文章编号:1007－6735(2026)04－0407－13　DOI: 10.13255/j.cnki.jusst.20260130001

# 速度目标导向的人形机器人风格化运动策略

马晓航[1]，　梁志远[2,3]，　甘文聪[2,3]，　朱玉迪[1,2]，　李清都[1,2,3]
（1. 中豫具身智能实验室，郑州 450046；2. 上海理工大学 机器智能研究院，上海 200093；
3. 上海卓益得机器人有限公司，上海 200082）

**摘要：**通过模仿学习赋予人形机器人拟人化运动技能，是构建多样化、高表现力机器人运动能力的重要途径。当前基于模仿学习的方法仍面临显著局限：以对抗运动先验（adversarial motion priors, AMP）为代表的对抗模仿方法虽能保留整体运动风格，却易丢失精细动作细节；而以 Mimic 为代表的精确模仿方法虽可高保真复现参考动作，却因策略与参考动作强耦合，难以响应实时控制指令。为此，提出一种新颖的模仿学习训练框架，通过观测空间与奖励机制的协同设计，显式解耦运动风格与任务目标，并实现二者高效协同。所提方法仅依赖高层速度指令驱动，成功将策略从“动作复现器”转变为“速度指令控制器”。实验表明，该策略在任意速度指令下均能实现高精度速度跟踪（整体误差< 0.13 m/s 或 rad/s），同时关键部位的位置和姿态相对原始参考动作的平均动态时间规整（dynamic time warping, DTW）距离分别为 0.837 与 1.135，充分验证了其在精准响应任务指令的同时，有效保留了参考运动的自然风格特征。

**关键词：**人形机器人；模仿学习；强化学习；风格化速度追踪器
**中图分类号：**TP 242.6　　**文献标志码：**A

# Velocity-guided stylized locomotion for humanoid robots

Ma Xiaohang[1], Liang Zhiyuan[2,3], Gan Wencong[2,3], Zhu Yudi[1,2], Li Qingdu[1,2,3]
(1. *Zhongyu Embodied Artificial Intelligence Laboratory, Zhengzhou 450046, China*; 2. *Institute of Machine Intelligence, University of Shanghai for Science and Technology, Shanghai 200093, China*; 3. *Shanghai Droid Robotics Co., Ltd., Shanghai 200082, China*)

**Abstract:** Enabling humanoid robots to acquire human-like locomotion skills through imitation learning is a key approach to building diverse and expressive motor capabilities. Current imitation learning approaches remain fundamentally limited: adversarial methods such as (adversarial motion priors) AMP preserve global motion characteristics but often lose fine-grained details, while precise imitation

**收稿日期:** 2026−01−30
**基金项目:** 上海市促进产业高质量发展专项资金—先导产业创新发展项目（RZ-RGZN-01-25-0673）
**第一作者:** 马晓航（1993—），男，硕士研究生．研究方向：机器人运控算法．E-mail：ma-mark@outlook.com
**通信作者:** 李清都（1980—），男，教授．研究方向：行走机器人理论与技术．E-mail：liqd@usst.edu.cn

**引文格式:** 马晓航，梁志远，甘文聪，等．速度目标导向的人形机器人风格化运动策略[J]．上海理工大学学报，2026，48（4）：407-419.
**Citation:** Ma Xiaohang, Liang Zhiyuan, Gan Wencong, et al. Velocity-guided stylized locomotion for humanoid robots[J]. Journal of University of Shanghai for Science and Technology, 2026, 48（4）：407-419.

frameworks like Mimic faithfully reproduce reference motions yet suffer from strong coupling between policy and demonstration, rendering them incapable of responding to real-time control commands. To address this trade-off, we proposed a novel imitation learning architecture through a principled redesign of the observation space and reward structure, explicitly disentangling motion style from task objectives while enabling their effective coordination. The resulting policy operates solely on high-level velocity commands—effectively transforming the original "motion reproducer" into a "velocity commanded controller." Experiments demonstrate that our method achieves accurate velocity tracking (overall error<0.13 m/s or rad/s) under arbitrary speed commands, while preserving stylistic fidelity: the average dynamic time warping (DTW) distances between generated and reference motions are 0.837 for key joint positions and 1.135 for orientations. These results confirm that our framework successfully reconciles precise task execution with natural motion style reproduction.

**Keywords:** *humanoid robots*; *imitation learning*; *reinforcement learning*; *stylized velocity tracker*

当前，任务性能与运动风格保真度之间的权衡是人形机器人运动控制策略中普遍面临的难题。以任务目标为导向的强化学习方法，尽管能够通过优化速度跟踪、能量效率或姿态稳定性等指标，生成高鲁棒性的控制策略，但其生成的步态常常缺乏自然性与流畅性，无法充分模拟生物运动的特点。相对而言，精确轨迹跟踪方法，如Mimic，能够高保真地复现参考动作，但由于策略与参考轨迹强耦合，难以实时响应任务指令。而BeyondMimic框架尽管尝试通过嵌套扩散模型引入目标条件以解决这一问题，但由于其高模型复杂度和推理延迟，无法满足实时应用需求，且存在运动风格退化的问题。对抗运动先验（adversarial motion priors, AMP）等对抗式模仿学习方法虽然在一定程度上能协同任务执行与风格保真，但容易陷入风格平均化与模式崩溃，难以保留精细动作特征。

为突破上述限制，本文提出了一种创新的模仿学习训练框架，旨在实现参考运动与任务目标的显式解耦与高效协同。该框架的核心在于对Actor-Critic架构中的观测空间设计进行重构：Actor网络仅接收高层速度指令、本体感知信息及其历史观测，彻底摆脱对参考动作的依赖；而Critic网络则保留完整的参考运动信息（包括历史与未来，共20帧），通过时序窗口隐式建模速度指令与参考风格之间的映射关系。基于此，框架结合了风格保持与速度跟踪的复合奖励机制，引导策略在训练过程中学习"由速度驱动、向风格对齐"的运动生成范式。实验结果表明，在不依赖参考动作作为在线输入的条件下，所训练的策略能够在响应任意速度指令的同时，有效复现参考运动的自然风格特征，其关键部位的位置、姿态与原始参考动作的平均动态时间规整（dynamic time warping, DTW）距离分别为0.837和1.135，显著优于AMP（分别为1.068和1.376），且接近Mimic（0.821和1.012）。

## 1 相关工作

传统的人形机器人步态控制多依赖于基于模型的控制方法，如零力矩点（zero moment point, ZMP）、线性二次型调节器（linear quadratic regulator, LQR）、模型预测控制（model predictive control, MPC）等。这类方法通过精确的动力学建模和轨迹规划实现稳定行走，具有良好的理论保证和实时性。然而，其性能高度依赖于模型准确性，在面对复杂地形、外部扰动或未建模动态环境时，鲁棒性显著下降。更重要的是，传统方法生成的步态往往呈现机械、僵硬的特征，难以复现人类运动中蕴含的自然协调性与风格多样性，限制了机器人在人机共融场景中的表现力与亲和力。

为克服上述局限，强化学习（reinforcement learning, RL）因其无需显式动力学模型、可通过试错自主优化策略等优势，逐渐成为足式机器人控制的主流范式。随着深度学习、高保真仿真及域随机化技术的发展，深度强化学习（deep reinforcement learning, DRL）成功实现了多种复杂步态的自主学习，并展现出优异的抗扰能力与环境适应性[1-3]。

然而，纯奖励驱动的强化学习策略通常仅优化任务目标（如目标速度、能耗），所生成的步态虽能高效鲁棒地达成任务目标，却普遍存在动作模式单一、缺乏生物合理性与自然协调性的问题。

在此背景下，模仿学习（imitation learning, IL）方法应运而生，旨在通过利用人类动作捕捉数据，引导策略学习兼具功能性与自然性的类人步态。相比强化学习，模仿学习通过引入专家示范作为先验知识，显著降低了策略搜索空间，提升了训练效率，并能有效复现参考动作中的精细时空结构，从而生成更流畅、更具生物合理性的运动。当前主流模仿学习方法主要分为两类。

### 1.1　Mimic 类模仿学习

Mimic 类模仿学习方法主要通过监督学习直接跟踪人类运动参考轨迹，减少奖励工程负担，专注于复现自然且流畅的机器人动作。该方法的核心是将人类动作捕捉数据重映射到机器人关节空间，使动作捕捉数据中人体骨骼关键点与机器人对应肢体在空间上按比例缩放并严格对齐，然后通过最小化关节角度和速度的误差来优化策略。2018 年，Peng 等[4] 提出的 DeepMimic 框架成为这一领域的里程碑工作，将强化学习与参考轨迹监督结合，通过定义轨迹匹配奖励和任务目标奖励的加权和，使机器人能够学习高动态动作，同时保持动作的自然度。

然而，Mimic 类模仿学习方法也存在明显的局限性。首先，对参考数据依赖性强，难以适应未见过的动作或环境；其次，动作库扩展性差，大多数方法仅限于单一动作的迁移，如 ASAP[5] 和 KungfuBot[6] 等方法只能在特定动作上表现良好；最后，泛化能力有限，在复杂地形或外部干扰下，步态质量显著下降。

由伯克利团队提出的 BeyondMimic 框架[7] 在 Mimic 模仿学习方法的基础上引入扩散模型，实现了从人类动作中学习技能并将其组合为面向任务执行控制策略的突破。但是，其扩散阶段的多步迭代采样带来推理延迟和实时性差等问题，额外的扩散模型增加了训练、调试和维护成本。

### 1.2　AMP 类对抗模仿学习方法

AMP 类对抗模仿学习方法[8-11] 是人形机器人步态控制领域的重大突破，它通过引入对抗训练机制，解决了传统 Mimic 类模仿学习方法与参考动作序列强耦合，无法泛化至既定任务的问题。AMP 的核心思想借鉴了生成对抗网络（generative adversarial network, GAN）框架：策略通过与环境交互生成状态转移序列，而判别器则学习区分这些序列与专家示范中的状态转移片段。策略的目标是最大化由判别器提供的模仿奖励与任务奖励所构成的复合奖励，从而在完成指定任务的同时，使其生成的运动分布尽可能逼近专家数据的分布。这种方法一定程度上实现了机器人在兼顾参考动作风格的同时实现既定任务，提高了步态的自然度和协调性。但也引入了动作风格平均化、动作细节丢失、模式崩溃不稳定[8] 等问题。

## 2　方　法

### 2.1　框架概述

为使机器人控制策略在复现人类自然行走动作风格的同时，能够准确响应外部给定的速度指令。本文提出了一种基于 DRL 面向风格化速度追踪的运控策略训练框架。如图 1 所示，训练框架由 4 个核心模块构成：a. 策略网络，负责根据传感观测生成目标关节位置；b. 价值网络，用于评估当前状态的长期回报；c. 底层 PD 控制器，以 1 000 Hz 的高频率将目标位置转换为关节扭矩；d. 物理仿真环境，提供真实的动力学反馈与状态信息。策略网络与价值网络采用近端策略优化（proximal policy optimization, PPO）算法进行端到端训练。训练过程中，智能体接收的总奖励由 3 类组成：$r_t = r_t^m + r_t^v + r_t^n$。式中：$r_t^m$ 为运动跟踪奖励类，旨在驱动机器人复现人类动作捕捉数据重定向至目标机器人后的参考动作数据的风格；$r_t^v$ 为速度跟踪奖励类，用于奖励策略跟踪动作数据的速度信息，$r_t^n$ 为常规奖励类覆盖了自碰撞、关节抖动惩罚及触地力平滑奖励等，引导策略学习更稳定、自然的运动步态，从而提升整体控制性能。该复合奖励信号驱动价值和策略网络协同优化：价值网络基于完整参考运动信息对状态−动作进行价值评价，联合判别运动合理性与指令一致性；策略网络则仅依赖本体感知与速度指令，通过价值网络提供的优势函数进行策略梯度更新，进而在无参考动作输入的情况下，自主生成符合高层语义的类人步态。

### 2.2　观测设计

为兼顾任务语义表达与物理动态建模，本文为策略网络与价值网络分别设计了差异化的观测输入，以实现部署可行性与学习效率的协同优化。

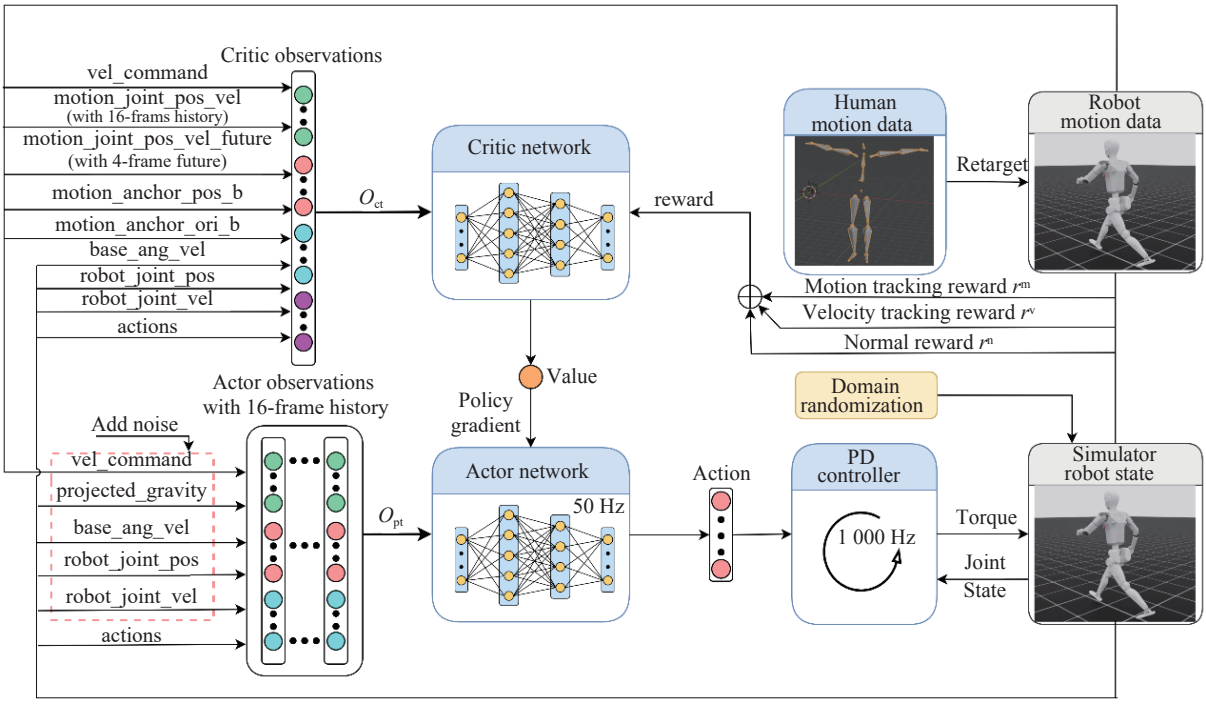


**图 1　训练框架示意图**

**Fig. 1　Schematic diagram of the training framework**

表 1 列出了价值网络的观测项，其聚焦于完整的状态监测。训练过程中仿真环境提供了完整的机器人状态监测，仿真环境中通过刚体动力学和运动学可得到现实环境中受限于传感器不足而无法获取的各肢体位置和姿态、速度、接触力等特权观测量。价值网络的观测项除了速度控制指令、基座角速度、机器人各关节位置和速度，还包含了参考运动的关节位置与关节速度(除当前和历史 16 帧外，还增加了未来 4 帧)、参考运动锚点位置和姿态等特权观测量。为便于对比参考运动和机器人的真实运动，选定机器人根部位为锚点部位，根部位于机器人的腰椎部位，以根部的位置、姿态和水平速度代表机器人的位置、姿态和速度。

策略网络的观测则严格限制为实际部署中可获取的传感器信号与速度指令，策略推理时其速度指令切换为手持遥控器发送，重力向量由机器人根部位置安装的惯性测量单元测量得到的欧拉角换算得来，根部位角速度通过直接读取惯性测量单元的测量值获取、机器人的关节角度和角速

**表 1　价值网络观测项**

**Tab. 1　Critic network observation items**

| 观测项 | 内容说明 | 时间范围 |
|---|---|---|
| vel_command | 参考运动在根部坐标系下的水平前进速度、横移速度、偏航角速度 | 当前时刻 |
| motion_joint_pos_vel | 参考运动历史关节角度与角速度 | 过去15帧+当前帧(共16帧) |
| motion_joint_pos_vel_future | 参考运动未来关节角度与角速度,提供短期运动意图上下文 | 未来4帧 |
| motion_anchor_pos_b | 参考运动锚点部位的位置 | 当前时刻 |
| motion_anchor_ori_b | 参考运动锚点部位的姿态 | 当前时刻 |
| base_lin_vel | 机器人根部位线速度 | 当前时刻 |
| base_ang_vel | 机器人根部位角速度 | 当前时刻 |
| robot_joint_pos | 机器人当前关节角度 | 当前时刻 |
| robot_joint_vel | 机器人当前关节角速度 | 当前时刻 |
| actions | 上一时间步策略网络输出的动作 | 当前时刻 |

度由驱动电机的编码器读值经速比换算获取。策略网络观测量均以 16 帧历史窗口（含当前帧）的形式输入，以支持时序推理与相位估计，见表 2。

值得注意的是，与 Mimic 及其变种不同，本文在策略网络观测中完全移除了参考动作的关节角度和角速度信息，仅保留从参考运动中提取的速度指令作为高层目标；而价值网络则仍保留完整的参考关节状态（含历史与未来状态）。该设计旨在引导价值网络学习高层速度指令语义与参考运动风格之间的隐式映射：通过 16 帧历史窗口推断当前步态相位，并结合未来 4 帧参考动作理解短期运动趋势，从而构建一个“速度→风格化步态”的内部模型。得益于此，训练完成的策略网络在仿真或实物环境中仅接收任意速度指令，即可自然涌现出与参考运动风格一致的人类步态，无需访问任何参考动作，显著提升了模仿学习方法的实用性与部署灵活性。此外，在训练阶段，对除 actions 外的所有观测注入均匀随机噪声，以模拟真实传感器读数中的不确定性，进一步增强策略网络对输入扰动鲁棒性及在现实场景中的泛化能力。

**表 2　策略网络观测项**

**Tab. 2　Actor network observation items**

| 观测项 | 内容说明 | 时间范围 |
|---|---|---|
| vel_command | 参考运动在根部坐标系下的水平前进速度、横移速度、偏航角速度 | 过去15帧+当前帧（共16帧） |
| projected_gravity | 重力向量在机器人根坐标系下的投影（反映躯干姿态倾角） | |
| base_ang_vel | 机器人根部位角速度 | |
| robot_joint_pos | 机器人当前关节角度 | |
| robot_joint_vel | 机器人当前关节角速度 | |
| actions | 上一时间步策略网络输出的动作 | |

## 2.3　奖励设计

奖励设计分为 3 大类：运动追踪类、速度跟踪类和常规奖励类。

奖励函数设计以 $L_2$ 范数量化跟踪误差经高斯核函数非线性变换得到未加权重的奖励值。$L_2$ 范数因其平滑性与可微性，能为策略优化器提供稳定且信息丰富的梯度信号。然而，直接使用 $L_2$ 范数作为奖励会导致梯度信号在误差较大时过于陡峭，在误差较小时又过于微弱。为了解决这个问题，引入了高斯核函数（Gaussian kernel）对 $L_2$ 误差进行非线性映射，以确保奖励函数的平滑性，避免因奖励突变导致策略震荡。最终形成的奖励函数的通用形式如下：

$$R = \exp\left(-\frac{\| \boldsymbol{e} \|^2}{\sigma^2}\right) \tag{1}$$

式中：$\boldsymbol{e}$为状态误差（如位置差、速度差等）；$\sigma$为控制奖励敏感度的超参数。

### 2.3.1　运动追踪奖励

运动追踪奖励用于引导策略控制机器人做出和参考运动尽可能一致的动作。首先，选定需要跟踪的身体部位集合，本文选定了机器人的根部以及各肢体部位作为跟踪目标集合 $\mathcal{B}_{\text{target}}$。通过 $R^t_{\text{anchor\_pos}}$、$R^t_{\text{anchor\_ori}}$ 2 个奖励项来驱动策略减小机器人锚点与参考运动锚点的位置和姿态误差。通过 $R^t_{\text{body\_pos}}$、$R^t_{\text{body\_ori}}$、$R^t_{\text{body\_vel}}$、$R^t_{\text{body\_w}}$ 4 个奖励项引导策略驱动机器人各部位减小与参考运动对应部位的位置、姿态、速度、角速度的误差，这些误差均在机器人的锚点部位坐标系下计算，通过对齐坐标系，消除机器人本体与参考运动间锚点位置和姿态差异的影响。坐标对齐涉及到平移对齐和偏航对齐。

平移对齐是将参考运动中的各肢体水平位置以锚点为基准，平移至仿真机器人的锚点水平位置，保留参考运动的原始高度，以消除参考运动和仿真机器人水平位置基准不一致的影响。平移对齐的平移量计算如下：

$$\Delta_{\boldsymbol{p}} = \begin{bmatrix} x_{\text{anchor}} - x_{\text{anchor,m}} \\ y_{\text{anchor}} - y_{\text{anchor,m}} \\ 0 \end{bmatrix} \tag{2}$$

式中：$x_{\text{anchor}}$、$y_{\text{anchor}}$为仿真机器人锚点的水平位置；$x_{\text{anchor,m}}$、$y_{\text{anchor,m}}$为参考运动中锚点的水平位置。

平移对齐后的参考动作肢体位置和姿态，再经过偏航对齐式 (4) 和式 (5)，将其位置和姿态绕 $z$ 轴（指向重力反方向）旋转 $\Delta\psi$，使其与仿真机器人偏航方向一致，以消除参考运动与仿真机器人

朝向不一致的影响。

$$\Delta\psi = \psi_{\text{anchor}} - \psi_{\text{anchor,m}} \tag{3}$$

式中：$\psi_{\text{anchor}}$为机器人当前锚点的偏航角；$\psi_{\text{anchor,m}}$为参考运动中锚点在对齐帧的偏航角。

$$\boldsymbol{p}_b^{\text{des}} = \boldsymbol{p}_{\text{anchor},m} + \Delta\boldsymbol{p} + \boldsymbol{R}_z(\Delta\psi)\left(\boldsymbol{p}_{b,m} + \boldsymbol{p}_{\text{anchor},m}\right) \tag{4}$$

式中：$\boldsymbol{p}_b^{\text{des}}$为对齐后身体部位 $b$ 的期望位置；$\boldsymbol{p}_{b,m}$为参考运动中身体部位 $b$ 的位置；$\boldsymbol{P}_{\text{anchor},m}$ 为参考运动中锚点部位的位置；$\Delta\boldsymbol{p}$ 为当前机器人锚点与参考运动锚点之间的水平位置偏移量；$\Delta\psi$为当前机器人锚点与参考运动锚点之间的偏航角差；$\boldsymbol{R}_z(\Delta\psi)$ 为绕世界坐标系 $z$ 轴旋转$\Delta\psi$的旋转矩阵；下标 $m$ 为参考运动（motion）中的对应状态；anchor 为目标身体部位，表示选取的锚点部位；$\boldsymbol{R}_z(\Delta\psi) = \begin{bmatrix} \cos\Delta\psi & -\sin\Delta\psi & 0 \\ \sin\Delta\psi & \cos\Delta\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}$为偏航对齐的旋转矩阵。

$$\boldsymbol{R}_b^{\text{des}} = \boldsymbol{R}_z(\Delta\psi)\boldsymbol{R}_{b,\text{m}} \tag{5}$$

式中，$\boldsymbol{R}_{b,\text{m}}$为参考运动中肢体 $b$ 的姿态。

表 3 中：$\mathcal{B}_{\text{target}}$为目标跟踪部位的集合；$\boldsymbol{p}_{\text{anchor}}^{\text{des}}$、$\boldsymbol{p}_{\text{anchor}}$分别为锚点部位的期望位置和实际位置；$\boldsymbol{R}_{\text{anchor}}^{\text{des}}$、$\boldsymbol{R}_{\text{anchor}}$分别为锚点部位期望姿态和实际姿态，通过$\boldsymbol{R}_{\text{anchor}}^{\text{des}}$、$\boldsymbol{R}_{\text{anchor}}^{\text{T}}$计算从实际锚点姿态到目标锚点姿态的相对旋转，经$\log(\boldsymbol{R}_{\text{anchor}}^{\text{des}}\boldsymbol{R}_{\text{anchor}}^{\text{T}})$将旋转矩阵转换为旋转向量，然后经 $L_2$ 范数度量实际姿态与目标姿态的误差；$\boldsymbol{p}_b^{\text{des}}$表示参考动作中身体部位 $b$ 的位置经锚点对齐（即平移至机器人当前锚点部位水平位置、保留参考高度，并绕 $z$ 轴旋转以匹配当前偏航角）后得到的目标位置表示；$\boldsymbol{p}_b$表示仿真环境中机器人部位 $b$ 的实际位置；$\boldsymbol{R}_b^{\text{des}}$表示参考动作中身体部位 $b$ 姿态经与机器人当前的锚点偏航对齐后得到的目标姿态；$\boldsymbol{R}_b$为机器人部位 $b$ 的实际姿态；$\boldsymbol{v}_b^{\text{des}}$为参考动作中身体部位 $b$ 的期望线速度；$\boldsymbol{v}_b$为机器人部位 $b$ 的实际线速度；$\omega_b^{\text{des}}$为参考动作中身体部位 $b$ 的期望角速度；$\omega_b$为机器人身体部位 $b$ 的实际角速度。

**表 3 运动追踪奖励项**

**Tab. 3 Motion tracking reward items**

| 奖励项 | 公式 | 权重 |
|---|---|---|
| $R_{\text{anchor\_pos}}^t$ | $\exp(-\|\boldsymbol{p}_{\text{anchor}}^{\text{des}} - \boldsymbol{p}_{\text{anchor}}\|^2/0.3^2)$ | 0.5 |
| $R_{\text{anchor\_ori}}^t$ | $\exp(-\|\log(\boldsymbol{R}_{\text{anchor}}^{\text{des}}\boldsymbol{R}_{\text{anchor}}^{\text{T}})\|^2/0.5^2)$ | 0.4 |
| $R_{\text{body\_pos}}^t$ | $\exp(-(\frac{1}{\lvert\mathcal{B}_{\text{target}}\rvert}\sum_{b\in\mathcal{B}_{\text{target}}}\|\boldsymbol{p}_b^{\text{des}} - \boldsymbol{p}_b\|^2)/0.3^2)$ | 0.5 |
| $R_{\text{body\_ori}}^t$ | $\exp(-(\frac{1}{\lvert\mathcal{B}_{\text{target}}\rvert}\sum_{b\in\mathcal{B}_{\text{target}}}\|\log(\boldsymbol{R}_b^{\text{des}}\boldsymbol{R}_b^{\text{T}})\|^2)/0.4^2)$ | 0.5 |
| $R_{\text{body\_vel}}^t$ | $\exp(-(\frac{1}{\lvert\mathcal{B}_{\text{target}}\rvert}\sum_{b\in\mathcal{B}_{\text{target}}}\|\boldsymbol{v}_b^{\text{des}} - \boldsymbol{v}_b\|^2)/1.0^2)$ | 1.0 |
| $R_{\text{body\_w}}^t$ | $\exp(-(\frac{1}{\lvert\mathcal{B}_{\text{target}}\rvert}\sum_{b\in\mathcal{B}_{\text{target}}}\|\boldsymbol{\omega}_b^{\text{des}} - \boldsymbol{\omega}_b\|^2)/3.14^2)$ | 1.0 |

### 2.3.2 速度跟踪奖励

速度跟踪奖励分为线速度跟踪误差奖励$R_{\text{lin_vel_}xy}^t$和 $z$ 轴偏航角速度误差奖励$R_{\text{ang_vel_}z}^t$，分别用来引导策略驱动机器人追踪水平面方向的线速度指令和 $z$ 轴方向偏航角速度指令，见表 4。

表 4 中：$\boldsymbol{v}_{\text{cmd},xy}$为机器人根节点在偏航对齐下的水平线速度指令；$\boldsymbol{v}_{\text{robot},xy}^{\text{yaw}}$为机器人根节点在偏航对齐坐标系下的实际水平线速度；$\omega_{\text{cmd},z}$为机器人根节点的 $z$ 轴角速度指令；$\omega_{\text{robot},z}$为机器人根节点的实际 $z$ 轴角速度。

### 2.3.3 常规奖励

常规奖励项涵盖了机器人动作变化率惩罚项、关节角度超限位惩罚项、自身碰撞惩罚项以及脚与地面接触平滑奖励项组成，它们分别用于限制策略输出的关节角度变化率过大与超出关节限位、减少身体的自碰撞、鼓励脚和地面平滑地接触。权重大小设置借鉴 BeyondMimic 项目的相关奖励项权重为基础值，训练过程中根据效果做针对性调整，最终在多次训练中迭代得到表现最稳定的权重配置，见表 5。

**表 4　速度跟踪奖励项**

**Tab. 4　Velocity tracking reward items**

| 奖励项 | 公式 | 权重 |
|---|---|---|
| $R^t_{\text{lin\_vel\_}xy}$ | $\exp\left(-\dfrac{\| \boldsymbol{v}_{\text{cmd},xy}-\boldsymbol{v}^{\text{yaw}}_{\text{robot},xy} \|^2}{1.0^2}\right)$ | 1.5 |
| $R^t_{\text{ang\_vel\_}z}$ | $\exp\left(-\dfrac{\| \omega_{\text{cmd},z}-\omega_{\text{robot},z} \|^2}{3.14^2}\right)$ | 1.5 |

**表 5　常规奖励项**

**Tab. 5　Regular reward items**

| 奖励项 | 公式 | 权重 |
|---|---|---|
| $R^t_{\text{a}_{\text{smooth}}}$ | $\| \boldsymbol{a}_t-\boldsymbol{a}_{t-1} \|^2$ | −0.15 |
| $R^t_{\text{joint\_limit}}$ | $\sum_{j=1}^{N}[\max(l_j-\theta_j,0)+\max(\theta_j-u_j,0)]$ | −10 |
| $R^t_{\text{undesired\_contacts}}$ | $\sum_{b\notin\mathcal{B}_{\text{Bee}}} 1[\| \boldsymbol{f}^{\text{self}}_b \|>1\text{N}]$ | −0.1 |
| $R_{\text{feet\_force\_smooth}}$ | $\exp(-\alpha\dfrac{1}{(H-1)N_{\text{f}}}\sum_{t=1}^{H-1}\sum_{k=1}^{N_{\text{f}}}\lvert F_{t+1,k,z}-F_{t,k,z}\rvert)$ | 0.001 |

表 5 中：$\boldsymbol{a}_t$、$\boldsymbol{a}_{t-1}$分别为当前时间步和上一时间步策略网络输出的动作向量；$\theta_j$、$l_j$、$u_j$分别为第 $j$ 个关节的角度、下软限位角度、上软限位角度；$\mathcal{B}_{\text{Bee}}$为机器人末端部位的集合；$\boldsymbol{f}^{\text{self}}_b$表示机器人自身碰撞的接触力；$F_{t,k,z}$为第 $t$ 个时间步第 $k$ 个脚的 $z$ 向接触力；$N_{\text{f}}=2$为脚的个数；$H=3$表示接触传感器历史数据长度；$\alpha$=0.8 为平滑权重参数。

## 2.4　域随机化

在训练过程中对环境参数(如物理属性、初始状态、观测噪声等)进行随机采样。从而扩大策略的训练分布，提升策略的鲁棒性，并减小仿真和实物环境之间的差异。

对策略网络的除 actions 外的所有观测项均注入均匀噪声来模拟实际传感器中的测量不确定性，防止策略过拟合于“理想干净”的仿真观测，增强真实世界中传感器噪声的鲁棒性。观测噪声范围，以实际传感器的噪声范围(参考传感器技术手册确定)为基础值取值，然后根据训练效果进行迭代调优，最终取值结果见表 6。

**表 6　策略网络观测噪声注入**

**Tab. 6　Actor network observation noise injection**

| 噪声加载项 | 噪声范围 | 单位 |
|---|---|---|
| vel_ command | (−0.1,0.1) | m/s、rad/s |
| projected_gravity | (−0.05,0.05) | — |
| base_ang_vel | (−0.2,0.2) | rad/s |
| joint_pos | (−0.01,0.01) | rad |
| joint_vel | (−0.5,0.5) | rad/s |

另外，训练过程中加入了真实世界可能出现的接触材质、初始状态、执行器参数、根部质量、关节零位、质心位置、随机速度冲击随机化事件项，并依据真实世界可能出现的取值范围和工程经验设定随机化参数，见表 7。旨在模拟真实世界中各种不确定性，提升策略对真实环境的适应能力。

**表 7　随机化事件项**

**Tab. 7　Randomize event items**

| 随机化事件 | 随机化参数 | 作用 |
|---|---|---|
| physics_material | 静摩擦因数:(0.3, 1.6)<br>动摩擦因数:(0.3, 1.2)<br>恢复系数:(0, 0.5) | 随机化机器人各连杆及地面的摩擦因数与恢复系数，模拟不同材质表面间的接触、摩擦与碰撞 |
| reset_base | 位置范围:(−0.5～0.5 m)<br>速度范围:(−0.5～0.5 m/s) | 每次重置时随机设置根部位置与速度，避免策略过拟合特定起始条件 |
| randomize_actuator_gains | 缩放系数:(0.8,1.2) | 随机缩放关节PD增益(刚度/阻尼)，模拟电机性能偏差、老化、性能变化等不一致情况 |
| add_base_mass | 质量增量:(−2.0 kg+2.0 kg) | 为指定机器人根部位加减随机质量，模拟负载变化 |
| add_joint_default_pos | 关节零位偏移:(−0.01 rad+0.01 rad) | 随机偏移关节默认位置，模拟编码器安装误差或标定偏差 |
| base_com | 质心偏移:(−0.025 m+0.025 m) | 随机躯干质心位置，增强对重心偏移(如携带物品)的鲁棒性 |
| push_robot | 时间间隔:(1.0 s, 3.0 s)<br>$x$/y向速度范围:(−0.5～0.5 m/s)<br>$z$向速度范围:(−0.2～0.2 m/s) | 以一定范围内的时间间隔施加随机速度冲击，模拟碰撞或人为推搡，提升抗干扰能力 |

## 2.5　终止判定条件

为提升训练效率，设定提前终止条件，判定当状态超出设定条件阈值时提前终止探索。本框架设定 3 个终止判定条件，见表 8。当机器人锚点部位高度或姿态偏离参考运动期望值超出阈值、机器人末端执行器(双手、双脚)高度偏离参考运动超出阈值，则提前重置该探索回合。因目标机器人身高 1.7 m，设定当机器人锚点部位、手、脚

表 8 终止判定条件

Tab. 8 Termination judgment conditions

| 终止判定项 | 设定阈值 |
|---|---|
| bad_anchor_pos_z | 0.25 m |
| bad_anchor_ori | 1.0 rad |
| bad_body_pos_z | 0.25 m |

的高度跟踪误差超出 0.25 m 或者机器人锚点姿态跟踪误差超出 1 rad 即判定跟踪失败，并提前终止探索以提升训练效率。

### 2.6 自适应采样

在策略探索过程中如超出终止判定条件会被判定为探索失败。参考动作序列中各动作切片的难易程度不同。传统模仿学习方法（如 Mimic）通常对参考动作序列进行均匀采样，这种策略容易导致容易的动作片段被过度采样，而困难动作区域则相对采样不足，导致训练效率低下。为优化该问题，本文借鉴 Beyond Mimic 的自适应采样机制，通过统计历史失败率对参考动作动态采样以提升训练效率。

将动作序列按顺序划分成 $S$ 条动作切片，每个切片包含涵盖 1 s 时长的动作序列，在训练过程中，统计各运动切片对应的失败情况，并采用指数滑动平均更新每个切片的历史失败统计量，以降低单个训练批次中随机失败对采样分布的影响。设第 $s$ 个切片在当前更新中的失败统计量为 $f_s$，则有历史失败统计量 $\bar{f}_s \leftarrow 0.999 f_s + 0.001 \bar{f}_s$ 通过指数滑动平均更新。在此基础上，为使发生失败的运动片段及其邻近片段获得更高的采样概率，同时保留一定的均匀采样概率，采用长度为 $K$ 的指数衰减核对各切片的历史失败统计量进行平滑。定义第 $s$ 个切片的平滑采样权重 $r_s$ 为

$$r_s = \sum_{u=0}^{K-1} \frac{\rho^u}{\sum_{v=0}^{K-1} \rho^v} \left( \bar{f}_{s+u} + \frac{\lambda}{S} \right) \tag{6}$$

式中：$S$ 为参考运动划分的切片总数；$\bar{f}_s$ 为第 $s$ 个切片经指数滑动平均更新后的历史失败统计量；$K$ 为平滑核长度；$\rho$ 为核权重衰减系数；$\lambda$ 为均匀采样比例；$r_s$ 为第 $s$ 个切片经邻域平滑后得到的采样权重。本文取 $K$=3、$\rho$=0.8、$\lambda$=0.1。对于序列末端不足 $K$ 个切片的情况，采用末端值复制方式进行边界填充。

根据式 (6) 得到各运动切片的平滑采样权重后，对所有切片的采样权重进行归一化，得到各切片的最终采样概率。第 $s$ 个切片采样概率为

$$p_s = \frac{r_s}{\sum_{j=1}^{S} r_j} \tag{7}$$

式中：$p_s$ 为第 $s$ 个运动片段被采样的最终概率；$r_s$ 为式 (6) 计算得到的平滑采样权重。

## 3 实验与结果分析

### 3.1 实验设置

本框架在 Droid X2C 型号机器人进行验证，其全身有 22 个自由度，腿部的 12 个自由度足以支持复杂且类人的步态运动。参考动作选择 lafan1 的 walk1subject5，训练和实验仿真在 IsaacLab/IsaacSim 平台完成。为了验证本框架训练的风格化速度追踪器策略（下称本策略）能够在外部速度指令下有效地复现原始动作风格：首先，采样 200 s 时长任意速度指令下本策略的速度数据评估速度追踪效果；然后，在仿真平台分别提取风格化速度追踪器策略、AMP 速度追踪器策略、Mimic 模仿策略、原始参考动作在相同的速度指令序列下的关键身体部位的相对位置和姿态序列，通过 DTW 算法分别计算 3 种策略与原始参考数据的相似度；另外，为了验证速度指令对动作风格具有引导作用，将策略网络中的速度指令全部置为 0 进行训练以对比消融；最后，将策略部署在实物机器人上，分别测试本策略在实机上跟踪参考动作速度序列和任意速度序列的效果。

### 3.2 主要结果

#### 3.2.1 速度追踪性能评估

在仿真平台上，采集了本策略驱动机器人在 200 s 内响应指令端任意速度指令的运动数据，并将其根部速度与对应指令进行对比，使用平均绝对误差（mean absolute error，MAE）计算速度跟踪误差：

$$\mathrm{MAE}_d = \frac{1}{T} \sum_{t=1}^{T} \| \boldsymbol{v}_{c,t}^{d} - \boldsymbol{v}_{r,t}^{d} \|, \ d \in \{x, y, \omega_z\} \tag{8}$$

式中：$T$ 为测试序列总时刻数；$\boldsymbol{v}_{c,t}^{d}$ 为 $t$ 时刻第 $d$ 个方向的速度指令；$\boldsymbol{v}_{r,t}^{d}$ 为机器人对应速度响应，分别为 $x$ 方向线速度、$y$ 方向线速度和绕 $z$ 轴的偏航角速度。

计算速度跟踪误差结果见图 2。图中，RESE 为均方根误差(root mean square error， RMSE)。结果表明，3 个方向的跟踪误差均维持在较低水平：$x$ 方向线速度 MAE 为 0.123 m/s，$y$ 方向 MAE 为 0.112 m/s，角速度 MAE 为 0.108 rad/s，整体误差均小于 0.13 m/s 或 rad/s。误差分布以 0 为中心对称波动，无系统性偏差，体现出良好的稳态精度与抗扰能力。尽管存在由仿真噪声或执行器动态引起的高频抖动，平滑后的误差趋势仍保持平稳，未出现累积漂移或发散现象，充分验证了所提策略在长时间运行中兼具高精度与强鲁棒性的速度跟踪性能。

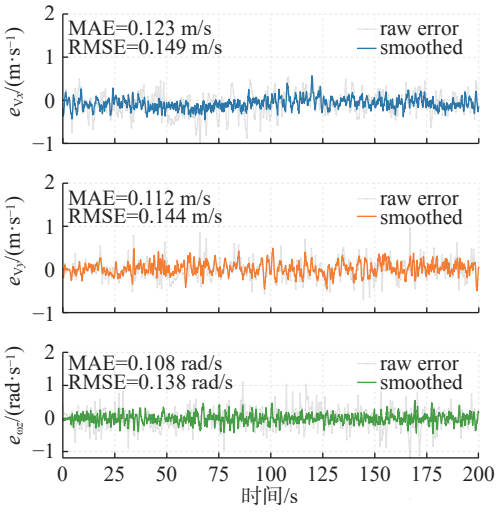


**图 2　速度追踪误差**

**Fig. 2　Velocity tracking error**

### 3.2.2　风格复现效果对比

为系统性地评估本策略的风格复现能力，基于 DTW 距离对 3 类策略(AMP、Mimic 与本策略)在参考动作上的风格复现性能进行量化分析。

$$\mathrm{DTW}\left(\boldsymbol{s}_{\mathrm{ref}}^{b},\boldsymbol{s}_{\mathrm{gen}}^{b}\right)=\min_{\pi\in\Pi}\sum_{(i,j)\in\pi}\left\|\boldsymbol{s}_{\mathrm{ref},i}^{b}-\boldsymbol{s}_{\mathrm{gen},j}^{b}\right\|_{2} \tag{9}$$

式中：$\boldsymbol{s}_{\mathrm{ref}}^{b}$、$\boldsymbol{s}_{\mathrm{gen}}^{b}$分别为肢体部位 $b$ 的参考运动特征序列和策略生成运动特征序列，运动特征分别取相对位置和相对姿态进行 DTW 距离计算；$\Pi$为所有满足单调性、连续性和边界条件的有效对齐路径集合；$\pi\in\Pi$为一条具体路径，$(i,j)\in\pi$为参考序列第$i$帧与生成序列第$j$帧在该对齐下相匹配。

分析涵盖两类运动特征：关键身体部位在锚点部位坐标系下的相对位置与相对旋转(转化为旋转矩阵计算)，覆盖上肢(大臂、手)、下肢(小腿、脚)共 8 个风格敏感部位。

表 9 为 3 种策略各身体部位相对参考动作的 DTW 值。Mimic 在位置与旋转两个维度上均取得最低的平均 DTW 距离，表明其生成动作在时空轨迹与姿态上最接近参考数据。其大臂与手部的 DTW 值显著低于其他方法的(如左大臂位置为 0.803 而 AMP 为 1.341)，反映出对上半身精细协调运动的高保真复现能力。同时，其下肢关节(小腿、脚)亦表现出最优姿态一致性(平均姿态 DTW 为 0.75)，说明其对参考运动的复现能力优异。

相比之下，本策略在下肢位置控制上表现突出(小腿和脚掌均为三者最低)，这与其显式速度跟踪目标一致，表明其在满足运动指令约束方面具有优势。然而，其上半身(尤其是大臂姿态)与参考动作的偏差较大(左大臂姿态 DTW=1.529)，显示在任务奖励约束后，牺牲了部分上肢自然性以换取任务可行性。

**表 9　3 种策略各身体部位相对参考动作的 DTW 值**

**Tab. 9　DTW values of relative reference movements for body parts under three strategies**

| 数据源 | 左大臂 | 右大臂 | 左手 | 右手 | 左小腿 | 右小腿 | 左脚掌 | 右脚掌 | 平均值 |
|---|---|---|---|---|---|---|---|---|---|
| AMP策略位置 | 1.341 | 1.411 | 0.876 | 1.396 | 0.904 | 0.866 | 0.881 | 0.872 | 1.068 |
| Mimic策略位置 | 0.803 | 1.099 | 0.62 | 1.036 | 0.758 | 0.731 | 0.767 | 0.758 | 0.821 |
| 本策略位置 | 1.001 | 1.077 | 0.7 | 1.135 | 0.68 | 0.69 | 0.725 | 0.695 | 0.837 |
| AMP策略姿态 | 1.763 | 1.997 | 1.38 | 1.642 | 1.059 | 1.12 | 1.029 | 1.023 | 1.376 |
| Mimic策略姿态 | 1.159 | 1.575 | 1.083 | 1.296 | 0.801 | 0.781 | 0.738 | 0.666 | 1.012 |
| 本策略姿态 | 1.529 | 1.677 | 1.257 | 1.428 | 0.808 | 0.849 | 0.781 | 0.758 | 1.135 |

AMP 方法在所有指标上均表现最弱，各身体部位轨迹的 DTW 值普遍最高(如右大臂姿态达 1.997)，反映出标准对抗模仿框架难以精确捕捉参考动作的细节结构，可能受限于判别器设计或奖励稀疏性。

综上，Mimic 展现出最优的整体运动保真度，而本策略则在任务导向与风格可控性之间实现了有效的平衡。

### 3.2.3 消融实验

为验证速度命令在风格化运动生成中的引导作用。在对照条件下训练了两组策略，基准组使用完整观测输入，消融组将策略网络的速度指令观测设为 0 向量，即移除速度指令的引导信息。

结果如图 3 所示，所有奖励项均表现出显著

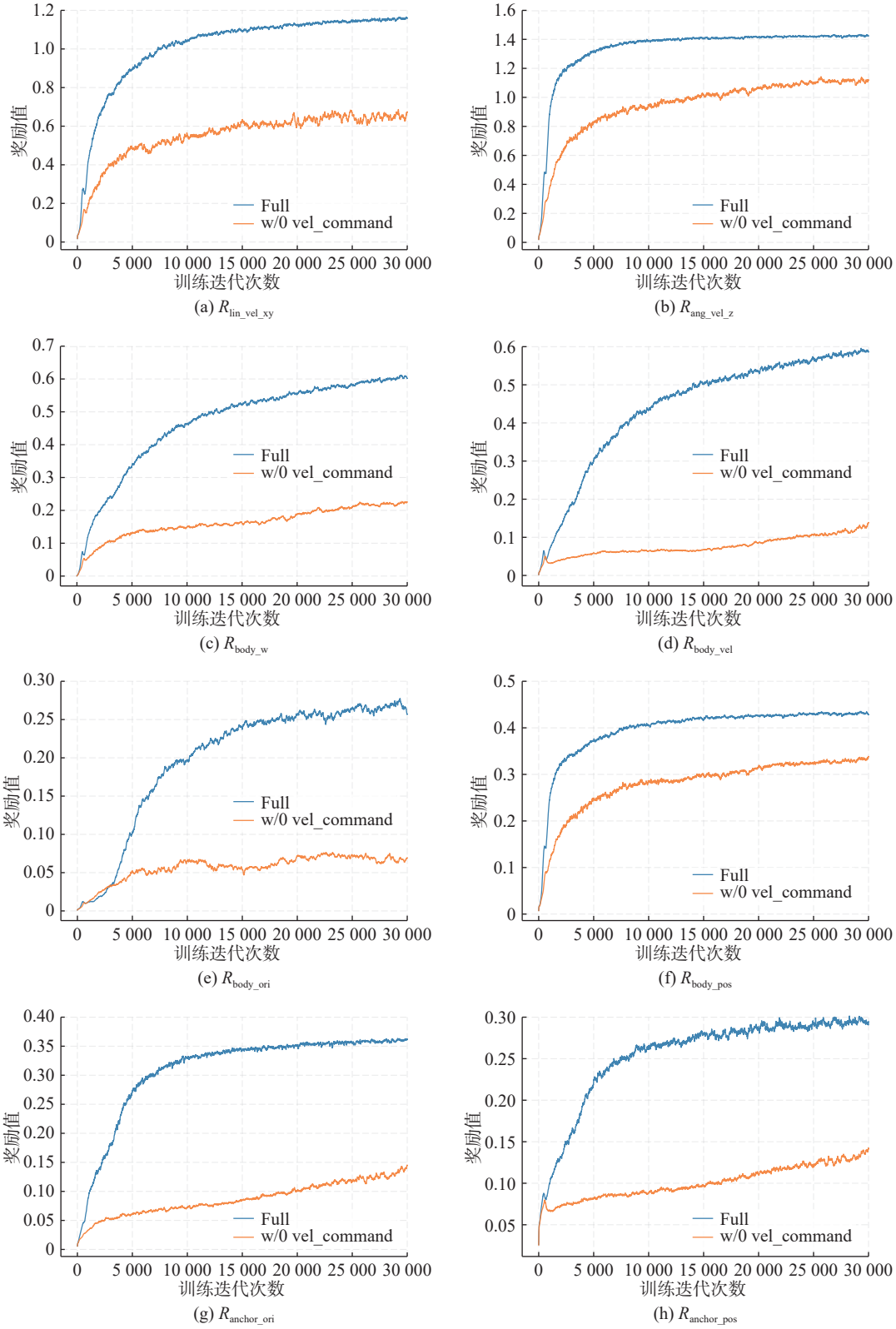


**图 3 消融前后奖励曲线对比**

**Fig. 3 Reward curves before and after ablation**

差异。在基准组 Full 中，机器人能够有效追踪目标速度指令，其在水平速度和偏航速度的奖励值迅速收敛至 1.2 及 1.4 附近，表明运动趋势与期望高度一致。相比之下，消融组 w/0 vel_command 的线速度奖励值始终低于 0.7，说明缺乏速度引导导致运动轨迹偏离目标。更重要的是，在身体姿态与位置相关奖励上，基准组在身体部位位置、姿态和锚点部位位置、姿态等维度均显著优于消融组。这表明速度指令不仅影响整体运动速度，还通过隐式反馈机制引导身体各部位的协调运动与空间定位，从而实现更自然、更具风格的运动表现。综上所述，本框架重构的速度指令观测在风格化运动控制中起到了关键的引导作用，它不仅决定了运动的速度，还通过奖励约束间接塑造了全身的运动风格。移除该信号后，策略无法有效学习到复杂的肢体协同动作，导致整体运动质量急剧下降。

### 3.2.4　实物复现参考动作风格实验

为验证本策略在实物平台上复现参考动作风格的能力，将参考动作的速度序列直接作为实物机器人的速度指令，在真实环境中驱动机器人执行行走任务。实验过程中，实时采集各关节角度，并通过正向运动学计算关键身体部位位置和姿态。然后，采用 DTW 距离对复现效果进行量化评估(结果见表 10)。

实验表明，实物机器人在参考速度序列引导下能够生成与参考动作风格相近的行走动作。然而，受仿真−现实差距(sim-to-real gap)影响，实物的风格复现性能表现弱于仿真环境，尤其体现在小腿和脚掌部位的 DTW 距离显著增大。这主要是因为：在速度跟踪行走任务中，腿部需频繁调整以支撑躯干并维持动态平衡，其工作于更高的扭矩区间，导致建模误差、摩擦、延迟等现实因素被进一步放大。相比之下，手臂等非承重部件受此类影响较小。

尽管如此，实物机器人关键部位的平均位置和姿态 DTW 值仍低于 AMP 策略在仿真中的对应值，表明本方法在风格复现能力上具有更优的迁移性能。

### 3.2.5　实物随机速度指令测试

为验证本策略在真实硬件系统中的泛化能力与控制鲁棒性，在人形机器人实物平台上部署策略测试。图 4、图 5 分别展示了实物测试和相似速度区间参考动作的半个步态周期内 5 个关键帧的动作序列，对应于右脚着地至左脚着地的过程。实验结果显示，机器人在无外部参考动作轨迹引导的情况下，能够稳定响应高层速度指令，并生成自然流畅的步态轨迹。实物机器人的步态自然

**表 10　实物机器人各身体部位相对参考动作的 DTW 值**

**Tab. 10　DTW values of each body part of physical robot relative to reference actions**

| 数据源 | 左大臂 | 右大臂 | 左手 | 右手 | 左小腿 | 右小腿 | 左脚掌 | 右脚掌 | 平均值 |
|---|---|---|---|---|---|---|---|---|---|
| 本策略位置 | 1.172 | 1.132 | 0.862 | 1.163 | 0.763 | 0.868 | 0.919 | 0.871 | 0.968 |
| 本策略姿态 | 1.296 | 1.433 | 1.137 | 1.415 | 1.028 | 1.166 | 0.956 | 0.985 | 1.177 |

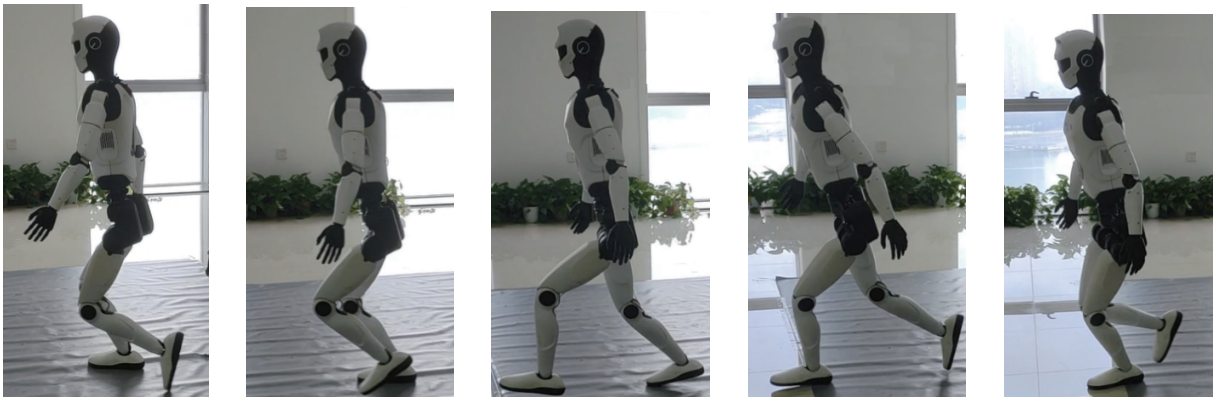

**图 4　实物机器人半步态周期关键帧序列**

**Fig. 4　Key frames of a half gait cycle in real-world locomotion**

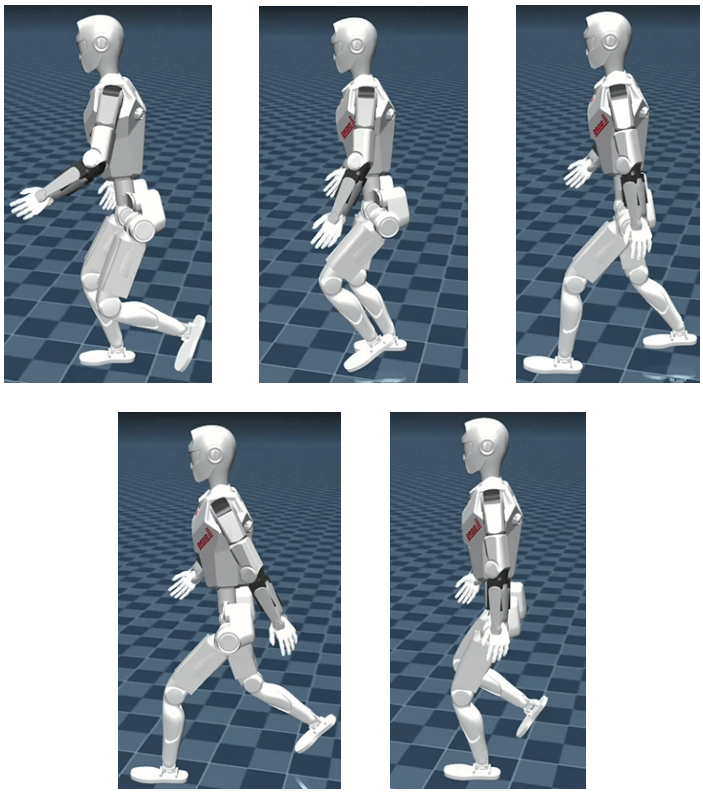

**图 5　参考动作半步态周期关键帧序列**

**Fig. 5　Key frames of the reference motion over a half gait cycle**

且协调：双足交替支撑与摆动过程平滑，膝关节屈伸角度合理，髋部摆动幅度适中，符合人类行走的动力学规律；身体姿态随着速度大小及转向幅度实时调整以维持身体平衡；双臂随着步态自然摆动且完美复现了参考动作中大幅度摆动左臂的风格特点。对比可见，该步态风格与参考数据高度一致，体现了策略在任务执行（速度跟踪）与风格还原之间的有效权衡。尽管存在轻微的响应延迟，但整体运动节奏、重心转移趋势及肢体协调性均达到预期水平，验证了所提框架在真实物理环境下的可行性与实用性。

## 4　结　论

本文针对现有模仿学习方法难以兼顾运动风格保真度与任务响应能力的问题，提出一种新颖的目标导向风格化运动策略训练框架。通过重构观测空间，该框架实现了运动风格与任务目标的显式解耦：策略网络仅依赖本体感知信息与高层速度指令，完全摆脱对参考动作序列的依赖；而价值网络则利用完整的参考运动上下文，隐式建模速度指令与风格化步态之间的映射关系。结合精心设计的复合奖励机制，所提方法成功将策略从“动作复现器”转变为“速度指令控制器”。

实验结果表明，在仅接收任意速度指令的条件下，本策略在 200 s 仿真轨迹上的整体速度跟踪误差均低于 0.13 m/s 或 rad/s。同时，本策略关键部位的位置、姿态相对于参考动作的平均 DTW 距离分别为 0.837 和 1.135，显著优于 AMP (1.068, 1.376)，并十分接近高保真但缺乏响应能力的 Mimic 方法的 (0.821, 1.012)。这充分证明，所提框架在实现精准速度跟踪的同时，有效保留了参考运动的自然风格特征。

值得注意的是，本框架在任务执行与风格保真之间取得了有效平衡，但这种平衡并非无代价。如表 9 所示，上半身（特别是大臂姿态）的 DTW 距离相对较高，表明策略为了优先保证速度跟踪这一核心任务的可行性，主动牺牲了部分上肢的精细自然性。研究认为，这是一种合理的权衡，因为对于行走任务，下肢的动力学可行性和稳定性是首要的。未来工作将探索更精细化的奖励机制，例如引入分层或自适应的奖励权重：在低速或静态场景下，可以动态提升姿态奖励的权重以保留更多风格细节；而在高速或高动态场景

下，则侧重于任务奖励以确保稳定性。

虽然本文以行走动作为例验证了所提框架的有效性，但其核心思想即通过将策略网络与价值网络观测解耦实现任务输入与动作风格的分离是具有普适性的。该框架并不依赖于行走特有的动力学模型或状态表示。理论上，只要能从参考动作中提取出高层任务指令（如舞蹈中的节奏、轨迹或音律等），并设计相应的任务奖励，本方法即可扩展至更复杂的全身动作。在更复杂的环境与任务场景中验证策略的泛化性与鲁棒性，以进一步拓展其在实际人形机器人系统中的应用潜力，这将是未来研究的重要方向。

**参考文献：**


[1] Tan J, Zhang T N, Coumans E, et al. Sim-to-real: learning agile locomotion for quadruped robots[C]//Robotics: Science and Systems XIV. Pittsburgh: Robotics: Science and Systems Foundation, 2018.

[2] Haarnoja T, Ha S, Zhou A, et al. Learning to walk via deep reinforcement learning[C]//Robotics: Science and Systems XV. Freiburg im Breisgau: Robotics: Science and Systems Foundation, 2019.

[3] Hwangbo J, Lee J, Dosovitskiy A, et al. Learning agile and dynamic motor skills for legged robots[J]. Science Robotics, 2019, 4(26): eaau5872.

[4] Peng X B, Abbeel P, Levine S, et al. DeepMimic: example-guided deep reinforcement learning of physics-based character skills[J]. ACM Transactions on Graphics (TOG), 2018, 37(4): 143.

[5] He T R, Gao J W, Xiao W L, et al. ASAP: aligning simulation and real-world physics for learning agile humanoid whole-body skills[C]//Robotics: Science and Systems XXI. Robotics: Science and Systems Foundation, 2025.

[6] Xie W J, Han J R, Zheng J K, et al. KungfuBot: physics-based humanoid whole-body control for learning highly-dynamic skills[EB/OL]. [2025-10-27]. https://arxiv.org/abs/2506.12851.

[7] Liao Q Y, Truong T E, Huang X Y, et al. BeyondMimic: from motion tracking to versatile humanoid control via guided diffusion[EB/OL]. [2025-11-13]. https://arxiv.org/abs/2508.08241.

[8] Peng X B, Ma Z, Abbeel P, et al. AMP: adversarial motion priors for stylized physics-based character control[J]. ACM Transactions on Graphics (TOG), 2021, 40(4): 144.

[9] Vollenweider E, Bjelonic M, Klemm V, et al. Advanced skills through multiple adversarial motion priors in reinforcement learning[C]//2023 IEEE International Conference on Robotics and Automation (ICRA). London: IEEE, 2023: 5120–5126.

[10] Tang A N, Hiraoka T, Hiraoka N, et al. HumanMimic: learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama: IEEE, 2024: 13107–13114.

[11] Peng X B, Guo Y R, Halper L, et al. ASE: large-scale reusable adversarial skill embeddings for physically simulated characters[J]. ACM Transactions on Graphics (TOG), 2022, 41(4): 94.


（编辑：毕莉明）