自动化学报
ACTA AUTOMATICA SINICA

大模型驱动的云边协同多智能体具身学习

王涓 赵大伟 朱琪 刘华平

Cloud-edge Collaborative Multi-agent Embodied Learning Driven by Large Models

WANG Juan, ZHAO Da-Wei, ZHU Qi, LIU Hua-Ping

在线阅读 View online: https://doi.org/10.16383/j.aas.c260096

您可能感兴趣的其他文章

面向具身操作的视觉语言动作模型综述

Survey of Vision-Language-Action Models for Embodied Manipulation

自动化学报. 2026, 52(1): 18−51 https://doi.org/10.16383/j.aas.c250394

基于大模型的具身智能系统综述

Embodied Intelligence Systems Based on Large Models: A Survey

自动化学报. 2025, 51(1): 1−19 https://doi.org/10.16383/j.aas.c240542

面向多智能体协作的注意力意图与交流学习方法

Attentional Intention and Communication for Multi-agent Learning

自动化学报. 2023, 49(11): 2311−2325 https://doi.org/10.16383/j.aas.c210430

具身智能自主无人系统技术

Embodied Intelligence Autonomous Unmanned Systems Technology

自动化学报. 2025, 51(4): 762−777 https://doi.org/10.16383/j.aas.c240456

端边云协同的PID整定智能系统

PID Tuning Intelligent System Based on End-edge-cloud Collaboration

自动化学报. 2023, 49(3): 514−527 https://doi.org/10.16383/j.aas.c230055

工业垂域具身智控大模型构建新范式探索

An Exploration of a New Paradigm for Constructing Industrial Domain-specific Embodied Intelligent Control Large Models

自动化学报. 2025, 51(11): 2454−2472 https://doi.org/10.16383/j.aas.c250247

大模型驱动的云边协同多智能体具身学习

王 涓[1] 赵大伟[2] 朱 琪[2] 刘华平[1]

摘 要 具身学习通过智能体与环境的主动交互与数据采集, 结合模型迭代更新, 实现自主智能. 然而, 随着环境复杂度和任务规模增加, 传统单智能体方法面临数据采集效率低、样本利用效率低、训练效率低等瓶颈, 严重制约了系统可扩展性. 针对上述问题, 提出一种云边协同的多智能体具身学习框架, 充分利用视觉−语言模型的高级推理能力, 在无需额外训练的前提下实现多智能体高效在线探索与协同数据采集. 具体地, 各智能体通过视觉−语言模型对观测数据及感知信息进行解析, 构建融合语义与空间状态的好奇心地图, 以指导短期目标选择, 实现语义驱动的高价值区域利用与空间驱动的未知区域探索的协同推进; 在短期探索任务完成后, 智能体能够基于全局空间状态与共享语义信息开展长期探索规划, 以保障探索的全面性与战略性. 所有智能体完成全场景探索后, 边缘端利用采集数据进行本地模型训练, 并将更新的参数和少量数据上传至云端, 由云端进行多源知识聚合, 生成全局优化模型. 实验结果表明, 所提方法显著优于基线方法, 为大规模复杂环境下多机器人自主智能学习提供新的有效范式.

关键词 具身学习; 多智能体; 云边协同; 视觉−语言模型; 自主探索

引用格式 王涓, 赵大伟, 朱琪, 刘华平. 大模型驱动的云边协同多智能体具身学习. 自动化学报, 2026, **52**(10): 2184−2199

DOI 10.16383/j.aas.c260096 **CSTR** 32138.14.j.aas.c260096

Cloud-edge Collaborative Multi-agent Embodied Learning Driven by Large Models

WANG Juan[1] ZHAO Da-Wei[2] ZHU Qi[2] LIU Hua-Ping[1]

Abstract Embodied learning enables autonomous intelligence through active interaction and data collection between agents and the environment, combined with iterative model updates. However, as the complexity of the environment and the scale of tasks increase, conventional single-agent methods face critical bottlenecks, including low data collection efficiency, low sample utilization efficiency, and low training efficiency, which severely limit system scalability. To address these challenges, we propose a cloud-edge collaborative multi-agent embodied learning framework that leverages the high-level reasoning capabilities of vision-language model (VLM) to achieve multi-agent efficient online exploration and collaborative data collection without additional training. Specifically, each agent uses a VLM to interpret observation data and perception information, constructing curiosity maps that integrate semantic and spatial states to guide short-term goal selection, thereby enabling synergistic advancement of semantic-driven utilization of high-value regions and spatial-driven exploration of unknown regions; Upon completing short-term exploration tasks, agents perform long-term exploration planning based on global spatial states and shared semantic information to ensure comprehensive and strategic exploration. After all agents complete full-scene exploration, edge agents train local models using the collected data, and the updated parameters with a small amount of data are uploaded to the cloud for multi-source knowledge aggregation to generate a globally optimized model. Experimental results demonstrate that the proposed method significantly outperforms baseline methods, providing an effective paradigm for multi-robot autonomous intelligence learning in large-scale complex environments.

Keywords embodied learning; multi-agent; cloud-edge collaboration; vision-language model; autonomous exploration

Citation Wang Juan, Zhao Da-Wei, Zhu Qi, Liu Hua-Ping. Cloud-edge collaborative multi-agent embodied learning driven by large models. *Acta Automatica Sinica*, 2026, **52**(10): 2184−2199

具身学习强调智能体通过与真实或模拟环境的主动交互获取数据, 并在感知、决策与行动的闭环中不断更新自身模型, 从而实现对复杂环境的自适应理解与自主智能行为[1]. 相较于传统依赖静态数据集的离线学习范式, 具身学习能够充分利用交互带来的信息增益, 在未知或动态环境中展现出更强

收稿日期 2026-01-31 录用日期 2026-04-12
Manuscript received January 31, 2026; accepted April 12, 2026
国家自然科学基金 (62120106005), 教育部基础学科和交叉学科突破计划 (JYB2025XDXM109), 国家重点研发计划 (2025YFF0522500) 资助
Supported by National Natural Science Foundation of China (62120106005), Fundamental and Interdisciplinary Disciplines Breakthrough Program of the Ministry of Education of China (JYB2025XDXM109), and National Key R&D Program of China (2025YFF0522500)
本文责任编委 温广辉
Recommended by Associate Editor WEN Guang-Hui
1. 清华大学计算机科学与技术系 北京 100084 2. 国防科技创新研究院 北京 100073
1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084 2. Defense Innovation Institute, Beijing 100073

的泛化能力与适应性, 已成为机器人感知、导航与操作等领域的重要研究方向.

然而, 随着应用场景逐渐向大规模、复杂环境拓展, 单智能体具身学习范式在探索效率和系统可扩展性方面面临显著瓶颈. 一方面, 单个智能体在有限时间与计算资源约束下, 难以高效完成对复杂环境中关键状态的探索; 另一方面, 现有方法多依赖强化学习进行探索与决策策略训练, 往往需要大量交互样本, 并对环境变化高度敏感, 难以在未知场景中保持稳定性. 这些问题在一定程度上限制了具身学习方法在开放环境中的实用性.

为提升系统的探索效率与鲁棒性, 多智能体系统 (multi-agent systems) 被引入具身智能研究. 现有多智能体研究大多从任务驱动的角度出发设计智能体的协同行为. 例如, 操作任务重点研究物理层的动作协调, 探索任务则主要聚焦于空间覆盖率 (spatial coverage) 的提升. 在此类范式下, 感知模型的输出通常作为环境状态的一部分用于策略决策, 但感知模型自身的学习状态或学习需求并未被显式建模为交互策略的优化目标, 智能体的交互行为仍主要围绕任务或探索指标进行设计. 相比之下, 本文将感知模型的学习需求显式纳入多智能体协同决策过程, 使交互行为能够根据感知模型的反馈进行调整, 以获取更有利于感知模型训练的数据, 从而形成感知–交互–学习的闭环机制.

在大规模或复杂环境中, 仅依靠边缘端多智能体仍存在显著局限: 一方面, 边缘端单个智能体可获取的全局信息有限, 导致探索策略局部最优; 另一方面, 边缘端智能体在处理复杂推理或高层规划时的计算能力受限. 与此同时, 大语言模型 (large language model, LLM) 和视觉–语言模型 (vision-language model, VLM) 的快速发展为具身学习提供了新的发展思路. 大模型往往具备强大的推理能力, 能够在无需任务特定训练的情况下理解开放词汇语义、空间关系以及高层任务目标, 为构建通用、可迁移的具身智能系统奠定基础. 已有研究表明, 大模型在机器人高层规划、空间推理和任务泛化方面展现出显著潜力. 为此, 本文引入云边协同机制: 云端统一部署大模型, 利用其强大的语义理解与推理能力进行全局策略规划, 并将决策指导下发至边缘端智能体; 边缘端智能体则在局部环境中执行策略并采集数据, 同时将更新后的本地模型和少量数据上传至云端, 实现边缘采集与云端整合的闭环协作. 该机制既解决了局部信息有限的问题, 又充分发挥大模型在高层规划和跨场景泛化的能力, 为多智能体具身学习提供高效、可扩展的支撑.

如图 1 所示, 本文提出一种云边协同的多智能体具身学习框架, 充分利用视觉–语言模型的通用推理能力, 在无需额外训练的前提下实现多智能体高效在线协同探索采集数据. 具体地, 在探索阶段, 云端大模型基于边缘端语义–空间状态信息生成全局策略以统一指导边缘端智能体探索. 在学习阶段, 各智能体基于探索过程中采集的数据在边缘端进行本地模型更新, 并将更新后的模型参数和少量样本上传至云端, 以减少通信开销; 云端通过多源知识聚合生成统一综合模型. 在部署阶段, 该综合模型可应用于新环境, 通过在线感知与决策验证其在未知场景下的泛化能力与感知性能.

本文的主要研究工作包括:

1) 提出一种大模型驱动的云边协同多智能体具身学习框架, 实现探索与学习一体化的多智能体协同机制, 提升了系统在大规模复杂环境下的可扩展性;

2) 基于视觉–语言模型构建融合语义–空间信息的环境表征, 并结合跨智能体的共享语义信息,

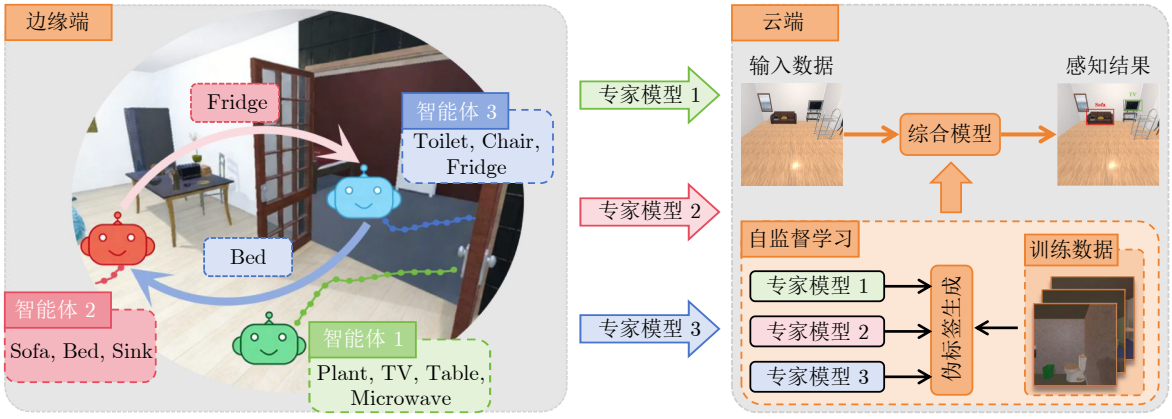


图 1　云边协同具身学习框架

Fig. 1　An embodied learning framework for cloud-edge collaboration

引导多智能体在无需额外训练的条件下开展高效的在线协同探索;

3) 设计边缘端本地学习与云端多源知识聚合相结合的协同学习流程, 在降低通信开销的同时显著提升模型的学习效率与跨场景泛化能力.

1 相关工作

1.1 具身学习

具身学习的核心在于感知、决策与行动的紧密耦合, 强调智能体通过主动交互实现自主学习[2]. 早期研究主要围绕具身学习在机器人视觉感知中的应用展开. 研究者通过引入语义好奇奖励机制, 度量多视角下感知预测的不一致性, 引导智能体主动探索高信息量区域[3]. 有学者利用多视角熵与视图散度联合样本选择与视图规划, 借助超像素级区域选择以显著降低数据标注成本[4]. 此外, 文献 [5] 通过在探索未知区域与采集高价值样本间进行动态权衡, 实现了语义分割模型的在线迭代更新; 在此基础上, 文献 [6] 将该架构拓展至终身学习场景, 赋予智能体在跨环境探索中持续积累与迁移视觉知识的能力.

为进一步优化探索效率, 信息路径规划 (informative path planning, IPP) 被引入具身学习框架. 现有方法利用感知模型的不确定性驱动路径规划, 实现了语义分割模型的主动域自适应[7]. 文献 [8] 将不确定性显式建模为地图信息增益以引导探索与数据采集. 进一步地, 有学者通过半监督主动学习与自适应路径规划策略[9], 以及基于语义分布不确定性与不一致性奖励的难样本挖掘机制[10], 实现未知环境中样本利用效率和模型性能的提升.

研究者们引入自监督与自改进机制摆脱对人工标注数据的依赖. 通过在探索阶段引导智能体主动改变观测视角获取跨视图观测, 并结合多视角一致性约束生成高质量伪标签 (pseudo-label, PL), 实现了模型的持续自主演进[11−14]. 文献 [15] 通过聚类、推理与建图的联合优化, 实现了开放世界语义场景理解, 为开放场景中未知类别学习提供了支撑.

为增强模型视觉表征能力, 有学者将运动信息融入视觉表征学习, 通过联合优化强化学习策略与逆动力学预测提升下游任务的泛化能力[16]. 文献 [17] 提出好奇心驱动的交互式环境表征框架, 利用快速概念映射降低少样本学习对标注的依赖. 文献 [18] 在推理阶段通过交互收集多帧信息, 并借助自监督损失持续更新模型. 在此基础上, 有学者进一步引入高层语义建模以提升智能体的环境理解能力, 通过结合外部知识与场景图结构[19], 或利用具身探索动态构建语义场景表示[20], 使智能体能够实现复杂语义推理与环境理解.

大语言模型和视觉−语言模型为具身学习提供了新范式. 早期研究主要聚焦于模块化的策略, 将冻结参数的预训练多模态大模型作为高级视觉特征提取器或推理组件[21−22]. 为构建通用基础模型, 有学者联合互联网数据、机器人行为数据, 通过预训练、微调等技术建模多模态信息[23−29]. 然而, 大模型往往面临算力消耗高和延迟高等瓶颈. 为此, 文献 [30] 部署低频运行的大模型与轻量级底层处理网络以降低计算开销. 此外, 一些新型具身世界模型的出现, 使智能体能够通过预测未来三维空间的状态演变来推演自身动作可能产生的结果[31−33], 有效弥合了视觉表征与真实场景之间的鸿沟, 提升了智能体在未知开放环境中的泛化能力.

尽管上述方法取得了显著进展, 但通常依赖任务特定的不确定性建模, 且聚焦于单智能体具身学习范式, 在环境规模扩大或任务复杂度提升时, 往往面临探索效率不足、训练周期过长及复杂开放环境中泛化能力不足等问题. 本文引入视觉−语言模型作为通用推理引擎, 在无需额外训练的条件下构建语义−空间融合的多智能体协同探索机制, 为多智能体协同具身学习提供了新的解决思路.

1.2 多智能体协同

为缓解大规模复杂环境下单智能体在探索效率和可扩展性方面的局限性, 有学者尝试引入多智能体系统, 通过并行感知、协作决策以及跨机器人的信息共享, 提高自主系统在未知环境中的探索效率和任务覆盖能力.

早期研究主要依赖几何启发式方法, 通常计算地图覆盖率或信息增益实现任务分配[34−35]. 针对通信受限场景, 文献 [36] 通过共享由新增顶点与边构成的轻量化拓扑信息实现高效协作探索. 进一步地, 有学者开始关注几何建模误差与语义不确定性的联合优化[37]. 文献 [38] 通过去中心化的协作框架实时共享语义地图以增强导航与协作能力. 文献 [39] 通过意图信息模块和注意力机制改善智能体间的协作与交流.

随着多智能体强化学习 (multi-agent reinforcement learning, MARL) 的兴起, 集中式训练与分布式执行 (centralized training with decentralized execution, CTDE) 范式被广泛应用于协同决策与控制任务. 文献 [40] 利用全局状态或联合观测支持协作策略的有效学习. 文献 [41] 提出 MADDPG 框

架, 通过集中式评论家引入状态与动作信息, 缓解了多智能体环境的非平稳性问题. 文献 [42] 提出的 MAPPO 则进一步融合 PPO (proximal policy optimization) 算法, 通过集中式价值函数提升训练稳定性. 为进一步优化长期协作一致性与泛化能力, 研究者提出基于动力学模型的角色发现与多样性增强框架, 利用互信息最大化与对比学习促进角色特异性行为[43]. 文献 [44] 引入共享循环记忆机制增强长期路径规划中的协作泛化. 文献 [45] 提出了基于深度强化学习的多机协同空战决策流程框架, 为多机协同空战智能决策提供了有效途径. 然而, 基于多智能体强化学习的方法通常依赖大量交互样本和长期离线训练, 对奖励设计和环境分布高度敏感, 实际应用时仍面临样本效率低、泛化能力受限等问题.

云边协同为多智能体系统提供了低通信开销与隐私保护的新方案. 文献 [46] 利用云–边–端协同控制机制缓解规模扩大导致的协同交互难题. 此外, 联邦学习 (federated learning, FL) 作为最具代表性的方法之一, 让各智能体在本地训练模型并仅交换模型参数或梯度, 而无需共享原始数据, 从而实现跨个体协同学习[47]. 进一步地, 研究者尝试将联邦学习与 MARL 融合, 通过联邦强化学习 (Fed-MARL) 框架, 实现零样本泛化规划、任务卸载与资源分配联合优化, 以及边缘设备分散式控制策略学习, 均兼顾隐私保护与网络可扩展性[48−50].

与此同时, 大模型技术的发展为多智能体协同提供了新机遇. 文献 [51] 利用大模型将高层自然语言指令转换为多机器人任务计划. 文献 [52] 借助大模型整合结构化空间信息与非结构化语言线索. 文献 [53] 利用大模型的推理能力实现任务规划、在线适应与人在环验证. 然而, 上述方法主要聚焦于任务层面的协同, 与感知层面的耦合仍然较弱.

尽管面向任务的多智能体协同研究已取得长足进展, 但底层感知与语义理解的协同学习方面的研究仍相对不足. 本文提出一种云边协同的多智能体具身学习框架, 通过引入视觉–语言模型进行语义–空间联合推理, 进而实现多源知识聚合, 为大规模复杂环境下的多智能体自主学习提供了一种新的可扩展方向.

2 问题定义

具身学习的核心目标是为智能体在未知环境 $\mathcal{E}$ 中设计主动探索与数据采集策略, 使其在有限交互预算下生成对感知模型最具信息价值的交互数据, 并通过模型更新提升其在未知环境中的泛化性能.

在云边协同多智能体具身学习场景下, 系统由 N 个部署于边缘端的智能体组成, 记为 $\{\mathcal{A}^i\}_{i=1}^N$. 每个智能体 $\mathcal{A}^i$ 搭配一个初始感知模型 Φ_{θ^i}, 定义在其专属语义类别集合 $\mathcal{C}^i = \{c_1^i, c_2^i, \cdots\}$ 上, 其中 θ^i 为模型参数.

在探索阶段, 云端大模型生成高层策略 Π_H, 用于规划多智能体在时间步 k 时的探索目标 $H_k = \{H_k^1, H_k^2, \cdots, H_k^N\}$, 每个智能体在边缘端执行对应的低层控制策略 π_L^i, 实现目标 H_k^i 以采集数据. 具体地, 智能体在探索过程中维护以下信息:

1) 局部探索地图 $M_{exp}^{i,k}$: 记录智能体 i 已探索区域的空间信息;

2) 局部语义地图 $M_{sem}^{i,k}$: 基于智能体 i 的观测图像 I_k^i, 记录局部区域内语义信息;

3) 共享语义信息 info_k^i: 记录截至时间步 k 时, 其他智能体向智能体 i 传递的包含语义类别 $\mathcal{C}^i$ 的目标在全局坐标系下的位置信息:

$$\text{info}_k^i = \cup_{t=1}^k \{\text{info}_{t,\,\mathcal{C}^i}^{j\to i} \mid j \neq i\} \tag{1}$$

其中, $\text{info}_{t,\,\mathcal{C}^i}^{j\to i}$ 表示时间步 t 时, 智能体 j 向智能体 i 传递的包含语义类别 $\mathcal{C}^i$ 的语义信息.

高层目标 H_k^i 包含短期探索目标 s_k^i 与长期探索目标 g_k^i, 通过高层策略 Π_H 生成:

1) 短期探索目标 s_k^i: 表示局部观测图像 I_k^i 中感兴趣目标在世界坐标系下的投影位置, 旨在引导智能体在局部范围内高效采集信息和执行导航任务. 短期探索目标的生成依赖好奇心地图:

$$M_{cur}^{i,k} = f(M_{sem}^{i,k},\ M_{exp}^{i,k}) \tag{2}$$

其中, 函数 $f(\cdot)$ 为融合函数. 在此基础上, 短期目标 s_k^i 由函数 $f_s(\cdot)$ 构建:

$$s_k^i = f_s(M_{cur}^{i,k},\ I_k^i,\ D_k^i,\ p_k^i,\ \Phi_{\theta^i}) \tag{3}$$

其中, D_k^i 为智能体观测到的深度图像; p_k^i 为智能体当前位姿; 函数 $f_s(\cdot)$ 用于评估局部区域内的语义信息和空间新颖性, 以实现对局部区域的有效探索.

2) 长期探索目标 g_k^i: 表示局部探索地图 $M_{exp}^{i,k}$ 上未探索且信息量丰富的区域, 旨在规划全局范围内的探索路径, 以探索高价值区域. 通过函数 $f_g(\cdot)$ 构建长期探索目标 g_k^i:

$$g_k^i = f_g(M_{exp}^{i,k},\ \text{info}_k^i) \tag{4}$$

其中, 函数 $f_g(\cdot)$ 用于评估全局区域内的语义信息和空间新颖性, 以实现对全局区域的有效探索.

对于高层目标 H_k^i, 边缘端低层控制策略 π_L^i 基于当前位姿 p_k^i 生成连续控制动作:

$$\boldsymbol{a}^i = \pi_L^i(H_k^i,\ p_k^i) = \{a_1^i, \cdots, a_t^i, \cdots, a_{T_{\max}}^i\} \tag{5}$$

其中, a_t^i 为第 t 步时的动作; $T_{\max}$ 为最大探索时间步. 交互过程中将采集训练数据集 $\mathcal{D}_{train}^i$ 用于模型知识更新.

在学习阶段, 各边缘端智能体基于其本地采集的数据更新模型:

$$\Phi_{\theta^{i\prime}} = \mathcal{U}^i(\Phi_{\theta^i},\ \mathcal{D}_{train}^i) \tag{6}$$

其中, $\mathcal{U}^i$ 为模型更新函数; $\theta^{i\prime}$ 为优化后的模型参数, 并将更新后的模型和少量数据上传至云端. 云端对来自不同智能体的本地模型进行整合, 得到融合多源知识的全局模型:

$$\Phi_{\theta^*} = \mathcal{G}(\{\Phi_{\theta^{i\prime}},\ \mathcal{D}_{cloud}^i\}_{i=1}^N) \tag{7}$$

其中, $\mathcal{D}_{cloud}^i$ 为第 i 个智能体上传至云端的数据; $\mathcal{G}(\cdot)$ 表示云端模型整合函数.

在部署阶段, 融合后的模型 Φ_{θ^*} 直接用于未知测试环境 $\mathcal{E}_{test}$ 中的性能评估. 最终目标是在测试数据集 $\mathcal{D}_{test}$ 上最大化模型 Φ_{θ^*} 的感知性能 $P(\cdot)$, 从而实现多智能体协同探索与学习的整体最优.

$$\Pi_H^* = \arg\max_{\Pi_H}\ \mathrm{E}_{\{\boldsymbol{a}^i\ \sim\ \pi_L^i(H_k^i,\ p_k^i)\}_{i=1}^N}\left[P(\Phi_{\theta^*},\ \mathcal{D}_{test})\right] \tag{8}$$

3 云边协同多智能体具身学习框架

本节介绍所提出的云边协同多智能体具身学习框架. 该框架以视觉−语言模型为核心决策单元, 在云端统一规划多智能体的高层探索目标, 并在边缘端完成低层动作执行与数据采集, 从而实现高效、可扩展的协同探索与感知模型学习. 整体流程如图 2 所示.

3.1 探索阶段

探索阶段的目标是在有限交互预算下, 引导多智能体采集对感知模型学习最具信息价值的数据. 为此, 云端基于视觉−语言模型生成高层策略 Π_H, 各边缘端智能体根据接收到的高层目标执行低层控制策略并与环境交互. 探索阶段的整体流程如图 3 所示.

3.1.1 高层策略

高层策略 Π_H 规划由云端视觉−语言模型完成, 其作用是在语义信息利用与空间探索之间进行权衡, 为智能体 i 在时间步 k 生成高层目标 $H_k^i = \{s_k^i,\ g_k^i\}$. 高层目标的生成综合考虑智能体的局部观测、历史探索状态以及跨智能体间共享的语义信息.

短期目标. 视觉−语言模型通过预测短期目标引导智能体在局部区域内优先访问语义信息密集或

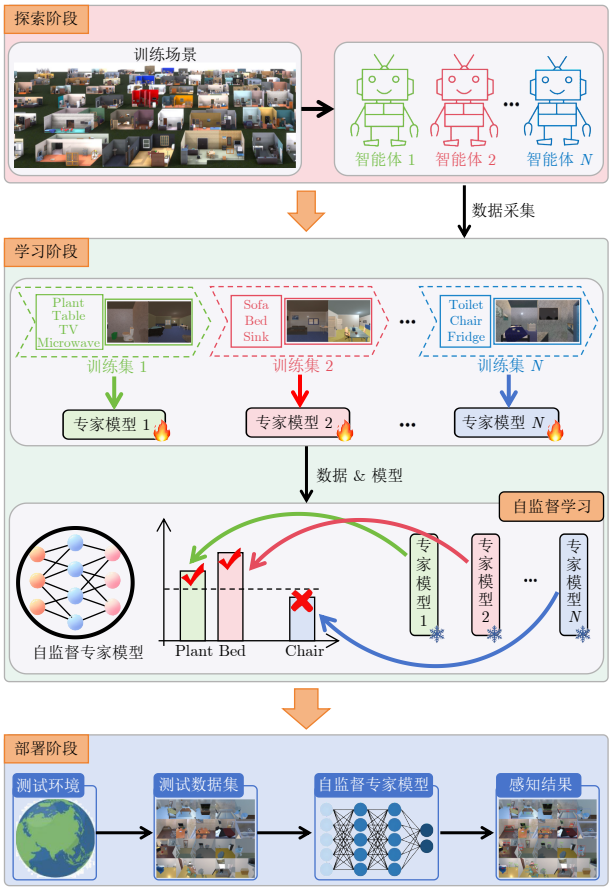


图 2 云边协同具身学习框架流程图

Fig. 2 A cloud-edge collaborative embodied learning framework flowchart

潜在信息增益较高的区域. 具体地, 智能体 i 在每一时间步维护并更新以下环境表征:

- 局部探索地图. 探索地图 $M_{exp}^{i,k}$ 记录了智能体自初始时刻至时间步 k 的空间探索状态, 用于刻画环境的空间新颖性. 具体地, 在时间步 k 时, 智能体利用映射模块, 基于当前深度观测 D_k^i 和位姿信息 p_k^i 构建世界坐标系下的点云表示 pcd_k^i, 并将全局点云投影至俯视平面, 从而获得当前可感知区域. 基于与智能体位置的距离, 感知区域被细分为以下三类区域:

1) 已访问区域. 与智能体距离小于阈值 D_{vis} 的区域被视作已访问区域, 并在探索图上用深灰色进行标识;

2) 已探索区域. 智能体已通过传感器观测到, 但与智能体距离大于阈值 D_{vis} 的区域被视作已探索区域, 并在探索图上用浅灰色进行标识;

3) 未探索区域. 历史与当前未观测到的区域被视作未探索区域, 并在探索图上用白色进行标识. 未探索区域能代表潜在的空间新颖性.

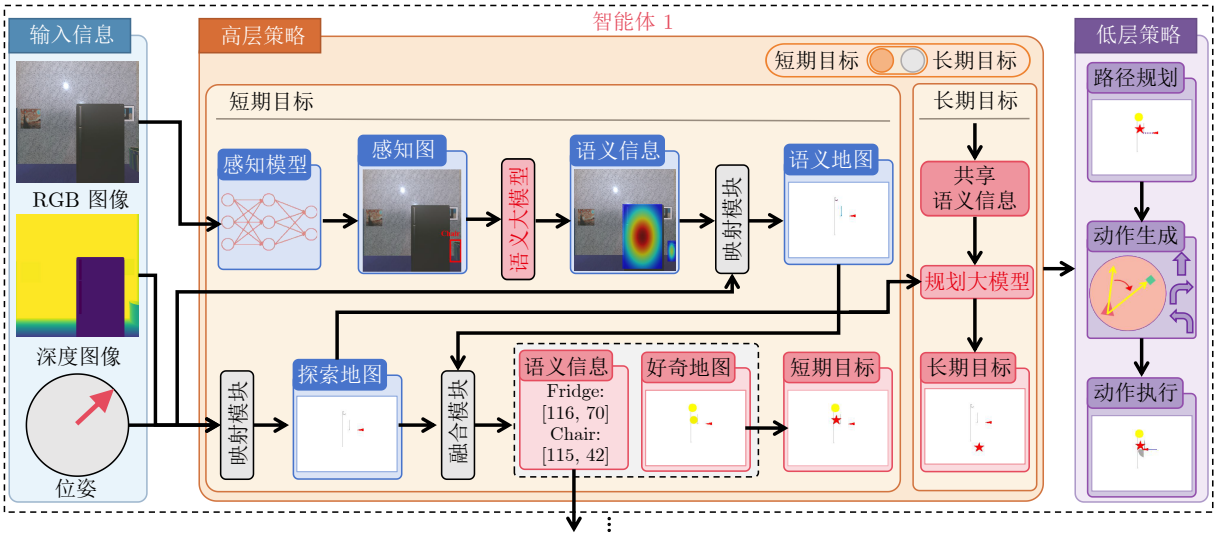


图 3　探索阶段流程图

Fig. 3　Flowchart of the exploration stage

基于地图区域与智能体的距离进行细粒度区域划分, 能清晰地展现智能体运动轨迹与环境中不同区域的探索状态, 为后续空间新颖性评估和目标规划提供基础.

● 局部语义地图. 语义地图 $M_{sem}^{i,k}$ 用于记录环境中不同空间位置所蕴含的语义信息价值. 具体地, 引入思维链技术, 利用视觉−语言模型推理生成语义信息. 在时间步 k 时, 视觉−语言模型接收两类关键信息作为输入: 当前时刻采集的 RGB 观测 I_k^i; 感知模型 Φ_{θ^i} 生成的感知结果 $y_k^i = \Phi_{\theta^i}(I_k^i)$, 其中, y_k^i 包含目标物体的边界框、语义类别标签以及置信度分数.

如图 4 所示, 视觉−语言模型对上述输入执行逐步推理过程, 以生成用于更新语义地图的高层语义信息. 该推理过程包括以下三个步骤:

1) 感知结果分析. 视觉−语言模型首先对当前场景进行语义理解, 并将其与感知模型输出 y_k^i 进行交叉验证. 通过该过程, 视觉−语言模型能够识别潜在的感知不确定性或偏差, 例如漏检目标或低置信度检测结果, 从而为后续语义区域筛选与评分提供依据.

2) 显著语义区域生成. 在感知结果分析的基础上, 视觉−语言模型利用其蕴含的知识与语义关联能力, 生成一组具有潜在探索价值的显著语义区域, 包括感知失败区域、具有导航意义的可通行区域 (如通向新空间的门) 以及在语义上与目标类别高

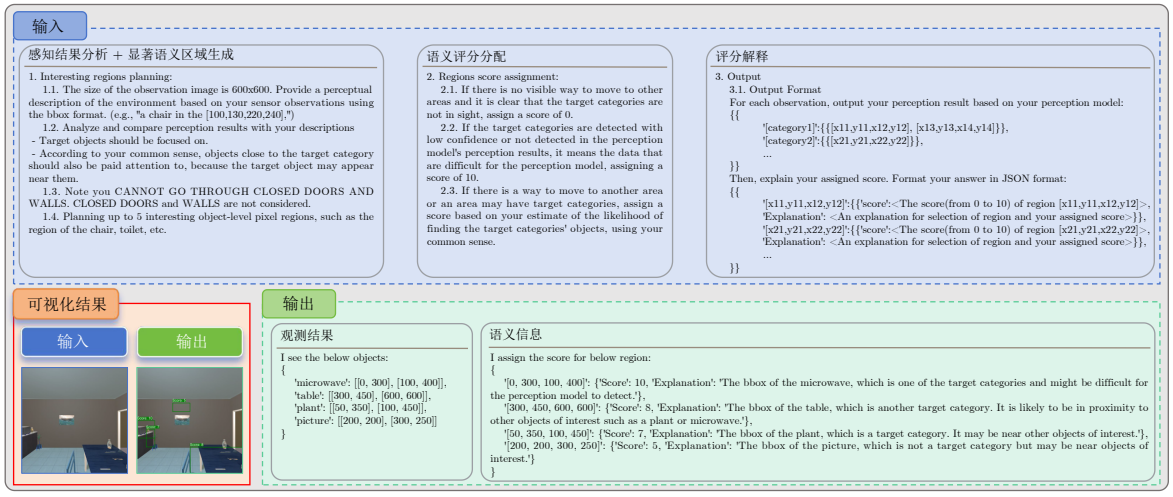


图 4　短期目标推理输入输出示例

Fig. 4　Examples of input and output for short-term goal reasoning

度相关的区域, 从而推理出候选语义区域 $\boldsymbol{r}_k^i = \{r_k^{i,1}, r_k^{i,2}, \cdots, r_k^{i,N_k}\}$, 其中, N_k 为候选区域数量. 每个候选区域用图像坐标系下的边界框坐标标识.

3) 语义评分分配. 视觉−语言模型对每个候选语义区域进行系统化评估, 并为其分配一个数值化的语义评分 $\mathcal{S}_k^i = \{S_k^{i,1}, S_k^{i,2}, \cdots, S_k^{i,N_k}\}$, 以标明该区域对感知模型学习的潜在价值. 评分过程遵循以下规则: a) 感知失败的区域被赋予最高评分; b) 具有潜在导航意义或语义关联丰富的区域获得较高评分; c) 不可通行区域 (如墙) 则被赋值为零分. 对于每个评分, 视觉−语言模型将同时生成对应的文本化推理解释, 以增强评分结果的可解释性.

带有语义评分的候选区域将用于更新局部语义地图 $M_{sem}^{i,k}$. 对于每个候选区域 $\boldsymbol{r}_k^i$, 首先利用深度观测 D_k^i 将区域内的像素点投影为智能体局部坐标系下的三维点云; 再结合当前位姿信息 p_k^i, 将局部三维点云转换至全局坐标系下, 并通过投影映射到俯视视角的二维语义地图中得到语义区域; 计算该区域的中心点坐标, 并维护其对应的语义评分 $\mathcal{S}_k^i$ 作为该候选区域的语义好奇值.

将当前时间步投影得到的语义信息与上一时间步的语义地图 $M_{sem}^{i,k-1}$ 通过聚合操作 $\mathrm{Agg}(\cdot)$ 进行融合, 生成更新后的语义地图:

$$M_{sem}^{i,k} = \mathrm{Agg}\left(M_{sem}^{i,k-1}, \mathrm{Proj}(\boldsymbol{r}_k^i, \mathcal{S}_k^i, p_k^i, D_k^i)\right) \quad (9)$$

其中, Proj 为投影操作, 将语义信息投影到俯视图上; 聚合操作 $\mathrm{Agg}(\cdot)$ 采用逐单元取最大值的方式, 以确保语义地图在时间上始终保留最显著的语义线索, 从而形成具有长期记忆能力的全局语义表示.

短期目标 s_k^i 用于引导智能体在局部范围内优先访问语义价值高且尚未充分探索的区域. 为此, 将局部语义地图 $M_{sem}^{i,k}$ 与局部探索地图 $M_{exp}^{i,k}$ 进行融合, 构建好奇心地图, 以同时考虑语义显著性与空间新颖性.

$$M_{cur}^{i,k}(x, y) = f\left(M_{sem}^{i,k}(x, y), M_{exp}^{i,k}(x, y)\right) = M_{sem}^{i,k}(x, y) \cdot \mathbb{I}\left(M_{exp}^{i,k}(x, y) \in \mathcal{U}_{exp}\right) \quad (10)$$

其中, (x, y) 为局部二维地图上的空间坐标; $\mathcal{U}_{exp}$ 表示未探索区域的空间坐标集合; $\mathbb{I}(\cdot)$ 为指示函数, 当且仅当目标区域属于未探索区域时, 该指示函数的值为 1, 否则为 0. 该机制能够有效过滤掉已充分探索区域的语义得分, 从而引导智能体优先前往具有高语义价值的新颖区域.

高层策略在好奇心地图上选择具有最大值的候选位置作为短期目标 s_k^i, 从而实现对高价值语义区域的优先探索.

长期目标. 将在全局尺度上分析空间拓扑结构, 评估不同未探索区域的潜在战略价值, 并预测最可能揭示新房间结构或高价值新物体的像素坐标位置, 作为长期子目标 g_k^i.

在探索过程中, 智能体 i 将其感知结果与视觉−语言模型推理得到的高层语义信息进行结构化编码, 并传递给负责对应语义类别的其他智能体. 具体地, 解析感知模型和视觉−语言模型预测的目标物体的语义类别与像素坐标, 并投影到俯视图中得到目标物体在全局坐标系下的空间位置, 再将物体语义类别 $\mathcal{C}^j$ 与空间位置编码为语义信息 $\mathrm{info}_{k,\mathcal{C}^j}^{i\to j}$ 传递给智能体 j. 智能体 i 可利用共享语义信息 info_k^i 补充环境语义线索. 该共享信息在探索阶段持续累积, 并作为长期目标规划的重要先验.

当短期目标已被充分执行, 即局部好奇心地图中不存在显著候选区域时, 高层策略将自适应切换生成长期探索目标 g_k^i, 以实现对全局未覆盖区域的战略性探索. 此时, 视觉−语言模型基于局部探索地图 $M_{exp}^{i,k}$ 记录的空间探索状态, 以及智能体间共享的语义信息 info_k^i, 推理当前尚未探索且高价值的区域作为长期目标 g_k^i, 以实现跨智能体协同决策, 提高探索效率和数据采集价值. 图 5 为长期目标推理示例.

最终, 短期目标 s_k^i 与长期目标 g_k^i 通过自适应切换机制共同指导智能体的运动决策. 短期目标确保在局部语义丰富区域进行高价值探索; 长期目标则引导智能体访问未探索空间, 实现全局环境的有效探索.

3.1.2 低层策略

低层策略 π_L^i 负责将高层生成的目标 H_k^i 转化为智能体可执行的动作序列 $\{a_k^i, a_{k+1}^i, \cdots\}$, 以实现从当前位置到目标位置的导航.

路径规划. 为实现当前位置到目标位置的最短路径规划, 利用广度优先搜索 (breadth-first search, BFS) 算法计算从当前位置 pos_k^i 到目标 H_k^i 的最短路径 $\mathcal{P}_k^i$. 其中, $\mathcal{P}_k^i = \{p_{k,1}^i, p_{k,2}^i, \cdots\}$ 表示沿最短路径依次经过的空间位置序列.

动作生成. 根据路径 $\mathcal{P}_k^i$, 智能体在每一时间步根据当前位姿 p_k^i 与路径中下一个位置 $p_{k,next}^i = (x_{k,next}^i, y_{k,next}^i)$ 的相对角度 $\Delta\theta_k^i$ 决定动作类型:

$$\Delta\theta_k^i = \mathrm{atan2}(y_{k,next}^i - y_k^i, x_{k,next}^i - x_k^i) - \theta_k^i \quad (11)$$

其中, (x_k^i, y_k^i) 为当前位置坐标; θ_k^i 为智能体当前朝向, 基于相对角度值选择动作:

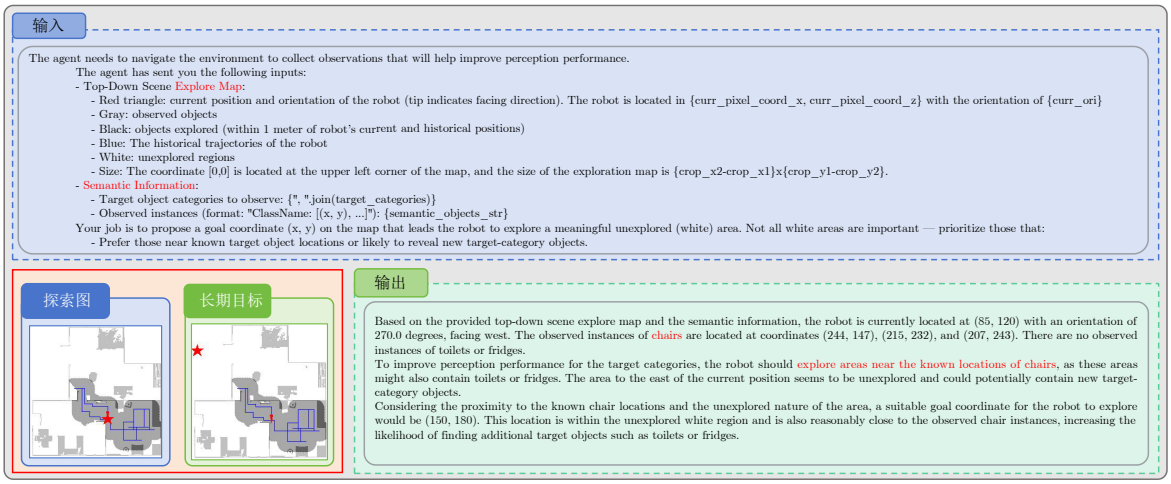


图 5　长期目标推理输入输出示例

Fig. 5　Examples of input and output for long-term goal reasoning

$$a_k^i = \begin{cases} \text{左转}, & \Delta\theta_k^i > 30^\circ \\ \text{右转}, & \Delta\theta_k^i < -30^\circ \\ \text{直行}, & -30^\circ \le \Delta\theta_k^i \le 30^\circ \end{cases} \tag{12}$$

智能体沿路径依次执行动作, 实时更新位姿, 并在每步重新评估相对角度差以确定下一个动作, 直至到达目标位置. 该策略保证了低层动作执行的连续性和局部可行性, 同时与高层目标规划紧密耦合, 实现语义驱动与空间驱动的协同探索.

3.2　学习阶段

在学习阶段, 通过云边协同的模型更新与知识整合机制, 在降低通信开销的同时, 实现多智能体采集知识的高效融合. 具体地, 边缘端负责基于本地采集数据进行感知模型更新, 并将更新后的模型和少量数据上传至云端; 云端将对来自不同智能体的专家模型进行统一整合与推理.

专家更新. 探索结束后, 对于第 i 个智能体, 利用其在环境交互过程中采集的数据集 $\mathcal{D}_{train}^i$ 对本地感知模型 Φ_{θ^i} 进行更新. 模型参数通过最小化损失函数进行优化:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box} + \lambda_{dfl}\mathcal{L}_{dfl} \tag{13}$$

其中, $\mathcal{L}_{cls}$、$\mathcal{L}_{box}$ 和 $\mathcal{L}_{dfl}$ 分别为分类损失、边界框回归损失和分布焦点损失; λ_{cls}、λ_{box} 和 λ_{dfl} 分别为对应的权重.

边缘端完成本地模型更新后, 智能体 i 将其更新后的模型参数 Φ_{θ^i} 及少量数据 $\mathcal{D}_{cloud}^i$ 上传至云端.

自监督学习. 为在云端高效整合来自不同边缘端的异构专家模型, 本文通过多专家协同生成伪标签数据的方式, 利用自监督学习模式, 将分散在不同专家模型中的知识统一蒸馏至单一综合模型中.

具体地, 云端收集来自各智能体的专家模型 $\Phi_\theta^c = \{\Phi_{\theta^1}, \Phi_{\theta^2}, \cdots, \Phi_{\theta^N}\}$, 并将其应用于云端数据集 $\mathcal{D}_{cloud} = \{\mathcal{D}_{cloud}^1, \mathcal{D}_{cloud}^2, \cdots, \mathcal{D}_{cloud}^N\}$. 对于数据集 $\mathcal{D}_{cloud}$ 中任一输入图像 I, 各专家模型分别给出预测结果, 包括目标类别、置信度以及边框坐标. 云端利用置信度筛选策略, 对多专家预测结果进行融合, 仅保留高置信度的可信结果, 作为该图像的伪标签 $\hat{y}$.

基于上述过程, 构建包含图像–伪标签对的伪标签数据集 $\mathcal{D}_{cloud}^{pl}$. 随后, 利用伪标签数据集对感知模型 Φ_θ 进行微调得到 Φ_{θ^*}. 需要注意的是, 该模型在结构上与专家模型保持一致, 但其分类头类别数被拓展为所有专家模型类别的集合 $\mathcal{C} = \{\mathcal{C}^1, \cdots, \mathcal{C}^N\}$. 训练过程中, 综合模型同样采用与专家模型一致的多任务损失函数, 以确保任务定义与优化目标的一致性.

3.3　部署阶段

在部署阶段, 将学习阶段得到的综合感知模型 Φ_{θ^*} 直接应用于未知测试环境中执行感知任务以测试模型在未知环境中的性能, 无需再维护或调度多个专家模型. 得益于多专家伪标签蒸馏机制, 综合模型在保持推理效率的同时, 能够有效继承各专家模型在不同场景与语义上的判别能力, 从而实现高效、稳定且具备良好泛化性能的云端感知服务.

具体地, 融合后的全局模型 Φ_{θ^*} 被部署于测试环境 $\mathcal{E}_{test}$ 中, 对采样自测试环境的测试数据集 $\mathcal{D}_{test}$ 中的观测样本进行推理. 对于每个输入观测, 云端利用综合模型预测最终感知结果. 部署阶段能有效验证模型在未知环境中的泛化能力.

4 实验设置

为验证所提出云边协同多智能体具身学习框架的有效性, 在仿真环境中开展实验.

4.1 实验环境

实验在多场景室内仿真环境 ProcTHOR-10K 中进行, 该环境包含多种典型室内布局与语义目标类别, 能覆盖复杂空间结构与多样化语义分布. 每个场景由若干相互连通的房间构成, 包含家具、电器等常见物体, 为具身探索与感知学习提供具有挑战性的测试条件. 需要注意的是, ProcTHOR-10K 包含 10 000 个训练场景与 1 000 个测试场景, 能保证训练与测试场景在空间布局与语义分布上互不重叠, 以评估模型在未知环境中的泛化能力. 具体地, 根据场景中房间数量对场景复杂度进行排序, 对于每一类复杂度, 选择编号最小的两个场景作为实验环境. 选取场景编号如下:

1) 训练环境: 15, 56, 668, 679, 553, 749, 210, 644, 435, 565, 99, 102

2) 测试环境: 0, 12, 73, 87, 98, 115, 144, 182, 190, 586, 650, 651

4.2 模型配置

4.2.1 感知模型

本文采用 YOLOv12 架构作为智能体的基础感知模型, 并加载在 COCO 数据集上进行预训练的轻量级模型 YOLOv12n 作为预训练模型, 通过目标检测任务评估所提方法的有效性. 目标检测任务损失函数由分类损失、边界框回归损失及分布焦点损失共同构成, 各项损失权重设置如下: 分类损失权重 $\lambda_{cls}=0.5$, 边界框回归损失权重 $\lambda_{box}=7.5$, 分布焦点损失权重 $\lambda_{dfl}=1.5$. 在学习阶段, 边缘端智能体基于各自采集的数据对感知模型进行本地更新, 每个模型的训练迭代次数设定为 1 000 次. 云端自监督学习阶段利用伪标签数据集进行综合模型更新, 具体地, 从边缘端随机选取 10% 数据上传到云端构建伪标签数据. 对于任一专家模型, 采用固定经验置信度阈值策略 (设定为 0.7) 对标签进行筛选, 仅保留置信度高于设定阈值的可信结果作为伪标签, 以减少低置信度预测带来的噪声影响, 在保证标签质量的同时, 避免了复杂自适应阈值搜索带来的额外计算开销. 综合模型的训练迭代次数设定为 1 000 次.

4.2.2 视觉–语言模型

为验证所提出方法在不同视觉–语言模型下的通用性, 本文通过调用 API 的形式, 选取不同预训练视觉–语言模型进行实验, 包括 Qwen2.5-VL-72B-Instruct 以及 Moonshot-V1-32K-Vision-Preview. 所有模型推理阶段的超参数统一设置为: $top_p=0$, $max_token=2\,000$, $frequency_penalty=0$, $temperature=0$.

4.3 智能体配置

实验中共部署 3 个智能体, 并行执行探索任务. 每个智能体对应一个独立的感知模型实例, 各感知模型语义类别划分如下:

1) 智能体 1: Plant, Table, TV, Microwave

2) 智能体 2: Sofa, Bed, Sink

3) 智能体 3: Toilet, Chair, Fridge

每个智能体均配备一台 RGB-D 相机, 视场角为 90°, 图像分辨率为 600×600 像素. 智能体动作空间包含三个离散动作: 向前移动: 前进 0.25 m; 左转: 逆时针旋转 90°; 右转: 顺时针旋转 90°. 每个探索回合的最大低层时间步数设定为 $T_{\max}=200$.

4.4 基线方法

为评估所提出方法的有效性, 本文选取以下基线方法进行对比:

- 随机策略: 每个智能体在探索过程中随机选择动作执行, 不利用任何环境信息或语义先验.
- 规则策略: 每个智能体采用预定义的探索策略, 即在无障碍物情况下持续向前移动, 当检测到障碍物时, 随机选择旋转方向执行转向动作. 该方法不具备语义感知能力, 也不进行多智能体协同.

5 结果与分析

5.1 定量结果

本文针对目标检测任务对所提出方法进行评估, 采用平均精度 (average precision, AP) 作为主要定量评价指标. 具体地, 本文报告 AP@0.50 (简称 AP50) 和 AP@0.50:0.95 (简称 AP) 两项指标. 其中, AP50 表示在交并比 (intersection over union, IoU) 阈值为 0.50 时计算得到的平均精度, 用于衡量模型在较宽松匹配条件下的检测能力. AP 则表示在 IoU 阈值从 0.50 开始, 以 0.05 步长增加到 0.95 时多个阈值下的平均精度, 该指标对定位精度要求更高, 能全面地反映模型在不同检测难度下的综合性能.

表 1 和表 2 给出了目标检测性能对比结果, 其中 pl 代表从边缘端随机选取 10% 数据上传到云端构建伪标签数据, 利用该伪标签数据训练得到的自

表 1 测试集上 AP50 性能 (%)
Table 1 The AP50 performance on the test dataset (%)

方法	数据	Plant	Table	TV	Microwave	Sofa	Bed	Sink	Toilet	Chair	Fridge	mAP50
Pretrain	—	73.6	36.1	50.0	3.6	**52.7**	53.3	23.0	50.0	**43.0**	24.6	41.0
Random	—	60.5	22.5	62.0	**21.9**	30.6	52.0	39.8	57.6	24.5	49.9	42.1
Random	pl	53.2	19.2	49.5	7.4	26.3	33.1	17.5	41.9	17.3	38.8	30.4
Rule	—	50.5	13.4	60.5	7.5	11.3	35.4	56.2	58.9	18.7	50.0	36.2
Rule	pl	69.0	19.0	69.3	2.2	27.9	53.1	32.9	53.9	30.2	59.0	41.7
Moonshot	—	84.0	36.6	74.5	17.1	1.6	65.7	63.0	73.9	33.6	**77.1**	52.7
Moonshot	pl	80.6	**38.6**	76.5	19.0	26.2	**73.0**	**69.3**	68.2	40.4	53.0	**54.5**
Qwen	—	**85.2**	35.9	**83.0**	11.1	12.8	68.6	59.6	**74.2**	33.7	67.3	53.1
Qwen	pl	81.7	38.3	80.2	0.6	29.8	68.0	50.9	72.5	36.0	63.5	52.1

表 2 测试集上 AP 性能 (%)
Table 2 The AP performance on the test dataset (%)

方法	数据	Plant	Table	TV	Microwave	Sofa	Bed	Sink	Toilet	Chair	Fridge	mAP
Pretrain	—	49.4	27.9	29.9	2.8	**50.3**	33.7	20.5	41.4	30.6	23.8	31.0
Random	—	46.7	15.7	47.9	8.4	25.9	38.9	35.6	55.3	13.5	43.7	33.2
Random	pl	37.1	12.6	30.8	2.8	21.9	24.6	17.0	36.6	8.9	34.3	22.7
Rule	—	61.8	18.2	**75.0**	**13.8**	14.5	44.2	61.2	64.4	**32.1**	52.6	43.8
Rule	pl	48.4	13.2	49.5	1.5	23.7	37.1	30.2	44.1	16.3	54.0	31.8
Moonshot	—	64.1	27.9	56.2	8.4	1.0	57.4	54.3	63.7	18.0	**69.4**	42.0
Moonshot	pl	60.4	28.3	57.6	8.5	23.0	55.4	**64.9**	56.1	23.3	49.7	42.7
Qwen	—	**64.9**	28.0	64.7	4.1	10.6	**58.3**	50.4	**66.2**	17.0	63.6	**42.8**
Qwen	pl	54.1	**29.1**	61.3	0.3	26.6	51.2	48.3	61.6	20.6	56.0	40.9

监督综合模型; “—”表示不采取任何操作, 直接将多个专家模型感知结果进行拼接. 从实验结果可以观察到, 大模型驱动的数据采集策略在平均精度均值 (mean average precision, mAP) 和 mAP50 上均显著优于其他采样方式. 在 AP50 指标下, Moonshot (Moonshot-V1-32K-Vision-Preview) 方法取得最高 mAP50 (54.5%), 较预训练模型 (Pretrain) 提升 13.5%. 在 AP 指标下, Qwen (Qwen2.5-VL-72BInstruct) 取得最高 mAP (42.8%), 相比预训练模型提升 11.8%. 这表明基于视觉–语言模型的数据自主采集策略能有效提升模型在目标检测任务中的泛化能力.

从各类别性能分析, 大模型驱动的方法在大多数类别上均表现优异. 例如, Moonshot 在 Fridge 上 AP50 指标达到 77.1%, 相比 Pretrain 提升 52.5%. 值得注意的是, 在 Microwave 这一难例类别上, 大多数方法相较于预训练模型有提升, 但大模型驱动的方法表现低于随机 (Random) 方法, 这是数据采集多样性不足导致的. 此外, 随机方法在部分类别上表现较好, 但整体稳定性不足, 这是由于随机策略采样时容易局限于局部范围内, 导致采集数据存在重复. 而规则 (Rule) 方法在部分类别如 TV 和 Chair 上有较好表现, 但存在类别间波动大的问题, 表明规则化策略采集数据的局限性, 导致模型鲁棒性不足.

5.2 定性结果

图 6 展示在未知测试环境中的目标检测可视化结果对比. 可以观察到, 预训练模型在复杂场景中容易出现漏检现象, 且在目标边界定位上存在偏差, 尤其在遮挡严重或目标分布稠密的区域表现较为不稳定. 相比之下, 本文方法在多种场景下能检测到更多目标实例, 并在目标边界的完整性与一致性方面表现更优. 特别是在存在局部遮挡或远距离的场景中, 所提出方法仍能保持较为稳定的检测效果, 显示出更强的环境适应能力与感知鲁棒性.

此外, 为直观展示本文方法在多智能体协同探索过程中的行为优势, 图 7 补充了探索轨迹的可视

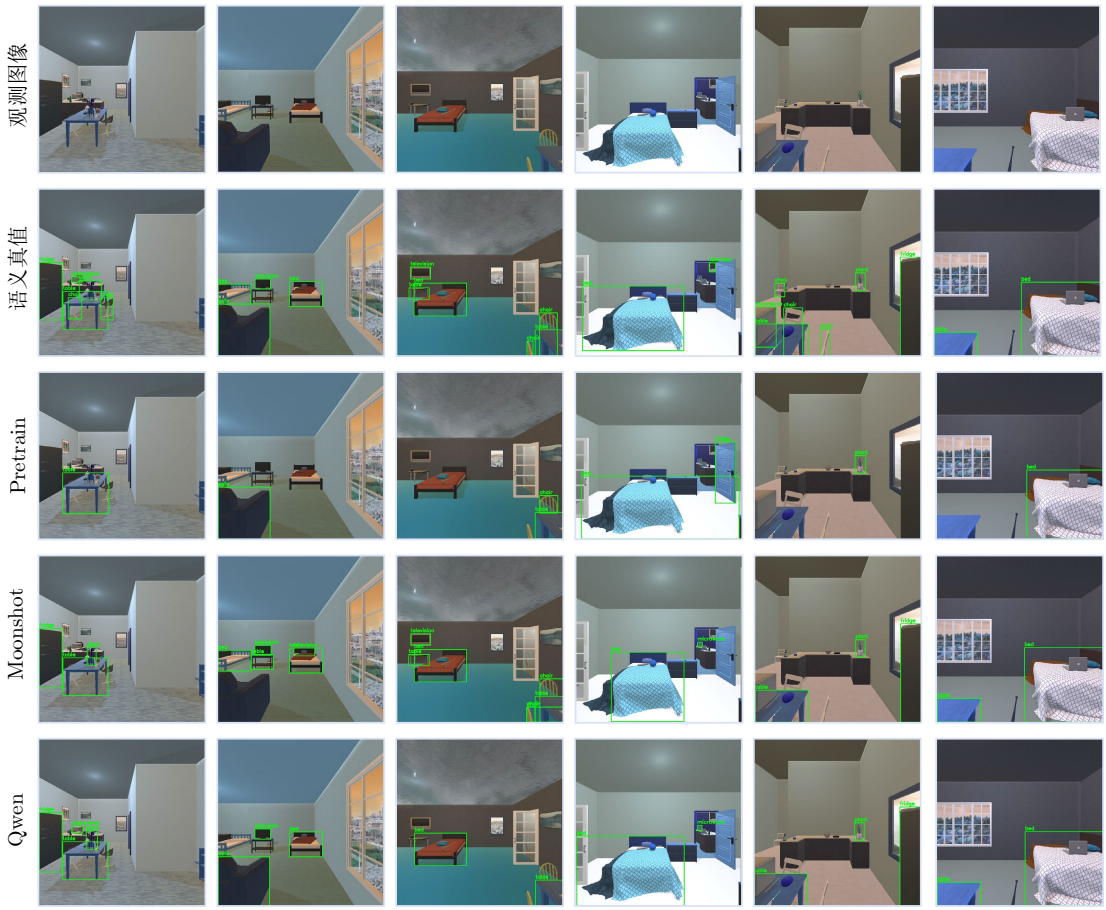


图 6 目标检测结果示例

Fig. 6 Examples of object detection results

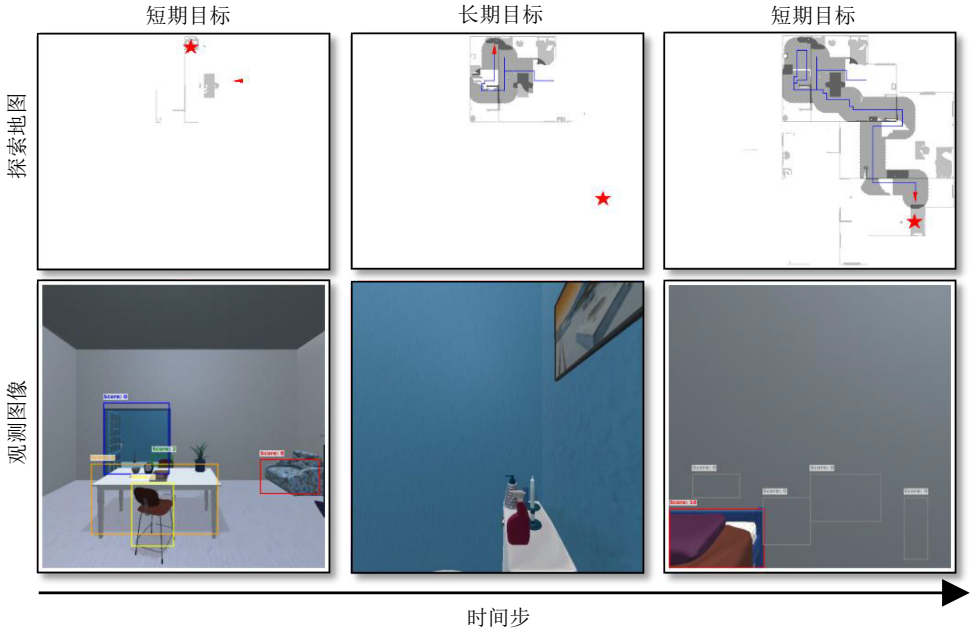


图 7 探索轨迹可视化

Fig. 7 Visualization of exploration trajectories

化结果. 可以观察到, 本文所提策略能够引导智能体在语义先验与空间新颖性的共同作用下, 协同且高效地覆盖高价值未知区域, 显著提升了有效数据的采集效率.

5.3　消融实验结果

为分析语义信息在探索策略中的作用, 对大模型在探索决策阶段是否引入跨智能体共享语义信息进行通信作评估. 实验结果如表 3 和表 4 所示, 其中, Sem 表示大模型在生成探索策略时, 通过通信引入跨智能体共享语义信息进行探索引导.

实验结果表明, 引入跨智能体共享语义信息进行探索决策能够采集更有效的数据, 从而显著提升感知模型的性能. 在 AP50 指标下, Moonshot 从 49.7% 提升至 52.7%, Qwen 从 48.4% 提升至 53.1%. 在更为严格的 AP 指标下, Moonshot 从 40.2% 提升至 42.0%, Qwen 从 39.1% 提升至 42.8%. 上述结果验证了引入跨智能体语义信息共享在探索策略改善上的有效性. 通过共享语义信息, 智能体能快速导航至目标类别附近采集有效数据, 进而改善感知性能.

5.4　结果分析

如表 1 和表 2 所示, 对比不同数据采集方法发现, 基于视觉–语言模型的主动采样策略在整体与绝大多数类别上, 均显著优于基线方法. 这一优势在部分困难类别 (如 Fridge) 以及类别间性能均衡性上更突出. 这表明, 视觉–语言模型能够深入理解视觉任务的数据需求, 引导智能体主动探索出信息量更大、更具挑战性的样本, 从而引导模型学习到更鲁棒的特征表示. 此外, Microwave 类别检测性能较低. 导致该类别性能低的因素为: 1) 样本数量较少: 在数据集中, Microwave 类别的样本数量相对较少, 导致模型在训练过程中难以学习到充分的特征表示; 2) 外观差异较大: Microwave 在不同场景中的外观、尺寸及摆放位置存在差异, 比如从侧面与正面观测时外观不一致, 增加了模型识别的难度; 3) 遮挡或尺度问题: 在室内环境中, Microwave 常出现在厨房台面或橱柜附近, 容易受到遮挡或呈现较小目标尺度, 从而影响检测效果. 图 8 为 Microwave 类样本示例.

如图 9 所示, 尽管视觉–语言模型展现出强大的语义推理能力, 但在开放环境中, 它们偶尔也会产生错误的语义推理. 在本文的框架中, 此类错误对整体探索行为的影响有限. 具体而言, 探索决策并非直接由单一的视觉–语言模型推理决定. 相反, 系统会结合语义信息和探索新颖性来计算好奇心地图. 一旦机器人访问过某个区域, 其探索新颖性就会降低, 从而避免重复探索可能出错的区域. 因此, 偶尔出现的视觉–语言模型推理错误不会显著影响整体探索效率.

表 3 和表 4 表明, 在有限时间步内, 跨智能体共享语义信息有助于探索策略更快更有效地覆盖关键目标区域, 提高高价值样本的采集比例, 从而增强检测模型对复杂场景的建模能力. 如图 5 所示, 智能体接收到其他智能体传递的 Chair 类信息, 云端制定长期目标探索策略时会优先引导智能体向目标物体靠近, 有效节约探索时间. 此外, 观察到个别类别在引入语义信息后存在轻微性能波动, 这可能与探索过程中语义引导导致的数据分布偏移有关. 然而, 从整体检测性能来看, 语义感知探索带来的

表 3　跨智能体共享语义信息消融实验 AP50 性能 (%)

Table 3　The AP50 performance of the ablation experiment on cross-agent shared semantic information (%)

方法	Sem	Plant	Table	TV	Microwave	Sofa	Bed	Sink	Toilet	Chair	Fridge	mAP50
Moonshot	—	84.0	41.8	79.6	31.3	7.2	68.3	23.5	74.5	29.0	58.0	49.7
Moonshot	√	84.0	36.6	74.5	17.1	1.6	65.7	63.0	73.9	33.6	77.1	52.7
Qwen	—	79.5	25.8	65.1	19.2	10.2	64.2	62.9	68.2	27.5	61.1	48.4
Qwen	√	85.2	35.9	83.0	11.1	12.8	68.6	59.6	74.2	33.7	67.3	53.1

表 4　跨智能体共享语义信息消融实验 AP 性能 (%)

Table 4　The AP performance of the ablation experiment on cross-agent shared semantic information (%)

方法	Sem	Plant	Table	TV	Microwave	Sofa	Bed	Sink	Toilet	Chair	Fridge	mAP
Moonshot	—	63.9	33.6	66.8	18.0	6.1	56.3	22.4	67.4	15.8	51.7	40.2
Moonshot	√	64.1	27.9	56.2	8.4	1.0	57.4	54.3	63.7	18.0	69.4	42.0
Qwen	—	53.5	18.0	52.5	9.4	8.8	55.5	58.6	62.3	14.4	58.1	39.1
Qwen	√	64.9	28.0	64.7	4.1	10.6	58.3	50.4	66.2	17.0	63.6	42.8

收益明显超过局部类别的波动影响. 实验结果验证了在探索策略中引入语义信息的有效性. 通过语义感知的探索决策, 大模型能够采集到更加有效的训练数据, 从而显著提升后续目标检测模型的整体性能.

本文还对比了分别使用伪标签与真值 (ground-truth, GT) 标签训练模型时的性能, 表 5 和表 6 给出了实验结果. 其中, gt 表示采用真值标签进行训



图 8 Microwave 类样本示例

Fig. 8 Examples of Microwave class samples

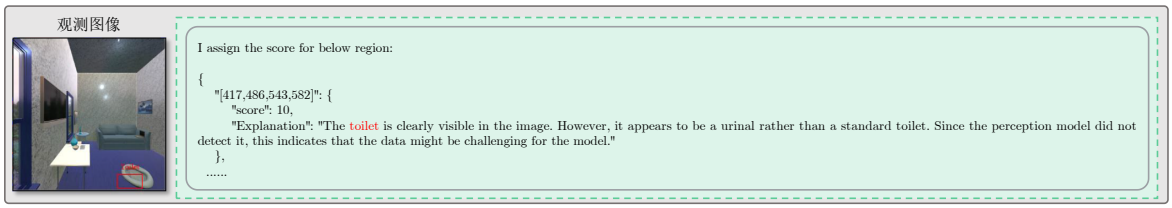


图 9 视觉–语言模型错误视觉推理示例

Fig. 9 Examples of wrong visual reasoning by vision-language models

表 5 伪标签消融实验 AP50 性能 (%)

Table 5 The AP50 performance in the pseudo-label ablation experiment (%)

方法	数据	Plant	Table	TV	Microwave	Sofa	Bed	Sink	Toilet	Chair	Fridge	mAP50
Random	pl	53.2	19.2	49.5	7.4	26.3	33.1	17.5	41.9	17.3	38.8	30.4
Random	gt	58.6	25.7	60.1	9.4	27.5	52.7	13.9	43.2	29.7	61.0	38.2
Rule	pl	69.0	19.0	69.3	2.2	27.9	53.1	32.9	53.9	30.2	59.0	41.7
Rule	gt	64.9	20.4	66.5	3.0	29.0	57.2	47.7	57.9	30.9	69.4	44.7
Moonshot	pl	80.6	38.6	76.5	19.0	26.2	73.0	69.3	68.2	40.4	53.0	54.5
Moonshot	gt	74.5	37.3	79.6	22.5	36.3	77.2	70.9	78.5	43.5	69.5	59.0
Qwen	pl	81.7	38.3	80.2	0.6	29.8	68.0	50.9	72.5	36.0	63.5	52.1
Qwen	gt	81.8	38.7	78.3	11.8	28.9	72.7	67.5	76.7	40.1	58.7	55.5

表 6 伪标签消融实验 AP 性能 (%)

Table 6 The AP performance in the pseudo-label ablation experiment (%)

方法	数据	Plant	Table	TV	Microwave	Sofa	Bed	Sink	Toilet	Chair	Fridge	mAP
Random	pl	37.1	12.6	30.8	2.8	21.9	24.6	17.0	36.6	8.9	34.3	22.7
Random	gt	44.1	17.0	44.0	4.2	24.3	41.0	13.3	35.8	17.8	54.3	29.6
Rule	pl	48.4	13.2	49.5	1.5	23.7	37.1	30.2	44.1	16.3	54.0	31.8
Rule	gt	48.5	14.1	47.1	0.8	25.3	43.8	45.5	54.6	17.2	64.2	36.1
Moonshot	pl	60.4	28.3	57.6	8.5	23.0	55.4	64.9	56.1	23.3	49.7	42.7
Moonshot	gt	59.2	28.9	61.8	8.8	33.2	62.2	65.9	69.3	25.2	60.5	47.5
Qwen	pl	54.1	29.1	61.3	0.3	26.6	51.2	48.3	61.6	20.6	56.0	40.9
Qwen	gt	57.6	30.6	62.3	4.9	24.6	60.3	64.4	68.1	23.8	53.3	45.0

练. 通过对比实验结果发现大模型方法显著优于基线方法, 甚至优于基线方法采用真值数据训练模型的性能, 为目标检测任务提供了一种有效且可扩展的训练范式. 需要注意的是, 采用真实标注训练始终优于伪标签训练. 此外, 结合表 1 和表 2 中引入伪标签后各类别性能的波动情况可以发现, 虽然伪标签总体上提升了模型的平均精度, 但在某些类别 (如 Fridge、Plant) 上性能却出现了下降. 这主要是由于原始专家模型在这些复杂类别上的原始预测本身存在不确定性, 导致生成的伪标签包含噪声或漏标错误. 这些低质量的伪标签作为监督信号引入了误差, 从而对模型在部分类别上的性能产生了明显的负面影响. 这进一步说明低质量伪标签会对模型性能产生明显负面影响, 未来可考虑引入针对难例类别引入更鲁棒的去噪机制.

6 结束语

本文围绕未知环境下目标检测任务中的数据高效采集与感知模型持续优化问题, 提出了一种大模型驱动的云边协同多智能体具身学习框架. 该框架在云端引入大模型驱动的高层策略, 并利用多智能体共享语义信息对边缘端探索目标进行统一规划, 同时在边缘端执行低层控制策略, 从而在有限的探索预算下获取具有更高语义价值与多样性的训练数据.

在通信方面, 边缘端仅需将本地更新后的模型参数及少量数据上传至云端, 从而有效降低系统的通信开销. 针对来自不同智能体的异构感知模型, 本文在云端进一步引入自监督学习模式实现模型整合, 有效减少计算量.

通过上述云边协同设计, 本文构建了一种交互决策–数据采集–模型优化–模型整合的闭环框架, 为长期自主运行的感知系统提供了一种可行且具备扩展性的解决方案.

参考文献

1 Liu H P, Guo D, Cangelosi A. Embodied intelligence: A synergy of morphology, action, perception and learning. *ACM Computing Surveys*, 2025, **57**(7): Article No. 186

2 Liu Hua-Ping, Guo Di, Sun Fu-Chun, Zhang Xin-Yu. Morphology-based embodied intelligence: Historical retrospect and research progress. *Acta Automatica Sinica*, 2023, **49**(6): 1131−1154 (刘华平, 郭迪, 孙富春, 张新钰. 基于形态的具身智能研究: 历史回顾与前沿进展. 自动化学报, 2023, **49**(6): 1131−1154)

3 Chaplot D S, Jiang H, Gupta S, Gupta A. Semantic curiosity for active visual learning. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 309−326

4 Siddiqui Y, Valentin J, Niessner M. ViewAL: Active learning with viewpoint entropy for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 9430−9440

5 Nilsson D, Pirinen A, Gärtner E, Sminchisescu C. Embodied visual active learning for semantic segmentation. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI, 2021. 2373−2383

6 Nilsson D, Pirinen A, Gärtner E, Sminchisescu C. Embodied learning for lifelong visual perception. arXiv preprint arXiv: 2112.14084, 2021.

7 Zurbrügg R, Blum H, Cadena C, Siegwart R, Schmid L. Embodied active domain adaptation for semantic segmentation via informative path planning. *IEEE Robotics and Automation Letters*, 2022, **7**(4): 8691−8698

8 Rückin J, Magistri F, Stachniss C, Popović M. An informative path planning framework for active learning in UAV-based semantic mapping. *IEEE Transactions on Robotics*, 2023, **39**(6): 4279−4296

9 Rückin J, Magistri F, Stachniss C, Popović M. Semi-supervised active learning for semantic segmentation in unknown environments using informative path planning. *IEEE Robotics and Automation Letters*, 2024, **9**(3): 2662−2669

10 Jing Y, Kong T. Learning to explore informative trajectories and samples for embodied perception. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 6050−6056

11 Fang Z Y, Jain A, Sarch G, Harley A W, Fragkiadaki K. Move to see better: Self-improving embodied object detection. arXiv preprint arXiv: 2012.00057, 2021.

12 Scarpellini G, Rosa S, Morerio P, Natale L, del Bue A. Look around and learn: Self-training object detection by exploration. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2024. 72−88

13 Scarpellini G, Rosa S, Morerio P, Natale L, del Bue A. Self-improving object detection via disagreement reconciliation. arXiv preprint arXiv: 2302.10624, 2023.

14 Chaplot D S, Dalal M, Gupta S, Malik J, Salakhutdinov R. SEAL: Self-supervised embodied active learning using exploration and 3D consistency. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 1002

15 Blum H, Müller M G, Gawel A, Siegwart R, Cadena C. SCIM: Simultaneous clustering, inference, and mapping for open-world semantic scene understanding. *Robotics Research*. Cham, Switzerland: Springer, 2022. 119−135

16 Liang X R, Han A, Yan W, Raghunathan A, Abbeel P. ALP: Action-aware embodied learning for perception. arXiv preprint arXiv: 2306.10190, 2023.

17 Clay V, Pipa G, Kühnberger K, König P. Development of few-shot learning capabilities in artificial neural networks when learning through self-supervised interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(1): 209−219

18 Kotar K, Mottaghi R. Interactron: Embodied adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 14840−14849

19 Tan S N, Ge M M, Guo D, Liu H P, Sun F C. Knowledge-based embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(10): 11948−11960

20 Li X H, Guo D, Liu H P, Sun F C. Embodied semantic scene graph generation. In: Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR, 2022. 1585−1594

21 Shridhar M, Manuelli L, Fox D. CLIPort: What and where

pathways for robotic manipulation. In: Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR, 2022. 894−906

22 Ji Z H, Lin H X, Gao Y. DyNaVLM: Zero-shot vision-language navigation system with dynamic viewpoints and self-refining graph memory. arXiv preprint arXiv: 2506.15096, 2025.

23 Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: An embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 8469−8488

24 Jiang Y F, Gupta A, Zhang Z C, Wang G Z, Dou Y Q, Chen Y J, et al. VIMA: General robot manipulation with multimodal prompts. arXiv preprint arXiv: 2210.03094, 2023.

25 Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, et al. RT-1: Robotics Transformer for real-world control at scale. arXiv preprint arXiv: 2212.06817, 2023.

26 Zitkovich B, Yu T H, Xu S C, Xu P, Xiao T, Xia F, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In: Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR, 2023. 2165−2183

27 Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, et al. OpenVLA: An open-source vision-language-action model. In: Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR, 2024. 2679−42713

28 Black K, Brown N, Darpinian J, Dhabalia K, Driess D, Esmail A, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In: Proceedings of the 9th Conference on Robot Learning. Seoul, Republic of Korea: PMLR, 2025. 17−40

29 NVIDIA. GR00T N1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv: 2503.14734, 2025.

30 Han B, Kim J, Jang J. A dual process VLA: Efficient robotic manipulation leveraging VLM. arXiv preprint arXiv: 2410.15549, 2024.

31 Zhen H Y, Qiu X W, Chen P H, Yang J C, Yan X, Du Y L, et al. 3D-VLA: A 3D vision-language-action generative world model. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 61229−61245

32 Cen J, Yu C H, Yuan H J, Jiang Y M, Huang S T, Guo J Y, et al. WorldVLA: Towards autoregressive action world model. arXiv preprint arXiv: 2506.21539, 2025.

33 Ye S, Ge Y H, Zheng K Y, Gao S Y, Yu S, Kurian G, et al. World action models are zero-shot policies. arXiv preprint arXiv: 2602.15922, 2026.

34 Simmons R G, Apfelbaum D, Burgard W, Fox D, Moors M, Thrun S, et al. Coordination for multi-robot exploration and mapping. In: Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence. Austin, USA: AAAI, 2000. 852−858

35 Yamauchi B. Frontier-based exploration using multiple robots. In: Proceedings of the 2nd International Conference on Autonomous Agents. Minneapolis, USA: ACM, 1998. 47−53

36 Zhang Z L, Yu J C, Tang J H, Xu Y F, Wang Y. MR-TopoMap: Multi-robot exploration based on topological map in communication restricted environment. *IEEE Robotics and Automation Letters*, 2022, **7**(4): 10794−10801

37 Liu X, Prabhu A, Cladera F, Miller I D, Zhou L F, Taylor C J, et al. Active metric-semantic mapping by multiple aerial robots. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 3282−3288

38 Liu X, Lei J Z, Prabhu A, Tao Y Z, Spasojevic I, Chaudhari P. SlideSLAM: Sparse, lightweight, decentralized metric-semantic SLAM for multirobot navigation. *IEEE Transactions on Robotics*, 2025, **41**: 6529−6548

39 Yu Wen-Wu, Yang Xiao-Ya, Li Hai-Chang, Wang Rui, Hu Xiao-Hui. Attentional intention and communication for multi-agent learning. *Acta Automatica Sinica*, 2023, **49**(11): 2311−2325
(俞文武, 杨晓亚, 李海昌, 王瑞, 胡晓惠. 面向多智能体协作的注意力意图与交流学习方法. 自动化学报, 2023, **49**(11): 2311−2325)

40 Foerster J N, Farquhar G, Afouras T, Nardelli N, Whiteson S. Counterfactual multi-agent policy gradients. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. Article No. 363

41 Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. Article No. 30

42 Yu C, Velu A, Vinitsky E, Gao J X, Wang Y, Bayen A, et al. The surprising effectiveness of PPO in cooperative multi-agent games. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1787

43 Goel H, Omama M, Chalaki B, Tadiparthi V, Pari E, Chinchali S P. R3DM: Enabling role discovery and diversity through dynamics models in multi-agent reinforcement learning. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. 19600−19620

44 Sagirova A, Kuratov Y, Burtsev M. SRMT: Shared memory for multi-agent lifelong pathfinding. arXiv preprint arXiv: 2501.13200, 2025.

45 Shi Wei, Feng Yang-He, Cheng Guang-Quan, Huang Hong-Lan, Huang Jin-Cai, Liu Zhong, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(7): 1610−1623
(施伟, 冯旸赫, 程光权, 黄红蓝, 黄金才, 刘忠, 等. 基于深度强化学习的多机协同空战方法研究. 自动化学报, 2021, **47**(7): 1610−1623)

46 Huang Shuai, Feng Yu-Hang, Zheng Tai-Xiong, Li Yong-Fu. Centralized cooperative control of mixed vehicle groups considering multi-vehicle influence under cloud-edge-end collaboration. *Acta Automatica Sinica*, 2026, **52**(1): 172−190
(黄帅, 冯雨航, 郑太雄, 李永福. 云−边−端协同下考虑多车影响的混行车群集中式协同控制. 自动化学报, 2026, **52**(1): 172−190)

47 Gummadi S, Gasparino M V, Vasisht D, Chowdhary G. Fed-EC: Bandwidth-efficient clustering-based federated learning for autonomous visual robot navigation. *IEEE Robotics and Automation Letters*, 2024, **9**(12): 11841−11848

48 Yuan Z Y, Xu S Y, Zhu M H. Federated reinforcement learning for robot motion planning with zero-shot generalization. *Automatica*, 2024, **166**: Article No. 111709

49 Lei S X, Tang H J, Li C Y, Zhang X Y, Xu C L, Wu H M. Federated MADDPG-based collaborative scheduling strategy in vehicular edge computing. *IEEE Transactions on Mobile Computing*, 2026, **25**(1): 54−66

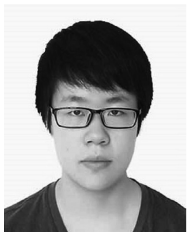
50 Andong F J E N, Min Q. Federated multi-agent reinforcement learning for privacy-preserving and energy-aware resource management in 6G edge networks. arXiv preprint arXiv: 2509.10163, 2025.

51 Kannan S S, Venkatesh V L N, Min B C. SMART-LLM: Smart multi-agent robot task planning using large language models. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 12140−12147

52 Wang R Y, Tsu H, Hunt D, Luo S C, Kim J, Pajic M. LLM-

MCoX: Large language model-based multi-robot coordinated exploration and search. arXiv preprint arXiv: 2509.26324v1, 2026.

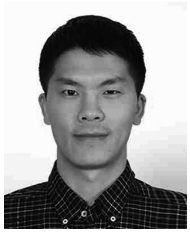
53 Zhu Y X, Chen J F, Zhang X T, Guo M, Li Z K. DEXTER-LLM: Dynamic and explainable coordination of multi-robot systems in unknown environments via large language models. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE, 2025. 10182−10189



王 涓 清华大学计算机科学与技术系博士研究生. 主要研究方向为具身学习.
E-mail: wangjuan20@mails.tsinghua.edu.cn
(**WANG Juan** Ph.D. candidate in the Department of Computer Science and Technology, Tsinghua University. Her main research interest is embodied learning.)



赵大伟 国防科技创新研究院副研究员. 2018 年获得国防科技大学博士学位. 主要研究方向为计算机视觉与自动驾驶.
E-mail: adamzdw@163.com
(**ZHAO Da-Wei** Associate researcher at the Defense Innovation Institute. He received his Ph.D. degree from National University of Defense Technology in 2018. His research interests include computer vision and autonomous driving.)



朱 琪 国防科技创新研究院副研究员. 2017 年获得国防科技大学博士学位. 主要研究方向为运动规划与自动驾驶.
E-mail: zhuqiwk@126.com
(**ZHU Qi** Associate researcher at the Defense Innovation Institute. He received his Ph.D. degree from National University of Defense Technology in 2017. His research interests include motion planning and autonomous driving.)



刘华平 清华大学计算机科学与技术系教授. 2004 年获得清华大学博士学位. 主要研究方向为具身感知与学习. 本文通信作者.
E-mail: hpliu@tsinghua.edu.cn
(**LIU Hua-Ping** Professor in the Department of Computer Science and Technology, Tsinghua University. He received his Ph.D. degree from Tsinghua University in 2004. His research interests include embodied perception and learning. Corresponding author of this paper.)