

• 专题二:双清论坛“工程科学融合人工智能的关键前沿基础科学问题” •

DOI: 10.3724/BNSFC-2025-0025

具身智能的关键挑战和技术*

郝建业** 汤宏焱 郑岩

天津大学 智能与计算学部,天津 300350

[摘要] 近年来,具身智能研究已成为多模态感知、大模型推理、智能决策等研究方向交汇融合的重要领域,在拓宽智能边界与赋能现实世界问题求解方面展现出巨大潜力。然而,由于具身操作的复杂性与任务场景的多样性,具身智能研究面临着高质量数据难采集、具身大模型难构建、训练与推理难优化等严峻的挑战。这些挑战阻碍了具身智能向规模更大、通用性更强、应用面更广的方向发展。本文首先介绍通用大模型与具身大模型之间的演进关联,并探讨具身智能在“数据—模型—优化”三方面的关键挑战。随后,本文系统性梳理具身智能的核心技术,包含具身数据与仿真、具身大小脑模型、具身训练与推理三条脉络,重点剖析技术发展趋势。最后,本文对开放问题与未来发展方向进行讨论与展望,包括虚实结合的规模化数据采集、通用具身推理与操控、少样本快速适应与实时决策等角度,旨在推动具身智能的技术突破与应用落地。

[关键词] 人工智能;智能决策技术;具身智能;大模型;机器人

具身智能是指智能体通过自身物理实体(如躯体、传感器、执行器等)与环境进行实时交互,在动态感知和行动中学习、演进,从而形成认知、推理与任务完成能力的智能模式。相较于传统人工智能,具身智能强调以物理本体作为承载智能的实体,以能够改变现实世界的行动作为决策内容。由此,具身智能研究有望面向现实世界实现更高的智能决策水平,拓宽智能技术落地应用的范围,推动通用人工智能的实现进程。

国内外的具身智能已跨越概念萌芽期,步入技术快速迭代、应用加速拓展阶段。先进具身本体硬件设计、大模型技术、智能决策算法等关键技术的突破,推动其快速发展。中国人形机器人与具身智能产业大会发布的《2025人形机器人与具身智能产业研究报告》显示,2025年中国具身智能市场规模预计达52.95亿元,占全球约27%;2030年全球具身智能市场规模预计达2 326.3亿元,复合年增长率达64.18%。此外,我国高度重视具身智能发展。2025年国务院政府工作报告^[1]首次将“具身

智能”明确列为国家未来产业重点培育方向,标志着具身智能正式上升为国家战略,为后续政策制定提供了顶层指引。工业和信息化部印发的《人形机器人创新发展指导意见》将具身智能产业定位为“科技竞争新高地、未来产业新赛道、经济发展新引擎”,为具身智能技术研发和产业落地提供了系统性框架。

在具身智能蓬勃发展的浪潮下,领域内的研究内容与技术路径在百花齐放的同时,已逐渐凝聚出初步的研究共识与焦点。然而,当前具身智能领域的关键挑战与科学问题、主流技术路径的发展脉络、远景目标与潜在前沿方向,仍缺乏系统梳理与深入讨论。因此,本文系统探讨具身智能研究中的关键挑战与核心技术,旨在明确研究目标、剖析技术发展趋势、深入讨论开放问题,推动具身智能的技术突破与应用落地。

1 大模型时代下的具身智能

具身智能是传统人工智能研究迈向具象化、现实化

收稿日期:2025-07-30; 修回日期:2025-12-04

* 本文根据国家自然科学基金委员会第409期“双清论坛”讨论的内容整理。

本文受到国家自然科学基金项目(62422605,92370132)的资助。

引用格式: 郝建业,汤宏焱,郑岩. 具身智能的关键挑战和技术. 中国科学基金,2026,40(1):107-117.

Hao JY, Tang HY, Zheng Y. Key challenges and technologies of embodied intelligence. Bulletin of National Natural Science Foundation of China, 2026, 40(1):107-117. (in Chinese)

的产物。凭借物理具身本体与现实决策交互,具身智能拓宽了人工智能求解现实问题的边界,也展现出应用落地的巨大潜力。与此同时,这对具身智能体提出了更高的智能水平要求:智能体需对多模态信息(如视觉观测、语言指令、本体感知等)进行精准感知与高效融合,理解并分解任务、推理理解任务的正确步骤与规划(被称为“具身大脑”),进而做出实时、鲁棒的操控决策(被称为“具身小脑”),并从产生的交互数据中学习演进。由此,具身智能已成为多模态感知、大模型推理、智能决策、机器人学等方向交汇融合的研究领域。

尽管现阶段具身智能研究已在机械臂操控、具身导航、整身运动控制等问题中取得一定成果,但这些方法与模型通常仅适用于较为特定的场景、较小范围的任务、较为专用的本体。在面向现实开放的问题场景时,这些方法与模型的通用性与泛化能力严重不足。因此,如何实现通用可泛化的具身智能是现阶段研究的关键科学问题。

1.1 通用大模型在具身问题中的缺陷

在具身智能研究发展的同时,大模型研究在国内外备受关注,大模型技术日新月异。目前,大语言模型(Large Language Model, LLM)^[2,3]与多模态大模型(Multi-modal Large Language Model, MLLM)^[4,5]已展现出极强的通用问答与工具调用能力,涵盖数理推断、

科学问答、代码生成、视频理解、内容创作、深度研究等多种任务类型。因此,以大模型作为基座并利用大模型的通识与推理能力,成为实现通用可泛化的具身智能的自然选择。

然而,通用大模型的能力并不足以直接完成通用可泛化具身智能的构建。大量的研究表明通用大模型在具身智能中存在两方面主要缺陷:(1)通用大模型缺乏对物理世界场景的理解:由于数据和领域知识的巨大差异,通用大模型对具身场景任务的空间、时间与动作理解能力不佳。(2)通用大模型欠缺与物理世界精细化交互的能力。通用大模型仅能够输出粗粒度语言指令,无法输出原生动作空间中的决策(如位姿、力控信号)以实现具身场景的精细化操作。这意味着通用可泛化具身智能的实现,需要以通用大模型为基座面向具身场景进行微调学习,并构建“感知—规划—决策”框架,实现多模态理解到具身操控的能力转化。

1.2 具身大模型训练范式

规模定律(Scaling Law)^[6]是大模型研究的“黄金法则”。规模定律指导下的大规模预训练与后训练范式是大模型获取通用能力的关键。类比之下,建立具身大模型训练范式是实现通用可泛化具身智能的必要途径。如图1所示,具身大模型的训练范式在实现逻辑上应包含四个部分构成的环路:(1)规模化、多样化的具身数据

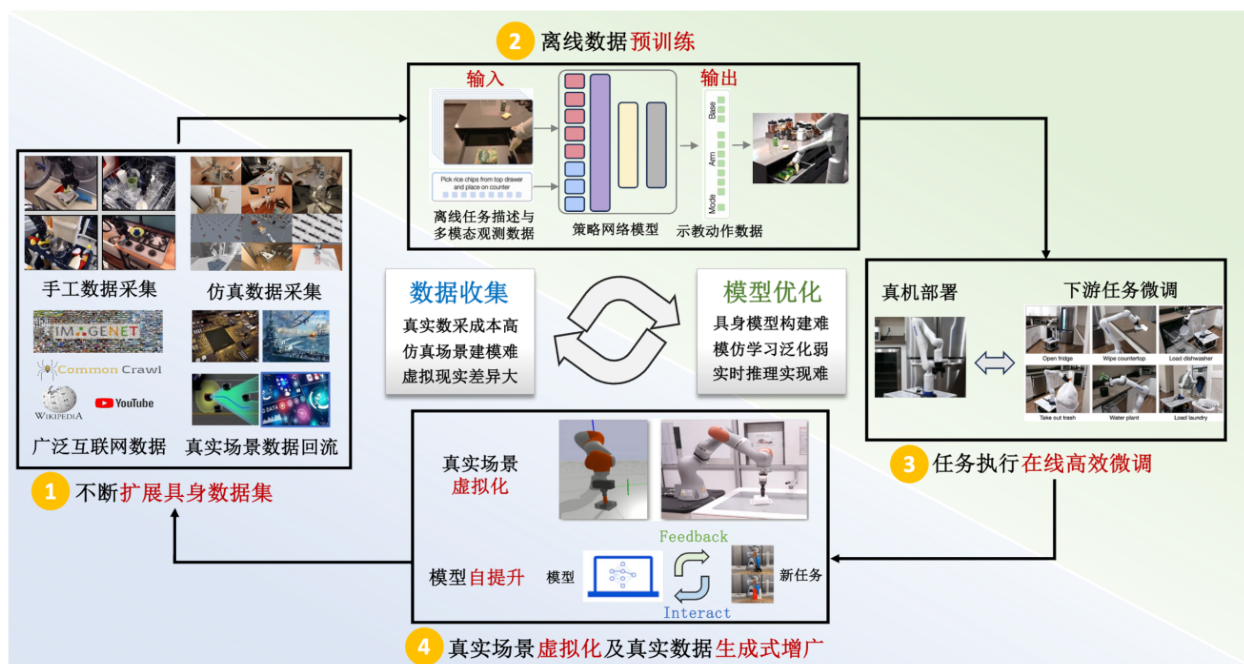


图1 具身大模型训练范式示意图:数据构建—离线训练—在线微调—数据回流四部分构成完整环路,整体形成数据收集与模型优化交替进行、迭代演进的具身大模型训练范式

Fig.1 Illustration of the Training Paradigm of Large Embodied Model: Four Components—Data Construction, Offline Training, Online Fine-tuning, and Data Reflow—Forming a Complete Loop, Establishing the Training Paradigm of Large Embodied Model where Data Collection and Model Optimization Proceed Alternately with Iterative Evolution

集,涵盖广泛的具身理解与决策知识;(2)基于离线具身数据的大规模预训练,优化模型的具身理解能力与操控策略;(3)任务执行过程中的在线高效微调,迁移预训练习得的具身能力并弥合训练测试差距;(4)真实场景虚拟化以及数据生成式增广,实现仿真与数据的飞轮,进而完成训练范式的闭环。上述具身大模型的训练范式由数据收集与模型优化两方面迭代构成,具身模型作为逻辑中心连接数据与优化。

尽管上述范式为研究通用可泛化的具身智能提供了具体的实现思路指导,目前具身智能研究仍未在通用性与泛化性方面取得显著的进展。在后续章节中,本文将探讨具身智能在“数据—模型—优化”三方面所面临的关键挑战,梳理现阶段具身智能的核心技术与发展趋势,并总结开放问题与未来研究方向。

2 具身智能的关键挑战

2.1 高质量具身数据难收集

高质量、大规模的具身数据是实现通用可泛化具身智能的基础。具身数据由“具身大脑数据”和“具身小脑数据”两部分构成,共同驱动机器人“感知—规划—决策”闭环;与大语言模型可直接从互联网爬取海量文本和图像语料不同,用于具身智能的多模态数据极度匮乏。具身大脑侧须覆盖视觉、语言与任务规划等长序列多模态信息,真实场景下不仅需要大规模采集,还需手工或半自动构造多样化标注,人工与设备成本高。具身小脑侧关注具身本体的关节位置、力—触觉等精细操控轨迹;真机采集依赖大量硬件并需通过遥操作、动作捕捉等途径为每个任务充分采集高质量演示数据,受人力与设备约束,规模化成本高、效率低。一种主流的替代方案是借助仿真,实现更加高效且自动化的数据构建,然而此方案存在虚拟到现实迁移的难题;另一方面,具身场景多样,物理属性与环境光照等组合近乎无穷,难以全面覆盖,使得模型遇到微小变化即可能性能骤降。此外,各实验室硬件缺乏统一,数据难以跨机共享。

2.2 高性能通用具身模型难构建

通用可泛化具身智能系统的核心,在于构建能够处理复杂多样任务的高性能具身模型。具身模型的复杂性主要体现在架构设计、多模态融合、跨场景泛化等层面,深刻影响着具身模型的性能表现和实际部署。架构设计方面,端到端具身模型直接基于多模态感知信息输出动作决策,范式简洁、易于实现,但性能依赖高质量数据与先进训练方案;新兴分层架构将负责认知决策的具身大脑与运动执行的具身小脑分离,在提升任务处理能力的同时,不可避免地引入了协调同步等技术难题。多

模态融合方面,模型需同时处理视觉、深度、点云、语言等异构数据,面临模态对齐和统一表示学习的挑战。当前3D视觉语言动作模型(Vision-Language-Action Model, VLA)等架构处理高维异构数据时多模态融合难度大、模型参数大、计算复杂度高。跨场景泛化方面,任务需求差异、环境变化以及机器人本体差异对模型跨任务、跨环境、跨本体泛化能力提出巨大挑战。当前具身模型虽实现了一定程度的跨相似物体、背景迁移的能力,但面对显著不同的任务时性能下降明显,机器人本体差异、跨场景数据统一处理与训练的复杂性进一步增加具身模型泛化的难度。

2.3 高效具身训练与推理难实现

与仅处理数字信息的非具身智能不同,具身智能必须通过与物理世界直接互动来学习,这种根本差异导致了其在训练和推理上的低效性。具身训练上的挑战主要源于真实世界学习的样本低效与泛化困难。基于强化学习训练需要海量的交互,过程缓慢、损耗硬件且有风险,而模仿学习的性能上限则受限于专家数据,难以超越和泛化。另一方面,机器人领域缺乏大规模、多样化的数据集,导致模型极易过拟合于训练数据中的虚假相关性(例如,将“擦除红色笔迹”的任务错误地理解为只针对“红色”),而无法学到抽象层级更高的通用概念。尽管以大型视觉语言模型为基座的后训练有望将海量常识迁移到机器人决策中,这种训练范式带来了新问题:机器人控制所需的物理操控与模型原有的二维视觉理解任务存在本质区别,导致知识遗忘和能力冲突。在具身推理端,大型具身模型的计算开销巨大,在资源受限的机器人硬件上实现实时推理是一个严峻的工程挑战。当前主流VLA模型推理频率普遍在3~15 Hz之间,远低于高精度机器人控制所需的100 Hz以上要求,因此与现实应用中的实际要求之间存在巨大的推理效率鸿沟。

3 具身智能的核心技术

围绕具身智能“数据—模型—优化”三方面关键挑战,本部分从具身数据与仿真、具身大小脑模型、具身训练与推理三方面系统性梳理核心技术脉络,重点剖析技术发展趋势。具身智能技术结构如图2所示。

3.1 具身数据与仿真

3.1.1 具身大脑数据

具身大脑作为负责高层次认知决策的核心模块,其数据构建和评估需面向多模态场景理解、空间推理和任务规划等多方面、多层次的具身能力。区别于通用视觉文本数据可从互联网便捷爬取,具身大脑数据需要专门

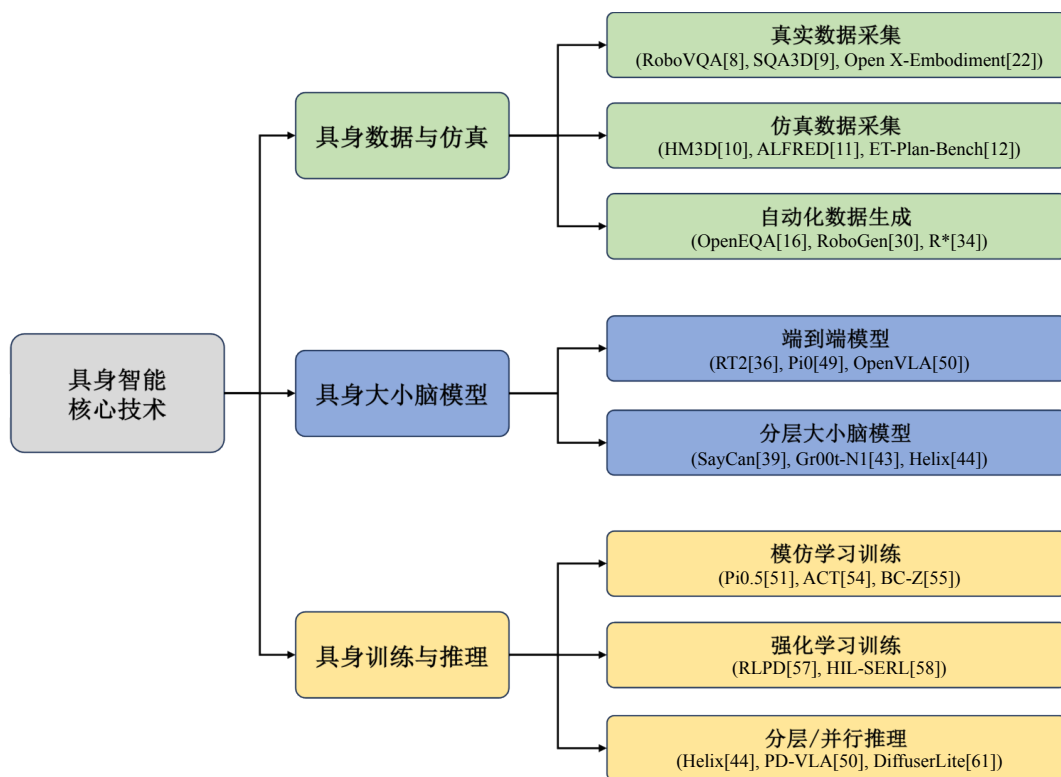


图2 具身智能技术结构图

Fig.2 Diagram of the Architecture of Embodied Intelligence Technologies

的数据集构建方法和评估体系来支撑其能力发展。从数据来源和构建流程看,具身大脑数据经历了从真实环境采集到仿真环境扩展,再到自动化构造的发展历程,并建立了相应的自动化评估机制。

真实环境数据为具身大脑提供了最贴近实际应用场景的训练基础。在具身智能研究兴起之前, Matterport3D^[7]数据集提供了高保真度的真实室内场景3D扫描数据,涵盖住宅、办公楼等场景,支持后续研究中的长距离导航、3D场景理解等复杂具身任务。近年来,面向具身大脑研究的数据集如RoboVQA^[8]基于真实室内场景和长时序具身操控任务,构建了包含约8万条问答对的大规模问答数据集,涵盖物体识别、空间定位、功能推理、场景描述、状态判断、行动规划等多维度具身能力;对于具身导航任务, SQA3D^[9]基于3D扫描构建了包含约6.8万条问答样本的大量真实数据,具有高度的场景真实性,但采集成本高昂。另一方面,仿真环境通过可控的虚拟环境可以实现大规模数据生成, HM3D^[10]、ALFRED^[11]、ET-Plan-Bench^[12]等基准在Habitat^[13]、AI2-THOR^[14]、VirtualHome^[15]等仿真器中构建了涵盖导航、操控、任务规划的多样化场景,支持任务复杂度和场景多样性的精确控制。上述两种具身大脑数据构造的渠道均依赖人工设计与干预,为此,自动化数据构造技术旨在利用多模态大模型,对源数据进行密集标注并

自动构造开放式问答对。例如, OpenEQA^[16]生成了1636条高质量具身问答数据,充分发挥了仿真数据的规模优势和真实数据的场景丰富性,在任务复杂度、场景多样性、语言表达丰富度等维度实现了有机结合,显著降低了人工标注成本并提高了数据构建效率。表1提供了代表性具身大脑数据集关于数据内容、规模、形式等维度的对比。然而,相较于通用大模型基于数十万亿tokens的训练数据规模(如GPT-4^[17]采用超过15T tokens的训练数据),现有具身大脑数据仍存在2~3个数量级的差距,数据稀缺性严重制约了具身智能模型的能力发展。

随着数据构建方法的发展,自动化评估体系的建立为具身大脑能力测评提供了标准化框架。传统评估严重依赖人工判断,成本高且一致性差。新兴的自动化评估机制基于大语言模型构建评分体系,如OpenEQA设计的基于LLM的评分器能够自动评估问答质量, UniEQA^[18]构建了最为完整的多层次评估框架,涵盖对象理解、时空感知、具身知识、具身推理、任务规划五大核心能力维度,实现了基础感知—高层推理—任务执行的分层能力评估。这种评估体系通过图片、视频等多模态数据和选择题、排序、开放式问答等多种指令类型,为具身大脑模型提供了全面的性能测评标准。综上,具身大脑数据构建正朝着虚实结合、自动化标注、标准化评

表1 代表性具身大脑数据集对比
Table 1 Comparison among Representative Datasets of Embodied Cerebrum

数据集	年份	现实/仿真	观测模态	规模	任务/范围	指令形式	硬件/平台
Matterport 3D	2017	现实	RGB-D/点云等+语言	10.8k室内场景	视觉语言导航/ 3D场景理解	自然语言	相机采集与重建
HM3D	2020	仿真	RGB-D/点云等+语言	7.5k训练/2k测试 目标导航任务	具身目标导航	自然语言	Habitat
ALFRED	2020	仿真	RGB+语言	21k训练/1.6k验证/ 3k测试导航/操控 任务规划数据	视觉语言导航、 家庭机器人 任务规划等	自然语言	AI2-THOR
RoboVQA	2023	现实	视频+语言	798k训练/18k验证 问答对	具身推理能力评估	自然语言	现实视频数据
SQA3D	2023	现实	3D扫描/鸟瞰图BEV等 +语言	26k训练/3k验证/3k 测试问答对	具身场景理解	自然语言	ScanNet
OpenEQA	2023	现实+仿真	RGB/视频+语言	1 636测试问答对	具身感知、推理等 多维度能力评估	自然语言	Habitat/ScanNet
ET-Plan-Bench	2025	仿真	多视角RGB+ 语言	10k+导航/操控任务 规划数据	家庭机器人任务规划	自然语言	VirtualHome/Habitat

估的方向发展。

3.1.2 具身小脑数据

具身小脑作为基于观测、推理输入做出实时操控决策的最终模块,其数据需要包含动作的精细化操控演示,通常通过真机或仿真收集实现。前者需要人工操控真机在真实世界生成专家示范;后者通过数据生成管道在高保真仿真中收集,两种方案互补,共同支撑具身小脑模型的演进。

对于基于真机的具身小脑数据集构建,最直接的方式是采用遥操控。Mobile ALOHA^[19]通过遥操控主臂来控制从臂,实现分钟级高质量数据收集;AgiBot World^[20]则采用全身动捕与力触同步采集,在厨房、仓储等多场景进行收集。在此类技术推动下,一批公开大规模真机数据集相继出现:谷歌RT-1^[21]发布了13万条专家演示数据,涵盖13台机器人、700多类桌面任务;Open X-Embodiment^[22]汇总22个硬件平台、60个数据源,构建超过百万条跨平台轨迹;RoboSet^[23]、BridgeData^[24,25]、DROID^[26]在真实家庭等场景中构建了基于深度图像与操控动作的数据集。真实数据使训练得到的具身模型能更好地解决现实世界的任务,但由于人力与设备成本高昂,数据收集难以规模化、覆盖程度低,难以支撑通用泛化具身操控的训练。基于仿真的具身小脑数据构建成为低成本的另一途径。Isaac^[27]和SAPIEN^[28]利用GPU加速的PhysX和光线追踪技术生成高保真操控轨迹;Genesis^[29]则通过超高速物理求解与自动化数据生成引擎加速多模态数据生成。基于上述仿真,具身小脑数据自动构建有两种主要方案:一是结合运动规划和位姿估计批量生成轨迹,二是基于深度强化学习在并行环

境中求解任务获得数据采集策略。基于上述仿真,代表性的仿真数据集LIBERO^[30]和RoboCasa^[31]提供了多样化的具身操控任务与用于具身模型训练的示教数据。表2展示了代表性具身小脑数据集关于数据内容、规模、形式等维度信息的对比情况。

随着大语言模型的应用,数据生成逐步向全流程自动化发展。RoboGen^[32]引入大语言模型构建仿真数据流水线,支持从任务生成到奖励函数设计的自动化;Eureka^[33]结合GPT-4^[17]与并行强化学习,相较RoboGen提升了数据采集成功率;R*^[34]则通过拆解奖励设计为逻辑结构与参数对齐,进一步加速收敛。整体上,仿真数据构建已形成了“任务生成—数据采集—奖励设计—策略优化”的闭环范式,但仿真数据需要其场景与真实场景高度对齐。目前,针对仿真到现实迁移的主流方案主要包含领域随机化以及领域适应两类,前者通过在仿真中加入参数(如摩擦系数)与视觉等随机化来提升仿真策略迁移的鲁棒性,后者则需要获取真实世界信息,并对仿真设置进行优化对齐。

3.2 具身大小脑模型

3.2.1 具身端到端模型

具身智能模型接收环境与任务的多模态信息(如视觉观测、语言指令、本体感知等),做出可执行的决策,通过与环境交互来完成任务。直观地,具身模型需对多模态信息进行融合,理解任务并拆分任务、生成任务执行规划,进而按规划做出实时操控决策。凭借大模型与深度学习的强大能力,构建具身模型最直接的方式是采用端到端模型,即以网络模型直接拟合输入输出间的函数映射,决策产生的过程隐含在网络隐藏层处理中。RT-

表2 代表性具身小脑数据集对比
Table 2 Comparison among Representative Datasets of Embodied Cerebellum

数据集	年份	现实/仿真	观测模态	规模	任务/范围	指令形式	硬件/平台
BridgeData V2	2023	现实	RGB-D/多视角	60 096轨迹/24环境/ 13技能	多场景泛化	目标图像/ 自然语言	WidowX 250机械臂
RoboCasa	2024	仿真	多视角RGB/状态 (可配语言)	120厨房场景/2.5k+ 物(150+类)/100任务/ 大于100K轨迹(人类 示范+自动生成)	家居类操控 (原子/复合 长时序任务)	自然语言	MuJoCo/robosuite (RoboCasa仿真)
DROID	2024	现实	多视角RGB-D +语言	76k轨迹/350小时/ 564场景/86任务	多机构采集	自然语言	Franka Panda
Open-X Embdiment	2023—2024	现实	RGB-D/多源	22种机器人/527技能/ 160 266任务	跨平台大一 统训练	自然语言/ 目标图像等	21个机构多机械臂
RoboSet	2023	现实	多视角RGB/状态	约7 500轨迹/12技能/ 38任务	厨房多任务 操控	任务描述	Franka-Emika/Robotiq
Mobile-ALOHA	2024—2025	现实	多视角RGB/状态	静态ALOHA 825+ 移动数据,每任务约 50条人类示范	移动+双臂长 时序	脚本/任务 描述	移动底盘+ALOHA
LIBERO	2023	仿真	RGB/状态+语言	130个任务,每个任务 50条人类示范	终身学习/知识 迁移(桌面/ 厨房等)	自然语言	MuJoCo/robosuite (Franka Panda仿真)

l^[21]是代表性的具身端到端模型,面向具身抓取任务,RT-1基于Transformer^[35]网络构建,从操控演示数据中通过模仿学习进行端到端训练。由于RT-1基于初始化的网络模型从零训练,其语言和视觉理解能力不足,泛化性较弱。对此,RT-2^[36]以视觉语言模型(Vision-Language Model,VLM)作为基座,并基于具身操控数据进行后训练微调,旨在将VLM内含的海量语义与常识知识迁移到具身端到端决策中。具身导航方面,Navid^[37]、UniNavid^[38]等模型同样采用端到端框架,融合时空视觉特征与指令语义直接输出动作,摒弃传统具身导航方法对人工设计的语义图、动作选择机制的依赖,实现端到端导航。

具身端到端模型基于多模态感知信息直接输出动作决策,范式简洁、易于实现,但其性能高度依赖高质量、规模化数据以及先进的具身端到端模型训练方法。另一方面,随着具身智能研究焦点从单一场景、简单任务向现实场景、长序复杂任务的转移,具身端到端模型的能力不足与其对数据、训练方法的依赖问题逐渐加剧,使得基于端到端模型实现通用可泛化具身智能的技术路线面临巨大挑战。

3.2.2 具身大小脑模型

在具身端到端模型显现出长序任务求解能力不足的同时,根据具身决策的职能差异,采取具身大小脑分层构建的具身模型应运而生。分层具身大小脑模型将复杂的具身任务分解为高层规划(具身大脑模型)和低层执行(具身小脑模型)两个层次,实现解耦式的协同决

策。具身大脑模型负责任务理解、规划拆解和高层推理,具身小脑模型基于具身大脑模型输出的拆解规划执行具体动作并与环境交互。

具身大脑作为分层架构中的高层规划与推理模块,在真实应用场景中需要面对复杂多变的环境,通过多模态理解和语义融合,将抽象的任务指令转化为具体的执行规划。从技术路线来看,当前主要形成了三种任务拆解与推理范式:(1)纯文本拆解与推理范式,基于子任务文本描述进行任务分解,如SayCan^[39]、L3MVN^[40],通过自然语言接口激活大语言模型的常识推理能力实现零样本泛化;(2)混合模态拆解与推理范式,结合文字和图像双模态生成更具表意清晰度的多模态子指令,如PERIA^[41]通过“感知—推理—想象—执行”链路生成包含目标状态图像的分步指令,VLFM^[42]将原始图像与导航指令直接映射至共享语义空间,构建融合语义关联度和物理可达性的动态价值地图增强导航规划;(3)隐层拆解与推理范式,采用隐式表征进行具身大小脑连接,如GR00T N1^[43]和Helix^[44]等工作通过“快”—“慢”双系统架构实现语义更丰富的隐式表征传递。上述三类范式各具特点:纯文本拆解与推理范式擅长利用基座大模型的文本能力,但存在多模态信息损耗;混合模态范式规划指引清晰、可解释性强但推理效率较低;隐层范式支持快速推理和端到端优化,但依赖训练数据且可解释性差。未来研究需聚焦多范式融合,结合基座模型知识先验与任务相关的丰富信息,提升具身大脑模型的任务拆解规划能力。

具身小脑模型基于上层的具身大脑模型输出的拆解规划与场景信息,做出底层操控决策,执行动作与环境交互。早期具身小脑模型在具身架构方面可以分为基于Transformer^[35]的架构和基于扩散模型^[45]两类。基于Transformer的骨架网络(如ACT^[46])使用自回归动作预测,表征能力强且天然契合各类通用大模型的网络架构,但缺乏轨迹拼接能力和多峰分布拟合能力;相比之下,基于扩散模型的骨架网络(如Diffusion Policy^[47]和RDT^[48])则直接利用扩散模型的去噪过程进行动作生成,利用扩散模型卓越的分布拟合能力取得优越的具身操控性能,然而受限于扩散模型的迭代去噪过程,导致决策速度缓慢。融合两种架构的优势,近期提出的具身小脑模型^[49]使用Transformer骨架(大参数量)和扩散模型动作头(小参数量)的组合,使用预训练的VLM对观测和文本指令进行表征后,用小参数的扩散模型动作头进行决策,这样不仅能利用VLM强大的视觉和语言理解能力,也能受益于扩散模型的复杂分布拟合能力,精准地进行动作决策。尽管主流具身小脑模型均基于预训练的VLM进行微调训练,其泛化能力依然不及预期,无法针对新场景或者新任务实现有效的零样本或者少样本决策。例如,在OpenVLA^[50]的评估中,模型在面对与训练数据风格迥异的新厨房环境或新物体时,其零样本任务成功率接近于0,这暴露了模型对特定训练场景的过拟合问题。因此,近期的研究着眼于将视觉观测和具身小脑决策联系起来,通过针对视觉感知(检测、分割)和面向底层操控的推理能力(物体关联、运动轨迹)进行后训练,进一步提升具身小脑模型的泛化性能。Pi0.5^[51]在训练数据中加入大量通用视觉数据,保留视觉理解能力;Gemini-ER^[52]基于大量思维链数据构建支持具身推理的VLM模型,并在此基础上进行Gemini Robotics VLA后训练;FSD^[53]则使用视觉标记作为具身无关的表征,提出基于空间关系的思维链,使用视觉标记桥接视觉—语言理解与动作生成,通过模拟人类思维模式的方式,实现空间视觉轨迹推理和实现跨本体跨任务的泛化操作。

综上,具身模型的构建存在端到端具身模型与分层具身大小脑模型两条主要技术路径。前者范式简洁但缺乏对长序任务的显式推理拆分与子任务规划,端到端训练受限于具身数据规模化的瓶颈,此外样本高效的训练方法也是此技术路线亟需突破的关键;后者决策过程直观、易理解,然而在具身大小脑信息传递、高层规划与底层动作的协同、分层推理的效率等方面均有待取得进一步技术发展,以提升具身模型在现实场景下的通用决策能力。

3.3 具身训练与推理方法

具身训练方法的演进趋势是从逐步降低对海量、高质量专家数据的依赖,转向让智能体通过与环境自主交互进行学习,从而突破模仿学习的性能上限,实现更强的泛化能力和自主学习能力。此发展趋势可分解为三个阶段:(1)基于离线专家数据的模仿学习^[54,55],收集大量专家演示数据,并基于这些离线、固定的数据集训练具身策略模型模仿专家数据。尽管先进的模型结构具备学习复杂多模态动作分布的能力,但此方法的性能上限受限于专家数据的“质”和“量”,模型难以泛化到专家数据没覆盖到的新场景,且无法进一步突破专家决策水平。(2)为克服对静态数据集的依赖,研究者们提出在模仿学习之上构建“数据飞轮”^[56],实现数据自我增强:让模型自主尝试完成任务,从而生成大量新的训练数据;随后将新生成的数据与旧数据合并,训练出一个更强大的新模型。重复此循环,模型在迭代中逐渐提升,极大地扩展了数据规模和多样性,使模型能够探索更广阔的状态空间,从而提升泛化能力。(3)结合强化学习,减少对专家数据的需求^[57]。此方式旨在最大化地利用环境的反馈信号,进一步减少对专家数据的需求。如HIL-SERL^[58]方法将模仿学习与强化学习深度结合,同时从极少量的专家演示和模型自身的探索经验中学习,极大降低了对专家演示数据的需求量;而RLPT-VLA^[59]仅仅使用5个专家演示,通过强化学习后训练在多个环境中大幅超越了基座模型。由此,具身智能体能学会“从反馈中辨别好坏”,而不仅仅是“复制行为”,因而具备通过自我探索学习并超越专家决策水平的潜力,真正突破模仿学习的性能天花板。

随着具身智能逐渐迈向通用大模型时代,主流的具身VLA大模型虽然能力强大,但由于参数量巨大(动辄数十亿甚至上百亿),导致其推理速度很慢,无法满足实时控制需求。以首次动作延迟作为模型推理频率计算标准,主流的具身大模型,例如OpenVLA (7B)^[50]和RDT-1B (1B)^[48]推理频率仅为6 Hz,而RT-2 (55B)^[36]甚至只有3 Hz。对于需要与物理世界进行流畅、精细交互的任务(如快速抓取、灵巧操作)而言,理想的控制频率往往需要达到几十甚至上百赫兹。对此,研究者通过分层和并行解码等方法,分别提升了自回归架构和扩散模型架构下的具身模型推理速度。Helix^[44]构建分层式的混合推理系统,将复杂的语义理解与快速的反应控制分离开,通过隐空间变量连接快(80 M参数量)—慢(7 B参数量)系统,通过分层提升决策频率至200 Hz。PD-VLA^[60]着眼于自回归预测的效率提升,将具身智能决策中的动作块建模为并行解码过程,无需额外训练提升

自回归预测框架的决策频率,实现2.5倍推理效率提升。而对于扩散模型架构,DiffuserLite^[61]使用从粗到细的思想,分层规划,降低生成难度,解决Diffusion框架决策频率过低问题,相比斯坦福大学提出的主流算法Decision Diffuser^[62],提升决策频率120倍,决策频率可达到100 Hz以上。

4 开放问题与未来方向

尽管具身智能领域内研究已从数据收集、模型构建、训练与推理优化方面提出了多种新技术,“数据—模型—优化”三方面关键挑战仍未被完全解决,通用可泛化具身智能的实现仍面临以下开放问题。

数据作为制约具身智能向更强能力演进的基础因素,如何实现虚实结合的全流程自动化数据收集,实现多样化、高质量数据的高效规模化是首要的开放问题。未来研究具体可围绕以下三方面展开:(1)结合“现实到虚拟”(Real-to-Sim)方法构建高质量、多样化的资产库,生成更多样、全面的高保真仿真场景^[63];(2)基于生成模型或并行强化学习等方法构建更高效的具身大小脑数据生产流程,实现多样高质量数据规模化^[29,32,34];(3)解决“虚拟到现实”(Sim-to-Real)迁移难题,充分发挥仿真数据的能力^[64]。通过从以上方面取得技术突破,有望构建高效的数据收集闭环体系,推动具身智能迈上新台阶。

尽管具身规划与操控、多模态融合等研究方向已取得显著进展,进一步面向现实世界实现通用具身模型仍面临诸多开放挑战。分层具身大小脑架构虽实现了高层拆解规划与底层运动执行的解耦,如何在保证推理深度的同时满足推理高效性,并在复杂动态环境中维持决策稳定性,仍需进一步突破^[65]。多模态空间表征融合方面,现有方法虽能处理2D/3D异构数据,但在跨模态对齐、统一表示学习等方面仍存在技术瓶颈^[66]。此外,当前研究缺乏通用、统一的动作表示机制,导致跨本体、跨任务的策略泛化能力有限,难以实现具身智能从专用性向通用性的根本跃升。面向通用具身智能研究异构本体、多样化场景的本征表示^[67],以及由其驱动的泛化决策是一个重要的未来方向。

具身训练与推理方面,现有研究已展现出从以模仿学习为主的训练范式,向结合强化学习的混合训练范式演进的趋势。然而,在面对泛化能力之外的新任务时,主流具身大模型仍需收集较多专家数据进行微调,收集成本高、计算资源消耗大,亟需在少样本快速适应方法方面取得突破。在推理方面,尽管现有研究针对推理执行效率提出了多种优化方法,主流具身大模型的决策速

度仍然远低于100 Hz。近期工作探索了异步推理和执行的VLA实时系统^[68],但如何实现模型普适、决策可靠的推理加速机制,进一步优化具身大模型的推理性能,满足实时、可靠的闭环控制需求,依然是至关重要的开放问题。

此外,现有具身模型在训练与评测数据构造、架构设计、模型规模、推理方式等方面存在较大差异,使得模型性能无法在较大范围内直接比较。大部分现有模型与方法缺乏在同一训练与评测任务集、同一硬件与软件栈、同一数据与度量下的标准化对比,使得系统性的定量对比研究难以开展。因此,设计一套面向各类具身模型的统一评测协议和任务集,对促进具身模型性能优化、推动实际落地需求至关重要。

5 总结

本文聚焦具身智能研究,系统性地介绍了“数据—模型—优化”三方面的关键挑战,并从具身数据与仿真、具身大小脑模型、具身训练与推理三个方面梳理了核心技术与发展趋势。当前具身智能的数据与训练规模落后于同期大语言模型、多模态大模型研究,因此“规模定律”在具身智能研究中尚未得到充分验证与利用;具身模型能力方面呈现出专用性尚可、通用性薄弱的现状,具身应用仍处于较为早期的探索阶段。面向通用可泛化具身智能这一目标,本文从虚实结合的规模化数据采集、通用具身推理与操控、少样本快速适应与实时决策、面向具身模型的标准化评测等角度,探讨了开放问题并展望了未来方向。这些方向的技术突破将推动具身智能的广泛应用,助力现实问题的解决并产生落地价值。

参考文献

- [1] 新华社. 具身智能如何走向未来? (2025-03-06)/[2025-07-30]. https://www.gov.cn/zhengce/202503/content_7010979.html.
- [2] Guo DY, Yang DJ, Zhang HW, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 2025, 645: 633—638.
- [3] Qwen. 突破极限: Qwen3-Max-Thinking 的能力跃迁. (2026-01-26)/[2026-01-26]. <https://qwen.ai/blog?id=qwen3-max-thinking>.
- [4] OpenAI. Introducing OpenAI o3 and o4-mini. (2025-04-16)/[2025-07-30]. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- [5] Google DeepMind. Gemini 2.5: Updates to our family of thinking models. (2025-06-17)/[2025-06-17]. <https://developers.googleblog.com/gemini-2-5-thinking-model-updates/>.
- [6] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models// *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*. NY, USA: OpenReview.net, 2022: 30016—30030.
- [7] Chang A, Dai A, Funkhouser T, et al. Matterport3D: Learning from

- RGB-D data in indoor environments// Proceedings of the 5th International Conference on 3D Vision (3DV). Qingdao, China; IEEE, 2018:667—676.
- [8] Sermanet P, Ding TL, Zhao J, et al. RoboVQA: Multimodal long-horizon reasoning for robotics// Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan; IEEE, 2024:645—652.
- [9] Ma X, Yong S, Zheng Z, et al. SQA3D: Situated question answering in 3D scenes// Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda; OpenReview.net, 2023.
- [10] Santhosh R, Aaron G, Erik W, et al. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D environments for embodied AI// Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS). Virtual; Curran Associates, 2021.
- [11] Shridhar M, Thomason J, Gordon D, et al. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA; IEEE, 2020:10737—10746.
- [12] Zhang LF, Wang YN, Gu HJ, et al. ET-plan-bench: Embodied task-level planning benchmark towards spatial-temporal cognition with foundation models// Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China; IEEE, 2025:21566—21573.
- [13] Savva M, Kadian A, Maksymets O, et al. Habitat: A platform for embodied AI research// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea; IEEE, 2019:9338—9346.
- [14] The Allen Institute for AI. Speed up your embodied AI training with AI2-THOR 2.7.0. (2021-02-10)/[2021-02-10]. <https://medium.com/ai2-blog/speed-up-your-training-with-ai2-thor-2-7-0-12a650b6ab5e>.
- [15] Puig X, Ra K, Boben M, et al. VirtualHome: Simulating household activities via programs// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA; IEEE, 2018:8494—8502.
- [16] Majumdar A, Ajay A, Zhang XH, et al. OpenEQA: Embodied question answering in the era of foundation models// Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA; IEEE, 2024:16488—16498.
- [17] OpenAI. GPT-4. (2023-03-14)/[2023-03-14]. <https://openai.com/index/gpt-4-research/>.
- [18] Zhang M, Fu X, Hao J, et al. UniEQA & UniEval: A Unified Benchmark and Evaluation Platform for Multimodal Foundation Models in Embodied Question Answering. (2025-05-16)/[2025-07-30]. <https://github.com/TJURL-Lab/UniEQA>.
- [19] Fu ZP, Zhao TZ, Finn C. Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation// Proceedings of the 9th Annual Conference on Robot Learning (CoRL). Seoul, Republic of Korea; PMLR, 2025:4066—4083.
- [20] AgiBot-World-Contributors, Bu QW, Cai JS, et al. AgiBot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems// Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2025:3549—3556.
- [21] Brohan A, Brown N, Carbajal J, et al. RT-1: Robotics Transformer for Real-World Control at Scale// Proceedings of the 19th Robotics: Science and Systems (RSS). Daegu, Republic of Korea; Robotics Proceedings, 2023.
- [22] O'Neill A, Rehman A, Maddukuri A, et al. Open X-embodiment: Robotic learning datasets and RT-X models// Proceedings of the 41st IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan; IEEE Computer Society, 2024:6892—6903.
- [23] Oliveira LBR, Amaral FA, Martins DB, et al. RoboSeT: A tool to support cataloging and discovery of services for service-oriented robotic systems. Robotics. Berlin, Heidelberg; Springer, 2015:114—132.
- [24] Ebert F, Yang YL, Schmeckpeper K, et al. Bridge data: Boosting generalization of robotic skills with cross-domain datasets// Proceedings of the 18th Robotics: Science and Systems (RSS). New York City, NY, USA; Robotics Proceedings, 2022.
- [25] Homer W, Kevin B, Abraham L, et al. Bridge data V2: A dataset for robot learning at scale// Proceedings of the 7th Annual Conference on Robot Learning (CoRL). Atlanta, GA, USA; PMLR, 2023: Robotics Proceedings, 1723—1736.
- [26] Khazatsky A, Pertsch K, Nair S, et al. DROID: A large-scale in-the-wild robot manipulation dataset// Proceedings of the 20th Robotics: Science and Systems (RSS), Delft, The Netherlands, 2024.
- [27] Mittal M, Guo K, State G, et al. Isaac lab: A GPU-accelerated simulation framework for multi-modal robot learning. (2025-09)/[2025-10-31]. https://research.nvidia.com/publication/2025-09_isaac-lab-gpu-accelerated-simulation-framework-multi-modal-robot-learning.
- [28] Xiang FB, Qin YZ, Mo KC, et al. SAPIEN: A Simulated part-based interactive Environment// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA; IEEE, 2020:11094—11104.
- [29] Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond. (2024-12)/[2025-07-30]. <https://github.com/Genesis-Embodied-AI/Genesis>.
- [30] Soroush N, Abhiram M, Lance Z, et al. RoboCasa: Large-scale simulation of everyday tasks for generalist robots// Proceedings of the 20th Robotics: Science and Systems (RSS). Delft, Netherlands; RSS Foundation, 2024.
- [31] Alexander K, Karl P, Suraj N, et al. DROID: A large-scale in-the-wild robot manipulation dataset// Proceedings of the 20th Robotics: Science and Systems (RSS). Delft, Netherlands; RSS Foundation, 2024.
- [32] Wang YF, Xian Z, Chen F, et al. RoboGen: Towards unleashing infinite data for automated robot learning via generative simulation. Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria; PMLR, 2024:51936—51983.
- [33] Ma YJ, Liang W, Wang G, et al. Eureka: Human-level reward design via coding large language models// Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria; Open Review.net, 2024.
- [34] Li P, Hao J, Tang H, et al. R*: Efficient reward design via reward structure evolution and parameter alignment optimization with large language models// Proceedings of the 42nd International Conference on Machine Learning (ICML). Vancouver, Canada; PMLR, 2025:34509

- 34527.
- [35] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need// Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Long Beach, California, USA; ACM, 2017; 6000—6010.
- [36] Zitkovich B, Yu T, Xu S, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control// Proceedings of the 7th Annual Conference on Robot Learning (CoRL). Atlanta, GA, USA; PMLR, 2023; 2165—2183.
- [37] Zhang J, Wang K, Xu R, et al. NaVid: Video-based VLM plans the next step for vision-and-language navigation// Proceedings of the 20th Robotics: Science and Systems (RSS). Delft, Netherlands; Robotics Proceedings, 2024.
- [38] Zhang JZ, Wang KY, Wang SA, et al. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks// Proceedings of the 21st Robotics: Science and Systems (RSS). Los Angeles, California, USA; Robotics Proceedings, 2025.
- [39] Ichter B, Brohan A, Chebotar Y, et al. Do as I can, not as I say: Grounding language in robotic affordances// Proceedings of the 6th Annual Conference on Robot Learning (CoRL). Auckland, New Zealand; PMLR, 2022; 287—318.
- [40] Yu B, Kasaei H, Cao M. L3MVN: Leveraging large language models for visual target navigation// Proceedings of the 36th IEEE/RSJ International Conference on Intelligent Robots and Systems. Detroit, MI, USA; IEEE Computer Society, 2023; 3554—3560.
- [41] Ni F, Hao JY, Wu SG, et al. PERIA: perceive, reason, imagine, act *via* holistic language and vision planning for manipulation// Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada; OpenReview.net, 2024; 17541—17571.
- [42] Yokoyama N, Ha S, Batra D, et al. VLFM: Vision-language frontier maps for zero-shot semantic navigation// Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan. IEEE, 2024; 42—48.
- [43] Nvidia. NVIDIA Isaac GR00T N1: An Open Foundation Model for Humanoid Robots. (2025-03-17)/[2025-03-17]. https://research.nvidia.com/publication/2025-03_nvidia-isaac-gr00t-n1-open-foundation-model-humanoid-robots.
- [44] Figure AI. Helix: A vision-language-action model for generalist humanoid control. (2025-02-20)/[2025-07-30]. <https://www.figure.ai/news/helix>.
- [45] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models// Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, BC, Canada. ACM, 2020; 6840—6851.
- [46] Gao CX, Wu CY, Cao MJ, et al. ACT: Empowering decision transformer with dynamic programming via advantage conditioning// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Vancouver, Canada, 2024; 38(11): 12127—12135.
- [47] Chi C, Feng S, Du Y, et al. Diffusion policy: Visuomotor policy learning via action diffusion// Proceedings of the 19th Robotics: Science and Systems (RSS). Daegu, Republic of Korea; Robotics Proceedings, 2023.
- [48] Liu S, Wu L, Li B, et al. RDT-1B: A diffusion foundation model for bimanual manipulation// Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore; OpenReview.net, 2025.
- [49] Physical Intelligence. $\pi 0$: Our First Generalist Policy. (2024-10-31)/[2024-10-31]. <https://www.pi.website/blog/pi0>.
- [50] Kim MJ, Pertsch K, Karamcheti S, et al. OpenVLA: An open-source vision-language-action model// Proceedings of the 9th Annual Conference on Robot Learning (CoRL). PMLR, 2025; 2679-2713.
- [51] Intelligence P, Black K, Brown N, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization// Proceedings of the 9th Annual Conference on Robot Learning (CoRL). Seoul, Republic of Korea; PMLR, 2025; 17—40.
- [52] Google. Gemini Robotics brings AI into the physical world. (2025-03-12)/[2025-03-12]. <https://deepmind.google/blog/gemini-robotics-brings-ai-into-the-physical-world/>.
- [53] Yuan Y, Cui H, Chen Y, et al. From seeing to doing: Bridging reasoning and decision for robotic manipulation// Proceedings of the 14th International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil; OpenReview.net, 2026.
- [54] Zhao T, Kumar V, Levine S, et al. Learning fine-grained bimanual manipulation with low-cost hardware// Proceedings of the 19th Robotics: Science and Systems (RSS). Daegu, Republic of Korea; Robotics Proceedings, 2023.
- [55] Jang E, Irpan A, Khansari M, et al. BC-Z: Zero-shot task generalization with robotic imitation learning// Proceedings of the 6th Conference on Robot Learning (CoRL). Auckland, New Zealand; PMLR, 2022; 991—1002.
- [56] Bousmalis K, Vezzani G, Rao D, et al. RoboCat: A self-improving generalist agent for robotic manipulation. Transactions on Machine Learning Research. 2024.
- [57] Ball PJ, Smith L, Kostrikov I, et al. Efficient online reinforcement learning with offline data// Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, Hawaii, USA; ACM, 2023; 1577—1594.
- [58] Luo JL, Xu C, Wu J, et al. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. Science Robotics, 2025, 10(105).
- [59] Tan SH, Dou KR, Zhao Y, et al. Interactive post-training for vision-language-action models// Workshop on Foundation Models Meet Embodied Agents at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025.
- [60] Song WX, Chen JY, Ding PX, et al. Accelerating vision-language-action model integrated with action chunking *via* parallel decoding// Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China; IEEE, 2025; 13162—13169.
- [61] Dong ZB, Hao JY, Li PY, et al. DiffuserLite: Towards real-time diffusion planning// Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, BC, Canada; OpenReview.net, 2025; 122556—122583.
- [62] Ajay A, Du Y, Gupta A, et al. Is conditional generative modeling all you need for decision-making// Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda; OpenReview.net, 2023.

- [63] Pfaff N, Fu E, Binaglia J, et al. Scalable Real2Sim: Physics-aware asset generation *via* robotic pick-and-place setups// Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China; IEEE, 2025: 6296—6303.
- [64] Höfer S, Bekris K, Handa A, et al. Sim2Real in robotics and automation: Applications and challenges. *IEEE Transactions on Automation Science and Engineering*, 2021, 18(2): 398—400.
- [65] The Allen Institute for Artificial Intelligence. MolmoAct: An Action Reasoning Model that reasons in 3D space. (2025-08-12)/[2025-08-12]. <https://allenai.org/blog/molmoact>.
- [66] Shang JH, Schmeckpeper K, May BB, et al. Theia: Distilling diverse vision foundation models for robot learning// Proceedings of the 8th Conference on Robot Learning (CoRL). Munich, Germany: PMLR, 2025: 724—748.
- [67] Yuan YF, Cui HQ, Huang YT, et al. Embodied-R1: Reinforced embodied reasoning for general robotic manipulation// Proceedings of the 14th International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil: OpenReview.net, 2026.
- [68] Black K, Galliker MY, Levine S. Real-time execution of action chunking flow policies// Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS). San Diego, CA: OpenReview.net, 2025.

Key Challenges and Technologies of Embodied Intelligence

Jianye Hao* Hongyao Tang Yan Zheng

College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

Abstract In recent years, embodied intelligence has emerged as a critical interdisciplinary domain integrating multimodal perception, large model reasoning, and intelligent decision-making, demonstrating tremendous potential to expand the boundaries of intelligence and empower problem-solving in the real world. However, due to the complexity of embodied manipulations and the diversity of task scenarios, embodied intelligence research faces severe challenges, including difficulties in high-quality data collection, large embodied model construction, and training and inference optimization. These challenges hinder the development of embodied intelligence toward larger scales, stronger generality, and broader applications. This paper first introduces the relationship between general large models and large embodied models, and discusses the key challenges of embodied intelligence in three aspects: “data-model-optimization”. Then, this paper systematically reviews the core technologies of embodied intelligence, covering three main threads—embodied data and simulation, embodied “cerebellum-cerebrum” models, and embodied training and inference—with a focus on analyzing the trends in technological development. Finally, the paper discusses open challenges and future directions, including scalable data collection via integrated virtual-real collection, general embodied reasoning and manipulation, few-shot adaptation and real-time decision-making, aiming to promote technological breakthroughs and real-world applications of embodied intelligence.

Keywords artificial intelligence; intelligent decision-making technologies; embodied intelligence; large models; robotics

郝建业 天津大学智能与计算学部教授。主要研究方向为强化学习、具身智能。在*Nature Communications*等期刊, International Conference on Machine Learning(ICML)、Annual Conference on Neural Information Processing Systems (NeurIPS)、International Conference on Learning Representations(ICLR)等会议发表论文100余篇。主持国家自然科学基金青年科学基金项目(B类)等10余项科研项目。相关成果在国产工业基础软件智能化、自动驾驶、5G网络优化、工业物流调度等领域广泛落地应用。

(责任编辑 贾祖冰 张强)