Wenxuan Huang (黄文轩)

osilly0616@gmail.com · +86-153-9960-7206 · https://github.com/Osilly

Google Scholar: https://scholar.google.com/citations?user=6Ys6HgsAAAAJ

Home Page: https://osilly.github.io

Education&Industry Experience

China Three Gorges University 2019.09-2023.06

School of Computer Science and Information Technology Bachelor

East China Normal University 2023.09-Present

School of Computer Science and Technology, Advisor: Shaohui Lin Master (Third-year)

Xiaohongshu 2024.06-Present

NLP group Research Intern

The Chinese University of Hong Kong 2025.05-Present

MMLab, Advisor: Wanli Ouyang Research Assistant

Research Interest & Proposal (Up to 2026/02)

Over the past two-plus years, my primary research interests have remained in **Model Acceleration** and AI4Geophysic, and I have completed several intriguing works. Early on, my research journey was both fortunate and unfortunate: I had no senior peers to guide me, and my advisor's help was minimal. In many respects, I had to rely entirely on myself, *i.e.*, from finding ideas, crafting a coherent story, implementing codes, and writing the paper, to interpreting reviewer comments and composing rebuttals. Luckily, although this was no easy feat, I mastered every step of the process, enabling me to rapidly complete multiple works such as *[CVPR2024] TokenExpansion*, *KDFS*, and [TGRS] *G-LA-MAG*. Through these challenges, I became a **"freedom" researcher**, capable of autonomously choosing and driving my research agenda.
Later, I joined a resource-rich industrial algorithms group where I secured formal research privileges and substantial support. This environment allowed me to broaden my research scope into the hot field of **Multimodal Large Language Models (MLLMs)**. Drawing on my independent research experience, I partly transitioned from executor to project leader, proposing ideas and guiding collaborators toward top-conference publications. Projects include *[ICLR2026] Vision-R1*, *Vision-DeepResearch*, *Vision-DeepResearch Benchmark (VDR-Bench)*, *[ICLR2026] IRG*, *[ICLR2025] Dynamic-LLaVA*, *LLaVA-RadZ*, *[ACMMM2025] TimeSoccer*, *[NeurIPS2025] Actial*, *[ICLR2026] AGILE*, *CompBench*, and *CLIP-Map*. These experiences have shaped my research style: I lead a project from inception to publication, avoiding rigid timelines and advancing at my own pace-an approach that, while sometimes causing schedule extensions, aligns with my vision of a **flexible, curiosity-driven research life**.
Currently, I remain focused on advancing AGI, with some topics below:

1. **Reasoning MLLMs & Agentic Reasoning (*Interleaving Reasoning*) MLLMs & Agentic RL**:
   - *[ICLR2026] Vision-R1*: First & Leader, text-based reasoning MLLM.
   - *Vision-DeepResearch*: First & Leader, agentic reasoning MLLM.
   - *Vision-DeepResearch Benchmark (VDR-Bench)*: Co-First (second) & Leader & Corresponding, multimodal deep-research benchmark.
   - *[ICLR2026] IRG*: First, interleaving reasoning MLLM (unified).
   - *[NeurIPS2025] Actial*: Co-First (second) & Leader, text-based reasoning MLLM.
   - *[ICLR2026] AGILE*: Co-First (second), agentic RL.
2. **Multimodal Understanidng**:
   - *[ACMMM2025] TimeSoccer*, Co-Firsr (second), Video MLLM.
   - *LLaVA-RadZ*, Co-Firsr (second) & Corresponding, MLLM.
3. **Efficient Model (Inference & Training Acceleration)**:
   - *[ICLR2025] Dynamic-LLaVA*, First, MLLM Inference Acceleration.
   - *[CVPR2024] TokenExpansion*, First, Vision-Transformer & MLLM Training Acceleration.
   - *CLIP-Map*, Co-First (second), CLIP Training Acceleration.
   - *KDFS*, First Student author (First author is my advisor), CNN Inference Acceleration.
4. **Others**:
   - [TGRS] *G-LA-MAG*, Firsr, AI4Science.
   - *CompBench*, Co-Firsr (second), image editing benchmark.

ℹ Selected Papers (Up to 2026/02)

**: Equal contribution, #: Project leader, ✉: Corresponding author*

Selected Paper

- **[Reasoning MLLM] Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models.**
  - ***Wenxuan Huang#***, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, Shaohui Lin.
  - [**Citation 400+&Star 700+ in less than a year**] Accepted to **ICLR 2026**, first author & project leader.
  - This is the first paper to explore how to effectively use R1-like RL for MLLMs and introduce Vision-R1, a reasoning MLLM that leverages cold-start initialization and RL training to incentivize reasoning capability.
  - arXiv: `https://arxiv.org/abs/2503.06749`
  - Github repo: `https://github.com/Osilly/Vision-R1`
- **[Agentic Reasoning MLLM] Vision-DeepResearch: Incentivizing DeepResearch Capability in Multimodal Large Language Models.**
  - ***Wenxuan Huang#***, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Yao Hu, Philip Torr, Feng Zhao, Wanli Ouyang.
  - [**Star 150+ in only one day**] preprint, first author & project leader.
  - We introduce the first long-horizon multimodal deep-research MLLM. It supports multi-turn, multi-entity, and multi-scale visual/textual search, extending both the number of reasoning turns and search-engine interactions to dozens of turns and hundreds of queries.
  - arXiv: `https://arxiv.org/abs/2601.22060`
  - Github repo: `https://github.com/Osilly/Vision-DeepResearch`
- **[Multimodal DeepResearch Benchmark] Vision-DeepResearch Benchmark: Rethinking Visual and Textual Search for Multimodal Large Language Models.**
  - Yu Zeng*, ***Wenxuan Huang*#***✉, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, Shiting Huang, Yiming Zhao, Yao Hu, Philip Torr, Wanli Ouyang, Shaosheng Cao.
  - [**Star 150+ in only one day**] preprint, co-first author (second) & project leader & corresponding author.
  - We propose the Vision-DeepResearch Benchmark (VDR-Bench) to address two key limitations of existing multi-modal deep-research benchmarks: (1) they are not centered on visual search, and (2) evaluation is conducted in overly idealized settings that overlook noisy real-world search engines.
  - arXiv: `https://arxiv.org/abs/2602.02185`
  - Github repo: `https://github.com/Osilly/Vision-DeepResearch`
- **[AIGC/Interleaving Reasoning/Unified MLLM] Interleaving Reasoning for Better Text-to-image Generation.**
  - ***Wenxuan Huang***, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, Junbo Qiao, Hangyu Guo, Yao Hu, Zhenfei Yin, Philip Torr, Yu Cheng, Wanli Ouyang, Shaohui Lin
  - Accepted to **ICLR 2026**, first author.
  - This is an early exploration to introduce Interleaving Reasoning to Text-to-image Generation field and achieve the SoTA benchmark performance. It also significantly improves the quality, fine-grained details and aesthetic aspects of generated images.
  - arXiv: `https://arxiv.org/abs/2509.06945`
  - Github repo: `https://github.com/Osilly/Interleaving-Reasoning-Generation`
- **[Interleaving Reasoning Postion/Survey] Interleaving Reasoning: Next-Generation Reasoning Systems for AGI.**
  - [**Star 200+**] In progress, first author.
  - ***Wenxuan Huang*** *et al.*
  - Recently, the introduction of *OpenAI o3*, *Deep research*, *Zochi*, and *BAGEL* has established an alternative reasoning formulation, which we designate as Interleaving Reasoning. In contrast to standard reasoning, Interleaving Reasoning is characterized by multi-turn interactions and exhibits sophisticated reasoning dynamics. This reasoning modality has empirically demonstrated superior accuracy in addressing complex problems. Consequently, we posit that Interleaving Reasoning potentially constitutes the Next-Generation Reasoning Systems for AGI.
  - Github repo: `https://github.com/Osilly/Awesome-Interleaving-Reasoning`

Accepted Paper

- **[Efficient MLLM] Dynamic-LLaVA: Efficient Multimodal Large Language Models via Dynamic Vision-language Context Sparsification.**
  - ***Wenxuan Huang***, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, Shaohui Lin.
  - Accepted to **ICLR 2025**, first author.
  - Dynamic-LLaVA is the first MLLM acceleration framework that simultaneously sparsifies both vision and language contexts while integrating inference efficiency optimization across different MLLM inference modes into a unified framework.
  - arXiv: `https://arxiv.org/abs/2412.00876`
  - Github repo: `https://github.com/Osilly/dynamic_llava`
- **[Transformer Training Acceleration] A General and Efficient Training for Transformer via Token Expansion.**
  - ***Wenxuan Huang***, Yunhang Shen, Jiao Xie, Baochang Zhang, Gaoqi He, Ke Li, Xing Sun, Shaohui Lin.
  - Accepted to **CVPR 2024**, first author.

- We proposed one plug-and-play Transformer training acceleration framework, without twisting the original training hyper-parameters, architecture, and introducing additional training strategies.
- arXiv: https://arxiv.org/abs/2404.00672
- Github repo: https://github.com/Osilly/TokenExpansion

- **[AI4Geophysic] An Intelligent First Arrival Picking Method of Microseismic Signals Based on the Small Sample Expansion.**
  - ***Wenxuan Huang***, Guanqun Sheng, Xingong Tang, Kai Ma, Jingyi Lu, Hang Sun.
  - Accepted to **IEEE Transactions on Geoscience and Remote Sensing (TGRS, Q1 Top SCI&CCF-B)**, first author.
  - We proposed one GAN to generation the microseismic samples under unsupervised conditions to expand the microseismic data having a limited number of samples. Then we use the enhanced first arrival picking network to improve the accuracy of first arrivals of low SNR microseismic signals.
  - Paper link: https://ieeexplore.ieee.org/abstract/document/10972295
  - Github repo: https://github.com/Osilly/G-LA-MSG-and-AOG-PSPNet
- **[Reasoning MLLM] Actial: Activate Spatial Reasoning Ability of Multimodal Large Language Models.**
  - Xiaoyu Zhan*, ***Wenxuan Huang*#***, Hao Sun*, Xinyu Fu, Changfeng Ma, Shaosheng Cao, Bohan Jia, Shaohui Lin, Zhenfei Yin, Lei Bai, Wanli Ouyang, Yuanqi Li, Jie Guo, Yanwen Guo.
  - Accepted to **NeurIPS 2025**, co-first author (second) & project leader.
  - We attempt to solve that current MLLMs cannot effectively capture the detailed spatial information required for robust real-world performance, especially cross-view consistency, a key requirement for accurate 3D reasoning.
  - arXiv: https://arxiv.org/abs/2511.01618
  - Github repo: https://github.com/warmsnow-sh/Actial
- **[Agentic RL/Agency task] Agentic Jigsaw Interaction Learning for Enhancing Visual Perception and Reasoning in Vision-Language Models.**
  - Yu Zeng*, ***Wenxuan Huang****, Shiting Huang*, Xikun Bao, Yukun Qi, Yiming Zhao, Qiuchen Wang, Lin Chen, Zehui Chen, Huaian Chen, Wanli Ouyang, Feng Zhao
  - Accepted to **ICLR 2026**, co-first author (second).
  - We introduce an agentic jigsaw interaction learning for enhancing visual perception and reasoning in MLLM without VQA label during training, and it demonstrates strong generalization across 9 general vision tasks.
  - arXiv: https://arxiv.org/abs/2510.01304
  - Github repo: https://github.com/yuzeng0-0/AGILE
- **[MLLM] TimeSoccer: An End-to-End Multimodal Large Language Model for Soccer Commentary Generation.**
  - Ling You*, ***Wenxuan Huang****, Xinni Xie, Xiangyi Wei, Bangyan Li, Shaohui Lin, Yang Li, Changbo Wang.
  - Accepted to **ACMMM 2025**, co-first author (second).
  - We propose the first end-to-end MLLM for soccer commentary generation, specifically designed for Single-anchor Dense Video Captioning (SDVC) in full-match soccer videos. The model jointly predicts timestamps and generates captions in a single pass, enabling global context modeling over 45-minute matches.
  - arXiv: https://arxiv.org/abs/2504.17365

Under-review Paper

- **[CNN Inference Acceleration] Filter Pruning for Efficient CNNs via Knowledge-driven Differential Filter Sampler.**
  - Shaohui Lin, ***Wenxuan Huang***, Jiao Xie, Baochang Zhang, Yunhang Shen, Zhou Yu, Jungong Han, David Doermann.
  - Submission to **IJCV** (major revision), first student author (first author is my advisor).
  - We proposed a unified CNN pruning framework directly optimized in an end-to-end manner in combination with global pruning constraint.
  - arXiv link: https://arxiv.org/abs/2307.00198
  - Github repo: https://github.com/Osilly/KDFS
- **[MLLM] LLaVA-RadZ: Can Multimodal Large Language Models Effectively Tackle Zero-shot Radiology Recognition?**
  - Bangyan Li*, ***Wenxuan Huang****✉, Zhenkun Gao, Yeqiang Wang, Yunhang Shen, Jingzhong Lin, Ling You, Yuxiang Shen, Shaohui Lin, Wanli Ouyang, Yuling Sun.
  - Preprint, co-first author (second) & corresponding author.
  - *Label in paper:* ***Wenxuan Huang*** *proposed the main idea and designed the experiments, contributing to the discussion of this paper.* ***Bangyan Li*** *refined and finalized the idea, implemented the code and experiments, and was responsible for writing the manuscript.*
  - Convert generative models to discriminative models to solve the problem of that MLLM cannot effectively tackle zero-shot radiology recognition.
  - arXiv: https://arxiv.org/abs/2503.07487
- **[Image Editing Benchmark] CompBench: Benchmarking Complex Instruction-guided Image Editing.**
  - Bohan Jia*, ***Wenxuan Huang****, Yuntian Tang*, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, Lei Bai, Wanli Ouyang, Lin Chen, Zihan Wang, Yuan Xie, Shaohui Lin.
  - Preprint (NeurIPS 2025 score 4444), co-first author (second).
  - We propose the first complex image editing benchmark.

- arXiv: https://arxiv.org/abs/2505.12200
- Github repo: https://github.com/BhJia/CompBench

- **[CLIP Inference Acceleration] CLIP-Map: Structured Matrix Adaptation for Parameter-Efficient CLIP Compression.**
  - Kangjie Zhang*, ***Wenxuan Huang****, Xin Zhou, Boxiang Zhou, Dejia Song, Yuan Xie, Baochang Zhang, Lizhuang Ma, Nemo Chen, Xu Tang, Yao Hu, Shaohui Lin.
  - Under-review, co-first author (second).
  - ***Wenxuan Huang*** *proposed the main idea and designed the experiments, contributing to the discussion of this paper.* ***Kangjie Zhang*** *refined and finalized the idea, implemented the code and experiments, and was responsible for writing the manuscript.*
  - We propose the first mapping-based CLIP compression framework that maps the parameters of CLIP to a smaller representation, thereby accelerating inference.
- **[Reasoning MLLM] Exploring End-to-End Paradigms for Visual Chinese Grammatical Error Correction.**
  - Xiaoman Wang, ***Wenxuan Huang***, Wenbiao Tao, Yike Zhao, Yaohui Liu, Yunshi Lan, Weining Qian
  - Under-review, second author.
  - We explore how to fine-turn one MLLM to solve the complex perceptron task.

In Progress

- **[Transformer Training Acceleration (journal version of ToE)] Feature Sparsification Training Paradigm: Toward Fast and Memory-Efficient General Transformer Training.**
  - In progress, may submit to **TPAMI**, first author.
  - ***Wenxuan Huang*** *et al.*
  - Try to accelerate MLLM, DiT, Segmentation, Object Detection, Classification and so on.
  - arXiv: https://arxiv.org/abs/2404.00672
  - Github repo: https://github.com/Osilly/TokenExpansion
- **[Efficient MLLM (journal version of Dynamic-LLaVA)] Dynamic-MLLM.**
  - In progress, may submit to **TPAMI**, first author.
  - ***Wenxuan Huang*** *et al.*
  - Try to accelerate the inference of Video/Audio MLLM, while combine Token Skipping and Token Merging to Achieve more efficiency in both prefill and decoding inference stages.
  - arXiv: https://arxiv.org/abs/2412.00876
  - Github repo: https://github.com/Osilly/dynamic_llava

Others

- **Weakly Supervised Semantic Segmentation via Progressive Confidence Region Expansion.**
  - Accepted to **CVPR 2025**.
  - Xiangfeng Xu, Pinyi Zhang, ***Wenxuan Huang***, Yunhang Shen, Haosheng Chen, Jingzhong Lin, Wei Li, Gaoqi He, Jiao Xie, Shaohui Lin.

GPA & Honor

- Rank 1st in undergraduate (1/56).
- Yangtze River Power Scholarship for the 2019-2020 academic year.
- National Scholarship for the 2020-2021 academic year (Top 0.2%).
- "Panshi" Scholarship for the 2023-2024 academic year.
- National Scholarship for the 2024-2025 academic year (Top 0.2%).