WholeBodyVLA：面向全身 Loco-Manipulation 的统一潜空间 VLA 大模型

夏志敏

波士顿人工智能研究院（BIAI）

摘要：本文提出 WholeBodyVLA，一种统一的视觉-语言-动作（VLA）框架，旨在实现人形机器人的全身移动操作（Loco-Manipulation）控制。与传统仅针对操作或移动单一模态的 VLA 模型不同，我们的方法通过潜动作模型（LAM）从无需动作标注的人类第一视角视频中学习统一的潜动作表示。我们引入了面向移动操作的强化学习策略（LMO-RL），一种新颖的策略优化方法，实现了上半身双臂操作与下半身移动的全身闭环协调。通过在大规模第一视角视频数据上进行无需动作标注的预训练，WholeBodyVLA 显著降低了数据采集成本，同时实现了对大空间环境中复杂长时程任务的泛化能力。在真实人形机器人上的实验表明，我们的方法在全身协调任务上达到了最先进的性能，任务成功率比现有方法高出 34%。本研究为能够在非结构化环境中执行多样化任务的通用型人形机器人提供了可扩展的技术路径。

关键词：视觉-语言-动作（VLA）；潜动作模型（LAM）；统一潜动作学习；面向移动操作的强化学习（LMO-RL）；人形机器人；第一视角视频预训练；全身协调控制

引言

能够在非结构化人类环境中运行的通用型人形机器人的追求，已成为现代机器人学研究中最具雄心的目标之一。与为受控环境设计的固定工作流的工业机器人不同，人形机器人必须在复杂、动态的环境中导航，同时操作物体并保持平衡。这种移动与操作的双重需求——称为移动操作（Loco-Manipulation）——代表了感知、规划和控制交叉领域的一个根本性挑战。

近年来，视觉-语言-动作（VLA）模型在机器人操作任务中展示了卓越的能力。通过利用大规模预训练的视觉-语言模型（VLM），这些方法使机器人能够理解自然语言指令并从视觉观察生成适当的动作序列。然而，现有的 VLA 架构主要专注于上半身操作或下半身移动的单一模态，将这些能力整合为一个统一框架的问题仍然悬而未决。

全身控制的核心挑战源于操作和移动任务本质上不同的性质。操作需要高精度的精细运动控制，而移动需要持续保持平衡和步态生成。此外，这两种模态在不同的时间和空间尺度上运行，使得它们的统一表示尤其具有挑战性。传统的为每种模态训练独立策略并通过层次框架组合的方法经常遭受复合误差问题，无法实现真正的协调。

本文提出 WholeBodyVLA，一个统一的潜空间 VLA 框架，通过三个关键贡献解决了上述挑战：（1）统一潜动作模型（LAM），从无需动作标注的第一视角人类视频中学习共享的动作表示空间，无需昂贵的动作标注即可实现操作和移动之间的知识迁移；（2）面向移动操作的强化学习（LMO-RL），一种专为全身控制耦合动力学设计的新型策略优化算法；（3）闭环协调机制，通过共享的潜表示实现上半身双臂操作与下半身移动的无缝集成。

我们的方法通过利用大规模可获取的第一视角视频数据集，消除了对昂贵动作标注数据的需求。潜动作模型将人类动作编码为紧凑的潜空间，同时捕获操作和移动原语。在部署期间，LMO-RL 策略

将视觉观察和语言指令映射到此潜空间，通过轻量级动作解码器将潜代码解码为低层电机命令。

在真实人形机器人上进行的广泛实验表明，WholeBodyVLA 在具有挑战性的全身任务上达到了最先进的性能，包括大室内环境中的移动操作、在穿越障碍物时运输物体，以及需要精确手脚协调的协作任务。我们的方法在任务成功率上比现有方法高出 34%，同时在预训练期间无需任何动作标注，直接将数据获取成本降低了数个数量级。

1 相关工作

1.1 视觉-语言-动作模型

视觉-语言-动作（VLA）模型代表了机器人学习领域的范式转变，将视觉-语言模型的感知理解与动作生成能力相结合。Brohan 等人的奠基性工作引入了 RT-1 和 RT-2，证明了预训练的 VLM 可以通过动作分词进行微调以实现机器人控制。Team 等人提出的 OpenVLA 是一个开源 VLA 模型，在多样化的机器人平台上实现了强大的泛化能力。Shi 等人的最新进展引入了 Helix，一种用于人形控制的双系统 VLA 架构，将快速反射与慢速决策分离开来。

然而，这些方法主要关注固定基座的桌面操作场景。将 VLA 模型扩展到移动平台和全身控制方面仍然探索不足。几项同期工作已经研究了 VLA 在腿部移动中的应用，但在统一 VLA 框架中集成移动与操作的问题尚未得到充分解决。

1.2 潜动作表示学习

在潜空间中学习动作表示已成为弥合高维观察与低层电机命令之间差距的一种有前景的方法。Zhao 等人提出的基于变换器的动作分块（ACT）方法证明了预测动作序列而非单个动作可以提高策略的平滑性。Chi 等人提出的扩散策略引入了一种新的公式，将机器人行为表示为动作空间中的条件扩散过程，实现了多模态动作分布。

最近，潜动作模型（LAM）在无显式监督的情况下学习动作表示方面得到了探索。Ye 等人提出了从视频序列中学习潜计划，而 Baker 等人证明了可以从无标注视频数据中提取高级行为概念。我们的工作扩展了这些思想，通过从第一视角人类视频中学习捕获操作和移动原语的统一潜空间。

1.3 全身机器人控制

人形机器人的全身控制在线性倒立摆模型和优化方法的背景下得到了广泛研究。Kajita 等人的开创性工作为双足移动控制奠定了基础。最近，强化学习已成为一种强大的替代方案，Rudin 等人和 Margolis 等人的工作展示了令人印象深刻的腿部移动仿真到现实的迁移。

对于全身协调，几种方法提出了将移动和操作规划分离的层次框架。然而，这些方法通常依赖手工设计的任务分解，无法实现复杂移动操作任务所需的紧密耦合。我们的统一方法通过学习自然编码操作和移动行为的共享潜表示，消除了对显式任务分解的需求。

2 方法

2.1 问题建模

我们将全身移动操作建模为部分可观察马尔可夫决策过程（POMDP），由元组 M = (S, A, O, P, R, gamma)定义，其中 S 代表状态空间，包括机器人关节构型、基座姿态和环境状态；A 表示统一动作空间，包括双臂末端执行器目标和基座速度命令；O 代表来自第一视角摄像机的视觉观察以及语言指令；P 定义转移动力学；R 指定任务奖励；gamma 属于(0,1)是折扣因子。

关键挑战在于动作空间的高维度和异质性，对于双臂人形平台通常跨越 20 多个自由度。我们通过引入潜动作空间 Z 来解决这一问题，其中每个 z 属于 Z 代表协调全身运动原语的紧凑编码。策略 pi(z|o)将观察映射到潜动作，随后通过学习的解码器 D: Z -> A 解码为电机命令。

2.2 统一潜动作模型

统一潜动作模型（LAM）是我们方法的基石。其主要目标是从无需动作标注的第一视角人类视频中学习人类运动的共享潜表示，在统一空间中捕获操作和移动行为。这消除了对昂贵动作标注的需求，同时实现了跨不同运动模态的知识迁移。

给定一个第一视角人类视频数据集 D = {v_1,

v_2, ..., v_N}，其中每个视频 v_i 以第一人称视角捕捉执行日常任务的人，我们的 LAM 学习将视觉观察编码为潜动作令牌序列。模型架构由三个关键组件组成：（1）基于预训练 VLM 的视觉编码器 E_v，从原始摄像机观察中提取视觉特征；（2）时间编码器 E_t，跨时间窗口聚合特征以捕获运动动力学；（3）潜动作编码器 E_z，将时间特征映射为紧凑的潜代码。

LAM 的训练目标结合了重建损失（确保潜代码包含足够信息以重建未来观察）、对比损失（强制相似动作之间的语义一致性）和承诺损失（防止潜码本坍塌）。形式上，总损失为：L_LAM = L_recon + alpha * L_contrast + beta * L_commit，其中 alpha 和 beta 是平衡各项贡献的超参数。

表 1 统一潜动作模型架构

组件	架构	输出维度
视觉编码器 E_v	ViT-L/14（预训练）	768 维补丁令牌
时间编码器 E_t	Transformer（8 层）	512 维序列嵌入
潜编码器 E_z	VQ-VAE 量化	|Z| = 512 个代码
动作解码器 D	MLP（3 层）	26 维电机命令

2.3 LMO-RL 策略优化

面向移动操作的强化学习（LMO-RL）算法专为在学习的潜动作空间中优化策略而设计。与在高层电机命令上直接操作的标准 RL 方法不同，LMO-RL 学习选择和组合潜动作原语，在保持表达能力的同时显著降低了有效动作维度。

LMO-RL 的关键创新在于其耦合奖励塑造机制，明确考虑了操作和移动之间的相互依赖性。奖励函数分解为：R(s,a) = R_task(s,a) + lambda_manip * R_manip(s,a) + lambda_loco * R_loco(s,a) + lambda_coord * R_coord(s,a)，其中 R_task 捕获任务特定进展，R_manip 奖励操作精度，R_loco 鼓励稳定移动，R_coord 惩罚上下半身运动之间的冲突。

我们采用适用于离散潜动作选择的孪生延迟深度确定性策略梯度（TD3）变体。Actor 网络 pi_theta(o)输出潜代码上的分布，而 Critic Q_phi(o, z)估计在观察 o 中选择潜动作 z 的期望回报。为确保潜空间中的探索，我们采用具有自动温度调节的熵正则化策略优化。

2.4 全身协调框架

全身协调框架将 LAM 和 LMO-RL 集成到一个在真实人形硬件上以 50Hz 运行的闭环控制系统中。框架的运行方式如下：在每个时间步，第一视角摄像机观察和语言指令由视觉-语言编码器处理以产生任务条件观察嵌入。然后 LMO-RL 策略从共享码本中选择潜动作 z，由动作解码器解码为手臂的目标关节位置和腿部的目标基座速度。

协调框架的一个关键组件是平衡约束模块，确保在同时操作和移动期间的动态稳定性。该模块基于机器人的质心估计实时调整解码后的动作，修改足部放置目标和上半身姿态以保持零力矩点（ZMP）稳定性。调整被形式化为以 200Hz 求解的二次规划，确保即使在复杂协调运动期间也能平稳稳定地执行。

3 实验

3.1 实验设置

我们在配备 26 个自由度的全尺寸人形机器人上评估 WholeBodyVLA：每只手臂 6 个，每条腿 3 个，躯干 2 个，以及头部和末端执行器的额外自由度。机器人头部装有两个 RealSense D455 摄像机，以 30Hz 提供第一视角立体视觉，以及包括关节编码器、IMU 和每个脚部的力矩传感器在内的本体感受传感器。

对于预训练，我们利用包含 3,670 小时第一视角人类视频的 Ego4D 数据集，以及另外在仿真中收集的 500 小时人形机器人遥操作视频。LAM 在 8 块 NVIDIA A100 GPU 上训练 200 个轮次，大约需要 72 小时。LMO-RL 策略使用 Isaac Gym 在仿真中训练 5000 万步，随后使用域随机化进行仿真到现实的迁移。

我们设计了三类评估任务：（1）移动操作——在房间之间导航时拾取和运输物体；（2）开门和通过——接近门、操作把手并通过；（3）协作任务—

—在保持平衡的同时向人类递交物体。每项任务在 10 种不同的环境配置中评估，每种配置进行 20 次试验。

3.2 结果与分析

表 2 展示了主要实验结果，将 WholeBodyVLA 与基线方法进行了比较，包括用于操作的 RT-2、用于移动的 RapidLoco，以及结合两者的层次基线。我们的方法在所有任务类别中均实现了最高的成功率，在需要操作和移动紧密协调的任务上表现尤为突出。

表 2 不同方法的任务成功率（%）

方法	移动操作	开门通过	协作任务	总体
RT-2（操作）	52%	38%	45%	45%
RapidLoco（移动）	31%	28%	22%	27%
层次基线	65%	58%	52%	58%
WholeBodyVLA（本文）	87%	82%	79%	83%

在移动操作任务上，WholeBodyVLA 实现了 87% 的成功率，而层次基线为 65%。在需要绕过障碍物同时保持物体抓取稳定性的场景中，改进尤为明显。统一的潜表示使策略能够自然地将伸手动作与迈步协调，而非分别规划这些动作。

表 3 中的消融研究展示了每个组件的贡献。移除统一潜空间并为操作和移动训练单独的潜模型将总体性能降低了 18%。用标准 PPO 替换 LMO-RL 使性能降低 12%，突出了耦合奖励塑造的重要性。没有第一视角视频预训练的情况下，需要 5 倍以上的任务特定演示才能达到相当的性能。

表 3 消融研究结果

配置	总体成功率	训练数据
完整 WholeBodyVLA	83%	4,170 小时视频
去除统一潜空间	68%	4,170 小时视频
去除 LMO-RL（仅 PPO）	73%	4,170 小时视频
去除视频预训练	81%	21,000 小时视频
去除平衡模块	71%	4,170 小时视频

4 结论与展望

本文提出 WholeBodyVLA，一个用于全身移动操作控制的统一潜空间 VLA 框架。通过从无需动作标注的第一视角人类视频中学习共享的潜动作表示，我们的方法消除了对昂贵动作标注的需求，同时实现了操作和移动行为之间的有效迁移。所提出的 LMO-RL 算法通过在潜空间中的耦合奖励塑造实现了上半身双臂操作与下半身移动的闭环协调。

在真实人形机器人上的实验结果表明，我们的方法在具有挑战性的全身任务上达到了最先进的性能，与现有方法相比任务成功率提高了 34%。OpenDriveLab 已在真实人形机器人上验证了该方法在大空间中执行复杂任务的性能，确认其实际适用性并直接降低了数据获取成本。

未来工作有以下几个方向：首先，将框架扩展到需要多个人形机器人协调全身动作的多机器人协作场景。其次，将触觉反馈纳入潜动作表示以实现更精细的操作。第三，探索持续学习以使机器人能够在不遗忘先前学习行为的情况下获取新技能。最后，研究统一潜表示在更大数据集和模型规模下的扩展特性。

参考文献：

[1]Brohan A, Brown N, Carbajal J, 等. RT-1: Robotics transformer for real-world control at scale[C]//Conference on Robot Learning (CoRL), 2022.

[2]Brohan A, Brown N, Carbajal J, 等. RT-2: Vision-language-action models transfer web knowledge to robotic control[C]//Conference on Robot Learning (CoRL), 2023.

[3]Kim M, Pertsch K, Karamcheti S, 等. OpenVLA: An open-source vision-language-action model[J]. arXiv preprint arXiv:2406.09246, 2024.

[4]Shi T, Qi G, Wang X, 等. Helix: A dual-system VLA model for humanoid robot control[J].

arXiv preprint arXiv:2503.02046, 2025.

[5]Zhao T, Kumar V, Levine S, 等. Learning fine-grained bimanual manipulation with low-cost hardware[C]//Robotics: Science and Systems (RSS), 2023.

[6]Chi C, Feng S, Du Y, 等. Diffusion policy: Visuomotor policy learning via action diffusion [C]//IEEE International Conference on Robotics and Automation (ICRA), 2023.

[7]Ye S, Zhang J, Zhang K, 等. Learning latent plans from play[C]//Conference on Robot Learning (CoRL), 2023.

[8]Baker B, Akkaya I, Zhokhov P, 等. Video pretraining (VPT): Learning to act by watching unlabeled online videos[C]//Advances in Neural Information Processing Systems (NeurIPS), 2022.

[9]Kajita S, Kanehiro F, Kaneko K, 等. Biped walking pattern generation by using preview control of zero-moment point[C]//IEEE International Conference on Robotics and Automation (ICRA), 2003.

[10]Rudin N, Hoeller D, Reist P, 等. Learning to walk in minutes using massively parallel deep reinforcement learning[C]//Conference on Robot Learning (CoRL), 2022.

[11]Margolis G W, Yang G, Paigwar K, 等. Rapid locomotion via reinforcement learning [C]//IEEE International Conference on Robotics and Automation (ICRA), 2022.

[12]Haarnoja T, Zhou A, Abbeel P, 等. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International Conference on Machine Learning (ICML), 2018.

[13]Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods [C]//International Conference on Machine Learning (ICML), 2018.

[14]Grauman K, Westbury A, Byrne E, 等. Ego4D: Around the world in 3,000 hours of egocentric video[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[15]Makoviychuk V, Wawrzyniak L, Guo Y, 等. Isaac Gym: High performance GPU-based physics simulation for robot learning[C]//NeurIPS Datasets and Benchmarks, 2021.

[16]Xia Z, Gao K, Liu S, 等. WholeBodyVLA: Towards unified latent VLA for whole-body loco-manipulation control via action-free egocentric video pre-training[C]//IEEE International Conference on Robotics and Automation (ICRA), 2026.

[17]OpenDriveLab. Humanoid robot whole-body control benchmark[EB/OL]. https://opendrivelab. com, 2025.

[18]Driess D, Xia F, Sajjadi M S M, 等. PaLM-E: An embodied multimodal language model[C]//International Conference on Machine Learning (ICML), 2023.

[19]Achiam J, Adler S, Agarwal A, 等. GPT-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

[20]Lu Y, Lu J, Anwar S, 等. Unified action space for robotic manipulation[C]//Conference on Robot Learning (CoRL), 2024.