特色专题

VLA 架构下的智能体演化:从机理建构到应用拓展

张慧[1],谢东锦[2],梁姝彤[1],李明轩[1],贾晓丰[3*],田永林[4],马思吉[5],李浩然[4],李浥东[1]

摘要 具身智能作为人工智能发展的新阶段,正在实现从“感知–认知”到“感知–认知–行动”一体化的跃迁。视觉–语言–动作(vision–language–action, VLA)模型通过统一视觉感知、语言理解与动作生成,为智能体在真实世界中的自主操作提供了关键技术路径。系统梳理了 VLA 技术的发展脉络与典型成果,总结了其架构范式,包括多模态感知输入、语义融合机制、强化与模仿学习、世界模型和多层次动作输出。结合自动驾驶、人机交互和工业装备等应用场景,进一步分析了 VLA 发展面临的核心挑战,包括数据资源匮乏、泛化与迁移能力不足、可解释性与算力压力等,并展望了未来趋势。

关键词 视觉–语言–动作模型;多模态学习;具身智能;大语言模型

人工智能正经历从“感知智能”“认知智能”向“具身智能(embodied intelligence)”的深刻演进。传统的人工智能系统能够识别图像、理解语言,但往往止步于“知”的层面,无法进一步完成“行”的环节。实现从“能感知”“会思考”到“可行动”的跃迁,是智能体走出虚拟空间、进入真实世界的关键所在。正如 Brooks 在 1991 年所提出的“智能源于与环境交互”的经典观点[1],具身智能的本质是通过感知、认知与行动三大能力的闭环融合,使智能体能够在开放环境中自主感知世界、理解意图并执行任务。围绕这一目标,视觉–语言–动作(vision–language–action, VLA)模型的提出成为支撑具身智能落地的核心技术路径。

1. 北京交通大学计算机科学与技术学院,北京 100044
2. 新疆大学软件学院,乌鲁木齐 830046
3. 北京市大数据中心,北京 101117
4. 中国科学院自动化研究所,北京 100190
5. 澳门科技大学创新工程学院,澳门 999078

在 VLA 概念提出之前,人工智能的发展长期呈现“分域演进”的格局:计算机视觉系统通过卷积神经网络(CNN)实现目标检测和分类,但无法理解语义;大语言模型(LLM)如 GPT–3/4 在自然语言理解和生成方面取得突破,却无法感知物理环境;机器人控制系统依赖强化学习或手工策略实现动作执行,却难以应对环境的不确定性和任务的开放性。尽管视觉–语言模型(VLM)通过大规模图文对齐学习打破了感知与语义之间的壁垒(如 OpenAI 的 CLIP[2]和 DeepMind 的 Flamingo[3]),实现了“看得懂语义”的跨模态理解,但仍然无法将语义指令转化为可执行的物理行为,智能体依旧停留在“能说不能做”的阶段。

具身智能研究的范式转向始于 2022 年 ChatGPT[4]的发布,其展现出的强大语义推理能力启发了研究者:如果能够将 LLM 的语义推理、VLM 的环境感知与机器人控制系统的动作生成统一到一个模型之中,智能体或许可以真正实现从理解到执行的全链条智能。这一设想在 2023 年由 Google DeepMind 提出的

收稿日期:2025–09–11;修回日期:2025–10–18
基金项目:国家自然科学基金青年项目(62203040);国家自然科学基金重点项目(62436010)
作者简介:张慧,副教授,研究方向为多智能体协同、具身智能等,电子信箱:huizhang1@bjtu.edu.cn;贾晓丰(通信作者),教授级高工,研究方向为复杂系统下的数据治理与数据智能,电子信箱:jiaxf@jxj.beijing.gov.cn
引用格式:张慧, 谢东锦, 梁姝彤, 等. VLA 架构下的智能体演化:从机理建构到应用拓展[J]. 科技导报, 2025, 43(20): 48–61; doi:10.3981/j.issn.1000-7857.2025.10.00077

Robotic Transformer 2[5](RT−2)首次实现。RT−2 将视觉、语言与动作 3 类 token 融合到统一的 Transformer 架构中,将机器人控制问题重构为自回归序列建模问题,从而显著提升了模型对未知物体与任务的零样本泛化能力。UC Berkeley 发布了 Octo[6]模型,利用超过 80 万条机器人演示数据实现了视觉−语言−动作联合训练,进一步推动了大规模数据驱动的具身智能发展。同年,相关研究在推动 VLA 技术向垂直领域深化应用方面取得系列突破,OpenDriveLab 发布的 DriveLM[7]将视觉语言模型与自动驾驶系统深度融合,通过构建语言问答式推理链与图结构化决策流程,显著提升了系统在未知场景的零样本适应能力;2024 年,理想汽车联合清华大学团队提出的 DriveVLM[8]建立了从场景描述到决策的递进推理机制,提升了复杂交通环境下的决策可靠性与透明度。针对动态环境中感知信息冗余度高、历史信息利用不足导致规划效率低等问题,北京大学团队提出的 Uni−NaVid[9]引入了在线令牌合并机制与多尺度观测编码策略实现动态导航任务中实时视频流的高效理解与鲁棒规划,为具身智能在动态环境中的长期自主运行奠定基础。

随着 VLA 技术的持续演进,其发展路径呈现出清晰的阶段性特征:2022—2023 年为奠基阶段,该阶段以实现视觉−语言−动作的端到端融合为主要目标,CLIPort[10]、Gato[11]、RT−1[12]等早期系统首次将视觉感知、语言理解与动作执行整合进统一模型,奠定了 VLA 的基础架构,并初步验证了多模态信息对任务执行的增强作用;2024 年进入认知增强阶段,研究重点转向推理能力与环境语义理解,VIMA[13]、VoxPoser[14]等模型采用链式推理和可供性理解机制,将语言推理与环境语义结合,使智能体能够更有效地进行任务分解、策略规划与决策解释,推动 VLA 从单纯感知控制向初步认知智能的跃升;2025 年以来进入泛化与安全部署阶段,SafeVLA[15]、Humanoid−VLA[16]等系统在形式化验证、全身控制、人机共融等方面不断突破,将发展重心拓展至模型安全性、跨场景适应能力与长期作业鲁棒性,标志着 VLA 从研究实验走向真实世界应用的关键跨越。与此同时,架构范式也逐步多样化:从早期融合模型(如 EF−VLA[17])到模仿人类双系统认知结构的 Groot N1[18],再到以 DCT 动作离散化与高效自回归实现 50 Hz 级高频控制与快速训练收敛的 π_0[19]、FAST[20],VLA 模型正在不断逼近“通用具身智能体”的形态。如 LeCun[21]所言:“真正的智能,不是预测下一个词,而是能够与世界互动、达成目标。”VLA 正是在这一理念的驱动下,逐步将深层语义推理与物理行动控制相融合,构建起从感知理解到环境交互的完整智能通路。

随着中国等国家在具身智能领域的战略布局不断加速,VLA 技术的基础性作用愈发凸显:它不仅是服务机器人、工业装备、自动驾驶等关键领域突破的技术支点,也是推动人工智能从“技术研发”走向“产业落地”的重要引擎。在方法论层面,VLA 强调的“感知−认知−行动”闭环与平行智能(parallel intelligence)[22−23]的“人工系统−计算实验−平行执行”闭环在框架上保持一致,并在具身交互、多模态融合和动态环境适应等关键场景中进一步实现了机制层的细化与技术落地。2 者在虚实互动、智能体协同与人机共融等方面形成互补,从而共同推动具身智能体系的系统化演进,为其构建新的理论基础与工程范式提供了支撑[24−25]。然而,尽管发展迅速,VLA 仍面临跨域泛化、长时序稳定性、可解释性及数据与算力成本等多重挑战。系统梳理 VLA 的核心技术、关键挑战与未来趋势,对于构建具身智能的理论体系与工程化路径具有重要意义。

1 VLA 架构范式

VLA 模型是具身智能的重要研究方向,其核心目标是将视觉感知、语言理解与动作执行整合为统一架构,从而实现多模态输入到多层次动作输出的端到端映射(图 1)。具体而言,VLA 系统通过视觉编码器、语言编码器与动作解码器的协同处理,融合来自机器人本体、外部环境及语言交互的多源感知信息,并结合强化学习、模仿学习与世界模型等优化机制,逐步生成从低阶控制命令到高层轨迹规划的动作策略,以支持智能体在动态环境中的自适应任务执行(图 2)。

1.1 多模态感知输入

VLA 模型通过整合机器人本体、环境感知及语言交互等多源信息,为任务执行提供了丰富的上下文支持,并构建了统一的语义表示空间,其输入模态主要包括 3 类:视觉感知、辅助传感信息与自然语言。

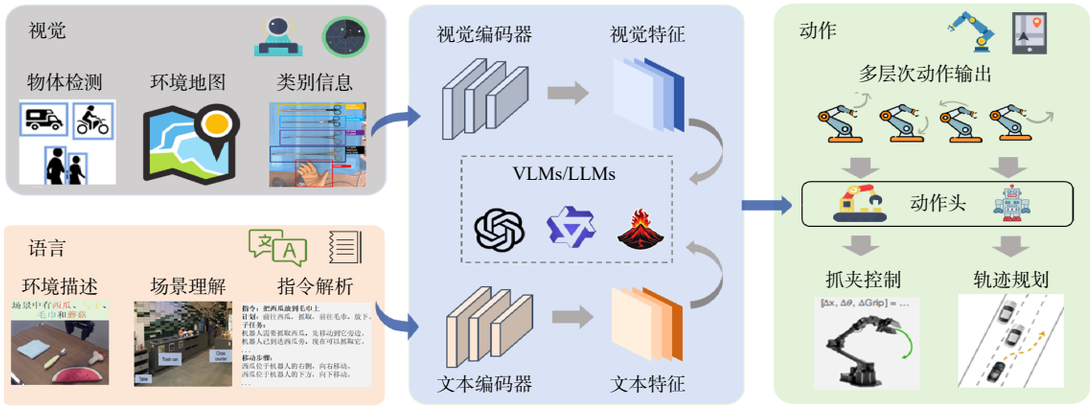

图 1 VLA 的架构范式

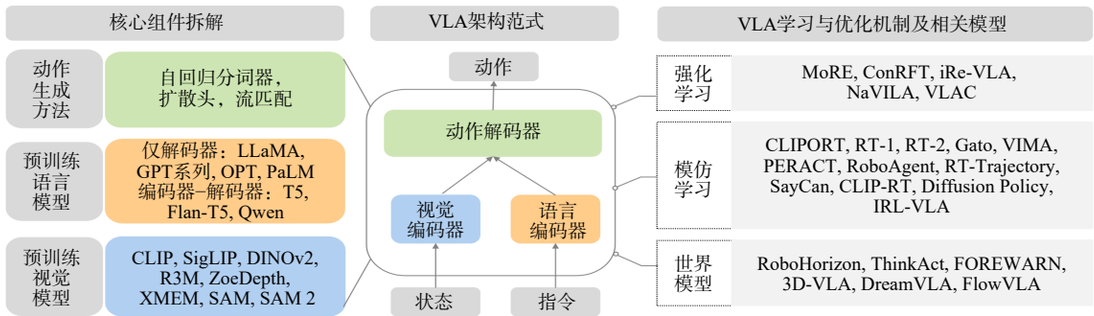

图 2 VLA 核心组件与学习优化机制

视觉感知是环境感知的核心输入，VLA 通常采用多模态视觉数据以全面捕捉场景的空间、几何与动态特征。具体而言，RGB(红绿蓝)图像提供物体的外观、颜色与纹理信息；RGB−D(红绿蓝−深度)图像在 RGB 基础上引入深度通道，补充三维位置与距离信息；点云数据通过稠密或稀疏的三维点集直接刻画物体表面几何形态；视频数据通过帧间时序关联记录场景动态演化，为长时序任务的理解与规划提供支持。为构建更全面的环境感知能力，VLA 系统还需整合来自机器人本体及外部环境的辅助传感器信息，以实现从感知到动作的连续校准与优化。本体感知数据包括关节角度、电机扭矩与夹具状态等，直接反映机器人的运动状态；触觉传感器数据捕捉微小的力变化与接触状态，适用于对力控精度要求高的精细操作场景。通过将辅助传感器信息与视觉数据时序同步，可有效提升系统感知准确性与动作协调性。自然语言输入作为人类意图的核心载体，在 VLA 中需支持多样化指令形式以满足不同交互需求。从指令粒度看，任务指令(如“整理书桌”)通常具有抽象性与高层次语义，需由模型自主分解为具体子任务，从而逐步执行以完成整体目标。相较之下，动作指令(如“把蓝色盒子放在书架第二层”)则具有明确的操作性，模型需解析其空间关系与操作约束，以实现语言到具体动作参数的直接映射。

1.2 核心组件拆解

VLA 的核心组件包括视觉编码器、语言编码器与动作解码器，3 者协同工作，完成从多模态输入至机器人动作的端到端映射。视觉编码器负责自高维图像或点云中提取紧凑的语义特征；语言编码器将自然语言指令转换为机器可理解的语义向量；动作解码器进一步将融合后的多模态表征转化为具体动作。

视觉编码器作为 VLA 环境感知的核心组件，通常采用预训练的视觉编码器，利用其在大规模数据集获得的泛化能力，降低对特定机器人场景数据的依赖。基于图像−文本对比学习的 CLIP[2]和 SigLIP[26]具备较强的语义对齐能力，已被广泛应用于 CLIPort[10]、SayCan[27]、CogACT[28]和 RDT−1B[29]等模型中。DINOv2[30]凭借其大规模自监督预训练策略与空间特征表达能力，成为 OpenVLA[31]、A3VLM[32]和 Hybrid VLA[33]等模型的核心视觉编码器。此外，针对特定感

知需求,VLA 还可集成多种预训练的专用视觉模型(如 ZoeDepth[34]、XMEM[35]、SAM[36]和 SAM 2[37]等),通过组合不同模型在空间感知、语义理解、深度估计和实例分割等方面的优势,全面满足机器人在操作任务中对环境感知的多维度需求。

VLA 的语言编码器作为指令理解的核心组件,通常基于预训练的 LLM 和预训练的 VLM 构建,能够将自然语言指令转化为结构化的语义嵌入表示,实现对指令意图的理解、复杂任务的解析及逻辑推理,为后续动作生成提供语义指导。目前 VLA 主流的语言编码器架构主要分为 2 类:仅解码器架构(decoder-only)和编码器-解码器架构(encoder-decoder)。仅解码器架构(如 LLaMA[38]、GPT 系列[39]及 PaLM[40]等)擅长处理开放式指令与多轮对话交互,其自回归生成机制能够基于上下文逐步预测词元,保持指令间的逻辑连贯性,因而被应用于 Gato[11]、Perceiver-Actor[41]、ACT[42]和 JARVIS-VLA[43]等模型中。编码器-解码器架构(如 T5[44]、Flan-T5[45]和 Qwen[46]等)在指令改写与任务分解方面表现优异,能够将抽象指令转化为具体操作流程,其代表性应用包括 Octo[6]、RT-1[12]、VIMA[13]以及 BLIP-2[47]等模型。

VLA 的动作解码器负责将融合后的多模态特征转化为机器人可执行的动作信号,其设计需根据任务需求选择合适的生成范式,以实现从低层控制命令到高层策略规划的灵活输出。目前主流的动作生成方法主要包括自回归分词器[9, 11-13, 48-50]、扩散头[6, 28-29, 51]以及流匹配[19, 52-54]等多种范式。自回归分词器借鉴了大语言模型的文本生成机制,将连续动作空间离散化为有限的“动作词元”,并通过自回归方式逐步生成动作序列,适用于桌面物品摆放等离散控制场景。以 RT-1[12]为例,该方法将机械臂动作拆解为机械臂运动,基座运动以及模式切换的 11 个不同的动作维度,并将每个动作维度离散化为 256 个区间,最后通过 Transformer 解码器逐词元预测动作。扩散头利用扩散模型的概率生成机制,将动作生成建模为“逐步去噪”的过程,尤其适用于高维、多模态的动作空间。代表性工作 Diffusion Policy[51]在动作隐空间中通过迭代去噪,生成符合物理约束的连续动作序列,显著提升了多种机器人(从 2 自由度到 14 自由度)在执行复杂任务时(如高精度操作与多模态决策)的任务完成率与泛化能力。流匹配通过学习动作空间的概率流场,能够直接生成符合物理规律的动作分布,从而避免扩散模型的多步迭代过程,特别适用于对实时性要求较高的任务场景。例如,π_0 模型[19]基于流匹配框架实现了 50 Hz 以上的高频动作生成,有效满足机器人实时控制需求。

1.3 学习与优化机制

为实现从多模态输入到动作输出的高效映射,VLA 模型不仅依赖于强大的感知与语言表征能力,还融合了多种学习与优化机制,共同推动动作策略的持续优化。当前的研究主要聚焦于 3 类方法:强化学习通过与环境交互来优化决策策略;模仿学习通过借鉴专家示范数据以快速掌握任务;世界模型则通过内部状态推演来提升规划的可靠性。

强化学习通过“尝试-反馈-调整”的循环机制,使机器人能够在与环境的持续交互中自主学习如何采取行动,以获取最优的长期回报。在 VLA 框架下,强化学习主要包含基于价值函数的方法[55-57]和基于策略梯度的方法[58-60]。基于价值函数的方法通过估计状态-动作对的期望累积奖励来指导策略优化,代表性方法如 MoRE[55]和 ConRFT[56]。该类方法核心在于训练 Q 函数评估每个状态-动作对的价值,并利用该信息优化策略网络,从而在多样化的视觉-语言输入下选择最优动作。基于策略梯度的方法通过梯度上升来直接更新策略网络,以提升期望的累积奖励,代表性方法如 NaVILA[58]和 iRe-VLA[59]。该类方法通过近端策略优化迭代更新策略参数,实现对复杂任务的高效学习。

模仿学习通过利用人类专家的示范轨迹,使 VLA 系统能够在有限探索下掌握任务策略。其核心思想是模仿专家在类似任务中的行为模式与决策分布,减少强化学习因大量随机探索而产生高昂代价,显著降低训练成本与时间开销。该方法适用于具备高质量专家数据且需快速部署的应用场景。行为克隆[10-13, 27, 41, 48-49, 51, 61]作为 VLA 模型中最核心的模仿学习方法之一,直接学习“状态-动作”之间的映射关系以复现专家行为。以 RT-1[12]与 RT-2[5]为例,将人类远程操作机械臂的轨迹视为〈图像,语言指令,动作〉样本,用于训练机械臂执行多种语言指令驱动的机器人操作任务。然而,行为克隆存在分布偏移与无法超越

专家水平的固有局限。为此,IRL-VLA[62]借鉴逆强化学习的思想,不再直接克隆动作,而是从大规模人类驾驶轨迹中反推隐含的奖励函数,并在强化学习过程中主动探索未见但回报更高的轨迹,实现更安全、高效的端到端自动驾驶。

世界模型通过“状态-动作→未来状态”的预测机制,使智能体能够在内部模拟环境进行推演与规划。这种机制相当于为智能体构建了一个“虚拟沙盘”,使其能够在执行前通过多次试错与策略评估,预测不同动作的影响,从而在现实环境中做出更优且更可靠的决策。在 VLA 模型中,常见的世界模型主要包括基于大语言模型的世界模型[63-64]和视觉世界模型[65-68]。大语言模型中蕴含着丰富的世界常识与因果知识,因而常被用于增强 VLA 模型,使智能体在推理和规划时能够引入更强的常识支撑。例如,RoboHorizon[64]通过使用 LLM 将长时任务拆解为子任务并生成对应的密集奖励函数,在潜在状态空间中构建循环世界模型,滚动推演“状态-动作→下一状态-奖励”序列,实现低样本、长时程的强化学习与策略优化。与基于 LLM 的世界模型不同,视觉世界模型更强调对环境动力学的显式建模,通过生成图像、视频或 3D 场景的未来状态来模拟环境动态变化,更贴近物理世界的真实过程。这类模型不仅能够捕捉连续的视觉演化,还能辅助具身智能体在潜在空间中进行更直观地环境预测与任务规划。例如,DreamVLA[67]显式预测未来世界的动态区域、深度图与语义特征,辅助机器人进行动作规划与环境推理。

1.4 多层次动作输出

VLA 系统的输出方式体现了其处理复杂任务时的抽象层次与决策粒度。随着研究的发展,输出形式已从早期的低阶动作控制逐渐扩展至更高层次的运动策略规划。

早期 VLA 系统主要输出关节角度、末端执行器位姿等低阶动作指令,通过将其建模为连续数值或离散动作编码,实现与机器人底层控制器或实时控制回路集成。这类方法虽然能够实现高精度的直接控制,但对感知误差敏感且在长时序任务中缺乏高层语义指导。随着任务复杂度的提升,研究重点逐渐转向运动策略规划,生成满足运动约束(如关节活动范围与速度限制),环境约束(如障碍物规避)及任务约束(如轨迹平滑性要求)的连续动作序列。运动策略规划显著增强了系统在长周期任务中的推理能力,并更好地融合了环境结构与任务语义信息。VLA 的分层输出实现了动作控制精度与策略规划能力的有效平衡:低阶动作控制确保执行层面的精确性,运动策略规划提供任务层面的决策指导,2 者协同使智能体具备应对复杂环境所需的连续空间推理与高精度执行能力。

2 VLA 场景工程与典型应用

2.1 场景工程

VLA 场景工程是为具身智能体设计和搭建一个集“视觉-语言-动作”交互于一体的“环境舞台”与“综合试验场”。其核心是在受控、接近真实的环境中,构建一个集“视觉-语言-动作”时序闭环、可执行动作空间与可量化评测于一体的集成化框架,用以系统性地训练、验证和复现智能体的复杂行为,使其不再局限于孤立的感知或语言任务。

1)环境构建。环境构建是场景工程的基石,其首要任务是根据任务领域选择合适的“环境底座”,这一选择直接决定了研发的上限与效率。针对家庭环境的移动导航与物体整理任务,可选用轻量级、高帧率的 Habitat[69];针对需要大规模并行训练的接触丰富、精细操控任务,ManiSkill3[70]以其 GPU 并行仿真与丰富的基准提供了强大支持;针对工业级应用和高保真度的仿真到现实(Sim2Real)迁移,NVIDIA Isaac Sim/Isaac Lab 凭借其稳定高效的物理引擎、逼真渲染和强大的合成数据生成能力成为首选方案。选定平台后,还需进行精细的工程化设置,包括统一全局坐标系、配置精确的物理属性(如质量、摩擦系数),并为环境中的对象赋予丰富的语义标签与可交互状态[71](如“可开启”“可容纳”),这是实现复杂任务程序化、规模化生成的前提[72]。

2)“感知-理解-行动”时序闭环构建。在可靠的环境建模基础上,搭建智能体的“感知-理解-行动”时序闭环是场景工程的核心。该闭环的构建始于多模态感知数据的同步,即所有输入流(如相机图像、深度图、语言指令)必须拥有严格对齐的时间戳。毫秒级的延迟或错位都可能导致任务失败[73-74],例如语言指令中的指代目标与视觉画面中的物体失配。其次是

鲁棒的语言接口设计,其需确保指令的可执行与可重现,并能处理多轮对话的上下文指代[75]。通过借鉴ALFRED[76]等成熟基准,可将模糊的自然语言任务系统性地分解为精确的机器可执行的目标状态序列。最后是分层的动作空间定义,这直接关系到决策的效率与精度。动作空间通常采用2层的定义:底层包含精细的关节力矩、末端速度等物理控制指令,保证动作的真实性与平滑度;高层封装了“抓取”“放置”等抽象技能,使顶层的规划算法(常与LLM结合)能高效地专注于任务的逻辑流程,从而有效提升解决长时程、复杂任务的能力[77-78]。

3)评估指标与策略。为科学评估并驱动智能体性能提升,需建立一套统一、全面的可量化指标体系。这套体系不仅要包含简单的“任务成功率”,还要涵盖更丰富的维度:对于导航任务,引入“成功率加权路径长度”(SPL)[79]来评估其效率;对于长序列操控任务,需记录子目标的完成率、操作时长乃至安全性(如碰撞次数)[80];同时,还需评估其决策与语言指令的逻辑一致性[76]。同时,在评估指标的基础上,还需要采用一系列系统化的评估策略,例如:分阶段评估,通过大规模仿真筛选与小样本物理验证相结合的方式,来平衡评估的效率与可信度[81];域自适应评估,其核心在于量化算法从仿真迁移到现实世界时的性能保持与恢复效率[82];虚实协同评估,利用与物理世界高度同步的虚拟副本,在保证安全和可控的前提下复现极端或危险场景[83];在线自适应评估,旨在实时测量智能体在任务执行过程中动态调整策略的反应能力[84]。这些方法结合评测指标共同构成了一套全面的评测体系,能够从多维度深入识别算法的优势与短板,从而有效指导其后续的优化迭代。此外,详尽的数据记录对实现实验可复现与数据驱动迭代至关重要。工程上要求为每一个任务回合(episode)保存完整的“数字档案”,内容包括全部多模态观测、语言指令、模型输出的动作序列与环境随机种子等[85],以便于调试和确定性重放。

2.2 数据资源

VLA具身智能的发展离不开大规模、高质量的数据资源支持。这些数据根据来源可主要分为真实环境数据、仿真环境数据,以及作为核心驱动力的自然语言与视觉指令数据。

1)真实环境数据。真实环境机器人数据集通过物理机器人在现实世界中采集,包含多模态传感信息及其控制指令。此类数据固有地反映了真实世界的复杂性,是检验和提升模型鲁棒性与泛化能力不可或缺的一环,但其采集成本高昂且伴随安全风险。

真实数据集的发展历程清晰地体现了具身智能研究的演进。早期的探索首先通过人类遥操作示教,系统性地验证了从演示中进行行为克隆(imitation learning)范式的可行性[86]。在此基础上,研究进入规模化阶段,数据集的轨迹数量扩大至数万级别,并引入自然语言作为任务指令,成功证实了在真实环境中训练“语言到动作”策略的有效性[87]。随着模仿学习范式的确立,研究重心转向提升模型的泛化能力:一方面通过融合多种机械臂平台的数据来解决跨设备视觉控制难题[88],另一方面则利用自动化策略采集近百万规模的多任务轨迹,为离线强化学习等前沿方法提供了理想的测试平台[89]。近年来,真实数据集的发展呈现出更加多元化和精细化的趋势,专注于更前沿的挑战。包括聚焦于家庭环境中的长尾细粒度任务[3]、为评估模型快速适应能力而设计的“一次示例学习”(one-shot learning)场景[90],以及通过分布式众包方式在数百个高度复杂的真实环境中采集数据,以验证模型在未知环境下的鲁棒性[91]。作为里程碑式的工作,大规模聚合数据集OXE[92]通过整合全球多个机构的数据,统一来自22种不同机器人形态的超百万段操控轨迹,实验证明,在其上训练的通用模型性能显著优于在单一数据集上训练的专家模型。这一工作有力证实了“跨具身正迁移”的可行性,标志着具身智能正迈向“大数据、大模型”的通用智能新时代。

2)仿真环境数据。仿真机器人数据集在虚拟环境中以程序化方式生成,具备成本低、可控性强、易于规模化等核心优势,为具身智能大模型的训练提供了数据基础。然而,其固有局限性主要在于“现实鸿沟”,即仿真器在物理保真度、视觉渲染等方面与现实世界存在差异。因此,在仿真训练中常常采用域随机化技术并结合真实数据微调,是当前提升模型迁移能力的主流范式。

仿真数据集正在快速发展,其趋势主要体现在任务复杂性、环境逼真度、数据规模化及生成自动化4个方面。首先,为了更好地评测智能体的推理与泛化能力,研究者致力于设计更复杂的任务接口,如引

入“多模态提示+自然语言”来支持组合式指令[13],并构建包含逼真场景和数千个高质量 3D 物体的大规模仿真框架[93]以缩小“现实鸿沟”。同时,面向抓取等核心技能,构建利用域随机化技术的超大规模合成数据集(达百亿帧级别)也逐渐成为一个重要方向,旨在通过海量数据训练出泛化能力强的基础模型[94]。近年来,一个更为前沿的趋势是利用大模型自动化生成高质量的专家数据。通过结合多模态大语言模型(MLLM)和“仿真闭环”的反馈机制,研究者能够为复杂的双臂操作等任务自动生成专家级轨迹[95],甚至能从极少数的人类演示中为数据稀缺的灵巧手自动合成大量高质量的操作轨迹[96]。这些自动化数据生成技术极大地降低了对人工示教的依赖,为智能体学习提供了更高效、可扩展的数据来源。

3)自然语言与视觉指令数据。将语言与视觉等高层语义指令融入机器人控制,是实现通用具身智能的核心。此类数据集旨在建立高层指令与底层感知及动作之间的深度映射,其核心挑战在于实现“指令到动作”的语义对齐与泛化推理。因此,其数据标注通常包含“提示–观测–动作”的完整信息链,以支持端到端的训练与评测。

在真实与仿真数据集中,融合开放词汇的语言与视觉指令已成为主流趋势。真实数据集如 RT−1[12]和聚合数据集 OXE[92]已验证了语言指令驱动真实机器人的可行性。仿真数据集如 VIMA[13]通过“多模态提示”机制,将文本与视觉标记结合,极大地增强了指令的表达力。为进一步丰富指令数据的规模与多样性,一个核心的前沿方向是利用 LLM 和 VLM 进行自动化的数据扩充[97]。这一探索主要有 2 种思路:一种是进行“后验式”标注,即利用 VLM 为海量已有的视觉–动作轨迹自动生成贴切、自然的文本指令,从而极大地丰富了语言–动作对的数量[98]。另一种更前沿的思路是构建完全自动化的数据生成流程,让智能体自主探索和学习。在这一范式中,VLM 负责观察真实场景(视觉)、主动提出有意义的任务(语言),并调用机器人自主执行和收集数据,形成“感知–思考–行动”的迭代闭环[99]。这些探索标志着由机器自动生成的、深度融合语言与视觉的指令数据,将成为未来数据集构建的核心方向,以覆盖更广阔、更复杂的任务语义空间。

2.3 典型场景应用

1)自动驾驶。VLA 模型作为多模态 AI 的前沿技术,正为自动驾驶领域带来一场关于“信任”与“协作”的根本性变革。它通过深度融合车辆的视觉感知、语言理解与行为决策,旨在解决传统自动驾驶系统可解释性差、泛化能力弱及人机交互难的长期困境。其核心价值在于将“端到端”的黑盒转变为一个可以沟通的“玻璃盒”,赋予车辆“自我解释”的能力。例如,系统不再沉默地执行决策,而是能清晰地叙述其决策链条:“前方有施工,建议提前并道”。如 DriveLM[7]等项目所展示的,这种将驾驶行为拆解为多步推理的能力,让每一次操作都有理有据,极大提升了乘坐的安全感。在此基础上,VLA 模型使车辆真正具备了“听懂人话、办对事”的能力,将人机交互从固定的命令集提升至流畅的自然语言对话。乘客可以提出“我容易晕车,请开得平稳一些”这类细致的驾驶风格偏好,系统能在确保安全的前提下动态调整策略[100−101]。这一趋势在消费级市场尤为明显,以华为问界、小鹏等为代表的头部品牌,均在新车型中将“AI 代驾”或“智慧司乘”作为核心卖点,其系统不仅能执行导航,更能就路线选择、实时路况进行主动的语音沟通,正逐渐成为行业标配。如 DriveGPT4[102],已开始将这种可解释的语言能力融入车辆的闭环控制,推动技术从演示走向实际评测[103]。

这种高级交互的背后,是 VLA 模型对开放世界更深层次的理解力。真实道路充满了传统检测器无法覆盖的“长尾场景”(如“一个车门半开的快递货车”),借助 DriveVLM[8]等工作的开放词汇理解能力,系统能灵活处理这些稀有但关键的目标。同时,ORION[104]等模型还将“记忆”纳入决策环路,能记住几分钟前的路况信息,从而做出更连贯、更具预见性的判断,其最直观的改善就是车辆大幅减少了无端的犹豫和急刹。这些日益成熟的技术正走出实验室,开始接受市场与法规的双重检验。例如,上海浦东新区自 2025 年 8 月起已向公众开放全无人 Robotaxi 试运营,车内会主动向乘客解释行程与路况;Waymo 也宣布 2025 年持续扩张其公开服务规模。这意味着“可解释、可沟通”的体验正被更大范围的真实乘客检验。同时,国家标准《智能网联汽车 自动驾驶数据记录系统》预计于 2026 年实施,其对自动驾驶行为的数

据记录要求,将从法规层面推动“能讲清楚、留痕可查”成为强制性要求。

2)机器人控制。在机器人控制领域,VLA 模型的核心在于将人类的自然语言指令与机器人的视觉环境理解深度融合,打通从多模态感知到物理执行的统一闭环,让机器人能“按人话把事办到位”。当前,这一理念正通过多种主流架构落地:既有像谷歌 RT-2[5]那样将“看与说”直接映射为行动的端到端路径,高效处理日常任务;也有类似 SayCan[27]的思路,先将复杂指令分解为稳妥的子步骤,确保长序列任务的完成度;还有如同 VoxPoser[14]强调的,先在三维空间中进行推理再规划轨迹,以更好地应对陌生环境。

这些在实验室验证的技术能力正快速渗透到产业一线。在制造与物流领域,Covariant 的 RFM-1[105]基础模型被用于处理复杂的分拣作业;亚马逊仓内机器人数量已突破 100 万台,并上线了生成式 AI 调度模型,其规模化运营为人机协作提供了最好的试金石。在家居领域,美的集团发布面向家务场景的大模型“美言”,智元机器人等通用人形机器人公司也已积累百万级真实操作数据,致力于打造通用服务机器人。在医疗领域,Ekso Bionics 与 ReWalk 等康复机器人企业融合感知-决策-控制闭环,为卒中及脊髓损伤患者提供个性化步态训练方案;在护理场景,Mabu 等社交机器人具备情感交互与用药提醒能力,并逐步接入家庭健康监测系统。在教育领域,机器人正从辅助工具升级为具备个性化教学能力的交互伙伴。以软银机器人旗下的 NAO 和 Pepper 为例,它们通过整合多模态情感识别与自适应学习算法,能够根据学生情绪状态动态调整教学节奏;乐高教育推出的 SPIKE Prime 机器人套件将计算思维训练融入实体搭建,使抽象编程概念在动手实践中具象化。与此同时,这种强劲的产业势头正与日益明确的行业共识相呼应。在 2025 年 8 月举办的世界机器人大会(WRC 2025)上,“AI+Robotics”的应用生态成为焦点,强调机器人不仅要会执行,更要“可对话、可理解”。紧接着,国际机器人联合会(IFR)的立场文件更明确指出“人形将是机器人领域的下一件大事”。这些动向表明一个强调“沟通能力”“执行有效性”与“行为可追溯”的评测及运营环境正在快速形成。

3)GUI 智能体交互。在 GUI(图形用户界面)智能体交互领域,VLA 模型的目标是实现从高级人类指令到低级界面操作的无缝转换,即“看懂屏幕、听懂人话、办好事情”。它需要像人类一样解析图形用户界面,并将“帮我预订一张明天去上海的特价机票”这类口语化目标,自主拆解为一连串精准的点击、输入和滚动等跨应用操作。因此,与自动驾驶或物理机器人不同的是,其评估重点并非物理世界的稳定性,而是界面理解的准确性与任务流程的完成度。

为确保模型真正“会用电脑”而非利用代码捷径,学界与业界已建立起一系列贴近现实的评测基准。针对早期智能体无法处理图形验证码等视觉元素的问题,现有工作已提供了真实可交互的网站环境,强制模型必须从原始截图中定位并操作元素,极大地推动了模型的视觉能力[106-107]。同时,高质量的数据同样至关重要,Mind2Web[108]项目构建了涵盖多个领域、附带完整专家操作轨迹的数据底座,让模型能学习长程规划。同时,为实现公平对比,还需要将分散的网页基准统一到标准化的生态中[109]。智能体在这些基准中的决策质量,高度依赖其感知前端的准确性。一方面,Google 的 ScreenAI[110]这类面向用户界面的视觉语言模型,能作为强大的“感知前端”,为智能体提供精准的版面解析和“屏幕问答”等可靠的视觉依据。另一方面,VLA 的操作执行能力正从网页扩展到移动端。AppAgent[111]证明了仅用基础动作就能操控复杂应用程序(APP),而 ShowUI[112]进一步将“视觉-语言-行动”统一到单个模型中,验证了纯视觉驱动的通用 GUI 智能体是可行的。在产业侧,Apple 通过 App Intents 加强 Siri 对第三方 APP 的操作,正推动一个由自然指令闭环控制的移动应用生态趋于成型。

3 挑战与展望

3.1 当前挑战

3.1.1 数据资源匮乏与成本高昂

VLA 的发展严重依赖大规模跨模态数据,在真实机器人环境中,获取包含语言、视觉与动作对应关系的高质量数据难度极大。尤其是长时序、多步骤的复杂任务,以及带有失败样本的操作过程,往往很难被系统化采集。即便已有部分数据集被用于训练 VLA 模型,但存在场景单一、任务覆盖不足和标注噪声等

问题,使得模型在面对开放环境时缺乏适应性。此外,跨设备、跨传感器采集的数据常常缺乏一致性,进一步加剧了泛化难度。这种数据瓶颈限制了 VLA 的可扩展性和落地速度。

3.1.2 泛化与迁移能力不足

尽管现有 VLA 模型在特定实验场景中表现优异,但其泛化能力仍显不足。在面对新的环境、新的物体,或者在不同硬件平台上迁移时,模型性能往往急剧下降。更重要的是,当前 VLA 在组合泛化和因果抽象方面能力有限,难以通过已有经验灵活应对未见过的复杂任务。例如,一个能够完成“搬起杯子”的模型,未必能顺利完成“先搬开书本再拿起杯子”的组合任务。这种迁移困难使得 VLA 在现实世界的鲁棒性和适应性仍然不足。

3.1.3 可解释性、安全性与算力压力

VLA 强调端到端的统一建模,但这也带来了可解释性不足的问题。模型的决策过程大多是黑箱式的,难以为人类提供透明的依据,这在医疗、交通等高风险领域极易引发安全隐患。与此同时,大规模模型的训练和推理对算力依赖极高,既增加了开发和部署成本,也在机器人本体上带来实时性和能耗压力。对于需要低延迟响应和长时间独立运行的应用场景,这些问题尤其突出。

3.2 未来展望

3.2.1 数据驱动与知识驱动结合

未来 VLA 的突破有赖于数据与知识的深度融合。一方面,可以通过仿真环境大规模生成多样化数据,再结合真实场景示教与合成数据扩充,降低采集成本并提高覆盖率。另一方面,引入符号推理、物理规律和人类常识,有助于缓解单纯依赖数据驱动的局限,提升模型对新任务的泛化能力。通过数据与知识的互补,VLA 有望逐步具备更强的学习效率与跨场景适应性。

3.2.2 世界模型与 VLA 的深度耦合

世界模型为具身智能提供了对环境的预测与推演能力,与 VLA 的感知−语言−动作一体化建模结合后,将显著提升智能体的长程规划与风险评估能力。未来的 VLA 不应仅停留在“感知−语言−动作”的水平,而是要能够“预测−推演−规避风险”,在动态环境中展现更强的稳健性。这种融合将为自动驾驶、手术辅助等需要前瞻性决策的任务提供新的解决方案。

3.2.3 可信赖与高效的智能体构建

未来 VLA 的应用前景依赖于其在安全性、透明性与效率上的提升。研究者需要开发可解释的决策机制,使系统在执行任务时能够清晰地展示判断依据,并在出现异常时具备回退或纠错机制。同时,结合软硬件协同优化,通过模型压缩、参数高效微调与云−边−端协作,降低能耗并提升运行效率。在保证可信性和高效性的前提下,VLA 有望在医疗、教育、交通和服务机器人等高风险或高实时性领域实现大规模落地。

3.2.4 开放基准与评测体系建设

随着 VLA 模型在具身智能领域的快速发展,如何系统、客观地评估模型的理解、推理与执行能力成为关键问题。近年来,EmbodiedBench[113]、REAL−Bench[114]、VLABench[115]等开放基准相继推出,覆盖从语言指令理解、空间感知到动作规划与物理执行的多维能力,为模型能力对比与瓶颈分析提供了统一框架。然而,现有基准仍主要依赖模拟环境,缺乏真实世界交互与跨场景迁移能力的评测。未来还需进一步完善任务多样性与复杂性,引入 Sim−to−Real 验证、安全与伦理指标,并建立开放、公正、可复现的评测生态,以推动通用具身智能体的可持续发展。

参考文献(References)

[1] Brooks R A. Intelligence without representation[J]. Artificial Intelligence, 1991, 47(1/2/3): 139−159.

[2] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. Oxford: PMLR, 2021: 8748−8763.

[3] Alayrac J B, Donahue J, Luc P, et al. Flamingo: A visual language model for few−shot learning[C]//Conference on Neural Information Processing Systems. New Orleans, Louisiana, US: Curran Associates, Inc., 2022: 23716−23736.

[4] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J/OL]. OpenAI, [2025−09−18]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[5] Zitkovich B, Yu T, Xu S, et al. RT−2: Vision−language−

action models transfer web knowledge to robotic control[C]// Conference on Robot Learning. Atlanta: PMLR, 2023: 2165−2183.

[6] Ghosh D, Walke H R, Pertsch K, et al. Octo: An open−source generalist robot policy[C]//Proceedings of Robotics: Science and Systems XX. Robotics: Science and Systems Foundation, 2024: 1−10.

[7] Sima C H, Renz K, Chitta K, et al. DriveLM: Driving with-Graph visual question answering[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2025: 256−274.

[8] Tian X, Gu J, Li B, et al. DriveVLM: The convergence of autonomous driving and large vision−language models[J]. arXiv preprint, 2024, arXiv: 2402.12289.

[9] Zhang J, Wang K, Wang S, et al. Uni−NaVid: A video−based vision−language−action model for unifying embodied navigation tasks[J]. arXiv preprint, 2024, arXiv: 2412.06224.

[10] Shridhar M, Manuelli L, Fox D. CLIPort: What and where pathways for robotic manipulation[C]//Conference on Robot Learning. Auckland: PMLR, 2022: 894−906.

[11] Reed S, Zolna K, Parisotto E, et al. A generalist agent[J]. arXiv preprint, 2022, arXiv: 2205.06175.

[12] Brohan A, Brown N, Carbajal J, et al. RT−1: Robotics transformer for real−world control at scale[C]//Proceedings of Robotics: Science and Systems XIX. Robotics: Science and Systems Foundation, 2023: 1−18.

[13] Jiang Y, Gupta A, Zhang Z, et al. VIMA: General robot manipulation with multimodal prompts[J]. Conference on Machine Learning, 2023, 202: 14975−15022.

[14] Huang W, Wang C, Zhang R, et al. VoxPoser: Composable 3D value maps for robotic manipulation with language models[J]. Proceedings of Machine Learning Research, 2023(229): 1461−1476.

[15] Zhang B, Zhang Y, Ji J, et al. SafeVLA: Towards safety alignment of vision−language−action model via constrained learning[J]. arXiv preprint, 2025, arXiv: 2503.03480.

[16] Ding P, Ma J, Tong X, et al. Humanoid−VLA: Towards universal humanoid control with visual integration[J]. arXiv preprint, 2025, arXiv: 2502.14795.

[17] Huang H, Liu F C, Fu L T, et al. Early fusion helps vision language action models generalize better[J/OL]. arXiv, [2025−09−18]. https://arxiv.org/abs/2410.15310.

[18] Bjorck J, Castañeda F, Cherniadev N, et al. GR00T N1: An open foundation model for generalist humanoid robots[J]. arXiv preprint, 2025, arXiv: 2503.14734.

[19] Black K, Brown N, Driess D, et al. π_0: A vision−language−action flow model for general robot control[J]. arXiv preprint, 2024, arXiv: 2410.24164.

[20] Chen Z, Wang J, Wang W, et al. Fast: Faster arbitrarily−shaped text detector with minimalist kernel representation[J]. arXiv preprint, 2021, arXiv: 2111.02394.

[21] LeCun Y. A path towards autonomous machine intelligence version 0.9. 2, 2022−06−27[J]. Open Review, 2022, 62(1): 1−62.

[22] 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.

[23] Wang X X, Yang J, Liu Y H, et al. Parallel intelligence in three decades: A historical review and future perspective on ACP and cyber−physical−social systems[J]. Artificial Intelligence Review, 2024, 57(9): 255.

[24] 李柏, 郝金第, 孙跃硕, 等. 平行智能范式视角下的视觉−语言−动作模型发展现状与展望[J]. 智能科学与技术学报, 2025, 7(3): 290−303.

[25] 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.

[26] Zhai X H, Mustafa B, Kolesnikov A, et al. Sigmoid loss for language image pre−training[C]//Proceedings of IEEE/CVF International Conference on Computer Vision(ICCV). New York: IEEE, 2023: 11975−11986.

[27] Ahn M, Brohan A, Brown N, et al. Do as i can, not as i say: Grounding language in robotic affordances[J]. arXiv preprint, 2022, arXiv: 2204.01691.

[28] Li Q, Liang Y, Wang Z, et al. CogACT: A foundational vision−language−action model for synergizing cognition and action in robotic manipulation[J]. arXiv preprint, 2024, arXiv: 2411.19650.

[29] Liu S, Wu L, Li B, et al. RDT−1B: A diffusion foundation model for bimanual manipulation[J]. arXiv preprint, 2024, arXiv: 2410.07864.

[30] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning robust visual features without supervision[J]. arXiv preprint, 2023, arXiv: 2304.07193.

[31] Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: An open−source vision−language−action model[J]. arXiv preprint, 2024, arXiv: 2406.09246.

[32] Huang S, Chang H, Liu Y, et al. A3VLM: Actionable articulation−aware vision language model[J]. arXiv preprint, 2024, arXiv: 2406.07549.

[33] Liu J, Chen H, An P, et al. HybridVLA: Collaborative diffusion and autoregression in a unified vision−language−action model[J]. arXiv preprint, 2025, arXiv: 2503.10631.

[34] Bhat S F, Birkl R, Wofk D, et al. ZoeDepth: Zero−shot transfer by combining relative and metric depth[J]. arXiv preprint, 2023, arXiv: 2302.12288.

[35] Cheng H K, Schwing A G. XMem: Long−term video object segmentation withanAtkinson−shiffrin memory model[C]//

Computer Vision−ECCV 2022. Cham: Springer Nature Switzerland, 2022: 640−658.

[36] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]// Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2023: 4015−4026.

[37] Ravi N, Gabeur V, Hu Y T, et al. SAM 2: Segment anything in images and videos[J]. arXiv preprint, 2024, arXiv: 2408.00714.

[38] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint, 2023, arXiv: 2302.13971.

[39] Achiam J, Adler S, Agarwal S, et al. GPT−4 Technical Report[J]. arXiv preprint, 2023, arXiv: 2303.08774.

[40] Chowdhery A, Narang S R, Devlin J, et al. PaLM: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1−113.

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

[42] Zhao T Z, Kumar V, Levine S, et al. Learning fine−grained bimanual manipulation with low−cost hardware[J]. arXiv preprint, 2023, arXiv: 2304.13705.

[43] Li M Y, Wang Z H, He K C, et al. JARVIS−VLA: Post−training large−scale vision language models to play visual games with keyboards and mouse[J]. arXiv preprint, 2025, arXiv: 2503.16365.

[44] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text−to−text transformer[J]. Journal of Machine Learning Research, 2020, 21(140): 1−67.

[45] Chung H W, Hou L, Longpre S, et al. Scaling instruction−finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1−53.

[46] Bai J, Bai S, Chu Y, et al. Qwen technical report[J]. arXiv preprint, 2023, arXiv: 2309.16609.

[47] Li J, Li D, Savarese S, et al. BLIP−2: Bootstrapping language−image pre−training with frozen image encoders and large language models[C]//International Conference on Machine Learning. Honolulu: PMLR, 2023: 19730−19742.

[48] Bharadhwaj H, Vakil J, Sharma M, et al. RoboAgent: Generalization and efficiency in robot manipulation *via* semantic augmentations and action chunking[C]//Proceedings of IEEE International Conference on Robotics and Automation (ICRA). New York: IEEE, 2024: 4788−4795.

[49] Li S L, Wang J, Dai R, et al. RoboNurse−VLA: Robotic scrub nurse system based on vision−language−action model[J]. arXiv preprint, 2024, arXiv: 2409.19590.

[50] Gu J Y, Kirmani S, Wohlhart P, et al. RT−trajectory: Robotic task generalization *via* hindsight trajectory sketches[J]. arXiv preprint, 2023, arXiv: 2311.01977.

[51] Chi C, Xu Z J, Feng S Y, et al. Diffusion policy: Visuomotor policy learning *via* action diffusion[J]. The International Journal of Robotics Research, 2025, 44(10/11): 1684−1704.

[52] Intelligence P, Black K, Brown N, et al. $\pi_{0.5}$: A vision−language−action model with open−world generalization[J]. arXiv preprint, 2025, arXiv: 2504.16054.

[53] Shukor M, Aubakirova D, Capuano F, et al. SmolVLA: A vision−language−action model for affordable and efficient robotics[J]. arXiv preprint, 2025, arXiv: 2506.01844.

[54] Driess D, Springenberg J T, Ichter B, et al. Knowledge insulating vision−language−action models: Train fast, run fast, generalize better[J]. arXiv preprint, 2025, arXiv: 2505.23705.

[55] Zhao H, Song W X, Wang D L, et al. MoRE: Unlocking scalability in reinforcement learning for quadruped vision−language−action models[J]. arXiv preprint, 2025, arXiv: 2503.08007.

[56] Chen Y H, Tian S, Liu S G, et al. ConRFT: A reinforced fine−tuning method for VLA models *via* consistency policy[J]. arXiv preprint, 2025, arXiv: 2502.05450.

[57] Xu K, Zhao S, Zhou Z, et al. A joint modeling of vision−language−action for target−oriented grasping in clutter[J]. arXiv preprint, 2023, arXiv: 2302.12610.

[58] Cheng A C, Ji Y, Yang Z, et al. NaVILA: Legged robot vision−language−action model for navigation[J]. arXiv preprint, 2024, arXiv: 2412.04453.

[59] Guo Y, Zhang J, Chen X, et al. Improving vision−language−action model with online reinforcement learning [J]. arXiv preprint, 2025, arXiv: 2501.16664.

[60] Zhai S, Zhang Q, Zhang T, et al. A vision−language−action−critic model for robotic real−world reinforcement learning[J]. arXiv preprint, 2025, arXiv: 2509.15937.

[61] Kang G C, Kim J, Shim K, et al. CLIP−RT: Learning Language−Conditioned Robotic Policies from Natural Language Supervision[J]. arXiv preprint, 2024, arXiv: 2411.00508.

[62] Jiang A, Gao Y, Wang Y, et al. IRL−VLA: Training an vision−language−action policy via reward world model[J]. arXiv preprint, 2025, arXiv: 2508.06571.

[63] Huang C P, Wu Y H, Chen M H, et al. ThinkAct: Vision−language−action reasoning via reinforced visual latent planning[J]. arXiv preprint, 2025, arXiv: 2507.16815.

[64] Chen Z X, Huo J, Chen Y T, et al. RoboHorizon: An LLM−assisted multi−view world model for long−horizon robotic manipulation[J]. arXiv preprint, 2025, arXiv: 2501.06605.

[65] Wu Y, Tian R, Swamy G, et al. From foresight to forethought: VLm−in−the−loop policy steering via latent alignment[J]. arXiv preprint, 2025, arXiv: 2502.01828.

[66] Zhen H, Qiu X, Chen P, et al. 3D−VLA: A 3D vision−language−action generative world model[J]. arXiv preprint, 2024, arXiv: 2403.09631.

[67] Zhang W Y, Liu H S, Qi Z K, et al. DreamVLA: A vision−language−action model dreamed with comprehensive world knowledge[J]. arXiv preprint, 2025, arXiv: 2507.04447.

[68] Zhong Z, Yan H, Li J, et al. FlowVLA: Thinking in motion with a visual chain of thought[J]. arXiv preprint, 2025, arXiv: 2508.18269.

[69] Szot A, Clegg A, Undersander E, et al. Habitat 2.0: Training home assistants to rearrange their habitat[C]//Conference on Neural Information Processing Systems. New York: Curran Associates, Inc. , 2021, 34: 251−266.

[70] Tao S, Xiang F, Shukla A, et al. ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI[J]. arXiv preprint, 2024, arXiv: 2410.00425.

[71] Li C S, Xia F, Martín−Martín R, et al. iGibson 2.0: Object−centric simulation for robot learning of everyday household tasks[J]. arXiv preprint, 2021, arXiv: 2108.03272.

[72] Li C, Zhang R, Wong J, et al. BEHAVIOR−1K: A human−centered, embodied ai benchmark with 1, 000 everyday activities and realistic simulation[J]. arXiv preprint, 2024, arXiv: 2403.09227.

[73] Bray N, Boeding M, Hempel M, et al. A latency composition analysis for telerobotic performance insights across various network scenarios[J]. Future Internet, 2024, 16(12): 457.

[74] Kamtam S B, Lu Q, Bouali F, et al. Network latency in teleoperation of connected and autonomous vehicles: A review of trends, challenges, and mitigation strategies[J]. Sensors, 2024, 24(12): 3957.

[75] Shridhar M, Thomason J, Gordon D, et al. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 10740−10749.

[76] Padmakumar A, Thomason J, Shrivastava A, et al. TEACh: Task−driven embodied agents that chat[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 2017−2025.

[77] Vezhnevets A S, Osindero S, Schaul T, et al. FeUdal networks for hierarchical reinforcement learning[C]//International Conference on Machine Learning. Oxford: PMLR, 2017: 3540−3549.

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

[79] Anderson P, Chang A, Chaplot D S, et al. On evaluation of embodied navigation agents[J]. arXiv preprint, 2018, arXiv: 1807.06757.

[80] Ray A, Achiam J, Amodei D. Benchmarking safe exploration in deep reinforcement learning[J]. arXiv preprint, 2019, arXiv: 1910.01708.

[81] Peng X B, Andrychowicz M, Zaremba W, et al. Sim−to−real transfer of robotic control with dynamics randomization[C]//Proceedings of IEEE International Conference on Robotics and Automation (ICRA). New York: IEEE, 2018: 3803−3810.

[82] Tobin J, Fong R, Ray A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York: IEEE, 2017: 23−30.

[83] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An open urban driving simulator[C]//Conference on Robot Learning. Mountain View: PMLR, 2017: 1−16.

[84] Kumar A, Fu Z, Pathak D, et al. RMA: Rapid Motor adaptation for legged robots[J]. arXiv preprint, 2021, arXiv: 2107.04034.

[85] Henderson P, Islam R, Bachman P, et al. Deep reinforcement learning that matters[C]//AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018.

[86] Sharma P, Mohan L, Pinto L, et al. Multiple interactions made easy (MIME): Large scale demonstrations data for imitation[C]//Conference on Robot Learning. Zürich: PMLR, 2018: 906−915.

[87] Jang E, Irpan A, Khansari M, et al. BC−Z: Zero−shot task generalization with robotic imitation learning[C]//Conference on Robot Learning. Auckland: PMLR, 2022: 991−1002.

[88] Dasari S, Ebert F, Tian S, et al. RoboNet: Large−scale multi−robot learning[J]. arXiv preprint, 2019, arXiv: 1910.11215.

[89] Kalashnikov D, Varley J, Chebotar Y, et al. MT−Opt: Continuous multi−task robotic reinforcement learning at scale[J]. arXiv preprint, 2021, arXiv: 2104.08212.

[90] Fang H S, Fang H J, Tang Z Y, et al. RH20T: A comprehensive robotic dataset for learning diverse skills in one−shot[J]. arXiv preprint, 2023, arXiv: 2307.00595.

[91] Khazatsky A, Pertsch K, Nair S, et al. DROID: A large−scale in−the−wild robot manipulation dataset[J]. arXiv preprint, 2024, arXiv: 2403.12945.

[92] Vuong Q, Levine S, Walke H R, et al. Open X−embodiment: Robotic learning datasets and rt−x models[C]//Conference on

Neural Information Processing Systems. New York: Curran Associates, Inc., 2023.

[93] Nasiriany S, Maddukuri A, Zhang L, et al. RoboCasa: Large−scale simulation of everyday tasks for generalist robots[J]. arXiv preprint, 2024, arXiv: 2406.02523.

[94] Deng S L, Yan M, Wei S L, et al. GraspVLA: A grasping foundation model pre−trained on billion−scale synthetic action data[J]. arXiv preprint, 2025, arXiv: 2505.03233.

[95] Chen T, Chen Z, Chen B, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation[J]. arXiv preprint, 2025, arXiv: 2506.18088.

[96] Jiang Z Y, Xie Y Q, Lin K, et al. DexMimicGen: Automated data generation for bimanual dexterous manipulation *via* imitation learning[C]//Proceedings of IEEE International Conference on Robotics and Automation (ICRA). New York: IEEE, 2025: 16923−16930.

[97] Duan J, Yuan W, Pumacay W, et al. Manipulate−anything: Automating real−world robots using vision−language models[J]. arXiv preprint, 2024, arXiv: 2406.18915.

[98] Xiao T, Chan H, Sermanet P, et al. Robotic skill acquisition *via* instruction augmentation with vision−language models[J]. arXiv preprint, 2022, arXiv: 2211.11736.

[99] Ahn M, Dwibedi D, Finn C, et al. AutoRT: Embodied foundation models for large scale orchestration of robotic agents[J]. arXiv preprint, 2024, arXiv: 2401.12963.

[100] Cui C, Ma Y S, Cao X, et al. Drive as you speak: Enabling human−like interaction with large language models in autonomous vehicles[C]//Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). New York: IEEE, 2024: 902−909.

[101] Cui C, Yang Z C, Zhou Y P, et al. Personalized autonomous driving with large language models: Field experiments[C]//Proceedings of IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). New York: IEEE, 2024: 20−27.

[102] Xu Z H, Zhang Y J, Xie E Z, et al. DriveGPT4: Interpretable end−to−end autonomous driving *via* large language model[J]. IEEE Robotics and Automation Letters, 2024, 9(10): 8186−8193.

[103] Xu Z H, Bai Y, Zhang Y J, et al. DriveGPT4−V2: Harnessing large language model capabilities for enhanced closed−loop autonomous driving[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2025: 17261−17270.

[104] Fu H, Zhang D, Zhao Z, et al. ORION: A holistic end−to−end autonomous driving framework by vision−language instructed action generation[J]. arXiv preprint, 2025, arXiv: 2503.19755.

[105] Sohn A, Nagabandi A, Florensa C, et al. Introducing RFM-1: Giving robots human-like reasoning capabilities[EB/OL]. (2024−03−11)[2025−09−11]. https://covariant.ai/insights/introducing-rfm-1-giving-robots-human-like-reasoning-capabilities.

[106] Zhou S, Xu F F, Zhu H, et al. WebArena: A realistic web environment for building autonomous agents[J]. arXiv preprint, 2023, arXiv: 2307.13854.

[107] Koh J Y, Lo R, Jang L, et al. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks[J]. arXiv preprint, 2024, arXiv: 2401.13649.

[108] Deng X, Gu Y, Zheng B, et al. Mind2Web: Towards a generalist agent for the web[C]//Conference on Neural Information Processing Systems. New York: Curran Associates, Inc. , 2023, 36: 28091−28114.

[109] Chezelles D, Le Sellier T, Shayegan S O, et al. The browsergym ecosystem for web agent research[J]. arXiv preprint, 2024, arXiv: 2412.05467.

[110] Baechler G, Sunkara S, Wang M, et al. ScreenAI: A vision−language model for UI and infographics understanding[J]. arXiv preprint, 2024, arXiv: 2402.04615.

[111] Zhang C, Yang Z, Liu J X, et al. AppAgent: Multimodal agents as smartphone users[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2025: 1−20.

[112] Lin K Q, Li L, Gao D, et al. ShowUI: One Vision−Language−Action Model for GUI Visual Agent[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville: IEEE, 2025: 19498−19508.

[113] Yang R, Chen H, Zhang J, et al. EmbodiedBench: Comprehensive benchmarking multi−modal large language models for vision−driven embodied agents[J]. arXiv preprint, 2025, arXiv: 2502.09560.

[114] Jin P, Huang D, Li C, et al. RealBench: Benchmarking verilog generation models with real−world ip designs[J]. arXiv preprint, 2025, arXiv: 2507.16200.

[115] Zhang S, Xu Z, Liu P, et al. VLABench: A large−scale benchmark for language−conditioned robotics manipulation with long−horizon reasoning tasks[C]// International Conference on Computer Vision. Honolulu, Hawaii, 2025: 11142−11152.

Agent evolution under the VLA architecture: From mechanistic construction to application expansion

ZHANG Hui[1], XIE Dongjin[2], LIANG Shutong[1], LI Mingxuan[1], JIA Xiaofeng[3*], TIAN Yonglin[4], MA Siji[5], LI Haoran[4], LI Yidong[1]

1. School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China
2. School of Software, Xinjiang University, Urumqi 830046, China
3. Beijing Big Data Centre, Beijing 101117, China
4. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
5. Faculty of Innovation Engineering, Macau University of Science and Technology, Macau 999078, China

Abstract Embodied intelligence represents a new stage in the evolution of artificial intelligence, marking a transition from "perception−cognition" to an integrated paradigm of "perception−cognition−action." The Vision−Language−Action (VLA) model provides a critical technological pathway for enabling autonomous agent operation in the real world by unifying visual perception, language understanding, and action generation. This paper systematically reviews the development trajectory and representative achievements of VLA technologies, and summarizes their architectural paradigm, which includes multi−modal perception, semantic fusion mechanisms, reinforcement and imitation learning, world models, and hierarchical action output. By considering application scenarios such as autonomous driving, human–computer interaction, and industrial equipment, we further analyze the core challenges faced by VLA development, including the scarcity of data resources, limited generalization and transferability, insufficient interpretability, and increasing computational demands, and we outline the future development trends.

Keywords vision−language−action model; multi−modal learning; embodied intelligence; large language model

（责任编辑　傅雪）