

MLM: Learning Multi-task Loco-Manipulation Whole-Body Control for Quadruped Robot with Arm

Xin Liu¹, Bida Ma¹, Chenkun Qi^{*1}, Senior Member, IEEE, Yan Ding^{†2}, Nuo Xu¹, Zhaxizhuoma², Guorong Zhang², Peng Chen², Kehui Liu², Zhongjie Jia², Chuyue Guan², Yule Mo¹, Jiaqi Liu¹, Feng Gao¹, Jiangwei Zhong⁴, Bin Zhao², and Xuelong Li^{2,3}, Fellow, IEEE

Abstract—带有机械臂的四足机器人进行全身运动操作仍然是一个具有挑战性的问题，特别是在实现多任务控制方面。为了解决这个问题，我们提出了 MLM，一种由真实世界和模拟数据驱动的强化学习框架。它使配备六自由度机械臂的四足机器人能够自主或在人类远程操作下执行多任务的全身运动操作。为了解决在运动操作学习过程中平衡多任务的问题，我们引入了一个具有自适应、基于课程的采样机制的轨迹库。这种方法使得策略能够高效利用真实世界收集的轨迹来学习多任务的运动操作。为了应对只有历史观测值的部署场景，并提高不同空间范围任务中策略执行的性能，我们提出了一个轨迹-速度预测策略网络。它可以预测不可观测的未来轨迹和速度。通过利用广泛的模拟数据和基于课程的奖励，我们的控制器在模拟中实现了全身行为，并能够无须调整直接迁移到真实世界的部署。在模拟中的消融研究验证了我们方法的必要性和有效性，而在配备 Airbot 机械臂的 Go2 机器人上的真实世界实验展示了策略在多任务执行中的良好性能。

Index Terms—Reinforcement learning, legged robot, multi-task loco-manipulation, whole-body control.

I. 介绍

四足动物机器人的运动控制，特别是通过强化学习 (RL)，已经获得了显著的关注和发展，提升了它们多样且令人印象深刻的能力 [1]–[6]。六自由度 (DoF) 机械臂可以为四足机器人执行操作提供更多的可能性。通过将四足机器人的运动和机械臂的操作能力结合，与仅使用四足机器人相比，可以完成更广泛的任务 [7], [8]。

然而，对于四足机器人来说，使用单一策略完成多种行走操控任务是一个挑战。跨多个任务采用单一策略可以通过消除对特定任务策略的需求来降低控制复杂性。然而，平衡多项任务的问题仍未被现有方法充分解决。因此，一个自适应平衡机制是必不可少的。实现自适应平衡的方法需要进一步研究。在模拟驱动的强化学习范式中，不同任务特定的虚拟资产和奖励通常各不相同，而仔细的个性化设计通常导致显著的工作量且缺乏普遍性。此外，由于模拟和现实世界之间在纹理、颜色、光照等因素上存在显著差异，处理操作所需的视觉信息差异通常很难解决。一个可行的解决方案是将多个任务提炼为多个轨迹学习。这避免了将任务目标的观测数据（比如视觉信息）纳入控制中，而是用轨迹取而代之。

轨迹的获取可以通过多种方法实现，包括类似 UMI 的方法 [9], [10]、遥操作 [11], [12] 和基于外骨骼的系统 [13]。需要注意的是，我们不收集用于四足行走操纵的全身轨迹数据，这是我们的方法与模仿学习的区别。这是由于物种选择以进行数据收集的挑战，以及无法从真实机器人中获得此类数据。因此，我们使用类似 UMI 的方法仅获取末端执行器的 6D 操纵轨迹，这是一种与机器人无关的轻量化解决方案。

将现实世界中的多任务操控轨迹整合到仿真中，并利用仿真来有效生成与环境交互的训练数据是一种可行的方法。此外，通过将多个任务抽象为多条轨迹并设计统一的任务奖励，可以指导四足机器人在单一策略下生成协调的全身行为以完成任务。这些来自现实世界的多任务数据提供了与物理环境动态一致的空间轨迹，并分布在机器人的操控空间中，这对于学习全身行为和推广到未知轨迹至关重要 [14]。这与在仿真中随机采样空间曲线形成对比，因为后者缺乏速度和加速度等现实世界动态。

需要注意的是，在仅能访问历史轨迹的部署方案中，如远程操作，未来轨迹信息不能作为观察使用。然而，未来信息对于全身执行任务的稳定性至关重要，因为它指示了动作中下一步的方向 [15]。因此，需要一个轨迹预测机制。

为了解决上述挑战，我们引入了 MLM，一种基于学习的多任务运动操作方法。我们整合了 Fast-UMI [10] 来提供操作轨迹。在模拟中，我们训练控制策略并将其转移到现实世界中。我们的贡献总结如下：

- 我们提出了一种基于轨迹-速度预测的框架，用于训练多任务运动操控控制器，该控制器通过真实世界和仿真数据驱动，支持通过遥操作和扩散策略进行轨迹接口部署。
- 我们引入了一个多任务现实世界轨迹数据库，其中自适应和基于课程的采样使机器人能够在各任务之间平衡性能，并通过单一策略网络学习多任务操作。
- 我们部署了一台带有机械臂的真实四足机器人，并验证了我们的方法在多任务中的全身运动操控控制方面的出色性能。

相较于基于模型的方法，RL（强化学习）驱动的腿式机器人运动控制器的成功归功于以下关键优势。RL 使用从模拟的机器人环境交互中收集的大量数据来训练控制策略，捕捉这些交互的动态特性，同时利用比现实世界训练允许更多的数据量。此外，基于奖励驱动的策略网络提供了从状态到关节动作的端到端映射，消除了推断过程中昂贵优化的需求。

* [†] Corresponding author

¹ School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China. liu15764167516@sjtu.edu.cn

² Shanghai AI Laboratory, Shanghai, 200030, China.

³ TeleAI, China Telecom Corp Ltd., Shanghai, China.

⁴ Lenovo Corporation, Shanghai, China.

基于并行动态模拟器 Isaac Gym, 一个四足机器人被训练在几分钟内行走成功 [16]。采用执行器网络来模拟 ANYmal 机器人的复杂执行器动力学, 从而实现从模拟到现实的转移 [17]。在扩展这种方法时, 引入了特权学习来训练 ANYmal 机器人, 成功实现了针对复杂地形的盲目运动控制器 [1]。此外, 还提出了一种名为 RMA 的方法以开发稳健的四足行走控制器 [3]。为了促进步态学习和提高对复杂地形的适应性, 引入了步态先验 [2], [18], [19], 使机器人能够利用平坦地形的运动数据并快速收敛到所需的步态模式。尽管这些努力已经显著提高了腿式机器人的运动能力 [20]–[23], 但它们在运动操控能力的发展方面仍需进一步探索。

为了赋予腿式机器人的操控能力, 研究人员将腿式平台的移动能力与机器人手臂的操作能力相结合。例如, 在 [24] 中, 解决了步态规划问题, 并开发了一种基于分层优化的控制器, 使四足机器人具有一个六自由度的机器人手臂来执行操作。在 [7] 的研究中, 提出了一种方法, 该方法通过搜索全身轨迹并确定接触计划, 在预定义环境中解决移动操作任务。然而, 这些基于优化的方法通常需要特定任务的设计, 这在将多个任务的执行整合到单一控制策略中时提出了重大挑战。此外, 由于涉及的优化过程, 它们还会产生大量的计算开销。

近年来, 多个研究聚焦于采用 RL 实现拥有机械臂的腿式机器人的端到端控制, 并探索在多任务间执行策略。在 [25] 中, 机械臂的运动被建模为可预测的外部力矩, RL 训练的运动策略利用这些力矩来补偿机械臂的影响。然而, 这种非全身性的方法限制了机械臂利用四足机器人移动性和姿态调整能力以扩展其操作范围的能力。为了实现全身控制, 提出了一种基于 RL 的正则化在线自适应方法, 成功实现了四足机动操作的模拟到真实转移 [26]。尽管如此, 手臂基座下定义的操作空间容易受到身体晃动的影响, 并且不够稳定。我们使用世界空间来解决这个问题, 因为操作目标与身体运动无关。在 [27] 中, 提出了一个非线性奖励融合模块, 以实现轮式四足操控器的全身行为。然而, 它没有整合任务的自动执行。在 [28] 中, 提出了一种基于 RL 的方法, 该方法结合视觉信息用于自主机动操作执行。然而, 由于手臂基座受到身体运动的影响, 逆运动学 (IK) 策略也容易受到运动干扰。在 [15] 中, 实现了一种以任务空间中基于轨迹的视觉驱动的机动操作, 但它没有考虑在单一策略中集成多个任务和各任务间性能的平衡。此外, 这仅适用于涉及未来轨迹推理的轨迹生成方法, 如扩散策略 (DP)。仅涉及历史轨迹的遥操作不适用。

II. 全身控制方法

我们开发了一种基于真实环境和模拟的双数据驱动强化学习框架, 用于四足机器人全身运动操控控制。在训练期间, 引入了一个包含多任务真实环境操控轨迹的库, 并为控制训练了一个端到端轨迹-速度预测策略网络。该流程如图 1 所示。

A. 全身运动操控策略

我们消除了对仿真中物体和环境感知的需求, 这加快了训练速度并避免了由视觉感知引起的现实和仿真之间的巨大差距。由于信息是部分可观测的, 我们将步态操控任务建模为部分可观测的马尔可夫决策过程 (POMDP) [19], [29]。我们使用强化学习来训练一个腿-臂全身策略 π_θ , 旨在找到最优策略参数 θ , 以最大化贴现期望回报

$\mathbb{E}_{\pi_\theta}[\sum_{t=0}^T \gamma^t r_t]$, 其中 r_t 是时间 t 的奖励, γ 是折扣因子 ($\gamma = 0.99$)。

我们引入了一种非对称 Actor-Critic (AAC) 框架来训练行走操控策略。该方法旨在解决 POMDP 问题, 同时避免复杂的两阶段训练过程 [4]。

状态空间: 策略的完整状态包括四足机器人的本体感觉 $\mathbf{o}_t^q \in \mathbb{R}^{42}$ 、六自由度机械臂的本体感觉 $\mathbf{o}_t^a \in \mathbb{R}^{18}$ 、轨迹姿态序列 $\mathbf{o}_{t-3:t+4}^{tcp} \in \mathbb{R}^{72}$ 、特权状态 $\mathbf{s}_t^p \in \mathbb{R}^{21}$ 以及地形高程数据 $\mathbf{s}_t^e \in \mathbb{R}^{187}$ 。

具体来说, $\mathbf{o}_t^q$ 包含四足机器人的身体角速度 $\boldsymbol{\omega}_t^q \in \mathbb{R}^3$, 表示其方向的重力向量 $\mathbf{g}_t^q \in \mathbb{R}^3$, 腿部关节位置 $\boldsymbol{\theta}_t^q \in \mathbb{R}^{12}$ 和速度 $\dot{\boldsymbol{\theta}}_t^q \in \mathbb{R}^{12}$, 以及前一个时间步长的腿部关节动作 $\mathbf{a}_{t-1}^q \in \mathbb{R}^{12}$ 。手臂的本体感知 $\mathbf{o}_t^a$ 包含手臂关节位置 $\boldsymbol{\theta}_t^a \in \mathbb{R}^6$, 速度 $\dot{\boldsymbol{\theta}}_t^a \in \mathbb{R}^6$, 和上一个手臂关节动作 $\mathbf{a}_{t-1}^a \in \mathbb{R}^6$ 。轨迹姿态序列 $\mathbf{o}_{t-3:t+4}^{tcp}$ 包括相对于时间 t 的过去和后续步骤中的工具中心点 (TCP), 每个都包含 3D 位置和旋转矩阵的前六个元素。涵盖机器人和环境的特权状态 $\mathbf{s}_t^p$ 包括四足机器人的身体线速度 $\mathbf{v}_t^q \in \mathbb{R}^3$, 接触力 $\mathbf{f}_t^c \in \mathbb{R}^{12}$, 外力 $\mathbf{f}_t^e \in \mathbb{R}^3$, 以及它在机器人上的位置 $\mathbf{p}_t^e \in \mathbb{R}^3$ 。最后, $\mathbf{s}_t^e \in \mathbb{R}^{187}$ 表示在机器人坐标框架周围均匀分布的 187 个地形点的高度。

动作空间: 全身策略输出腿部的 12 个关节位置偏移量 $\mathbf{a}_t^q \in \mathbb{R}^{12}$, 和手臂的 6 个关节位置偏移量 $\mathbf{a}_t^a \in \mathbb{R}^6$, 均相对于它们的默认位置。这些偏移量被添加到默认位置上, 并作为比例-导数 (PD) 控制 [27] 的目标位置指令发送到电机。

奖励: 保持四足机器人协调的运动操控具有挑战性。在 [19] 的基础上, 我们使用对抗性运动先验 (AMP) 通过风格奖励来促进四足步态学习。在策略训练过程中, 同时训练一个对抗性判别网络, 以提供定义为 $r^s = \max[0, 1 - 0.25 \times (d^{\text{score}} - 1)^2]$ 的风格奖励。

策略的主要奖励 r^g 是跟踪机器人手臂的 TCP 位置 and 方向:

$$\begin{aligned} r^g &= r^{g_p} \cdot r^{g_o} \\ r^{g_p} &= \exp(-(e_1^p/\sigma_1^p \oplus e_2^p/\sigma_2^p \oplus \dots \oplus e_N^p/\sigma_N^p)) \\ r^{g_o} &= \exp(-(e_1^o/\sigma_1^o \oplus e_2^o/\sigma_2^o \oplus \dots \oplus e_N^o/\sigma_N^o)) \end{aligned} \quad (1)$$

其中 N 是任务总量, $\oplus$ 表示连接操作。我们已经建立了一种课程来更新不同任务的奖励因子 σ_n^p 和 σ_n^o , 更多细节在 II-C 中描述。约束奖励 r^l 用于调节机器人的合理操作、能量限制和硬件约束。所有奖励在表 I 中显示。

TABLE I: 全身机动操作学习的奖励机制。

Term	Equation	Weight
r^g	(1)	2.0
r^l	$-\ \boldsymbol{\tau}\ _2$	$1E-4$
	$-\ \ddot{\mathbf{q}}\ _2$	$2.5E-7$
	$-\ \mathbf{a}_{t-1} - \mathbf{a}_t\ _2$	0.1
	$-n_{\text{collision}}$	0.1
	$-\ \max(\boldsymbol{\tau} - \boldsymbol{\tau}^{\text{limit}}, 0)\ _2$	0.05
r^s	$-\ \max(\dot{\mathbf{q}} - \dot{\mathbf{q}}^{\text{limit}}, 0)\ _2$	0.5
	$\max[0, 1 - 0.25 \times (d^{\text{score}} - 1)^2]$	0.5

地形课程: 除了增强其以更多自由度调整姿态的能力外, 四足机器人的腿部配置还能灵活适应复杂地形。为了使四足机器人能够在复杂地形中执行行走操作任务, 我们引入

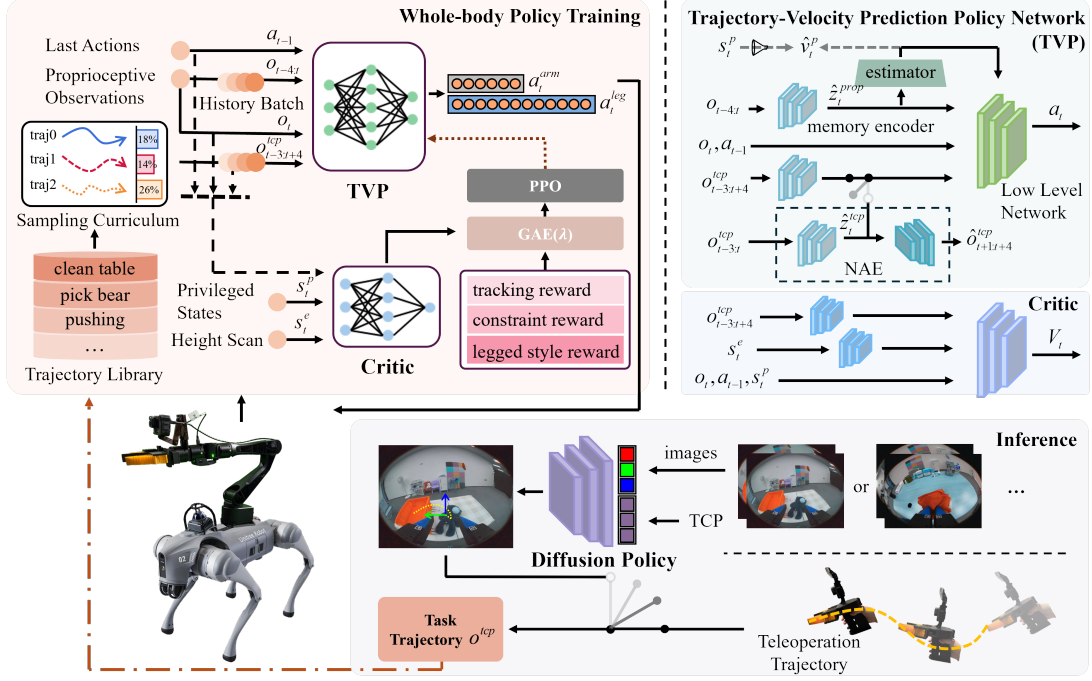


Fig. 1: 训练和推理流程概述。我们使用强化学习来训练一个全身控制策略。一个轨迹库与自适应采样课程相结合，为策略提供了一组多样的真实世界操作轨迹进行跟踪。一个标准化的自动编码器对历史目标轨迹进行编码，隐式传递有关即将到来的目标的信息；或者，该策略可以直接接收未来的轨迹，从而启用两种部署模式：遥操作和 DP 生成。我们还通过监督学习训练一个估计器来估计基础线性速度 $\hat{v}_t^p$ 。

了地形课程。这种方法允许机器人逐步学习在几种地形上进行全身行为，从较简单的条件开始到较困难的条件。

这些地形包括随机起伏的表面、斜坡、离散的障碍物、楼梯。与纯粹的运动任务不同，我们移除了地形穿越更新机制，而是使用奖励阈值更新。具体来说，我们在一个情节中监控 TCP 跟踪奖励，当它达到阈值时，机器人在下一集里进入课程的下一级；否则，它会回到上一级。我们发现这种机制有效地确保了机器人在保持基础稳定性的同时，维持良好的跟踪性能。

领域随机化：为了减少模拟到真实的差距，我们在训练过程中在指定范围内随机化机器人硬件、控制和环境参数，以确保模拟涵盖或近似现实条件。如表 II 所示。这包括 PD 增益、连杆质量、载荷质量和位置、地面摩擦力以及电机强度。此外，引入了 0 到 0.02 秒的观察延迟，以减轻系统的非实时行为。

TABLE II: 随机化参数及其范围。

Parameters	Range [Min, Max]	Unit
Leg Joint Stiffness	$[0.8, 1.2] \times 30$	-
Leg Joint Damping	$[0.8, 1.2] \times 0.8$	-
Arm Joint Stiffness	$[0.8, 1.2] \times 20$	-
Arm Joint Damping	$[0.8, 1.2] \times 0.5$	-
Initial Joint Positions	$[0.5, 1.5] \times \text{nominal value}$	rad
Link Mass	$[0.8, 1.2] \times \text{nominal value}$	Kg
Payload Mass	$[0, 2]$	Kg
Payload Position	$[-0.05, 0.05]$ relative to base origin	m
Ground Friction	$[0.05, 2.0]$	-
Motor Strength	$[0.8, 1.2]$	-
Observation Delay	$[0, 0.02]$	s

B. 网络

训练框架由一个策略网络和一个评论者网络组成。策略网络作为演员和控制器，仅接收可观察状态作为输入。

轨迹-速度预测策略网络：我们提出了一种轨迹-速度预测策略网络 (TVP)，用于预测不可观察的目标轨迹和身体速度，整合额外的本体感受，并输出关节动作。一个轨迹编码器 E_{θ_1} 从历史姿态帧中提取特征。然后，这些特征被输入到一个解码器 D_{θ_2} 中，以预测后续轨迹序列，如 (2) 所示。这种预测在某些情境中至关重要，特别是在例如遥操作这样的情况下，当仅有历史轨迹记录可用时。另外，在采用使用 DP 的策略时，未来轨迹可以直接提供，并且 TVP 将被跳过。

$$\hat{z}_t^{tcp} = E_{\theta_1}(o_{t-3:t}^{tcp}) \quad \hat{o}_{t+1:t+4}^{tcp} = D_{\theta_2}(\hat{z}_t^{tcp}) \quad (2)$$

我们引入了一个归一化自编码器 (NAE) 来构建轨迹预测器的结构和损失函数。重建损失衡量 NAE 从潜在嵌入 $\hat{z}_t^{tcp}$ 中重建未来轨迹的效果：

$$\mathcal{L}_{NAE} = \mathcal{L}_{rec} = \text{MSE}(\hat{o}_{t+1:t+4}^{tcp}, o_{t+1:t+4}^{tcp}) \quad (3)$$

具体而言，使用 MSE 来最小化预测的未来轨迹与真实值之间的差异。重要的是，我们将 $\tanh$ 激活函数应用于编码的潜在嵌入进行归一化，把它们限制在一个有限范围内。这有助于防止超出预期分布的异常值，从而有助于稳定训练过程。此外，我们更新由 NAE 构建的轨迹预测器，同时与其他网络一起进行。一方面，这有助于准确预测未来轨迹；另一方面，它生成有用的潜在嵌入供低层网络使用，从而促进所需动作的学习。

记忆编码器对最后五帧本体感知观察进行编码。通过监督学习来训练估算器，以估计身体的线速度：

$$\begin{aligned} \hat{z}_t^{prop} &= E_{\theta_3}(o_{t-4:t}) & \hat{v}_t^p &= EST_{\theta_4}(\hat{z}_t^{prop}) \\ \mathcal{L}_{est} &= \text{MSE}(\hat{v}_t^p, v_t^p) \end{aligned} \quad (4)$$

。估算是重要的，因为机动操作任务涉及远距离运动和短距离局部操作。身体的估算线速度有效地捕捉了这两种任务之间的区别，从而使得这些不同的机动操作任务能够更准确地执行。编码后的嵌入与其他本体感知数据和之前的动作一起输入到低级网络中，该网络输出高斯分布 $\mathbf{a}_t^b \sim \mathcal{N}(\boldsymbol{\mu}_t^b, \boldsymbol{\sigma}_t^b)$ 的均值向量 $\boldsymbol{\mu}_t^b \in \mathbb{R}^{18}$ ，其中 $\boldsymbol{\sigma}_t^b \in \mathbb{R}^{18}$ 表示动作方差，方差通过近端策略优化 (PPO) 进行优化。

评论网络：评论网络包括一个轨迹编码器 E_{ψ_1} 和一个高度编码器 E_{ψ_2} ，用于进行降维和表示。 ψ_1 和 ψ_2 代表评论网络参数。这些编码的特征与本体观察、先前动作和特权信息相结合，用于计算状态值函数 V_t ，以评估当前状态的预期奖励：

$$V_t = E_{\psi_3}(E_{\psi_1}(o_{t-3:t+4}^{tcp}) \oplus E_{\psi_2}(s_t^e) \oplus o_t \oplus a_{t-1} \oplus s_t^p) \quad (5)$$

C. 轨迹生成与采样

轨迹生成包括在训练期间学习的轨迹和在部署期间跟踪的轨迹。以下是这些轨迹的生成和采样策略的详细信息。

训练中的轨迹库：我们在全身策略训练中引入了一个轨迹库，其中包含来自多个操作任务的真实世界 TCP 位姿数据。在训练过程中，这些轨迹的初始点被设定为末端执行器的位姿，作为情节重置之后的起始状态，防止在早期阶段出现低质量样本。

我们从若干典型操作任务中选择轨迹用于学习 [10]。每个任务包括 200 个轨迹，例如“清洁桌子”、“倒可乐”、“打开容器”、“拿起小熊”、“拿起杯子”和“拔掉充电器”。可以添加更多的轨迹来扩展多任务学习。我们还结合了来自 [15] 的“推”数据。这些任务涵盖了重复运动、抓取和放置动作、旋转开关以及长距离移动，使它们成为常见现实操作任务的代表。数据被存储为独立的 pickle 文件以供训练。此外，我们发现，在 TCP 轨迹跟踪奖励 r^g 的指导下，机器人在一定程度上可以推广到跟踪其他未见过的轨迹。

轨迹采样更新：不同任务的轨迹复杂程度不同：有些涉及位置和方向的更多变化，而其他的则变化较少。如果在所有任务中对轨迹进行均匀采样，策略可能倾向于专注于跟踪更简单的轨迹，因为它们带来更高的奖励，同时在更具挑战性的任务上表现不佳。这可能导致策略陷入局部最优。

为了解决这个问题，我们提出了一种自适应轨迹采样和奖励因子更新机制，这与整个训练过程中的均匀采样不同。在初始阶段，所有任务的轨迹以相同的概率被采样。同时，记录每个任务的平均跟踪奖励。当某个任务的跟踪奖励发生变化时，其采样概率将相应更新，如下所示，优先选择更具挑战性的轨迹以增强学习。

$$P'_n = \frac{1}{1 + \exp(k_n(\bar{r}_n^g - \lambda_n))} \quad (6)$$

$$P_n = \frac{P'_n}{\sum_{i=1}^N P'_i}, i, n = 1, \dots, N$$

其中 $\bar{r}_n^g$ 表示每个任务轨迹的平均跟踪奖励。为防止遗忘，该机制为每个轨迹保留了一个最小概率。

同时，每个任务轨迹采用其自身的奖励系数因子。具体来说，经过反复训练和调整，建立了一组预定义的平均姿态误差阈值和任务奖励因子，称为 $\{(T_{k,n}, \sigma_{k,n})\}$ 。每个任务的平均姿态误差计算如下，包括位移空间中的欧几里得距离和旋转空间中的角度误差。

$$e_n^p = \|\mathbf{p}_{n,tcp}^{\text{target}} - \mathbf{p}_{n,tcp}\|_2 \quad (7)$$

$$e_n^o = \arccos(0.5 \sum_i (\mathbf{R}_{n,tcp}^{\text{target}} \cdot \mathbf{R}_{n,tcp}^T)_{ii} - 0.5)$$

，其中 σ_n^p 和 σ_n^o 根据 (8) 进行更新。

$$\sigma_n^p = \begin{cases} \sigma_{k,n}^p & \text{if } k = \min \{j : \bar{e}_n^p < T_{j,n}^p \text{ exists}\} \\ \sigma_{max,n}^p & \text{otherwise} \end{cases} \quad (8)$$

$$\sigma_n^o = \begin{cases} \sigma_{k,n}^o & \text{if } k = \min \{j : \bar{e}_n^o < T_{j,n}^o \text{ exists}\} \\ \sigma_{max,n}^o & \text{otherwise} \end{cases}$$

该因子根据每个任务轨迹的奖励值变化自适应更新。随着训练的进行，该因子增加了每个任务的跟踪要求。任务特定奖励因子专门关注于每个单独任务轨迹的跟踪性能。

在部署阶段：在部署过程中，目标轨迹可以通过预先训练的 DP 提供，也可以通过手持夹钳进行人工远程操作提供。

DP 输出的目标轨迹被应用于自动化四足机器人运动操作任务中。策略输入包括多模态数据，例如每个时间步的手臂本体感受和相应的鱼眼图像，TCP 轨迹定义了动作空间。我们使用基于 ViT 的 CLIP 视觉编码器 (ViT Base Patch 16, 输入 224×224) 训练 DP，方法参考了 [30] 和 [31]。使用手持夹具进行遥操作是另一种选择，该选项允许人类直接控制带有手臂的四足机器人的全身行为。如前所述，全身策略可以在合理的操作范围内推广以跟踪先前未见过的轨迹。因此，我们开发了一种遥操作方案以控制四足机器人，扩展其执行全身行为的能力。

手持设备上的 Realsense T265 记录相对于其启动时原点的位姿数据。我们将机器人启动后的初始 TCP 点设置为其轨迹的原点，以确保手持设备和机器人的 TCP 初始点之间的一致性。当操作员移动设备时，记录的历史位姿数据被输入到策略中进行追踪，使得在远程操作时可以实现全身运动行为。

在全身策略训练期间，我们利用 Isaac Gym 从 4,096 个代理并行收集数据。策略每 50 个模拟步骤更新一次，整个训练——在单个 NVIDIA RTX 3090 Ti GPU 上运行——大约需要八小时才能完成 10,000 次迭代。策略控制器的操作频率为 50。我们通过多线程控制实现 DP 或遥操作、全身策略和机器人控制的异步通信，如图 2 所示。

该控制器部署在一个 12 自由度的四足机器人 Uniree Go2 上，该机器人配备有一个 6 自由度的 Airbot Play 机械臂。机械臂的末端执行器安装有一个单自由度的平行夹持器 [10]，如图 2 所示。机器人的总重量为 19.9 kg。机械臂和四足机器人均由 Go2 上的电池供电。整个身体策略的推理在 Go2 的 NVIDIA Jetson Orin NX 上运行，而自动模式下的 DP 则在配备 NVIDIA RTX 3090Ti 的 PC 上进行推理，目标轨迹通过以太网发送。在遥操作模式下，手持式夹持器通过同一链接传输轨迹和夹持器命令。

在部署过程中，需要获取目标姿态和当前 TCP 姿态在世界坐标系中的位置。我们在机器人的机身上安装了 Livox Mid-360 Lidar，并部署了一种里程计 [32] 来构建世界坐标系，其中 Lidar 的初始位置作为原点。在系统运行过程中，通过正向运动学 (FK) 计算出 TCP 的姿态，并将其转换到世界坐标系中。目标姿态也以类似的方式进行处理。

我们通过模拟中进行的一系列消融实验，展示了所提出方法的必要性和有效性。对于具有相同设置的实验，我们进行了 20 次使用不同种子的试验来记录数据。

无 TVP 的策略网络：以研究中基于单个 MLP (sMLP) 网络的 Actor 作为基线，我们验证了 TVP 策略网络对控制的必要性。在没有外部感知的情况下，四足机器人运动依赖本体感觉信息来感知环境。我们的训练结果表明，当

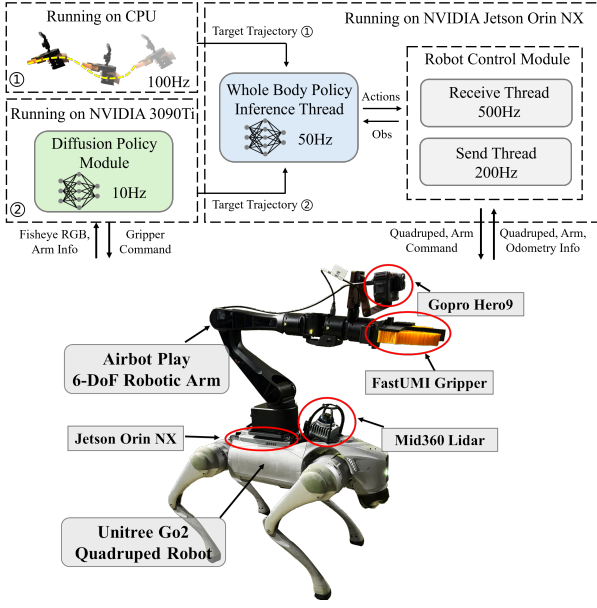


Fig. 2: 系统部署架构：我们开发了一个多计算机多线程系统用于部署。主机电脑运行 DP (10 Hz)，使用 Realsense T265 和 GoPro 生成或捕获手持夹具的轨迹和命令 (100 Hz)。频率影响任务执行速度。这个参考轨迹通过以太网发送到下位机。安装在机器人上的 Jetson Orin NX 负责处理多线程任务，包括接收机器人状态 (500 Hz)、全身策略推断 (50 Hz)，以及发送控制命令 (200 Hz)。

用历史本体感觉信息估计状态的组件被移除，且一个单一的 MLP 将历史本体感觉观测映射到动作时，行走操控任务的任务跟踪奖励在训练中的提高更为缓慢，且收敛于较低水平，如图 ?? 所示。进一步评估表明，这种策略导致了更大的关节扭矩和位置波动，导致关节运动产生更明显的抖动，如图 3 (a) 和 (b) 所示。

TVP 策略网络没有 NAE：当未来目标轨迹不可见（远程操作）时，我们进行了消融实验来评估 NAE 所带来的性能，NAE 可以根据过去轨迹数据预测未来轨迹。在这种设置下，只对历史和当前轨迹进行编码，结果的潜在嵌入直接输入到低层网络中生成动作输出。因为策略缺乏后续追踪轨迹的信息，其追踪性能恶化：训练期间，追踪奖励始终低于装备了基于 NAE 预测机制的策略所取得的奖励，如图 ?? 所示。图 3 (c) 展示了通过 NAE 解码器预测的未来轨迹与其真实值的 MSE。重建误差由位置和姿态矩阵的前两行表示，针对四个帧进行绘制并以热图展示。结果表明重构误差极小，这表明 NAE 的潜在嵌入包含预测未来轨迹所需的信息。

表格 III 展示了七个任务的政策评估。在经过 800 个时间步骤后，记录了末端执行器轨迹跟踪的位姿误差。位置误差使用欧几里得距离（厘米）测量，而方向误差是指末端执行器与目标轨迹方向之间的最小旋转角度（弧度）。我们计算所有时间步骤的平均误差以评估整体跟踪性能。结果表明，具有未来轨迹预测机制的 TVP 优于其他政策，而单一基于 MLP 的政策由于其有限的信息提取和表示能力表现最差。

自适应轨迹采样：在此实验中，我们设置所有机器人在整个训练过程中均匀采样（统一采样）所有任务轨迹，以评估我们自适应轨迹采样策略的必要性。我们记录了训练过程中两个典型任务的奖励曲线，如图 4 所示。结果表明，

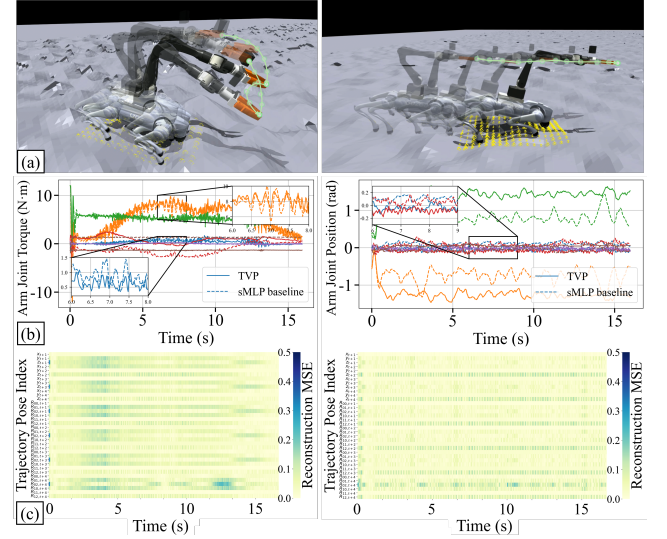


Fig. 3: 策略网络消融分析：(a) 仿真中“拔掉充电器”和“推动”动作序列的示意图。(b) 在任务期间，TVP 和单一 MLP 网络的手臂关节力矩和位置曲线，sMLP 显示出更多的抖动。(c) 通过 TVP 的 NAE 重建未来轨迹的 MSE 热图，两项任务均适用。

均匀采样无法有效平衡任务性能。最初，不同姿态变化任务的奖励曲线接近，但随着时间推移，奖励出现分歧，要求更多姿态变化的任务因为注意力不足（即较高采样率）而收敛到更低的奖励。相比之下，自适应采样在各任务之间维持了相似的奖励，平衡了跟踪性能。表 III 表明，对于轨迹变化更频繁的任务（例如，“打开容器”），跟踪精度明显下降。

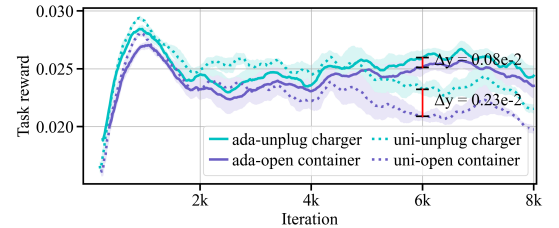


Fig. 4: 两种任务在自适应采样和均匀采样下的轨迹跟踪奖励。

D. 实际世界实验

我们通过实际实验进一步验证了我们方法的可行性和有效性。实验包括通过手持夹持器进行远程操作和基于 DP 的推理执行。所有任务都是使用统一训练的全身策略完成的。我们的方法在所有场景中展示了出色的目标处理性能，实现了从仿真到现实世界的零次转移。请参阅补充材料中的视频。

手持夹具遥操作：图 ?? I-V 展示了几个使用遥操作进行全身移动操作的实际演示，例如清理纸屑、捡拾垃圾、存放玩具、拔下充电器以及移除螺栓来打开门。图 ?? III 演示了一个远程任务。可以观察到，当执行底层任务时，例如捡拾垃圾，四足机器人通过向下倾斜来辅助手臂，使末端执行器能够达到更低的位置。相反，身体高度以上的任务需要抬高身体。当目标位置远超出当前工作空间时，四足机器人利用其运动能力重新定位自身，从而到达遥远的地点。这些行为展示了卓越的全身协调能力。我们还评估

TABLE III: 模拟消融研究：每个任务的平均 6D 跟踪误差及所有任务中这些平均误差的方差。指标：欧氏距离/旋转矩阵角度差。(厘米/弧度)

	pushing	unplug charger	pick bear	pick cup	open container	clean table	pour coke	Variance
TVP (Ours)	1.32 / 0.08	1.08 / 0.07	0.87 / 0.11	0.66 / 0.05	0.74 / 0.06	0.94 / 0.09	1.07 / 0.15	0.050 / 0.0012
TVP wo NAE	1.78/0.11	1.49/0.07	1.30/0.13	1.26/0.07	1.43/ 0.05	1.31/0.10	1.78/0.22	-
sMLP (baseline)	2.23/0.12	2.02/0.14	2.31/0.37	2.24/0.14	2.20/0.13	2.17/0.22	3.13/0.54	-
Uni-sample	1.30/0.07	1.11/0.08	0.96/0.13	0.79/0.06	1.69/0.10	1.18/0.11	1.41/0.21	0.088/0.0026

了训练任务，并将追踪性能记录在表 IV 中。结果表明，实际部署的追踪性能与模拟的追踪性能相当。

TABLE IV: 真实环境实验结果。(厘米/弧度/%)

		EE pos err	EE rot err	Success Rate
Teleoperation	unplug charger	1.39	0.054	98
	open container	1.02	0.047	100
DP	unplug charger	1.37	0.039	80
	open container	1.24	0.036	85

通过深度预测的推理：如图 ?? VI-VII 所示，我们展示了该方法通过预训练的深度预测自主执行任务的能力。具体来说，我们将安装在机器人手臂腕部的 GoPro 摄像头获取的信息，以及手臂的状态输入到深度预测中。然后，深度预测会自主生成目标物体的轨迹。随后，使用强化学习的全身策略来控制关节执行操作。其性能主要受限于深度预测推理的结果。实验结果表明，如表 IV 所示，我们的方法可以利用深度预测来规划并自主执行多任务操作。

III. 结论

我们提出了一种强化学习框架 MLM，该框架利用真实世界和模拟数据来解决四足机器人在多任务全身行走操控中的问题。为了在学习多个任务轨迹时平衡性能，我们集成了一个具有自适应采样机制的数据库。为了解决在远程操作中只能观察到历史轨迹的情况，我们提出了一种轨迹速度预测策略网络来预测未来轨迹，从而提高策略执行性能。速度估计支持跨不同空间范围的任务执行。最后，在 Go2 和 Airbot Play 机器人上的模拟和实际实验突出了所提出方法的良好性能。事实上，我们的方法可以扩展使四足机器人在新环境中执行更多任务，充分利用它们的行走和操控能力。这取决于更高级别的决策和规划。未来视觉语言模型和导航的集成可能会解决这一限制。

REFERENCES

- [1] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, p. eabc5986, 2020.
- [2] J. Wu, G. Xin, C. Qi, and Y. Xue, "Learning robust and agile legged locomotion using adversarial motion priors," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4975–4982, 2023.
- [3] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," in *Proceedings of Robotics: Science and Systems*, 2021.
- [4] G. Ji, J. Mun, H. Kim, and J. Hwangbo, "Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4630–4637, 2022.
- [5] Y. Li, J. Li, W. Fu, and Y. Wu, "Learning agile bipedal motions on a quadrupedal robot," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 9735–9742.
- [6] T. Peng, L. Bao, J. Humphreys, A. M. Delfaki, D. Kanoulas, and C. Zhou, "Learning bipedal walking on a quadruped robot via adversarial motion priors," in *Annual Conference Towards Autonomous Robotic Systems*, 2024, pp. 118–129.
- [7] J.-P. Sleiman, F. Farshidian, and M. Hutter, "Versatile multicontact planning and control for legged loco-manipulation," *Science Robotics*, vol. 8, no. 81, p. eadg5014, 2023.
- [8] N. Yokoyama, A. Clegg, J. Truong, E. Undersander, T.-Y. Yang, S. Arnaud, S. Ha, D. Batra, and A. Rai, "Asc: Adaptive skill coordination for robotic mobile manipulation," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 779–786, 2024.
- [9] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Proceedings of Robotics: Science and Systems*, 2024.
- [10] Zhaxizhuoma, K. Liu, C. Guan, Z. Jia, Z. Wu, X. Liu, T. Wang, S. Liang, P. Chen, P. Zhang, H. Song, D. Qu, D. Wang, Z. Wang, N. Cao, Y. Ding, B. Zhao, and X. Li, "Fastumi: A scalable and hardware-independent universal manipulation interface with dataset," *arXiv preprint arXiv:2409.19499*, 2025.
- [11] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in *Proceedings of Robotics: Science and Systems*, 2023.
- [12] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "RT-1: Robotics Transformer for Real-World Control at Scale," in *Proceedings of Robotics: Science and Systems*, 2023.
- [13] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 15 031–15 038.
- [14] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *arXiv preprint arXiv:2502.01143*, 2025.
- [15] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "Umi-on-legs: Making manipulation policies mobile with manipulation-centric whole-body controllers," in *Proceedings of the 8th Conference on Robot Learning*, vol. 270, 2024, pp. 5254–5270.
- [16] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Proceedings of the 5th Conference on Robot Learning*, vol. 164, 2022, pp. 91–100.
- [17] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, p. eaau5872, 2019.
- [18] A. Escontrela, X. B. Peng, W. Yu, T. Zhang, A. Iscen, K. Goldberg, and P. Abbeel, "Adversarial motion priors make good substitutes for complex reward functions," in *IEEE International Conference on Intelligent Robots and Systems*, 2022, pp. 25–32.
- [19] X. Liu, J. Wu, Y. Xue, C. Qi, G. Xin, and F. Gao, "Skill latent space based multitask learning for a legged robot," *IEEE Transactions on Industrial Electronics*, vol. 72, no. 2, pp. 1743–1752, 2025.
- [20] F. Shi, C. Zhang, T. Miki, J. Lee, M. Hutter, and S. Coros, "Rethinking Robustness Assessment: Adversarial Attacks on Learning-based Quadrupedal Locomotion Controllers," in *Proceedings of Robotics: Science and Systems*, 2024.
- [21] D. Hoeller, N. Rudin, D. Sako, and M. Hutter, "Anymal parkour: Learning agile navigation for quadrupedal robots," *Science Robotics*, vol. 9, no. 88, p. eadi7566, 2024.

- [22] T. Miki, J. Lee, L. Wellhausen, and M. Hutter, “Learning to walk in confined spaces using 3d representation,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 8649–8656.
- [23] T. He, C. Zhang, W. Xiao, G. He, C. Liu, and G. Shi, “Agile But Safe: Learning Collision-Free High-Speed Legged Locomotion,” in *Proceedings of Robotics: Science and Systems*, 2024.
- [24] C. D. Bellicoso, K. Krämer, M. Stäuble, D. Sako, F. Jenelten, M. Bjelonic, and M. Hutter, “Alma - articulated locomotion and manipulation for a torque-controllable robot,” in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8477–8483.
- [25] Y. Ma, F. Farshidian, T. Miki, J. Lee, and M. Hutter, “Combining learning-based locomotion policy with model-based manipulation for legged mobile manipulators,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2377–2384, 2022.
- [26] Z. Fu, X. Cheng, and D. Pathak, “Deep whole-body control: Learning a unified policy for manipulation and locomotion,” in *Proceedings of the 6th Conference on Robot Learning*, vol. 205, 2023, pp. 138–149.
- [27] K. Jiang, Z. Fu, J. Guo, W. Zhang, and H. Chen, “Learning whole-body loco-manipulation for omni-directional task space pose tracking with a wheeled-quadrupedal-manipulator,” *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 1481–1488, 2025.
- [28] M. Liu, Z. Chen, X. Cheng, Y. Ji, R. Qiu, R. Yang, and X. Wang, “Visual whole-body control for legged loco-manipulation,” in *Proceedings of the 8th Conference on Robot Learning*, 2024.
- [29] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, “Learning robust perceptive locomotion for quadrupedal robots in the wild,” *Science Robotics*, vol. 7, no. 62, p. eabk2822, 2022.
- [30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.
- [31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.
- [32] K. Chen, R. Nemiroff, and B. T. Lopez, “Direct lidar-inertial odometry: Lightweight lio with continuous-time motion correction,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 3983–3989.