

# Kaiwen Zhou

Updated February 19, 2025

**Email:** kzhou35@ucsc.edu

**Github:** //github.com/KevinZ-01

**Webpage:** https://kevinz-01.github.io/

## Research interests

AI Agents, (Multimodal) Large Language Models, AI Safety

## Education

**University of California, Santa Cruz**

Ph.D. in Computer Science and Engineering

Sep. 2021 – Present

Advisor: Prof. Xin Eric Wang.

**Zhejiang University**

B.S. in Statistics

Sep. 2017 – June 2021

## Work experience

**Samsung Research America**

Research intern

June 2024 – Sep. 2024

Mentor: Yilin Shen

**Honda Research Institute**

Research intern

April 2023 – Dec. 2023

Mentor: Kwonjoon Lee

**Samsung Research America**

Research intern

June 2022 – Sep. 2022

Mentor: Yilin Shen

## Publications

**Multimodal Situational Safety**

**Kaiwen Zhou\***, Chengzhi Liu\*, Xuandong Zhao, Anderson Compalas, Dawn Song, Xin Eric Wang

*ICLR 2025, NeurIPS Workshop on RBFM 2024* *Oral*

**The Hidden Risks of Large Reasoning Models: A Safety Assessment of R1**

**Kaiwen Zhou**, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Gaowen Liu, Jayanth Srinivasa, Dawn Song, Xin Eric Wang

*Arxiv 2025*

**Muffin or Chihuahua? Challenging Large Vision-Language Models with Multi-panel 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*

## **FedVLN: Privacy-preserving Federated Vision-and-Language Navigation**

**Kaiwen Zhou**, Xin Eric Wang.

*ECCV 2022*

## **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.

*Socal NLP 2022*

### Selected research

#### **Multimodal Situational Safety**

Advisor: Prof. Xin Eric Wang

April. 2024 – Sep. 2024

- We propose a novel safety problem where the situation indicated by the visual input is a critical factor that influences the safety of the user's intent behind the query.
- We benchmark SOTA MLLMs and perform in-depth analysis.
- We propose multi-agent pipelines to improve situational safety performance.

#### **Visual Commonsense Reasoning with LLM and VLMs**

Advisor: Dr. Kwonjoon Lee, Prof. Xin Eric Wang

Mar. 2023 – Sep. 2023

- We studied the problem of visual commonsense reasoning and defined it into two sub-tasks: visual commonsense inference and visual commonsense understanding.
- We proposed a framework maximizing the capability of LLMs and VLMs to solve them.

#### **LLM Commonsense Reasoning for Zero-shot Object Navigation**

Advisor: Prof. Xin Eric Wang, Dr. Yilin Shen

June 2022 – Jan. 2023

- We proposed a framework that combines the commonsense reasoning of pre-trained LLM and classical navigation methods via Probabilistic Soft Logic (PSL) for zero-shot object navigation.
- We achieve SOTA zero-shot object navigation performance.

#### **Privacy-preserving Federated Vision-and-Language Navigation**

Advisor: Prof. Xin Eric Wang

Sep. 2021 – Mar. 2022

- We study the data privacy problem of VLN and propose a federated learning framework for vision and language navigation.
- We preserve the training and inference data privacy with comparable results with centralized training and achieve the best performance in pre-exploration.

### Skills

#### **Programming**

Python, Pytorch.

### Other experience

#### **Amazon Alexa Prize SimBot Challenge**

Advisor: Prof. Xin Eric Wang

Jan. 2022 – Apr. 2023

We investigated the problem of dialog-based embodied instruction following on TEACH benchmark and won the **first place** public challenge in the first phase. In the second phase, we are building an interactive embodied agent that can finish diverse tasks cooperating with human players. We won the **third place** in the second phase.