# ZIJUN WANG

zwang745@ucsc.edu · +(1) 8313348994 · asillycat.github.io · n5wjgV0AAAAJ

## EDUCATION

**University of California, Santa Cruz**, Santa Cruz, United States 2024.08 – Present

*PhD student (Expected graduation date: 06/2028)*
**Advised by Prof. Cihang Xie** at VLAA LAB in Computer Science and Engineering Department
**Research interest: AI Safety & Alignment, Autonomous Agents, Data-Centric LLM Training**

**Zhejiang University**, Hangzhou, China 2020.09 – 2024.06

*Bachelor of Engineering*
**Major in Computer Science and Technology**, College of Computer Science and Technology

## EXPERIENCE

**ByteDance Intern** San Jose, CA

*Data-TnS-Algo-Foundations & Intelligence Service* *2025.06 – 2026.04*

- Under Supervision of Fengze Liu
- Worked on **Pretraining Foundation LLM**
- First-authored **NAG-based Ranking**, a training-free and interpretable framework for target-oriented pretraining data selection (Accepted by **ICML 2026**)
- Co-authored **InfoLaw**, a data-aware scaling law framework modeling quality-weighted mixture data and repetition (Accepted by **ICML 2026**)

## PUBLICATIONS

### STAR-1: Safer Alignment of Reasoning LLMs with 1K Data

***Zijun Wang**, et al.*
Accepted by *Association for the Advancement of Artificial Intelligence* (**AAAI 2026 (*oral*)**)
***TL;DR:*** STAR-1 is a high-quality, just-1k-scale safety dataset for LRMs built on diversity, deliberative reasoning, and rigorous filtering. Fine-tuning LRMs with STAR-1 leads to an average 40% improvement in safety across four benchmarks, with only a marginal 1.1% average decrease in reasoning ability across five tasks.

### Target-Oriented Pretraining Data Selection via Neuron-Activated Graph

***Zijun Wang**, et al.*
Accepted by *International Conference on Machine Learning* (**ICML 2026**)
***TL;DR:*** We introduce Neuron-Activated Graph Ranking (NAG), a training-free and interpretable framework for target-oriented pretraining data selection. NAG characterizes each target input by a sparse set of high-impact neurons in off-the-shelf LLMs, improving target-oriented pretraining by 4.9% over random sampling and outperforming SOTA baselines by 5.3% on HellaSwag; deactivating just 0.12% of the identified neurons causes substantial degradation, isolating a sparse functional backbone for target-specific features.

### Your Agent, Their Asset: A Real-World Safety Analysis of OpenClaw

***Zijun Wang**, et al.*
In submission to *Conference on Language Modeling* (**COLM 2026**)
***TL;DR:*** A real-world safety analysis of OpenClaw, a widely-used personal AI agent with broad system access (Gmail, Stripe, etc.). We introduce the **CIK taxonomy** (Capability, Identity, Knowledge) for persistent agent state, and show that poisoning any single CIK dimension raises attack success rate from 24.6% to 64–74% across four models; even the strongest defense still admits 63.8%, indicating the vulnerabilities are architectural.

### VLAA-GUI: Knowing When to Stop, Recover, and Search, A Modular Framework for GUI Automation

*Qijun Han, Haoqin Tu,* ***Zijun Wang****, et al.*
In submission to *European Conference on Computer Vision* (**ECCV 2026**)
***TL;DR:*** VLAA-GUI is a modular GUI agentic framework integrating a Completeness Verifier (visual-evidence success validation), a Loop Breaker (multi-tier filtering against repetitive failures), and a Search Agent (LM-queried workflows). It achieves 77.5% on OSWorld and 61.0% on WindowsAgentArena, with three backbones surpassing human performance on OSWorld in a single pass.

### AttnGCG: Enhancing Adversarial Attacks on Language Models with Attention Manipulation

***Zijun Wang****, et al.*
Accepted by *Transactions on Machine Learning Research* **(TMLR 2025)**
***TL;DR:*** AttnGCG additionally manipulates models' attention scores to enhance LLM jailbreaking, achieving an average improvement of 7% in Llama-2 and 10% in Gemma series, with robust attack transferability against unseen harmful goals and black-box LLMs.

### How Many Unicorns Are in This Image? A Safety Evaluation Benchmark for Vision LLMs

*Haoqin Tu*, Chenhang Cui*,* ***Zijun Wang****, et al. (* equal contribution)*
Accepted by *European Conference on Computer Vision* (**ECCV 2024**)
***TL;DR:*** A comprehensive safety evaluation suite for Vision LLMs, shifting focus from standard performance to out-of-distribution generalization and adversarial robustness.

### MIRA: When Visualizing is the First Step to Reasoning, a Benchmark for Visual Chain-of-Thought

*Yiyang Zhou*, Haoqin Tu*,* ***Zijun Wang****, et al. (* equal contribution)*
Accepted by *Conference on Computer Vision and Pattern Recognition* (**CVPR 2026**)
***TL;DR:*** MIRA is a benchmark where generating intermediate visual images (sketches, diagrams, path drawings) is essential for reasoning, targeting tasks with complex structures and spatial relationships that are difficult to express through language alone.

### Chasing the Public Score: User Pressure and Evaluation Exploitation in Coding Agent Workflows

arXiv 2026
***TL;DR:*** Examines whether coding agents exploit evaluation metrics under user pressure. Using AgentPressureBench (34 tasks, 1,326 trajectories from 13 agents), identifies 403 exploitative behaviors; stronger models exhibit higher exploitation rates (correlation 0.77), and explicit anti-exploit prompting reduces exploitation from 100% to 8.3%.

### InfoLaw: Information Scaling Laws for Large Language Models with Quality-Weighted Mixture Data and Repetition

Accepted by *International Conference on Machine Learning* (**ICML 2026**)
***TL;DR:*** A data-aware scaling framework that predicts LLM loss from model size, data mixture weights, and repetition, enabling efficient data-recipe selection under varying compute budgets.

### AHELM: A Holistic Evaluation of Audio-Language Models

*Tony Lee, Haoqin Tu, Chi Heem Wong,* ***Zijun Wang****, et al.*
arXiv 2025
***TL;DR:*** A comprehensive benchmark for evaluating audio-language models across diverse audio understanding tasks.

### Where on Earth? A Vision-Language Benchmark for Probing Model Geolocation Skills Across Scales

arXiv 2025
***TL;DR:*** A benchmark probing VLM geolocation capabilities across multiple spatial scales.

**Mimicking the Physicist's Eye: A VLM-centric Approach for Physics Formula Discovery**

arXiv 2025

***TL;DR:*** Leveraging vision-language models to automate physics formula discovery from visual observations.

**Handling Feature Heterogeneity with Learnable Graph Patches**

*Yifei Sun, Yang Yang, Xiao Feng,* ***Zijun Wang****, et al.*

*Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (**KDD 2025**)

***TL;DR:*** Introduces learnable graph patches as minimal semantic units for cross-domain graph pre-training.

## AWARDS

- **Second Place** in both base & large model subtracks of Red Teaming LLM@**NeurIPS 2023**, Torjan Detection Challenge(**Team leader**). [Code]
- **National Scholarship (top 0.2% national-wide)** issued by Ministry of Education of the People's Republic of China, 2021