MINGSHENG LI

Fudan University | H: +86 199 4623 3989 | limc22@m.fudan.edu.cn
Homepage

About Me

I am currently a **final-year** Master's student in **Artificial Intelligence** at Fudan University, advised by Prof. Tao Chen. I am also fortunate to work with Dr. Hongyuan Zhu from A*STAR, Singapore, Dr. Gang Yu, Dr. Xin Chen, and Dr. Chi Zhang from Tencent, and Dr. Bo Zhang from Shanghai AI Lab. Before this, I received my bachelor's degree in Information Engineering from Fudan University in 2022.

I work in the fields of deep learning and computer vision, focusing on **large models**, **multi-modal learning**, and **AIGC**. My research pursues to develop robust and scalable general-purpose AI systems to solve complex problems. Besides, I was awarded the **National Scholarship** as the **1st-ranked** student out of 244.

Research Interests

AIGC, Multi-modal Learning, Large Models, Multi-Agents, Embodied AI and Continual Learning.

Education

Masters in Artificial Intelligence Sep. 2022 - Jun. 2025
Fudan University. Shanghai, China
Bachelor in Information Engineering Sep. 2018 - Jun. 2022
Fudan University. Shanghai, China

Experience

Shanghai AI Lab. Feb. 2024 - Oct. 2024
Research Intern at the Center for the Frontiers of AI, working on multi-modal large models.

Tencent Tech. Oct. 2024 - Present
Research Intern in the Foundation Model Group, working on multi-agent systems for scene generation.

Selected Publications

- GeoX: Geometric Problem Solving Through Unified Formalized Vision-Language Pre-training.
  Mingsheng Li, Renqiu Xia, Hancheng Ye, Wenjie Wu, Hongbin Zhou, Jiakang Yuan, et al.
  [Under Review by ICLR 2025 | Score: 8866 (Top 2%) | OpenReview]
  [Summary]: Introducing formalized pre-training and a multi-modal large model for automated GPS.
- Chimera: Improving Generalist Model with Domain-Specific Experts.
  Tianshuo Peng, **Mingsheng Li**$^{\text{co-first}}$, Renrui Zhang, Song Mao, Conghui He, Aojun Zhou, Xiangyu Yue, et al.
  [Under Review by CVPR 2025 | OpenReview]
  [Summary]: Enhancing LMMs for specific tasks via cost-effective generalist-specialist collaboration.
- 3DET-Mamba: State Space Model for End-to-End 3D Object Detection.
  Mingsheng Li, Jiakang Yuan, Sijin Chen, Lin Zhang, Anyu Zhu, Xin Chen, Tao Chen.
  [NeurIPS 2024 | project]
  [Summary]: An end-to-end 3D detection model with Mamba.
- M3DBench: Let's Instruct Large Models with Multi-modal 3D Prompts.
  Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, Tao Chen.
  [ECCV 2024 | project | arxiv | github]
  [Summary]: A large-scale dataset instructing 3D VLMs with interleaved multi-modal prompts.
- WI3D: Weakly Incremental 3D Detection via Vision Foundation Models.
  Mingsheng Li, Sijin Chen, Shengji Tang, Hongyuan Zhu, Yanyan Fang, Xin Chen, et al.
  [T-MM 2024 | paper]
  [Summary]: Introducing new categories to pre-trained 3D detectors with 2D foundation models.
- Lightweight Model Pre-training via Language Guided Knowledge Distillation.
  Mingsheng Li, Lin Zhang, Mingzhen Zhu, Zilong Huang, Gang Yu, Tao Chen, Jiayuan Fan.
  [T-MM 2024 | paper | arxiv | github]

[Summary]: Using language to refine knowledge distillation between teacher and student.

- Vote2Cap-DETR++: Decoupling Localization and Describing for End-to-End 3D Dense Captioning.
  Sijin Chen, Hongyuan Zhu, **Mingsheng Li**, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, Tao Chen.
  [T-PAMI 2024| paper | arxiv | github]
  [Summary]: Decoupled feature extraction for localizing and describing objects in 3D scenes.

- LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning.
  Sijin Chen, Xin Chen, Chi Zhang, **Mingsheng Li**, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, Tao Chen.
  [CVPR 2024 | project | paper | github]
  [Summary]: 3D-LLMs respond to visual and text interactions in complex 3D scenes.

- EgoExoLLM: Dual-level Memory-enhanced Ego-Exo Learning with Large Language Models.
  Lei Gao, Jiakang Yuan, Bizhe Bai, **Mingsheng Li**, Bo Zhang, Tao Chen
  [Under Review by CVPR 2025 | OpenReview]
  [Summary]: A general model focusing on ego-exo video learning with global and local memory banks.

- DocGenome: A Large Benchmark for Multi-Modal Language Models in Real-World Academic Document Understanding.
  Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, Shiyang Feng, **Mingsheng Li**, Bin Wang, et al.
  [Under Review by ICLR 2025 | Rating: 8653 | OpenReview]
  [Summary]: A structured document dataset covering 500K scientific documents from 165 disciplines.

Recent Projects

- **AIGC** May. 2023 - Present
  Vision Language Models. Presented **GeoX**, a multi-modal large model for automated geometric problem solving with unified formalized pre-training and a GS-Former module, submitted to ICLR 2025 (Score: 8866, ranking in the top 2% globally). Put forward **Chimera**, a scalable pipeline improving existing Multi-modal Large Models on domain-specific tasks via Generalist-Specialist Collaboration, achieving SOTA performance in chart, table, math, and document domains, submitted to CVPR 2025.

  Embodied Foundation Model. Presented **LL3DA**, a multi-modal 3D large model responding to both text and visual interactions with complex 3D scenes, accepted by CVPR 2024. Unveiled **M3DBench**, a comprehensive 3D-language dataset covering 327k interleaved multi-modal instructions for 10 tasks, and developed an interactive embodied assistant, accepted by ECCV 2024. Proposed **EgoExoLLM**, an embodied generalist with dual memory banks that learns from ego-exo videos, capturing both global and detailed information, submitted to CVPR 2025.

  3D Scene Generation. Developing multi-agent AI systems to generate diverse, coherent, and interactive 3D scenes, with the capability to edit and arrange objects within customized scenes.

- **Multi-modal Understanding and Reasoning.** Jan. 2023 - May. 2024
  Presented **3DET-Mamba**, a state-space architecture for indoor 3D object detection, integrating local- and global- feature learning with a novel decoder, accepted by NeurIPS 2024. Proposed **Vote2Cap-DETR++**, which decouples learnable queries into localization and caption queries to capture task-specific features for object caption generation, accepted by T-PAMI 2024. Put forth **WI3D**, a label-efficient approach for continual learning that utilizes cost-effective 2D foundation models to introduce new categories to 3D detectors, accepted by T-MM 2024.

- **Lightweight Models Pre-training.** Aug. 2022 - Jan. 2023
  Proposed **LGD**, focusing on language-guided lightweight model pre-training, accepted by T-MM 2024.

Scholarships and Awards

National Scholarship (rank 1/244).	2024
Second Prize of Graduate Academic Scholarship.	2023
Outstanding Graduate of Fudan University.	2022
Second Prize of China Undergraduate Mathematical Contest in Modeling, Shanghai.	2020
STEM (Science, Technology, Engineering, and Mathematics) Scholarship.	2020
First Prize of Chinese Mathematics Competitions, Shanghai (Top 20).	2019

Skills

Programming:	Python (primary), C, Matlab, SQL, etc.
Tools:	PyTorch, Visual Studio, MeshLab, Jupyter Notebook, etc