arXiv:2603.00926v1 [cs.RO] 1 Mar 2026

DAM-VLA：基于动态动作模型的视觉-语言-动作框架，用于机器人操作

Xiongfeng Peng[1], Jiaqian Yu[1], Dingzhe Li[1],Yixiang Jin[1],Lu Xu[1],Yamin Mao[1],Chao Zhang[1],Weiming Li[1], Sujin Jang[2,3],Dongwook Lee[2],和 Daehyun Ji[2]

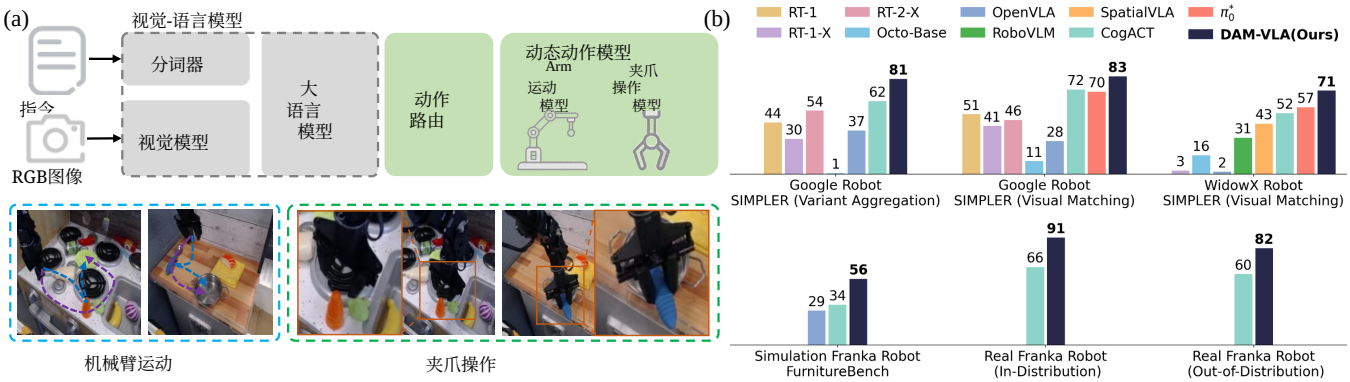

图 1：DAM-VLA 框架和实验结果。(a) 我们提出了 DAM-VLA 框架，将视觉语言模型的内在推理能力与专为机械臂运动和夹爪操作设计的基于扩散的动作模型动态结合起来。在各种机器人任务中，机械臂运动通常覆盖的空间范围大于夹爪操作。因此，在观测图像中，机械臂运动的轨迹往往占据大部分区域，而夹爪操作通常局限于一小块区域；(b) 在广泛评估中，与最先进的视觉-语言-动作方法相比，DAM-VLA 实现了更高的平均成功率，在 SIMPLER 仿真中的抓取与放置任务、FurnitureBench 仿真中的长时域任务，以及真实世界的抓取与放置评估中均展现出性能提升。

摘要——**在仓库、医院和家庭等动态环境中，机器人必须在大幅运动与精细操作之间无缝切换，才能完成复杂任务。然而，当前的视觉-语言-动作（VLA）框架大多改编自预训练的视觉-语言模型（VLM），往往难以兼顾任务适应性与复杂操作所需的专门精确度。为应对这一挑战，我们提出 DAM-VLA，一种基于动态动作模型的视觉-语言-动作框架。DAM-VLA 将视觉语言模型推理与专门用于机械臂和夹爪控制的基于扩散的动作模型相结合。具体而言，该框架引入了：（i）一种动作路由机制，利用特定任务的视觉和语言线索选择合适的动作模型（例如机械臂运动或夹爪操作）；（ii）一种动态动作模型，将高层视觉语言模型认知与低层视觉特征融合，以预测动作；（iii）一种双尺度动作加权机制，实现机械臂运动模型与夹爪操作模型之间的动态协调。在广泛评估中，与当前最先进的视觉-语言-动作方法相比，DAM-VLA 在模拟环境（SIMPLER、FurnitureBench）和真实场景中均取得了更高的成功率，展现出从标准抓取与放置任务到高难度长时程任务和高接触任务的稳健泛化能力。**

[1]Xiongfeng Peng、Jiaqian Yu、Dingzhe Li、Yixiang Jin、Lu Xu、Yamin Mao、Chao Zhang 和 Weiming Li 任职于中国三星研发中心北京分部（SRCB）高级研究实验室[2]Sujin Jang、Dongwook Lee 和 Daehyun Ji 任职于韩国三星人工智能中心 DS部门[3]Sujin Jang 还任职于韩国汉阳大学ERICA校区

I. 引言

机器人学面临的一项核心挑战，是让机器人能够在动态环境中执行各种任务。传统的机器人学习方法通常针对特定机器人和任务整理数据集，并据此训练策略。由此得到的策略往往是专用型策略，例如广受欢迎的 ACT [1] 和扩散策略 [2]。尽管这些方法在目标场景中能够达到较高的精确度，但在不同环境和任务之间的泛化能力较差。

近年来，视觉-语言-动作模型因能够通过将连续动作离散化为多个区间来进行动作预测，从而将预训练视觉语言模型扩展至机器人学领域，受到了广泛关注。RT-2 [3] 和 OpenVLA [4]等代表性工作已在多任务学习和泛化能力方面展现出出色的性能。视觉-语言-动作模型使机器人能够理解视觉观测和语言指令，并生成具有泛化能力的动作序列。因此，利用视觉语言模型的固有能力开发 VLA框架，对于在动态环境中兼顾特定任务的精确度和广泛的泛化能力至关重要。

一方面，若干现有的视觉-语言-动作方法，如 RT-H [5], RT-Affordance [6], 和 ECOT [7], , 引入思维链（CoT）推理来分析夹爪与物体之间的空间关系，并预测相应的动作。尽管思维链增强了



图 2：以将胡萝卜放在盘子上的任务为例，我们指出机械臂运动与夹爪操作之间的三点差异：路径约束、视觉注意力和数据集表征。

视觉语言模型的推理能力并提升泛化能力，但它会引入大量额外的推理词元，并显著增加推理时间。另一方面，基于扩散的视觉语言动作方法，包括 $\pi 0$ [8]、TinyVLA [9], RDT-1B[10],RoboDual [11],CogACT [12],和 HybridVLA [13],，在视觉语言模型之后附加一个独立的扩散头。这些方法要么以视觉语言模型提取的特征为条件，要么在扩散过程中将去噪时间步与带噪动作联合嵌入词元序列。尽管能够实现更精确的操作，但仅依赖视觉语言模型提取的特征会限制更丰富多模态线索的融合。

为了更好地利用视觉语言模型的固有能力，服务于兼顾特定任务操作与通用操作、并适用于动态环境的视觉-语言-动作模型，我们首先指出机械臂运动与夹爪操作之间的若干关键差异。在许多机器人任务中，机械臂运动覆盖的空间范围通常大于夹爪操作。因此，机械臂轨迹往往在观测图像中占据主导，而夹爪操作则局限于较小的局部区域。图 2 以将胡萝卜放在盘子上的任务为例展示了这种差异，其中非灰色区域的视觉注意力突显了两者的不同。具体而言，机械臂运动需要全局注意力，而夹爪操作则需要局部聚焦。其根本差异可概括如下：**(1) 路径约束**。机械臂运动的约束相对较少，因为机器人可以沿多种轨迹到达胡萝卜。相比之下，夹爪操作受到严格约束，必须采用精准的抓取姿态才能成功。**(2) 视觉注意力**。机械臂运动依赖于对场景的全局理解，而夹爪操作则需要细粒度的局部视觉注意力。**(3)数据集表征**。数据集中的机械臂运动片段通常远多于夹爪操作片段。尽管夹爪操作片段数量较少，但它们对任务成功至关重要，而且往往更为复杂。

基于这些区别，我们利用视觉语言模型的推理能力来区分动作类型（机械臂运动与夹爪操作），并应用相应的动作模型来执行所需的操作。我们没有将视觉语言模型与独立的动作模型松散耦合，而是提出了**DAM-VLA**框架（图1），充分发挥视觉语言模型的优势，以支持动态环境中的任务特定精确度和泛化能力。本工作的主要贡献总结如下：（1）**动作路由**。视觉语言模型引导的路由器解读任务——特定的视觉和语言线索，以选择合适的动作模型（例如，机械臂运动或夹爪操作）。（2）**动态动作模型**。一种双头扩散模型，将视觉语言模型的高层认知与低层视觉信息相结合，以通过不同模型预测动作。（3）**双尺度动作加权**。一种双尺度加权机制（即轨迹级和动作块级）可实现机械臂运动模型与夹爪操作模型之间的动态协调。（4）**广泛评估**。与最先进的视觉-语言-动作方法相比，DAM-VLA在SIMPLER仿真中的抓取与放置任务 [14] 、FurnitureBench仿真中的长时域任务 [15],以及真实世界抓取与放置实验中，均取得了更高的平均成功率。

II. 相关工作

视觉-语言-动作模型。 大语言模型 [16], [17], [18]和视觉语言模型 [19], [20], [21], [22], [23] 推动了视觉-语言-动作模型的发展，这类模型通过整合动作生成能力扩展了视觉语言模型。RT-2 [3] 将七维动作离散化为离散词元，并采用视觉语言模型 PaLI-X [24] 进行预测；OpenVLA [4] 则采用类似方法，使用 Prismatic视觉语言模型 [25]。RT-H [5]、RT-Affordance [6], 和 ECoT [7] 引入思维链（CoT）推理，在预测动作之前分析夹爪与物体之间的空间关系。尽管这些方法能够更有效地利用预训练视觉语言模型的推理能力，从而提升泛化能力，但大量额外的推理词元（例如，OpenVLA 中为 7 个，而 ECoT 中为 350 个）显著降低了推理速度并减少了控制频率。此外，它们未能充分满足精确动作估计所要求的连续性、准确性和特异性。

扩散动作模型。 扩散策略 [2] 将扩散模型 [26], [27] 应用于机器人学习，展现了对多模态动作分布进行建模的能力。后续研究 [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37] 通过将扩散策略扩展到三维环境、拓展其能力、提升效率并引入架构创新，进一步推进了这一研究方向。其中，Octo [37] 在基于 Transformer 的骨干网络中加入紧凑的扩散头，使其能够适配不同机器人的动作输出。尽管基于扩散的模型提升了其在多种任务和机器人平台上的效率与性能，但它们尚未利用预训练的大语言模型和视觉语言模型，而这些模型的

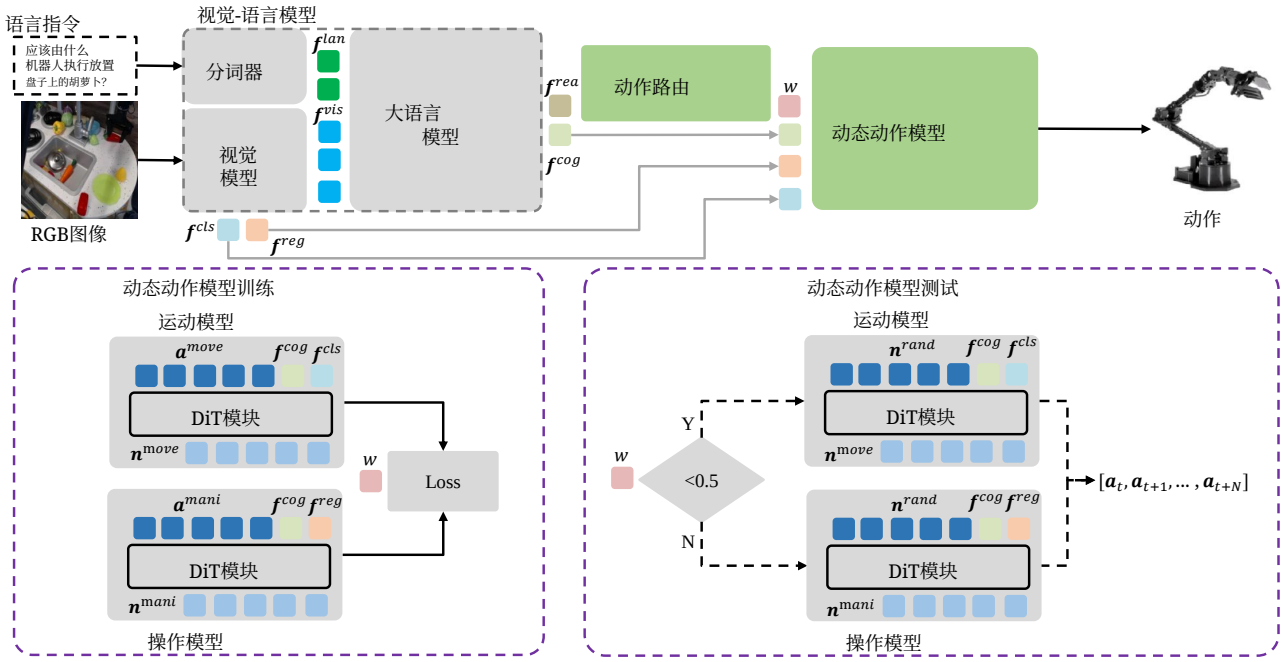

图3：我们的DAM-VLA架构。给定RGB图像观测和任务描述，模型预测一段时序动作序列。该过程包含三个关键组件：1）视觉-语言模型将观测编码为视觉词元、类别词元和寄存器token，并将视觉词元与一组语言词元融合，生成认知与推理潜变量；2）动作路由模块生成权重并将其输入动态动作模型；3）动态动作模型通过将来自视觉模型的低层类别词元或寄存器词元与来自视觉语言模型的高层认知潜变量相结合，动态执行不同的动作模型，以预测动作序列。

强大泛化能力和推理能力有望进一步增强策略鲁棒性。

基于扩散的视觉-语言-动作模型。为结合视觉语言模型与基于扩散的动作模型的优势，$\pi0$ [8] 引入独立的扩散头，通过流匹配生成动作，而TinyVLA [9]则在紧凑的视觉语言模型之后附加轻量级扩散头。CogACT [12]与DiVLA [38] 将推理与动作预测解耦，分别交由视觉语言模型和扩散头完成。HybridVLA [13] 进一步在单一大语言模型中整合扩散与自回归动作预测。然而，这些方法并未充分利用视觉语言模型的固有能力来构建VLA模型，以在动态环境中同时实现任务特定的精确度与可泛化的操作。

III. 方法

在本节中，我们首先在第III-A节描述DAM-VLA的整体架构。随后，我们在第III-B节介绍视觉语言模型。为充分利用不同动作模型的特定操作能力以及视觉语言模型固有的推理能力，我们在第III-C节引入动作路由机制与动态动作模型。为进一步增强鲁棒性，我们在第III-D节提出双尺度动作加权机制。

A. 整体架构

我们的目标是开发一种基于动态动作模型的视觉-语言-动作框架，使不同的机器人能够实际执行多种动态环境中的任务，同时接收 RGB图像观测和以语言指令形式给出的任务描述。形式化地，给定时间 t的语言指令 $\boldsymbol{l}$ 和视觉观测 $\boldsymbol{o}_t$ ，模型 π会预测一个时序动作序列 $[\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, ..., \boldsymbol{a}_{t+N}] = \pi(\boldsymbol{l}, \boldsymbol{o}_t)$。动作空间 $\boldsymbol{a}_t = [\delta \boldsymbol{x}, \delta \boldsymbol{\theta}, s^{grip}]$ 对应于具有 7 个自由度（DoF）的夹爪，其中 $\delta \boldsymbol{x}$表示末端执行器的相对平移偏移量，$\delta \boldsymbol{\theta}$ 表示旋转变化，而 $s^{grip} \in \{0, 1\}$表示夹爪的张开或闭合状态。

在图3中，DAM-VLA的架构由三个关键组件组成：1）视觉-语言模型，它将观测 $\boldsymbol{o}_t$ 中的信息编码为视觉词元 $\boldsymbol{f}^{vis}$、类别词元 $\boldsymbol{f}^{cls}$和寄存器词元$\boldsymbol{f}^{reg}$，并将视觉词元 $\boldsymbol{f}^{vis}$与来自语言指令 $\boldsymbol{l}$的一组语言词元 $\boldsymbol{f}^{lan}$ 进行整合，生成认知潜变量 $\boldsymbol{f}^{cog}$和推理潜变量 $\boldsymbol{f}^{rea}$；2）动作路由模块，生成权重 w 并将其输入动态动作模型；3）动态动作模型，通过将来自视觉模型的低层类别词元 $\boldsymbol{f}^{cls}$或寄存器词元 $\boldsymbol{f}^{reg}$与来自视觉语言模型的高层认知潜变量 $\boldsymbol{f}^{cog}$ 相结合，动态执行不同的动作模型，以预测动作序列 $[\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, ..., \boldsymbol{a}_{t+N}]$。

B. 视觉-语言模型

视觉模型将RGB图像输入处理为一组词元，其中不仅包括视觉词元 $\boldsymbol{f}^{vis}$，还包括来自DINOv2 [39]的类别词元 $\boldsymbol{f}^{cls}$和寄存器词元 $\boldsymbol{f}^{reg}$。视觉模型由强大的

视觉变换器、DINOv2和SigLIP [40],组成，这些模型在互联网规模图像数据上进行了预训练，以捕捉丰富的低级视觉特征和高级语义理解。在每个时间步 t，图像观测 $\boldsymbol{o}_t$ 同时输入手臂运动模型和夹爪操作模型，二者分别生成下采样后的特征图。随后，这些特征图沿通道维度进行拼接，经过线性投影层，并序列化为一组词元，其中包括视觉词元 $\boldsymbol{f}^{vis} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_{NV}]$、类别词元 $\boldsymbol{f}^{cls}$和寄存器词元 $\boldsymbol{f}^{reg}$。

大语言模型负责整合视觉信息和语言指令，并估计推理词元。我们使用 LLaMA-2 模型 [18] 作为骨干网络。语言指令 $\boldsymbol{l}$ 被分词为一组语言词元 $\boldsymbol{f}^{lan} = [\boldsymbol{l}_1, \boldsymbol{l}_2, ..., \boldsymbol{l}_{NL}]$。随后，这些词元 $\boldsymbol{f}^{lan}$与视觉词元 $\boldsymbol{f}^{vis}$以及一个额外的可学习推理词元拼接在一起，并由大语言模型通过因果注意力机制进行处理。最终输出分别为认知潜变量和推理潜变量 $\boldsymbol{f}^{cog}$和 $\boldsymbol{f}^{rea}$。 $\boldsymbol{f}^{rea}$和 $\boldsymbol{f}^{cog}$分别由大语言模型第二层和最后一层 Transformer 层的隐藏特征得到。 $\boldsymbol{f}^{rea}$ 作为后续动作路由模块的输入，用于选择合适的动作模型； $\boldsymbol{f}^{cog}$ 则作为动作模型的输入，用于预测动作。

C. 动作路由机制与动态动作模型

为了确定动作状态属于手臂运动还是夹爪操作，我们设计了一种动作路由机制。该机制利用视觉语言模型的推理能力来学习权重 w ，并通过我们设计的标注权重 $\hat{}w \in 0, 1$ 进行监督，以动态执行手臂运动模型或夹爪操作模型。标注权重 $\hat{}w$ 根据机器人夹爪的状态确定。当机器人夹爪的状态从张开变为闭合，或从闭合变为张开时，我们将其定义为机器人动作从手臂运动（$\hat{}w = 0$）向夹爪操作（$\hat{}w = 1$）的转换。$\hat{}w$ 的详细计算过程将在双尺度动作加权部分中说明。为了利用视觉语言模型的推理能力，我们将来自视觉语言模型的推理潜变量 $\boldsymbol{f}^{rea}$ 输入动作路由模块，并通过全连接层和归一化层进行处理。输出为预测权重 w，并通过以下交叉熵损失进行监督：

$$L_{class} = || - (\hat{w} \log(w) + (1 - \hat{w}) \log(1 - w))||^1. \quad (1)$$

为了充分发挥不同扩散动作模型各自特定的操作能力，以及视觉语言模型固有的推理能力，我们提出了动态动作模型。在训练阶段，动态动作模型利用标注权重$\hat{}\boldsymbol{w}^{move}$和$\hat{}\boldsymbol{w}^{mani}$，有针对性地训练手臂运动模型和夹爪操作模型。这些权重由轨迹权重和动作块权重计算得出。关于$\hat{}\boldsymbol{w}^{move}$ 和$\hat{}\boldsymbol{w}^{mani}$ 设计的详细说明，请参见双尺度动作加权部分。除了将视觉语言模型提取的高层认知潜变量 $\boldsymbol{f}^{cog}$ 作为条件外，两个动作模型还分别将低层视觉词元 $\boldsymbol{f}^{cls}$或 $\boldsymbol{f}^{reg}$ 作为条件。由于两个模型在图像中关注的注意力区域不同，因此我们输入不同的视觉词元。手臂运动模型需要更多全局注意力，因此我们将类别词元$\boldsymbol{f}^{cls}$ 作为条件。相比之下，夹爪操作模型更关注局部注意力，因此我们输入寄存器词元 $\boldsymbol{f}^{reg}$。对于动作模型中的扩散 Transformer（DiT） [27] 块，我们参考了 [12]。手臂运动模型和夹爪操作模型的损失函数分别通过以下方式进行监督：

$$L_{move} = ||\boldsymbol{n}_i^{move} - \hat{\boldsymbol{n}}^{move}||^2_{\sum \hat{\boldsymbol{w}}^{move}}, \quad (2)$$

$$L_{mani} = ||\boldsymbol{n}_i^{mani} - \hat{\boldsymbol{n}}^{mani}||^2_{\sum \hat{\boldsymbol{w}}^{mani}}, \quad (3)$$

其中 $||\cdot||_{\sum}$ 表示马氏距离，它根据标注权重$\hat{}\boldsymbol{w}^{move}$和$\hat{}\boldsymbol{w}^{mani}$对误差项进行加权。 $\boldsymbol{n}_i^{move}$和 $\boldsymbol{n}_i^{mani}$分别表示手臂运动模型和夹爪操作模型在加噪动作序列的第 i个去噪步骤上的预测噪声值。$\hat{}\boldsymbol{n}^{move}$ 和$\hat{}\boldsymbol{n}^{mani}$ 分别对应于运动和操作的扩散动作模型的真实噪声值，并且是随机生成的。两个动作模型的输入包括加噪动作 $\boldsymbol{a}^{move}$和 $\boldsymbol{a}^{mani}$。 $\boldsymbol{a}^{move}$ 使用真实噪声$\hat{}\boldsymbol{n}^{move}$ 和真实动作$\hat{}\boldsymbol{a}$计算，而 $\boldsymbol{a}^{mani}$ 使用真实噪声$\hat{}\boldsymbol{n}^{mani}$ 和真实动作$\hat{}\boldsymbol{a}$计算。总损失计算为运动损失、操作损失和分类损失的加权和。相应的超参数 w_1、 w_2和 w_3 分别设置为1.0、1.0和0.0001。

$$L = w_1 * L_{move} + w_2 * L_{mani} + w_3 * L_{class}. \quad (4)$$

在测试阶段，动态动作模型会根据预测的权重 w 选择并执行相应的动作模型。如果 $w < 0.5$，模型将运行手臂运动模型，以高层视觉语言模型提取的认知潜变量 $\boldsymbol{f}^c$ 和全局类别词元 $\boldsymbol{f}^{cls}$ 为条件，预测一系列多步动作。否则，如果 $w \geq 0.5$，模型将运行夹爪操作模型，以认知潜变量 $\boldsymbol{f}^c$ 和局部寄存器词元 $\boldsymbol{f}^{reg}$ 为条件。此外，两个模型都会接收随机噪声 $\boldsymbol{n}^{rand}$ 作为输入，以促进扩散过程。

D. 双尺度动作加权

为增强区分手臂运动与夹爪操作的鲁棒性，我们提出了一种用于模型训练的双尺度动作加权机制，如图4所示。该机制的核心目标是从全局（轨迹级）和局部（动作块级）两个角度调节不同动作类型的重要性，从而自适应地监督学习过程。

轨迹级权重（$\boldsymbol{w}^t$）：这一全局视角根据机器人夹爪的二元状态变化，将整个任务轨迹划分为不同的机械臂运动和夹爪操作阶段

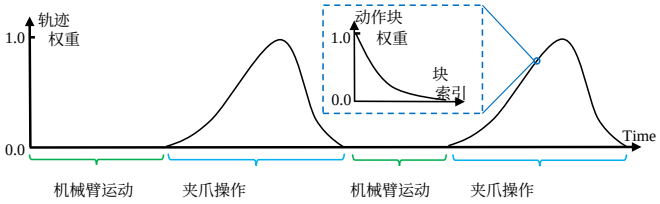

图 4：双尺度动作加权机制示意图。轨迹权重通过非对称高斯分布突出关键操作阶段。在每个预测块内，动作块权重采用指数衰减，以优先保证近期的时间精度。最终权重融合两个尺度的信息，以指导模型监督。

。对于每个夹爪操作过程 k，我们采用非对称高斯分布 $\{\mathcal{N}(u,\ \sigma_l^2)\ \mathcal{N}(u,\ \sigma_r^2)\}$ 来定义权重 $\boldsymbol{w}_k^t$。两个分布具有相同的均值 u，表示夹爪状态发生转换的时间中点（例如，从打开变为闭合）。为体现这样的先验：动作模型在状态变化前需要更高的精确度和监督重点，我们为前沿（$\sigma_l = 6$）赋予较大的方差，为后沿（$\sigma_r = 2$）赋予较小的方差。聚合轨迹权重定义为 $\boldsymbol{w}^t = \text{Nor m}(\sum_k \boldsymbol{w}_k^t)$，表示所有与操作相关的权重之和经归一化后的结果。

动作块级权重（$\boldsymbol{w}^a$）：从局部视角出发，我们考虑动作序列固有的时间不确定性。鉴于预测置信度通常会随着与当前状态的时间距离增加而降低，我们采用指数递减函数：$\boldsymbol{w}_j^a = \gamma^j$，其中 j 表示动作块内的索引，$\gamma = 0.8$ 为衰减因子。

多尺度集成：最终权重表示为 $\boldsymbol{w}^{move} = (1 - \boldsymbol{w}^t) \odot \boldsymbol{w}^a$ 和 $\boldsymbol{w}^{mani} = \boldsymbol{w}^t \odot \boldsymbol{w}^a$。将这些权重应用于 L_{move} 和 L_{mani}后，我们提出的双尺度动作加权机制通过逐点调制，动态协调手臂运动模型与夹爪操作模型。

此外，标注权重 $\hat{w}$ 由轨迹权重导出：若 $\hat{\boldsymbol{w}}^t > 0.5$，则为 $\hat{w} = 1$；否则为 $\hat{w} = 0$。该权重 $\hat{w}$ 作为预测权重 w的真实标签，并通过交叉熵损失 L_{class}进行监督。

IV. 实验

我们进行了大量实验，全面评估所提出方法的性能，并清晰展示其在特定任务和通用操作场景中的有效性。具体而言，IV-A 节详细介绍了训练和微调流程。随后，我们在 IV-B 节展示了 SIMPLER [14]和 FurnitureBench [15] 上的实验结果。我们还基于抓取与放置任务进行了真实场景评估，并在 IV-C 节展示了结果。IV-D 节通过消融研究分析了框架中各个组成部分的作用。

方法 / Google(VA)	不同任务上的成功率				Avg
	PCC	MN	OCD	ODPA	
RT-1 [3]	90%	46%	35%	3%	44%
RT-1-X [44]	49%	33%	29%	10%	30%
RT-2-X [44]	82%	79%	35%	21%	54%
Octo-Base [37]	1%	4%	1%	0%	1%
RoboVLM [45]	76%	60%	11%	-	-
SpatialVLA [46]	88%	73%	42%	-	-
OpenVLA [4]	64%	64%	19%	1%	37%
CogACT [12]	96%	**84%**	29%	40%	62%
DAM-VLA（我们的方法）	**98%**	74%	**68%**	**84%**	**81%**

表 I：在 SIMPLER 模拟评估的变体聚合（VA）设置下，我们的方法与现有视觉-语言-动作方法在 Google 机器人上执行四项任务的成功率比较。（PCC：拾取可乐罐，MN：移至附近，OCD：打开/关闭抽屉，ODPA：打开抽屉并放入苹果，平均值：平均值）

方法 / Google（视觉匹配）	不同任务的成功率				Avg
	PCC	MN	OCD	ODPA	
RT-1 [3]	87%	39%	72%	8%	51%
RT-1-X [44]	59%	33%	56%	17%	41%
RT-2-X [44]	79%	78%	25%	4%	46%
Octo-Base [37]	18%	4%	24%	0%	11%
OpenVLA [4]	14%	51%	48%	0%	28%
RoboVLM [45]	77%	62%	44%	-	-
SpatialVLA [46]	86%	78%	57%	-	-
CogACT [12]	92%	82%	**75%**	39%	72%
π_0^* [8]$^+$	89%	81%	55%	53%	70%
DAM-VLA（我们的方法）	**96%**	**84%**	**75%**	**78%**	**83%**

表 II：在 SIMPLER 模拟评估的视觉匹配（VM）设置下，Google 机器人在四项任务上的成功率比较。（PCC：拾取可乐罐，MN：移至附近，OCD：打开/关闭抽屉；ODPA：打开抽屉并放置苹果，Avg：平均值；$^+$：open-pi-zero，在 Fractal 数据集上训练。）

A. 训练与微调细节

大型 Open X-Embodiment 数据集 [41] 包含来自 22 种不同机器人形态的逾 100 万条机器人操作轨迹。我们主要使用其中两个重要子集 Fractal [42] 和 Bridge- DataV2 [43],作为预训练数据集。我们的视觉-语言-动作模型在 8 块 NVIDIA H100 GPU 上以恒定学习率 2×10^{-5}和 256 的批量大小训练，耗时约两天。

此外，我们还使用仿真和真实场景数据集对 DAM-VLA 模型进行微调。在仿真实验中，我们采用 FurnitureBench 基准，该基准包含接触密集型的长时程操作任务。具体而言，我们使用“单腿”装配任务中的 500 条专家轨迹对 DAM-VLA 模型进行微调。在真实世界评估中，我们构建了一个抓取与放置场景，由操作员远程操控 Franka 机器人抓起一个杯子并将其放入一个碗中。我们总共采集了 50 条演示轨迹用于微调。微调过程采用与预训练相同的超参数：学习率为 2×10^{-5}，批量大小为 256，并使用 8 个 NVIDIA H100 GPU。为公平比较，我们也使用相同的数据集对 OpenVLA 和 CogACT 基线模型进行微调。

方法 / WidowX(VM)	不同任务的成功率				Avg
	SoT	CoP	GoY	EiB	
RT-1-X [44]	4%	8%	0%	0%	3%
Octo-Base [37]	7%	13%	0%	44%	16%
Octo-Small [37]	47%	8%	1%	51%	27%
OpenVLA [4]	4%	0%	0%	4%	2%
ECOT [7]	4%	8%	0%	0%	3%
RoboVLM [45]	29%	25%	13%	58%	31%
SpatialVLA [46]	17%	25%	**29%**	**100%**	43%
CogACT[12]	63%	50%	25%	71%	52%
π_0^*[8]$^+$	62%	59%	24%	81%	57%
DAM-VLA（我们的方法）	**88%**	**71%**	25%	**100%**	**71%**

表 III：在 SIMPLER 仿真评估的视觉匹配（VM）设置下，我们的方法与现有视觉-语言-动作方法在 WidowX 机器人上四项任务的成功率比较。（SoT：将勺子放在毛巾上；CoP：将胡萝卜放在盘子上；GoY：将绿色积木叠放在黄色积木上；EiB：将茄子放入黄色篮子；平均值：平均值；$^+$：open-pi-zero，使用 BridgeDataV2 训练。）

B. 仿真评估

我们首先使用 SIMPLER 仿真 [14],进行评估。SIMPLER 是一套开源仿真环境，旨在复现常见的真实世界机器人操作设置。与真实场景评估相比，这种真实到仿真的方法具有可扩展、可复现且信息丰富等优点，可作为高质量真实世界评估的补充。对各种 VLA 模型的大量测试表明，在 SIMPLER 仿真环境中的评估结果与真实世界表现高度相关。SIMPLER 支持两种机器人形态：Google 机器人和 WidowX机器人。对于 Google 机器人，我们在四项任务中分别采用视觉匹配（VM）和变体聚合（VA）设置进行评估；而 WidowX机器人仅在 VM 设置下进行评估。VM 设置通过尽可能减少仿真环境与真实世界环境之间的差异，紧密复现真实世界任务。相比之下，VA 设置在 VM 设置的基础上，引入背景、光照、干扰物和桌面纹理等因素的变化。所有 VLA 模型均以任务完成成功率作为评估指标。值得注意的是，我们参照 CogACT [12]确定在 SIMPLER 中进行的试验次数。

表 I 和表 II 对比了我们的方法与现有视觉-语言-动作方法在 Google 机器人上的表现，涵盖四项任务。我们的模型表现最佳，平均成功率在 VM 和 VA 设置下分别达到 83% 和 81%。值得注意的是，在需要特定任务操作的“打开抽屉并放入苹果”任务中，我们取得了显著提升。此外，在 VA 设置下，我们的成功率明显高于其他方法，表明 DAM-VLA 能够减轻动态环境中的性能下降。

表 III 报告了在 VM 设置下，我们的方法与现有视觉-语言-动作方法在 SIMPLER 中的 WidowX机器人上执行四项任务时的成功率。我们的模型取得了最高的平均成功率，达到 71%，大幅领先于其他方法。尤其是在“将勺子放在毛巾上”和“将胡萝卜放在盘子上”任务中，DAM-VLA 表现出显著提升，这些任务涉及更多样的物体姿态变化。

除了在 SIMPLER 上进行评估外，我们还评估了我们的方法在

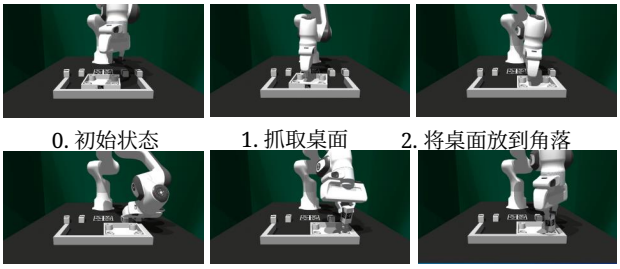

图 5：FurnitureBench 环境中“单腿”装配任务的完整过程。

方法 / FurnitureBench	各步骤的成功率				
	1	2	3	4	5
OpenVLA [4]	96%	94%	78%	53%	29%
CogACT [12]	98%	96%	96%	56%	42%
DAM-VLA（本文）	**100%**	**100%**	**100%**	**62%**	**56%**

表 IV：FurnitureBench 中“单腿”装配任务各步骤的成功率，并与现有的 OpenVLA 和 CogACT 方法进行了比较。第五步表示该任务的最终成功率。

Furni- tureBench 的 “One-Leg” 组装任务上的表现，该任务涉及高接触、长时域操作。完整的任务流程如图 5 所示。我们在随机设置家具初始位置的条件下进行了 50 次评估试验。表 IV 展示了与 OpenVLA 和 Cog ACT 模型相比，“One-Leg” 组装任务各步骤的成功率。我们的方法在这项具有挑战性的家具组装任务中始终优于以往的视觉-语言-动作方法，展现出在高接触操作场景下更强的泛化能力和精确度。具体而言，在“抓取桌面”、“将桌面放到角落”和“拿起桌腿”前三个步骤中，我们的成功率达到 100%；在高接触操作“拧紧桌腿”步骤中，我们也取得了更高的成功率。

C. 真实场景评估

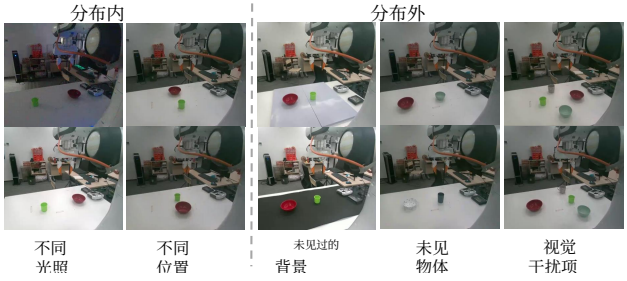

图 6：评估涵盖分布内场景和分布外场景。分布内设置包括与训练数据一致的物体位置和光照条件变化，而分布外设置则引入了训练期间未出现的新背景、未见物体和视觉干扰项。

我们的真实世界数据集是在多种物体摆放方式和光照条件下采集的。为评估鲁棒性，我们将评估分为分布内和分布外场景，如图 6 所示。分布

方法 / 真实世界	不同场景下的成功率 分布内	分布外	平均值
CogACT	65.7%	60.0%	62.9%
DAM-VLA（我们的方法）	**91.4%**	**82.2%**	**86.8%**

表 V：真实世界评估中，我们的方法与 CogACT 在抓取与放置任务上的成功率。ID：分布内，OOD：分布外。

组件					成功率			
VT	ACW	TW	DAM	DL	Google (VM)	Google (VA)	WidowX (VM)	Avg
-	-	-	-	-	64%	61%	50%	58%
✓	✓	-	-	-	78%	68%	53%	66%
✓	✓	✓	-	-	76%	63%	51%	63%
✓	✓	✓	✓	-	82%	72%	43%	66%
✓	✓	✓	✓	✓	83%	**81%**	**71%**	**78%**
-	✓	✓	✓	✓	**84%**	75%	58%	73%
-	-	-	✓	✓	66%	64%	49%	60%

表 VI：在 SIMPLER环境的视觉匹配（VM）设置下，对 WidowX机器人进行消融研究；并在视觉匹配（VM）和变体聚合（VA）设置下，对 Google 机器人进行消融研究。

内场景包含的物体位置和光照变化均与训练分布一致。相比之下，分布外场景引入了训练期间未出现过的背景、新颖物体和视觉干扰项。

表 V 报告了在真实世界抓取与放置任务中，DAM-VLA 与基线 CogACT 的成功率。为确保比较公平，每种方法在 16 种不同条件下均针对每种条件进行 5 次试验，涵盖分布内和分布外场景，总计 80 次试验。结果表明，DAM-VLA 始终优于 CogACT，在两种评估设置下都取得了更高的成功率。

D. 消融研究

在消融实验中，我们采用 SIMPLER 评估环境，在 VM 设置下使用 WidowX机器人，并在 VM 设置和变体聚合设置下使用 Google 机器人。为评估本方法中不同组件的贡献，我们对其相应成功率进行了详细分析。如表 VI 所示，所考察的主要组件包括：（1） **视觉词元（视觉标记）**：由视觉模型输出的视觉类别词元和寄存器token，供动作模型用于预测动作。（2）**动作块权重（ACW）**：用于监督 DAM-VLA 的动作块权重。（3）**轨迹权重（轨迹权重）**：用于监督 DAM-VLA 的轨迹权重。（4） **动态动作模型（DAM）**：由手臂运动模型和夹爪操作模型组成。（5） **双潜变量（双潜变量）**：从大语言模型的不同 Transformer 层提取的推理潜变量和认知潜变量。从仅包含单一动作模型的基线视觉-语言-动作模型出发，加入视觉标记和 ACW 可提升 Google 机器人和 WidowX机器人上的性能。由于当前视觉-语言-动作仅包含单一扩散动作模型，此处的视觉标记仅包含类别词元。引入专为双动作模型设计的轨迹权重后，两款机器人上的性能均略有下降，这是意料之中的。随后加入 DAM 和双潜变量，取得了 78% 的最高平均成功率，证明了我们 DAM-VLA 框架的有效性。当视觉-语言-动作配置为双动作模型时，视觉标记同时包含类别词元和寄存器token。最后，从 DAM-VLA 中移除视觉标记以及双尺度动作加权机制（ACW 和轨迹权重）会导致性能大幅下降，这进一步凸显了动态动作模型与双尺度动作加权机制的关键作用。值得注意的是，移除视觉标记意味着同时移除类别词元和寄存器token。

V. 结论

我们提出的 DAM-VLA 方法将视觉语言模型的内在推理能力与专为机械臂运动和夹爪操作设计的专用扩散动作模型进行动态整合。大量实验表明，我们的方法不仅显著优于现有视觉-语言-动作方法，而且在动态环境中的特定任务和通用操作场景下均能取得稳定表现。DAM-VLA 根据视觉语言模型引导的线索，在专用动作模型之间动态路由，为将语义理解融入具身决策开辟了新路径。该方法在多种任务和动态环境中展现出的泛化能力与稳定性，凸显了其作为下一代适应性机器人系统基础框架、应用于真实世界场景的潜力。

尽管 DAM-VLA 已取得出色的性能，但仍有几个方面有待进一步改进。首先，尽管我们的实验评估已经涵盖了接触密集型和长时域任务，但目前主要集中于抓取放置和家具组装；扩展到更多样化的任务类别将提供更广泛的验证。其次，尽管 DAM-VLA 显著提升了大多数任务的成功率，但其在“将绿色积木叠放在黄色积木上”场景中的提升相对有限，这表明精细协调和稳定性仍有增强空间。最后，DAM-VLA 目前仅在两个动作模型之间进行路由；将这一机制推广到能够处理多种动作类型并利用更丰富路由信号的场景中，将进一步拓展其在复杂现实场景中的适用性。

参考文献

[1] T. Z. Zhao、V. Kumar、S. Levine 和 C. Finn，《使用低成本硬件学习精细双手操作》，*arXiv*预印本 *arXiv:2304.13705*，2023年。[2] C. Chi、Z. Xu、S. Feng、E. Cousineau、Y. Du、B. Burchfiel、R. Tedrake 和 S. Song，《扩散策略：通过动作扩散进行视觉运动策略学习》，《国际机器人研究杂志》，第 02783649241273668 页，2023年。[3] A. Brohan、N. Brown、J. Carbajal、Y. Chebotar、X. Chen、K. Choromanski、T. Ding、D. Driess、A. Dubey、C. Finn 等，《RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制》，*arXiv*预印本 *arXiv:2307.15818*，2023年。[4] M. J. Kim、K. Pertsch、S. Karamcheti、T. Xiao、A. Balakrishna、S. Nair、R. Rafailov、E. Foster、G. Lam、P. Sanketi 等，《OpenVLA：开源视觉-语言-动作模型》，*arXiv*预印本 *arXiv:2406.09246*，2024年。[5] S. Belkhale、T. Ding、T. Xiao、P. Sermanet、Q. Vuong、J. Tompson、Y. Chebotar、D. Dwibedi 和 D. Sadigh，《RT-H：运用语言构建动作层级》，*arXiv*预印本 *arXiv:2403.01823*，2024年。

[6] S. Nasiriany、S. Kirmani、T. Ding、L. Smith、Y. Zhu、D. Driess、D. Sadigh 和 T. Xiao，《RT-Affordance：可供性是机器人操作的多功能中间表征》，arXiv预印本 arXiv:2411.02704，2024年。[7] M. Zawalski、W. Chen、K. Pertsch、O. Mees、C. Finn 和 S. Levine，《通过具身思维链推理实现机器人控制》，arXiv预印本 arXiv:2407.08693，2024年。[8] K. Black、N. Brown、D. Driess、A. Esmail、M. Equi、C. Finn、N. Fusai、L. Groom、K. Hausman、B. Ichter 等，《π0：用于通用机器人控制的视觉-语言-动作流模型，2024》，网址 https://arxiv.org/abs/2410.24164。[9] J. Wen、Y. Zhu、J. Li、M. Zhu、Z. Tang、K. Wu、Z. Xu、N. Liu、R. Cheng、C. Shen 等，《TinyVLA：迈向快速、数据高效的机器人操作视觉-语言-动作模型》，IEEE《机器人与自动化快报》，2025年。[10] S. Liu、L. Wu、B. Li、H. Tan、H. Chen、Z. Wang、K. Xu、H. Su 和 J. Zhu，《RDT-1B：用于双手操作的扩散基础模型》，arXiv预印本 arXiv:2410.07864，2024年。[11] Q. Bu、H. Li、L. Chen、J. Cai、J. Zeng、H. Cui、M. Yao 和 Y. Qiao，《迈向协同、通用且高效的机器人操作双系统》，arXiv预印本 arXiv:2410.08001，2024年。[12] Q. Li、Y. Liang、Z. Wang、L. Luo、X. Chen、M. Liao、F. Wei、Y. Deng、S. Xu、Y. Zhang 等，《CogACT：用于协同机器人操作中认知与行动的基础视觉-语言-动作模型》，arXiv预印本 arXiv:2411.19650，2024年。[13] J. Liu、H. Chen、P. An、Z. Liu、R. Zhang、C. Gu、X. Li、Z. Guo、S. Chen、M. Liu 等，《HybridVLA：在统一的视觉-语言-动作模型中融合扩散与自回归》，arXiv预印本 arXiv:2503.10631，2025年。[14] X. Li、K. Hsu、J. Gu、K. Pertsch、O. Mees、H. R. Walke、C. Fu、I. Lunawat、I. Sieh、S. Kirmani 等，《在仿真中评估真实世界机器人操作策略》，arXiv预印本 arXiv:2405 05941，2024年。[15] M. Heo、Y. Lee、D. Lee 和 J. J. Lim，《FurnitureBench：用于长时程复杂操作的可复现真实世界基准》，《国际机器人研究杂志》，第 02783649241304789 页，2023年。[16] J. Achiam、S. Adler、S. Agarwal、L. Ahmad、I. Akkaya、F. L. Aleman、D. Almeida、J. Altenschmidt、S. Altman、S. Anadkat 等，《GPT-4 技术报告》，arXiv预印本 arXiv:2303.08774，2023年。[17] H. Touvron、T. Lavril、G. Izacard、X. Martinet、M.-A. Lachaux、T. Lacroix、B. Roz`ière、N. Goyal、E. Hambro、F. Azhar 等，《Llama：开放且高效的基础语言模型》，arXiv预印本 arXiv:2302.13971，2023年。[18] H. Touvron、L. Martin、K. Stone、P. Albert、A. Almahairi、Y. Babaei、N. Bashlykov、S. Batra、P. Bhargava、S. Bhosale 等，《Llama 2：开放式基础模型与微调聊天模型》，arXiv预印本 arXiv:2307.09288，2023年。[19] J.-B. Alayrac、J. Donahue、P. Luc、A. Miech、I. Barr、Y. Hasson、K. Lenc、A. Mensch、K. Millican、M. Reynolds 等，《Flamingo：用于少样本学习的视觉语言模型》，神经信息处理系统进展，第 35 卷，第 23 716–23 736 页，2022年。[20] G. Team、R. Anil、S. Borgeaud、J.-B. Alayrac、J. Yu、R. Soricut、J. Schalkwyk、A. M. Dai、A. Hauth、K. Millican 等，《Gemini：一系列能力强大的多模态模型》，arXiv预印本 arXiv:2312.11805，2023年。[21] J. Bai、S. Bai、S. Yang、S. Wang、S. Tan、P. Wang、J. Lin、C. Zhou 和 J. Zhou，《Qwen-VL：具有多样能力的前沿大型视觉-语言模型》，arXiv预印本 arXiv:2308 3，2023 12966，1，2，. 卷号 期号 页码。[22] J. Li、D. Li、S. Savarese 和 S. Hoi，《BLIP-2：利用冻结图像编码器和大型语言模型引导语言-图像预训练》，载于《国际机器学习大会》。PMLR，2023年，第 19 730–19 742 页。[23] H. Liu、C. Li、Q. Wu 和 Y. J. Lee，《视觉指令微调》，神经信息处理系统进展，第 36 卷，第 34 892–34 916 页，2023年。[24] X. Chen、J. Djolonga、P. Padlewski、B. Mustafa、S. Changpinyo、J. Wu、C. R. Ruiz、S. Goodman、X. Wang、Y. Tay 等，《PaLI-X：扩展多语言视觉与语言模型》，arXiv预印本 arXiv:2305.18565，2023年。[25] S. Karamcheti、S. Nair、A. Balakrishna、P. Liang、T. Kollar 和 D. Sadigh，《Prismatic视觉语言模型：探索视觉条件语言模型的设计空间》，载于第41届国际机器学习大会，2024年。[26] J. Ho、A. Jain 和 P. Abbeel，“去噪扩散概率模型”，神经信息处理系统进展，第33卷，第6840–6851页，2020年。[27] W. Peebles 和 S. Xie，“基于变换器的可扩展扩散模型”，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第4195–4205页。[28] K. Black、M. Nakamoto、P. Atreya、H. Walke、C. Finn、A. Kumar 和 S. Levine，“利用预训练图像编辑扩散模型实现零样本机器人操作”，arXiv预印本 arXiv:2310.10639，2023年。[29] S. Dasari、O. Mees、S. Zhao、M. K. Srirama 和 S. Levine，“机器人扩散变换器的构成要素”，arXiv预印本 arXiv:2410.10088，2024年。[30] A. Z. Ren、J. Lidard、L. L. Ankile、A. Simeonov、P. Agrawal、A. Majumdar、B. Burchfiel、H. Dai 和 M. Simchowitz，“扩散策略策略优化”，arXiv预印本 arXiv:2409.00588，2024年。[31] M. Reuss，¨O. E. Ya˘gmurlu、F. Wenzel 和 R. Lioutikov，“多模态扩散变换器：从多模态目标中学习多样化行为”，arXiv预印本 arXiv:2407.05996，2024年。[32] M. Uehara、Y. Zhao、K. Black、E. Hajiramezanali、G. Scalia、N. L. Diamant、A. M. Tseng、T. Biancalani 和 S. Levine，“将连续时间扩散模型微调为熵正则化控制”，arXiv预印本 arXiv:2402.15194，2024年。[33] M. Uehara、Y. Zhao、K. Black、E. Hajiramezanali、G. Scalia、N. L. Diamant、A. M. Tseng、S. Levine 和 T. Biancalani，“反馈高效的扩散模型在线微调”，arXiv预印本 arXiv:2402.16359，2024年。[34] T. Z. Zhao、J. Tompson、D. Driess、P. Florence、K. Ghasemipour、C. Finn 和 A. Wahid，“Aloha unleashed：机器人灵巧性的一种简单方案”，arXiv预印本 arXiv:2410.13126，2024年。[35] T.-W. Ke、N. Gkanatsios 和 K. Fragkiadaki，“3D扩散执行器：基于3D场景表示的策略扩散”，arXiv预印本 arXiv:2402.10885，2024年。[36] X. Jia、Q. Wang、A. Donat、B. Xing、G. Li、H. Zhou、O. Celik、D. Blessing、R. Lioutikov 和 G. Neumann，“Mail：利用选择性状态空间模型改进模仿学习”，载于第8届机器人学习年会，2024年。[37] O. M. Team、D. Ghosh、H. Walke、K. Pertsch、K. Black、O. Mees、S. Dasari、J. Hejna、T. Kreiman、C. Xu 等，“Octo：一种开源通用机器人策略”，arXiv预印本 arXiv:2405.12213，2024年。[38] J. Wen、M. Zhu、Y. Zhu、Z. Tang、J. Li、Z. Zhou、C. Li、X. Liu、Y. Peng、C. Shen 等，“Diffusion-vla：通过统一扩散与自回归扩展机器人基础模型”，arXiv预印本 arXiv:2412.03293，2024年。[39] M. Oquab、T. Darcet、T. Moutakanni、H. Vo、M. Szafraniec、V. Khalidov、P. Fernandez、D. Haziza、F. Massa、A. El-Nouby 等，“DINOv2：无监督学习稳健视觉特征”，arXiv预印本 arXiv:2304.07193，2023年。[40] X. Zhai、B. Mustafa、A. Kolesnikov 和 L. Beyer，“用于语言图像预训练的Sigmoid损失”，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第11 975–11 986页。[41] A. O’Neill、A. Rehman、A. Maddukuri、A. Gupta、A. Padalkar、A. Lee、A. Pooley、A. Gupta、A. Mandlekar、A. Jain 等，“Open X-Embodiment：机器人学习数据集与RT-X模型：Open X-Embodiment协作0”，载于《2024 IEEE国际机器人与自动化会议（ICRA）》。IEEE，2024年，第6892–6903页。[42] A. Brohan、N. Brown、J. Carbajal、Y. Chebotar、J. Dabis、C. Finn、K. Gopalakrishnan、K. Hausman、A. Herzog、J. Hsu 等，“RT-1：面向大规模现实世界控制的机器人变换器”，arXiv预印本 arXiv:2212.06817，2022年。[43] H. R. Walke、K. Black、T. Z. Zhao、Q. Vuong、C. Zheng、P. Hansen-Estruch、A. W. He、V. Myers、M. J. Kim、M. Du 等，“BridgeData v2：用于大规模机器人学习的数据集”，载于《机器人学习会议》。PMLR，2023年，第1723–1736页。[44] Q. Vuong、S. Levine、H. R. Walke、K. Pertsch、A. Singh、R. Doshi、C. Xu、J. Luo、L. Tan、D. Shah 等，“Open X-Embodiment：机器人学习数据集与RT-X模型”，载于《通用机器人：面向可扩展技能获取的学习范式 CoRL2023》，2023年。[45] X. Li、P. Li、M. Liu、D. Wang、J. Liu、B. Kang、X. Ma、T. Kong、H. Zhang 和 H. Liu，“迈向通用机器人策略：构建视觉-语言-动作模型时哪些因素至关重要”，arXiv预印本 arXiv:2412.14058，2024年。[46] D. Qu、H. Song、Q. Chen、Y. Yao、X. Ye、Y. Ding、Z. Wang、J. Gu、B. Zhao、D. Wang 等，“SpatialVLA：探索视觉-语言-动作模型的空间表示”，arXiv预印本 arXiv:2501 15830，2025年。