Kaishen Wang

wangks@stu.scu.edu.cn | kaishenw7322@gmail.com | https://github.com/tunantu

Education

Sichuan University, BS in Computer Science Sep. 2021 – Jun. 2025

- **GPA**: 3.9/4.0; **Average Score**: 91.75/100; **Rank**: 3
- **Coursework:** Machine Learning, Introduction to Deep Leaning, Data Structure and Algorithm Analysis, Principles of Computer Organization, Computer Network, Theory of Optimization, Operating System

Publication

Strengthening Layer Interaction via Dynamic Layer Attention

Kaishen Wang, Xun Xia, Jian Liu, Zhang Yi, Tao He

IJCAI 2024

Zer0-Jack: A memory-efficient gradient-based jailbreaking method for black box Multi-modal Large Language Models

Kaishen Wang*, Tiejin Chen*, Hua Wei

Accepted to NeurIPS Workshop 2024, ICLR 2025 under review

DAMO: Decoding by Accumulating Activations Momentum for Mitigating Hallucinations in Vision-Language Models

Kaishen Wang*, Hengrui Gu*, Meijun Gao, Kaixiong Zhou

ICLR 2025 under review

Experience

Internship at **Purdue**, supervised by Prof. **Ruqi Zhang** Oct. 2024 - Ongoing

- We focus on addressing hallucination issues in VLMs by aligning visual and textual modalities through RLHF.

Internship at **NCSU**, supervised by Prof. **Kaixiong Zhou** Mar. 2024 - Oct. 2024

- We developed the DAMO technique, which accumulates layer-wise activations to enhance visual semantics consistency and mitigate hallucinations during inference.

Internship at **ASU**, supervised by Prof. **Hua Wei** Feb. 2024 - Oct. 2024

- We proposed Zer0-Jack, a zeroth-order optimization method to jailbreak black-box MLLMs efficiently, with a low memory usage cost similar to inference procedure.

Internship at **SCU**, supervised by Prof. **Tao He** Jun. 2023 - Jan. 2024

- We introduced Dynamic Layer Attention (DLA) architecture to restore the dynamic context representation to facilitate layer interaction.

Research Interests

- Building **reliable** models, particularly in the context of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs).
- **Alignment** research, specifically in aligning models with human preferences, and aligning textual and visual modalities in MLLMs.

Skills

Languages: Python, C, C++, SQL, Java, C#, JavaScript

Software: Docker, Git, Microsoft SQL Server