中国科学: 信息科学 2024 年 第 54 卷 第 9 期: 2035–2082
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

大模型驱动的具身智能: 发展与挑战

白辰甲[1,2], 许华哲[3], 李学龙[2,1*]

1. 中国电信人工智能研究院 (TeleAI), 上海 200232
2. 中国电信人工智能研究院 (TeleAI), 北京 100033
3. 清华大学交叉信息学院, 北京 100084
* 通信作者. E-mail: xuelong_li@ieee.org

收稿日期: 2024–03–11; 修回日期: 2024–06–25; 接受日期: 2024–08–12; 网络出版日期: 2024–08–29

国家自然科学基金 (批准号: 61871470, 62306242) 资助项目

摘要 大模型驱动的具身智能是涵盖人工智能、机器人学和认知科学的交叉领域, 重点研究如何将大模型的感知、推理和逻辑思维能力与具身智能相结合, 提升现有模仿学习、强化学习、模型预测控制等具身智能框架的数据效率和泛化能力. 近年来, 随着大模型能力的不断提升, 以及具身智能中示教数据、仿真平台、任务集合的不断完善, 大模型和具身智能的结合将成为人工智能的下一个浪潮, 有望成为人工智能迈向实体机器人的重要突破口. 本文围绕大模型驱动的具身智能这一研究领域, 从 3 个方面进行了系统的调研、分析和展望. 首先, 回顾了大模型和具身智能的相关技术背景, 以及具身智能现有的学习框架. 其次, 按照大模型赋能具身智能的方式, 将现有研究分为大模型驱动的环境感知、大模型驱动的任务规划、大模型驱动的基础策略、大模型驱动的奖励函数、大模型驱动的数据生成等 5 类范式. 最后, 总结了大模型驱动的具身智能中存在的挑战, 对可行的技术路线进行展望, 为相关研究人员提供参考, 进一步推动国家人工智能发展战略.

关键词 具身智能, 大模型, 环境感知, 任务规划, 基础策略

1 引言

具身智能 (embodied AI) 是人工智能、机器人学、认知科学的交叉领域, 主要研究如何使机器人具备类似人类的感知、规划、决策和行为能力 [1]. 具身智能可以追溯到 20 世纪 50 年代, 艾伦 · 图灵首次提出具身智能的概念, 探索如何使机器感知和理解世界, 并作出相应的决策和行动 [2,3]. 随后在 80 年代对符号主义的反思中, 以罗德尼 · 布鲁克斯为代表的研究者逐渐认识到, 智能不应该只在对数据的被动学习中得到, 而应该通过与环境进行主动交互中获取, 应当重点研究如何让机器人主动适应环境 [4]. 近年来, 在高性能算力平台和大规模标注数据的支持下, 深度学习方法通过挖掘数据模式, 在图

引用格式: 白辰甲, 许华哲, 李学龙. 大模型驱动的具身智能: 发展与挑战. 中国科学: 信息科学, 2024, 54: 2035–2082, doi: 10.1360/SSI-2024-0076
Bai C J, Xu H Z, Li X L. Embodied-AI with large models: research and challenges (in Chinese). Sci Sin Inform, 2024, 54: 2035–2082, doi: 10.1360/SSI-2024-0076

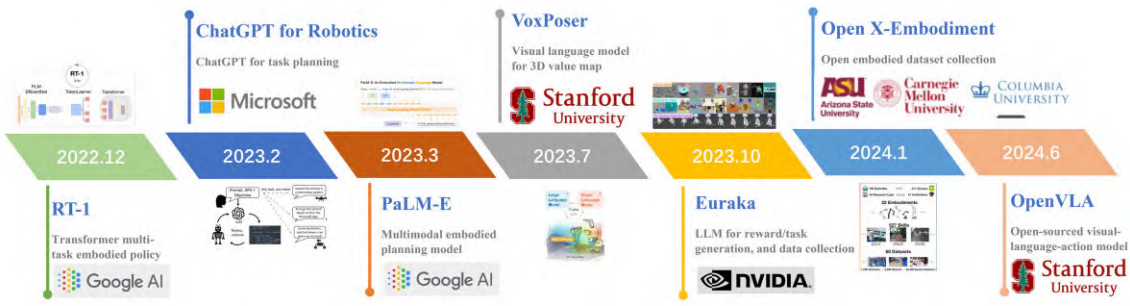


图 1 (网络版彩图) 领域典型进展

Figure 1 (Color online) The significant progress in the field

像识别、语言处理、围棋、蛋白质结构预测等任务中取得了一系列突破性的进展. 然而, 这些非具身智能体缺乏与环境交互学习的经验, 无法直接驱动机器人实体完成特定任务. 相比较而言, 具身智能强调感知 - 运动回路 (perception-action loop), 使用物理实体来感知和建模环境, 根据任务目标和实体能力进行规划和决策, 最后使用实体的运动能力来完成任务. 具身实体对任务的完成结果将作为反馈进一步优化智能体的策略, 从而使智能体的行为能够适应变化的环境, 这一过程与人类的学习和认知过程有很高的相似性. 具身智能在研究中更多体现智能的理念, 在具身实体中融合了视觉、语言、决策等多方面的技术来提升智能体的通用型和泛化性 [5].

近年来, 以 ChatGPT 为带代表的大语言模型 (large language model, LLM) [6] 技术取得了突破性的进展, 通过在大规模网络对话数据中进行学习, ChatGPT 能够实现包括自动问答、文本分类、自动文摘、机器翻译、聊天对话等各种自然语言理解和自然语言生成任务, 同时具备在少样本和零样本场景下达到了传统监督学习方法的性能, 并具有较强的泛化能力 [7]. 通过先进的思维链 (chain-of-thought, CoT) [8] 等提示技术, 大语言模型的逻辑推理能力获得了大幅提升, 从而有望解决复杂具身智能场景中的任务分解和推理问题. 视觉基础模型 (visual foundation model, VFM) [9] 通过自监督的学习目标可以获得强大的视觉编码器, 能够解决如图像分类、语义分割、场景理解等视觉感知任务. 在具身智能任务中, 强大的视觉编码器能够对视觉传感器获得的周围环境信息进行分析和理解, 从而帮助智能体进行决策. 在此基础上, 视觉 - 语言模型 (visual-language model, VLM) [10] 通过引入预训练视觉编码器和视觉 - 语言模态融合模块, 使得大语言模型能够获取视觉输入, 同时根据语言提示进行视觉问答. 在具身智能中, 引入视觉 - 语言模型能够使智能体根据任务语言指令和环境的视觉观测进行推理和决策, 从而提升智能体对环境的感知和理解能力. 多模态大模型 (large multimodal model) [11,12] 通过引入视频、音频、肢体语言、面部表情和生理信号等更多模态, 可以分析更丰富的传感器输入并进行信息融合, 同时结合具身智能体中特有的机器人状态、关节动作等模态信息, 帮助解决更复杂的具身智能任务. 大模型通过充分利用大规模数据集中学习到的知识, 结合特定的具身智能场景和任务描述, 为智能体提供环境感知和任务规划的能力. 图 1 列举了近年来大模型驱动的具身智能领域的代表性成果.

在赋能感知和规划之外, 大模型能够和具身智能的经典框架结合, 提升策略的泛化能力和对环境的适应能力. 具身智能的传统框架主要包括模仿学习 (imitation learning, IL) [13] 、强化学习 (reinforcement learning, RL) [14] 、模型预测控制 (model-predictive control, MPC) [15] 等. 具体地, 模仿学习遵循监督学习的范式, 通过直接从专家轨迹数据中学习策略, 但往往受限于专家数据的规模和协变量偏移 (covariate shift) 问题而容易产生较高的泛化误差; 强化学习通过在环境交互中试错来获得样本, 通过

最大化奖励来获得策略和值函数, 但在机器人任务中受限于复杂的奖励设计和长时间的环境交互; 模型预测控制通过使用环境模型产生对未来策略执行情况的预测, 结合策略搜索方法获得当前最优的动作, 但依赖于对环境的先验知识和环境模型的泛化能力. 近年来, 许多研究尝试了大模型技术与上述框架的结合, 从而克服现有框架面临的问题 [16]. 具体地, 在模仿学习中, 大语言模型和视觉语言模型能够作为基础策略使智能体利用大模型对环境的理解和泛化能力, 同时, 大模型对任务的分解能够产生的任务短期目标来降低模仿学习的难度 [17]; 在强化学习中, 大模型能够根据对任务和场景的理解产生合适奖励函数来引导强化学习中价值函数和策略函数的学习, 同时强化学习能够作为大模型的基础策略和人类偏好对齐的工具, 引导策略的输出符合人类偏好 [18]; 在模型预测控制的框架下, 大模型能够利用从大量训练数据中获取的对物理世界的理解构建环境模型, 进而使智能体能够使用环境模型进行交互和策略搜索 [19]. 此外, 视觉生成模型和语言生成模型可以根据任务需求生成机器人交互环境供强化学习算法进行交互, 或生成交互数据来扩充特定任务下的专家样本, 用于缓解真实机器人任务中普遍存在的数据稀缺问题 [20].

本文围绕大模型驱动的具身智能, 首先介绍相关技术背景, 包括具身智能的基本概念, 大模型相关技术, 以及强化学习、模仿学习、模型预测控制等策略学习框架. 随后, 从学习范式的角度, 将大模型驱动的具身智能算法进行分类, 主要包括大模型驱动的环境感知、任务规划、基础策略、奖励函数和数据生成等 5 个方面. 其中, (1) 大模型驱动的环境感知从冗余的多传感器观测中进行特征抽取和信息融合, 能够提取对策略学习有用的信息, 从而使具身智能学习框架普遍受益; (2) 大模型对宏观任务的规划使用大模型的逻辑推理能力对复杂任务进行分解, 允许使用灵活的底层学习框架对分解后的任务进行策略学习; (3) 大模型驱动的基础策略可以与模仿学习框架进行结合并作为模型学习的初始策略, 在使用少量机器人的任务数据微调后, 大模型能够将通用的环境理解能力和特定的具身应用场景结合, 减少策略训练对机器人数据的需求量并提升策略的泛化能力; (4) 大模型驱动的奖励函数可与强化学习算法进行结合, 减少机器人场景中人为进行奖励函数设计的难度, 降低奖励函数设计对物理先验知识的依赖, 克服强化学习算法在机器人任务中面临的稀疏奖励问题; (5) 大模型驱动的数据生成根据学习框架的不同分为两类: 一方面, 大模型可作为环境模型生成智能体的未来轨迹预测, 与模型预测控制算法和基于模型的强化学习算法相结合进行策略搜索; 另一方面, 大模型可以生成机器人数据用于具身策略训练, 作用于无模型强化学习算法和模仿学习算法, 从而缓解机器人任务的数据缺乏问题.

在对研究现状进行总结和分析的基础上, 本文提出了大模型驱动的具身智能研究中存在的 5 大挑战, 主要包括: (1) 大模型在特定具身场景中的适应问题. 从宏观上看, 大模型是广泛意义上的 “通才”, 而在特定具身任务中往往需要能解决该任务的 “专才” 智能体, 如何使用大模型中涌现的通用知识在机器人任务中达到精确的物体操作和稳定的运动控制, 仍然是一项长期的挑战. (2) 大模型策略和人类偏好的对齐问题. 具身任务的策略偏好和大模型中使用人类偏好往往有所不同, 例如, 面对具身智能规划问题, 大语言模型往往趋向于给出多样的、全面的回答, 而智能体执行任务需要准确的、可安全执行的指令分解. 如何将大模型能力和人类偏好在具身智能任务中进行对齐是一项重要的研究问题. (3) 具身策略的跨域泛化问题. 大模型能够对不同的任务指令进行解析, 对多样化的视觉场景进行识别. 然而, 具身智能同时面临着跨域泛化的难题, 如环境参数改变、机器人动力学改变, 跨形态学实体的泛化等机器人特有的问题, 目前大模型尚不具备直接解决问题的能力. (4) 大模型驱动多智能体协作的能力. 在解决复杂任务中往往需要多个智能体进行协作, 其中涉及到的任务分配、合作博弈、沟通反馈等传统多智能体合作问题在大模型背景下缺乏相关研究, 如何使大模型驱动多智能体进行高效协作在未来是重要的研究问题. (5) 大模型具身策略的决策实时性问题. 机器人策略在执行过程中环境

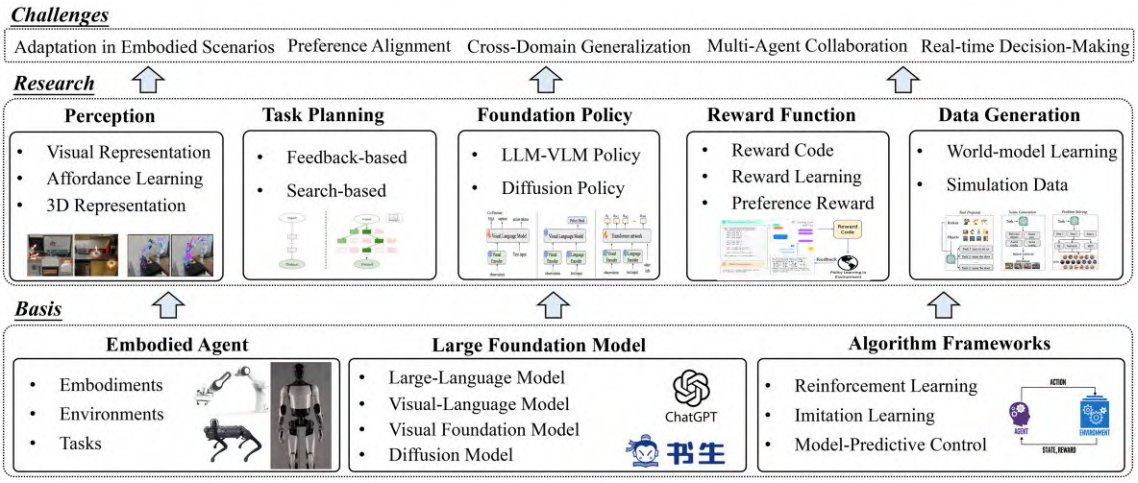


图 2 (网络版彩图) 综述整体框架

Figure 2 (Color online) The overall framework of this survey

观测是快速变化的, 具身策略需要保持较高的决策频率. 而大模型在进行单次推理时需要较高的计算代价, 如何解决大模型在规划和决策时的实时性是大模型在实体机器人应用的重要问题. 本文将对以上挑战进行分析和总结, 对可能的研究路线进行展望, 为大模型在具身智能中的广泛应用提供系统性参考. 本文的主要框架如图 2 所示.

2 背景

本节将从具身知识基本概念、具身智能学习框架、大模型相关技术等几个方面介绍本文的背景知识. 其中, 具身智能基本概念主要包括具身实体定义、传感器、任务定义等; 具身智能学习框架主要包括模仿学习、强化学习、模型预测控制等; 大模型技术主要包括大语言模型、视觉 - 语言大模型、多模态大模型、扩散生成模型等.

2.1 具身系统基本概念

具身智能系统的基本结构如图 3 所示, 主要包括实体、任务、环境三个部分. 其中, 具身实体是系统的核心, 主要包括机器人、传感器、执行器等部分. 在特定任务中, 机器人通过传感器获取对环境的感知, 随后由具身智能算法产生当前合适的动作, 将动作传输给执行器, 执行器产生底层机器人指令与环境进行交互, 获得环境的反馈和更新后的场景感知信息并循环进行上述过程. 机器人的类型往往决定了具体可以使用的传感器和执行器, 常用的机器人类型包括机械臂、四足机器人、移动机器人、灵巧手、人形机器人等, 如图 4 所示.

机械臂是最为常见的具身实体类型, 用来执行物体操作和抓取等任务. 人工智能算法驱动的机械臂研究拥有较长的历史, 常用的机械臂类型包括 Franka, xArm, Sawyer, Kuka, UR5 等. 针对常用的机械臂类型, 研究人员开发了许多高效率的仿真平台, 例如 MuJoCo [21], Deepmind Control Suite [22], Franka Kitchen [23], RoboSuite [24], ManiSkill [25] 等. 仿真环境提供了物理仿真和环境渲染的功能, 前者通过机械臂自身的物理结构进行动力学转移的数学模型构建, 后者通过图像渲染等工具获得机械臂及其周围环境的 2D/3D 观测用于机器学习算法的训练. 通过仿真平台, 智能体能够快速的与环境交

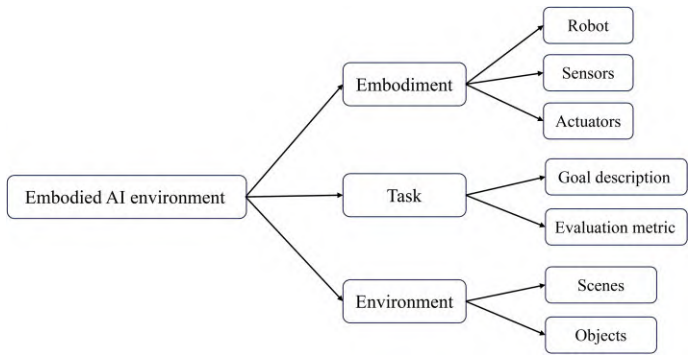


图 3 具身系统基本概念示意图

Figure 3 An overview of embodied AI systems

图 4 (网络版彩图) 常见的具身实体机器人

Figure 4 (Color online) Common embodied robots

互并获得大量的交互样本, 同时可以在标准任务集合上快速验证算法的有效性. 同时, 现有的高性能仿真环境如 ManiSkill2 [26] 等提供了便捷的仿真 - 真实环境迁移通道, 能够将仿真训练得到的策略迁移到真实机械臂中. 然而, 仿真 - 真实迁移中由于光照、背景、视角、操作物体、物理引擎的变化, 仍然存在较大的难度, 因此也有许多研究选择直接使用真实机械臂上采集的数据集进行训练. 相比较而言, 直接使用实体机械臂进行数据采集和策略学习需要昂贵的时间成本, 同时需要解决策略安全性等问题. 但随着许多高质量在真实机械臂专家示教数据集的公开, 研究者可以使用公开数据集进行策略模仿学习, 降低了数据获取的成本. 常用的机械臂专家示教数据集包括 MT-Opt [27], BridgeData [28], RH20T [29], ManiSkill2 [26], Open-X [20] 等. 例如, 最新公布的 Open-X [20] 数据集包含了来自 22 个不同的机器人实体, 涵盖 500 多个不同的机器人任务, 共计超过 1 M 条机器人轨迹, 大大促进了领域的发展. 通用操作接口 (universal manipulation interface, UMI) [30] 使用手持夹爪和鱼眼镜头, 提供了一种廉价的数据采集方案.

四足机器人可以在复杂地形条件下进行稳定行走、奔跑、跳跃和避障等任务. 同时, 四足机器人在传统控制算法控制和人工智能算法驱动的控制方式上都取得了显著的成就, 是具身实体的重要组成之一. 常见的四足机器人类型有宇树科技开发的 Aliengo, Go1, A1 等, 麻省理工学院研发的 MiniCheetah, 波士顿动力研发的 SpotMini, 苏黎世联邦理工学院的 Anymal 等. 在运动时, 四足机器人需要对不同的地形使用不同的运动策略, 而采集涵盖所有不同地形的数据集是较为困难的. 同时, 四足机器人具有较高的关节自由度, 人类使用遥操作和预定义策略的方式来控制四足机器人运动具有较高的难度. 使用强化学习方法训练四足机器人的高自由度运动策略需要大量的交互数据和精心设计的奖励函数, 因此依赖高度并行化的仿真环境进行算法调优. 常用的仿真环境包括 IsaacGym [31], Pybullet [32],

MuJoCo [21], Raisim [33], Webots [34] 等强大的仿真平台. 仿真环境在提供相对真实的物理动力学环境和各种传感器的同时, 规避了真机实验的危险性, 节约了开发成本. 主流的仿真环境使用物理仿真引擎来对环境中的物体进行数学建模, 从而计算其中由于碰撞摩擦等力的作用带来的影响. 但由于现实环境的复杂性, 仿真中的渲染始终与实际存在差距, 在涉及环境彩色信息的任务中从仿真到真机的迁移应用仍然是一大挑战. 此外, 研究人员构建了采集自动物或者通过传统控制算法生产的参考动作, 通过模仿学习来使得四足机器人具有学习步态和其他运动技能, 这种方式可以减少由于仿真环境的渲染不足带来的影响 [35, 36].

移动机器人拓展了固定机械臂的使用场景, 与固定底座的机械臂相比, 移动机器人通过可移动的底座进行运动, 随后在自主选择的场景下执行抓取和物体操作任务. Google 开发的 EveryDay Robot 将移动机器人作为家庭生活和办公室工作的助手, 能够进行倒垃圾、清洁桌子、整理房间等. 移动机器人在室内的移动使用视觉定位和导航技术, 移动机械臂的操作使用大模型驱动的具身智能框架进行训练, 有效扩展了机械臂的能力. Google 提出的 Robot Transformer [37, 38] 采集了移动机器人完成日常任务的轨迹片段, 构成了真实移动机器人的专家数据集, 包含了 700 多个任务, 包括移动物体、拉开抽屉、厨房操作、开罐子等, 学习到的策略在新的任务指令上有一定的泛化能力. 近期斯坦福大学提出的 Mobile ALOHA 移动机器人 [39] 以低廉的成本和复杂的操作能力受到广泛关注, 通过大量专家示教数据的模仿学习, 展示的 Mobile ALOHA 机器人能够完成滑蛋虾仁、干贝烧鸡、蚝油生菜等菜品的制作, 同时能够擦拭桌子、整理碗柜等. 该机器人的移动速度可以和人类行走的速度相媲美, 大约 1.42 m/s, 能够在操作较大的家用物品时保持稳定. 在远程操控时, 手臂和移动底座均可以用来进行遥操作, 从而由人类专家完成数据的采集. 在未来, 移动机器人有望替代人类完成常见的家务操作, 具有广泛的应用前景. 然而, 相比于四足机器人, 常用的移动机器人使用轮式结构, 暂不具备复杂地形的运动能力.

灵巧手是一种新兴的具身实体类型, 用来执行复杂的灵巧操作任务. 常用的灵巧手类型包括 Adroit hand, Allegro, Shandow Hand, Realman Hand 等. 灵巧手对硬件设计有着较高的要求, 包括电机、丝杠、减速器、驱动控制系统、位置传感器、力传感器、触觉控制等新型部件. 对于常用的灵巧手类型, 研究人员使用 Isaac Gym [31] 和 MuJoCo [21] 等平台进行仿真, 模拟真实世界中的灵巧操作场景, 从而避免直接使用昂贵的硬件设备所带来的风险. 通过仿真平台可以降低研究成本, 保证实验的可重复性和可控性, 帮助研究人员进行系统性的测试和比较. 灵巧手从仿真到现实的迁移是较为困难的, 在抓握任务中, 由于涉及连续的表面接触与受力, 让模拟与现实有很大的差距; 而对于如接抛物体的动态任务, 摩擦、惯性等物理特性均会增大仿真到现实的迁移难度. 部分研究直接使用实体灵巧手进行数据采集和策略学习, 利用已有的功能模型、数据集或通过遥控操作构成专家演示数据集, 或通过自主重置环境的能力进行自主操作, 降低数据采集与训练的难度与时间成本. 常用的灵巧手示教数据集包括 UniDexGrasp 抓取数据 [40]、Handversim 人类手腕轨迹模式数据集 [41]、DAPG 复杂任务演示数据 [42] 等.

人形机器人指模仿人类外观和行为的机器人, 其形态更适合人类社会场景, 具有全地形行为能力, 可以覆盖人类活动的方方面面. 人形机器人发展的重要节点包括 2000 年 ASIMO (本田) 、2013 年 Atlas (波士顿动力) 、2022 年 Optimus (特斯拉) 等, 此外国产机器人也逐步追赶, 包括宇树科技、智源机器人、傅立叶智能、优必选机器人、小米机器人等均推出了人形机器人的初代版本. 整体来看, 当前人形机器人硬件本体系统可满足人形机器人基本需求, 核心零部件国产化趋势增强. 但在机器人硬件基础初步具备的情况下, 当前人形机器人仍局限于结构化已知环境下的少量有限动作控制, 全身运动控制算法稳定性与泛化性差, 主动环境感知与人机交互能力不足. 随着大模型驱动的具身智能技术

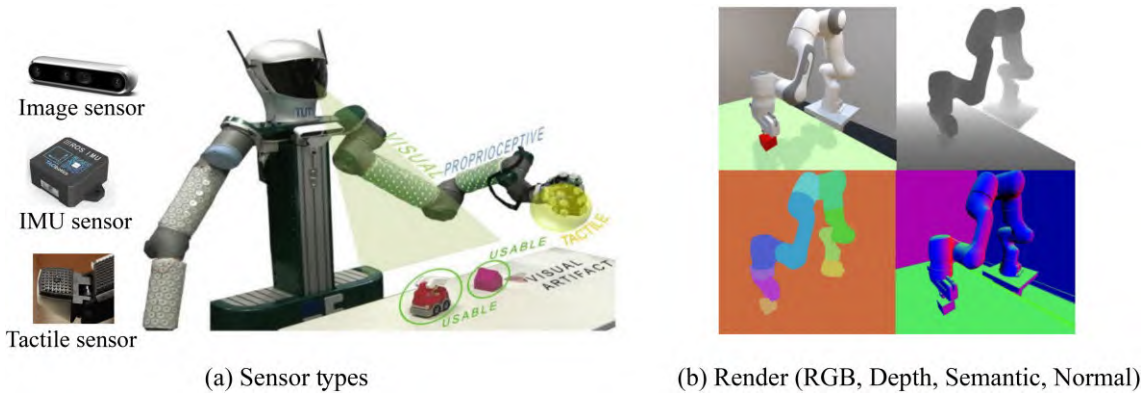


图 5 (网络版彩图) 机器人常见的传感器类型和仿真环境观测

Figure 5 (Color online) Sensor types and rendering in simulators of robots

的不断发展, 大模型将有机会赋能人形机器人的运动控制和任务规划, 使人形机器人技术成熟. 在未来, 人形机器人可以提供更好的医疗保健服务, 帮助老年人和残障人士生活, 可以替代部分工人工作, 提高工作安全性. 同时有望在军事、娱乐、教育等领域产生应用. 人形机器人将改变人类与机器的互动方式, 重新定义人机关系, 对社会结构和文化产生深远的影响.

在传感器方面, 具身机器人使用的传感器主要包括视觉传感器、机器人本体感知 (proprioception) 传感器、触觉/力学 (haptic/tactile perception) 传感器、GPS 和 IMU 传感器等, 如图 5 (a) 所示. 大模型驱动的具身智能中由于大模型主要用于处理视觉和语言, 因此对视觉传感器较为关注. 如图 5 (b) 所示, 通过视觉传感器或仿真环境中的视觉渲染功能, 可以获得 RGB、深度图、语义分割图、点云等视觉输入信息. 机器人的本体感知传感器根据机器人类型而有所不同. 例如, 机械臂的本体感知包括各个关节的位置、速度、扭矩和机械臂末端位置等, 四足机器人的本体感知信息包括自身线速度、角速度, 身体关节位置、力传感器等各种传感器信息, 动作通常设置为身体关节的目标位置等, 而灵巧手的本体感知包括关节位置、速度、力矩, 指尖位置、力传感, 手腕的位置与角度等. 触觉/力学传感器主要用于灵巧手和人形机器人手部关节中, 用于灵巧手感受抓握物体的力度, 估计物体的形状等. GPS/IMU 传感器在机器人运动中进行定位, 多用于移动机器人、人形机器人等. 此外, 由于仿真器中可以获取预设的环境参数, 仿真环境中机器人常拥有 “特权” 观测. 例如, 仿真机械臂抓取任务中可以获得准确的抓取点或物体位置, 仿真四足机器人运动中可以获取地形信息和环境摩擦参数等. 特权信息能够帮助策略在仿真环境中快速收敛, 但在真实环境中是缺失的, 需要通过额外传感器或特权信息蒸馏等架构来进行仿真 - 真实 (sim-to-real) 的策略迁移 [43].

在执行器方面, 根据机器人类型的不同, 机器人底层分别采用不同的控制器进行执行. 在机械臂中, 常见的动作输出有关节位置、δ 关节位置、末端执行器姿态和 δ 末端执行器姿态等. 通过控制器驱动相应电机来进行关节位置、速度、扭矩的调整, 从而执行相应的动作. 四足机器人的动作一般为足部关节的位置, 通过 PD 控制器来产生相应的电机扭矩. 灵巧手的动作一般为手指关节的力矩、手腕的位置、方向等, 随后转换为关节扭矩进行执行. 人形机器人拥有更多的关节和更高的自由度, 一般采用多个控制器分别控制手部关节、双臂系统、腿部关节、躯干等. 一般而言, 大模型更关注于高层的任务执行和动作选择, 底层由传统的执行器进行执行.

具身智能的任务由常见的抓取任务逐步演变为更加复杂的任务集合. 在机械臂和移动机器人中, 以 ManiSkill2 仿真平台为例, 任务集合包括精细操作任务 (如钻孔装配) 、刚性物体操作 (如 Pick-and-

place) 、铰链物体操作 (如开门、开箱等) 、柔性物体操作 (倒水、挂毛巾) 等. 其中, 铰链物体和柔性物体都具有复杂的模型, 在策略学习中需要隐式的估计位移和形变. 四足机器人学习的目标是适应不同的地形, 包括粗糙地面、沙地、石子路、台阶等 [44], 同时有研究使四足机器人学习不同的步态, 如行走、长跳、奔跑、双脚站立、跨越障碍物等 [45]. 灵巧手用于测试的任务包括转魔方、旋转物体、开门、倒水等 [46]. 人形机器人由于包含了双足运动、双臂操作、双灵巧手等系统, 可执行的任务更为多样, 几乎可以涵盖现有具身实体的所有任务集. 为了和人类进行更好的交互, 许多任务包含了语言描述的目标, 如 VIMA [47], Calvin [48] 等, 常见的语言任务描述如 “将红色方块放进抽屉”, “按下蓝色按钮打开柜子” 等, 任务的执行情况使用执行成功率和时间效率进行衡量.

具身智能任务所使用的仿真环境已经在对不同具身实体的介绍中进行了详细介绍. 目前, 仿真环境的发展仍然十分迅速, 英伟达公司近期开发的 Isaac-sim [31] 通过和 CUDA 和英伟达显卡的高度耦合, 能够通过硬件加速和高速并行对复杂的动力学模型进行快速仿真, 包括四足机器人、人形机器人等. 此外, 现有仿真器对复杂柔性物体、流体、触觉传感器的仿真仍然存在不足, 同时仿真器环境和真实环境仍然存在较大的区别, 在未来仍然有很大的发展空间. 此外, 新型具身实体不断发展, 如光动无人机 [49] 实现了对无人机的全天时智能视觉跟瞄和自主远程能量补充, 开启了无限续航无人机的探索.

2.2 具身智能学习框架

常用的具身学习框架主要包括模仿学习、强化学习、模型预测控制等. 具身智能的策略学习问题可以描述为马尔科夫决策过程 (Markov decision process, MDP), 表示为 $\mathcal{M}=(\mathcal{S},\mathcal{O},\mathcal{A},\mathcal{P},r,\mathcal{L})$. 其中, (1) $\mathcal{O}$ 作为观测空间, 用于表示机器人传感器的输入. 如前文所述, 观测空间中可能包含了视觉观测、本体感知、触/力传感、定位信息等, 取决于具体的机器人硬件设置和任务设置. 一般而言, 多种传感器之间会存在较多的信息冗余. (2) $\mathcal{S}$ 表示状态空间, 由历史的观测和当前观测进行感知学习的方法获取. 视觉传感器可以捕捉周围环境的图像和视频信息, 触觉传感器能够让机器人感知物体的形状、质地和温度, 力学传感器则可以测量机器人在执行任务时施加的力和受到的反作用力. 感知学习也包括对自身状态的感知, 机器人需要了解自身的姿态、位置和运动状态, 以便在执行任务时做出正确的决策. 一般认为状态具有马尔科夫性 (Markov property), 即机器人后续的状态将完全由当前状态和动作决定, 和历史状态无关. (3) $\mathcal{A}$ 表示动作空间, 不同机器人的动作空间定义如上节所述, 一般为机械臂末端姿态或关节位置等. (4) 状态转移函数 $\mathcal{P}(s'|s,a)$ 表示在当前状态和动作下, 智能体到达下一个状态 s' 的概率, 是由机器人动力学和环境决定的. 传统的方法一般假设已知系统的动力学方程或使用线性函数进行动力学建模, 而在复杂具身智能系统中动力学方程是较为复杂的, 往往需要强大的神经网络结构或大模型来进行构建. (5) 奖励函数 $r(s,a)$ 用于衡量动作执行后得到的下一个状态 s' 的好坏, 例如当前机械臂末端和目标位置的距离, 四足机器人的行进速度和目标速度的差距等, 一般用于强化学习和模型预测控制中. 宏观上, 奖励函数作为环境的反馈来帮助策略进行迭代, 强化学习算法通过奖励反馈进行策略优化, 以期望在未来获得更大的奖励. (6) 任务描述空间 $\mathcal{L}$ 一般使用自然语言的方式进行任务的描述, 方便使用大模型进行理解和规划. MDP 中的奖励函数和任务描述密切相关, 例如 “关闭抽屉” 和 “打开抽屉” 的奖励函数在数值上具有较大差异. 智能体策略表示为 $\pi(a|s,l)$, 策略根据任务描述和当前状态采取合适的动作.

2.2.1 模仿学习

行为克隆 (behavior cloning, BC) 是模仿学习的基本框架, 机器人的策略 $\pi(a|s,l)$ 通过模仿专家数据得到. 专家数据集表示为 $\mathcal{D}=\{\tau_i,l_i\}_{i\in[N]}$, 其中 $\tau_i=[s_0,a_0,s_1,a_1,\ldots,s_{t-1},a_{t-1},s_T]$ 代表专家轨迹,

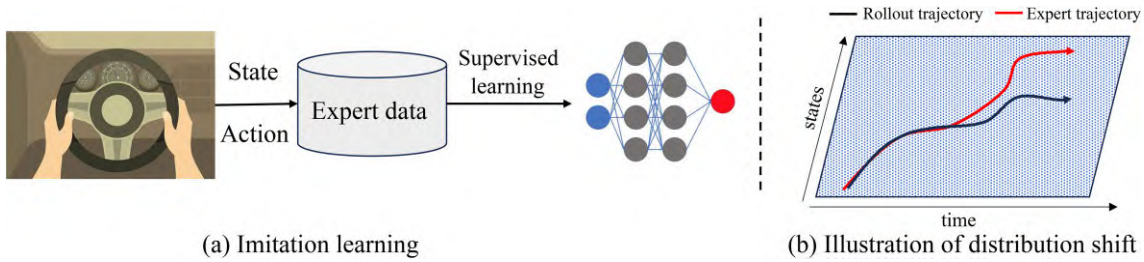


图 6 (网络版彩图) 模仿学习和分布偏移示意图

Figure 6 (Color online) Illustration of the process of imitation learning and distribution shift problem

l_i 代表任务描述. 行为克隆的损失函数表示为

$$\mathcal{L}(\theta) := -\mathbb{E}_{(\tau,l)\sim\mathcal{D}}\left[\sum_{t=0}^{T-1}\log\pi_\theta(a_t|s_t,l)\right]. \tag{1}$$

模仿学习的优势在于不需要建模环境的状态转移, 仅需要通过专家的状态 - 动作集合可以直接进行策略学习, 如图 6 (a) 所示. 模仿学习经过了数十年的发展, 有许多细分的研究领域, 下面进行简要阐述.

分布偏移问题. 模仿学习存在固有的分布偏移问题, 在模仿策略的实际执行中会导致泛化误差, 如图 6 (b) 所示. 具体地, 专家数据集 $\mathcal{D}$ 中的状态是有限的, 当智能体在任务执行时遇到的新状态 s_t^{new} 和数据集中的状态差距较大, 则模仿学习策略 $\pi(a|s,l)$ 将会产生不可预估的动作, 环境转移到下一个新状态 s_{t+1}^{new}, 交互轨迹和数据集轨迹产生分布偏移. 随着分布偏移的累积, 策略在真实环境中的轨迹将会大大偏离数据集中的专家轨迹, 从而由泛化能力不足导致任务失败. DAgger [50] 分析了这种状态分布偏移导致的问题, 并提出在策略执行过程中由专家进行介入, 对新采集的状态进行专家动作标记, 从而扩大数据对状态和动作空间的覆盖. 该思想在后续的 LazyDAgger [51], SafeDAgger [52], ThriftyDAgger [53] 中进一步发展, 这些方法显著降低了专家介入的次数.

动作分布建模. 由于人类专家的策略往往较为复杂, 产生的轨迹往往呈现出随机性和多模态. 为了建模复杂的人类专家策略, 研究人员提出使用表达能力较强的生成模型来进行策略建模. 生成对抗网络和扩散模型在最初被用于建模复杂的图像分布, 可以生成高分辨率的图像, 后被引入模仿学习中建模复杂的策略分布. GAIL [50,54] 提出使用生成对抗网络对动作分布进行建模, 使用生成器建模策略分布, 使用判别器来判断动作是否来源于专家动作. 根据生成对抗网络的损失, 生成器输出的动作将逐步接近于复杂的专家动作. 扩散生成模型近期也被用于建模人类专家或离线数据集的动作分布 [55~58], 使用多步的逆向去噪过程, 由随机噪声逐步恢复出多模态的专家动作.

无动作轨迹模仿. 在如式 (1) 所示的模仿学习损失函数中假设专家对轨迹中的每个状态都进行了动作标记. 然而, 许多机器人模仿的数据并没有记录动作, 而仅记录了机器人的轨迹. 例如, 互联网上有大量的机器人操作物体的视频, 可以认为每段视频包含一个周期的机器人视频观测. 如何从无动作标记的轨迹中进行策略模仿在近期受到关注. VPT [59] 使用逆环境模型为大量的无动作标记数据打上动作标签, 随后使用模仿学习方法进行策略学习. 其他研究通过学习智能体状态和专家轨迹之间的相似度, 将其作为奖励函数引导智能体产生和专家轨迹相似的轨迹. 已有的度量包括在图像表征空间中的相似度 [60,61], 最优传输机制导出的相似度矩阵 [62] 等, 从专家轨迹中挖掘奖励函数后使用强化学习算法进行优化. 此外, APV [63] 和 GR-1 [64] 设计了一种视频预测的结构, 从大量无动作标签的视频中构建视频预测模型.

逆强化学习. 逆强化学习旨在从观察到的智能体的行为中推断出其背后的潜在目标或奖励函数.

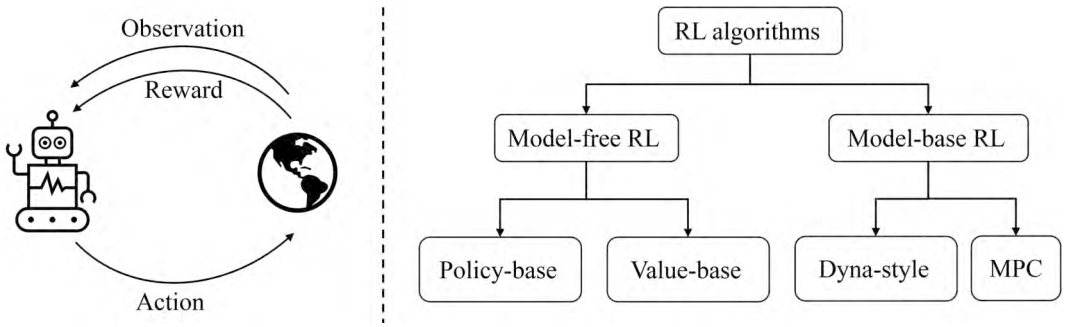


图 7　强化学习结构和基本算法分类

Figure 7　Illustration of the RL agent and the taxonomy of basic RL algorithms

早期的逆强化学习方法为边际类算法, 旨在从专家演示中学习奖励函数, 获得与专家策略相近的策略, 随后基于熵的逆强化学习算法产生了最大熵逆强化学习、相对熵逆强化学习、贝叶斯逆强化学习等 [65]. 在深度学习的背景下, 结合神经网络的感知力和强化学习的决策力提出基于神经网络的逆强化学习, 包含学徒学习逆强化学习、最大边际规划深度逆强化学习、最大熵深度逆强化学习、生成对抗逆强化学习等. 通过逆强化学习, 智能体可以从专家的行为中学习, 并在没有明确奖励信号的情况下做出智能决策, 使智能体更好地适应复杂和不确定的环境, 提高其决策能力和智能水平.

2.2.2　强化学习

强化学习的目标是通过智能体与环境的交互来最大化奖励 [14]. 在时间步 t, 智能体根据当前的环境的状态 s_t 来选择动作 a_t, 环境通过执行该动作得到下一个状态 s_{t+1}, 并反馈给智能体一个奖励信号 r_{t+1}. 定义 t 时刻的 "回报" 为从该时间步开始到周期结束的 "折扣" 奖励之和, 形式化为

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^{T-t-1} r_{T-1}. \tag{2}$$

其中, 奖励权重随着时间步以 γ 的速度衰减. 由于策略和环境存在随机性, 从 (s_t, a_t) 出发按照同一策略进行交互, 每次得到的 R_t 往往是不同. 强化学习使用值函数 (value function) 表示多次交互中的回报水平. (s, a) 在策略 π 下的值函数 $Q^\pi(s, a)$ 衡量了从该状态动作出发按照策略 π 选择动作, 在整个周期结束时获得的回报的期望.

$$Q^\pi(s, a) = \mathbb{E}_\pi[R_t | s_t = s, a_t = a] = \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\Big]. \tag{3}$$

$Q^\pi(s, a)$ 又被称为动作值函数. 对 $Q^\pi(s, a)$ 在动作上取期望, 可得状态值函数 $V^\pi(s) = \mathbb{E}[Q^\pi(s, a)]$. 在有限动作空间下, 可以根据值函数的估计选择最优动作 $a_t = \arg\max_a Q(s_t, a)$. 在估计值函数中, 值函数可以根据定义写为 $Q(s, a) = \mathbb{E}[r + \gamma \mathbb{E}_{s' \sim p(s'|s,a), a' \sim \pi(a'|s')} Q(s', a')]$, 使用立即奖励 r 和下一个时间步的 Q 值来更新当前时间步的 Q 值估计, 称为时序差分 (temporal-difference, TD) 估计法. 强化学习结构和基本算法分类如图 7 所示, 下面将介绍几种强化学习研究的算法分支, 这些分支均可以和大模型结合并在具身智能中应用.

值函数学习. Q- 学习 (Q-learning) 是一种时序差分为基础的最优值函数求解方法, 在计算 Q_{target} 时选择使 $Q(s', a')$ 最大的动作, 即 $Q_{\text{target}} = r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a')$, 随后使用 $Q(s, a) = Q(s, a) + \alpha[Q_{\text{target}} - Q(s, a)]$ 更新值函数. 深度 Q 首次将这种从状态动作到值函数的映射关系用神经网络表示,

网络的权重记为 θ. 在 Q 学习中, 根据时序差分方法的思路, 当前时刻的 Q 值可以根据立即奖励和下一时刻的 Q 值来得到. 深度 Q 学习在基本原理的基础上, 使用包括深度 Q 网络、经验池和目标网络来提升优化的稳定性 [66]. 在此基础上, 许多方法分别从解决值函数的高估问题 [67,68], 提升值函数学习的样本效率 [69,70], 提升策略的探索能力等方面进行了研究 [71,72].

策略学习. 策略梯度法是强化学习进行策略求解的另一类重要方法 [73]. 策略梯度中对累积奖励的梯度进行求解, 表示为 $\nabla_\theta J(\pi_\theta)$, 形式化为

$$\nabla_\theta J(\pi_\theta) = \nabla_\theta \mathbb{E}_{\tau\sim P(\tau|\theta)}\big[R(\tau)\big] = \nabla_\theta \int_\tau P(\tau|\theta)R(\tau) = \int_\tau \nabla_\theta P(\tau|\theta)R(\tau), \tag{4}$$

其中 $\nabla_\theta P(\tau|\theta) = P(\tau|\theta)\nabla_\theta \log P(\tau|\theta)$, 且

$$\log P(\tau|\theta) = \log p(s_0) + \sum \Big(\log P(s_{t+1}|s_t,a_t) + \log \pi_\theta(a_t|s_t)\Big). \tag{5}$$

代入后得到策略梯度的形式为 $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim P(\tau|\theta)}\Big[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)R(\tau)\Big]$. 策略梯度法引入基线来降估计的方差 [74]. 当使用 $V^\pi(s_t)$ 来作为基线时, 梯度中 $Q^\pi(s_t,a_t)$ 的项转化为 $Q^\pi(s_t,a_t) - V^\pi(s_t)$. 其中, $A^\pi(s_t,a_t) = Q^\pi(s_t,a_t) - V^\pi(s_t) = r + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)$ 定义为优势函数 (advantage) [75], 策略梯度为 $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta} \sum_{t=0}^{T}[A_\phi(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)]$.

近端策略优化 (proximal policy optimization, PPO) [76,77] 是从基本的策略梯度法中发展而来, 具有较好的稳定性和广泛的应用. PPO 引入需要描述两个策略 $\pi_{\theta_{\text{old}}}$ 和 π_θ, 其中, $\pi_{\theta_{\text{old}}}$ 是上一个时间步的策略, π_θ 是当前策略. PPO 一方面要求 π_θ 相对于策略 $\pi_{\theta_{\text{old}}}$ 能够有稳定的性能提升, 另一方面要求 π_θ 和 $\pi_{\theta_{\text{old}}}$ 之间的距离不能太远. PPO 算法在具身智能任务和大模型中有着广泛的应用.

基于模型的方法. 基于模型的方法希望智能体具有规划的能力, 在做出行为之前能够提前预测执行该行为会带来的后果, 从而快速找到最优动作 [78]. 基于模型的方法需要建立环境模型, 该模型是利用真实交互数据拟合而来的, 模型输入状态动作对 s,a, 输出 s' 的预测. Dyna-Q 是一种典型的基于模型的强化学习方法 [79], 假设环境是确定的并且状态动作空间是离散可数的, 这样可以把环境模型想象成一个类似 Q 函数的表格, 利用环境模型产生了虚拟样本进行额外的 Q 函数损失, 加速了 Q 函数的收敛. 基于模型方法的主要问题是, 模型预测的误差会在模型向后推演中累积, 从而对值函数学习产生负面影响. MBPO 等方法分析了误差累积导致的性能差距 [80,81], IVE 算法分析了模型预测过程的不确定性 [82], Dreamer 算法通过隐空间的环境建模进行多步预测和推理 [83,84].

离线强化学习. 强化学习中智能体与环境的交互代价较高, 同时具身智能任务中许多场景很难进行在线的样本收集. 解决该问题的一个思路是利用已有的数据集进行学习, 称为 "离线" 强化学习 [85], 数据集可以来自于其他智能体采集的数据或由人类采集的数据. 然而, 由于数据集所含的数据量和覆盖范围是有限的, 无法覆盖真实世界中的状态空间和动作空间. 离线强化学习致力于解决使用固定数据集来进行策略学习中出现的问题. 现有方法通过策略约束 [86,87]、不确定性估计 [88,89]、悲观值函数等思路 [90,91], 限制策略对离线数据集所覆盖以外的动作的访问来解决该问题.

2.2.3 模型预测控制

在模型预测控制 (MPC) 中, 环境模型可以使智能体无需与环境交互而得到下一步状态和奖励. 这样的决策过程与人类相似, 人类在做出决策之前, 通常会思考做出这步决策会带来的后果, 根据不同的后果来选择最有利的决策. 智能体同样可以利用环境模型来进行推演, 从而选择出最优的动作, 即

找到如下动作序列:

$$(a_t, ..., a_{t+H-1}) = \underset{a_t,...,a_{t+H-1}}{\arg\max} \sum_{t'=t}^{t+H-1} \gamma^{t'-t} r(s_{t'}, a_{t'}). \tag{6}$$

随后, 智能体将执行第一步的动作 a_t. 执行后再次使用规划方法得到下一步的最优规划动作.

随机打靶法. 随机打靶法 (random shooting) 是一种通过随机采样的方式帮助智能体决策的方法, 给出当前状态 s_0 和长度为 T 的随机动作序列 $[a_0, a_1, a_2, ..., a_T]$, 利用训练的环境模型可以得到仿真轨迹 $[s_0, a_0, \hat{r}_0, \hat{s}_1, a_1, \hat{r}_1, \hat{s}_2, a_2, \hat{r}_2, ..., \hat{s}_T, a_T, \hat{r}_T]$. 通过多次采样随机动作序列 $[a_0, a_1, a_2, ..., a_T]$, 可以得到不同的状态动作对的值函数估计, 记为 $\hat{Q}(s, a) = \sum_{t=0}^{T} \gamma^t \hat{r}_t$, 然后根据 $\pi(s) = \arg\max_a \hat{Q}(s, a)$ 选择下一步的动作. 智能体执行动作 $\pi(s)$ 之后, 再次重复采样过程进行规划. 随机打靶法的缺点是不同轨迹方差较大, 可能无法采样到高回报的动作. 有研究提出使用交叉熵方法 (cross-entropy method, CEM) 进行采样 [92]. CEM 不使用随机的动作序列, 而是从某个动作分布中采样动作序列, 根据动作取得的累积回报调整动作采样的分布, 这样有更大概率采样到高回报的动作序列.

集成概率轨迹采样法. 集成概率轨迹采样法 (probabilistic ensembles with trajectory sampling, PETS) 是对随机打靶法的改进 [93], 通过将不确定性感知的概率环境模型和轨迹采样相结合, 实现了和无模型强化学习方法接近的效果. PETS 可以衡量任意不确定性 (aleatoric uncertainty) 和认知不确定性 (epistemic uncertainty), 其中任意不确定性是由随机系统本身带来的不确定性, 比如观测噪声或状态转移噪声; 认知不确定性是指由于数据缺失带来的不确定性, 可以随着训练数据量的增加而减少. 实现上, PETS 使用了集成概率环境模型方法, 建立了若干个概率环境模型, 每个环境模型使用参数化高斯分布来拟合. PETS 使用基于 CEM 的随机打靶法的方式进行规划, 可以区分两种不确定性, 任意不确定性使用一条轨迹的预测方差来衡量, 而认知不确定性通过多条轨迹之间的预测方差得到, 从而对环境未来的预测更加全面.

2.2.4 具身框架和认知智能的联系

认知智能是指人类在感知、理解、记忆、推理、判断和解决问题等方面的能力, 涉及从外部环境中获取信息, 通过大脑处理这些信息并形成对世界的认识 [94]. 具身智能理论强调, 认知过程不仅仅依赖于大脑, 而是与整个身体及其与环境的互动密切相关. 身体的动作、感知和环境的交互对认知能力的发展和实现具有重要的影响和作用. 在具身智能的框架下, 认知智能通过身体与环境的动态交互形成. 例如, 触觉和运动感知帮助人们理解空间关系和物体属性, 这些直接的感知体验进一步影响复杂的认知过程, 如推理和决策. 在具身智能的框架下, 认知智能不仅仅是大脑的产物, 而是通过身体与环境的动态交互形成的. 人类通过身体的感知与运动获取信息, 这些信息不仅帮助形成对周围世界的理解, 还影响思维和判断过程.

2.3 大模型技术

2.3.1 大语言模型

语言模型从统计学的角度计算词序列 $w_1, w_2, \ldots, w_m$ 的联合概率分布 $p(w_1, w_2, \ldots, w_m)$. 为了降低联合概率分布计算的难度, n-gram 语言模型按照从左到右的顺序对联合概率分布进行链式法则分解, 将第 i 个词的条件概率表示为 $p(w_i|w_{i-n+1}, \ldots, w_{i-1})$. 在语言模型的发展中, 先后使用前馈神经网络、循环神经网络、卷积神经网络、记忆网络等结构条件概率分布进行建模, 有效地挖掘词之间的长序列依赖关系 [95, 96]. 2018 年提出的 Bert 语言模型 [97] 使用双向 Transformer 结构, 设计了掩码预测和未来句子预测等自监督预测任务进行语言模型的预训练, 在下游的自然语言标准任务中获得了巨

大提升. 2020 年 OpenAI 提出 GPT-3 模型 [98], 使用更多的参数量进行模型的预训练, 将上下文长度扩展到 2048 词元 (token) , 每个注意力层包含了 96 个注意力头. GPT-3 学习得到的模型具有零样本的泛化能力, 刷新了许多自然语言处理的标准任务中的成绩.

近年来, 大语言模型进行了快速发展. OpenAI 提出了 ChatGPT, 获得了巨大成功 [6], 根据已经公开的技术细节, ChatGPT 的训练过程主要包含三个阶段. 首先, 进行基础语言模型的预训练, 根据大量预料构建无监督的预测任务, 进行长序列模型预测训练; 随后, 使用人工标注数据和开源数据进行指令微调 (instruction tuning), 使 GPT 的输出能够符合问答习惯, 同时能够完成多种任务; 最后, 使用人类偏好数据训练奖励模型, 使用强化学习中的 PPO 策略梯度算法对语言模型进行优化, 使模型的输出和人类偏好进行对齐. ChatGPT 通过大量数据的预训练获得上下文的理解能力和对现实世界的常识知识, 通过指令微调能够根据问题来解决不同的任务, 通过强化学习优化训练能输出符合人类偏够好的结果, 避免了有害回答的产生. 随后, OpenAI 发布 GPT-4 模型, 提升了模型的上下文长度, 同时具备了多模态理解能力. GPT-4 能够识别和提取图像信息, 回答更加专业的领域知识, 进行数学推理和编程等, 为实现通用人工智能迈出了重要的一步. 此外, 常用的开源大模型还包括 LLaMA [99,100], Vicuna [101], Phi-2 [102] 等. 近期, 国产的大语言模型也纷纷发布, 包括上海人工智能实验室书生通用的大模型、百度文心大模型、百川智能百川大模型、阿里通义千问大模型等, 这些模型具有更好的中文理解能力.

2.3.2 视觉基础模型

视觉基础模型一般以卷积神经网络 (如 ResNet 等 [103]) 或者视觉 Transformer (ViT) 等 [104] 为基础模型, 通过自监督学习的方式提取图像的特征表示, 随后将特征提取器用于下游任务中. 早期的视觉特征提取器使用分类任务进行预训练, 在许多下游任务中取得了不错的效果. MAE [105] 提出自监督的学习方式, 使用掩码后的视觉输入来还原原始的图像, 可以从大量无标记的图像中进行预训练. CLIP [9] 使用大量配对的图像和文本描述来学习图像的语义特征. 首先, 分别使用图像编码器和文本编码器得到图像和文本特征; 随后, 使用对比学习的目标 [106], 在特征空间内最大化配对的图像 - 文本特征的相似度, 同时增大不配对的图像 - 文本在特征空间中的距离, 获得和文本理解匹配的视觉特征. 在此基础上, 许多工作分别从训练数据 [107]、模型架构 [108]、损失函数 [109]、生成式辅助任务 [110] 等方面进行改进 [111]. Segment Anything (SAM) 是近期提出的视觉分割大模型 [112], 从给定语言或视觉提示从图像中分割出对应的区域. SAM 使用 1100 万张图像的大规模数据集上进行预训练, 具有很强的语义理解和泛化能力. 例如, SAM 能通过少量数据微调解决医学图像分割问题 [113]. 近期有研究使用 SAM 获得对具身任务的场景分割, 提升对具身场景的理解能力 [114].

2.3.3 视觉 - 语言模型

视觉 - 语言模型 (VLM) 同时融合了大语言模型和视觉基础模型, 使模型能够同时接收图像和语言作为输入, 并根据语言指令和图像信息产生输出, 进行图像问答任务. SimVLM [115] 提出了一种编码 - 解码的结构, 将图像和文本输入编码器, 从解码器中输出后续文本或回复. BLIP2 [116,117] 提出了一种典型的 VLM 架构, 使用预训练的图像编码器对图像提取特征, 随后使用提出的 QFormer 结构从冻结的编码器中提取视觉特征. QFormer 使用多模态信息匹配的方式进行预训练, 从而对齐视觉和文本表征. 提取的特征经过线性投影后和语言指令作为大语言模型的输入. Flamingo [12,118] 使用类似的结构, 使用少量例子作为输入使模型具备少样本泛化能力. 图 8 对视觉基础模型和 VLM 的结构进行了对比. 在此基础上, LLaVa [119], Mini-GPT4 [120], instructBLIP [121] 等模型分别从模型结构、预训练任

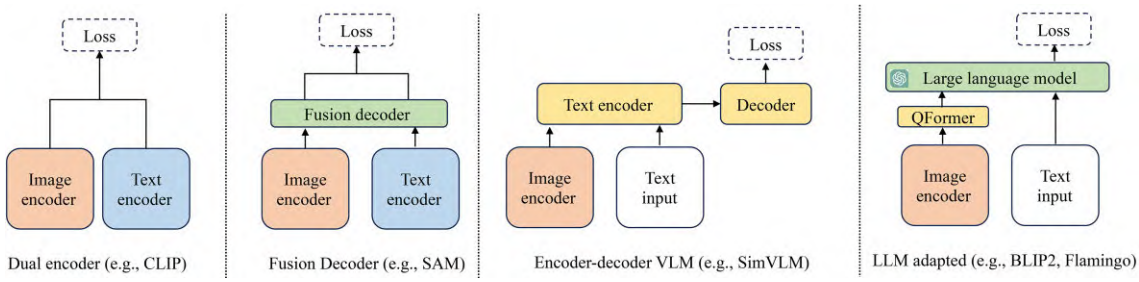


图 8 (网络版彩图) 视觉基础模型和视觉 - 语言模型框架的典型架构示意图

Figure 8 (Color online) Comparision of basic architectures in visual foundation models and visual-language models

务、语言基础模型等角度进一步提升了 VLM 的理解和推理能力. Video-Chat [122], VideoLLaMa [123] 等将图片输入扩展为视频输入, 使大模型能够根据语言指令和视频输入进行问答. 在训练中使用类似 QFormer 的结构进行模态转换, 同时引入音频来对齐视听信号, 增强对视频的整体理解. 然而, 由于视频信息的复杂性, VideoLLaMa 在处理长序列视频中仍然存在困难, 同时音频 - 视频 - 文本对齐的高质量数据集规模有限, 在未来将会有进一步发展.

2.4 生成式大模型

扩散模型是一种重要的图像生成模型 [124], 在生成高质量图像方面的能力超出了传统的自编码器 [125]、显式概率模型网络 [126] 和对抗生成网络 [127]. 将真实图像记为 x^0, 扩散模型使用参数化的扩散过程函数 $p_\theta(x^0) = \int p(x^T) \prod_{t=1}^{T} p_\theta(x^{t-1}|x^t)|\ x^{1:T}$, 使用去噪过程将高斯噪声 $x^T = \mathcal{N}(\mathbf{0}, \mathbf{I})$ 通过 T 个时间步转化为原始图像 x^0. 去噪序列 $x^{T:0}$ 是一个马尔可夫链, 每一个时间步产生的中间图像用参数化的高斯分布表示: $p_\theta(x^{t-1}|x^t) = \mathcal{N}(\mu_\theta(x^t, t), \Sigma(x^t, t))$. 相反, 前向加噪过程使用固定参数的分布逐步添加高斯噪声: $x^t = \sqrt{\alpha^t} x^{t-1} + \sqrt{1-\alpha^t}\epsilon_t$, 其中 $\alpha^t = 1 - \beta^t$, $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. 根据贝叶斯定理, 可以推导逆向过程的解为

$$q(x^{t-1}|x^t) = \mathcal{N}\left(\frac{1}{\sqrt{\alpha^t}}(x^t - \frac{\beta^t}{\sqrt{1-\bar{\alpha}^t}}\epsilon(x^t, t)), \beta^t \frac{1-\bar{\alpha}^{t-1}}{1-\bar{\alpha}^t}\right) . \tag{7}$$

其中, $\bar{\alpha}^t = \prod_1^t \alpha^i$. 扩散模型通过最大化概率的变分下界 (ELBO) 训练 [128], 损失函数可以简化为

$$\mathcal{L}(\theta) = \mathbb{E}_{x^0,\epsilon,t}[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}^t}x^0 + (\sqrt{1-\bar{\alpha}^t})\epsilon, t)\right\|^2] , \tag{8}$$

其中 ϵ_θ 为参数化的网络模型. 扩散模型已经在可控图像生成 [129]、文本到图像生成 [130] 等方面, 并有 DALL-E [131], Sora [132] 等在图像和视频生成方面的成功应用. 在具身智能领域, 扩散模型已经被用于强化学习策略学习 [57]、动作规划 [133, 134]、图像目标生成 [135]、轨迹生成 [136] 等方面有成功应用.

多模态生成大模型 (multimodal generative models) 能够处理和生成多种类型的数据 (例如文本、图像、音频、视频等), 从而实现更加丰富和复杂的内容创作. 例如, DALL-E 模型通过大规模预训练, 学习了文本和图像之间的复杂关系, 展示了从文本描述生成高质量图像的能力 [131]. 另一个典型模型 Video-LaVIT [137] 通过自回归方式预测下一个图像或文本词元, 在统一的生成目标下同时处理图像和文本. 同时, 该模型提出了动态长度的离散视觉词元, 减少了图像块之间的相互依赖性, 增强了在大型语言模型中图像和文本表示的兼容性. 多模态生成大模型可以用于模拟复杂的环境和任务场景, 为具身智能体提供丰富的训练数据, 提高其在真实环境中的适应性和鲁棒性.

3 大模型驱动的具身环境感知

3.1 图像观测特征学习

在具身智能任务中, 智能体对周围环境的观测一般以视觉信息呈现. 在模仿学习和强化学习任务中, 需要根据任务指令和场景信息来从视觉观测中提取与任务和环境相关的特征. 在具身智能框架中, 传统的视觉观测的特征提取方法主要包含以下四个方面: (1)**数据增广**. 在进行强化学习的值函数或策略学习中, 使用图像裁剪、移动、对抗增强等方式 [138,139] 进行数据增广, 对原有图像状态进行扩充 $s \to \{s_{\text{aug}}\}$. 在强化学习和模仿学习的训练过程中, 使用数据增广能够使 $V(s)$ 和 $V(s_{\text{aug}})$ (或 $\pi(s)$ 和 $\pi(s_{\text{aug}})$) 具有相似的值, 从而增强值函数或策略函数在状态邻域内的平滑性. (2)**对比学习**. 对比学习 [140] 是一种自监督视觉特征提取方法, 在具身智能中常使用智能体轨迹构建正负样本. 例如, 现有方法使用状态的图像增强 [141]、相邻时间步的状态观测 [142]、环境状态转移 [143,144] 作为正样本, 使用随机采样的两个状态作为负样本, 从而学习轨迹时序或状态转移相关的特征. (3)**环境模型**. 通过当前状态表征和动作来重建下一个状态表征, 能够提取与环境转移相关的表征, 而忽略背景等无关的要素 [145,146]. 互模拟 (bisimulation) 算法 [147] 构造了一个表征空间, 希望在表征空间中两个状态的特征距离正比于奖励映射后的距离和环境转移模型映射后的距离, 能够在理论上提升强化学习的样本效率. SimSR 将互模拟表征扩展到离线强化学习中 [148]. (4)**值函数学习**. 相比于单步环境模型的预测, 值函数能获取较长时中未来的奖励的环境转移信息, 通过从离线数据集中学习值函数的预测能得到有意义的视觉表征 [149]. Successor 表征 [150] 隐式的使用值函数学习估计长期环境转移导致的状态概率密度. 虽然以上方法通过表征学习能够提升样本效率, 但仍然存在以下不足: (1) 在新的环境和任务中, 需要使用这些表征学习的目标进行重新训练, 具有较高的计算的代价; (2) 表征主要针对特定任务和环境, 在多任务、多样化环境中的泛化能力差; (3) 由于学习表征所用的训练数据有限, 表征的鲁棒性较差, 容易受环境扰动的影响.

为了解决上述问题, 近期有方法引入视觉预训练模型来提升表征的泛化性和鲁棒性, 基本结构如图 9 所示. PIE-G 算法 [151] 提出将预训练的 ResNet 编码器对视觉观测进行编码, 使用编码后的向量学习值函数和策略, 大大提升了策略的视觉泛化能力. PVR [152] 对自然图像中学习得到的监督和自监督表征在多种强化学习框架下进行了对比, 结果表明使用预训练表征能够在大部分场景中提升具身策略的效果. ViGen [153] 对比了传统的表征学习方法和 PIE-G 在机械臂操作、导航、灵巧手操作中的性能, 并使用背景、光照、相机视角、场景结构变换的方式的测试视觉编码器的鲁棒性, 结果表明预训练表征在能够在平均水平上提升策略的鲁棒性. 在常规的图像预训练数据集 (如 ImageNet [154] 等) 之外, 具身智能的表征学习中常用第一人称视角的人类物体操作数据集. 主要原因是人类操作的视频数据量大, 且涵盖了人类日常生活中场景的物体类型和交互方式, 同时具有语言标注. 人类第一人称示教视频和机械臂操作视频有很大的相似性, 容易学习到可迁移的表征, 常用的人类操作数据集有 Ego4D [155], Epic-Kitchen [156], Something-something [157] 等. R3M [158] 是一个专为机械臂操作任务设计的预训练视觉提取器, 使用 ResNet 作为基本结构, 使用视频序列的时序关系和语言 - 视频的相关关系构造对比学习损失. R3M 在大规模 Ego4d 数据集中进行训练, 得到的预训练模型能够和模仿学习或强化学习算法集合, 直接用于机械臂场景决策. 在常用的机械臂操作任务集合中, R3M 能够获得超越 CLIP [9], MVP [159] 等预训练模型的效果, 表明 R3M 能够从人类操作视频中能够提取机械臂操作相关的可迁移知识. VIP [160] 在人类数据集上学习表征, 将表征学习问题转换为目标导向的强化学习问题, 通过对比学习实现了隐式的值函数学习, 在下游任务中能同时产生状态表征和值函数估计. Voltron 模

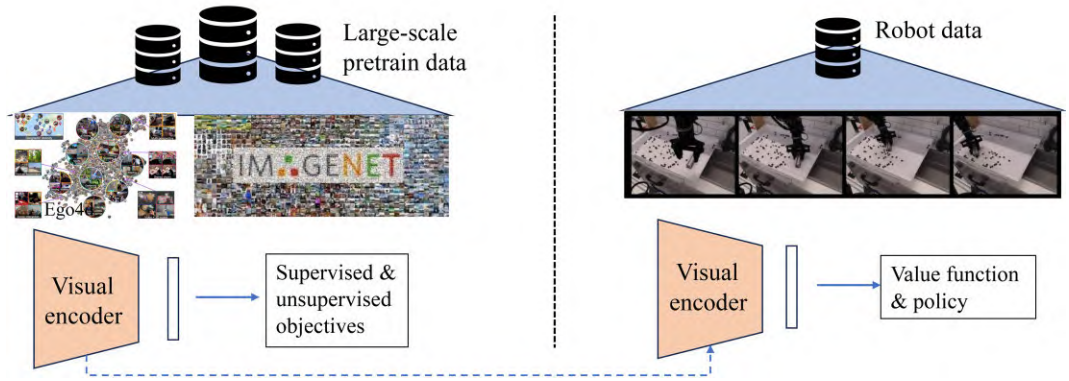


图 9 (网络版彩图) 预训练表征用于具身决策示意图

Figure 9 (Color online) An illustration of pre-traind perception module for embodied decision-making

型 [161] 使用语言描述和相应的人类操作视频作为输入, 使用 MAE 学习目标从掩码图像中恢复原始图像, 从中学习低层次的语义特征. 同时, Voltron 使用掩码上下文图像信息预测视频对应的文本描述, 从而获取高层的人类意图的相关特征. Huo 等提出了人类导向的表征学习方法 [162], 从人类交互数据中进行状态变化分类、时序定位、目标检测、手部检测等任务的构造, 通过学习这些辅助任务来获得有意义的表征. RoboMP^2 算法 [163] 通过微调多模态大模型实现基于物体属性、空间关系和知识推理的物体定位, 从而根据定位产生规划.

在具身表征学习中, 已有论文对现有基准方法进行了对比和分析. ExploreVisual [164] 在不同的训练数据、模型架构、预训练目标上得到的表征对下游任务在策略学习中的影响, 结果表明: (1) 人类交互数据 (如 Ego4D) 相比于自然图像能够更好的挖掘和具身决策相关的特征; (2) ResNet 相比于 ViT 结构能够更好的保留能够在下游任务泛化的特征; (3) 图像增强和对比学习相比于其他自监督任务能够更好的挖掘具身场景特征. 在此基础上, 提出了 Vi-PRoM 算法, 结合对比学习、视觉语义预测和动力学预测获得可泛化的具身特征提取器. Hansen 等 [165] 对比了现有大模型驱动的视觉特征预训练方法和直接策略学习方法在不同背景、光照、颜色变化下的视觉泛化能力, 结果表明使用视觉预训练方法虽然能够提升学习效率, 但相比于直接策略学习方法在最终性能上并没有明显的优势, 表明挖掘对具身智能体具有更强泛化能力的表征仍然具有很大的挑战.

3.2 Affordance 提取

预训练的视觉表征虽然能够在一定程度上提取上提升策略的泛化能力, 但紧凑的表征往往缺乏可解释性, 很难理解表征中编码了哪些信息. 机器人的 Affordance 是一种对操作任务更具有解释性的通用特性. 具体地, Affordance 一般指机器人和物体交互时, 物体表现出的 “怎样使用” 的性质, 即 “被交互的物体应该怎么使用”, 代表对于其功能的视觉提示, 比如 “茶壶手柄是被握着的” “门是从外向里推开的” 等等. 在机器人场景中, 获取物体的交互方式能够给机器人更加直观的提示, 使机器人在交互中遵循 Affordance 的提示, 避免了完全通过尝试解决任务. 在介绍 Affordance 提取之前, 现有方法也从更粗粒度的角度获取场景的相关知识. 例如, VIMA [47] 使用预训练的目标检测器对场景中的物体进行检测, 随后将分割出的物体区域进行编码作为策略的输入. 使用目标检测结果能使策略直接关注到所抓取物体的信息, 不需要通过模仿学习损失进行目标相关的特征提取. Instruct2Act [166] 使用了最新开放物体检测模型 (包括 SAM [112], CLIP [9] 物体检索等) , 根据任务指令从场景中对相应的物体区域进

图 10 (网络版彩图) 典型的 Affordance 表示方式 [169, 170, 174]

Figure 10 (Color online) Representative affordance representation in existing methods

行检索, 随后通过坐标关系获取物体在三维空间中的位置, 使用机械臂直接进行抓取. 然而, 这些方法仅获取了粗粒度的物体位置信息, 而缺乏 "怎样使用" 物体这些细粒度的信息, 在复杂功能物体和铰链物体操作中存在不足.

现有研究将 Affordance 作为一种细粒度的物体功能描述, 一般包含两个方面: 物体的交互位置 (如门把手, 茶壶手柄等) , 物体的交互轨迹 (如向内推、向上提等) . Affordance 描述对于复杂功能物体或铰链物体操作是非常重要的提示, 可以直接指导机械臂的行为. Nagarajan 等 [167] 提出, 由于人类操作视频中广泛存在人类和物体的交互模式, 可以考虑从人类操作视频中挖掘物体的 Affordance. 具体地, 该方法根据人类操作视频构建预测模型来预测一系列动作 (如可推拉, 可拿起, 可倾倒等) , 随后将预测模型用于估计非活动对象是如何进行交互的. TSC 算法 [168] 通过在人类操作视频中检测人手、交互对象和接触状态来提供监督信号, 随后屏蔽手部位置构建预测模型对监督信号进行建模, 从而对静止操作物体产生接触点和作用方式的预测. OCT 算法 [169] 提出从人类交互视频中同时学习人手轨迹和人手与物体未来的接触点, 并提出了一种自动化方法来对大量视频数据进行标记, 构建了以目标为中心的 Transformer 结构来对这些信息进行建模. VRB 算法 [170] 使用类似的思路, 从人类第一视角的数据中预测人手和物体的接触点, 以及在接触后对物体操作的方向, 并将这些知识成功的迁移到实体机器人中. 在机器人任务中, Mimicplay 算法 [171] 通过人类示教数据来检测人手操作物体的轨迹, 并将该轨迹作为隐式的规划结果迁移到机器人策略中, 从而引导机器人解决长周期的复杂决策任务. 此外, 还有一些研究利用人手运动轨迹和人手关键点等信息引导机器人决策 [172, 173]. ATM 算法 [174] 构建了更加精细的轨迹预测模型, 该方法不仅检测视频中机械臂/人手和物体接触点的运动轨迹, 而可以直接预测视频帧内任意点的未来轨迹, 从而为机械臂操作提供更详细的控制指导, 用最少的动作标记数据学习稳健的视觉运动策略. 图 10 可视化了代表性的 Affordance 信息.

此外, 在机械臂任务中如果能获取 6D 抓取位姿, 则可以通过机械臂自身的逆动力学解算来驱动机械臂夹爪产生相应的精确 6D 姿态. GraspNet [175, 176] 提出了一个大规模 6D 姿态抓取数据集, 包含 190 个杂乱和复杂的场景, 对总计 97280 张图像中的所有对象进行了精确 6D 姿势和密集抓取姿势的标记, 包含 88 个物体和超过 11 亿个抓握姿势. 在大规模数据集的基础上, 可以通过监督学习方法估计 6D 抓取位姿, 从而有效的结合机械臂自身底层规划控制来完成抓取任务.

3.3 3D 视觉表征提取

在机械臂、移动机器人、人形机器人任务操作中, 仅使用单个相机观测产生的 2D 图像作为输入往往难以处理复杂场景中的抓取问题. 在机械臂操作任务中, 需要获取操作物体的的 3D 场景特征, 包括目标位置、方向、遮挡、物体之间的堆叠关系等. 为了解决该问题, 现有研究中使用多个 RGB-D 相机观测还原场景的三维场景, 从而更好的表示复杂场景信息. 然而, 目前在大模型领域, 3D 大模型的研究仍处于起步阶段, 尚未有在具身智能领域的典型应用.

表 1 具身环境感知内容架构和代表算法

Table 1 Category and representative algorithms for embodied perception

Content structure	Category	Representative algorithms
Traditional method	Data augmentation	S4RL [138], DrQ [139]
	Contrastive learning	CURL [141], TCN [142], TPC [143], DB [144]
	Environment model	Bisimulation [147], SimSR [148]
	Value function	V-PTR [149], Successor Feature [150]
Visual representation	Natural-image based	ViGen [153], PIE-G [151], CLIP [9], MVP [159], PVR [152], RoboMP^2 [163]
	Human-data based	R3M [158], VIP [160], Voltron [161], Human [162], Vi-PRoM [164], LfS [165]
Affordane learning	Objects & contact	VIMA [47], Instruct2Act [166], Nagarajan et al. [167], TSC [168]
	Trajectory & plan & pose	OCT [169], VRB [170], Mimicplay [171], ATM [174], GraspNet [175, 176]
3D representation	3D Feature	Q-attention [178], Peract [179], PolarNet [180], SGR [181], FlowBot3D [182]
	Multi-view representation	where2act [185], Wu et al. [186]

在 3D 观测的机器人抓取中, 一般使用 3D 数据从头开始进行策略学习. 3D 输入相对于 2D 图像输入具有更高的复杂度, 在输入中具有更高的维度. 例如, 对于 100×100 的图像输入, 扩展为 3D 表示后输入维度为 100^3, 因此需要开发适用于 3D 特征提取的参数高效的网络结构. Deepmind 提出的 Perceiver-IO [177] 是一种参数高效的网络结构, 在低维的隐空间中计算注意力, 使用隐向量进行解码操作, 具有更少的计算和内存需求. Q-attention [178] 和 Peract [179] 使用 Perceiver 网络对 3D 体素 (voxel) 的信息进行特征提取, 通过策略的模仿学习损失进行训练, 在 3D 抓取任务中取得了很好的效果. PolarNet [180], SGR [181], FlowBot3D [182] 等方法使用还原的 3D 点云作为输入, 随后通过 PointNet [183, 184] 点云特征提取网络进行特征提取, 使用策略模仿学习损失进行训练. 在 Affordance 提取方面, where2act [185] 对机器人抓取中的 3D 点云输入进行特征提取, 为每一个点推断是否是功能部位 (如抽屉把手) 以及如何对该功能部位进行操作 (如垂直把手拉动). Wu 等 [186] 考虑 3D 场景的遮挡问题, 引入对比学习设计机器人目标条件遮挡场, 在学习中忽略不重要的物体, 同时保持对重要场景的敏感度, 可以推广到具有许多新颖遮挡物组合的场景. 算法总结如表 1 所示.

4 大模型驱动的具身任务规划

4.1 开环和闭环反馈

大语言模型在大规模语言数据中训练后, 对现实世界环境和机器人任务具有丰富的高层次先验知识. 例如, 在服务机器人中, 面对 “如何倒一杯牛奶” 的任务, 大语言模型能够对任务进行分解并给出规划: 1. 从冰箱中拿出牛奶; 2. 拧开牛奶瓶的盖子; 3. 找一个杯子; 4. 把牛奶倒进杯子中. 随后, 机器人可以根据底层技能库对大模型任务规划进行顺序执行, 从而完成整体的任务. 基于大语言模型的具身任务规划有以下的特性: (1) 大语言模型能够利用思维链 [8] 能力, 对复杂任务进行逐步分解; (2) 大语言模型生成的规划是自然语言描述的, 需要有执行语言指令的机器人底层技能库; (3) 大语言模型需要对场景信息有充分的描述并作为 LLM 的提示 (prompt), 需要充分的上下文信息来指导决策. 微软推出的 ChatGPT for Robotics [187] 是大模型用于机器人规划的典型应用. 该方法在 prompt 中对任务信息和机器人能够执行的操作函数进行详细描述, 包括每个机器人底层函数的定义和能够执行的任务. 大语言模型根据任务描述和底层技能库依次调用相应的函数来完成任务, 能够在机械臂、无人机

等实体上应用. Wenlong 等 [188] 评价了几种大语言模型在 VirtualHome 中的任务规划能力, 并设计了规划校正方法避免产生的规划不可执行. 结果表明, 在 prompt 中给定任务描述和任务规划的例子能够有效的提升大语言模型的规划能力, 帮助对新任务产生合理规划. ReAct 算法 [189] 提出, 在进行任务规划中, 大语言模型在输出动作的同时输出显式的推理过程, 将推理过程转变成想法显式的作为输入. ReAct 的思考过程符合思维链的原理, 同时使动作的决策更具有可解释性. 同时, ReAct 指出, 使用部分最优轨迹对模型进行微调能够显著的提升规划能力. 这些方法一般被认为是开环 (open-loop) 任务规划的方法, 主要依赖 prompt 来描述场景信息, 使用思维链的进行任务规划. 然而, 在具身任务中, 上述方法容易出现大模型规划和现实世界不匹配的问题. 例如, 当询问大模型该如何打扫卫生时, 大模型的回答的第一步很可能是 "去拿吸尘器", 而可能屋子可能只有扫帚可以作为清洁工具, 导致大语言模型的规划无法执行.

开环任务规划中存在的问题能够通过闭环反馈的方法进行解决, 现有的解决方法主要分为 3 类: 大模型自我反馈、环境反馈、值函数反馈. 下面进行分别介绍.

大模型自我反馈是一种新颖的方法, 引入额外的大模型评价器 (Evaluator) 来对大模型规划器 (Planner) 的规划进行评价. Self-Refine [190] 使用 Planner 来对输出的规划进行反馈, Planner 根据反馈结果对上次规划结果进行改进. 在结构方面, Self-Refine 并不需要引入额外的大模型, 通过多轮迭代能够不断提升规划的合理性. Reflexion [191] 设计了多种反馈机制, 其中包括使用大语言模型对自身规划的反思和评价, 同时设计了长期记忆模块将评价的结果进行存储. Reflexion 的过程非常类似于强化学习, 大语言模型规划器可以看作 Actor, 而评价器作为 Critic. 与强化学习的区别在于自我反馈方法不需要进行模型参数更新, 而是利用大模型的能力根据 prompt 反思来进行策略更新, 该方法在 AlfWorld 具身规划任务中达到了更好的效果. BeamSearch 算法 [192] 将大模型反思引入随机集束搜索中, 使用集束搜索解决长序列决策问题. 在每个时间步, 该方法引入大语言模型进行反馈, 从中选择当前最好的搜索结果进行扩展, 从而提升集束搜索的效率. SwiftSage 算法 [193] 同时使用了小语言模型和大语言模型, 在给定的任务规划示教数据集上使用离线模仿学习的方法直接模仿得到策略, 从而进行快速决策. 同时, 在进行最终决策时, SwiftSage 设计了启发式算法灵活选择小模型或大模型进行决策. ViLA [194] 引入 GPT-4V 作为机械臂操作中的规划器. GPT-4V 可以接收视觉输入, 识别场景中物体的堆叠关系、空间结构、环境约束等, 从而作为视觉反馈提升大语言模型的决策能力. 然而, 根据自我反馈进行策略改进过度依赖于大模型自身的能力. Deepmind 在最近的研究中指出 [195], 大模型自我反馈有一定的局限性, 使用自我反馈进行错误纠正特别依赖于对反馈和 prompt 的精心设计, 且纠错效率有限.

外部环境反馈是另一种高效的规划迭代方法. 基本原理是, 使用大模型产生的规划与外部环境进行交互, 由外部环境对规划执行的结果进行反馈. 随后, 将反馈结果转化成文本描述, 大模型对反馈结果进行分析后改进规划内容. Inner Monologue [196] 在机械臂操作任务中利用不同的感知模型设计了多种反馈, 包括碰撞检测、场景描述反馈、是否成功的反馈等. 例如, 机械臂在打开柜子失败后, 外部感知模型会检测到失败并告知上次打开柜门失败的原因, 随后大模型根据反馈结果进行重新规划. RoCo [197] 提出了在基于环境反馈和智能体沟通的多智能体协作框架, 使用仿真环境的碰撞检查器对机械臂规划路径进行合理性检测, 同时对合作智能体在每个时间步的目标完成情况进行反馈, 帮助智能体进行错误纠正和反思. DoReMi [198] 引入了一个额外的视觉语言模型作为检查器, 在执行大模型规划的每个时间步检查当前的场景是否满足约束. 例如在 "如何倒一杯牛奶" 的任务中, 在第二步拧开盖子之前机械臂必须已经把牛奶拿到桌子上, 视觉语言模型会检查桌子上是否有牛奶来判断上一步是否执行成功. 如果没有成功则检查器打断执行, 对错误规划进行快速纠正. 环境的外部反馈需要引

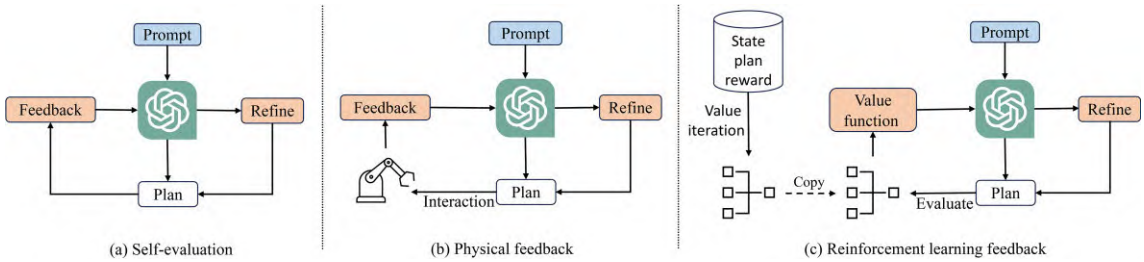


图 11 (网络版彩图) 大模型规划方法基本结构

Figure 11 (Color online) Several basic architectures of LLM-based planning

入额外的检测模块或调用具身仿真中的检测器进行, 同时需要设置一定的规则将反馈结果转化为自然语言描述.

强化学习反馈方法通过引入强化学习概念对大模型的规划结果进行反馈. SayCan 算法 [199] 是一个典型的架构, 通过引入了强化学习中值函数 (value function) 的概念, 对智能体的底层调用技能进行策略学习的同时学习动作的值函数. 在最终决策中, 综合考虑大语言模型输出的动作和动作值函数的大小, 从而避免输出不合理的动作. 具体地, 由于动作值函数衡量了累积回报的未来期望, 如果动作值函数的值很小, 则表明执行该动作在未来可能会导致总体的任务失败. Text2Motion [200] 在机械臂任务执行时提出了几何可行性的概念, 在堆叠不同的物体时考虑物体的几何关系. 对大语言模型产生的多个可能的规划, 检查器分别计算几何可行的概率, 并与大模型原始的概率结合作出最终的规划. Remember 算法 [201] 设计了一个记忆模块, 用值函数评估方法对大模型预测的结果计算值函数, 将状态 - 动作 - 值函数的评价结果作为记忆进行存储. 在与环境交互中, 对于规划产生的新动作, 大模型从记忆中检索相似的状态 - 动作元组并根据这些元组对应的值函数对规划动作是否能够成功进行估计. ReAD 算法 [202] 在大模型多智能体协同中引入了强化学习反馈机制, 通过序列优势函数评价单个智能体行为在总体的贡献, 通过最大化优势函数调整智能体的策略.

图 11 对比了三种大模型规划方法的不同, 其中大模型自我反馈在架构上最为简洁, 然而在复杂具身环境中不能保证准确性. 基于外部环境反馈的结构最为有效, 但需要引入额外的检查器或在动作执行完之后进行反馈, 试错成本较高. 基于强化学习的反馈方法能够在动作执行前根据值函数反馈调整策略, 但需要预训练可靠的值函数网络来对动作进行评价.

4.2 规划搜索算法

在长序列决策任务中, 每个时间步的规划不仅需要考虑当前时间步的最优性, 同时需要考虑由该时间步的动作选择而带来的未来长时间决策的最优性. 与强化学习类似, 在长周期规划过程中, 往往需要牺牲某些时间步的最优性而换取长期的最大回报. 在大模型规划中, 思维链的输出方式往往更多考虑每个时间步规划的最优性. 同时, 思维链规划过程中对未来存在隐式的推理流程, 但随着周期越长, 对未来的规划将会更不可靠. 例如, 大模型很难通过思维链的方式输出决策周期 $T = 100$ 步的超长动作规划, 因为很难判断在较长时间步后环境的变化情况. 然而, 在具身智能中存在许多具有挑战性的长周期决策任务, 例如需要组装许多零件的装配任务、具有严密逻辑关系的决策任务、调度多智能体协作的规划任务等. 为了解决具身智能中的长周期决策问题, 需要引入规划搜索算法来对未来的可行解进行搜索, 从而保证长序列决策周期的最优性.

基于树搜索的方法. 大语言模型中树搜索的方法的基本原理遵循蒙特卡洛树搜索 (MCTS) [203] 的思路. MCTS 是一种经典的树搜索方法, 在 AlphaGo 围棋智能体 [204] 中有着非常成功的应用. 在解决

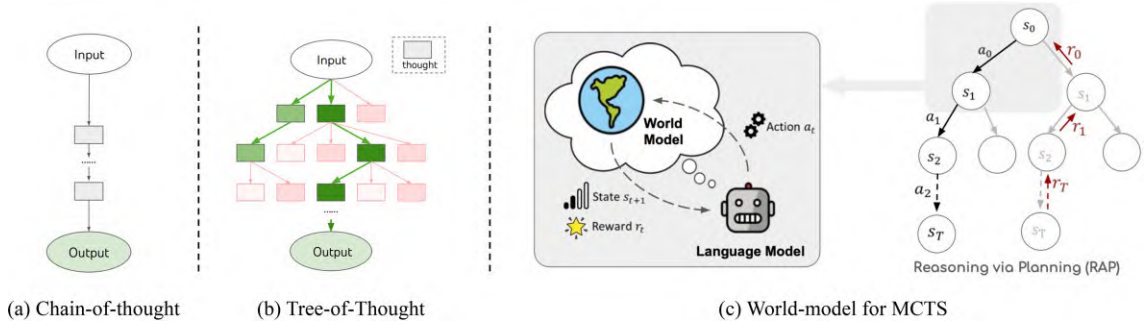


图 12 (网络版彩图) 大模型驱动的树搜索方法

Figure 12 (Color online) The LLM-based tree search architectures

围棋等大规模搜索问题中, 使用树搜索的方法通过构建搜索树, 并遵循选择节点、扩展节点、蒙特卡洛搜索估计、回溯更新四个步骤不断扩展搜索树, 根据搜索树选择当前状态的最优动作. 在大语言模型规划中, Yao 等 [205] 将思维链通过蒙特卡洛树搜索的思路扩展为 “思维树” (tree of thoughts, TOT). 在树的每一层使用大语言模型采样来扩展不同的分支节点, 在叶子节点处使用启发式的值函数评价方法计算值函数, 同时结合深度和广度优先搜索算法进行最优路径搜索. FAFA 算法 [206] 算法将大语言模型融入树搜索的框架下, 使用强化学习值函数的更新的方法对叶子节点的值函数分布进行估计, 同时根据树结构进行值函数回溯计算最优路径. TS-LLM 算法 [207] 针对在树搜索过程中使用启发式方法对值函数估计不准确的问题, 提出使用强化学习值函数估计方法显式的对奖励函数和值函数进行估计, 对蒙特卡洛树搜索过程中扩展的叶节点价值进行评价. 在经过回溯后, TS-LLM 更新路径节点的值函数, 结合深度/广度搜索算法和 MCTS 搜索算法选择最优路径. TreePlanner [208] 提升了大语言模型进行树搜索规划的询问效率和搜索效率, 能够高效的进行错误纠正. REX 算法 [209] 在基于大语言模型的蒙特卡洛树搜索过程重点讨论了如何平衡探索和利用的问题, 通过扩展非最优动作节点来保持对环境的探索. 上述树搜索的方法需要预知环境模型, 例如在数学推理和围棋等任务, 可以使用规则式的方法推理采取动作后的下一个状态. 然而, 在许多非结构化的具身智能场景中, 构建搜索树需要显式的对环境模型进行构建. LLM-MCTS 算法 [19] 在 VirtualHome 中利用大语言模型对环境自身的理解能力, 使大语言模型同时作为环境模型和策略规划器, 并将二者统一到蒙特卡洛树搜索的架构中. RAP 算法 [210] 引入了强化学习中世界模型的概念, 使用大语言模型作为世界模型对下一个状态的分布 $p(s'|s,a,c)$ 的分布进行建模, 其中 c 为引导大语言模型进行状态预测的上下文. 随后, 将世界模型融入蒙特卡洛树搜索的过程中, 设计了启发式的奖励函数对叶节点进行值函数估计, 使用置信区间上界 (UCB) 算法在扩展节时平衡探索与利用. 与上述使用大模型直接作为世界模型的方法不同, Dynalang 算法 [211] 显式的构建了参数化的环境模型来预测未来时间步的环境语言描述和图像观测变化, 使用类似 Dreamer 结构将语言融入世界模型的构建中, 利用语言包含的丰富的世界描述知识来使环境模型的预测更加准确, 同时能够融入视频预测、奖励预测等模块. Dynalang 算法在大规模具身数据集上进行预训练, 可以和树搜索方法进行结合. 图 12 对比了典型的树搜索结构, 树搜索提升了大语言模型在长周期任务中的规划能力.

基于 PDDL 语言的搜索. 规划领域定义语言 (planning domain definition language, PDDL) 是传统规划领域中常用的语言. PDDL 常用于解决长周期的规划问题, 例如组装一个产品需要许多不同的零件, 各种零件之间有复杂的功能依赖关系. PDDL 语言能够通过任务定义和目标定义, 结合谓词和动作定义, 在可行空间中搜索达到目标状态的可行解决方案, 该方法可以包含成百上千个时间步. PDDL

表 2 大模型驱动的具身任务规划

Table 2 Category and representative algorithms for LLM-based embodied planning

Content structure	Category	Representative algorithms
Feedback-based planning	Open-loop planning	CoT [8], ChatGPT-for-Robotics [187], Zero-Shot [188], ReAct [189]
	Self-evaluation feedback	Self-Refine [190], Reflexion [191], BeamSearch [192], SwiftSage [193], ViLA [194]
	External env. feedback	Inner Monologue [196], RoCo [197], DoReMi [198]
	Value function feedback	SayCan [199], Text2Motion [200], Remember [201], ReAD [202]
Search-based planning	Tree-based search	ToT [205], FAFA [206], TS-LLM [207], Tree-Planner [208], REX [209], LLM-MCTS [19], RAP [210], Dynalang [211]
	PDDL search	LLM+P [212], ISR-LLM [213], LLM+PDDL [214]

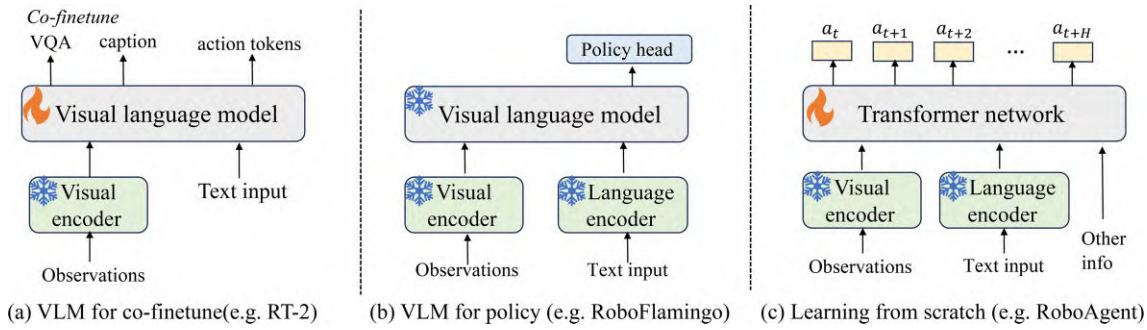
的优势在于已有许多先进的可用规划求解器, 且通过求解器搜索到的可行解都是完全符合约束要求的. LLM+P [212] 将大语言模型和 PDDL 求解相结合, 主要包含三个步骤. 首先, 使用语言问题描述和 PDDL 问题描述的例子作为大模型 prompt, 大模型能够学习如何现在的面临的问题转化为 PDDL 语言; 随后, 大模型使用 PDDL 求解器对该问题进行求解, 生成 PDDL 语言描述的规划; 最后, 大模型将 PDDL 规划转换为自然语言进行执行. 将大模型和 PDDL 相结合能够很大程度上弥补大模型在长序列解空间搜索问题中的不足, 为解决长周期决策问题提供了新的思路. ISR-LLM [213] 提出在将问题转换为 PDDL 语言后, 可以利用大模型自身对 PDDL 的理解能力产生任务规划, 随后使用使用 PDDL 合理性检查和大模型自身检查作为反馈, 帮助大模型改进之前的规划结果. LLM+PDDL [214] 采取了另一种思路, 使用 GPT-4 对 PDDL 问题进行自然语言归纳, 并编写代码程序进行解决. 对于解决情况, GPT-4 引入自我反馈机制进行纠正并重复上述过程. 算法总结如表 2 所示.

5 大模型驱动的具身基础策略

5.1 LLM/VLM 驱动的基础策略

在基于大模型的任务规划中, 虽然反馈机制和搜索机制能在一定程度上提升规划性能, 但仍然存在两点不足: (1) 大模型在执行具身任务的规划中大模型权重是不进行调整的, 主要依靠任务描述和场景 prompt 来进行泛化, 有一定的局限性; (2) 大模型需要将任务描述转换成规划, 然后使用机器人底层技能库将规划转换为动作进行执行, 较为依赖机器人底层技能库的设计. 本节介绍大模型驱动的具身基础策略, 其基本结构如图 13 所示, 主要特点是将原始的大模型参数作为基础策略, 并利用具身数据对大模型参数进行微调, 使大模型能够更加适应于具身决策场景, 减轻底层技能库定义的依赖, 提升决策效率.

大模型微调的决策规划. 在大语言模型直接产生任务规划时依赖模型中编码的知识. 由于大模型缺乏具身任务规划的相关知识, 且在具身任务规划时不对大模型参数进行调整, 大模型需要使用额外的反馈模块来对产生的不合理规划进行迭代. 现有研究指出, 一种更为直接的方式是使用具身智能数据对大模型原有的预训练参数进行微调, 使其适应于具身智能任务场景. 此时, 可以认为预训练的大语言模型/视觉语言模型将作为具身智能的基础策略, 在进行微调后得到具身大模型. Palm-E [215] 是 Google 提出的一种具身规划大模型. 由于具身数据含有状态、图像等输入, Palm-E 首先将这些输入编



(a) VLM for co-finetune(e.g. RT-2) (b) VLM for policy (e.g. RoboFlamingo) (c) Learning from scratch (e.g. RoboAgent)

图 13 (网络版彩图) LLM/VLM 驱动的基础策略基本结构

Figure 13 (Color online) The basic architectures of LLM/VLM-based policy

码到和大语言模型输入相同的潜变量层, 随后基于大语言模型的自注意力机制将这些输入和文本输入以相同的形式进行处理. Plam-E 最终输出语言描述的任务规划, 与收集的数据集中的任务规划结果计算损失, 进行端到端的训练. Palm-E 能够处理多模态的输入数据, 将具身智能数据和视觉 QA、语言问答任务等数据进行共同训练, 能够处理多种复杂问题. Embodied-GPT 算法 [216] 收集了基于 Ego4D 的人类操作问答数据集, 并指出该数据集能够更好的迁移到具身智能的任务中. 同时, Embodied-GPT 使用了更轻量级的 7B 语言模型进行微调, 在底层使用模仿学习的方法将单步的规划转换成底层策略进行执行, 在 Franka Kitchen, Meta-World 等测试任务中超过了 R3M, BLIP-2 等通过预训练表征来进行策略学习的方法.

大模型微调的策略学习. 大模型驱动的具身任务规划能够在语言层面产生指令, 随后将语言指令转化为机器人动作进行执行, 存在任务 - 规划 - 动作 (Task-Plan-Action) 的两步映射关系. 其中, 在进行规划 - 动作映射时需要根据机器人已经构建的底层技能库来完成. 相比较而言, 模仿学习和强化学习算法中只存在着任务 - 动作 (Task-Action) 的直接映射关系, 能够根据任务和当前智能体的状态而直接输出动作, 其中任务分解流程是隐式完成的. 为了将大模型驱动的决策任务简化为由任务到动作的直接映射, LaMo 算法 [217] 尝试了使用小规模 GPT-2 语言模型作为离线强化学习的基础策略, 使用 Decition Transformer (DT) [218] 的条件模仿学习框架进行训练. 在训练中, 不对大语言模型的参数进行全量学习, 而是使用大模型轻量级微调技术 LoRA [219] 对模型参数进行微调, 尽可能多的保留大模型自身编码的知识结构. 在经过微调后, LaMo 能够将语言模型编码的知识和强化学习决策任务结合, 使 GPT-2 以较小的微调代价适应于离线决策任务. Google 提出的 Robot Transformer (RT) 系列使用了更大规模的语言模型和更多的具身智能任务数据进行, 在许多具身智能任务中都获得了出色的效果. RT-1 算法 [37] 使用预训练的 EfficientNet-B3 网络初始化, 以机器人状态和历史图片作为输入, 通过 EfficientNet 特征提取后直接输出动作. RT-1 中将机器人动作的每个维度均匀离散化将动作词元化, 使用监督学习的损失进行训练. RT-2 算法 [38] 整体使用大规模预训练的视觉 - 语言模型结构, 模型参数可以达到 55B 的参数量, 远超 RT-1 的参数规模, 同时利用大规模预训练 VLM 模型中编码的丰富视觉问答知识来帮助具身模型的训练. RT-2 将输出的动作进行和 RT-1 相同的离散化操作后将词元加入 VLM 原先的词表中, 可以把动作词元视为另外一种语言进行处理, 无需改变原有 VLM 结构设计. 由于 RT-2 已经在海量的视觉问答任务中进行预训练, 在对图片和任务指令的理解上有更加丰富的经验, 在任务集合上具有更强的泛化能力. RT-X 架构 [20] 在 RT-2 模型的基础上, 收集并整理了来自全球 60 多个实验室的丰富的机械臂数据, 覆盖了 22 个具身实体, 527 种不同的具身实体类型, 共计 16 万种不同任务. 训练得到的具身大模型对场景和任务具备更丰富的理解和泛化能力, 同时由于

具身实体的动作空间不同, 得到的策略在动作空间中具有更强的泛化能力. RT 系列模型虽然取得了好的效果, 但对于模型结构和训练细节的开源层次不高, 研究人员往往难以完全复现其性能. 面对该问题, RobotFlamingo 架构 [220] 基于开源的视觉 - 语言模型 Flamingo [12, 118] 构建具身决策模型. 在输入中使用预训练的视觉编码器和语言编码器将视觉观测和语言指令进行特征提取, 使用预训练的开源 Flamingo 模型进行特征融合并产生针对任务的回答, 在 Flamingo 的输出层加入策略预测网络直接输出动作. SAM-E 算法 [221] 利用视觉分割模型 SAM 强大的可提示感知能力, 通过解析文本指令使模型关注到场景中的操作物体, 从而理解三维操作空间. Leo 算法 [222] 提出了三维场景的理解和语言问答大模型, 通过三维场景 - 语言问答和场景 - 语言 - 动作对齐, 能够产生三维空间中的具身策略. 3D-VLA 算法 [223] 在 3D 大语言模型基础上加入目标生成能力, 使其进行多模态目标生成和场景理解, 同时能够进行目标定位、动作预测等. OpenVLA [224] 和 Octo [225] 以大语言模型为基座, 提出了开源的视觉 - 语言 - 动作模型, 能够用于真实机械臂的控制.

大模型驱动的直接策略学习. 虽然上述基于大模型的策略取得了不错的效果, 但由于大模型具有较大的参数量, 在机器人任务中需要更大的计算和时间消耗, 具有较低的决策频率. 在上述结构的启发下, 有部分研究采取自行设计的 Transformer 网络结构, 直接使用机器人数据从头开始训练网络, 在部分具身决策任务中取得了不错的效果. 斯坦福大学提出 ALOHA 结构 [226] 使用 Transformer 编码 - 解码网络结构, 以不同方位的观测图像作为输入, 通过解码器直接输出机械臂动作. 为了解决长周期决策问题, ALOHA 使用动作分块的概念, 一次预测多个时间步的动作序列, 增强了长周期任务中动作预测的整体性. 在硬件方面, 该研究搭建了低廉的 ALOHA 开源双臂机器人实验平台, 使人类能够完成便捷的示教数据采集, 仅使用采集的机械臂数据进行训练. 进一步地, 斯坦福大学团队搭建了 Mobile ALOHA [39] 移动平台, 通过专家示教数据的模仿学习能够完成滑蛋虾仁、干贝烧鸡、蚝油生菜等菜品的制作, 其出色的效果获得了广泛关注. HiveFormer [227], RVT [228] 等算法相继提出了 3D 空间中的机械臂模仿学习架构, 以多视角的机器人观测数据作为输入, 使用 CLIP 和 Transformer 结构进行语言和视觉特征提取和融合, 直接预测机器人在 6D 空间的抓取位姿, 在 RLBench 测试平台和真实 3D 机械臂抓取任务中达到了当前最好的效果. Google 提出的 RoboCat 结构 [229] 使用跨实体、跨任务的具身模仿学习框架, 在使用 VQ-GAN [230] 对视觉输入词元化之后, 使用标准的 DT 回归损失根据历史的状态、观测、目标信息对未来的智能体动作和观测进行预测. 同时, RoboCat 能够生成部分虚拟数据训练, 不断提升智能体的能力. 在新任务上, RoboCat 仅需 100∼1000 个示教样本就能完成快速策略泛化. FAIR 提出 RoboAgent 算法 [231] 使用 SAM 模型 [112] 对场景中的物体进行检测并进行数据增强, 快速提升场景中物体的丰富程度, 从而得到在不同物体类型上有更好的泛化效果. 此外, RoboAgent 使用 Transformer 编码 - 解码框架, 使用视觉观测、任务描述、机器人状态等作为输出, 对动作序列进行预测, 最小化动作序列预测的误差.

大模型驱动的世界模型策略. 大模型能够帮助智能体构建环境转移模型, 也被称为世界模型 (world model, WM). 世界模型如 Dreamer [83], TD-MPC [232], IRIS [233] 等可以建模环境, 随后在隐空间或原始轨迹中生成智能体轨迹, 并利用这些轨迹结合基于模型的强化学习或模型预测控制等方法获得最优策略. 由于该过程同时属于大模型数据生成的一部分, 相关内容将在后续章节介绍.

5.2 扩散模型驱动的基础策略

基于扩散模型的动作规划. 扩散模型作为一种图像生成模型, 通过前向的噪声扩散过程得到高斯噪声, 通过多步逆向的去噪过程恢复出原始图像. 在图像生成领域, 扩散模型已经被验证能够建模高维度的复杂数据, 因此在具身智能任务中被用于建模高维度的决策序列. 具体地, 扩散模型可以直接

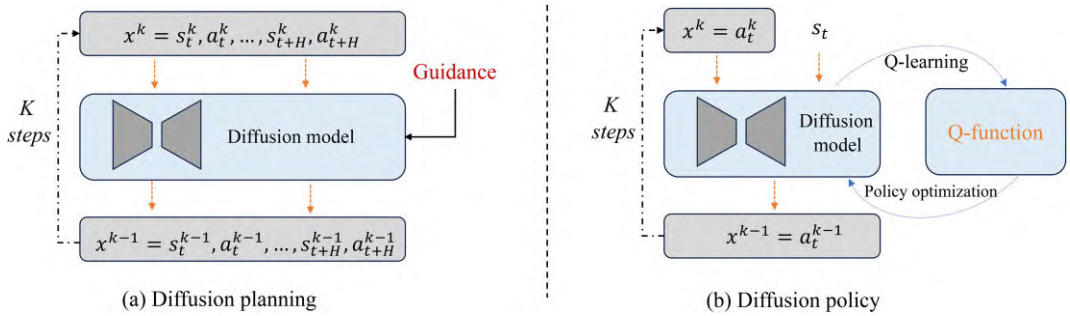


图 14 (网络版彩图) 扩散模型驱动的动作规划和策略学习

Figure 14 (Color online) Diffusion-based action planning and policy learning

作为策略规划器 (Planner), 通过对状态 - 动作序列 $[(s_0, a_0), \ldots, (s_T, a_T)]$ 的整体建模, 能够从原始噪声还原出整条决策轨迹, 从而在执行时作为规划器来生成未来的轨迹. Diffuser [133] 是一种典型的将扩散模型作为规划器的算法, 在建模中使用历史轨迹作为条件, 使用奖励引导进行轨迹生成, 能够根据历史轨迹生成未来的高奖励轨迹规划. Decision Diffuser [234] 进一步增强了条件生成的能力, 使用无分类器 (classifier-free) 引导的条件生成方法, 将回报函数、任务描述、技能描述、约束条件等作为条件进行可控的扩散动作生成, 使扩散模型生成的轨迹满足多种混合的条件, 提升了扩散模型在决策任务中的规划能力. MetaDiffuser [134] 使用 Meta-Learning 的框架, 对任务轨迹使用上下文编码器进行编码, 并将编码向量作为条件进行轨迹生成. 在扩散模型生成中显式的使用环境模型预测损失引导模型生成符合环境动力学的轨迹. 在测试中, 通过将测试任务的示教轨迹进行上下文编码, 该模型能够扩展到新任务的轨迹规划中. MTDiffuser [136] 增强了扩散模型在多任务规划中的泛化能力, 将任务的示教数据经过编码后作为扩散模型的提示, 使用条件生成过程引导扩散模型生成具有显著差异的多任务轨迹, 在具身多任务规划问题中取得了出色的效果. ChainedDiffuser [235] 增强了扩散模型在长序列决策任务规划中的能力, 使用层次化的方法首先产生高层任务的规划, 如需要达到的关键目标位置等, 随后使用底层扩散过程生成状态到目标规划的动作轨迹. 目前, 由于大规模预训练扩散模型主要用于图像生成、代码生成、文本生成等具有大量开源数据的场景, 暂未有典型的将预训练扩散模型参数用于具身任务的成果.

基于扩散模型的策略生成. 扩散模型在直接进行动作规划之外, 可以作为智能体的策略和模仿学习/强化学习算法进行结合, 在现有学习框架中进行训练. 在一般的学习框架中, 策略使用确定性函数或高斯分布进行建模, 策略输出仅能表示单峰或多峰的策略分布. 然而, 在许多人类示教数据中, 不同人类专家在相同状态下采取的策略分布是非常多样的, 策略建模具有高度的复杂性, 使用扩散模型进行策略建模在复杂决策任务中更有优势. Pearce 等 [56] 使用扩散模型在模仿学习框架下对人类的复杂动作分布进行建模, 能够在大规模机器人决策任务中达到更好的效果. BESO 算法 [236] 在目标导向的模仿学习框架中使用随机微分方程来进行动作建模, 同时使用非最优多样化轨迹进行训练. SfBC 算法 [237] 将策略学习过程分解为基于扩散模型的动作建模和基于强化学习的动作评价, 使用值函数加权的重要性采样方法从扩散模型中进行动作选择. Diffusion-QL 算法 [55] 在强化学习框架下使用扩散模型作为基础策略, 使用最大化值函数的学习目标对动作扩散模型进行训练, 更好的建模了离线强化学习数据集的动作分布. EDP 算法 [238] 提出更加高效的动作采样流程, 从中间噪声状态可以一步恢复到原始动作, 可以和不同的离线强化学习框架结合. IDQL 算法 [239] 将扩散模型和模仿学习策略进行结合, 用于最大化值函数或优势函数, 提升离线强化学习的效果. Zhu 等 [240] 对扩散模型在决策问题中

表 3 大模型驱动的具身基础策略

Table 3 Category and representative algorithms for policies based on large-scale foundation models

Content structure	Category	Representative algorithms
LLM/ VLM-based	LLM/VLM-fintuned planning	Palm-E [215], Embodied-GPT [216]
	LLM/VLM-fintuned policy	LaMo [217], RT-1 [37], RT-2 [38], RT-X [20], RobotFlamingo [220], SAM-E [221], Leo [222], 3D-VLA [223], OpenVLA [224], Octo [225]
	Direct Transformer policy	ALOHA [226], Mobile ALOHA [39], HiveFormer [227], RVT [228], RoboCat [229], RoboAgent [231]
Diffusion-based	Diffusion-based planning	Diffuser [133], Decision Diffuser [234], MetaDiffuser [134], MTDiffuser [136], ChainedDiffuser [235]
	Diffusion-based policy	Pearce [56], BESO [236], SfBC [237], Diffusion-QL [55], EDP [238], IDQL [239]

的研究进行了综述. 扩散模型驱动的动作规划和策略学习的基本结构如图 14 所示.

虽然大模型提供了可泛化的控制方法, 但仍可以与传统控制方法的结合: (1) 传统控制方法具有精确性和稳定性的优势, 例如, 比例 - 积分 - 微分 (PID) 控制、状态反馈控制和模型预测控制 (MPC) 等有明确的数学模型, 能提供高精度和鲁棒的性能. (2) 传统控制算法通常是实时运行的, 计算复杂度较低, 且工作原理和结果可通过数学推导和物理直观解释, 便于调试和验证. 在应用中可以采用混合控制架构, 使用传统控制方法处理低层次的、高频的控制任务 (如稳定性和精度控制), 而使用大模型驱动的方法处理高层次的、低频的决策任务, 提高策略的收敛性和稳定性. 算法总结如表 3 所示.

6 大模型驱动的具身奖励函数

6.1 奖励函数代码生成

在马尔科夫决策过程 (MDP) 的定义中, 奖励函数用于评价在状态 s 处执行动作 a 后, 智能体达到的下一个状态所带来的收益, 记为 $r(s,a)$. 在考虑多个时间步的决策任务中, 在状态 s_t 处执行动作 a_t 的回报函数定义为 $R(s_t,a_t)=\sum_t \gamma^t r(s_t,a_t)$. 在使用 Q 函数对回报函数估计的前提下, 最优动作可以通过贪心的动作选择 $a=\arg\max Q(s,a)$ 或采用优化的方式使策略输出最大化 Q 函数的动作. 模型预测控制 (MPC) 通过评价多个动作序列的回报值, 从中选择使得未来回报最大的动作. 在强化学习和 MPC 中, 奖励函数的定义和获取是非常重要的, 智能体的策略通过优化奖励函数产生. 然而, 具身奖励函数的定义往往需要一定的先验知识. 例如, 在机械臂开门的任务中, 可以直接使用稀疏奖励的定义 (0 或 1) 表示最后是否成功打开了门, 然而该奖励由于其稀疏性往往难以学习. 为了解决该问题, 可以模拟人类开门的过程来设计奖励: 首先握住门把手, 随后转动, 最后拉开门. 通过完成每个任务的步骤设计密集的奖励函数, 从而引导智能体完成最终目标. 在现有的基于强化学习和 MPC 的具身智能研究中, 奖励函数往往由机器人领域专家设计, 在设计中同时考虑任务完成度、能量损耗、安全性等诸多因素, 随后根据专家经验知识对各个奖励函数项进行加权得到最终的奖励函数, 具有一定的设计难度.

在大模型的背景下, 由于大模型对机器人系统和状态有一定的先验知识, 可以使用大语言模型/视觉语言模型进行奖励图或奖励函数的生成. VoxPoser 算法 [241] 尝试使用大语言模型编写奖励函数的代码. 首先根据任务的语言描述, 使用视觉检测模型对场景中感兴趣的物体进行定位. 随后 VoxPoser 对可能产生奖励的物体和可能产生危险的物体分别对待, 在 3D 机器人操作空间内生成奖励图和约束

图, 得到最终价值分布图. 在 MPC 中, 可以使用现有的运动规划器在价值分布图中规划一条无碰撞的最优路线进行执行. VoxPoser 的优势在于, 由于大语言模型能够从场景 prompt 和文本描述中对可能产生奖励的物体和可能产生危险的物体进行推理, 而视觉检测模型已经在大量的物体类型上训练过, 能够对大部分场景中的物体进行识别, 从而使 VoxPoser 的奖励图生成机制能够适应于未见过的场景任务. Text2Reward [18] 算法利用大模型对环境的理解能力编写奖励函数定义的代码, 通过提供场景描述、环境介绍、奖励函数例子等作为 prompt, 大语言模型能够根据任务要求生成完成该任务的密集奖励函数. 同时, 由于大语言模型具备理解代码的能力, 对于环境和场景等定义可以直接以代码定义的格式进行提供, 无需转换为自然语言进行输入. 强化学习算法根据大模型产生的奖励函数进行策略训练, 在环境中进行交互产生轨迹, 人类通过观察交互轨迹给出反馈, 大模型根据反馈提升现有的奖励函数的设计. Eureka 框架 [242] 是英伟达提出的奖励生成框架, 能够和英伟达的高效并行仿真环境 Omniverse 进行结合, 对自定义的任务进行奖励函数生成. Eureka 利用了 GPT-4 中更为强大的任务规划能力、代码编写能力和反馈优化能力, 以环境的代码定义和任务描述作为提示直接产生任务的奖励函数. 随后, 使用强化学习算法对奖励函数进行策略优化, 根据策略的执行结果对 GPT-4 进行反馈, GPT-4 分析原因并改进奖励函数中不合理的部分. 在多个复杂的机械臂和灵巧手操作任务中的实验结果表明, Eureka 具有强大的奖励函数编写能力, 能够根据结果反馈不断的改进奖励函数设计, 取得了超越人类专家的效果. DeepMind 在近似工作 [243] 中将大模型生成奖励函数用于更加复杂的四足机器人运动控制中, 为四足机器人的各种技能学习提供奖励函数, 能够驱动四足机器人学习奔跑、跳跃、直立行走等复杂的技能. 同时, 该方法也能基于人类反馈来对奖励函数的规划进行改进. 总体而言, 奖励函数生成相比于直接进行任务规划和策略学习具有更低的难度, 无需理解机器人底层动作的执行逻辑, 仅需完成较为上层的奖励函数设计. 在奖励函数的定义下, 可以根据仿真或真实机器人系统中已经具备的求解器类型, 结合强化学习或模型预测控制等算法对策略进行求解, 具有更高的求解效率. 然而, 目前的奖励函数生成算法均需通过反馈和多次迭代优化, 仅适应于有限的机器人实体任务, 在未来有很大的发展空间.

6.2 奖励函数学习和计算

在使用大模型直接生成奖励函数的代码之外, 其他研究方法通过构建奖励模型对奖励函数进行估计, 主要包含三个类别: 视频预测模型、语言 - 视频匹配模型、预训练视觉 - 语言模型. (1) 在视频预测模型方面, VIPER 算法 [244] 通过对专家给出的大量智能体执行视频轨迹进行学习, 构建专家轨迹的条件概率生成模型 $\log p(x_{1:T}) = \sum_t \log p(x_t|x_{1:t})$. 由于该模型描述了专家轨迹中状态的概率分布, 使用专家轨迹进行评价时能获得较大的概率估计值. 相反, 如果当前智能体的策略水平较低, 则使用概率密度函数对智能体的交互视频进行评价将得到较低的概率密度. VIPER 算法使用从专家轨迹中学习的概率分布对当前智能体的轨迹进行评价, 将奖励函数定义为 $r = \log p(s_t|s_{1:t})$. Diffusion Reward [245] 使用更为先进的 VQ-Diffusion 扩散模型对智能体交互视频进行建模, 并指出专家轨迹在扩散模型的预测中会呈现出更大的确定性, 而非最优轨迹可能更加多样, 最终使用扩散模型预测的条件熵来作为奖励. (2) 在语言 - 视频匹配模型方面, LIV 算法 [246] 通过对比学习和值函数优化在隐空间中对任务的文本描述和任务的交互视频的相似性进行最大化, 通过大规模训练得到的文本和视频编码器对任务描述和相应的轨迹视频的相关程度进行判别. 在新任务执行时, 对新任务的文本描述和当前智能体的视频轨迹相关程度进行度量, 能够间接的评价当前智能体轨迹和专家轨迹的相似程度, 从而作为智能体的奖励. (3) 视觉 - 语言模型奖励使用预训练的视觉模型和语言模型来判断任务语言描述和视频轨迹的相似性, 并将其作为奖励函数. VLM-RM 算法 [247] 使用 CLIP 模型的语言和视觉编码器

来判断 Humanoid 中智能体的行为是否和文本描述一致, 并将该度量作为奖励函数. LAMP 算法 [248] 使用奖励函数 $r = G_\theta(F_\phi(o_1), F_\phi(o_i), L_\alpha(x))$ 作为奖励, 其中 F_ϕ 是 R3M 算法 [158] 从大规模人类操作数据集中训练出的编码器, G_θ 衡量了从状态 o_1 到 o_i 的智能体变化情况和任务描述编码 $L_\alpha(x)$ 的匹配程度, 对符合人类专家的轨迹产生较高的奖励. LAMP 同时使用了强化学习中由不确定性定义内在激励函数, 从而引导智能体高效的探索环境. 目前奖励函数的学习由传统的逆强化学习理论扩展而来, 基于专家视频轨迹, 结合视觉基础模型和视觉 - 语言模型对视频和文本描述的理解, 根据当前轨迹和专家轨迹或文本描述的匹配程度来计算奖励.

6.3 偏好驱动的奖励函数

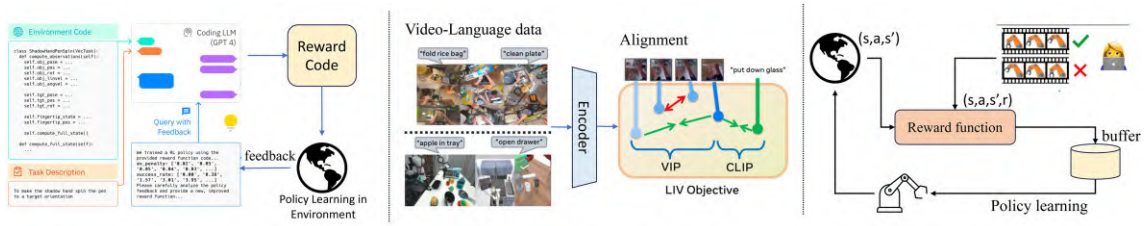
大语言模型在经过预训练的指令微调后, 一个重要的步骤是将大语言模型的输出和人类的偏好进行对齐, 进行人类反馈的强化学习 (reinforcement learning from human feedback, RLHF) [249]. 在人类反馈的强化学习中, 需要根据人类偏好的标记数据建模奖励函数. 近期, 人类偏好 (Preference) 的概念也被引入具身智能中, 由人类专家或预定义的规则对轨迹偏好进行打分, 根据打分结果训练奖励函数模型, 随后将奖励函数模型用于强化学习算法训练 [250]. 具体地, 从智能体的离线轨迹数据集中采样两个轨迹片段 τ_1 和 τ_2, 随后人类对两个轨迹进行比较并根据自身偏好给出偏好标签 $y \in \{0, 1, 0.5\}$, 其中 $y = 1$ 代表人类更加偏好轨迹 τ_1 (即 $\tau_1 \succ \tau_2$) , $y = 0$ 代表 $\tau_1 \prec \tau_2$, 而 $y = 0.5$ 则代表两条轨迹拥有相似的偏好. 现有方法一般使用 Bradley-Terry (BT) 模型 [251] 来从偏好数据中学习奖励函数.

给定特定具身任务中的偏好数据集 [252]$\mathcal{D}_\tau = \{(\tau_1, \tau_2, y)\}$, 奖励函数 $\hat{r}$ 通过最大化交叉熵损失函数进行学习:

$$\mathcal{L}_{\hat{r}} = \mathbb{E}_{\mathcal{D}_\tau}\left[y \log P[\tau_1 \succ \tau_2] + (1 - y) \log P[\tau_2 \succ \tau_1]\right], \tag{9}$$

其中 $P[\tau_1 \succ \tau_2] = \exp \sum_t \hat{r}(s_t^1, a_t^1) / \sum_{i \in \{0,1\}} \exp \sum_t \hat{r}(s_t^i, a_t^i)$. 根据学习得到的奖励函数 $\hat{r}$, 使用强化算法进行策略训练. PEBBLE 算法 [253] 提出了反馈高效的偏好强化学习框架, 将偏好策略学习分为两个阶段. 在第一阶段中, 利用强化学习的高效探索算法学习探索策略, 通过和仿真环境交互来获得多样化的轨迹数据. 在多样的轨迹数据中进行人类反馈的偏好标注更具有代表性, 能够使学到的奖励函数具有更强的鲁棒性和泛化性, 提升了人类反馈的效率. 在第二阶段, 使用学习到的奖励函数和强化学习算法进行策略训练, 在多个具身任务数据集上达到了和原始外部奖励训练类似的效果. SURF 算法 [254] 提出了额外训练的偏好预测器, 对大量未进行偏好标记的数据集进行偏好重标记, 随后使用片段随机裁剪的方法进行数据增强, 从而在奖励学习中提升人类反馈数据的利用效率. RUNE 算法 [255] 提出了奖励不确定性的概念, 通过拟合多个奖励函数并衡量预测的方差来作为不确定性度量, 利用不确定性来指导智能体的探索. 智能体通过探索能够采集多样化的轨迹, 从而有利于偏好奖励的学习. Preference Transformer (PT) 模型 [256] 使用更复杂的 Transformer 结构来进行偏好奖励的学习, 从而建模更加复杂的奖励函数分布. 近期, 许多研究尝试绕过奖励函数的建模, 直接使用偏好数据来进行策略学习. DPPO 算法 [257] 定义了策略和轨迹的距离度量, 通过最小化策略和偏好轨迹之间的距离能够使用 BT 模型进行直接的策略优化. CPL 算法 [258] 引入优势函数的概念, 通过对偏好和非偏好数据的对比学习进行策略优化. OPPO 算法 [259] 将标量的奖励函数学习扩展为偏好表征学习, 利用三元组损失在表征空间中将具有类似偏好的轨迹表征进行拉近, 而增大了不同偏好表征的距离. 在进行数据生成时使用最优轨迹的偏好表征作为条件, 使用扩散模型进行条件策略建模.

图 15 对 3 种奖励函数生成的机制进行了对比. 其中, 使用大语言模型进行奖励代码生成的方式最为直接, 依赖于大语言模型对环境和任务定义的自我理解, 并通过实际执行情况的改进不断反馈; 奖励学习的机制需要从大规模具身视频和文本的数据集中进行学习, 在实际任务表现中具有较好的效果.



(a) Reward code generation (e.g., Eureka) (b) Video-language alignment for reward learning (c) Preference-based reward

图 15 (网络版彩图) 基于大模型的奖励生成框架对比

Figure 15 (Color online) Comparison of reward generation methods based on foundation models

表 4 大模型驱动的具身奖励函数

Table 4 Category and representative algorithms for reward function modeling based on large-scale foundation models

Content structure	Category	Representative algorithms
Reward code generation	LLM/VLM-based generation	VoxPoser [241], Text2Reward [18], Eureka [242], Lang2Reward [243]
Reward learning from videos	Video prediction	VIPER [244], Diffusion Reward [245]
	Language-video alignment	LIV [246]
	Pre-trained VLM/VFM	VLM-RM [247], LAMP [248]
Preference-based reward	Reward modeling	PbRL [249], PEBBLE [253], SURF [254], RUNE [255], PT [256]
	Direct policy learning	DPPO [257], CPL [258], OPPO [259]

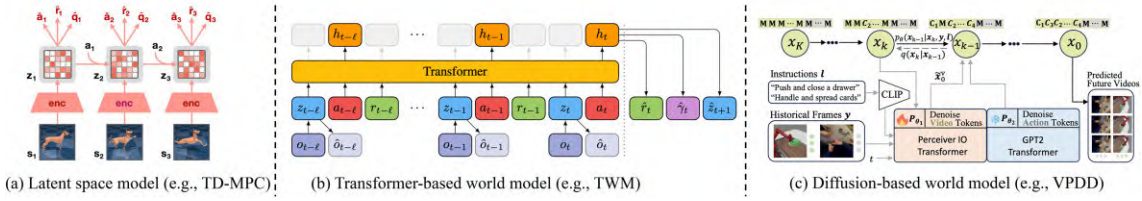
与奖励代码生成不同的是, 视频学习到的奖励函数属于黑盒模型, 而大语言模型生成的奖励代码的定义是白盒模型, 人类能够进行阅读和改进. 偏好奖励函数的学习使用特殊的偏好数据集, 随着众包标注和各种偏好数据集的开放 [260], 在未来将有助于具身场景中学习个性化和多样化的奖励函数, 使机器人的行为与人类的特定偏好进行对齐 [261]. 算法总结如表 4 所示.

7 大模型驱动的具身数据生成

7.1 世界模型的数据生成

世界模型 (world model) 的构建是具身智能研究的重要内容 [262]. 在狭义上, 世界模型能够帮助智能体对未来的状态和轨迹进行预测, 对基于模型的强化学习算法和模型预测控制算法进行帮助; 在广义上, 世界模型可以揭示物理世界的运行规律, 包括智能未来的视觉观测如何改变, 状态如何根据策略和环境动力学函数进行转移, 以及最终智能体在世界模型中将会收敛到的状态等. 类比人类对世界的理解能力, 人类对世界的运行规律具有基本的常识 (基于经验、知识、物理定律等), 能够在大脑中快速推演执行某个动作后产生的后果. 世界模型致力于复刻人类的此种能力, 通过理解复杂物理世界的运行规律, 对执行动作产生的未来状态转移进行估计, 从而在当前状态进行最优动作的选择. 同时, 在世界模型中进行推演能够产生大量的推演数据轨迹, 这些数据能够帮助智能体理解策略的执行情况, 丰富强化学习或模仿学习的训练数据, 并进行策略的改进.

从架构上, 现有的世界模型的构建方法主要包括以下 3 类: 隐空间世界模型、Transformer 世界模型、扩散世界模型. 图 16 对比了 3 种世界模型的典型结构, 下面将进行分别介绍.



(a) Latent space model (e.g., TD-MPC) (b) Transformer-based world model (e.g., TWM) (c) Diffusion-based world model (e.g., VPDD)

图 16 (网络版彩图) 具身智能中世界模型的典型框架

Figure 16 (Color online) The classical architecture of world model for embodied agents

隐空间世界模型以 Dreamer 系列工作为代表. 为了避免在图像观测空间中直接构建环境模型 $p(o_{t+1}|o_t, a_t)$ 带来的建模困难, Dreamer [83] 首先将视觉观测 o_t 映射到表征空间中, 随后在表征空间中进行环境模型 $q(s_{t+1}|s_t, a_t)$ 的构建. 在隐空间中进行建模和推理能够降低模型在多步环境模型预测中的误差, 避免使环境模型关注于图像重建的细节. Dreamer 在学习隐空间环境模型的同时学习如何从隐空间重建原始状态, 以及从隐空间预测奖励函数, 可以根据对未来多个时间步的状态和奖励的预测结果获得更为准确的价值估计. Dreamer 在后续改进中将性能进行了提升 [84, 263], 包括使用离散的隐变量空间便于建模环境中的离散环境转移过程, 进行跨域任务训练等. 在 Dreamer 基础上, 研究人员相继提出了多种隐空间的环境模型构建方法. TD-MPC [232] 针对连续控制问题, 使用非观测重建的方式在隐空间内构建环境状态转移模型, 随后使用模型预测控制算法进行策略求解, 使用隐空间模型预测的短期奖励和值函数学习结合对回报进行评估. TD-MPC2 [264] 通过改进损失函数提升了鲁棒性, 同时通过提升模型容量和结构设计使世界模型能够容纳来自多种具身实体任务的数据, 能够在 80 多个来自不同实验平台的仿真环境中达到了出色的效果. TD-MPC 随后被扩展到离线场景下 [265], 可以从机器人采集的离线数据中学习环境模型, 通过离线强化学习值函数学习的方式避免了泛化误差的产生, 能够在真实机器人中进行应用. MV-MWM 算法 [266] 将隐空间环境模型的构建扩展到多视角观测中, 利用机器人多个视角的相机输入学习表征, 随后在表征空间中构建跨视角的环境模型, 学习到的策略能够对视角的变化鲁棒. APV 算法 [63] 和 ContextWM 算法 [267] 针对机器人数据缺乏的问题, 提出利用大规模人类操作数据集进行联合训练, 首先在无动作标记的数据集中训练隐空间预测模型, 随后加入动作标记数据进行隐空间环境模型预测和奖励预测进行微调, 大大提升了无标签数据的利用率. Dynalang [211] 将模型扩展到多模态数据, 能够同时预测图像和文本表示的状态变化.

Transformer 世界模型的雏形可以追溯到 Trajectory Transformer (TT) [268]. TT 对连续控制任务中的状态和动作以分桶的形式将其转化为词元, 再通过每个词元查询获得其对应的嵌入表示, 有效地增加了轨迹序列在嵌入空间里的表达长度和语义. 随后 TT 使用 Transformer 结构根据当前状态和动作对未来的状态词元进行预测, 在进行最优策略求解时使用 Beam-Search 规划方法在多种未来的可能轨迹中搜索最优动作, 验证了 Transformer 世界模型与规划方法结合的优越性. Trans Dreamer 算法 [269] 使用 Transformer 替换了 Dreamer 世界模型中使用的循环神经网络, 并提出了 Transformer 状态空间模型. TWM [270] 参考了 Dreamer 系列工作的设计思路, 保留使用离散的隐变量空间刻画环境跳转过程的模块, 同时加入了均衡采样避免 Transformer 世界模型在早期采样的轨迹上过度学习. 同期, IRIS 算法 [233] 进一步释放 Transformer 精确捕捉数据中长时序依赖关系的潜力, 首次正式提出了基于词元的 Transformer 世界模型, 舍弃了离散的隐变量空间刻画环境跳转过程的模块, 使用 VQ-VAE 对当前视觉观测进行离散化的词元表征, 以自回归的方式直接预测未来的视觉观测词元, 而非预测出未来的隐变量表征后再间接预测重建的视觉观测. IRIS 使用更为直接的观测预测带来了更为精确的观测重建, 在十分苛求样本效率的 Atari-100K 环境里给出了卓越的表现. REM 算法 [271] 考

虑了 Transformer 在生成轨迹时自回归串行的推理效率瓶颈, 提出了并行观测预测机制, 将推理效率提升至 IRIS 的 7∼8 倍. STORM 算法 [272] 在 TWM 的工作基础上借鉴 IRIS 优势, 并融合观测和动作两种模态形成单一词元, 在 Atari-100K 上取得了最好的结果. MARIE [273] 算法提出了多智能体领域的 Transformer 世界模型, 通过结合分散式共享的世界模型和集中式智能体表征获得了更好的多智能体协同表现. iVideoGPT [274] 提出了高效的词元压缩表示, 通过人类操作和机器人视频预测进行 Transformer 模型预训练, 随后添加动作条件和奖励预测进行微调. iVideoGPT 可用于视频轨迹生成和规划, 能够和强化学习算法结合进行策略学习. Genie 算法 [275] 使用大量视频数据训练词元编码器和隐动作模型, 随后构建时空 Transformer 网络进行视频预测的环境模型训练, 能够进行以动作为条件的长序列视频生成.

Diffusion 世界模型在近期受到了广泛关注, OpenAI 提出的 Sora 视频生成模型 [132] 被认为是世界模拟器. 与隐空间世界模型不同, Sora 可以根据语言描述在原始的图像空间中生成多步的图像预测, 组成长达 60s 的内容连贯的视频. 在实现上, Sora 使用编码网络将视频和图像表示为词元, 随后使用超大规模的扩散模型在编码中进行加噪和去噪流程, 随后将去噪后的词元映射到原始的图像空间中. Sora 在具身智能任务中有着广泛的应用前景, 可以根据机器人任务的描述和轨迹先验生成智能体在后续时间步的轨迹视频, 将生成的视频序列用于基于模型的强化学习、蒙特卡洛树搜索、MPC 算法中. 在 Sora 大规模扩散模型提出之前, 已有多个小规模的扩散模型用于具身智能数据生成. SynthER 算法 [276] 使用扩散模型学习低维的强化学习离线轨迹数据集, 可以生成轨迹数据对原始数据进行增强. 在相同的离线强化学习算法训练下, SynthER 产生的数据能够获得比原始数据集训练更好的效果, 表明扩散模型可以学习轨迹数据中的状态表示和动力学方程, 生成的数据和原始轨迹保持了高度一致性. MTDiff 算法 [136] 将扩散模型用于多任务轨迹生成, 提出使用任务专家轨迹作为 prompt 来指导生成符合该任务目标和动力学的智能体轨迹. MTDiff 可以直接使用混合的多任务智能体轨迹进行学习, 根据任务提示生成多样化的轨迹, 从而有望将扩散模型用于大规模数据集上的通用决策问题. UniPi 算法 [277] 提出直接在图像空间对智能体的轨迹进行建模, 使用扩散模型根据语言输入和初始图像对未来的关键视频帧进行生成, 随后在时间序列进行超分辨获得一致性增强的密集图像序列. UniPi 通过训练逆环境模型对生成的视频进行动作补全, 可以直接用于智能体决策. UniSim 算法 [278] 进一步增强了扩散模型在轨迹预测方面的性能, 使用互联网数据和机器人交互视频进行联合训练, 得到的模型能够根据高层和低层任务指令对长序列视频轨迹进行预测, 类似于真实世界的模拟器. RoboDreamer 算法 [279] 将组合指令分解为单个指令并分别使用扩散模型进行视频轨迹生成, 同时可以扩展到多模态的指令集合上, 如目标图像、目标草图等. VPDD 算法 [280] 使用大规模人类操作数据集训练轨迹预测模型, 随后加入少量动作标签数据微调动作生成模块, 减轻了对大量机器人交互数据的需求. Pandora 架构 [281] 结合了预训练的大语言模型和扩散模型视频生成器, 通过以动作为条件的视频预测训练和指令微调进行可控视频生成.

7.2 仿真环境的数据生成

在使用世界模型中编码的知识进行数据生成之外, 大模型可以借助现有的仿真环境进行自动化的环境生成和数据采集. 大语言模型的使用可以大大提升任务的多样性, 降低任务仿真环境编写的难度, 提升机器人数据的多样性. GenSim 框架 [282] 是 MIT 提出使用大语言模型进行自动任务提出、自动环境构造、自动任务解决、自动数据采集的全流程框架. 首先, 大语言模型首先根据简短的任务描述和任务需求来产生相应任务的仿真场景代码搭建, 同时提供了一套自动化的流程来验证仿真环境的可行性并进行迭代的修正. 其次, 根据不同任务生成的仿真环境构建一个高质量大模型生成的任务库, 用

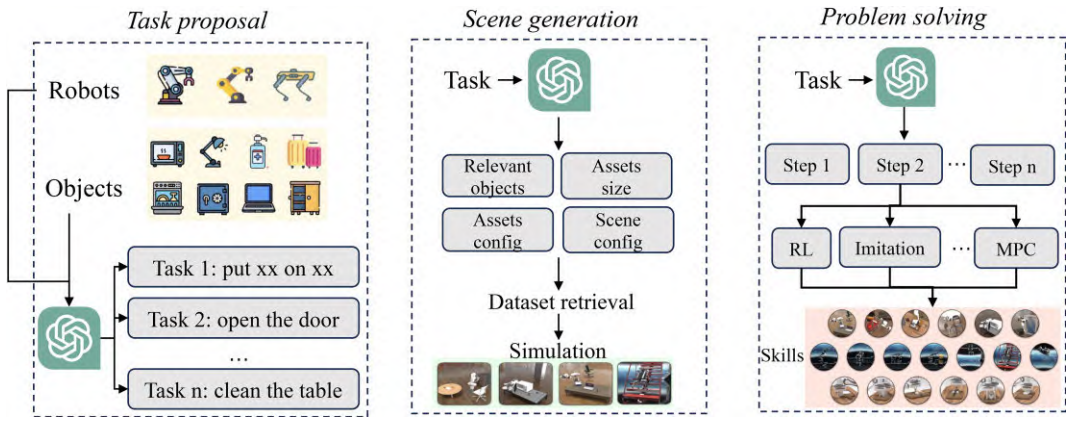


图 17　(网络版彩图) RoboGen 仿真环境和数据生成框架

Figure 17　(Color online) The data generation framework of RoboGen

表 5　大模型驱动的具身数据生成

Table 5　Category and representative algorithms for LLM-based embodied data generation

Content structure	Category	Representative algorithms
World Model	Latent-space model	Dreamer [83], Dreamer-v2 [84], Dreamer-v3 [263], TD-MPC [232], TD-MPC2 [264], TD-MPC-Offline [265], MV-MWM [266], APV [63], ContextWM [267], Dynalang [211]
	Transformer model	TT [268], Trans-Dreamer [269], TWM [270], IRIS [233], REM [271], STORM [272], MARIE [273], GR-1 [64], iVideoGPT [274], Genie [275]
	Diffusion model	SORA [132], SynthER [276], MTDiff [136], UniPi [277], UniSim [278], RoboDreamer [279], VPDD [280], Pandora [281]
Simulation data	LLM-generated	GenSim [282], RoboGen [283]

于在构建新任务时进行检索和反馈优化. 随后, 根据任务搭建的流程可以从中采集大量专家数据, 在现有模仿学习架构的基础上训练模仿学习策略. GenSim 框架可以用于根据目标任务来搭建仿真环境并产生目标任务的数据, 也可以利用大模型的能力进行探索来产生新颖的任务和数据. GenSim 的不足之处是仅面向 Ravens 仿真器中的机械臂抓取任务. RoboGen 框架 [283] 进一步提出更为通用的仿真环境生成器, 可以在机械臂、移动机器人、四足机器人、灵巧手等主流的具身实体上生成仿真环境. 首先, RoboGen 使用现有仿真环境中存在的物体, 大语言模型根据对物体功能及如何进行交互的理解来提出有意义的任务; 随后, 大模型根据任务中需要的场景和目标, 通过调用仿真器底层函数来搭建符合该任务描述的仿真环境; 大模型对任务进行分解, 对子任务选择强化学习、模仿学习、轨迹优化等算法对任务进行求解, 最终产生能够解决不同任务的策略和数据. RoboGen 框架的主体框架如图 17 所示. 算法总结如表 5 所示.

8　现有研究的联系和存在的挑战

大模型驱动的具身智能已经环境感知、任务规划、基础策略、奖励函数和数据生成方面获得了蓬勃发展, 有望使 “数字” 大模型在 “实体” 机器人中生根发芽. 现有技术存在诸多联系, 总结如下: (1) **环境感知**能够赋能下游的任务规划和策略学习过程, 通过对场景的理解分析, 将场景信息转为自然语言描述或词元表示, 随后使用大模型进行任务分解和策略学习. 如近期提出的 ManipLLM [284] 框

架将环境感知和任务执行结合在一起, 通过大模型问答使大模型输出哪些像素点可以用来操作物体, 并根据视觉输入和 CoT 流程引导大模型输出机械臂末端的操作向量对物体进行操作. (2) **任务规划**模块是大模型解决复杂任务的核心, 能够和其他方面进行广泛结合. 首先, 任务规划模块依赖环境感知模块将对环境的理解转换为自然语言描述, 从而更好的利用大语言模型的推理能力; 其次, 任务规划模块获得的基础任务单元依赖于机器人行为策略进行执行, 其中可能调用大模型基础策略. 例如 VoxPoser [241] 框架中大模型产生的规划依赖于底层 MPC 控制器进行执行. (3) **大模型基础策略**框架往往隐含的进行了环境感知理解和任务规划部分, 例如 OpenVLA [224] 框架使用预训练的视觉编码器进行环境感知, 将环境信息词元和语言指令混合输入大语言模型, 随后大语言模型完成隐含的任务规划并输出代表动作的词元进行动作执行; LCB [285] 框架利用了大语言模型的规划能力生成了隐含目标用于指导底层策略的学习. (4) **大模型奖励函数**能够和下游的强化学习算法进行结合, 用于具身策略学习. 同时, 偏好驱动的奖励函数生成能够用于大模型任务规划和基础策略于人类偏好的对齐, 进一步提升大模型策略的安全性和对人类指令的理解能力. 例如, RL4VLM [286] 在仿真卡牌任务中使用大语言模型产生的规划和真实环境进行交互, 根据奖励和强化学习算法对大模型进行微调, 从而提升大模型在特定环境中的适应能力. (5) **数据生成**是大模型具身智能的基础模块, 能够强有力的赋能其他模块. 通过对特定场景的具身数据生成, 使大模型能够获得更多的数据进行策略学习和任务规划的微调, 从而提升大模型的能力. 大模型环境感知也依赖于仿真环境提供的感知数据, 从而获得精细的环境感知 (如触觉、力觉等) 指导智能体完成复杂任务.

然而, 现有大模型驱动的具身智能研究仍然处在初级阶段, 面临着如下挑战.

大模型在特定具身场景中的适应问题. 在特定的应用场景中, 机器人的硬件构造和需要完成的任务有一定的特殊性, 并非在大多数研究中使用的标准机器人和典型任务. 在大模型规划中, 现有的闭环反馈方法虽然能够根据环境反馈和值函数反馈纠正大模型在特定具身任务规划中产生的 "幻觉", 但环境反馈纠正需要较长的交互轮次, 且特定机器人数据可能会较为缺乏, 往往难以训练准确的值函数. 此外, 大模型具身核心部分是虚实一体, 需要建立具有物理特性的数字孪生场景. 使得大模型不仅能够在虚拟环境中进行复杂任务的学习和优化, 还能通过数字孪生场景的反馈机制, 不断调整和提升其在现实世界中的表现. 然而, 高保真数字孪生的构建、模型训练和优化、感知与操作的集成、知识库的构建与维护以及系统的适应与泛化能力等方面存在诸多挑战. 随着虚实一体技术的不断发展和完善, 大模型驱动的具身智能将能够实现更高水平的自主优化和自适应能力.

大模型策略和人类偏好的对齐问题. 大模型在训练中使用了大量人类对话数据, 对于特定问题能够给出全面的、多样化的回答. 然而, 具身任务在执行中更需要大模型给出简洁的、准确的、安全的指令用于执行. 如何将大模型的规划能力和在具身智能的特定偏好需求进行对齐是未来需要解决的问题. 现有任务规划和奖励生成中使用人类反馈对大模型生成的规划和奖励代码进行调整, 在基础策略中使用专家数据进行微调, 能够在一定程度上将使大模型与人类偏好进行对齐. 在未来, 具身智能有望依托大模型中更为先进的人类反馈强化学习 (RLHF) 方法对大模型进行微调 [287], 通过构造类似大模型安全对齐的具身安全对齐数据集 [288], 使大模型在具身规划中的输出更加符合人类偏好. 在具身任务中定义风险损失函数, 通过在策略学习和微调中构建安全约束的优化学习目标 [261], 对模仿学习或强化学习得到的策略进行进一步的约束, 可以提升大模型策略的安全性.

具身策略的跨域泛化问题. 与视觉和自然语言理解中数据拥有标准的模式不同, 具身任务往往涉及多样的实体类型和动态的环境变化. 在策略学习后, 只要智能体和环境的动力学参数稍作改变, 原有的具身策略将很难直接适用. 例如, 四足机器人在光滑路面训练得到行走策略后, 在颠簸和沙石路面将由于摩擦力的改变而难以行走. 对于此类策略跨域泛化的问题, 直接的解决方案是在大模型的提

示或训练数据中包含所有可能发生的变化情况, 并给出这些情况下智能体策略应该如何进行调整的规划, 然而这些设计需要人类的专家知识, 且往往无法覆盖可能的环境变化. 受传统的跨域动力学研究启发, 一种可行的解决是借鉴强化学习的动力学泛化方法, 对于小规模的环境扰动或仿真 - 真实迁移带来的动力学变化使用正激励噪声 [289]、域随机化、数据生成、鲁棒表征等方法进行数据增强 [290]; 对于动力学变化较大的跨域变化, 使用域分类器和值函数跨域动力学变化进行度量 [291]. 通过显式的跨域度量将标准大模型策略迁移到特定的跨域任务中, 解决智能体策略在仿真 - 真实泛化和跨域泛化中存在的问题 [292].

大模型驱动多智能体协作问题. 在解决任务中往往需要多个智能体进行协作, 其中涉及到复杂的多智能体任务分配问题. 具体地, 大模型需要根据不同智能体的角色和能力进行合理任务分配. 同时, 当智能体的数量不断增加时会给任务分配带来相当的困难. 现有研究提出了对话沟通式框架来用大模型解决多智能体问题, 然而从实际表现来看仍然有很大提升的空间 [293]. 在智能体任务分配上, 传统的中心式评价 - 分布式执行 (CTDE) 多智能体框架有望带来更多启发 [294], 使用中心大模型对智能体协作行为进行评价, 使用分布式小模型来控制单个智能体的任务执行, 并通过中心智能体反馈来最大化团队收益 [295]. 在沟通方面, 随着智能体数量的增多, 群聊式沟通往往会带来交互轮次多、分配不均衡等问题 [197]. 近期提出的 ReAD 算法通过优势函数对个体在团队中的贡献进行评价, 使用大模型元提示优化优势函数, 实现了个体策略之间的高效协同 [202]. 竞争式的多智能体结构在大模型驱动的具身智能中仍然缺乏研究, 如何使智能体在博弈和对抗中学习是未来重要的研究课题, 在智能体博弈中引入更多的博弈框架有望使大模型智能体演化出复杂的均衡态 [296]. 此外, 多智能体协作会衍生更多的人工智能伦理问题, 现有研究已经对伦理计算问题进行了相关讨论 [297].

具身智能在真实环境中所面临的挑战. 在数据获取方面, 具身智能系统需要处理多样且复杂的传感数据, 包括视觉、听觉、触觉等多种信息源. 真实环境中的数据往往是非结构化、动态变化的, 如何获取全面的、高质量的感知数据是重要的问题, 对实时采集和处理提出了更高的要求. 在平台安全性方面, 具身智能系统依赖网络进行数据传输和通信, 这使其容易成为网络攻击的目标, 确保通信的安全性和数据的隐私性至关重要. 另外, 智能机器人在物理环境中操作时, 必须确保能够避免对人类和周围环境的伤害. 这要求具身智能系统具备精确的运动控制和障碍物检测能力, 以避免意外碰撞. 在真实环境反馈方面, 真实环境中的各种干扰因素, 如光线变化、噪声、障碍物等, 也可能影响传感器的性能和数据的准确性, 系统必须具备处理和适应这些干扰的能力. 同时, 具身策略需要根据环境的负反馈进行适当调整, 从而提升对环境的适应能力.

大模型具身策略的决策实时性问题. 在现有的大模型任务规划中, 往往假设整个规划在执行过程中环境是静态的, 而环境的动态改变可能会导致之前的规划无法适用. 在具身任务中, 机器人的任务规划和策略选择都需要保持较高的频率来适应动态的环境变化. 直接调用 GPT-4 的开放接口进行决策往往难以满足实时性的要求, 许多情况下需要使用开源大模型并在充足的算力支持下进行快速推理. 此外, 为了适应具身应用, 大模型有时需要直接在具身机器人上进行运行, 为大模型的轻量化提出了更高的要求. 在未来, 如何有效的保证大模型的轻量化是重要的研究课题, 包括对大模型进行模型剪枝 [298]、模型量化 [299], 模型低秩分解 [300] 等手段, 同时需要设计大模型的具身能力度量标准, 使大模型在降低模型的复杂度的同时保持在具身推理和决策中的能力.

9 总结与展望

本文围绕大模型驱动的具身智能这一前沿交叉领域, 总结了大模型和具身智能的相关技术背景,

涵盖了基本概念、学习框架、大模型技术等, 随后从 5 个方面依次介绍了大模型驱动的环境感知、任务规划、基础策略、奖励函数、数据生成中的前沿研究工作, 并总结了目前研究的成果和挑战.

大模型驱动的具身智能是一项有着重大应用前景的研究方向, 将日益成熟的大模型技术与机器人进行结合, 用于解决机器人在感知、规划、决策中的各种问题. 虽然目前关于采取何种路线能够达到通用具身智能仍很难定论, 但大模型的成功切实的推动了具身智能的发展. 在大模型和机器人领域的不断发展下, 未来大模型驱动的具身智能研究将带来很多新的研究方向, 包括: (1) **统一具身数据平台**. 构建涵盖尽可能多的具身实体、任务类型、环境、场景、动力学的多模态快速仿真和真实数据平台, 推动数据基础设施建设. (2) **通用具身数据表征**. 机器人状态描述往往涵盖了除了视觉和语言之外的其他信息, 包括本体感知、力学/触觉传感器、位置描述等, 构建统一机器人多模态观测的具身表征是重要的问题. (3) **鲁棒具身控制策略**. 提升现有大模型进行具身决策时的安全性和鲁棒性, 避免环境干扰和外界因素带来的机器人交互危险, 提升对机器人系统对环境变化的适应能力. (4) **可控具身策略生成**. 提升大模型具身智能策略的可控性, 建立安全保障机制, 明确的安全边界和行为规范, 确保机器人的行动不会超越这些预设的限制. (5) **人机合作具身智能**. 机器人作为人类的助手需要理解人类意图, 进行增强的人类意图识别, 帮助机器人提前准备响应, 提高合作效率. (6) **异构智能体协同**. 使用大模型控制不同类型的异构智能体进行高效协同, 结合多智能体强化学习进行合作策略学习和引导, 建立统一的伦理标准和操作准则, 确保所有智能体协同工作时遵循相同的伦理约束. (7) **轻量化具身策略**. 推动大模型和小模型结合的方式, 以轻量化模型和较少的计算代价解决具身智能任务. (8) **人形机器人**. 推动运动控制、抓取操作、导航、灵巧操作等具身技能集于一身, 更好地为人类服务.

参考文献

1 Smith L, Gasser M. The development of embodied cognition: six lessons from babies. Artif Life, 2005, 11: 13–29
2 Turing A M. Computing machinery and intelligence. Mind, 1950, 59: 433–460
3 Turing A M. Computing Machinery and Intelligence. Springer, 2009
4 Brooks R A. New approaches to robotics. Science, 1991, 253: 1227–1232
5 Duan J F, Yu S, Tan H, et al. A survey of embodied AI: from simulators to research tasks. IEEE Trans Emerg Topic Computat Intell, 2022, 6: 230–244
6 OpenAI. GPT-4 Technical Report. 2023
7 Dong Q X, Li L, Dai D M, et al. A survey for in-context learning. arXiv, 2022. arXiv:2301.00234
8 Wei J, Wang X Z, Schuurmans D, et al. Chain of thought prompting elicits reasoning in large language models. In: Neural Information Processing Systems, 2022
9 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
10 Li J N, Li D X, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning, 2023
11 Li C Y. Large multimodal models: notes on CVPR 2023 tutorial. arXiv, 2023. arXiv:2306.14895
12 Alayrac J-B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. Neur Inform Process Syst, 2022
13 Zare M, Kebria P M, Khosravi A, et al. A survey of imitation learning: algorithms, recent developments, and challenges. arXiv, 2023. arXiv:2309.02473
14 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. MIT Press, 2018
15 Moerland T M, Broekens J, Plaat A, et al. Model-based reinforcement learning: a survey. Found Trend Mach Learn, 2023, 16: 1–118
16 Firoozi R, Tucker J, Tian S, et al. Foundation models in robotics: applications, challenges, and the future. arXiv,

2023. arXiv:2312.07843
17 Song C H, Wu J M, Washington C, et al. LLM-planner: few-shot grounded planning for embodied agents with large language models. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2023
18 Xie T B, Zhao S H, Wu C H, et al. Text2Reward: dense reward generation with language models for reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2024
19 Zhao Z R, Lee W S, Hsu D. Large language models as commonsense knowledge for large-scale task planning. In: Proceedings of NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023
20 Open X. Embodiment Collaboration. Open X-Embodiment: robotic learning datasets and RT-X models. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2024
21 Todorov E, Erez T, Tassa Y. Mujoco. A physics engine for model-based control. In: Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012. 5026–5033
22 Tassa Y, Doron Y, Muldal A, et al. Deepmind control suite. arXiv, 2018. arXiv:1801.00690
23 Gupta A, Kumar V, Lynch C, et al. Relay policy learning: solving long-horizon tasks via imitation and reinforcement learning. arXiv, 2019. arXiv:1910.11956
24 Zhu Y, Wong J, Mandlekar A, et al. robosuite: a modular simulation framework and benchmark for robot learning. arXiv, 2020. arXiv:2009.12293
25 Mu T Z, Ling Z, Xiang F B, et al. Maniskill: generalizable manipulation skill benchmark with large-scale demonstrations. In: Proceedings of Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021
26 Gu J Y, Xiang F B, Li X L, et al. Maniskill2: a unified benchmark for generalizable manipulation skills. In: Proceedings of International Conference on Learning Representations, 2022
27 Kalashnikov D, Varley J, Chebotar Y, et al. MT-Opt: continuous multi-task robotic reinforcement learning at scale. arXiv, 2021. arXiv:2104.08212
28 Walke H R, Black K, Z Zhao T Z, et al. Bridgedata v2: a dataset for robot learning at scale. In: Proceedings of Conference on Robot Learning, 2023. 1723–1736
29 Fang H-S, Fang H J, Tang Z Y, et al. Rh20t: a comprehensive robotic dataset for learning diverse skills in one-shot. In: Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition, 2023
30 Chi C, Xu Z J, Pan C, et al. Universal manipulation interface: in-the-wild robot teaching without in-the-wild robots. arXiv, 2024. arXiv:2402.10329
31 Makoviychuk V, Wawrzyniak L, Guo Y R, et al. Isaac Gym: high performance GPU based physics simulation for robot learning. In: Proceedings of Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021
32 Coumans E, Bai Y. Pybullet Quickstart Guide, 2021
33 Hwangbo J, Lee J, Hutter M. Per-contact iteration method for solving contact dynamics. IEEE Robot Autom Lett, 2018, 3: 895–902
34 Michel O. Webots: symbiosis between virtual and real mobile robots. In: Proceedings of Virtual Worlds: First International Conference, Paris, 1998. 254–263
35 Peng X B, Coumans E, Zhang T, et al. Learning agile robotic locomotion skills by imitating animals. In: Robotics: Science and Systems, 2020
36 Han L, Zhu Q, Sheng J, et al. Lifelike agility and play on quadrupedal robots using reinforcement learning and generative pre-trained models. arXiv, 2023. arXiv:2308.15143
37 Brohan A, Brown N, Carbajal J, et al. RT-1: robotics transformer for real-world control at scale. In: Robotics: Science and Systems, 2023
38 Zitkovich B, Yu T, Xu S, et al. RT-2: vision-language-action models transfer web knowledge to robotic control. In: Proceedings of Annual Conference on Robot Learning, 2023
39 Fu Z, Zhao T Z, Finn C. Mobile Aloha: learning bimanual mobile manipulation with low-cost wholebody teleoperation. arXiv, 2024. arXiv:2401.02117

40 Xu Y, Wan W, Zhang J, et al. Unidexgrasp: universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 4737–4746

41 Chao Y-W, Paxton C, Xiang Y, et al. Handoversim: a simulation framework and benchmark for human-to-robot object handovers. In: Proceedings of 2022 International Conference on Robotics and Automation (ICRA), 2022. 6941–6947

42 Rajeswaran A, Kumar V, Gupta A, et al. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In: Robotics: Science and Systems, 2018

43 He H, Bai C, Lai H, et al. Privileged knowledge distillation for sim-to-real policy generalization. arXiv, 2023. arXiv:2305.18464

44 Shi J, Bai C, He H, et al. Robust quadrupedal locomotion via risk-averse policy learning. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2024

45 Long J, Wang Z, Quanyi Li Q, et al. Hybrid internal model: a simple and efficient learner for agile legged locomotion. arXiv, 2023. arXiv:2312.11460

46 Ze Y, Zhang G, Zhang K, et al. 3D diffusion policy. arXiv, 2023. arXiv:2312.11460

47 Jiang Y, Gupta A, Zhang Z, et al. VIMA: general robot manipulation with multimodal prompts. In: Proceedings of International Conference on Machine Learning, 2023

48 Mees O, Hermann L, Rosete-Beas E, et al. Calvin: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robot Autom Lett, 2022, 7: 7327–7334

49 Li X L, Huang G, Wang Z G, et al. Optics-driven drone. Sci China Inf Sci, 2024, 67: 124201

50 Ho J, Ermon S. Generative adversarial imitation learning. In: Neural Information Processing Systems, 2016

51 Hoque R, Balakrishna A, Putterman C, et al. Lazydagger: reducing context switching in interactive imitation learning. In: Proceedings of 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), 2021. 502–509

52 Zhang J, Cho K. Query-efficient imitation learning for end-to-end simulated driving. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2017

53 Hoque R, Balakrishna A, Novoseller E, et al. Thriftydagger: budget-aware novelty and risk gating for interactive imitation learning. In: Proceedings of Conference on Robot Learning, 2022. 598–608

54 Orsini M, Raichuk A, Hussenot L, et al. What matters for adversarial imitation learning? In: Neural Information Processing Systems, 2021. 14656–14668

55 Wang Z, Hunt J J, Zhou M. Diffusion policies as an expressive policy class for offline reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2023

56 Pearce T, Rashid T, Kanervisto A, et al. Imitating human behaviour with diffusion models. In: Proceedings of International Conference on Learning Representations, 2023

57 Chi C, Feng S, Du Y, et al. Diffusion policy: visuomotor policy learning via action diffusion. In: Neural Information Processing Systems, 2023

58 Pearce T, Rashid T, Kanervisto A, et al. Imitating human behaviour with diffusion models. In: Proceedings of International Conference on Learning Representations, 2023

59 Baker B, Akkaya I, Zhokov P, et al. Video pretraining (VPT): learning to act by watching unlabeled online videos. In: Neural Information Processing Systems, 2022. 24639–24654

60 Aytar Y, Pfaff T, Budden D, et al. Playing hard exploration games by watching YouTube. In: Neural Information Processing Systems, 2018

61 Sermanet P, Lynch C, Chebotar Y, et al. Time-contrastive networks: self-supervised learning from video. In: Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018. 1134–1141

62 Haldar S, Mathur V, Yarats D, et al. Watch and match: supercharging imitation with regularized optimal transport. In: Proceedings of Conference on Robot Learning, 2023. 32–43

63 Seo Y, Lee K, James S L, et al. Reinforcement learning with action-free pre-training from videos. In: Proceedings of

International Conference on Machine Learning, 2022. 19561–9579
64 Wu H, Jing Y, Cheang C, et al. Unleashing large-scale video generative pre-training for visual robot manipulation. In: Proceedings of International Conference on Learning Representations, 2024
65 Arora S, Doshi P. A survey of inverse reinforcement learning: challenges, methods and progress. Artif Intell, 2021, 297: 103500
66 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
67 Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2016
68 Lan Q, Pan Y, Fyshe A, et al. Maxmin Q-learning: controlling the estimation bias of Q-learning. In: Proceedings of International Conference on Learning Representations, 2020
69 Lee S Y, Sungik C, Chung S-Y. Sample-efficient deep reinforcement learning via episodic backward update. In: Neural Information Processing Systems, 2019
70 Chen X, Wang C, Zhou Z, et al. Randomized ensembled double Q-learning: learning fast without a model. In: Proceedings of International Conference on Learning Representations, 2020
71 Bai C J, Wang L X, Han L, et al. Principled exploration via optimistic bootstrapping and backward induction. In: Proceedings of International Conference on Machine Learning, 2021. 577–587
72 Li Z, Li Y, Zhang Y, et al. HyperDQN: a randomized exploration method for deep reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2022
73 Sutton R S, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation. In: Neural Information Processing Systems, 1999
74 Mei J, Chung W, Thomas V, et al. The role of baselines in policy gradient optimization. In: Neural Information Processing Systems, 2022. 17818–17830
75 Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation. arXiv, 2015. arXiv:1506.02438
76 Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization. In: Proceedings of International Conference on Machine Learning, 2015. 1889–1897
77 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv, 2017. arXiv:1707.06347
78 Moerland T M, Broekens J, Plaat A, et al. Model-based reinforcement learning: a survey. Found Trend Machine Learn, 2023, 16: 1–18
79 Sutton R S. Dyna, an integrated architecture for learning, planning, and reacting. ACM Sigart Bullet, 1991, 2: 160–163
80 Luo Y, Xu H, Li Y, et al. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In: Proceedings of International Conference on Learning Representations, 2019
81 Janner M, Fu J, Zhang M, et al. When to trust your model: model-based policy optimization. In: Neural Information Processing Systems, 2019
82 Filos A, Vertes E, Marinho Z, et al. Model-value inconsistency as a signal for epistemic uncertainty. In: Proceedings of International Conference on Machine Learning, 2022. 6474–6498
83 Hafner D, Lillicrap T, Ba J, et al. Dream to control: learning behaviors by latent imagination. In: Proceedings of International Conference on Learning Representations, 2020
84 Hafner D, Lillicrap T, Norouzi M, et al. Mastering Atari with discrete world models. In: Proceedings of International Conference on Learning Representations, 2021
85 Lange S, Gabel T, Riedmiller M. Batch reinforcement learning. In: Reinforcement Learning: State-of-the-Art, 2012. 45–53
86 Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: Proceedings of International Conference on Machine Learning, 2019. 2052–2062
87 Fujimoto S, Gu S S. A minimalist approach to offline reinforcement learning. In: Neural Information Processing

Systems, 2021. 20132–20145
88 Wu Y, Zhai S, Srivastava N, et al. Uncertainty weighted actor-critic for offline reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2021. 11319–11328
89 An G, Moon S, Kim J-H, et al. Uncertainty-based offline reinforcement learning with diversified qensemble. In: Neural Information Processing Systems, 2021. 7436–7447
90 Kumar A, Zhou A, Tucker G, et al. Conservative Q-learning for offline reinforcement learning. In: Neural Information Processing Systems, 2020. 1179–1191
91 Yu T, Thomas G, Yu L, et al. MOPO: model-based offline policy optimization. In: Neural Information Processing Systems, 2020. 14129–14142
92 Bharadhwaj H, Xie K, Shkurti F. Model-predictive control via cross-entropy and gradient-based optimization. In: Learning for Dynamics and Control, 2020. 277–286
93 Chua K, Calandra R, McAllister R, et al. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In: Neural Information Processing Systems, 2018
94 Li X L. Multi-modal cognitive computing. Sci Sin Inform, 2023, 53: 1–32 [李学龙. 多模态认知计算. 中国科学: 信息科学, 2023, 53: 1–32]
95 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Neural Information Processing Systems, 2014
96 Neubig G. Neural machine translation and sequence-to-sequence models: a tutorial. arXiv, 2017. arXiv:1703.01619
97 Devlin J, Chang M-W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186
98 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Neural Information Processing Systems, 2020. 1877–1901
99 Touvron H, Lavril T, Izacard G, et al. LLAMA: open and efficient foundation language models. arXiv, 2023. arXiv:2302.13971
100 Touvron H, Martin L, Stone K, et al. LLAMA 2: open foundation and fine-tuned chat models. arXiv, 2023. arXiv:2307.09288
101 Chiang W-L, Li Z, Lin Z, et al. Vicuna: an open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org (accessed 14 April 2023), 2023
102 Javaheripi M, Bubeck S, Abdin M, et al. Phi-2: the surprising power of small language models. Microsoft Research Blog, 2023
103 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
104 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2021
105 He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 16000–16009
106 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of International Conference on Machine Learning, 2020. 1597–1607
107 Cherti M, Beaumont R, Wightman R, et al. Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 2818–2829
108 Shen S, Li C, Hu X, et al. K-lite: learning transferable visual models with external knowledge. In: Neural Information Processing Systems, 2022. 15558–15573
109 Yao L, Huang R, Hou L, et al. FILIP: fine-grained interactive language-image pre-training. In: Proceedings of International Conference on Learning Representations, 2022
110 Bao H, Dong L, Piao S, et al. Beit: bert pre-training of image transformers. In: Proceedings of International Conference on Learning Representations, 2022

111 Gan Z, Li L, Li C, et al. Vision-language pre-training: basics, recent advances, and future trends. Found Trend Comput Graph Vis, 2022, 14: 163–352

112 Kirillov A, Mintun E, Ravi N, et al. Segment anything. arXiv, 2023. arXiv:2304.02643

113 Mazurowski M A, Dong H, Gu H, et al. Segment anything model for medical image analysis: an experimental study. Med Imag Anal, 2023, 89: 102918

114 Wang Z, Ze Y, Sun Y, et al. Generalizable visual reinforcement learning with segment anything model. arXiv, 2023. arXiv:2312.17116

115 Wang Z, Yu J, Yu A W, et al. SimVLM: simple visual language model pretraining with weak supervision. In: Proceedings of International Conference on Learning Representations, 2022

116 Li J, Li D, Xiong C, et al. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of International Conference on Machine Learning, 2022. 12888–12900

117 Li J, Li D, Savarese S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning, 2023

118 Li X, Liu M, Zhang H, et al. Vision-language foundation models as effective robot imitators. In: Proceedings of International Conference on Learning Representations, 2024

119 Liu H, Li C, Wu Q, et al. Visual instruction tuning. In: Neural Information Processing Systems, 2023

120 Zhu D, Chen J, Shen X, et al. Minigpt-4: enhancing vision-language understanding with advanced large language models. arXiv, 2023. arXiv:2304.10592

121 Dai W, Li J, Li D, et al. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In: Proceedings of Thirty-seventh Conference on Neural Information Processing Systems, 2023

122 Li K, He Y, Wang Y, et al. Videochat: chat-centric video understanding. arXiv, 2023. arXiv:2305.06355

123 Zhang H, Li X, Bing L. Video-llama: an instruction-tuned audio-visual language model for video understanding. arXiv, 2023. arXiv:2306.02858

124 Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. In: Neural Information Processing Systems, 2021. 8780–8794

125 Kingma D P, Welling M. Auto-encoding variational bayes. arXiv, 2013. arXiv:1312.6114

126 Van den Oord A, Kalchbrenner N, Espeholt L, et al. Conditional image generation with pixelcnn decoders. In: Neural Information Processing Systems, 2016

127 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Neural Information Processing Systems, 2014

128 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Neural Information Processing Systems, 2020. 6840–6851

129 Epstein D, Jabri A, Poole B, et al. Diffusion self-guidance for controllable image generation. In: Neural Information Processing Systems, 2024

130 Clark K, Jaini P. Text-to-image diffusion models are zero shot classifiers. In: Neural Information Processing Systems, 2024

131 Betker J, Goh G, Jing L, et al. Improving Image Generation with Better Captions. https://cdn.openai.com/papers/dall-e-3.pdf

132 OpenAI. Video generation models as world simulators. Open AI Blog, 2024

133 Janner M, Du Y, Tenenbaum J B, et al. Planning with diffusion for flexible behavior synthesis. In: Proceedings of International Conference on Machine Learning, 2022. 9902–9915

134 Ni F, Hao J, Mu Y, et al. Metadiffuser: diffusion model as conditional planner for offline Meta-RL. In: Proceedings of International Conference on Machine Learning, 2023. 26087–26105

135 Gao J, Hu K, Xu G, et al. Can pre-trained text-to-image models generate visual goals for reinforcement learning? In: Neural Information Processing Systems, 2024

136 He H R, Bai C J, Xu K, et al. Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning. In: Neural Information Processing Systems, 2023

137 Jin Y, Sun Z, Xu K, et al. Video-laVIT: unified video-language pre-training with decoupled visual-motional tokenization. In: Proceedings of the Forty-first International Conference on Machine Learning, 2024

138 Sinha S, Mandlekar A, Garg A. S4rl: surprisingly simple self-supervision for offline reinforcement learning in robotics. In: Proceedings of the 5th Annual Conference on Robot Learning, 2021

139 Kostrikov I, Yarats D, Fergus R. Image augmentation is all you need: regularizing deep reinforcement learning from pixels. arXiv, 2020. arXiv:2004.13649

140 Van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv, 2018. arXiv:1807.03748

141 Laskin M, Srinivas A, Abbeel P. Curl: contrastive unsupervised representations for reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2020. 5639–5650

142 Sermanet P, Lynch C, Chebotar Y, et al. Time-contrastive networks: self-supervised learning from video. In: Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018. 1134–1141

143 Nguyen T D, Shu R, Pham T, et al. Temporal predictive coding for model-based planning in latent space. In: Proceedings of International Conference on Machine Learning, 2021. 8130–8139

144 Bai C J, Wang L X, Han L, et al. Dynamic bottleneck for robust self-supervised exploration. In: Neural Information Processing Systems, 2021. 17007–17020

145 Yang M, Nachum O. Representation matters: offline pretraining for sequential decision making. In: Proceedings of International Conference on Learning Representations, 2021. 11784–11794

146 Seo Y, Hafner D, Liu H, et al. Masked world models for visual control. In: Proceedings of the 6th Annual Conference on Robot Learning, 2022

147 Zhang A, McAllister R T, Calandra R, et al. Learning invariant representations for reinforcement learning without reconstruction. In: Proceedings of International Conference on Learning Representations, 2021

148 Zang H, Li X, Zhang L, et al. Understanding and addressing the pitfalls of bisimulation-based representations in offline reinforcement learning. In: Neural Information Processing Systems, 2023

149 Bhateja C, Guo D, Ghosh D, et al. Robotic offline RL from internet videos via value-function pre-training. In: Proceedings of NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023

150 Lehnert L, Littman M L. Successor features combine elements of model-free and model-based reinforcement learning. J Mach Learn Res, 2020, 21: 8030–8082

151 Yuan Z, Xue Z, Yuan B, et al. Pre-trained image encoder for generalizable visual reinforcement learning. In: Neural Information Processing Systems, 2022. 13022–13037

152 Parisi S, Rajeswaran A, Purushwalkam S, et al. The unsurprising effectiveness of pre-trained vision models for control. In: Proceedings of International Conference on Machine Learning, 2022. 17359–17371

153 Yuan Z, Yang S, Hua P, et al. Rl-vigen: a reinforcement learning benchmark for visual generalization. In: Neural Information Processing Systems, 2023

154 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis, 2014, 115: 211–252

155 Grauman K, Westbury A, Byrne E, et al. Ego4D: around the world in 3,000 hours of egocentric video. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 18995–19012

156 Damen D, Doughty H, Farinella G M, et al. The epic-kitchens dataset: collection, challenges and baselines. IEEE Trans Patt Anal Mach Intell, 2021, 43: 4125–4141

157 Goyal R, Kahou S E, Michalski V, et al. The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 5842–5850

158 Nair S, Rajeswaran A, Kumar V, et al. R3m: a universal visual representation for robot manipulation. In: Proceedings of the 6th Annual Conference on Robot Learning, 2022

159 Xiao T, Radosavovic I, Darrell T, et al. Masked visual pre-training for motor control. arXiv, 2022. arXiv:2203.06173

160 Ma Y J, Sodhani S, Jayaraman D, et al. VIP: towards universal visual reward and representation via value-implicit pre-training. In: Proceedings of International Conference on Learning Representations, 2023

161 Karamcheti S, Nair S, Chen A S, et al. Language-driven representation learning for robotics. In: Robotics: Science and Systems (RSS), 2023
162 Huo M, Ding M, Xu C, et al. Human-oriented representation learning for robotic manipulation. arXiv, 2023. arXiv:2310.03023
163 Lv Q, Li H, Deng X, et al. Robomp2: a robotic multimodal perception-planning framework with multimodal large language models. In: Proceedings of International Conference on Machine Learning, 2024
164 Jing Y, Zhu X, Liu X, et al. Exploring visual pre-training for robot manipulation: Datasets, models and methods. In: Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023. 11390–11395
165 Hansen N, Yuan Z, Ze Y, et al. On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline. In: Proceedings of International Conference on Machine Learning, 2023. 12511–12526
166 Huang S, Jiang Z, Dong H, et al. Instruct2act: mapping multi-modality instructions to robotic arm actions with large language model. arXiv, 2024. arXiv:2305.11176
167 Nagarajan T, Feichtenhofer C, Grauman K. Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 8688–8697
168 Goyal M, Modi S, Goyal R, et al. Human hands as probes for interactive object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 3293–3303
169 Liu S, Tripathi S, Majumdar S, et al. Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 3282–3292
170 Bahl S, Mendonca R, Chen L, et al. Affordances from human videos as a versatile representation for robotics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 13778–13790
171 Wang C, Fan L, Sun J, et al. Mimicplay: long-horizon imitation learning by watching human play. In: Proceedings of the 7th Annual Conference on Robot Learning, 2023
172 Bahl S, Gupta A, Pathak D. Human-to-robot imitation in the wild. In: Robotics: Science and Systems, 2022
173 Xiong H, Li Q, Chen Y-C, et al. Learning by watching: physical imitation of manipulation skills from human videos. In: Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021. 7827–7834
174 Wen C, Lin X, So J, et al. Any-point trajectory modeling for policy learning. arXiv, 2023. arXiv:2401.00025
175 Fang H-S, Gou M, Wang C, et al. Robust grasping across diverse sensor qualities: the graspnet-1billion dataset. Int J Robot Res, 2023
176 Fang H-S, Wang C, Gou M, et al. Graspnet-1billion: a large-scale benchmark for general object grasping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 11444–11453
177 Jaegle A, Borgeaud S, Alayrac J-B, et al. Perceiver IO: a general architecture for structured inputs outputs. In: Proceedings of International Conference on Learning Representations, 2022
178 James S, Wada K, Laidlow T, et al. Coarse-to-fine Q-attention: efficient learning for visual robotic manipulation via discretisation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 13739–13748
179 Shridhar M, Manuelli L, Fox D. Perceiver-actor: a multi-task transformer for robotic manipulation. In: Proceedings of the 6th Annual Conference on Robot Learning, 2022
180 Chen S, Pinel R G, Schmid C, et al. Polarnet: 3D point clouds for language-guided robotic manipulation. In: Proceedings of the 7th Annual Conference on Robot Learning, 2023
181 Zhang T, Hu Y, Cui H, et al. A universal semantic-geometric representation for robotic manipulation. In: Proceedings of the 7th Annual Conference on Robot Learning, 2023
182 Eisner B, Zhang H, Held D. Flowbot3D: learning 3D articulation flow to manipulate articulated objects. In: Robotics: Science and Systems (RSS), 2022
183 Qi C R, Yi L, Su H, et al. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Neural Information Processing Systems, 2017

184 Qian G, Li Y, Peng H, et al. Pointnext: revisiting Pointnet++ with improved training and scaling strategies. In: Neural Information Processing Systems, 2022. 23192–23204
185 Mo K, Guibas L J, Mukadam M, et al. Where2act: from pixels to actions for articulated 3D objects. In: Proceedings of International Conference on Computer Vision, 2021. 6793–6803
186 Wu R, Cheng K, Zhao Y, et al. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. In: Neural Information Processing Systems, 2023
187 Vemprala S, Bonatti R, Bucker A, et al. ChatGPT for robotics: design principles and model abilities. Microsoft Auton Syst Robot Res, 2023, 2: 20
188 Huang W, Abbeel P, Pathak D, et al. Language models as zero-shot planners: extracting actionable knowledge for embodied agents. In: Proceedings of International Conference on Machine Learning, 2022. 9118–9147
189 Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models. In: Proceedings of International Conference on Learning Representations, 2023
190 Madaan A, Tandon N, Gupta P, et al. Self-refine: iterative refinement with self-feedback. In: Neural Information Processing Systems, 2023
191 Shinn N, Cassano F, Berman E, et al. Reflexion: language agents with verbal reinforcement learning. In: Neural Information Processing Systems, 2023
192 Xie Y, Kawaguchi K, Zhao Y, et al. Self-evaluation guided beam search for reasoning. In: Neural Information Processing Systems, 2023
193 Lin B Y, Fu Y, Yang K, et al. Swiftsage: a generative agent with fast and slow thinking for complex interactive tasks. In: Neural Information Processing Systems, 2023
194 Hu Y, Lin F, Zhang T, et al. Look before you leap: unveiling the power of GPT-4v in robotic vision-language planning. arXiv, 2023. arXiv:2311.17842
195 Huang J, Chen X, Mishra S, et al. Large language models cannot self-correct reasoning yet. arXiv, 2023. arXiv:2310.01798
196 Huang W, Xia F, Xiao T, et al. Inner monologue: embodied reasoning through planning with language models. In: Proceedings of Annual Conference on Robot Learning, 2022
197 Mandi Z, Jain S, Song S. Roco: dialectic multi-robot collaboration with large language models. arXiv, 2023. arXiv:2307.04738
198 Guo Y, Wang Y-J, Zha L, et al. Doremi: grounding language model by detecting and recovering from plan-execution misalignment. arXiv, 2023. arXiv:2307.00329
199 Ahn M, Brohan A, Chebotar Y, et al. Do as I can, not as I say: grounding language in robotic affordances. In: Proceedings of Annual Conference on Robot Learning, 2022
200 Lin K, Agia C, Migimatsu T, et al. Text2motion: from natural language instructions to feasible plans. Auton Robot, 2023, 47: 1345–1365
201 Zhang D, Chen L, Zhang S, et al. Large language models are semi-parametric reinforcement learning agents.In: Neural Information Processing Systems, 2023
202 Zhang Y, Yang S, Bai C J, et al. Towards efficient LLM grounding for embodied multi-agent collaboration. arXiv, 2024. arXiv:2405.14314
203 Browne C B, Powley E, Whitehouse D, et al. A survey of Monte Carlo tree search methods. IEEE Trans Comput Intell AI Games, 2012, 4: 1–43
204 Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search. Nature, 2016, 529: 484–489
205 Yao S, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. In: Neural Information Processing Systems, 2023
206 Liu Z, Hu H, Zhang S, et al. Reason for future, act for now: a principled architecture for autonomous LLM agents. In: Proceedings of NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023
207 Feng X, Wan Z, Wen M, et al. Alphazero-like tree-search can guide large language model decoding and training. In:

Proceedings of NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023
208 Hu M, Mu Y, Yu X, et al. Tree-planner: efficient close-loop task planning with large language models. arXiv, 2023. arXiv:2310.08582
209 Murthy R, Heinecke S, Niebles J C, et al. Rex: rapid exploration and exploitation for AI agents. arXiv, 2023. arXiv:2307.08962
210 Hao S, Gu Y, Ma H, et al. Reasoning with language model is planning with world model. In: Empirical Methods in Natural Language Processing, 2023
211 Lin J, Du Y, Watkins O, et al. Learning to model the world with language. arXiv, 2023. arXiv:2308.01399
212 Liu B, Jiang Y, Zhang X, et al. LLM+P: empowering large language models with optimal planning proficiency. arXiv, 2023. arXiv:2304.11477
213 Zhou Z, Song J, Yao K, et al. ISR-LLM: iterative self-refined large language model for long-horizon sequential task planning. arXiv, 2023. arXiv:2308.13724
214 Silver T, Dan S, Srinivas K, et al. Generalized planning in PDDL domains with pretrained large language models. arXiv, 2023. arXiv:2305.11014
215 Driess D, Xia F, Sajjadi M S M, et al. PaLM-E: an embodied multimodal language model. In: Proceedings of International Conference on Machine Learning, 2023. 8469–8488
216 Mu Y, Zhang Q, Hu M, et al. Embodiedgpt: vision-language pre-training via embodied chain of thought. In: Neural Information Processing Systems, 2023
217 Shi R, Liu Y, Ze Y, et al. Unleashing the power of pre-trained language models for offline reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2024
218 Chen L, Lu K, Rajeswaran A, et al. Decision transformer: reinforcement learning via sequence modeling. In: Neural Information Processing Systems, 2021. 15084–15097
219 Hu E J, Shen Y, Wallis P, et al. LORA: low-rank adaptation of large language models. arXiv, 2021. arXiv:2106.09685
220 Li X, Liu M, Zhang H, et al. Vision-language foundation models as effective robot imitators. In: Proceedings of International Conference on Learning Representations, 2024
221 Zhang J J, Bai C J, He H R, et al. SAM-E: leveraging visual foundation model with sequence imitation for embodied manipulation. In: Proceedings of International Conference on Machine Learning, 2024. 58579–58598
222 Huang J, Yong S, Ma X, et al. An embodied generalist agent in 3D world. In: Proceedings of International Conference on Machine Learning, 2024. 20413–20451
223 Zhen H, Qiu X, Chen P, et al. 3D-VLA: a 3D vision-language-action generative world model. In: Proceedings of International Conference on Machine Learning, 2024. 61229–61245
224 Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: an open-source vision-language-action model. arXiv, 2024. arXiv:2406.09246
225 Team O M, Ghosh D, Walke H, et al. OCTO: an open-source generalist robot policy. arXiv, 2024. arXiv:2405.12213
226 Zhao T Z, Kumar V, Levine S, et al. Learning fine-grained bimanual manipulation with low-cost hardware. In: Robotics: Science and Systems, 2022
227 Guhur P L, Chen S, Pinel R G, et al. Instruction-driven history-aware policies for robotic manipulations. In: Proceedings of Conference on Robot Learning, 2023. 175–187
228 Goyal A, Xu J, Guo Y, et al. RVT: robotic view transformer for 3D object manipulation. In: Proceedings of the 7th Annual Conference on Robot Learning, 2023
229 Bousmalis K, Vezzani G, Rao D, et al. RoboCAT: a self-improving generalist agent for robotic manipulation. Trans Mach Learn Res, 2023
230 Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 12873–12883
231 Bharadhwaj H, Vakil J, Sharma M, et al. RoboAgent: generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2024

232 Hansen N, Wang X, Su H. Temporal difference learning for model predictive control. In: Proceedings of International Conference on Machine Learning, 2022

233 Micheli V, Alonso E, Fleuret F. Transformers are sample-efficient world models. In: Proceedings of International Conference on Learning Representations, 2023

234 Ajay A, Du Y, Gupta A, et al. Is conditional generative modeling all you need for decision making? In: Proceedings of International Conference on Learning Representations, 2023

235 Xian Z, Gkanatsios N, Gervet T, et al. ChainedDiffuser: unifying trajectory diffusion and keypose prediction for robotic manipulation. In: Proceedings of the 7th Annual Conference on Robot Learning, 2023

236 Reuss M, Li M, Jia X, et al. Goal-conditioned imitation learning using score-based diffusion policies. arXiv, 2023. arXiv:2304.02532

237 Chen H, Lu C, Ying C, et al. Offline reinforcement learning via high-fidelity generative behavior modeling. In: Proceedings of International Conference on Learning Representations, 2023

238 Kang B, Ma X, Du C, et al. Efficient diffusion policies for offline reinforcement learning. In: Neural Information Processing Systems, 2023

239 Hansen-Estruch P, Kostrikov I, Janner M, et al. IDQL: implicit q-learning as an actor-critic method with diffusion policies. arXiv, 2023. arXiv:2304.10573

240 Zhu Z, Zhao H, He H, et al. Diffusion models for reinforcement learning: a survey. arXiv, 2023. arXiv:2311.01223

241 Huang W, Wang C, Zhang R, et al. Voxposer: composable 3D value maps for robotic manipulation with language models. In: Proceedings of Annual Conference on Robot Learning, 2023

242 Ma Y J, Liang W, Wang G, et al. Eureka: human-level reward design via coding large language models. In: Proceedings of International Conference on Learning Representations, 2024

243 Yu W, Gileadi N, Fu C, et al. Language to rewards for robotic skill synthesis. In: Proceedings of Annual Conference on Robot Learning, 2023

244 Escontrela A, Adeniji A, Yan W, et al. Video prediction models as rewards for reinforcement learning. In: Neural Information Processing Systems, 2023

245 Huang T, Jiang G, Ze Y, et al. Diffusion reward: learning rewards via conditional video diffusion. arXiv, 2023. arXiv:2312.14134

246 Ma Y J, Kumar V, Zhang A, et al. LIV: language-image representations and rewards for robotic control. In: Proceedings of International Conference on Machine Learning, 2023. 23301–23320

247 Rocamonde J, Montesinos V, Nava E, et al. Vision-language models are zero-shot reward models for reinforcement learning. In: Proceedings of NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023

248 Adeniji A, Xie A, Sferrazza C, et al. Language reward modulation for pretraining reinforcement learning. arXiv, 2023. arXiv:2308.12270

249 Christiano P F, Leike J, Brown T, et al. Deep reinforcement learning from human preferences. In: Neural Information Processing Systems, 2017

250 Wirth C, Akrour R, Neumann G, et al. A survey of preference-based reinforcement learning methods. J Mach Learn Res, 2017, 18: 1–46

251 Bradley R A, Terry M E. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 1952, 39: 324–345

252 Shin D, Dragan A D, Brown D S. Benchmarks and algorithms for offline preference-based reward learning. Trans Mach Learn Res, 2023

253 Lee K, Smith L M, Abbeel P. PEBBLE: feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In: Proceedings of International Conference on Machine Learning, 2021. 6152–6163

254 Park J, Seo Y, Shin J, et al. SURF: semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2022

255 Liang X, Shu K, Lee K, et al. Reward uncertainty for exploration in preference-based reinforcement learning. In:

Proceedings of International Conference on Learning Representations, 2022
256 Kim C, Park J, Shin J, et al. Preference transformer: modeling human preferences using transformers for RL. In: Proceedings of International Conference on Learning Representations, 2023
257 An G, Lee J, Zuo X, et al. Direct preference-based policy optimization without reward modeling. In: Neural Information Processing Systems, 2024
258 Hejna J, Rafailov R, Sikchi H, et al. Contrastive preference learning: learning from human feedback without reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2024
259 Kang Y, Shi D, Liu J, et al. Beyond reward: offline preference-guided policy optimization. arXiv, 2023. arXiv:2305.16217
260 Yuan Y, HAO J Y, Ma Y, et al. Uni-RLHF: universal platform and benchmark suite for reinforcement learning with diverse human feedback. In: Proceedings of International Conference on Learning Representations, 2024
261 Dai J, Pan X, Sun R, et al. Safe RLHF: safe reinforcement learning from human feedback. In: Proceedings of International Conference on Learning Representations, 2024
262 Ha D, Schmidhuber J. World models. arXiv, 2018. arXiv:1803.10122
263 Hafner D, Pasukonis J, Ba J, et al. Mastering diverse domains through world models. arXiv, 2023. arXiv:2301.04104
264 Hansen N, Su H, Wang X. TD-MPC2: scalable, robust world models for continuous control. In: Proceedings of International Conference on Learning Representations, 2024
265 Feng Y, Hansen N, Xiong Z, et al. Finetuning offline world models in the real world. In: Proceedings of the 7th Annual Conference on Robot Learning, 2023
266 Seo Y, Kim J, James S, et al. Multi-view masked world models for visual robotic manipulation. In: Proceedings of International Conference on Machine Learning, 2023
267 Wu J, Ma H, Deng C, et al. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. In: Neural Information Processing Systems, 2023
268 Janner M, Li Q, Levine S. Offline reinforcement learning as one big sequence modeling problem. In: Neural Information Processing Systems, 2021
269 Chen C, Yoon J, Wu Y-F, et al. Transdreamer: reinforcement learning with transformer world models. In: Proceedings of Deep RL Workshop NeurIPS, 2021
270 Robine J, Hoftmann M, Uelwer T, et al. Transformer-based world models are happy with 100k interactions. In: Proceedings of International Conference on Learning Representations, 2023
271 Cohen L, Wang K, Kang B, et al. Improving token-based world models with parallel observation prediction. arXiv, 2024. arXiv:2402.05643
272 Zhang W, Wang G, Sun J, et al. Storm: efficient stochastic transformer based world models for reinforcement learning. In: Neural Information Processing Systems, 2023
273 Zhang Y, Bai C J, Zhao B, et al. Decentralized transformers with centralized aggregation are sample-efficient multi-agent world models. arXiv, 2024. arXiv:2406.15836
274 Wu J, Yin S, Feng N, et al. iVideoGPT: interactive videogpts are scalable world models. arXiv, 2024. arXiv:2405.15223
275 Bruce J, Dennis M D, Edwards A, et al. Genie: generative interactive environments. In: Proceedings of International Conference on Machine Learning, 2024
276 Lu C, Ball P, Teh Y W, et al. Synthetic experience replay. In: Neural Information Processing Systems, 2024
277 Du Y, Yang S, Dai B, et al. Learning universal policies via text-guided video generation. In: Neural Information Processing Systems, 2023
278 Yang M, Du Y, Ghasemipour K, et al. Learning interactive real-world simulators. arXiv, 2023. arXiv:2310.06114
279 Zhou S, Du Y, Chen J, et al. Robodreamer: learning compositional world models for robot imagination. In: Proceedings of International Conference on Machine Learning, 2024
280 He H R, Bai C J, Pan L, et al. Large-scale actionless video pre-training via discrete diffusion for efficient policy learning. arXiv, 2024. arXiv:2402.14407

281 Xiang J, Liu G, Gu Y, et al. Pandora: towards general world model with natural language actions and video states. arXiv, 2024. arXiv:2406.09455

282 Wang L, Ling Y, Yuan Z, et al. Gensim: generating robotic simulation tasks via large language models. In: Proceedings of International Conference on Learning Representations, 2024

283 Wang Y, Xian Z, Chen F, et al. RoboGen: towards unleashing infinite data for automated robot learning via generative simulation. arXiv, 2023. arXiv:2311.01455

284 Li X, Zhang M, Geng Y, et al. ManiPLLM: Embodied multimodal large language model for object-centric robotic manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 18061–18070

285 Shentu Y, Wu P, Rajeswaran A, et al. From LLMs to actions: latent codes as bridges in hierarchical robot control. arXiv, 2024. arXiv:2405.04798

286 Zhai Y, Bai H, Lin Z, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. arXiv, 2024. arXiv:2405.10292

287 Ji J, Qiu T, Chen B, et al. AI alignment: a comprehensive survey. arXiv, 2023. arXiv:2310.19852

288 Ji J, Liu M, Dai J, et al. Beavertails: towards improved safety alignment of LLM via a human-preference dataset. In: Neural Information Processing Systems, 2023

289 Li X L. Positive-incentive noise. IEEE Trans Neural Network Learn Syst, 2022

290 Chen X, Hu J, Jin C, et al. Understanding domain randomization for sim-to-real transfer. In: Proceedings of International Conference on Learning Representations, 2022

291 Xu K, Bai C J, Ma X T, et al. Cross-domain policy adaptation via value-guided data filtering. In: Neural Information Processing Systems, 2024

292 Zhao W, Queralta J P, Westerlund T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In: Proceedings of 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020. 737–744

293 Zhang H, Du W, Shan J, et al. Building cooperative embodied agents modularly with large language models. In: Proceedings of NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023

294 Chen Y, Arkin J, Zhang Y, et al. Scalable multi-robot collaboration with large language models: centralized or decentralized systems? arXiv, 2023. arXiv:2309.15943

295 Kuba J G, Chen R, Wen M, et al. Trust region policy optimisation in multi-agent reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2022

296 Hao J Y, Shao K, Li K, et al. Research and applications of game intelligence. Sci Sin Inform, 2023, 53: 1892–1923 [郝建业, 邵坤, 李凯, 等. 博弈智能的研究与应用. 中国科学: 信息科学, 2023, 53: 1892–1923]

297 Gao Y L, Zhang R, Li X L. Artificial intelligence ethical computation. Sci Sin Inform, 2024, 54: 1646–1676 [高漪澜, 张睿, 李学龙. 人工智能伦理计算. 中国科学: 信息科学, 2024, 54: 1646–1676]

298 Sun M J, Liu Z, Bair A, et al. A simple and effective pruning approach for large language models. arXiv, 2023. arXiv:2306.11695

299 Liu Z, Oguz B, Zhao C, et al. LLM-QAT: data-free quantization aware training for large language models. arXiv, 2023. arXiv:2305.17888

300 Xu Y, Xie L, Gu X, et al. QA-LORA: quantization-aware low-rank adaptation of large language models. In: Proceedings of International Conference on Learning Representations, 2024

Embodied-AI with large models: research and challenges

Chenjia BAI[1,2], Huazhe XU[3] & Xuelong LI[2,1*]

1. *Institute of Artificial Intelligence (TeleAI), China Telecom Corp. Ltd., Shanghai* 200232*, China*;
2. *Institute of Artificial Intelligence (TeleAI), China Telecom Corp. Ltd., Beijing* 100033*, China*;
3. *Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing* 100084*, China*
* Corresponding author. E-mail: xuelong_li@ieee.org

Abstract Embodied artificial intelligence (AI) driven by large-scale models is a cross-disciplinary field covering AI, robotics, and cognitive science, focusing on how to combine the perception, reasoning, and logical thinking abilities of large-scale models with embodied AI to improve the data efficiency and generalization ability of existing embodied AI frameworks such as imitation learning, reinforcement learning, and model predictive control. In recent years, with the continuous improvement of the capabilities of large-scale models and the continuous improvement of expert datasets, simulation platforms, and task sets in embodied robots, the combination of large-scale models and embodied AI will become the next wave of AI and is expected to become an important breakthrough for AI to move towards physical robots. This article focuses on the research field of embodied AI driven by large-scale foundation models (LFM), conducting systematic research, analysis, and prospects. Firstly, we review the relevant technical backgrounds of large models and embodied intelligence, as well as the existing learning frameworks of embodied intelligence. Secondly, according to how large models empower embodied intelligence, we divide the existing research into five paradigms: LFM-driven environmental perception, LFM-driven task planning, LFM-driven basic strategy, LFM-driven reward function, and LFM-driven data generation. Finally, we summarize the challenges in existing research, look forward to feasible technical routes, provide references for researchers, and further promote the national AI development strategy.

Keywords embodied AI, large-scale models, environment perception, task planning, foundation policy