大模型与智能机器人的融合：智能感知、导航与操作

谢远龙[1]，李文龙[1]，张友民[2]，唐剑[3]，周雪峰[4]，
郑世祺[5]，杨恺伦[6]，谢胜泉[7]，王书亭[1]

（1. 华中科技大学机械科学与工程学院，湖北 武汉 430074；
2. 加拿大康考迪亚大学机械与航天航空学院，加拿大 蒙特利尔 H3G 1M8；
3. 国家地方共建具身智能机器人创新中心，北京 101100；
4. 广东省科学院智能制造研究所，广东 广州 510650；
5. 中国地质大学（武汉）自动化学院，湖北 武汉 430074；
6. 湖南大学机器人学院，湖南 长沙 410082；
7. 英国利兹大学电子与电气工程学院，英国 利兹 LS2 9JT）

摘　要：随着人工智能与机器人技术的深度融合，智能机器人正从依赖预设规则的结构化环境向开放、动态的复杂场景演进。传统方法在语义理解、长时序任务及人机自然交互方面面临泛化不足的挑战。本综述系统梳理了大语言模型及多模态大模型赋能机器人智能感知、导航与操作的关键进展：在感知层面，通过多模态融合与语言-空间联合推理，增强了对环境语义与几何属性的深度理解；在导航层面，利用思维链任务分解与常识推理，实现了对模糊指令的解析与在未知环境中的自主探索；在操作层面，借助视觉-语言动作模型与物理常识耦合，提升了复杂交互任务的灵巧性与适应性。研究表明，大模型的引入推动了机器人技术从“感知驱动”到“认知驱动”的范式转变，显著提升了系统的上下文推理与自主决策能力。然而，跨模态对齐精度、实时性能、安全可靠性及仿真到现实的泛化等核心问题仍有待突破。本文为构建通用化、认知增强的智能机器人系统提供了系统的技术参考与发展路线。

关键词：智能机器人；人工智能；多模态大模型；智能导航；灵巧操作

Integration of Large Models and Intelligent Robots: Technologies for Intelligent Perception, Navigation and Manipulation

XIE Yuanlong[1], LIWenlong[1], ZHANG Youmin[2], TANG Jian[3], ZHOU Xuefeng[4],
ZHENG Shiqi[5], YANG Kailun[6], XIE Shengquan[7], WANG Shuting[1]

(1. *School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan* 430074, *China*;
2. *Department of Mechanical, Industrial and Aerospace Engineering, Concordia University, Montreal H3G 1M8, Canada*;
3. *National and Local Joint Innovation Center for Embodied Intelligent Robots, Beijing* 101100, *China*;
4. *Guangdong Institute of Intelligent Manufacturing, Guangzhou* 510650, *China*;
5. *School of Automation, China University of Geosciences (Wuhan), Wuhan* 430074, *China*;
6. *School of Robotics, Hunan University, Changsha* 410082, *China*;
7. *School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, UK*)

Abstract: With the deep integration of artificial intelligence and robotics, intelligent robots are evolving from structured environments reliant on pre-defined rules towards open, dynamic, and complex scenarios. Traditional methods face challenges in generalization when dealing with semantic understanding, long-horizon tasks, and natural human-robot interaction. This survey systematically reviews key advancements in intelligent perception, navigation, and manipulation empowered by large language models and multimodal large models. At the perception level, multimodal fusion and language-spatial joint reasoning enhance deep understanding of environmental semantics and geometric attributes. For navigation, techniques such as chain-of-thought task decomposition and commonsense reasoning enable the parsing of ambiguous instructions and autonomous exploration in unknown environments. In manipulation, vision-language-action models coupled with physical commonsense improve the dexterity and adaptability of complex interactive tasks. Research indicates that the introduction of large models drives a paradigm shift in robotics from “perception-driven” to “cognition-driven,” significantly enhancing the system's capabilities in contextual reasoning and autonomous decision-making. However, core challenges remain, including cross-modal alignment accuracy, real-time performance, safety, reliability, and Sim2Real generalization. A systematic technical reference and development roadmap are provided for building general-purpose, cognition-enhanced intelligent robotic systems.

Keywords: intelligent robots, artificial intelligence, multimodal large model, intelligent navigation, dexterous manipulation

人工智能（AI）推动机器人逐步从封闭、预设的结构化环境向复杂、非结构化且动态演化的开放场景演进[1]，如图 1 所示。传统机器人多部署于高度可控的工业场景，依赖精确环境建模、确定性路径规划与预编程行为逻辑执行任务，如基于规则的运动规划（如 A*、RRT 系列算法）[3]与基于几何建模的感知系统（如 SLAM）[4]。在面对真实世界中高度不确定的语义环境、动态障碍物、长时序依赖任务以及人类意图的隐式表达时，此类系统存在泛化能力不足与适应性缺陷。

近年来，机器人应用场景从工厂流水线延伸至家庭服务、医疗康复与应急救援等复杂场景，其对环境理解、自主决策与实时交互能力提出了前所未有挑战。在此背景下，传统的基于环境模型、预设规则、演示示教的技术路线显示出固有的局限性：环境感知方法局限于对环境几何特征的提取与表征，缺乏对高维度场景语义以及环境变化上下文关系的深层理解；行为决策方面依赖显式的策略流程与离线优化，难以应对环境突发变化和自然语言指令等新形态任务需求；物体操作技术方面则多基于预设流程和人工示教，难以对真实世界的多维度属性及其动态变化做出响应。由于缺少对语言、语义、图像等复杂高维信息的有效处理和利用，上述方案在各种新的应用场景和应用需求中效果不佳。特别是涉及多模态输入融合（如视觉—语言—触觉）、长程时序推理与跨模态语义对齐等任务时，传统方法的可扩展性与鲁棒

基金项目：国家自然科学基金（52275488，52575573，52205524，52188102）；国家重点研发计划（2023YFB4703800）.
通信作者：王书亭，wangst@hust.edu.cn
收稿/录用/修回：2025-10-31/2026-03-01/2026-03-13

性显著下降。

为突破上述瓶颈，满足行业发展对相关支撑性技术的需求，学术界与工业界聚焦“智能感知与导航驱动”范式，以强化智能体在物理世界中的感知—导航—行动闭环与情境化认知能力为核心，着力推动机器人从“感知驱动”向“认知驱动”转变。其中的重点是将认知过程嵌入物理交互，使机器人不仅能理解真实世界的多维度信息，还能够有效处理和利用它们，实现基于语义理解、常识推理与情境感知的自主决策与适应性行为生成[5]。以经典机器学习算法（如深度学习和强化学习）为核心的技术雏形，让机器人开始具备处理复杂高维数据并生成动作指令来完成导航和操作等高级任务。但受限于基础性算法本身的上限，缺少强大的神经网络基础模型作为支撑，相关的研究成果和技术多停留于仿真和简单真实场景，未达到人们对“智能化”的期许。近年来，以大语言模型（LLMs）[6]为代表的生成式人工智能技术，特别是基于 Transformer 架构的自回归模型（如 GPT 系列、PaLM、LLaMA 等），在自然语言理解、逻辑推理与知识表征方面展现出超越传统符号系统的能力。通过引入多模态大模型（MLLMs），如 Flamingo、Kosmos-2、LLaVA 与 Qwen-VL，机器人可实现跨模态语义对齐与联合推理，在复杂场景中完成“感知—导航—操作”一体化任务。

在此技术演进下，大模型与智能机器人的融合已成为推动智能感知、导航与操作技术发展的关键路径：大模型为机器人提供了类人级的常识知识库、上下文感知与任务分解能力，机器人则为大模型提供了物理世界中的感知输入与行动反馈机制，二者共同实现从感知真实世界到与真实世界产生互动的信息循环。这一融合不仅重塑了传统机器人的系统架构，更催生出基于大模型的具身推理[7]、具身强化学习[8]、可解释性决策生成、动态任务重规划与人—机—环境协同交互等新兴研究方向。大模型的引入，尤其是基于大规模无监督预训练的多模态大模型，如 Flamingo、Kosmos-2、LLaVA、Qwen-VL 与 Pix2Seq，可在海量图文对、视频—文本对等多源异构数据上进行联合表征学习，显著提升机器人对开放场景语义信息的抽象与泛化。进一步地，知识增强型模型的引入，大幅提升了机器人的认知能力。通过引入外部知识库与上下文学习机制，机器人可在执行任务过程中动态调用常识知识，对物理规律、社会规范与因果逻辑实现隐式建模。如语言模型驱动的自然语言指令执行[9]使机器人能直接解析人类的模糊、非结构化指令，实现从“指令—动作映射”到“语义—行为对齐”的跨越，增强人机协作的自然性与可操作性。

图 1 智能机器人及其支撑性技术发展历程

Fig.1 The development history of intelligent robots and their supporting technologies

当前围绕大模型与智能机器人融合的研究仍处于早期的快速演化阶段。若以应用场景或应用对象为标准进行整理，相关成果广泛分布于计算机视觉、自然语言处理、机器人学、决策理论等多个领域，研究焦点高度分散且未形成权威的主流技术路线。具体的，较为热门的研究工作集中于大模型在视觉语言理解中的表征能力（如 CLIP、BLIP 系列），部分聚焦于模型与控制接口的适配（如 PROMPT-RL[10]），还有一部分则更倾向于探索大模型在具身决策中的可解释性与安全性（如 Chain-of-Thought、Tree-of-Thought 和 Reasoning with Uncertainty），而另一些则致力于构建端到端的具身智能框架（如 Agent-based Embodied AI、RoboCrafter 和 VLM-Driven Embodied Agents）。这些研究展现出远超早期经典机器学习方法的强大功能和应用潜力。与此同时，大量新技术成果的集中涌现往往也伴随着系统性的整合与横向分析对比的缺乏。因此，系统梳理和分析大模型与智能机器人在智能感知、导航与操作技术等关键环节的技术进展，对于推动智能机器人向更高层次的自主性、适应性与通用性发展具有重要理论与实践意义。本文旨在回答如下关键问题：大模型如何重构传统机器人技术链路？其在提升机器人泛化能力、降低对人工标注依赖方面实现何种范式突破？在引入多模态理解、自我反思与自然语言交互等新能力的同时，如何应对推理延迟、可解释性缺失、安全边界模糊等挑战？为此，本

文将进一步梳理当前主流技术路径的共性框架与差异特征，归纳面向通用智能体的系统架构演进方向，为未来大模型驱动的智能机器人提供理论支撑与技术路线。

1 智能感知（Intelligent perception）

1.1 智能感知的核心目标与挑战

智能感知的核心目标在于使机器人能通过多模态传感器实时获取环境信息，在语义、几何与物理层面实现精准的环境理解与交互支持。其根本任务是突破传统技术路线中被动感知方案的局限性，为智能体在动态、开放环境中的自主决策与灵巧操作提供必要的基础信息[1]。这一目标可分解为如下 3 个关键维度：一是多模态融合与协同感知，即整合视觉、语言、触觉等异构模态信息，构建统一且互补的场景表征，以支持对物体属性、空间关系以及物理特性等多维度属性的联合推断；二是语义与空间的联合定位，不仅识别物体和场景的语义类别，还需在 3 维空间中精确理解其位置、姿态、运动状态等几何与物理信息，为后续动作生成提供可操作的空间参考；三是环境交互与动态适应，通过与环境的实时物理交互，主动减少感知过程中的不确定性，适应光照变化、遮挡及物体形变等挑战，最终实现智能体在仿真和现实之间的高效泛化。

智能感知的核心挑战贯穿感知、推理、执行全流程，既存在单一模态的固有局限，也面临跨模态与跨场景的深层矛盾，主要面临以下一系列挑战：

1) Sim2Real 场景下，仿真与真实世界在渲染、内容、传感器及动力学模型上的差异难以弥合。具体表现为，由于仿真与真实环境的特征分布不一致，导致跨模态数据对齐失效，进而在多目标关联与动态空间处理中引发语义空间感知偏差。同时，模型在陌生真实场景中更易诱发语言幻觉与空间感知断层，显著削弱推理的可靠性。

2) 在空间信息表征问题中，图像驱动的方法往往受限于传感器精度与外部干扰，而点云驱动的方法则经常面临语义几何匹配难题。除此之外，在跨模态数据融合过程中，紧耦合与松耦合策略又难以平衡计算效率与精度。

3) 现有方法多适配静态环境，难以应对物体移动、光照变化等扰动，模型感知与关联更新滞后，导致高度动态场景中响应效率不足。在实时性与硬件成本层面，当前高精度模型延迟高而轻量模型精度低，专用硬件门槛与跨硬件适配难题还限制了规模化应用。

1.2 大模型驱动的智能感知关键技术方法

大模型驱动的具身感知技术体系以“环境感知—语义解析—动作生成—现实落地”的全链路能力构建为核心目标，其中多模态感知融合、视觉—语言—动作（VLA）解码与表征、语言—空间联合主动推理以及 Sim2Real 感知差异与泛化 4 大关键技术，形成了从底层数据获取到高层智能决策、再到现实场景适配的逐层递进、紧密耦合的有机整体。如图 2 所示。通过各技术环节的协同联动实现能力跃升：多模态感知融合作为感知入口，为后续技术环节提供全面且精准的环境数据支撑；VLA 解码与表征承担决策中枢功能，将感知信息转化为可执行的动作指令；语言—空间联合主动推理则强化语义与物理空间的关联认知以提升决策合理性；Sim2Real 感知差异与泛化用以解决虚拟训练与现实应用的鸿沟问题，确保整个技术体系的实用性与泛化性。

1.2.1 大模型赋能的多模态感知融合

与传统多模态感知仅关注静态数据处理不同，多模态感知融合的核心特征在于感知与反馈的动态闭环，智能体通过物理动作改变感知视角或与物体交互，同步获取视觉、触觉、听觉等多模态反馈信息，再经大模型处理实现环境认知的迭代优化。作为智能机器人系统的核心支撑，动态融合技术赋予机器人对操作对象多维度属性（物理、空间、语义）的深度认知能力，从而确保精确抓取、灵巧操作和人机协同交互等任务的可靠执行。基于不同数据源的动态交互特性与数据表示特征，适用于智能机器人系统的多模态融合感知方法可分为以下 3 类。

基于点云数据的融合方法。是机器人 3 维环境感知的主流技术路径，主要分为传统方法与端到端方法两大类，主要应用于机器人对操作场景的 3 维几何建模与目标定位。传统方法基于成熟的计算机视觉技术，利用 2 维语义分割[12] 、目标检测等手段，将图像转换为带有语义信息的伪点云[13]，进而与激光雷达采集的真实点云融合，构建出完整的 3 维语义环境模型。该方法对机器人系统的优势在于：可直接适配现有视觉模型，模块化程度高，且降低了 3 维标注数据的依赖成本；端到端方法则通过卷积或 Transformer 网络直接对点云数据进行深层特征提取[14]，并引入图像模态信息挖掘点云的语义关联与精细深度特征，实现对机器人操作环境的精准感知与建模[15]。尤其在复杂场景下的目标识别与定位任务中表现更优，为机器人精确抓取提供了更高精度的 3 维环境信息。

基于图像数据融合方法。聚焦机器人视觉感知的高效性与语义理解能力，分为基于 RGB 图像的方法与基于 BEV 图像的方法。基于 RGB 图像的方法通过跨模态交互机制实现图像与文本语义的关联理解，以自然语言监督训练或自监督学习为核心构建模型，融合多模态数据，使机器人能够通过自然语言指令定位目标对象，同时降低了大规模机器人感知数据集的构建成本[16]。基于 BEV 表征的融合方法将摄像头、激光雷达等异构传感器数据统一映射至 BEV 空间[17]，有效消除不同传感器在高度信息上的不一致性，实现多模态数据的统一编码处理。该类方法能够为移动操作机器人的路径规划与实时抓取决策提供高效的感知输出。

基于异构表征的融合方法。旨在解决机器人复杂操作时的感知难题。该方法摒弃了统一的表征范式，针对视、听、触等模态的异构特性分别提取特征[20]，并利用协同融合技术提升机器人的环境感知水平。该方法聚焦各模态数据的本质优势，有针对性地提取关键特征并进行融合，从而有效弥补单一模态的感知局限[21]，实现更稳定的抓取姿态调整；同时避免了统一编码可能带来的信息损耗或模态偏差问题，保障了关键感知信息的完整性。但该方法对机器人系统的模型设计要求较高，需针对触觉、视觉、听觉等不同模态的维度特性单独设计特征提取逻辑，导致模型结构复杂度相对较高[22]，对传感

器同步与数据预处理提出了更高要求。

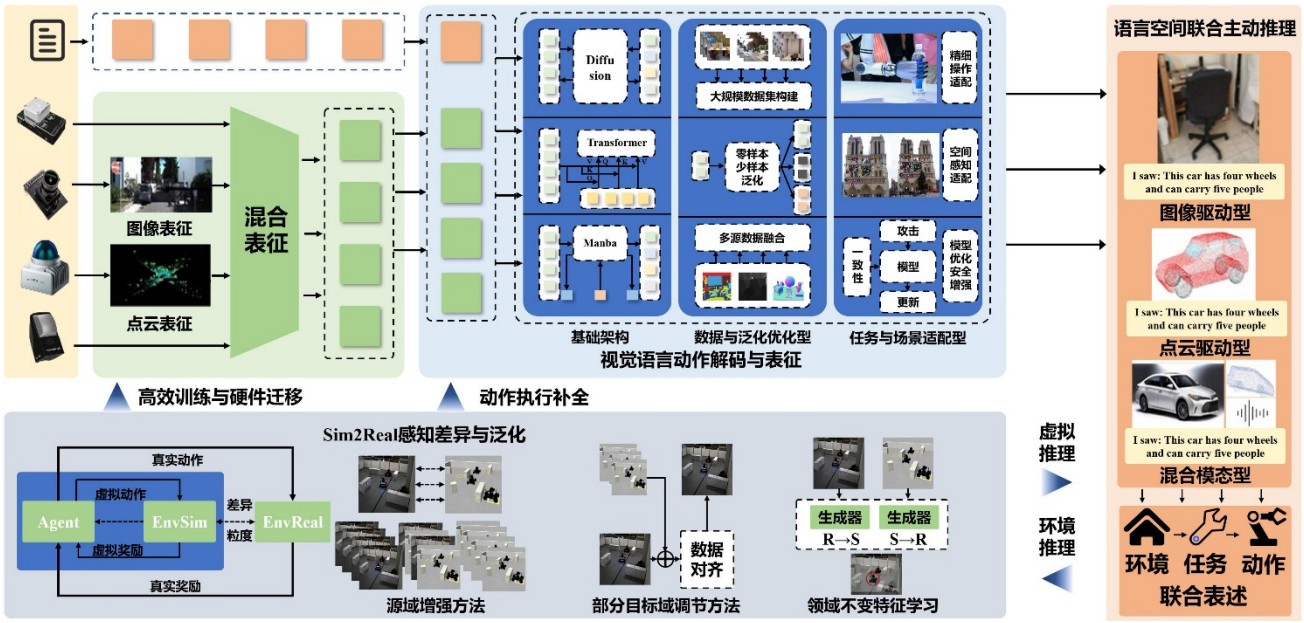

图 2 具身感知系统框架图

Fig.2 Framework of the embodied perception system

1.2.2 视觉语言动作解码与表征

目前依赖手工规则或单一模态的操控方法难以满足智能体对“感知—理解—决策—执行”全链路协同的需求[23]。如图 3 所示，VLA 模型通过深度耦合视觉感知、语言理解与动作生成，成为突破传统机器人操控局限的关键方向[24]。针对早期端到端方法任务单一、泛化能力有限的痛点，VLA 模型不仅能灵活解析任务指令，还能结合实时的视觉感知动态调整策略，有效克服了传统预编程方法的刚性束缚。

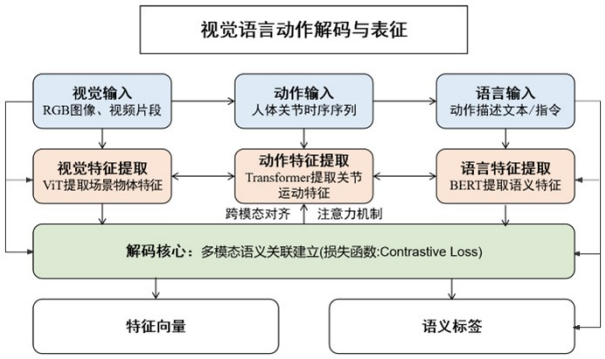

图 3 视觉语言动作解码与表征框架图

Fig.3 Framework of visual language action decoding and representation

1) VLA 基础模型

生成式架构：生成式 VLA 模型将动作生成视为连续空间的概率建模问题，主要分为扩散模型和流匹配模型两大类。扩散模型通过前向加噪与反向去噪的迭代机制，学习动作分布并生成连续动作序列。Diffusion Policy[25]创新性地将视觉场景特征与语言指令语义映射至同一特征空间，为动作优化过程提供场景适配与任务目标约束。π 0 模型[26] 进一步将扩散模型与视频预测相结合，实现了长时程任务的稳定执行。流匹配模型则通过直接学习概率密度函数的变换来生成动作，在保持生成质量的同时显著提升了推理速度。生成式架构的优势在于能够生成平滑连续的轨迹，避免了离散化带来的精度损失，但需要大量高质量的动作数据进行训练。

自回归架构：自回归 VLA 模型将动作生成视为序列预测问题，通过逐步生成离散的动作标记来完成任务。这类方法依托 Transformer 架构的强大多模态理解能力，将视觉感知和语言理解统一编码，实现高精度的动作决策。代表性工作 RT-1 模型[27] 是首个大规模真实场景 Transformer-based VLA 模型，其通过共享编码器将视觉帧序列与语言 token 序列编码为统一特征，再利用轻量化动作头输出关节角度、末端状态等控制信号。OpenVLA[28]进一步扩展了这一范式，在 970K 机器人 episodes 数据上预训练，实现了更强的泛化能力。自回归架构的优势在于能够充分利用预训练 VLM 的推理能力，实现复杂的任务分解和逻辑控制，但存在误差累积和推理延迟较高的固有缺陷。

混合架构：为了结合自回归架构的推理能力与生成式架构的连续控制优势，研究者提出了多种混合架构。Hybrid VLA 模型[29]创新性地在单一 LLM 中同时集成扩散和自回归动作生成能力，通过协作集成实现更稳定的执行。Diffusion VLA 模型[30]采用自回归推理与扩散生成的层次化设计，上层负责任务推理，下层负责轨迹生成。这类架构试图平衡效率、准确性和适应性，是当前 VLA 研究的重要方向。

2) 数据与泛化优化型 VLA 模型

数据稀缺与泛化能力不足是 VLA 模型落地的核心瓶颈：传统方法依赖大规模、高质量的 VLA 配对样本，而机器人样本采集成本高、场景覆盖有限。数据与泛化优化型 VLA 模型通过构建大规模数据集、设计零样本/少样本学习策略以及融合多源数据知识，旨在突破数据约束，提升模型在未见过场景中的适配能力。

大规模 VLA 数据集构建：构建方法主要分为跨平台标准化和真实开放场景两大类。跨平台标准化方法依赖统一动作表示、多场景采集以及数据集与模型适配技

术，可提高数据的复用率、有效降低训练成本、提高模型的适配性，但同时容易丢失动作样本的细节。如O'Neill 等[31]针对跨平台数据复用难问题，将工业机械臂、服务机器人等动作统一为通用格式，同步关联场景、指令、动作数据，并使用三元组进行表达和存储，形成大规模的 VLA 数据集与统一框架[32]。

零样本/少样本泛化：核心是迁移预训练知识，用来解决大场景下的泛化问题，实现少量/无机器人样本适配新任务，主要分为任务零样本和语义驱动零样本两种主要类型。任务零样本泛化依赖任务无关动作表示与指令视觉协同。如 Jang 等[33]提出首个方案：将动作拆为基础运动单元与动态参数，预训练单元后，通过指令解析目标、视觉定位位置，填入参数生成新动作，无需新任务机器人样本。语义驱动零样本泛化依托预训练 VLM[34]，把指令拆解为目标检测、空间推理、动作规划逻辑链，结合机器人运动学模型转化为动作参数，无需预学动作规律。

多源数据融合 VLA：通过整合不同模态、场景、来源的数据，弥补单一数据源的局限性，强化模型对复杂场景的理解与动作生成能力。该方法依托多模态特征统一编码与跨源数据协同技术，如 Reed 等[35]在通用 VLA 模型研究中，通过整合机器人操控数据、人类对话数据、图像识别数据及视频动作数据，实现了从对话理解到动作生成的完整链路。多源数据的互补性提升了复杂场景下的建模精度，端到端建模减少了特征损失；但异构数据格式差异大导致融合难度较高，同时数据质量参差不齐易引入噪声，且对硬件算力提出了较高要求。

3) 任务与场景适配型 VLA 模型

任务与场景适配型 VLA 模型通过定制特征提取模块、优化动作生成策略，让模型更贴合具体的应用需求，最大化特定场景下的性能。

精细操作适配。通过 VLA 闭环协同，满足精细操作对动作精度、力控的高要求，避免动作误差导致任务失败。该方法依托高分辨率视觉定位、实时力控反馈与动态动作修正技术，如双机械臂精细协作任务[36]中通过高分辨率视觉定位目标细节、力控反馈调整交互力度，在提升操作精度的同时降低硬件成本；在动作执行中[37]，由 VLM 实时分析视觉反馈并生成调整建议，进而驱动动作生成模块来修正后续轨迹。这种动态修正机制不仅提升了精细操作的精度，还有效增强了系统的容错性。

空间感知适配。要求机器人在复杂空间场景中，精准理解物体间的空间位置关联，避免因空间理解偏差引发的动作错误，解决因空间认知不足引发误操作的问题。依托场景体积编码和 3 维空间特征映射技术，Liu 等[38]虽以导航任务为研究重点，但其提出的体积环境表征思路可迁移至 VLA 的操控场景中：将场景编码为融合视觉语义与空间体积特征的网格，减少点云重复计算，提升空间理解效率。Qu 等[39]则创新引入 3 维空间特征编码方案，通过 RGB-D 相机获取点云以构建 3 维网格，再将语言指令中的空间描述映射为网格空间约束，以此指导动作生成，强化模型的空间推理能力。空间特征编码技术提升了复杂场景下的认知精度，跨任务的技术迁移则降低了研发成本。

模型优化与安全增强。旨在通过强化学习提升性能，并引入安全机制以规避潜在风险。该方法不仅满足了场景扩展对系统可靠性、训练效率及推理速度的严苛要求，还有效解决了复杂环境下的性能波动与高安全风险问题。该方向依托强化学习性能优化、人类反馈安全机制，在微调阶段引入动作一致性奖励函数[40]，在相同场景指令下，若动作结果一致则给予高奖励，若结果波动大则施加惩罚。

1.2.3 sim2real 下感知差异与泛化

仿真训练是当前机器人及通用智能体感知系统开发中的主流范式，然而，如何克服“Sim2Real”之间的本质差异，仍是其面临的根本性挑战：仿真训练的模型迁移到物理世界时性能会显著下降。该问题的核心是仿真与现实世界之间存在固有的现实差距，直接导致了模型的感知差异与泛化难题。此差距主要体现在视觉渲染、场景内容、传感器建模和物理动力学模型等方面。因此，系统性地减小此差距，是在现实中实现仿真模型可靠部署的关键。

源域增强方法。通过优化仿真数据的生成过程与分布，实现与真实数据的对齐，主要分为高保真重建、域随机化与通用世界模型三类。高保真重建遵循 Real2Sim2Real 的思想，旨在从源头消除感知差异。早期工作利用图形学技术精确模拟物理过程与材质。随着神经渲染技术的发展，研究重心转向利用神经辐射场和 3D 高斯泼溅等技术直接从真实图像中重建逼真场景[41]。此方法主要局限于其高昂的计算成本和复杂的建模流程，难以满足大规模数据生成与场景多样性的实际需求。域随机化则通过扩展源域分布，迫使模型学习鲁棒特征，Horváth 等[42]在物体检测任务中通过极端纹理扰动实现零样本迁移，通用世界模型则代表了最新范式，利用生成式 AI 完成数据生成。内容生成已从程序化规则进化为利用大语言模型自动编写仿真内容，以及利用扩散模型生成语义丰富的增强数据。此类方法利用海量数据训练的先验知识，使得模型能够在直接预测动力学演变。尽管研究前景广阔，但生成内容中不可控的幻觉与长时序物理一致性的缺失，容易导致误差累积，限制了其在复杂长周期任务中的部署能力。

领域不变特征学习方法。从模型层面出发，通过特定训练策略对齐仿真与真实数据的特征分布，主要分为无监督域自适应与领域泛化两类。无监督域自适应在训练中同时使用带标签的仿真源域数据和无标签的真实目标域数据，利用后者引导模型缩小域间分布差异。典型方法包括对抗域适应框架构建域对齐机制和梯度加权对抗对齐技术。然而，此类方法高度依赖无标签真实数据的可用性，若获取困难，则效果会大幅受限。相比之下，领域泛化在训练阶段模型仅能访问一个或多个源域，不接触目标域，旨在学习通用表示以泛化至未见目标域。该方法仅使用带标签的仿真源域数据，通过特殊训练策略提升模型泛化能力，如通过对源域特征进行多尺度扰动训练以提升对未知域差异的容忍度[43]，构建双向域适应框架，通过源域到中间域、中间域到潜在目标域的双向迁移，提升策略泛化性[44]，以及结合任务物理特性设计多模态特征融合策略，实现仿真技能到真实场景的迁移。该方法的最大优势是完全不依赖目标域数据，

部署灵活，但因缺乏目标域信息指导，难以精准建模域差异，导致通用特征学习难度大。

基于部分目标域调节方法。利用少量真实数据对模型或仿真环境进行针对性调整，主要包括微调、系统辨识与元学习三类。微调是先在大规模仿真数据上预训练模型，再结合去噪架构与迁移学习方案应对噪声干扰，随后对少量真实数据微调，通过设计预训练与参数预测联动机制，使模型能单次预测最优仿真配置[45]，最后通过真实场景小样本调整超参数。该方法流程简单、迁移效果直接，但对数据分布偏移敏感。系统辨识则通过少量真实物理实验收集数据，以估计仿真中的关键动力学参数。面对动态复杂任务，研究者采用基于采样的参数识别与主动探索框架[79]，高效筛选关键物理参数以提升估计精度。该方法适配性强，适合依赖动力学的任务，但极端场景的实验数据采集难度大，且易受实验噪声影响。元学习则是在多样仿真环境中训练快速适应能力，使模型学会利用少量数据快速更新参数[46]，进入真实场景后，仅需少量样本即可快速适配。

1.2.4 语言-空间联合主动推理

随着 3D 计算机视觉在空间信息感知能力的提升，引入大语言模型来实现跨模态融合成为构建高智能上下文感知系统的核心方向。LLM 与 3D 空间表征的融合，本质是实现语言语义与空间结构的双向映射机制：一方面，LLM 为 3D 数据赋予高层语义与上下文逻辑；另一方面，3D 空间表征为 LLM 提供物理世界的接地约束，从而有效抑制模型的幻觉现象。

图像驱动型语言-空间联合感知推理，以图像为空间信息的核心载体，构建语言语义与图像像素的关联映射。其中，单目图像驱动方法核心挑战在于解决深度歧义问题，这往往需要结合语言语义引导才能有效避免对空间关系的错误判断[47]。为此，通过单目深度估计、目标检测与语义分割的协同，建立语言描述与图像中特定区域的空间关联，可以有效避免上述问题。而多视图图像驱动方法则通过融合同一场景的多视角图像，利用视角互补性弥补单目图像的空间信息缺陷[48]。该方法通过提取多视图图像获取 3D 特征，让语言语义能关联到更精确的 3 维空间位置。除此之外，RGB-D 图像驱动方法依托 RGB 图像的语义信息与深度图像的几何信息，实现语义与空间直接关联[49]，相较于单目与多视图方法，其可直接获取像素级的 3 维坐标，能更高效地建立语言描述与空间实体的映射关系，在人机交互、机器人抓取等需精确空间度量的任务中具有显著优势[50]。但在多模态融合阶段，文本语义与几何特征的对齐往往难以保持高精度的尺度一致性，且现有模型对局部几何细节的感知能力依然受限。

点云驱动型语言-空间联合感知推理，以 3 维点云为空间信息载体，依托点云直接表征 3 维几何形态的优势，更真实地还原物理空间结构，分为直接对齐、分布对齐与任务定制对齐等方法。直接对齐方法聚焦于建立语言语义元素与点云几何元素的直接关联。其思路是先对语言进行语义解析，然后将编码的点云和 3D 类别的文本对齐，融合几何、外观与语言信息，让 LLM 直接理解点云[51]，以实现从 2D 到 3D 的知识转移。分布对齐方法针对点云无序性、稀疏性及语义模糊性，从概率分布层面实现语言与空间信息的匹配。任务定制对齐方法以具体应用任务需求为导向，设计个性化的语言-点云对齐逻辑。该方法通过定制特征提取模块，利用掩码解码功能[52]，引入对象标识符和以对象为中心的表示[53]，构建以对象为中心的 3D 指令数据集及对应交互提示，使对齐过程更适配任务目标，最大化特定场景下联合推理的性能。

混合模态型语言-空间联合感知推理，融合了图像与点云的优势，主要分为紧耦合、松耦合两种策略。紧耦合策略强调图像与点云信息的深度交互融合，通常在特征提取的早期阶段即开展模态信息的深度融合，构建跨 3D 点云与图像、语言、音频等多模态的联合嵌入空间[54]，先将图像纹理语义特征嵌入点云几何特征，再通过端到端预训练对齐点云特征与图文对齐特征[55]，形成语义-几何一体化的特征表示。这种方式能充分挖掘两种模态的互补性，减少模态间信息丢失，适用于对空间精度与语义丰富度均有较高要求的任务。松耦合策略先对图像与点云分别进行独立的特征提取及语义、空间信息解析，再在决策层将两种模态的结果与语言语义相关联，进而实现语言、动作与感知的协同关联[56]。

1.3 小结

现有的智能感知技术主要作为机器人导航与操作的基础，提供可靠的显式或隐式表征，根据机器人作业场景、作业任务等差异性，从不同的维度提供针对性感知信息，如表 1 所示。然而，智能感知技术目前仍存在一些问题，主要集中在多模态融合、硬件适配、数据质量以及安全可靠四大方面[57]：1) 多模态融合层面，目前的方法对于不同模态数据的对齐精度停留在任务级，尚未实现像素级动作的精准关联。这导致语言的模糊表述难以转化为定量的动作参数，且基于静态数据训练的模型在动态场景中鲁棒性不足，在人流密集等复杂环境下的任务成功率显著下降[47]。2) 硬件适配方面，高性能感知依赖激光雷达、高分辨率触觉传感器等高价硬件，边缘设备难以支撑 Point-Bind、GPT4Point 等大模型的推理需求。轻量化模型难以在低算力成本下兼顾高性能，现有方案在处理精细任务时仍存在明显的精度折损[58]。3) 数据质量层面，高质量多模态配对数据稀缺，复杂动态场景覆盖不足、标注质量参差不齐且成本高。工业与户外场景样本占比低，跨平台数据复用困难，不同机器人的机械结构差异导致动作表示无法统一，缺乏标准化的数据集构建规范。4) 模型安全方面，物理交互安全机制薄弱，模型缺乏危险动作预判能力，传感器故障时无有效的容错策略，易引发设备损坏或人员受伤。恶意指令拦截不足，且工业场景依赖物理限位来规避风险的方式与通用化目标矛盾。LLM 的语言幻觉与 3D 感知的空间偏差叠加，使动作生成的逻辑可解释性差，人类信任度不足。

表1 智能感知方法

Tab.1 Intelligent perception methods

研究方向	方法名称	主要场景	方法特点	关键性能指标
大模型赋能的多模态感知融合	Sparse fusedense	机器人环境建模	直接或间接构建高精度的 3 维环境模型，核心在于解决机器人对 3 维几何信息的精确感知与物理建模	在 Car(Easy)类别上 3D 检测精度（AP）达到 89.17%
	EMHI	人机协同交互	利用丰富的视觉纹理与语义信息，实现高效的目标定位与多传感器数据在统一空间下的编码	在自然语言指令下目标定位的准确率提升至 85% 以上。
	RPEFlow	机器人灵巧操作	针对视觉、触觉、听觉等不同模态的物理维度特性，设计独立的特征提取逻辑后进行协同融合	在光照剧烈变化场景下，抓取成功率保持在 92%。
视觉语言动作解码与表征	RT-1	居家服务/操控	构建从感知到决策转化的底层数学模型与推理引擎	单模型覆盖 130k+不同任务变体
	Diffusion VLA	长程任务执行	构建跨平台标准化的大规模数据集、设计零样本/少样本的任务迁移策略以及融合人类视频等多源数据	零样本新任务成功率 56%
	ConRFT	精细化操控	将 VLA 模型的能力定制转化到具体高难度或高风险的实际作业场景中	动作抖动率：轨迹平滑度提升，末端执行器抖动幅度降低 35%
Sim2Real 下感知差异与泛化	Blenderproc	仿真场景重建	在仿真端搭建出接近真实或涵盖真实变异的数据，从源头消除 Sim 与 Real 的分布差异	真实场景检测 mAP 65%
	TL-SRND	跨域物体检测	在训练过程中通过对抗对齐或多尺度扰动，强制模型忽略仿真与现实的非本质差异	域不变特征判别准确率 >85%
	Task2sim	动力学参数辨识	采用微调、系统辨识或元学习技术，利用少量的真实物理世界数据对预训练模型或仿真参数进行校准	仿真与真实轨迹误差降低 50%
语言-空间联合主动推理	GPro3D	辅助驾驶/室外感知	利用成熟的 2D 视觉-语言大模型能力，通过深度估计或多视几何辅助，解决语言到空间的关联	在自动驾驶语义感知任务中，3D 目标召回率提升 10%
	Pointclip	3 维场景理解	通过直接对齐、分布对齐或任务定制对齐，建立语言与 3 维点云几何元素的直接映射	3D 语言定位 mIoU 提升 8%～12%
	Point-Bind	多模态语义对齐	采用紧耦合或松耦合策略，构建统一的跨模态嵌入空间	跨模态嵌入空间余弦相似度一致性 >0.9

2 智能导航（Intelligent navigation）

2.1 智能导航的核心目标与挑战

21 世纪以来，基于深度强化学习的智能导航方法由于其不依赖预建环境模型的特性，解决了居家服务、智能制造、水域勘探等动态场景下的诸多问题[59]-[61]。深度强化学习能够通过与环境交互而学习，被认为是智能导航技术的重要奠基方法，最早在 Atari 游戏[62]、围棋[63]等领域展现了超越人类的决策与博弈能力。但基于强化学习的导航决策方法逐渐暴露出场景泛化、虚拟到现实迁移困难等问题，严重限制了其在实际场景中的应用[64]。与此同时，随着多模态大模型技术迎来了迅速发展且各类导航任务的数据集不断丰富[65]-[67]，导航的核心目标逐步从路径规划、动作选择等单一决策问题，转变为依据模糊决策目标进行上层目标拆分与下层动作执行、环境约束等多维信息融合的复杂决策问题[68]-[70]。研究者逐渐探索将具备海量基础常识的多模态基底模型融入机器人决策系统，以提升系统的交互性与泛化性[71]-[73]。如 Dorbala 等[75]则将 LLM 模型应用于移动机器人的零样本导航任务，实现了无预训练、无先验地图条件下未知环境中的自主目标搜索。与传统点到点导航相比，在缺乏精确几何建模的条件下，智能导航利用视觉、语言等非结构化信息开展语义驱动的探索、避障与识别[76]。

基于大语言模型的智能导航方法研究在跨模态语义融合、任务层级分解、用户意图理解及实时推理等关键技术方向已取得突破性进展，为构建通用化、自主化智能体提供了创新性理论框架与技术实现路径。然而，若要实现具备高可靠性与强环境适应性的决策系统，仍需攻克一系列核心科学问题与工程技术难题。

1) 智能机器人在感知阶段采集的数据普遍呈现高维性、异构性和不确定性等特征。如何实现实时感知数据与先验知识的高效融合，构建精准且具备动态更新能力的环境认知模型，并据此快速生成可靠决策，避免因感知-决策链路失耦引发的认知偏差与决策延迟，仍是当前亟待攻克的关键科学问题。

2) 多模态大模型初始训练目标并非针对机器人决策任务，其通用性优势难以充分适配不同导航任务的特异性需求，在真实环境部署中决策稳定性与可靠性欠佳。如何在维持导航决策模型通用性与泛化能力，提升其在特定作业场景下的专用性与任务适配能力，已成为导航决策领域的主要技术瓶颈之一。

3) 决策模型在实际场景中易受环境噪声、感知误差

及数据扰动等因素干扰，产生偏离预期的决策输出，在长时序任务中会形成风险累积，对系统安全性与任务连续性构成威胁。如何构建含自检测与自校正功能的决策反馈机制进行错误识别、动态修正与优化，是提升导航决策可靠性与安全性的核心挑战。

2.2 大模型驱动的智能导航关键技术方法

在导航决策前，智能体通常需要融合多模态信息，对视觉、语言及环境状态数据进行协同处理和任务拆分。决策过程中，通过大模型的常识推理能力构建对感知信息的基础语义理解，通过因果建模分析动作与观测间的逻辑关联，确保决策序列的连贯性与合理性。同时，为保证导航过程的可控性与稳定性，需根据实时传感数据对导航过程进行误差评估，并进行实时的反馈纠错[77]-[78]。针对上述环节，当前主流的智能导航方法主要设计了基于思维链的任务拆分、融合常识的决策推理与基于闭环反馈的决策纠错 3 类关键技术，本章将对这 3 类关键技术进行详细介绍。

2.2.1 基于思维链的任务拆分

当前导航决策规划领域的核心挑战在于，如何借助通用大模型实现人类非结构化自然语言指令到机器人可执行程序语言的有效映射，即机器人通过环境感知信息对人类自然语言指令进行语义解析，明确指令中的决策目标，并通过逻辑推理生成实现该目标的步骤序列[80]。2022 年，谷歌研究团队[79]提出了思维链（CoT）方法，该方法提出的示例引导链式思考机制为大模型驱动的任务分解提供了理论基础与方法论支撑。根据相关研究，该方法在提示工程中通过融入问题解构实现了子任务拆解及验证优化，有效激发了大模型的时序推理能力，进而解决了零样本导航场景下因任务复杂性导致大模型输出错误动作指令的问题。CoT 方法通过结构化语言分析，将全局自然语言描述的长程导航任务，转化为一系列连续且可执行的子目标序列。该方法为实现指令、语义与动作的跨模态对齐，建立了自然语言指令与空间语义的深度映射关系，并对导航指令中关键实体、空间拓扑关系及动作逻辑进行结构化解析，最终构建了符合机器人认知范式的形式化思维链标签体系。

基于上述理论框架，诸多研究者针对智能体导航决策问题开展了系统性研究，并取得了一系列阶段性成果，以 NavGPT 模型[80]为例，该模型将显式推理作为思维链的核心载体，以 GPT 模型为核心支撑实现导航任务的分步拆解。在导航执行过程中，模型并非直接生成并输出动作指令，而是通过思维链先完成“指令解析—子目标分解—环境匹配—动作决策”的推理过程。该显式分步推理机制本质上是思维链在导航决策场景中的直接应用，既充分利用了 LLM 内置的海量常识知识，又有效规避了传统 VLM 模型因采用隐式推理模式而产生的决策序列不连贯问题，使每一步动作输出均具备明确的逻辑支撑。后续，Zhou 等[81]进一步提出改进版本 NavGPT-2，该模型将思维链与策略网络相融合，构建了推理指导动作、动作反馈推理的闭环。在训练阶段，NavGPT-2 通过 GPT-4V 生成海量导航思维链示例数据，用于引导 VLM 形成“观测-推理-动作”的逻辑范式。思维链在此类模型中的融入，既保留了大模型的可解释性优势，又弥补了传统策略网络端到端决策的固有缺陷，尤其在复杂环境中，推理链可辅助策略网络规避错误决策方向，显著提升动作选择的精准度。

为进一步提升思维链在导航场景中的决策精度与鲁棒性，研究者们进一步提出多维度优化策略。在语义解析维度，通过融合环境感知语义信息，实现对含环境语义地标的导航指令的精准解析。例如，C-Instructor[82]提出带地标的思维链（CoTL）方法，将思维链机制应用于导航指令生成场景，有效解决传统导航指令存在的步骤连贯性不足、关键地标缺失等问题。在空间约束维度，依托大规模 3 维场景数据集构建空间先验知识图谱，为子目标拆分提供地理逻辑支撑，典型如 GVNav[83]通过引入 HM3D 数据集的场景连接图，在子目标规划阶段提前规避空间逻辑冲突，保障规划合理性。在环境动态不确定性维度，通过引入动态调整机制以自适应方式调控任务拆分粒度，当环境感知信息模糊时采用粗粒度拆分以确保决策效率，当环境信息充分时切换至细粒度拆分以提升决策精度。以 EvolveNav[84]为例，其创新性设计动态思维链生成机制，摒弃对预设子目标模板的依赖，基于实时感知的环境信息动态调控拆分粒度，该模式从根本上解决了固定子目标设定与实时环境状态脱节的核心问题。

思维链方法在导航领域相关研究中的应用场景虽存在差异，但其核心思想存在共通性。基于思维链的任务拆分方法均通过对复杂任务进行降维处理，将导航过程涉及的语言理解、视觉匹配、动作规划等多维度需求，转化为线性化、可分步验证的推理序列[85]。该方法不仅充分发挥了大模型在常识储备与逻辑推理方面的核心优势，还能为底层动作执行模块提供明确的语义导向，同时显著提升了导航系统的可解释性与鲁棒性，构成了大模型与导航任务之间的关键衔接载体。

2.2.2 融合常识的决策推理

在当前决策过程中，导航方法的一大瓶颈在于模型训练过程中难以捕捉不同任务间的逻辑共同点，只是通过暴力搜索的方式求解复杂问题，这严重影响了方法的泛化性。而融合常识的决策推理方法通过融入人类对物理世界、空间关系及任务逻辑的先验认知，填补了机器人在感知受限、场景迁移过程中的信息缺口，进而降低了模型对特定场景标注数据的依赖度，提升了决策系统的通用性与推理合理性[86]-[87]。大模型的常识推理能力是其区别于传统机器学习、神经网络方法的核心优势，该能力通过激活预训练阶段所习得的人类常识、环境规律、物体关联及社会规范等隐性知识，为导航决策提供非显式指令约束与场景适配逻辑，有效缓解传统导航模型因依赖标注数据而难以适配未知场景或模糊指令的核心痛点[88]-[89]。具体而言，大模型在接收导航指令与环境观测信息后，可从预训练知识库中激活与任务相关的常识要素（如物体空间分布规律、社会交互规范等），随后将常识信息与当前场景观测数据、指令目标进行多维度关联匹配，筛选出符合常识约束的候选动作集，并以常识为核心依据优化决策动作的优先级排序，剔除违背常识的错误决策选项，最终实现导航决策效率与准确性的双重提升[90]-[91]。

按常识类型维度划分，导航决策所需的常识可归纳为 3 大核心类别，各类常识分别对应解决特定场景下的决策瓶颈问题。第 1 类为空间布局常识，即人类对不同场景下物体空间分布模式的固有认知，该类常识可有效弥补机器人局部感知的视野局限。例如，中国科学院大学 2024 年提出的 CoNav 框架[92]，通过接入 SceneVerse 大规模空间常识库，构建“场景-物体-位置”三元关联图谱。当机器人接收到导航目标后，会基于该图谱优先搜索目标物体高频出现的区域，而非采用传统的遍历式搜索，该类空间常识的融入使目标搜索效率提升 35%以上。第 2 类为物理交互常识，即描述物体物理属性及交互规律的认知体系，其核心作用在于降低物理环境不确定性对决策的干扰。以 RT-X 模型为例，该模型在原有 100 余万条机器人轨迹预训练的基础上，新增 20 万条标注有物理交互常识的轨迹，涵盖物体抓取、放置、碰撞规避等物理交互过程中的细节规则。针对仿真到现实（Sim2Real）迁移中的物理参数差异，模型通过常识模糊匹配机制，摒弃对固定摩擦系数等定量参数的依赖，转而采用人类常识中的定性规则，使真实场景决策成功率较未融合常识的基础版 RT-X 提升 23%，有效缩小了仿真与现实间的物理域鸿沟。第 3 类为任务逻辑常识，即人类完成复杂任务时的步骤性规律与逻辑约束，该类常识可优化长程任务的决策序列排序，规避逻辑错误导致的任务失效。斯坦福大学于 2024 年提出的 TaskCommonsense[93]模型，通过解析 15 万条人类日常任务记录，提炼出“前置条件-核心动作-后置处理”的标准化任务逻辑链，并将其编码为可解释的决策规则；在长程导航任务中，该模型可依据任务逻辑常识优化决策步骤排序，使任务执行失效率降低 18%，显著提升长程任务的逻辑连贯性。

从常识与决策的融合路径来看，当前主流技术范式可划分为预训练注入与实时推理调用两类，且二者常采用协同方式，以实现效率与灵活性的平衡。预训练注入模式通过在模型预训练阶段嵌入常识数据，使模型固化基础常识认知能力。例如，美国乔治梅森大学研究团队提出的 VLM-Social-Nav 方法[94]，面向人机交互的社会性导航场景，将 GPT-4V 作为社会常识的核心承载模块，引导机器人遵循靠右通行等隐性社会规范。同时，通过大模型识别人类行为、手势等非语言信号推断行人意图，依据意图判断结果降低人机交互冲突。实时推理调用模式则更适配场景特异性较强的常识需求，其核心机制为构建轻量化常识知识库，机器人在实时感知场景标签后，通过语义匹配完成相关常识的快速检索与调用。例如，斯坦福大学研究团队[95]针对动态场景规划问题提出的动态推理机制，即通过该模式实现基于实时反馈的环境信息检索适配常识，进而动态调整子目标与动作决策逻辑。

大模型的常识推理能力在导航决策场景中扮演隐性知识引擎的核心角色，既填补显式指令未涵盖的场景规则空白，又破解传统数据驱动型导航模型的固有局限。常识推理推动导航决策从被动响应式决策向主动引导式决策跃迁，最终达成决策效率、执行准确性与人类兼容性的协同优化，这亦是智能导航决策区别于传统方法的核心优势。值得关注的是，融合常识的决策推理仍面临常识泛化性与场景动态性两大关键挑战：前者表现为不同地域、文化语境下的常识异质性，致使单一常识库在跨地域场景中的适配性能衰减；后者源于动态场景中常识的时效性演化，以典型家居场景为例，沙发的常规功能为坐卧，但其实际使用中常被临时用于堆放杂物，而固化的常识知识易弱化“杂物”与“沙发”间的关联属性，进而导致机器人无法精准定位沙发上的目标物品。针对上述挑战，未来可通过采集机器人实时交互数据，实现特定场景下常识条目的自动更新及关联规则权重的动态调整，为长期服役的服务机器人提供常识知识迭代的可行路径。

2.2.3 基于闭环反馈的决策纠错

基于闭环反馈的决策纠错方法聚焦构建感知-决策-评估-反馈的动态循环机制。通过实时捕获导航过程中的不确定性干扰，量化决策偏差程度并生成针对性修正策略，可有效规避传统开环导航系统因决策失效导致的任务中断问题，提升机器人在复杂动态环境中的导航鲁棒性[96]-[97]。大模型初始训练目标并非面向智能导航决策任务，其输出结果在实际部署中存在固有不确定性。同时，环境噪声、传感器精度局限等因素易引发感知误判，双重影响下决策准确性难以保障[98]。针对此核心问题，研究者提出导航决策场景下的实时纠错方案：决策系统通过实时采集环境观测数据、动作执行反馈及状态定位信息，与预设任务目标进行多维度偏差比对，一旦检测到偏差超出阈值，立即触发决策调整流程，形成闭环决策链路[99]。该技术路径可有效应对导航场景中的环境动态性、执行误差及约束违反等问题，通过抑制错误累积效应，保障导航任务的可行性、准确性与鲁棒性。

偏差评估机制是闭环反馈系统的核心环节，其评估精度与实时性是该领域的研究关键。当前主流评估方法通过构建量化偏差指标体系与动态评估窗口的协同机制实现这一目标[100]-[102]。针对大模型通用性输出与特定场景约束难以匹配的核心问题，Paul 等[103]提出辅助驾驶模型 LeGo-Drive，该模型基于实时预测的目标位置，引入环境约束与车辆动力学约束，通过对导航决策输出结果的实时监测与评估实现决策反馈，有效解决了视觉语言模型因忽略车辆运动学特性及场景约束而产生错误决策、导致任务失败的问题。针对导航决策过程中动作执行误差累积的难题，Brown 等[104]提出端到端视觉-语言-动作模型 RT-2，该模型接收视觉观测信息与语言指令后直接输出动作参数，同时通过视觉传感器实时观测动作执行效果并转化为反馈信号，用于微调模型的动作预测模块，抑制误差累积。Sathyamoorthy 等[105]在 CoNVOI 方法中提出基于上下文一致性与路径反馈的双重实时纠错机制，该机制在实时监测路径可行性与目标可达性的基础上，构建上下文一致性路径检测模块，通过判定动作执行过程与子目标上下文的匹配度，规避决策偏移问题。

闭环反馈实时纠错技术的具体实现方案因应用场景的差异化需求而呈现多样性，但均通过实时反馈机制通过闭环链路将导航场景中的环境不确定性、动作执行误差及约束违反等问题转化为可量化的决策调整信号，从而抑制错误累积效应并实现对动态场景的自适应适

配[106]-[107]。该技术路径不仅显著提升了智能体导航决策的准确性与鲁棒性，更为大模型与导航决策系统的深度融合提供了核心技术支撑，为提升系统整体性能奠定了重要基础。

2.3 小结

现有的基于大模型导航研究多面向室内导航场景，以导航成功率与路径效率作为核心指标，借助大模型提升目标寻找、环境探索、动态避障等任务的表现，如表2 所示。基于大模型的智能导航决策系统在少样本及零样本任务中表现显著，但当前研究仍面临三大核心瓶颈：1） 现有方法多采用提示工程技术简化决策任务，通过约束模型输出空间以获取可用结果，但其本质仍是将决策与感知过程解耦——模型仅能依托内置常识开展决策推理，难以根据具体任务约束与场景条件进行动态适配。因此，如何针对特定导航决策任务设计专用模型架构，或通过定向微调等方式构建决策输出与感知输入间的物理关联，进而提升模型任务适配能力，已成为该领域亟待突破的核心难题。2） 大模型庞大的参数量及输出优化过程导致输出延迟较高，难以满足决策任务的实时性需求。当前主流解决方案为通过降低决策频率提升实时性，但该策略会弱化系统对动态场景的响应能力，加剧决策滞后风险。如何平衡模型推理实时性与决策安全性间的矛盾，是制约导航决策技术落地应用的关键瓶颈。3）实时纠错机制的工程化实现面临显著技术挑战，而在动作执行前完成先验性错误检测与修正，对提升导航决策技术的落地可行性具有重要现实意义。未来可通过融合大模型领域的前沿技术（如知识蒸馏、常识知识图谱构建等）优化导航决策模型，或采用决策流程的深度解耦策略，为解决多目标优化冲突问题提供可行技术路径。

表2 智能导航方法

Tab2. Intelligent navigation methods

研究方向	方法名称	主要场景	方法特点	关键技术指标
基于思维链的任务拆分	PaLM-E[89]	居家服务	通过端到端训练实现单一模型同时支持机器人多类具身任务	训练后 12B 模型下任务成功率>50%
	SayNav[95]	室内探索导航	以增量构建的 3D 场景图作为 LLM 的物理接地载体，动态生成分步含条件语句与备用选项的多目标导航计划	GPT-4 模型下导航成功率为 61.6%
	NavGPT[80]	室内 VLN	首次在视觉-语言导航中实现了全流程可解释的 LLM 原生显式推理，并揭示其零样本场景下的规划能力	MP3D 数据集，无训练条件下成功率达 34%
融合常识的决策推理	E^2 BA	室内环境探索	设计含双级联触发机制的回溯判别器动态优化路径以减少无效探索	零样本、GPT2-large 模型下导航成功率达 70.2%
	VLM-Social-Nav[94]	人机交互场景	提出社交合规性作为导航优化目标，提高了人机交互的安全性与规范性	GPT-4 模型下碰撞概率为 4.76%
	ESC	室内环境探索	创新性地提出概率软逻辑（PSL）将常识转化为软约束融入前沿探索，全程无需任何导航数据训练	MP3D 数据集，无训练条件下成功率达 28.7%
	VLFM	室内环境探索	通过 VLM 直接从图像中提取语义价值生成含置信度的双通道价值地图，以此选择高价值前沿高效探索	MP3D 数据集，无训练条件下成功率达 36.4%
	OVER-NAV	室内环境探索	引入 Omnigraph 结构化表示整合多模态导航数据，通过专属融合机制提取关键知识优化决策	IR2R 数据集，无训练条件下成功率达 65%
基于闭环反馈的决策纠错	LM-Nav	室外导航	通过整合预训练的 LLM、VLM 和 VNM，实现了机器人在复杂户外环境中依据自然语言指令完成长距离导航	室外场景下导航成功率达 80%
	LeGo-Drive[103]	辅助驾驶	首创联合端到端训练机制，将初始非可导航目标修正为实际可达位置并生成平滑无碰撞路径	室外场景下导航成功率达 81.2%
	MapNav	室内 VLN	实时构建并动态更新显式语义地图，使 VLM 能精准理解空间语义并实现端到端导航	R2R 数据集，无训练条件下成功率达 39.7%

3 智能操作（Intelligent manipulation）

3.1 智能操作的核心目标与挑战

智能操作是指智能机器人基于对环境的感知与理解，通过末端执行器主动改变物理世界状态的过程。这一概念已从早期的单一“抓取”延展至推、拉、旋转、装配及工具使用等复杂交互行为。其核心目标是赋予机器人在非结构化、开放场景中类人级的通用操作能力：既能理解抽象的自然语言指令（如“把那个红色的杯子收起来”），又能处理接触过程中的精细物理反馈[108]-[109]。

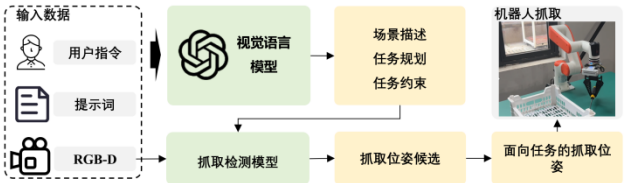

图 4 基于视觉-语言的操作姿态预测与策略生成

Fig4. Visual-language-based prediction of operational postures and generation of strategies

纵观其发展脉络，智能操作正经历从“几何驱动”向“语义驱动”，再向“物理与常识驱动”的范式转变。

早期系统多依赖精确的几何建模与预编程逻辑，虽在工业场景表现稳定，但在面对未见物体或柔性形变物体时，常因缺乏语义泛化能力而失效。随着大模型技术的引入，当前的研究重点转向利用LLMs和MLLMs的推理能力，解决“意图解析—可供性发现—动作生成”的全链路协同问题[110]。然而，要实现真正通用的智能操作，当前技术仍面临3大核心挑战：

1) 开放语义与操作动作的精准映射挑战：如何将抽象、模糊的自然语言指令转化为机器人可执行的、满足几何约束的6维位姿与力控参数。这要求系统不仅能识别物体，还能理解物体的功能部件（如杯柄、开关）及其对应的操作方式。

2) 接触丰富任务中的物理适应性挑战：在涉及摩擦、形变和多点接触的精细操作（如插拔、擦拭）中，单纯依靠视觉感知难以获取准确的物理参数。如何在缺乏精确物理模型的情况下，利用多模态感知（视觉-触觉）与物理常识进行自适应调整，是提升操作鲁棒性的关键。

3) 动态非平稳环境下的闭环控制挑战：真实环境充满了动态变化（如物体滑动、人类干扰）。传统的“规划-执行”开环模式难以应对执行误差，系统需要具备高频的实时感知与在线纠错能力，在毫秒级时间内完成从感知到动作的闭环映射。

3.2 大模型驱动的智能操作关键技术方法

3.2.1 基于视觉-语言的操作姿态预测与策略生成

基于视觉-语言的操作姿态预测与策略生成主要解决智能操作中“如何理解任务”与“如何规划动作”的问题，核心在于利用大模型的语义对齐能力，弥补传统几何方法在语义缺失和泛化性上的不足。如图4所示的典型框架中，系统首先将接收的RGB-D视觉数据与用户的自然语言指令作为多模态输入。在处理流程上，视觉语言模型（VLM）不仅仅进行被动的语义理解，而是主动生成场景描述并进行层级化的任务规划，从而明确具体的任务约束。与此同时，抓取检测模型并行工作，基于几何信息生成抓取姿态候选。最终，该框架通过核心的解码核心将高层的语义规划与底层的几何检测进行跨模态对齐，筛选出符合意图的“面向任务的抓取位姿”，从而驱动机器人完成精准抓取。围绕这一架构，目前的研究主要围绕以下几个路径展开。

跨模态语义对齐与可供性发现：传统抓取检测算法多基于几何形态（如平面、边缘）生成候选框，无法理解物体部件的功能属性。大模型驱动的方法通过大规模图文预训练，实现了视觉特征与语言语义的深度对齐。例如，Point-Bind与PointLLM等工作通过将3维点云特征投影至大语言模型的嵌入空间，使机器人具备了细粒度的语义理解能力[111]。当接收到“握住马克杯把手递给我”的指令时，模型不再仅仅识别“马克杯”整体，而是能结合几何与语义信息，精准推理出“把手”这一特定区域的可供性，从而生成符合人类操作习惯的姿态。这种能力显著提升了机器人对异形物体和零样本场景的适应性，但目前仍面临计算开销大、复杂场景下跨域对齐精度衰减以及端侧部署难等问题。

语言驱动的视觉定位与动作建模：该路径致力于利用语义信息从复杂场景中锁定目标，并将操作姿态预测与后续动作序列生成进行一体化建模。系统首先将自然语言描述作为查询条件，在视觉感知数据中检索符合语义要求的物体实例[112]。随后通过构建视觉-语言-动作联合模型，融入任务导向逻辑。例如，OVGNet[113]和OVMR[114]等模型以开放词汇建模为核心，在家庭服务等存在大量未知物体的环境中具有较高的操作成功率。然而，该方法高度依赖初始语义定位的准确性，一旦识别出现偏差，后续操作规划将失效；且在多目标外观相似或高动态交互任务中，策略的一致性仍需加强。

模型轻量化与实时性优化：为满足动态场景的快速响应与端侧部署需求，该路径致力于通过模型蒸馏、参数剪枝及低维表征学习等技术，在保证预测精度的前提下实现低延迟推断[115]-[117]。虽然这类方法能够有效提升“感知-决策”的协同效率，但通常依赖大规模、高质量的示范数据来构建动作先验，且分段式处理流程容易受到特征传播误差的影响。

3.2.2 基于物理常识与多模态触觉的精细交互控制

基于物理常识与多模态触觉的精细交互控制旨在将力平衡、力矩守恒等物理规律与机器人的感知、决策模块深度耦合，从而在非结构化、多样化环境中实现抓取力的精准分配与动态调节。这一过程的具体逻辑如0所示：首先在“物理量信息获取”阶段，系统利用多模态模型并行提取接触状态、摩擦系数等特征；随后进入“物理属性估计”模块，通过对比特征信息处理，对物体的物理属性概率（如软硬度、滑移趋势）进行推断；最后在“闭环映射”环节，结合物理规则与安全边界，将推断结果转化为实时的控制指令，并根据反馈动态调整策略。这种架构主要解决智能操作中“如何接触”与“如何用力”的问题。针对视觉感知在接触瞬间存在的遮挡与深度误差问题，则通过引入物理常识与触觉反馈，构建精细的力位混合控制模型。

以高分辨率触觉传感器为核心感知单元的力控制方法，通过捕获接触面的力/压力分布、切向力及局部滑动特征，结合传感器硬件与控制算法的协同优化，构建从触觉信号到物体物理属性、抓取力指令的映射机制。该方法受限于触觉传感器对环境干扰的敏感性、复杂触觉信号的解析难度及高昂的计算成本，在物体形态极端不规则或接触区域存在遮挡时，传感器的感知覆盖范围易存在不足，进而影响力控制精度。

针对软质、不规则形态物体的抓取需求，基于软抓手与多触点设计的力控制路线，通过优化执行器结构与感知集成方案提升适配性。Zhou等[117]研发了集成传感单元的三指软机器爪，结合深度学习完成物体识别与闭环力控策略构建；Xu等[119]模仿人类手指多模态感知机制，整合力、温度、滑动传感数据动态调整抓取力分布；Deng等[120]则基于力矩平衡原理优化多触点作用力分配以保障抓取稳定性。该方法存在软抓手耐用性不足等问题，在高强度抓取任务场景中的应用仍存在一定局限。

为应对物体尺度跨度大、物理特征难以完整观测的技术瓶颈，多尺度与多物理特征建模方向通过融合全局-局部几何特征及多维度物理量构建综合建模框架[121]-

[124]，在异质材料并存的复杂场景中表现出更优的力控制性能，其力控准确率较传统单一特征方法提升约5.8%-16.83%[125]，但模型复杂度的增加也导致了计算资源需求的显著上升。此类方法通常采用多尺度卷积网络提取不同层级的几何特征，结合注意力机制强化关键区域感知，或设计专门的物理量回归分支实现对物体刚度、质量等属性的精准估计，从而提升多尺度物体抓取时的力控制鲁棒性。但该类方法的模型结构复杂度较高，难以在高速动态抓取任务中维持实时性，限制了其在时间敏感场景中的应用。

为突破对已知物体的依赖、提升对未知物体的泛化能力与物理推理质量，大模型与物理常识耦合的研究尝试将归纳性物理常识嵌入视觉-触觉-动作跨模态大规模模型。已有工作将 Transformer 架构应用于抓取姿态与力预测任务，通过多模态数据融合将模型决策意图转化为物理可执行的控制信号；另有研究通过微调 VLM 模型，使其能够根据物体的视觉材质推理出相应的物理特性，进而生成自适应的抓取力度预设值。例如，在“倒水”任务中，模型能根据容器的材质与液面高度，动态调整手臂的加速度与倾斜角度。这种物理常识的注入，有效解决了传统方法在未知物体动力学参数辨识上的冷启动难题，提升了柔性物体操作的成功率[126-128]。另一方面，构建视触融合的大规模模型，如 Tactile-VLA[129]等模型将高分辨率触觉传感器的信号转化为图像或序列化 Token，与视觉、语言特征共同输入 Transformer 架构进行多模态联合训练。该类模型能够实时感知接触面的滑动、受力分布及纹理变化。该类模型能够实时感知接触面的滑动、受力分布及纹理变化。一旦检测到打滑趋势或异常接触力，模型能利用其推理能力快速触发调整策略，形成了“感知-推理-修正”的微秒级触觉闭环，极大增强了操作的鲁棒性。其优势在于强化了多模态信息协同能力，同时降低了对大规模标注数据的依赖，但受制于大模型固有的高计算开销、端侧设备资源适配难度，且在多物体复杂碰撞、物体非线性变形等场景中，力控制精度易出现衰减。

依托数据驱动的力控制优化方法，通过大规模标注数据的离线学习与在线反馈，实现对力控制策略的迭代优化，具体包括基于触觉反馈的自适应力学习方法、无需人工标注的 “轻柔抓取” 策略演示学习方案，以及将预测模型与触觉/视觉反馈相结合的模型预测控制技术[130-131]。该类方法能快速适配动态场景的变化，在实时性表现上具有显著优势，但对高质量、多样化训练数据的依赖性较强；当物体物理属性处于训练数据分布之外或训练样本稀缺时力控制性能易出现下滑，难以保证抓取的稳定性。

3.2.3 动态场景下的端到端决策与自适应闭环控制

动态场景下的端到端决策与自适应闭环控制旨在使机器人能够对运动目标或变化环境作出快速、可靠的抓取响应。其核心是构建一个高效的“感知-规划-控制”闭环，解决智能操作中“如何适应变化”与“如何抗扰动”的问题，致力于构建高频、实时的感知-执行闭环，以应对动态环境下的不确定性。当前，该领域面临多重挑战：动态目标的实时感知易受遮挡与运动模糊影响；抓取规划与执行控制的协同性难以匹配高速机器人的动态需求；面对多样化的未知运动模式，决策策略的泛化能力不足；此外，抓取过程中的滑动、惯性等物理干扰也增加了控制精度保障的难度。这些挑战共同制约了动态场景下抓取系统的稳定性与可靠性。

作为“感知-规划-控制”闭环的基础环节，动态目标的视觉感知与定位依托视觉感知算法优化与多传感技术融合，实现动态目标运动状态的实时捕获与精准定位，其核心通过改进目标跟踪算法的抗遮挡能力、融合多模态感知信号补偿单一视觉的信息缺失，从而解决遮挡与跟踪滞后问题。例如，改进后的 RT-DETR 目标检测算法在高速动态场景中可同时实现 97.4%的检测精度与 67 帧/秒的高帧率；协同规划策略则通过多关节运动耦合优化实现 ZMP 偏移收敛与能耗降低。环境自适应搜索算法在障碍物动态变化场景中展现出更优的轨迹平滑性[132]-[134]。尽管该类方法在动态场景下表现出优异的实时性与定位精度，可有效应对堆叠遮挡导致的感知盲区问题，但受限于高帧率感知对硬件算力的高需求，其部署成本较高，且在极端速度场景下，目标运动轨迹的非线性特征易引发预估偏差，影响力定位的准确性，难以完全满足高速动态抓取的感知需求。

基于动态目标感知结果的轨迹规划与协同控制，主要解决运动过程中的机械臂晃动、控制延时及高速抓取适配性问题[135-137]。该技术方向通过两种核心路径实现优化：一是基于动态目标运动状态的实时感知结果，采用迭代优化的在线建模方法构建目标运动预测模型，结合“move-and-grasp”协同策略实现抓取规划与机械臂运动控制的动态适配；二是通过费拉里法求解机械臂运动学逆解并引入动态插补算法，或设计恒定时间规划策略，分别满足高速抓取的轨迹精度需求与传送带等连续动态场景的抓取时效性需求。然而，该类方案在复杂动态场景中，预规划轨迹难以覆盖目标运动的随机性变化，导致适应性较弱；同时，多关节协同运动过程中的耦合干扰易降低轨迹跟踪精度，而平滑轨迹规划对机械臂硬件响应速度的高要求，进一步限制了其在硬件资源受限场景中的应用。

学习与大模型驱动的动态决策，通过深度学习、强化学习及大语言/视觉-语言模型的融合应用，提升抓取决策策略对复杂动态场景的适应性与泛化能力[138]。典型研究如 GraspGPT[110]、RT-Grasp[139]等，通过将开放式语义理解、视觉-语言注意力机制或跨模态推理融入抓取决策框架，使系统能够基于自然语言指令或视觉场景语义，实现对未知动态目标的零样本泛化抓取。但该技术方向存在显著瓶颈：强化学习方法对大规模仿真训练数据的强依赖，导致其在真实动态场景中迁移时易产生域适应偏差，难以精准匹配真实物理环境的特性；大模型复杂的推理过程导致决策延迟较高，无法满足极端动态场景对毫秒级决策响应的需求，限制了其在时间敏感型任务中的应用。多模态融合与物理自适应控制作为“感知-规划-控制”闭环的最终执行保障，通过将视觉动态感知信息、力觉接触信号、触觉滑差感知数据及物体物理参数估计结果进行多模态时空融合，构建物理干扰自

适应补偿机制，从而实现对抓取过程中滑差、惯性误差及接触非线性等物理现象的主动补偿与鲁棒控制[138-143]。该类方法能够显著提升动态场景下抓取控制的稳定性，尤其在目标运动状态突变场景中表现出较强的适应性，但多模态数据来源的异质性导致时空同步精度难以保障，易引发控制指令的时序偏差；同时，多源信息中的冗余特征处理会增加计算开销，可能影响控制闭环的实时性，而物体物理参数推断过程易受感知噪声干扰，进一步降低自适应补偿的精度。

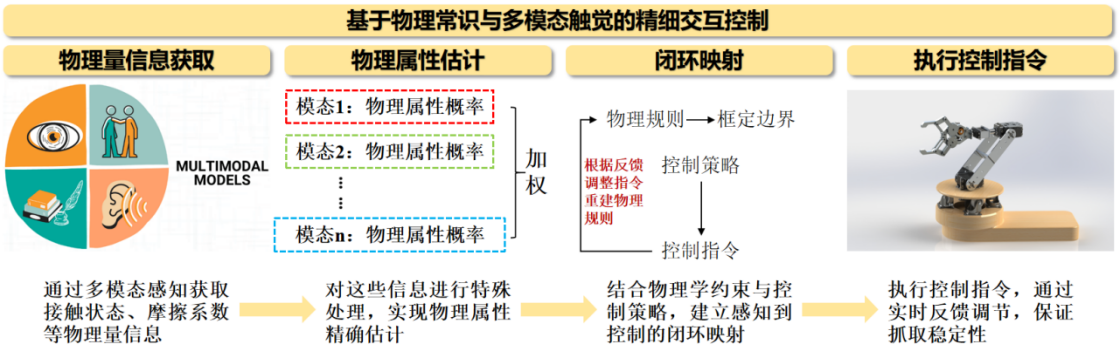

图 5 基于物理常识与多模态触觉的精细交互控制

Fig5. Fine interaction control based on physical principles and multimodal tactile sensations

表3 智能操作方法

Tab3. Intelligent manipulation methods

研究方向	方法名称	主要场景	方法特点	性能指标
基于视觉-语言的抓取姿态预测	开放词汇建模与零样本学习（如OVGNet）	开放场景未知类别物体抓取	摒弃了传统的固定类别训练模式，利用大模型的语义泛化能力，实现对训练集中未见过的物体或抽象语言指令的直接理解与操作	100 个新类别物体测试中抓取成功率 **64.4%**；颜色与语言组合泛化任务完成率超 **80%**
	3D 视觉与大模型深度融合（如QwenGrasp）	工业零件精密装配复杂物料分拣	针对工业场景对精度的极高要求，将大模型的语义理解与 3D 点云/深度图的几何特征深度对齐，解决了传统视觉在大批量、多姿态零件识别中的精度瓶颈	类别级 6-DOF 姿态预测精度达**毫米级**；满足工业级装配公差要求
	多模态协同姿态建模	多机器人协作人机交互抓取	在多臂协同或人机共融场景下，通过共享语义空间与坐标系，使得机器人不仅能预测自身抓取点，还能根据人类或队友的动作预测避让姿态与协作时机	显著提升多智能体协作效率；交互意图理解准确率提升
	模型轻量化与端侧部署优化	移动机器人无人机动态抓取	通过剪枝、量化及算法架构优化，解决了大模型参数量大、推理慢的问题，使其能够部署在算力受限的移动终端上，提高实时性	推理延迟大幅降低，满足**实时性**需求；适配嵌入式计算平台
基于物理常识的抓取力控制	物理常识推理大模型（如 Tactile-VLA）	未知物体泛化抓取易碎品操作	突破了传统力控依赖物理参数辨识的局限，利用 VLA 模型直接从视觉图像中推理物体的物理属性，实现基于常识的“无接触”力参数预设	零样本学习任务成功率 **75%～80%**；力控策略生成时间略有增加
	视触觉时空同步融合	杂乱桌面清理堆叠物体抓取	针对视觉在接触瞬间易受遮挡的问题，引入高频触觉反馈与视觉信号进行时空对齐，利用触觉补充视觉盲区的几何信息，大幅提升了密集堆叠场景下的抓取稳定性	杂乱场景抓取成功率提升 **15%～55%**；实现 **10～30Hz** 高频同步采样
	高精度/上下文感知力位控制（如SITT，CAGE）	工业3C部件装配多材质线缆插拔	结合上下文感知（Context-Aware）与高灵敏度传感器，能够根据操作阶段动态调整力控策略，特别适用于插拔电线、焊接移除等精细工艺	动态响应灵敏度达 **115mV/N**；复杂装配任务成功率 **96.8%～100%**
	柔顺阻抗控制与环境刚度辨识	农业果蔬采摘水下动态作业	在非结构化环境中（如波动的流体或脆弱的生物组织），通过在线辨识环境刚度并调整机械臂阻抗参数，实现了“刚柔并济”的自适应无损抓取	果蔬损伤率降至 **2%以下**；力跟踪平均偏差仅 **0.014N**
动态场景下的抓取决策与控制	语义驱动的动态决策（如GraspGPT）	家庭服务开放动态环境	将大语言模型的逻辑推理能力引入动态决策回路，使机器人能够理解模糊的自然语言指令，并在环境动态变化时，基于语义一致性自动修正抓取目标	未知动态目标抓取成功率 **80%～83.7%**；语义理解 mAP 达 **79.32**

	高速视觉检测与并联规划（如改进 RT-DETR）	工业传送带分拣高速流水线	结合了 RT-DETR 的高精度实时检测与 Delta 机器人的高动态响应能力，在 90mm/s 传送带速度下，将抓取中心平均偏差控制在 1cm 左右，解决了高速运动目标的时延补偿问题	检测精度 **97.4%**，帧率 **67FPS**；实现 **90 次/min** 高速抓取
	动态 3D 场景图长时序跟踪	人员跟踪服务动态场景交互	构建包含时间维度的 3D 场景图谱，不仅记录物体位置，还持续跟踪物体状态变化与人员轨迹，为长周期的服务任务提供历史记忆与决策依据	人员跟踪精度 **93%**；场景变化检测精度（SCDA）**94.23%**
	全身协同控制（Whole-Body Control）	四足机器人/移动底盘复杂地形作业	针对移动操作基座不稳的难题，采用强化学习（PPO）或模型预测控制，实现底盘运动与机械臂动作的全身耦合规划，在崎岖地形或移动中保持抓取末端的极高稳定性	运动中末端晃动**<0.025m**；动态抓取成功率 **0.89**

3.3 小结

依托大模型的语义理解与零样本泛化能力，现有方法不仅在工业场景的精密装配与高速分拣中实现了毫米级精度与高频响应，更在家庭等非结构化场景中，针对未知物体与模糊指令展现出了超越传统方法的适应性。这一量化对比为后续分析当前技术在极端环境下的鲁棒性瓶颈提供了数据支撑，如表 3 所示。智能操作技术虽已在多模态感知融合、物理属性自适应匹配、动态场景响应等方面取得阶段性突破，但迈向类人化自主抓取仍面临 3 大核心问题。

1) 在视觉-语言姿态预测任务中，自然语言的抽象指令难以精准映射到 3 维抓取约束，尤其在遮挡、稀疏点云及复杂背景场景下，该映射过程易失准。同时，模型对未见物体或开放词汇的泛化能力不足，导致零样本场景下的性能下降；此外，高精度推理所需的复杂流程，与端侧设备低延迟、低算力的需求形成尖锐矛盾。

2) 基于物理常识的力控制技术，在感知噪声、复杂接触几何及未知材质的干扰下，难以可靠且实时地估计摩擦系数、质量分布等关键参数。现有力映射模型难以直接迁移至千差万别的非结构化环境，而多模态传感数据融合所需的严格时空同步与高频处理，进一步增加了系统延迟，进而影响控制的实时性与鲁棒性。

3) 动态场景中，目标的运动、遮挡及光照变化，易导致感知与决策环节出现滞后。规划层与控制层在高扰动环境中的耦合效应，暴露了系统对延时的敏感性，最终造成抓取不稳定。现有方法普遍缺乏对新物体、新运动模式的快速在线适配能力，而大模型固有的推理延迟，也限制了其在需即时反应任务中的应用。

4 未来研究展望（Prospects）

4.1 大模型加持的多模态智能感知

当前大模型驱动的智能感知技术在多模态融合精度、硬件适配性、数据质量、安全可靠性及 Sim2Real 泛化能力等方面存在核心短板。未来研究应以提升不同粒度的多模态数据间的对齐问题为切入点，重点攻克复杂动态非结构化场景中的语义-空间关联、语言模糊表述向定量动作参数转化等难题，从而增强大模型对真实物理世界状态以及跨时域变化情况的有效感知与精确表达能力。除此之外，通过构建轻量化、高效能的专用大模型，优化跨硬件适配，平衡感知精度与硬件成本，提升跨平台数据复用率降低标注成本，增强动作生成的逻辑可解释性，可有效推动边缘设备规模化落地。同时，深化 Sim2Real 泛化技术研究，融合高保真重建、领域不变特征学习与少量真实数据调节方法破解仿真与现实的感知鸿沟，进而推动大模型与智能感知技术在服务机器人、工业操控、人机协同等领域的深度融合，拓展其在复杂动态场景中的应用边界，助力智能体实现更精准、高效、安全的自主感知与灵巧操作，推动智能机器人产业的高质量发展。

4.2 面向复杂非结构化场景的智能导航

当前大模型对机器人导航技术在各类复杂任务上的应用虽取得了不少突破，但还存在一些比较突出的问题，比如感知与决策过程存在一定的脱节，大模型的通用性难以适配不同导航场景的具体需求，推理速度慢导致实时性不足，实时纠错机制的工程化落地难度大，常识知识在跨场景、动态环境中适配性欠佳等等，这进而导致机器人导航与决策时依然需要依赖于预设规则作为“保底策略”，这些预设规则包括但不限于显式任务分解机制与静态导航决策规划框架，这限制了机器人完成更为动态灵活以及模糊任务指令的导航任务，因此未来的研究方向可以围绕这些痛点展开。例如，构建具备意图理解、因果推理、风险评估与策略自适应能力的导航专用大模型，以此实现对复杂任务环境的深层认知与高效响应，具体而言，可通过解析自然语言指令并将其转化为可执行的动态目标函数，在此基础上让系统在任务执行中开展多路径策略探索与实时评估，兼顾决策语义可解释性与导航、操作策略性能的持续优化，还可以尝试将任务执行中的经验以结构化形式存储于外部知识库或记忆网络以实现模型的在线动态迭代，进而将导航技术拓展到更复杂场景，真正实现通用化、自主化的智能导航。

4.3 意图理解与自我纠错的智能操作

当前大模型驱动的智能操作技术虽然已经可以让机器人实现传统方法难以做到的灵活的物体抓取与操作，但仍面临例如语义指令与 3 维抓取约束映射失准、力控技术在复杂环境中鲁棒性不足、动态场景下感知决策滞后、大模型端侧部署算力矛盾等核心问题。同时，当应对真实世界中多物体动态协同与交互的普遍需求时，机器人精准识别物体、理解其空间拓扑关系、物理属性及操作语义逻辑与因果依赖的能力显得更为重要。

未来研究可以优先考虑融合大语言模型与多模态大模型的语义理解能力，实现上下文理解解析自然语言指令中的高层语义意图，结合视觉输入与场景先验推理物体间空间关系、潜在冲突与操作依赖性，再完成抓取点位的动态路径规划与精准抵达。这期间，大模型既作为策略生成器又提供可解释推理路径，理论上可以为任务失败或环境扰动时的策略回溯与动态修正提供支撑，可以用于解决未见物体泛化、复杂接触力控及实时响应等难题，这些在鲁棒性、可解释性与环境适应性上的提升，未来不仅能广泛应用于工业自动化柔性生产、家庭服务机器人、特种作业救援等场景，还能支撑更复杂的多物体交互任务落地，推动机器人深入日常生活场景和复杂任务。

5 结论（Conclusion）

本文系统综述了大模型与智能机器人融合的智能感知、导航与操作技术最新研究进展，从感知、导航与操作 3 个核心层面，深入分析了大语言模型与多模态大模型在认知推理、任务规划与物理交互中的作用机制与技术路径。研究结果表明，大模型的引入推动了智能机器人系统由“任务驱动”向“认知驱动”的范式演进，显著提升系统在跨模态语义理解、上下文推理、动态环境适应与自主策略生成等方面的综合能力。未来大模型与智能机器人融合研究将聚焦于“认知增强-协同闭环-通用泛化”三大核心方向，持续推进系统性深化。一方面，应融合知识图谱、因果建模与自反学习机制，构建具备长期记忆、情境感知与自我修正能力的智能体架构，以增强模型在复杂语境下的理解与适应能力；另一方面，需优化模型轻量化设计、提升计算能效比，推进边缘部署与实时响应能力，构建涵盖人-机-环境协同的多层级交互体系，实现虚拟认知与物理行为的动态映射与双向反馈。

参考文献（Reference）

[1] 刘华平，郭迪，孙富春，等. 基于形态的具身智能研究:历史回顾与前沿进展[J]. 自动化学报，2023，49(6)：1131-1154.
LIU H P, GUO D, SUN F C, et al. Morphology-based embodied intelligence: Historical retrospect and research progress[J]. Acta Automatica Sinica, 2023, 49(6): 1131-1154.

[2] 兰沣卜，赵文博，朱凯，等. 基于具身智能的移动操作机器人系统发展研究[J]. 中国工程科学，2024，26(1)：139-148.
LAN F B, ZHAO W B, ZHU K, et al. Development of mobile manipulator robot system with embodied intelligence[J]. Engineering Sciences, 2024,26(1): 139-148.

[3] ZHANG H Y, LIN W M, CHEN A X. Path planning for the mobile robot: A review[J]. Symmetry (Basel), 2018, 10(10):450.

[4] W. RONE AND P. BEN-TZVI. Mapping, localization and motion planning in mobile multi-robotic systems[J]. Robotica, 2013, 31(1): 1-23.

[5] 沈甜雨，陶子锐，王亚东，等. 具身智能研究的关键问题:自主感知、行动与进化[J]. 自动化学报，2025，51(1)：43-71.
SHEN T Y, TAO Z R, WANG Y D, et al. Key problems of embodied intelligence research: autonomous perception, action, and evolution[J]. Acta Automatica Sinica, 2025, 51(1): 43-71.

[6] 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.
WANG W S, TAN N, HUANG K, et al. Embodied intelligence systems based on large models: A survey[J]. Acta Automatica Sinica, 2025, 51(1): 1-19.

[7] HUANG W L, XIA F, XIAO T, et al. Inner monologue: Embodied reasoning through planning with language models[DB/OL]. (2022-07-12) [2026-03-09]. https://arxiv.org/abs/2207.05608.

[8] EPPE M, GUMBSCH C, KERZEL M, et al. Hierarchical principles of embodied reinforcement learning: A review[DB/OL]. (2022-08-18) [2026-03-09]. https://arxiv.org/abs/2012.10147.

[9] SHE L, CHENG Y, CHAI J Y, et al. Teaching robots new actions through natural language instructions[C]//IEEE International Symposium on Robot and Human Interactive Communication. Piscataway, USA: IEEE, 2014: 868-873.

[10] TEJESH V, SUPRIYA M. Enhanced image captioning using cnn and blip models[C]//International Conference on Multi-Agent Systems for Collaborative Intelligence. Piscataway, USA: IEEE, 2025: 859-863.

[11] 刘华平，郭迪，孙富春，等. 基于形态的具身智能研究:历史回顾与前沿进展[J]. 自动化学报，2023，49(6)：1131-1154.
LIU H P, GUO D, SUN F C, ZHANG X Y, et al. Morphology based embodied intelligence: historical retrospect and research progress[J]. Acta Automatica Sinica, 2023, 49(6):1131-1154.

[12] WU H, WEN C, SHI S, et al. Virtual sparse convolution for multimodal on 3d object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023: 21653-21662.

[13] WU H, WEN C, LI W, et al. Sparse fuse dense: towards high quality 3d detection with depth completion[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 4508-4517.

[14] HONG Y, ZHEN H, CHEN P, et al. 3D-LLM: injecting the 3D World into large language models[C]//NIPS Conference on Neural Information Processing Systems. California, USA: NIPS, 2023, 36: 20482-20494.

[15] CHEN Z, CHEN Z, JIAO Z, et al. LL3DA: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE. 2024: 26418-26428.

[16] FAN Z, DAI P, SU Z, et al. EMHI: A multimodal egocentric human motion dataset with HMD and body-worn IMUs [C]//AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI. 2025: 2879-2887.

[17] LIU Z, TANG H, AMINI A, et al. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2023: 2774-2781.

[18] WEI Y, ZHAO L, ZHENG C, et al. BEV-CFKT: A lidar-camera cross-modality-interaction fusion and knowledge transfer framework with transformer for bev 3D object detection[J]. Neurocomputing, 2024, 582: 127527.

[19] WANG H, TANG H, SHI S, et al. UniTR: A unified and efficient multi-modal transformer for bird's-eye-view representation[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2023: 2740-2750.

[20] XIE L, XIANG H, XU Z, et al. AMMF: Attention-based multi-phase multi-task fusion for small contour object 3d detectionp[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2022: 2858-2864.

[21] TIAN Y L, ZHANG X J, WANG X, et al. ACF-Net: As

ymmetric cascade fusion for 3D detection with lidar point clouds and images[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(2): 3360-3371.
[22] WAN J, LI Z, LIU Y, et al. RPEFlow: Multimodal fusion of RGB-PointCloud-Event for joint optical flow and scene flow estimation[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2023: 8222-8228.
[23] FENG R, HU J, XIA W, et al. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors[DB/OL]. (2025-04-01) [2026-03-09]. https://arxiv.org/abs/2502.12191.
[24] 张慧，梁姝彤，李明轩，等. 视觉–语言–动作模型综述: 从前史到前沿[J]. 自动化学报，2025，51(9)：1922-1950.
ZHANG H, LIANG S T, LI M X, et al. Vision-language-action models: From the early foundations to the state-of-the-art[J]. Acta Automatica Sinica, 2025, 51(9): 1922-1950.
[25] CHI C, FENG S, DU Y, et al. Diffusion policy: Visuomotor policy learning via action diffusion[J]. The International Journal of Robotics Research, 2025, 44(10-11): 1684-1704.
[26] YE S, JANG J, JEON B, et al. Latent action pretraining from videos[DB/OL]. (2025-05-15) [2026-03-05]. https://arxiv.org/abs/2410.11758.
[27] BROHAN A, BROWN N, CARBAJAL J, et al. RT-1: Robotics transformer for real-world control at scale[DB/OL]. (2022-08-11) [2026-03-09]. https://arxiv.org/abs/2212.06817.
[28] KIM C,PERTSCH K, SERMANET P, et al. OpenVLA:An open-source vision-language-action model[DB/OL]. (2024-09-05) [2026-03-05]. https://arxiv.org/abs/2406.09246.
[29] JIANG Z, ZHANG T, JANNER M, et al. HybridVLA:Collaborative diffusion and autoregression in vision-language-action models[DB/OL]. (2025-06-23) [2026-03-05]. https://arxiv.org/abs/2503.10631.
[30] WEN J, ZHU M, ZHU Y, et al. DiffusionVLA: Autoregressive reasoning and diffusion policies for generalizable vision-language-action models[C]//International Conference on Machine Learning. New York, USA: ACM, 2025, 267: 1-17.
[31] O'NEILL A, REHMAN A, MADDUKURI A, et al. Open X-Embodiment: robotic learning datasets and rt-x models: Open xembodiment collaboration[C]//IEEE International Conference on Robotics and Automation, Piscataway, USA: IEEE, 2024: 6892-6903.
[32] KHAZATSKY A, PERTSCH K, NAIR S, et al. DROID: A large-scale in-the-wild robot manipulation dataset[DB/OL]. (2024-03-19) [2026-03-09]. https://doi.org/10.48550/arXiv.2403.12945.
[33] JANG E, IRPAN A, KHANSARI M, et al. BC-Z: Zero-shot task generalization with robotic imitation learning[C]//Conference on Robot Learning. New York, USA: ACM, 2021: 991-1002.
[34] KIM M J, PERTSCH K, KARAMCHETI S, et al. OpenVLA: An open-source vision-language-action model [C]//Conference on Robot Learning. New York, USA: ACM, 2024: 2679-2713.
[35] REED S, ZOLNA K, PARISOTTO E, et al. A generalist agent[DB/OL]. (2022-11-11) [2026-03-05]. https://arxiv.org/abs/2205.06175.
[36] ZHAO T Z, KUMAR V, LEVINE S, et al. Learning fine-grained bimanual manipulation with low-cost hardware[C]//19th Robotics: Science and Systems. California, USA: MIT Press, 2023.
[37] WU Y, TIAN R, SWAMY G, et al. From foresight to forethought: VLM-in-the-loop policy steering via latent alignment[C]//21st Robotics: Science and Systems, California, USA: MIT Press, 2025.
[38] LIU R, WANG W, YANG Y. Volumetric environment representation for vision-language navigation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2024: 16317-16328.
[39] QU D, SONG H, CHEN Q, et al. SpatialVLA: Exploring spatial representations for vision-language-action model [C]//21st Robotics: Science and Systems, California, USA: MIT Press, 2025.
[40] CHEN Y, TIAN S, LIU S, et al. ConRFT: A reinforced fine-tuning method for vla models via consistency policy [C]//21st Robotics: Science and Systems, California, USA: MIT Press, 2025.
[41] DENNINGER M, SUNDERMEYER M, WINKELBAUER D, et al. Blenderproc: Reducing the reality gap with photorealistic rendering[C]//16th Robotics: Science and Systems. .
[42] HUBER J, HÉLÉNON F, WATRELOT H, et al. Domain randomization for Sim2Real transfer of automatically generated grasping datasets[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2024: 4112-4118.
[43] KIM Y, SOH J W, PARK G Y, et al. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE. 2020: 3482-3492.
[44] ZHAO Y, JING X, QIAN K, et al. Skill generalization of tubular object manipulation with tactile sensing and sim2real learning[J]. Robotics and Autonomous Systems, 2023, 160: 104321.
[45] MISHRA S, PANDA R, PHOO C P, et al. Task2Sim: Towards effective pre-training and transfer from synthetic data[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 9194-9204.
[46] ARND T K, HAZARA M, GHADIRZADEH A, et al. Meta reinforcement learning for Sim-to-Real domain adaptation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 2725-2731.
[47] YANG F, XU X, CHEN H, et al. GPro3D: deriving 3D bbox from ground plane in monocular 3D object detection[J]. Neurocomputing, 2023, 562: 126894.
[48] ZHANG J, XIA Z, DONG M, et al. CoMatcher: Multi-view collaborative feature matching[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2025: 21970-21980.
[49] ZHU C, WANG T, ZHANG W, et al. LLAVA-3D: A simple yet effective pathway to empowering lmms with 3d-awareness[DB/OL]. (2025-8-27) [2026-3-5]. https://arxiv.org/abs/2409.18125.
[50] MA C, LU K, CHENG T Y, et al. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3D priors[J]. Advances in neural information processing systems, 2024, 37: 68803-68832.
[51] ZHANG R, GUO Z, ZHANG W, et al. Pointclip: Point cloud understanding by clip[C]//Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 8552-8562.
[52] HUANG H, CHEN Y, WANG Z, et al. Chat-scene: Bridging 3D scene and large language models with object identifiers[J]. Advances in Neural Information Processing Systems, 2024, 37: 113991-114017.
[53] WANG Z, HUANG H, ZHAO Y, et al. Chat-3D: Data-efficiently tuning large language model for universal dialogue of 3D scenes[DB/OL]. (2023-08-17) [2026-03-05]. https://arxiv.org/abs/2308.08769.
[54] GUO Z, ZHANG R, ZHU X, et al. Point-bind & point-LLM: aligning point cloud with multi-modality for 3D understanding, generation, and instruction following[DB/O

L]. (2023-09-01) [2026-03-05]. https://arxiv.org/abs/2309.00615.
[55] ZHOU J, WANG J, MA B, et al. Uni3D: exploring unified 3D representation at scale[DB/OL]. (2023-10-10) [2026-03-05]. https://arxiv.org/abs/2310.06773.
[56] HONG Y, ZHENG Z, Chen P, et al. Multiply: A multisensory object-centric embodied large language model in 3d world[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2024: 26406-26416.
[57] 杨静，王晓，王雨桐，等. 平行智能与 CPSS: 三十年发展的回顾与展望[J].自动化学报，2023，49(3)：614-634.
YANG J, WANG X, WANG Y T, et al. Parallel intelligence and CPSS in 30 years: An acp approach[J]. Acta Automatica Sinica, 2023, 49(3): 614-634.
[58] 朱威，洪力栋，施海东，等. 结合优势结构和最小目标 Q 值的深度强化学习导航算法[J]. 控制理论与应用，2024，41(4)：716-728.
ZHU W, HONG L D, SHI H D, et al. Deep reinforcement learning navigation algorithm combining advantage structure and minimum target Q-value[J]. Control Theory and Applications, 2024, 41(4): 716-728.
[59] CHENG J Y, FAN J H, LI X L, et al. Asymmetric information enhanced mapping framework for multirobot exploration based on deep reinforcement learning[J]. IEEE Transactions on Robotics, 2025, 36: 6250-6266.
[60] DEVO A, MEZZETTI G, COSTANTE G, et al. Towards generalization in target-driven visual navigation by using deep reinforcement learning[J]. IEEE Transactions on Robotics, 2020, 36(5): 1546-1561.
[61] 孟怡悦，郭迟，刘经南. 采用注意力机制和奖励塑造的深度强化学习视觉目标导航方法[J]. 武汉大学学报（信息科学版），2023，49(7)：1100-1108,1119.
MENG Y Y, GUO C, LIU J N. Deep reinforcement learning visual target navigation method based on attention mechanism and reward shaping[J]. Geomatics and Information Science of Wuhan University, 2023, 49(7): 1100-1108,1119.
[62] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518, 529-533.
[63] SILVER D, HUANG A, MADDISON C, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529: 484-489.
[64] QI Y K, WU Q, ANDERSON P, et al. REVERIE: Remote embodied visual referring expression in real indoor environments[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 9982-9991.
[65] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. New York, USA: ACM, 2021, 139: 8748-8763.
[66] DEITKE M, VANDERBILT E, HERRASTI A, et al. PROCTHOR: large-scale embodied ai using procedural generation[J]. Advances in Neural Information Processing Systems, 2022,35: 5982-5994.
[67] XIA F, AMIR Z R, HE Z Y, et al. Gibson Env: Real-world perception for embodied agents[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 9068-9079.
[68] 谢远龙，王书亭，程祥，等. 大模型驱动的具身智能机器人导航技术综述[J]. 华中师范大学学报（自然科学版），2025，59(5)：677-693.
XIE Y L, WANG S T, CHENG X, et al. An overview of large model-driven embodied intelligent navigation[J]. Journal of Central China Normal University (Nat. Sci.), 2025, 59(5): 677-693.
[69] 白辰甲，许华哲，李学龙. 大模型驱动的具身智能: 发展与挑战[J]. 中国科学：信息科学，2024，54(9)：2035-2082.
BAI C J, XU H Z, LI X L. Embodied-AI with large models: Research and challenges[J]. Sci Sin Inform, 2024, 54(9): 2035-2082.
[70] 赵博涛，亢祖衡，瞿晓阳，等. 基于多模态大模型的具身智能体研究进展与展望[J]. 大数据，2025，11(3)：108-138.
ZHAO B T, KANG Z H, QU X Y, et al. Review and emerging trends of embodied agent based on multimodal large language models[J]. Big Data Research, 2025, 11(3): 108-138.
[71] XU R T, ZHANG J G, SUN J X, et al. MRFTrans: Multimodal representation fusion transformer for monocular 3d semantic scene completion[J]. Information Fusion, 2024, 111: 102493.
[72] ZHANG Z, LIN A, WANG C X, et al. Interactive navigation in environments with traversable obstacles using large language and vision-language models[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2024: 7867-7873.
[73] AN D, WANG H Q, WANG W G, et al. ETPNav: Evolving topological planning for vision-language navigation in continuous environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(7): 5130-5145.
[74] LI H, LI M, CHENG Z Q, et al. Human-aware vision-and-language navigation: bridging simulation to reality with dynamic human interactions[C]//Conference on Neural Information Processing Systems. New York, USA: ACM, 2024, 37: 119411-119442.
[75] DORBALA V S, MULLEN J F, MANOCHA D. Can an embodied agent find your “cat-shaped mug”? llm-based zero-shot object navigation[J]. IEEE Robotics and Automation Letters, 2024, 9(5): 4083-4090.
[76] QIAO Y, LIU Q, LIU J, et al. LLM as copilot for coarse-grained vision-and-language navigation[C]//European Conference on Computer Vision. Berlin, German: Springer, 2024, 15063: 459-476.
[77] GAO F, SHI L, TANG J F, et al. Visual and textual commonsense-enhanced layout learning for vision-and-language navigation[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 21311-21324.
[78] WEN C C, HUANG Y S Y, HUANG H, et al. Zero-shot object navigation with vision-language models reasoning [C]//International Conference on Pattern Recognition. Berlin, German: Springer, 2025, 15318: 389-404.
[79] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Neural Information Processing Systems. California, USA: NIPS, 2022, 35: 24824-24837.
[80] ZHOU G Z, HONG Y C, WU Q. NavGPT: Explicit reasoning in vision-and-language navigation with large language models[C]//Conference on Innovative Applications of Artificial Intelligence. Palo Alto, USA: AAAI, 2024, 849: 7641-7649.
[81] ZHOU G Z, HONG Y C, WANG Z, et al. NavGPT-2: Unleashing navigational reasoning capability for large vision-language models[C]//European Conference on Computer Vision. Berlin, Germany: Springer, 2024, 260-278.
[82] KONG X H, CHEN J Y, WANG W G, et al. Controllable navigation instruction generation with chain of thought prompting[C]//European Conference on Computer Vision. Berlin, Germany: Springer, 2024: 37-54.
[83] LI Z R, ZHOU G Z, HONG H D, et al. Ground-level viewpoint vision-and-language navigation in continuous environments[DB/OL]. (2025-02-26) [2025-10-13]. https://arxiv.org/abs/2502.19024.
[84] LIN B Q, NIE Y S, ZAI K L, et al. EvolveNav: self-improving embodied reasoning for llm-based vision-language

navigation[DB/OL]. (2025-07-18) [2025-10-13]. https://arxiv.org/abs/2506.01551.
[85] MU Y, ZHANG Q L, HU M K, et al. EmbodiedGPT: vision-language pre-training via embodied chain of thought [DB/OL]. (2023-09-13) [2025-10-20]. https://arxiv.org/abs/2305.15021.
[86] CHI F Y, WANG Y X, NASIOPOULOS P, et al. Multiagent collaborative decision-making using small vision–language models for autonomous driving[J]. IEEE Internet of Things Journal, 2025, 12(24): 55344-55355.
[87] DAGAN G, KELLER F, KELLER A L. Dynamic planning with a LLM[C]//Conference on Neural Information Processing Systems. California, USA: NIPS, 2024: 1-14.
[88] 王浩，范俊铭，郑湃. 基于大语言模型的人机交互移动检测机器人导航方法[J]. 计算机集成制造系统，2024，30(5)：1587-1594.
WANG T, FAN J M, ZHENG P. Large language model-based approach for human mobile inspection robot interactive navigation[J]. Computer Integrated Manufacturing Systems, 2024, 30(5): 1587-1594.
[89] DRIESS D, XIA F, SAJJADI M S M, et al. PaLM-E: an embodied multimodal language model[C]//International Conference on Machine Learning. New York, USA: ACM, 2023, 340: 8469-8488.
[90] MOHAMMADI B, ABBASNEJAD E, QI Y K, et al. Parameter-efficient action planning with large language models for vision-and-language navigation[J]. Pattern Recognition, 2026, 172: 112462.
[91] CHIANG H T L, XU Z, FU Z P, et al. Mobility VLA: multimodal instruction navigation with long-context VLMs and topological graphs[C]//Conference on Robot Learning. New York, USA: ACM, 2025, 270: 3866-3887.
[92] HAO H H, HAN M F, LI C L, et al. CoNav: collaborative cross-modal reasoning for embodied navigation[DB/OL]. (2025-5-22) [2025-10-13]. ）.
[93] IKEUCHI K, WAKE N, SASABUCHI K, et al. Semantic constraints to represent common sense required in household actions for multimodal learning-from-observation robot[J]. The International Journal of Robotics Research, 2024, 43(2): 134-170.
[94] SONG D, LIANG J, PAYANDEH A, et al. VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models[J]. IEEE Robotics and Automation Letters, 2025, 10(1): 508-515.
[95] RAJVANSHI A, SIKKA K, LIN X, et al. SayNav: Grounding large language models for dynamic planning to navigation in new environments[C]//International Conference on Automated Planning and Scheduling. Palo Alto, USA: AAAI, 2024: 464-474.
[96] LIU H K, MA Z Q, LI Y N, et al. Hierarchical language models for semantic navigation and manipulation in an aerial-ground robotic system[J]. Advanced Intelligent Systems, 2025: 1-18.
[97] DU Y, WU C Z, FENG M T, et al. Free-form instruction guided robotic navigation path planning with large vision-language model[C]//Intelligent Robotics and Applications: 17th International Conference. Berlin, Germany: Springer, 2024, 15209: 381-396.
[98] HE Y, ZHOU K, TIAN T L. Multi-modal scene graph inspired policy for visual navigation[J]. The Journal of Supercomputing, 2025, 81. DOI: 10.1007/s11227-024-06541-8.
[99] ZHU F D, LIANG X W, ZHU Y, et al. SOON: Scenario-oriented object navigation with graph-based exploration [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2021: 12689-12699.
[100] WANG X H, WANG W G, SHAO J Y, et al. Learning to follow and generate instructions for language-capable navigation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5): 3334-3350.
[101] WEN S H, ZHANG Z Y, SUN Y X, et al. OVL-MAP: An online visual language map approach for vision-and-language navigation in continuous environments[J]. IEEE Robotics and Automation Letters, 2025, 10(4): 3294-3301.
[102] ANDERSON P, WU Q, TENEY D, et al. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 3674-3683.
[103] PAUL P, GARG A, CHOUDHARY T, et al. LeGo-Drive: language-enhanced goal-oriented closed-loop end-to-end autonomous driving[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2024: 10020-10026.
[104] BROWN A, BROWN N, CARBAJAL J, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control[DB/OL]. (2023-7-28) [2025-7-10]. https://arxiv.org/abs/2307.15818.
[105] SATHYAMOORTHY A J, WEERAKOON K, ELNOOR M, et al. CoNVOI: Context-aware navigation using vision language models in outdoor and indoor environments[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2024: 13837-13844.
[106] JIANG J J, ZHU Y M, WU Z R, et al. DualMap: Online open-vocabulary semantic mapping for natural language navigation in dynamic changing scenes[J]. IEEE Robotics and Automation Letters, 2025, 10(12): 12612-12619.
[107] YOKOYAMA N, HA S, BATRA D, et al. VLFM: Vision-language frontier maps for zero-shot semantic navigation [C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2024: 42-48.
[108] SAHBANI A, EL-KHOURY S, BIDAUD P. An overview of 3D object grasp synthesis algorithms[J]. Robotics & Autonomous Systems, 2012, 60(3):326-336.
[109] ZENG A, SONG S, YU K T, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching[J]. International Journal of Robotics Research, 2022, 41(7): 690-705.
[110] TANG C, HUANG D, GE W, et al. Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping[J]. IEEE Robotics and Automation Letters, 2023, 8(11): 7551-7558.
[111] GUO Z, ZHANG R, ZHU X, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3D understanding, generation, and instruction following[DB/OL]. (2023-09-01) [2026-03-09]. https://arxiv.org/abs/2307.15818.
[112] VAN V T, VU M N, HUANG B, et al. Language-driven grasp detection with mask-guided attention[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2024: 7492-7498.
[113] Li M, Zhao Q, Lyu S, et al. Ovgnet: a unified visual-linguistic framework for open-vocabulary robotic grasping[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2024: 7507-7513.
[114] MA Z, ZHANG S, WEI L, et al. OVMR: Open-vocabulary recognition with multi-modal references[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2024: 16571-16581.
[115] VASU P K A, FAGHRI F, LI C L, et al. FastVLM: Efficient vision encoding for vision language models[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2025: 19769-19780.
[116] SHUKOR M, AUBAKIROVA D, CAPUANO F, et al. Smolvla: A vision-language-action model for affordable and efficient robotics[DB/OL]. (2025-06-02) [2026-03-09]. https://arxiv.org/abs/2506.01844.
[117] AGARWALLA M, KUMAR H, DANDEKAR R, et al. NanoVLMs: How small can we go and still make coherent

Vision Language Models? [DB/OL]. (2025-02-11) [2026-03-09]. https://arxiv.org/abs/2502.07838.
[118] ZHOU Z, ZUO R, YING B, et al. A sensory soft robotic gripper capable of learning-based object recognition and force-controlled grasping[J]. IEEE Transactions on Automation Science and Engineering, 2022, 21(1): 844-854.
[119] XU J, XU B, ZHAN H, et al. A soft robotic system imitating the multimodal sensory mechanism of human fingers for intelligent grasping and recognition[J]. Nano Energy, 2024, 130. DOI: 10.1016/j.nanoen.2024.110120.
[120] DENG Z, JONETZKO Y, ZHANG L, et al. Grasping force control of multi-fingered robotic hands through tactile sensing for object stabilization[J]. Sensors, 2020, 20(4). DOI: 10.3390/s20041050.
[121] KAN Y, LIU X, LIU H. An Image and State Information-Based PINN with Attention Mechanisms for the Rapid Prediction of Aircraft Aerodynamic Characteristics[J]. Aerospace, 2025, 12(5). DOI: 10.3390/aerospace12050434.
[122] LI M J, LIAN Y, LIU X, et al. Adaptive multi-physics finite element-material point method with two-way conversion between elements and particles for metal additive manufacturing[J]. Computer Methods in Applied Mechanics and Engineering, 2025, 446: 118263.
[123] LI M, LI H, GE J, et al. A hierarchical multi-scaling method for modeling the mesoscale geometry of 3D woven composite preform with twisted structure[J]. Composite Structures, 2025, 354. DOI: 10.1016/j.compstruct.2024.118778.
[124] LIU W, DARUNA A, CHERNOVA S. Cage: Context-aware grasping engine[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 2550-2556.
[125] 高翔，谢海晟，朱博，等. 基于多尺度特征融合和抓取质量评估的抓取生成方法[J]. 仪器仪表学报，2023，44(7)：101-111.
GAO X, XIE H S, ZHU B, et al. Grasp generation method based on multiscale features fusion and grasp quality assessment[J]. Chinese Journal of Scientific Instrument, 2023, 44(7): 101-111.
[126] WANG S, ZHOU Z, KAN Z. When transformer meets robotic grasping: exploits context for efficient grasp detection[J]. IEEE robotics and automation letters, 2022, 7(3): 8170-8177.
[127] GAO J, SARKAR B, XIA F, et al. Physically grounded vision-language models for robotic manipulation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2024: 12462-12469.
[128] GUO D, XIANG Y, ZHAO S, et al. Phygrasp: generalizing robotic grasping with physics-informed large multimodal models [DB/OL].(2024-2-26) [2025-10-13] https://arxiv.org/abs/2402.16836
[129] HUANG J, WANG S, LIN F, et al. Tactile-VLA: Unlocking vision-language-action model's physical knowledge for tactile generalization[J]. IEEE Access,2025,13:162467-162504.
[130] HAN Y, YU K, BATRA R, et al. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer[J]. IEEE/ASME Transactions on Mechatronics, 2024, 30(1): 554-566.
[131] CHENG Y, LIU S, ZHA F, et al. A2G: leveraging intuitive physics for force-efficient robotic grasping[J]. IEEE Robotics and Automation Letters, 2024, 9(7): 6376-6383.
[132] CHEN P, LU W. Deep reinforcement learning-based moving object grasping[J]. Information Sciences, 2021, 565: 62-76.
[133] WU C, WANG R, SONG M, et al. Real-time whole-body motion planning for mobile manipulators using environment-adaptive search and spatial-temporal optimization[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2024: 1369-1375.
[134] WANG J, JIN Y, SHI J, et al. EHC-MM: embodied holistic control for mobile manipulation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2025: 13330-13336.
[135] 苏婷婷，张好剑，王云宽，等. 基于费拉里法的 Delta 机器人动态目标抓取算法[J]. 华中科技大学学报（自然科学版），2018，46(6)：128-132.
SU T T, ZHANG H J, WANG Y K, et al. Dynamic picking algorithm based on ferrari's method for delta robot [J]. Journal of Huazhong University of Science and Technology (Nature Science Edition), 2018, 46(6): 128-132.
[136] LIU B, WEI S, YAO M, et al. Dynamic grasping of manipulator based on realtime smooth trajectory generation [J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2024, 238(1): 188-202.
[137] ISLAM F, SALZMAN O, AGARWAL A, et al. Provably constant-time planning and replanning for real-time grasping objects off a conveyor belt[J]. International Journal of Robotics Research, 2021, 40(12-14): 1370-1384.
[138] CHEN H, KIYOKAWA T, WAN W, et al. Adaptive grasping of moving objects in dense clutter via global-to-local detection and static-to-dynamic planning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2025: 14059-14065.
[139] XU J, JIN S, LEI Y, et al. Rt-grasp: reasoning tuning robotic grasping via multi-modal large language model[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2024: 7323-7330.
[140] XIE W, LAVERING J, CORRELL N. DeliGrasp: Inferring object mass, friction, and compliance with LLMs for adaptive and minimally deforming grasp policies[C]//Conference on Robot Learning. New York, USA: ACM, 2024: 1290-1309.
[141] 王建强，黄开启，苏建华. 基于前馈径向基网络的动态抓取参数估计方法[J]. 深圳大学学报（理工版），2022，39(3)：334-342.
WANG J Q, HUANG K Q, SU J H. Dynamic grasping parameter estimation based on feedforward radial basis function network[J]. Journal of Shenzhen University (Science & Engineering), 2022, 39(3): 334-342.
[142] 赵源麒，王清珍. 基于视觉协同的双机械臂物品动态抓取系统设计与实现[J]. 仪器与设备，2024，12(2)：107-116.
ZHAO Y Q, WANG Q Z. Design and realization of dual robotic arm item dynamic gripping system based on visual collaboration[J]. Instrumentation and Equipments, 2024, 12(2): 107-116.
[143] YAN Z, LI S, WANG Z, et al. Dynamic open-vocabulary 3D scene graphs for long-term language-guided mobile manipulation[J]. IEEE Robotics and Automation Letters, 2025, 10(5): 4252-4259.

作者简介：

谢远龙（）
王书亭（）