朱晓萌 Xiaomeng Zhu

xzhubr@connect.ust.hk 130-8666-5226 Google Scholar

研究兴趣

多模态大模型 • 具身智能 • 主动感知与决策 • 视频理解与预测 • 场景图与时序推理 • 任务导向的物体功能推理（Affordance Reasoning）

教育背景

香港科技大学 (HKUST) 2024.09 – 至今

计算机科学与工程学院，博士研究生，导师：林方真教授 GPA: 3.85/4.00

中国科学院大学 (UCAS) 2021.09 – 2024.06

中国科学院自动化研究所 (CASIA)，模式识别专业，多模态人工智能系统全国重点实验室

硕士（推荐免试），导师：李书晓研究员 GPA: 3.83/4.00

电子科技大学 (UESTC) 2017.09 – 2021.06

自动化工程学院，自动化专业，雄鹰班

本科 GPA: 3.99/4.00，排名：1/138（前 1%）

发表论文

** 表示共同第一作者 / equal contribution*

主动响应智能体

1. **ProAct: A Benchmark and Multimodal Framework for Structure-Aware Proactive Response**
   Xiaomeng Zhu, F. Zhu, W. Zhou, Y. Tian, Z. Hu, Y. Huang, Y. Guo, X. Wu, Z. Zhang, et al.
   ICML 2026.
2. A Computable Game-Theoretic Framework for Multi-Agent Theory of Mind
   F. Zhu, Y. Pan, **Xiaomeng Zhu**, F. Lin
   AAAI 2025 Workshop on Multi-Agent AI.

场景理解

3. **OOTSM: A Decoupled Linguistic Framework for Effective Scene Graph Anticipation**
   Xiaomeng Zhu, C. Wang, H. Wang, X. Liu, F. Lin
   IEEE Transactions on Multimedia (TMM) (一审 AQE, Minor Revision).
4. **Afford-X: Generalizable and Slim Affordance Reasoning for Task-Oriented Manipu-**

lation
Xiaomeng Zhu, Y. Li, L. Cui, P. Li, H. Gao, Y. Zhu, H. Zhao
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (Major Revision).

5. Listening with the Eyes: Benchmarking Egocentric Co-Speech Grounding across Space and Time
W. Zhou, X. Xiong, Z. Hu, **Xiaomeng Zhu**, C. Zhao, H. Dong, Z. Zhang, M. Tang, J. Wang
arXiv:2603.07966, 2026.

6. Plan Right, Then Plan Tight: Symbolic RL for Efficient Embodied Reasoning
X. Shi, **Xiaomeng Zhu**, Y. Huang, Y. Tian, Y. Guo, Z. Sun, L. Yin, Y. Zhou
ACL ARR 2026 May (in submission to EMNLP).

其他

7. Aeroengine Performance Prediction Using a Physical-Embedded Data-Driven Method
T. Mo, S. Dai, A. Fu, **Xiaomeng Zhu**, S. Li
IEEE Transactions on Aerospace and Electronic Systems (TAES), 2025.

8. MGML: Momentum Group Meta-Learning for Few-Shot Image Classification
Xiaomeng Zhu, S. Li
Neurocomputing, Vol. 514, pp. 351–361, 2022.

研究工作

研究方向聚焦**视频场景理解**，涵盖主动感知、场景图预测、任务导向 affordance reasoning 及其在具身智能数据生产中的应用。

ProAct：面向主动响应的基准与多模态框架 ICML 2026

现有具身智能系统以被动指令执行为主，限制了机器人的自主性和协作效率。ProAct 旨在训练智能体从连续视频中持续监测环境、自主判断介入时机并选择行动。主要贡献包括：（1）构建 ProAct-75 基准，覆盖辅助、维修和安全监控三大场景，包含 75 个任务、5,383 段视频和 91,581 条步骤级标注，并提供了包含 AND/OR 依赖与并行线程的显式任务图以支持结构化决策；（2）提出 ProAct-Helper 框架，基于 MLLM 并以 Hierarchical Binding Module（HBM）增强跨层级状态感知，利用熵驱动启发式搜索在任务图约束下选择 proactive action，支持并行线程执行而非仅跟随人类下一步意图。实验表明，ProAct-Helper 在触发检测 mF1 上超越强闭源模型 6.21 个百分点，平均节省 0.25 步，并行动作率提升 15.58 个百分点。

面向世界模型训练的视频场景理解与数据生产管线 Tencent Robotics X, 专利已提交

在视频场景理解处理框架基础上构建数据标注管线，将原始视频切分为有意义的 action segment，精准定位 action segment 首尾帧，生成结构化训练样本，解决 World Model 训练中大规模视频转化为训练数据的生产瓶颈。技术细节涉密暂不公布，见后续技术报告。该管线已实际投入数据生产。

OOTSM：基于 LLM 解耦的场景图预测 IEEE TMM, Minor Revision

Scene Graph Anticipation 需要根据历史视频预测未来帧中的对象集合及关系。传统端到端视觉方法在长尾关系、未来对象动态变化和语义一致性上存在明显不足。本工作将 SGA 拆为两阶段：GOA 先利用 LLM 推断未来可能出现或保留的对象，OORA 在预测对象集合上进行对象导向的关系推理，结合 relation BCE loss 和 transition 约束建模关系的时间动态变化。在 Action Genome 上完成对象集

合动态分析、关系预测评估、消融实验和效率分析。

Afford-X：任务导向的物体功能推理 IEEE TPAMI, Major Revision

传统 open-vocabulary detection 擅于根据 noun query 找物体，但任务导向的 affordance reasoning 需要回答” 哪个对象能完成给定的动作或功能”。例如任务为 clean bottle with something 时，模型应选择 napkin/cloth 等工具，而非被名词 bottle 误导。利用 GPT-4 生成 task-object pair 并通过人工 4 级量表验证，构建 COCO-Tasks/Aff 和 LVIS-Aff 数据集。在轻量 VLM 框架上引入 VA 模块抑制名词偏置、BF 模块增强跨模态 mid-level alignment，通过 noun-pronoun 蒸馏提升 unseen 任务泛化能力。与 Detect-before-Reasoning、Reason-before-Detection 等多类 MLLM pipeline 对比，证明专用轻量模型在效率与稳定性上的优势。

实习经历

腾讯 Robotics X 2025.06 – 至今

研究实习生

从事具身智能世界模型训练数据生产管线（视频标注与质量检查）的设计与实现。

北京大学，人工智能学院 2023.05 – 2025.02

算法实习生，指导老师：Yixin Zhu, Hao Zhao

负责开放世界任务导向目标检测项目 Afford-X 的实验验证与论文撰写；搭建相关算法平台并参与机器人仿真环境下的算法验证。

获奖与荣誉

- 国家奖学金（本科，2 次）；中国科学院大学研究生学业奖学金
- 全国大学生智能汽车竞赛全国一等奖（西部赛区一等奖、省一等奖）
- 四川省优秀毕业生；唐立新奖学金