球面潜在 基于物理的模拟人形控制运动先验 l

Jing Tan *1 **Weisheng Xu** * 1 **Xiangrui Jiang** * 1 **Jiaxi Zhang** [1] **Kun Yang** [1] **Kai Wu** [1] **Jiaqi Xiong** [2]**Shiting Chen** [1] **Yangfan Li** [1] **Yixiao Feng** [1] **Yuetong Fang** [1] **Yujia Zou** [1] **Yiqun Song** [1] **Renjing Xu** †1

项目主页： https://colin-jing.github.io/SLMP/

arXiv:2603.01294v1 [cs.RO] 1 Mar 2026

摘要

基于物理的模拟人形控制中的运动先验学习是一个活跃的研究课题。现有方法主要包括变分自编码器（VAE）和对抗运动先验（AMP）。VAE 会引入信息损失，而随机潜变量采样有时会产生无效行为。AMP 则存在模式崩溃问题，难以捕捉多样化的运动技能。我们提出了球面潜在运动先验（SLMP），一种用于学习运动先验的两阶段方法。第一阶段，我们训练一个高质量的运动跟踪控制器。第二阶段，我们将跟踪控制器蒸馏到球面潜在空间中。通过结合蒸馏、判别器以及判别器引导的局部语义一致性约束，构建出结构化的潜在动作空间，从而在不造成信息损失的情况下实现稳定的随机采样。为评估 SLMP，我们收集了一个包含两小时人类格斗动作捕捉数据的数据集，并表明 SLMP 能够在不造成信息损失的情况下保留精细的运动细节，且随机采样可以产生语义有效且稳定的行为。当应用于双智能体基于物理的格斗任务时，SLMP 仅使用简单的基于规则的奖励，就能生成类人且符合物理规律的格斗行为。此外，SLMP 能够泛化到不同的人形机器人形态，展现出超越单一模拟化身的迁移能力。

1. 引言

近年来，基于物理的模拟人形机器人控制取得了显著进展，使模拟人形机器人能够模仿大规模动作捕捉数据集，执行敏捷动作（Peng 等人，2018年；Luo 等人，2021；2023；Tessler 等人，2024年；Zhang 等人，2025年；Jing 等人，2025年），并与环境互动（Gao 等人，2024年；Pan 等人，2025年；Wang 等人，2025年）。这些进展推动了计算机图形学、游戏、虚拟现实和机器人学领域的应用，而这些领域都离不开物理上合理的人类运动。尽管取得了这些成果，通过原始关节力矩控制高维人形机器人仍然面临挑战，因为其动力学不稳定，而且低层控制对奖励设计十分敏感（Luo 等人，2024a）。为缓解这些问题，近期许多方法引入运动先验，将动作限制在物理上合理且协调的运动流形上，从而简化下游任务的策略学习（Peng 等人，2022；Tessler 等人，2023；Luo 等人，2024a；b；Tessler 等人，2024年）。

学习运动先验的一项关键挑战，是同时实现多样性与有效性。基于变分自编码器（Kingma &Welling，2013）的方法（Won等人，2021；Zhu 等人，2023；Luo 等人，2024a）将全身运动轨迹压缩为潜在编码，但这一瓶颈往往会丢失重要的运动细节，而随机潜变量采样可能产生不合理的行为或导致跌倒。基于对抗运动先验（Peng 等人，2021）的方法（Peng 等人，2022；Tessler 等人，2023）引入对抗奖励塑造，利用判别器（Goodfellow等人，2020）区分真实动作与模拟动作，并引导策略趋向运动数据分布。然而，这些方法往往会遭遇模式崩溃和语义覆盖范围有限的问题，难以扩展到多样化的动作数据集。因此，现有的先验难以支持广泛的运动库、稳定的随机采样以及可靠的层次化复用。

这些局限源于底层方法的结构特性。基于变分自编码器的先验则通过高斯潜在瓶颈压缩高维、多模态的人体运动，这会促成单峰潜在分布并模糊细粒度变化，同时使潜在空间中的大片区域缺乏训练信号，导致随机采样可能生成不合乎情理的动作。基于对抗运动先验的先验依赖在对抗奖励下进行强化学习，将高方差的策略梯度与判别器反馈耦合，而判别器反馈会强调少数几种高

*共同贡献 [1]香港科技大学（广州），中国广州 [2]牛津大学，英国牛津。通讯作者：Renjing Xu < renjingxu@hkust-gz.edu.cn>。

预印本。*2026*年*3*月*3*日。

奖励行为。这往往会使策略转向少数几种模式，而不是覆盖充分的运动库。由于人形机器人运动库本质上具有多模态且依赖状态的特性，合适的潜在空间应体现有助于控制的有效局部结构，同时避免过度的信息压缩和无界的低密度区域，以免随机采样变得不稳定。

受这些需求启发，我们提出了球面潜在运动先验（SLMP），这是一种两阶段方法，可为基于物理的人形机器人构建结构化潜在动作空间。第一阶段，我们训练一个高质量的动作跟踪控制器。第二阶段，我们通过以潜在编码为条件来训练策略，将该控制器蒸馏到单位球潜在空间中。蒸馏目标由三个部分组成：从专家控制器进行模仿蒸馏、用于区分专家动作与非专家动作的判别器损失，以及由判别器引导的局部语义一致性损失，用以塑造球面上连贯的邻域。这种方法无需重建瓶颈即可保留动作细节，同时支持稳定且多样的随机采样。

为评估 SLMP，我们采集了一个为期两小时的格斗动作捕捉数据集，涵盖打击动作、闪避动作、状态转换和步法。实验表明，SLMP 能保留精细的动作细节；与基于 VAE 和 AMP 的基线方法相比，其随机采样性能更强，并能生成依赖状态的潜在邻域，以编码可行动作集。在双智能体基于物理的格斗任务中，将 SLMP 用作先验后，简单的高层策略便能利用轻量级奖励信号生成物理上合理且多样的行为。最后，我们证明，同一流程也能推广至真实的人形机器人平台。我们的贡献总结如下：

- 我们收集并将发布一个时长两小时的人类格斗动作捕捉数据集，其中包含多样的击打、闪避和步法行为，适用于基于物理的人形控制研究。
- 我们提出了 SLMP，这是一种两阶段蒸馏框架，可为人形构建结构化的球形潜在动作空间，从而实现稳定且多样的潜变量条件控制。
- 通过大量实验，我们证明 SLMP 仅使用简单的基于规则的奖励函数，便能实现符合物理规律的双智能体对抗，并能泛化至不同的人形机器人形态。

2. 相关工作

2.1. 基于物理的人形动作跟踪。

DeepMimic 等早期的基于物理的系统（Peng 等人，2018年）证明了强化学习能够在模拟环境中跟踪参考动作捕捉片段。UHC（Luo 等人，2021）实现了在 AMASS（Mahmood 等人，2019年）序列上进行通用全身跟踪，而 PHC（Luo 等人，2023）移除了外力模块，并在大规模数据集上实现了更稳定的跟踪。PHC+（Luo 等人，2024a）进一步改进了训练，使日常动作的跟踪近乎完美；MaskedMimic（Tessler 等人，2024年）则引入了掩码运动修复，使单个跟踪模型能够支持多种条件模式。最近，FARM（Jing 等人，2025年）专注于高动态动作，并结合帧加速增强与残差混合专家架构，以更好地处理基于物理的跟踪中的快速姿态转换。

2.2. 基于物理的人形运动潜在空间

基于物理的控制所用的潜在运动空间主要遵循两类设计思路。基于 AMP 的方法，例如 ASE（Peng 等人，2022）在球面流形上学习对抗性技能嵌入，以用于下游控制；而 CALM（Tessler 等人，2023）则学习条件对抗潜在模型，以实现可控的角色操控。这些方法支持技能组合，但在高度多样化的数据集上，往往存在潜在空间覆盖度有限、采样稳定性不足的问题。基于 VAE 的方法则通过重建目标来学习运动先验。PULSE（Luo 等人，2024a）和 MaskedMimic（Tessler 等人，2024年）采用 VAE 风格的潜在动作模型进行角色控制，而 Neural Categorical Priors（Zhu 等人，2023）则使用 VQ-VAE 风格的（Van Den Oord 等人，2017年）离散潜变量来编码行为模式。然而，基于重建的目标函数可能会造成信息瓶颈，并降低采样可靠性。相比之下，我们的方法通过控制器蒸馏学习连续球面潜在空间，着重实现稳定且语义明确的采样，以用于基于物理的人形控制。

3. 方法

我们的方法遵循图 1所示的流程。我们首先收集一个时长两小时的人类格斗动作捕捉数据集，并将其转换为 SMPL（Loper 等人，2015）动作片段。基于这些片段，我们通过动作跟踪（第 π_{track} 3.1节）训练一个目标条件专家控制器。接着，我们将该专家蒸馏到球面潜在空间中，并利用判别器和我们提出的判别器引导的局部语义一致性损失来塑造潜在流形（第3.2节）。最后，我们在双智能体格斗任务中使用 SLMP 作为结构化低层先验。高层策略输出驱动 SLMP 控制器的潜在编码，而自博弈（Silver 等人，2017年；Berner 等人，2019年）则通过简单的基于规则的奖励来优化这些潜在决策（第 3.3节）。

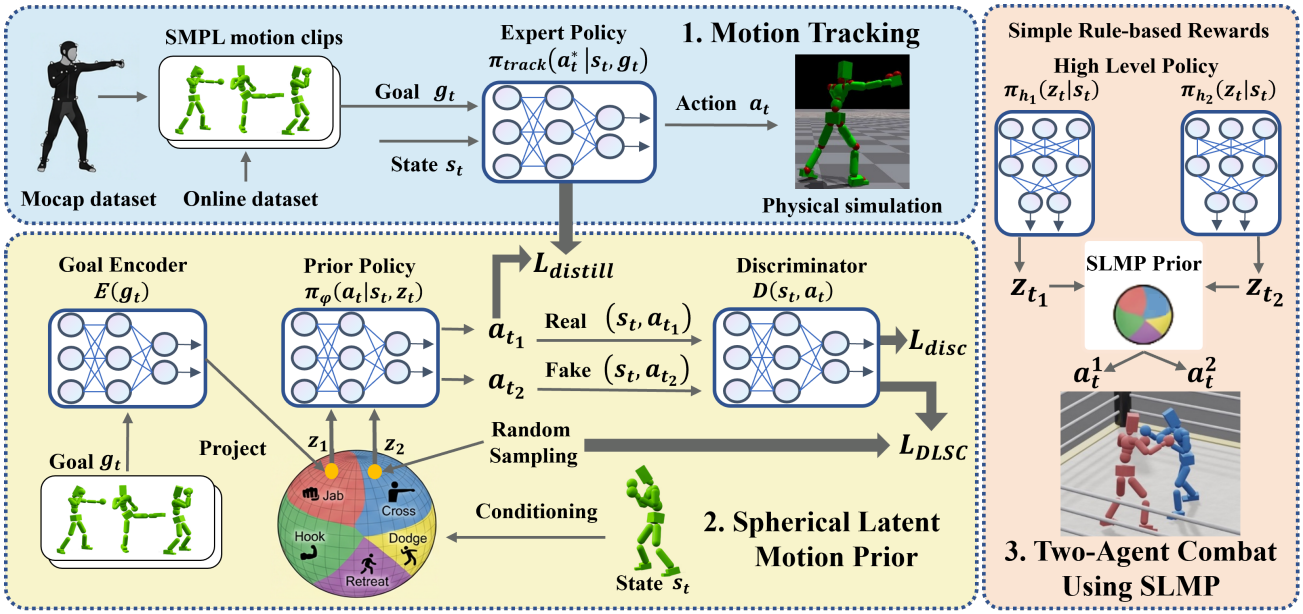


图 *1.* 球面潜在运动先验（SLMP）概述。我们采集了时长两小时的格斗动作捕捉数据集，并将其转换为 SMPL 动作片段用于训练。我们训练一个目标条件动作跟踪控制器，然后使用三种损失将其 Distill 到单位球潜在空间：模仿、判别器，以及我们的判别器引导的局部语义一致性损失（L_{DLSC}）。SLMP 支持有意义的随机采样，并通过简单奖励驱动双智能体对抗等下游任务。

3.1. 通过目标条件强化学习进行动作跟踪

我们训练一个目标条件动作跟踪控制器π_{track}，将人形机器人的本体感知状态和参考驱动目标映射为各关节的目标姿态。这种形式遵循目标条件强化学习（GCRL）框架，该框架广泛应用于大规模人形机器人控制器。在每个时间步 t，策略都会观测人形机器人的状态 s_t 以及一个 g_t 由 SMPL 参考动作导出的目标。策略为每个驱动关节输出 PD 目标角度 a_t，模拟器通过 PD 控制器施加关节力矩。我们使用近端策略优化（Schulman等人，2017年）优化 π_{track}，以最大化跟踪奖励，从而鼓励对参考动作进行物理模仿。奖励包括针对关节位置、关节方向、线速度和角速度的指数匹配项，以及用于抑制过度扭矩消耗的能量惩罚。训练完成后，π_{track}在我们的格斗动作数据集上实现了可靠的跟踪性能，并作为潜在蒸馏阶段的专家控制器。

3.2. 球面潜在运动先验（SLMP）

我们学习一个由潜变量条件化的运动先验，将 π_{track}蒸馏到球面潜在空间中。SLMP 的训练流程总结于算法1。先验策略$\pi_\varphi(s_t, z_t)$ 接收人形状态 s_t 和潜在编码$z_t \in \mathbb{R}^d$，其中 $\|z_t\|_2 = 1$，并输出一个 PD 动作。在每次训练迭代中，我们展开先验以获得 s_t，查询专家跟踪控制器以获得监督动作$a_t^* = \pi_{\text{track}}(s_t, g_t)$，使用从数据集中采样的参考目标 g_t，并将该目标编码为

$$z_1 = E(g_t), \quad \|z_1\|_2 = 1. \tag{1}$$

然后我们生成先验动作

$$a_{t1} = \pi_\varphi(s_t, z_1). \tag{2}$$

为了填充球面潜在空间，我们采样 $\epsilon \sim \mathcal{N}(0, I)$，将其归一化到单位球面上，并将其视为随机潜变量

$$z_2 = \frac{\epsilon}{\|\epsilon\|_2}, \quad a_{t2} = \pi_\varphi(s_t, z_2). \tag{3}$$

我们使用三种损失来优化 π_φ。模仿蒸馏损失使 a_{t1} 与专家动作相匹配：

$$L_{\text{distill}} = \|a_{t1} - a_t^*\|_2^2. \tag{4}$$

为了提供分布内信号，我们引入一个判别器 $D(s_t, a_t)$，用于区分专家动作与非专家动作。我们使用二元交叉熵目标训练 D，其中 a_{t1} 被视为正样本，a_{t2} 被视为负样本：

$$L_{\text{disc}} = -\mathbb{E}\big[\log D(s_t, a_{t1}) + \log\big(1 - D(s_t, a_{t2})\big)\big]. \tag{5}$$

判别器最小化 L_{disc}，而 π_φ 不会直接优化这一对抗损失。

为塑造潜在流形，我们引入一种由判别器引导的局部语义一致性损失。对于这对 $(z_1,\ z_2)$，我们计算球面距离和邻域权重

$$d_{12} = \|z_2 - z_1\|_2, \qquad w_d = e^{-\beta\, d_{12}}. \tag{6}$$

以及基于判别器的语义权重

$$w_c = 1 + \big|\min\big(0,\ D(s_t, a_{t2})\big)\big|. \tag{7}$$

随后，由判别器引导的局部语义一致性损失定义为

$$L_{\text{DLSC}} = w_d\, w_c\, \|a_{t2} - a_t^*\|_2^2. \tag{8}$$

最终的 SLMP 目标在最小化蒸馏与语义一致性的同时，判别器单独更新：

$$L_{\text{SLMP}} = \lambda_{\text{distill}} L_{\text{distill}} + \lambda_{\text{DLSC}} L_{\text{DLSC}}. \tag{9}$$

仅 L_{disc} 更新 D，而 L_{SLMP} 更新 π_φ。为保持训练稳定性，我们首先仅使用 w_d 目标训练直至收敛，然后引入 w_c 进行联合优化。需要注意的是，尽管 π_φ 并不直接优化 L_{disc}，语义权重 w_c 仍将 π_φ 与判别器耦合，从而在潜在空间中引发隐式对抗交互。

直观理解。纯蒸馏能够保留专家行为，但不支持随机采样，因为单位球面上的大多数潜在方向缺乏语义意义，因而无法诱导出有效动作。也可以改用标准的对抗生成器损失，以提高随机潜变量所诱导动作的判别器得分，但此类对抗反馈仅提供粗粒度的分布内梯度，无法提供复现精确关节空间行为所需的细粒度监督。因此，我们采用 L_{DLSC} 作为由判别器隐式引导的塑形目标。几何权重 w_d 鼓励局部平滑性，而基于判别器的权重 w_c 则提高分布外样本的权重，以施加更强的约束。这些项共同在球面上刻画出结构化潜在动作空间。该结构以当前状态 s_t 为条件而出现。具有多个合理未来的状态（例如站立）会产生多模态潜在区域，而高度受约束运动（例如腾空踢击）则在整个球面上表现出较低的语义变化性。

3.3. 高层任务：双智能体对抗

与 NCP（Zhu 等人，2023）和 Smplolympics（Luo 等人，2024b）类似，SLMP 被部署为结构化的低层控制先验，而高层策略在潜在空间中运行并输出潜在动作编码。对于每个智能体 $i \in \{1, 2\}$，高层策略会采样

$$z_{t_i} \sim \pi_{h_i}(z \mid s_{t_i}), \tag{10}$$

算法 1 球面潜在运动先验（SLMP）

1：冻结专家控制器 π_{track}；初始化先验策略 π_φ，编码器 E, 判别器 D
设置使用 $_w_c \leftarrow$ **假**
重复
  从数据集中采样参考目标 g_t；先验展开以获取状态 s_t
  计算专家动作 $a_t^* = \pi_{\text{track}}(s_t,\ g_t)$
  编码潜变量 $z_1 = E(g_t)$ 并归一化到单位球面
  采样随机潜变量 $z_2 \sim \mathcal{S}^{d-1}$
  计算先验动作 $a_{t1} = \pi_\varphi(s_t,\ z_1)$ 和 $a_{t2} = \pi_\varphi(s_t,\ z_2)$
  计算蒸馏损失 L_{distill} (公式4)
  如果使用$_w_c$ =**为假，则**
    计算 L_{DLSC}，不使用 w_c
    通过最小化 L_{SLMP} 来更新 $(\pi_\varphi,\ E)$
    如果已收敛**则**
      设置使用$_w_c \leftarrow$ **真**
    结束如果
  else
    计算语义一致性损失 L_{DLSC}（式 8）
    计算判别器损失 L_{disc}（式5）
    更新 (π_φ, E) 通过最小化 L_{SLMP}（式 9）
    更新 D 通过最小化 L_{disc}（式5）
  结束如果
直至收敛

而 SLMP 则通过以下方式生成最终的 PD动作

$$a_{t_i} = \pi_\varphi(s_{t_i}, z_{t_i}). \tag{11}$$

这样便能清晰地划分职责：高层策略在紧凑空间中选择潜变量行为，而 SLMP 则提供符合物理规律的全身执行。

双智能体对战场景通过自博弈进行训练，其中每个智能体都会观察自身的本体感知状态以及对手的相对姿态。奖励完全基于任务且十分稀疏，仅包含基于规则的命中和击倒事件，不包含任何额外的奖励塑形。当智能体之间距离过远或任一智能体倒地时，训练会提前终止，这促使智能体在稀疏奖励下积极交战。尽管奖励设计十分简约，高层策略仍学会了协调的击打和闪避行为，而 SLMP 则能保持符合物理规律的运动执行和状态转换。

4. 实验

我们的所有实验均基于第4.1节介绍的两小时格斗动作捕捉数据集，该数据集为评估提供了参考动作和跟踪目标。随后，我们在第 4.2节评估学得的潜在空间，

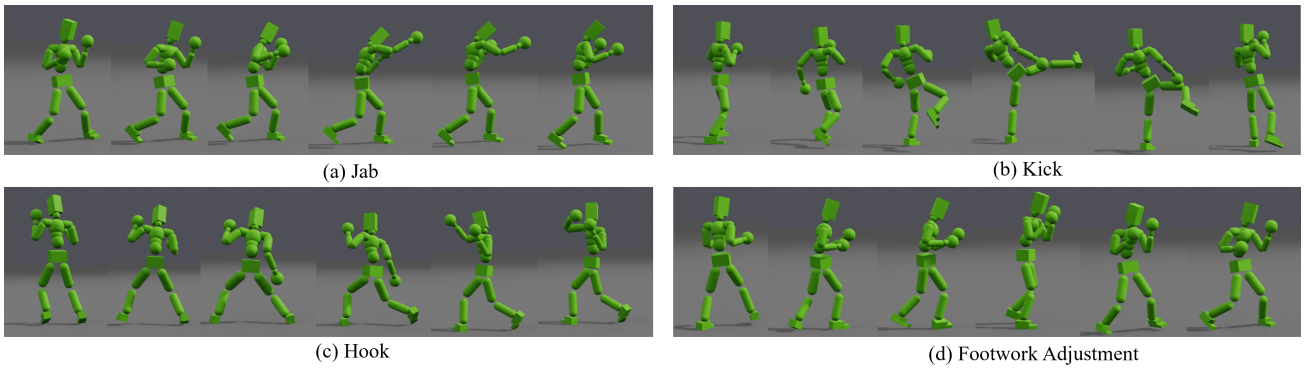


图2. 随机潜变量条件化轨迹的定性示例，由 SLMP 生成。均匀采样的潜在编码产生多样、稳定且符合物理规律的全身动作。补充视频中提供了更多推演。

在该节中，我们量化运动跟踪性能、评估随机展开稳定性、评估语义真实性、可视化球面潜在流形，并针对潜在表示和损失组件进行消融实验。我们通过使用简单的基于规则的奖励运行双智能体格斗任务，展示学得的潜在空间如何支持高位决策。最后，我们展示 SLMP 也可应用于真实的人形机器人平台（参见附录A）。

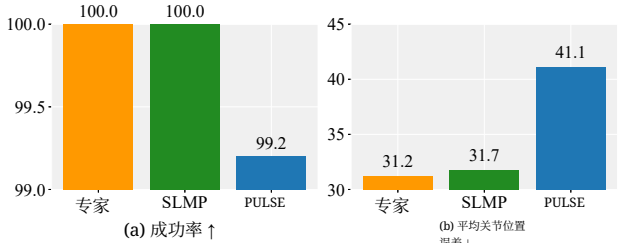


图3. 潜在空间中的动作跟踪性能。我们通过在潜在空间中跟踪参考片段来评估信息损失。与 PULSE 相比，SLMP 的成功率更高、平均关节位置误差更低，并且表现接近专家控制器。

4.1. 格斗动作捕捉数据集

本研究的一项关键贡献是构建了一个专为基于物理的人形控制设计的新型人类格斗动作捕捉数据集。我们招募了一名拥有三年以上踢拳经验的志愿者。运动数据由 Xsens MVN Link 惯性动作捕捉服以 180 Hz 的频率记录（Xsens，2013）。

参与者被要求执行全面的击打和防守技术。该数据集涵盖站姿和步法（包括不同的速度和不同的重量分布）、拳法（刺拳、后手直拳、勾拳、上勾拳和摆拳）、踢击（前踢、回旋踢、侧踢、低踢和旋转踢）、膝击、肘击，以及多种防守动作（闪避、下潜、后仰和格挡）。至关重要的是，为了支持多样化的学习格斗策略，采集的动作包含明确的速度变化（慢速、正常和快速/爆发式）以及目标高度（头部、躯干和腿部）。此外，我们还从 AMASS 数据集中选取了一小组拳击动作片段，总时长约为 8 分钟。总计而言，我们收集了约两小时的数据，并将其分割成 502 个片段，每个片段的时长约为14 秒。对于所有实验，数据集均下采样至 30 Hz。更多详情请参阅附录 B。

4.1.1. 运动跟踪性能

我们通过使用潜在编码跟踪参考片段，并在物理仿真器中展开运动，评估潜在空间引入的信息损失。对于每个参考帧，我们计算 $z_t = E(g_t)$，并重建动作$a_t = \pi_\varphi(s_t, z_t)$，随后在模拟中执行该动作以实现潜空间跟踪。我们将其与两个基准进行比较：（1）专家跟踪控制器 π_{track}（上界），以及（2）PULSE（Luo 等人，2024a），一种基于 VAE 潜在空间的先进运动先验。我们衡量成功率（跟踪过程中未跌倒或发散的片段比例，越高越好）和平均关节位置误差 （模拟身体标记点与参考身体标记点之间的平均每关节位置误差，越低越好）。图 3 显示，SLMP 的成功率高于 PULSE，平均关节位置误差低于 PULSE，并接近专家上界。这表明 SLMP 在潜空间重建过程中引入的信息损失更少。

4.2. 潜在空间评估

4.2.1. 随机展开稳定性

为评估随机展开稳定性，我们对潜在编码进行均匀采样 $z \sim \mathcal{S}^{d-1}$ 并执行由此生成的轨迹$a_t = \pi_\varphi(s_t,\ z)$ 在 Isaac Gym（Makoviychuk 等人，2021）中以固定时间范围进行测试，并与 PULSE 进行比较。由于 PULSE 使用高斯潜在分布，其潜在空间的外围区域语义密度较低，从这些区域随机采样往往会产生无效的关节

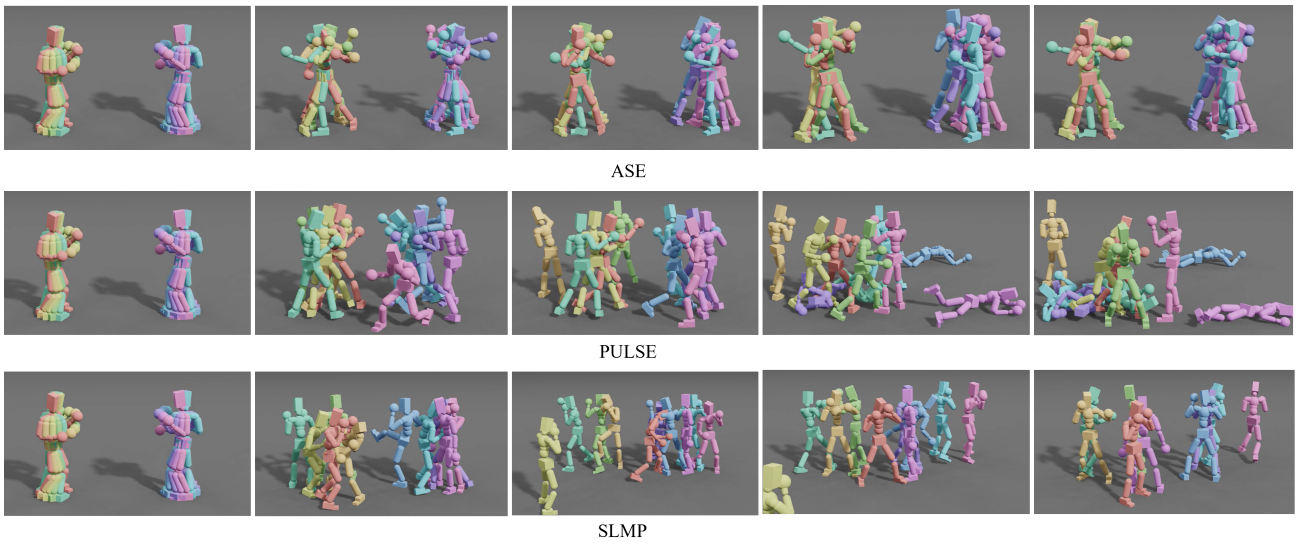


图 *4.* 随机潜变量条件化轨迹的定性比较。ASE表现出重复且多样性低的行为，PULSE偶尔会崩溃，而SLMP则能生成多样且稳定的动作。完整推演请参阅补充视频。

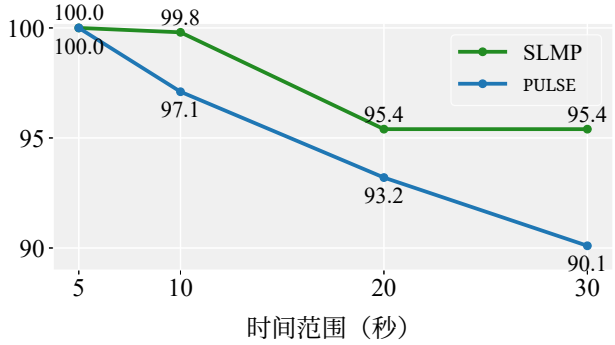


图 *5.* 在IsaacGym中进行1000次试验得到的随机潜变量展开存活曲线。SLMP在较长时间范围内保持显著更高的存活率，而PULSE的存活率则因其VAE潜在空间尾部的语义稀疏性而下降。

动作，导致角色迅速摔倒。我们对每种方法进行 1000 次随机推演，并报告在 5 秒、10 秒、20 秒和 30 秒后未摔倒的试验比例。如图 5所示，SLMP 在所有时间范围内的存活率都显著更高，而 PULSE 的存活率会随着展开时长增加而下降，这表明在随机采样下，SLMP 能为高动态格斗行为提供分布更均匀、有效性更高的潜在动作空间。

4.2.2. 随机潜变量推演的语义真实性

我们从球面先验中采样潜在编码，并从中立站姿开始展开 SLMP。图 2 展示了具有代表性的示例。SLMP 能生成丰富多样、符合物理规律且语义清晰的格斗动作，包括格挡、刺拳、踢击和步法调整

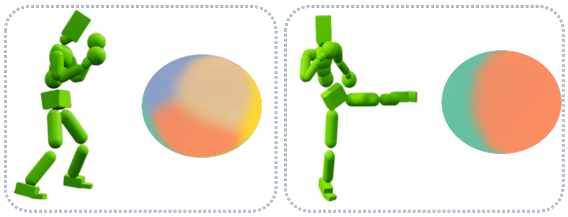


图 *6.* 状态相关的球面潜在流形。左：格挡站姿会产生多样的可行后续动作，并在球面上形成多个区域。右：空中踢击仅产生少数可行延续，使球面塌缩为一个或两个区域。

动作。我们没有观察到明显瑕疵，例如过度的脚部滑动或不自然的关节姿态。图 4 展示了 PULSE 和 ASE 的定性结果。PULSE 偶尔会失去平衡并倒下，而 ASE 虽能保持直立，却退化为重复且多样性较低的动作，例如小幅度摆臂。相比之下，SLMP 能生成多样且连贯的动作。请参阅补充视频，以更清晰地了解动作真实性和多样性的表现。

4.2.3. 球面潜在流形可视化

我们使用状态条件可视化来展示 SLMP 如何根据当前姿态改变潜在球面的结构。对于固定的人形状态 s_t，我们在单位球面上均匀采样潜在编码 z，将每一对(s_t, z) 输入先验策略以获得动作$a_t = \pi_\varphi(s_t, z)$，然后在关节空间中对所得动作进行聚类。接着，我们为每个簇分配一种颜色，并将对应的潜在点绘制在球面上，由此揭示 SLMP 认为从该状态出发有多少种不同的运动模式是可行的（图 6）。

我们对两个具有代表性的状态进行了可视化。对于中立防守姿势，SLMP 会产生多种多样的后续动作，例如刺拳、摆拳和闪避步，球面则分解为多个带颜色的区域，且区域边界平滑。这种模式表明，从该姿势出发可以执行许多性质不同但均有效的动作。相比之下，对于腾空踢击状态，大多数潜在编码都会解码为非常相似的落地或恢复动作，只有一两个簇就覆盖了几乎整个球面。球面塌缩为少数几种颜色，反映出角色在空中时几乎没有符合物理规律的选择。总体而言，这些可视化结果表明，SLMP 在潜在球面上编码了一个依赖状态的可行动作集合，从而确保随机采样在不同姿态下都保持稳定且符合物理规律。

4.2.4. 消融研究

表1. 针对潜在表示的消融实验。成功率和平均关节位置误差用于衡量潜在跟踪质量。存活率衡量均匀采样下随机展开的稳定性，定义为在 10 秒内未发生跌倒的 1000 次随机采样推演所占的百分比。

方法	成功率（%）↑	平均关节位置误差↓	存活率(%) ↑
VAE	99.8	33.7	10.2
向量量化变分自编码器	99.0	36.9	12.4
球面	**100.0**	**31.5**	0.1
SLMP（我们的方法）	**100.0**	31.7	**99.8**

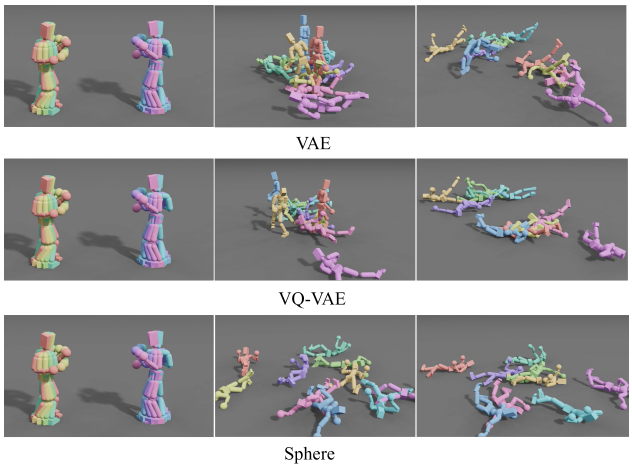


图7。 每一行展示从不同潜在结构中采样得到的一条展开轨迹。

潜在表示。我们通过将三种变体与 SLMP 进行比较，消融研究潜在表示的效果：(i) 变分自编码器潜变量，(ii) 向量量化变分自编码器潜变量，以及 (iii) 未采用我们提出的语义一致性目标的单位球面潜变量。如表 1所示，变分自编码器和向量量化变分自编码器都受到重构导致的信息损失影响，并且在随机潜变量采样下表现不稳定。尽管二者都能达到合理的跟踪精度，但存活率会大幅下降，这表明许多随机采样的潜变量并不对应物理上有效或语义有意义的运动。仅施加球面约束虽然移除了高斯先验，并略微提升了跟踪效果，但潜在空间仍缺乏语义结构。因此，随机采样的潜变量通常对应不连贯或协调性较差的行为，导致频繁跌倒，存活率接近于零。

相比之下，SLMP 保持了结构化球面潜在流形，其中相邻的潜变量对应语义一致的动作。这种结构在保持跟踪保真度的同时，显著提升了随机采样稳定性，实现了近乎完美的存活率。图 7 进一步以可视化方式展示了定性滚动轨迹。有关更清晰的演示，请参阅补充视频。

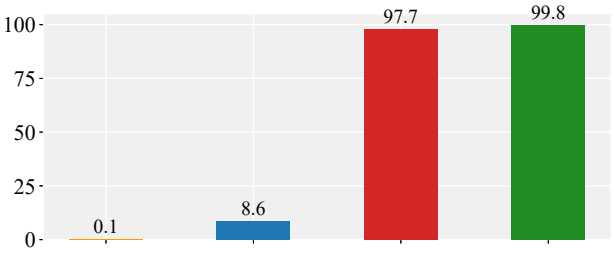


图8. 随机潜变量采样下10秒时的存活率。

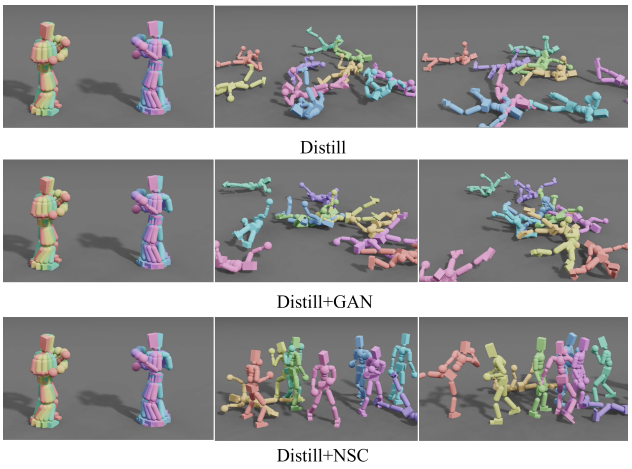
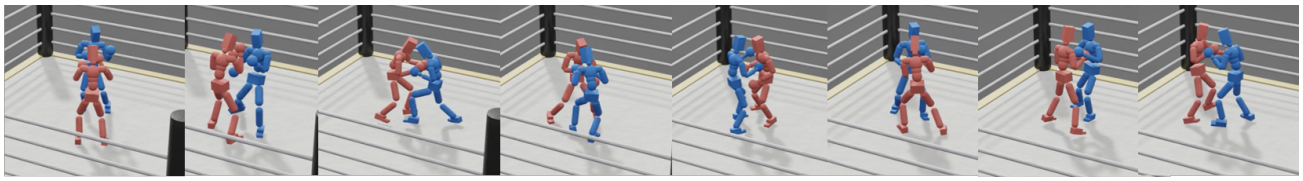
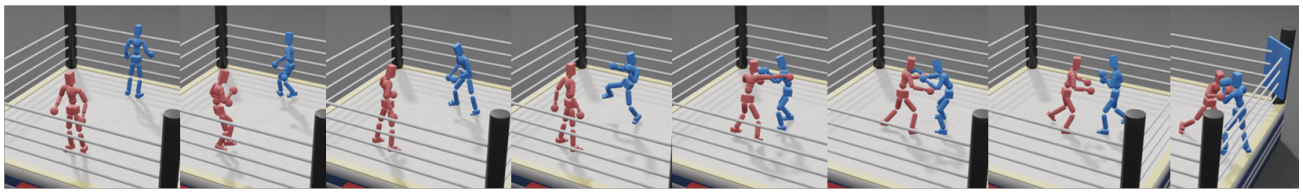


图9. 每一行展示了在不同训练损失配置下从潜在球面采样得到的展开。

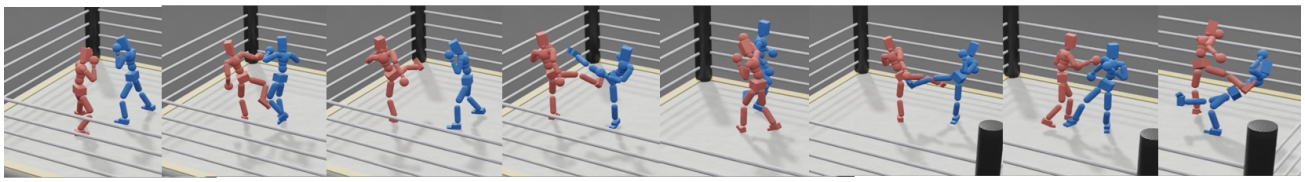
损失组件。 为分离各学习信号的贡献，我们比较了四种配置：(i) Distill、(ii) Distill+生成对抗网络、(iii) Distill+最近语义一致性（不包含判别器语义）以及 (iv)SLMP。由于四种配置均实现了近乎完美的潜变量



NPC-Style Reward



NPC-Style + AMP Reward



Simple Rule-based Reward

图 *10.* 以 SLMP 作为运动先验的双智能体格斗任务定性结果。我们比较了稠密 NCP 式奖励、Smplolympics 式奖励（以 AMP 奖励增强的 NCP 式奖励）以及稀疏规则奖励。NCP 式奖励会导致退化的静止交互，而 Smplolympics 式奖励会使策略偏向重复的出拳行为。相比之下，结合 SLMP 的稀疏规则奖励能够产生多样且类人的战斗行为。

跟踪精度，我们重点关注随机展开稳定性，并报告多个时间范围内的存活率。

我们在图 8 中可视化了 10 秒时域上的生存函数，其中条形图报告了保持未跌倒的 rollout 比例。这一差距表明，在随机采样下，SLMP 维持了显著更稳定且整体有效的潜在动作空间。

如图 9 所示，三种变体在随机采样下均表现出不同程度的不稳定性。Distill 对潜在流形没有任何正则化，因此会产生无约束动作。Distill+生成对抗网络仅从对抗性 logit 接收粗粒度的分布级梯度，无法为连续控制提供可操作的指导。Distill+NSC 为大多数样本提供了有意义的约束，但当语义不同的行为争夺球面上的邻近区域时，仍存在歧义。SLMP 通过将几何邻近性与判别器提供的语义线索相结合，在所有时域上实现了显著更高的存活率。更清晰的演示请参阅补充视频。

4.3. 高层双智能体格斗

我们进一步评估所学潜在空间在竞争环境中对高层决策的支持程度。我们将其与先前工作中常用的两种奖励设计进行比较。第一种遵循 NCP（Zhu 等人，2023），其依赖于密集塑形奖励，包括朝向、邻近性、运动速度以及末端执行器接触。第二种采用 Smplolympics（Luo 等人，2024b）中使用的奖励设计，在 NCP 式塑形奖励的基础上额外加入基于 AMP 的对抗奖励，以强制保证动作真实性。

我们的设定仅使用稀疏的规则型格斗事件：智能体击中对手时给予命中奖励，对手倒地时给予击倒奖励。不使用模仿、对抗或塑形项。尽管没有密集塑形，SLMP 仍使高层策略能够通过自博弈发现多样的进攻和闪避行为。如图 10所示，采用 NCP 式奖励训练的策略往往会收敛到退化策略：两个智能体彼此靠近，却无法执行连贯的打击动作。Smplolympics 式奖励虽然提升了动作真实性，却使策略偏向少数行为，导致反复采用拳击模式，战术多样性有限。使用 SLMP 和稀疏规则奖励时，智能体则会表现出有条理的击打、步法和反击行为。这些结果表明，SLMP 提供了强大的运动先验，减少了对复杂奖励塑形的需求，并能实现丰富多样的多智能体行为。

5. 结论

我们提出了球面潜在运动先验（SLMP），这是一种两阶段框架，旨在学习高保真且具有语义结构的运动先验，用于基于物理的人形控制。SLMP 将专家蒸馏与球面潜在流形及判别器引导的语义一致性目标相结合，从而缓解基于变分自编码器的先验所造成的信息损失，以及对抗方法中的模式崩溃倾向。在大规模格斗动作捕捉数据集上开展的实验表明，SLMP 能够保留精细的运动细节、支持稳定的随机采样，并使智能体在稀疏奖励信号下展现出类人的多智能体格斗行为。除了仿真人形化身之外，我们还在物理仿真中的逼真人形机器人模型上验证了 SLMP，证明其与实际机器人形态和控制约束相兼容。这些结果凸显了结构良好的运动先验对于减少奖励工程，以及简化复杂人形任务中的高层控制的重要性。未来的工作将探索把 SLMP 扩展到规模更大、更多样化的动作数据集，将其应用于需要丰富环境交互的高层任务，并在实体人形机器人上开展真实世界验证。

参考文献

Berner, C., Brockman, G., Chan, B., Cheung, V., D.ebiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., 等。Dota 2 大规模深度强化学习。*arXiv* 预印本 *arXiv:1912.06680*，2019年。ENGINEAI Robotics。Engineai pm01 人形机器人。
h
t
t
ps://www.engineai.com.cn/product-pm01.html。访问日期：2026-03-01。Gao, J., Wang, Z., Xiao, Z., Wang, J., Wang, T., Cao, J., Hu, X., Liu, S., Dai, J., 和 Pang, J。Coohoi：利用操纵物体动力学学习协作式人类—物体交互。神经信息处理系统进展，37:79741–79763，2024年。Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., 和 Bengio, Y。生成对抗网络。*ACM*通讯，63(11):139–144，2020年。Jing, T., Chen, S., Li, Y., Xu, W., 和 Xu, R。FARM：面向基于物理的高动态人形控制的帧加速增强与残差混合专家。*arXiv* 预印本 *arXiv:2508.19926*，2025年。Kingma, D. P. 和 Welling, M。自编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013。

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., 和 Black, M. J. SMPL：一种多人物蒙皮线性模型。*ACM* 图形学汇刊（*SIGGRAPH Asia* 论文集），34(6): 248:1–248:16，2015年10月。Luo, Z., Hachiuma, R., Yuan, Y., 和 Kitani, K. 用动力学调节的运动学策略进行自我中心姿态估计。神经信息处理系统进展，34: 25019–25032，2021。Luo, Z., Cao, J., Kitani, K., Xu, W. 等。面向实时模拟化身的持续人形控制。载于*IEEE/CVF*计算机视觉国际会议论文集，第10895–10904页，2023。Luo, Z., Cao, J., Merel, J., Winkler, A., Huang, J., Kitani, K. M. 和 Xu, W. 面向基于物理的控制的通用人形运动表示。载于第十二届国际表征学习会议，2024a。网址
https://openreview.net/forum?id=OrOd8PxOO2。Luo, Z., Wang, J., Liu, K., Zhang, H., Tessler, C., Wang, J., Yuan, Y., Cao, J., Lin, Z., Wang, F. 等。Smplolympics：面向物理模拟人形机器人的体育环境。*arXiv*预印本 *arXiv:2407.00187*，2024b。Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G. 和 Black, M. J. AMASS：以表面形状形式存档的动作捕捉数据。载于*IEEE/CVF*计算机视觉国际会议论文集，第5442–5451页，2019。Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A. 等。Isaac Gym：用于机器人学习的高性能GPU物理仿真。*arXiv*预印本 *arXiv:2108.10470*，2021。Pan, L., Yang, Z., Dou, Z., Wang, W., Huang, B., Dai, B., Komura, T. 和 Wang, J. Tokenhsi：通过任务标记化统一合成物理人体—场景交互。载于计算机视觉与模式识别会议论文集，第5379–5391页，2025年。Peng, X. B., Abbeel, P., Levine, S. 和 Van de Panne, M. DeepMimic：通过示例引导的深度强化学习习得基于物理的角色技能。*ACM*图形学汇刊（*TOG*），37(4):1–14，2018年。Peng, X. B., Ma, Z., Abbeel, P., Levine, S. 和 Kanazawa, A. AMP：用于风格化、基于物理的角色控制的对抗运动先验。*ACM*图形学汇刊（*ToG*），40(4):1–20，2021。Peng, X. B., Guo, Y., Halper, L., Levine, S. 和 Fidler, S. ASE：面向物理模拟角色的大规模可复用对抗性技能嵌入。*ACM*图形学汇刊（*TOG*），41(4):1–17，2022。

Ross, S., Gordon, G. 和 Bagnell, D. 将模仿学习和结构化预测归约为无遗憾在线学习。载于第十四届国际人工智能与统计学会议论文集，第627–635页。JMLR研讨会与会议论文集，2011。Schulman, J., Wolski, F., Dhariwal, P., Radford, A. 和 Klimov, O. 近端策略优化算法。*arXiv*预印本 *arXiv:1707.06347*，2017。Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. 等。在没有人类知识的情况下精通围棋。《自然》，550(7676):354–359，2017。Tessler, C., Kasten, Y., Guo, Y., Mannor, S., Chechik, G. 和 Peng, X. B. CALM：用于可引导虚拟角色的条件对抗潜在模型。载于*ACM SIGGRAPH 2023*会议论文集，第1–9页，2023。Tessler, C., Guo, Y., Nabati, O., Chechik, G. 和 Peng, X. B. MaskedMimic：通过掩码运动修复实现统一的基于物理的角色控制。*ACM*图形学汇刊（*TOG*），43(6):1–21，2024。宇树科技。Unitree G1 人形机器人。https://www.unitree.com/cn/g1。访问日期：2026-03-01。Van Den Oord, A., Vinyals, O. 等。神经离散表示学习。神经信息处理系统进展，30，2017。Wang, Y., Zhao, Q., Yu, R., Tsui, H. W., Zeng, A., Lin, J., Luo, Z., Yu, J., Li, X., Chen, Q. 等。Skillmimic：从演示中学习篮球交互技能。载于计算机视觉与模式识别会议论文集，第17540–17549页，2025。Won, J., Gopinath, D. 和 Hodgins, J. 物理模拟角色进行双人竞技运动时的控制策略。*ACM*图形学汇刊（*TOG*），40(4):1–11，2021。Xsens, M. 使用微型惯性传感器进行完整的六自由度人体运动跟踪。*Daniel RoetenbergLuingeHenk*，2013。Zhang, Z., Bashkirov, S., Yang, D., Shi, Y., Taylor, M. 和 Peng, X. B. 通过对抗式微分判别器进行基于物理的运动模仿。载于*SIGGRAPH Asia 2025*会议论文集（*SIGGRAPH Asia '25*会议论文集），2025。Zhu, Q., Zhang, H., Lan, M. 和 Han, L. 用于基于物理的角色控制的神经类别先验。*ACM*图形学汇刊（*TOG*），42(6):1–16，2023。

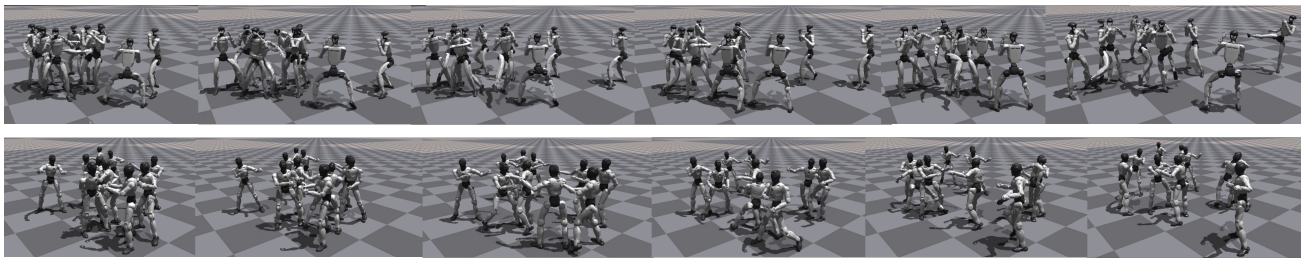


图 *11.* 在仿真中对逼真的人形机器人模型进行随机潜变量采样。上：Unitree G1。下：ENGINEAI PM01。两个模型都能在随机潜变量输入下稳定地产生多样化的动作。

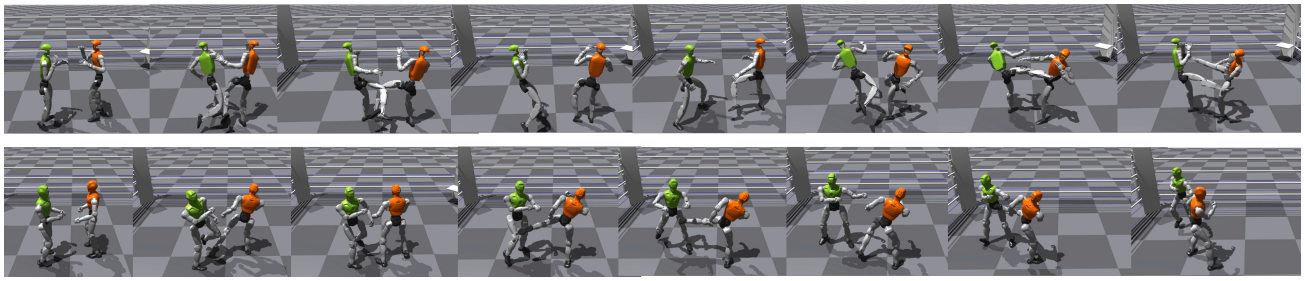


图 *12.* 双智能体对抗：类似人类的仿真中对逼真的人形机器人模型进行评估。上：Unitree G1。下：ENGINEAI PM 对抗行为，遵循 SLMP 先验。

A. 在真实人形机器人模型上的验证

为了进一步评估 SLMP 的泛化能力，我们在物理仿真中的真实人形机器人模型上验证了所学运动先验。与主要实验中使用的基于 SMPL 的虚拟形象不同，这些模型反映了实际的机器人运动学结构。具体而言，我们在 Unitree G1（宇树科技）和 ENGINEAI PM01（ENGINEAIRobotics）人形机器人模型上开展了实验。

A.1. 随机潜变量采样

我们首先评估随机潜变量采样在两种机器人模型上的表现。如图11所示，这些机器人能够稳定地生成在随机条件下的多样化运动域采样，表明 SLMP 能够泛化到不同的人形形态。

A.2. 双智能体战斗评估

我们进一步使用逼真的机器人模型，在双智能体对抗场景中评估 SLMP。如图12所示，机器人能够生成类人战斗行为，表明 SLMP 支持不同人形形态之间的高层交互。

总体而言，这些结果表明，SLMP 并不局限于单一的虚拟人形表示，而是能够在物理仿真中泛化到逼真的机器人形态。这一验证显示，SLMP 未来有望部署于实体人形平台。

B. 数据集

表2总结了数据集的详细统计信息。

C. 实现细节

C.1. 网络架构

所有网络均采用多层感知机（MLP）实现，其深度和宽度因组件而异，汇总见表 3。两个高层战斗控制器 π_{h_1} 和 π_{h_2} 采用相同的架构。

表2。统计　格斗动作捕捉数据集（502个片段）的统计信息。该数据集涵盖多种速度和目标级别。

类别	动作类型	属性与目标	片段	Dur.
站姿	正架（静止、格挡）、重心转移	速度：慢速、正常	20	11s
步法	迈步、转身（90°）、滑步	步幅：小、大	62	11s
直拳	刺拳、后手直拳	目标：头部、躯干、向下	60	11s
勾拳与摆拳	勾拳、摆拳	目标：头部、躯干；类型：平勾／上勾	50	11s
上勾拳	上勾拳（左/右）	目标：头部、躯干	25	11s
前踢	前踢（左/右）	目标：躯干；包括重型沙袋	40	11s
回旋踢	回旋踢（左/右）	目标：低位、中位、高位；包括重型沙袋	55	11s
其他踢法	侧踢、低位、旋转踢	目标：腿部、躯干、头部；包含重型沙袋	55	11s
近距离	膝击、肘击	目标：躯干、头部；方向：多方向	30	11s
防御	格挡、闪躲、下潜躲闪、后仰闪避、拦截	强度：由低到高	70	11s
组合拳	刺拳-直拳-勾拳变式，反击	序列：2-5个打击动作；混合防御	35	10至20秒
总计	**所有类别**	**–**	**502**	**~2.0小时**

C.2. 训练超参数

我们使用近端策略优化训练专家跟踪控制器 π_{track}。潜在先验通过 DAgger 风格的在线模仿过程进行训练（Ross 等人，2011）：学生策略执行轨迹采样，专家则提供同策略监督。表 4汇总了关键超参数。

表 *3.* 我们的框架中使用的网络架构。所有网络均实现为多层感知机（多层感知机）。

组件	隐藏单元	激活函数
π_{track}	[2048, 1536, 1024, 1024, 512, 512]	SiLU
$E(\cdot)$	[512, 256]	ReLU
π_{φ}	[4096, 2048, 1024, 1024, 512, 512]	SiLU
$D(\cdot)$	[4096, 2048, 1024, 512]	ReLU
π_{h_1}, π_{h_2}	[2048, 1024, 512]	SiLU

表 *4.* 我们实验中使用的关键超参数。

近端策略优化超参数					
批量大小	环境数量	LR	γ	片段 ϵ	
32768	1024	5×10^{-5}	0.99	0.2	
SLMP 超参数					
λ_{distill}	λ_{disc}	判别器学习率	λ_{DLSC}	β	d
1.0	1×10^{-4}	5×10^{-5}	1.0	0.1	64

C.3. 软件与硬件配置

所有仿真均在 Isaac Gym 中实现，物理仿真时间步长为 60 Hz，比例微分控制也以相同频率运行。训练时，运动数据会降采样至 30 Hz。实验在搭载 Ubuntu 22.04、Python 3.8 和 PyTorch 2.1 的机器上进行。训练使用的服务器配备了两块 NVIDIA RTX 4090 GPU。训练跟踪控制器通常需要 24 小时，训练潜在先验需要 12 小时，训练双智能体对战策略则约需 8 小时。

D. 任务详细表述

我们补充介绍三个训练阶段的实现细节，重点说明状态/观测设计和奖励函数，因为它们对保证实验可复现性最为关键。

D.1. 阶段 1：运动跟踪控制器

我们按照 PHC（Luo 等人，2023）采用目标条件强化学习的形式训练专家运动跟踪控制器。该任务是一个马尔可夫决策过程，其中策略为每个驱动自由度输出 PD 目标，模拟器通过 PD 控制器施加力矩。

状态。仿真状态定义为

$$s_t \triangleq (s_t^p, s_t^g), \tag{12}$$

其中 s_t^p 是人形机器人的本体感知，s_t^g 是参考驱动目标。本体感知状态为

$$s_t^p \triangleq (q_t, \dot{q}_t), \tag{13}$$

其中，q_t 和 $\dot{q}_t$ 分别表示仿真中的人形机器人姿态和速度。

目标状态采用下一步参考量与仿真对应量之间的差异：

$$s_t^{g\text{-rot}} \triangleq (\hat{\theta}_{t+1} \ominus \theta_t,\ \hat{p}_{t+1} - p_t,\ \hat{v}_{t+1} - v_t,\ \hat{\omega}_{t+1} - \omega_t,\ \hat{\theta}_{t+1},\ \hat{p}_{t+1}), \tag{14}$$

其中，$\ominus$用于计算旋转差。

所有量 (s_t^p、s_t^g) 均根据人形机器人的当前朝向和根位置进行归一化，与 PHC（Luo 等人，2023）一致。

奖励。逐步奖励由任务模仿项、AMP 风格判别器项和能量惩罚组成：

$$r_t = 0.5\, r_t^g + 0.5\, r_t^{\text{amp}} + r_t^{\text{energy}}. \tag{15}$$

对于动作跟踪，任务奖励的形式为

$$\begin{aligned} r_t^{g\text{-imitation}} = {} & w_{\text{jp}} \exp\big(-100\|\hat{p}_t - p_t\|\big) + w_{\text{jr}} \exp\big(-10\|\hat{q}_t \ominus q_t\|\big) \\ & + w_{\text{jv}} \exp\big(-0.1\|\hat{v}_t - v_t\|\big) + w_{\text{j}\omega} \exp\big(-0.1\|\hat{\omega}_t - \omega_t\|\big), \end{aligned} \tag{16}$$

该奖励衡量模拟动作与参考动作之间在连杆平移、旋转、线速度和角速度方面的差异。

能量惩罚定义为

$$r_t^{\text{energy}} = -0.0005 \sum_{j \in \text{joints}} \|\mu_j \omega_j\|^2, \tag{17}$$

其中 μ_j 和 ω_j 分别为关节力矩和关节角速度。风格奖励 r_t^{amp} 使用 AMP 风格判别器计算，其观测设计和训练目标与 PHC（Luo 等人，2023）相同。

D.2. 阶段 2：SLMP 潜在先验

状态与目标。策略输入采用与阶段 1 相同的本体感知状态定义。目标 g_t 源自参考动作，并由目标编码器按照主论文中的描述进行编码。本阶段不引入其他观测。

D.3. 阶段 3：双智能体对抗

观测。观测设计基本遵循 SMPLOlympics 中的拳击设置（Luo 等人，2024年b）。每个智能体都会在自我中心坐标系中接收自身信息和对手信息，包括：

- 自身本体感知：关节姿态、速度和根状态，

- 相对于智能体的对手根部姿态和速度，
- 智能体的攻击肢体（手部和足部）与对手的得分区域（头部和躯干）之间的相对位置，
- 关键身体部位所受的接触力大小。

所有量均以航向对齐的局部坐标系表示。

奖励函数。 我们采用简单的基于规则的稀疏奖励，重点鼓励有效击打。

命中奖励。 当智能体的手部或足部进入距对手头部或躯干 0.3 米以内，且相应的接触力超过 30 时，给予正向奖励。奖励大小与测得的接触力成正比。

命中惩罚。一个 在相同条件下，智能体自身的头部或躯干被击中时，也会施加对称惩罚 s.

击倒奖励。对手倒下时奖励 $+50$。反之，智能体倒下时会施加 -50 的惩罚。

终止条件。出现以下情况时，回合将终止：

- 任一智能体被击倒，
- 骨盆间距离低于 0.3 米且持续超过 1 秒，
- 攻击肢体与对手的得分区域之间的距离低于 0.3 米且持续超过 1 秒（以防止奖励投机），
- 在早期训练期间，如果骨盆距离超过 1.2 米。

自我对弈设置。 我们采用一种交替式自我对弈方案，使用两个独立的策略实例。在任意时刻，一个策略被指定为学习智能体，另一个则作为固定对手。训练期间，使用近端策略优化更新学习智能体，同时冻结对手策略。每训练 250 轮，双方互换角色：此前固定的策略变为可训练策略，另一个则保持固定。

这种交替更新可防止两个智能体同时偏移，从而稳定竞争性学习，并鼓励智能体持续适应不断进步的对手。两个策略的初始化完全相同，但在训练过程中独立演化。