doi:10.3969/j.issn.1003-3106.2025.11.014
引用格式:袁霆宇,刘凯,关标良,等.具身智能大模型综述及展望[J].无线电工程,2025,55(11):2256-2273.[YUAN Tingyu, LIU Kai, GUAN Biaoliang, et al. A Comprehensive Review and Future Perspectives on Embodied AI Large Models[J]. Radio Engineering, 2025, 55(11): 2256-2273.]

具身智能大模型综述及展望

袁霆宇[1,2],刘　凯[3],关标良[3],叶　雯[2,4],赵雅萃[5],赵朝阳[1,6],王金桥[1,2]
(1.中国科学院自动化研究所 紫东太初大模型研究中心,北京 100083;
2.中国科学院大学 人工智能学院,北京 100083;
3.西安交通大学 软件学院,陕西 西安 710049;
4.中国科学院自动化研究所 模式识别实验室,北京 100083;
5.帝国理工学院 哈姆林中心,伦敦 SW7 2AZ;
6.中科视语(北京)科技有限公司,北京 100083)

摘　要:视觉-语言-行动(Vision-Language-Action,VLA)模型是实现通用具身人工智能的核心技术,旨在在统一的端到端框架内融合视觉感知、语言理解与动作决策。对 VLA 模型的研究现状与发展脉络进行了全面而系统的梳理。追溯了 VLA 模型的理论起源,阐明了其从分离式模块向统一架构演进的范式变迁。针对 VLA 的演进路线,以多模态融合与认知分层为重点阐述了 SpatialVLA、TLA 与 GR00T N1 等工作。构建了一个详尽的 VLA 模型分类体系,从宏观架构和系统分层 2 个核心维度,深入剖析了从 RT-1 等开创性工作到引入大规模知识迁移的 RT-2、OpenVLA、ECOT 等工作,再到双系统架构的 Helix、OpenHelix、DexVLA、DexGraspVLA 等前沿模型的关键技术与设计思想。系统性地整合与评述了支撑 VLA 研究的主流仿真环境、核心数据集与基准,并探讨了其在机器人操作、自主导航、工业自动化等领域的应用现状与前景。深入剖析了当前 VLA 研究在泛化性与数据效率、长时程任务规划、实时响应速度等方面面临的核心挑战,并对融合世界模型、提升数据效率等未来研究方向进行了展望。

关键词:视觉-语言-行动模型;大模型;具身智能;机器人学习;多模态学习
中图分类号:TP391.41　　**文献标志码**:A　　**开放科学(资源服务)标识码(OSID)**:
文章编号:1003-3106(2025)11-2256-18

A Comprehensive Review and Future Perspectives on Embodied AI Large Models

YUAN Tingyu[1,2], LIU Kai[3], GUAN Biaoliang[3], YE Wen[2,4], ZHAO Yacui[5], ZHAO Chaoyang[1,6], WANG Jinqiao[1,2]
(1. *Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing* 100083, *China*;
2. *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing* 100083, *China*;
3. *School of Software Engineering, Xi'an Jiaotong University, Xi'an* 710049, *China*;
4. *New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100083, *China*;
5. *The Hamlyn Centre, Imperial College London, London SW7 2AZ, United Kingdom*;
6. *Objecteye. Inc, Beijing* 100083, *China*)

Abstract: Vision-Language-Action (VLA) models are a core technology for achieving general embodied artificial intelligence, aiming to integrate visual perception, language understanding, and action decision-making within a unified end-to-end framework. The current research status and development trajectory of VLA models are comprehensively and systematically reviewed. The theoretical origins of VLA models are traced, and the paradigm shift from modular designs to unified architectures is clarified. Along the evolutionary path of VLA, representative works such as SpatialVLA, TLA, and GR00T N1 are presented with a focus on multimodal

收稿日期:2025-08-02
基金项目:国家自然科学基金面上项目(62276260,62176254)
Foundation Item: General Program of National Natural Science Foundation of China (62276260,62176254)

fusion and cognitive hierarchies. A detailed taxonomy of VLA models is constructed from two key dimensions—macro architecture and system hierarchy. Key technologies and design principles are deeply analyzed, ranging from pioneering works such as RT-1, to models introducing large-scale knowledge transfer such as RT-2, OpenVLA, and ECOT, and further to cutting-edge dual-system architectures such as Helix, OpenHelix, DexVLA, and DexGraspVLA. Mainstream simulation environments, core datasets, and benchmarks supporting VLA research are systematically integrated and reviewed. The application status and prospects of VLA models in robotic manipulation, autonomous navigation, and industrial automation are explored. Core challenges in current VLA research are analyzed, including generalization and data efficiency, long-horizon task planning, and real-time responsiveness. Future research directions are discussed, including integration with world models and enhancement of data efficiency.

Keywords: VLA models; large models; embodied AI; robot learning; multimodal learning

0 引言

具身智能的核心目标是创建能够在物理世界中感知、推理并与环境进行实时交互的智能体[1]。通用人工智能(Artificial General Intelligence, AGI)旨在创建具备全面认知能力和强大泛化能力的人类级别智能系统。而具身人工智能被认为是通往AGI的关键路径,因为它能通过与环境的持续交互,学习并构建抽象的语言符号与多模态传感器数据的知识表征。这种在感知-行动-反馈闭环中的学习过程,使其能够真正理解世界,从而逐步迈向AGI。

实现上述宏伟愿景仍面临诸多关键挑战,主要体现在以下几个方面:多模态感官输入的有效融合、抽象语言指令向物理世界具体行为的映射机制,以及智能体在复杂真实环境中的安全性与鲁棒性保障。近年来,研究人员提出了VLA模型,旨在构建一个统一的端到端计算框架,融合感知、理解与控制能力,使智能体能够“看到”“听懂”并自主生成“可执行动作”,从而成为真正具备认知能力和行动能力的一体化智能系统。

在VLA模型出现之前,机器人领域的研究很大程度上是在视觉、自然语言处理和机器人控制等领域独立进行的。一个典型的机器人任务流程可能是:一个独立的视觉模块,如卷积神经网络(Convolutional Neural Networks, CNN)[2],负责“看”,例如识别出一个苹果;一个独立的自然语言处理模块负责“听”,例如解析“把苹果递给我”的指令;最后,一个独立的运动规划与控制模块负责“做”,即执行预先编程的抓取动作。这种分离式的流水线架构存在明显的“集成鸿沟”:感知模块缺乏对任务上下文的理解,语言模块不了解物理世界的约束,而控制模块则难以泛化到新的物体或场景。VLA模型的提出,旨在从根本上解决这一分裂状态。它致力于构建一个统一的、端到端的计算框架,能够联合处理视觉输入、理解语言指令,并直接生成可执行的机器人动作序列。通过将感知、推理和行动这3个核心要素紧密耦合在单一模型中,VLA模型使得机器人不仅能够“看到”和“听到”,更能够“理解”并“行动”,这是迈向真正自适应、可泛化、高智能的具身智能体的革命性一步。

近年来,VLA模型的研究呈现出爆炸式增长,大量新模型不断涌现,使得对这一快速发展的领域进行系统性梳理变得至关重要。

本文的贡献总结如下:

① 构建了VLA模型的系统性知识框架与分类体系:全面梳理了VLA模型从分离式模块到端到端统一架构的演进历程,并基于模型的宏观架构和动作表征方式,创建了一个详尽的分类体系。

② 整合并评述了VLA研究的核心资源:系统性地总结、对比并评述了支撑VLA研究的关键资源,涵盖了主流的仿真环境、核心数据集与基准,以及关键的评估指标。

③ 深入剖析了核心挑战并展望了前沿发展方向:系统性地论述了当前VLA研究面临的核心挑战,结合最新进展,对未来研究方向进行了前瞻性展望。

1 VLA模型定义与理论框架

1.1 VLA模型的定义

VLA模型的概念由RT-2[3]模型的研究团队首次提出,其定义为:任何能够处理来自视觉和语言的多模态输入,并生成机器人动作用以完成具身任务的模型。

一个典型的VLA模型框架通常包含以下核心组件:

① 视觉编码器:利用视觉基础模型,如视觉转换器(Vision Transformer, ViT[4]),提取当前环境状态的预训练视觉表征,包含物体的类别、姿态和几何形状等信息。

② 语言编码器：利用大语言模型（Large Language Model，LLM）[5]来编码用户的自然语言指令。

③ 多模态融合模块：将视觉和语言表征进行对齐与融合，采用的技术包括交叉注意力机制或将视觉特征直接注入 LLM 的嵌入空间等。

④ 动作解码器：基于融合后的多模态表征，生成机器人动作。动作可以被表示为离散的词元（token）或连续值。

图 1 展示了典型的 VLA 模型架构，包含了视觉模型、语言模型以及动作模型，分别处理图片输入、任务输入以及动作输出。图中指令以“Pick the sushi piece from the bowl and place it on the table”为例，VLA 模型的工作流程可概括为以下几个关键步骤：首先，视觉传感器获取的图像信息，例如 sushi（寿司）、bowl（碗）和 table（桌子）等，会被输入至视觉编码器进行特征提取。与此同时，移动任务指令则被送入语言编码器进行语义理解。随后，多模态融合模块将这 2 种信息整合，形成一个统一的、面向任务的意图表征，并输出动作词元。这些动作词元代表了机器人手臂末端的姿态，包括平移、旋转和开合状态等，并以离散编码形式表示。动作解码器负责将这些动作词元解析成具体的末端执行器位姿。接着，解析出的位姿信息会发送至底层机器人控制器，控制器通过逆运动学求解，计算出机器人各关节需要旋转的角度，从而进行精确控制。值得注意的是，视觉语言模型（Vision-Language Model，VLM）[6]的推理频率（通常为 1～10 Hz）远低于机器人伺服控制器的控制频率（通常为 50～1 000 Hz）。在每个高频控制周期内，控制器驱动电机，最终平滑地完成将寿司从碗中移动到桌面等物理操作。

1.2 多模态表征与融合

VLA 模型的核心在于其处理与整合视觉、语言和动作的异构模态信息的能力[1]，这一能力决定了其在多模态感知、任务理解与行动规划中的智能水平。由于图像与语言在结构、语义密度、时间动态等方面存在显著差异，如何建立跨模态的一致性语义空间，成为模型设计中的关键挑战。这一过程依赖于高表达能力的模态编码器以捕捉各自模态的深层语义信息，并依赖高效而鲁棒的模态融合机制以实现特征对齐与协同推理。随着 Transformer[7]架构的广泛应用，近年来的研究逐渐趋向使用统一的序列建模范式处理多模态信息，推动了视觉-语言融合策略的快速演进。

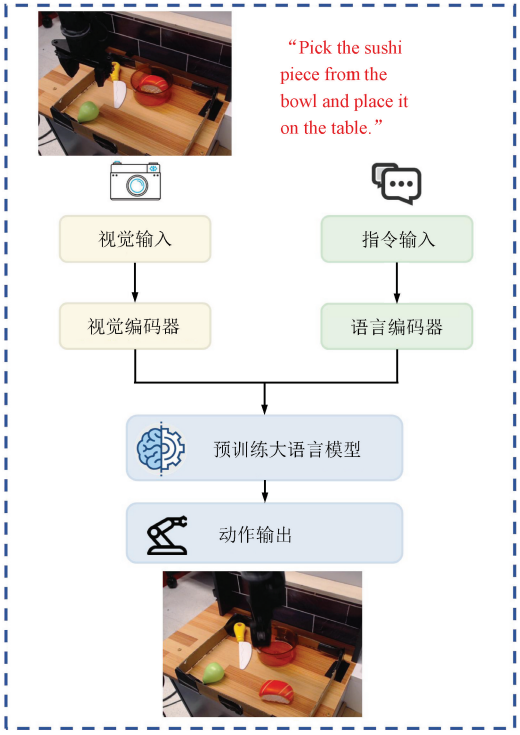


图 1 典型 VLA 模型架构，包含语言模型、视觉模型、动作模型

Fig. 1 Typical VLA model architecture comprising language, visual, and action models

1.2.1 视觉编码器

视觉编码器的任务是从输入的图像（通常是机器人视觉传感器捕获的 RGB 或者 RGBD 图像）中提取有助于任务的特征表示。

（1）早期方法：早期的 VLA 模型，如 RT-1[8]，采用了在 ImageNet[9]等大规模图像数据集上预训练的 CNN 作为视觉主干，例如 EfficientNet[10]。CNN 通过其固有的卷积和池化操作，擅长捕捉图像的局部特征和空间层次结构。

（2）现代方法：随着 Transformer 架构在视觉领域的成功，近期的 VLA 模型普遍转向使用 ViT[4]及其变体作为视觉编码器。ViT 的革命性在于它将图像视为一个序列。该方法首先将输入的图像划分为一系列固定大小的图像块（patches），随后对这些图像块进行线性嵌入处理，并辅以位置编码信息。由此，构建得到一个“图像词元”（image tokens）序列，该序列随即被送入一个标准的 Transformer 编码器进行深度特征提取与处理。ViT 的核心优势在于其自注意力机制，该机制能够计算序列中任意 2 个图像块之间的依赖关系，从而捕捉全局上下文信息。相比于 CNN 的局部感受野，ViT 能够更好地理解物

体间的长距离关系和场景的整体布局。这对于需要复杂空间推理的机器人任务至关重要,例如理解场景布局与物体间关系。想象一个杂乱的桌面场景,机器人需要理解“杯子在键盘左侧,但在笔记本电脑前面”这类复杂的拓扑关系,以便规划正确的抓取路径。ViT 的全局注意力机制能够直接建模“杯子”与“键盘”“笔记本电脑”等所有物体间的空间关系,而 CNN 则需要通过极深的网络才能间接推断这些关系。如表 1 所示[4],ViT 在所有数据集上都优于 CNN 的 SOTA(State-of-the-Art)模型(如 ResNet[11]),这充分证明了其在特征提取能力上的根本性优势。显然,一个能够提供更全面、更具关系感知的视觉表征的编码器,必然能显著提升 VLA 模型在复杂任务中的表现。

RT-2 采用的 DINOv2[12] 和 SigLIP[13] 等先进的视觉 Transformer 变体在自监督预训练下表现出强大的表征能力和良好的泛化性,它们已成为当前 VLA 模型常用的视觉编码器选择之一。

表 1　ViT 与 CNN 架构在不同数据集的对比

Tab. 1　Comparison of ViT and CNN architectures on different datasets

Dataset	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152×4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55**±0.04	87.76±0.03	85.30±0.02	87.54±0.04	88.4/88.5*
ImageNet Real	**90.72**±0.05	90.54±0.03	88.62±0.05	90.54	90.55
CIFAR-10	**99.50**±0.06	99.42±0.03	99.15±0.03	99.37±0.06	—
CIFAR-100	**94.55**±0.04	93.90±0.05	93.25±0.05	93.51±0.08	—
Oxford-IIIT Pets	**97.56**±0.03	97.32±0.11	94.67±0.15	96.62±0.23	—
Oxford Flowers-102	99.68±0.02	**99.74**±0.00	99.61±0.02	99.63±0.03	—
VTAB (19 tasks)	**77.63**±0.23	76.28±0.46	72.72±0.21	76.29±1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

注:事项 * 是 Touvron 等的一个稍微改进的结果。

1.2.2　语言编码器

语言编码器负责理解用户的自然语言指令,并为整个模型提供高级的语义和常识推理能力。VLA 模型的“大脑”通常是一个预训练好的 LLM,如 Google 公司推出的 PaLM[14]、T5,Meta 公司推出的 LLaMA[15],或开源的 Gemma[16]等,通过在数万亿级别的文本语料上进行训练,学习到了丰富的语言规律、世界知识和推理能力。在 VLA 框架中,LLM 扮演着中心枢纽的角色。LLM 不仅需解析指令的字面语义,更需结合视觉信息进行基础性关联,以理解指令所提及的对象和地点在当前视觉场景中的具体对应关系,例如,当接语言编码器接受抽象语言指令“小心拿起玻璃杯移到橱柜里”时,首先由 LLM 对指令进行语义解析,将“小心”理解为“需要控制抓取力度”以及“优先选择易碎物体”,并抽取出关键目标“玻璃杯”及其属性“易碎”。随后,通过视觉编码器在视觉场景中检测出物体并识别其类别及属性标签(如 glass、fragile),LLM 将二者进行匹配,实现语言与视觉的语义接地。LLM 常基于这种跨模态的理解,进行决策和规划,生成下一步应该执行的动作[17]。

1.2.3　多模态融合策略

如何有效融合来自视觉与语言 2 种模态的异构信息,是 VLA 模型设计的核心挑战之一。理想的融合机制应能将模态间的语义对齐至统一表示空间,从而支持跨模态的感知接地与推理决策。如图 2 所示,当前主流的融合策略主要包括以下几类:

① 特征层调制:早期模型如 RT-1[8] 采用特征线性调制(Feature-wise Linear Modulation,FiLM)机制[18],将语言嵌入作为调制因子,注入至视觉编码器如 EfficientNet[10] 的中间特征层。具体而言,模型学习从语言中生成一组缩放与偏置参数,对视觉特征图的每一通道进行仿射变换,从而实现语言对视觉感知过程的早期引导。这种方法实现了低成本、深层次的融合方式,适用于轻量化部署。

② 跨模态注意力:基于 Transformer 架构的跨注意力机制已成为当前多模态融合的主流选择[19]。在此机制中,一种模态的表示(如语言指令)作为查询,主动对另一模态的表征(如视觉特征)进行选择性聚合,实现动态的语义对齐。如 VIMA[20]、PaLM-E[14] 模型采用层叠式交叉注意力机制来处理视觉词元与语言词元间的关联,显著增强了多模态上下文

理解能力。

③ 投影与拼接:许多最新的通用 VLA 模型(如 OpenVLA[21])采用将不同模态投影到同一维度后进行拼接的融合策略。具体做法是使用多层感知机(Muiti-Layer Perceptron,MLP)将视觉编码器输出的图像块表征映射至与语言词元相同的向量空间,然后将视觉词元与语言词元拼接为统一的输入序列送入 LLM。此策略结构简洁,融合效率高,并可充分利用 LLM 内部自注意力机制实现模态间交互。

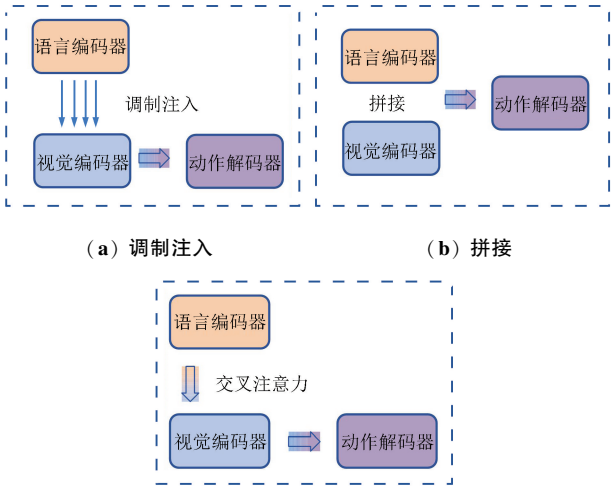


图 2 常见的多模态融合策略

Fig. 2 Prevalent multimodal fusion strategies

VLA 模型并非传统 VLM 的延伸,在核心目标上,传统 VLM 的主要目标是对多模态信息进行被动理解和表征,其任务通常为视觉问答(Visual Question Answering,VQA)、图像描述和跨模态检索,输出空间限于文本、图像、类别标签或相似度得分。而 VLA 模型的核心目标则是将理解转化为主动决策与物理行动,生成一个面向任务的、可执行的动作。因此,VLM 的融合表征只需具备良好的语义区分度即可,VLA 的融合表征则还可能要考虑时序性,不仅需要融合当前时刻的视觉语言信息、机械臂位姿状态等,还常需要整合历史的观测与动作序列,从而将语言指令高效地转化为物理世界中精确连贯的动作序列。

1.3 动作解码与表征

动作解码是 VLA 模型区别于传统 VLM 的关键环节,其任务是将模型内部的决策意图转化为机器人可执行的物理动作指令[22]。动作表征方式的演进,不仅体现了机器人控制复杂性的本质,也标志着 VLA 研究从迁移语言模型能力向动作生成机制的转变[23]。

1.3.1 离散化动作分词

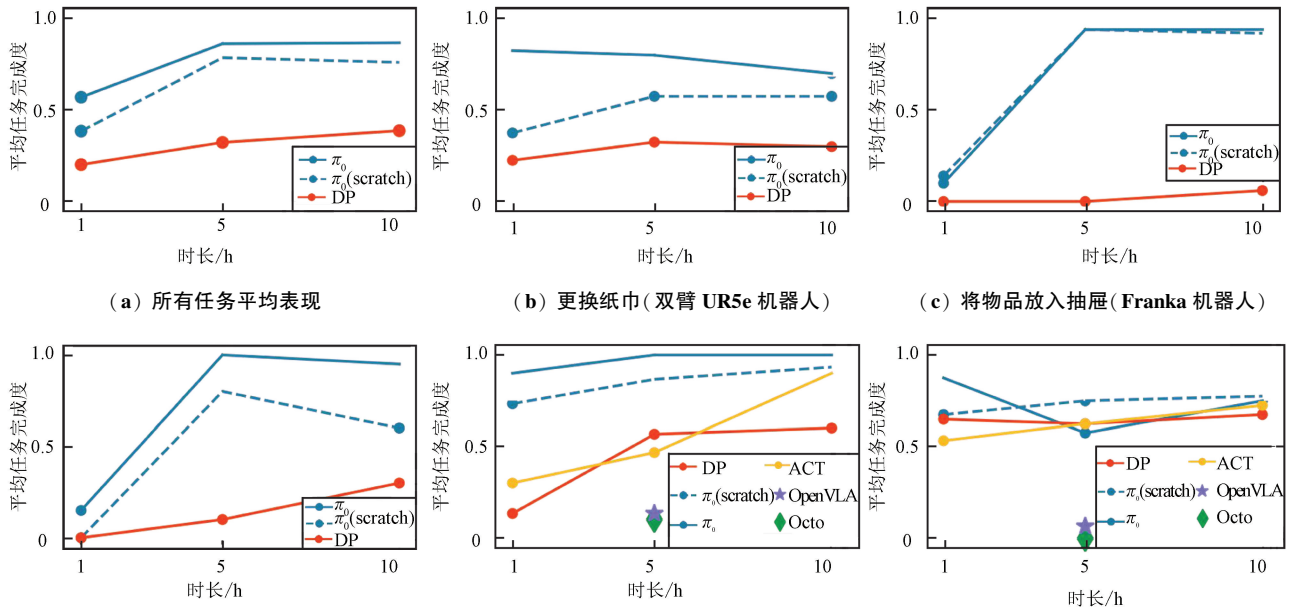
该方法是 VLA 模型早期阶段的主流方案,被 Gato[24]、RT-1[8]、RT-2[3]等代表性工作采用。其核心思想在于对连续动作空间进行离散化处理。具体而言,该方法将每个动作维度划分为预设数量(例如 256 个)的离散区间,并为每个区间赋予唯一的标记符(token ID),从而将原始的连续动作向量转换为一个多维词元序列。以 7 自由度的机械臂为例,其动作通常由末端执行器的 6 维位姿(x、y、z、roll、pitch、yaw)和夹爪状态组成。通过对每个维度进行独立量化,最终得到一个由 7 个词元构成的动作序列。在此基础上,LLM 即可自回归生成动词元,这个过程类似自然语言建模。该方法的优势在于实现简洁、易于与 LLM 架构无缝对接,使得机器人任务能够直接受益于大模型的序列建模能力。然而,其主要问题在于引入量化误差、难以保留高精度运动信息;此外,随着自由度和离散区间数量的增加,动作词元的词表规模呈指数增长,增加了模型学习与泛化的难度。此外,这种粗粒度的离散编码难以应对高频、光滑、精细的控制需求。

1.3.2 连续动作生成

为克服离散化策略在建模精度与动作平滑性上的限制,近期研究开始采用去噪扩散概率模型(Denoising Diffusion Probabilistic Models,DDPM)[25]直接生成高维连续动作序列。扩散模型通过逐步去噪过程生成目标动作序列:在训练阶段,模型学习将真实的动作序列逐步添加噪声至高斯分布;在生成阶段,则从高斯噪声初始化,通过多步迭代的去噪过程逐步恢复出符合条件语义的平滑动作序列。这种条件生成机制确保了输出动作不仅在运动学上连贯、平滑,而且严格遵循输入语言指令的语义约束。

代表性工作包括 Diffusion Policy[26]和 π_0[27],二者均在机器人控制任务中取得了显著的性能提升。在此类模型中,扩散过程是条件生成的,去噪网络不仅输入时间步和噪声化动作,还接收由上游 VLM 编码的任务语义表征作为条件,完成从任务意图到动作序列的映射。该结构使得模型在处理复杂自然语言指令时具备更强的语义一致性与动作生成灵活性。Octo 团队[28]使用扩散模型,使用条件扩散解码头来预测连续的、多模态的动作分布,实现了高质量、平滑且符合语言指令语义的连续轨迹控制。但

扩散模型的主要瓶颈在于推理效率低下，其逐步去噪的生成机制通常需要数十至上百次前向传播，难以满足机器人系统对实时性的要求。为提升效率，π_0 引入了条件流匹配技术[29]，用单步采样替代传统多步去噪过程，在多个机器人操作基准任务中显著缩短了推理时间，同时保持了动作生成的准确性和语义一致性，为扩散模型在 VLA 场景中的部署提供了可行路径。如图 3 所示[28]，π_0 的任务成功率显著高于自回归式的 OpenVLA、ACT 等模型，充分验证了其生成动作的连续性优于自回归方法，并在保持语义一致性的同时提升了整体性能。



（a）所有任务平均表现　（b）更换纸巾（双臂 UR5e 机器人）　（c）将物品放入抽屉（Franka 机器人）

（d）折叠毛巾（双臂 ARX 机器人）　（e）堆叠碗（UR5e 机器人）　（f）将特百惠放入微波炉（双臂 ARX 机器人）

图 3　π_0 与其他模型的对比

Fig. 3　Comparison between π_0 and other models

在 π_0 的基础上，$\pi_{0.5}$[30] 通过协同训练多源异构数据，成功实现了在未知环境中的“开放世界”泛化。其核心在于采用了层次化推理架构：先规划高层语义子任务，再指导底层连续动作的生成，从而显著提升了复杂任务的成功率与鲁棒性。

2　VLA 模型架构的演进与分类

VLA 模型在近几年经历了快速的架构演进，经历了从分离式系统向统一建模范式的转变，体现出感知、理解与控制的深度融合。本节梳理 VLA 模型的代表性发展阶段与分类框架，并通过对典型工作进行剖析，展示其技术路径的转型趋势。

2.1　VLA 起源与演变

2.1.1　VLA 模型诞生前的技术图景

在 VLA 概念提出之前，机器人智能系统长期遵循“分而治之”的模块化设计范式，分别在计算机视觉、自然语言处理与机器人控制等子领域独立发展 VIMA[31]。视觉模块通常专注于从图像或视频中执行目标检测、语义分割等感知任务，以构建对环境状态的理解；语言模块致力于解析人类的文本或语音指令，提供语言层面的语义分析与反馈；而控制模块（即策略网络或控制器）则聚焦于根据任务目标执行运动规划与底层物理控制。尽管这种模块化设计在早期推动了多项突破性进展，但其固有的分离性也暴露出明显局限。以 VLN-BERT[32] 为例，该方法首次将 BERT[33] 引入视觉导航任务中，实现了视觉信息、语言指令与路径状态的统一编码，推动了语言驱动导航决策的研究方向。又如 CLIPort[34]，借助 CLIP[35] 在图文对齐上的预训练能力，使得模型可以通过自然语言指令直接生成操控动作，并展现出在少量样本条件下的良好泛化能力。上述工作初步建立了视觉、语言与动作之间的表征桥梁，为 VLA 模型的提出奠定了方法论基础。然而，这类分离式架构在语义融合层面存在根本瓶颈：各模块之间缺乏统一的表示空间，难以实现语言、视觉与动作之间的深层语义对齐。这种语义鸿沟使得机器人

系统难以将抽象高层指令映射到高维视觉观测的具体对象,并进一步转化为机械臂所需的精确动作轨迹。

模块之间语义脱节、端到端可微训练缺失以及对长链任务的弱泛化性,逐渐成为制约机器人系统通用化能力的关键瓶颈[36]。这种背景催生了将感知、语言理解与控制决策统一建模的 VLA 模型,标志着具身智能系统由分离式向统一式的发展。

2.1.2 “动作语言化”范式的提出与实践

作为一种统一感知、语言理解与动作控制的端到端框架,VLA 模型真正迈入研究前沿,始于 Google 公司在 2022 年发布的 Robotic Transformer 系列工作。RT-1[8] 首次将 Transformer 架构引入低维机器人控制任务,实现了图像、语言与动作的端到端建模,揭示了统一多模态输入与动作输出的可能性。随后发布的 RT-2[3] 更进一步融合了 CLIP[35] 等大规模视觉语言预训练模型,使机器人具备从互联网上迁移跨模态知识的能力,显著提升了模型在开放环境下的泛化控制能力。RT-2 的关键创新在于提出“动作语言化”(Action-as-Language)范式:将原本连续的机器人动作(如末端执行器的位移、旋转及夹爪开合)离散化为若干词元,使其与语言序列统一到相同的建模空间。按照“Δpos_x Δpos_y Δpos_z Δrot_x Δrot_y Δrot_z”(若不考虑夹爪以及基座运动)表示原始动作,将动作空间的每一维都均匀离散成 256 个离散量从而实现动作的离散化表示。通过这种方式,就可以得到 6 个整数例如“132,114,128,5,25,156”,将这些动作词元视作操作机器人的语言,与给定指令所转换的词元空间一同微调,从而实现对齐。这一设计突破了传统机器人控制与语言处理之间的语义壁垒。如图 4 所示[3],RT-2 模型将语言、动作和图像整合至一个统一的输出空间,这使得模型能够利用 VLM 的强大能力来生成语言,同时也可将动作视为一种特殊的语言表达形式进行处理。

在具体实现中,RT-2 采用 PaLI-X[37] 和 PaLM-E[14] 等强大的多模态预训练模型作为编码主干,联合微调互联网上的大规模视觉语言任务数据(如图像描述、VQA)与机器人操控数据(包括第一视角图像、自然语言指令及对应的机器人动作)。训练过程实现了 VLA 三者的统一建模。在推理阶段,模型接收当前图像和语言指令后,以自回归方式逐步生成动作词元,最终通过反序列化还原为连续控制信号,即将动作词元转换为原始动作(末端执行器的平移、旋转等),从而完成从感知到执行的闭环操控流程。这一范式验证了 LLM 在动作生成领域的可迁移性,为后续 VLA 模型的发展奠定了方法论基础。

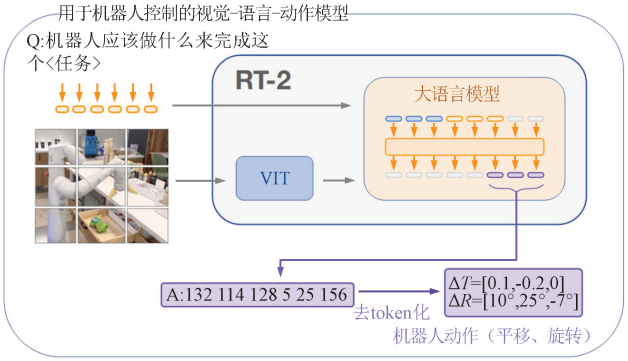


图 4 RT-2 的 VLA 架构

Fig. 4 VLA architecture of the RT-2

2.1.3 从统一建模到开放泛化的关键跃迁

尽管 RT 系列模型(RT-1、RT-2)在推动 VLA 架构统一方面取得了重要进展,其应用仍主要局限于封闭任务分布。随着研究者对具身智能泛化能力的关注不断提升,VLA 模型正迈入开放泛化阶段,即在未知环境与全新任务指令下生成合理行为的能力成为核心研究议题。为应对这一挑战,近期工作在模型结构与训练范式上提出了多项关键创新。其中,OpenVLA[21] 模型首次构建了真正意义上的开源 VLA 控制框架,基于 LLaMA 2-7B[38] 语言模型,融合 SigLIP[13] 和 DINOv2[12] 等视觉主干,并在包含 97 万条机器人演示轨迹的 Open X-Embodiment[39] 数据集上进行训练,覆盖了摆放、整理、工具使用等多类任务。该模型支持 LoRA[40] 微调和 4-bit 量化,能够在单张消费级 GPU 上实现实时部署,显著降低了 VLA 研究与应用的门槛。同时,π[27,30] 系列模型提出了基于轨迹流分布的连续动作建模方法,采用生成式流匹配替代传统自回归生成机制,提升了长时序动作的建模能力与推理效率。通过融合异构多源训练数据(仿真轨迹、家庭视频、网络图文)显著增强了模型的跨任务、跨环境与跨机器人平台的泛化能力。为进一步解决 VLA 模型在微调后泛化至全新物体能力下降的问题,ObjectVLA[41] 提出了一种创新解决方案。其核心是联合训练机器人演示数据与带有定位元数据(如边界框)的视觉-文本数据,把物体的通用语义知识与机器人特定动作有效链接,其实现了将“递出苹果”等技能零样本泛化至上百种未见过物体的能力,还支持仅用少量手机照片对模型进行快速微调,从而高效适应现实世界中的全新目标。

2.1.4 多模态融合与认知分层的演进路线

在实现视觉、语言与动作统一建模的基础上,VLA 研究分化出多种技术路径[42],包括多模态感知增强、认知分层等方向。

感知模态的多样性已成为提升 VLA 模型环境理解能力的关键途径,研究者开始引入点云、深度

图、轨迹等以增强模型的空间理解能力。VoxPoser[43]引入体素化表示以实现对操控空间的显式建模。PointVLA[44]将稀疏点云编码进 Transformer 结构,有效提升了机器人对物体形状与姿态的识别能力。SpatialVLA[45]更进一步地融合空间语言与几何结构建模,在动态场景中实现了对复杂空间关系的显式推理。OmniManip[46]为解决精细操作任务的空间理解问题,将机器人动作抽象成具有空间约束的交互原语。在接触密集型任务中,仅凭视觉感知难以满足精确操作的需求。此时,整合触觉反馈能够显著提升机器人操作技能的灵活性与鲁棒性[47]。例如,TLA[48]模型通过构建专用的“触觉-动作-指令”数据集,并有效处理序列化触觉反馈,在栓孔装配等任务中展现出对不同装配间隙和几何形状的强大泛化能力。

在推理与控制机制层面,研究者提出了双系统分层架构设计,以增强系统的认知能力与任务规划灵活性。代表性工作如 Hi-Robot[49],其采用 VLM 负责高层任务分解与推理,并将低层动作控制交由 VLA 模块完成,实现了“任务-语义-动作”之间的语义对齐与层次映射,具备对复杂多步骤任务的处理能力。GR00T N1[50]也采用了双系统架构。其中,系统二是一个基于 Eagle-2 VLM[51]的视觉-语言模块,负责对环境进行推理并解析接收到的指令,进而生成行动规划。而系统一则是一个扩散变换器(Diffusion Transformers,DiT)[52]的动作模块,其功能是将上述规划转化为精确且连续的机器人动作。这 2 个系统通过交叉注意力机制紧密耦合,并可实现端到端联合训练。

Figure AI 团队提出的 Helix[53]模型同样采用了双系统分层架构,并在这一分层基础上引入了快慢思考的机制。这 2 个系统分别负责处理场景的视觉线索与语义目标,以及高频率地生成连续控制信号。上层的 VLM 能够以相对较低的频率理解环境,而下层的策略模块则能以更高的反应速度产生连续动作。DexVLA[54]则提出将 VLM 与扩散动作专家模块化解耦的方案,并通过具身课程学习策略提升跨机器人平台的泛化与适应能力。其三阶段训练机制显著降低对大规模任务标注的依赖,实现了端到端复杂任务的自主执行。

2.2 VLA 的分类

当前 VLA 模型的发展趋势可沿 2 个关键维度进行系统分类,如图 5 所示。其一是模型的宏观架构范式可分为基于自回归的 VLA 模型、基于扩散模型的 VLA 模型。其二是从认知分层角度,分为单系统和双系统分层架构。

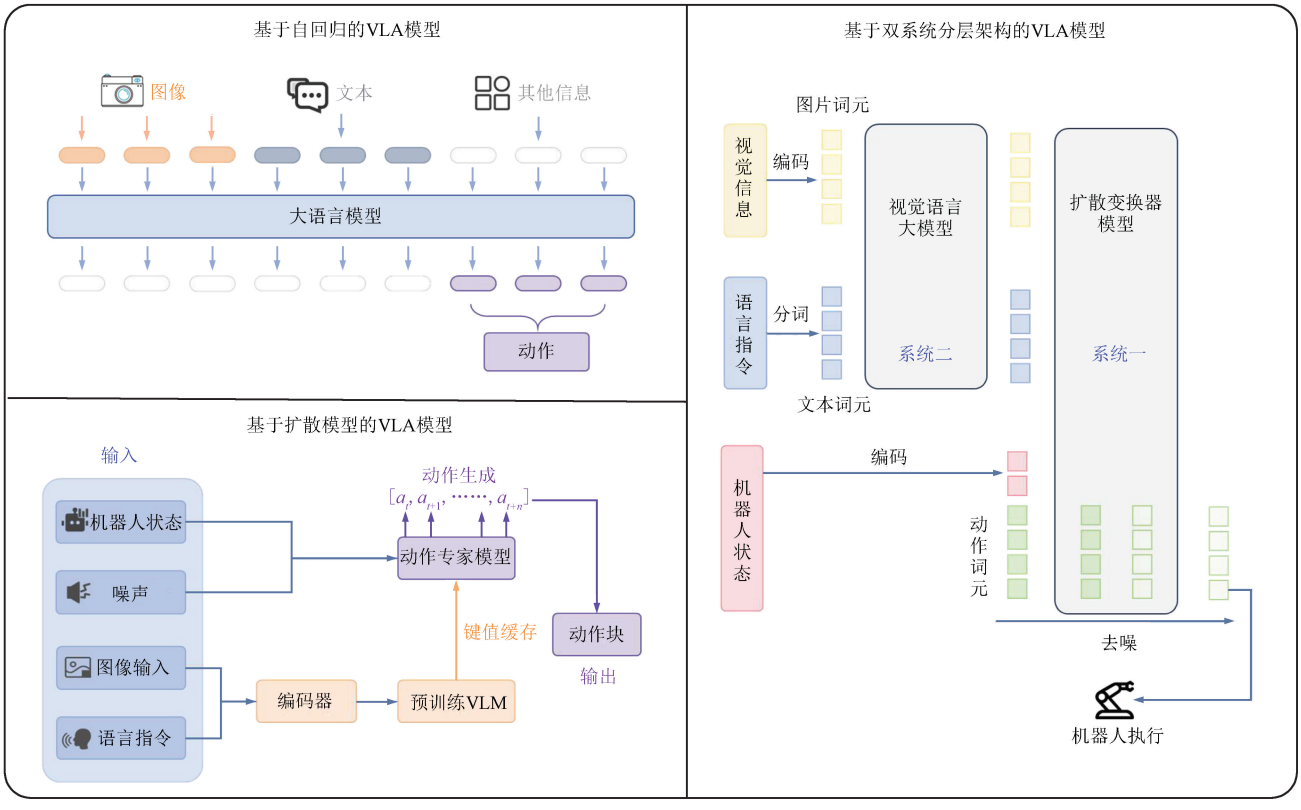


图 5 VLA 模型的分类

Fig. 5 Classification of VLA models

2.2.1 基于宏观架构的分类

VLA 模型的宏观架构依据其核心的动作生成机制,可划分为 2 种主要的技术范式。如表 2 所示,分为预测离散动作词元的自回归式以及生成连续动作序列的生成式。离散动作词元的自回归序列建模,虽然解码速度快,更适用于对实时性要求高的场景,但其量化误差较大,导致难以实现平滑精细的控制。这可能需要通过设计特定的技巧(如掩码、分块等)来缓解误差影响。相比之下,采用连续化动作类型的 VLA 模型通过生成式建模实现连续动作的生成,从而能够进行平滑控制。然而,这种方法会显著增加计算开销和延迟。

表 2 基于宏观架构的分类

Tab. 2 Classification based on macro architecture

分类	核心思想	代表模型
预测离散动作词元的自回归式	将动作映射为离散词元	RT-2、ACT、ECOT、UniPi、OpenVLA-OFT、WorldVLA
生成连续动作序列的生成式	学习动作轨迹的条件概率分布	π_0、Diffusion Policy、RDT-1B、TacAR、HybridVLA

(1) 预测离散动作词元的自回归式

自回归式源于自回归模型,其通过序列预测离散词元,每个词元生成都依赖前一个词元[55]。VLA 则将预测问题重构为一个基于多模态模型的序列预测任务[56]。RT-2 首次通过引入结构化的动作词元(如 move_to(x,y,z)、grasp 等)实现了语言、视觉与控制的统一建模。在 OpenVLA 的基础上,ECOT[57]则将思维链(Chain of Though,CoT)[58]融入到推理中,通过训练 VLA 进行计划、子任务、动作以及物体边界框等多步骤推理提高任务成功率以及泛化性。UniPi[59]则将顺序决策问题建模为文本条件的视频生成过程,通过合成未来帧实现隐式动作规划,从中提取动作词元,有效统一了语言、视觉与策略表征。

自回归式模型生成连续动作时,一旦前面预测动作出现错误,后续动作的预测会随时间不断放大此错误而导致性能下降,为解决误差累计问题,ACT[60]提出动作分块(Action Chunking)的思想,对每次推理的动作取加权结果减少错误累计。为解决泛化问题,OpenVLA-OFT[61]采用并行解码,并优化动作表示避免离散化造成的细节损失。WorldVLA[62]则提出动作注意力掩码策略,在生成当前动作时选择性地屏蔽掉之前的动作信息,有效缓解了错误累积的问题。而 Dense Policy 提出一种新型自回归策略,通过双向自回归学习扩展稀疏关键帧序列从而生成高精度连贯的密集动作序列,解决传统自回归单向预测难以捕捉动作序列的双向依赖关系的难题。

(2) 生成连续动作序列的扩散生成式

自回归方法模仿 VLM 的词元预测,但量化过程中会破坏动作姿态的连续性。而基于生成模型的动作分布学习不再直接回归单一确定性动作,而是将策略学习问题视为条件生成建模任务,即对动作轨迹的条件概率分布 $p(A|O)$ 进行建模。其中 A 表示动作序列,O 表示观测信息[63]。在该范式中,DDPM[25]因其在高维连续数据建模上的表现,受到越来越多研究的关注。Diffusion Policy[26]是该方向的代表作,通过条件扩散模型从随机噪声中逐步生成轨迹。π_0[27]采用流匹配而非传统扩散模型,以显式学习轨迹生成向量场。SmolVLA[64]同样基于流匹配生成连续动作块并通过视觉令牌缩减以及交叉与自注意力交替方法降低计算量、提升动作生成的平滑性。RDT-1B[65]提出机器人扩散(Robot Diffusion Transformer,RDT)有效表示多模态,并且以可扩展的 Transformer 来处理多模态输入的异质性。TacTAR[66]提出反应扩散策略(Reactive Diffusion Policy,RDP),并结合视觉触觉的反馈进行复杂的操控。HybridVLA[67]将自回归和扩散策略无缝集成到单个 LLM 的自回归下一个词元预测中,实现更强大的控制。

2.2.2 认知分层

如表 3 所示,VLA 模型在认知分层的角度可分为单系统端到端架构和双系统分层架构。前者直接由多模态输入到动作,侧重统一建模、紧耦合推理;后者则将高层任务规划与低层动作控制解耦,通过分层协作提升任务的执行能力和泛化性。

表 3 基于认知分层的分类

Tab. 3 Classification based on cognitive hierarchies

分类	核心思想	代表模型
单系统端到端	端到端多模态到动作	RT-1、ACT、RT-2、OpenVLA、ECOT、UniPi
双系统分层架构	高层任务规划与低层控制结合	Hi-Robot、GR00T N1、Helix、DexVLA、Psi-R1、DexGraspVLA

(1) 单系统端到端架构

单系统端到端架构在统一的、端到端的模型完成从多模态信息输入到动作输出的过程,内部可能有不同组件但其紧密耦合作为一个整体进行训练和

推理[68]，包括基于经典 Transformer 结构的方法 RT-1、ACT 等，自回归式的 RT-2、OpenVLA、ECOT、UniPi 等；基于扩散的方法，通过 DDPM 生成动作，包括 Diffusion Policy、π_0、RDT-1B、TactAR、HybridVLA 等。

（2）双系统分层架构

双系统分层架构则将高层任务规划与低层控制分离，提高长时域任务的执行效率。代表性工作有 VLM 与 LLM 结合的 Hi-Robot[49]、GR00T N1[50]，Helix[53]、DexVLA[54] 等。灵初智能团队发布的 Psi-R1[69]也采取了双系统架构，其负责推理规划的上层 VLM 做环境感知时，额外将动作分词一同输入从而强化多模态融合能力，有助完成长程任务的灵巧操作。DexGraspVLA[70]在双系统的基础上，将视觉和语言输入转换为域不变表示并高效利用基于扩散的模仿学习以捕捉数据集动作分布，实现强大的泛化性能。针对双系统理解能力有限以及延迟问题，OneTwoVLA[71]采取在推理和行动之间根据任务动态切换，实现推理与行动的反馈闭环，解决了双系统推理与执行解耦导致的理解不足、延迟等问题。

3 关键资源：数据集、仿真器与基准

VLA 系统的构建与评估高度依赖于高质量的训练数据、具有高度真实感的仿真环境以及标准化的评测基准。这些资源不仅为模型提供训练支持，也在实践中定义了研究任务的结构与挑战。

3.1 仿真环境

在具身 AI 研究中，仿真器扮演着过去数据集在互联网 AI 中的核心角色。由于在真实世界中进行大规模机器人训练和测试成本高昂、效率低下且存在安全风险，仿真环境成为了不可或缺的工具[72]。一个优秀的仿真器能够在保证一定真实性的前提下，提供安全、高效、可复现、可大规模并行的实验平台。表 4 展示了当前具身 AI 领域最主流的几个仿真器。

表 4 主流具身仿真器

Tab. 4 Mainstream embodied simulators

类别	开发机构	核心特点
AI2-THOR	AI2	专注室内场景
Habitat	MetaAI	室内场景，关注人机交互
iGibson	斯坦福大学	强调真实的物理交互
SAPIEN	加州大学圣迭戈分校	高精度物理操控
Isaac Sim / Gym	NVIDIA	照片级渲染与 GPU 加速
RoboHive	斯坦福大学	标准化多任务基准

AI2-THOR[73]：由艾伦人工智能研究所（AI2）开发，AI2-THOR 提供了一系列高度逼真、可交互的 3D 室内场景，包括厨房、客厅、卧室和浴室等。其突出特点是支持丰富的物体交互，物体不仅可以被移动，还具有多种状态，如开/关、冷/热、干净/脏等。这使得 AI2-THOR 非常适合用于需要复杂物体操作和状态追踪的视觉导航与交互任务，例如 ALFRED[74] 基准是在该平台之上构建的。

Habitat（Habitat-Sim）[75]：由 Meta AI 公司主导开发，具有极致的渲染速度和模拟效率。适合需要海量样本的强化学习算法的大规模并行训练。Habitat 支持加载多种通过 3D 扫描技术构建的真实世界场景数据集，如 Matterport3D[76] 为智能体的导航任务提供了高度逼真的视觉环境。

iGibson[77]：源于斯坦福大学，同样基于真实世界的扫描数据构建场景，但更加强调物理交互的真实性。iGibson 2.0[78] 版本引入了更为复杂的非运动学对象状态，如温度、湿度、可切割等，并支持通过 VR 接口采集人类演示数据，这使其成为研究复杂、长时程家庭任务的有力工具。

SAPIEN[79]：由加州大学圣迭戈分校等机构开发，核心优势在于其高精度的物理模拟，其基于强大的物理引擎（如 PhysX），支持精细的关节控制和逼真的接触物理学。SAPIEN 适用于研究需要精细操作和复杂动力学交互的机器人任务，如灵巧手操作和工具使用。

NVIDIA Isaac Sim/Isaac Gym[80]：由 NVIDIA 公司基于其 Omniverse 平台打造，Isaac Sim 是面向机器人研究的下一代仿真工具。其最大特点是提供了极致的真实感和高性能的 GPU 加速物理模拟。Isaac Gym 作为其核心组件之一，能够直接在 GPU 上实现大规模并行的强化学习训练，极大地缩短了训练时间。Isaac Sim 强大的虚实迁移（Sim-to-Real）能力和对工业级机器人（如 Franka Emika、Universal Robots）的广泛支持，使其成为学术界和工业界进行前沿机器人研究的热门选择。

RoboHive[81]：由斯坦福大学开发，是一个基于 MuJoCo[82] 物理引擎的机器人模拟框架。它不追求场景的视觉真实感，而是专注于提供一个模块化、可扩展的环境，用于研究机器人学习的核心问题。RoboHive 提供了一系列标准化的、经过精心设计的机器人操控任务（例如开门、堆叠积木、使用工具），这些任务被广泛用作评估机器人多任务学习和元学习能力的基准。

在选择仿真器时,往往需要在多个维度上进行权衡,如物理真实性、视觉真实感、模拟效率和任务多样性。例如,当研究重点为大规模导航策略时,Habitat 的速度优势是首选;当任务涉及复杂的物理交互和物体状态变化时,iGibson 或 SAPIEN 更为合适;当追求极致的视觉真实感和高效的并行强化学习训练时,NVIDIA Isaac Sim 是理想选择;而当需要在标准化的多任务环境中验证算法的泛化性时,RoboHive 则提供了绝佳的平台。

3.2 核心数据集与基准

高质量、大规模、多样化的数据集是训练 VLA 模型泛化能力并进行公平评估的基础[83]。近年来,一系列精心设计的基准不仅推动了模型的发展,也定义了领域内的核心研究问题,如表 5 所示,数据集覆盖了从语言理解到物理操控、多任务泛化到终身学习等多样化研究维度。

表 5 具身智能任务数据集与基准

Tab. 5 Embodied intelligence task dataset and benchmark

仿真器	核心聚焦点
OXE	高保真模拟与跨平台部署能力,支持 Sim2Real 的模仿学习研究
BridgeData V2	多样化机器人操作数据集
ALFRED	视觉-语言任务中多阶段目标分解与长指令理解执行
BEHAVIOR-1K	面向日常家庭活动的大规模任务库与常识泛化评估框架
CALVIN	多任务混合训练下的高鲁棒性连续操控与零样本泛化能力
ManiSkill	强物理属性建模支持下的操控策略物体类别泛化研究
RLBench	语言+视觉+动作模态融合的多任务模仿学习与低样本适应
LIBERO	多任务语言指令下的持续学习评估框架,关注抗遗忘与泛化能力

其中,Open X-Embodiment[39](OXE)是目前规模最大、覆盖最广的多机器人真实交互轨迹集,由 Google DeepMind 联合 21 家研究机构发布。它汇集了来自 22 种具身平台、超过 100 万条轨迹,统一了 VLA 三模态格式,成为推动通用型机器人策略学习的重要里程碑。OXE 的前身之一 Bridge Dataset[84] 聚焦于厨房场景的多任务操控,展示了真实数据在训练高性能模仿学习模型方面的潜力。

BridgeData V2 将原来的数据集扩展到 24 个环境、13 种技能,收录近 54 000 条高质量、标注良好的 VLA 示范。其中,大部分数据来源于基础技能,如拾取与放置、推动、重新定位物体。在此基础上,数据集还包含了一些更复杂的技能,包括:打开和关闭门和抽屉、擦拭表面、折叠布、堆叠积木、扭动旋钮、拨动开关以及转动水龙头等。为了保证学习到的技能具备良好的泛化能力,数据中每种技能均在多样化的目标和环境条件下采集示例,从而提升技能在不同场景下的可迁移性和适应性。

在复杂任务与语言规划能力方面,ALFRED[74] 提供了一个结合高层指令与细化执行步骤的家庭任务平台,包含 25 743 条英语指令,描述了 8 055 个专家演示,每个演示平均有 50 个步骤,产生了 428 322 个图像动作对。其多步、不可逆的任务结构对模型的长时程记忆与推理能力构成挑战,已成为衡量 VLA 综合智能水平的重要基准,其评价指标包括任务成功率、目标条件成功率(例如完成某个任务需要多个步骤,计算动作结束时完成的子任务数量与需要完成子任务数量的比例)与路径加权指标等。同类任务中,CALVIN[85] 强调多指令连续执行,要求模型具备持续状态感知与高鲁棒性;而 BEHAVIOR-1K[86] 则以 1 000 项人类日常活动为目标,突出模型在层级规划、常识理解与复杂环境泛化上的能力,是朝向真正具身智能体的关键挑战集。

在仿真环境中,ManiSkill[87] 与 RLBench[88] 提供了丰富的对象交互任务与可控演示数据。其中,ManiSkill 强调跨物体泛化与物理推理,而 RLBench 则因其专家轨迹生成机制,在少样本与模仿学习研究中广泛应用。

此外,LIBERO[89] 专门设计面向机器人持续学习的基准和数据集,通过涵盖物体、空间与程序知识迁移的任务体系,考察模型的持续学习、正向迁移与抗遗忘能力。LIBERO 共包含 130 个机器人操作任务,分为 4 个子集,用于分析不同任务变化维度的影响:LIBERO-SPATIAL 主要包括物体空间位置变化的数据,LIBERO-OBJECT 主要包括物体类别变化的数据,LIBERO-GOAL 主要包括任务目标变化的数据,LIBERO-LONG 主要包括更长、更复杂任务的数据。每个任务子集中均包含多个任务示例,并提供演示数据。其评估体系基于任务成功率,前向迁移能力衡量利用旧知识学习新任务的效率,负后向迁移衡量学习新任务后对旧知识的遗忘程度,成功率曲线下降面积衡量前向迁移和后向迁移的综合表现。

这些数据集不仅反映了 VLA 模型在控制精度、任务理解与泛化能力上的不同要求，也构成了评估具身智能系统多维能力的基本参考框架。随着基准数据与评测体系的不断完善，VLA 研究正处在从技术探索走向系统性综合智能的发展阶段[90]。

3.3 评估指标

对 VLA 模型性能的量化评估是推动该领域发展的重要基础。最核心的指标通常是任务成功率，它衡量智能体在多次测试中成功完成任务的比例。任务成功通常由预设的条件判定，如目标物体是否被正确放置、动作是否达到预期效果等。

在涉及空间移动的任务中，导航误差（Navigation Error）也被广泛使用[91]，用于衡量任务结束时智能体当前位置与目标位置之间的距离，反映了模型在导航和定位上的精度。相比之下，目标进度（Goal Progress）则关注智能体在执行过程中向目标有效靠近的程度，更能体现任务执行的连续过程。对于需要准确识别和定位目标物体的任务，交并比（Intersection over Union，IoU）是常见的评价手段，用来衡量模型预测的目标区域与真实目标区域的重合度，直观反映视觉定位的准确性。另外，在需要从多个候选动作或选项中选取正确结果的场景中，平均排名（Mean Rank）和 Top-K 准确率则用于衡量模型对正确答案的排序能力和检索性能。

具身智能任务往往具有环境复杂、任务多样且目标不唯一的特点，单一指标难以全面反映模型能力。因此，研究者也在尝试引入更丰富的评价方式，如人类偏好评分以及大模型评分、轨迹相似度、执行效率等，以多维度综合评估模型的表现和实际应用价值[92]。

4 VLA 模型的应用领域

VLA 模型通过将高级语义理解与低级物理执行无缝连接，极大地拓展了机器人的应用边界。其能力不再局限于执行预设的程序，而是能够理解人类的意图并在真实世界中灵活地完成任务。这使得 VLA 模型的应用正在从受控的实验室环境迅速走向开放、动态的现实世界服务场景。如图 6 所示，VLA 模型的应用从自主导航到机器人动作与家庭服务，再到工业自动化落地场景等均有应用。

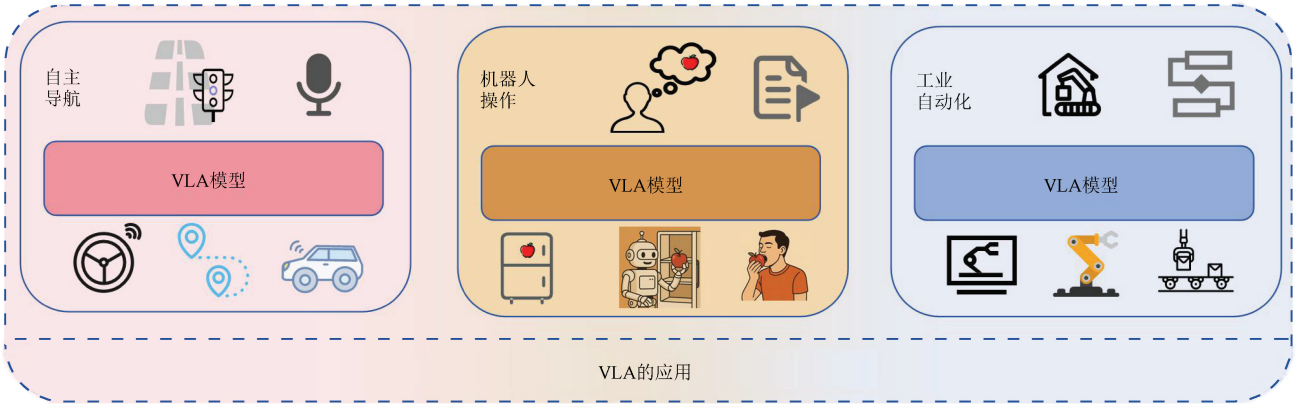


图 6 VLA 模型的应用领域

Fig. 6 Application fields of VLA models

4.1 机器人操作与家庭服务

机器人操作是 VLA 模型最核心且研究最深入的应用领域。家庭环境中的任务种类繁多且充满不确定性，VLA 模型赋予机器人前所未有的灵活性和通用性，使其能够完成各种复杂的日常家务。它不仅能够执行通用的物体操作，如从简单的“拾取-放置”到打开和关闭抽屉、柜门，将物品放入容器或整理桌面，还具备处理非刚性物体的能力。例如，在折叠衣物和铺平餐巾这类任务中，机器人需要对物体的物理特性有更深刻的理解。VLA 模型还需能够理解并执行基于用户意图的间接指令。以“我想吃苹果”这一任务为例，模型首先需要进行解析，将用户的愿望解析为“获取一个苹果并递给我”的具体操作目标。随后，模型结合常识知识（苹果通常存

放于冰箱)与视觉感知,自主规划出一个包含导航、交互与操作的复杂任务序列:① 导航至冰箱;② 打开箱门;③ 通过视觉识别并抓取一个苹果;④ 关闭箱门;⑤ 返回用户位置并完成递送。在此过程中,模型能够将语言概念(如“苹果”“冰箱”)与视觉实体精准关联,并生成连贯的底层动作指令,充分展现了其在家庭服务场景下处理多步复杂任务的能力。

4.2 自主导航

在自主导航领域,VLA 模型提供了全新的解决方案,尤其适合处理需要与环境和指令进行复杂交互的场景。模型能够融合几何指令和周围环境的视觉环境进行分析,并动态响应生成一系列动作,最终导航到目标[93]。这种端到端的方法有望取代传统复杂的模块化系统,更好地应对复杂导航环境。VLA 模型被广泛应用于无人机[94]、移动机器人等,实现基于自然语言指令的自主导航。模型需要理解指令、识别目标并规划路径。RaceVLA[95]模型展示了 VLA 在无人机中实时处理视觉反馈并调整飞行策略的能力,而专为腿式机器人设计的 NaVILA[96]框架,则通过分层结构将高级语言指令转化为精确的步态控制,使机器人能够在崎岖地形中稳定导航。OpenDriveVLA[97]则采取分层视觉语言对齐,将 2D 和 3D 结构化的视觉词元投影到统一的语义空间中,在开环轨迹规划以及驾驶问答方面取得了先进成果。

4.3 工业自动化

VLA 模型正推动工业机器人发生范式转变,将传统依赖刚性编程、只能在高度结构化环境中运行的机器人,逐步演化为能够进行高级推理、灵活执行多样化任务并与人类自然交互的智能机器人[98]。这一转型不仅降低了机器人在新产线、新产品变型下的重新编程和调试成本,还提高了在动态制造场景下的适应性和协作性,为实现真正的自主、智能化自动化工厂奠定了基础。借助 VLA 模型,机器人可以感知并理解视觉输入、语言指令和自身状态,推理任务意图并实时生成合理的控制决策,更好地完成高复杂度的工业任务。CogACT[99]引入基于扩散动作的 Transformer 生成精细电机动作,在多步装配、螺丝紧固以及分拣零件等高精度任务中表现优异。

5 核心挑战

尽管 VLA 模型取得了令人瞩目的进展,但将其从实验室推向广泛的现实世界应用仍面临诸多严峻挑战。当前的研究前沿正围绕这些挑战展开,其核心目标已从早期的“能力实现”转向追求部署时的“可靠性、安全性和效率”。这标志着该领域正在从“概念验证”迈向“工程落地”的成熟阶段。

5.1 泛化性与数据效率

当前 VLA 模型,特别是主要依赖行为克隆训练的模型,对训练数据的分布极为敏感。当面对训练中见过的物体、背景、光照条件或任务指令时,其性能往往会急剧下降[100]。这种泛化能力的不足是阻碍其在开放、非结构化环境中可靠部署的主要障碍。不同于可以直接使用海量互联网图文数据训练的 LLM 和 VLM,VLA 模型需要包含精确动作标签和环境反馈的多模态机器人交互数据[101]。而在现实环境中采集此类数据,不仅需要昂贵的硬件系统和严格的安全保障,还面临大量人力与时间成本。这种数据获取上的瓶颈,已成为当前 VLA 系统扩展能力和跨任务适应性的核心限制因素。更为根本的问题在于,VLA 系统模型能力不足限制了对高质量交互数据的采集,而数据不足又反过来限制了模型能力的提升。

5.2 长时程任务与推理

现实世界中的许多任务通常具有长时程特性,需要执行数个相互关联的步骤才能完成。对此,VLA 模型在记忆保持、任务规划和复杂推理方面面临极大挑战[102]。当前主流的自回归动作生成方法,由于逐步预测动作,容易导致误差累积,表现为“顾此失彼”的困境,最终影响任务的整体完成效果。仅依赖自回归模型难以捕捉任务间的长期依赖与全局约束,缺乏对未来状态的预判和整体策略规划。对此,未来 VLA 架构可能需要引入世界模型,以模拟环境动态、预测未来状态,同时结合长期记忆模块(如 3D 场景图、任务历史记录)实现对历史信息的存储与调用。通过综合当前观测、历史记忆和对未来可能情景的“想象”,模型才能实现更具前瞻性和鲁棒性的决策制定[103-104]。尽管这一思路已有部分初步尝试,但现阶段大多数 VLA 模型仍未系统集成长期推理和世界建模,导致其在复杂多步骤任务中的表现受限。显然,突破长时程推理瓶颈,是推动 VLA 从“动作生成”向“智能规划”迈进的关键路径。

5.3 实时响应速度

不同于主要处理文本、图像等静态数据的 AI 范式,VLA 模型的核心应用场景是与动态变化的物理世界进行持续的实时交互,对模型的决策实时性提

出了极高的要求。一旦模型的推理速度不及环境的变化速度,系统便可能基于陈旧的状态感知,持续生成无效甚至危险的动作指令。然而,当前 VLA 模型在性能与速度之间面临着一种固有矛盾。为追求更强的泛化能力与长时程推理能力,研究者倾向于采用更大规模的 VLM 或集成更复杂的任务规划架构,但这几乎不可避免地导致了推理时间的延长。因此,如何在保证高层决策能力的前提下,通过模型压缩、架构优化或设计高效推理算法等手段,在计算性能和响应速度之间达成最佳平衡,是推动 VLA 模型从理论验证走向工程落地必须攻克的关键瓶颈。

6 结束语

VLA 模型作为连接高级认知与物理执行的关键技术,已成为具身智能领域的核心。通过端到端统一视觉感知、语言理解与动作生成,VLA 突破了传统模块化架构的局限,展现出卓越的通用性和灵活性。现代 VLA 模型以大规模预训练为基础,实现了从海量互联网知识向机器人操作的有效迁移,极大提升了泛化能力和复杂行为的涌现。生态系统的完善,包括多样化的仿真平台、丰富的数据集及标准化基准,为 VLA 的发展提供了坚实的支撑。与此同时,泛化性、数据效率、长时规划等难题仍亟待突破。

参考文献

[1] MA Y E, SONG Z X, ZHUANG Y Z, et al. A Survey on Vision-Language-Action Models for Embodied AI[EB/OL]. (2025-03-04)[2025-07-10]. https://arxiv.org/abs/2405.14093.

[2] SHIN H C, ROTH H R, GAO M C, et al. Deep Convolutional Neural Networks for Computer-aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning[J]. IEEE Transactions on Medical Imaging, 2016, 35(5):1285-1298.

[3] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control[C]//Proceedings of Conference on Robot Learning. Atlanta:PMLR,2023:2165-2183.

[4] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale[EB/OL]. (2020-10-22)[2025-07-10]. https://arxiv.org/abs/2010.11929.

[5] CHANG Y P, WANG X, WANG J D, et al. A Survey on Evaluation of Large Language Models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3):1-45.

[6] ZHANG J Y, HUANG J X, JIN S, et al. Vision-Language Models for Vision Tasks: A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2024,46(8):5625-5644.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[EB/OL]. (2017-06-12)[2025-07-10]. https://arxiv.org/abs/1706.03762.

[8] BROHAN A, BROWN N, CARBAJAL J, et al. RT-1: Robotics Transformer for Real-world Control at Scale[EB/OL]. (2022-12-13)[2025-07-10]. https://arxiv.org/abs/2212.06817.

[9] DENG J, DONG W, SOCHER R, et al. ImageNet: A Large-scale Hierarchical Image Database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009:248-255.

[10] TAN M X, LE Q V, et al. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]//International Conference on Machine Learning. Long Beach: PMLR, 2019:6105-6114.

[11] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016:770-778.

[12] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: Learning Robust Visual Features Without Supervision[EB/OL]. (2023-04-14)[2025-07-10]. https://arxiv.org/abs/2304.07193.

[13] ZHAI X H, MUSTAFA B, KOLESNIKOV A, et al. Sigmoid Loss for Language Image Pre-training[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris:IEEE/CVF,2023:11941-11952.

[14] DRIESS D, XIA F, SAJJADI M S M, et al. PaLM-E: An Embodied Multimodal Language Model[C]// Proceedings of the 40th International Conference on Machine Learning. Honolulu:PMLR,2023: 8469-8488.

[15] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and Efficient Foundation Language Models[EB/OL]. (2023-02-27)[2025-07-10]. https://arxiv.org/abs/2302.13971.

[16] MESNARD T, HARDIN C, DADASHI R, et al. Gemma: Open Models Based on Gemini Research and Technology[EB/OL]. (2024-03-13)[2025-07-10]. https://arxiv.org/abs/2403.08295.

[17] SAPKOTA R, CAO Y, ROUMELIOTIS K I, et al. Vision-Language-Action Models: Concepts, Progress, Applications and Challenges[EB/OL]. (2025-05-07)[2025-07-10]. https://arxiv.org/abs/2505.04769.

[18] PEREZ E, STRUB F, DE VRIES H, et al. FiLM: Visual Reasoning with a General Conditioning Layer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018: 3942-3951.

[19] DU Y, LIU Z, LI J, et al. A Survey of Vision-Language Pre-trained Models[C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022). Vienna: IJCAI Organization, 2022: 5582-5590.

[20] JIANG Y F, GUPTA A, ZHANG Z C, et al. VIMA: General Robot Manipulation with Multimodal Prompts [C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023: 1-42.

[21] KIM M J, PERTSCH K, KARAMCHETI S, et al. OpenVLA: An Open-source Vision-Language-Action Model [EB/OL]. (2024-06-13) [2025-07-10]. https://arxiv.org/abs/2406.09246.

[22] WEN J J, ZHU Y C, LI J M, et al. TinyVLA: Towards Fast, Data-efficient Vision-Language-Action Models for Robotic Manipulation[EB/OL]. (2024-09-19) [2025-07-10]. https://arxiv.org/abs/2409.12514.

[23] ZHEN H Y, QIU X W, CHEN P H, et al. 3D-VLA: A 3D Vision-Language-Action Generative World Model[EB/OL]. (2024-03-14) [2025-07-10]. https://arxiv.org/abs/2403.09631.

[24] REED S, ZOLNA K, PARISOTTO E, et al. A Generalist Agent[EB/OL]. (2022-05-12) [2025-07-10]. https://arxiv.org/abs/2205.06175.

[25] HO J, JAIN A, ABBEEL P. Denoising Diffusion Probabilistic Models[EB/OL]. (2020-06-19) [2025-07-10]. https://arxiv.org/abs/2006.11239.

[26] CHI C, XU Z J, FENG S Y, et al. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion [EB/OL]. (2023-03-07) [2025-07-10]. https://arxiv.org/abs/2303.04137.

[27] BLACK K, BROWN N, DRIESS D, et al. π_0: A Vision-Language-Action Flow Model for General Robot Control [EB/OL]. (2024-10-31) [2025-07-10]. https://arxiv.org/abs/2410.24164.

[28] OCTO MODEL TEAM, GHOSH D, WALKE H, et al. Octo: An Open-source Generalist Robot Policy [EB/OL]. (2024-05-20) [2025-07-10]. https://arxiv.org/abs/2405.12213.

[29] LIPMAN Y, CHEN R T Q, BEN-HAMU H, et al. Flow Matching for Generative Modeling[EB/OL]. (2022-10-06) [2025-07-10]. https://arxiv.org/abs/2210.02747.

[30] PHYSICAL INTELLIGENCE, BLACK K, BROWN N, et al. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization[EB/OL]. (2025-04-22) [2025-07-10]. https://arxiv.org/abs/2504.16054.

[31] JIANG S C, HUANG Z L, QIAN K G, et al. A Survey on Vision-Language-Action Models for Autonomous Driving[EB/OL]. (2025-06-30) [2025-07-10]. https://arxiv.org/abs/2506.24044.

[32] HONG Y C, WU Q, QI Y K, et al. A Recurrent Vision-and-Language BERT for Navigation[EB/OL]. (2020-11-26) [2025-07-10]. https://arxiv.org/abs/2011.13922.

[33] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.

[34] SHRIDHAR M, MANUELLI L, FOX D. CLIPort: What and Where Pathways for Robotic Manipulation[EB/OL]. (2021-09-24) [2025-07-10]. https://arxiv.org/abs/2109.12098.

[35] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models from Natural Language Supervision[C]// Proceedings of the 38th International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748-8763.

[36] KAWAHARAZUKA K, OH J, YAMADA J, et al. Vision-Language-Action Models for Robotics: A Review Towards Real-world Applications [J]. IEEE Access, 2025, 13: 162467-162504.

[37] CHEN X, DJOLONGA J, PADLEWSKI P, et al. PaLI-X: On Scaling up a Multilingual Vision and Language Model [EB/OL]. (2023-05-29) [2025-07-10]. https://arxiv.org/abs/2305.18565.

[38] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open Foundation and Fine-tuned Chat Models[EB/OL]. (2023-07-18) [2025-07-10]. https://arxiv.org/abs/2307.09288.

[39] O'NEILL A, REHMAN A, MASSUKURI A, et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models[EB/OL]. (2023-10-13) [2025-07-10]. https://arxiv.org/abs/2310.08864.

[40] HU E J, SHEN Y L, WALLIS P, et al. LoRA: Low-rank Adaptation of Large Language Models [EB/OL]. (2021-06-17) [2025-07-10]. https://arxiv.org/abs/2106.09685.

[41] ZHU M J, ZHU Y C, LI J M, et al. ObjectVLA: End-to-End Open-World Object Manipulation Without Demonstration[EB/OL]. (2025-02-26)[2025-07-10]. https://arxiv.org/abs/2502.19250.

[42] HAN X F, CHEN S P, FU Z H, et al. Multimodal Fusion and Vision-Language Models: A Survey for Robot Vision [EB/OL]. (2025-04-03)[2025-07-10]. https://arxiv.org/abs/2504.02477.

[43] HUANG W L, WANG C, ZHANG R H, et al. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models[EB/OL]. (2023-07-12)[2025-07-10]. https://arxiv.org/abs/2307.05973.

[44] LI C M, WEN J J, PENG Y, et al. PointVLA: Injecting the 3D World into Vision-Language-Action Models[EB/OL]. (2025-03-10)[2025-07-10]. https://arxiv.org/abs/2503.07511.

[45] QU D L, SONG H M, CHEN Q Z, et al. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model[C]//Robotics: Science and Systems (RSS 2025). Utrecht: Robotics: Science and Systems Foundation, 2025:1-19.

[46] PAN M Y, ZHANG J M, WU T S, et al. OmniManip: Towards General Robotic Manipulation via Object-centric Interaction Primitives as Spatial Constraints[EB/OL]. (2025-01-07)[2025-07-10]. https://arxiv.org/abs/2501.03841.

[47] ZHANG C H, HAO P Z, CAO X K, et al. VTLA: Vision-Tactile-Language-Action Model with Preference Learning for Insertion Manipulation[EB/OL]. (2025-05-14)[2025-07-10]. https://arxiv.org/abs/2505.09577.

[48] HAO P, ZHANG C H, LI D Z, et al. TLA: Tactile-Language-Action Model for Contact-rich Manipulation[EB/OL]. (2025-03-11)[2025-07-10]. https://arxiv.org/abs/2503.08548.

[49] SHI L X Y, ICHTER B, EQUI M, et al. Hi Robot: Open-Ended Instruction Following with Hierarchical Vision-Language-Action Models[EB/OL]. (2025-02-26)[2025-07-10]. https://arxiv.org/abs/2502.19417.

[50] BJORCK J, CASTAÑEDA F, CHERNIADEV N, et al. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots[EB/OL]. (2025-03-18)[2025-07-10]. https://arxiv.org/abs/2503.14734.

[51] LI Y H, WEI F Y, ZHANG C, et al. EAGLE-2: Faster Inference of Language Models with Dynamic Draft Trees [EB/OL]. (2024-06-24)[2025-07-10]. https://arxiv.org/abs/2406.16858.

[52] PEEBLES W, XIE S N. Scalable Diffusion Models with Transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 4172-4182.

[53] Figure. Helix: A Vision-Language-Action Model for Generalist Humanoid Control [EB/OL]. (2025-02-20)[2025-07-10]. https://www.figure.ai/news/helix.

[54] WEN J J, ZHU Y C, LI J M, et al. DexVLA: Vision-Language Model with Plug-in Diffusion Expert for General Robot Control[EB/OL]. (2025-02-10)[2025-07-10]. https://arxiv.org/abs/2502.05855.

[55] XIONG J Z, LIU G C, HUANG L, et al. Autoregressive Models in Vision: A Survey[EB/OL]. (2024-11-08)[2025-07-10]. https://arxiv.org/abs/2411.05902.

[56] LU J S, CLARK C, LEE S, et al. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 26429-26455.

[57] ZAWALSKI M, CHEN W, PERTSCH K, et al. Robotic Control via Embodied Chain-of-Thought Reasoning[EB/OL]. (2024-07-11)[2025-07-10]. https://arxiv.org/abs/2407.08693.

[58] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought Prompting Elicits Reasoning in Large Language Models[C]//Advances in Neural Information Processing Systems 36 (NeurIPS 2022). New Orleans: NeurIPS, 2022: 24824-24837.

[59] DU Y L, YANG M J, DAI B, et al. Learning Universal Policies via Text-guided Video Generation[C]//Advances in Neural Information Processing Systems 37 (NeurIPS 2023). New Orleans: NeurIPS, 2023: 1-17.

[60] ZHAO T Z, KUMAR V, LEVINE S, et al. Learning Fine-grained Bimanual Manipulation with Low-cost Hardware[EB/OL]. (2023-04-23)[2025-07-10]. https://arxiv.org/abs/2304.13705.

[61] KIM M J, FINN C, LIANG P, et al. Fine-tuning Vision-Language-Action Models: Optimizing Speed and Success [C]//Robotics: Science and Systems (RSS 2025). Utrecht: Robotics: Science and Systems Foundation, 2025: 1-24.

[62] CEN J, YU C H, YUAN H J, et al. WorldVLA: Towards Autoregressive Action World Model [EB/OL]. (2025-06-26)[2025-07-10]. https://arxiv.org/abs/2506.21539.

[63] JIANG A X, GAO Y, SUN Z G, et al. DiffVLA: Vision-Language Guided Diffusion Planning for Autonomous Driving[EB/OL]. (2025-05-26)[2025-07-10]. https://arxiv.org/abs/2505.19381.

[64] SHUKOR M, AUBAKIROVA D, CAPUANO F, et al. SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics[EB/OL]. (2025-06-02)[2025-07-10]. https://arxiv.org/abs/2506.01844.

[65] LIU S M, WU L X, LI B G, et al. RDT-1B: A Diffusion Foundation Model for Bimanual Manipulation[EB/OL]. (2024-10-10)[2025-07-10]. https://arxiv.org/abs/2410.07864

[66] XUE H R, REN J J, CHEN W D, et al. Reactive Diffusion Policy: Slow-fast Visual-tactile Policy Learning for Contact-rich Manipulation[EB/OL]. (2025-03-04)[2025-07-10]. https://arxiv.org/abs/2503.02881.

[67] LIU J M, CHEN H R, AN P J, et al. HybridVLA: Collaborative Diffusion and Autoregression in a Unified Vision-Language-Action Model[EB/OL]. (2025-03-13)[2025-07-10]. https://arxiv.org/abs/2503.10631.

[68] GHOSH A, ACHARYA A, SAHA S, et al. Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions[EB/OL]. (2024-04-12)[2025-07-10]. https://arxiv.org/abs/2404.07214v2.

[69] PsiBot. The Second Wave of Real VLA: Psi R1 Achieves Generalized Intelligence at the Brain Level![EB/OL]. (2025-04-29)[2025-07-10]. https://www.psibot.ai/en/008_en/.

[70] ZHONG Y F, HUANG X C, LI R C, et al. DexGraspVLA: A Vision-Language-Action Framework Towards General Dexterous Grasping[EB/OL]. (2025-02-28)[2025-07-10]. https://arxiv.org/abs/2502.20900.

[71] LIN F Q, NAI R Q, HU Y D, et al. OneTwoVLA: A Unified Vision-Language-Action Model with Adaptive Reasoning[EB/OL]. (2025-05-17)[2025-07-10]. https://arxiv.org/abs/2505.11917.

[72] LI Z X, WU X Y, DU H Y, et al. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges[EB/OL]. (2025-01-04)[2025-07-10]. https://arxiv.org/abs/2501.02189.

[73] KOLVE E, MOTTAGHI R, HAN W, et al. AI2-THOR: An Interactive 3D Environment for Visual AI[EB/OL]. (2017-12-14)[2025-07-10]. https://arxiv.org/abs/1712.05474.

[74] SHRIDHAR M, THOMASON J, GORDON D, et al. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. eattle:IEEE,2020:10737-10746.

[75] SAVVA M, KADIAN A, MAKSYMETS O, et al. Habitat: A Platform for Embodied AI Research[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9339-9347.

[76] CHANG A, DAI A, FUNKHOUSER T, et al. Matterport3D: Learning from RGB-D Data in Indoor Environments[C]//Proceedings of the IEEE International Conference on Computer Vision. Qingdao:IEEE,2017:667-676.

[77] SHEN B K, XIA F, LI C S, et al. iGibson 1.0: A Simulation Environment for Interactive Tasks in Large Realistic Scenes[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague: IEEE, 2021:7520-7527.

[78] LI C S, XIA F, MARTÍN-MARTÍN R, et al. iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks[EB/OL]. (2021-08-06)[2025-07-10]. https://arxiv.org/abs/2108.03272.

[79] XIANG F B, QIN Y Z, MO K C, et al. SAPIEN: A Simulated Part-based Interactive Environment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle:IEEE,2020:11097-11107.

[80] MAKOVIYCHUK V, WAWRZYNIAK L, GUO Y R, et al. Isaac Gym: High Performance GPU-based Physics Simulation for Robot Learning[EB/OL]. (2021-08-24)[2025-07-10]. https://arxiv.org/abs/2108.10470.

[81] KUMAR V, SHAH R, ZHOU G Y, et al. RoboHive: A Unified Framework for Robot Learning[C]//37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. New Orleans: NeurIPS,2023:44323-44340.

[82] TODOROV E, EREZ T, TASSA Y. MuJoCo: A Physics Engine for Model-based Control[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura-Algarve: IEEE, 2012: 5026-5033.

[83] GURUPRASAD P, SIKKA H, SONG J W, et al. Benchmarking Vision, Language, & Action Models on Robotic Learning Tasks[EB/OL]. (2024-11-04)[2025-07-10]. https://arxiv.org/abs/2411.05821.

[84] EBERT F, YANG Y F, SCHMECKPEPER K, et al. Bridge Data: Boosting Generalization of Robotic Skills with Cross-domain Datasets[EB/OL]. (2021-09-27)[2025-07-10]. https://arxiv.org/abs/2109.13396.

[85] MEES O,HERMANN L,ROSETE-BEAS E,et al. CALVIN:A Benchmark for Language-Conditioned Policy Learning for Long-horizon Robot Manipulation Tasks[J]. IEEE Robotics and Automation Letters,2022,7(2):7327-7334.

[86] LI C, ZHANG R, WONG J, et al. Behavior-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation[C]//Conference on Robot Learning. Atlanta:PMLR,2023: 80-93.

[87] GU J Y,XIANG F B,LI X L,et al. ManiSkill2:A Unified Benchmark for Generalizable Manipulation Skills[C]// International Conference on Learning Representations (ICLR 2023). Kigali:PMLR,2023:1-30.

[88] 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.

[89] LIU B, ZHU Y F, GAO C K, et al. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning [C]// NIPS'23: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc. ,2023:44776-44791.

[90] GURUPRASAD P,WANG Y Y,CHOWDHURY S,et al. Benchmarking Vision,Language,& Action Models in Procedurally Generated, Open Ended Action Environments [EB/OL]. (2025-05-08)[2025-07-10]. https://arxiv.org/abs/2505.05540.

[91] GU J, STEFANI E, WU Q, et al. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2022). Dublin: Association for Computational Linguistics, 2022: 7606-7623.

[92] JI Y H, TAN H J, SHI J Y, et al. RoboBrain: A Unified Brain Model for Robotic Manipulation from Abstract to Concrete[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville:IEEE, 2025: 1724-1734.

[93] 司马双霖,黄岩,何科技,等. 视觉语言导航研究进展[J]. 自动化学报,2023,49(1):1-14.

[94] 杨玉琪,王梦云,刘运卓,等. 具身智能及其在自主无人系统的应用研究[J]. 无人系统技术,2024,7(5):99-110.

[95] SERPIVA V,LYKOV A,MYSHLYAEV A,et al. RaceVLA: VLA-based Racing Drone Navigation with Human-like Behaviour[EB/OL]. (2025-03-04)[2025-07-10]. https://arxiv.org/abs/2503.02572。

[96] CHENG A C, JI Y Z, YANG Z J, et al. NaviLa: Legged Robot Vision-Language-Action Model for Navigation[EB/OL]. (2024-12-05)[2025-07-10]. https://arxiv.org/abs/2412.04453.

[97] ZHOU X C, HAN X Y, YANG F, et al. OpenDriveVLA: Towards End-to-End Autonomous Driving with Large Vision Language Action Model[EB/OL]. (2025-03-30)[2025-07-10]. https://arxiv.org/abs/2503.23463.

[98] QIAN K,SUN T Y,WANG W H. Exploring Large Vision-Language Models for Robust and Efficient Industrial Anomaly Detection[EB/OL]. (2024-12-01)[2025-07-10]. https://arxiv.org/abs/2412.00890.

[99] LI Q X,LIANG Y B, WANG Z Y, et al. CogAct: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation[EB/OL]. (2024-11-29)[2025-07-10]. https://arxiv.org/abs/2411.19650.

[100] FRANCIS J, KITAMURA N, LABELLE F, et al. Core Challenges in Embodied Vision-Language Planning[J]. Journal of Artificial Intelligence Research, 2022, 74: 459-515.

[101] 邓鹏,唐文涛,罗静. 机器人大模型发展与挑战[J]. 电子测量与仪器学报,2024,38(12):12-25.

[102] MOGADALA A, KALIMUTHU M, KLAKOW D. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods[J]. Journal of Artificial Intelligence Research,2021,71:1183-1317.

[103] CHERN E,HU Z L,CHERN S,et al. Thinking with Generated Images[EB/OL]. (2025-05-28)[2025-07-10]. https://arxiv.org/abs/2505.22525.

[104] LI C Z, WU W S, ZHANG H Y, et al. Imagine While Reasoning in Space: Multimodal Visualization-of-Thought [EB/OL]. (2025-01-13)[2025-07-10]. https://arxiv.org/abs/2501.07542.

作者简介

袁霆宇 男,(2002—),博士研究生。主要研究方向:具身智能、多模态大模型、3D生成、强化学习等。

刘 凯 男,(2002—),硕士研究生。主要研究方向:具身智能、多模态大模型、计算机视觉等。

关标良 男,(2002—),硕士研究生。主要研究方向:具身智能、强化学习、计算机视觉、多模态大模型等。

叶 雯 女,(2003—),博士研究生。主要研究方向:具身智能、多模态大模型智能体、多智能体系统、AIGC、计算机视觉等。

赵雅萃 女,(2002—),硕士研究生。主要研究方向:具身智能、多模态大模型等。

赵朝阳 男,(1985—),博士,副研究员。主要研究方向:视频图像分析、多模态大模型、具身智能等。

王金桥 男,(1978—),博士,研究员。主要研究方向:具身智能、视频图像分析、多模态大模型、自监督学习、目标检测与跟踪、细粒度识别、行为识别。