** Oct. 2025 – Jan. 2026

Develop a safety evaluation dataset with safety risks in professional-level agentic tasks. Build an agent safety evaluation framework. Identify safety gaps of current AI models.

**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.

**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.

## AI TECHNICAL SKILLS

---

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

## MISCELLANEOUS

---

- Dissertation-Year Fellowship, UCSC (2025-2026)
- Area Chair: ARR Oct 2025
- Reviewer: NeurIPS 2023, ICLR 2024, ICML 2024, ICLR 2025, ICLR 2026