arXiv:2603.01452v1 [cs.AI] 2 Mar 2026

扩展任务，而非样本：通过多任务基于模型的强化学习掌握 HumanoidControl

刘少怀[1*]
德克萨斯农工大学
liushaohuai5@tamu.edu

叶伟睿[2*]
MIT
ywr@csail.mit.edu

杜逸伦[3†]
哈佛大学
ydu@seas.harvard.edu

谢乐[3†]
哈佛大学
xie@seas.harvard.edu

[*]贡献相同 [†]指导贡献相同

摘要

开发能够掌握多种技能的通用机器人，仍然是具身人工智能领域的一项核心挑战。近期研究虽强调扩展模型参数和离线数据集，但这类方法在机器人学领域受到限制，因为学习需要主动交互。我们认为，有效的在线学习应扩展任务数量，而非每项任务的样本数量。这种学习模式揭示了基于模型的强化学习（MBRL）的结构性优势。由于不同任务间的物理动力学保持不变，共享世界模型能够汇集多任务经验，学习稳健且与任务无关的表征。相比之下，当不同任务在相似状态下要求采取相互冲突的动作时，无模型方法会受到梯度干扰。因此，任务多样性可作为 MBRL 的正则化器，提升动力学学习效果和样本效率。我们通过 **EfficientZero-Multitask（EZ-M）**实现了这一理念。这是一种面向在线学习、具有高样本效率的多任务基于模型的强化学习算法。在 **HumanoidBench**这一具有挑战性的全身控制基准上进行评估时，EZ-M 以显著高于强基线方法的样本效率达到了最先进的性能，且无需大幅扩展参数规模。这些结果表明，任务扩展是实现可扩展机器人学习的关键维度。项目网站请见此处。

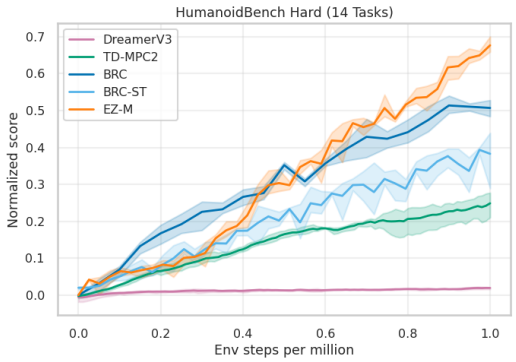


图*1.* **HumanoidBench-困难任务平均归一化得分**。在环境交互次数限制为 100 万次的情况下，EZ-M 达到并超越了强劲的基线方法。所有跑步均使用 3 个随机种子。EZ-M 的表现显著优于所有基线方法。

1. 引言

能够实现稳健泛化与适应的通用智能体，始终是具身人工智能研究的核心目标。受大语言模型（LLMs）成功的启发，近期机器人学的发展趋势大多优先采用基础模型范式——扩展模型参数并汇集海量离线数据集（Reed等人，2022年；Brohan等人，2022年；Kim等人，2024年；Black等人，2024年；O'Neill等人，2024年）。尽管这种方法能够广泛覆盖各类任务，但它面临着机器人学独有的一项根本局限：与被动文本生成不同，具身智能需要与物理环境进行主动交互。仅依赖离线数据，会限制智能体纠正错误、适应动态变化以及通过反馈优化行为的能力。因此，扩展在线学习——提升主动交互的效率与效能——与扩展离线数据同等重要。

我们提出，有效的在线学习需要拓展任务的广度，这一范式称为在线任务扩展，而非仅仅增加每项任务的样本数量。这一转变揭示了基于模型的强化学习（MBRL）的一项关键结构优势。无模型（MF）策略往往会在不同任务要求相似状态下采取相互冲突的动作时遭遇梯度干扰，而 MBRL 则专注于学习底层转移动力学。由于物理定律不随任务而变，共享的世界模型可以汇集多样的多任务数据，从而完善对环境的理解。在此背景下，任务多样性可充当动力学正则化器：它能防止模型对单项任务狭窄的状态分布过拟合，

确保所学到的物理规律能够泛化到全局。因此，MBRL能够有效地将多任务扩展转化为稳健的表征学习。

在这项工作中，我们提出了 EfficientZero-Multitask (EZ-M)，这是一种基于 EfficientZero 架构（Ye等人，2021年）构建的高样本效率多任务强化学习算法。EZ-M 融入了有针对性的架构改进和一致性正则化，以便在在线设置中有效利用动力学不变性。我们在 Humanoid-Bench（Sferrazza等人，2024年）上评估了我们的方法。该领域具有高维特性，对复杂密集接触动力学进行建模是其中的主要瓶颈。

我们的实证结果表明，EZ-M 不仅达到了最先进水平（SoTA）的性能，还展现出鲜明的正迁移特性：随着任务基数增加，无模型（MF）基线方法的性能会下降或停滞，而 EZ-M 则利用任务多样性提升样本效率。这表明机器人学习存在一条可扩展的发展路径：除了完全依赖大规模离线数据集之外，我们还设想一种大规模并行在线学习范式，让多样化任务协同构建统一、通用的世界模型。我们的主要贡献如下：

- 我们提出 EfficientZero-Multitask（EZ-M），通过新颖的一致性机制缓解任务间的负干扰，将基于模型的强化学习（MBRL）的样本效率扩展到多任务在线设定。
- 我们提供了理论分析，阐明基于模型的强化学习（MBRL）的结构性优势。我们表明，尽管无模型（MF）目标在多任务设定中会遭遇梯度冲突，MBRL 却能受益于动力学不变性；随着任务数量增加，其样本效率渐近地更优。
- 我们在综合性基准 HumanoidBench 上取得了最先进的性能，表明 MBRL 中的在线任务扩展为高维机器人控制提供了一条可行途径。

2. 相关工作

人形机器人控制。由于欠驱动和接触复杂性，人形机器人控制仍是一项严苛的基准挑战。以往工作涵盖基于模仿的跟踪（Penget al.，2018年）和对抗强化学习（Peng et al.，2021年），并常借助 GPU加速仿真（Makoviychuketal.，2021年）；HumanoidBench（Sferrazza et al.，2024年）等综合性基准凸显了在线学习在稳定性和样本效率方面持续存在的挑战。

多任务学习。为缓解诸如梯度干扰（Yu等人，2020年）和异构奖励尺度（Hessel 等人，2019年），已有工作采用了从梯度投影（Yu等人，2020年；Chen等人，2018年）到蒸馏（Teh等人，2017年）的多种技术。并行研究利用高容量视觉-语言-动作（VLA）模型（Kim等人，2024年），将监督学习或离线强化学习应用于大规模遥操作数据集（O'Neill等人，2024年）以实现泛化。有别于这些资源密集型范式，我们优先采用**样本高效的在线学习**。我们并不依赖大规模模型扩展或预先收集的数据集，而是利用紧凑的共享动力学模型进行基于模型的强化学习。我们证明，动力学与价值学习中的结构正则化能够有效替代单纯的规模，从而在有限的参数预算内实现稳定、任务可扩展的在线强化学习。

基于模型的强化学习。以往研究证实了通过学习动力学进行规划的有效性，方法涵盖经验增强（Sutton，1991）和潜在空间展开（Hafner等人，2019年；2023年），到基于搜索的轨迹选择（Hansen 等人，2022年；Ye等人，2021年）以及价值重估计（Schrittwieser 等人，2020年；Wang 等人，2025年）。近期进展还进一步融合了扩散模型（Hafner等人，2025年）。然而，这些创新主要着眼于单任务性能。有限的多任务基于模型的强化学习文献仅限于模型预测控制方法（Hansen 等人，2023年；2025年），而在线多任务学习中蒙特卡洛树搜索的潜力仍未得到探索。

3. 预备知识

多任务强化学习（MTRL）旨在针对一组任务 $\mathcal{T}$ 训练单个智能体。每个任务 $\tau \in \mathcal{T}$ 引入一个马尔可夫决策过程 $\mathcal{M}_\tau = (\mathcal{S}_\tau, \mathcal{A}_\tau, P_\tau, r_\tau, \gamma)$，其折扣因子 γ 共享，但状态空间 $\mathcal{S}_\tau$、动作空间 $\mathcal{A}_\tau$、状态转移 $s_{\tau t+1} = P_\tau(s_\tau, t, a_\tau, t)$ 和奖励函数 r_τ 则因任务而异。我们学习一个任务条件策略 $\pi_\theta(a|s, \tau)$，使在 $\mathcal{T}$ 上取平均的预期回报最大化，其数学定义为 $J(\theta) = \mathbb{E}_{\tau\in\mathcal{T}}\left[\mathbb{E}_{\pi_\theta(\cdot|s,\tau)}\left[\sum_{t\geq 0}\gamma^t r_\tau(s_\tau, t, a_\tau, t)\right]\right]$ 与单任务强化学习相比，MTRL 必须通过共享表示拟合异质的学习信号，这可能导致梯度冲突、负迁移和不稳定，尤其是当共享评论家面对不同的奖励价值尺度时。实际训练通常通过任务均衡采样和任务条件机制混合来自多个任务的经验，同时借助归一化或其他稳定化技术控制跨任务干扰。

蒙特卡洛树搜索和 Gumbel 搜索 是 MuZero 系列和 Effi- cientZero 系列中用于生成改进策略和筛选动作的两种树搜索技术。MCTS(Coulom，2006年；Kocsis & Szepesv´ari，2006年) 通过迭代扩展叶节点并递归更新搜索树中节点的价值来构建搜索树，沿着

下潜路径前进，每次选择上置信界分数最高的节点向下遍历（Auer等人，2002年）。MCTS 根据根节点下潜路径的访问计数分布生成改进策略，并将访问次数最多的动作作为筛选动作。因此，MCTS 通常需要大量节点扩展才能生成稳定的改进策略，而且无法并行化也会导致计算效率问题（Ye等人，2022年）。相比之下，Gumbel 搜索（Danihelka 等人，2022年）利用 Gumbel 无放回采样（Kool等人，2019年）和序贯减半（Karnin 等人，2013年）迭代地二分动作空间。它用 Gumbel 分数取代基于计数的上置信界来选择节点，而 Gumbel 分数更加依赖价值估计。这些改动不仅减少了节点扩展，更重要的是保证了策略改进。

4. 方法

为了跨多个任务训练强化学习智能体，算法应满足：（1）随任务数量增加而扩展；（2）能够适应随任务变化的观测空间与动作空间；（3）对随任务变化的奖励与价值尺度具有鲁棒性；（4）在训练和推理方面均具有计算效率。EZ-M 建立在 EfficientZero-v2 的成功之上，后者已被证明在具有视觉/本体感觉输入以及连续/离散动作的任务上兼具样本效率与计算效率。我们将 EfficientZero-v2 扩展到在线多任务强化学习，并通过如下改进设计来稳定多任务学习。我们的方法可总结于图2。

4.1. 模型、搜索与学习

模型组件。EZ-M通过联合进行时间一致性建模、分类奖励/价值预测以及真实/想象时序差分学习，在隐状态空间中对候选动作进行规划。与Dreamer系列方法（Hafner等人，2019年；2020年；2023年）不同，这些方法经过训练以重构原始输入，而EZ-M不使用重构损失，使模型学习始终专注于控制目标。与 EfficientZero-v2不同，EZ-M采用独立的奖励预测路径，该路径与状态转移预测相同，而非在下一状态预测之后连接奖励预测头 $r_t = \mathcal{R}(h_{t+1})$。为避免任务间变化的动作维度和尺度造成干扰，动作编码器会将不同任务的动作投影到同一动作嵌入空间，从而稳定共享任务的模型训练。具体而言，EZ-M架构由六个经过学习的子网络组成：

$$\begin{aligned} h_{\tau,t} &= \mathcal{H}(s_{\tau,t}), & u_{\tau,t} &= \mathcal{U}(a_{\tau,t}), \\ \hat{h}_{\tau,t+1} &= \mathcal{G}(h_{\tau,t}, u_{\tau,t}), & \hat{r}_{\tau,t} &= \mathcal{R}(h_{\tau,t}, u_{\tau,t}), \\ \hat{v}_{\tau,t} &= \mathcal{V}(h_{\tau,t}), & p_{\tau,t} &= \mathcal{P}(h_{\tau,t}), \end{aligned} \tag{1}$$

其中，τ 表示任务标识符，t 表示时间步。变量$h_{\tau,t}$和$u_{\tau,t}$分别对应从原始状态$s_{\tau,t}$和动作$a_{\tau,t}$中得到的隐状态表示和动作嵌入。动力学网络 $\mathcal{G}$ 以自回归方式预测后续潜在状态$\hat{h}_{\tau,t+1}$，而$\hat{r}_{\tau,t}$和$\hat{v}_{\tau,t}$则估计即时奖励和价值函数。

利用模型先验进行搜索。我们使用上述模型输出作为先验信息来引导树搜索，并应用 Gumbel 技巧（Danihelka 等人，2022年；Wang 等人，2024年）。
在根节点处，我们使用模型输出计算每个根节点下降在动作 a 下的 Gumbel 分数 $g(a)$，其中

$$\begin{aligned} g(a) &= p(a) + \sigma(\hat{q}(a)) \\ \hat{q}(a) &= \mathbb{E}_{\{s,a\}\sim\mathcal{G},\mathcal{P}}\left[\sum_{t\geq 0}^{H}\gamma^t\hat{r} + \gamma^H\hat{v}\right] \end{aligned} \tag{2}$$

其中 σ 是一个遵循（Danihelka 等人，2022年）的单调函数。 $\{s,a\}$ 表示在节点扩展预算有限的情况下，按照动态网络和策略网络的预测得到的搜索树中所有状态-动作集合。 H 表示节点扩展过程中每条潜入轨迹的长度。 $\hat{q}(a)$ 是沿潜入路径的自举价值估计的访问次数平均值，并使用最小-最大归一化将其归一化到 [0,1]。顺序减半（Karnin 等人，2013年）用于在对剩余下降路径进行均等扩展后，迭代地将根动作候选集一分为二，定义如下

$$\mathcal{A}_n = \underset{|\mathcal{A}_{n-1}|/2}{\text{argtop}}\ g(a) \tag{3}$$

然后，最终将通过以下方式生成筛选后的动作 A 和改进后的策略 π

$$A = \underset{a\in\mathcal{A}_{\text{remain}}}{\arg\max}\ g(a) \tag{4}$$

$$\pi := (\mathcal{A}, \text{softmax}(g), A) \tag{5}$$

在非根节点处，动作选择遵循提出的策略 p，以避免意外的探索噪声最终影响根节点价值估计。这带来一项原则：访问次数与策略先验相匹配。

策略学习。与基于模型的演员-评论家强化学习算法中所采用的最大Q策略梯度不同（Hansen 等人，2022年；2023年），我们使用监督学习，让策略头输出 p，以模仿改进策略 π，该策略由树搜索生成。这里的一个关键问题是，EZ系列是基于状态价值的，因此应用最大Q学习并不自然。一项相关工作表明，这种对改进策略的蒸馏可以在TD-MPC框架中带来性能提升（Wang 等人，2025年）。

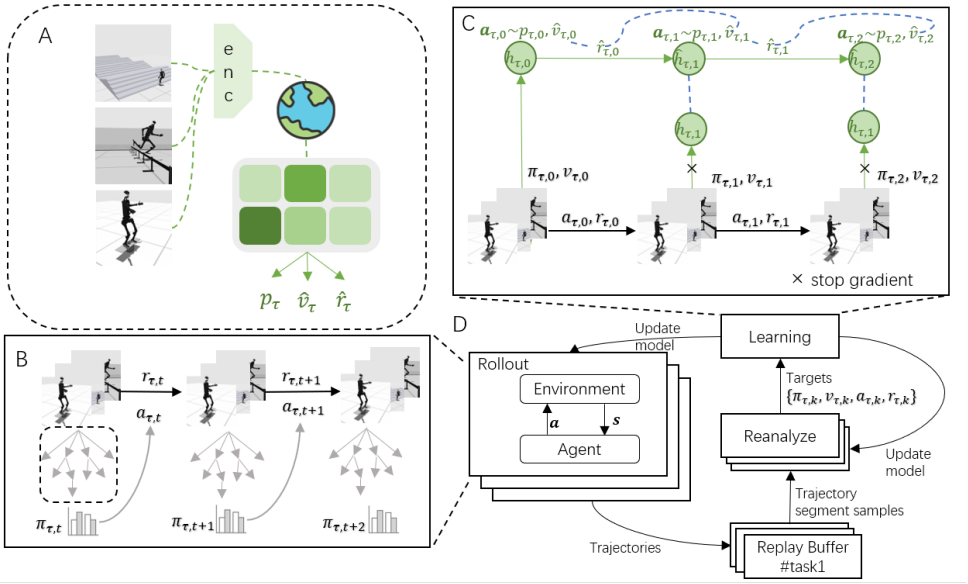


图 2. EfficientZero-Multitask（EZ-M）概览。绿色箭头和标注表示模型推理与预测。黑色标注和粗体箭头表示目标/真实价值与转移。(A) 任务共享模型架构。EZ-M 使用共享模型，在训练多个任务时预测策略、价值和奖励，并以任务嵌入作为各组件的条件。(B) 均衡数据收集。展开工作器按均衡的计划采样任务，并执行相应的任务策略与环境交互。这可以防止因数据不平衡导致学习崩溃，并随着时间推移保持多样化的任务覆盖。(C) 多任务展开学习。对于每个采样的轨迹片段，模型会在以任务索引为条件的潜在空间中展开，在想象展开过程中依次生成 $\hat{r}_{\tau,k}$、$\hat{v}_{\tau,k}$ 和 $\hat{p}_{\tau,k}$。监督信号来自真实转移（$a_{\tau,k}$,$r_{\tau,k}$）和重新分析的目标（$\pi_{\tau,k}$, $v_{\tau,k}$）；交叉熵损失使单个网络能够学习跨任务一致的预测。蓝色虚线表示时间一致性和路径一致性。(D) 分布式实现。数据收集（展开）将轨迹送入与任务无关的回放缓冲区，而重新分析会定期使用最新模型重新计算基于搜索的目标。学习器使用重新分析后的批次更新参数，并将更新后的权重广播回展开/重新分析工作器，以实现可扩展的异步训练。

我们使用完全重新分析，对经验回放中每个采样状态应用树搜索，以生成目标策略，因为EZ-M遵循离策略强化学习设定。具体而言，改进策略 π 由采样动作候选集 $\mathcal{A}$、其对应的Gumbel 分数 g以及一个经过筛选的动作 A组成。我们发现，一个简单的最大似然估计（MLE）损失，定义为

$$\mathcal{L}_p = -\log p(A) \tag{6}$$

在实践中能够产生更加稳定的性能，尤其是在具有高维动作空间的任务中。

奖励预测。在任务间奖励尺度差异显著的情况下，准确预测环境奖励是多任务基于模型的强化学习中的一项关键挑战。由异构奖励尺度引起的任务间 Q 值尺度变化问题，也给多任务无模型强化学习带来了障碍（Hessel 等人，2019年；Nauman 等人，2025年）。不同于采用类似研究中的复杂奖励归一化方法，我们在奖励预测中仅使用分类建模（Schritwieser 等人，2020年），而不进行归一化或参数扩展。（Bellemare 等人，2017年）还报告称，分类建模中的交叉熵损失能够产生更加均衡的梯度和损失尺度，因为分类损失对尺度变化并不敏感。

价值学习。精准预测价值对于我们的离策略、基于价值的强化学习算法至关重要。因此，我们对价值预测进行了大量改进。由于我们没有使用奖励归一化，因此在价值估计中继续采用分类建模；事实证明，这种方法在多任务行动者-评论家强化学习中能够有效预测随任务变化的价值（Nauman 等，2025）。我们还采用了 Hansen 等人（2023）提出的价值集成，以稳定价值预测并缓解过高估计。多个价值头也改进了 Wang 等人（2024）提出的混合价值目标。该方法在从回放缓冲区采样的过时转移上，使用基于模型的价值估计，即重新分析树搜索的根节点自举价值。（Wang 等人，2024）采用基于规则的方法来确定何时使用基于模型的目标，以及如何界定过时转移。我们可以根据多个价值估计得出价值预测方差；这一方法此前用于设计内在奖励并增强探索（Osband 等人，2016；Pathak 等人，2017）。现在，我们将价值方差用作时序差分目标与基于模型的目标之间的权重平衡，其定义为 $v_{\text{mix}} = \alpha v_{\text{td}} + (1-\alpha)$ v_{model}，其中权重系数 $\alpha = (\text{Var}(\hat{v}) - \text{Var}(\hat{v}))/(\text{Var}(\hat{v}) - \text{Var}(\hat{v}))$。$\text{Var}(\hat{v})$ 和 $\text{Var}(\hat{v})$ 分别是经验价值方差的最大值和最小值。随着价值方差增大，尤其是在那些新鲜转移中，算法应更多地依赖

时序差分目标。这些过时转移的方差相对较低，因此我们可以更多地使用基于模型的目标。

潜在状态上的时间一致性。为稳定多步展开，并缓解想象过程中的表征漂移，我们在潜在空间中施加时间一致性正则化。直观而言，这一约束会使预测潜在状态（由动力学生成）与编码潜在状态（直接从相应的未来观测中得出）保持对齐。通过促使这两条路径收敛，我们确保想象轨迹始终锚定于真实数据所诱导的表征。具体来说，我们最小化预测转移与目标表征之间的投影距离（例如，负余弦相似度（Chen 和 He，2021年）），并对目标应用停止梯度，以防止坍塌。对于多任务学习而言，这种正则化至关重要，因为它可以防止动力学模型在多样化任务混合所引入的分布偏移下发生偏离。

基础训练目标。总而言之，我们将整体训练目标定义为奖励预测损失、策略（或经搜索改进的策略）损失、价值预测损失，以及在 K 个展开步上的潜在一致性正则化之和：

$$\mathcal{L}^{\text{base}} = \sum_{k=0}^{K} \Big(\mathcal{L}_{\text{CE}}(r_k, \hat{r}_k) + \mathcal{L}_p(\pi_k, p_k) + \mathcal{L}_{\text{CE}}(v_k, \hat{v}_k) + \mathcal{L}_{\text{SimSiam}}(h_k, \hat{h}_k) \Big) \tag{7}$$

其中，r_k，π_k，v_k 表示展开步k 处的训练目标（例如自举的 n 步奖励/价值目标和搜索策略目标），而 $\hat{}r_k$，p_k，$\hat{}v_k$ 则是相应的网络预测。项 $\mathcal{L}_{\text{SimSiam}}(\cdot)$ 汇总了对潜在动力学/表征的时间一致性正则化。这个统一目标使 EZ-M 能够从回放轨迹中端到端地学习任务条件世界模型以及与规划器对齐的策略/价值预测，同时保持稳定的潜在展开，这对有效规划至关重要。

4.2. 路径一致性

Wang 等人提出的方法（2024年）在损失设计上的一个关键局限是，奖励和价值预测器通过独立监督进行训练，缺乏对想象展开期间相互一致性的明确约束。因此，学习得到的价值 $\hat{}v(h_k)$ 常常偏离其递归分解 $\hat{}r(h_k, u_k) + \gamma\hat{}v(h_{k+1})$。这种差异会随展开步数不断累积，导致树搜索过程中估计误差不断加剧。在多任务设定下，这种不一致性会进一步恶化，因为共享模型必须同时泛化到异质的任务动力学和不同的奖励尺度。

为解决这一问题，我们引入路径一致性正则化器，以在异构奖励尺度下稳定多任务价值学习，其灵感来自（Farquhar等人，2021年）。对于任务 τ，给定一个 K-步轨迹片段 $(s_k, a_k, r_k, s_{k+1})_{k=0}^{K-1}$ 从相应的回放缓冲区中采样，我们最小化以下目标

$$\mathcal{L}^{pc} = \sum_{k=0}^{K-1} \mathcal{L}_{\text{CE}}\left(\text{sg}\left(\hat{r}(h_k, u_k) + \gamma\hat{v}(h_{k+1})\right), \hat{v}(h_k)\right) \tag{8}$$

停止梯度算子 sg(·) 可防止奖励与价值学习之间产生不稳定的目标耦合，因为路径一致性损失不会更新奖励预测网络。这一特性令人满意，因为网络可以专注于跨任务的挑战性奖励预测。最终的训练目标可以表示为对各任务求和：

$$\mathcal{L} = \sum_{\tau \in \mathcal{T}} \mathcal{L}_\tau^{\text{base}} + \mathcal{L}_\tau^{\text{pc}} \tag{9}$$

4.3. 多任务设计选择

可学习任务嵌入。为使 EZ-M 适用于多任务学习，我们引入可学习任务嵌入，根据任务身份显式地对智能体的所有主要组成部分进行条件化（Hansen 等人，2023年）。具体而言，每个任务索引都映射为一个可训练向量，并通过拼接注入表征/动力学/预测模块。这一设计以轻量的方式实现任务特定特化，同时让大部分参数保持共享；当不同任务具有不同的转移模式或相似模式时，这一点尤为重要。在实践中，这种嵌入使网络能够形成任务感知隐状态以及价值/奖励预测，无需为每个任务设置独立的输出头；同时，它还提供了一个清晰的接口，便于通过学习新的嵌入并复用共享主干网络来扩展到新任务。

动作掩码与观测填充。多任务环境的动作空间大小以及本体感觉控制中的观测维度往往各不相同，这可能导致单一统一的网络接口无法正常工作。我们通过使用动作掩码和观测填充，对不同任务的输入和输出进行标准化。对于动作，我们维护一个全局动作空间，并在策略评估和规划期间应用逐任务二值掩码，使非法动作无效。对于观测，我们使用全零张量将任务特定观测填充至固定的最大维度。这种简单的统一方式避免了临时的架构分支，保留了向量化训练，并确保以明确、安全的方式处理各任务之间的差异（不会选择无效动作，填充后的输入也不会泄露虚假信号）。

独立经验回放。 为缓解任务不平衡，并尽可能减少由异质数据分布导致的负迁移，我们采用独立的

经验回放策略。在标准的单体缓冲区中，奖励量级较高或动力学较简单的任务往往会主导优先采样机制。相比之下，我们为每个任务维护专用的回放缓冲区，从而能够显式控制任务间的采样比例。这种解耦确保复杂任务或稀疏奖励任务不会被占主导地位的任务边缘化。此外，这种设计简化了架构：采样同质批次有助于保持任务条件化的一致性，并简化任务专属动作掩码的管理。

5. 理论分析

我们形式化了基于模型的强化学习（MBRL）在在线任务扩展方面的结构性优势。我们表明，尽管无模型（MF）基线方法在相关任务间的梯度相似度较低，如图3所示，EZ-M 将任务多样性用作动力学正则化器，从而获得更优的渐近样本效率。

5.1. 问题形式化

考虑共享同一物理环境的多任务设定 $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$。我们将全局状态空间$\mathcal{S}$ 和动作空间 $\mathcal{A}$ 定义为所有任务中可达状态与动作的并集（或一个共享的潜在流形）。尽管各个任务可能只使用这些空间的子集，物理转移动力学 $P(s'|s,\ a)$ 仍保持平稳。每个任务 τ_k 由唯一的奖励函数 $r_k(s,\ a)$ 定义。在 EZ-M 中，我们通过 $h = \mathcal{H}(s)$ 和 $u = \mathcal{U}(a)$ 将这些映射到共享潜在空间。

5.2. 干扰与不变性

命题 5.1 (无模型学习中的梯度干扰 (Yu 等人，2020年))。 设 θ 为无模型策略的共享参数。如果不同任务 τ_i，τ_j 的最优策略在同一潜在状态下要求采取互不相交的动作，则其梯度的期望余弦相似度为负：

$$\mathbb{E}\left[\langle \nabla_\theta \mathcal{L}_{\tau_i}, \nabla_\theta \mathcal{L}_{\tau_j} \rangle\right] < 0. \tag{10}$$

备注：在标准无模型多任务学习中，这要求进行梯度投影（例如 PCGrad），或大幅扩展参数规模，以分离相互冲突的任务子空间。

引理 5.2 (动力学不变性)。 物理转移动力学 P 在各个 $\mathcal{T}$ 之间保持不变。因此，存在唯一的最优参数集 ϕ^*，对应的动力学模型 $\mathcal{G}_\phi$ 能够同时最小化所有任务的预测误差：

$$\phi^* \in \bigcap_{k=1}^{K} \arg\min_{\phi} \mathcal{L}_{dyn}^{(k)}. \tag{11}$$

备注：与策略学习不同，动力学学习不存在相互冲突的目标。尽管不同任务可能要求在同一状态下采取相互矛盾的动作（例如，任务 A 要求“左转”，而任务 B 要求“右转”），但任何动作所产生的物理结果都是客观的，且不受任务意图影响。因此，任务多样性实际上起到了正则化器的作用，促使表征捕捉环境的因果结构，而非任务特定相关性。

5.3. 样本复杂度分析

设 d_{dyn} 和 d_{rew} 分别表示学习转移动力学和单个奖励函数所需的样本复杂度。

假设 1（动力学主导）。 在高维机器人控制中，学习物理动力学所需的样本量远高于学习特定任务奖励所需的样本量：$d_{\text{dyn}} \gg d_{\text{rew}}$。

备注：这与无奖励探索领域的理论发现一致（Jin 等人，2020年；Agarwal 等人，2020年），这些研究表明，学习转移算子决定了极小极大样本复杂度（随 $O(S^2A)$ 增长），而奖励估计的成本相对较低（随 $O(SA)$ 增长）。

定理 5.3（渐近任务扩展效率）。 设N_{MF} 和 N_{MB} 分别表示掌握 K 个任务所需的总样本复杂度。在无模型方法存在最坏情况干扰、而基于模型方法具有共享动力学的条件下，每任务样本复杂度满足：

$$\lim_{K\to\infty} \frac{N_{MB}(K)}{K} = d_{rew} \ll \lim_{K\to\infty} \frac{N_{MF}(K)}{K} = d_{dyn} + d_{rew}. \tag{12}$$

证明。无模型智能体会隐式地将动力学学习与奖励学习纠缠在一起。由于梯度干扰（命题5.1），它们无法有效共享结构知识，因此将每个控制问题都作为独立实例来解决：$N_{\text{MF}} \propto K(d_{\text{dyn}} + d_{\text{rew}})$。相比之下，EZ-M 将这两个过程解耦。不变动力学 $\mathcal{G}$ 只需学习一次（引理5.2）。对于任何新增任务 $k > 1$，智能体都可以利用成熟的世界模型，只需付出学习奖励头的成本。因此，$N_{\text{MB}} \approx 1 \cdot d_{\text{dyn}} + K \cdot d_{\text{rew}}$。将其除以 K 并取极限即可得到结论。

□

6. 实验

本节将介绍用于评估我们所提出的 EZ-M 方法的实验。这些实验旨在回答以下问题：（1）EZ-M 的性能与主流基线相比如何？（2）随着同时训练的任务数量增加，模型性能如何变化？（3）世界模型的各个组件如何在任务之间共享知识？（4）所提出的方法（如路径一致性、任务嵌入和独立经验回放）的必要性是什么？

这些问题的答案将在以下小节中列出，我们的实验的更多细节和曲线见附录。A。

6.1. 实验设置

基准测试。我们在 HumanoidBench（Sfer- razza 等人，2024年）上评估 EZ-M。HumanoidBench 是一套具有挑战性的全身人形机器人控制任务，要求机器人在密集接触动力学条件下协调数十个自由度。与充斥着基线方法的标准运动基准测试（如 DMControl（Tassa 等人，2018年））相比，HumanoidBench 包含楼梯、跨栏和迷宫等困难任务，这些任务仍存在相当大的待解决差距。遵循（Nauman 等人，2025年），我们考虑两个难度等级：**中等**和**困难**。中等难度采用 H1-* 设置（不包含手部控制），涵盖 9 项运动任务：站立、行走、跑步、楼梯、爬行、滑行、杆上行走、跨栏和迷宫。困难难度采用 H1hand-*（包含手部控制），涵盖 14 项任务：上述 9 项运动任务，以及简单坐姿、sit-hard、简单平衡、困难平衡和伸手。两种设置的环境步数均限制为 100 万。不使用动作重复。

基线方法。 我们将 EZ-M 与涵盖单任务（ST）和多任务（MT）场景，以及无模型（MF）和基于模型（MB）方法的成熟基线进行比较。主要基线包括 **TD-MPC2** (ST，基于模型)（Hansen 等人，2023年）、**DreamerV3** (ST，基于模型)（Hafner等人，2023年）、**MH-SAC** (MT，无模型)（Yang等人，2020年），以及 **BRC**(MT，无模型)（Nauman 等人，2025年）。在可行的情况下，基线分数直接取自已发表结果；否则，我们报告重新评估官方代码所得的结果。值得注意的是，BRC 依靠一个庞大的十亿参数模型来实现其性能。相比之下，EZ-M 采用紧凑架构，仅包含 1600 万参数，这表明结构化世界模型无需蛮力扩展参数规模，也能实现更高的样本效率。

评估指标。 为评估总体性能，我们报告归一化分数，具体计算方法详见附录 B。此外，我们在图 6和图 7（附录A）中提供了完整的训练曲线，展示原始回报随环境步数的变化。

6.2. 性能分析与知识共享

EZ-M 在具有挑战性的 HumanoidBench 套件上实现了最先进水平（SoTA）的性能。如图 6所示，与基线方法相比，EZ-M 在 9 个中等难度任务中展现出更高的样本效率。此外，在维度更高的设置中，EZ-M 依然保持优势，在 14 个接触密集型 H1Hand 任务上的表现优于竞争方法（图7）。从量化结果来看，在中等难度设置下，EZ-M 在 9 项任务中有效解决了 7 项或取得了 SoTA 结果（表 1）；在困难设置下，14 项任务中有 10 项达到这一水平（图 7），凸显了其在复杂控制领域中的鲁棒性。

梯度相似性分析。为探究其内在机制，我们分析任务间的梯度相似性，将 EZ-M 的世界模型组件（动力学、奖励、策略/价值）与 BRC 的演员/评论家进行比较。我们选择（*h1-walk*，*h1-run*）作为相关任务对以及（*h1-walk*，*h1*-爬行）作为关系较远的一组。如图3所示，EZ-M 对相关任务对始终保持较高相似性（颜色较深的橙色线条），同时对关系较远者保持较低相似性（颜色较浅的橙色线条）。这表明 EZ-M 能够有效利用正迁移，同时缓解干扰。相比之下，即便对于相关任务，BRC 的相似性也始终较低，说明基于模型的架构可通过共享动力学实现更优的知识迁移。

失败案例分析。尽管 EZ-M 在运动和动态操作任务中占据主导，但我们观察到在静态任务（如*sit-hard*）上，其性能相较 BRC 存在差距（表2）。我们推测，坐姿中复杂的自身接触动力学更难被世界模型准确预测，这为未来在接触建模方面的改进留下了空间。

计算效率。尽管蒙特卡洛树搜索规划本身存在开销，EZ-M 仍通过异步分布式架构（Ye等人，2021年）将轨迹生成与模型优化解耦，从而保持较高的效率。对于中等任务套件，EZ-M 在 2 块 NVIDIA A40 GPU 上约需 **10 小时**。相比之下，最强的基线方法 BRC 要在单块 A40 GPU 上耗费超过 **40 小时**才能达到相同的步数。这些结果表明，EZ-M 的并行化设计有效抵消了规划成本，使研究迭代速度显著快于顺序执行的无模型基线方法。

6.3. 随任务数量扩展的性能

我们进一步研究 EZ-M 随并发任务数量变化的扩展特性。我们评估了任务数量分别为 $N \in \{1, 4, 9\}$ 的训练配置，其中 $N = 1$作为单任务基线。对于$N = 4$设置，我们将任务划分为两个互不相交的子集： {*h1-stand*、*h1-walk*、*h1-run*、*h1-stair*} 和 {*h1-crawl*、*h1-pole*、*h1-slide*、*h1-hurdle*}。为严格分离多任务扩展带来的优势，我们报告所有设置下这 8 项任务的平均性能。如图 4所示，在总环境步数预算相同的情况下，EZ-M 呈现出正向扩展定律：性能随着任务数量增加而持续提升。这一趋势从经验上验证了多任务在线强化学习作为提升样本效率的一种机制的有效性。

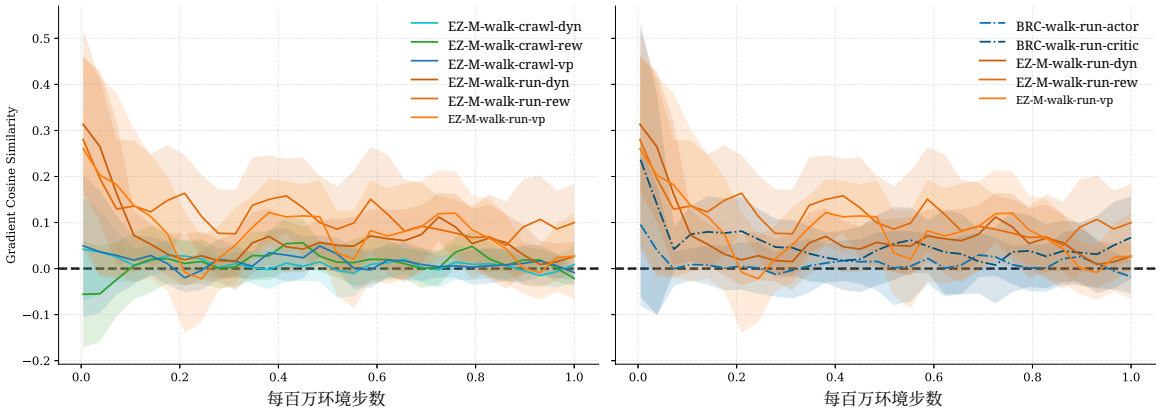


图*3*。 **训练过程中的梯度相似度**。我们比较了 BRC 和 EZ-M 在不同模型模块上的梯度相似度。我们选择了两对任务（*h1-walk*、*h1-run*）和（*h1-walk*、*h1-crawl*），分别代表相关任务和不相关任务。*dyn*、*rew* 和 *vp* 分别代表 EZ-M 中的动力学、奖励和价值策略模型。**（左）**我们验证了在整个训练过程中，相关任务对之间的梯度相似度高于不相关任务对，表明存在积极的知识迁移。**（右）**我们展示了在相关任务对中，EZ-M 模块的梯度相似度高于 BRC 模块。为便于观察，对曲线进行了轻微平滑处理。

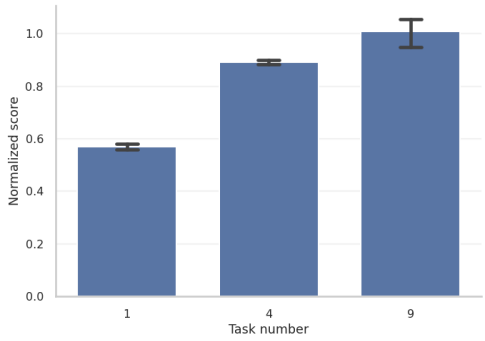


图*4*。 **性能随任务数量的扩展**。 *y*轴表示任务平均的归一化回合回报，*x*轴表示同时参与训练的任务数量。我们使用了 3 个不同的随机种子，并报告了 95% 的置信区间。

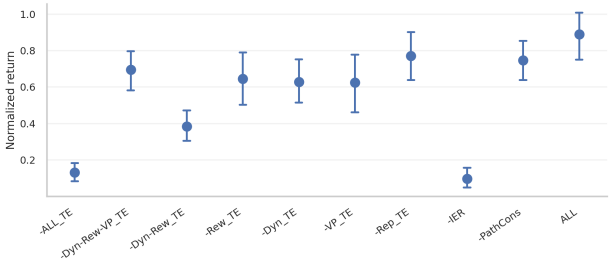


图*5*。 **对模型组件进行消融**。 *y*轴表示任务平均归一化回合回报，*x*轴表示不同设置。 – 表示移除一个组件。*TE* 表示任务嵌入。*Dyn*、*Rew*、*VP*、*Rep* 分别表示动力学、奖励、价值策略和表征模型。*-Dyn-Rew-VP TE* 表示从动力学、奖励和价值策略模型中移除任务嵌入。*IER*表示独立经验回放。*PathCons* 表示路径一致性损失。我们使用 3 个不同的随机种子，并报告 95% 置信区间。

6.4. 消融研究

最后，我们通过消融研究，分别考察 EZ-M 中引入的核心机制的作用：路径一致性（PC）损失、独立经验回放（IER）和任务嵌入。鉴于世界模型的架构较深，我们还研究了任务条件化的最佳注入位置。所有实验均在 HumanoidBench-Medium 设置下进行，步数预算为 1M。如图 5所示，**独立经验回放**是最关键的组成部分；移除它会导致性能大幅下降，凸显了它在缓解数据不平衡方面的作用。**路径一致性损失**也带来了显著提升，因为它能确保奖励预测与价值预测保持一致。此外，我们的结构分析证实，任务嵌入不可或缺，尤其是在对动力学模型和奖励模型进行条件化时，它们能够捕捉任务特定的物理规律和目标。

7. 讨论

在这项工作中，我们证明了基于模型的强化学习在扩展在线多任务机器人控制方面的有效性。我们在 HumanoidBench 上的结果验证了任务扩展是一种强大的正则化手段：通过利用不变物理动力学，EZ-M 将多样化的多任务数据转化为稳健的表征，其样本效率优于无模型基线方法。这证实，共享世界模型能够有效衔接不同的行为，同时缓解以策略为中心的方法中普遍存在的梯度干扰。树搜索带来的计算开销仍是一个局限，可能会给高频率的真实世界部署带来延迟挑战。此外，我们的框架假设具备共享的物理具身；在动力学迥异或异质的任务之间，迁移效果可能会减弱。

未来的工作将聚焦于：（1）将规划器蒸馏为轻量级策略，以实现实时推理；（2）整合语言条件目标，推动 EZ-M 实现开放式指令跟随。

参考文献

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe：低秩马尔可夫决策过程的结构复杂度与表征学习。神经信息处理系统进展，33:20095–20107，2020年。

Auer, P., Cesa-Bianchi, N., and Fischer, P. 多臂老虎机问题的有限时间分析。机器学习，47(2):235–256，2002年。Bellemare, M. G., Dabney, W., and Munos, R. 强化学习的分布视角。载于国际机器学习会议，第449–458页。PMLR，2017年。Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., 等 $\pi 0$：用于通用机器人控制的视觉-语言-动作流模型。corr，abs/2410.24164，2024年。doi: 10.48550.*arXiv*预印本 *ARXIV.2410.24164*，2024年。Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., 等。RT-1：用于大规模现实世界控制的机器人变换器。*arXiv*预印本 *arXiv:2212.06817*，2022年。Chen, X. and He, K. 探索简单的孪生表征学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15750–15758页，2021年。Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. GradNorm：用于深度多任务网络自适应损失平衡的梯度归一化。载于国际机器学习会议，第794–803页。PMLR，2018年。Coulom, R. 蒙特卡洛树搜索中的高效选择性与回传算子。载于国际计算机与游戏会议，第72–83页。Springer，2006年。Danihelka, I., Guez, A., Schrittwieser, J., and Silver, D. 通过基于Gumbel分布的规划改进策略。载于国际表征学习会议，2022年。Farquhar, G., Baumli, K., Marinho, Z., Filos, A., Hessel, M., van Hasselt, H. P., and Silver, D. 自洽模型与价值。神经信息处理系统进展，34:1111–1125，2021年。Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. 梦想并控制：通过潜在想象学习行为。*arXiv*预印本 *arXiv:1912.01603*，2019年。

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. 使用离散世界模型掌握 Atari。*arXiv* 预印本 *arXiv:2010.02193*，2020年。Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. 通过世界模型掌握多样化领域，2024年。网址 *https://arxiv.org/abs/2301.04104*，2023年。Hafner, D., Yan, W., and Lillicrap, T. 在可扩展世界模型内部训练智能体。*arXiv* 预印本 *arXiv:2509.24527*，2025年。Hansen, N., Wang, X., and Su, H. 用于模型预测控制的时序差分学习。*arXiv* 预印本 *arXiv:2203.04955*，2022年。Hansen, N., Su, H., and Wang, X. TD-MPC2：用于连续控制的可扩展、稳健世界模型。*arXiv* 预印本 *arXiv:2310.16828*，2023年。Hansen, N., Su, H., and Wang, X. 学习用于连续控制的大规模多任务世界模型。*arXiv* 预印本 *arXiv:2511.19584*，2025年。Hessel, M., Soyer, H., Espeholt, L., Czarnecki, W., Schmitt, S., and Van Hasselt, H. 使用 PopArt 的多任务深度强化学习。载于*AAAI*人工智能会议论文集，第33卷，第3796–3803页，2019年。Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. 强化学习中的无奖励探索。载于国际机器学习会议，第4870–4879页。PMLR，2020年。Karnin, Z., Koren, T., and Somekh, O. 多臂老虎机中的近最优探索。载于国际机器学习会议，第1238–1246页。PMLR，2013年。Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., 等。OpenVLA：一种开源视觉-语言-动作模型，2024年。网址 *https://arxiv.org/abs/2406.09246*，2024年。Kocsis, L. and Szepesv´ari, C. 基于老虎机的蒙特卡洛规划。载于欧洲机器学习会议，第282–293页。Springer，2006年。Kool, W., Van Hoof, H., and Welling, M. 随机束搜索及其所在：用于无放回序列采样的 Gumbel-top-k 技巧。载于国际机器学习会议，第3499–3508页。PMLR，2019年。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基于GPU的物理仿真。*arXiv* 预印本 *arXiv:2108.10470*，2021年。Nauman, M., Cygan, M., Sferrazza, C., Kumar, A., and Abbeel, P. 更大规模、经过正则化且采用分类形式：高容量价值函数是高效的多任务学习器。*arXiv* 预印本 *arXiv:2505.23150*，2025年。Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. 通过自举DQN实现深度探索。神经信息处理系统进展，第29卷，2016年。

O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., 等。Open X-Embodiment：机器人学习数据集与 RT-X模型：Open X-Embodiment 合作项目 0。载于*2024*年 *IEEE*机器人与自动化国际会议 *(ICRA)*，第 6892–6903 页。IEEE，2024年。

Pathak, D., Agrawal, P., Efros, A. A., 和 Darrell, T. 通过自监督预测进行好奇心驱动探索。载于国际机器学习会议，第2778–2787页。PMLR，2017年。

Peng, X. B., Abbeel, P., Levine, S., 和 Van de Panne, M. DeepMimic：通过示例引导，深度强化学习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., and Kanazawa, A. AMP：用于风格化物理角色控制的对抗性运动先验。*ACM*图形学汇刊 *(ToG)*，40(4):1–20，2021年。

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., et al. 通才智能体。*arXiv* 预印本 *arXiv:2205.06175*，2022年。

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. 通过基于学习得到的模型进行规划，精通 Atari、围棋、国际象棋和将棋。*Nature*，588(7839): 604–609，2020年。

Sferrazza, C., Huang, D.-M., Lin, X., Lee, Y. 和 Abbeel, P. HumanoidBench：用于全身运动和操作的模拟人形机器人基准测试。*arXiv* 预印本 *arXiv:2403.10506*，2024年。

Sutton, R. S. Dyna：一种集学习、规划和反应于一体的架构。*ACMSigart Bulletin*，2(4):160–163，1991。

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., 等. DeepMind 控制套件。*arXiv* 预印本 *arXiv:1801.00690*，2018年。Teh, Y., Bapst, V., Czarnecki, W. M., Quan, J., Kirkpatrick, J., Hadsell, R., Heess, N., 和 Pascanu, R. Distral：稳健的多任务强化学习。神经信息处理系统进展，30，2017年。Wang, S., Liu, S., Ye, W., You, J., 和 Gao, Y. EfficientZero v2：利用有限数据掌握离散与连续控制。*arXiv* 预印本 *arXiv:2403.00564*，2024年。

Wang, Y., Guo, H., Wang, S., Qian, L. 和 Lan, X. 自举式模型预测控制。*arXiv* 预印本 *arXiv:2503.18871*，2025年。

Yang, R., Xu, H., Wu, Y. 和 Wang, X. 通过软模块化实现多任务强化学习。神经信息处理系统进展，33:4767–4777，2020年。

Ye, W., Liu, S., Kurutach, T., Abbeel, P. 和 Gao, Y. 用有限数据掌握 Atari 游戏。神经信息处理系统进展，34：25476–25488，2021年。

Ye, W., Abbeel, P. 和 Gao, Y. 明智地利用思考时间：通过虚拟扩展加速蒙特卡洛树搜索。神经信息处理系统进展，35：12211–12224，2022年。

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K. 和 Finn, C. 用于多任务学习的梯度手术。神经信息处理系统进展，33：5824–5836，2020年。

A. 训练曲线

图 6 和图 7 分别展示了中等和困难任务套件上的训练动态。我们绘制评估回报随环境步数变化的曲线，每条曲线均汇总了多个随机种子的结果（阴影区域表示各次跑步之间的变化）。这些图补充了论文主体中的总结，展现了整个训练过程中的学习速度和稳定性。

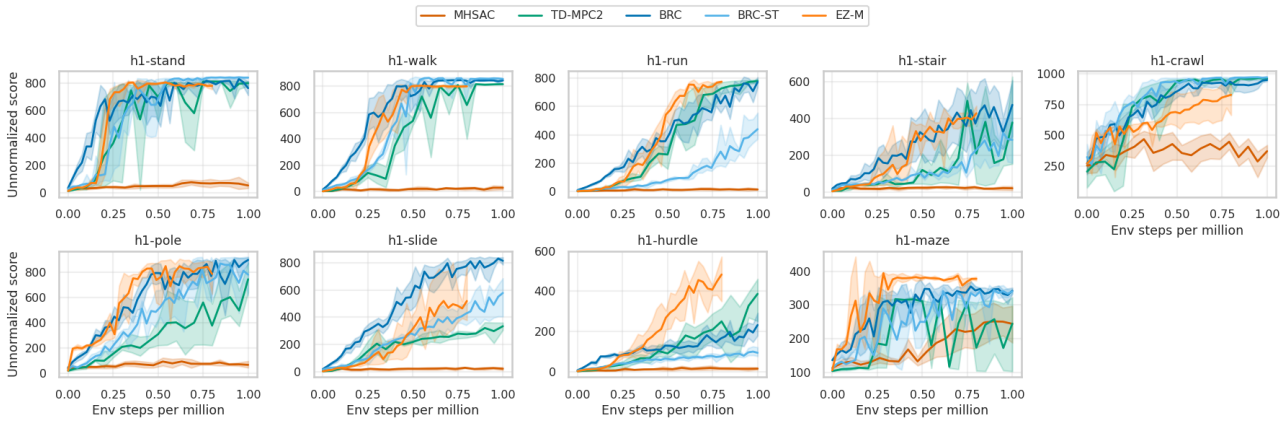


图 *6*。 **HumanoidBench-Medium 基准上的未归一化曲线**。 *y*轴表示回合回报，*x*轴表示每百万次环境交互。所有实验均使用 3 个不同的随机种子。我们报告 95% 置信区间。

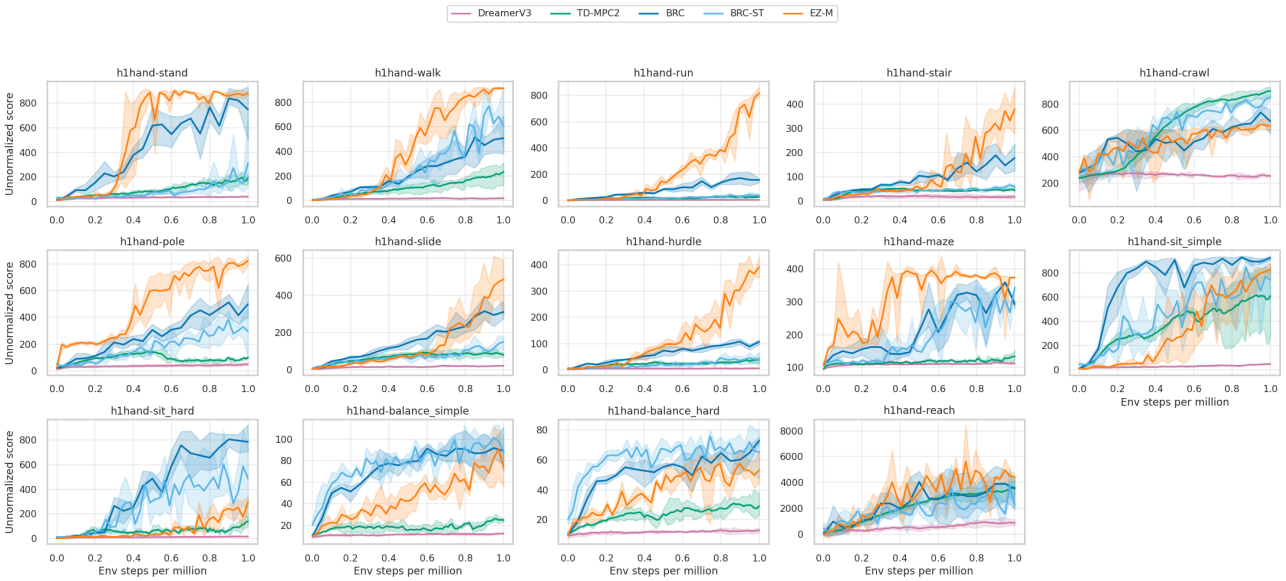


图 *7*。 **HumanoidBench-Hard 基准上的未归一化曲线**。 *y*轴表示回合回报，*x*轴表示每百万次环境交互。所有实验均使用 3 个不同的随机种子。我们报告 95% 置信区间。

B. 分数归一化与表格性能

我们使用以下公式对回合回报进行归一化：

$$\text{Return}_{\text{norm}} = \frac{\text{Return} - \text{Random Score}}{\text{Success Score} - \text{Random Score}} \tag{13}$$

每个任务的随机分数和成功分数可在表中找到。1 和 2。

表 *1.* **HumanoidBench-Medium（1M）任务的分数**。 我们将超过成功分数或达到最先进水平的结果以**粗体** 显示。

Task	随机	成功	MHSAC	TD-MPC2	BRC	BRC-ST	EZ-M（我们的方法）
h1-stand	10.545	800.0	74.904	**810.688**	**829.358**	**841.246**	**806.034**
h1-walk	2.377	700.0	29.352	**814.249**	**844.784**	**858.926**	**802.969**
h1-run	2.020	700.0	17.434	**778.845**	**772.299**	439.246	**775.867**
h1-stair	3.112	700.0	25.421	**497.040**	474.852	322.088	459.234
h1-crawl	272.658	700.0	449.088	**963.360**	**949.326**	**971.295**	**830.971**
h1-slide	3.191	700.0	29.004	334.359	**831.837**	578.478	577.740
h1-pole	20.090	700.0	93.175	**744.285**	**897.267**	**869.307**	**878.993**
h1-hurdle	2.214	700.0	19.038	387.131	232.031	99.793	**542.286**
h1-迷宫	106.441	1200.0	251.297	311.648	354.939	344.027	**380.795**

表*2.* HumanoidBench-Hard（1M）任务的分数。我们将**达到成功分数或达到最先进水平**的结果加粗。

Task	随机	成功	DreamerV3	TD-MPC2	BRC	BRC-ST	EZ-M（我们的）
h1hand-站立	11.973	800.0	40.938	196.511	**840.061**	315.990	**891.780**
h1hand-行走	2.505	700.0	19.819	234.061	515.383	**771.471**	**919.960**
h1hand-跑步	1.927	700.0	6.580	32.551	171.790	44.512	**818.397**
h1hand-楼梯	3.161	700.0	16.193	46.537	188.810	64.864	**380.349**
h1hand-爬行	278.868	800.0	261.801	**897.086**	735.992	**849.207**	644.386
h1hand-滑行	3.142	700.0	20.625	89.454	313.723	148.877	**535.404**
h1hand-杆	19.721	700.0	48.306	98.703	512.936	386.245	**841.353**
单手跨栏	2.371	700.0	6.381	38.890	108.067	46.112	**405.686**
单手迷宫	106.233	1200.0	116.332	133.406	358.803	342.268	**380.642**
单手简单坐姿_	10.768	750.0	47.866	612.656	**926.804**	**843.850**	**814.762**
单手困难坐姿_	2.477	750.0	15.589	139.181	**805.502**	605.966	304.836
单手简单平衡 _	10.170	800.0	12.615	26.401	91.749	**101.260**	98.004
单手困难平衡 _	10.032	800.0	13.110	29.886	**72.918**	72.199	57.565
单手伸展	-50.024	12000.0	894.156	3610.454	3895.343	3729.537	**5353.093**

C. 超参数与模型架构

表 3 汇总了　　我们实验中使用的关键超参数。除非另有说明，我们保持相　　e
超参数配　　置在所有任务中的设置，以确保公平比较并减少逐环境调优。

表4 详细介绍了 EZ-M 在 HumanoidBench-Hard 上使用的网络架构。该模型采用模块化设计，包含共享表征骨干网络、任务嵌入条件模块，以及独立的动力学/奖励/价值/策略预测组件。我们采用轻量级多层感知机骨干网络，并使用层归一化和残差块以实现稳定优化；同时使用离散支持集输出头（51 个分箱）进行奖励和价值预测，以提高学习信号的质量。总体而言，整个网络仅包含 **16M** 个参数，规模比采用 **BRC** 基线的 **1B** 规模模型小几个数量级，凸显了我们方法的参数效率。

表 *3.* HumanoidBench 实验中使用的关键超参数

超参数	**值**
环境	HumanoidBench
折扣因子 γ	0.99
展开步数	5
时序差分步数	5
批次大小	512
回放缓冲区大小	1×10^6 (中等) / 2×10^6 (困难)
训练步数	2×10^5 (中等) / 6×10^5 (困难)
网络隐藏层维度	512
任务嵌入维度	128
任务嵌入用途	表征 / 动力学 / 奖励 / 策略-价值
残差块数量	3
蒙特卡洛树搜索模拟	32
排名靠前的动作	16
采样动作	16
优化器	Adam
学习率	1×10^{-4}
权重衰减	2×10^{-5}

表 *4.* EfficientZero 网络架构（用于 HumanoidBench-Hard）。

模块	架构
表征 f_θ	多层感知机(305→512) + 层归一化 +双曲正切； 3×残差块(512)
任务嵌入	嵌入(14, 128)，最大范数=1
动力学 g_θ	层归一化(701)；多层感知机(701→512→512)； 3×残差块(512)
奖励头 r_θ	动作多层感知机(189→64)+层归一化+双曲正切； 动力学主干网络：LN(704)+MLP(704→512→512)+3×残差块(512)；
	奖励： [线性层(512→512)+批归一化+ReLU]×2; 线性层(512→51)
价值头 v_θ	主干网络维度=640； 3×残差块(640)；
	5 个头： [线性层(640→512)+随机失活+批归一化+ReLU]×2; 线性层(512→51)
策略头 π_θ	[线性层(640→512)+批归一化+ReLU]×2; 线性层(512→122)
投影	线性层(512→1024)+层归一化+ReLU; 线性层(1024→1024)+层归一化+ReLU;
	线性层(1024→512)+层归一化
投影头	线性层(512→1024)+层归一化+ReLU; 线性层(1024→512)
总参数量	**16M**