DiffGR00T：面向人形机器人全身视觉-语言-动作模型的执行感知可微训练

Sichen Wu、Qian Sun、Mengfan Liu、Weikai Qin、Bike Zhu 和 Hesheng Wang†上海交通大学自动化系，上海 200240，中国

arXiv:2611.32665v1 [cs.RO] 24 Nov 2026



图 1：**DiffGR00T 概览**。在 SONIC 和世界模型保持冻结的情况下更新 VLA动作头；训练期间，世界模型取代模拟器。仅在仿真中，平均任务成功率从 59.2% 提高至 71.7%。

摘要—**视觉-语言-动作模型在以手臂为中心的操作中取得了出色成果，但人形机器人全身控制（WBC）需要动态平衡和紧密耦合的全身协调。现有的分层人形机器人视觉-语言-动作模型通常将运动生成与物理执行分离，导致跟踪和平衡误差无法直接用于优化视觉-语言-动作模型生成器。我们提出 DiffGR00T，这是一种端到端可微训练框架，将 GR00T N1.7 与预训练全身行为基础模型 SONIC 相结合。为实现执行感知可微训练，我们用一个学习得到的动作条件世界模型取代不可微模拟器，使多步身体执行损失能够通过冻结的全身控制器和流匹配采样器反向传播，直接传递至 VLA动作头。在四项 HumanoidArena 仿真任务中，与监督式基线视觉-语言-动作模型相比，DiffGR00T 将平均任务成功率提高了 12.5 个百分点（59.2% → 71.7%），这表明执行感知梯度能够有效地使高层运动生成与闭环全身动力学相协调。**

I. 引言

视觉-语言-动作模型（VLA）在以手臂为中心的操作领域取得了显著进展，其预测的末端执行器位姿或关节空间目标可通过低级伺服控制器执行 [1], [2]。然而，将这些能力扩展到人形机器人，会带来本质不同的控制挑战。与固定基座机械臂不同，人形机器人具有欠驱动的浮动基座动力学，在不断变化的

†通讯作者：Hesheng Wang（电子邮箱：wanghesheng@sjtu.edu.cn）。

接触和平衡约束下，必须协调运动、操作与姿态调节。直接预测关节目标并不能保证全身执行的稳定性，而在大型视觉-语言-动作模型内部直接学习高频稳定控制仍然困难。因此，许多人形机器人系统采用分层架构：由视觉-语言-动作模型生成高层运动意图，再由预训练的行为基础模型（BFM）或全身控制器通过本体感觉反馈来实现这些意图。

这种分层范式推动了人形机器人运动与操作领域的显著进展。SONIC 通过共享潜在接口构建了一个具有泛化能力的全身运动基础模型，而 WholeBodyVLA 和 OpenHLM 等系统则将任务条件动作生成与预训练运动控制器相结合 [3]–[5]。由于这些架构分别训练 VLA 来输出运动参考、训练控制器来跟踪这些参考，控制器在部署时能够提供反应式全身稳定控制，但仍存在动作-执行训练差距：跟踪误差和平衡误差无法沿直接梯度路径反馈到 VLA 的运动生成目标函数中。要通过端到端可微训练弥合这一差距，就必须在策略-控制器接口上建立可微梯度通路。

弥合这一差距并非易事，因为稳健的全身控制与可微优化传统上遵循不同的训练范式。人形机器人运动跟踪器通常采用无模型强化学习进行训练，其物理执行反馈以标量奖励的形式提供，而非可微状态转移。此外，富接触人形机器人动力学涉及非光滑的支撑切换，因此通过传统刚体模拟器直接求导在数值上容易不稳定，计算开销也很大。尽管 DiffMimic 等可微物理方法已探索解析运动模仿梯度 [6]，但将此类优化扩展到高维全身控制和大型 VLA 生成器，计算开销仍高得难以承受。

一种替代通过模拟器求导的方法，是通过学习到的世界模型求导。动作条件神经世界模型通过拟合从模拟器或机器人收集的状态转移数据而得到，处处可微——包括跨越接触转换时也可微——因此，无须调用接触求解器，就能通过预测状态对执行目标进行反向传播，同时利用较短的展开窗口限制预测误差的累积。SuperMimic [7] 展示了这一范式在人形机器人跟踪中的应用：在学习到的世界模型上，通过短时域时间反向传播（BPTT）训练跟踪策略；在一项速度跟踪实验中，还将梯度进一步经由共享解码器传递到轻量级目标编码器。然而，将这一执行梯度路径扩展至穿过冻结的令牌条件控制器和多步流匹配采样器，并应用于人形机器人视觉-语言-动作模型，仍鲜有探索。

我们提出**DiffGR00T**，一种端到端可微的训练框架，将 GR00T N1.7 与预训练的 SONIC 行为基础模型相结合。我们首先在携带已记录 SONIC 运动标记的 HumanoidArena 演示数据上微调 GR00T N1.7 动作头 [8],[9]，使 VLA 与 SONIC 的潜在运动接口对齐。在执行感知可微训练期间，学习得到的动作条件世界模型提供多步状态转移，以评估身体执行损失，使执行梯度能够通过世界模型、冻结的 SONIC 解码器和流匹配采样器，直接反向传播至 VLA动作头。

我们的方法利用了视觉-语言-动作模型动作块的时间结构。流匹配视觉-语言-动作模型生成长度为 40 步的运动令牌序列，其中前 20 个令牌在闭环中执行。在每个 50Hz 控制步中，冻结的 SONIC解码器将当前令牌和本体感知历史映射为 29 个关节目标，而世界模型则预测由此产生的机器人状态，并将其作为本体感知反馈给下一步。随后，多步身体执行损失通过一个十步展开窗口、冻结的解码器以及四步流匹配采样器进行反向传播，只更新视觉-语言-动作模型的动作头（通过 LoRA 适配器和输出投影）；世界模型和仅用于训练的评论家网络则始终保持固定。在 HumanoidArena 的四项仿真任务中，与监督训练的基线视觉-语言-动作模型（59.2% $\rightarrow$ 71.7%）相比，DiffGR00T 将平均任务成功率提高了 12.5 个百分点。

我们的主要贡献如下：

1) **面向全身执行的可微世界模型**。我们利用控制器激励回放和少量闭环策略数据，为一台29自由度人形机器人学习动作条件化的动力学模型，并在执行感知可微训练期间固定其参数，使其在短时展开中充当模拟器的可微替代模型。

2) **将执行梯度传入流匹配视觉-语言-动作模型**。我们通过学习到的世界模型、冻结的 SONIC 解码器和四步流匹配采样器反向传播多步身体执行损失，以更新 VLA动作头，同时固定视觉-语言主干网络、控制器、世界模型和仅用于训练的评论家网络。

3) **执行对齐监督**。我们确定了实现有效人形机器人执行反馈所需的关键监督项——根部位置和漂移，用于抑制每个窗口内的漂移；针对任务阶段变化的过渡帧加权；以及按速度覆盖度进行窗口采样，以确保运动在梯度预算中得到保留，同时校准各目标，使其保持尺度一致。

II. 相关工作

A. 流匹配策略与策略改进

π_0 将预训练视觉-语言骨干网络与流匹配动作专家相结合，以生成连续的

动作块 [1]。基于这一范式，近期研究探索了用于监督式策略学习的替代梯度路径：DIAL通过预测的视觉意图连接视觉语言模型（VLM）与反应式动作专家，并在联合微调过程中，将监督式流匹配损失经由这一潜在意图反向传播至可训练的VLM组件 [2]；VITA是一种将图像编码映射为潜在动作的视觉运动策略，它直接通过流匹配常微分方程（ODE）求解器的欧拉步，对监督式动作重建损失求导 [10]。

另一条互补的研究路线是将机器人执行经验融入流匹配策略更新。RECAP 对执行展开拟合价值函数，并依据价值函数导出的优势指标来条件化监督式视觉-语言-动作模型更新 [11], ；而 Q-VGM 则从展开奖励中学习动作块评论家网络，并将评论家网络相对于去噪动作块的梯度转换为局部速度目标，从而避免沿着完整去噪轨迹进行反向传播 [12]。尽管这些方法利用了常微分方程微分、优势条件化或价值引导，但都未将多步身体执行目标经由全身控制器和学习到的动作条件动力学传递至流匹配生成器。

B. 人形机器人全身动作生成与控制

人形机器人运动与操作系统在高层动作生成与底层控制之间采用了多种接口。SONIC 利用动作捕捉数据训练出具有泛化能力的全身运动跟踪器，并通过共享的动作标记接口展示了由 GR00T N1.5 驱动的视觉-语言-动作模型运动与操作 [3]。HumanoidArena 对分层策略与跟踪器的组合进行了基准测试，并介绍了一种直接预测模式：视觉-语言-动作模型输出 64 维 SONIC 身体潜变量 [9], ，本研究采用的正是这一接口和基准设置。其他分层框架则采用不同的中间表征：WholeBodyVLA 从无动作的第一人称视频中学习运动和操作潜变量，并将其解码为上半身关节目标；与此同时，强化学习控制器执行运动指令[4],; OpenHLM 则将关节空间遥操作与异构数据协同训练相结合，生成由独立底层策略跟踪的全身关节参考值 [5]。

除单阶段参考生成之外，HAF-VLA 还通过分阶段流匹配与跨阶段条件生成累积式全身指令，并由现有强化学习控制器执行运动和腰部指令 [13]。不同于显式参考跟踪，SENTINEL 将经过模拟器训练的全身控制器克隆为一种语言与本体感知策略，直接预测低层动作目标；随后冻结该策略，并使用 PPO 训练残差动作头[14]。在这些接口——共享运动令牌、关节空间参考、分阶段指令以及直接的低层目标——中，执行偏差由下游跟踪器或独立的残差动作头吸收，而不是通过控制器反向传播来更新高层视觉-语言-动作模型生成器。

C. 可微全身控制与执行感知学习

将物理执行梯度直接传递到策略学习中，这一方法已从可微物理发展到学习得到的动力学代理。DiffMimic 通过在可微刚体模拟器中对模仿损失进行微分，训练人形机器人运动跟踪策略 [6]。为避免对接触求解器进行微分，ControlVAE 通过冻结的潜在技能控制器和学习到的世界模型反向传播任务损失，从而优化模拟角色的任务策略 [15]。MoConVQ 同样使用学习到的世界模型来训练离散运动表示及其解码器，同时通过预测词元单独训练其文本条件生成器[16]。

与我们的人形机器人设置最接近的是 SuperMimic 和 HWAM。SuperMimic 根据模拟器状态转移拟合可微世界模型，并通过短预测时域的想象展开来训练人形机器人跟踪策略；在下游速度跟踪实验中，执行梯度也会经过世界模型和共享解码器（该分支上的解码器权重被冻结），传递到速度目标编码器 [7]。HWAM 通过联合建模运动学参考与对齐的实测执行后状态来纳入物理执行，并结合辅助目标函数：根据关节轨迹预测未来视频，以及根据视觉状态转移重建关节轨迹[17]。

相比之下，DiffGR00T 在流匹配人形机器人视觉-语言-动作模型上进行执行感知的可微训练：多步身体执行损失通过学习得到的动作条件动力学的短程展开、冻结的 SONIC 解码器以及四步流匹配采样器反向传播，仅更新视觉-语言-动作模型的动作头（通过 LoRA 适配器和输出投影），而视觉语言模型骨干网络和控制器均保持固定。我们从三个方面评估这一训练路径：流头的梯度连通性、与模拟器回报的方向一致性，以及仿真中的闭环任务表现。

III. 方法

A. 总体架构与控制接口

如图 2 所示，该框架由三个相互耦合的组件构成：一个流匹配视觉-语言-动作模型、一个冻结的全身基础模型 SONIC，以及一个动作条件可微世界模型。视觉-语言-动作模型将自我中心观测、语言指令和机器人状态映射为一个运动令牌块；身体基础模型将每个令牌与本体感知历史一起解码为关节级指令；世界模型则预测机器人状态在这些指令作用下如何演变。

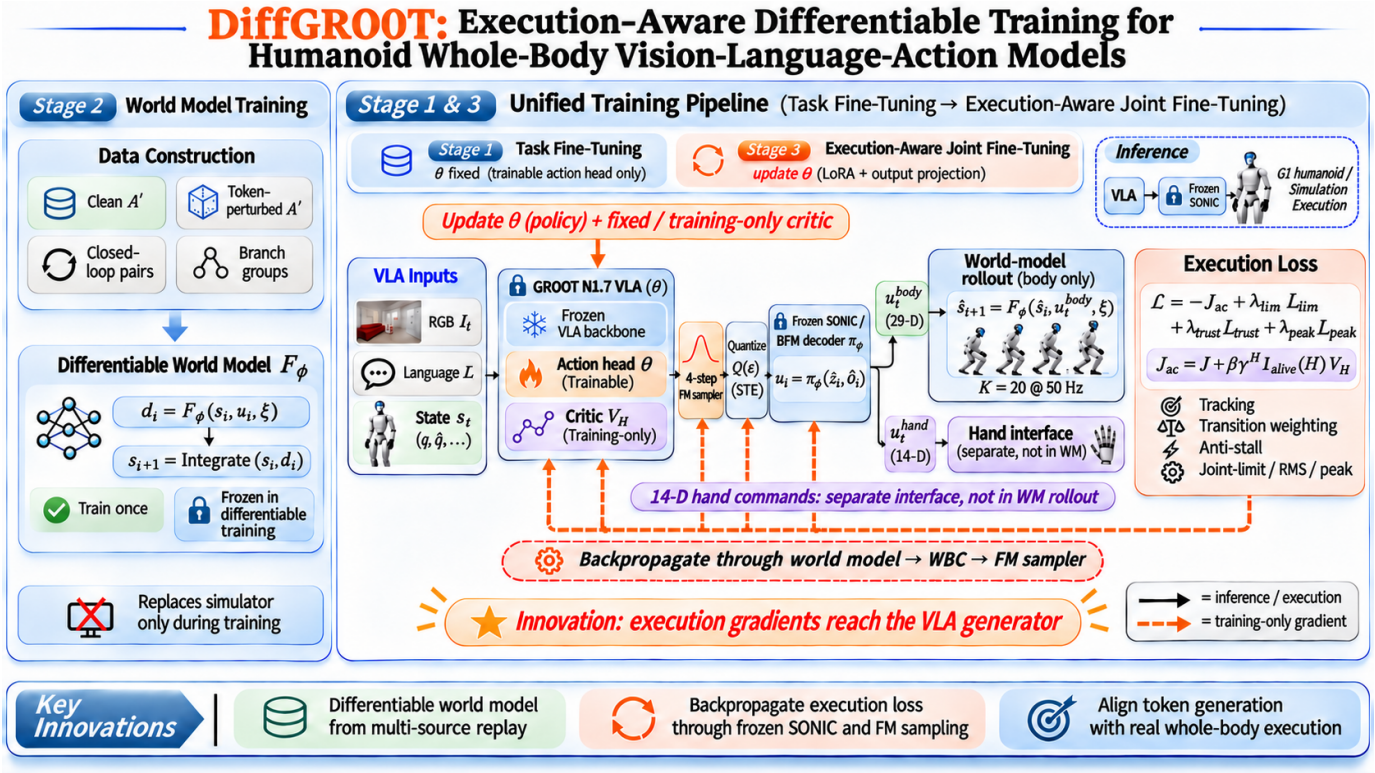


图 2：**DiffGR00T架构**。训练的是 GR00T N1.7 的视觉-语言-动作模型动作头；视觉-语言-动作模型骨干网络、SONIC解码器、动作条件世界模型和仅用于训练的评论家网络均被冻结。执行损失通过世界模型、SONIC解码器和四步流匹配采样器反向传播。评论家网络提供固定的尾部价值，不属于部署控制器。世界模型基于多源回放训练一次，仅在训练期间替代模拟器；推理时不会使用它。

所有任务共享同一架构，并有两种截然不同的执行模式。在部署时，只运行视觉-语言-动作模型、冻结的 SONIC 解码器和手部接口：视觉-语言-动作模型生成一个动作块，SONIC 将其身体令牌和本体感知历史解码为 29 维关节指令，而独立的手部接口则执行 14 条手指指令。此时不运行世界模型和仅用于训练的评论家网络，因此学习得到的代理不会增加推理开销。

训练期间，模拟器由动作条件世界模型替代。对于身体分支，同一个块会经过冻结的 SONIC 解码器和世界模型，随后产生的执行成本会通过世界模型、冻结的解码器和四步流匹配采样器反向传播至视觉-语言-动作模型动作头。由于冻结模块仍保留在反向传播计算图中，其参数不会改变，但其雅可比矩阵会改变。这种冻结但可微的路径，是该架构与标准视觉-语言-动作模型微调之间的核心区别。世界模型基于多源回放训练一次，绝不会部署；评论家网络仅用于训练，用于概括展开的尾部。

视觉-语言-动作模型通过四步流匹配采样器生成一个包含 40 步的块。在时间 t，该块

$$A_t = [Z_t, H_t] \in \mathbb{R}^{40\times 78},$$

其中身体部分 $Z_t \in \mathbb{R}^{40\times 64}$ 收集与 SONIC 兼容的运动令牌 $z_{t+i|t} \in \mathbb{R}^{64}$，用于 $i = 0, \ldots, 39$；手部部分 $H_t \in \mathbb{R}^{40\times 14}$ 则为每只手的每个时间步收集七条手指指令。因此，每个时间步都是一个 78 维向量：64 维身体令牌加 14 维手指指令。

对于每个身体时间步 $j = t + i$，冻结的身体基础模型 π_ψ 接收令牌 z_j 和包含十帧的本体感觉历史 $O_j \in \mathbb{R}^{10\times 93}$，并输出一个 29 维的归一化身体动作

$$u_j = \pi_\psi(z_j, O_j). \tag{1}$$

u_j 中的 29 个条目与 29 个身体关节一一对应；它们是归一化动作坐标，尚不是关节角度。O_j 中的每个历史帧都堆叠了相对于默认配置的 29 个关节位置偏移、29 个关节速度、3 个身体角速度分量、3 个投影重力分量，以及 29 维的上一时刻动作（$29 + 29 + 3 + 3 + 29 = 93$）。

随后，将归一化动作转换为关节位置目标，

$$q_j^{\mathrm{cmd}} = q^{\mathrm{default}} + \boldsymbol{\alpha} \odot u_j, \tag{2}$$

其中，$q^{\mathrm{default}} \in \mathbb{R}^{29}$ 为标称关节构型，$\boldsymbol{\alpha} \in \mathbb{R}^{29}$ 为逐关节动作尺度，$\odot$ 为逐元素乘法；低级 PD 控制器跟踪 q_j^{cmd}。14 个手指指令被压缩为 HumanoidArena 的二值开合约定，并通过独立的手部接口执行，因此不会进入 29 维身体动作。

VLA 从 GR00T N1.7 初始化，并在包含已记录运动令牌的任务演示数据上进行微调，同时保持视觉-语言主干网络和 BFM 固定，从而仅训练动作头。该阶段生成任务条件化令牌块，但从未观察执行这些令牌所产生的物理后果；第 III-C 节将这一后果转化为梯度。

B. 可微世界模型

在训练过程中，模拟器被一个动作条件化世界模型 F_ϕ 所替代，该模型预测身体动作引起的状态转移。模型预测的是单步状态增量，而非绝对下一状态：绝对目标受机器人全局位姿主导，而 20 毫秒步长内的状态转移较小且近似平稳。因此，网络输出d_t，并通过积分恢复下一状态。令 $\mathcal{I}$表示积分算子，该算子将状态和预测的增量映射为下一状态。于是

$$\begin{aligned} d_t &= F_\phi(s_t, a_t, \xi) \in \mathbb{R}^{70}, \\ \hat{s}_{t+1} &= \mathcal{I}(s_t, d_t), \end{aligned} \tag{3}$$

其中，$s_t \in \mathbb{R}^{71}$ 表示根位姿（位置和单位四元数）、29 个关节位置、根线速度和角速度，以及 29 个关节速度。动作$a_t \in \mathbb{R}^{29}$ 是 BFM 生成的归一化身体动作，而 $\xi \in \mathbb{R}^{33}$ 是一个仿真上下文向量（摩擦力、编码器偏差和质心偏移），在评估时保持为标称值。状态为 71 维，增量为 70 维，这是因为方向在 s_t 中以四元数存储，但在 d_t 中预测为 3-D 旋转向量。

积分分为两个阶段。首先，速度增量生成下一个根部线速度、根部角速度和关节速度。其次，这些更新后的速度在 20 毫秒步长内推进根部位置、根部朝向和关节角度。预测残差增量 Δp、$\Delta\theta$、Δq作为局部修正，叠加在这次基于速度的更新之上。对于方向，三维旋转向量通过指数映射转换为旋转，并与当前方向组合。与根部相关的增量均在当前航向坐标系中表示，因此目标不依赖于机器人的全局位置或偏航角。因此，该增量较小

表 I：**世界模型数据混合**。阶段长度分别为 880 和 1650；第二阶段使用 $K = 20$ 自回归窗口。占比按批次计算。

来源	阶段 1	阶段 2	Role
干净 A′	0.15	0.20	名义动力学
令牌扰动的 A′	0.28	0.45	偏离轨迹状态
闭环配对数据	0.27	0.35	部署分布
分支组	0.30	0.00	动作对比

且处于坐标系局部范围内，而完整的下一个状态则通过积分恢复。

动作既通过多层感知机输入进入，也通过一个残差动作头进入，

$$\begin{aligned} F_\phi(s_t, a_t, \xi) = \mathrm{MLP}_\phi(s_t, a_t, \xi) \\ + G_\phi(s_t)\,(a_t - a_{\mathrm{ref}}), \end{aligned} \tag{4}$$

其中，a_{ref} 是参考动作。残差项为动作提供了直接的状态相关路径，可在状态相同但动作不同的分支组上进行监督。多层感知机保留了非线性动作依赖，而 G_ϕ则捕捉 a_{ref}附近的局部线性修正。 G_ϕ 采用零初始化，因此训练从多层感知机开始，并逐步学习残差修正。留出集检查证实，学到的动作依赖超过了噪声底。

模型使用四种状态转移来源进行训练（表 I）。我们将 A′ 记为从冻结的 BFM 收集的控制器激励回放语料库；干净的 A′ 和令牌扰动后的 A′ 是其中两个子集。第一种是干净的控制器激励回放：在仿真中使用参考令牌驱动冻结的 BFM，并记录由此产生的状态–动作–下一状态三元组，整个过程中不让视觉-语言-动作模型参与。它提供名义跟踪锚点和广泛的转移数据池。第二种是在相同的采集过程中加入时间相关的高斯令牌噪声，噪声幅度为 0.02、0.05 和 0.10，并在一个 20 步数据块内保持不变；由于 BFM 执行的是受扰动的令牌，存储的动作仍然是产生该状态转移的动作。这扩大了对偏离轨迹状态和恢复状态的覆盖度。每隔 28–37 步进行一次周期性重置，位移最大为 0.3 米，从而进一步增加恢复状态。我们扰动令牌而不是动作，以确保存储的动作仍与实际执行的状态转移相匹配。

第三种来源是在视觉-语言-动作模型参与闭环时收集的闭环配对数据。这些数据覆盖部署分布，包括接触和漂移。第四种是分支组：同一状态对应四种不同的令牌序列，为残差动作头和增益项提供同状态动作对比。参考运动库不是动力学来源；它仅用于定义参考坐标系。

训练分两个阶段进行。阶段 1 使用单步状态转移（$K = 1$）；阶段 2 使用 $K = 20$自回归窗口。对于从 s_0 开始的展开 $\hat{s}_0, \ldots, \hat{s}_K$，令 $R(q)$ 和 $H(q)$ 分别表示根四元数的旋转矩阵和航向坐标系。令 $SO(3)$ 表示旋转群，$\mathrm{Log}: SO(3) \to \mathbb{R}^3$ 表示矩阵对数，它将旋转矩阵映射为其旋转

向量。那么，$\|\mathrm{Log}(R(q)^\top R(\hat{q}))\|_2$ 就是预测方向与记录方向之间的测地角。定义分块状态误差

$$\mathrm{Err}(\hat{s}, s) = \begin{bmatrix} H(q)^\top(\hat{p}-p)/\sigma_p \\ \mathrm{Log}\big(R(q)^\top R(\hat{q})\big)/\sigma_R \\ (\hat{q}_{\mathrm{joint}}-q_{\mathrm{joint}})/\sigma_q \\ R(q)^\top(\hat{v}-v)/\sigma_v \\ R(q)^\top(\hat{\omega}-\omega)/\sigma_\omega \\ (\hat{\dot{q}}-\dot{q})/\sigma_{\dot{q}} \end{bmatrix}, \tag{5}$$

其中，$\hat{p}$、$\hat{q}$、$\hat{v}$、$\hat{\omega}$、$\hat{q}_{\mathrm{joint}}$、$\hat{\dot{q}}$ 分别为预测的位置、四元数、根部线速度、根部角速度、关节位置和关节速度。状态损失为

$$\mathcal{L}_{\mathrm{state}} = \frac{1}{K}\sum_{j=1}^{K} \|\mathrm{Err}(\hat{s}_j, s_j)\|_2^2, \tag{6}$$

其中，σ_p、σ_R、σ_q、σ_v、σ_ω、$\sigma_{\dot{q}}$ 分别为各块的归一化因子。方向块采用四元数测地误差，位置和速度则使用帧局部约定表示。这避免了在单一原始 L_2 范数中混合米、弧度和四元数组分量。

分支组会增加增益损失。对于包含 $K_b = 4$ 个分支（包括参考分支）的组，设 $\hat{s}^{(k)}$ 和 $s^{(k)}$ 分别为执行分支 k 后的预测状态和记录状态。增益损失使相对于第一分支的预测状态差与真实状态差相匹配：

$$\mathcal{L}_{\mathrm{gain}} = \mathbb{E}_{\mathcal{B}}\left[\frac{1}{K_b}\sum_{k=1}^{K_b-1}\left\|\frac{(\hat{s}^{(k)}-\hat{s}^{(0)})-(s^{(k)}-s^{(0)})}{\sigma_\Delta}\right\|_2^2\right], \tag{7}$$

其中$\mathcal{B}$ 表示分支组，σ_Δ 是位置、方向、关节和速度各块的缩放因子。增益损失基于记录状态坐标中的原始状态增量计算。在收集的分支组中，四元数符号与参考分支保持一致，因此在此设置下无需进行符号规范化。总目标函数为

$$\mathcal{L}_{\mathrm{WM}} = \mathcal{L}_{\mathrm{state}} + \lambda_g\,\mathcal{L}_{\mathrm{gain}}, \tag{8}$$

其中包含 $\lambda_g = 2.0$ 以及针对四个来源的指数移动平均自适应损失混合。分支占比仅用于阶段 1（表 I）。模型仅训练一次，只有在留出窗口上的多步误差没有退化时才予以接受，并在策略更新期间保持冻结。我们并不将其视为精确的模拟器；它只需为策略更新提供近似的执行梯度。

世界模型并未按照在线策略进行更新。刷新并非只需额外采集一次数据：随着策略不断变化，刷新必须反复进行，而且每一轮都需要在多个任务和多个回合中进行闭环 VLA 推理，随后重新训练世界模型并设置新的门控。因此，运行 300 次更新就需要进行多轮此类刷新，总成本远高于一小时的微调。反复刷新还会使目标函数变得非平稳，并可能降低动作效果校准的准确性；早期测试中的在线更新变体就出现了这种情况。因此，我们只训练一次世界模型，并保持其固定。闭环配对数据是固定的校准数据源，而非实时更新的数据流；固定的世界模型使闭环对比结果能够归因于策略更新。

表 II：**在线策略世界模型刷新成本**。一次刷新是一个重复进行的轮次，而不是一次性操作：每个轮次都需要跨多个任务的闭环 VLA 数据、世界模型重新训练以及新的门控。所有报告的数值均使用固定的世界模型。

	固定	在线策略
闭环 VLA 数据	–	多任务、多个回合
WM 重新训练	–	p每轮
WM 门控	–	每轮
策略更新	59–66 分钟	每轮
刷新轮次	0	多个
单次运行总计	≈1 h	≫1 h

C. 执行感知可微训练

可微训练保留了从块经过 BFM、世界模型和采样器的计算图，使得在想象执行上定义的目标函数能够传递至生成器。在更新过程中，除 VLA 的动作头外，其他部分均保持固定：BFM、SONIC 编码器、世界模型和仅用于训练的评论器均保留其参数。评论家网络提供固定的尾部价值，而策略是唯一会改变部署控制器的模块。

执行目标函数定义在一个 0.4 秒的片段上：在 50 Hz 下，策略执行一个包含 20 步的区块，而目标函数则在其中 10 步的可微前缀上进行评估。剩余的 10 步不进行逐步求导；相反，一个固定的、仅用于训练的评论器根据第 10 步的想象状态估计其尾部回报。该值会纳入损失，并且相对于想象状态仍保持可微。评论器在初始化后保持固定；策略训练期间不会更新它。参考轨迹是仅用于训练的局部片段参考，用于定义漂移目标。在部署时，视觉-语言-动作模型既不会接收参考轨迹，也不会接收评论器。因此，策略学习的是在每个片段内减少漂移，而不是让完整的闭环轨迹拟合全局参考。

策略起始状态仅从 A′ 片段中抽取。不使用参考库中的帧，因为它们是理想化的参考状态，而不是机器人实际到达过的状态，因此执行漂移为零，也无法覆盖用于纠正的偏离轨迹状态或恢复状态。闭环视觉-语言-动作模型轨迹也不纳入其中，因为它们无法为每个到达的帧提供对齐的局部参考片段，这意味着无法构造漂移目标。此外，它们的漂移会在整个回合中累积，无法在 0.4 秒的片段内纠正。因此，闭环数据用于训练世界模型，而不作为策略起始状态。

如果 A′ 中的某一帧曾被访问过、其相对于片段参考的根部偏移不超过 0.10 米、至少有 10 帧历史记录，且 20 步参考窗口未超出片段范围，则该帧可作为策略起始帧。我们仅使用训练片段，按照训练覆盖权重对符合条件的帧进行采样，并且只读取帧和状态，不读取动作。0.10 米的界限构成一个信赖域，因为 10 步窗口只能纠正几厘米的偏差；更大的偏移会表现为近乎恒定的误差。界限以外的帧仍用于训练世界模型，但不作为策略起始帧。

a) 前向传播：一次更新就是一个可微程序。它采样一个令牌块，通过冻结的 BFM 对其解码，展开世界模型，并在想象状态上评估执行成本。给定条件上下文 c_t 和噪声样本 ε，令牌块通过四个欧拉步生成，

$$z^{k+1} = z^k + \Delta\tau_k \, v_\theta(z^k, \tau_k, c_t), \qquad k = 0, \dots, 3, \quad (9)$$

其中 z^k 是第 k步后的块， v_θ 是流匹配速度场， $\Delta\tau_k$ 是步长。这四步生成策略所用的 40 步块。

随后，将身体通道量化到 BFM 使用的 $1/16$ 网格，

$$\tilde{z} = z + \mathrm{sg}\big(Q(z) - z\big), \quad (10)$$

其中 $Q(z)$ 表示网格量化，sg 表示停止梯度。前向路径使用与部署时相同的量化令牌，而反向路径则原样传递梯度。

从观测状态 $s_{t|t}$ 和历史记录 $O_{t|t}$出发，BFM 和世界模型在执行块上展开。记$\hat{s}_i \equiv \hat{s}_{t+i|t}$ 和 $O_i \equiv O_{t+i|t}$,

$$\begin{aligned} u_i &= \pi_\psi(\tilde{z}_i, \hat{O}_i), \\ d_i &= F_\phi(\hat{s}_i, u_i, \xi_{t+i}), \\ \hat{s}_{i+1} &= \mathcal{I}(\hat{s}_i, d_i), \\ \hat{O}_{i+1} &= \mathcal{U}(\hat{s}_{i+1}, \hat{O}_i, u_i), \end{aligned} \quad (11)$$

对于 $i = 0$、.....、$H-1$ 以及 $H = 10$。第一行将量化后的令牌与本体感知历史一起解码为归一化身体动作 $u_i \in \mathbb{R}^{29}$。第二行利用世界模型预测状态增量，并应用积分算子 $\mathcal{I}$ ，以获得在该动作和仿真上下文 ξ_{t+i}下的下一个想象状态$\hat{s}_{i+1}$ 。第四行使用新的状态和动作更新本体感知历史，因此下一个令牌将在与部署相同的历史约定下进行解码。策略更新使用 $H = 10$ 个步骤。固定评论家在第 H步进行评估；其值概括了剩余的$K - H = 10$ 个步骤。

每步成本汇总经过归一化且封顶的误差，

$$\begin{aligned} \ell_{k,i} &= \frac{\min(e_{k,i},\, \kappa\, \sigma_k^2)}{\sigma_k^2}, \\ \ell_{\mathrm{st},i} &= \frac{\min(e_{\mathrm{st},i},\, \kappa\, \sigma_{\mathrm{st}}^2)}{\sigma_{\mathrm{st}}^2}, \\ c_i &= w(f_{t+i})\, \Delta t \left(\sum_{k \in \mathcal{A}} w_k \, \ell_{k,i} + w_{\mathrm{st}} \, \ell_{\mathrm{st},i} \right), \\ J &= -\sum_{i=0}^{H-1} \gamma^i c_i, \end{aligned} \quad (12)$$

其中， c_i 是正成本， J 是对应的回报，$\Delta t = 0.02$s 是控制周期， γ 是折扣因子，κ 是上限乘数，而$\mathcal{A}$ 是跟踪项的活跃集合。七个候选误差 $e_{k,\ i}$ 涵盖身体位置、身体朝向、身体线速度和角速度、根部位置、根部朝向以及根部水平速度；权重为零的项处于非活跃状态。在任务转换处，状态转移权重为 $w(f_{t+i}) = 1.5$ ，其他位置则为 1。对于根部位置，$\ell_{k,\ i}$ 采用相同的截断形式，但在上限之外使用对数尾部，因此即使漂移较大，其梯度也不会消失。

世界模型会预测所有状态通道，包括速度，因为积分需要这些信息。其速度通道的可靠性低于位置：在 $H = 10$处，关节速度相关性约为 0.2，且根部速度存在系统性高估。因此，我们对角速度、根部线速度和偏航角速度赋予零成本，仅保留较小的身体线速度权重。对于防停滞项，我们使用三维根位移，包括其垂直分量。

防停滞误差针对根部运动停滞。令 $p_i \in \mathbb{R}^3$ 表示第 i 步的根部位置，并定义

$$\begin{aligned} m_i &= \frac{\|p_i - p_{i-1}\|}{\Delta t}, \\ m_i^{\mathrm{ref}} &= \frac{\|p_i^{\mathrm{ref}} - p_{i-1}^{\mathrm{ref}}\|}{\Delta t}, \end{aligned} \quad (13)$$

其中， m_i 和 m_i^{ref}分别为实际根部速度和参考根部速度。防停滞误差为

$$\begin{aligned} e_{\mathrm{st},i} = \ & \left(\min\{[m_i^{\mathrm{ref}} - m_i - \delta]_+,\, \mu_{\mathrm{st}}\}\right)^2 \\ & \times \mathbf{1}\{m_i^{\mathrm{ref}} > m_{\mathrm{gate}}\} \mathbf{1}\{m_i < \rho\, m_i^{\mathrm{ref}}\}, \end{aligned} \quad (14)$$

其中， $\delta = 0.05$ 米/秒、 $m_{\mathrm{gate}} = 0.1$ 米/秒、 $\mu_{\mathrm{st}} = 0.6$ 米/秒，以及 $\rho = 0.10$。这四个标量在验证单元上选定一次，随后在各任务中保持固定。仅当参考根部运动非零且实际根节点运动低于参考值的 ρ 倍时，门控才会开启。防停滞项采用单一的通用形式，并在各任务之间共享。它能够检测根部运动停滞，包括 SitSofa 中最后的下蹲动作。其观测到的最大影响出现在 SitSofa 上。

策略更新会展开 $H = 10$ 步。在第 H 步评估固定评论家，其价值概括了剩余的 $K - H = 10$ 步。

该更新会最小化执行成本，以及由其生成的关节目标所涉及的三组可微约束，

$$\begin{aligned}\mathcal{L} &= -\bar{J}_{\mathrm{ac}} + \lambda_{\mathrm{lim}}\mathcal{L}_{\mathrm{lim}} + \lambda_{\mathrm{trust}}\mathcal{L}_{\mathrm{trust}} + \lambda_{\mathrm{peak}}\mathcal{L}_{\mathrm{peak}},\\ \bar{J}_{\mathrm{ac}} &= J + \beta\,\gamma^{H}\,\mathbf{1}_{\mathrm{alive}}(H)\,V_H,\end{aligned} \tag{15}$$

其中 $\mathcal{L}_{\mathrm{lim}}$ 会对关节限位违规进行惩罚，$\mathcal{L}_{\mathrm{trust}}$ 是一个同噪声信赖域惩罚，用于惩罚关节目标相对于原始FT策略和仅投影策略的偏差，而$\mathcal{L}_{\mathrm{peak}}$ 则限制逐关节峰值偏差。标量 β 对评论家价值进行加权， γ 是折扣因子。这里，$V_H = V(y_H)$是对剩余 $K-H$ 步的回报进行的评论家估计，而 $\mathbf{1}_{\mathrm{alive}}(H)$是第 H步的存活指示器。在我们的设置中，$K=20$ 和$H=10$，因此 V_H 估计剩余10步的尾部回报；目标是来自 y_H的、与梯度分离的 $K-H$步蒙特卡洛回报。向量 y_H 依次堆叠了第 H步的想象状态、参考状态、二者之差、仿真上下文、本体感知历史以及规划令牌片段。y_H 中的规划令牌特征与梯度分离，因此评论家梯度不会绕过想象状态。评论家在初始化后保持固定；其输出仍然相对于 y_H可微，并且其梯度的范数相对于 J的梯度范数受到限制。因此，前10步会接收直接的执行梯度，而后10步仅通过评论家价值参与计算。

b) 反向传播：策略梯度通过在展开图上进行一次反向遍历获得。我们直接使用正的逐步成本 c_i ，因此执行回报为

$$J = -\sum_{i=0}^{H-1}\gamma^{i}c_i.$$

下面的递推式对第一-H 执行回报 J求导。评论家网络尾项在 y_H处单独求导，并累加到同一次反向遍历中；因此，对于 J，边界伴随变量为 $\delta_H = \mu_H = 0$。

定义状态伴随变量和历史伴随变量

$$\delta_i = \frac{\partial J}{\partial \hat{s}_i}, \qquad \mu_i = \frac{\partial J}{\partial \hat{O}_i},$$

以及局部雅可比矩阵

$$\begin{aligned}A_i &= \frac{\partial \hat{s}_{i+1}}{\partial \hat{s}_i}, \quad B_i = \frac{\partial \hat{s}_{i+1}}{\partial u_i},\\ C_i &= \frac{\partial u_i}{\partial \hat{O}_i}, \qquad D_i = \frac{\partial \hat{O}_{i+1}}{\partial \hat{s}_i},\\ E_i &= \frac{\partial \hat{O}_{i+1}}{\partial \hat{O}_i}, \quad F_i = \frac{\partial \hat{O}_{i+1}}{\partial u_i}.\end{aligned}$$

这里， A_i 传播状态敏感度， B_i 将动作变化映射为下一状态变化， C_i 将历史变化映射为动作变化， D_i 将状态变化映射为历史变化，E_i 传播历史敏感度，而 F_i 捕获直接的动作到历史路径。最后一项是必需的，因为历史更新不仅使用状态和先前历史，也会使用动作。

对于 J的 $\delta_H = \mu_H = 0$ ，伴随变量满足，对于 $i = H-1, \ldots, 0$,

$$\begin{aligned}\delta_i &= -\gamma^{i}\frac{\partial c_i}{\partial \hat{s}_i} + A_i^{\top}\delta_{i+1} + D_i^{\top}\mu_{i+1},\\ \mu_i &= C_i^{\top}\left(-\gamma^{i}\frac{\partial c_i}{\partial u_i} + B_i^{\top}\delta_{i+1} + F_i^{\top}\mu_{i+1}\right) + E_i^{\top}\mu_{i+1}.\end{aligned} \tag{16}$$

第一个方程将折扣后的直接成本对状态的敏感度，与未来状态和历史的敏感度相结合。第二个方程对历史执行相同操作，通过C_i将动作敏感度转换为历史敏感度，并纳入直接的动作到历史路径F_i。位置差运动估计还使 c_i依赖于$\hat{}$ s_{i-1}；实现会通过完整方案对这种依赖关系求导，而上面的递推式为便于阅读省略了这一点。

令牌梯度通过经由BFM投影动作伴随变量得到，

$$\frac{\partial J}{\partial \tilde{z}_i} = \left(\frac{\partial u_i}{\partial \tilde{z}_i}\right)^{\top}\left[-\gamma^{i}\frac{\partial c_i}{\partial u_i} + B_i^{\top}\delta_{i+1} + F_i^{\top}\mu_{i+1}\right]. \tag{17}$$

矩阵 B_i 是步骤 i处积分转移的动作雅可比矩阵：

$$B_i = \frac{\partial \mathcal{I}(\hat{s}_i, d_i)}{\partial d_i}\,\frac{\partial F_\phi(\hat{s}_i, u_i, \xi_{t+i})}{\partial u_i}.$$

因此，它包含积分步骤，而不只是 F_ϕ的原始输出。对于式 (4) 中的残差动作头，它还同时包含 MLP动作导数和 $G_\phi(\hat{s}_i)$，而不仅仅是G_ϕ 。

由于直通量化器的导数为 1，传递到动作头参数的梯度为

$$\nabla_\theta J = \sum_{i=0}^{H-1}\left(\frac{\partial z_i}{\partial \theta}\right)^{\top}\frac{\partial J}{\partial \tilde{z}_i}. \tag{18}$$

采样器雅可比矩阵展开为

$$\begin{aligned}\frac{\partial z_i}{\partial \theta} &= \sum_{k=0}^{3} P_k\,\Delta\tau_k\,\frac{\partial v_k}{\partial \theta},\\ P_k &= \prod_{m=k+1}^{3}\left(I + \Delta\tau_m\frac{\partial v_m}{\partial z_m}\right),\end{aligned} \tag{19}$$

其中， P_k 是块级采样器雅可比矩阵，它将后续采样器步数的敏感度反向传播至步骤 k； $\partial z_i/\partial\theta$ 表示其第 i个词元块。只有 θ（动作头参数，包括低秩适配器和输出投影）会接收策略更新； ψ 和 ϕ不受影响。评论家网络保持固定。其尾部价值在 y_H处添加边界梯度，并将其纳入同一次反向传播；评论家参数不会更新。

世界模型仅训练一次，使用激励回放以及通过初始策略收集的闭环回合；更新策略时，其系数保持固定。评估始终在模拟器中进行。算法1总结了一次更新，其计算成本包括四次

算法1 一次执行感知可微训练更新。

要求： 窗口 (s_t, O_t, ξ_t)，上下文 c_t
采样 ε；展开采样器（4 步）
$\tilde{z} \leftarrow Q(z)$ 前向传播，反向传播时采用恒等映射
对 $i = 0$ 到 $H-1$ **执行**
  $u_i \leftarrow \pi_\psi(\tilde{z}_i,\ \hat{}O_i)$
  $d_i \leftarrow F_\phi(\hat{s}_i,\ u_i,\ \xi_{t+i})$
  $\hat{s}_{i+1} \leftarrow \mathcal{I}(\hat{s}_i,\ d_i)$
  $O_{i+1} \leftarrow \mathcal{U}(\hat{s}_{i+1},\ O_i, u_i)$
结束循环
$\sum_{i<H} \gamma^i c_i$；形成 $\mathcal{L}$
反向传播；更新 θ
在 H处评估固定评论家网络

采样器的 DiT 前向传播、每个展开步骤各一次 BFM 和世界模型前向传播、一次固定评论家评估，以及一次反向遍历；其激活占用随 H 而非 K 增长。

IV. 实验

A. 实验设置

平台与任务。我们在 Isaac Lab 仿真环境中的 HumanoidArena 上使用 Unitree G1 进行评估，采用 29 个身体自由度和 14 个手指关节，身体控制频率为 50Hz，视觉-语言-动作模型块长度为 40 步，重规划预测时域为 20 步。仿真任务集包含四项 HumanoidArena 任务：OpenDoor、SitSofa、Boxing 和 VisionNavi。主要比较涵盖这四项任务，并使用完整的基线网格；本文所有定量闭环数据均来自模拟器。

数据与划分。训练数据包含 400 个记录的 HumanoidArena 回合，其中带有 SONIC 运动标记（每项任务 100 个）。在进行时间窗口划分前，先将场景实例和指令分开，因此留出场景和释义改写均不会出现在训练数据中。视觉-语言-动作模型使用这些回合进行微调。世界模型采用第 III-B 节所述的四来源混合数据（干净的 A′、令牌扰动的 A′、闭环配对数据和分支组），仅训练一次，并在执行感知可微训练期间保持固定；未来状态仅用于构造其回归目标，在部署时不可用。

指标与实验方案。所有成功率比较均采用相同的实验方案：4 个任务、随机种子 0–2，以及重复次数 0–19，因此每个任务有 60 个回合，每个实验组有 240 个回合，且各组的回合随机种子相匹配。任务成功率是主要指标。只有执行的回合成功，单元格才算成功；仅当测试框架无法启动（环境步数为零）时，才会重新运行一次该单元格；失败或超时回合仍计入所有分母。跌倒是指因跌倒而终止的回合。配对单元格采用 McNemar 检验。我们比较两个实验组：使用冻结的 SONIC 解码器进行任务微调的视觉-语言-动作模型（基线）以及采用执行感知可微训练训练 LoRA 适配器的同一模型（完整方案）；两组共享检查点、随机种子和回合。完整方案组在所有任务中使用相同的架构；基线不含适配器。

表 III：**共享训练方案**。 防停滞项具有单一通用形式，并在各任务间共享；它定义于三维根位移，并使用参考相对停滞门控。所有其他组件均为共享。

VLA	GR00T N1.7，8k 步任务微调
适配	LoRA 秩 16，$\alpha = 32$，16.9M 可训练参数
Data	400 回合
世界模型	动作条件化，训练一次后固定
评论家网络	初始化后固定
块 / 展开	40 / 20 步
预测时域 / 采样器	$H = 10$ / 4 欧拉步
更新次数 /批次 / 学习率	300 / 32 / 10^{-4} 余弦调度
$\lambda_{\lim}$ / β	0.01855 / 0.975
硬件	2× RTX 4090

表 IV：**仿真中的闭环结果**。随机种子 0–2 ×，重复次数 0–19，每个任务共 60 个回合，每个机械臂共 240 个回合。跌倒率是因跌倒而终止的回合所占的百分比。各项以基线→我们的方法的顺序显示；若仅给出一个跌倒率值，则表示两个机械臂的跌倒率均为 0.0%。

Task	基线 我们的方法 (%) (%)	Δ (pp)	Fall (%)
OpenDoor	65.0 78.3	+13.3	0.0
Boxing	71.7 81.7	+10.0	0.0
SitSofa	48.3 65.0	+16.7	0.0
VisionNavi	51.7 61.7	+10.0	15.0→8.3
平均值	59.2 71.7	+12.5	3.8→2.1

训练方案。所有任务共用相同的架构和训练方案（表 III）。视觉-语言-动作模型采用 GR00T N1.7 检查点，并在 HumanoidArena 令牌上微调 8k 步。LoRA 应用于注意力模块、前馈模块和输出投影（秩为 16，$\alpha = 32$，含 16.9M 个可训练参数，占 1.64B 参数模型的 1.0%）。世界模型仅训练一次，之后保持固定。每次更新使用 40 步块、20 步展开、10 步可微窗口和 4 个欧拉采样器步数。训练共进行 300 次更新，批量大小为 32，学习率采用余弦调度，从 10^{-4} 开始，在两块 RTX 4090 GPU 上运行。

B. 闭环基准比较

我们首先比较完整的闭环控制。可微分训练使平均成功率提高了 12.5 个百分点（表 IV），并且每个任务均有所提升；增幅在 10.0 至 16.7 个百分点之间，其中绝对增幅最大的是接触丰富的 SitSofa 任务。我们采用了温和的覆盖度重加权，以确保行走帧不会在初始分布中占比过低；下面的消融实验将单独分析其影响。

表 IV 展示了主要对比结果。平均成功率从 59.2% 提升至 71.7%（+12.5 个百分点），且每项任务均有所提升：SitSofa 的成功率从 48.3% 提升至 65.0%（+16.7 个百分点），OpenDoor 从 65.0% 提升至 78.3%（+13.3 个百分点），Boxing 和 VisionNavi 则分别从 71.7% 提升至 81.7%、从 51.7% 提升至 61.7%（各提升 +10.0 个百分点）。表 VII 单独分析了反停滞贡献：移除

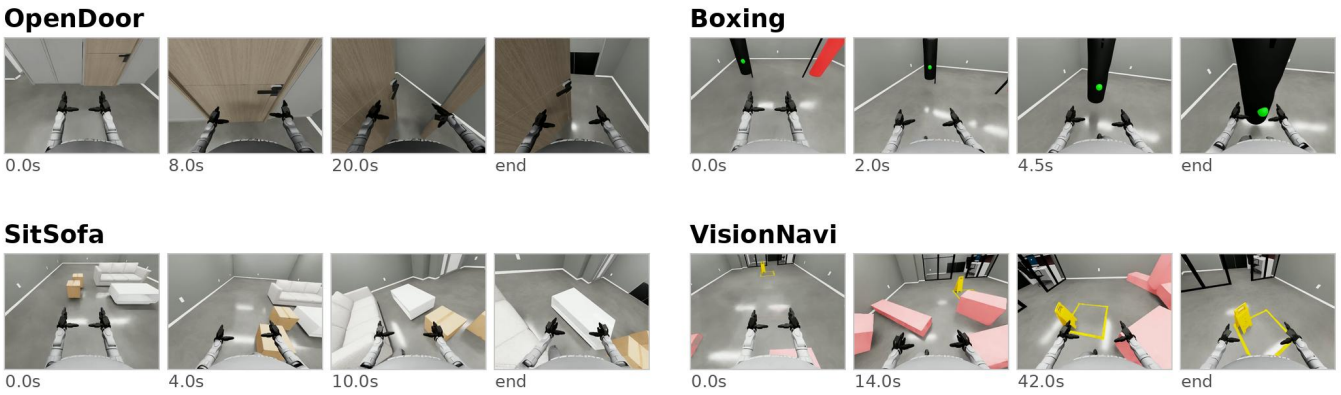


图 3：**仿真评估任务**。每个单元格代表一个任务；帧展示第一帧、两个关键时刻和最终帧（“结束”），并按时间或“结束”标注。OpenDoor：接近、把手接触和开门。Boxing：接近以及两次出拳/接触时刻。SitSofa：接近、接触和坐姿。VisionNavi：穿越障碍场到达目标。帧仅用于示意，并不表示成功标注。

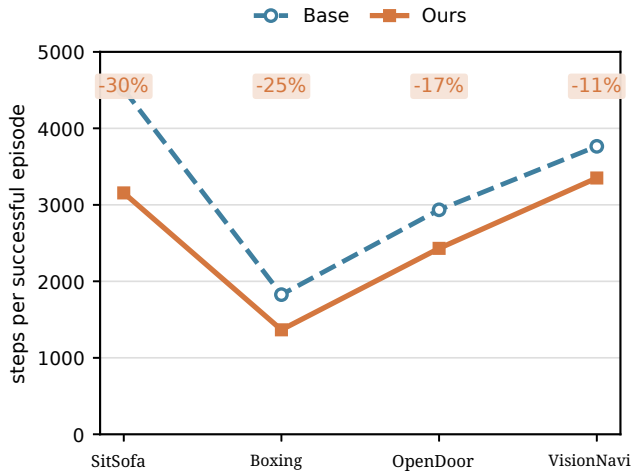


图 4：**成功回合的成本**。每个成功回合的模拟器步数，即 60 个回合的总步数除以成功次数。百分比表示从基线到我们的方法的变化；完整策略在每项任务上的成本都更低。

后，OpenDoor、Boxing 和 VisionNavi 的成功率仍与完整方案相同，而 SitSofa 的成功率则从 65.0% 降至 50.0%。仅在 VisionNavi 中观察到跌倒率，且其数值从 15.0% 降至 8.3%，同时该任务的成功率提高了十个百分点，因此这一增益不能归因于更冒险的行为。所有失败回合和超时回合均计入每项任务的分母。在 240 个配对单元上，整体增益具有统计显著性（McNemar检验 $p < 0.01$）。

图4比较了两个机械臂每次成功回合所耗费的模拟器步数，该数据通过对全部60个回合的步数求和，再除以成功次数得到。完整策略在每项任务上所需步数都更少：在OpenDoor上，成本从2935步降至2431步（−17%）；在VisionNavi上从3765步降至3351步（−11%）；在SitSofa上从4508步降至3156步（−30%）；在Boxing上从1826步降至1366步（−25%），平均值则从3102步降至2490步（−20%）。节省最多的是基线最常超时的任务，即SitSofa和Boxing。因此，额外的成功并非以增加交互时间为代价：相同的预算带来了更多成功，而且每次成功的成本更低。

C. 代理与梯度分析

连通性是必要的。仅执行目标函数就能在流头上产生有限且非零的梯度，全部四个采样器步数都能接收到梯度，并且世界模型的输入雅可比矩阵在其参数被冻结时仍保持活跃。

我们将代理用作方向，而不是校准回报。两项检查支持这一点。首先，它对 405 组反事实动作序列的排序表现显著高于随机水平（四选一准确率为 70.4%，$\tau = 0.67$ 在 $H = 10$，随机水平为 25%）。其次，在针对 87 个留出窗口的 $\pm\delta$ 探测实验中（$H = 10$、$\delta = 0.03$），沿着世界模型梯度移动，真实回报平均提升了 $+0.014$，且预测回报变化与真实回报变化的相关系数为 0.689。检查点根据留出验证窗口上的可微执行回报进行选择。在采样器梯度估计器中，路径式方法将令牌误差保持在 0.115；离散伴随在回报方面与其相当，但保真度不及；截断方法存在偏差，并使令牌误差升至 0.174。

D. 微分时域的选择

式 (12) 中的预测时域 H 决定 20 步执行块中有多少内容能获得信用分配，以及更新对代理的信任程度。较短的窗口更容易分配信用，但执行块的大部分内容仍无法获得信用分配；较长的窗口覆盖执行块的更多内容，但会累积

表 V：**梯度估计器保真度**。在相同的估计器探测后得到的令牌误差；回报列表示该估计器是否匹配路径式回报。

估计器	令牌误差	回报
路径式	0.115	参考
离散伴随	0.121	匹配
截断	0.174	有偏

表VI：**微分时域的选择**。各候选窗口中代理回报与模拟器回报之间的四选一准确率和肯德尔 τ。最后一列标出两项标准（$\geq 70\%$ 和 $\tau \geq 0.65$）。

H	四选一准确率（%）	Kendall τ	Pass
1	62.7	0.563	
2	72.3	0.700	✓
5	76.8	0.732	✓
10	70.4	0.669	✓
20	64.2	0.596	

代理偏差。因此，我们根据代理选择 H，采用两个不涉及闭环成功率的准则：在反事实组中，代理在至少 70% 的四选一比较中必须选出模拟器也更偏好的动作序列（随机猜测的概率为 25%）；并且代理对四个候选项的排序必须达到至少 0.65 的 Kendall τ。在满足这两个条件的窗口中，我们选择最大的一个，以便梯度尽可能覆盖代理能够支持的执行块范围。这是排序准则，而非绝对准确率准则：代理被用作方向信号，而不是校准后的模拟器。

排序测试使用了从由扰动运动令牌驱动的身体控制器滚动中收集的 405 个反事实组。每组都从同一状态开始，并包含四个候选动作序列；代理模型和模拟器在相同的H步数上为每个序列分配回报，然后对所得排序进行比较。如表 VI 所示，$H = 1$ 低于两个标准（62.7%，$\tau = 0.56$）。在 $H = 5$ 时，准确率达到峰值（76.8%，$\tau = 0.73$）；在 $H = 10$ 时仍高于两个标准（70.4%，$\tau = 0.67$），但在 $H = 20$ 时低于两个标准（64.2%，$\tau = 0.60$）。

$H = 10$ 是同时通过两个标准的最大窗口。$H = 5$ 在相同测试中的排序最佳，但只为执行块的四分之一分配信用；而 $H = 20$覆盖整个执行块，却低于两个标准。因此，我们为每个可微训练分支固定使用 $H = 10$；闭环结果是在所选预测时域上的确认，而不是预测时域扫描。

所有闭环数值均采用相同的架构和训练流程。防停滞项采用单一通用形式，并对所有任务启用；其相对于参考方案的门控会将有效梯度集中于 SitSofa。架构、数据、检查点、预测时域、展开和优化器均保持一致。在 $H = 10$，最终方案在四项任务上的得分分别为 78.3%、81.7%、65.0% 和 61.7%，

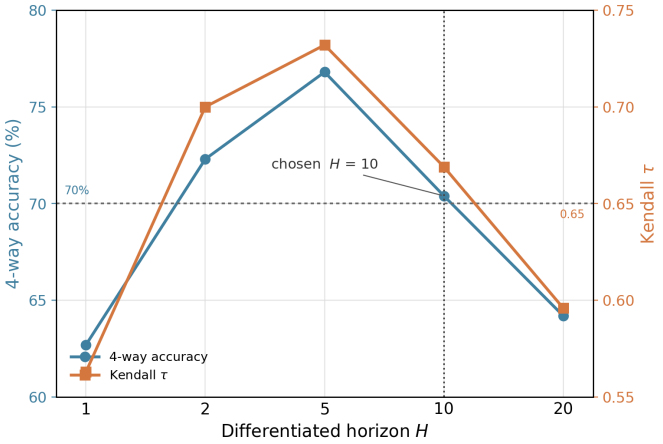


图 5：**代理排序与预测时域的关系**。在 405 个反事实组中，代理回报与模拟器回报之间的四向选择准确率（圆形）和 Kendall τ（方形）；虚线表示用于选择窗口的阈值（70% 和 0.65）。选定的窗口为 $H = 10$。为便于阅读，H 值按等间距显示。

表 VII：**四项任务上的留一成功率**。完整方案是标准方案；基线作为参考。所有任务均启用防停滞，但其效果集中于 SitSofa；覆盖度和状态转移均保持一致。VisNavi 是 VisionNavi 的缩写。

配置	OpenDoor (%)	Boxing (%)	SitSofa (%)	VisNavi (%)	平均值 (%)
基线（参考）	65.0	71.7	48.3	51.7	59.2
完整方案 −覆盖度	73.3	75.0	56.7	55.0	65.0
完整方案 −状态转移	71.7	76.7	58.3	56.7	65.8
完整方案 −防停滞	78.3	81.7	50.0	61.7	67.9
Full	78.3	81.7	65.0	61.7	71.7

平均值为 71.7%，而基线为 59.2%（表 IV）。

E. 消融研究

本节报告两组互补的消融实验。首先，我们每次从规范的完整方案中移除一个策略目标组件，并比较任务级成功率；基线作为参考。这可以将一个门控防停滞项与两个共享项——覆盖度和转移加权——区分开来。其次，我们改变世界模型数据混合，并在排序、闭环根误差、位置动作效应和联合动作增益方面比较所得模型。两部分分别回答不同的设计问题：在部署时哪些目标函数项发挥作用，以及哪些数据源能使代理对梯度有用。

1) 策略目标组件：我们每次从标准完整方案中移除一个组件，并比较任务级成功率；基线作为参考列出。该表将门控反停滞项与两个共享目标项区分开来，因此既展现了任务局部性，也揭示了跨任务影响。

a) 覆盖度：覆盖度呈现出广泛的运动特征。移除该项后，OpenDoor 的成功率从 78.3% 降至 73.3%，Boxing 从 81.7% 降至 75.0%，SitSofa 从 65.0% 降至 56.7%，VisionNavi 从 61.7% 降至 55.0%。

b) 状态转移：转移加权的影响范围最广。移除该项后，OpenDoor 的成功率从 78.3% 降至 71.7%，Boxing 从 81.7% 降至 76.7%，SitSofa 从 65.0% 降至 58.3%，VisionNavi 从 61.7% 降至 56.7%。在把手接触、开始坐下以及接近至接触的转移阶段，效果最为明显，因此 OpenDoor 和 SitSofa 的成功率下降约 6–7 个百分点，而 Boxing 和 VisionNavi 则下降 5 个百分点。持续坐姿诊断结果一致：采用转移加权时，56 个回合中有 11 个持续保持坐姿；不采用时则为 4 个（分别为 20% 和 7%；McNemar检验 $p = 0.092$）。我们将其视为趋势性证据，而非显著性证据。

c) 防停滞：防停滞项适用于所有任务。其门控基于三维根位移：当参考根部运动非零，且实际根节点运动低于参考值的一定比例 ρ 时，门控便会开启。移除该项后，OpenDoor、Boxing 和 VisionNavi 的成功率与完整方案相同；SitSofa 的成功率则从 65.0% 降至 50.0%。该结果仍略高于基线的 48.3%，因为梯度路径和其他目标函数项仍然存在。观察到的最大影响出现在 SitSofa 上；未报告各任务的门控激活统计数据。

d) 总体模式：每一行中间方案在每项任务上的表现都介于基线和完整方案之间，且没有任何单个组件能够解释四项任务的全部差距。这些行的差异在于作用位置：覆盖度偏重运动，状态转移加权最为均衡，而防停滞的作用则主要体现在 SitSofa 上。这些差距具有任务特异性，而非均匀分布，这与三个术语的不同作用相一致。

2) 世界模型数据源：世界模型数据混合与上述策略目标组件是相互独立的消融轴。我们依据多个代理指标进行选择，而不仅仅依据排序（表VIII）。闭环配对数据是唯一能够校准部署分布的数据源：H=10 闭环根节点误差从 2.1 cm 降至 0.45 cm。令牌扰动使位置动作的作用效果 R^2 从 0.00 提升至 0.16。分支组改善了联合动作敏感度，使联合增益从 0.604 提升至 0.634，并带来小幅排名增益（0.710 至 0.717）。完整混合数据并非排名最高——A′+令牌的排名高出 2.8 个百分点——但它在四项代理指标上均达到帕累托最优，并且与最佳闭环根节点和位置动作R^2相比，分别相差不超过 0.02 cm 和 0.02 R^2。因此，我们将其视为一种均衡的设计选择，而不是宣称其具有普遍优越性；生产版世界模型在此混合数据的基础上，增加了 SitSofa 时间窗口。

F. 真实机器人闭环推理

此外，我们还在实体 Unitree G1 上以 50Hz 进行真实机器人数据收集和部署，并采用 20 步重规划。我们首先在目标场景中收集演示数据，再使用这些真实机器人轨迹对视觉-语言-动作模型进行进一步微调，随后运行相同的执行感知可微训练

表 VIII：**世界模型数据源**。 所有机械臂均使用相同的初始化、阶段设置和 420 组排序。闭环根节点是在 4,000 个留出的闭环时间窗口上计算的 H=10；位置动作是在 400 个分支组上计算的 H=10 R^2；关节增益是预测的关节动作效应与模拟器关节动作效应之间的 H=1 斜率。生产版世界模型增加了坐沙发时间窗口，并非这个受控的完整机械臂方案。

Arm	四路闭环根节点位置 (%)	R^2 (cm)	关节增益	
干净 A′ 仅	73.8	2.14	0.00	0.619
A′+词元	74.5	2.16	0.16	0.626
A′+词元+闭环	70.2	0.45	0.19	0.599
无分支	71.0	0.46	0.20	0.604
full	71.7	0.47	0.18	0.634

训练配方中包含真实机器人轨迹。适配后的策略被部署用于闭环推理；因此，这是一次适配后的真实机器人运行，而不是零样本迁移。这些运行结果用于定性验证，不纳入成功率表；定量比较仍仅限于仿真。图 6 展示了四段具有代表性的序列：(a) 拾起一个垃圾袋；(b) 踩下脚踏垃圾桶并投入垃圾；(c) 从跪姿将一支试管放到桌上；(d) 将球踢向目标。

V. 局限性

我们的定量证据仅限于四个模拟任务。真实机器人运行是在真实机器人数据收集和适配之后进行的定性闭环推理，而不是真实机器人成功率基准。

学习到的世界模型是一个粗粒度的动作条件模拟器。它展开执行块并提供可微执行梯度，但其精度不足以取代模拟器进行评估，也不足以提供校准回报。在留出策略转移探针上，在四任务探针中，其 H=10 误差约为根部位置 3–4cm 、旋转 7–8° 。这支持在执行窗口内进行粗粒度的信用分配，但无法支持精确的长预测时域预测。数据源表中的 CL-root列是在留出的闭环时间窗口上进行的另一种测量，并非同一个探针。

可微更新也并非完整的在线策略循环：世界模型基于固定数据混合进行训练，因此策略无法跟踪自身不断变化的状态分布。这限制了可实现的增益；仅靠延长可微训练不太可能弥合剩余差距。

执行目标函数涵盖 29 个身体关节。14 个手部通道从 8000 步的任务微调中初始化，不由执行目标函数直接优化；我们不评估灵巧手操作。以物体为中心的接触不在当前仅状态世界模型的建模范围内，该模型不会显式建模物体状态或接触动力学。对留出场景、新的物体摆放、指令释义、动作延迟、观测噪声、摩擦力和外部推力的泛化仍有待未来研究。

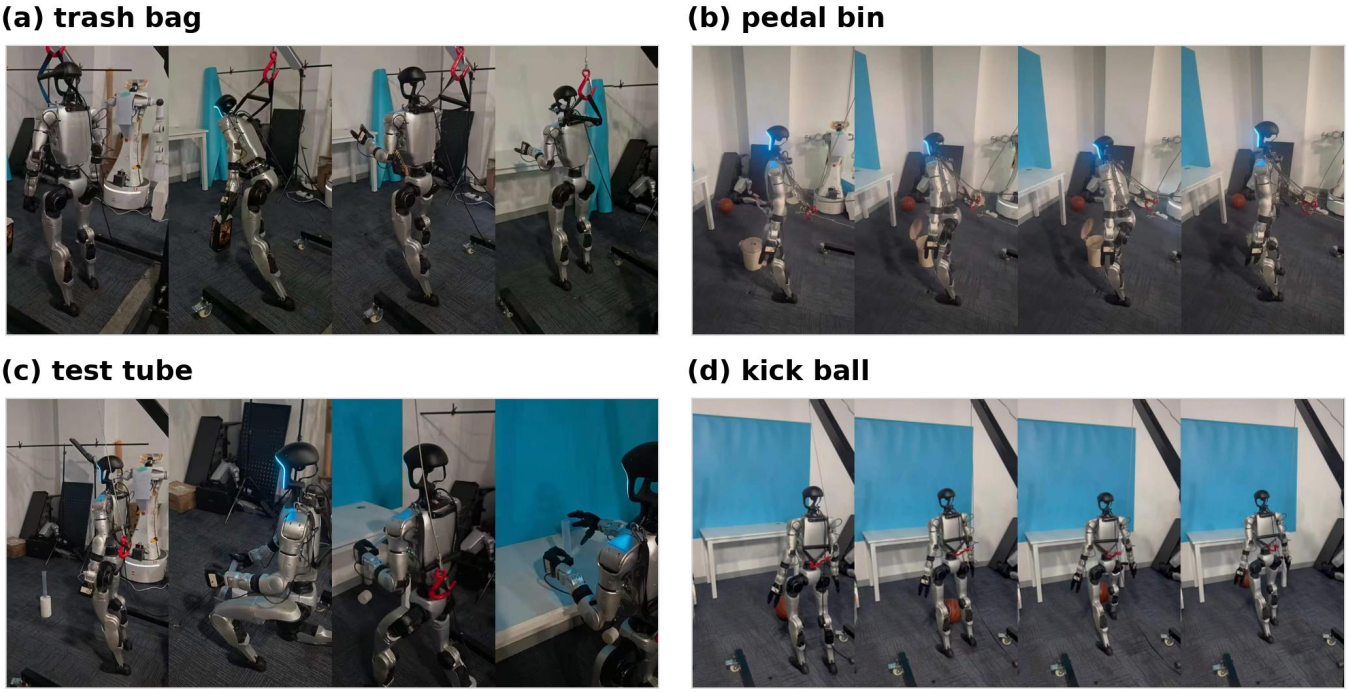


图6：**在 Unitree G1 上进行真实机器人闭环推理**。(a) 弯腰捡起一个垃圾袋。(b) 走到脚踏式垃圾桶旁，踩下踏板并将垃圾扔进去。(c) 跪下，捡起一支试管并将其放在桌上。(d) 将球踢向目标位置。每个面板均展示完整的录制序列；这些帧仅作示意，不代表任何定量成功率。

一个自然的下一步是改进世界模型本身，而非只改进策略优化器。在架构方面，速度感知预测头、运动学一致性约束和接触感知状态表示，或许能提高速度通道的准确性，使其足以纳入执行成本，而不必降低其权重。在数据方面，提升现有转移数据源的覆盖度和均衡性，或许能更好地表示恢复过程和接触状态。通过更好的校准，并依据当前策略分布定期刷新模型，可进一步降低代理误差。更准确的模型能让更长的可微时间范围和更大的信赖域更加安全，从而减少执行梯度对代理误差的依赖。将同样的方法扩展到 14 个手部关节，则是另一个独立方向。

VI. 结论

我们提出了一种执行感知可微训练框架，通过学习得到的动作条件世界模型，将流匹配人形机器人视觉-语言-动作模型与冻结的全身控制器结合起来。其核心思想是将执行成本通过控制器和四步采样器反向传播至动作头，从而使策略针对其实际控制的身体进行优化，而不只是针对令牌进行优化。在四项 HumanoidArena 任务上，该 $H = 10$方案将平均成功率从 59.2% 提高至 71.7%，并将每个成功回合的成本从 3102 降至 2490 步，其中 SitSofa 的提升最大；该架构保持不变，仅加入门控根部运动防停滞项，就足以改变接触行为，而无需修改模型。经过真实机器人数据适配后，我们还在实体 Unitree G1 上定性地完成了推理闭环。主要局限仍包括粗粒度代理、固定的数据混合以及仅身体梯度；物体感知世界模型、策略在线刷新，以及留出数据或鲁棒性评估将留待未来工作。

参考文献

[1]K. Black, N. Brown, D. Driess, A. Esmail 等，“π_0：一种用于通用机器人控制的视觉-语言-动作流模型”，*arXiv* 预印本 *arXiv:2410.24164*，2024年。 [在线]。可访问：https://arxiv.org/abs/2410.24164 。Y. Chen, Y. Ge, H. Zhou, M. Ding, Y. Ge 和 X. Liu，“DIAL：通过潜在世界建模解耦意图与动作，实现端到端视觉-语言-动作模型”，*arXiv* 预印本 *arXiv:2603.29844*，2026年。 [在线]。可访问：https://arxiv.org/abs/2603.29844 。Z. Luo, Y. Yuan, T. Wang, C. Li 等，“SONIC：扩展运动跟踪规模，实现自然人形机器人全身控制”，*Science Robotics*，第11卷，第117期，页码 eaed4592，2026年。 [在线]。可访问：https://arxiv.org/abs/2511.07820 。H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng 和 H. Li，“WholeBodyVLA：面向全身移动操作控制的统一潜在视觉-语言-动作模型”，*arXiv* 预印本 *arXiv:2512.11047*，2025年，亦于 ICLR 2026 [在线]。可访问：https://arxiv.org/abs/2512.11047 。Y. Hu, H. Zhu, B. Zheng, Y. Hu, T. Zhang, Z. Chen, J. Zhao, R. Nai 和 Y. Gao，“OpenHLM：全身人形机器人移动操作的实证方案”，*arXiv* 预印本 *arXiv:2606.22174*，2026年。 [在线]。可访问：https://arxiv.org/abs/2606.22174

[6]J. Ren, C. Yu, S. Chen, X. Ma, L. Pan 和 Z. Liu，“DiffMimic：利用可微物理实现高效运动模仿”，*arXiv* 预印本 *arXiv:2304.03274*，2023年。 [在线]。可访问：https://arxiv.org/abs/2304 03274 。Q. Liao, Y. Shao, J. [7]Long, K. Zakka, R. Horowitz, P. Abbeel 和 K. Sreenath，“SuperMimic：借助单个可微世界模型扩展人形机器人全身控制”，发表于机器人学习会议，2026年。 [在线]。可访问：https://arxiv.org/abs/2610.28480[8] 英伟达，“Isaac GR00T：基础模型与代码”，https: // github.com/NVIDIA/Isaac-GR00T，2026年，公开软件和模型文档，访问于2026年10月。[9] T. Wang, Z. Xie, B. Yang 等，“HumanoidArena：自我中心分层全身学习基准测试”，*arXiv* 预印本 *arXiv:2606.17833*，2026年。 [在线]。可访问：https://arxiv.org/ abs/ 2606 17833 。D. Gao, B. [10]Zhao, A. Lee, I. Chuang, H. Zhou, H. Wang, Z. Zhao, J. Zhang 和 I. Soltani，“VITA：视觉到动作流匹配策略”，*arXiv* 预印本 *arXiv:2507.13231*，2025年。[在线]。可访问：https://arxiv.org/abs/2507.13231[11] Physical Intelligence，“$\pi^*_{0.6}$：一种能够从经验中学习的视觉-语言-动作模型”，2025年。 [在线]。可访问：https://arxiv.org/abs/ 2511.14759[12] Z. Wang, R. Zhang, Y. Liu, X. Mao, M. Wang 和 Y. Mu，“Q-VGM：面向离线到流匹配视觉-语言-动作策略的在线强化学习，” *arXiv* 预印本 *arXiv:2606.08015*，2026年。 [在线]。可访问：https://arxiv.org/abs/2606 08015 。L. Gu、C. Hou、M. Li、X. Wang、[13]X. Lv、S. Bai 等，“HAF-VLA：面向人形机器人视觉-语言-动作控制的分层动作流框架”，*arXiv* 预印本 *arXiv:2610.26584*，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2610 26584 。Y. [14]Wang、H. Jiang、S. Yao、Z. Ding 和 Z. Lu，“用于人形机器人全身控制的端到端语言-动作模型”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2026年，第 38 399–38 409 页，arXiv:2511.19236，预印本标题为 SENTINEL。 [在线]。可访问：https://arxiv.org/abs/2511 19236 。H. Yao、Z. Song、B. [15]Chen 和 L. Liu，“ControlVAE：基于模型学习用于基于物理的角色的生成式控制器”，*ACM*图形学汇刊，第 41 卷，第 6 期，第 1–16 页，2022年。[16] H. Yao、Z. Song、Y. Zhou、T. Ao、B. Chen 和 L. Liu，“MoConVQ：通过可扩展的离散表示实现统一的基于物理的运动控制”，*ACM*图形学汇刊，第 43 卷，第 4 期，第 1–21 页，2024年。[17] Y. Yang、J. Rong、M. Zhu、Z. Zhao、Q. Hu、Z. Lan、W. Mao、Y. Li、Z. Fu 和 H. Chen，“通过联合状态-动作生成实现的人形机器人世界动作模型”，*arXiv* 预印本 *arXiv:2610 12026*，2026 2610 12026。。 [在线]。可访问：https://arxiv.org/abs/ 。