Haocun Ye 叶浩存

PhD student • Multimodal large models

Institute of Computing Technology, Chinese Academy of Sciences, Beijing | yehaocun23s@ict.ac.cn | github.com/yhc98002-bit

scholar.google.com/citations?user=CaPPENgAAAAJ | openreview.net/profile?id=~Haocun_Ye1 | orcid.org/0009-0007-2242-3054

Education

Institute of Computing Technology, Chinese Academy of Sciences (UCAS) | PhD student, Artificial Intelligence (master's-to-PhD track) Sep 2023 – Jun 2028 (expected)

Advisor: Prof. Yiqiang Chen | Beijing Key Laboratory of Mobile Computing and Pervasive Devices | **ICT Outstanding Master's Student Award, 2025**

Yanshan University | B.Eng. in Computer Science and Technology Sep 2019 – Jun 2023

Research Interests

Training, post-training and trustworthy evaluation of multimodal large models — from controllable generation with diffusion models, to data construction and SFT of domain LLMs, to shortcut audits of clinical-text benchmarks and the mechanisms of RLVR post-training. The recurring question: **does a benchmark gain correspond to a genuinely acquired capability?**

Publications & Patents

[1] **Haocun Ye**, Xinlong Jiang, Chenlong Gao, Bingyu Wang, Wuliang Huang, Yiqiang Chen. VersaFusion: A Versatile Diffusion-Based Framework for Fine-Grained Image Editing and Enhancement. **AAAI 2025** first author

[2] **Haocun Ye**, Xinlong Jiang, Yuanyuan Hu, et al. The Diagnosis Is Already in the Notes: A Channel-Wise Audit of Diagnostic Shortcuts in Clinical-Text Benchmarks. **EMNLP 2026** first author code ↗

[3] Tingyu Wu, Guangyu Cao, Qizhen Lan, Zhisheng Chen, Ziyan Weng, Bingkun Zhu, Miao Su, Chenglong Li, Zhengwei Xie, Huacan Wang, Sen Hu, **Haocun Ye**, Nan An, Zaoqu Liu, Ronghao Chen. KnowMeBenchV2: Evidence-Grounded Person-Centric Long-Video Understanding. **EMNLP 2026** (main conference) co-author

[4] Learning Without Looking: Image-Dependent Gains from Image-Free RLVR first author • in preparation

[5] Invention patent (granted): A Training Method for Large Models in Ophthalmology (一种眼科领域大模型训练方法)

Research & Project Experience

RLVR post-training mechanisms in vision-language models — Learning Without Looking

2026 – present

in progress • first author

- **Problem:** When RLVR raises a VLM's visual-reasoning score, did it learn new visual ability, or learn to better use visual information the pretrained model already had? Standard benchmarks cannot tell.
- **Method & contribution:** Designed a causal train × test visual-visibility matrix (real image / caption / no image / gray square); implemented GRPO training of Qwen2.5-VL-3B/7B with EasyR1 plus multi-seed automated evaluation; designed and built **FlipTrack**, a counterfactual benchmark of paired images whose answers must flip (blind accuracy is 0 by construction); validated across model families (InternVL3-9B, Gemma-3).
- **Result:** Training with no images at all still recovers ~49% (3B) → 78% (7B) of the image-dependent gain, growing with scale; a +24.4 pp gain on standard benchmarks shrinks to +2.5 pp on FlipTrack; identified a "visual grounding corrosion" effect — a visual-necessity check for multimodal RL.

Scalable high-quality ophthalmology data & trustworthy evaluation | industry-sponsored • lead of the evaluation line

2025 – present

EMNLP 2026 code

- **Problem:** Medical multimodal benchmarks treat diagnosis-redacted clinical records as neutral context, so image–text fusion gains may come from text shortcuts rather than cross-modal reasoning.
- **Method & contribution:** Proposed and implemented **CRAFT**, an audit protocol that renders each record into 20 redaction levels and channel-isolated views and measures per-channel label recoverability with frozen-encoder linear probes; ran all experiments on 5,904 multi-site ophthalmic visits (8 cohorts, 3 seeds) and reproduced the rankings with Qwen3.5-9B; led uveitis image-data governance (irrelevant-image / QR-code / screenshot filtering, UWF/OCT audit subsets).
- **Result:** Redacted records still reach 0.944 macro-AUROC (0.998 unredacted); the most aggressive dictionary scrub closes <1 pp; narrative text alone 0.978, field-presence pattern alone 0.854 — fusion gains must be reported alongside text-branch shortcut floors. Code open-sourced; being engineered into a platform audit module.

Ophthalmology LLM training: data construction & instruction tuning| industry-sponsored • core member, led the final wrap-up granted patent 2025

- **Task:** Build domain corpora and post-train Qwen2.5-14B / Qwen2-VL-7B for ophthalmic consultation and report generation; took over the project for its final wrap-up and delivery.
- **Action:** Built the data pipeline (LLM domain filtering → semantic cleaning → OCR-based QA synthesis from 212 textbooks → paraphrase augmentation → SFT formatting), yielding **9.0M** samples and 27,878 image–text pairs; ran multi-node full-parameter SFT with DeepSpeed ZeRO-3 on 2×8 A800 (bf16, global batch 1024, RDMA/NCCL, weight-aggregated checkpoints) and compared full FT vs LoRA (frozen vision encoder); built three-dataset automatic evaluation and a four-dimension LLM-judge protocol.
- **Result:** Report-generation CIDEr 0.0004 → 0.557 (DeepEyeNet); LLM-judge average 2.6 → 4.1; the method was granted an invention patent.

Diffusion-based fine-grained image editing & crop-growth simulation| MADA platform • principal contributor AAAI 2025 2024 – 2025

- **Problem:** T2I models struggle to change content (e.g., a crop's growth stage) while preserving subject identity and background; drag-style editors need per-image fine-tuning and long preparation.
- **Method & contribution:** Proposed and implemented **VersaFusion**, a tuning-free dual-branch (guidance / generation) editing framework: proximal negative-prompt inversion (PNI) recovers the latent, KV memory injection and classifier guidance provide cross-branch control, and an IP-Adapter-style image prompt enables reference-driven editing; deployed it as the crop-growth-stage simulation feature of the MADA platform.
- **Result:** 5.23 s preparation (DragDiffusion 63.56 s), feature-alignment MSE 11.45 (vs 16.99), FID 32.98 (better than DragonDiffusion).

Skills

Languages: Python, C++
Model training: PyTorch; LLM SFT (full / LoRA), multi-node training (DeepSpeed ZeRO-3), multimodal RLVR (GRPO / EasyR1)
Models & frameworks: Qwen2/2.5-VL, Qwen2.5, InternVL, Gemma; Stable Diffusion inversion and editing
Data & evaluation: LLM-based data filtering / cleaning / synthesis pipelines; benchmark construction and trustworthy evaluation (counterfactual evaluation, shortcut & leakage audits)