基于视觉语言模型的具身智能机器人操作研究

陆　丽
广西产研院人工智能与大数据应用研究所有限公司　广西　南宁　530000

摘　要：本文综述了视觉语言模型（VLMs）在具身智能机器人操作中的关键作用，重点阐述了三类核心研究方法：一是基于VLM的端到端策略学习，直接映射多模态输入到动作输出；二是分层任务分解与符号-神经混合推理，将高层语义指令逐步转化为可执行子任务；三是基于提示工程与上下文学习，利用VLM的ICL能力，以提示引导避免参数微调；四是人机协同交互范式，依托VLM实现自然语言引导下的实时干预与反馈。研究表明，VLMs显著提升了机器人在零样本泛化、跨任务迁移和语义理解方面的能力，然而，当前研究仍面临具身对齐偏差、动作空间离散化、实时性限制及安全伦理等挑战[1]。

关键词：具身智能；视觉语言模型；机器人操作；多模态学习；任务规划

引言

传统机器人依赖模块化架构（如感知–规划–控制分离），虽在封闭场景有效，但泛化性差、开发成本高且误差累积严重。与此不同，具身智能强调智能源于智能体与环境的持续交互，其核心特征是感知、认知与行动的闭环耦合，要求机器人具备情境理解、意图推理与自适应执行能力。基于VLM的具身系统通常包含三大核心技术模块：多模态环境感知、语义驱动的任务规划以及具身动作执行。本文系统综述VLM赋能具身智能机器人的主流研究范式与代表性工作、分析其核心挑战，旨在厘清该方向的发展脉络，为构建通用、灵活、可交互的服务机器人提供理论支撑与实践指引。

1 机器人操作的传统范式探讨

传统机器人操作主要分为三类范式：（1）示教再现：通过人工引导或离线编程预定义运动轨迹，虽在高度结构化的工业场景中具备稳定性，但缺乏适应任务变化的灵活性。（2）基于模型的规划：通过构建环境与对象的精确动力学或运动学模型，通过搜索算法（如RRT*）规划动作序列，然而该范式对建模精度要求高，难以应对建模误差与环境的不确定性。（3）端到端模仿学习：旨在建立“状态—动作”的端到端映射，直接从专家演示中学习，但是受限于高质量数据集采集成本及分布偏移问题，泛化能力不足。

2 视觉语言模型在机器人操作中的核心作用机制

VLMs的引入，从根本上改变了机器人操作的范式，其核心机制在于打破传统“感知—规划—控制”分治架构的刚性壁垒。这不仅赋予了机器人解析复杂场景语义与抽象意图的能力，更关键的是建立了互联网大规模先验知识向具身动作策略迁移的通道，从而克服了传统机器人严重依赖预定义规则、难以应对环境变化的局限性。

2.1 输入模块

输入模块承担着采集与预处理多模态原始数据的关键任务，旨在为后续决策提供高保真的视觉与语言信息。在视觉层面，该模块整合RGB摄像头、深度相机及3D点云传感器等多源异构数据，并通过去噪、畸变校正及目标检测初筛等预处理手段，构建高精度的环境表征；在语言层面，则利用先进的语音识别技术（如Whisper、Wav2Vec）将非结构化的语音信号转化为文本，进而通过意图识别与指代消解（如结合视觉上下文解析“左边”等空间指代）完成指令的语义解析

2.2 处理模块

处理模块作为系统的认知核心，主要承担多模态信息的深层融合与语义解析任务。该模块依托基于Transformer架构的视觉语言模型（如Flamingo、PaLI、RT-2等），利用对比学习或跨注意力机制将视觉特征与语言词元进行细粒度对齐，从而构建统一的语义表征空间。在此基础上，模块不仅具备零样本或少样本推理能力，能够泛化理解未见过的物体组合（如“带条纹的马克杯”），还能结合CLIP等模型实现高精度的空间语义解析，将“左/右/上/下”等抽象方位词映射至具体的视觉区域。这种深度融合机制赋予了机器人强大的上下文理解能力，使其能够在杂乱桌面场景中精准定位“最右边那个没盖盖子的杯子”，或依据常识推理解析模糊指令“拿一个看起来能装水的容器”，通过视觉属性匹配有效解决了开放环境下语义歧义与目标识别的难题。

2.3 决策模块

决策模块作为连接高层语义认知与底层物理执行的

桥梁，负责将抽象的任务目标转化为可执行的运动规划。该模块首先将VLM输出的语义目标（如“抓取红色圆柱体”）解析为具体的动作基元，并结合6D位姿估计技术（如PVN3D、或基于Perceiver IO等架构构建的位姿估计模型）确定操作目标的精确空间坐标。针对复杂任务，系统引入任务分解机制，利用PDDL或基于LLM的规划框架（如SayCan）将长程指令拆解为有序的子任务序列；随后，集成RRT*、CHOMP或MoveIt!等运动规划算法，在满足避障约束与物理规则的前提下生成最优路径。

2.4 执行模块

执行模块作为物理实现的终端，负责将决策层生成的动作序列转化为精准、稳定的机械运动。该模块基于ROS（Robot Operating System）架构，通过底层节点驱动机械臂（如UR5、Franka Emika）、自适应夹爪及移动底盘等硬件单元，并引入力控与柔顺控制算法，以确保在插拔USB、拧瓶盖等精细操作中实现力矩的平滑控制，有效避免刚性碰撞风险。同时，依托EtherCAT、CAN总线等低延迟通信技术以及样条轨迹插值算法，系统有力保障了动作执行的实时性与平滑度。这种高可靠性的执行能力，使得机器人能够在工厂装配线上完成“将蓝色电阻安装至PCB第3排第2列焊点”的毫米级精密作业，亦能在老年护理场景中轻柔地将水杯递送到用户手中，实现了刚性机械系统对语义指令的柔性响应。

2.5 反馈模块

反馈模块构建了“感知-决策-执行”的闭环控制机制，是保障系统鲁棒性与持续进化能力的关键。该模块整合了视觉重识别（ReID）、物体状态检测及触觉传感反馈，能够实时监测任务执行状态（如物体是否被抓取稳固、是否发生倾倒）。在短期纠错层面，系统一旦检测到执行失败（如抓空或误抓），即刻触发重规划流程，并能响应人机交互中的自然语言纠正指令（如“不是这个，是旁边那个！”）动态更新指代锚点；在长期优化层面，通过强化学习（如PPO、SAC）或模仿学习对底层动作策略进行微调，并结合经验回放机制优化决策模块中的任务调度逻辑。这种分层闭环学习机制使得机器人不仅能从“误抓白色杯子而非透明玻璃杯”的错误中逐步提升判别与操作精度，更能在整理书架等持续任务中习得用户的个性化偏好（如“科幻小说放最上层”），从而实现从单一任务执行向自适应智能体的演进[2]。

3 主流研究范式与代表性工作

目前，基于VLM的具身智能研究主要沿着以下几条技术路线展开，如下表1所示：

表1：基于VLM的具身智能主流研究范式对比

范式	核心特点	优势	局限	代表工作
端到端策略学习	输入直接映射到动作	高效、端到端优化	数据需求大、难解释	RT-2, PaLM-E
分层任务分解	高层规划 + 底层执行	可解释、模块化、数据高效	依赖技能库	SayCan, Inner Monologue
提示与上下文学习	通过Prompt引导行为，免微调	快速适应、低开销	稳定性弱、依赖提示设计	RT-2 (ICL), Prompt-based VLMs
人机协同学习	人类参与交互与修正	支持模糊指令、持续进化	需用户介入	Interactive Language, PaLM-E (HITL)

3.1 端到端VLM策略学习

该范式构建一个统一的神经网络架构，将原始多模态输入（如RGB图像、深度图、语言指令）直接映射为低层机器人动作（如关节角速度、末端执行器位姿）。该范式的核心逻辑在于将机器人动作空间为离散化并嵌入VLM的词表空间，使得动作生成退化为序列预测问题。其典型技术路径依托大规模预训练模型的语义先验，通过在包含真实或仿真交互数据的“观测-动作”对上进行协同微调，实现了“感知-决策-执行”的一体化映射。

RT-2（Robotic Transformer 2）是该范式的典型代表，其创新性地将机器人动作向量量化为类似文本的token序列，使VLM能够以预测下一个单词的方式预测下一个动作，从而将互联网规模的语义知识有效迁移至具身操作任务中。尽管该架构具有极高的系统简洁性，并能有效捕捉多模态输入与动作输出间的复杂非线性映射，但其对海量高质量机器人交互数据的依赖、黑盒模型固有的可解释性缺失以及在安全关键场景下的验证难题，仍是当前制约其广泛落地的关键瓶颈。

3.2 分层任务分解与推理

该范式借鉴经典人工智能的分层思想，构建了“高层语义规划—底层技能执行”的解耦架构，旨在解决端到端模型在长程任务中推理能力不足的问题。在此架构中，VLM扮演认知引擎的角色，负责宏观层面的任务理解、环境解析与逻辑推演，将复杂的自然语言指令递归分解为原子级技能序列；底层控制器则精准运动控制

与执行。其典型工作流程涵盖任务解析与意图理解、可行技能检索、价值函数排序以及基于状态反馈的迭代规划。代表性工作如：SayCan利用VLM评估技能可行性与任务相关性，结合LLM先验知识实现零样本任务泛化；Inner Monologue则引Inner Monologue 引入内部推理链（chain-of-thought），使系统能在失败后自主反思并调整策略。该范式的核心优势在于架构的可解释性强、模块化设计便于复用既有机器人技能库，且显著降低了对昂贵机器人交互数据的依赖。

3.3 基于提示工程与上下文学习

该范式充分利用VLM的上下文学习（In-Context Learning, ICL）能力，通过精心构造的提示（Prompt）引导模型行为，避免参数微调。技术路径包括：（1）构建任务演示库（指令–观察–动作三元组）；（2）设计结构化提示模板（如Few-shot、Chain-of-Thought）；（3）动态注入机器人状态信息（如关节角度、接触力）以增强具身感知。代表性工作如：将机器人本体感知信号编码为视觉特征，无缝嵌入VLM输入空间，提升对物理交互的理解。

3.4 人机协同与交互式学习

该范式将人类视为智能系统的一部分，通过自然语言实现双向交互。技术流程包括：（1）模糊指令澄清；（2）执行过程中的实时反馈接收；（3）基于人类修正的在线策略更新；（4）长期记忆积累形成个性化协作模式。PaLM-E in Human-in-the-loop 实现了人在回路中的持续学习，显著提升长周期任务成功率。3.5范式融合：协同演进的新趋势

近期研究多采用多范式融合策略，兼顾效率、鲁棒性与适应性：（1）RT-2 + SayCan 混合架构：Google团队将 RT-2 端到端动作生成能力与 SayCan 分层规划机制结合，保持高执行效率，引入可干预推理层，提升任务成功率与安全性。（2）上下文学习 + 人机交互：如 CoPa 系统，用 Few-shot 提示初始化策略，执行中主动向用户提问优化后续动作，形成“提示驱动 + 交互修正”闭环学习。（3）分层规划 + 端到端微调：部分研究高层用 VLM 分解任务，底层用微调后的端到端策略执行技能，兼顾语义理解与精细控制。

4 面临的挑战与问题

当前基于视觉语言模型（VLM）的具身智能面临多维度挑战，主要体现在四个层面：（1）模型层：VLM在静态互联网图文数据预训练，缺乏具身体验，存在“具身对齐偏差”；其离散token架构难以自然表达机器人连续高维动作空间，离散化策略易在精度与复杂度间失衡。（2）数据层：高质量具身交互数据稀缺，构建需同步记录多模态信息，精细标注成本高，严重制约模型泛化能力。（3）工程层：大模型推理延迟高，难以满足实时控制需求；系统适配异构机器人硬件难，在非结构化环境中保持鲁棒性对部署是严峻考验。（4）安全层：VLM可能因指令误解等产生危险行为，需建立可验证的决策护栏、伦理约束机制与人机协同监督框架，确保行为可靠可控[3]。

5 结语

视觉语言模型（VLM）给具身智能机器人操作领域带来新活力，突破了传统机器人系统在语义理解和泛化能力上的局限，让机器人能以更自然的方式与人类和环境交互。以VLM为“认知引擎”，研究者正打造能理解复杂指令、适应开放环境、与人类协同的下一代智能机器人。未来研究可聚焦这些方向：一是具身预训练，构建大规模机器人交互数据集，让VLM在真实或仿真环境预训练以解决具身对齐偏差；二是混合架构与神经符号系统，融合VLM统计学习和符号AI逻辑推理优势；三是高效具身表示学习，设计具有物理意义的紧凑表示并探索新架构；四是主动感知与探索，使机器人能主动获取信息；五是个性化与长期学习，让机器人学习用户偏好，构建终身学习框架。

参考文献

[1]杜国锋,邵士博,李尚霖,等.融合视觉语言模型与近端策略优化算法的人形机器人步态切换方法[J].机械工程学报,2025,61(21):204-212.

[2]刘政鑫.跨维智能：“AI定义机器人”的具身实践[J].机器人产业,2026,(01):67-71.

[3]孔祥鑫.基于轻量化网络设计的具身智能复合机器人控制与AI视觉系统研发[J].科学技术创新,2025,(18):203-206.