arXiv:2611.33011v1 [cs.CV] 24 Nov 2026

MESSI：基于物理的角色控制的多专家足球技能整合

彭小刚[1*]，**邹佳琪**[2*]，**钟磊**[3]，**SiruiXu**[4]，**蒋怀祖**[1]1东北大学 2杜克大学 3爱丁堡大学 4伊利诺伊大学厄巴纳-香槟分校

https://github.com/neu-3dvi/MESSI/ https://neu-3dvi.github.io/MESSI/

摘要

基于物理的足球角色必须执行多样的运动技能，并在技能之间实现平滑过渡，例如从带球衔接到射门。然而，单独执行时表现良好的技能，在连续执行时可能会失败，因为前置行为可能使角色和球处于不熟悉的状态。我们提出**MESSI**（**M**ulti-**E**xpert**S**occer**S**kill **I**ntegration，多专家足球技能集成）框架，通过共享适应实现多种足球技能过渡，同时在过渡之外保留原始技能策略。MESSI 首先利用运动先验和任务奖励训练八个具备任务能力的专家，然后将其蒸馏为统一的技能条件化 Transformer。我们没有针对不同转换重新训练单个专家，而是通过在线强化学习，在冻结学生策略上训练一个共享的上下文条件化低秩适配器。该适配器在技能切换时激活，帮助目标技能处理不熟悉的进入状态；同时，学习式退出机制决定何时将控制权交还给原始学生策略。通过在原生技能执行期间禁用适配器，MESSI 从机制上保留了原始学生策略。在单项技能、技能转换和连续技能链上的实验表明，相较于直接切换，该方法取得了显著提升，其中带球到射门的成功率从 19.8% 提高至 61.8%。

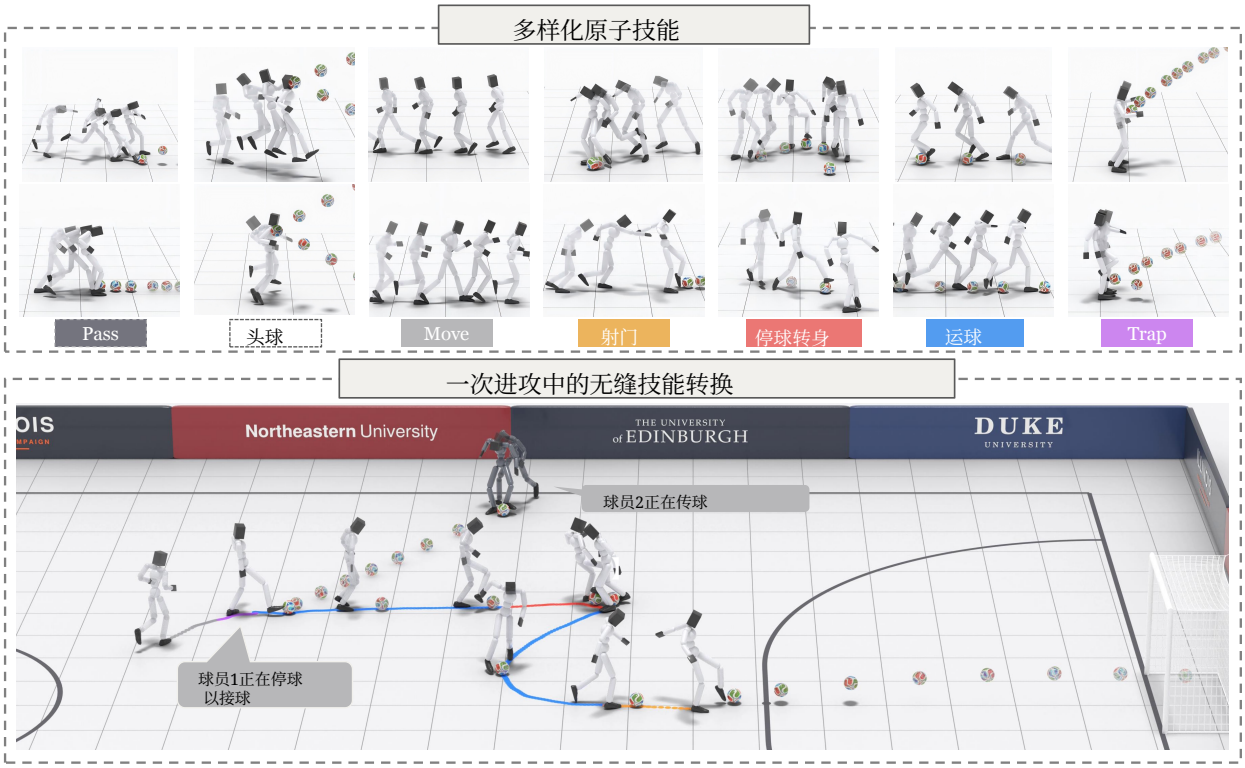


图 1 | MESSI：基于物理的可组合足球控制。 我们的统一控制器整合了八种足球技能，并能在它们之间实现平滑转换。**上：** 具有代表性的原子技能。**下：** 一段连贯的比赛过程，结合了传球、停球、带球和射门。

* 表示贡献相同

1. 引言

足球为基于物理的角色控制提供了一个具有挑战性的场景，将全身协调与同动态移动的球进行精确交互结合起来 [Haarnoja 等人，2024年，Hong 等人，2019年，Kim 等人，2025年，Liu 等人，2022年，Ye 等人， 2026]。一个多才多艺的足球角色必须执行多种技能，包括运动、带球、射门、传球和停球，同时还要快速在这些技能之间进行转换。例如，球员可能在带球后立即射门，或在跑动中接到传球。这类转换之所以困难，是因为角色的姿态、动量和球状态取决于前置行为，可能与目标技能训练期间遇到的状态存在显著差异。因此，即使经过充分训练的单项技能，在连续执行时也可能失败，如图 2所示。

近年来，基于物理的角色控制取得了进展，通过动作跟踪和大规模动作预训练实现了可复用的运动技能 [Juravsky 等人，2024年，Peng 等人，2018年、2022年，Shi 等，2026，Tessler 等人， 2024]。然而，仅跟踪参考动作并不能保证实现任务特定的控球，也无法确保在动作数据中缺失的可靠转换。尽管独立训练的、具备任务能力的专家能够优化各自的技能目标，但当目标技能遇到陌生的身体与球的状态时，组合它们的行为仍然具有挑战性。现有方法以多种方式处理技能转换。一些方法联合学习技能与转换，例如通过统一的多技能策略 [Ye 等人， 2026] ，或通过增强的交互轨迹 [Yu 等人， 2025]。引入新的转换时，这类方法可能需要重新进行联合策略训练或准备转换数据，从而可能影响此前学习到的行为。另一些方法则利用过渡感知初始化，或使用前序技能生成的状态进行微调 [Avinash 等人，2026年，Kim 等人， 2025], , 将转换适应与单个技能策略相结合。对于更具挑战性的转换，策略组合方法还会在必要时引入中间行为 [Xu 等人，2025a],, 这可能增加针对特定转换的训练与维护开销。尽管这些方



图 2 | 基于物理的足球控制中技能过渡的挑战。（a）在原生初始化下，射门成功。（b）从带球直接切换到射门时，由于身体和球状态不利而失败。（c）借助我们的适配器，在切换时处于激活状态，相同的交接最终实现了成功射门。叠加的姿态展示了运动序列，虚线箭头表示目标射门方向。

方法展现了有效的技能组合能力，但在无需反复修改单个专家或引入独立转换模块的情况下，适应不断扩展的技能库仍然颇具挑战。这促使我们采用一种共享适应策略，在限制对既有技能行为干扰的同时，支持多样化的技能转换。

为此，我们提出了 **MESSI** （**M**多**E**专家**S**足球**S**技能**I**集成），这是一个整合具备任务能力的足球技能，并通过统一的基于物理的控制器实现技能间转换的框架。MESSI 包含八种技能：无球移动、带球、射门、传球、头球、停球、

急停转身和起身。我们首先利用运动先验、任务奖励和定制的课程学习训练技能专属专家。随后，通过统一观测空间，将这些专家的行为蒸馏到一个共享的技能条件学生模型中。将专家整合为单一的技能条件策略，为过渡适应提供了共同的基础。这使不同的技能对能够共享一个适应模块，而不必为各个专家策略分别进行适应。

我们的核心洞见是，只有在技能切换后才需要进行过渡适应，因为此时目标技能会遇到由其前序技能生成的陌生状态。我们不在所有情况下修改目标技能，而是在冻结学生策略上引入一个共享的、带门控的上下文条件化低秩适配器。技能切换时门控开启，学习式退出机制则决定何时关闭门控，使冻结学生策略恢复控制。在转换之外，适配器绝不发挥作用。我们通过在线强化学习，利用前序技能生成的状态和目标技能奖励来训练适配器。根据技能对和目标指令对适配器进行条件化，使单个适应模块能够服务于多个转换。新的转换技能对可以纳入共享适应过程，而无需修改原有专家或学生策略骨干网络，也就不必维护单独的、针对特定转换的模块。这一设计提升了转换鲁棒性，同时完全保持原生技能不变。

我们的主要贡献总结如下：

- 我们开发了一种基于物理的足球控制框架，将八个经过独立训练且具备任务能力的专家整合为统一的技能条件化策略，为过渡适应提供共享骨干网络。
- 我们提出一种使用上下文条件化低秩适配器的在线共享过渡适应方法，使多个技能对能够共同适应，而无需重新训练各个专家。
- 我们引入一种学习式退出机制，动态确定过渡适应的持续时间，使控制器能够在不再需要适应时返回冻结学生策略，同时在原生技能执行期间保留原始策略。

2. 相关工作

基于物理的交互技能。基于物理的角色控制已从运动跟踪 [Peng 等人， 2018] 和对抗性运动先验 [Peng 等人， 2021, 2022]发展到可扩展的统一控制器 [Shi 等，2026年，Tessler 等人， 2024]。运动型球类交互还带来了额外挑战：需要协调全身运动与动态物体，并精确把握接触时机。早期研究通过轨迹优化和强化学习 [Liu 和 Hodgins， 2018], 学习篮球带球，并通过耦合的角色—球控制 [Hong 等人， 2019]合成足球带球和射门。近期方法展示了篮球交互技能 [Wang 等人， 2025], 敏捷的人形机器人足球 [Haarnoja 等人， 2024], 精准射门 [Zhong 等人， 2026],以及基于视觉的足球控制 [Jin 等人， 2026]。这些进展推动了具备任务能力的技能学习，使其优化物理交互目标，而不局限于复现参考动作。

技能组合与转换。 以往工作通过多种策略处理技能组合问题。一些方法联合学习技能与转换。 [Xie 等人， 2022]采用分层专家混合模型实现足球颠球和技能切换，而 SkillMimic-V2 [Yu 等人， 2025] 构建增强交互轨迹和状态转换场。SkillX 在统一策略中学习多种足球技能 [Ye 等人， 2026] , 而 SoccerBot 将潜在动作先验与转换状态初始化相结合 [Shao 等人， 2026]。其他方法则使目标技能适应由其前序技能生成的状态，包括 PhysicsFC [Kim 等人， 2025] 以及 Joga [Avinash 等人， 2026]。Learning to Ball 进一步将策略适应与中间策略相结合，以应对具有挑战性的转换 [Xu 等人，2025年a]。尽管这些方法效果良好，但将其扩展到更多技能和转换可能需要更新

共享策略、增强转移数据，或维护专门的适应流程。相较之下，我们研究的是在独立训练的技能整合到单个控制器之后进行的过渡适应。

专家蒸馏与策略适应。策略蒸馏 [Rusu 等人， 2016] 和 DAgger [Ross 等人， 2011] 为整合专家行为奠定了基础。SuperPADL 逐步将专门化的强化学习专家蒸馏到可扩展的控制器中 [J uravsky 等人， 2024], ；InterMimic 则将师生学习与强化学习精炼相结合，以处理多样的人物与物体交互 [Xu 等人，2025b]。MaskedManipulator 将运动跟踪控制器蒸馏为目标条件策略 [Tessler 等人，2025], ，而 TokenHSI 则使用任务标记化实现统一的交互控制 [Pan 等人， 2025]。训练后的适应也通过策略精炼 [Xu 等人，2023b], 、条件低秩更新 [Shi 等人， 2026], 、物体感知残差学习 [Zhao 等人， 2025], ，以及对蒸馏交互先验进行强化学习微调 [Xu 等人， 2026]等方式得到探索。相比之下，MESSI 侧重于通过单个上下文条件适配器，在冻结的学生策略上适应多种技能转换，并利用学习式退出将适应限制在技能边界处，同时保留原生策略的行为。

3. 方法

如图 3所示，MESSI 包含三个阶段：技能专属专家学习、多技能蒸馏和统一转移适配。测试时，控制器接收当前角色和球的状态、技能标签以及任务指令。用户或高层规划器会在预定的技能切换时指定下一个技能及其指令，无需参考动作。

3.1. 技能专属专家学习

如图 3 (i)所示，我们首先针对每项技能 k 训练一个专家策略 $\hat{}\pi^k(\hat{}\mathbf{a}_t \mid \hat{}\mathbf{s}_t, \mathbf{g}_t^k)$。状态 $\hat{}\mathbf{s}_t$包含角色的本体感觉以及相对于角色的球状态，对所有技能均相同。任务指令 $\mathbf{g}_t^k$ 则针对技能 k 而定。动作是供 PD控制器使用的目标关节角度。每个专家都在各自的环境中训练，使用各自的动作数据、初始状态分布和终止规则。这些设置针对每项技能分别确定，不受后续集成阶段的约束。

奖励与优化。 每个专家都会最大化多个目标 j：由技能定义的任务奖励 r_{task} ，以及来自相应对抗运动先验的风格奖励$\mathcal{M}^k$ [Peng 等人，2021]，

$$r_t^{\text{style}} = -\log\left(1 - D(\mathbf{o}_{t-H+1:t})\right), \tag{1}$$

其中 D 是一个针对 H 角色状态窗口的判别器。对于踢球和头球技能， D 还会以技术标签为条件，并且这些技能还会加入一个较小的腿部姿势目标，使踢击腿保持接近参考动作。将各项目标加总为一个奖励，可以使其有效权重在训练期间随奖励尺度变化。受 [Xu 等人， 2023a], 的启发，我们不将任务奖励和风格奖励合并为单一标量目标，而是分别优化归一化优势，以在训练期间保持它们各自的相对贡献。

射门。 射门指令指定球的目标初速度、击球脚和技术（脚背或脚内侧射门）。球被击中时记录发射速度。从那时起直到回合结束，任务奖励为 $\exp(-k\|\mathbf{v}_{\text{launch}} - \mathbf{v}^*\|^2)$；如果指定脚完成了击球，还会获得额外奖励。因此，策略能够获得关于单次短暂接触质量的稠密信号。击球前，接近项会将击球脚引向球。击球后，恢复项会奖励平衡站姿。指令从技术特定的速度、仰角和方向范围内采样，而球的起始状态要么静止，要么处于滚动状态。

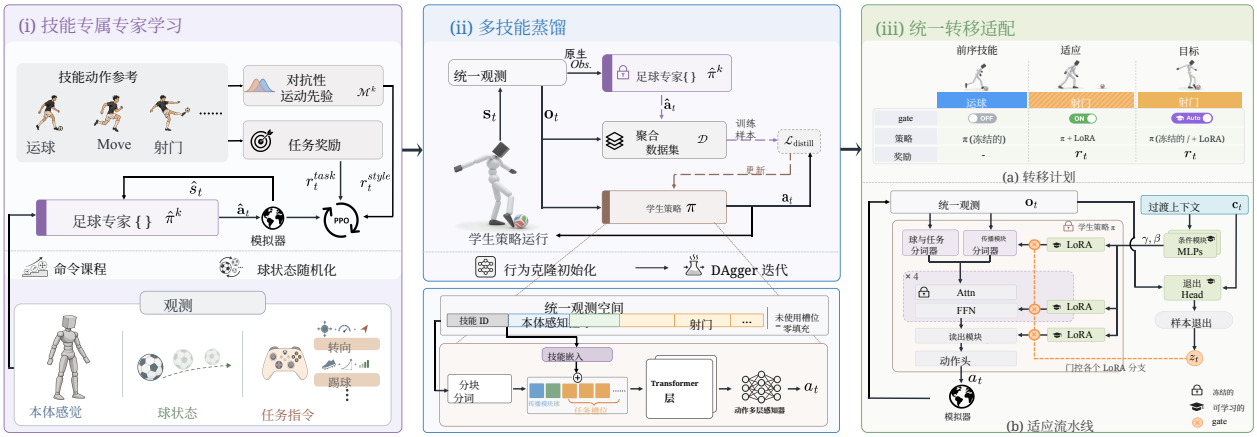


图 3 | MESSI 概览。 (i) 技能专属专家学习： 八种技能中的每一种都通过运动先验和各自的任务奖励进行训练。(ii) 多技能蒸馏： 通过行为克隆再结合 DAgger，将专家策略蒸馏为一个在统一观测空间中运行的技能条件学生模型。(iii) 统一转移适配： 冻结学生策略上的共享条件 LoRA 适配器仅在技能切换后发挥作用，并基于前序技能生成的状态进行在线训练。

带球。 带球指令指定目标方向和速度。任务奖励将速度跟踪与前导项相结合，使球沿目标方向保持在角色前方的固定距离处。对球的横向偏差的惩罚强于对纵向偏差的惩罚。速度跟踪奖励会乘以一个作用于角色–球距离的门控值。当球处于可触及范围内时，该门控值保持不变；超出该范围后则逐渐衰减，因此策略无法通过跑到球前方而将球甩在身后来获得任何收益。

其他技能。 其余技能遵循相同的模板，采用相应的指令和任务目标；八种技能的奖励、指令分布、初始化和终止条件等详细信息见附录。

3.2. 多技能蒸馏

这些专家的指令、网络输入和环境各不相同。我们将它们蒸馏为一个统一的学生策略 π，使其能够在一个观测空间中执行所有技能，并通过标签选择技能。

统一观测空间。学生策略观测到

$$\mathbf{o}_t = \left[\mathbf{s}_t,\ \mathbf{b}_t^1, \ldots, \mathbf{b}_t^M,\ \mathbf{e}_{k_t}\right], \tag{2}$$

其中 $\mathbf{s}_t$ 是第 3.1 节中的共享状态，$\mathbf{b}_t^1$、……、$\mathbf{b}_t^M$是任务块，$\mathbf{e}_{k_t}$ 是当前技能的独热标签 k_t。每个技能的指令都会写入一个任务块。指令类型相同的技能共享一个区块：例如，带球和无球移动都使用转向块。其他技能对应的任务块则设为零。通过固定的选择映射 P_k 可以恢复专家 k 的原生观测 P_k $\mathbf{o}_t = (\mathbf{s}_t,\ \mathbf{g}_t^k)$，因此当前技能的专家能够为学生策略访问的每个状态提供标签。

学生架构。学生模型是一个 Transformer [Vaswani 等人提出的模型， 2017] 作用于观测块。一个本体感知分词器、一个球分词器，以及每个任务块对应的一个任务分词器，分别将其所在区块映射为一个令牌，并将技能标签的嵌入添加到每个令牌中。注意力采用掩码机制：本体感知令牌关注所有令牌，而其他每个令牌只能关注自身和本体感知令牌。因此，由多个技能共享的任务块（例如转向）不会直接关注只有其中部分技能使用的区块（例如球）。读出层将输出令牌拼接起来，动作层再将其映射为平均动作 $\boldsymbol{\mu}(\mathbf{o}_t)$。

蒸馏。 学生策略经过训练，以在由观测和技能标签组成的聚合数据集 D 上匹配专家的平均动作，

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{(\mathbf{o},k)\sim\mathcal{D}}\left[\left\|\boldsymbol{\mu}(\mathbf{o}) - \hat{\boldsymbol{\mu}}_k(P_k\,\mathbf{o})\right\|^2\right], \tag{3}$$

其中 $\hat{\boldsymbol{\mu}}_k$ 是专家 $\hat{\pi}^k$的平均动作。我们首先用每位专家在其各自环境中的滚动轨迹填充 $\mathcal{D}$ ，其中包含该专家训练所用的初始状态、指令和终止条件，并通过行为克隆初始化学生策略。随后，学生策略的误差会使其进入专家很少访问的状态。因此，我们采用 DAgger [Ross 等人， 2011]。在每一轮中，学生策略在每种技能对应的环境中执行其平均动作，该技能的专家为每个访问状态提供标签，将新样本添加到 $\mathcal{D}$中，然后使用 $\mathcal{D}$中的全部数据重新训练学生策略。每种技能贡献相同数量的样本，且专家保持冻结。附录介绍了架构和训练计划。

每条滚动轨迹都从该技能自身的初始状态出发，执行单一技能。因此，蒸馏过程从不训练学生策略从前序技能留下的状态开始执行某项技能。第 3.3 节将讨论这些转换。

3.3. 统一转移适配

尽管蒸馏学生保留了各项技能的能力，但直接切换技能可能会使目标策略面对陌生的身体与球状态。针对这些状态进行微调或许能改善技能转移，但也可能改变目标技能在原生初始化下的表现。我们的做法是通过单个共享模块来适配冻结学生策略，并采用两项关键设计：根据技能转移进行条件化适配，以及学习何时停用适配器。

转移计划。 如图 3(iii) 所示，学生策略在禁用适配器的情况下执行前序技能（$z_t = 0$）。技能切换时，目标技能变为激活状态，并启用适配器（$z_t = 1$）以处理陌生的进入状态。随后，学习式退出策略会决定何时停用适配器，并将控制权交还给原始学生策略。原生技能执行始终绕过适配器。

共享条件适配器。 我们不针对每个技能对分别训练策略或适配器，而是使用一个由过渡上下文 c_t 条件化的共享低秩适配器，该上下文编码前序技能和目标技能、目标指令，以及自切换以来的经过时间。遵循条件低秩适配 [Hu 等人， 2021， Shi 等人， 2026],每个经过适配的线性层计算

$$y = W_0 x + z_t B\left[(Ax)\odot(1+\gamma(c_t)) + \beta(c_t)\right], \tag{4}$$

其中 W_0 是冻结的， A 和 B 是可训练的低秩矩阵，而 $\gamma(c_t)$和 $\beta(c_t)$则根据过渡上下文调节更新。这样，同一组适配器参数便可用于多个技能对，同时 $z_t = 0$ 可还原原始学生策略。我们将其应用于本体感知分词器、Transformer 前馈层和读出模块。

学习式退出策略。 所需的适应时长会因转换和遇到的状态而异。固定时间窗口可能会过早终止适应，也可能让适配器保持激活状态的时间超过必要时长。因此，我们学习一种状态依赖退出策略，根据观测 o_t 和过渡上下文 c_t预测停用适配器的概率。退出策略与适配器通过 PPO 联合训练，在目标任务奖励与每步适配器使用成本之间进行权衡。退出后获得的奖励也计入回报，因此退出决策会考虑之后由冻结学生策略执行的情况。退出决策仅允许在 0.2 秒后进行，首次采样到退出时，门控便会关闭，直到下一次技能切换。

表格 1 | 原子技能与蒸馏保真度。 每项技能包含 1024 个配对原生案例。比率以 % 表示，速度误差以 m/s 表示。学生策略在适配器关闭的情况下运行。绿色：最佳；橙色：次佳。

方法	无球			运球		射门	Pass	头球	Trap	急停转身	起身
	成功率（%）↑	速度误差（米/秒）↓	失败（%）↓	速度误差（米/秒）↓	失败（%）↓	方位（%）↑	方位 (%) ↑	合法触球 + 方位 (%) ↑	成功 (%) ↑	停球锁存器（%）↑	恢复（%）↑
TokenHSI* [Pan 等人， 2025]	34.47	1.10	15.04	2.830	**11.43**	2.73	7.03	3.12	23.63	43.85	95.02
SkillX* [Ye 等人， 2026]	56.48	0.726	5.79	2.324	19.31	3.91	9.96	5.18	30.76	72.07	94.63
仅专家	**69.43**	**0.617**	**1.76**	**1.989**	12.60	**79.88**	**67.58**	22.46	**97.46**	**91.70**	**99.12**
仅行为克隆学生	29.98	0.919	43.16	2.281	28.03	59.77	55.08	22.36	70.21	88.38	98.73
行为克隆+DAgger 学生（我们的方法）	68.95	0.639	3.12	2.104	17.77	77.93	63.77	**22.66**	95.02	91.21	99.02

训练。 技能切换后，我们通过适配器使用成本和失败惩罚来优化目标技能奖励：

$$r_t = r_l^{\text{task}}(s_t) - 0.02z_t - p_t, \tag{5}$$

其中， p_t 发生摔倒或球丢失时取 1，其他失败时取 0.5。使用 PPO 联合优化适配器和退出策略，重要性比率根据动作和退出决策的联合对数概率计算。只有适配器处于激活状态的步骤才会对策略梯度产生贡献。评论家以 c_t 为条件。所有转换均通过在线仿真，使用共享适配器参数进行联合训练。

表 2 | 技能转换性能与原生策略保持。 基于相同前序技能生成的状态计算的转换成功率（%），以及原生技能性能。

方法	转换成功率（%）↑					原生策略	
	运球 →射门	Stop →运球	Trap →运球	起身 →运球	无球 →运球	射门 (%) ↑	运球速度 误差（米/秒）↓
SkillX* [Ye 等人， 2026]	2.7	95.8	47.9	91.3	83.9	-	-
学生策略 (直接切换)	19.8	44.1	20.7	32.6	62.7	77.93	2.104
专家 + STI	54.1	63.8	27.7	55.7	59.8	63.09	2.662
全量微调	57.3	96.9	49.1	91.0	95.0	54.82	2.152
Per-trans.	**67.8**	94.0	47.8	90.8	92.6	77.93	2.104
Ours	61.8	**97.0**	**50.4**	**91.7**	**96.3**	77.93	2.104

4. 实验

我们从三个层次评估 MESSI：来自其原生初始状态的单项技能（第 4.3 节）、由前序技能生成的状态开始的单次转换（第 4.4 节和第 4.5 节），以及技能的连续链（第 4.6 节）。GPU PhysX 推演在不同机器之间存在差异，因此每次配对比较都在同一台机器上、从完全相同的初始状态开始运行。

4.1. 实现细节

所有技能均在 Isaac Lab 中进行仿真 [Mittal 等人， 2023] 使用 GPU PhysX 以 240 Hz 运行，所有策略则以 60 Hz 运行，与参考动作的帧率一致。角色是一个由 PD控制器驱动的 28自由度人形机器人，脚部采用鞋形碰撞网格。刚体求解器既不模拟空气动力学，也不模拟滚动阻力，因此我们在球飞行时为其添加空气阻力和马格努斯力，并在球于地面滚动时施加恒定的滚动阻力。每个专家均使用 4 块 NVIDIA RTX A6000 GPU 训练 13 至 53 小时。将八个专家蒸馏为拥有 530 万参数的学生策略，使用一块 GPU 约需 4.4 小时（先进行行为克隆，再进行四轮 DAgger）；训练共享适配器则需使用四块 GPU 11.2 小时。附录列出了物理参数、动作数据、网络架构和训练超参数。

4.2. 基线方法

技能学习。 我们将技能专属专家和蒸馏学生策略与 TokenHSI* 进行比较 [Pan 等人， 2025] 以及 SkillX* [Ye 等人， 2026], 这两种改编基线均采用统一的 Transformer 策略，从头开始针对全部八种技能进行联合训练。两种基线均遵循其官方网络架构和 PPO 配置，并使用我们的观测、任务奖励和动作数据。不过，它们不采用我们的技能专属训练策略，包括课程学习、失败加权采样、多目标优势和判别器标签。为单独评估 DAgger 的作用，我们还评估了仅通过行为克隆训练的学生策略。

转换。 直接切换无需适配器即可激活目标技能。专家 + STI 遵循 PhysicsFC 的基于技能转换的初始化，从前序技能生成的状态开始，使用 PPO 对目标专家进行微调 [Kim 等人， 2025]。回合以 0.7 的概率从技能切换（交接）状态开始，否则从原生初始状态开始；微调使用与我们的适配器相同的每次转换样本预算。一个经过微调的带球专家服务于所有四个进入带球技能的转换。这是我们对已发表方法的重新实现，而非 PhysicsFC 的代码。为研究退出机制，我们还运行了禁用退出头的适配器检查点（个性化方式），以及采用事后截断的版本，即在交接后的固定时间关闭门控。为评估参数共享的作用，我们还与逐转换适配器基线（Per-trans.）进行比较，该基线在同一个冻结学生策略上为每个技能对分别训练一个适配器。我们还纳入了全量微调（Full FT），该方法直接更新统一学生策略以进行过渡适应。

4.3. 原子技能与蒸馏

每项技能均在从其专家训练所用的初始状态和指令中抽取的 1,024 个案例上进行评估，且所有策略均运行相同的案例。无球移动在没有失败，且回合平均速度误差和朝向误差分别低于 0.75 m/s 和 35° 时视为成功。带球没有完成事件，因此我们报告其速度误差和失败率。发射方位角与指令的偏差在 10° 以内，且发射速度至少达到指定速度的一半时，射门和传球即视为成功；射门还要求使用指定脚。头球要求头部完成首次合法触球，且方位角误差低于 25°。停球使用其成功位；急停转身要求球在脚下停止的瞬间完成转身；起身则使用恢复标志。学生策略运行时关闭适配器，因为这些回合开始前没有发生转换。

专家。表1 显示，我们的技能专属专家在射门、传球和停球等接触敏感任务上显著优于联合训练的基线方法。然而，即使是专门训练的专家，头球仍然具有挑战性，这表明仅靠特定任务训练并不能解决所有困难的球交互。带球结果也揭示了一种权衡：TokenHSI* 的失败次数较少，但速度跟踪误差大得多，这说明保持控球权并不一定意味着能够准确跟随指令。

蒸馏。 单独使用行为克隆会使多项技能的性能显著下降，而 DAgger 能够恢复大部分损失的性能。这凸显了让学生策略在由其自身动作诱导的状态上接受监督的重要性，而不是仅仅依赖专家示范。剩余的差距，尤其是在头球方面，很大程度上反映了原始专家的局限性，而非蒸馏本身的问题。

射门准确度。表1 中的射门方向指标评估水平发射方向、最低发射速度以及是否使用了正确的踢球脚，但不评估球的最终落点。为了进一步评估射门准确度，我们测量球相对于指令目标穿过球门平面的位置。我们指令策略进行射门，并瞄准一个7.32×2.44米宽、距离11米的球门上的点，如图5 所示。球在使用指定脚踢出且速度不低于指定速度一半的情况下，穿过球门平面并进入球门范围内，即视为射门成功。在4,096次射门中，有3,479次瞄准

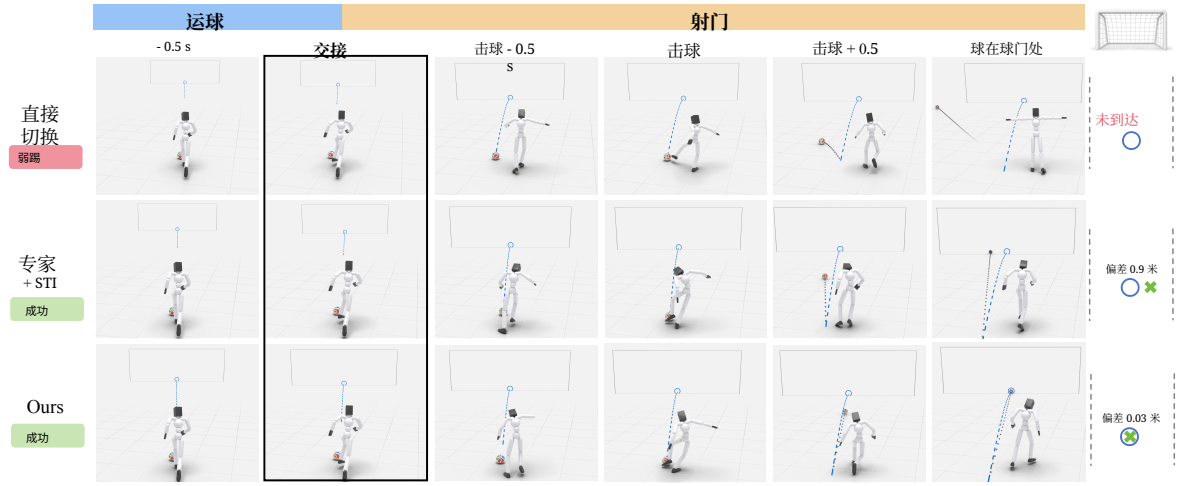


图 4 | 运球→从同一个交接点射门。 直接切换时踢球力度较弱；STI 微调后的专家距离目标 0.9 米，我们的方法距离目标 0.03 米。

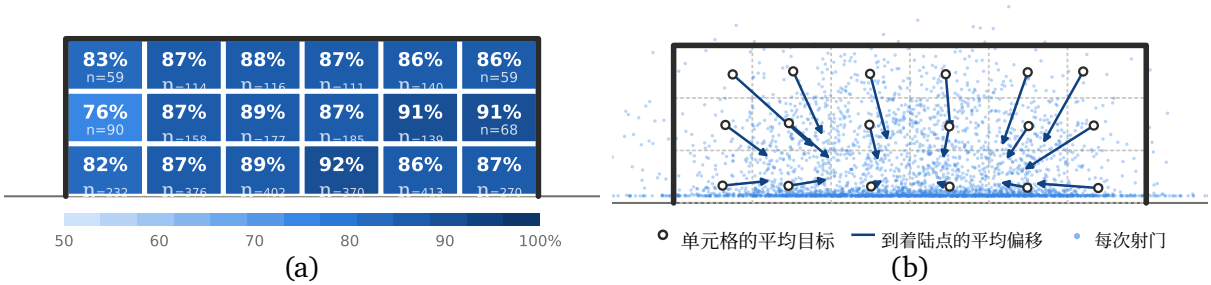


图 5 | 我们的射门策略的射门落点，目标是 11 米外的球门（点球）。（a）每个指令目标单元格的成功率（瞄准该处的 $n =$ 次射门）。（b）着陆点，以及每个单元格从目标到着陆点的平均偏移。

球门范围内，其中87.2%成功。球门各处的成功率差异很小，每个目标单元格的成功率介于76%至92%之间。剩余误差具有系统性，而非随机性。发射中位数达到指定速度的89%，因此瞄准高位目标的射门平均会低约1.0米，而瞄准门柱的射门则会向内偏约0.9米。

4.4. 技能过渡

我们评估了五种转换，这些转换来自冻结学生策略执行前序技能直至满足其交接条件后产生的交接状态，每种转换约有 400 至 510 个样本。每种方法都从完全相同的状态接管。Dribble→Shoot 在上述原生射门规则下取得成功。当角色未摔倒，并在 4 秒截止时间前稳定进入带球状态时，进入带球的转换即视为成功：最后一秒内的速度误差低于 0.75 m/s，且球位于 2.5 m 范围内。我们还在原生技能初始化下评估每个已部署策略，以考察转换适应是否会影响原有技能。对于 MESSI，在这些回合中适配器门控始终保持关闭，因此根据构造，其动作与蒸馏学生策略的动作完全相同。该评估验证了原生模式策略保持，而冻结学生策略在适配器退出后继续执行的能力，则通过转换滚动评估进行考察。

表格 3 | 退出规则。 一个适配器检查点在来自相同交接状态的全部 11 个转换中的任务成功率（%）。截止时间会在交接后经过固定时间时关闭门控。退出：学习式退出在截止时间前返回控制的交接比例（%）；开启时长：适配器处于激活状态的平均时间（秒）。技能标签包括：无球、带球、射门、传球、头球、停球、急停转身、起身。绿色／橙色：每行的最佳／次佳。

	任务成功率（%）↑							学习式退出	
	截断点（秒）								
	直接	0.25	0.5	1	2	个性化方式	我们的方法	退出	On
博 →	19.8	26.9	45.5	61.0	**62.0**	61.8	61.8	10	1.61
博 →	22.2	28.0	37.1	37.6	**38.0**	**38.0**	**38.0**	1	1.46
博 →	82.4	95.7	96.3	97.1	97.8	97.6	**99.2**	100	0.96
ST →	44.1	48.4	69.3	88.2	96.1	**97.0**	**97.0**	1	3.98
特 →	20.7	32.4	34.5	42.2	47.3	**50.4**	**50.4**	2	3.77
古 →	32.6	53.3	73.8	83.1	90.5	**91.7**	**91.7**	2	3.90
Mv → Dr	62.7	78.0	84.7	91.2	95.9	**96.3**	**96.3**	1	3.98
Sh → Mv	69.5	89.6	95.2	97.8	98.2	94.2	**98.4**	100	0.60
帕 →	70.2	89.2	93.7	95.7	**97.2**	**97.2**	96.3	100	0.66
Gu → Mv	68.3	84.8	88.2	91.1	**92.3**	90.9	91.6	99	0.70
Hd → Mv	61.1	74.0	81.1	83.1	84.8	**87.5**	84.1	99	0.69

定量结果。 表 2凸显了单项技能能力与转换鲁棒性之间的差距。尽管蒸馏学生策略在原生射门回合中的成功率达到 77.9%，但从带球直接切换会使成功率降至 19.8%，暴露出进入状态分布之间存在显著不匹配。全量微调提升了转换性能，但使原生射门成功率从 77.93% 降至 54.82%，体现了适应与保持既有行为之间的权衡。我们的门控适配器通过保持原始学生策略不变，避免了原生执行中的这一权衡。与逐转换适应相比，我们的共享适配器在五种转换中的四种上取得了最高成功率，尽管专门化方法在带球→射门转换中受益。这些结果表明，共享适应参数能够在不分别修改每个技能策略的情况下，提供具有竞争力的性能。

定性比较。 由于 TokenHSI* 和 SkillX* 即使采用原生技能初始化，表现仍然不佳，因此我们将定性比较重点放在应用于训练所得技能的三种主要过渡策略上。图 4 展示了三种方法下从带球到射门的转换。直接切换会产生一次弱踢，球未能到达球门。STI微调专家成功射门，但射偏了指令目标 0.9 米。相比之下，我们的共享适配器实现了更精准的射门，球到达的位置距目标仅 0.03 米。

4.5. 学习的退出时机

我们在不同退出规则下，使用相同的技能切换状态，对适配器在全部 11 个转换上的表现进行了评估。除学习式退出和直接切换外，持久（Pers.）变体从不退出，而事后截断会在技能切换后 0.25 至 2 秒关闭门控，无论退出头部为何。每个转换的截止时间为 4 秒，踢球则在发射后 1 秒进行判定。对于急停转身，成功标准是表 1 中的停球锁存器。最后两列分别报告学习式退出在截止时间前返回控制的频率，以及适配器平均控制角色的时长。

定量结果。 表3 显示，不同目标技能所需的适应时长差异显著。转换为无球移动时，通常只需 0.7 秒，冻结学生策略便可恢复

表格 4 | 适应成本和模型占用空间。 所有方法均使用369个MPPO样本。参数量统计可训练策略权重，包括条件化和退出模块。存储空间报告特定于适应的FP32权重；GPU小时涵盖转移训练。

方法	模块	参数量 (M)	GPU 小时	存储空间 (MiB)
专家 + STI	5 个策略	14.28	44.1	53.5
全量微调	1 个策略	5.34	42.7	20.4
Per-trans.	11个适配器	4.33	43.0	16.5
Ours	**1个适配器**	**0.39**	44.7	**1.5**

控制；而转换为带球时，则很少触发提前退出。这表明，是否需要继续适应，很大程度上取决于目标技能。

学习式退出能改善某些转换中的表现，例如射门→无球，但并不能持续提升所有技能对的成功率。其主要优势在于能够调整干预时长，而无需为每种转换手动选择截断点。

适应效率。 在相同的369M PPO交互预算下，我们的共享适配器仅使用0.39M个可训练策略参数（Param.），并仅需1.5MiB的额外存储空间，比逐转换适应少约 11× 。尽管资源占用更少，MESSI在所有评估的技能链上仍实现了具有竞争力的转换性能和更高的完成率。训练成本仍然相近（42.7–44.7 GPU小时），这表明共享的主要优势在于提高参数和存储效率，而非减少训练计算量。

表 5 | 连续技能链完成率。 所有阶段依次完成且期间没有环境重置的回合百分比（每条技能链512个回合）。阶段完成情况依据技能特定标准进行评估，而踢球指令准确率则单独评估。上方的技能标签保持一致。

技能链	学生策略 (直接切换)	专家 +STI	全量微调	Per-trans.	SkillX*	Ours
博士 → → →	41.8	51.6	97.3	35.0	62.5	**97.5**
博士 → → →	38.5	58.6	97.1	35.2	63.5	**97.7**
移动 → → → →	25.8	33.2	29.5	48.8	6.6	**51.0**
移动 → → → →	23.8	34.6	28.5	48.0	7.8	**50.6**
Gu → Dr → Sh → Mv	31.2	51.6	50.0	82.4	58.0	**82.6**

4.6. 连续技能链

我们在连续仿真中评估五种预定义的技能链，不重置角色或球。只有当所有阶段依次满足其完成标准时，技能链才算成功：停球和急停转身需要停球锁存器，起身需要恢复，带球需要持续 2 秒且至少前进 0.5 米，踢球需要踢出球后保持站立 1 秒，无球移动则需要跟踪速度和朝向。所有方法使用相同的标准，并在阶段完成后切换技能。该指标衡量的是连续执行，而非精确的踢球指令；后者将在第 4.3 节和4.4节中单独评估。每种技能链在 512个回合（两个随机种子）中进行评估。专家 + STI 使用微调后的运球和射门专家。

定量结果。表5 表明，孤立的技能转换表现出色，并不一定能转化为可靠的技能链。逐转换适配器在这两种

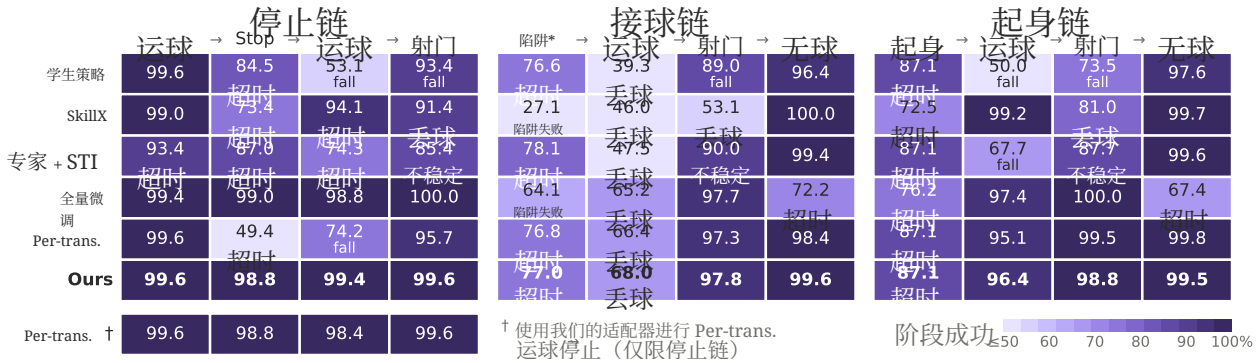


图 6 | 链条失败的位置。 每个单元格报告条件阶段成功率（%），该数值基于抵达该阶段的回合计算（每种方法初始为 512 个回合）。对于失败率至少为 5% 的阶段，显示占主导地位的失败类型。混合基线仅将 Per-trans. 的带球至停止适配器替换为我们的方法，从而大幅恢复停球阶段及后续阶段的表现。接球阶段包括此前的无球接近。为简洁起见，省略传球链。

运球–停球–运球–射门链上的完成率仅为 35.0% 和 35.2%，而 MESSI 的完成率为 97.5% 和 97.7%。全量微调在这些技能链上的表现相近，但在涉及停球的较长技能链上表现下降。这些结果促使我们考察连续执行过程中失败发生的位置。

失败分析。 图6按阶段定位失败。Per-trans. 的逐转换适应在停球阶段的成功率仅为 49.4%，随后在带球阶段的成功率为 74.2%。仅将其带球至停止适配器替换为我们的方法后，这两项成功率分别提升至 98.8% 和 98.4%，几乎恢复到 MESSI 的表现。这表明链条中的一个转换是主要瓶颈。对于接球链，MESSI 的失败主要集中在停球（77.0%）和后续带球（68.0%）阶段，而后续射门和无球阶段的成功率均超过 97%。全量微调在射门后的无球延续阶段也表现不佳（67.4%，而我们的起身链为 99.5%）。这些发现揭示了单个阶段的失败如何限制长时域技能组合。

5. 结论

我们提出了 MESSI，这是一种统一的基于物理的足球控制器，整合了八种具备任务能力的技能，并通过共享门控适配器调整它们之间的转换。我们的方法在保留原始学生策略原生执行能力的同时，提高了转换鲁棒性。实验进一步表明，可靠的技能组合取决于克服特定的转换瓶颈，而非孤立地优化各个转换。

参考文献

activeanimation. Aa 足球运动员动画包。 https://assetstore.unity.com/packages/3d/animations/aa-soccer-player-animations-package-237656?srsltid=AYCX9vKiPm1cYYCB25nomkbdf0-p2q_a3ZxbpWMA-uX_JasopIIApx5r，2022年。

Narula Avinash、Nguyen Khai、Fey Nolan、Parsuram Venki、Marangola John、Peticco Martin 和 Agrawal Pulkit。Joga：利用主动视觉学习敏捷人形机器人足球技能。2026年。

John Eric Goff 和 Matt J. Carré。足球轨迹分析。《美国物理学杂志》，77(11)：1020–1027，2009年。

Tuomas Haarnoja、Ben Moran、Guy Lever、Sandy H. Huang、Dhruva Tirumala、Jan Humplik、Markus Wulfmeier、Saran Tunyasuvunakool、Noah Y. Siegel、Roland Hafner、Michael Bloesch、Kristian Hartikainen、Arunkumar Byravan、Leonard Hasenclever、Yuval Tassa、Fereshteh Sadeghi、Nathan Batchelor、Federico Casarini、Stefano Saliceti、Charles Game、Neil Sreendra、Kushal Patel、Marlon Gwira、Andrea Huber、Nicole Hurley、Francesco Nori、Raia Hadsell 和 Nicolas Heess。利用深度强化学习为双足机器人学习敏捷足球技能。《科学·机器人学》，9(89)：eadi8022，2024年。doi：10.1126/scirobotics.adi8022。

Félix G. Harvey、Mike Yurick、Derek Nowrouzezahrai 和 Christopher Pal。稳健的动作插值。*ACMTransactions on Graphics*，39(4)，2020年。

Seokpyo Hong、Daseong Han、Kyungmin Cho、Joseph S. Shin 和 Junyong Noh。基于物理的全身足球动作控制：用于带球和射门。*ACMTransactions on Graphics*，38(4)，2019年。doi：10.1145/3306 346.3322963。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适应。*arXiv preprint arXiv:2106.09685*，2021年。URL https://arxiv.org/abs/2106.09685。

Jiakang Jin、Yixiao Huo、Pengyuan Wang、Yinan Han、Tingxuan Zhang、Zhuobing Zhao、Xuanxin Zhou、Zhangchen Ye、Enxuan Ruan、Yifei Bao、Jiankun Yang、Chenghao Sun、Wenhao Cui、Xiaoyu Tian 和 Yiming Li。DAVIS：面向人形机器人足球技能的仅深度端到端主动视觉框架。*arXiv preprintarXiv:2609.28175*，2026年。URL https://arxiv.org/abs/2609.28175。

Keller Jordan、Yuchen Jin、Vlado Boza、Jiacheng You、Franz Cesista、Laker Newhouse 和 Jeremy Bernstein。Muon：用于神经网络隐藏层的优化器。 https://kellerjordan.github.io/posts/muon/，2024年。

Jordan Juravsky、Yunrong Guo、Sanja Fidler 和 Xue Bin Peng。SuperPADL：通过渐进式监督蒸馏扩展语言指导的基于物理的控制。在 *ACM SIGGRAPH 2024 Conference Papers* 中，2024年。网址：https://arxiv.org/abs/2407.10481。

Minsu Kim、Eunho Jung 和 Yoonsang Lee。PhysicsFC：为基于物理的足球运动员控制器学习用户控制技能。*ACM Transactions on Graphics*，2025年。doi：10.1145/3731425。网址： https://arxiv.org/abs/2504.21216。

Libin Liu 和 Jessica K. Hodgins。使用轨迹优化和深度强化学习学习篮球运球技能。*ACM Transactions onGraphics*，37(4)：142:1–142:14，2018年。doi：10.1145/3197517。3201315。

Siqi Liu、Guy Lever、Zhe Wang、Josh Merel、S. M. Ali Eslami、Daniel Hennes、Wojciech M. Czarnecki、Yuval Tassa、Shayegan Omidshafiei、Abbas Abdolmaleki、Noah Y. Siegel、Leonard Hasenclever、Luke Marris、Saran Tunyasuvunakool、H. Francis Song、Markus Wulfmeier、Paul Muller、Tuomas Haarnoja、Brendan D. Tracey、Karl Tuyls、Thore Graepel 和 Nicolas Heess。从运动控制到模拟人形机器人足球中的团队协作。《科学·机器人学》，7(69)：eabo0235，2022年。doi：10.1126/scirobotics.abo0235。

Naureen Mahmood、Nima Ghorbani、Nikolaus F. Troje、Gerard Pons-Moll 和 Michael J. Black。AMASS：表面形状动作捕捉档案。在*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集中，第5442–5451页，2019年。

Maxdesign-3D。终极足球运动动作集。 https://www.fab.com/listings/a273e2ac-554f-4d90-b0ed-9ac4b9cb65b0，2025年。Mayank Mittal、Calvin Yu、Qinxi Yu、Jingzhou Liu、Nikita Rudin、David Hoeller、Jia Lin Yuan、Ritvik Singh、Yunrong Guo、Hammad Mazhar、Ajay Mandlekar、Buck Babich、Gavriel State、Marco Hutter 和 Animesh Garg。Orbit：面向交互式机器人学习环境的统一仿真框架。*IEEE Robotics andAutomation Letters*，8(6)：3740–3747，2023年。Liang Pan、Zeshi Yang、Zhiyang Dou、Wenjia Wang、Buzhen Huang、Bo Dai、Taku Komura 和 Jingbo Wang。TokenHSI：通过任务标记化统一生成物理人-场景交互。在*IEEE/CVF*计算机视觉与模式识别会议论文集中，第5379–5391页，2025年。URLhttps://arxiv.org/abs/2503.19901。Xue Bin Peng。MimicKit：用于动作模仿与控制的强化学习框架。 https://github.com/xbpeng/MimicKit，2025年。Xue Bin Peng、Pieter Abbeel、Sergey Levine 和 Michiel van de Panne。DeepMimic：示例引导的基于物理的角色技能深度强化学习。*ACMTransactions on Graphics*，37(4)，2018年。doi：10.1145/3197517.3201311。Xue Bin Peng、Ze Ma、Pieter Abbeel、Sergey Levine 和 Angjoo Kanazawa。AMP：用于风格化基于物理的角色控制的对抗运动先验。*ACMTransactions on Graphics*，40(4)，2021年。doi：10.1145/345062 6.3459670。Xue Bin Peng、Yunrong Guo、Lina Halper、Sergey Levine 和 Sanja Fidler。ASE：面向物理仿真角色的大规模可复用对抗性技能嵌入。*ACM Transactions on Graphics*，41(4)，2022年。doi：10.1145/3528223.3530110。Stephane Ross、Geoffrey J. Gordon 和 J. Andrew Bagnell。将模仿学习和结构化预测约化为无遗憾在线学习。在国际人工智能与统计学会议中，2011年。URL https://arxiv.org/abs/1011.0686。Andrei A. Rusu、Sergio Gomez Colmenarejo、Caglar Gulcehre、Guillaume Desjardins、James Kirkpatrick、Razvan Pascanu、Volodymyr Mnih、Koray Kavukcuoglu 和 Raia Hadsell。策略蒸馏。*arXiv*预印本 *arXiv:1511.06295*，2016年。URL https://arxiv.org/abs/1511.06295。John Schulman、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel。使用广义优势估计进行高维连续控制。在国际学习表征会议（*ICLR*）中，2016年。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。Haoyu Shao、Yunshen Wang、Jingze Zhang、Shaohang Zhu、Yong-Lu Li、Tianzhu Zhang、Baoxiong Jia 和 Siyuan Huang。像人类一样踢足球：通过潜在动作先验学习灵活敏捷的人形机器人足球。在机器人学习会议（*CoRL*）中，2026年。URL https://arxiv.org/abs/2611.06564。Zehong Shen、Huaijin Pi、Yan Xia、Zhi Cen、Sida Peng、Zechen Hu、Hujun Bao、Ruizhen Hu 和 Xiaowei Zhou。通过重力视角坐标实现以世界为基础的人体动作恢复。在*SIGGRAPH Asia 2024 ConferencePapers*中，2024年。Yi Shi、Yifeng Jiang、Chen Tessler 和 Xue Bin Peng。GPC：用于可迁移运动控制的大规模生成式预训练。在 *ACM SIGGRAPH 2026 Conference Proceedings*中，2026年。URL https://arxiv.org/abs/2606.29148。Chen Tessler、Yunrong Guo、Ofir Nabati、Gal Chechik 和 Xue Bin Peng。MaskedMimic：通过掩码动作修复实现统一的基于物理的角色控制。*ACMTransactionson Graphics*，43(6)，2024年。doi：10.1145/3687951。

Chen Tessler、Yifeng Jiang、Erwin Coumans、Zhengyi Luo、Gal Chechik 和 Xue Bin Peng。MaskedManipulator：适用于移动操作的多功能全身控制。载于 *ACM SIGGRAPH 亚洲 2025* 会议论文集，2025年。doi：10.1145/3757377.3763934。URL https://arxiv.org/abs/2505.19086。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的。神经信息处理进展，第 30 卷，2017年。

Yinhuai Wang、Qihan Zhao、Runyi Yu、Hok Wai Tsui、Ailing Zeng、Jing Lin、Zhengyi Luo、Jiwen Yu、Xiu Li、Qifeng Chen、Jian Zhang、Lei Zhang 和 Ping Tan。SkillMimic：从示范数据中学习篮球交互技能。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 17540–17549 页，2025年。URL https://arxiv.org/abs/2408.15270。

Zhaoming Xie、Sebastian Starke、Hung Yu Ling 和 Michiel van de Panne。利用分层混合专家学习足球颠球技能。载于 *ACM SIGGRAPH 2022 Conference Proceedings*，2022年。doi：10.1145/352823 3.3530735。

Pei Xu、Xiumin Shang、Victor Zordan 和 Ioannis Karamouzas。结合任务控制的 Composite Motion Learning。*ACMTransactionson Graphics*，42(4)，2023a。doi：10.1145/3592447。

Pei Xu、Kaixiang Xie、Sheldon Andrews、Paul G. Kry、Michael Neff、Morgan McGuire、Ioannis Karamouzas 和 Victor Zordan。AdaptNet：面向基于物理的角色控制的策略适应。*ACMTransactions onGraphics*，42(6)，2023b。URL https://motion-lab.github.io/AdaptNet/。

Pei Xu、Zhen Wu、Ruocheng Wang、Vishnu Sarukkai、Kayvon Fatahalian、Ioannis Karamouzas、Victor Zordan 和 C. Karen Liu。Learning to Ball：组合用于长时域篮球动作的策略。*ACM Transactionson Graphics*，2025a。doi：10.1145/3763367。URL https://arxiv.org/abs/2509.22442。

Sirui Xu、Hung Yu Ling、Yu-Xiong Wang 和 Liang-Yan Gui。InterMimic：迈向基于物理的人与物体交互的通用全身控制。载于*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，第12266–12277页，2025年b。doi：10.1109/CVPR52734.2025.01145。

Sirui Xu、SamuelSchulter、Morteza Ziyadi、Xialin He、Xiaohan Fei、Yu-Xiong Wang 和 Liangyan Gui。InterPrior：扩展基于物理的人与物体交互的生成式控制。*arXiv*预印本 *arXiv:2602.06035*，2026年。URL https://arxiv.org/abs/2602.06035。

Hongdi Yang、Chengyang Li、Zhenxuan Wu、Gaozheng Li、Jingya Wang、Jingyi Yu、Zhuo Su 和 Lan Xu。Smgdiff：使用扩散概率模型进行足球动作生成。载于*2025*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），第11807–11817页。IEEE，2025年。

Zhangchen Ye、Enxuan Ruan、Yifei Bao、Runhan Huang、Jiankun Yang、Jiakang Jin、Yixiao Huo、Pengyuan Wang、Yinan Han、Huaxing Huang、Wenhao Cui、Yiming Li 和 Xiaoyu Tian。SkillX：面向人形机器人足球的统一多技能策略学习。载于机器人学习会议，2026年。URL https://arxiv.org/abs/2609.06718。

Runyi Yu、Yinhuai Wang、Qihan Zhao、Hok Wai Tsui、Jingbo Wang、Ping Tan 和 Qifeng Chen。Skillmimic-v2：从稀疏且带噪声的示范中学习稳健且可泛化的交互技能。载于计算机图形学与交互技术特别兴趣小组会议论文集，第 1–11 页，2025年。

Siheng Zhao、Yanjie Ze、Yue Wang、C. Karen Liu、Pieter Abbeel、Guanya Shi 和 Rocky Duan。ResMimic：通过残差学习，从通用动作跟踪迈向人形机器人全身移动操作，2025年。URL https://arxiv.org/abs/2510.05070。

Yichao Zhong、Yidan Lu、Yuhang Lu、Tianyang Tang、Haoguang Mai、Yixuan Pan、Tianyu Li、Li Chen、Jingbo Wang、Zhongyu Li、Peng Lu 和 Hongyang Li。《RoboNaldo：通过动作引导的课程强化学习实现精准、稳定且强力的人形机器人足球射门》。*arXiv* 预印本 *arXiv:2606.11092*，2026年。URLhttps://arxiv.org/abs/2606.11092。

附录

内容

A 实现细节 .2

A.1 仿真、角色与球 .2

A.2 动作数据.3 A.3 专家训练.A.4 学生策略蒸馏 .4

A.5 TransitionAdapter.5 A.6 超参数与计算资源 .6

A.7 基线实现.7

B 评估指标 .8

B.1 原子技能. .8

B.2 转换 .9

B.3 原生保持能力。.10

B.4 连续链评估.10

C 特定技能专家详情.11

C.1 无球移动 .11

C.2 带球.12 C.3 射门 .13

C.4 传球.13 C.5 头球.14 C.6 停球 .15

C.7 急停转身 .16

C.8 起身.16**D 补充实验**.17 D.1 完整技能转换评估.17 D.2 更多技能转换可视化结果 .17

D.3 过渡适应消融.18 D.4 蛇形带球分析.19 D.5 技能链的其他结果 .19

E 局限性与未来工作 .22

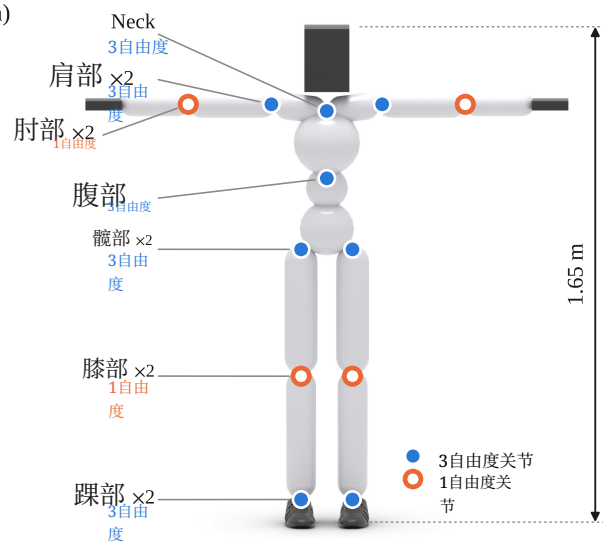
A. 实现细节

A.1. 仿真、角色和球

仿真。 所有实验均在 Isaac Lab 中进行 [Mittal 等人， 2023](Isaac Sim 5.1) 使用 GPU PhysX，代码库基于 MimicKit 构建 [Peng， 2025]。仿真以 240 Hz 运行，所有策略均以 60 Hz 进行查询。物理求解器采用四次位置迭代和一次速度迭代，并启用了连续碰撞检测。

角色。 该角色是一个人形体，包含15个刚体，总质量为52.1 kg，站立高度为1.65 m（图 7a）。它具有28个驱动自由度：腹部、颈部、肩部、髋部和脚踝各有三个，肘部和膝部各有一个。每个关节均由PD控制器驱动，策略指定目标角度。刚度k_p范围为100至1000 N m/rad，阻尼为 $k_d = k_p/10$，扭矩限制范围为50至200 N m；图 7b按关节列出了这些参数。脚部不使用方盒，而是采用运动鞋网格的凸包作为碰撞形状，因此与球的接触会遵循鞋子的形状（图 7c）。

(a)



(b)15个刚体，52.1千克，身高1.65米，28个驱动自由度；PD伺服器，目标来自策略，$k_d = k_p/10$

关节	DoF	k_p (N 米/弧度)	$\tau_{\text{最大值}}$ (N m)
腹部	3	1000	200
Neck	3	100	50
肩部 ×2	3	400	100
肘部 ×2	1	300	70
髋部 ×2	3	500	200
膝部 ×2	1	500	150
踝部 ×2	3	400	90
总计	28	100–1000	50–200

(c) 足部碰撞体：运动鞋网格（绿色）的凸包，而不是盒体（灰色）

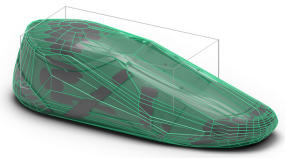


图 7 | 角色。 (a) 零姿态，即T字姿态，以及驱动关节及其自由度。(b) 各关节的刚度 k_p 和扭矩上限 $\tau_{\text{最大值}}$ （$k_d = k_p/10$）。(c) 足部碰撞体：运动鞋网格的凸包，用于替代方框。

球。 该球为半径$R = 0.11$ m、质量为 0.45 kg 的 5号足球。由于物理引擎既不模拟空气动力学，也不模拟滚动阻力，因此我们在每个策略步都对球施加三种附加力。设$\mathbf{v}$ 和 $\boldsymbol{\omega}$ 分别表示球的线速度和角速度， ρ_a表示空气密度， $A = \pi R^2$ 表示横截面积。当球的速度超过 1 m/s 时，会施加二次阻力

$$\mathbf{F}_d = -\tfrac{1}{2}\rho_a A C_d \|\mathbf{v}\|\mathbf{v}$$

。当球同时处于空中时，即球心高于其静止高度超过 3 cm 时，会施加马格努斯力

$$\mathbf{F}_m = \tfrac{1}{2}\rho_a A C_l \|\mathbf{v}\|^2 \widehat{\boldsymbol{\omega}\times\mathbf{v}}, \quad C_l = \min\left(C_l^{\max}, R\|\boldsymbol{\omega}\|/\|\mathbf{v}\|\right)$$

同样适用，其中升力系数遵循足球的风洞测量结果 [Goff and Carré， 2009]。在地面上，恒定减速度 0.15 g 会抵消水平速度，直到球停止。作为参考，空气阻力会使一次 25 m/s 的踢球在

表 6 | 球和地面的物理参数。空气密度为 $\rho_a = 1.2\text{kg/m}^3$。当两种材料接触时，它们的摩擦系数取平均值，并采用较大的恢复系数。角色使用默认材料（摩擦系数 1.0，恢复系数 0），因此球与地面以及脚的接触具有 0.725 的有效摩擦系数和 0.65 的恢复系数。

系数	Ball	地面
摩擦系数	0.45	1.0
恢复系数	0.65	0.0
线性阻尼	0.0	–
角阻尼	0.05	–
滚动阻力	$0.15\,g$	–
阻力系数 C_d	0.25	–
最大升力系数 $C_{\max_l}$	0.35	–

16° 仰角从约 34 m 降至 26 m。我们还
将球的最大角速度提高到 6×10^4 °/s，使其旋转不会被模拟器截断。表6总结了相关参数。

A.2. 动作数据

所有参考动作都已重定向到我们的角色，并以 60 fps 的速度回放。唯一的例外是一组经过减速处理的 22 个 LaFAN1 行走片段，它们以 30 fps 存储，并在回放时进行插值。足球动作采集自 Soccer-X [Yang等人、 2025], 两个商用动作捕捉数据集 [activeanimation、2022年、Maxdesign-3D，以及 2025], 使用 GVHMR 从视频重建的动作 [Shen等人， 2024]。这些片段中的脚部悬空问题均已在 Blender 中手动修正。部分片段还记录了球的轨迹，我们会在回合从参考状态开始时用它来初始化球。

无球数据集由 127 个运动片段组成：其中 65 个来自 AMASS 的 ACCAD 子集 [Mahmood等人，2019], 24 来自 LaFAN1 [Harvey等人， 2020], 23 来自 MimicKit 的示例动作，以及 15 个足球片段（包括镜像副本）。停球专家重复使用这些片段。起身数据集包含 297 个片段，其中 290 个是 LaFAN1 起身片段，一半经过镜像处理。起身回合从另一组 584 个摔倒片段初始化，这些片段不会提供给判别器。八位专家总共使用 526 个不同的片段（32 分钟）。表 7 列出了各项技能所用的数据。

A.3. 专家训练

每个专家均使用 PPO 进行训练 [Schulman 等人， 2017] 并采用对抗运动先验 [Peng 等人，2021]。遵循 Composite Motion Learning [Xu 等人，2023a], 任务奖励和风格奖励被视为独立目标。每个目标都有各自的价值头，其优势也分别进行归一化，因此决定各目标相对影响的是目标权重，而非奖励尺度。我们将这种 AMP 的多目标变体称为 MO-AMP（表10）。任务目标和风格目标的权重相等。射门和传球会增加一个腿部姿势目标，任务、风格和姿势的权重分别为 0.4、0.4 和 0.2。头球会增加一个姿势目标，该目标还涵盖颈部、腹部和肩部，权重分别为 0.5、0.3 和 0.2（附录C.5）。风格奖励为 $r\text{style} = -\log \max(1 - D,\ 10^{-4})$。

判别器只观察角色，不观察球。其输入是以最后一帧的根节点为参照的一段身体状态窗口；对于射门、停球、无球移动和起身，还包括关节速度。对于射门和传球，判别器还以

表 7 | 技能专家的训练设置。Len. 表示最大回合长度。Clips 和 Data 指判别器使用的动作数据，Samples 和 Time 指每位专家的训练预算。

技能	Len. (s)	片段	数据 (分钟)	样本 (B)	时间 (h)
无球	20.0	127	14.6	5.90	21.0
带球	20.0[a]	26	0.9	0.73	13.2
射门	10.0	10	0.5	2.12	24.4
传球	3.0	22	0.6	4.17	52.9
头球	2.5	8	0.1	3.41	30.8
停球	2.2[b]	158	15.2	5.82	39.4
急停转身	4.0	30	1.1	2.83	19.4
起身	6.0	297	14.5	6.55	35.5

[a] 在训练期间，回合也会在其参考片段结束时（最长 4.5 秒）结束。 [b] 回合会在球计划到达时间之后 0.5 秒结束。

踢球技术为条件，并通过腿部姿势目标鼓励动作与参考动作保持一致。头球采用姿势目标，重点关注颈部、腹部和肩部。停球判别器以球类别为条件，无球判别器则以角色正在移动还是站立为条件。

策略输出归一化动作。对于每个单轴关节，动作会被映射为关节范围中点附近的 PD目标，而每个三轴关节则将其三个动作解释为指数映射。表10 中的动作标准差以这些归一化单位表示。观测使用运行统计量进行归一化，并在前 10^8 个样本之后冻结。未应用领域随机化。

每个回合都从技能专属时间窗口内采样的参考状态初始化，例如，对于踢球技能，采样击球之前的帧。当除脚以外的任何身体部位接触地面时，将触发提前终止；但起身技能例外，因为它从地面开始。当球距离超过 2.5 米时，急停转身回合也会终止。无球、传球和停球专家均基于较早的检查点进行微调，附录C 介绍了每项技能的最终设置。每个专家均使用单个随机种子进行训练。表10 列出了超参数。

A.4. 学生策略蒸馏

学生策略观测各专家观测的并集（表8）。所有量均以角色的朝向坐标系表示，并且在归一化之前，非活动技能的观测块均填充为零。除技能独热编码外，每个区块都通过各自的线性层投影为一个256维令牌，并加上学习得到的令牌类型嵌入。遵循 TokenHSI [Pan 等人的工作，2025], 技能独热编码经嵌入后加到每个令牌上。随后，令牌由 Transformer编码器处理，其注意力掩码设置为：除本体感觉令牌外，每个令牌只能关注自身和本体感觉令牌。最后，将输出令牌拼接起来，并通过一个含有1024个单元隐藏层的多层感知机映射为28维动作均值。

首先，使用每位专家在其自身环境中的滚动轨迹，通过行为克隆训练学生策略。这些滚动轨迹采用每位专家的最后一个课程阶段，且不使用自适应裁剪采样。随后，我们进行四轮 DAgger [Ross 等人，2011]。每一轮中，执行学生策略的平均动作，并将每个访问状态标记为相应专家的动作。各轮数据会汇总起来。每轮结束后，重新初始化学生策略，并从头开始使用所有已收集的数据进行训练

表格 8 | 学生策略的观测块。每位专家都会观测本体感觉、球以及自身的指令块。急停转身没有指令。

区块	Dim.	内容
本体感觉	140	根节点高度、旋转和速度；关节旋转；关节速度；头部、手部以及足部位置
Ball	7	存在标志；相对位置和速度
转向	5	方向、速度和朝向（无球、带球）；带球将其方向重复为朝向
射门	9	发射速度、脚、技术；击球标志位和击球后的经过时间
传球	7	发射速度、技术；击球标志位和击球后的经过时间
起身	2	恢复标志位、经过时间
头球	4	发射速度；技术标签（单一类别）
停球	3	身体部位，成功标记
技能	8	技能独热编码（仅学生策略）
总计	185	

，同时重新拟合观测归一化器，并留出 5% 的样本用于验证集。损失为学生策略与专家的归一化动作均值之间的均方误差。

A.5. Transition Adapter

上下文。 过渡上下文 $\mathbf{c}_t$ 是一个 80 维向量。交接后，它包含源技能和目标技能的独热标签（16 维）、源阶段、开放门控和目标阶段的标志位（3 维），以及自交接以来的时间（除以 5 秒并截断至 1，1 维）。它还包含目标技能的命令块，放置在统一观测的 30 个命令维度中对应的位置，速度和发射速度均除以 10（30 维），并配有一个用于标记该区块的掩码（30 维）。交接前，目标条目存放源技能及其命令，时间条目则从源技能的起点开始计时。

架构。 公式 4 中的低秩矩阵秩为 $r = 16$，LoRA 缩放系数设为$\alpha = r$，因此应用更新时不进行重新缩放。LoRA 应用于学生策略的十个线性层：本体感知分词器、每个 Transformer 块中的两个前馈层，以及读出层。注意力投影、球和任务分词器、技能嵌入以及动作头保持不变。每个适配层都有各自的条件器，这是一个大小为 $80 \to 64 \to 32$、采用 SiLU 激活函数的多层感知机。下投影矩阵使用 Kaiming 均匀初始化，上投影矩阵和每个条件器的最后一层则初始化为零，因此训练开始时，适配后的策略与学生策略完全相同。退出头是一个包含两个隐藏层（分别有 256 和 128 个单元）的多层感知机，以归一化后的观测和 $\mathbf{c}_t$ 为输入。其输出权重初始化为零，输出偏置设为 -7，因此训练初期很少提前退出；且只有在经过 $\tau_{\text{至少}} = 0$。2 秒后才允许退出。评论家是一个独立的 Transformer，其架构与学生策略相同。其输出特征与$\mathbf{c}_t$ 拼接后，传入一个具有 1024 个单元并采用 SiLU 激活函数的层，再得到价值输出。

回合。 表9 列出了用于训练的 11 种转换，以及各自的交接方式和成功条件。每个回合都从源技能在其专家所处场景中开始，初始化方式与训练时相同（附录 C），处于最终球类课程阶段，且不使用自适应裁剪采样，但有三种例外。无球移动从带球场景开始，球静止放置在距角色 2–4 米处，位于其朝向的 范围内，并指令角色以 2 米/秒的速度朝球跑去。停球从朝交汇点的无球接近开始，持续至球到达前 0.6 秒，但最长为

表 9 | 用于适配器训练的技能过渡。 全部 11 种转换中源技能的交接条件及目标技能的成功标准。

转换：无球→运球，运球→{射门、传球、停止}，
停止→运球，停球→运球，起身→{运球、无球}，
{射门、传球、头球}→无球。

源	交接时机
无球	距离放置在 2–4 米外的无人控制的球 1 米以内
带球	起点后 1 秒，球可触及
急停转身	球已停止（停球事件，附录C.7），球可触及
停球	成功位置位，球可触及
起身	已恢复（进入带球状态：球也可触及）
射门、传球	击球后 1 秒
头球	头球后 0.3 秒，最晚不超过 0.6 秒
目标	**成功条件**
带球	沿指令前进 2 秒并移动 0.5 米后，球可触及
射门、传球	击球后 1 秒（射门：使用指定脚）
急停转身	球已停止，球可触及
无球	速度（0.3 秒平均值）在 0.5 m/s 以内，且朝向在 30° 以内与指令一致，持续 0.3 秒，至少在交接后

0.3 秒。对于起身→运球，球静止放置在距身体 1–1.5 米处。指令带球源沿初始朝向以 1.5 米/秒的速度移动。当门控关闭时，执行学生策略的平均动作，不添加探索噪声，与测试时相同。交接后，如果角色摔倒、球丢失（持续 0.25 秒距离超过 4 米；急停转身时为 2.5 米；无球移动期间或击球后不检查），或目标未能及时成功（无球移动为 3 秒，带球和急停转身为 4 秒，踢球则须在 5 秒内完成射门），回合即告失败。摔倒或丢球的惩罚p_t 为 1，其他失败（例如超时或用错误脚射门）的惩罚为 0.5。门控开启期间，每一步的使用成本为 $\kappa = 0.02$。成功回合会被截断并进行价值自举，而非视为终止状态。这样可以避免在任务奖励随时间累积时惩罚过早成功。

交接条件。 在表 9 中，若球位于 2.5 m 范围内且高度低于 1.5 m，则认为可触及。除无球追踪外，所有条件都要求直立站姿（骨盆高于 0.75 m，且除脚之外没有其他地面接触）；对于带球、无球和头球，这一姿势须在交接帧检查，对于其他情况则须在此前 0.15 秒内检查。起身本身已要求保持站立 0.3 秒，而头球会在接触后 0.6 秒内完成交接，除非角色摔倒。评估标准在第 4.4 节和 4.3 节中另行规定。

优化。 每个转换都在独立的工作进程中进行训练，使用 1,024 个环境。11 个工作进程共享四个 GPU，并在每次更新时对梯度求平均。每个转换接收 33.6M 个样本，总计 369M 个样本。其余超参数列于表 11中。

A.6. 超参数和计算资源

表 10 和 11列出了训练超参数。所有策略均在 NVIDIA RTX A6000 GPU 上训练。每个专家使用四块 GPU 训练 13 至 53 小时（表 7）。蒸馏在单块 GPU 上耗时 4.4 小时（不包括评估），适配器则使用四块 GPU 训练 11.2 小时。GPU PhysX 仿真在固定机器上具有可复现性，但在不同机器之间会有所差异。因此，我们在同一个作业中，从完全相同的初始状态开始，执行每一组配对比较，例如比较适配器与直接切换。

A.7. 基线实现

统一多技能基线。 TokenHSI* [Pan 等人、 2025] 和 SkillX* [Ye 等人、 2026] 均使用我们的观测、任务奖励和动作数据，在全部八项技能上进行联合训练。两者均使用四块 RTX A6000 GPU、8,192 个环境，以及一个由 $0.5r_{\text{task}} + 0.5r_{\text{style}}$组成的组合奖励，其中$r_{\text{style}} = -2\log(1 - D)$。它们不使用我们的技能特定课程学习、失败加权采样、多目标优势、Muon 优化器或技术特定判别器标签。TokenHSI 使用一个共享判别器，并嵌入 8 个技能的独热编码标签；SkillX 则为每项技能使用 8 个不同的判别器。两者均使用确定性动作进行评估。

TokenHSI*。 我们遵循官方实现 [Pan 等人的 2025], 设置，包括其 64 维任务标记、四层 Transformer、PPO/AMP 设置以及两阶段训练。第一阶段从原生初始化开始联合训练八项技能（0.82B 个样本，30 小时）。第二阶段冻结共享演员网络，并在六条固定技能链上训练任务分词器和动作头适配器（0.12B 个样本，4 小时）。

SkillX*。 由于官方代码不可用，我们重新实现了 SkillX [Ye 等人的实现 2026] ，采用其五帧观测历史、双层 Transformer、多层感知机演员网络和 KL 自适应 PPO。每个 20 秒的回合每隔 5–10 秒切换一次技能，且不会重置角色或球。为在我们的任务上获得稳定训练，我们采用参考状态初始化，将判别器的最小二乘损失替换为二元交叉熵，将动作标准差限制在 1.0 以内，并终止失败技能。我们还引入了仅用于训练的带球球损失条件。训练使用了 0.77B 个样本，耗时 26.1 小时，并评估最终检查点。按计划进行的失败加权技能采样以及后续切换课程学习在该训练预算下尚未达到。

转换适应基线。 全量微调和 Per-trans. 使用与 MESSI 相同的 11 个在线转换、交接条件、目标奖励和 PPO 样本预算（每次转换 33.6M 个样本）。冻结学生策略执行前序技能，并在交接时开始适应。

Per-trans. 每种转换都单独训练一个适配器，其架构、退出策略和超参数与 MESSI 相同。每个适配器使用 3360 万个样本，并且仅部署于对应的技能对。图 6 中的混合基线仅将其从运球到停止的适配器替换为我们的方法。

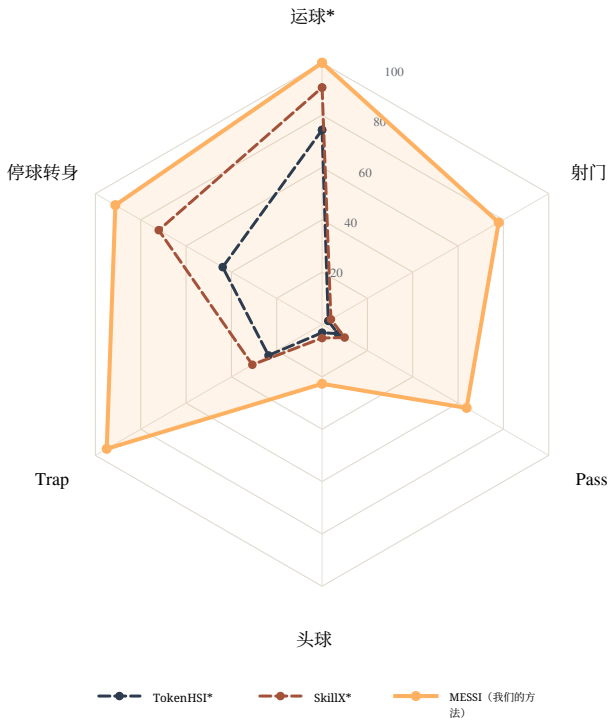


图 8 | 六项足球技能的原子技能性能。 在原生初始化条件下，我们的蒸馏学生在全部六项技能上均优于 TokenHSI* 和 SkillX*。带球技能通过逆速度误差进行归一化（$100 \times e_{\text{最小值}}/e$）；其他坐标轴表示各技能特定的成功率（%）。

全量微调。 交接后，我们对学生演员网络参数进行全量微调，不使用适配器或退出策略。我们使用 10^{-5} 的学习率；更高的学习率 5×10^{-5} 会导致向无球移动的转换崩溃。评估期间，微调策略会完全取代原始学生策略，包括原生

表 10 | 技能专家的超参数，除非另有说明，否则由全部八项技能共享。三元组分别指演员、评论家和判别器。

参数	值
算法	MO-AMP（PPO）
演员和评论家	多层感知机 4×1024[a]
判别器	多层感知机 2×1024[b]
判别器窗口	20 帧[c]
梯度惩罚	5–20
Logit正则化	0.01
回放缓冲区	200k[d]
动作速率惩罚	0–0.1
动作边界惩罚	10
动作标准差（固定）	0.055[e]
优化器	Muon[f] / Adam / SGD[g]
学习率	5×10^{-4} / 10^{-4} / 2.5×10^{-4} [h]
权重衰减	0.01 / 0 / 10^{-4}
每次更新的训练轮数	5 / 2 / 2
每轮的小批量数	8 / 16 / 16
梯度裁剪	50
折扣因子 γ / 广义优势估计 λ [Schulman 等人， 2016]	0.99 / 0.95
PPO 裁剪比率	0.2
归一化优势裁剪	4
环境	8,192[i]
展开长度	32
GPUs	4

[a] 无球技能使用（1024，512）。 [b] 2×256 传球、头球和起身技能使用 [c] 10 起身技能使用 [d] 100头球使用 k。 [e] 0头球使用 .05。[f] [Jordan 等人， 2024], 应用于权重矩阵；偏置使用 AdamW，学习率为 10^{-4}。 [g]动量为 0.9。[h] 1。25×10^{-4} 用于停球和急停转身的判别器。 [i] 16无球和停球使用 384，传球使用 6,144。

回合和连续技能链。

专家 + STI。 遵循 PhysicsFC [Kim 等人的工作， 2025], 我们使用其原始的 PPO/AMP 配置，对五个目标技能专家进行微调。训练回合以 0.7 的概率从前序技能生成的交接状态开始，否则从原生初始化开始。对于具有多个前序技能的目标技能，单个专家会针对所有相应的转换进行联合微调。每个转换贡献 3,360 万个 PPO 样本，与我们的适配器训练预算相匹配。总计而言，五个微调专家覆盖全部 11 个转换，使用约 3.69 亿个样本。

B. 评估指标

所有策略均采用其平均动作。每项比较均在同一台机器上，从相同的初始状态开始运行（附录 A.6）。技能控制中的技能规则规定的终止情况均视为失败（附录 C）：所有技能中，摔倒均算作失败；此外，各技能还有自身的失败条件，例如急停转身中球距离超过 2.5 米。表 13 列出了每项技能的成功规则。

B.1. 原子技能。

每项技能均在 1,024 个固定案例上进行评估，这些案例取自其专家训练时的初始状态和指令。专家的运行轨迹生成这些案例，其他所有策略只有在其初始物理状态、任务状态和指令序列与这些案例完全匹配时才会被接受。一个回合会持续到该

表 11 | 学生策略和适配器的超参数。各组数值分别对应学生策略的行为克隆和 DAgger，以及适配器的演员和评论家。

	学生策略	适配器
算法	行为克隆 +DAgger[a]	PPO
网络	Transformer[b]	LoRA + 特征线性调制[c]
可训练参数 5.3M		0.39M[d]
动作标准差（固定）	0.055	0.055
优化器	AdamW[e]	Adam
学习率	3×10^{-4}，余弦[f]	3×10^{-4} / 10^{-4}
折扣因子 / 广义优势估计	–	0.99 / 0.995
PPO 裁剪比率	–	0.2
环境	512 / 256[g]	1,024[h]
展开长度	600 / 400	64
每次更新的训练轮数	40 / 30	5 / 2
小批量大小	2,048	8,192 / 4,096[h]
退出头	–	多层感知机 (256, 128)
退出$\tau_{\text{最小值}}$，成本 κ	–	0.2 秒，0.02
样本	250万 + 4×0.8M	3360万[h]
GPUs	1	4

[a] 4 DAgger 轮次。 [b] 4 层，宽度 256，4 个头部，前馈宽度 1024。 [c] 在 10 层上秩为 16。 [d] 0LoRA 和条件器中有 .29 M，退出头中有 0.10M。 [e]不使用权重衰减；梯度范数裁剪至 1.0。 [f] 每轮重新启动。[g] 每项技能。[h]小时每次转换。

表 12 | 状态转移基线。全量微调和 Per-trans. 在我们的适配器的 11 个转换上进行训练；专家 + STI 微调涵盖全部 11 个转换的五个目标技能专家。样本按转换计。

方法	训练对象	回合起点	样本	Exit
Ours	一个共享适配器（0.39M）	在线交接	已学习 33.6M	
Per-trans.	每次转换使用一个适配器（11×0.39M）	在线交接	已学习 33.6M	
全量微调	所有学生演员权重（5.3M）	在线交接	33.6M	–
专家 + STI	五个目标技能专家	70% 已存储交接，30% 原生	3360 万	–

技能自身的终止条件或时域（附录 C）。对于无球移动和带球，速度误差是平面根速度与指令速度之间的距离，而朝向误差是朝向与指令朝向之间的角度。两者均在回合的所有帧上取平均，包括每次指令变化后的瞬态过程。失败率统计以失败告终的回合数。对于踢球，发射速度是检测到的击球时球的速度（附录 C.3–C.5），方位角误差是发射速度与指令速度之间的平面角。

B.2. 转换。

一个案例对应一次交接。冻结学生策略在自己的场景中执行源技能，直至满足交接条件（表 9）；我们通过对交接帧的完整状态进行哈希校验，确保每种方法都从完全相同的交接状态接管。使用两个随机种子发出 256 个请求，可为每种转换产生约 400 至 510 次交接；源技能未能交接的请求不计入目标技能。从交接开始，目标技能以原生回合运行，并在首次失败时确定结果。带球目标的指令速度为 1.5m/s，方向取决于源技能：急停转身后朝进入时的反方向，停球后朝初始朝向，无球移动后朝交接时的朝向，起身后朝向球。无球目标的指令为沿均匀

表 13 | 成功规则。原生：表 1。转换目标：表2 和 3，其中每条规则还要求在截止时间之前未发生失败。速度和朝向误差取给定帧数内的平均值。

技能	原生	作为转换目标
无球	无失败；速度误差 < 0.75 米/秒，且朝向误差 $< 35°$ 在整个回合中	速度误差 < 0.75 米/秒，且朝向误差 $< 35°$ 在前 4 秒
带球	速度误差和失败率（无成功规则）	球始终在 2.5 米以内；速度误差 < 0.75 米/秒，持续 3–4 秒
射门	方位角误差 $< 10°$，发射速度为 $\geq 50\%$ 指令，指定脚	作为原生；发射后 1 秒内不会失败
传球	方位角误差 $< 10°$，发射速度为 $\geq 50\%$ 的指令	作为原生；发射后 1 秒内不会失败
头球	首次检测到的合法触球为头球（附录C.5），方位角误差 $< 25°$	–
停球	成功位（附录C.6）	–
停转停球锁存器（附录C.7）		停球锁存器
起身	已恢复（附录 C.8）	–

随机方向移动，并朝向行进方向，速度为 1–3 m/s。踢球指令要求沿带球方向踢出，力度取该技术范围的中间值，并在各请求中均衡分配技术和支撑脚。成功与否根据 4 s 的截止时间判定；踢球则在发射后 1 s 判定。带球的速度误差取截止时间前最后一秒的平均值，因此急停→运球的转身阶段不计入平均值；无球移动的速度误差则从交接时开始取平均值，以此衡量技能建立得有多快。在表 3 中，退出表示在截止时间前触发学习式退出的交接比例；开启表示从交接开始到门控保持开启的平均时间，上限为截止时间。事后截断会在交接后的指定时间精确关闭门控。

B.3. 原生保持能力。

表 2 中的原生列使用表 1 中的案例和相同指标评估每个已部署策略。可关闭的策略（如我们的适配器）会在该处运行学生策略；微调策略则运行自身。

统计学。 当两种方法在相同案例上接受评估时，我们使用双侧精确符号检验对其进行比较。两种方法都成功或都失败的案例会被舍弃。若两种方法同样出色，则每个剩余案例支持任一方法的概率均为 1/2；在这一假设下，若 $p < 0.05$，我们便称差异显著。该检验仅考虑评估案例之间的变化。每个训练策略都来自单个训练随机种子，因此该检验未涵盖不同训练运行之间的变化。

B.4. 连续链评估

每条技能链在连续仿真中运行，最长为 20 秒，期间不会重置角色或球。初始技能遵循其对应的专家环境，后续技能则在阶段完成后切换。表 14 列出了完成和失败标准。

停球链以带球开始，接球链以无球接近来球开始，起身链则从倒地状态开始。后续技能的指令根据转换评估协议进行采样。

所有方法共享相同的初始状态和指令。每种方法都使用自己的策略执行第一个技能，因此后续的交接状态可能有所不同。对于 MESSI，每次技能切换都会激活

表格 14 | 连续链的阶段测试。阶段在摔倒或满足所列条件时失败。可达：球在 2.5 米以内且低于 1.5 米；站立：骨盆高于 0.75 米，躯干保持直立，且除脚之外身体没有其他部位接触地面，并在此状态下持续最后 0.15 秒。

阶段	完成条件	失败条件
第一阶段		
带球	起点后 1 秒，球可触及，站立（在该帧）	不在 2 秒内；球远于 4 米，持续 0.25 秒
停球	成功位已设置，球可触及，站立	停球失败（附录 C.6）；计划到达时间后 0.5 秒
起身	已恢复，球可触及	在 6 秒内未恢复
后续阶段		
急停转身	球已停住（陷阱锁存）、球可触及、站立	球距离超过 2.5 米达 0.25 秒；4 秒内未恢复
带球	沿指令前进 2 秒且推进 0.5 米后，球可触及，站立	球距离超过 4 米达 0.25 秒；4 秒内未恢复
射门，传球	击球后 1 秒，保持站立（任一只脚）	此时未站立；球距离超过 4 米，持续 0.25 秒；或 5 秒内没有击球
无球	速度（0.3 秒平均值）在 0.5 米/秒以内，且朝向持续 0.3 秒处于指令的 30° 以内，至少 0.5 秒交接后	不在 3 秒内

带有新过渡上下文的适配器，而学习式退出则决定何时将控制权交还给冻结学生策略。

C. 技能专家

本节介绍每个技能专家的指令、奖励、初始化和终止条件。任务奖励是以下各项的加权和 $[0,1]$，但停球的基于势能的接近项除外；权重为负的项是惩罚项。我们将根节点和球的水平位置分别记为 $\mathbf{p}$ 和 $\mathbf{b}$，将它们的速度分别记为 $\dot{\mathbf{p}}$ 和 $\dot{\mathbf{b}}$，并将角色的朝向记为 $\mathbf{h}$。距离单位为米，且 $[x]_+ = \max(x,0)$。摔倒是指身体中除脚以外的任何部位与地面发生接触。多个专家通过分阶段课程进行训练，或基于更早的检查点进行微调；我们介绍每个专家的最终阶段。目标权重、判别器设置和 PPO 超参数见附录 A。

C.1. 无球移动

指令。 指令由移动方向 $\mathbf{u}$、速度 v^* 和朝向 $\mathbf{w}$组成，均以朝向坐标系表示。每隔 4–12 秒采样一次新指令。方向 $\mathbf{u}$ 和$\mathbf{w}$独立且均匀地抽取，而 v^* 则在 0.5 米/秒与一个最大速度之间均匀抽取；该最大速度取决于$\mathbf{u}$ 和$\mathbf{w}$之间的夹角。在 0°、45°、90°、135° 和 180° 处，最大速度分别为 5.0、3.5、2.5、2.3 和 3.5 米/秒，其间采用线性插值。角色有 0.1 的概率会被指令要求原地站立（$v^* = 0$）。若某条指令要求的移动方向与朝向偏差超过 45°，则下一条指令有 0.2 的概率为停止，并保持相同朝向。

奖励。 令 $\bar{\mathbf{v}}$表示根节点速度 $\dot{\mathbf{p}}$在 0.3 秒的指数窗口内取平均后的值，并令 $c = [\mathbf{h}\cdot\mathbf{w}]_+$表示朝向对齐程度。任务奖励为

$$r^{\text{task}} = 0.7\, r^{\text{vel}} + 0.3\, r^{\text{face}} + 0.3\, r^{\text{smooth}}. \tag{6}$$

对于移动指令，速度项为

$$r^{\text{vel}} = \exp\!\left(-0.5\,\|v^*\mathbf{u} - \bar{\mathbf{v}}\|^2/\sigma^2\right) \cdot \left(f + (1-f)\,c\right) g_{\text{step}}, \tag{7}$$

其中 $\sigma =$ 最小值$(0.5\,v^*,\ 1)$ 米/秒；如果角色相对于指令向后移动（$\bar{\mathbf{v}}\cdot\mathbf{u} < 0$），则 r为速度。使用 σ 对 v^* 进行缩放，可以防止缓慢的指令通过静止不动来满足：在 $v^* = 0.5$ 米/秒时，站立不动可获得 e^{-2}，而固定 $\sigma = 1$时则为 $e^{-0.125}$。朝向系数 f 为 0.2。对于侧向指令，即相对于朝向偏转 45° 至 135° 的指令，当v^* 从 1 米/秒增加到 2 米/秒时，f 线性增加至 1。步态门控 g步为 $=$最小值$(1,\ \Delta z/0.03$ 米$)$，当角色不抬脚而拖步时会暂不发放奖励，其中 Δz 是过去 1 秒内两只脚高度变化范围中较小的值。当 v^* 从 1 米/秒增加到 1.5 米/秒时，该门控会线性淡出。

对于站立指令，$r\text{vel} = \exp(-128\,\|\bar{\mathbf{v}}\|^2)\,(0.2 + 0.8\,c)\,(0.5 + 0.5\,r\text{stand})$。站姿项 $r$stand 是多个高斯函数的乘积，用于使腹部和腿部关节角度、骨盆高度及骨盆俯仰角接近站立片段的均值，使脚保持水平，并从指令开始 1.5 秒后使负载在双脚之间保持平衡。朝向项为 $r\text{face} = c\,g\text{step}$。平滑度项 $r\text{smooth} = \exp(-k\,\ddot{q}^2)$ 会惩罚 14 个腿部关节的平均平方加加速度，其中 $k = 2\times10^{-8}/(1 + (v^*/1.5)^2)$。判别器同样以移动/站立标志作为条件。对于策略，该标志由指令给出；对于示范窗口，当平均根节点速度低于 0.1 m/s 时将其设定为移动/站立标志。

初始化。 每个回合都从 127 个跑步、行走、侧向迈步、倒退行走、转向和站立片段之一中的随机时刻所对应的参考状态开始，其中包括镜像副本。每个片段被采样的概率与其数据集权重乘以难度分数成正比；难度分数根据该片段近期的跌倒率及其偏离参考状态的程度计算得出。判别器的示范窗口也从相同分布中抽取。球被静止放置在 10m 外，其观测被设为零。

终止。 当角色摔倒或经过 20 s 后，一个回合结束。

C.2. 带球

指令。 该指令由目标方向 $\mathbf{u}$ 和速度 $v^* \in [0。5, 5]$ m/s 组成，每 4–7 s 重新采样一次。不存在朝向指令，也不会对朝向进行奖励。

奖励。 令 $\mathbf{b}^* = \mathbf{p} + 0。45\mathbf{u}$ 表示目标球位置，即沿指令方向位于角色前方 0.45 m 处。误差 $\mathbf{e} = \mathbf{b} - \mathbf{b}^*$ 被分解为纵向分量 $e_{\parallel} = \mathbf{e}\cdot\mathbf{u}$和横向分量$e_{\perp}$。任务奖励为

$$r^{\text{lead}} = \exp\left(-(e_{\parallel}^2 + 4\,e_{\perp}^2)\right), \tag{8}$$

$$r^{\text{vel}} = \exp\left(-0.5\,\|v^*\mathbf{u} - \dot{\mathbf{p}}\|^2\right), \tag{9}$$

$$g_{\text{ball}} = \exp\left(-1.5\,[\|\mathbf{b} - \mathbf{p}\| - 0.6]_+\right), \tag{10}$$

$$r^{\text{task}} = g_{\text{ball}}\,r^{\text{vel}} + 0.6\,r^{\text{lead}}, \tag{11}$$

其中 $r\text{vel} = 0 \text{ if } \dot{\mathbf{p}}\cdot\mathbf{u} < 0$。当球距根节点在 0.6 m 以内时，门控 $g$ball 等于 1，因此不会惩罚正常运球的球距。超过这一距离后，门控值会逐渐衰减：距离为 1.6 m 时降至 0.22，距离为 2.6 m 时降至 0.05。

初始化。 每个回合都从运球片段中随机时刻的参考状态开始。球的位置取自该片段，其速度则根据该片段中球轨迹的 10 帧拟合得出。

终止。 当角色摔倒或经过 20 秒后，一个回合结束。在训练期间，一个回合也会随着其参考片段（表 7）结束。

C.3. 射门

指令。 该指令由朝向坐标系中的目标初速度 $\mathbf{v}^*$ 、击球脚和技术组成，后两者均为独热向量。脚和技术取自重置时采样的参考片段。脚背射门的速度为 22–35 m/s，仰角为 0–20°；脚内侧射门的速度为 16–26m/s，仰角为 0–15°。对于脚背射门，方位角在 ±15° 范围内均匀采样；对于脚内侧射门，则在 ±20° 范围内均匀采样，采样中心为参考片段击球帧时角色的朝向。策略还会观测球是否已被击出，以及击球后时间；后者除以 1 秒后截断为 1。

奖励。 当球速首次超过 3 m/s 时，检测到击球，并将该时刻的球速度保存为 $\mathbf{v}_{\text{launch}}$。令 t 为击球后时间，并令 $\rho_{a,\,b} = \text{片段}((t-a)/(b-a),\ 0,\ 1)$为一个在击球前值为零的渐变函数。任务奖励为

$$r^{\text{task}} = 0.3\, r^{\text{reach}} + 0.6\, r^{\text{launch}} + 0.15\, r^{\text{foot}} + 0.3\, r^{\text{settle}} + 0.3\, r^{\text{rec}}. \tag{12}$$

击球前，触球项 $r\text{reach} = \exp(-2\, d^2)$ 将一只脚引向球，其中 d 是较近的脚到球的距离。击球后，$r\text{发射} = \exp(-0.01\, \|\mathbf{v}_{\text{launch}} - \mathbf{v}^*\|^2)$ 对发射速度进行评分，而 $r\text{foot} = 1$ 则在指定脚完成击球时得分。这两项在回合剩余时间内均保持击球时的数值。稳定项

$$r^{\text{settle}} = \rho_{0.25,0.5}\, \exp\left(-[\|\dot{\mathbf{p}}\| - 1.2]_+^2\right) \cdot \exp\left(-100\, (z - 0.86)^2\right) \tag{13}$$

要求根节点 t 使其减速至不超过 1.2 m/s，且高度 z 接近 0.86 m。恢复项

$$r^{\text{rec}} = \rho_{0.6,1.0}\, \exp\left(-[(z_f - 0.3)/0.15]_+^2 - [(s_f - 4)/2]_+^2\right) \tag{14}$$

要求脚部回落，其中 z_f是较大的脚部高度， s_f是相对于根节点的较大脚部速度。姿势目标（权重为 0.2）为 $r\text{姿势} = \exp\left(-10 \sum_i w_i \delta_i^2 / \sum_i w_i\right)$，其中 δ_i 是关节 i的模拟旋转与参考旋转之间的角度。权重 w_i 对脚踝为 3，对膝盖和髋部为 2，对其他关节为 1，因此腿部在姿势目标中占主导地位。

初始化。 每个回合都从片段击球帧之前采样的参考状态开始。球会在朝向坐标系中向前偏移最多 ±0。4 m，并向侧面偏移最多 ±0。6 m。随着采样时间接近击球帧，该偏移会线性缩小至零，因此角色始终有时间调整其接近方式。在 40% 的回合中，球从静止开始。否则，球会以最高 3 m/s 的速度滚动，横向速度最高为 0.9 m/s，其方向与参考球的运动方向或（若参考球静止）头部方向的偏差不超过 ±90°。

终止。 当角色摔倒、参考片段结束或经过 10 秒后，回合即告结束 .

C.4. 传球

指令。 指令由朝向坐标系中的目标初速度 $\mathbf{v}^*$ 以及作为独热向量的技术组成，该技术取自重置时采样的参考片段。指令中不包含击球脚。数据集包含每个片段及其镜像，因此任一只脚都可以击球。地滚球传球使用 4.7–13.5 m/s 的速度和 0–8° 的仰角，同时还会限制仰角，使垂直发射速度保持在 1m/s 以下。挑传使用 13.5–24m/s 的速度和 18–35° 的仰角。对于地滚球传球，方位角在参考片段发射方向周围的 ±30° 范围内均匀采样；对于挑传，则在 ±35° 范围内均匀采样。对于地滚球传球，该发射方向通常与朝向相距较远。击球标志位和时间的观测方式与射门相同。

奖励。 击球的检测方式与射门相同，另外还需满足球速度在一个控制步内变化超过 1.5m/s。任务奖励为

$$r^{\text{task}} = 0.03\, r^{\text{reach}} + 0.2\left(r^{\text{speed}} + r^{\text{azim}} + r^{\text{elev}}\right) + 0.3\, r^{\text{land}} + 0.3\, r^{\text{settle}} + 0.3\, r^{\text{rec}}, \tag{15}$$

其中， r到达点、 r稳定项和 r恢复项的定义与射门中的相同。对于每个通道，发射过程分别进行评分，其中 r速度= $\exp(-(\Delta s/3.5\ \text{m/s})^2)$、 r方位角= $\exp(-(\Delta\psi/25^\circ)^2)$以及 r仰角= $\exp(-(\Delta\theta/12^\circ)^2)$；其中， Δs、 $\Delta\psi$和 $\Delta\theta$分别是相对于 $\mathbf{v}^*$而言， $\mathbf{v}_{\text{launch}}$ 的速度、方位角和仰角误差。落地点项为 r落地 = $\frac{1}{2}\sum_{\sigma\in\{3,8\}\ \text{m}} \exp(-e^2/\sigma^2)$，其中 e 是使用包含重力、空气阻力和滚动阻力的球模型预测的 $\mathbf{v}_{\text{launch}}$ 与 $\mathbf{v}^*$ 的到达点之间的距离。地滚球传球在球停止的位置到达，挑传则在球首次触地的位置到达。四个发射项在击球前均为零，之后保持其击球时的数值。姿势目标（权重 0.2）的定义与射门中的相同。

初始化。 回合的初始化方式与射门相同，但有三点不同：在 33% 的回合中，球从静止开始；其横向滚动速度至多为 0.68m/s；此外，挑传片段会额外获得 0.3–1.0 m 的向前位移，该位移会以相同方式缩小。

终止 。 当角色摔倒、参考片段结束或经过 3 s.

C.5. 头球

指令。 指令是朝向坐标系中的目标出球速度 $\mathbf{v}^*$ 。速度在 5.5–16.5m/s 范围内均匀采样，仰角在 -32° 至 -5° 范围内采样，因此每次指令头球都会向下。方位角在参考片段接触帧角色朝向周围的 $\pm 110^\circ$ 范围内均匀采样。

奖励。 当球速度在一个控制步内变化超过 3 m/s、球速超过 3 m/s，且考虑身体尺寸后，头部是距离球最近的身体部位并与球的距离在 0.35 m 以内时，即检测到一次头球。因此，与其他身体部位的接触不会触发头球事件。该控制步的球速度被存储为 $\mathbf{v}_{\text{h}}$。任务奖励为

$$r^{\text{task}} = 0.05\, r^{\text{reach}} + 0.1\, r^{\text{hit}} + 0.2\, r^{\text{speed}} + 0.25\, r^{\text{azim}} + 0.15\, r^{\text{elev}} + 0.2\, r^{\text{settle}}. \tag{16}$$

头球前， r达到= $\exp(-3\,[d-0.28]_+^2)$，其中 d 是头部到球的距离。头球后， r击中= 1，出球速度 $\mathbf{v}_{\text{h}}$ 与传球一样，针对每个通道分别进行评分， r速度、 r方位角和 r仰角的容差分别为 4m/s、 40°和 20° 。稳定项的定义与射门相同。姿势目标（权重为 0.2）的定义与射门相同，但颈部、腹部和肩部的权重分别为 3、2 和 1.5，其余部位的权重为 1。

初始化。 对于每个片段，参考状态取自触球前至少 0.3 秒的一帧，在该帧中双脚均低于 0.16m；如果不存在这样的帧，则取第 0 帧。因此，对于包含跳跃的片段，策略必须自行完成起跳。球按照片段自身的传中轨迹运动。球的速度是根据接下来的 10 个参考帧进行考虑重力的最小二乘拟合得到的，并且会移动球的起始位置，使其在无阻力弹道飞行下于参考触球时间抵达参考接触点。

术语 **国家**。 当角色摔倒、头球后 0.8 秒，或经过 2.5 s.

C.6. 停球

指令。 该指令以独热向量的形式指定接球身体部位，即脚或胸部，同时策略还会观测一个成功位。身体部位由发球决定。球以 0.9 米或更高的高度到达角色时，属于胸部球，其首次触球必须由躯干完成。较低的球必须先接触脚或胫部。

奖励。 设 t_a 和 $\mathbf{p}_c$ 分别为发球的计划到达时间和交汇点。任务奖励为

$$\begin{aligned} r^{\text{task}} = {} & 0.3\, r^{\text{app}} + 0.3\, r^{\text{face}} + 2\, r^{\text{touch}} + 0.6\, r^{\text{ctrl}} \\ & + 0.5\, r^{\text{traj}} + 0.2\, r^{\text{stand}} - 0.2\, c^{\text{arm}} - 0.3\, c^{\text{plant}}. \end{aligned} \tag{17}$$

接近项 $r\text{app} = 10\,(0.99\,\phi_t - \phi_{t-1})$ 是一种基于势函数的塑形，其形式为 $\phi = \exp(-(d/0.5)^2)$，其中 d 表示较近的脚到$\mathbf{p}_c$的距离；对于胸部停球，则表示躯干到$\mathbf{p}_c$之外 0.22 m 处的距离。该项持续计入奖励，直到 t_a 前 0.25 s。朝向项 $r\text{face} = \exp(-([\alpha - \alpha_0]_+/25^\circ)^2)$ 从 t_a 前 1 s 起计入奖励，直到首次触球或 $t_a + 0.2$ s，以先到者为准。其中，α 是朝向与来球之间的夹角；脚部停球时 α_0 为对应值，胸部停球时 30° 为对应值。合法首次触球会一次性获得 $r\text{touch} = 1$。触球后 0.25 s，测量一次球的平面速度与来球速度之比 η，以及球的朝向 ψ 相对于触球时角色朝向的角度。从此之后，控制项为

$$\begin{aligned} r^{\text{ctrl}} = {} & \exp\!\left(-([|\psi| - 85^\circ]_+/15^\circ)^2\right) \\ & \cdot \exp\!\left(-([\eta - 0.8]_+/0.2)^2\right) \\ & \cdot \left(0.5 + 0.5\exp(-((\eta - \eta^*)/0.2)^2)\right), \end{aligned} \tag{18}$$

其中 η^*（片段中保留的速度比）对脚部停球为 0.2，对胸部停球为 0.6。若球速低于 1 m/s，则第一项因子设为 1，且 $r\text{ctrl}$ 按一个在触球后 0.3 s 达到 1 的斜坡进行缩放。若 $\eta \le 0.8$，并且 $|\psi| \le 90^\circ$ 或球速低于 1 m/s，则置位成功位；若回合随后失败，则清除该位。

轨迹项 $r\text{traj}$ 跟踪相同球类别（地面、空中球或胸部）片段的膝部与骨盆俯仰轨迹中位数，范围从 0.3 s 之前 t_a 到触球后 0.5 s。它还将朝向球的根节点速度上限设为片段中位数加 0.2 m/s。该项在站立起始时发放，而在片段起始中仅对空中球和胸部停球发放。站姿项 $r\text{stand}$ 对胸部触球后 0.5 s 起的静止直立站姿给予奖励。由于回合在计划到达后 0.5 s 结束，该项在最终阶段很少生效。惩罚项 $c\text{arm}$ 和 $c\text{plant}$ 适用于手高出骨盆超过 0.8 m（脚部停球）或 0.6 m（胸部停球）的情况，以及在空中球脚部接球的 0.5 s 内抬起超过 2.5 cm 或滑动超过 16 cm 的支撑脚。判别器以球类别为条件，其取值为低、高或附近无球。

初始化。 大多数回合从 30 个接球片段中的一个开始，起始时间位于该片段首次触球前 0.2 至 1.2 秒之间，但不会早于片段的起点。片段按照与无球移动相同的难度加权概率进行采样，其中六个胸部接球片段占基础权重的 54%。发球会替换片段中的球，但保留其类别。球会到达片段中角色前方 0.2–1.0 米、侧方最多 0.5 米的随机点（胸部停球：分别为 0.1–0.35 米和 0.2 米），其来球方向位于角色朝向的 $\pm 90^\circ$（$\pm 45^\circ$）范围内。球会在片段规定的触球时间到达；对于从初始朝向的 60° 至 90° 方向飞来的球，到达时间最多延后 0.5 秒。地滚球以 4–12 米/秒的速度滚动。空中脚球以 4–12 米/秒的速度到达，并以 10–60° 的速度下降；胸部停球则以 7–12 米/秒的速度到达，并以 10–45° 的速度下降。其到达高度为片段的触球高度 ±0。25 米；对于空中脚球，该高度限制在 0.3–0.85 米之间，对于胸部停球，则限制在 1.0–1.35 米之间。发射状态通过根据球模型（包含空气阻力和滚动阻力）从交汇点向后积分求得。在 30% 的回合中，角色会改为从 14 个站立片段中的一个开始，并以相等概率接收脚球或胸部停球，

球会在 0.5–1.2 秒后从 ±90° 范围内到达。它会到达角色前方最多 1.5 米、侧方最多 1.0 米的点（胸部停球：分别为 0.8 米和 0.3 米）。发球会偏向较难的案例。约三分之一的发球会以范围外侧的 10° 为方位，20% 的训练重置会重放最近的一次失败，另外 30% 会使用更窄的范围。这些范围是两个成功门控发球课程的顶层范围，最终专家将在此基础上进行微调。

终止。当角色摔倒、手或手臂触碰球，或在 $t_a + 0.5$ 秒前未发生合法首次触球时，回合失败。发生合法首次触球后，回合会在 $t_a + 0.5$ 秒时因超时结束，其数值通过自举法估计。失去球不会结束回合。

C.7. 急停转身

指令。没有指令。离场方向固定为 $\mathbf{u} = -\mathbf{u}_0$，重置时角色朝向的反方向 $\mathbf{u}_0$。

奖励。奖励分为两个阶段，由停球事件分隔。设 d 为较近的脚到球的距离。当球在 $d < 0.35$ 米以内，并且其减速度超过滚动阻力所造成的减速度 5 m/s²时，记录一次制动检查。制动检查后，在满足 $\|\mathbf{b}\| < 1\text{m/s}$、$d < 0.45\text{m}$，且球沿 $\mathbf{u}_0$ 方向位于根节点前方不超过 0.15m 的第一个步，停球指示器 τ 锁定为 1。任务奖励为

$$\begin{aligned} r^{\text{task}} = (1-\tau)\big(&0.5\, r^{\text{trap}} + 0.05\, r^{\text{brake}}\big) \\ &+ 0.25\, r^{\text{check}} \\ &+ g_{\text{exit}}\left(0.4\, r^{\text{exit}} + 0.35\, r^{\text{carry}} + 0.15\, r^{\text{face}}\right). \end{aligned} \tag{19}$$

停球前，$r_{\text{停球}} = \exp(-0.5\,\|\dot{\mathbf{b}}\|^2 - 8\,d^2)$ 会奖励脚边缓慢移动的球，而 $r_{\text{制动}} = \exp(-[\|\dot{\mathbf{p}}\| - 0.6]_+^2 - 2\,[d - 0.8]_+)$ 会奖励在球仍然靠近时减速。每一步中，$r_{\text{制动检查}} = \min(1, [a - 1.47]_+/10)\,\mathbb{1}[d < 0.35]$ 会奖励超过滚动阻力所提供的 1.47 m/s² 的球的减速度 a。停球后，门控 g 退出会在 0.15 到 0.4 秒之间从 0 线性上升到 1，离场项为 $r_{\text{退出}} = \text{clip}(\dot{\mathbf{p}} \cdot \mathbf{u}/1.5, 0, 1)$、$r_{\text{朝向}} = [\mathbf{h} \cdot \mathbf{u}]_+$ 和 $r_{\text{带球}} = \exp(-2\,e_{\parallel}^2 - 4\,e_{\perp}^2)$。其中，$e_{\parallel}$ 和 $e_{\perp}$ 分别是 $\mathbf{b} - (\mathbf{p} + l\,\mathbf{u})$ 沿 $\mathbf{u}$ 方向和横跨 $\mathbf{u}$ 方向的分量，前导量 l 为 $= 0.4 + 0.08\,[\dot{\mathbf{p}} \cdot \mathbf{u}]_+$ m。没有门控时，最容易采取的策略是立即转身，把球留在身后。

初始化。重置状态取自根节点平滑速度达到 2 m/s 的带球片段，时间在首次达到该速度的帧与之后速度降至 0.6 m/s 以下的帧之间均匀选取。六个急停转身片段仅用于判别器数据，占样本的 60%。根节点速度设为 $v_0\mathbf{u}_0$，其中 $v_0 \sim \text{U}[2.5, 4.5]$ m/s；关节速度则按参考速度的 v_0 倍缩放，并限制在 $[0.5, 2.5]$之间。球的起点位于前方 $12\,v_0 \cdot \text{U}[0.55, 1.55]$ m 处，侧向偏移最多为 0.25 m，以 $v_0 \cdot \text{U}[0.7, 1.3]$ 的速度滚动，方向与 ±25° 相差不超过 $\mathbf{u}_0$。

终止。当角色摔倒、球与根节点的距离超过 2.5 米，或经过时间达到 4 秒时，回合结束。

C.8. 起身

指令。没有指令。策略会观测恢复标志，以及除以 3 秒后截断至 1 的经过时间；球的观测值设为零。

奖励。当角色的骨盆连续 0.3 秒保持在 0.75 米以上、躯干向上轴的垂直分量 u_z 高于 0.85，且除脚以外的身体部位均未接触

表 15 | 技能指令和任务目标。每位专家也会获得风格奖励。

技能	指令 $\mathbf{g}^k$	任务目标
无球	方向、速度、朝向	跟踪速度和朝向
带球	方向、速度	跟踪速度，让球保持在前方
射门	发射速度、脚、技术	以指令速度和脚进行击球
传球	发射速度、技术	以指令速度进行地面传球或挑传
头球	发射速度	以指令速度用头部顶传中球
停球	身体部位（脚、胸部）	用该部位接球并使球减速
急停转身	–	停住球，转身，然后把球运回来
起身	–	从摔倒中站起来

地面时，即视为恢复。随后标志位保持开启。任务奖励为

$$r^{\text{task}} = 0.3\, r^{\text{height}} + 0.2\, r^{\text{up}} + r^{\text{prog}} + 0.5\, r^{\text{stand}} - 0.5\, c^{\text{late}} - 0.2\, c^{\text{head}}, \tag{20}$$

当 $r_{\text{高度}} = \min(z_{\text{头部}}/1.3, 1)$ 且 $r_{\text{起身}} = [u_z]_+$ 时。恢复之前，$r_{\text{进度}} = \min\left(20\,[\min(z, 0.85) - z_{\max}]_+, 1\right)$ 奖励仅针对骨盆高度超过其回合最大值 z 的全新纪录 $z_{\max}$，因此上下摇晃不会获得任何奖励。恢复之后，$r_{\text{站立}} = \exp(-2\,\|\dot{\mathbf{p}}\|^2 - 4\,\dot{z}^2 - 0.5\,\|\boldsymbol{\omega}\|^2)$ 会奖励静止的根节点，其中 $\dot{\mathbf{p}}$ 是根节点的水平速度，$\boldsymbol{\omega}$ 是根部角速度。角色尚未在 3 秒后恢复时，惩罚 c_{late} 为 1；头部接触地面时，c_{head} 为 1。判别器数据中不包含摔倒。在从摔倒过程中的某个时刻开始的回合里，风格奖励会被其批次均值取代，直到除脚以外的身体部位累计触地 0.2 秒，或经过 1.5 秒。

初始化。 重置状态取自 584 个片段（27.4 分钟，包含镜像副本）：290 个 LaFAN1 摔倒片段，采样权重为 0.5；290 个 LaFAN1 起身片段，权重为 0.4；以及四个足球起身片段 [activeanimation, 2022] 权重为 0.1。摔倒片段在最后 1.5 秒内采样，并在头部首次降至 0.6 米以下后最多 1 秒处结束，因此角色从摔倒过程中的某个时刻开始，速度沿用片段中的速度。起身片段在前 40% 内采样，此时角色仍在地面上。判别器使用 297 个片段（14.5 分钟）：LaFAN1 起身片段占比 0.75，涵盖俯卧、仰卧、侧卧和跪姿起点；足球起身片段占 0.1；两个 AMASS 躺卧到蹲姿片段占 0.05；以及一个来自 Soccer-X Yang 等人 [Yang等人, 2025] 占 0.1 的站立片段。片段均在角色站起后的任何跳跃动作之前截断。

终止。 恢复之前，摔倒不会结束回合。恢复之后，摔倒则会结束回合。回合也会在 6 秒后结束。

D. 补充实验

D.1. 完整技能转换评估

表 16 比较了全部 11 种训练转换。MESSI 在 11 种转换中的 9 种上取得了最高成功率，平均成功率为 82.3%，而全量微调和 Per-trans. 分别为 73.1% 和 72.5%。例外是运球到射门和头球到无球，Per-trans. 和 SkillX* 在这两种转换上分别表现更好。

D.2. 更多技能转换可视化结果

图9 和10 展示了停止到运球以及起身到运球转换的更多定性比较。所有方法均从相同的交接状态开始评估。对于停止到运球，直接

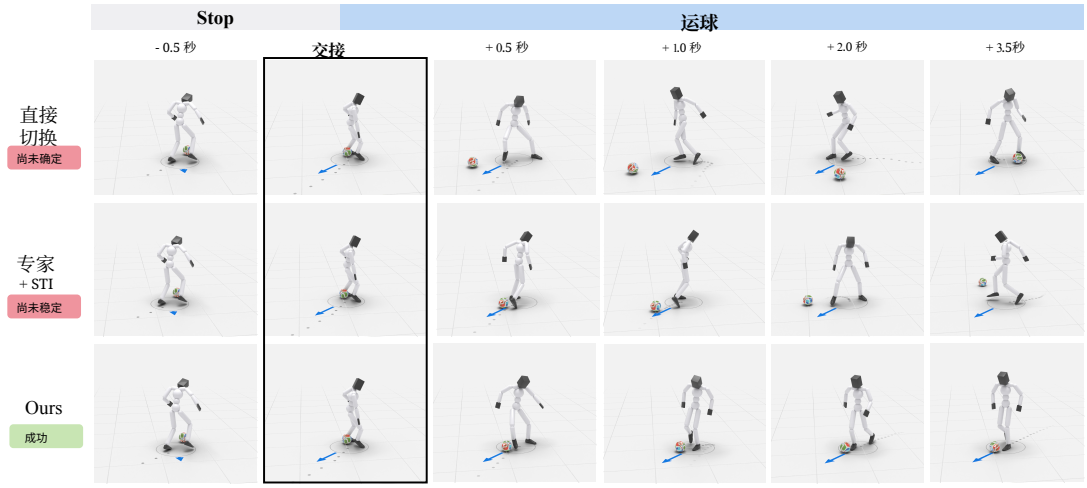


图 9 | 停止→运球，来自同一交接。

切换和专家 + STI 在随后的运球阶段失去控球，而 MESSI 保持控球权并平稳恢复运球。对于起身到运球，直接切换未能建立稳定的运球，而专家 + STI 失去平衡并摔倒。相比之下，MESSI 成功地从恢复转换到受控运球。这些示例进一步说明了适应前序技能生成的状态对于可靠技能组合的重要性。

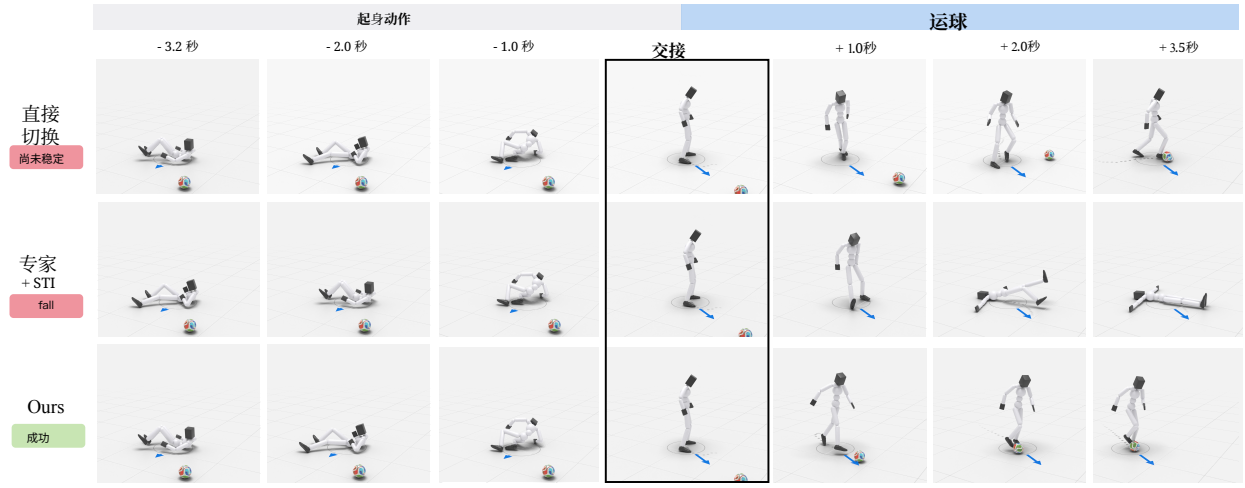


图 10 | 起身→运球，来自同一交接。

D.3. 转换适应消融实验

表 17 评估转换条件化与学习式退出的效果。所有变体在 11 个转换中共享一个适配器，并使用相同的训练预算。条件化在无上下文、技能对标签和完整过渡上下文之间变化。

条件化的影响。 在没有学习式退出的情况下，条件化并不能持续提升转换成功率。目标技能和指令已经编码在学生观测中，这可能会削弱显式条件化带来的额外收益。

表 16 | 完整的技能转换比较。 在全部 11 个训练过的转换中，从相同的前序技能生成的交接状态进行评估，成功率（%）。所有适应基线均在相同的 11 个转换集上训练。

转换	SkillX*	直接-切换	专家 + STI	Full FT	Per-翻译	**Ours**
博 →	2.7	19.8	54.1	57.3	**67.8**	61.8
博 →	17.3	22.2	31.6	35.7	35.1	**38.0**
博 →	75.7	82.4	90.2	95.9	43.5	**99.2**
ST →	95.8	44.1	63.8	96.9	94.0	**97.0**
特 →	47.9	20.7	27.7	49.1	47.8	**50.4**
古 →	91.3	32.6	55.7	91.0	90.8	**91.7**
Mv → Dr	83.9	62.7	59.8	95.0	92.6	**96.3**
Sh → Mv	83.4	69.5	83.1	78.0	82.5	**98.4**
Pa → Mv	83.5	70.2	76.7	72.8	81.9	**96.3**
Gu → Mv	81.4	68.3	72.5	68.3	86.2	**91.6**
Hd → Mv	**87.4**	61.1	69.8	64.5	74.9	84.1
平均值	68.2	50.3	62.3	73.1	72.5	**82.3**

学习式退出的影响。 采用学习式退出进行训练能够显著提升平均成功率，尤其是转换为无球移动时。在没有退出机制的情况下，持续适应通常会导致初始转换后运动能力下降。学习式退出使冻结学生策略能够恢复原生控制。由于仅在测试时禁用退出机制也会产生相似结果，因此性能提升主要源于使用退出机制进行训练，而不是源于在测试时将其停用。

D.4. 蛇形带球分析

我们在杆间距不同的蛇形赛道上评估连续转向和控球能力。在 4.0 米间距下，我们的方法中的学生策略达到 48% 的成功率，相比之下，SkillX* 为 33%，TokenHSI* 为 0%。然而，在 3.0 米和 2.5 米间距的更紧凑赛道上，SkillX* 的表现更好（47%/30%，而我们的方法为 5%/0%）。我们的失败与从带球专家继承的航向偏差和延迟转向响应一致，这表明局限性源于底层技能，而非过渡适应。

D.5. 技能链的其他结果

图 12 展示了长时域技能组合的另外四个定性示例。每个序列都包含多个连续执行的足球技能，在连续仿真过程中不会重置角色或球。叠加显示的角色姿态和球轨迹展现了连续技能转换过程中的协调全身运动与球交互。这些示例通过可视化连续多技能执行，对论文正文中的定量结果进行了补充。

严格踢球标准。 表 18 重复了表 5，但采用了严格的踢球测试。射门时还必须使用指定脚，在 10° 以内发射至指定方位角，并达到至少一半的指定速度。随后，链式成功率受踢球准确度限制，因此每种方法完成的技能链都大幅减少。在这一标准下，我们的控制器不再是最佳方法。全量微调在两条停步转向技能链上领先（85.0% 和 28.7%，而我们的结果为 64.8% 和 18.8%），Per-trans.

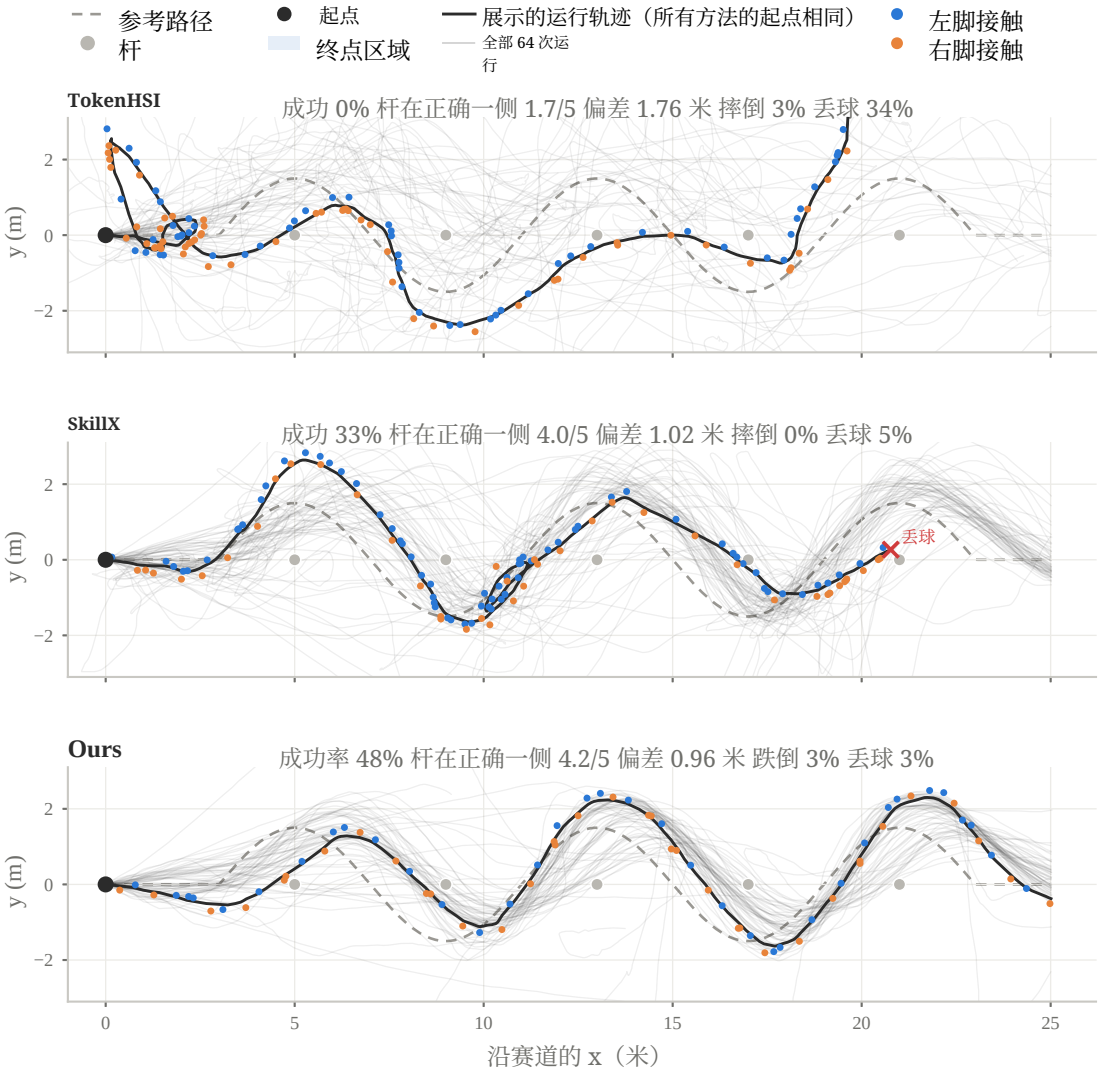


图 11 | 蛇形带球（俯视图）。所有方法都在相同的纯追踪转向指令下，沿着相同的参考蛇形路线（虚线；5 根杆，间距 4 米， ±1.5 米，1.8 米/秒）运行，起始状态均来自相同的 64 个重置状态。每种方法展示的运行轨迹都从相同状态开始；细线表示全部 64 次运行；圆点标示左脚／右脚触地。如果一次运行在未摔倒或丢球的情况下完成，并从每根杆的同一侧通过，则视为成功。

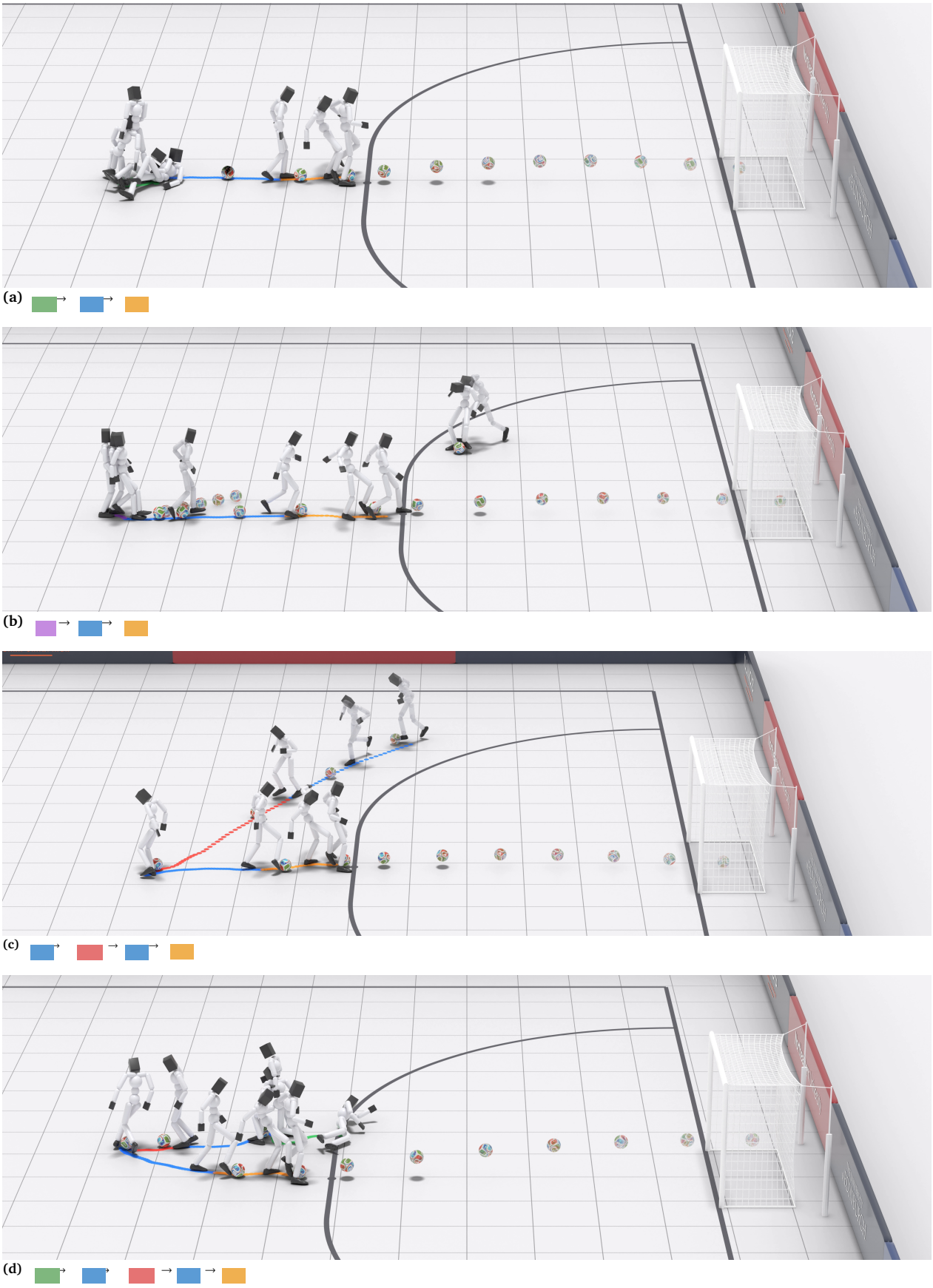


图 12 | 长时域技能组合的其他定性结果。 技能标签为：带球，射门，停球，急停转身，[illegible]身。

表 17 | 过渡适应消融。 在所有 11 种转换中，从相同的交接状态出发测得的任务成功率（%）。每个变体都是一个由所有转换共享的适配器，且使用相同的训练预算进行训练。各变体的区别在于训练时是否采用学习式退出，以及适配器的条件化所读取的上下文：无、源技能和目标技能（成对），或完整上下文$\mathbf{c}_t$。退出头始终读取$\mathbf{c}_t$。绿色／橙色：每行的最佳／次佳结果。

		不使用退出			使用退出	
	直接	无	成对	完整	无	我们的方法
博→	19.8	61.2	53.1	58.4	56.9	**61.8**
博→	22.2	39.4	39.4	**41.6**	35.7	38.0
博→	82.4	91.8	94.3	92.7	95.7	**99.2**
ST→Dr	44.1	96.5	96.7	96.1	**97.3**	97.0
传→	20.7	**51.6**	48.8	49.9	49.1	50.4
古→	32.6	90.5	91.3	90.2	91.1	**91.7**
Mv→Dr	62.7	97.4	**97.6**	97.4	**97.6**	96.3
Sh→Mv	69.5	54.3	44.1	29.6	97.0	**98.4**
帕→	70.2	49.0	42.4	24.0	95.6	**96.3**
Gu→Mv	68.3	44.4	41.8	25.3	91.2	**91.6**
Hd→Mv	61.1	48.1	39.0	21.9	83.2	**84.1**
平均值	50.3	65.8	62.6	57.0	80.9	**82.3**

在以停球或起身开头的技能链上表现更好（分别为 26.8%、15.6% 和 55.3%，而我们的数值为 19.1%、11.3 % 和 35.2%）。在带球后的踢球准确度方面，两者都优于我们的共享适配器。然而，两者在主要标准下都不占优势：全量微调完成以停球或起身开头的技能链的次数少得多，而 Per-trans. 完成停步转向技能链的次数少得多（表5）。SkillX 在主要标准下完成了许多技能链，但在严格标准下几乎无法完成，因此其踢球很少符合指令。我们的控制器在严格比率上的表现与持久适配器相当，因此退出操作不会降低踢球准确度。

E. 局限性与未来工作

局限性。 MESSI 仅适配了 11 种训练过的转换，未见过的技能切换可能会失败。若干训练过的转换仍然具有挑战性，尤其是停球到盘带和盘带到传球。学生策略也继承了其专家的局限性，包括头球能力较弱以及带球响应延迟。

我们的链完成度指标强调连续执行，而非精确的踢球指令；急停转身通过停球而非完整的转身动作进行评估。此外，学习式退出在带球和踢球过程中很少停用适配器。

训练需要特定技能奖励、课程学习以及手动定义的交接条件。添加新技能需要专家训练、学生策略重新蒸馏和适配器重新训练。所有实验均在仿真中进行。

未来工作。 未来方向包括学习高层次的技能选择，将适应扩展到未见过的转换，改进罕见行为，以及支持多人控制和实体人形机器人。

表 18 | 连续链，严格踢球标准。 完整链成功率（%），每个单元格包含 512 条链，如表 5所示，但踢球还必须使用指定脚（射门），在指定方位角的 10° 范围内发射，并达到至少为指定速度一半的速度。各列对应表 5中的链，自上而下读取。绿色／橙色：每条链的最佳／次佳。

	Dr ST Dr Sh	Dr ST Dr Pa	Mv Tr Dr Sh Mv	Mv Tr Dr Pa Mv	Gu Dr Sh Mv	Avg.	Main avg.
专家	6.4	12.5	4.1	5.3	3.3	6.3	32.8
学生策略	4.9	9.2	2.9	5.9	2.9	5.2	32.2
SkillX [Ye 等人， 2026]	0.0	5.5	0.2	1.2	0.0	1.4	39.6
专家 +STI	21.3	16.4	10.9	9.8	18.2	15.3	45.9
全量微调	**85.0**	**28.7**	17.2	8.8	37.1		60.5
Per-trans.	21.5	8.8	**26.8**	**15.6**	**55.3**	25.6	49.8
Per-trans.†	66.4	17.2	**26.8**	**15.6**	**55.3**		74.4
持久化	64.6	20.3	18.9	11.3	35.5	30.2	75.4
Ours	64.8	18.8	19.1	11.3	35.2	29.8	**75.9**

平均值：五个严格比率的非加权平均值；主要平均值：表 5†在主要标准下的相同平均值。Per-trans. 中将其运球→停止适配器替换为我们的方法。只有前两条链包含该转换，因此最后三个单元格重复显示 Per-trans.。