BOSHEN XU

website: xuboshen.github.io
(+86) 13881861005
boshenx@ruc.edu.cn

EDUCATION

Renmin University of China *Sep. 2023 - now*
Ph.D in Computer Science Advisor: Qin Jin
University of Electronic Science and Technology of China *Sep. 2019 - Jun. 2023*
B.S in School of Computer Science and Engineering Advisor: Qin Jin
GPA: 3.97/4.00(90.41/100), ranked 2/65

RESEARCH INTERESTS

computer vision and egocentric AI.

EXPERIENCE

MiLM Plus, Xiaomi Inc. *April. 2025 - now*

- **Long-form video understanding, e.g., temporal video grounding.** **Research Intern**

PROJECTS

Multi-Modal Large Language Model *Mar. 2025- now*

- **Post-Train LVLM for Temporal Video Grounding (arxiv 2025)**. We introduce Time-R1 framework, TimeRFT training, and TVGBench evaluation, to advance the field of using LVLM for temporal video grounding. Time-R1 achieves state-of-the-art performance on TVG using only 2.5K general data for RL fine-tuning, with improved performance on 4 video QA benchmarks.
- **Task-Specific LVLM for Temporal Video Grounding (CVPRW 2025)**. We introduce TimeZero, a reasoning-guided LVLM designed for the temporal video grounding task. This task requires precisely localizing relevant video segments within long videos based on a given language query. TimeZero tackles this challenge by training the MLLM through RL with CoT, enabling the model to reason about video-language relationships.

2D Human-Object Interaction Understanding *Nov. 2022- now*

- **3D-Aware Egocentric Video-Language Pretraining (arxiv 2025)**. Humans develop spatial awareness through perceiving and interacting in the 3D world. To compensate 3D spatial understanding of egocentric video-language models, we propose EgoDTM, developed by multi-modal pertaining on millions of egocentric videos with depths and texts.
- **Egocentric Vision-Language Pretraining (ICLR 2025)**. We discover that current Egocentric Vision-Language Models (EgoVLM) are lack of open-vocabulary EgoHOI recognition ability on our constructed benchmark EgoHOIBench. To address this issue, we propose EgoNCE++, an asymmetric contrastive learning pretraining objective to pretrain the EgoVLMs.
- **Egocentric Hand-Object Interaction (ACM MM 2023)**. The egocentric data is expensive to fetch. We propose a visual prompt-based method to effectively pretrain model with third-person videos to achieve viewpoint adaptation and viewpoint generalization to Ego-HOI recognition.
- **Human-Object Interaction Detection (CVPR 2023)** The HOI task has long been plagued by the lack of data and supervision due to the large number of possible combinations of verbs and objects in real life. Models trained on limited data can only predict HOIs within a set of fixed

categories. We therefore propose an open-category HOI pretraining model with specially designed proxy tasks that can well generalize to novel interaction categories.

Object Relation Understanding in 3D Space. *Aug. 2023- Nov. 2023*

- **Autonomous 3D Object Assembly (3DV 2025)** We addresses the combinatorial explosion challenge, where the number of possible combinations rises beyond exponentially with increasing parts, by leveraging weak constraints from assembly sequences, effectively reducing the solution space's complexity.

PUBLICATIONS

1. Ye Wang*, Ziheng Wang*, **Boshen Xu*‡**, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, Qin Jin. Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding. In arxiv, 2025
2. **Boshen Xu**, Yuting Mei, Xinbi Liu, Sipeng Zheng, Qin Jin. EgoDTM: Towards 3D-Aware Egocentric Video-Language Pretraining. In arxiv, 2025
3. Ye Wang*, **Boshen Xu***, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, Qin Jin. TimeZero: Temporal Video Grounding with Reasoning-Guided LVLM. In CVPR ViSCAL Workshop, 2025
4. **Boshen Xu**, Ziheng Wang*, Yang Du*, Zhinan Song, Sipeng Zheng, Qin Jin. Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions? In ICLR, 2025
5. **Boshen Xu**, Sipeng Zheng, and Qin Jin. SPAFormer: Sequential Part Assembly with Transformers. In 3DV, 2025
6. Liangyu Chen, Zihao Yue, **Boshen Xu**, Qin Jin. Unveiling Visual Biases in Audio-Visual Localization Benchmarks. In ECCV AVGenL Workshop, 2024
7. **Boshen Xu**, Sipeng Zheng, and Qin Jin. POV: Prompt-Oriented View-Agnostic Learning for Egocentric Hand-Object Interaction in the Multi-View World. In ACM MM, 2023
8. Sipeng Zheng, **Boshen Xu**, and Qin Jin. Open-Category Human-Object Interaction Pre-Training via Language Modeling Framework. In CVPR, 2023

SERVICE

- Conference Reviewer: CVPR, NeurIPS, ICLR, ACM MM, ACL, ACCV.
- Journal Reviewer: TOMM.

AWARDS AND SCHOLARSHIPS

Outstanding Innovative Talents Cultivation Funded Programs of RUC.	*2024*
First Class Scholarship for Ph.D Students.	*2023-2024*
Provincial Outstanding Graduate.	*2023*
National Second Prize in China Undergraduate Mathematical Contest in Modeling.	*2021*
Tencent Special Scholarship.	*2021*
National Scholarship for Undergraduates.	*2020*

June 5, 2025