计算机应用
*Journal of Computer Applications*
ISSN 1001-9081,CN 51-1307/TP

# 《计算机应用》网络首发论文

题目：面向具身智能的视觉-语言-动作模型动作表征和生成策略综述
作者：张文涛，孙奥兰，瞿晓阳，张旭龙，王健宗
收稿日期：2025-08-07
网络首发日期：2025-09-30
引用格式：张文涛，孙奥兰，瞿晓阳，张旭龙，王健宗．面向具身智能的视觉-语言-动作模型动作表征和生成策略综述[J/OL]．计算机应用．
https://link.cnki.net/urlid/51.1307.TP.20250929.2133.016

Journal of Computer Applications　ISSN 1001-9081
计算机应用　CODEN JYIIDU　http://www.joca.cn

DOI:10.11772/j.issn.1001-9081.2025070900

# 面向具身智能的视觉-语言-动作模型动作表征和生成策略综述

张文涛[1,2]，孙奥兰[1]，瞿晓阳[1]，张旭龙[1]，王健宗[1*]

(1.平安科技(深圳)有限公司，广东 深圳 518046;
2.清华大学 深圳国际研究生院，广东 深圳 518071)
(通信作者电子邮箱 jzwang@188.com)

**摘 要:** 视觉－语言－动作模型是实现具身智能的核心路径，其核心在于将多模态感知理解无缝转化为物理世界的具体行动。然而，动作表征与生成策略作为连接“感知”与“执行”的枢纽环节，面临着高维连续空间、动作多样性与机器人实时控制需求间的复杂挑战。该综述系统性地梳理和总结了 VLA 模型中动作表征和生成策略的演进脉络、核心技术与未来方向，内容详细剖析了离散和连续两种动作表征方式，以及自回归、非自回归和混合生成策略，并深入探讨了它们在动作精度、生成多样性与推理效率之间的内在权衡。此外，综述还涵盖了面向实时控制的新兴高效策略，如混合生成架构等。最后，通过比较分析对现有技术图景进行了总结，并展望了未来在与世界模型结合、跨机器人形态通用表征等方向上的前沿挑战与研究机遇，旨在为构建更通用、更高效的具身智能体提供参考。

**关键词:** 具身智能；视觉－语言－动作模型；动作表征；生成策略

**中图分类号:**TP18　　文献标志码：**A**

# Survey on action representation and generation strategies in Vision-Language-Action models for embodied intelligence

ZHANG Wentao[1,2], SUN Aolan[1], QU Xiaoyang[1], ZHANG Xulong[1], WANG Jianzong[1*]

(1. Ping An Technology (Shenzhen) Company Limited, Shenzhen Guangdong 518046, China;
2. Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen Guangdong 518071, China)

**Abstract:** Vision–Language–Action (VLA) models constitute a critical pathway toward embodied intelligence, with their core function being the seamless transformation of multimodal perception and understanding into concrete actions in the physical world. Action representation and generation strategies, serving as the pivotal bridge between perception and execution, face significant challenges stemming from high-dimensional continuous spaces, the diversity of action modalities, and the stringent demands of real-time robotic control. This survey provides a systematic review of the evolution, key methodologies, and future directions of action representation and generation in VLA models. We analyze discrete and continuous representations in depth, and examine autoregressive, non-autoregressive, and hybrid generation strategies, highlighting their inherent trade-offs in terms of precision, diversity, and efficiency. In addition, we cover emerging high-efficiency approaches designed for real-time control, such as hybrid generation architectures. Finally, we present a comparative synthesis of the current technological landscape and outline frontier challenges and research opportunities, including integration with world models and the development of generalizable representations across heterogeneous robotic embodiments. This work aims to provide a comprehensive reference for advancing more general, efficient, and reliable embodied agents.

**Keywords:** embodied intelligence; visual-language-action model; action representation; generation strategy

## 0　引言

具身智能(Embodied AI)通过控制智能体在物理世界中感知、交互并完成任务，被认为是构建通用人工智能的重要途径。近年来，随着大语言模型(Large Language Model, LLM)与视觉语言模型(Visual Language Model, VLM)的巨大成功，一类全新的多模态模型——视觉－语言－动作(Vision-Language-Action, VLA)模型应运而生，其旨在通过融合视觉感知、语言理解与动作生成能力，解决语言条件下的机器人任务[1]。VLA 模型通过将互联网规模的语义知识与真实世界的物理交互能力相结合，在处理复杂、非结构化环境中的任

**收稿日期：**2025-08-07； **修回日期：**2025-09-19； **录用日期：**2025-09-22。
**基金项目：**深港联合资助项目（A 类）（SGDX20240115103359001）。
**作者简介：**张文涛(2001—)，男，江西上饶人，硕士研究生，主要研究方向：大模型、具身智能； 孙奥兰（1996—），女，河北唐山人，工程师，硕士，主要研究方向：大模型、信号处理、具身智能； 瞿晓阳（1988—），男，湖北随州人，博士，CCF 会员，主要研究方向：大模型，体系结构、具身智能； 张旭龙（1988—），男，河南许昌人，博士，CCF 会员，主要研究方向：大模型、具身智能、跨模态智能计算； 王健宗（1983—），男，湖北天门人，正高级工程师，博士，CCF 高级会员，主要研究方向：大模型、联邦学习、深度学习。

务时，展现出超越传统强化学习方法的通用性与灵活性，成为当前机器人与人工智能领域的研究前沿[2]。

VLA 领域在过去数年间取得了飞速发展，涌现出大量模型与方法。从早期的 RT-1[3]，到引入大规模网络知识的 RT-2[4]，再到开源社区推动的 OpenVLA[5]以及采用扩散策略的 Octo[6]等模型，VLA 的架构与能力不断演进。针对这一新兴领域，已有学者对机器人基础模型及大型语言模型在机器人中的应用进行了综述[7-10]。然而，这些工作多为宏观概述，对于 VLA 模型如何将多模态“理解”转化为物理“行动”这一核心环节，即动作的表征与生成策略，缺乏系统且深入的剖析。为了弥补现有综述中的空白，并帮助研究人员更迅速深入地掌握该领域的前沿进展，本文旨在对面向具身智能的 VLA 模型动作表征和生成策略进行全面系统的综述。

## 1 VLA 模型的发展与现状

VLA 模型的通用架构通常由三个核心部分组成，分别是视觉编码器、语言编码器和动作解码器，它们协同工作以实现从输入到输出的完整流程[3]。视觉编码器负责处理来自摄像头等传感器的视觉输入，它通常采用强大的预训练视觉基础模型(如 ViT[11])，将场景中的像素信息转化为包含物体类别、位置、姿态和几何关系的结构化特征表示。语言编码器负责理解用户的自然语言指令，该模块通常基于大型语言模型或其变体，将文本指令编码为机器能够理解的向量表示[12]。动作分词器是 VLA 模型关键的组成部分，负责将融合后的视觉和语言信息转化为具体的机器人控制指令，它最常见的是采用自回归的方式，像生成文本一样，一步步地生成代表机器人动作的序列，这些动作可以是关节角度、末端执行器位姿或轮式速度等[13-14]。这种端到端的架构，克服了传统机器人系统中视觉、语言和控制模块各自为政、难以协作的“碎片化管道”问题，从而达到了更强的适应性、泛化能力和任务执行的流畅性[15]。

VLA 模型的发展始于对传统人工智能中视觉、语言和动作系统相互分离的碎片化架构的突破，在近年得到快速发展，如图 1 所示，该图是以“vision-language-action”为关键词谷歌学术搜索结果得到的发文量，2025 年的发文量截止 5 月，2025 年增加的白色部分为按前 5 个月平均发文量估计的后 7 个月发文量。2022 年，以 CLIPort 为代表的早期模型率先将预训练的视觉语言表征与机器人操作相结合，实现了基于语义的精确控制[16]；与此同时，谷歌的 RT-1 则通过大规模模仿学习和开创性的 Transformer 架构，首次将机器人动作离散化为词元进行预测，显著提升了真实世界控制的规模和能力[3]。2023 年，RT-2 的发布成为一个里程碑，它通过在大规模的互联网视觉问答数据和机器人轨迹数据上进行联合微调，证明了网络知识可以被直接迁移到机器人控制中，使模型涌现出“视觉思想链”等高级推理能力[4]。进入 2024 年，该领域向着更开放和高效的方向发展，Octo 模型整合了大规模的开放机器人数据集(Open X-Embodiment)并采用扩散生成策略，推动了通用机器人策略的进步[6]；而 OpenVLA 则作为一个强大的开源模型，证明了通过高效的参数调优，较小规模的模型也能达到甚至超越大型闭源模型的性能[5]。发展至今，最新的趋势是面向更复杂的系统和应用，如 Groot N1 模型，开始采用双系统架构，即结合 LLM 进行高层规划和快速扩散策略进行底层控制来驱动人形机器人，标志着 VLA 模型正朝着更通用、安全和与人类协作的具身智能方向迈进[17]

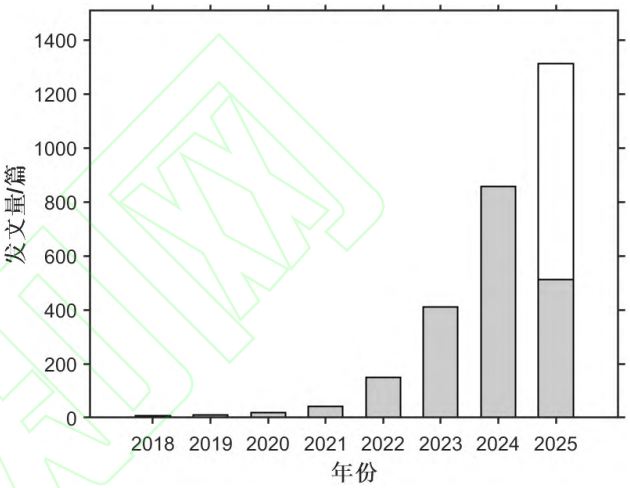


**图 1 近年 VLA 模型相关论文的发文量**

Fig. 1 Number of VLA Model Publications in Recent Years

在 VLA 模型中，动作表征(Action Representation)直接定义了机器人执行的物理指令，而生成策略(Generation Strategy)则是模型的核心决策机制，它负责将视觉和语言等多模态输入映射到相应的动作表征，从而决定在特定情境下应生成何种动作序列以完成任务，二者共同构成 VLA 模型中从“理解”到“执行”的关键桥梁，如图 2 所示。最初的 VLA 模型开始将连续的动作空间离散化为动作词元，并利用 Transformer 架构进行自回归预测，将机器人控制巧妙地转化为了一个序列生成任务[3]。为解决离散化带来的精度损失，以 ACT 模型为代表的工作转而去探索在连续空间中建模，其采用 CVAE 来学习动作的概率分布，从而生成更精确多样的动作[18]。紧接着，以 Diffusion Policy 为开创性工作的扩散模型成为新的研究热点，它通过去噪生成过程，在产生高质量、平滑且多样化的动作轨迹上表现卓越，但其迭代式的推理过程也带来了巨大的计算开销和延迟[19]，具体分类如图 3 所示。目前最新的 VLA 研究前沿聚焦于提升效率，如 $\pi_0$ – Fast 模型中提出的 FAST 技术，通过压缩动作表征，极大地提升模型的推理速度以满足实时控制的需求[20]。

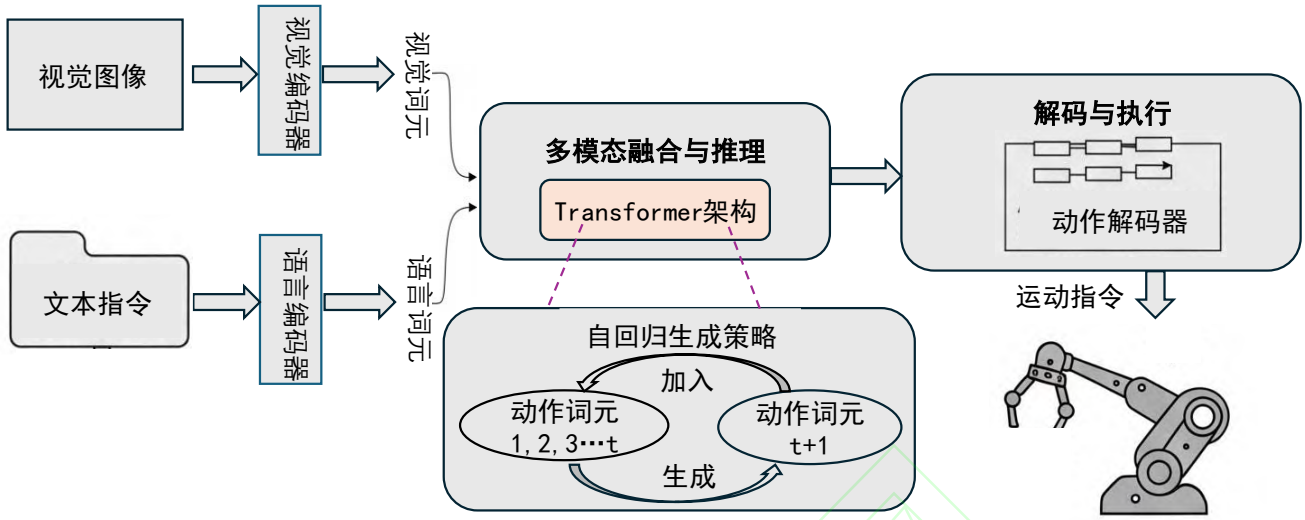


图 2 VLA 模型的经典框架

Fig. 2 Classical framework for VLA models

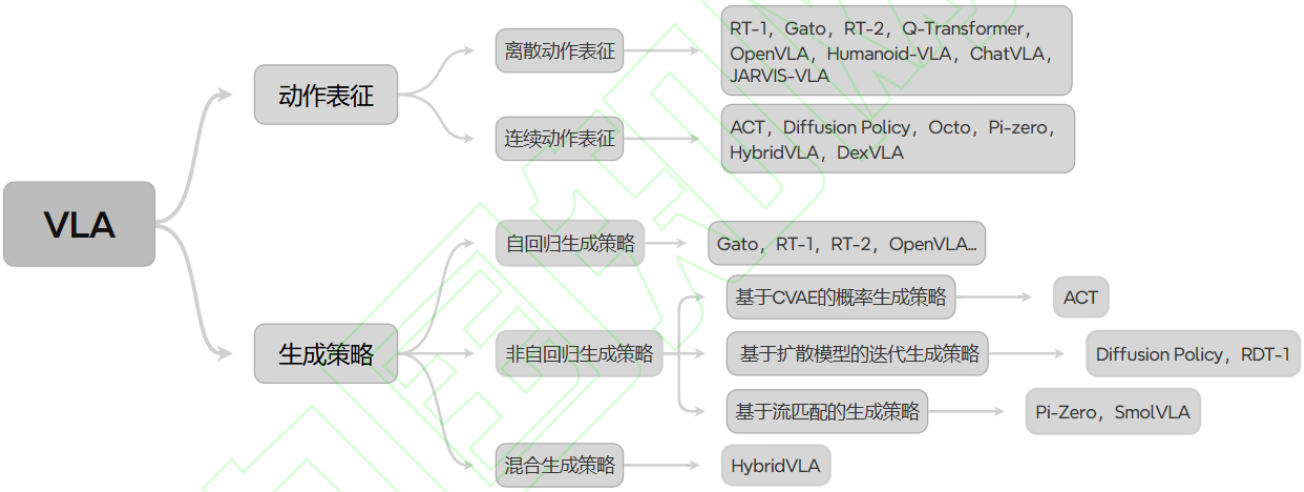


图 3 VLA 模型中的动作表征和生成策略

Fig. 3 Action representation and generation strategies for VLA model

## 2 VLA 模型的动作表征

在 VLA 模型中，动作表征是连接抽象的“理解”与具身的“执行”之间最为关键的部分之一[21]。一个模型即便能完美地感知世界并理解指令，但如果无法将这些信息转化为精确、有效的物理动作，其在现实世界中的价值便无从谈起[22-24]。因此，如何设计动作表征，以应对机器人原生动作空间的高维度、连续性以及一个任务可能存在多种有效解法的复杂挑战，是决定 VLA 模型性能的核心问题[25-26]。为了解决这一难题，领域内的研究逐渐分化为两大主流技术路径：离散化动作表征与连续动作表征。

### 2.1 离散动作表征

离散化动作表征是 VLA 模型发展初期的一种开创性核心思想，其目的是将机器人控制问题转化为强大的序列模型能够直接处理的“语言”问题。

传统的机器人动作(如机械臂末端的 *XYZ* 坐标、旋转角度、夹爪开合度等)是高维度的连续值。然而，以 Transformer 为代表的大型语言模型擅长处理的是离散的词元 (tokens)，为了利用这种技术，研究者们提出将连续的动作空间进行“量化”或“分箱”(Binning)，将其转化为一个有限的、离散的“动作词典”。以模型 RT-1 为例，它将机器人动作空间进行了如下处理：对于每个动作维度，将其均匀地划分为 256 个区间。这样，一个具体的连续值就会被映射到这 256 个区间中的某一个，并被赋予一个唯一的整数 ID。通过这种方式，原本复杂的连续控制信号就被转换成一个可以由模型预测的、离散的分类问题[3]。

离散化动作表征方法在近年得到了快速发展和广泛应用，如表 1 所示，其中 1 个动作维度对应机器人的 1 个可控

自由度。RT-1 (2022) 被认为是将 Transformer 架构成功应用于大规模真实世界机器人控制的开创性工作之一，它首次证明了通过将机器人动作词元化，一个标准的 Transformer 模型可以学习完成数百种不同的厨房操作任务[3]。紧接着，Gato(2022) 提出了更具雄心的“通用智能体”构想，其核心在于统一的词元化方案。它将所有类型的输入数据，无论是图像块、文本、按钮按键还是机器人动作，全部转换成一个扁平化的词元序列。这个序列随后被送入一个巨大的 Transformer 网络进行处理，从而展现了惊人的任务通用性[27]。在此基础

上，RT-2 (2023) 标志着 VLA 发展史上的一个重要里程碑。它不再将 VLA 视为一个独立的机器人模型，而是将其看作一个大型 VLM 能力的直接延伸。其核心思想是，一个足够强大的 VLM 本身就蕴含了世界的常识和推理能力，因此可以通过一种被称为“符号微调”(Symbol-tuning)的方法，让模型直接“说出”代表机器人动作的词元，从而将网络的抽象知识无缝迁移到物理世界的控制中[4]。为了突破行为克隆(Behavioral Cloning)的性能瓶颈，Q-Transformer (2023)作为 RT-1 架构的重要演进，将离散化动作的思路与更强大的离线强化学习(Offline RL)范式相结合。它通过学习一个自回归的 Q 函数，使模型能够从包含成功与失败经验的混合质量数据中进行学习，从而获得了远超传统模仿学习的鲁棒策略[28]。

**表 1 典型模型所用的离散化动作表征**

Tab. 1 Discretized action representation used in typical models

| 年份 | 模型 | 核心范式 | 平台 | 任务领域 | 动作空间类型 | 动作维度 | 离散区间 | 存在的问题 |
|---|---|---|---|---|---|---|---|---|
| 2022 | RT-1[3] | 模仿学习 | 移动机械臂 | 移动操作 | 末端执行器位置、姿态+移动盘 | 11 | 256 | 模仿学习上限、泛化局限 |
| 2022 | Gato[27] | 通用监督学习 | Sawyer 机械臂等 | 机器人操作、游戏、对话 | 末端执行器速度控制+夹爪控制 | 5 | 1024 | 上下文长度限制、推理速度慢 |
| 2023 | RT-2[4] | VLM 共同微调 | 移动机械臂 | 语义驱动的操作 | 末端执行器位置、姿态+移动盘 | 11 | 256 | 物理技能局限、计算成本高 |
| 2023 | Q-Transformer[28] | 线下强化学习 | 移动机械臂 | 多任务操作 | 末端执行器位置、姿态、夹爪 | 8 | 256 | 奖励函数局限、高维动作局限 |
| 2024 | OpenVLA[5] | VLM 微调 | 多种机械臂 | 跨物理形态的操作 | 末端执行器位置、姿态+夹爪 | 7 | 256 | 仅支持单图像、推理效率低 |
| 2025 | Humanoid-VLA[29] | 语言−运动对齐 | 人形机器人 | 移动−操作 | 全身运动姿态 | 24 | 1024 | 数据质量数量有限、依赖底层 RL 策略 |
| 2025 | JARVIS-VLA[30] | ActVLP | 虚拟智能体 | 游戏操作 | 键盘与鼠标 | —— | 51 | 推理速度慢、与顶尖人类玩家差距大 |

时至今日，离散化 VLA 范式仍在不断拓展其应用边界。Humanoid-VLA (2025) 标志着该技术从桌面机械臂向更复杂的人形机器人迈出的关键一步，致力于将自回归控制范式扩展到高自由度的全身姿态控制上[29]。而 JARVIS-VLA (2025) 则将其应用领域从物理世界拓展到了《我的世界》(Minecraft) 等虚拟游戏环境中，利用 VLM 丰富的世界知识来理解复杂指令，并生成离散的键盘和鼠标操作序列，展现了其在复杂交互式环境中的巨大潜力[30]。

分析表 1 中的模型可以发现，离散动作表征的核心范式经历了从早期的直接模仿学习与通用监督学习，向深度融合大型视觉语言模型先验知识的共同微调范式演变，近期模型则进一步发展出更为复杂的复合式训练框架。这种范式上的演进与模型所应对的平台和任务领域日益复杂化紧密相关，从最初的移动机械臂和 Sawyer 机械臂，扩展至跨物理形态的多类型机械臂，乃至人形机器人和虚拟智能体等前沿平台。与此同时，动作空间的维度和表征方式也愈发精细，从 Gato[27] 的 5 维速度控制，发展到 RT-2[4]的 11 维位姿底盘控制再到

Humanoid-VLA[29]的 24 维全身姿态控制。然而，这种技术复杂度的提升也伴随着持续的挑战：无论是早期模型的模仿学习性能上限，还是后期大模型普遍面临的推理效率低、计算成本高昂的问题，亦或是对高质量、大规模数据集的严重依赖，都反映出 VLA 模型在追求更高智能水平与应对现实世界部署约束之间的核心权衡。

离散动作表征方法的最大贡献在于它统一了机器人技术与大型序列模型，使得将在互联网数据上预训练的 VLA 模型的强大知识和泛化能力迁移到机器人控制上成为可能[2]。然而，其主要缺点也同样明显：离散化过程必然会导致精度损失。对于需要亚毫米级精度的任务(如精密装配)，这种量化误差是不可接受的，可能导致动作生硬或任务失败[31]。这一固有的局限性也直接推动了后续基于概率分布和扩散模型等连续空间动作生成策略的发展。

### 2.2 连续动作表征

连续动作表征的核心思想是用概率分布应对动作多模态性，直接在连续空间中预测动作面临一个核心挑战就是动作的多模态性。这意味着对于同一个任务，例如把杯子放到桌子中间，存在无数条同样有效的、细微差别各异的动作轨迹。如果使用简单的回归模型，如基于均方误差 MSE 损失进行训练，模型会试图去“平均”所有可能的正确答案，最终导致模式崩溃(mode collapse)——即模型只会生成一个模糊的、平均化的，甚至完全不可用的动作[32-33]。为了解决这个问题，连续动作表征的核心思想是：不预测一个单一的确定性动作，而是学习一个能够覆盖所有可能有效动作的完整概率分布。通过这种方式，模型能够捕捉到任务解法的多样性。在实际执行时，可以从这个学习到的分布中进行采样，从而得到一个具体的、连贯的、且有效的动作序列[34]。

当前，VLA 模型的研究中，连续动作表征已发展成为与离散化方法并行的核心技术范式之一，并在众多前沿模型中得到了广泛的实现与验证，如表 2 所示。这一演进并非简单的技术更替，而是在范式、架构和核心挑战上都呈现出深刻的变化。ACT[18](2023)模型是探索连续动作概率建模的早期代表性工作之一，它为高精度的灵巧操作任务提供了有效的解决方案。它通过一个条件变分自编码器(Conditional Variational Autoencoder, CVAE)架构来处理连续动作，CVAE 能学习到一个概率性的隐变量空间，这个空间捕捉了所有可能有效动作的分布，从而避免了简单回归模型导致的模式崩溃问题。在进行推理时，模型从这个隐空间采样，然后解码出多样化且高精度的连续动作序列[35]。Diffusion Policy(2023)是将扩散模型(Diffusion Models)成功引入机器人策略学习的开创性工作，为 VLA 的动作生成开辟了一条全新的、影响深远的技术路径，它将机器人策略学习重新定义为一个条件下的动作去噪过程。模型从一个随机噪声开始，在给定的视觉等条件下，通过一个神经网络逐步迭代地去噪，最终生成一条完整、平滑的连续动作轨迹[19]。这个过程是在连续空间中进行的，并且非常擅长学习和表示复杂、多模态的动作分布[36]。Octo(2024)采用了一个基于 Transformer 的扩散解码器来生成连续动作，它的主要贡献在于证明了扩散策略的可扩展性，能够生成适应不同种类机器人和多样化任务的连续动作指令，展示了其强大的泛化能力[6]。

$\pi_0$ (2024)是一个先进的 VLA 模型，它采用了不同于传统扩散模型的另一种新颖的生成式方法来处理连续动作。它采用了流匹配(Flow Matching)的生成式架构，并结合了一个专门的动作专家模块。流匹配学习一种将简单噪声分布平滑地变换为复杂真实动作分布的向量场，从而直接生成连续的控制信号[37]。HybridVLA (2025)的核心是一个统一的 LLM，它在同一层级集成了协作式的扩散和自回归动作生成器，两种策略并行预测同一个低层级动作，并通过协同训练和自适应融合，显著增强了动作的鲁棒性[38]。 DexVLA(2025)采用分层设计，其底层是一个可插拔的、十亿参数级别的扩散动作专家。这个专家模块专门负责根据上层规划生成高质量的、精细的连续动作指令，特别适合需要复杂指尖协调的灵巧操作任务[39]。

分析表 2 可知，随着时间的推移，采用连续动作表征的 VLA 模型展现出向更强大、更复杂的生成式方法演进的明确趋势，以应对日益精细化的机器人操控任务。早期的连续动作模型采用建模动作序列，有效处理了双臂协同操作中的多模态行为。紧随其后的一系列模型，则普遍转向了基于扩散或流匹配的生成策略。这一转变的核心优势在于，扩散类模型能够更稳定、更精确地学习和生成高维、多模态的连续动作分布，这对于完成精细操作至关重要。这种方法论上的统一，支撑了模型在平台多样性上的显著扩展，从单一的双臂或单臂机器人，拓展到了灵巧手等多物理形态平台。然而，这种能力的提升也带来了共同的挑战：复杂的生成过程导致了普遍存在的推理延迟问题，同时这些高性能模型对训练数据有强依赖性，这使得数据获取的成本和难度成为限制模型精度与部署效率进一步提升的关键瓶颈。

表 2 典型模型所用的连续动作表征

Tab. 2 Continuous action representation for typical models

| 年份 | 代表模型 | 核心范式 | 平台 | 任务领域 | 动作维度 | 表征方法类型 | 存在的问题 |
|---|---|---|---|---|---|---|---|
| 2023 | ACT[18] | 模仿学习 | 机械臂 | 精细双臂操作 | 14 | 条件变分 | 硬件局限、感知挑战 |
| 2024 | Octo[6] | 模仿学习 | 机械臂 | 跨物理形态的通用操作 | 7/14 | 条件扩散 | 手腕摄像头处理不佳、依赖演示数据 |
| 2024 | $\pi_0$ [37] | VLM 微调 | 机械臂、移动机器人 | 高灵巧度、长时程操作 | 18 | 条件流匹配 | 严重依赖大规模、部分不开源的高质量演示数据 |
| 2025 | HybridVLA[38] | 协同训练 | 机械臂 | 通用桌面操作 | 7/14 | 混合生成 | 推理速度受限 |
| 2025 | DexVLA[39] | 具身课程学习 | 机械臂、灵巧手机器人 | 跨物理形态的灵巧操作 | — | 多头扩散 | 富含接触的复杂场景中局限大 |

## 3 VLA 模型的动作生成策略

在 VLA 模型中，动作生成策略扮演着至关重要的角色，它是将模型对世界的抽象理解转化为物理世界中具体行动的核心引擎。一个 VLA 模型即便拥有超凡的视觉感知和语言理解能力，但若缺乏一个高效且精确的动作生成策略，其所有智能都将止步于虚拟的思考，无法在现实世界中产生任何有意义的交互，这也正是 VLA 模型区别于传统视觉语言模型的根本所在[40-41]。

动作生成策略直接决定了机器人行为的质量、效率和适应性，并且其设计本身就是在一系列关键性能指标之间进行的复杂权衡。一是精度与效率的权衡，机器人的任务，尤其是精细操作，对动作精度有极高要求。然而，能够生成高质量轨迹的策略通常计算开销巨大，推理速度慢[42-43]。反之，一些追求速度的策略(如经典的自回归生成)虽然稳定，但其串行解码方式严重限制了其频率(通常仅为 3-5 赫兹)，远不能满足机器人实时控制所需的 100 赫兹以上的频率[44]。二是多样性与稳定性的权衡，现实世界充满不确定性，一个任务通常有多种有效的完成方式，优秀的生成策略需要能够捕捉到这种多样性，以适应新颖或变化的环境[45]。

因此，选择何种动作生成策略，并非简单的技术选型，而是对 VLA 模型整体能力的一次深刻塑造。为了在这些相互制约的性能指标中找到最佳平衡，本章将深入探讨 VLA 主流的动作生成策略，如图 4 所示。

### 3.1 自回归生成策略

自回归生成是一种核心的序列数据生成方法，其根本原理是按顺序、一步一步地生成序列中的每一个元素。这个策略的名称来源于“自”(Auto)和“回归”(Regressive)，意为模型的预测依赖于其自身过去已生成的输出。在 VLA 模型下，一个自回归策略可以被形式化地表示为[9]

$$\pi(a_t \mid p, s_{\le t}, a_{<t}) \tag{1}$$

在给定语言指令 p、历史状态 $s_{\le t}$ 和历史动作 $a_{<t}$ 的条件下，生成当前动作 $a_t$ 的概率[3]。现代最先进的自回归模型几乎无一例外地基于 Transformer 的解码器架构，这一架构的核心是其“掩码多头自注意力机制”[46]。掩码在这里起到了至关重要的作用：它确保了在预测序列中第 t 个位置的元素时，模型的注意力机制只能关注到第 1 到 t-1 个位置的信息，而无法“偷看”到 t 位置之后的信息。这种机制强制模型遵循时间的单向因果流，是自回归生成得以实现的基础[47-48]。

在 VLA 模型中，自回归模型将一个由多个离散的动作词元组成的动作序列的联合概率分布，分解为一系列条件概率的乘积。根据 Reed[27]等的总结，给定一个动作序列 $s_{1:L}$ 和参数 θ，可以使用概率的链式法则对数据进行建模：

$$\log p_\theta(s_1,\ldots,s_L) = \sum_{l=1}^{L} \log p_\theta(s_l \mid s_1,\ldots,s_{l-1}) \tag{2}$$

令 $b$ 为一个训练批次 $\mathcal{B}$ 中各个序列的索引，定义一个掩码函数 $m$，其规则如下：如果索引为 $l$ 的词元来自文本或智能体记录的动作，则 $m(b,l)=1$，否则为 0，对于一个批次 $\mathcal{B}$ 的训练损失则为

$$\mathcal{L}(\theta,\mathcal{B}) = -\sum_{b=1}^{|\mathcal{B}|}\sum_{l=1}^{L} m(b,l)\log p_\theta(s_l^{(b)} \mid s_1^{(b)},\ldots,s_{l-1}^{(b)}) \tag{3}$$

这种将动作词元化的自回归框架构成了早期 VLA 模型的基石，它成功地将强大的序列建模能力引入机器人领域。目前很多 VLA 模型仍然使用自回归生成策略，例如 VIMA 采用了一个 Transformer 编码器来共同处理以物体为中心的视觉词元和指令词元，其动作解码器是自回归的 Transformer，能够根据复杂的、混合了图像和文本的提示，生成精确的动作序列[49-50];ChatVLA 构建了一个统一的视觉－语言－动作规

划器，该规划器基于自回归模型，能够根据上下文生成用于对话的文本回复或用控制的动作指令[51]。

因此，自回归生成策略不仅是 VLA 模型的开山之石，更在近年来的发展中展现出强大的生命力和可扩展性[52]。其核心优势在于，它革命性地将机器人控制这一物理问题，成功地统一到了强大且高度可扩展的序列模型框架下，然而为了应用该策略而进行的动作离散化，也不可避免地会损失动作的精度[53]。

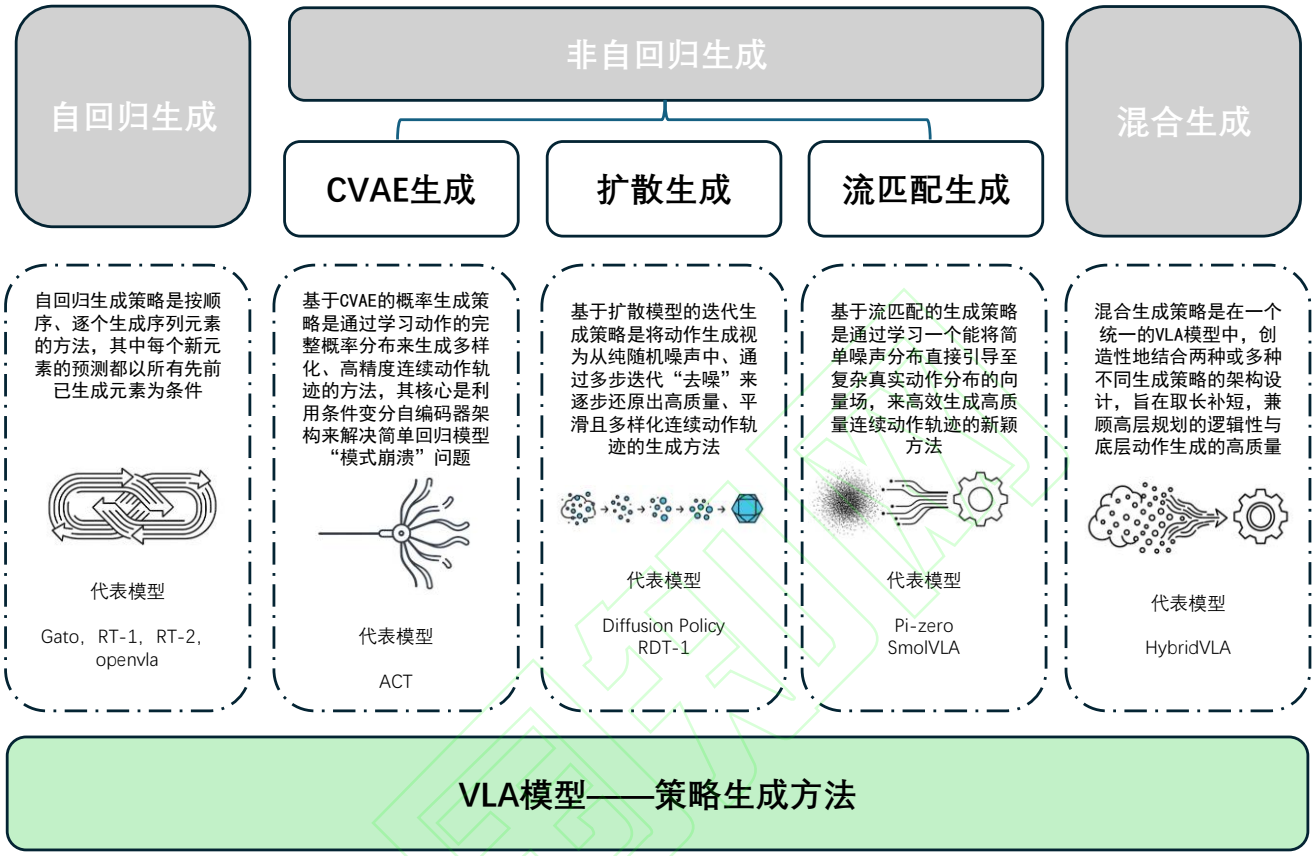


图 4 VLA 模型中的策略生成方法

Fig. 4 Generation strategies methods in VLA models

### 3.2 非自回归生成策略

非自回归生成是一种旨在解决自回归生成策略速度瓶颈的先进动作生成方法。在传统的自回归模型中，动作词元必须一个接一个地串行生成，这极大地限制了模型的推理速度，使其难以满足机器人实时控制的高频需求。非自回归生成策略的核心目标就是打破这种串行依赖，以实现显著的加速[54-55]。

#### 3.2.1 基于 CVAE 的概率生成策略

基于CVAE的概率生成策略是一种先进的连续动作表征生成方法，它的出现主要是为了解决传统回归模型在处理机器人动作时遇到的两大核心难题：动作多模态性和模式崩溃[56-57]。该策略的核心是利用条件变分自编码器的强大能力来对动作的概率分布进行建模，CVAE 通过学习一个低维度的隐变量空间来捕捉动作数据内在的多样性，这个隐空间中的每一个点，都对应着一个具体、有效的动作序列，通过在这个空间中进行采样，模型就能够生成多样化的动作输出[58]。

CVAE 框架包含两个关键部分：CVAE 编码器和 CVAE 解码器。编码器在训练时使用，它的作用是将一个具体的专家演示动作序列 $a_{t:t+k}$ 和对应的观测 $o_t$ 压缩成一个低维的隐变量 $z$。编码器输出的是一个高斯分布的均值和方差，即 $z$ 服从 $N(\mu,\sigma^2)$。解码器是模型在训练和推理时都会用到的核心部分。它接收当前的观测 $o_t$ 和一个从隐空间中采样的 $z$ 作为条件，然后生成一个预测的动作序列 $\hat{a}_{t:t+k}$。通过将 $z$ 作为输入，策略的行为就能够得到调节[18]。

CVAE 的训练目标是最大化在给定观测条件下，生成真实动作序列的对数似然。这通过优化证据下界(Evidence Lower Bound, ELBO)来实现，其损失函数通常包含重构损失和正则化损失。根据 Zhao 等[18]的推导，重构损失确保解码器(策略)能够根据观测和隐变量 $z$ 精确地重构出原始的动作序列，其目的是让模型学会基本的动作模仿，一般用 L1 损失来计算。

$$\mathcal{L}_{reconst} = \mathrm{L1}(\hat{a}_{t:t+k}, a_{t:t+k}) \tag{4}$$

其中：$\hat{a}_{t:t+k}$ 是预测的动作序列，$a_{t:t+k}$ 是专家演示动作序列。该损失项计算了生成动作与真实专家动作之间的 L1 范数，其目标是让模型学会基本的动作模仿。通过最小化这项损失，

解码器被激励去生成一个尽可能接近专家演示的动作序列，从而保证动作的准确性和有效性。正则化损失过计算 KL 散度来约束编码器生成的隐变量分布逼近一个预设的先验分布 p(z)。这可以防止编码器将所有信息都在隐变量 z 中，从而迫使解码器(策略)更多地依赖观测 $o_t$ 来做出决策，增强了模型的泛化能力，具体表示为

$$\mathcal{L}_{reg} = D_{KL}(q_\phi(z \mid a_{t:t+k}, o_t) \| \mathcal{N}(0, I)) \tag{5}$$

其中：$q_\phi$ 代表编码器神经网络。该损失项是 CVAE 模型的核心。它通过最小化 KL 散度，强制编码器生成的隐变量分布逼近一个简单的、结构化的先验分布，这极大地增强了模型的泛化能力并且解决了动作多模态性的问题。最终的总损失函数是这两部分的加权和，即:

$$\mathcal{L} = \mathcal{L}_{reconst} + \beta \mathcal{L}_{reg} \tag{6}$$

其中：β 是一个超参数，用于平衡重构的精确度和对先验分布的遵循程度。

ACT[18]模型是采用 CVAE 策略最典型的代表模型，它构建了一个 CVAE-Transformer 架构，专门用于处理高精度的双臂灵巧操作任务。ACT 还将 CVAE 策略与动作分块(Action Chunking)技术相结合，这意味着它的 CVAE 解码器一次性预测的是一小段未来的连续动作序列，而不是单个时间步的动作[18]。该策略能够生成高精度的连续动作，避免了离散化带来的量化误差，同时通过概率建模，能够有效处理动作多模态性并生成多样化的解决方案，但 CVAE 的架构和训练过程比简单的自回归模型更为复杂。

#### 3.2.2 基于扩散模型的迭代生成策略

基于扩散模型的迭代生成策略是当前 VLA 领域中用于生成连续动作的最前沿、性能最强大的方法之一。它是一种生成式模型，其核心思想新颖且强大：将动作的生成过程，重新定义为一个从纯粹的随机噪声中、逐步去噪以还原出清晰动作轨迹的迭代过程[59-61]。

该策略通过一个训练好的噪声预测网络 (noise prediction network) $\epsilon_\theta$ 来工作，以 Diffusion Policy[19]为例，首先生成一个与目标动作序列维度相同的纯高斯噪声向量 $A^K$，在接下来的 $K$ 个步骤中，模型反复执行去噪操作。在每一步 $k$，噪声预测网络 $\epsilon_\theta$ 会接收当前的噪声动作 $A^K$、观测信息(如图像和指令)以及当前的迭代步数 $k$ 作为输入，并预测出其中包含的噪声。之后模型从当前的噪声动作 $A^K$ 中减去预测出的噪声，并加入少量新的随机噪声，从而得到一个更“干净”的动作向量 $A^{k-1}$，以此经过 $K$ 轮迭代后，初始的纯噪声向量 $A^K$ 最终被精炼成一个无噪声的、可执行的动作序列 $A^0$。这整个迭代过程可以被描述为：

$$A_t^{k-1} = \alpha(A_t^k - \gamma\epsilon_\theta(O_t, A_t^k, k) + \mathcal{N}(0, \sigma^2 I)) \tag{7}$$

其中：$O_t$ 是机器人当前的观测信息，$\alpha, \gamma, \sigma$ 是噪声调度相关的超参数，它们会随着迭代步数 $k$ 的变化而调整，控制着每一步去噪的幅度和引入新噪声的量，$\epsilon_\theta$ 是噪声预测网络，它是通过一个简单的均方误差(Mean Squared Error, MSE)损失函数进行训练的。模型的目标是使预测的噪声 $\epsilon_\theta$ 与实际添加的噪声 $\epsilon^k$ 尽可能接近，其损失函数为

$$\mathcal{L} = \text{MSE}(\epsilon^k, \epsilon_\theta(O_t, A_t^0 + \epsilon^k, k) \tag{8}$$

其中：$A_t^0$ 是来自专家数据集的真实动作序列。训练时，会随机选择一个迭代步数 $k$，并向真实动作添加相应水平的噪声，然后让网络去预测并移除这些噪声。这个过程的本质是让网络做一个去噪的任务，通过反复练习，网络最终学会了从一个被噪声污染的动作中识别并分离出噪声部分。

自被引入机器人领域以来，扩散模型策略迅速发展，并在一系列先进的 VLA 模型中得到应用。MDT(2024)在扩散模型的架构上做出了重要创新，它将专为图像生成设计的、性能更优越的 DiT(Diffusion Transformer)架构成功应用于动作生成任务，取代了在早期扩散模型中常用的 U-Net 架构，展示了更好的性能和扩展性[62-63]。RDT-1B(2024) 是一个参数量高达 12 亿的大型扩散基础模型，专注于解决复杂的双臂协同操作(bimanual manipulation)任务，它同样基于强大的 DiT 架构，并通过在一个统一了动作格式的大型多机器人数据集上进行训练，获得了强大的零样本泛化能力[46]。CogACT 是一个专为工业机器人操作设计的 VLA 框架，其核心也基于扩散模型，它在多步装配、螺丝紧固等高精度工业任务中表现出色，真实世界任务成功率远超其他模型[64-65]。

目前，基于扩散模型的迭代策略生成的动作轨迹质量极高、平滑且极具多样性，是目前表征复杂动作分布最有效的方法之一，但迭代式的去噪过程计算开销巨大，推理速度慢，这成为其在需要高频反馈的实时机器人应用中的主要障碍[66-67]。一个关键的解决办法是通过知识蒸馏等技术，将一个需要多步采样的扩散模型的知识，提炼到一个仅需少量采样就能生成高质量结果的模型中，同时也可以针对扩散模型中大量的矩阵运算，设计专用的硬件加速器或在 FPGA/ASIC 上实现优化的计算核，通过软硬件协同设计来突破算力瓶颈。

#### 3.2.3 基于流匹配的生成策略

基于流匹配的生成策略是一种用于生成连续动作的新颖且强大的方法，在 VLA 模型的发展中，它作为扩散模型的一种高效替代方案而出现，旨在以更稳定、可能更快速的方式来学习复杂的数据分布并生成高质量的动作[68-69]。

与扩散模型逐步去噪的思路不同，流匹配通过一种更直接的回归目标来训练，通常训练效率更高且在数学上更易于处理，它的核心机制是概率路径与向量场。根据 Black[37]等的推导，概率路径 (Probability Path)是在纯噪声 $A^0 \sim \mathcal{N}(0, I)$ 和真实

数据 $A^1$ 之间定义一条连续的路径，在此之后学习一个向量场 $v_\theta(A^t, O_t, t)$，其中 $O_t$ 是机器人当前的观测信息，t 代表当前点在从噪声到真实数据的路径上所处的时间点。这个向量场能够描述路径上任意一点 $A^t$ 的流动方向和速度，使其能够沿着正确的轨迹移动到最终的真实动作 $A^1$。

模型的训练目标是让神经网络预测的向量场 $v_\theta$ 与目标向量场 $u_t$ 尽可能地接近，这是通过最小化一个简单的均方误差(MSE)损失函数来实现的，即流匹配损失，数学表达为

$$\mathcal{L}(\theta)=\mathbb{E}_{t,p_t(A),p(A^1)}[\| v_\theta(A^t,O_t,t)-(A^1-A^0)\|^2] \quad (9)$$

其中：$p(A^1)$ 表示从所有真实动作的数据集中，随机抽取一个动作 $A^1$ 作为本次训练的目标，$p_t(A)$ 表示根据随机选出的 $A^1$ 和 $A^0$，构造出一个位于两者路径中间的点 $A^t$。这个训练过程本质上是一个简单的回归任务，让模型学会如何根据当前状态(含噪动作和观测)计算出通往目标(真实动作)的正确方向。在推理阶段，动作的生成是一个常微分方程的求解过程。从一个随机采样的噪声点 $A^0 \sim \mathcal{N}(0,I)$ 开始，模型通过沿着学习到的向量场 $v_\theta$ 进行积分，逐步将噪声点“推送”到数据流形上，最终在 $t=1$ 时生成一个高质量的动作 $A^1$，这个积分过程通常使用数值方法(如欧拉法)来近似：

$$A^{t+\delta}=A^t+\delta\cdot v_\theta(A^t,O_t,t) \quad (10)$$

其中：$\delta$ 是积分的步长，一般来说，它使用 10 个积分步骤即可完成从噪声到动作的生成。

$\pi_0$ (2024)是一个用于通用机器人控制的先进VLA模型，它明确地将流匹配作为其核心生成技术。$\pi_0$ 模型采用了一个PaliGemma VLM 作为其强大的视觉语言基座，并结合了一个专门的、基于流匹配的动作专家模块，这个专家模块负责接收上层的规划意图，并通过流匹配的原理直接生成高质量的连续控制信号[37]。GraspVLA (2025) 是首个主要在合成数据上进行预训练的抓取 VLA 模型，GraspVLA 的动作解码器同样是一个基于流匹配的动作专家，它进一步将流匹配与一种名为渐进式动作生成(Progressive Action Generation, PAG)的技术相结合，以支持从零样本到少样本的泛化，并能适应特定物体的抓取偏好[70]。

流匹配策略是非常高效且强大的生成方法，它为 VLA 动作生成提供了一个有别于扩散模型的、极具前景的技术路径，作为一个相对较新的策略，其在 VLA 领域的全部潜力和局限性仍在被积极探索中，未来有望在推理速度和训练稳定性上展现出比传统扩散模型更大的优势。

### 3.3 混合生成策略

混合生成策略是 VLA 模型中一个前沿且强大的生成策略，它本身并非一种全新的底层生成原理，而是一种更高阶的架构设计思想。其核心是在一个统一的 VLA 模型中，创造性地融合两种或多种不同的基础生成策略，以期取长补短，在应对复杂任务时达到单一策略难以实现的效果。它旨在构建一个既能进行高效、长时程的规划，又能生成高质量、平滑、精确的底层动作的 VLA 系统，其最终目标是在机器人控制的速度、精度、多样性和鲁棒性之间取得一个更优的平衡[71-72]。

在混合策略中，非常经典的组合就是融合自回归策略和扩散策略。自回归模型非常擅长处理语言、进行逻辑推理和长序列规划，因此它非常适合用于高层级规划，生成离散的、有逻辑的子任务序列[73]；而扩散模型在生成高质量、平滑且多样化的连续动作轨迹方面表现卓越，因此它更适合用于低层级控制，负责将一个具体的子任务指令转化为精细、平滑的物理动作[74]。

HybridVLA (2025) 是采用混合策略的典型代表，它在一个统一的 LLM 中，采用协作式的扩散和自回归生成策略。HybridVLA 专注于复杂的机器人操作任务，包括单臂和双臂协同操作，它的优势在于能够形成一个自适应的动作集成(Adaptive Action Ensemble)，在复杂的真实和模拟任务中实现了鲁棒的控制和强大的泛化能力，其性能超越了此前的 SOTA 模型[38]。

通过结合不同策略的优点，混合策略能够更好地处理包含长时程规划和精细物理交互的复杂任务，在解决高级任务时，这种分层协同的模式通常比单一策略的端到端模型表现更好。随着 VLA 模型被应用于更复杂的场景(如人形机器人、家庭服务等)，混合与分层架构将成为一个越来越重要的发展方向，如何更高效、更无缝地融合不同策略，将是未来 VLA 研究的一个关键课题。但目前混合生成策略也有很多关键技术难点，首先是高层规划与底层控制的时空错配，其根源在于不同系统的设计目标冲突。高层规划追求的是逻辑推理的深度和广度，这需要庞大的模型和复杂的计算；而底层控制追求的是反应的速度和精度，这要求模型轻量且高效。其次是特征空间的对齐与接地，高层规划器生成的通常是抽象的、符号化的子任务指令，而底层控制器需要的是具体的、数值化的动作指令。如何确保高层的符号指令能够被底层控制器准确理解并接地到正确的物理动作上，是保证任务成功的关键。针对这些问题，未来解决方向是设计异步执行框架，允许底层控制器在一个缓冲计划的指导下持续高频运行，同时高层规划器在后台异步地生成和更新未来的计划，也可以可以探索如何构建一个统一的、跨越符号与连续值的表征空间，让高层规划器和底层控制器在这个共享的表征空间中进行交互，从而从根本上解决对齐问题。

### 3.4 从感知到执行的策略对比

为了具象化不同动作表征和生成策略的实际影响，本文以一个常见的机器人任务为例：擦掉桌子上的咖啡渍。对于离散动作表征与自回归生成策略，VLA 模型通过视觉编码器识别出“咖啡渍”、“桌子”等物体，并结合语言指令理解任务意图。通过采用自回归的方式，一步步地生成一系列代表机器人动作的离散词元并产生动作序列，机器人的底层控制器接收这些离散的动作指令，并依次执行。整个动作流程看起来会比较清晰，但可能会显得有些停顿和机械，不够连贯。而连续动作表征与非自回归生成策略是该策略将动作生成视为一个去噪过程，模型从一个完全随机的噪声开始，在当前视觉观测的引导下，通过数十个迭代步骤，逐步得到一条完整、平滑、高精度的机械臂末端运动轨迹。机器人直接流畅地执行这条生成好的轨迹，擦拭动作会非常连贯、平滑，并且能够根据视觉感知识别的污渍具体形状，生成贴合的擦拭路径。

## 4 模型评估

为了系统性地研究和评估 VLA 模型的可扩展性、泛化能力与终身学习能力，本章将详细介绍两个在具身智能领域具有代表性的基准数据集：LIBERO[75]和 Open X-Embodiment[56]。这两个数据集各有侧重，为评测不同维度的模型性能提供了坚实的基础。通过不同模型的性能指标，可以间接得到动作表征和生成策略对模型的影响。

### 4.1 LIBERO 数据集

LIBERO(Lifelong Learning Benchmark for Robot Manipulation, LIBERO) 是一个专为机器人终身学习设计的开创性基准，旨在系统性地研究智能体在其生命周期内学习和适应新任务的能力。它尤其关注两类知识的迁移：陈述性知识(关于实体和概念，如物体类别、空间关系)和程序性知识(关于如何行动和执行，如抓取、开门等行为)。LIBERO 通过一个可扩展的程序化生成流程，原则上可以创造出无限多的任务。为便于标准化评测，该项目目前提供了四个任务套件，共计 130 个任务。所有任务均提供了由人类专家通过遥操作收集的高质量演示数据，以支持样本高效的模仿学习。这四个核心任务套件分别专注于测试对空间关系的知识转移、对程序性知识的转移、对物体概念的知识转移和对混合多种知识的综合迁移能力。

为了量化不同模型的效果，LIBERO 引入三个核心评价指标，这些指标都基于任务成功率计算，分别是前向迁移、负向后迁移和成功率曲线下面积，但是目前研究针对 VLA 模型性能测试效果基本还是主要以任务成功率来衡量，下面为一些常见离散和连续动作 VLA 模型在 LIBERO 数据集测试的结果，其中成功率为模型在四个任务套件的平均成功率：

**表 3 LIBERO 数据集典型 VLA 模型评估**

Tab. 3 Evaluation of Typical VLA Models on LIBERO

| 动作类型 | VLA 模型 | 平均成功率/% |
|---|---|---|
| 连续 | Diffusion Policy[19] | 72.4 |
| | Octo[6] | 75.1 |
| | DiT Policy[76] | 82.4 |
| | OpenVLA-OFT[13] | 95.4 |
| | $\pi_0$ [37] | 94.2 |
| 离散 | OpenVLA[5] | 76.5 |
| | WorldVLA[77] | 79.1 |

分析表 3 可以明显看出，连续动作模型在平均成功率上展现出比离散动作模型更强的潜力，并且流匹配和扩散模型生成策略要远优于传统的自回归生成策略。其中：OpenVLA-OFT 在所有模型中表现最佳，是当前的 SOTA 模型。这表明，机器人学习领域正在通过更先进的连续动作生成策略来不断提升操控的精准度和任务成功率。离散动作模型虽然有效，但在精度和成功率上可能已遇到瓶颈。

### 4.2 Open X-Embodiment 数据集

Open X-Embodiment 是一个大规模、开放的机器人学习数据集，旨在推动通用机器人策略的研究。其核心思想是，通过汇集来自全球多个研究机构、在不同机器人上收集的数据，来训练一个能够适应新机器人、新任务和新环境的通用模型。传统机器人学习方法通常是为一个特定的机器人、一项特定的任务训练一个特定的模型。Open X-Embodiment 项目旨在打破这种局限，通过在极其多样化的数据上进行大规模训练，验证跨机器人平台的正向迁移效应，即一个模型可以从其他机器人的经验中学习，从而提升自身在未见过的任务上的能力。为了便于使用，所有来源不同的数据都被转换成统一的 RLDS (Robotics Learning Data Set, RLDS) 数据格式。这种标准化的格式支持在主流的深度学习框架中进行高效、并行的加载。针对这个数据集，目前的模型性能指标是任务的成功率，目前典型的 VLA 模型测试的结果如下：

**表 4 Open X-Embodiment 数据集典型 VLA 模型评估**

Tab. 4 Evaluation of Typical VLA Models on Open X-Embodiment

| 动作类型 | VLA 模型 | 平均成功率/% |
|---|---|---|
| 连续 | Octo-Base [6] | 16.8 |
| | $\pi_0$ [37] | 70.1 |
| 离散 | RT-1[3] | 6.8 |
| | TraceVLA[78] | 42.0 |
| | RT-1-X[56] | 53.4 |
| | RT-2-X[56] | 60.7 |
| | OpenVLA[5] | 27.7 |

分析上表可知，性能最好的连续模型($\pi_0$, 70.1%)优于性能

最好的离散模型(RT-2-X, 60.7%)，这表明先进的连续动作模型在该数据集上可能具有更高的性能上限，但是 Octo-Base 的成功率仅 16.8%，远低于大多数离散模型，这说明动作空间的类型(连续或离散动作表征)本身并不决定性能，模型的具体架构和训练方法更为关键。

## 5 挑战与机遇

在详细探讨了 VLA 模型中动作表征的两种范式以及与之对应的各类生成决策过程后，我们可以看到该领域取得了长足的进步。然而，在通往更通用、更鲁棒的具身智能的道路上，动作表征与策略生成仍然面临着深刻的挑战，同时也孕育着巨大的发展机遇，本节将对这些机遇与挑战进行总结与展望。

### 5.1 与世界模型深度融合

一个极具前景的发展方向是将 VLA 模型与学习型世界模型进行深度融合，这是推动具身智能从当前的反应式控制，向更高级的预测性规划演进的关键一步。世界模型能够学习环境的动态变化规律，并在其内部得到不同动作可能导致的未来结果。通过赋予 VLA 模型这种预测未来的能力，机器人将不再仅仅是根据当前感知进行决策，而是能够执行更复杂的长时程任务，例如通过预演来选择最优的工具使用顺序，或是在执行有风险的操作前预测到潜在的失败并主动规避，这将大幅推动具身智能领域的发展。

### 5.2 对传统范式机器人的突破

传统机器人系统普遍采用“感知–规划–控制”的模块化、分离式流程。这种架构的主要挑战在于，各个独立优化的模块之间存在集成鸿沟，使得系统整体性能受限于最薄弱的环节，且难以适应复杂的非结构化环境。VLA 模型带来的机遇在于其端到端的学习范式，它直接将高维的视觉、语言输入映射到动作输出，避免了中间状态的复杂转换和信息损失。这种统一的框架使得模型能够学习到感知、语言和动作之间更深层次、更隐式的关联，从而在泛化能力和鲁棒性上超越了传统模块化系统。

### 5.3 面向实时控制的高效生成

为了突破现有策略的速度瓶颈，高效生成技术正成为研究热点。以 Groot N1 为代表的并行解码策略，能够并发地生成动作词元块，将延迟降低降低 60%，使 100 赫兹的实时控制成为可能[17]；同时，以 $\pi_0$ – Fast 变体所采用的 FAST 等压缩动作表征技术，通过在频域上进行操作，将长序列压缩为少量词元，实现了高达 15 倍的推理加速，这使得在资源受限的机器人上部署强大的 VLA 模型成为现实[20]。无论是通过并行解码来加速生成过程，还是通过压缩表征来减少生成的数据量，这些高效生成技术都在推动 VLA 模型克服实时性的挑战。它们使得在资源受限的机器人上部署强大的 VLA 模型成为可能，是目前该领域最重要的技术难点，也是充满机遇的发展方向。

### 5.4 跨机器人形态的通用表征

跨机器人形态的通用表征是 VLA 模型未来发展中最具变革性的机遇之一，其目标是打破当前“一个模型服务一个机器人”的困境，训练出能够不限机器人型号、即插即用的机器人形态无关策略，实现这一目标的核心在于设计一种不依赖于特定机器人运动学的抽象动作表征[79-80]。例如，模型的输出不再是某个机器人手臂特定的关节角度，而是一个更通用的、抽象的指令。这一研究方向的最终愿景是，未来一个强大的 VLA 模型可以将其学会的技能，通过仅需几分钟的校准数据，就无缝迁移到全新的轮式平台、四足机器人或人形机器人上，从而极大地加速具身智能的部署和应用。同时当前主流的 VLA 模型本质上仍属于模仿学习的范畴，这意味着它们的性能上限受限于演示数据的质量和覆盖度，很难通过探索发现超越演示数据的新策略或更优解。如何让 VLA 模型在模仿的基础上，具备强化学习的探索和自我优化能力，从而弥补这一能力鸿沟，是该领域一个至关重要的前沿方向。目前，一些融合了离线强化学习思想的模型正在为此做出尝试，但这仍然是一个巨大的挑战。

### 5.5 开放世界中的安全性与可靠性

确保 VLA 模型在开放、动态且不可预测的真实世界中的安全性与可靠性，是其从实验室走向实际应用所面临的最严峻挑战之一[60,81]。目前的 VLA 系统在这一方面存在多重瓶颈：首先，许多安全机制仍依赖于预设的、硬编码的力或力矩阈值，这极大地限制了模型在遇到新颖或突发状况时的自适应能力[82]；其次，模型的感知鲁棒性不足，例如，在光照不佳或有阴影的场景下，其视觉模块的准确率会下降 20%~30%[83]；在杂乱环境中，机器人也常常会错误判断被部分遮挡物体的位置或姿态，从而导致任务失败更关键的是，专门用于碰撞预测的模型在动态空间中的准确率通常也只有 82%左右，这在家庭或工厂等人机共存的环境中带来了严重的安全隐患[84-85]。此外，即使是紧急停止这类核心安全功能，其本身也包含 200 到 500 毫秒的显著延迟，这在高速运动或关键操作中可能是致命的[86-87]。这些在感知、决策和执行层面上的可靠性与安全性缺陷，共同构成阻碍 VLA 模型在自动驾驶、医疗和家庭服务等安全关键领域广泛部署的核心障

碍。

### 5.6 计算与能源消耗问题

巨大的计算与能源消耗是阻碍先进 VLA 模型从实验室走向广泛实际应用的核心瓶颈，尤其是在移动机器人等边缘计算场景中。当前，最先进的 VLA 模型参数量可高达 70 亿，在未经压缩的情况下通常需要超过 28GB 的显存才能运行[88]。这些严苛的要求远远超出了 NVIDIA Jetson 等主流边缘 AI 平台的处理能力，从而将这些强大模型的应用限制在了拥有专业计算资源的场合[55]。此外，采用如扩散模型等高性能生成策略会带来额外的计算开销，其成本大约是传统自回归解码器的 3 倍[43-44]。除了计算能力，能源消耗也是一个关键制约因素，尤其对于依赖电池供电的移动平台而言，如何降低单次推理能耗已成为一个重要的研究方向。VLA 模型的端到端特性虽然带来了强大的泛化能力，但其“黑箱”的决策过程也带来了新的挑战。当模型执行失败时，我们很难判断问题出在视觉感知、语言理解还是动作决策上，这为系统的调试和安全性验证带来了困难。此外，VLA 模型的性能严重依赖于超大规模、高质量的演示数据集，这些数据的采集、标注和处理成本极其高昂，成为限制其发展和应用的关键瓶颈。

## 6 结语

从本文分析可知，VLA 模型呈现出一个清晰的技术演进趋势：VLA 的动作生成正从早期将离散表征与自回归策略的成功结合，逐步迈向采用更复杂的生成式建模来驱动连续表征，以实现更高的动作保真度和多样性。最后混合生成策略的出现，预示着未来的发展方向将是融合不同策略的优点，以期在机器人控制的“速度、质量、多样性”这一核心权衡中取得更优的平衡，最终赋能于能够在复杂、开放世界中可靠运行的通用智能体。

**参考文献**


[1] Roumeliotis K I, Tselikas N D. Chatgpt and Open-AI models: A preliminary review[J]. Future Internet, 2023, 15(6): 192.

[2] Stone A, Xiao T, Lu Y, et al. Open-World Object Manipulation using Pre-Trained Vision-Language Models[C]//Conference on Robot Learning. PMLR, 2023: 3397-3417.

[3] Brohan A, Brown N, Carbajal J, et al. RT-1: Robotics Transformer for Real-World Control at Scale[J]. Robotics: Science and Systems XIX, 2023.

[4] 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. PMLR, 2023: 2165-2183.

[5] Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: An Open-Source Vision-Language-Action Model[C]//Conference on Robot Learning. PMLR, 2025: 2679-2713.

[6] Ghosh D, Walke H R, Pertsch K, et al. Octo: An Open-Source Generalist Robot Policy[C]//Robotics: Science and Systems. 2024.

[7] Sapkota R, Cao Y, Roumeliotis K I, et al. Vision-language-action models: Concepts, progress, applications and challenges[EB/OL]. [2025-05-10]. https://arxiv.org/pdf/2505.04769.pdf.

[8] 赵博涛, 亢祖衡, 瞿晓阳, 等. 基于多模态大模型的具身智能体研究进展与展望[J]. 大数据, 2025, 11(3). [ZHAO Botao, KANG Zuheng, QU Xiaoyang, et al. Research Progress and Prospects of Embodied Intelligence Agents Based on Multimodal Large Models[J]. Big Data Research, 2025, 11(3).]

[9] Zhong Y, Bai F, Cai S, et al. A Survey on Vision-Language-Action Models: An Action Tokenization Perspective[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2507.01925.pdf.

[10] Liu Y, Chen W, Bai Y, et al. Aligning cyber space with physical world: A comprehensive survey on embodied ai[EB/OL]. [2025-027-11]. https://arxiv.org/pdf/2407.06886.pdf.

[11] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.

[12] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2023, 36: 34892-34916.

[13] Kim M J, Finn C, Liang P. Fine-tuning vision-language-action models: Optimizing speed and success[EB/OL]. [2025-07-02]. https://arxiv.org/pdf/2502.19645.pdf.

[14] Zhang Z, Zheng K, Chen Z, et al. GRAPE: Generalizing Robot Policy via Preference Alignment[C]//ICRA 2025 Workshop on Foundation Models and Neuro-Symbolic AI for Robotics.

[15] Katiyar N. A Model-Driven Framework for Domain-Specific Adaptation of Time Series Forecasting Pipeline[M]. McGill University (Canada), 2023.

[16] Shridhar M, Manuelli L, Fox D. Cliport: What and where pathways for robotic manipulation [C]// Conference on robot learning. PMLR, 2022: 894-906.

[17] Bjorck J, Castañeda F, Cherniadev N, et al. Gr00t n1: An open foundation model for generalist humanoid robots[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2503.14734.pdf.

[18] Zhao T, Kumar V, Levine S, et al. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware[J]. Robotics: Science and Systems XIX, 2023.

[19] Chi C, Xu Z, Feng S, et al. Diffusion policy: Visuomotor policy learning via action diffusion[J]. The International Journal of Robotics Research, 2023: 02783649241273668.

[20] Pertsch K, Stachowicz K, Ichter B, et al. Fast: Efficient action tokenization for vision-language-action models[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2501.09747.pdf.

[21] Gao J, Belkhale S, Dasari S, et al. A taxonomy for evaluating generalist robot policies[EB/OL]. [2025-07-12]. https://arxiv.org/pdf/2503.01238.pdf.

[22] Chen B, Xu Z, Kirmani S, et al. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 14455-14465.

[23] Fan C, Jia X, Sun Y, et al. Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2505.02152.pdf.

[24] Cheng H, Xiao E, Yu C, et al. Manipulation Facing Threats: Evaluating Physical Vulnerabilities in End-to-End Vision Language Action Models[EB/OL]. [2025-07-10]. https://arxiv.org/pdf/2409.13174.pdf.

[25] Firoozi R, Tucker J, Tian S, et al. Foundation models in robotics: Applications, challenges, and the future[J]. The International Journal of Robotics Research, 2025, 44(5): 701-739.

[26] Hu Y, Xie Q, Jain V, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2312.08782.pdf.

[27] Reed S, Zolna K, Parisotto E, et al. A generalist agent[EB/OL]. [2025-

07-08]. https://arxiv.org/pdf/2205.06175.pdf.

[28] Chebotar Y, Vuong Q, Hausman K, et al. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions [C]// Conference on Robot Learning. PMLR, 2023: 3909-3928.

[29] Ding P, Ma J, Tong X, et al. Humanoid-vla: Towards universal humanoid control with visual integration[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2502.14795.pdf.

[30] Li M, Wang Z, He K, et al. Jarvis-vla: Post-training large-scale vision language models to play visual games with keyboards and mouse[EB/OL]. [2025-06-20]. https://arxiv.org/pdf/2503.16365.pdf.

[31] Gu Z, Li J, Shen W, et al. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2501.02116.pdf.

[32] Nie Y, Li L, Gan Z, et al. MLP architectures for vision-and-language modeling: An empirical study[EB/OL]. [2025-06-05]. https://arxiv.org/pdf/2112.04453.pdf.

[33] Wang S. Roboflamingo-plus: Fusion of depth and rgb perception with vision-language models for enhanced robotic manipulation[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2503.19510.pdf.

[34] Ren A Z, Lidard J, Ankile L L, et al. Diffusion policy policy optimization[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2409.00588.pdf.

[35] Guo Y, Zhang J, Chen X, et al. Improving Vision-Language-Action Model with Online Reinforcement Learning[EB/OL]. [2025-07-04]. https://arxiv.org/pdf/2501.16664, 2025.

[36] Chiang H T L, Xu Z, Fu Z, et al. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs[EB/OL]. [2025-07-10]. https://arxiv.org/pdf/2407.07775.pdf.

[37] Black K, Brown N, Driess D, et al. $\pi_0$: A Vision-Language-Action Flow Model for General Robot Control[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2410.24164.pdf.

[38] Liu J, Chen H, An P, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2503.10631.pdf.

[39] Wen J, Zhu Y, Li J, et al. DexVLA: Vision-Language Model with Plug-In Diffusion Expert for General Robot Control[EB/OL]. [2025-06-20]. https://arxiv.org/pdf/2502.05855, 2025.

[40] Zhao Q, Lu Y, Kim M J, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models [C]// Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 1702-1713.

[41] Li Z, Wu X, Du H, et al. Benchmark evaluations, applications, and challenges of large vision language models: A survey[EB/OL]. [2025-06-18]. https://arxiv.org/pdf/2501.02189.pdf.

[42] Gbagbe K F, Cabrera M A, Alabbas A, et al. Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations [C]// 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2024: 2864-2869.

[43] Xiang T Y, Jin A Q, Zhou X H, et al. VLA Model-Expert Collaboration for Bi-directional Manipulation Learning[EB/OL]. [2025-08-01]. https://arxiv.org/pdf/2503.04163.pdf.

[44] Driess D, Springenberg J T, Ichter B, et al. Knowledge Insulating Vision-Language-Action Models: Train Fast, Run Fast, Generalize Better[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2505.23705.pdf.

[45] Liu S, Wu L, Li B, et al. Rdt-1b: a diffusion foundation model for bimanual manipulation[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2410.07864.pdf.

[46] 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, volume 1 (long and short papers). 2019: 4171-4186.

[47] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[48] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[49] Jiang Y, Gupta A, Zhang Z, et al. VIMA: General Robot Manipulation with Multimodal Prompts[C]//NeurIPS 2022 Foundation Models for Decision Making Workshop.

[50] Xu Z, Wu K, Wen J, et al. A survey on robotics with foundation models: toward embodied ai[EB/OL]. [2025-05-08]. https://arxiv.org/pdf/2402.02385.pdf.

[51] Zhou Z, Zhu Y, Zhu M, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model[EB/OL]. [2025-06-10]. https://arxiv.org/pdf/2502.14420.pdf.

[52] Guruprasad P, Sikka H, Song J, et al. Benchmarking Vision, Language, & Action Models on Robotic Learning Tasks[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2411.05821.pdf.

[53] O'Neill A, Rehman A, Maddukuri A, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0 [C]// 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 6892-6903.

[54] Ke T W, Gkanatsios N, Fragkiadaki K. 3d diffuser actor: Policy diffusion with 3d scene representations[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2402.10885.pdf.

[55] Ding P, Zhao H, Zhang W, et al. Quar-vla: Vision-language-action model for quadruped robots [C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 352-367.

[56] Ju C, Wang H, Li Z, et al. Turbo: informativity-driven acceleration plug-in for vision-language models[EB/OL]. [2025-06-10]. https://arxiv.org/pdf/2312.07408.pdf.

[57] Zhang T, Hu Y, Cui H, et al. A Universal Semantic-Geometric Representation for Robotic Manipulation[C]//Conference on Robot Learning. PMLR, 2023: 3342-3363.

[58] Xu S, Wang Y, Xia C, et al. VLA-Cache: Towards Efficient Vision-Language-Action Model via Adaptive Token Caching in Robotic Manipulation[EB/OL]. [2025-07-01]. https://arxiv.org/pdf/2502.02175.pdf.

[59] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[60] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models [C]// International conference on machine learning. PMLR, 2021: 8162-8171.

[61] Choi J, Kim S, Jeong Y, et al. ILVR: Conditioning method for denoising diffusion probabilistic models[C]//18th IEEE/CVF International Conference on Computer Vision, ICCV 2021. Institute of Electrical and Electronics Engineers Inc., 2021: 14347-14356.

[62] Reuss M, Yağmurlu Ö E, Wenzel F, et al. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2407.05996.pdf.

[63] Peebles W, Xie S. Scalable diffusion models with transformers [C]// Proceedings of the IEEE/CVF international conference on computer vision. 2023: 4195-4205.

[64] Li Q, Liang Y, Wang Z, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2411.19650.pdf.

[65] Wu J, Zhong M, Xing S, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks[J]. Advances in Neural Information Processing Systems, 2024, 37:

69925-69975.

[66] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12873-12883.

[67] Bolya D, Huang P Y, Sun P, et al. Perception encoder: The best visual embeddings are not at the output of the network[EB/OL]. [2025-06-10]. https://arxiv.org/pdf/2504.13181.pdf.

[68] Lipman Y, Chen R T Q, Ben-Hamu H, et al. Flow Matching for Generative Modeling[C]//The Eleventh International Conference on Learning Representations..

[69] Gat I, Remez T, Shaul N, et al. Discrete flow matching[J]. Advances in Neural Information Processing Systems, 2024, 37: 133345-133385.

[70] Deng S, Yan M, Wei S, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2505.03233.pdf.

[71] Wen J, Zhu Y, Li J, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation[EB/OL]. [2025-07-18]. https://arxiv.org/pdf/2409.12514.pdf.

[72] Zhang B, Zhang Y, Ji J, et al. SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Constrained Learning[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2503.03480.pdf.

[73] Li X, Liu M, Zhang H, et al. Vision-Language Foundation Models as Effective Robot Imitators[C]//The Twelfth International Conference on Learning Representations.

[74] Chen L, Lu K, Rajeswaran A, et al. 78 Reinforcement learning via sequence modeling[J]. Advances in neural information processing systems, 2021, 34: 15084-15097.

[75] Liu B, Zhu Y, Gao C, et al. Libero: Benchmarking knowledge transfer for lifelong robot learning[J]. Advances in Neural Information Processing Systems, 2023, 36: 44776-44791.

[76] Hou Z, Zhang T, Xiong Y, et al. Diffusion transformer policy[EB/OL]. [2025-07-12]. https://arxiv.org/pdf/2410.15959.pdf.

[77] Cen J, Yu C, Yuan H, et al. WorldVLA: Towards Autoregressive Action World Model[EB/OL]. [2025-06-12]. https://arxiv.org/pdf/2506.21539.pdf.

[78] Zheng R, Liang Y, Huang S, et al. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies[EB/OL]. [2025-06-18]. https://arxiv.org/pdf/2412.10345.pdf.

[79] Li X, Li P, Liu M, et al. Towards generalist robot policies: What matters in building vision-language-action models[EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2412.14058.pdf.

[80] Intelligence P, Black K, Brown N, et al. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization[EB/OL]. [2025-06-08]. https://arxiv.org/pdf/2504.16054.pdf.

[81] Fan L, Chen K, Xu Z, et al. Language Reasoning in Vision-Language-Action Model for Robotic Grasping [C]// 2024 China Automation Congress (CAC). IEEE, 2024: 6656-6661.

[82] Ma Y, Song Z, Zhuang Y, et al. A survey on vision-language-action models for embodied ai[EB/OL]. [2025-07-05]. https://arxiv.org/pdf/2405.14093.pdf.

[83] Zhou X, Han X, Yang F, et al. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model[EB/OL]. [2025-07-15]. https://arxiv.org/pdf/2503.23463.pdf.

[84] Zhen H, Qiu X, Chen P, et al. 3d-vla: A 3d vision-language-action generative world model[EB/OL]. [2025-06-18]. https://arxiv.org/pdf/2403.09631.pdf.

[85] Huang J, Yong S, Ma X, et al. An embodied generalist agent in 3d world[EB/OL]. [2025-06-17]. https://arxiv.org/pdf/2311.12871.pdf.

[86] Patel D, Eghbalzadeh H, Kamra N, et al. Pretrained language models as visual planners for human assistance [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 15302-15314.

[87] Zhou Y, Wang S, Dai S, et al. Chop: Mobile operating assistant with constrained high-frequency optimized subtask planning[EB/OL]. [2025-06-20]. https://arxiv.org/pdf/2503.03743.pdf.

[88] Wang Z, Zhou Z, Song J, et al. Towards testing and evaluating vision-language-action models for robotic manipulation: An empirical study[EB/OL]. [2025-06-18]. https://arxiv.org/pdf/2409.12894.pdf.

This work is partially supported by Hong Kong/Shenzhen Joint Funding Project (Category A) (SGDX20240115103359001)

**ZHANG Wentao**, born in 2001, M. S. candidate. His research interests include large models, embodied intelligence.

**SUN Aolan**, born in 1996, M. S. Her research interests include large models, signal processing, and embodied intelligence.

**Qu Xiaoyang**, born in 1988, Ph. D. His research interests include large models, architecture, and embodied intelligence.

**ZHANG Xulong**, born in 1988, Ph. D. His research interests include large models, embodied intelligence, and cross-modal intelligent computing.

**WANG Jianzong**, born in 1983, Ph. D. His research interests include large models, federated learning, and deep learning.