Wenjie Du

duwenjie24@gmail.com • Homepage • Google Scholar • Singapore

Education

Nanyang Technological University, Singapore — Jan. 2026 – Present

M.Sc. in Artificial Intelligence

Sichuan University, Chengdu, China — Sept. 2020 – June 2024

B.Eng. in Computer Science and Technology, **Average Scores:** 89.4/100, **GPA:** 3.74/4.00

Publications

- **Wenjie Du**, Li Jiang, Keda Tao, Xue Liu, Huan Wang. "Which Heads Matter for Reasoning? RL-Guided KV Cache Compression." **arXiv'25**. [Page], [PDF]
- Keda Tao, Yuhua Zheng, Jia Xu, **Wenjie Du**, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, Weiqiang Wang, Jian Liu, Can Qin, Yulun Zhang, Ming-Hsuan Yang, Huan Wang. "LVOmniBench: Pioneering Long Audio-Video Understanding Evaluation for Omnimodal LLMs." **arXiv'26**. [Page], [PDF]
- Keda Tao, **Wenjie Du**, Bohan Yu, Weiqiang Wang, Huan Wang. "OmniAgent: Audio-Guided Active Perception Agent for Omnimodal Audio-Video Understanding." **arXiv'25**. [Page], [PDF]
- Hao Wen, Shizuo Tian, Borislav Pavlov, **Wenjie Du**, Yixuan Li, Ge Chang, Shanhui Zhao, Jiacheng Liu, Yunxin Liu, Ya-Qin Zhang, Yuanchun Li. "AutoDroid-V2: Boosting SLM-based GUI Agents via Code Generation." **MobiSys'25**, **Best Artifact Award**. [PDF]
- Shanhui Zhao, Hao Wen, **Wenjie Du**, Cheng Liang, Yunxin Liu, Xiaozhou Ye, Ye Ouyang, Yuanchun Li. "LLM-Explorer: Towards Efficient and Affordable LLM-based Exploration for Mobile Apps." **MobiCom'25**. [PDF]

Experience

Nanyang Technological University — Singapore

Research Intern at MMLab, supervised by Ziwei Liu — Jan. 2026 – Present

- Designing and building an evaluation benchmark for **VLM reward models** and **VLM-as-judge** in RLHF.

Westlake University — Hangzhou, China

Research Intern at EncodeLab, supervised by Huan Wang — July 2025 – Jan. 2026

- Led **RLKV**, using RL as a probe to discover reasoning-critical attention heads by optimizing per-head cache usage against generation outcomes; allocates full KV cache to critical heads while aggressively compressing others, achieving **near-lossless 20–50% compression** on reasoning tasks, with generalization validated on knowledge QA.
- Wrote a **Triton** local-attention kernel and integrated it into the **AReaL + SGLang** RLVR training stack to enable head-level mixed full/local attention; extended full/local KV cache management in the **SGLang** runtime, achieving **1.94**× speedup at 50% compression.
- Co-developed **OmniAgent**, a **ReAct+reflection** agent decomposing omnimodal LLMs into 3 text-interfaced perception tools; surpasses **Gemini 2.5-Flash** by **+10% minimum** on three audio-video benchmarks.

Hong Kong University of Science and Technology — Hong Kong SAR

Research Assistant at MINSys Group, supervised by Xiaomin Ouyang — Sept. 2024 – Mar. 2025

- Applied semantic embeddings to improve cross-domain generalization in Human Activity Recognition tasks.

Institute for AI Industry Research, Tsinghua University — Beijing, China

Research Intern at AIoT Group, supervised by Yuanchun Li — Dec. 2023 – Aug. 2024

- Designed and implemented **AutoDroid-V2**, a script-based GUI agent that pre-constructs App API documentation; reduced token consumption by **90%** and latency by **93.1%** vs. traditional methods.
- Developed **LLM-Explorer**, a low-cost, high-efficiency Android app exploration agent for extracting UI interactions.

ByteDance Ltd. — Shanghai, China

Golang Software Engineer Intern — May 2023 – Nov. 2023

Uranus Quant Fin-tech Co. — Chengdu, China

C++ Software Engineer Intern — June 2022 – Sept. 2022

Selected Awards

- National Scholarship (Top 3/318) — 2023
- Outstanding Graduate Student of Sichuan Province — 2024

SKILLS

Research: Python, PyTorch, Triton, LaTeX. **Systems**: C, C++, Go, Linux, Git, JavaScript.