Hongchen Wei

(+86) 132-0180-7991 | hc_wei@whu.edu.cn | hcwei13.github.io

Education

Wuhan University — Wuhan, China
Ph.D. student, advised by Prof. Zhenzhong Chen (IIP Lab) — *Sep. 2023 – Jun. 2027*

Nanjing University of Science and Technology — Nanjing, China
M.S. student, advised by Prof. Yang Yang (PCA Lab) — *Sep. 2020 – Apr. 2023*

Xi'an Shiyou University — Xi'an, China
B.E. — *Sep. 2016 – Jun. 2020*

Research Interests

Executable virtual environments inspired by generative agents for real-world interaction simulation, dynamic evaluation, data synthesis, and RL-based post-training; Tool-augmented long-document/video multimodal agents; Multimodal model merging for perception–reasoning alignment and cross-modal knowledge transfer.

Research Experience

Microsoft Research Asia (MSRA) — Beijing, China
Research Intern, mentored by Dr. Bei Liu — *Dec. 2025 – present*

- Built executable multi-agent environments for workplace productivity scenarios, simulating role collaboration, task delegation, tool use, permission boundaries, and long-horizon state dependencies to support dynamic agent evaluation, synthetic data generation, and RL-based post-training.
- Designed **DocAtlas**, modeling long-document understanding as a mutable-state information-seeking process with Search/Read/Note/Review tools and a hierarchical evidence tree; the GPT-5.4 instantiation reaches **71.4%** on MMLongBench-Doc, surpassing the 65.8% human-expert reference, and end-to-end RL in the DocAtlas environment improves Qwen3.5-4B from a 54.4% direct-input baseline to **63.7%**.
- Built **XL-DocBench**, a fully human-verified benchmark for extra-long document understanding: 1,519 questions across 6 professional domains, contexts up to 2,303 pages, 12 reasoning labels, and evidence-page annotation by 194 human experts; designed the tree-guided synthesis pipeline and typed verification rules, and used the pipeline to synthesize **50K+** high-quality long-document, multi-document, and multi-hop QA examples to support post-training of Microsoft foundation models.

IIP Lab, Wuhan University — Wuhan, China
Doctoral Researcher — Multimodal Large Language Models — *Apr. 2023 – present*

- Identified and characterized **post-merge perception-reasoning asymmetry** in MLLMs, and proposed **PASA**, a self-alignment framework using detached-anchor asymmetric GRPO that recovers visual-grounding loss while improving reasoning without ground-truth answer labels in RL.
- Proposed **ANO**, a training-free, layer-specialized model-merging rule guided by MLLM visual-attention decay that injects reasoning into deep layers while preserving shallow visual grounding — no gradient updates or labeled multimodal reasoning data required.
- Released **FRANK-ZERO 38B**, an R1-style multimodal reasoner built by merging QwQ-32B with InternVL-38B; achieves **74.1%** on MathVista, exceeding OpenAI o1 (73.9%).
- Led a line of work on long-form multimodal understanding and generation, including visual-context-window extension for long-video understanding, open-world spatio-temporal video grounding, and long-caption generation / unsupervised prompt learning for richer visual-language generation.

PCA Lab, Nanjing University of Science and Technology — Nanjing, China
Master's Researcher — Semi-supervised Vision-Language Learning — *Sep. 2020 – Apr. 2023*

- Developed semi-supervised image-captioning methods exploiting cross-modal prediction & relation consistency.

Selected Publications

Summary: 17 papers in total, including **10 first-author** and 7 co-authored. Topics span: Multimodal LLMs & VLM Agents (8); Long-Document / Long-Video Understanding (5); Model Merging & Cross-Modal Transfer (2); Semi-Supervised Vision-Language Learning (2).

- **Hongchen Wei***, et al. "DocAtlas: Long-Document Understanding as Mutable-State Interaction." *NeurIPS 2026 (under review). Work done at MSRA.*

- **Hongchen Wei***, et al. "XL-DocBench: Benchmarking Evidence-Grounded Extra-Long Document Understanding." *NeurIPS 2026 (under review). Work done at MSRA.*
- **Hongchen Wei***, et al. "PASA: Post-Merge Perception–Reasoning Asymmetry as a Self-Alignment Signal for MLLMs." *NeurIPS 2026 (under review).*
- **Hongchen Wei***, et al. "Layer-Specialized Model Merging for Training-Free Cross-Modal Reasoning Transfer (ANO)." *NeurIPS 2026 (under review).*
- **Hongchen Wei***, et al. "LongCaption: Unlocking the Power of Long Caption Generation in Large Multimodal Models." *Under review.*

- **Hongchen Wei***, et al. "Visual Context Window Extension: A New Perspective for Long Video Understanding." *ACM MM 2025* (CCF-A).
- **Hongchen Wei***, et al. "RealVG: Unleashing MLLMs for Training-Free Spatio-Temporal Video Grounding in the Wild." *ACM MM 2025* (CCF-A).
- **Hongchen Wei***, et al. "Improving Generalization of Image Captioning with Unsupervised Prompt Learning." *ACM TOMM 2024.*
- **Hongchen Wei***, et al. "Exploiting Cross-Modal Prediction and Relation Consistency for Semi-Supervised Image Captioning." *IEEE TCYB 2023* (JCR Q1, top).
- **Hongchen Wei***, et al. "S2OSC: A Holistic Semi-Supervised Approach for Open Set Classification." *ACM TKDD 2022.*

Co-author (selected)

- "Remote Sensing Semantic Segmentation Quality Assessment based on Vision Language Model." *IEEE TGRS 2025* (JCR Q1, top).
- "See What We Cannot See: A Geo-guided Reasoning Benchmark for Object Counting under Adverse Earth Observation Conditions." *CVPR 2026.*
- "TDSAgent: A Task-Driven Sampling Agent for Long Video Question Answering." *Under review.*
- "LOP: Learning Optimal Pruning for Efficient On-Demand MLLMs Scaling." *Under review.*
- "GTC: Game-Theoretic Token Compression for Video Large Language Models." *Under review.*
- "ETC: Extreme Token Compression via Task-aware Visual Information Distillation in VLMs." *Under review.*
- "RSFAKE-1M: A Large-Scale Dataset for Detecting Diffusion-Generated Remote Sensing Forgeries." *Under review.*

Selected Open-Source Projects

FRANK-ZERO 38B | *R1-style multimodal reasoner*

- Built by merging QwQ-32B with InternVL-38B; achieves **74.1%** on MathVista, beating OpenAI o1 (73.9%); strong long-chain multimodal reasoning.

CLIP-RS | *Remote-sensing CLIP, 10M curated image–text pairs*

- Domain-adapted CLIP for remote sensing; SOTA on semantic-segmentation quality evaluation benchmarks.

RSFAKE-1M | *First large-scale remote-sensing forgery dataset*

- 500K diffusion-generated forgeries + 500K real images; detectors trained on it show substantially better generalization.

Selected Awards & Honors

1st Place, Baidu PaddlePaddle Paper-Reproduction Contest, Multimodal Track	2021
1st Prize, 3rd DIGIX Global Campus AI Algorithm Elite Contest (team leader)	2021
Top 1% (East-China region), 4th China College Computer Contest — AI Innovation	2021
2nd Prize, Baidu Cognitive AI Creativity Contest, Development Track	2022
Rank 8 / 2839, National AI Competition (AI+ Visual Feature Coding)	2023

Patent

H. Wei et al. "A Multi-View Deep-Learning Method for Sneaker Authentication based on Attention Mechanism." CN114186613A (*student first inventor*).