面向通用场景的基于视觉-语言-动作模型的具身移动操作方法研究

(清华大学攻读硕士学位研究生论文选题报告)

培 养 单 位： 深圳国际研究生院

项 目 名 称： 电子与通信工程

申 请 人： 刘 昱 杉

学 号： 2024214103

指 导 教 师： 张 晓 平 教 授

二〇二五年十一月

目 录

第 1 章　课题背景与意义

1.1　研究背景

近年来，随着人工智能基础模型（Foundation Models）和多模态大模型（Multi-modal Large Models）的快速发展，具身智能（Embodied Intelligence）逐渐被视为迈向通用人工智能的重要路径之一，其核心目标是在统一的“感知—理解—决策—执行”闭环中，使智能体能够在真实或高保真拟真环境中通过与物理世界的持续交互获取知识、完成复杂任务 [1-2]。在这一大背景下，具身移动操作（Embodied Mobile Manipulation）成为连接认知智能与物理执行能力的关键形态之一：机器人不仅需要具备在三维空间中自主移动的能力，还需在语义丰富、结构复杂的环境中完成抓取、放置、开关、插拔等多样化操作，从而实现从“可导航”到“可服务”的跨越 [3]。与传统仅关注机械臂操作或仅关注移动平台导航的研究不同，具身移动操作强调在统一任务视角下对移动与操作进行协同建模和联合优化，即在语言或任务指令驱动下，依据环境语义、物体状态和任务约束进行一体化的运动规划和执行。

近年来，基于通用模型 (Foundation Model) 驱动的具身移动和操作分别取得了有效的进展。在移动方面，视觉-语言-导航 (Visual-Language-Navigation, VLN) 在从语言到可达目标的路径规划方面实现了丰富的积累，在仿真和真机中呈现出令人印象深刻的效果；同时，无论是基于模仿学习 (Imitation Learning, IL) 还是离线强化学习 (Offline Reinforcement Learning, RL) 的操作方法，在对象识别、姿态估计、抓取规划等方面都取得了显著的进展，并在真实机器人平台上得到了验证。然而，在移动操作任务中，移动和操作之间存在不可忽视的强耦合：移动阶段所形成的视角、距离与遮挡关系直接决定了后续操作的观测质量与可达性，而操作对目标物体、姿态与约束的先验又会反过来影响移动的策略选择与路径代价。因此，仅在导航或操作单一子任务上取得进展，并不足以保证端到端任务的成功率、稳定性与泛化能力。与此同时，以视觉-语言-动作（Visual-Language-Action, VLA）为代表的大模型在指令理解、情景推理与跨任务迁移上展现出统一表达与泛化能力，通过将开放式语言意图与感知结果对齐为可执行中间表示并生成带约束的分层技能序列，从而支撑“从语言到动作”的一体化规划与在动态噪声下的鲁棒重规划，为“从语言到可执行子目标与技能序列”的一体化规划提供了新的可行性。因此，如何将大模型的语义优势落地为低层可控、可验证的移动与操作策略，并在扰动与不确定条件下保持鲁棒，是当前研究的重要方向。

从国家战略和社会发展需求的角度，明确提出要推动人工智能与实体经济深度融合，发展面向制造业、服务业、医疗和养老等领域的智能机器人与智能装备。《新一代人工智能发展规划》提出：到 2030 年人工智能理论、技术与应用总体达到世界领先水平，成为世界主要人工智能创新中心，智能经济、智能社会取得明显成效，为跻身创新型国家前列和经济强国奠定重要基础 [4]；《“十四五”机器人产业发展规划》中进一步聚焦重点推进工业机器人、服务机器人、特种机器人重点产品的研制及应用，拓展机器人产品系列，提升性能、质量和安全性，推动产品高端化智能化发展。[5]；《高等学校人工智能创新行动计划》中强调要“瞄准世界科技前沿，不断提高人工智能领域科技创新、人才培养和国际合作交流等能力”[6]。

在上述国际学术发展趋势和国家战略需求的双重驱动下，本文所聚焦的“面向通用场景的基于视觉—语言—动作模型的具身移动操作方法”，可以被视为顺应技术演进与应用需求的一项自然延伸与深化：一方面，前沿综述表明，具身智能、移动操作与基础模型三者的交汇已成为当前机器人学的重要发展方向，有必要在统一视角下系统梳理从感知、表示到决策、控制的完整链条，进而形成面向真实任务的通用方法论框架。另一方面，针对家庭、实验室、仓储等典型室内场景，现有系统在统一建模自然语言指令、环境语义信息与移动/操作决策方面仍存在明显局限，尚缺乏在保障安全性与可控性的前提下，兼顾任务多样性与环境变化的通用技术路线。在此背景下，本课题的研究问题可以被较为清晰地界定为：如何利用视觉—语言—动作模型，在复杂室内环境中统一处理指令理解、环境感知与具身移动操作决策，从而支撑一体化的任务执行流程：既对接具身智能和机器人学习在基础模型时代的发展脉络，又紧扣通用场景下典型具身移动操作任务的客观需求，为后续研究方案与技术路线的设计提供了问题基础与方向锚点。

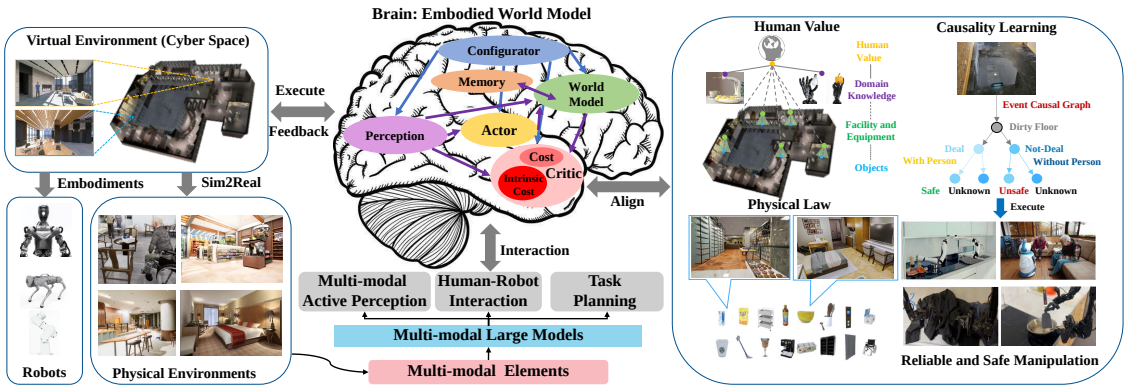

图 1.1　具身智能体整体框架示意图

1.2　研究意义

从学术研究角度看，围绕“面向通用场景的基于视觉-语言-动作模型的具身移动操作方法”开展系统研究，有望在多个层面形成具有前瞻性的理论与方法贡献。其一，通过引入统一的中间表示，将自然语言指令、环境语义信息与可执行动作在同一模型中进行对齐，使高层语义规划与低层几何/动力学控制之间的映射关系更加明确，为移动与操作的协同规划提供可解释、可分析的语义—几何桥接机制 [7]；其二，在 VLA 模型框架下，将指令理解、情境推理与分层技能序列生成整合为端到端的“语言—感知—动作”链路，有助于深化我们对基础模型在具身场景中“世界知识 + 具身经验”混合表征与推理能力的理解，并为构建具有多任务泛化能力的通用具身智能体提供可行路径 [8-11]；其三，面向典型通用场景，对移动与操作强耦合带来的观测可达性、操作可行性与安全约束等问题进行系统刻画，并在统一评测框架下分析任务成功率、鲁棒性与泛化能力，有望推动移动操作从“单场景工程系统”走向“可比较、可复现的研究基准”，为后续更大规模、更复杂场景下的具身移动操作研究提供方法与系统基础 [3]。

从应用需求角度看，本课题的研究有望为实验室助理机器人、仓储物流机器人以及家居服务机器人等典型应用场景中的具身移动操作任务提供更加通用、可落地的技术方案。通过在统一的视觉-语言-动作模型框架下提升机器人在通用场景中的整体能力，可以减少对环境的精细改造与大量人工规则设计，降低系统部署和维护成本，并提升在任务变化、物体变化和环境部分改造情况下的持续可用性与可扩展性。与此同时，得益于基础模型在多模态表征和知识迁移方面的优势，基于本课题形成的方法体系在扩展到新任务、新物体或新环境时，有望通过少量示例或在线交互快速适配，进一步支撑具身智能系统在真实应用中的长期演进与迭代 [12-15]。

第 2 章　国内外研究现状及分析

2.1　具身导航与通用室内场景理解方法

近年来，面向实际应用的导航技术经历了从传统移动机器人导航到具身导航（Embodied Navigation）的持续演进。传统的移动机器人导航，在建图、定位与路径规划等模块上已经形成了较为成熟的理论和工程体系 [16-17]，并在室内服务机器人、自动导引车（AGV）等场景中得到了广泛应用。然而，传统导航方法主要聚焦于几何可达性和路径最优性，虽然在简单环境下效果显著，但在面对复杂语义场景和动态交互时，往往对环境理解和复杂任务的支持有限，难以直接支撑指令驱动的移动操作。

随着具身智能和大规模多模态模型的快速发展，研究逐渐转向“在未知或半未知环境中一边感知、一边交互、一边优化导航策略”的具身导航新范式。该范式强调通过本体传感器主动探索环境，在线构建语义地图与对象级表示，从而使得机器人能够在动态、复杂的环境中进行更加智能的决策和执行 [18-20]。在感知模态方面，具身视觉导航（Embodied Visual Navigation, E-VN）任务如 PointNav、ImageNav 和 ObjectNav 等，已逐渐成为研究重点 [21-22]。这些任务主要依赖纯视觉输入，旨在提升机器人在无先验地图条件下的导航能力。同时，视觉-语言导航（Vision-and-Language Navigation, VLN）的出现进一步拓展了具身导航的任务范围，使得机器人能够在仿真和真实环境中根据自然语言描述进行跨房间导航和目标搜索 [7, 23]，并依托高保真平台如 Habitat 开展了广泛的系统评测 [24]，显著提升了从“到坐标”到“到物体/到语义位置”的导航能力。与此同时，针对复杂环境中单一传感器易受遮挡和噪声干扰的问题，研究者们提出了多源传感器融合方法，通过将激光雷达、相机、毫米波雷达和惯性测量单元（IMU）等传感器的数据进行联合建模，从而提高导航系统的精度和鲁棒性 [25-26]。这些多模态传感器的融合为具身导航提供了更为强大的环境感知能力，使得机器人能够在复杂、动态环境下实现精确的定位与安全的导航。

在通用室内场景理解方面，研究重点逐步从单帧 2D 语义分割扩展到面向机器人应用的三维语义建图与场景级理解。早期的工作如 SemanticFusion 将卷积神经网络（CNN）用于语义预测，并结合稠密 RGB-D SLAM 实现了实时语义标注，进而在三维网格上实现了在线语义标注 [27]；Panoptic Fusion 等系统进一步在体素地图中统一建模“背景类别”（stuff）与“前景物体”（things），支持大规模室内环境的全景语义映射与网格导出 [28]。与此同时，随着大规模 RGB-D 数据集的出

现，如 ScanNet 和 Matterport3D 等提供了丰富的三维重建与语义标注数据，为三维场景的语义分割、实例分割和场景补全等任务奠定了基准 [29-30]。近年来的研究开始关注开放词汇与高层关系建模：OpenScene、UniM-OV3D 等方法通过图文预训练模型与三维特征的对齐，实现了对任意文本类别的开放词汇三维语义理解 [31-32]；Hierarchical Open-Vocabulary 3D Scene Graphs 等工作则进一步构建了开放词汇三维场景图，将物体、功能区域及其语义关系编码为结构化表示，为语言约束的导航与操作提供了可查询的场景先验 [33]。

在产业界，以自动驾驶为代表的车企和出行公司在大规模道路数据和强算力平台的支撑下，已经形成了两条具有代表性的落地路线。一类是以多传感器冗余为特点（如摄像头 + 激光雷达 + 高精度地图）的模块化感知-规划-控制栈，强调安全冗余与可解释性；另一类则以特斯拉 FSD v12 及后续版本为代表，采用“从像素到控制”的端到端视频 Transformer，仅依赖多路摄像头进行感知与决策，并已在真实道路上大规模部署，依靠在线数据闭环持续迭代 [34-35]。

从传统导航到具身导航、从几何建图到语义/全景语义映射、从封闭类别到开放词汇三维场景图，现有研究为具身移动操作提供了坚实的技术基础。然而，这些工作大多将导航与场景理解作为独立的感知与决策模块，评价指标也主要聚焦于到达率、定位精度或语义分割精度等，对于“导航位置与视角选择如何影响后续操作的可达性与安全性”以及“如何在统一的视觉-语言-动作表示下，将通用室内场景理解直接服务于移动操作决策”的系统研究仍然存在不足，这为后续基于 VLA 大模型构建一体化具身移动操作方法留下研究空间。

2.2　基于学习的具身操作与移动操作控制方法

在具身操作（Embodied Manipulation）领域，近年来的控制方法大致经历了从“感知-规划-控制”模块化管线向端到端学习范式的转变，研究围绕模仿学习（Imitation Learning, IL）、强化学习（Reinforcement Learning, RL）及其融合展开了大量系统性研究 [36-37]。早期的代表性工作通过引入导引策略搜索（Guided Policy Search）等方法，直接从相机图像端到端学习将视觉映射到关节力矩或末端位姿的深度视觉运动策略，显著简化了传统操作系统中手工设计特征与控制器的流程 [38]；这些方法通过强化学习对系统进行训练，逐步实现了对多样化操作任务的有效支持。在此基础上，基于大规模深度强化学习框架（如 QT-Opt），利用数十万次真实抓取试验离线估计 Q 函数，在单摄像头输入下实现了对未知物体的高成功率闭环抓取，并自动涌现了试探、重抓、物体调整等复杂行为 [39]。此外，模仿学习方向逐渐从传统行为克隆（Behavior Cloning, BC）扩展到语言条件与多任务设定，如

BC-Z 通过统一处理多任务、多模态演示，在同一策略中实现了零样本任务泛化与语言条件操作，从而在一定程度上缓解了“单任务单策略”的局限性 [40]。为了解决在复杂任务中的稳定性和多模态动作分布问题，扩散策略（Diffusion Policy）应运而生，它提出将机器人视觉运动策略表示为条件去噪扩散过程，从而使得机器人能够在复杂环境下生成更加稳定且具有鲁棒性的动作序列。该方法在多项基准任务上相较于传统的 IL/RL 方法取得了显著性能提升 [41]。进一步的，3D Diffusion Policy（DP3）通过将稀疏点云编码为紧凑的三维表示，显著提升了在空间、视角和外观变化下的泛化能力 [42]，为复杂场景下的操作任务提供了强有力的支持。

在移动操作（Mobile Manipulation）控制方法方面，传统系统普遍采用底盘导航与机械臂操作解耦的架构：导航模块规划机器人基座的移动轨迹，在到达目标附近后再调用抓取或操作控制器完成局部交互。这种分而治之的设计便于工程集成与安全验证，但在需要频繁“边走边动”、对位姿耦合要求高的任务（如推门、从狭窄空间中取物）中，容易出现协同不充分的问题。为了解决这一问题，近年来一批端到端或分层一体化的移动操作控制方法开始涌现。例如，Skill Transformer 在统一的 Transformer 框架下同时预测高层技能（如导航、抓取、放置）和全身低层动作，实现了在长时序重排任务中的移动与操作联合建模，相比传统的分阶段方法显著减少了中间切换带来的误差累积 [43]；Deep Whole-Body Control 通过深度强化学习统一优化导航与操作，使机器人能够在仅使用 RGB 视觉的条件下，在仿真和真实环境中完成多种日常移动操作任务，验证了端到端策略在复杂家庭环境中的可行性 [44]。在数据驱动的移动操作模仿学习方面，Mobile ALOHA 提出了低成本全身远程示教平台，通过采集包含底盘、双臂和手部在内的整机演示轨迹，并与已有桌面操作数据联合训练行为克隆策略，在开关柜门、使用电梯、烹饪等长时序移动操作任务上取得了显著性能提升 [45]。这些工作通过利用多源数据和深度学习技术，有效解决了长时序操作任务中的时序依赖和控制精度问题。

目前，具身操作与移动操作方法在端到端 IL/RL、扩散策略到技能级分层控制等方面都取得了丰富进展，为机器人在复杂环境中完成操作任务提供了多种可选的技术路线。然而，现有研究多聚焦于单机位桌面场景或特定环境下的任务集，尚未充分考虑如何将这些方法应用于“通用室内场景中基于语言指令的具身移动操作”这一目标。具体来说，跨场景泛化、移动—操作协同决策以及与高层语义（如视觉—语言—动作模型）紧密耦合等方面仍然存在较大的提升空间。

2.3　大模型驱动的具身移动操作策略

近年来，大模型驱动的具身策略逐渐成为连接高层语义理解与低层控制的关键技术方向。随着基础模型（Foundation Models）在自然语言处理和计算机视觉等领域的突破，其在具身智能中的应用，特别是在移动操作和任务执行中的潜力，也得到了广泛关注。其技术路径大致可分为以下几类：“LLM+ 技能库规划”、“端到端视觉-语言-动作（VLA）策略”和“跨形态通用策略与系统落地”。

在“LLM+ 技能库规划”方向，以 PaLM-SayCan 为代表的工作，首次将大型语言模型（LLM）与价值函数或技能库相结合，通过语言模型进行任务分解与技能排序，再由预定义的导航与操作控制器执行，实现了从自然语言到多步机器人操作的闭环 [8]。这一方法不仅证明了语言模型的强大能力可以显著提升规划正确率和任务成功率，同时也为具身智能任务的语义推理和任务规划提供了新的思路。Code as Policies 等方法进一步利用代码生成类 LLM，将自然语言指令直接翻译为可执行的策略代码，程序结构可以直接表达感知处理与控制 API 的组合，显著提升了大模型参与具身规划的灵活性和可解释性 [46]。在端到端 VLA（视觉-语言-动作）方向，RT-1 首次系统性地将 Transformer 架构应用于从多任务真实机器人轨迹中学习“图像/文本到离散动作 token”的策略，在大规模厨房操作数据上展现了良好的任务泛化能力 [9]；其后提出的 RT-2 将互联网规模的视觉-语言预训练与机器人数据结合，将视觉-语言模型（VLM）扩展为 VLA，使单一模型既具备 web 语义知识又能输出可执行的机器人动作，在零样本泛化与语义推理方面取得了显著提升 [15]。这一研究表明，基于大模型的具身操作方法能够在更广泛的任务范围内实现较好的推理能力和泛化能力，尤其是在动态环境下的适应性显著增强。Open X-Embodiment 数据集与 RT-X 模型进一步整合了来自 22 种机器人形态、百万余真实轨迹的数据，表明在大规模跨平台数据上训练的高容量模型可以在多种机器人平台之间实现正迁移 [47]。这类研究拓展了具身智能领域的跨平台应用能力，使得大模型不仅限于特定硬件平台，也能够在不同平台上进行广泛部署和灵活适配。在开放社区方面，OpenVLA 在 Open X-Embodiment 等数据集上预训练了 7B 级 VLA 模型，提供了统一的“视觉编码器 + 语言主干 + 动作头”架构及可复用代码库，使研究者能够在多种机械臂与任务上进行高效微调与迁移 [11]；此外，Physical Intelligence 提出的 $\pi_{0.5}$ 模型将图像、文本与动作统一为连续流模型，在多机器人、多场景的测试中表现出了较好的泛化能力，成为“机器人基础模型”的代表之一 [48]。尽管在多个领域取得了显著进展，现有的 VLA 大模型在具身操作中的应用仍面临诸多挑战。首先，模型在开放词汇理解和长时序推理方面的能力仍然有限，尤其是需要在长时间跨度的任务中进行推理与决策时，模型的精度和稳定性仍需要进一步提升 [49]；

其次，虽然 VLA 模型在语义推理和任务规划方面表现出色，但在低层精细控制与实际操作过程中如何保持高效性与鲁棒性，仍然是一个重要的研究方向 [50]。

在具身移动操作的应用中，部分工作开始关注如何将固定基座操作 VLA 模型迁移到“可移动底盘 + 机械臂”的场景中。例如，MoManipVLA 提出了在不重新收集大规模移动操作数据的前提下，利用预训练的 VLA 模型预测高泛化能力的末端执行器 waypoint，并通过双层优化联合规划底盘与机械臂轨迹，显著提升了跨任务、跨环境的成功率 [51]。这一方法使得具身操作能够在更加复杂和动态的环境中进行高效执行，且不需要重新收集大量的训练数据。在工业界，Gemini Robotics 1.5 等系统通过引入动作通道和具身推理模块，使机器人能够通过“推理-规划-执行”的流程，从视觉与语言输入生成多平台可迁移的控制命令。此类系统进一步推出了轻量版的本地运行版本，以降低时延与对网络的依赖，从而提升了在现实环境中的应用效率 [52]。这些系统的成功应用证明了大模型驱动的具身策略不仅能提供较高的任务执行能力，还能解决网络依赖和实时性的问题。

大模型驱动的具身策略在统一建模视觉、语言与动作、提升跨任务与跨场景泛化方面已经取得了显著进展，尤其是在大规模数据集和高计算平台的支持下，模型的学习能力和应用范围不断拓展。然而，现有系统仍然多聚焦于固定基座操作或特定场景下的任务集，对“通用室内场景中导航与精细操作强耦合”的具身移动操作任务，尚缺乏能够同时兼顾语义推理、几何可达性、安全约束与实时性的统一方法框架。基于 VLA 的具身移动操作方法有望通过整合视觉感知、语言理解与运动控制，克服当前模型在任务迁移、跨平台部署以及多场景适应等方面的局限，为更复杂的具身智能任务提供强有力的支持。

2.4　开放场景基准与移动操作系统集成框架

在具身智能逐渐走向落地应用的过程中，开放场景基准与系统集成框架发挥了承上启下的关键作用：一方面，系统化的基准任务与评价指标为不同算法提供了可比较的实验环境与统一问题定义；另一方面，面向真实机器人系统的集成框架则将感知、导航与操作串联为可部署的工程流水线，为从仿真到真机的迁移提供了基础支撑。相关研究大致沿着“从桌面单点操作到多任务具身操作基准”、“从局部场景到全屋尺度开放场景移动操作基准”和“从仿真环境到端到端移动操作系统集成”的脉络逐步演进。

在基准任务层面，早期工作主要聚焦于机械臂单点操作与多任务强化学习评测。RLBench 提出包含近百个操作任务的标准基准与仿真环境，覆盖抓取、插拔、开关等多种典型操作，为基于视觉的深度强化学习与模仿学习算法提供了统一评测

平台 [53]。Meta-World 则面向多任务与元强化学习，定义了 50 个 MuJoCo 操作任务，并给出跨任务泛化与快速适应能力的系统评估方案 [54]。在此基础上，ManiSkill2 将任务规模扩展到包含刚体、软体、关节体等多类对象的 20 余种技能族，并显式区分“固定基座操作”和“移动基座操作”，同时提供大规模演示数据与高效视觉强化学习接口，为算法在多模态输入、任务多样性与仿真效率之间的折中提供了优良平台 [55]。这类基准多以桌面操作或局部场景为主，对环境尺度和语义开放性要求相对有限，但在统一任务接口、动作空间和评测指标方面奠定了重要基础。随着语言条件与长时序需求的提升，一批面向“语言–操作”一体化的具身基准开始出现。CALVIN 基准在逼真的模拟场景中定义了多阶段、长时序的语言条件操作任务，强调在视觉遮挡、物体重排等扰动下的策略鲁棒性与重规划能力 [56]；LIBERO 系列基准则通过构造多套家居场景中的语言条件操作任务，系统分析了知识迁移与终身学习在机器人操作中的可行性与挑战，为跨任务迁移和持续学习算法提供了标准化测试平台 [57]。近期的 VLABench 等工作在此基础上进一步扩大任务规模与语言表达范围，强调开放词汇指令、长时序逻辑约束和 VLA 模型评测，为大模型驱动的具身操作策略提供了系统化对比基准 [58]。这一类基准在动作粒度、任务复杂度和语言表达丰富性方面不断提升，但大多仍假设机械臂基座固定，对导航–操作强耦合的移动操作场景覆盖有限。面向更贴近真实住宅与服务场景的需求，近年的一条重要发展路线是构建“全屋尺度、物体可交互、任务多样化”的开放场景具身基准。Habitat 2.0 在 ReplicaCAD 等高保真室内模型基础上，提出了 Home Assistant Benchmark（HAB），定义“收拾房间、整理杂物、餐桌布置”等长时序移动操作任务，并系统比较了分层强化学习与传统 Sense–Plan–Act 管线在此类任务上的性能与泛化能力 [59]。iGibson 2.0 则强调物体中心建模和物理交互逼真度，面向日常家务任务构建了多种移动操作场景，支持机器人在复杂布置环境中进行导航、开关门、搬运等操作，为研究从视觉感知到物理推拉、抓取的闭环策略提供了可扩展环境 [60]。BEHAVIOR 与后续的 BEHAVIOR-1K 进一步从“人类日常活动库”的角度出发，将数以千计的日常活动分解为一系列具身任务模板，并在 OmniGibson 等高保真模拟器中实例化，覆盖从物体收纳、清洁到烹饪等多类家务活动，为人类中心具身智能与移动操作研究提供了前所未有的任务覆盖度 [61]。

针对移动操作这一具体方向，一些基准开始显式将底盘导航与机械臂操作纳入统一任务定义。ManiSkill2 中的 Push Chair、Open Cabinet 等任务将移动基座与操作目标耦合在一起，要求策略在大范围探索中寻找合理操作位姿 [55]；HomeRobot 项目提出的 Open-Vocabulary Mobile Manipulation（OVMM）基准，在仿真与真实

Hello Robot Stretch 平台上统一定义了“语言指令 → 导航 → 感知 → 操作”的端到端移动操作链路，强调开放词汇目标指定与跨环境泛化能力 [62]。LaNMP 等新近基准则从多维度系统刻画移动操作难度，包括场景规模、视角变化、物体遮挡与语义多样性等，并提出细粒度的分阶段成功率与安全性指标，推动了对“导航-操作-语义理解联动关系”的定量研究 [63]。这些工作在不同程度上体现出向“通用、开放场景移动操作”方向演化的趋势，但在大规模数据采集成本、真实场景复杂度与评测标准统一性方面仍面临不小挑战。

在系统集成与工程落地方面，主流开放平台也在逐步形成“仿真-算法-真机”一体的移动操作研究范式。一方面，Habitat-Lab、OmniGibson、ManiSkill2 等框架在仿真端提供了可扩展场景资源、统一任务接口与多种控制模式，使研究者能够在同一平台内比较 IL、RL、VLA 等多种算法，并支持与 ROS/ROS 2 等机器人中间件打通，简化了策略上板流程 [55, 59, 61]。另一方面，Robosuite、Isaac Gym 等面向机器人学习的仿真框架通过模块化任务定义与 GPU 加速物理仿真，显著降低了大规模策略训练与消融实验的门槛，并为后续在 Isaac Sim 等工业级仿真环境及真实机器人上的迁移提供了接口 [64-65]。以 HomeRobot 软件栈为代表的移动操作系统进一步在 LoCoBot、Stretch 等低成本平台上集成导航、抓取与开放词汇感知模块，形成了从仿真到真实家居环境的完整实验流程，为研究者验证语言驱动移动操作策略提供了开源工程基础 [62]。

综合来看，现有开放场景基准与系统集成框架已经在任务多样性、场景真实性和算法可复现性等方面取得了显著进展，为大规模评测具身智能算法和推动移动操作落地提供了重要支撑。然而，从“面向通用室内场景的、基于视觉-语言-动作模型的一体化具身移动操作方法”这一目标出发，仍然存在若干明显空缺：一是多数基准在任务定义上仍偏向固定基座或局部场景，对导航-操作强耦合下的安全视角选择、主动感知与失败恢复等问题覆盖有限；二是现有系统多将大模型作为高层规划或语义标注模块，缺乏在统一 VLA 表示下对导航策略与低层操作控制进行系统评测的基准与软件接口；三是针对开放词汇目标、动态场景扰动与跨场景迁移的联合评测仍处于起步阶段。围绕这些不足，本课题在后续研究中将结合前述开放基准与系统框架，进一步构建适用于通用室内场景的移动操作任务集与评测方案，并在此基础上设计和验证基于 VLA 模型的一体化具身移动操作方法。

2.5　本章小结

本章围绕具身移动操作的相关研究进展，对具身导航与通用室内场景理解、基于学习的具身操作与移动操作控制、大模型驱动的具身移动操作策略以及开放场

景基准与系统集成框架等方向进行了系统梳理。总体来看，现有工作在多个层面取得了显著成果：在感知与环境建模上，从传统几何建图逐步演进到语义映射、全景语义与开放词汇三维场景理解，为机器人在复杂室内环境中的稳健感知与语义理解奠定了坚实基础；在控制方法上，从模块化“感知-规划-控制”到端到端 IL/RL 及扩散策略，再到一体化的全身移动操作控制，显著提升了机器人在长时序、多阶段操作任务中的学习效率与表现能力；在大模型驱动的具身策略方面，以 VLA 为代表的模型初步展示了统一建模视觉、语言与动作、利用大规模跨平台数据实现跨任务与跨形态泛化的潜力；在基准与系统层面，各类开放场景基准与仿真-真机一体化软件栈，推动了具身智能算法的可复现性和系统化评估。

然而，现有研究也表明，距离“面向通用室内场景、基于视觉-语言-动作模型的一体化具身移动操作”仍存在差距。一方面，导航与操作在多数系统中仍然以弱耦合或后期集成的方式存在，对“导航视角选择如何影响操作可达性与安全性”“如何在移动-操作强耦合条件下实现主动感知与失败恢复”等关键问题缺乏系统刻画和统一建模；另一方面，当前 VLA 模型多聚焦于固定基座或局部场景，在长时序推理、低层精细控制、实时性与安全约束等方面，尚未形成适用于移动操作任务的成熟方法框架，与现有开放基准之间也缺少针对性强、接口友好的评测体系。此外，面向开放词汇目标、动态环境扰动与跨场景迁移的联合评估体系仍然不完善，难以全面反映方法在真实通用场景中的有效性与鲁棒性。基于上述成就与不足，可以看出：构建一种在统一视觉-语言-动作表示下，将通用室内场景理解、具身导航与精细操作紧密耦合的移动操作方法，并在开放基准与真实平台上予以验证，既是延展现有研究脉络的必然方向，也是支撑服务机器人、实验室助理机器人和仓储物流机器人等应用落地的现实需求。

第 3 章　课题主要研究内容

3.1　拟解决的关键问题

当前具身移动操作方法在通用场景下面临以下瓶颈：

1. 现有 VLA 模型多以“从图像与文本直接预测动作序列”的方式工作，缺少能够直接反映操作位置与姿态的中间表示。在复杂三维环境中，因缺乏显式可达性与安全约束建模，易出现物理不可达或存在碰撞风险的策略。
2. 多数系统沿用“先导航、后操作”的弱耦合范式。对于视角与姿态高度敏感的任务，这种设计难以同时兼顾视角质量、末端可达域和运动安全性，从而对任务成功率与执行效率形成明显制约。
3. 高质量移动操作数据的采集与标注成本较高，不同场景和机器人差异显著，在新环境中的性能急剧下降。尽管已有跨平台数据集和高保真仿真环境，如何有效利用异构数据与模拟数据，实现对通用室内场景的稳健泛化，缺乏成熟的方法体系。

3.2　拟研究内容

1. 在 VLA 框架下，探索将语言意图、场景语义结构、几何关系与动作可达性等信息统一编码的中间表征形式，形成适配移动操作任务的视觉编码器、语言主干与动作输出层设计，为后续协同决策与安全约束提供可解释的语义–几何桥接层。
2. 构建能够同时处理多模态观测、兼顾全局移动与局部精细操作需求的协同决策框架，在同一策略体系内协调基座位姿选择、视角配置与末端动作规划，从而在复杂室内环境中提升整体任务执行的稳定性与效率。
3. 基于典型通用室内场景，构建覆盖多类语言驱动移动操作任务的统一任务集，设计兼顾任务成功率、路径与操作效率以及跨场景迁移能力的评价指标；在仿真与真实机器人平台上开展系统集成与对比实验，验证所提出 VLA 移动操作方法在多任务、多场景条件下的有效性与适用范围。

第 4 章　引用文献的标注

模板支持 BibTeX 和 BibLaTeX 两种方式处理参考文献。下文主要介绍 BibTeX 配合 natbib 宏包的主要使用方法。

4.1　顺序编码制

在顺序编码制下，默认的 `\cite` 命令同 `\citep` 一样，序号置于方括号中，引文页码会放在括号外。统一处引用的连续序号会自动用短横线连接。

也可以取消上标格式，将数字序号作为文字的一部分。建议全文统一使用相同的格式。

4.2　著者-出版年制

著者-出版年制下的 `\cite` 跟 `\citet` 一样。

第 5 章　预期创新点与研究成果

5.1　预期创新点

1. 提出显式编码语言意图、场景语义结构、几何关系与动作可达性约束的 VLA 中间表征与模型架构，将传统“黑箱式”从图像/文本到动作的映射，提升为可解释的语义–几何桥接层，为安全约束和行为验证提供基础。
2. 在统一策略框架下，将基座位姿选择、视角配置与末端动作规划纳入同一优化过程，形成兼顾全局移动与局部精细操作需求的协同决策方法。
3. 面向通用室内场景构建语言驱动移动操作任务集与评价指标体系，结合异构真实数据和高保真仿真数据，探索数据高成本条件下的 VLA 训练与跨场景迁移策略。

5.2　预期成果

完成基于真实移动操作本体的原型系统集成与示范应用，在实验室场景和企业应用场景中验证所提方法的工程可行性，并在机器人与人工智能等方向的高水平国际会议或期刊上发表若干篇学术论文。

第 6 章　研究计划与进展

6.1　已完成工作概括

前期工作已在具身导航、多模态感知、风险识别、主动视觉精细操作以及轮式双臂移动操作本体等方面形成了较为扎实的技术与系统基础：在面向化学实验室场景，设计并搭建了自主巡检与风险排除机器人系统，实现了从语言/规则指令出发到“巡检-识别-处置”的闭环流程，并以共同第一作者撰写一篇期刊论文在投；面向双臂精细操作场景，搭建了基于主动视觉驱动的双臂操作平台，分别在仿真和真机上验证了：“细节观测”对于有限样本条件下提升了精细操作任务的稳定性与鲁棒性，并以第一作者撰写一篇会议论文在投；面向数据驱动的操作任务，参与设计了基于外骨骼的包含触觉反馈的数据采集系统，验证了其装置的便携性、数据质量的有效性，并作为第二作者撰写一篇会议论文，已被接收；同时作为某头部机器人公司算法方向实习生，深度参与了基于轮式双臂移动操作本体的重点落地项目，在仿真和真机平台上系统验证了数据驱动统一控制框架的可行性。这些工作为后续开展基于视觉-语言-动作模型的具身移动操作研究提供了可直接复用的数据管线和算法模块。

6.2　在学期间发表的学术论文与研究成果

[1] Shoujie Li*, **Yushan Liu***, et al. ALARMbot: Autonomous Laboratory Safety Inspection and Operable Hazard Intervention Robot Enabled by Foundation Models. IEEE Transactions on Automation Science and Engineering (T-ASE), Under Review. (共同一作，JCR 一区期刊)

[2] **Yushan Liu***, Shilong Mu*, et al. AVR: Active Vision-Driven Precise Robot Manipulation with Viewpoint and Focal Length Optimization. IEEE International Conference on Robotics & Automation (ICRA 2026), Under Review. (第一作者，清华大学 A 类会议)

[3] Xintao Chao*, Shilong Mu*, **Yushan Liu**, et al. Exo-ViHa: A Cross-Platform Exoskeleton System with Visual and Haptic Feedback for Efficient Dexterous Skill Learning. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025), Accepted. (第二作者，清华大学 B 类会议)

6.3　研究计划

表 6.1　研究工作计划

日期	研究内容
2025.09–2025.12	梳理视觉–语言–动作（VLA）、具身导航与移动操作相关文献，统一现有实验平台与代码框架，完成主要仿真环境、数据处理与训练评测流水线的搭建；
2026.01–2026.04	面向移动操作任务设计 VLA 中间表征与模型架构，在典型室内场景中完成基础策略的实现与对比，开展仿真实验与消融分析；
2026.05–2026.08	对拟提出的 VLA 驱动具身移动操作方法进行系统完善与工程优化，在多种真机和多类场景下验证其通用性与鲁棒性；
2026.09–2026.11	整理与归纳全部实验数据与代码，撰写至少一篇学术论文并完成投稿；
2026.12–2027.02	完成学位论文初稿；
2027.03–2027.06	完成学位论文终稿，准备毕业答辩。

参考文献

[1] Liu Y, Chen W, Bai Y, et al. Aligning cyber space with physical world: A comprehensive survey on embodied ai[J]. IEEE/ASME Transactions on Mechatronics, 2025, 30(3): 1-22.

[2] Xiao X, Liu J, Wang Z, et al. Robot learning in the era of foundation models: A survey[J]. Neurocomputing, 2025: 129963.

[3] Thakar S, Srinivasan S, Al-Hussaini S, et al. A survey of wheeled mobile manipulation: A decision-making perspective[J]. Journal of Mechanisms and Robotics, 2023, 15(2): 020801.

[4] 中华人民共和国国务院. 新一代人工智能发展规划 [EB/OL][EB/OL]. 2017. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[5] 工业和信息化部等. “十四五”机器人产业发展规划 [EB/OL][EB/OL]. 2021. https://www.gov.cn/zhengce/zhengceku/2021-12/28/content_5664988.htm.

[6] 中华人民共和国教育部. 高等学校人工智能创新行动计划 [EB/OL][EB/OL]. 2018. http://www.moe.gov.cn/srcsite/A16/s7062/201804/t20180410_332722.html.

[7] Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2018.

[8] Ahn M, Brohan A, Brown N, et al. Do as i can, not as i say: Grounding language in robotic affordances[A]. 2022.

[9] Brohan A, Brown N, Carbajal J, et al. Rt-1: Robotics transformer for real-world control at scale [A]. 2022.

[10] Driess D, Xia F, Sajjadi M S, et al. Palm-e: An embodied multimodal language model[A]. 2023.

[11] Kim M J, Pertsch K, Karamcheti S, et al. Openvla: An open-source vision-language-action model[A]. 2024.

[12] Jiang Y, Gupta A, Zhang Z, et al. Vima: General robot manipulation with multimodal prompts [C]//Fortieth International Conference on Machine Learning. PMLR, 2023.

[13] Shridhar M, Manuelli L, Fox D. Cliport: What and where pathways for robotic manipulation [C]//Conference on robot learning. PMLR, 2022: 894-906.

[14] Shridhar M, Manuelli L, Fox D. Perceiver-actor: A multi-task transformer for robotic manipulation[C]//Conference on Robot Learning. PMLR, 2023: 785-799.

[15] Zitkovich B, Yu T, Xu S, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control[C]//Conference on Robot Learning. PMLR, 2023: 2165-2183.

[16] 姚陈鹏, 石文博, 刘成菊, 等. 移动机器人导航技术综述[J]. 中国科学: 信息科学, 2023, 53 (12): 2303-2324.

[17] Abdulsaheb J A, Kadhim D J. Classical and heuristic approaches for mobile robot path planning: A survey[J]. Robotics, 2023, 12(4): 93.

[18] Wu Y, Zhang P, Gu M, et al. Embodied navigation with multi-modal information: A survey from tasks to methodology[J]. Information Fusion, 2024, 112: 102532.

[19] 王文晟, 谭宁, 黄凯, 等. 基于大模型的具身智能系统综述[J]. 自动化学报, 2025, 51(1): 1-19.

[20] 高超, 杨莹, 陈世超, 等. 多模态模型驱动的具身智能研究综述[J]. 智能感知工程, 2025, 2 (2): 1-12.

[21] Krantz J, Maksymets O, Gokaslan A, et al. Instance-specific image goal navigation: Training embodied agents to find object instances[A]. 2022.

[22] Chaplot D S, Gandhi D P, Gupta A, et al. Object goal navigation using goal-oriented semantic exploration[C]//Larochelle H, Ranzato M, Hadsell R, et al. Advances in Neural Information Processing Systems (NeurIPS). Curran Associates, Inc., 2020: 4247-4258.

[23] Gu J, Stefani E, Wu Q, et al. Vision-and-language navigation: A survey of tasks, methods, and future directions[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Dublin, Ireland: Association for Computational Linguistics, 2022.

[24] Savva M, Kadian A, Maksymets O, et al. Habitat: A platform for embodied ai research[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE Computer Society, 2019.

[25] Huang K, Shi B, Li X, et al. Multi-modal sensor fusion for auto driving perception: A survey [A]. 2022.

[26] 张燕咏, 张莎, 张昱, 等. 基于多模态融合的自动驾驶感知及计算[J]. 计算机研究与发展, 2020, 57(9): 1781-1799.

[27] McCormac J, Handa A, Davison A, et al. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks[C]//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017: 4628-4635.

[28] Narita G, Seno T, Ishikawa T, et al. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 4205-4212.

[29] Dai A, Chang A X, Savva M, et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 2432-2443.

[30] Chang A, Dai A, Funkhouser T, et al. Matterport3d: Learning from rgb-d data in indoor environments[C]//Proceedings of the International Conference on 3D Vision (3DV). IEEE, 2017: 667-676.

[31] Peng S, Genova K, Jiang C, et al. Openscene: 3d scene understanding with open vocabularies [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 815-824.

[32] He Q, Peng J, Jiang Z, et al. Unim-ov3d: uni-modality open-vocabulary 3d scene understanding with fine-grained feature representation[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI). IJCAI, 2024: 821-829.

[33] Werby A, Huang C, Büchner M, et al. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation[C]//Proceedings of Robotics: Science and Systems. Delft, Netherlands: RSS, 2024.

[34] Waymo. Waymo self-driving technology overview[EB/OL]. 2023[2025-11-13]. https://waymo.com/tech/.

[35] Karpathy A, Tesla Autopilot Team. Tesla full self-driving (fsd) v12 end-to-end neural network architecture[EB/OL]. 2023[2025-11-13]. https://www.tesla.com/AI.

[36] Han D, Mulyana B, Stankovic V, et al. A survey on deep reinforcement learning algorithms for robotic manipulation[J]. Sensors, 2023, 23(7): 3762.

[37] Celemin C, Pérez-Dattari R, Chisari E, et al. Interactive imitation learning in robotics: A survey [J]. Foundations and Trends® in Robotics, 2022, 10(1-2): 1-197.

[38] Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies[J]. Journal of Machine Learning Research, 2016, 17(39): 1-40.

[39] Kalashnikov D, Irpan A, Pastor P, et al. Scalable deep reinforcement learning for vision-based robotic manipulation[C]//Conference on robot learning. PMLR, 2018: 651-673.

[40] Jang E, Irpan A, Khansari M, et al. Bc-z: Zero-shot task generalization with robotic imitation learning[C]//Conference on Robot Learning. PMLR, 2022: 991-1002.

[41] Chi C, Xu Z, Feng S, et al. Diffusion policy: Visuomotor policy learning via action diffusion [J]. The International Journal of Robotics Research, 2025, 44: 1684-1704.

[42] Ze Y, Zhang G, Zhang K, et al. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations[A]. 2024.

[43] Huang X, Batra D, Rai A, et al. Skill transformer: A monolithic policy for mobile manipulation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2023: 10852-10862.

[44] Fu Z, Cheng X, Pathak D. Deep whole-body control: learning a unified policy for manipulation and locomotion[C]//Conference on Robot Learning. PMLR, 2023: 138-149.

[45] Fu Z, Zhao T Z, Finn C. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation[A]. 2024.

[46] Liang J, Huang W, Xia F, et al. Code as policies: Language model programs for embodied control[A]. 2022.

[47] O'Neill A, Rehman A, Maddukuri A, et al. Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 6892-6903.

[48] Intelligence P, Black K, Brown N, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization[A]. 2025.

[49] Ma Y, Song Z, Zhuang Y, et al. A survey on vision-language-action models for embodied ai [A]. 2024.

[50] Zhong Y, Bai F, Cai S, et al. A survey on vision-language-action models: An action tokenization perspective[A]. 2025.

[51] Wu Z, Zhou Y, Xu X, et al. Momanipvla: Transferring vision-language-action models for general mobile manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2025: 1714-1723.

[52] Team G R, *et al.* Gemini robotics: Bringing ai into the physical world[A]. 2025.

[53] James S, Ma Z, Arrojo D R, et al. Rlbench: The robot learning benchmark & learning environment[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3019-3026.

[54] Yu T, Quillen D, He Z, et al. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning[C]//Conference on robot learning. PMLR, 2020: 1094-1100.

[55] Gu J, Xiang F, Li X, et al. Maniskill2: A unified benchmark for generalizable manipulation skills[A]. 2023.

[56] Mees O, Hermann L, Rosete-Beas E, et al. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 7327-7334.

[57] Liu B, Zhu Y, Gao C, et al. Libero: Benchmarking knowledge transfer for lifelong robot learning [J]. Advances in Neural Information Processing Systems, 2023, 36: 44776-44791.

[58] Zhang S, Xu Z, Liu P, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2025: 11142-11152.

[59] Szot A, Clegg A, Undersander E, et al. Habitat 2.0: Training home assistants to rearrange their habitat[J]. Advances in neural information processing systems, 2021, 34: 251-266.

[60] Li C, Xia F, Martín-Martín R, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks[A]. 2021.

[61] Li C, Zhang R, Wong J, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation[C]//Conference on Robot Learning. PMLR, 2023: 80-93.

[62] Paxton C, Wang A, Shah B, et al. Homerobot: An open source software stack for mobile manipulation research[C]//Proceedings of the AAAI Symposium Series: Vol. 2. 2023: 518-525.

[63] Jaafar A, Raman S S, Wei Y, et al. Lanmp: A language-conditioned mobile manipulation benchmark for autonomous robots[A]. 2024.

[64] Zhu Y, Wong J, Mandlekar A, et al. Robosuite: A modular simulation framework and benchmark for robot learning[A]. 2020.

[65] Makoviychuk V, Wawrzyniak L, Guo Y, et al. Isaac gym: High performance gpu-based physics simulation for robot learning[A]. 2021.