基于大型视觉语言模型 (VLM) 的机器人操作：综述

机器人操作是机器人技术和具身人工智能 (embodied AI) 领域的关键前沿，需要精确的运动控制以及对动态环境中视觉和语义线索的综合理解。传统方法基于预先定义的任务规范和严格的控制策略，通常难以在非结构化的新场景中扩展或推广。近年来，基于在海量图像文本数据集上预训练的大型视觉语言模型 (VLM) 构建的视觉语言动作 (VLA) 模型已成为一种变革性范式。通过利用大型 VLM 在开放世界泛化、分层任务规划、知识增强推理和丰富的多模态融合方面的能力，这些模型使机器人能够解读高级指令、识别未知环境并执行复杂的操作任务。本综述首次系统地、以分类学为导向地综述了基于大型 VLM 的机器人操作 VLA 模型。我们首先明确定义基于 VLM 的大型 VLA 模型，并描述两种主要的架构范式：（1）单体模型，涵盖具有不同集成度的单系统和双系统设计；（2）分层模型，通过可解释的中间表示明确地将规划与执行分离。在此基础上，我们对基于 VLM 的大型 VLA 模型进行了深入研究：（1）与高级领域的集成，包括强化学习、无需训练的优化、从人类视频中学习以及世界模型集成；（2）综合各具特色的特征，整合其架构特性、操作优势以及支持其发展的数据集和基准；（3）识别有前景的方向，包括记忆机制、四维感知、高效自适应、多智能体协作以及其他新兴能力。本综述整合了最新进展，旨在解决现有分类法中的不一致问题，缓解研究碎片化，并通过系统地整合大型 VLM 与机器人操作交叉领域的研究来填补关键空白。我们提供了一个定期更新的项目页面来记录正在进行的进展：https://github.com/JiuTian-VL/Large 基于 VLM 的 VLA 用于机器人操控。

1 引言

机器人操控是机器人技术与具身人工智能融合的关键挑战 [1]–[5]。它的实现不仅需要精确的运动控制，还需要对复杂动态环境中各种视觉和语义线索的深刻理解。机器人操控在众多领域都具有广泛的应用，包括先进制造、高效物流、精准医疗保健和多功能家政服务 [6]–[8]。传统的操控方法 [9]–[16] 主要基于精心设计的控制策略和严格预定义的任务规范。然而，这些方法在非结构化的现实世界环境中——尤其是在面对新奇物体、模糊的自然语言指令或前所未有的环境配置时——明显存在缺陷，从而暴露出其在可扩展性和泛化能力方面的固有局限性。

近年来，大型视觉语言模型 (VLM) [17]– [25] 已成为一种变革性范式。大型 VLM 已在海量网络规模的图像文本数据集上进行预训练，展现出弥合视觉感知与自然语言理解之间语义鸿沟的卓越能力。这种创新能力使大型视觉语言模型 (VLM) 能够结合文本描述来解读复杂的视觉场景，超越单纯的物体识别，从而涵盖整体的情境理解。大型视觉语言模型 (VLM) 随后与机器人系统的集成，催生了一类新型模型：基于大型视觉语言模型 (VLM) 的视觉-语言-动作 (VLA) 模型 [26]– [32]。如图 1 所示，这种新兴范式在克服传统机器人流程的根本局限性方面具有巨大潜力。它使机器人能够解读高级人类指令，泛化至未知的物体和场景，推理复杂的空间关系，并在动态、非结构化环境中执行复杂的操作任务。例如，VLA 模型可以执行诸如“将笔记本电脑旁边的红色马克杯放在顶层架子上”之类的命令，这项任务需要将视觉基础、空间推理和顺序运动规划进行复杂的融合。

图 1：基于大型 VLM 的视觉-语言-动作 (VLA) 模型在机器人操作方面的核心优势图解。基于大型 VLM 的 VLA 模型充分利用了大型视觉-语言模型 (VLM) 的优势，包括 (1) 开放世界泛化能力、(2) 分层任务规划、(3) 知识增强推理以及 (4) 丰富的多模态融合。这些能力赋能了多样化的机械臂，并显著提升了机器人智能。

本文基于对近期研究 [26]–[37] 的广泛回顾以及对该领域 [38]–[43] 的深入理解，提出了基于大型 VLM 的 VLA 模型的统一定义：该模型 (1) 利用大型 VLM 理解视觉观察和自然语言指令，以及 (2) 执行直接或间接服务于机器人动作生成的推理过程。我们进一步区分了两大类基于 VLM 的 VLA 模型，如图 2 和图 3 所示：

• 单体模型（图 3 左）包含单系统和双系统实现。(1) 单系统模型 [26], [27], [44], [45] 将环境理解（包括视觉感知、语言理解和机器人状态感知）和动作生成集成在一个统一的架构中。相比之下，(2) 双系统模型 [29]–[32] 采用 VLM 主干网络进行场景解释，并采用动作专家进行动作决策，两者通过潜在表征的传播进行信息交换。

• 分层模型（图 3 右）[46]–[50] 明确地将规划与策略执行分离。这些方法与双系统端到端方法的区别在于两个定义性特征：(1) 结构化的中间输出，其中规划器模块生成可解释的表示，例如关键点检测、可供性图或轨迹提议，然后由策略模块处理以制定可执行的操作；(2) 解耦的训练范式，通过专门的损失函数或 API 介导的交互，实现分层模块的独立优化。

该分类法强调了 VLA 开发中的关键设计维度，尤其是在系统集成粒度和认知分解的明确性方面，同时保持了与现代表示学习范式的必要联系。

图 3：基于 VLM 的两大类 VLA 模型的比较。单体模型（第 3 节）在单系统或双系统架构中集成了感知、语言理解和动作生成，后者还包含一个额外的动作专家。相比之下，分层模型（第 4 节）通过可解释的中间输出（例如，子任务、关键点、程序、可供性）将规划与策略执行分离。

基于上述定义和分类法，我们对广泛的相关工作进行了全面调查，揭示了新兴 VLA 领域存在的几个关键差距，调查的组织结构如图 2 所示。首先，该领域的术语和建模假设仍然不一致，研究领域在各个学科（机器人技术、计算机视觉、自然语言处理等）之间呈碎片化。其次，现有的评论往往侧重于 VLM [51]–[55] 或孤立的机器人操作 [2]、[56]–[59]，缺乏对它们交叉领域中出现的独特挑战和进展的全面综合。因此，迫切需要进行一项系统且原则性的综述，阐明基于大型VLM的VLA模型的基础，梳理相关方法，并概述这一集成范式的未来发展方向。本综述旨在弥补这一空白。我们对基于大型VLM的VLA模型研究进展进行了结构化且深入的概述，旨在提供该领域的全景视角，以加深对该领域的理解并推动未来的突破。我们的主要贡献概括如下：

图 2：我们综合综述的组织结构概述（上图）以及基于大型 VLM 的机器人操作“视觉-语言-动作”模型显著发展的时序图（下图）。该时间线重点介绍了单体模型和分层模型的代表性里程碑，从而展现了该领域的最新进展。

• 基于大型 VLM 的 VLA 模型发展的纵向综述：我们系统地回顾了 VLM 的演进轨迹、操作学习的技术进步以及随后基于大型 VLM 的 VLA 范式的出现。此外，我们还研究了单体模型和分层模型的发展，识别了关键挑战并概述了未来发展方向。

• 基于大型 VLM 的 VLA 建模实践的横向综述：我们提供了更细粒度的单体模型和分层模型的比较分类，并从结构和功能角度对其进行了详细研究。我们进一步探索基于大型 VLM 的 VLA 模型的前沿研究，重点介绍其独特之处以及支撑其发展的数据集。

本综述提供了高层次的总结和概念路线图，有助于理解该领域的发展和结构组织。

本综述的其余部分组织如下：如图 2 所示，第 2 节介绍了 VLM 的演进和机器人操作基础方面的基本背景知识。第 3 节探讨了单体模型，详细介绍了单系统和双系统架构、它们各自的优势以及设计权衡。第 4 节探讨了分层模型，将其分为仅规划器和规划器策略框架，并进一步根据中间表示类型（例如子任务、关键点和程序）进行细分。第 5 节讨论了其他高级领域，包括基于强化学习的优化、免训练方法、从人类视频中学习以及基于世界模型的方法。第 6 节探讨了分层模型，将其分为仅规划器和规划器策略框架，并进一步根据中间表示类型（例如子任务、关键点和程序）进行细分。第 7 节讨论了其他高级领域，包括基于强化学习的优化、免训练方法、从人类视频中学习以及基于世界模型的方法。第6节分析了基于大型VLM的VLA模型的定义特征，涵盖多模态融合、指令遵循和多维泛化。第7节对基于大型VLM的VLA模型研究中使用的各种数据集和基准进行了分类和分析，涵盖了模拟数据、现实世界数据和人机交互数据。第8节探讨了关键的开放性挑战和有前景的未来研究方向。最后，第9节对全文进行了总结。

2 背景

2.1 视觉-语言模型的演进

近年来，大型VLM [17]–[25]的出现标志着其从特定任务架构向能够处理各种多模态任务的统一框架的转变 [60]–[64]。现代视觉语言模型 (VLM)，例如 LLaVA1.5 [18] 和 Qwen-VL [20]，通常采用三组件架构：用于编码视觉输入的视觉编码器、用于将视觉特征与文本嵌入对齐的投影器，以及用于介导多模态推理的大型语言模型。这种设计将多模态任务统一为文本输出。它使 VLM 不仅能够执行视觉问答或物体检测等传统任务，还能执行更高级的功能，包括组合推理和空间导航，这些功能对于实际应用至关重要。在此基础上，DriveVLM [65] 展示了 VLM 与自动驾驶系统的集成，而 CogAgent [66] 则展示了 VLM 与图形用户界面交互的能力。这些进步凸显了大型 VLM 在连接视觉和语言方面日益精进的实力，同时也揭示了利用大型 VLM 构建广泛实际应用的新兴潜力。

至关重要的是，大型视觉语言模型 (VLM) 在实际应用中的潜力 [67]–[76] 取决于其通过视觉指令调优所获得的泛化能力。通过在大规模、精心挑选的视觉指令跟踪数据集上进行训练，VLM 获得了理解抽象或开放式指令的灵活性，并能够泛化到前所未有的多模态场景。LLaVA [21] 率先采用了这一范式，利用 GPT-4 将原始图文对转换为 15 万个对话样本，显著增强了其处理开放式查询的能力。InternVL2 [22] 则通过引入 LLM 引导的自动过滤流程取得了进展。它有助于清理噪声注释，减少模型的异常行为。总而言之，这些进展展示了视觉指令调优如何使大型 VLM 做好遵循高级命令的准备。这种能力为将大型 VLM 扩展至用于机器人操作的 VLA 奠定了基础，这需要在现实世界中对模糊指令进行跨视觉、语言和动作的鲁棒推理。

视觉语言模型 (VLM) 的最新进展进一步强调了其可扩展性，从而增强了多模态感知和推理能力。LLaVA-OneVision [77] 等模型将图像和视频处理统一在一个框架内。Qwen-2VL [78] 引入了对任意输入大小的动态分辨率支持，而 Vision-R1 [79] 则利用强化学习来改进思路链推理。总而言之，这些创新赋予了 VLM 更强大的现实世界感知和决策能力，为 VLM 进一步整合动作生成奠定了基础。

基于这些突破，大型 VLM 的下一个发展方向是从被动观察过渡到与物理环境的主动交互。虽然 DriveVLM [65] 和 CogAgent [66] 等模型已经展示了 VLM 处理现实世界输入和生成高级规划的能力，但它们仍然无法以具身的、物理的方式直接与现实世界互动。

2.2 机器人操作的进展

早期的语言调节机器人系统通常将独立的视觉编码器与语言模块或规划器配对，而不是使用统一的多模态模型。例如，CLIPort [80] 使用 CLIP 视觉-语言编码器 [81] 进行语义基础构建，并使用 Transporter [82] 网络进行空间推理。同样，RT-1 [83] 使用基于 CNN 的视觉编码器 [84]，该编码器由源自通用句子编码器的独立语言嵌入进行调节，将摄像头图像和文本指令映射到运动动作。通过将独立的预训练视觉和语言编码器与多任务机器人操作数据相结合，这些早期方法实现了一定程度的多任务泛化。然而，它们的模块化设计难以推广到未知概念，并且在处理复杂或模糊指令时会遇到困难。

相比之下，最新的 VLA 模型将视觉、语言和动作控制紧密集成在一个统一的系统中。 RT-2 [27] 以预训练的大型 VLM 主干网络（例如 PaLM-E [85] 或 PaLI-X [86]）为起点，并在互联网规模的视觉语言任务和真实机器人轨迹上对其进行联合训练。至关重要的是，机器人动作被转换为文本标记，并与常规语言输出包含在同一个训练语料库中。这种简单的方案使模型能够将机器人控制“吸收”为另一项语言任务，从而形成一个统一的 VLA 模型。经验表明，RT-2 表现出显著增强的语义理解能力。与 RT-1 [83] 相比，RT-2 [27] 能够泛化到新物体和未知指令（例如，将物体放置在特定数字上），并能够执行基本推理（选择最小或最近的物体）。换句话说，大规模视觉语言预训练为机器人策略注入了现实世界的知识和更强大的语言基础，而这是早期模型所缺乏的。

继 RT-2 展示了互联网规模的 VLM 预训练如何提升机器人控制能力后，更多 VLA 模型应运而生。例如，π0 [29] 在预训练的 VLM 之上采用了流匹配架构，并在各种灵巧机器人数据集上进行训练，从而实现了强大的零样本泛化能力，并可通过微调轻松适应新任务。此外，OpenVLA [26] 已发布，这是一个规模达 70 亿的开源 VLA 模型，已在约 97 万个真实机器人演示上进行预训练。它实现了卓越的通用操作性能，并支持通过 LoRA [87] 等技术在消费级硬件上进行高效微调。这些进步共同彰显了 VLA 驱动的机器人操作向更高通用性和开放可及性迈进的趋势。

2.3 相关综述

人们对具身人工智能 (embodied AI) 的 VLA 模型日益增长的兴趣，催生了多项综述 [1]–[6], [8], [88], [89]，但大多数综述侧重于更广泛的架构范式或多样化的应用领域，而缺乏针对机器人操作的大型 VLM VLA 系统的系统性探索。例如，Ma 等人 [2] 对 VLA 架构进行了全面的综述，回顾了用于集成视觉、语言和动作模态的模块化、端到端和混合方法。然而，该综述缺乏对近期利用预训练 VLM 作为基础组件的趋势的重点分析。这些模型凭借其强大的跨模态对齐和零样本泛化能力，已成为主流范式。Sapkota 等人 [8] 对更广泛的 VLA 应用进行了广泛的研究，例如自动驾驶、增强现实导航和精准农业。然而，其广度削弱了应对机器人特定挑战（例如实时驱动约束、传感器噪声鲁棒性和长期决策）所需的深度。Wang 等人 [88] 的一项综述对将纯文本 LLM 集成到机器人任务规划中进行了初步概述。该综述强调了 LLM 根据自然语言指令生成精确行动计划的能力。然而，该研究主要侧重于高级推理，并未探讨将 VLM 应用于机器人视觉感知和动作决策的挑战，而最近的 VLA 模型旨在填补这一空白。

为了弥补这一新兴领域缺乏全面深入综述的不足，我们对近期的 VLA 研究进行了结构化概述。基于第一节中描述的 VLA 模型定义，我们追踪了 VLA 在机器人领域的发展，分析了关键架构和学习范式，并回顾了相关的数据集和基准。我们的目标是提供一个全面的视角，以支持更深入的理解并推动未来的发展。

3 单体模型

单体 VLA 模型主要以两种方式实现：单系统和双系统架构，如图 3 左图所示。在单系统设计中（见 3.1 节），视觉感知、语言指令和机器人状态被联合输入到一个统一的模型中，该模型处理所有模态，并通过自回归或并行解码来解码可执行动作。相比之下，双系统架构（见 3.2 节）将功能划分为两个协作模块：系统 2（VLM 主干）执行速度较慢但更广义的反射推理，而系统 1（动作专家）则专注于快速处理以支持反应行为。单系统 VLA 架构简洁、开发流程精简，并避免了复杂的模块间通信；而双系统 VLA 则利用分工将反应速度与精准度完美结合。

3.1 单系统模型

单系统 VLA 模型体现了单体设计理念。它们旨在通过统一模型将大型 VLM 的语义知识迁移到机器人操作任务中。本节回顾了单系统模型的范式，并探讨了两个主要研究方向：增强模型解决复杂任务的能力，以及提高推理效率以实现实际部署。所有模型均总结于表 1 中。

表 1：单系统 VLA 模型。在 LLM / VLM 列中，省略 V-Encoder 表示 VLM；否则表示 LLM。在 Learning 列中，“AD”表示自回归解码，“PD”表示并行解码。“SFT”表示微调，不同于动作预测模仿学习，在动作预测模仿学习中，字幕制作、VQA、推理等任务都属于 SFT。括号中的“A”和“B”表示 Action head 或 Backbone 使用的学习方法。

3.1.1 经典范式：自回归解码

自回归解码范式直接借鉴了 LLM 的序列生成能力。通过将机器人的连续动作空间离散化为 token 序列，该模型可以顺序预测动作 token。如图 4 中的“自回归解码”所示，VLM 接收视觉观察、自然语言指令以及可选的机器人状态作为输入。然后，它自回归生成动作 token，这些 token 可以通过下游的去 token 化器转换为可执行动作。

RT 系列 [27]、[83]、[90] 和 OpenVLA [26] 是此范式的典型示例。RT-1 [83] 将 Transformer [91] 引入 VLA 领域，并将连续的机器人动作编码为离散的 token。RT-2 [27] 继承了这种基于 token 的动作表示，并将其应用于更大规模的 VLM。它对互联网规模的视觉语言和机器人轨迹数据进行协同微调，成功地将丰富的视觉语义知识从网络数据迁移到机器人操作任务。RT-2-X [90] 通过在 Open X-Embodiment (OXE) 数据集 [90] 上应用协同微调，进一步提升了跨机器人技能迁移。更进一步，OpenVLA [26] 用 SigLIP [92] 和 DINOv2 [93] 的组合取代了 RT 系列中的大参数视觉编码器。通过在大规模真实机器人操作数据上进行微调，该模型以更少的模型参数实现了卓越的性能。其完全开源的特性使其成为后续研究中广泛使用的基准模型。

3.1.2 模型性能增强

VLA 模型的发展伴随着模型性能的提升。为了使模型能力更好地与人类智能相结合，性能增强工作主要集中在三个方向：

增强感知模态。如图 4 所示，这些模型通过扩展输入模态来改进真实世界的信息获取。(1) 3D 感知。由于现实世界是三维且连续的，基于 VLM 的大型 VLA 模型无法仅依靠 2D 图像来完全理解现实世界。因此，Leo Agent [94] 直接从模拟环境中获取以对象为中心的 3D 点云，并使用 PointNet++ [118] 对其进行编码；然后将编码后的特征输入空间变换器 [119] 以生成“空间”信息。在现实世界中，SpatialVLA [100] 提出了 Ego3D 位置编码，该编码从二维图像估计深度，并将其反向投影到以自我为中心的三维坐标系中，将逐像素的三维位置与二维特征融合，从而生成以自我为中心的“空间”信息。类似地，BridgeVLA [35] 从 RGB-D 图像重建三维点云，并生成三个正交投影视图，使“空间”信息能够以与大型 VLM 兼容的二维格式表示。

1. 四维感知。为了更好地捕捉机器人操作中的空间结构和时间动态，需要四维感知来实现更全面的场景理解。TraceVLA [97] 将采样的运动点轨迹叠加到当前图像上，从而创建图 4 中的“轨迹”信息。结合原始观测数据，它为大型 VLM 提供了时空理解能力。此外，4D-VLA [108] 将 3D 坐标集成到视觉特征中，以解决空间坐标不一致的问题，并利用历史关键帧的存储库采样来缓解时间状态模糊性，从而改进时空推理。ST-VLA [110] 引入了“空间轨迹”方法，将历史时间“轨迹”信息与场景的空间深度信息相结合，在某些机器人操作任务中超越了之前的 SpatialVLA [100] 和 TraceVLA [97] 模型。

（3）触觉和听觉感知。VTLA [103] 通过视觉编码器对触觉信息进行编码，使其与视觉序列对齐，然后将其与视觉和文本标记一起输入 LLM，从而构建“视觉触觉-语言-动作”模型。VLAS [102] 使用“耳语编码器”提取语音特征，并通过 MLP 将其投影到 LLM 嵌入空间，从而引入图 4 中的“音频”信息。VTLA 和 VLAS 都构建了特定于任务的数据集以进行微调。此外，FuSe [98] 在微调过程中学习将“触觉”和“音频”模态与语言概念对齐，从而在有限的模态特定标注下实现多模态融合。OE-VLA [104] 利用多样化的多模态数据和两阶段课程学习策略，将指令从文本扩展到交错图像、视频和光学指令。

图 4：单体单系统模型中代表性范式的比较。3.1.1 节展示了经典自回归解码的模型示意图。箭头表示信息流。3.1.2 节介绍了通过引入其他模态、利用思路链推理和增强泛化能力来增强模型能力的方法。3.1.3 节讨论了一些通过架构优化、参数设计和解码策略来提高推理效率的方法。

增强推理能力。为了将 VLA 模型从简单的被动控制转向更高级的审议决策，增强推理能力至关重要。如图 4 中“增强推理能力”所示，LLM 生成一个思路链式“推理”过程，并将其作为上下文信息用于生成最终的“动作”。ECoT [120] 生成一个推理链，该推理链将高级任务规划与基于视觉的特征依次结合，然后输出最终动作。CoT-VLA [33] 通过预测代表像素空间中计划状态的子目标观察值，进一步引入了视觉思路链式推理。与以推理为导向的视角相比，LoHoVLA [106] 采用“分层闭环控制”机制来处理规划错误、动作失败和外部干扰，从而处理长周期任务。ReFineVLA [105] 采用“选择性迁移微调”策略，并设定双重学习目标，仅对上层进行微调，从而使模型能够增强多模态理解。这些研究共同强调了推理在使 VLA 模型实现更可靠、更泛化的动作预测方面的核心作用。

增强泛化能力。模型的泛化能力是指其在不同平台和场景下执行不同任务的能力。如图 4 中的“增强泛化能力”所示，UniAct [99] 定义了一个“通用动作码本”，将异构的机器人动作抽象为统一的表示。这种统一的编码消除了动作空间的异质性，并实现了跨具体化学习和推理，从而增强了泛化能力。ReVLA [96] 采用可逆训练策略，逐渐将视觉编码器恢复到其原始的预训练状态，从而减轻了微调过程中的灾难性遗忘，并改善了分布外的视觉泛化能力。

另一种增强泛化的方法是生成更鲁棒、更可靠的动作，如图 4 所示。HybridVLA [34] 将扩散和自回归解码集成到一个统一的模型中。它进一步采用了“协作动作集成”机制，自适应地融合这些动作，为不同任务选择最合适的生成策略，从而提高控制鲁棒性。类似地，VOTE [109] 引入了一种自适应的“集成投票”策略，将过去的动作预测按与当前动作的相似性分组。它对多数集进行平均，以生成更鲁棒的动作，平衡响应性和稳定性。

使模型能够更好地理解物理动态可以促进生成更可靠的动作。WorldVLA [38] 将动作模型和世界模型集成在一个统一的自回归框架中，用于捕捉物理动态。同样，UnifiedVLA [107] 引入了一个世界模型作为训练后任务，使其能够从大规模未标记视频中学习物理世界的因果动态。UP-VLA [101] 提出了一个训练框架，通过在预训练期间预测下一帧图像来隐式学习世界的物理动态。因此，这些模型可以产生更可靠的动作，更好地遵循物理定律。

【此处省略模型优化和推理加速方法的部分】

3.2 双系统模型

为了同时满足深度推理和实时动作生成的需求，单体双系统 VLA 架构应运而生。这些模型采用两个不同的系统：一个更大、速度较慢的系统 2 用于高级推理和规划，另一个更小、速度较快的系统 1 用于低级动作生成。这种分离实现了更快的响应时间和更顺畅的实际部署。它为机器人操控提供了可靠的战略规划和高速控制。分离的低级控制模块也更适合特定任务的优化。这种方法与单系统模型的关键区别在于引入了动作专家来明确地解耦动作生成。与分层模型相比，它的主要区别在于它不产生可解释的中间输出。表 2 总结了具有代表性的双系统方法。图 5 展示了两种具有代表性的架构：基于级联的方法和基于并行的方法。

表 2：双系统 VLA 模型。 “系统2骨干”列列出了双系统方法中用作系统2组件的VLM骨干。“系统1学习”列列出了行动专家作为系统1使用的学习方法。“Diff”表示基于扩散的学习，“FM”表示流匹配，“MSE”表示均方误差，“BCE”表示二元交叉熵，“AR”表示自回归学习。

3.2.1 基于级联的方法

如图5上部所示，基于级联的双系统VLA模型以串行方式将高级语义推理与低级实时控制分离。高级系统（系统2）通常使用VLM来处理多模态输入、进行语义基础构建并生成行动计划。这些计划被编码为潜在的认知表征，而不是直接执行。然后，低级系统（系统 1）以更高的频率将这些表征解码为可执行的机器人动作，从而显著提高运行时效率并促进实际部署。图 5 展示了两种架构，它们的区别在于动作专家的设计方式。

为了将模型的“认知”和“动作”能力解耦，许多方法选择一个单独的模型作为动作专家。如图 5 左上角的“单独动作专家”部分所示，VLM 主干将特征传输给动作专家，以传达来自视觉、文本和机器人特定状态输入的信息。例如，CogACT [127] 引入了扩散变换器 (DiT) [142] 作为动作模型。特别地，它还提出了一种自适应动作集成算法，以实现更平滑、更高效的运动轨迹。GR00T N1 [32] 也采用了双系统架构，其中引入了 DiT 作为低级动作模型。这两个系统紧密耦合，并进行端到端联合训练，从而显著加快策略步骤。一些模型采用类似的架构，但不使用 DiT。例如，DP-VLA [124] 使用了行为克隆变换器 [143]。

许多模型在此基础上进行了改进。HiRT 模型 [128] 使用 VLM 以较低频率运行。此过程提取特征以进行对场景的长期理解。然后，以较高频率部署轻量级的视觉动作策略。这种设计可以提高机器人操作效率，并使系统在部署期间能够跟上现实环境的变化。TriVLA [132] 是一个三模块系统，包含一个用于预测未来帧的视频生成模型、一个用于解释指令的 VLM 和一个扩散动作专家。GF-VLA [133] 从人类演示中提取基于信息的手-物体和物体-物体场景图，然后将它们与 LLM 融合，生成可解释的行为树和用于双臂控制的低级笛卡尔动作。RationalVLA [134] 通过可学习的潜在接口将 VLM 与扩散策略耦合。VLM 发出一个令牌作为控制器生成动作的条件，或者发出一个令牌来拒绝不可行的命令。VQVLA [135] 采用卷积残差 VQ-VAE [144]（已用动作序列预训练）来取代 OpenVLA [26] 的分箱方法。该模型显示出随着模拟数据的增加而线性性能提升，并且模拟与实际之间的差距更小。一些模型将动作专家集成到 VLA 主干网络中，Fast-in-Slow [40] 就是一个典型的例子。如图 5 右上角的“统一动作专家”部分所示，动作专家利用了 VLM 主干网络的最终 Transformer 模块。它们以不同的频率运行，从而在单个预训练模型中实现两个系统之间的无缝协调。

3.2.2 基于并行的方法

如图 5 下半部分所示，基于并行的双系统 VLA 设计了一个与 VLM 主干并行运行的动作专家。这两个组件在推理过程中交互交换信息。根据动作专家的选择和交互机制，图 5 中定义了两个类别。

如图 5 左下角的“共享注意力架构”部分所示，该架构借鉴了混合专家 (MoE) 框架 [145] 的灵感。它们利用 VLM 主干和动作专家共享的自注意力层中的标记交互，将高级推理与低级执行分离。来自视觉和文本输入的信息与共享自注意力层中的噪声和机器人特定输入交互。这种设计有助于针对特定任务进行优化。例如，这种设计允许在动作专家中融入真正的 MoE 或其他专门设计的架构。π0 [29] 就是一个典型的例子。其骨干网络权重由预训练的 VLM 初始化。为了处理特定于机器人的输入和动作生成，引入了第二组独立权重，即基于流匹配的动作专家，并从头开始训练。ForceVLA [137] 使用 π0 [29] 作为基础模型，并使用具有 MoE 的 FVLMoE 模块将力模态引入 VLA。OneTwoVLA [138] 基于 π0 [29]，可以在两种模式之间切换：显式推理和基于最新推理生成动作。这种架构使两个系统更容易异步运行，并进一步提高了效率。

许多创新都建立在此基础上。π0.5 [30] 在 π0 [29] 的基础上引入了一个额外的步骤。 VLA 模块整合视觉信息，将高级提示转换为更细粒度的子任务预测，然后由 π0 [29] 的双系统架构进行处理。π0.5KI [37] 以 π0.5 [30] 为基础。在训练过程中，它可以防止来自动作专家的梯度流入 VLM 主干网络，以保持 VLM 的知识优势。π0-FAST [123] 引入了一种 DCT 驱动的动作标记方法，有助于 VLA 模型的高效自回归训练。villa-X [141] 将潜在动作置于机器人状态中，并通过联合扩散对潜在​​动作和机器人动作进行联合建模，实现结构化的视觉-动作整合。Tactile-VLA [139] 将触觉感知与混合位置-力控制相结合，融入 VLA 模型中。如图 5 右下角的“交叉注意力架构”部分所示，该架构将视觉、文本和状态输入馈送到 VLM，其中自注意力层生成键值对，并将键值对传递给动作专家的交叉注意力层。例如，SmolVLA [31] 就采用了这种架构。它利用轻量级且冻结的 VLM 主干网络，并仅训练下游的流匹配 Transformer 作为动作专家，从而提高了效率。另一个例子是 GR-3 [140]，它使用流匹配将视觉语言理解与机器人轨迹学习统一起来。

图 5：单体双系统模型中代表性范式的比较。第 3.2.1 节介绍了基于级联的方法，其中 VLM 主干网络与动作专家级联，以便将其输出特征一次性转发给动作专家。第 3.2.2 节介绍了基于级联的方法，其中 VLM 主干网络与动作专家级联，以便将其输出特征一次性转发给动作专家。 3.2.2 介绍了基于并行的方法，其中 VLM 主干网络和动作专家并行运行并以各种方式交互。

4 层次化模型

层次化建模是大型基于 VLM 的 VLA 模型的基础范式，尤其适用于需要长周期推理、空间抽象或动作分解的场景。这些模型通常由一个高级规划器和一个低级策略组成，如图 3 右图所示。规划器接收指令和观察结果，并将其转换为可解释的中间表示。然后，策略接受这些表示，并生成机器人可直接执行的动作序列或代码。表 3 提供了层次化模型的概述。

表 3：层次化 VLA 模型。 “类型”列表示规划器的输出类型，其中“K”表示关键点，“S”表示子任务，“P”表示程序。“学习”列指定模型采用的学习方法，其中“SFT”表示监督微调，“RL”表示强化学习，“IM”表示模仿学习，“API”是一种特殊情况，指调用现有模型。

重要的是，分层模型中的规划器和策略可以独立运行，而无需任何模块严格依赖于另一个模块。这种模块化设计实现了灵活的组合：许多研究仅关注规划器组件的设计，并利用现有的现成策略进行执行。因此，我们将分层模型分为两类：仅规划器（第 4.1 节）和规划器+策略（第 4.2 节）。与同样涉及多个模块的双系统 VLA 模型不同，分层模型中的中间表示可被人类明确解释。根据其性质，每个类别可进一步细分为基于子任务、关键点和程序的方法。4.3 节将对整体式模型和分层模型进行比较分析。

4.1 仅规划器

4.1.1 基于程序的方法

在这种方法中，规划器生成用于机器人操作的中间程序，这些程序分为两类：机器人可执行程序和辅助程序。机器人可执行程序基于机器人库构建，可直接执行以控制机器人。例如，Chainof-Modality [149] 采用多模态提示策略，其中 VLM 跨不同模态进行多轮对话，最终生成一个机器人可执行的 Python 程序来重现该任务。类似地，Instruct2Act [157] 生成 Python 代码，用于调用 API 来控制机器人动作。相比之下，辅助程序支持策略理解任务，但无法直接执行。ROVI [150] 就是这类程序的典型代表，它生成辅助程序来描述潜在动作，并通过平移和旋转成本解决实际执行问题。同样，ReLEP [151] 使用带有记忆库的大型视觉语言模型 (VLM)，将技能库中的任务分解为基本技能。它以辅助程序的形式生成规划，从而实现强大的长视界性能。

4.1.2 基于关键点的方法

基于关键点的规划器使用大型视觉语言模型 (VLM) 预测观测中的显著点，这些显著点通常对应于机械手应该到达的交互区域（例如，抽屉的把手）。一些方法通过预测路径点来处理机器人操控。MoManipVLA [146] 通过大型视觉语言模型 (VLA) 模型在每一步生成一个关键路径点，随后通过双层轨迹优化框架将其细化为可执行动作。其他方法则强调可供性驱动的关键点。 RoboPoint [147] 解释自然语言指令，生成指定精确操作目标的视觉关键点。此外，ManipLVM-R1 [46] 使用组相对策略优化 (GRPO) [171] 训练视觉语言模型 (VLM)，以预测抓取的可供性区域和目标物体的轨迹，从而生成更具泛化的规划器。类似地，RoboBrain [153] 通过将用于高级规划的 LLaVA [21] 与用于识别可交互区域的 A-LoRA 模块以及用于预测轨迹航点的 T-LoRA 模块相结合，将任务规划、可供性感知和轨迹估计集成在一起。这表明这些模型涵盖了可供性和航点预测，形成了一种混合表示。此外，一些研究探索了关键点的不同形式。例如，RoVI [150] 采用基于草图的界面，使用 YOLOv8 从手绘注释中提取起点、航点和终点，然后将其用作轨迹约束。

4.1.3 基于子任务的方法

在这种方法中，规划器通常是一个大型的视觉语言模型 (VLM)，它接收高级隐式指令（例如，清理桌子）和观察结果，并将其分解为逐步的文本命令。由于这些模型生成的是可解释的中间指令而非可执行操作，因此在实际部署中仍然需要低级控制策略。早期的研究成果，例如 PaLME [85]（一种使用机器人操作数据训练的 LLaVA 风格的视觉语言模型），证明了将通用的视觉问答 (VQA) 功能与机器人命令生成统一起来的可行性。在此基础上，Embodied-Reasoner [47] 引入了“观察-思考-行动”轨迹，以支持逐步规划过程中的空间分析、反思和验证。强化规划 [148] 保留了这种子任务分解的设置，但通过一个两阶段的流程来提升泛化能力：先进行 SFT，然后进行基于 GRPO 的强化微调。同时，一些系统将感知与推理分离：Embodied-R [172] 将用于感知的大型 VLM 与用于推理的小型 LM 相结合，用于在视频上进行具身空间推理，其推理轨迹可用于逐步生成用于操作的子任务提案。相比之下，ViLA [152] 利用 GPT-4V 作为外部规划器，使其根据文本和视觉输入生成候选任务计划，然后仅执行第一个命令，丢弃其余命令。

4.2 规划器+策略

4.2.1 基于关键点的方法

这类模型通常使用大型 VLM 将子目标作为空间基元（例如图像上的离散关键点或二维路径）进行落地。然后，低级策略使用这些基元来预测连续轨迹和控制。例如，HAMSTER [48] 首先根据指令和观察结果预测轨迹关键点。然后，它将这些关键点连接成一条带有渐变色的有序路径，并叠加此路径以指导策略执行。此外，ReKep [50] 利用 DINOv2 [93] 和 SAM [173] 生成关键点提案，然后使用 GPT-4o [174] 将这些关键点转换为成本函数。随后，优化器将这些函数求解为路径点和动作。此外，A0 [49] 采用了可供性感知的层次结构：规划器将接触点和接触后运动预测为与具体化无关的可供性表征，然后由动作策略将其转换为控制。

4.2.2 基于子任务的方法

基于子任务的层次模型通过指令连接规划器和策略。规划器的作用与基于子任务的仅规划器层次模型相同，并附加了一个低级策略来生成动作序列。

一个代表性的例子是 HiRobot [154]，其规划器接受开放式用户指令，然后将其分解为策略所需的原子命令。类似地，DexVLA [156] 也具有可变长度机器人 (VLM) 规划器和基于扩散的动作策略。该策略以 VLM 规划器给出的子任务标记为条件，在复杂且长视域的机器人操作任务中表现出色。PointVLA [159] 基于 DexVLA 构建，通过在策略中加入点云编码器和注入器来增强空间感知，使模型能够在几何复杂的场景中遵循规划器的指令。RoBridge [161] 提示规划器生成原始动作的文本指令，并为要执行的策略形成一个不变的可操作表示。此外，SkillDiffuser [165] 通过预测一组技能的高级模型将复杂任务分解为子任务，而低级扩散策略则实现具体的动作。类似地，RoboMatrix [158] 将执行组织成一个三层结构：模块化调度层生成子任务序列，技能层编码并选择可重用行为，硬件层实现机器人控制。HiBerNAC [167] 提出了一个异步的分层框架。在该框架中，多智能体神经结构首先将高级指令分解为结构化子任务。然后，异步流水线管理这些子任务，并协调反应式 VLA 执行最终的低级操作。此外，MALMM [169] 包含一个规划器、一个监督器和一个编码器。规划器为编码器生成子任务，监督器协调模块之间的转换。编码器充当策略，并将计划转换为可执行的机器人代码，包括动作和位置。

4.3 比较：单片式 vs. 分层式

在基于 VLM 的大型机器人操作 VLA 模型中，单片式和分层式架构之间的区别主要在于它们如何将视觉输入和语言指令映射到动作（通过统一方法还是模块化方法）。单片式模型强调单一集成的流水线，该流水线联合优化感知、推理和控制，将高级多模态语义直接转化为低级动作。这种设计使机器人操作能够实现整体且紧密耦合的学习。相反，分层式架构采用多阶段设计，明确地将高级规划与低级策略执行分离，从而提高了模块化和可解释性。这种系统分解允许组件独立设计、训练或替换，从而增强了灵活性，并简化了领域知识的集成或适应新的机器人操作任务。

另一个核心区别在于中间处理的固有特性。单体模型虽然可能嵌入复杂的内部推理，但其实现机制位于通常难以被外部检查的潜在空间中。这种隐式策略使它们能够充分利用大型模型的表征能力，从而有可能发现并非由人类明确设计的高效任务分解。另一方面，分层系统则明确地致力于生成清晰的、人类可理解的中间输出。这使得它们在需要可解释性、详细的任务监控或与传统机器人流程兼容的机器人操作场景中尤为有利，在这些场景中，高级计划可以独立验证或修改。

尽管存在这些差异，但这两种方法都为机器人操作领域不断发展的 VLA 领域贡献了独特的优势。单体模型突出了统一学习和最小化手动分解的强大功能，为跨不同任务的泛化提供了一条简化的途径。分层框架通过明确地对认知和控制进行分层，提供了更高的透明度和增强的模块灵活性，这对于复杂的多阶段任务或安全关键型部署至关重要。总的来说，这些范式展现了连接视觉语言理解和具体行动的互补策略，每个策略都为下一代智能机器人操作系统的设计提供了宝贵的见解。

5 其他先进领域

除了核心 VLA 架构之外，我们还重点介绍了四个在鲁棒性、效率和长期规划方面具有突破性进展的方向：基于强化学习的优化、无需训练的改进、从人类视频中学习以及基于世界模型的 VLA。表 4 总结了代表性方法。

表 4：代表性 VLA 方法分为四类：基于强化学习、无需训练、从真人视频学习和基于世界模型的方法。该分类遵循第五节中的方法论分析。

5.1 基于强化学习的方法

强化学习 (RL) 在提升 VLA 的泛化能力 [175] 和任务完成率 [176] 方面发挥着重要作用。它大致可以分为在线（通过实时交互优化策略）和离线（从预先收集的轨迹中学习）。大多数 VLA 方法都建立在传统的 RL 算法之上。我们在表 5 中列出了几种方法。

表 5：VLA 的代表性 RL 方法。✓ 代表在线，✗ 代表离线。“S”代表稀疏奖励，“D”代表密集奖励，“RM”表示预先训练的奖励模型，“GPT”表示由 GPT 提供的奖励，“TC”表示基于任务完成情况的奖励函数。

与通常在几轮内完成生成的 LLM 或 VLM [79]、[177]、[178] 中的 RL 不同，VLA 模型通常涉及一条轨迹中有数百步的长周期任务，这使得基于规则的奖励函数容易出现稀疏性和不稳定性 [175]、[179]。为了解决这个问题，许多研究结合了学习到的密集奖励信号。VLA-RL [179] 训练机器人过程奖励模型 (RPRM) 来预测动作序列的成功可能性。ReWiND [39] 将奖励建模为朝着最终目标状态的进展，为视觉上更接近任务完成的状态分配更高的奖励。除了训练奖励模型之外，Grape [175] 和 TGRPO [180] 还利用强大的 VLM [174]、[181]，通过提示来生成基于反馈的奖励信号。

除了奖励稀疏性之外，由于模拟速度慢，在线学习的样本效率低下。为了解决这个问题，一些方法采用了离线-在线混合训练。 ReWiND 将离线隐式 Q 学习 [182] 与在线软性 Actor-Critic [183]​​ 相结合，加速学习并增强实际部署的安全性。HIL-SERL [184] 基于 RLPD，并在训练过程中引入了人机交互。ConRFT [185] 采用两阶段训练方案：离线阶段使用 Cal-ConRFT（结合行为克隆和 Q 学习）初始化策略；然后是在线 HIL-ConRFT，通过人机交互平衡监督学习和 RL 损失。

此外，一些方法利用 RL 作为数据引擎来增强通用机器人模型。RLDG [186] 就是一个很好的例子，它通过 HIL-SERL [184] 训练专家策略，使其达到 100% 的任务成功率，然后通过 sft 将其提炼为基础模型。类似地，iRe-VLA [187] 在扩展的数据集上交替进行在线强化学习（用于收集新的成功轨迹）和稳定训练，通过迭代模仿和探索逐步改进模型。

5.2 无需训练的方法

无需训练的方法通常利用模块化和可扩展的设计来改进现有的 VLA 架构，而无需进行训练。这使得快速原型设计、简化研究和有针对性的增强成为可能，同时保留了模型的原始功能并避免了额外的成本。

为了提高基于 VLM 的大型 VLA 模型的效率，人们提出了一系列无需训练的方法，无需重新训练或更改架构。一个引人注目的例子是 FlashVLA [115]，它采用了一种触发机制，当动作和视觉线索保持稳定时，它会跳过完整的解码，从而选择性地重用或修剪视觉 token。类似地，EfficientVLA [188] 会修剪冗余的语言层，过滤与任务相关的视觉标记，并缓存中间特征；而 VLA-Cache [189] 会通过任务相关性过滤和层自适应重用，重用缓存的静态标记键值表示 [91]。此外，SP-VLA [190] 将空间语义标记修剪与动作感知调度器相结合，该调度器将直观的步骤路由到轻量级生成器，将复杂的步骤路由到完整的 VLA。同样，PD-VLA [113] 将动作分块下的自回归解码重新表述为并行定点迭代，从而实现同时进行标记预测。FAST [123] 通过离散余弦变换和字节对编码压缩动作序列，以减少高频控制中的冗余，使训练速度提高 5 倍。此外，RTC [41] 通过监控任务进度和调整控制频率来优化控制时间调度，减少推理过程中不必要的计算。

5.3 从人类视频中学习

VLA 模型的最新进展引入了几种超越传统策略学习策略的全新学习范式。其中，一个值得注意的趋势正在显现：利用人类视频数据指导机器人策略学习，利用人与物体以及机器人与物体交互中的结构相似性来对齐视觉和时间线索。

该方法旨在通过从丰富的人类视频数据中迁移与任务相关的知识来缩小体现差距。它使用人类视频来预训练感知或调整策略。例如，人机语义对齐 [42] 使用配对的人机视频跨领域对齐视觉编码器，而 UniVLA [192] 则从未标记的人类和机器人视频中学习以任务为中心的潜在动作，以统一策略规划。类似地，LAPA [193] 利用 VQ-VAE [144] 量化的潜在动作，在大规模视频-语言对上进行预训练，从而实现了从无动作的人类视频的迁移。此外，VPDD [194] 将离散扩散应用于统一的视频标记，通过未来动态预测促进跨领域知识迁移。3D-VLA [195] 将人机交互视频与机器人演示相结合，以实现更丰富的 3D 推理；Humanoid-VLA [196] 利用从在线视频中恢复的姿态运动轨迹来增强运动多样性。总而言之，这些方法表明，人类视频数据可以显著丰富机器人策略学习，即使在机器人数据有限的情况下也能实现鲁棒性和多功能性。

5.4 基于世界模型的 VLA

世界模型以其能够学习环境动态的紧凑潜在表示的能力而著称，已成为实现预测推理和长期规划的有力工具。近年来，将世界模型集成到 VLA 系统中，作为一种通过明确建模环境动态来增强行动规划的有前景的策略，已获得广泛关注。

这些方法并非直接从当前观察中生成动作，而是模拟未来状态，使智能体能够预测其动作的后果并相应地改进决策。例如，WorldVLA [38] 引入了一个自回归动作世界模型，该模型在基于标记的统一架构中联合学习预测视觉结果并生成动作。世界模型和动作模型之间的这种相互强化，提升了视觉想象力和动作保真度，从而支持更稳健的长期规划。World4Omni [43] 采用大规模世界模型生成描绘中间任务状态的子目标图像。然后，这些生成的视觉线索被用于指导模块化低级策略，从而实现跨不同环境和机器人实例的零样本操作。3DVLA [195] 采用生成式世界模型来预测基于指令的未来目标图像和点云，从而有效地模拟场景动态。 RIGVid [197] 使用基于扩散的世界模型生成候选任务视频，然后由 VLM 筛选以确定可行性，FoundationPose [198] 则提取用于执行的抓取器姿态。V-JEPA 2-AC [199] 在基于互联网规模数据训练的大规模自监督视频编码器之上构建了一个基于潜在动作条件的世界模型。在推理过程中，它通过模拟基于目标图像条件的未来潜在状态来执行基于模型的规划，从而将密集的视频理解与零样本机器人操作连接起来。

6. VLA 模型的特点

6.1 多模态融合

共享嵌入空间。基于大型 VLM 的 VLA 模型将视觉观察和语言指令嵌入到一个共享的、语义对齐的潜在空间中。与将感知、指令解析和控制模块化的传统机器人操作模型不同，基于大型 VLM 的 VLA 模型依赖于预先训练的 VLM 主干网络来生成两种模态的联合嵌入。这种统一的空间有助于感知和指令之间紧密的语义连接，从而提高动作解码的保真度并减少模块间语义漂移 [8]。

多模态 Token 级集成。许多 VLA 模型的核心特征是将视觉、语言、本体感觉和动作等连续模态离散化为由单个转换器处理的 token 序列 [40]、[107]、[201]。这种设计实现了 token 级别的多模态集成和细粒度协调：像“抓住红杯子”这样的语言 token 会立即将注意力引向相关的视觉 token，而本体感受 token 和动作 token 则会引导执行。交错的时间结构能够自然地捕捉感知-动作周期中的跨模态依赖关系。与在预定义阶段融合模态的传统方法相比，这种动态集成减少了延迟和语义碎片化。

全面的模态兼容性。VLA 模型的另一个显著特点是其固有的无缝兼容各种感知模态的能力，例如深度（例如点云）、触觉输入和环境音频。这种兼容性源于从预训练的 VLM 继承的与模态无关的语义对齐，这使得新的感知 token 能够无缝集成，而无需改变核心架构或进行完整的模型重新训练。例如，PointVLA [159] 将点云数据集成到预训练的 VLA 架构中，以丰富空间感知和泛化能力，而无需重新训练主干网络。

8. 未来方向

数据集和基准测试。随着 VLA 模型的发展，对模拟数据集的依赖暴露出巨大的现实差距。合成场景缺乏真实环境的视觉复杂性。相比之下，收集真实世界数据成本高昂，限制了数据的多样性和规模。同时，现有的基准测试通常侧重于短期拾取放置任务，并报告简单的成功率。这些设置几乎无法洞察诸如长期规划、移动操控和多智能体协作等实际挑战。未来的数据集可能会将大规模真实世界数据采集与能够捕捉实际挑战的任务套件相结合。评估应包含更丰富的指标，例如子任务成功率、时间效率和抗干扰能力。此类数据集和基准测试将有助于 VLA 系统在更广泛的功能方面取得系统性进展。

记忆机制和长期规划。有效的真实世界操控通常需要进行长远规划并回忆过去的观察结果。然而，目前大多数 VLA 依赖于逐帧推理，导致其行为短视，并且利用历史上下文的能力有限。为了解决这个问题，一个有前景的方向是设计将前瞻性规划和记忆机制与情景感知相结合的架构。通过将高级决策建立在情境记忆的基础上，智能体可以超越被动反应，转向连贯的、目标驱动的动作序列，从而在复杂场景中实现更一致、更有效的操控。

3D 和 4D 感知……机器人操控本质上涉及在三维和时间动态环境中与物体的交互。然而，现有的 VLA 模型主要基于静态的 2D 视觉输入。这限制了智能体推理复杂空间和时间信息的能力，包括深度、可供性、物体运动和人类动作。4D 感知超越了静态的 2D 快照，需要一个能够理解 3D 场景如何随时间演变的模型。从 2D 感知到 4D 感知的进展需要整合深度或点云观测数据，并将多模态输入融合成统一的表征。它还涉及嵌入基于时间背景的感知，并动态地重新规划行动。这些4D感知能力将共同实现现实世界中更稳健、更具适应性的操控。

移动操控。现实世界中的任务通常需要同时执行移动和操控，从而催生了移动操控这一专业领域。这项任务需要将导航能力和交互技能与环境协同整合，从而引入对感知和控制的紧密耦合需求。未来的 VLA 模型并非将移动和抓取视为独立的阶段，而是可能受益于学习集成策略，这些策略能够自适应地优先考虑移动和手臂的协调，最终实现更稳健、更灵活的移动操控。

多智能体协作。许多现实世界中的协作任务，例如移动大物体或使用关节工具，都需要智能体协商意图、适应队友的动作，并对多步骤目标进行联合推理。然而，当沟通和角色分配变得至关重要时，单智能体 VLA 往往会失效，这凸显了交互感知表征的重要性。将新兴对话协议与共享世界模型相结合，可以使代理同步计划并灵活分配子任务。这种合作能力有望将机器人团队从松散的个体提升为能够完成复杂群体目标的紧密协作者。

开放世界中的终身学习。机器人自主性需要持续的技能习得，并且不能发生灾难性的遗忘。然而，大多数基于静态数据集训练的 VLA 模型难以处理不熟悉的对象和新的交互模式。它们也难以融入新的经验。这揭示了静态训练范式的局限性。未来的 VLA 系统可能会受益于在开放世界中通过探索和反馈逐步积累知识的机制。融入随时间增长的记忆结构，并使代理能够将经验自组织成可重用的抽象概念，也可能为长期能力奠定基础。

模型效率。尽管 VLA 模型在理解视觉语言指令和规划动作方面表现出色，但它们通常受制于过高的计算和内存成本。在资源受限的机器人平台上部署这些模型会引发延迟、内存占用和持续运行方面的担忧。关键挑战在于如何在模型容量与实时推理需求之间取得平衡，并保持多模态对齐以在压缩过程中保持准确性。有前景的方向包括任务感知的动态 token 剪枝、用于动作块之间无缝转换的异步推理以及硬件友好的量化方案。

9. 结论

总而言之，本综述首次对基于大型 VLM 构建的 VLA 模型进行了原则性综合。我们追溯了 VLM 的演变，分析了它们与机器人操作的集成，并提出了一个由两部分组成的分类法：单体式（单系统和双系统）和层级式（仅规划器和规划器+策略）。该分类法构建了现有模型并指导未来的 VLA 设计。我们还将数据集、基准测试、学习范式和先进方法组织成一个连贯的框架；识别定义性特征，例如开放世界泛化、多模态推理和动态环境中的指令落地；并总结模块化规划与控制、更大规模的多模态语料库以及与强化学习和世界模型更紧密联系的趋势。展望未来，优先事项包括跨具体化适应、可扩展的真实世界部署、高级推理与低级执行之间更紧密的耦合，以及在互联网规模上使用人类演示。我们希望本综述能够为统一感知、语言和行动的具身人工智能奠定基础。