Hongrui Jia

jiahongrui@stu.pku.edu.cn GitHub Google Scholar hongruijia.github.io

Research Vision

My research focuses on **improving the reliability of multimodal reasoning in MLLMs and agents**. I study how to diagnose failure modes in multimodal reasoning, enforce evidence-grounded reasoning processes, and improve model capabilities through diagnostic-driven training. My goal is to **build multimodal AI systems that can reason reliably in complex open-world environments**.

Education

Jun, 2027 (Expected) **Peking University** Beijing, China
Master of Engineering in Software Engineering
Advisors: Prof. Wei Ye; Prof. Shikun Zhang
GPA: 3.73/4.00 **Citations:** 216 **H-index:** 6 (May 2026)

Jun, 2024 **South China University of Technology** Guangzhou, China
Bachelor of Engineering in Software Engineering
GPA: 3.89/4.00

Publications (Selected)

[1] **From Blind Spots to Gains: Diagnostic-Driven Iterative Training for Large Multimodal Models**
Hongrui Jia, Chaoya Jiang, Shikun Zhang, Wei Ye
International Conference on Machine Learning (ICML 2026) Code

[2] **OSWorld-MCP: Benchmarking MCP Tool Invocation In Computer-Use Agents**
Hongrui Jia, Jitong Liao, Xi Zhang, Haiyang Xu, Tianbao Xie, Chaoya Jiang, Ming Yan, Si Liu, Wei Ye, Fei Huang
International Conference on Learning Representations (ICLR 2026) Code

[3] **Mitigating Visual Context Degradation in Large Multimodal Models: A Training-Free Decoupled Agentic Framework**
Hongrui Jia, Chaoya Jiang, Shikun Zhang, Wei Ye
Computer Vision and Pattern Recognition Conference, Findings Track (CVPR 2026) Code

[4] **SymDPO: Boosting In-Context Learning of Large Multimodal Models with Symbol Demonstration Direct Preference Optimization**
Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, Shikun Zhang
Computer Vision and Pattern Recognition Conference (CVPR 2025) Code

[5] **MaVEn: An Effective Multi-Granularity Hybrid Visual Encoding Framework for Multimodal Large Language Model**
Chaoya Jiang*, **Hongrui Jia*** (*Equal Contribution), Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, Shikun Zhang
Advances in Neural Information Processing Systems (NeurIPS 2024) Code

[6] **Hal-eval: A Universal and Fine-grained Hallucination Evaluation Framework for Large Vision Language Models**
Chaoya Jiang*, **Hongrui Jia*** (*Equal Contribution), Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, Shikun Zhang
ACM International Conference on Multimedia (ACM MM 2024), **Oral**

[7] **ISC4DGF: Enhancing Directed Grey-box Fuzzing with LLM-Driven Initial Seed Corpus Generation**
Yijiang Xu, **Hongrui Jia**, Liguo Chen, Xin Wang, Zhengran Zeng, Yidong Wang, Qing Gao, Jindong Wang, Wei Ye, Shikun Zhang, Zhonghai Wu
Journal of computer science and technology (JCST)

[8] ***MaVEn*^{2}: A Unified Multi-Granularity Visual Encoding Framework for End-to-End Interleaved Vision-Language Reasoning**
Jiashu Lv*, Chaoya Jiang*, **Hongrui Jia***, Ruiyan Xu, Yang Han, Shikun Zhang, Wei Ye
Under Review (TPAMI)

Honors (Selected)

Oct, 2025 **National Scholarship** Ministry of Education, China

Oct, 2021 **National Scholarship** Ministry of Education, China

Industrial Experience

Tongyi Lab, Alibaba Group **Beijing**
Research Intern · Mentored by: Haiyang Xu, Ming Yan *May 2025 – Mar 2026*

- As the importance of MCP capabilities grows, existing GUI benchmarks lack MCP tool integration, leading to unfair evaluations. To address this, I constructed the **OSWorld-MCP** benchmark for comprehensively evaluating computer-use agents on MCP tool invocation, GUI interaction, and decision-making, with the paper accepted at **ICLR 2026**. I also integrated the OSWorld-MCP environment into a **large-scale model RL training framework** to enhance foundation models' tool-calling and GUI capabilities.
- To tackle the **sparse RL reward** problem in GUI scenarios, I designed a Reward Model training pipeline and built the critic framework in Mobile-Agent-v3. Through multi-round querying and voting along with a dual-channel consensus mechanism (text reasoning channel + multimodal reasoning channel), the critic achieved over 90% accuracy at both the step level and trajectory level. I also contributed to building an asynchronous RL framework.
- Constructed diverse synthetic environments for RL training of large-scale models (Qwen3-VL-235B-A22B), continuously improving **foundation models' long-horizon task capabilities, tool-calling capabilities, and GUI capabilities**.

JD EXPLORE ACADEMY, Jingdong Group **Beijing**
Research Intern · Mentored by: Zenan Zhou *Mar 2026 – Present*

- Built the web search capability for JoyAI: on the framework side, constructed an **AgenticSearch framework**; on the model side, improved search-while-answering ability via RL to reduce hallucinations and enhance timeliness. Designed and introduced Skills with a SkillRouter to enable domain-adaptive search and generation.
- Built a sandbox to improve foundation models' **image search, text search, and code-based image processing capabilities** via RL.
- As multimodal models are increasingly applied to complex reasoning tasks, traditional discrete-text-based reasoning paradigms suffer from visual information loss and reasoning drift. To address this, I explored the **latent reasoning** mechanism that incorporates continuous visual representations (latent tokens) during inference to strengthen the model's retention and utilization of visual semantics. The work focuses on resolving latent token rigidity by guiding the model to generate and utilize diverse latent representations at critical reasoning steps, thereby improving visual consistency and expressiveness in reasoning.

Academic Service

Conference Reviewer: CVPR 2026, ICLR 2026, EMNLP 2025, ACL 2025, NAACL 2025

Open-source Service: Main contributor to OSWorld-MCP Framework, DPE Framework and open-source VLM Training Methods SymDPO.