Lin Fu

19857109661 · 748544794@qq.com · Hangzhou
Current Status: Student

Education

Zhejiang University M.Eng. in Electronic Information Engineering, College of Integrated Circuits, Research Area: Multimodal Large Models 2024.09 – 2027.06

Zhejiang University B.S. in Information and Computational Science, School of Mathematical Sciences, GPA: 3.52 / 4.00, Recommended Admission 2020.09 – 2024.07

Internship Experience

Alibaba DAMO Academy, Research Intern 2026.02 – 2026.06

- **GUI Agent World Model Benchmark:** Contributed to the construction of a benchmark for mobile GUI world models by formulating user interactions as action-conditioned UI state prediction tasks. Designed two evaluation pipelines, including offline trajectory replay and online agent-loop evaluation, to assess contextual consistency, state persistence, action responsiveness, and task progress in multi-step interactions, providing reproducible benchmark support for GUI Agent simulation training and reliability evaluation. Submitted to EMNLP 2026.

Huawei Technologies Co., Ltd. Terminal BG, Algorithm Engineer Intern 2025.04 – 2025.08

- **AIGC-Assisted Creative Illustration Tool Demo:** Implemented image editing capabilities including outpainting, repainting, object erasing, automatic matting, color enhancement, and super-resolution based on large models and generative AI. Completed the full loop from frontend interaction, including upload, selection, prompt/parameter configuration, to backend inference invocation and result return. Adapted and encapsulated input/output formats, mask representations, resolutions, and color spaces across different models, and developed reusable invocation interfaces and demo workflows.
- **Automatic Illustration Workflow Optimization (Cover Cropping Badcase Mitigation):** Addressed core issues in cover generation such as poor keyframe selection and subject cropping by organizing badcase categories and trigger scenarios, then formulating optimization strategies. Introduced large models as decision modules to automatically determine key steps such as whether matting is needed and whether cropping is suitable / to what extent cropping is feasible. Integrated these decisions with existing rule-based and detection modules to upgrade cover cropping and subject preservation strategies, improving subject completeness and reducing online badcases.

Research Experience

VideoKR: Knowledge- and Reasoning-Intensive Video Understanding Data and Post-Training 2025.07 – 2026.02

- **Motivation:** Proposed VideoKR to address the lack of structured domain knowledge and high-quality reasoning supervision in existing video understanding datasets, which limits post-training for scientific scenarios. Built a knowledge-driven video understanding data framework.
- **Data and Experiments:** Constructed a course-level knowledge base covering 82 disciplines and 6K knowledge points, and collected 145K YouTube (CC) videos through knowledge-point-driven retrieval. Expanded and generated three types of samples, VIDR / KNOWVID / KNOWVIDR, obtaining 201K SFT data and 114K RL data. Conducted SFT+GRPO training based on Qwen2.5-VL-7B and Qwen3-VL-8B.
- **Evaluation and Analysis:** On benchmarks including Video-MME, MVBench, VideoMMMU, and SciVideoBench, SFT+RL achieved consistent gains over SFT-only and RL-only settings. The model showed improved ability to integrate visual evidence with domain knowledge for reasoning.
- **Outcome: Lin Fu***, Zheyuan Yang*, Yang Wang, Tingyu Song, Arman Cohan, Yilun Zhao. *VideoKR: Towards Knowledge- and Reasoning-Intensive Video Understanding*. **ICML 2026 Spotlight**.

TEXOCR: Document OCR Benchmark and Model Training for Compilable Page-to-LaTeX Reconstruction 2025.09 – 2026.01

- **Motivation:** Proposed TEXOCR to address the limitations of existing document OCR methods, which mainly focus on PDF-to-text or Markdown conversion and lack evaluation benchmarks and training methods for compilable page-level LaTeX reconstruction.
- **Data and Experiments:** Built TEXOCR-Bench with 3,000 expert-annotated documents, as well as TEXOCR-Train based on arXiv papers from 2022–2025, containing 52K papers and 384K page image–LaTeX aligned samples. Conducted two-stage SFT+RLVR training based on Qwen3-VL-2B, using verifiable rewards for text, formulas, tables, citations, and compilability.
- **Evaluation and Analysis:** Systematically evaluated 23 OCR/MLLM models and found that existing methods still suffer from clear limitations in formula transcription, table recovery, citation consistency, and compilation success rate. TEXOCR achieved leading performance among open-source models, and RLVR brought stable improvements over SFT-only training.

- **Outcome:** Chengye Wang*, **Lin Fu**, Zexi Kuang, Yilun Zhao. *TEXOCR: Benchmarking and Advancing Document OCR Models for Compilable Page-to-LaTeX Reconstruction*. **ACL 26**.

Sci-VBench: A Video Generation Benchmark for Scientific Reasoning 2025.12 – 2026.04

- **Motivation:** Proposed Sci-VBench to address the fact that existing video generation evaluations mainly emphasize visual quality and text alignment, while lacking assessment of scientific knowledge constraints, multi-step causal reasoning, and temporal consistency.
- **Data and Experiments:** Constructed 1,500 expert-annotated samples covering four major domains, natural sciences, medical and health sciences, humanities and social sciences, and engineering, across 78 disciplines. Each sample is paired with a high-level reference guide and a 1–5 anchored rubric. Systematically evaluated 11 frontier closed-source and open-source text-to-video models.
- **Evaluation and Analysis:** Proposed an evaluation framework combining automatic perceptual metrics with rubric-conditioned MLLM-as-Judge. Verified that introducing evaluation specifications improves the agreement of non-expert and MLLM ratings with expert ratings. Results show that current models still have clear shortcomings in scientific causal correctness and spatiotemporal consistency.
- **Outcome:** Diandian Zhang*, Tingyu Song*, **Lin Fu***, Zheyuan Yang, Yilun Zhao. *Sci-VBench: Benchmarking Knowledge-Grounded Scientific Reasoning in Video Generation*. **Submitted to COLM 2026**.

Technical Skills

Programming and Frameworks: Python, PyTorch, LLaMA-Factory, Verl

Training and Fine-Tuning: Familiar with fine-tuning methods for multimodal large models, including SFT, LoRA, and QLoRA

Post-Training and Alignment: Familiar with common post-training and alignment paradigms, including DPO, RLHF, and RLVR

Engineering and Data: Familiar with machine learning and deep learning fundamentals as well as common training workflows; experienced in large-scale dataset synthesis