

Shuo Gao

Haidian, Beijing, China | gaoshuo@ruc.edu.cn | +86 138 6335 3710

EDUCATION

Renmin University of China, M.S. in Computer Science Sept 2024 – Present

- GPA: 3.66/4.0
- **Study Area:** Vision-Language Models, Multimodal Large Language Models, Computer Vision
- **Honors and Awards:** Academic Scholarship

Central University of Finance and Economy, B.S. in Computer Science Sept 2020 – June 2024

- GPA: 3.89/4.0 (top 10%)
- **Relevant Coursework:** Data Structure, Computer Network, Machine Learning and Pattern Recognition
- **Honors and Awards:** Second Prize Scholarships (2 times), Academic Scholarships (2 times)

RESEARCH EXPERIENCES

Student Lab Member, AIMC Lab, Renmin University of China – Beijing, China Sept 2024 – June 2026

- Advanced Multi-modal Large Language Models (MLLMs) for Open-World Compositional Zero-Shot Learning (CZSL), exploring the innate reasoning capabilities of agents like Gemini and LLaVA.
- Proposed a baseline framework that decomposes complex compositional tasks into sequential attribute-object recognition steps, leveraging Chain-of-Thought (CoT) to outperform 10 specialized CZSL models without targeted training.
- Developed a Text-driven Multi-modal Prototype Optimization approach for ophthalmic multi-disease recognition, utilizing LLMs to generate hierarchical lesion descriptions (labels, lesions, details) for fine-grained semantic alignment.

Research Intern (remote), AIM Lab, University of Amsterdam – Amsterdam, The Netherlands Jan 2025 – June 2025

Advisor: Yingjun Du, Cees Snoek

- Proposed a test-time prompt adaptation framework for Vision-Language Models (e.g., CLIP) to enhance generalization on unseen domains.
- Designed an uncertainty-aware optimization strategy that estimates token-level epistemic uncertainty via Monte Carlo Dropout and leverages a self-gating mechanism to dynamically scale gradient updates, preventing overfitting during adaptation.
- Achieved state-of-the-art performance across 15 diverse datasets, demonstrating significant improvements in both cross-dataset and domain generalization tasks.
- Authored a research paper detailing the methodology and experimental results.

Research Intern, 360 AI Research – Beijing, China Oct 2025 – June 2026

Hosts: Jincheng Li, Chunyu Xie

- Proposed a unified MLLM framework that bridges the gap between semantic comprehension and precise localization by decoupling visual encoders and task-specific heads within a single Large Language Model.
- Designed a dual-encoder architecture integrating InternViT for multimodal understanding and OWLv2-ViT for detection, effectively resolving the feature conflict between high-level reasoning and fine-grained perception.
- Devised a three-stage training strategy (Alignment → Head Warm-up → Mixed Instruction Tuning) utilizing over 1.5M mixed data samples (image-text pairs, LMM instructions, and COCO), ensuring balanced optimization of LMLoss and DetLoss.
- Authored a research paper demonstrating our methodology and experimental results.

Research Intern (remote), RAIR Lab, UNC-Chapel Hill – Chapel Hill, NC, USA Feb 2026 – Present

Advisor: Zhongzheng (Jason) Ren

- Developed FramePack-U, a training-free long video understanding framework that compresses arbitrarily long

videos into a fixed-size visual representation with constant $O(1)$ memory complexity.

- Designed a relevance- and diversity-aware frame sampling strategy that uses text-frame semantic similarity and visual redundancy filtering to prioritize instruction-relevant yet complementary video frames.
- Proposed a hierarchical resolution packing scheme that allocates higher spatial resolution to important frames and progressively compresses less critical frames under a fixed token budget.
- Conducted extensive experiments on LongVideoBench, MLVU, and Video-MME, showing consistent improvements over vanilla sampling, token reduction, and keyframe selection baselines across multiple VLM backbones and memory budgets.
- Authored a research paper detailing the methodology and experimental results.

Intern, LinkerBot – Beijing, China

July 2026 – Present

Hosts: Langyong Shi, Xiaoxiao Hu

- Working on Embodied Foundation Models and World Action Models for Dexterous Manipulations.

RESEARCH PAPERS

Elephant in the Fridge: Constant-Memory Frame Packing for Long Video Understanding May 2026

Shuo Gao, Chenhao Zheng, Zhongzheng Ren

In *Neural Information Processing Systems (NeurIPS)*, 2026

Do Multi-modal LLMs possess Compositional Zero-Shot Recognition Capabilities? March 2025

Yipeng Xu, **Shuo Gao**, Gang Yang, Xinyue Ju, Yici Zhang, Ruixiang Zhao, Xirong Li

In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026

Text-Driven Multi-Modal Prototype Optimization for Multi-Disease Recognition on Retinal Fundus Images Aug 2025

Shuo Gao, Gang Yang, Yici Zhang, Jingjing Xu

In *International Conference on Bioinformatics and Biomedicine (BIBM)*, 2025

Not All Tokens Matter Equally: Uncertainty-Guided Test Time Prompt Adaptation for Vision-Language Models May 2025

Shuo Gao, Yingjun Du, Cees G. Snoek

Under Review

TransDet: Unifying Multimodal Understanding and Precise Object Detection in MLLMs Jan 2026

Shuo Gao, Jincheng Li, Chunyu Xie, Dawei Leng, Yuhui Yin

Under Review

Creating a 4D Sketch from a Single Image May 2026

Jingyu Liu, **Shuo Gao**, Zijie Xin, Ruixiang Zhao, Bangxiang Lan, Qingwei Shen, Yuchuan Deng, Qijie Wei, Yuhan Fu, Xirong Li

Under Review