Robotica（2026），**xx**:
xx 1–7 doi:10.1017/xxxxx

CAMBRIDGE
UNIVERSITY PRESS

研究论文



HierKick：用于视觉引导的足球机器人控制的分层强化学习

YizhiChen[1,2], Zheng Zhang[2,3], Zhanxiang Cao[2,3], Yihe Chen[4], Shengcheng Fu[1,2], Liyun Yan[2,3], YangZhang[2,3], Jiali Liu[2,5], HaoyangLi[2,3] 和 Yue Gao*[2,3]

[1]中国上海同济大学。[2]中国上海创新研究院，上海。
[3]中国上海交通大学，上海。[4]中国江西财经大学。
[5]中国上海华东师范大学。通讯作者。电子邮箱:
yuegao@sjtu.edu.cn。

关键词：分层强化学习，人形足球机器人，多模态感知

arXiv:2603.00948v1 [cs.RO] 1 Mar 2026

摘要

足球机器人的控制涉及多时间尺度决策，需要平衡长期战术规划与短期动作执行。传统的端到端强化学习（RL）方法在复杂动态环境中面临挑战。本文提出 HierKick，这是一种基于双频分层强化学习的视觉引导足球机器人控制框架。该框架采用分层控制架构，包含一个 5 Hz 高层策略，该策略集成 YOLOv8 进行实时检测，并通过教练模型选择任务；同时还包含一个经过预训练的 50 Hz 低层控制器，用于精确的关节控制。借助这一架构，框架实现了接近、对齐、带球和射门四个步骤。实验结果表明，该框架在 IsaacGym、MuJoCo 和真实世界中的成功率分别为 95.2%、89.8% 和 80%。HierKick 为复杂环境中的机器人控制提供了一种有效的分层范式，可扩展至多时间尺度任务；其模块化设计和技能复用为智能机器人控制开辟了新路径。

1. 引言

近期，人形机器人研究已从实现稳定行走[1, 2, 3] 发展到在真实场景中完成复杂任务[4, 5, 6]。这一进展与生物进化过程中精妙的神经调节机制相呼应。例如，当运动员追逐滚动的足球时，大脑就像战略指挥中心，迅速决定是带球突破对手，还是传球配合。与此同时，小脑则充当精密的执行器，不断调节肌肉张力，以维持平衡并校准踢球角度。高层决策与低层执行之间的协同，体现了生物体如何应对具有不同时间尺度的任务。

足球任务是一项出色的长序列执行验证任务[7]。由于球的轨迹难以预测，策略必须具有鲁棒性。视觉感知容易受到遮挡和光照变化的影响，使机器人无法获取完整的全局信息。因此，机器人必须规划战术，以接近球[8]、向前带球，并瞄准或射门，同时在高

© 作者 2026。

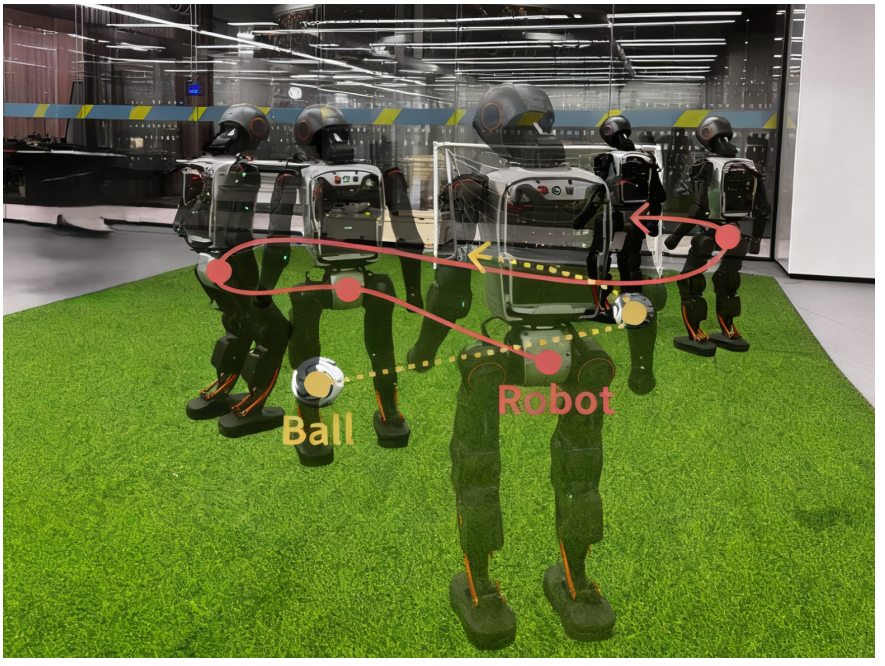


图 ***1.*** 机器人轨迹图。该图展示了机器人在足球场上的轨迹。在教练模型的引导下，机器人基于 *HierKick* 框架执行一系列动作，包括接近、对齐、带球和射门。

层面上进行操作，同时在低频层面维持平衡、调整步长并控制踢球力度。这两个层面一旦脱节，就会直接导致任务失败。

当前技术面临多项限制：奖励稀疏导致收敛困难[9, 10],策略缺乏稳定性，且控制延迟较高，难以满足低层动作要求[11]。此外，通常需要大量人工干预，而且难以适应动态环境变化[12]。为解决这些问题，本文提出了 HierKick，一种面向视觉引导足球机器人的双频分层强化学习框架。其核心思想是通过解耦并协同“战术决策—动作执行”，整合预训练技能，并利用多模态感知增强鲁棒性，从而平衡多种时间尺度需求。

我们提出了 HierKick 框架，该框架采用层次化频率设置。具体而言，我们训练了一个教练模型，引导机器人为不同任务选择合适的速度。教练模型输出加速度，并将其积分，以生成机器人在每个高层时间步的速度命令。该设计使 HierKick 在仿真环境中的接近—带球—射门复合任务中达到 95.2% 的成功率，并在真实世界部署中达到 80% 的成功率。这些结果验证了该框架的有效性和迁移能力。对比实验和消融实验进一步表明，双频分层控制增强了策略稳定性和实时性能，而复用预训练技能则加快了训练并改善了收敛性。此外，大规模并行训练和噪声建模提升了鲁棒性和样本效率。在 50 Hz 的控制频率下，系统的端到端控制延迟约为 20 ms，满足实时控制要求。

本文的主要贡献如下：

• 我们提出了 HierKick，这是一种采用层次化频率设置的双频分层强化学习框架，可实现机器人的自适应速度控制。
• 该框架采用通过课程学习训练的教练模型和分层奖励函数。该模型评估场上情形并输出加速度信号，再将这些信号积分，以便在每个高层时间步生成速度指令，使机器人逐步掌握各种技能。
• 我们成功地在 Booster T1 实体人形机器人上实现了分层控制策略，在 50 Hz 控制频率下实现了约 20 毫秒的端到端控制延迟，满足实时控制的实际需求。

2. 相关工作

2.1. 基于强化学习的人形机器人动态运动

近年来，强化学习（RL）在控制人形机器人的动态运动方面取得了显著成功[13, 14, 15]。许多研究采用无模型强化学习技术[16]，使机器人能够掌握多样且灵活的运动技能。通过采用多阶段训练方案，机器人在跳跃[17] 和快速恢复[18] 等任务中的鲁棒性得到了显著提升。此外，一些研究将视觉系统融入其中，帮助机器人在复杂地形中行进[19, 12, 20]。然而，这些研究主要关注人形机器人的运动方面。相比之下，我们的方法 HierKick 针对动态足球任务，该任务不仅要求运动稳定，还要求机器人能够与外部物体进行精确交互。

2.2. 基于学习的机器人足球技能

机器人足球长期以来一直是足式机器人运动与协调领域的基准任务，尤其是在 RoboCup 系列赛事中[21]。早期 RoboCup 参赛者通常依赖基于规则或脚本的控制策略[22],，而如今人们正日益转向基于学习的方法。尤其是四足机器人，已通过强化学习展现出出色的带球和守门技能[23]。然而，双足机器人的相关方法[24] 通常仅限于离散踢球动作，缺乏对球的连续控制。一些端到端学习方法虽已在仿真中实现灵活控制，但往往难以应对多时间尺度决策的挑战[25], , 导致策略收敛困难且控制延迟较高。

为了解决这些问题，我们提出的 HierKick 方法采用分层强化学习（HRL）[26], 这一有效范式，通过将任务分解为高层教练策略和低层策略来简化学习。与许多传统的分层强化学习方法不同，后者同时学习两种策略可能导致训练不稳定[27], , HierKick 使用预训练低层运动控制器。该设计将低层策略视为固定的运动生成器，从而缓解冷启动问题，提高训练稳定性和效率，并使高层策略能够专注于长期任务规划。

3. 方法

面向人形机器人动态足球任务的 HierKick 框架分为两个主要部分（见图 2）：高层控制和低层执行，从而实现战术决策与运动执行之间的时间尺度解耦。在以 5 赫兹运行的高层控制模块中，YOLOv8 可对球和球门进行实时视觉感知。教练策略将这些感知数据与历史加速度和速度指令相结合，生成加速度指令，从而实现自适应速度调节，以进行战略任务规划。此外，我们提出了一种在促进学习过程中发挥关键作用的多阶段奖励机制。该机制针对以下任务设计：

接近、对齐、带球和射门。为避免控制行为不稳定，并确保指令间平滑过渡，我们使用增量逻辑在奖励函数中加入正则化项。这种一体化设计避免了目标冲突，使技能能够逐步积累，并确保运动符合具体情境。

3.1. 问题陈述

我们将人形足球机器人的控制建模为一个具有两个不同决策层级的分层马尔可夫决策过程（MDP）[28]。高层MDP定义为$(S,\ A,\ f,\ r,\ S,\ \gamma)$，其中 S 表示高层状态空间，A 表示高层动作空间，f 表示高层系统动力学，r 表示高层奖励函数，S 表示初始高层状态，而 γ 表示折扣因子。

类似地，低层MDP[29]定义为„$S,\ A,\ f,\ r,\ S,\ \gamma$”。HierKick方法的目标是学习一个高层策略 π，使其最大化期望折扣回报 $G_t = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t r^t\right]$，同时利用预训练的低层策略 π 执行运动。我们采用带有非对称演员-评论家架构的近端策略优化（PPO）[30]，在训练期间使用特权信息，并在部署期间依赖部分观测。低层策略基于传统的运动训练方法，因此本文不再对此作进一步阐述。

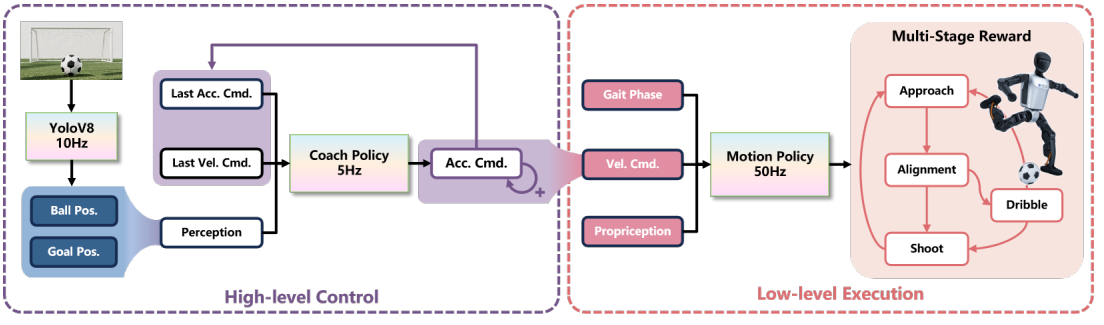


图 *2. HierKick* 的分层控制框架。该框架由三个相互连接的模块组成：*(1)* 感知：以 *10 Hz* 运行的 *YOLOv8* 检测器处理视觉输入，确定足球和球门的位置；*(2)* 高层控制：采用以 *5 Hz* 运行的“教练策略”，通过 *PPO* 强化学习融合感知结果和历史控制数据（上一次的加速度指令和速度指令），生成加速度指令；*(3)* 低层执行：以 *50 Hz* 运行的“运动策略”接收速度指令、步态相位和本体感觉等输入，执行一系列足球动作（接近、对齐、带球、射门），训练则由多阶段奖励机制引导。

3.2. 高层教练策略

高层教练策略以每10个仿真步一次的频率运行，对应200毫秒的间隔。该频率与人类战术决策的时间尺度一致，有助于避免可能破坏足球比赛稳定性的过度反应行为。在每个时间步，该策略接收足球特定观测，并输出速度增量命令以引导机器人的动作。

3.2.1. 观测空间

高层观测空间由三个主要类别组成：足球特定感知、机器人状态和命令历史。这些组成部分共同为决策过程提供信息。观测空间的形式化定义如下：

$$\mathbf{o}_{\text{high-level}} = \left[\mathbf{p}_{\text{robot-ball}}, \mathbf{p}_{\text{robot-goal}}, d_{\text{ball}}, d_{\text{goal}}, \mathbf{v}_{\text{robot}}, \mathbf{c}_{\text{prev}}\right]$$

这里，$\mathbf{v}_{\text{robot}}$ 表示机器人沿x轴和y轴方向的线速度，以及绕z轴的角速度。$\mathbf{p}_{\text{robot-goal}}$ 表示从机器人基座指向球门中心的向量，该向量是利用机器人视觉系统检测到的场地地标计算得出的。d_{ball}和 d_{goal} 分别表示机器人到球和球门的直线距离。最后，$\mathbf{c}^{\text{prev}}$ 表示在上一个时间步发出的速度增量命令，有助于保持机器人运动的连续性。

3.2.2. 动作空间

高层动作空间是一个由速度增量构成的三维连续向量 $\Delta\mathbf{v} = \{\Delta v_x, \Delta v_y, \Delta\omega_z\}$，其中 Δv_x、Δv_y 和 $\Delta\omega_z$ 分别表示前后向线速度（x轴）、侧向线速度（y轴）和偏航角速度（z轴）的增量变化。这些增量受到限制，以确保机器人安全运行：$\Delta v_x \in \{-0.2, 0.2\}$ m/s、$\Delta v_y \in \{-0.1, 0.1\}$ m/s，以及 $\Delta\omega_z \in \{-0.1, 0.1\}$ rad/s。速度命令通过递归方式更新，以保持运动平滑：

$$\mathbf{v}_{\text{cmd}}(t) := \mathbf{v}_{\text{cmd}}(t-1) + \Delta\mathbf{v}(t)$$

这种设定使速度能够逐步调整，同时保持机器人运动的时间连续性。

3.3. 低层运动控制器

低层运动控制器采用预训练踢球模型，该模型展现出稳健的行走和转向能力。它接收来自高层教练策略的速度指令，并以 50 Hz 的频率生成关节位置目标。该频率确保机器人运动得到精确控制，使其在动态任务中能够保持稳定且响应灵敏的动作。

3.3.1. 观测空间

低层观测空间着重捕捉机器人身体和运动的实时动态状态，为维持稳定性并实现精准的运动控制提供关键信息。该观测空间为 47 维，正式定义如下：

$$\mathbf{o}_{\text{low-level}} = \left[\mathbf{g}, \boldsymbol{\omega}, \mathbf{v}_{\text{cmd}}, \cos(2\pi\phi), \sin(2\pi\phi), \mathbf{q}, \dot{\mathbf{q}}, \mathbf{a}_{\text{prev}}\right]$$

这个 47 维空间整合了运动学、动力学和时间数据，使低层控制器能够以 50 Hz 的频率生成稳定且响应灵敏的关节目标。通过纳入高层速度指令并保持精准跟踪，系统确保了运动过程中的动态适应性和稳定性。

3.3.2. 动作空间

低层动作空间是一个 12 维连续空间，可直接控制人形机器人的腿部目标关节位置。这 12 个维度对应人形机器人每条腿的六个自由度。关节位置目标由 PD控制器 [31], 处理，PD控制器会将目标位置转换为实际的关节驱动信号。这一机制可确保准确跟踪

所需位置，减轻运动振荡，并在任务执行过程中保持机器人运动的稳定性。

3.4. 多阶段奖励机制

为解决足球任务不同阶段之间行为优先级相互冲突的问题，我们设计了一组距离自适应奖励函数。这些函数会根据机器人与球的距离 d_{ball} 动态调整，确保高层策略优先考虑各阶段特定的目标，同时与整体任务目标保持一致。

基于课程学习的原理 [32] 以及机器人学中的有限状态机 (FSM) [33] （它们将复杂任务分解为易于管理的子任务），我们提出了一种多阶段学习策略。该策略根据机器人与球的距离调整奖励优先级，使机器人能够依次习得适合各阶段的技能，同时不损害任务的连贯性。此外，这种方法还能缓解策略陷入局部最优的常见问题。

完整的分层奖励函数整合了所有阶段特定的组成部分，表示如下：

$$r_{\text{total}} = \mathbf{w} \cdot \mathbf{r}$$

其中，$\mathrm{w} = \text{»}\lambda_1,\ \lambda_2,\ \lambda_3,\ \lambda_4,\ \lambda_5\text{...}^{\top}$ 是权重向量，反映各阶段的相对重要性。更复杂的技能会被赋予更高的权重。向量 $\mathrm{r} = \left[r_{\text{approach}}, r_{\text{alignment}}, r_{\text{dribble}}, r_{\text{shoot}}, r_{\text{delta}}\right]$ 表示各阶段的奖励，其中 r_{delta} 作为正则化项。

3.4.1. 阶段*1*接近阶段奖励

当机器人距离球较远时，首要目标是高效地朝球定向运动。接近奖励优先鼓励机器人沿球的方向移动，同时尽量减少绕行。其定义如下：

$$r_{\text{approach}} = (\mathbf{v}_{\text{robot}} \cdot \hat{\mathbf{v}}_{\text{robot-ball}}) \cdot M_{\text{approach}}(d)$$

在该表达式中，$\mathrm{v}_{\text{robot}}$ 表示高层速度命令，$\hat{}\mathrm{v}_{\text{robot-ball}}$ 是从机器人指向球的单位向量，确保奖励鼓励机器人朝球移动。M_{approach}„d”是二值阶段掩码函数，仅在接近阶段激活该奖励，其定义如下：

$$M_{\text{approach}}(d) = \begin{cases} 1 & \text{if } d > u_1 \\ 0 & \text{otherwise} \end{cases}$$

其中，u_1 表示机器人与球之间“远距离”的阈值，确保仅当机器人需要接近球时才激活该奖励。当 $\mathrm{v}_{\text{robot}}$ 与 $\hat{}\mathrm{v}_{\text{robot-ball}}$方向一致时，会获得正向$r_{\text{approach}}$，从而鼓励机器人快速、直接地接近球。

3.4.2. 阶段*2*对齐阶段奖励

当机器人接近球时，重点转向将机器人调整到适合射门角度的位置。目标是调整机器人身体，使“机器人-球-球门”连线尽可能笔直，从而降低射偏的可能性。对齐奖励兼顾角度准确度与运动平滑性，定义如下：

$$r_{\text{alignment}} = \left[\cos(\theta_{\text{rbg}}) - \beta \cdot |\Delta\omega_{\text{desired}} - \Delta\omega_{\text{actual}}|\right] \cdot M_{\text{alignment}}(d)$$

这里，θ_{rbg} 是从机器人指向球的单位向量与从机器人指向球门的单位向量之间的夹角。cos„θ_{rbg}”的取值范围为 1（完全对齐）到 -1（方向相反），用于奖励直线定位。$\Delta\omega_{\text{desired}}$ 是对齐所需的角速度，$\Delta\omega_{\text{actual}}$ 是

指令角速度。$\mathrm{j}\Delta\omega_{\text{desired}} - \Delta\omega_{\text{actual}}\mathrm{j}$用于惩罚因转向不精确造成的失准。参数 $\beta = 0.5$ 用于平衡对齐精度与运动平滑性。 $M_{\text{alignment}}(d)$"是对齐阶段的阶段掩码，定义如下：

$$M_{\text{alignment}}(d) = \begin{cases} 1 & \text{if } u_2 < d \leq u_1 \\ 0 & \text{otherwise} \end{cases}$$

其中 u_2 和 u_1 分别表示与球"相对较近"和"较远"距离的阈值。当机器人需要对准球时，该奖励被激活。

3.4.3. 第三阶段带球阶段奖励

当机器人靠近球且远离球门时，重点转向轻柔地操控球朝向球门，避免过快的速度使机器人失稳或将球击飞。带球奖励结合了朝向目标的运动和速度调节：

$$r_{\text{dribble}} = \mathbf{v}_{\text{robot}} \cdot \hat{\mathbf{v}}_{\text{ball-goal}} \cdot M_{\text{dribble}}(d_1, d_2)$$

此处，$\hat{}\mathbf{v}_{\text{ball-goal}}$ 是从机器人指向球门的单位向量，用于奖励朝球门方向的运动。 $M_{\text{dribble}}(d_1, d_2)$"是带球阶段的阶段掩码，定义为：

$$M_{\text{dribble}}(d_1, d_2) = \begin{cases} 1 & \text{if } 0 < d_1 \leq u_2,\ d_2 > u_3 \\ 0 & \text{otherwise} \end{cases}$$

其中 d_1 和 d_2 分别表示机器人与球之间、以及球与球门之间的相对距离。当条件满足时，机器人因将球带向球门而获得奖励。

3.4.4. 阶段 *4* 射门阶段奖励

在射门阶段，奖励机制鼓励机器人对齐球门并高速移动，以完成有力射门：

$$r_{\text{shoot}} = \left[\hat{\mathbf{v}} \cdot \hat{\mathbf{v}}_{\text{ball-goal}} + \mu \cdot \min(v_{\max}, \|\mathbf{v}\|_2 \cdot \alpha)\right] \cdot M_{\text{shoot}}(d_1, d_2)$$

这里，$\hat{}\mathbf{v} = \mathbf{v}_{\text{robot}}/(\|\mathbf{v}_{\text{robot}}\|_2 + \varepsilon)$"用于归一化机器人速度方向，而$\hat{}\mathbf{v}_{\text{ball-goal}}$ 是从球指向球门的单位向量。参数 μ 是速度奖励系数， $v_{\max}$ 是最大速度奖励， α 是速度放大因子。阶段掩码 $M_{\text{shoot}}(d_1, d_2)$"定义如下：

$$M_{\text{shoot}}(d_1, d_2) = \begin{cases} 1 & \text{if } 0 < d_1 \leq u_2,\ d_2 < u_3 \\ 0 & \text{otherwise} \end{cases}$$

此奖励机制鼓励机器人对齐球门，并奖励更高的速度，从而实现更有力的射门 .

3.4.5. 命令平滑正则化

为防止出现不稳定行为并确保指令间平滑过渡，引入一个正则化项：

$$r_{\text{delta}} = -\zeta_1 \cdot \max(0, \|\Delta\mathbf{v}\|_2 - \delta_{\max}) - \zeta_2 \cdot \mathbf{1}[\|\Delta\mathbf{v}\|_2 < \delta_{\min}]$$

其中， ζ_1 会对超过最大阈值 $\delta_{\max}$ 的过高速度指令施加惩罚，而 ζ_2 会对过小的指令施加惩罚，从而确保动作受到控制且能产生实质影响。

这种多阶段奖励设计通过一次仅激活一个阶段来确保阶段一致性，避免目标相互冲突。各阶段层层递进，从基础接近到复杂射门，同时连续的距离阈值可避免奖励不连续性。每个阶段都会鼓励适合足球任务情境的特定运动模式。

表 ***1.*** 多阶段和标准奖励函数汇总。总奖励是任务特定奖励（仅在特定阶段启用）和标准正则化奖励（全局启用）的加权和。

类别	组成部分	数学公式	描述 / 条件
任务特定奖励（阶段相关）			
接近	r_{approach}	“$\mathbf{v}_{\text{robot}} \cdot \hat{}\mathbf{v}_{\text{robot-ball}}$” $\cdot M_{\text{approach}}$	鼓励向球移动（$d > u_1$）
对齐	$r_{\text{alignment}}$	»cos„θ_{rbg}” $-\beta$j$\Delta\omega_{\text{err}}$j... $\cdot M_{\text{alignment}}$	对齐机器人-球-球门向量（$u_2 < d \le u_1$）
带球	r_{dribble}	„$\mathbf{v}_{\text{robot}} \cdot \hat{}\mathbf{v}_{\text{ball-goal}}$” $\cdot M_{\text{dribble}}$	轻柔地引导球朝向球门（$d_1 \le u_2,\ d_2 > u_3$）
射门	r_{shoot}	»对齐$^{\text{goal}}$，$\mu \cdot$ 最小„$v_{\max},\ v_{\text{shoot}}$”..... $\cdot M_{\text{shoot}}$	最大化踢球速度和准确度（$d_1 \le u_2,\ d_2 < u_3$）
标准奖励与正则化奖励（全局）			
平滑度	r_{delta}	$-\zeta_1$kΔvk $-\zeta_2\mathbb{I}$„kΔvk $< \delta_{\min}$”	惩罚不规则或微不足道的指令变化
存活	r_{survival}	1.0 如果 $h_{\text{root}} > h_{\text{thresh}}$ 否则 0	鼓励机器人保持直立
朝向	r_{orient}	−千克$_{\text{proj}}$ −克$_{\text{vertical}}$k^2	惩罚基座倾斜（俯仰/横滚偏差）
能量	r_{energy}	$-\sum$ kτk^2	惩罚过大的关节力矩消耗

所有奖励分量的完整定义，包括用于确保训练稳定性的标准存活项与正则化项，详见表1。

3.5. 视觉感知

实时球检测采用YOLOv8目标检测算法实现，该算法运行于机器人的机载GPU上。检测流程以30赫兹的频率处理RGB图像，并输出所检测球的边界框。这些二维边界框坐标随后结合深度信息与相机标定参数，转换为三维位置。

球在机器人坐标系中的位置采用如下变换进行计算：

$$\mathbf{p}_{\text{ball}} = \mathbf{R}_{\text{cam}}^{\text{robot}} \cdot \mathbf{K}^{-1} \cdot [x \cdot d,\ y \cdot d,\ d]^T + \mathbf{t}_{\text{cam}}^{\text{robot}}$$

在该方程中，K表示相机的内参矩阵，“x, y, d”为检测到的二维球坐标，其中 d 表示深度（相机到球的距离）。$\mathbf{R}_{\text{cam}}^{\text{robot}}$和$\mathbf{t}_{\text{cam}}^{\text{robot}}$分别表示旋转矩阵和平移向量，用于将球的位置从相机坐标系变换到机器人坐标系。

4. 实验

4.1. 实验设置

HierKick 框架在三个不同的环境中进行评估，以考察其有效性和迁移能力：Isaac Gym、MuJoCo，以及在 Booster T1 人形机器人上的真实世界部署。Isaac Gym 用于高保真物理仿真，MuJoCo 用于快速原型开发，而真实世界部署则在 Booster T1 机器人上进行。如图所示，所有实验环境均配备标准化的足球场、球门和标准尺寸的足球。3，

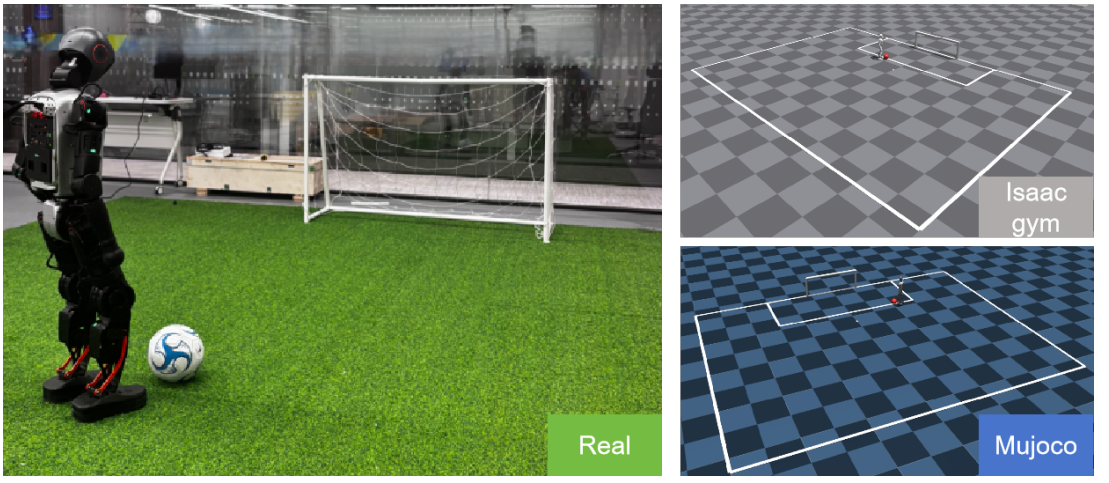


图*3.* 仿真与真实环境。足球机器人的踢球任务环境包括仿真和真实场景，用于算法开发、验证和部署。

4.1.1. 环境概述

IsaacGym 是主要的仿真平台，可实现 GPU 并行仿真，从而支持数据高效的近端策略优化（PPO）训练，并通过广泛的领域随机化提高泛化能力。*MuJoCo* 作为独立的物理引擎，用于评估模型应对物理建模差异的鲁棒性。真实世界平台 使用在人工草坪上的 Booster T1 人形机器人，并配备一个小型球门和一个标准足球（见图 3）。

4.1.2. 感知与控制架构

在感知方面，真实机器人上部署了 YOLOv8，以约 10 Hz 的频率运行，用于检测球和球门。相机经过内参和外参标定，检测到的二维边界框结合深度信息和相机标定参数转换为三维坐标（见第 3.5节）。控制架构由两个主要模块组成：教练策略和运动策略。这些模块与预训练的踢球技能集成，以 50 Hz 的频率运行，用于跟踪速度指令并生成 12 自由度的关节目标。从感知到动作的端到端延迟在 50 Hz 下约为 20 ms。

4.1.3. 初始化与随机化

每次试验开始时，球和机器人的位姿以及机器人朝向都会随机化。在仿真中，还会对地面摩擦力、球的质量、半径和弹性、相机外参、传感器噪声以及指令延迟抖动进行额外随机化，以增强模型的鲁棒性和泛化能力。实验中使用的 PPO 算法的具体超参数和网络架构列于表 2。

4.2. 评估与指标

每次试验都以随机生成的开球配置开始，并且必须在 20 秒内完成。满足以下两个条件，试验才算成功：第一，球必须在

表 *2.* 超参数与配置详情..

类别	参数	值
PPO 配置	折扣因子 (γ)	0.99
	广义优势估计参数 (λ)	0.95
	裁剪范围 (ϵ)	0.2
	KL目标	0.01
	学习率	3×10^{-4}
	批大小	4096
	小批量大小	1024
	优化轮次	5
环境与控制	高层频率	5 Hz
	低层频率	50 Hz
	最大回合长度	20 s
	延迟（仿真）	~20 毫秒
网络	策略多层感知机	»512, 256, 128...
	评论家多层感知机	»512, 256,128...
	激活函数	ELU

20 秒时限结束前完全越过球门线。第二，机器人必须在整个试验过程中保持直立。使用两个主要指标来量化性能。

4.2.1. 任务成功率

在 Isaac Gym、MuJoCo 和真实机器人上计算任务成功率（见图 4），该指标表示各环境中成功试验所占的比例。为验证 HierKick 中关键观测信息的必要性及输出设计的合理性，我们进行了两项消融实验。第一项是 HierKick Without $\mathbf{d}_{\text{ball}}$、$\mathbf{d}^{\text{goal}}$，该方法移除了策略中的观测信息 $\mathbf{d}_{\text{ball}}$和 $\mathbf{d}^{\text{goal}}$，迫使策略在缺少这些关键位置线索的情况下做出决策。第二项是 HierKick ($\mathbf{c}^{\text{prev}} \rightarrow \mathbf{v}_{\text{robot-ball}}$)，该方法将策略原有的加速度相关命令输出 $\mathbf{c}^{\text{prev}}$ 替换为 $\mathbf{v}_{\text{robot-ball}}$，以检验输出类型对任务表现的影响。这些消融变体与原始 HierKick 和端到端方法的成功率汇总于表 3。

表*3.* 组件变更及 ***HierKick*** 方法（相较于端到端方法）对模型成功率的影响。

对比方法	成功率
HierKick	**95.2%**
HierKick（不含 $\mathbf{d}_{\text{ball}}$、$\mathbf{d}^{\text{goal}}$）	23.2%
HierKick（$\mathbf{c}^{\text{prev}}$! $\mathbf{v}_{\text{robot-ball}}$）	8.2%
端到端	25.6%

4.2.2. 踢球距离

踢球距离定义为球的最终位置与球门线之间的距离，并且仅记录成功试验的数据。为确保统计可靠性，每个仿真配置均

通过两万次随机试验进行评估。对于真实机器人，则在不同光照条件和不同初始化设置下，分多个场次进行测试。

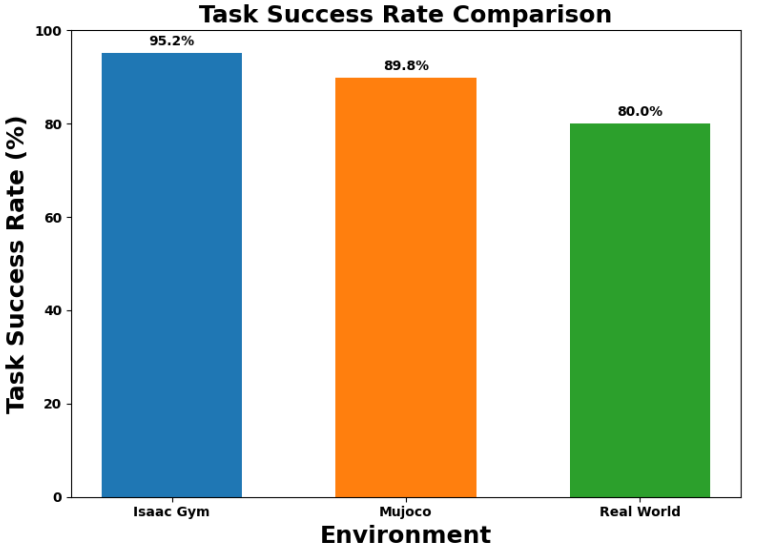


图 ***4.*** 成功率比较。此图展示了仿真环境与真实世界部署中的任务成功率比较。

4.3. 跨平台结果

图 4 总结了 HierKick 的跨平台性能。在 Isaac Gym 中，成功率为 95.2%，这得益于高保真仿真以及极小的物理建模误差。在 MuJoCo 中，成功率为 89.8%，较 Isaac Gym 下降了 5.4 %。这一下降主要源于接触和摩擦建模的差异，以及特定引擎的集成问题；这些因素会轻微影响带球和最后一步对齐的稳定性。在真实机器人上，成功率为 80.0%，较 MuJoCo 进一步下降了 9.8%。加入 $\mathbf{c}^{\text{prev}}$后，教练策略能够过滤偶发的误检测，而不会扰乱系统的指令。真实机器人性能下降可归因于真实世界不确定性，这些不确定性造成了仿真到现实差距。这些差距源于局部草皮摩擦、球反弹差异以及机器人与环境的相互作用等因素。领域随机化缓解了其中许多问题，但极端草皮区域仍会影响带球表现。导致性能下降的常见失效模式包括：部分球遮挡导致对齐延迟、低摩擦区域导致最后一步时脚部打滑，以及眩光引入深度偏差。尽管面临这些挑战，80.0% 的成功率仍展现了分层策略出色的仿真到现实迁移能力。

4.4. 踢球距离分析

我们分析了四种配置下的踢球距离（球的落地点与球门中心之间的水平距离）分布：完整的 HierKick、去除 $\mathbf{d}_{\text{ball}}$、$\mathbf{d}^{\text{goal}}$ 的 HierKick、端到端方法，以及 HierKick（$\mathbf{c}^{\text{prev}}$！ $\mathbf{v}_{\text{robot-ball}}$）。图 5 展示了该分布，其趋势与表 3中的消融结果一致。完整的 HierKick 展现出最集中的分布——方差最小

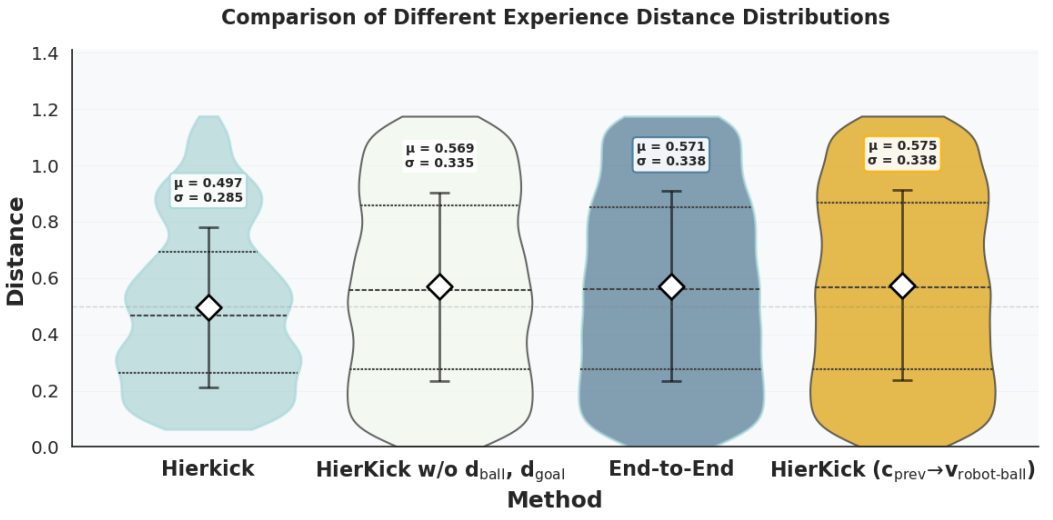


图*5.* 到球门中心的水平距离。三种条件下的踢球距离分布。*HierKick* 的均值最小，分布也最集中；而 *Remove-*和*Hier-Kick (*$\mathbf{c}^{prev} \rightarrow \mathbf{v}_{robot\text{-}ball}$*)* 实验的均值较大。白色菱形标记表示均值；误差线表示一个标准差。位于 *0.5* 米处的虚线表示参考阈值。每种条件下均有 *5280* 个样本。

且平均踢球距离最低，表明其踢球稳定且精准，这得益于其全面的观测组件，从而能够实现准确的对齐和速度控制。相比之下，端到端方法、HierKick（$\mathbf{c}^{prev}$！ $\mathbf{v}_{robot\text{-}ball}$）以及去除 $\mathbf{d}_{ball}$、$\mathbf{d}^{goal}$ 的 HierKick 均表现出更大的平均距离和更宽的分布。这证实了端到端方法和 HierKick 变体（省略关键感知组件或改变命令历史）会导致系统性失准或最后击球速度不足，这与其在表 3中的低成功率一致。

4.5. 不同方法的奖励比较

图 6 比较了训练过程中不同方法的奖励变化。该图比较了四种方法在训练过程中的奖励：HierKick、HierKick（不含 $\mathbf{d}_{ball}$，$\mathbf{d}^{goal}$、HierKick（$\mathbf{c}^{prev} \rightarrow \mathbf{v}_{robot\text{-}ball}$）以及端到端方法。总体而言，HierKick 获得的奖励最高且最稳定，而其他方法的奖励相对较低，这验证了 HierKick 方法的优越性。

5. 结论

我们提出了 HierKick，这是一种面向视觉引导的人形足球机器人的双频层次强化学习（RL）框架。该框架采用 5 Hz 教练策略来发出速度增量，同时由 50 Hz 预训练控制器利用 YOLOv8 进行感知并执行这些命令，从而形成闭环。这种频率分离设计在长时域战术规划与接近、对齐、带球和射门阶段的精准快速关节控制之间取得了平衡。多阶段奖励和领域随机化确保了训练稳定性和仿真到现实迁移，端到端延迟约为 20 ms。消融研究表明，相对球的位置几何关系和

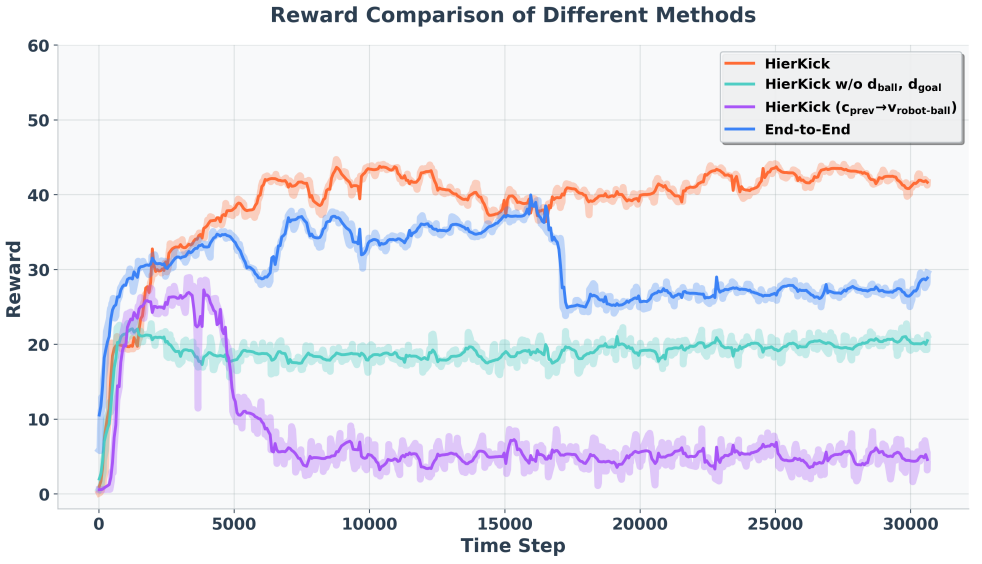


图 ***6.*** 奖励　　比较。端到端与 *HierKick* 的奖励变化比较。　　.

命令历史是取得成功的关键组成部分。复用预训练技能可缩短训练时间并改善收敛。分层方法能带来更精准、方差更低的踢球距离。不过，该方法仍存在局限，包括遮挡、草皮差异、最后接触不确定性、固定的阶段阈值/决策频率，以及感知保真度有限等挑战。

参考文献

[1] Marvin Zhang、Xinyang Geng、Jonathan Bruce、Ken Caluwaerts、Massimo Vespignani、Vytas SunSpiral、Pieter Abbeel 和 Sergey Levine。用于张拉整体机器人运动的深度强化学习。载于*2017*年 *IEEE* 国际机器人与自动化会议（*ICRA*），第 634–641 页。IEEE，2017年。[2] Seungmoon Song、Łukasz Kidziński、Xue Bin Peng、Carmichael Ong、Jennifer Hicks、Sergey Levine、Christopher G Atkeson 和 Scott L Delp。通过神经力学仿真中的深度强化学习对人体运动控制建模。《神经工程与康复杂志》，18(1):126，2021年。[3] Hanik Liskustyawati、Slamet Riyadi、Sri Santoso Sabarini、Agus Mukholid 和 Fatoni Yanuar Akhmad Budi Sunaryo。敏捷性训练对女子足球初学者带球技术技能敏捷性的影响：基于体质指数的考察。*Retos*：体育教育、运动与休闲新趋势，(52):509–520，2024年。[4] Jianlan Luo、Charles Xu、Jeffrey Wu 和 Sergey Levine。通过人在回路强化学习实现精准灵巧的机器人操作。《科学·机器人学》，10(105):eads5033，2025年。[5] Toru Lin、Kartik Sachdev、Linxi Fan、Jitendra Malik 和 Yuke Zhu。面向人形机器人的基于视觉的灵巧操作的仿真到现实强化学习。*arXiv* 预印本 *arXiv:2502.20396*，2025年。[6] Zhaoyuan Gu、Junheng Li、Wenlan Shen、Wenhao Yu、Zhaoming Xie、Stephen McCrory、Xianyi Cheng、Abdulaziz Shamsah、Robert Griffin、C Karen Liu 等。人形机器人运动与

操作：控制、规划与学习的当前进展及挑战。*arXiv* 预印本 *arXiv:2501.02116*，2025年。[7] Marco Aguiar、Bruno Gonçalves、Goreti Botelho、Ricardo Duarte 和 Jaime Sampaio。人际定位的规律性可区分小场地足球赛中的短序列与长序列比赛。《足球科学与医学》，1(3):258–264，2017年。[8] Luigi Giordano、Ario Federici、Manuela Valentini 和 Francesca D'Elia。足球中的带球：对学习理论的探讨。2019年。[9] Rati Devidze、Parameswaran Kamalaruban 和 Adish Singla。面向稀疏奖励强化学习的探索引导奖励塑造。神经信息处理系统进展，35:5829–5842，2022年。[10] Zijing Hu、Fengda Zhang、Long Chen、Kun Kuang、Jiahui Li、Kaifeng Gao、Jun Xiao、Xin Wang 和 Wenwu Zhu。迈向更好的对齐：针对稀疏奖励，以强化学习训练扩散模型。载于计算机视觉与模式识别会议论文集，第 23604–23614 页，2025年。[11] Ignacio Abadía、Francisco Naveros、Eduardo Ros、Richard R Carrillo 和 Niceto R Luque。基于小脑的方案：解决机器人控制中的非确定性时延问题。《科学·机器人学》，6(58):eabf2756，2021年。[12] Muleilan Pei、Hao An、Bo Liu 和 Changhong Wang。一种面向未知动态环境中移动机器人路径规划的改进 Dyna-Q 算法。*IEEE*《系统、人类与控制论汇刊：系统》，52(7):4415–4425，2021年。[13] Laura Smith、J Chase Kew、Tianyu Li、Linda Luu、Xue Bin Peng、Sehoon Ha、Jie Tan 和 Sergey Levine。通过迁移经验学习并适应敏捷运动技能。*arXiv* 预印本 *arXiv:2304.09834*，2023年。[14] Andrew Hundt、Julia Schuller 和 Severin Kacianka。迈向公平的人工智能与机器人敏捷研发。*arXiv* 预印本 *arXiv:2402.08242*，2024年。[15] Chuong Nguyen、Lingfan Bao 和 Quan Nguyen。通过迭代学习控制，从简单练习中掌握敏捷跳跃技能。载于*2025*年 *IEEE* 国际机器人与自动化会议（*ICRA*），第 10237–10244 页。IEEE，2025年。[16] Ryota Morimoto、Satoshi Nishikawa、Ryuma Niiyama 和 Yasuo Kuniyoshi。采用集成模型的软体连续机械臂无模型强化学习。载于*2021*年 *IEEE* 第四届软体机器人国际会议（*RoboSoft*），第 141–148 页。IEEE，2021年。[17] Soo-Hwan Chae、Sang-Min Baek、Jongeun Lee 和 Kyu-Jin Cho。通过运动模式快速转换和增强跳跃高度调节，实现敏捷且节能的跳跃—爬行机器人。*IEEE/ASME*《机电一体化汇刊》，27(6):5890–5901，2022年。[18] Dewen Liu、Changfei Li、Jieqiong Zhang 和 Weidong Huang。机器人服务故障与恢复：文献综述及未来方向。《国际先进机器人系统杂志》，20(4):17298806231191606，2023年。[19] Han Hu、Kaicheng Zhang、Aaron Hao Tan、Michael Ruan、Christopher Agia 和 Goldie Nejat。面向杂乱崎岖地形中自主机器人导航的深度强化学习仿真到现实流程。*IEEE*《机器人与自动化快报》，6(4):6569–6576，2021年。[20] Joonho Lee、Marko Bjelonic、Alexander Reske、Lorenz Wellhausen、Takahiro Miki 和 Marco Hutter。学习轮腿机器人的稳健自主导航与运动。《科学·机器人学》，9(89):eadi9641，2024年。[21] Michael Beukman、Branden Ingram、Geraud Nangue Tasse、Benjamin Rosman 和 Pravesh Ranchod。Robocupgym：RoboCup 中一项具有挑战性的连续控制基准。*arXiv* 预印本 *arXiv:2407.14516*，2024年。[22] Seung-Joon Yi、Stephen McGill、Dennis Hong 和 Daniel Lee。人形足球机器人团队的分层运动控制。《国际先进机器人系统杂志》，13(1):32，2016年

.

[23] Miguel Abreu、Luís Paulo Reis 和 Nuno Lau。通过强化学习探索技能集基元，设计一支高水平的 RoboCup 足球队。《神经计算与应用》，第 1–36 页，2025年。[24]Isaac Jesus Da Silva、Danilo Hernani Perico、Thiago Pedro Donadon Homem 和 Reinaldo Augusto da Costa Bianchi。用于机器人足球运动员的人形机器人深度强化学习。《智能与机器人系统杂志》，102(3):69，2021年。[25] Dhruva Tirumala、Markus Wulfmeier、Ben Moran、Sandy Huang、Jan Humplik、Guy Lever、Tuomas Haarnoja、Leonard Hasenclever、Arunkumar Byravan、Nathan Batchelor 等。利用自我中心视觉和深度强化学习学习机器人足球。*arXiv* 预印本 *arXiv:2405.02425*，2024年。[26] Yingkai Wang、Qingshan Wang、Qi Wang 和 Zhiwen Zheng。用于能量收集无线通信的基于分层强化学习的访问控制。《*IEEE*自动化科学与工程汇刊》，21(1):1000–1011，2023年。[27] Yao Dong、Yong He、Zhilong Su、JiangPeng Ni、Wei Feng 和 Xinyu Wu。一种基于分层强化学习的自主外骨骼运动规划方法。载于*2021*年 *IEEE*实时计算与机器人国际会议（*RCAR*），第 1076–1081 页。IEEE，2021年。[28] VE Bolshakov 和 Aleksandr Nikolaevich Alfimtsev。马尔可夫决策过程中的多智能体协作强化学习分层方法。载于《数学报告》，第 108 卷，第 S382–S392 页。Springer，2023年。[29] Seohong Park 和 Sergey Levine。用于无监督基于模型的强化学习的可预测MDP抽象。载于国际机器学习会议，第 27246–27268 页。PMLR，2023年。[30] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017年。[31] Victor Santibañez 和 Rafael Kelly。机器人操纵器的带前馈补偿PD控制：分析与实验。*Robotica*，19(1):11–19，2001年。[32] Rémy Portelas、Cédric Colas、Lilian Weng、Katja Hofmann 和 Pierre-Yves Oudeyer。深度强化学习的自动课程学习：简要综述。*arXiv* 预印本 *arXiv:2003.04664*，2020年。[33] Ren Liu、Nitish Sontakke 和 Sehoon Ha。PM-FSM：用于实现稳健四足运动的有限状态机调制策略。载于*2022*年 *IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第 4063–4069 页。IEEE，2022年。