arXiv:2303.02234v2 [cs.RO] 9 Mar 2023

事后状态：融合仿真与真实任务要素，实现高效强化学习

Simon Guist、Jan Schneider、Alexander Dittrich、Vincent Berenz、Bernhard Sch¨olkopf 和 Dieter B¨uchler

摘要—**当能够以较低成本生成大量数据时，强化学习在解决复杂任务方面展现出了巨大潜力。在机器人学领域，生成训练数据的一种方法是基于由第一性原理推导出的动力学模型进行仿真。然而，对于涉及复杂软体机器人等情况的任务，构建此类模型要困难得多。借助强化学习在日益复杂的场景中进行有效训练，可以充分利用软体机器人等复杂系统。在本文中，我们利用动力学复杂程度的不平衡来提高采样效率。我们（i）将任务抽象为不同的组成部分，（ii）将简单的动力学部分转移到仿真中，以及（iii）复制这些虚拟部分以生成更多数据**并进行事后处理。**我们提出的新方法——事后状态（HiS）——利用这些数据，并选择最有用的转移用于训练。该方法可与任意离策略算法结合使用。我们在多个具有挑战性的仿真任务上验证了该方法，结果表明，无论单独使用还是与现有的事后算法——事后经验回放（HER）——结合使用，它都能提升学习效果。最后，我们在一个物理系统上评估了 HiS，结果显示，它能够提升肌肉机器人在复杂乒乓球任务中的性能。实验视频和代码可在webdav.tuebingen.mpg.de/his/找到。**

I. 引言

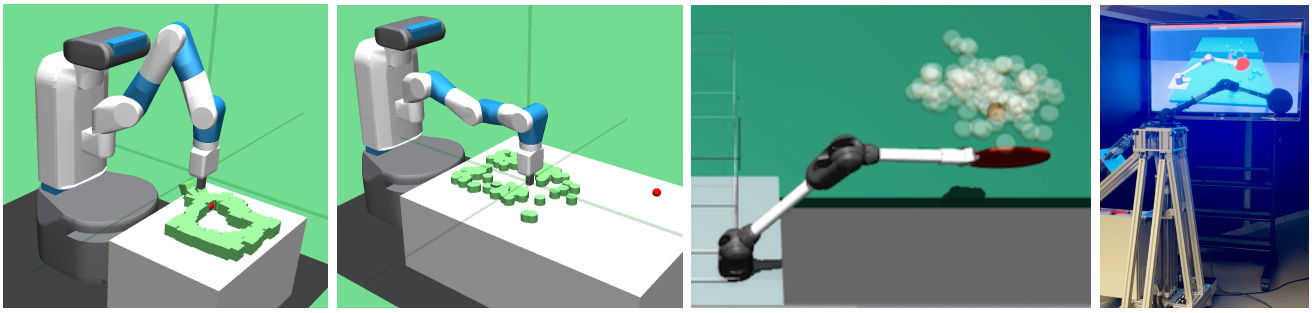
强化学习（RL）在为具有挑战性的机器人任务设计最优行为方面具有巨大潜力。这些任务的困难通常源于物体操作中的接触、执行快速而精确的运动的要求，或难以控制的系统（如软体机器人）。尽管已有研究表明，RL 能够处理此类问题，但 RL 方法需要与环境进行大量交互这一主要缺点依然存在。这使得在真实机器人上进行学习颇具挑战。

在仿真中训练并将由此得到的策略迁移到真实系统，是缓解这一挑战的一种方法。然而，即使仿真中存在微小误差，也可能导致仿真系统与真实系统的行为不同，这个问题称为现实差距。对于软体机器人等难以准确建模的复杂系统，这一问题更加严重。仿真到现实技术旨在减轻现实差距的影响。许多常用技术通过域随机化，使学得的策略能够更好地应对任务动力学的变化，从而实现更有效的策略迁移随机化 [26, 41]。然而，这些方法通常计算量较大，并且需要对参数进行精细调优，例如动力学参数的随机化幅度。随着任务复杂度不断提高，仿真到现实方法的这些缺点也会累积，因此有必要在真实系统上部分或完全地学习许多复杂任务。

仿真与真实混合（HySR） [5] 是一种简化真实系统上复杂任务训练的方法。该方法基于这样一种认识：对于某些任务，环境的某些部分比其他部分具有更简单的动力学。例如，许多机器人球类游戏由机器人和球组成，与球相比，机器人的动力学更为复杂。如果机器人配备了柔性组件，建模会变得更加困难，从而进一步加剧建模复杂度的不匹配。HySR 的核心思想是将任务中复杂的部分保留在真实环境中，而对较简单的部分进行仿真。这一策略在促进向完全真实系统迁移的同时，也带来了显著的实践优势。首先，由于需要处理的真实物体更少，任务设置得以简化；例如，这些物体可能会磨损，或需要考虑安全因素。其次，它能够大幅简化回合式强化学习任务中的重置机制。机器人由执行器驱动，通常可以自行重置，而环境重置往往需要手动将所有物体移回初始位置。第三，可以对虚拟物体添加在完全真实系统上无法实现的数据增强技术。最后，仿真能够提供虚拟物体的精确位置或所发生的接触等信息，而这些信息对于真实物体而言可能难以获取。这些数据可用于生成标签或奖励，从而辅助训练过程。尽管 HySR 能够简化训练的实践环节，但训练所需的大量真实机器人交互仍可能给真实世界强化学习带来问题。

在这项工作中，我们提出了一种方法，能够显著减少学习所需的真实系统交互次数。我们的方法“事后状态（HiS）”的核心思想是：将单个真实实例的数据与通过同时模拟虚拟部分的多个不同实例而生成的额外数据配对。以机器人球类运动为例，我们的方法会同时模拟机器人动作对多个虚拟球的影响。我们用这些额外的虚拟数据对每次滚动过程中的虚拟部分进行重新标记，以事后视角。直观而言，智能体所经历的是：如果它遇到的是另一个虚拟部分，但执行的动作与原始回合中的相同，

本研究得到马克斯·普朗克智能系统研究所的支持。所有作者均隶属于马克斯·普朗克智能系统研究所，地址为 Max-Planck-Ring 4, 72076 T¨ubingen, 德国。



(a) 推 (b) 滑动 (c) 模拟机器人乒乓球 (d) 真实机器人乒乓球

图 1：HiS 在本文所考虑任务中的可视化。HiS 并非使用单个物体进行训练，而是并行使用多个虚拟物体，以生成更多数据和经验，并在训练早期提高获得更高奖励的概率。

会发生什么。这一重新标记过程使强化学习智能体无需在真实系统上采集更多转移，就能生成额外经验。随后，任何离策略强化学习算法（例如软演员-评论家（SAC） [12]）都可以将这些事后数据与常规训练数据结合使用。特别是对于奖励稀疏的任务，额外数据非常有益，因为它能提高智能体在训练早期获得正奖励的概率。本文的贡献如下：

1) 扩展 *HySR*，以实现样本高效的训练
我们的主要贡献是提出一种在混合仿真与真实（HySR）设置下进行样本高效训练的方法。为此，我们将 HySR 训练形式化，并基于此开发了一种名为事后状态（HiS）的新算法。

2) 评估 *HER* 与 *HiS* 的相互作用
我们的实验表明，HiS 与 HER 结合使用比单独使用任一方法都能实现更高的样本效率。我们认为，HiS 和 HER 以不同且互补的方式提升任务性能。

3) 在多种任务上对 *HiS* 进行全面的实验验证我们展示了 HiS 在原始 HySR 真实机器人乒乓球任务上的性能提升。此外，我们还研究了 HiS 在多大程度上可应用于操作任务，并报告了在这些场景中更高效的训练结果（所有实验概览见图 1）。我们还发现，在完全仿真的环境中，HiS 能够缩短墙钟时间，实现更高效的训练。

II. 相关工作

我们提出的新方法 HiS 与其他研究有一些共同之处，也从中汲取了灵感。接下来，我们将讨论这些相似之处。

事后经验回放（HER） [1] 通过事后重新标记转移中的目标状态，提高稀疏目标条件强化学习任务的样本效率。该方法将原始目标替换为智能体到达的状态，然后使用原始数据与重新标记的数据训练离策略强化学习智能体。这样，智能体在到达重新标记的目标时会获得额外的正反馈；与未能达到原始目标所收到的负反馈相比，这种反馈可能更具指导意义。Rauber 等人 [31] 通过使用重要性采样，将这一思路扩展到了同策略算法。动态后见经验回放（DHER） [9] 是 HER 的一种变体，支持在回合中不断变化的动态目标。该方法使重新标记目标这一思路适用于抓取移动物体等任务。HER 对事后目标进行均匀采样，而近期方法则根据由此产生的转移对智能体的指导作用，对目标进行优先排序。这些方法会采样能够引导智能体趋近真实目标 [2, 32, 10] 或导致较大时序差分（TD）误差 [22] 的事后目标。其他方法则均匀采样事后目标，并在从回放缓冲区采样时优先选择信息量更大的转移，这与优先经验回放（PER） [33]类似。Zhao 和 Tresp [40] 优先选择展现困难行为的轨迹，并根据轨迹过程中系统能量的增量来量化难度。Beyene 和 Han [3] 主要采样会产生较大时序差分误差的事后转移。与 HER 及其扩展方法类似，我们的方法也会进行事后重新标记，但它着重于交换状态的虚拟部分，而非目标。

HiS 和仿真到现实技术都利用仿真来制定能够迁移到现实的策略。为减轻现实差距的影响，人们提出了多种技术，例如域随机化 [26, 28, 23]和域适应 [15,14, 39] 。另一方面，HiS 假设环境的某些组成部分的现实差距足够小，并提出了一种在仿真与现实混合场景中提高效率的方法。仿真到现实方法还可以与 HiS 结合使用，以弥合剩余差距。

其他研究也探索了利用同一任务的真实数据和仿真数据来提升学习性能。例如，Kang 等人 [16]利用仿真数据学习感知系统，并基于真实数据学习奖励模型。他们使用结合学习得到的奖励模型的模型预测控制来解决四旋翼导航任务。Di-Castro 等人 [8] 将成本低但不精确的仿真转移与成本高昂的真实转移相结合，为真实任务学习策略。相比之下，HiS 探索了通过在每个转移中混合仿真和真实组件来实现更高效的训练。

算法1 事后状态（HiS）

1：初始化离策略　　强化学习算法 $\mathbb{A}$，回放　缓冲区 R，选择判据 c

2：**对** 回合 $i = 1,\ 2,\ \ldots\ldots$ **执行**

3:　　采样 $\mathbf{s}^{\mathrm{v}}_{1:T} \sim \mathcal{D}_{\mathrm{rec}}$ 并将真实系统初始化为 $\mathbf{s}^{\mathrm{r}}_1$

4:　　**对于** $t = 1, \ldots, T$ **执行**

5:　　采样动作 $\mathbf{a}_t$，从 $\mathbb{A}$ 给定的 $\mathbf{s}_t = [\mathbf{s}^{\mathrm{r}}_t、\ \mathbf{s}^{\mathrm{v}}_t]$

6:　　将 $\mathbf{a}_t$ 应用于真实系统，并观察 $\mathbf{s}^{\mathrm{r}}_{t+1}$

7:　　$\mathbf{s}^{\mathrm{v}}_{t+1} \sim$仿真$([\mathbf{s}^{\mathrm{r}}_t、\ \mathbf{s}^{\mathrm{v}}_t]、\ \mathbf{a}_t)$　或获取$\mathbf{s}^{\mathrm{v}}{}_{t+1}$ 来自$\mathbf{s}^{\mathrm{v}}{}_{1:T}$

8:　　**结束循环**

9:　　**对于** $t = 1, \ldots, T$ **执行**　　▷ 标准回放

10:　　存储$([\mathbf{s}^{\mathrm{r}}_t, \mathbf{s}^{\mathrm{v}}_t], \mathbf{a}_t, r([\mathbf{s}^{\mathrm{r}}_t, \mathbf{s}^{\mathrm{v}}_t], \mathbf{a}_t), [\mathbf{s}^{\mathrm{r}}{}_{t+1}, \mathbf{s}^{\mathrm{v}}_{t+1}])$ 中 R

11:　　**结束循环**

12:　　初始化临时缓冲区 R^{temp}　　▷ HiS 重放

13:　　**对于** $j = 1, 2, \ldots$ **执行**

14:　　采样 $\mathbf{s}^{\mathrm{v}}{}_{m1:\,T} \sim \mathcal{D}_{\mathrm{rec}}$

15:　　**对于** $t = 1, \ldots, T$ **do**

16:　　$\mathbf{s}^{\mathrm{v}m}_{t+1} \sim$仿真$([\mathbf{s}^{\mathrm{r}}_t,\ \mathbf{s}^{\mathrm{v}m}{}_t\ ], \mathbf{a}_t)$ 或从 $\mathbf{s}^{\mathrm{v}m}{}_{t+1}$ 从 $\mathbf{s}^{\mathrm{v}m}{}_{1:T}$

17:　　$e^m_t = ([\mathbf{s}^{\mathrm{r}}_t,\ \mathbf{s}^{\mathrm{v}m}{}_t\ ], \mathbf{a}_t,\ r([\mathbf{s}^{\mathrm{r}}_t,\ \mathbf{s}^{\mathrm{v}m}{}_t\ ], \mathbf{a}_t), [\mathbf{s}^{\mathrm{r}}_{t+1},\ \mathbf{s}^{\mathrm{v}m}{}_{t+1}])$

18:　　将重新标记的转 e^m_t 存入 R^{temp}

19:　　**结束循环**

20:　　**结束循环**

21:　　每隔几回合：

　　将 R^{temp} 中的转移添加 R，使用判据 c

22:　　每隔几步：

　　使用回放缓冲区 R 对 $\mathbb{A}$ 进行优化

23: **结束循环**

利用离线强化学习 [19, 20] 也可以通过使用现有数据集来缓解在真实系统上训练所带来的一些问题。离线强化学习的核心问题在于，收集数据集的行为策略与训练策略之间存在分布偏移 [20]。近期研究通过限制训练策略，使其与行为策略相似 [36, 34,17], ；修改 Q函数的目标，使其相对于不在数据集中的样本更加保守 [18, 38], ；或惩罚学习得到的模型在轨迹展开中的不确定性 [37, 29]。

III. 方法

HiS 的核心组成部分包括并行虚拟轨迹的生成，以及从 HiS 生成的额外转移中进行选择的准则。本节将分别介绍这些组成部分，以及 HiS 所基于的 HySR 设置。

A. 强化学习基础

我们考虑将任务表述为折扣马尔可夫决策过程，该过程由元组 $\mathcal{M} = (\mathcal{S},\ \mathcal{A},\ \mathcal{P},\ \gamma,\ r)$ 定义，其中 $\mathcal{S}$ 表示状态空间，$\mathcal{A}$ 表示动作空间。转移动态 $\mathcal{P}(\mathbf{s}_{t+1}|\mathbf{s}_t,\ \mathbf{a}_t)$ 表示执行动作 $\mathbf{a}_t \in \mathcal{A}$ 后从状态 $\mathbf{s}_t \in \mathcal{S}$ 到达状态 $\mathbf{s}_{t+1} \in \mathcal{S}$ 的概率。此外，智能体在每个时间步都会获得标量奖励 $r_t = r(\mathbf{s}_t,\ \mathbf{a}_t)$。强化学习的目标是找到一个最优策略 $\pi^\star$，使折扣总回报最大化

$$R = \mathbb{E}_\pi \left[\sum_t^T \gamma^t r_t \right], \tag{1}$$

其中，T 是时间范围，$\gamma \in [0, 1]$ 是折扣因子。智能体通过与环境交互来收集转移 $e_t = (\mathbf{s}_t,\ \mathbf{a}_t,\ r_t,\ \mathbf{s}_{t+1})$，并利用这些转移迭代更新其策略 π。许多流行的离策略强化学习算法，例如 SAC [12] 或深度Q网络（DQN） [24], ，会将这些转移存储在缓冲区 $\mathcal{D}$ [21] 中，并在更新时重放它们，例如用于更新动作价值函数 $Q(\mathbf{s}_t,\ \mathbf{a}_t)$。这个回放缓冲区大小固定，包含最近的转移。环形缓冲区会用最新的转移替换最旧的转移，是一种常见的实现方式。Q函数的更新涉及最小化时序差分误差的一种变体

$$\delta(\mathbf{s}, \mathbf{a}, r, \mathbf{s}') = r + \gamma \max_{\mathbf{a}' \in \mathcal{A}} Q(\mathbf{s}', \mathbf{a}') - Q(\mathbf{s}, \mathbf{a}), \tag{2}$$

其中的变化可能包括选择下一个动作的另一种方式 $\mathbf{a}'$，或者采用 $\delta_n = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n \max_{\mathbf{a}'} Q(\mathbf{s}_{t+n}, \mathbf{a}') - Q(\mathbf{s}_t, \mathbf{a}_t)$ 的 $\delta_n = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n$ n 步时序差分误差公式。时序差分误差代表了一种惊奇程度的度量：δ的绝对值越高，该特定转移对Q函数更新的影响就越大。因此，时序差分误差指标通常用于对转移进行排序，例如用于优先级扫描 [25, 27]和优先经验回放（PER） [33]。Q函数损失使用回放缓冲区 $\mathcal{D}$ 以及平方时序差分误差

$$J(\theta) = \mathbb{E}_{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}') \sim \mathcal{D}} \left[\left(r + \gamma \max_{\mathbf{a}' \in \mathcal{A}} Q_{\theta^{\mathrm{old}}}(\mathbf{s}', \mathbf{a}') - Q_\theta(\mathbf{s}, \mathbf{a}) \right)^2 \right], \tag{3}$$

其中，旧参数 θ^{old} 在当前策略参数 θ 的优化过程中保持固定，并且根据一种传统上均匀、但也可变化（例如在PER中）的分布，从 $\mathcal{D}$ 中采样转移。

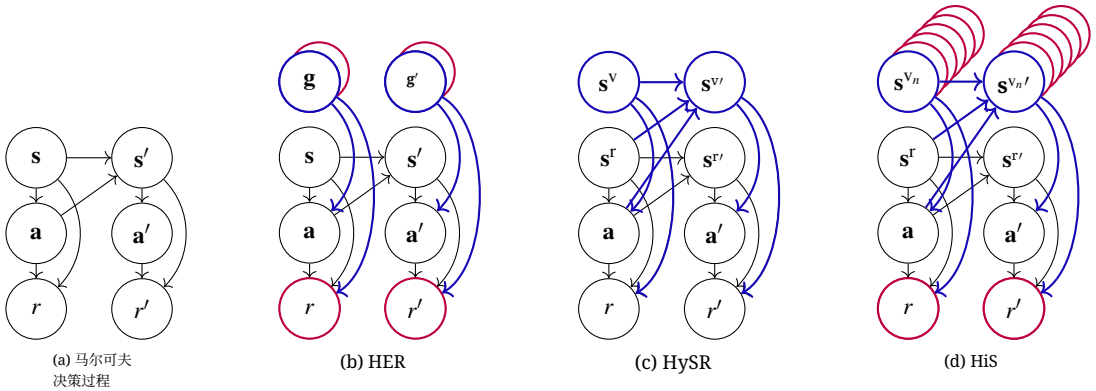
B. 仿真与真实混合训练

仿真与真实混合（HySR）训练 [5] 是事后状态（HiS）的基础。直观而言，HySR的理念是将任务中更易建模的部分交由仿真来处理，从而获得实际收益。同时，在训练过程中保留较难的部分为真实环境。更正式地说，HySR假设马尔可夫状态$\mathbf{s}$ 分为真实状态$\mathbf{s}^{\mathrm{r}}$ 和虚拟状态$\mathbf{s}^{\mathrm{v}}$

$$\mathbf{s} = [\mathbf{s}^{\mathrm{r}}, \mathbf{s}^{\mathrm{v}}] \tag{4}$$

并且，控制真实部分$p(\mathbf{s}^{\mathrm{r}}_{t+1} | [\mathbf{s}^{\mathrm{r}}_t, \mathbf{s}^{\mathrm{v}}_t], \mathbf{a}_t)$的动力学比控制虚拟部分$p(\mathbf{s}^{\mathrm{v}}_{t+1} | [\mathbf{s}^{\mathrm{r}}_t, \mathbf{s}^{\mathrm{v}}_t], \mathbf{a}_t)$的动力学更复杂。为了更好地将HySR 训练后学到的策略迁移到完全真实的环境中，HySR 保留复杂的真实动力学，并要求虚拟动力学足够准确。另一项要求是虚拟状态不会影响真实状态

$$p(\mathbf{s}^{\mathrm{r}}_{t+1} | [\mathbf{s}^{\mathrm{r}}_t, \mathbf{s}^{\mathrm{v}}_t], \mathbf{a}_t) = p(\mathbf{s}^{\mathrm{r}}_{t+1} | \mathbf{s}^{\mathrm{r}}_t, \mathbf{a}_t) \tag{5}$$



(a) 马尔可夫决策过程　(b) HER　(c) HySR　(d) HiS

图2：一般马尔可夫决策过程、HER、HySR 和 HiS 的图模型。所有蓝色和红色元素均表示与一般马尔可夫决策过程的差异。红色表示事后重新标记的组件。HER 在马尔可夫决策过程中加入目标状态，并重新标记目标。HySR 将状态分离为虚拟部分和真实部分。附加箭头表示动力学。HiS 结合了 HySR 和 HER 的关键思想，分别是状态分离与动力学，以及重新标记。

因为将力从虚拟部分映射到真实部分可能十分复杂，尤其是映射到复杂或未知的动力学时。

虚拟动力学不一定会在整个训练过程中模拟该部分。在真实部分与虚拟部分首次接触之前，HySR 会重放虚拟部分的记录数据，而不是对其进行模拟，以尽量减少模型误差的累积。为此，HySR 首先从包含N_{rec} 条记录轨迹的数据库 $\mathcal{D}_{\text{rec}}$中均匀采样一条虚拟轨迹 $\tau^{\text{v}_{\text{rec}}}$ 。

$$\mathcal{D}_{\text{rec}} = \left[\mathbf{s}_1^{\text{v}_{n}^{\text{rec}}}, \ldots, \mathbf{s}_T^{\text{v}_{n}^{\text{rec}}}\right]_{n=1}^{N_{\text{rec}}} \tag{6}$$

虚拟部分与真实部分接触后，模拟动力学决定后续运动，因为无法从记录中推断后续部分的任何信息。一条完整的 HySR 训练轨迹，

$$\tau^{\text{HySR}} = [e_1^{\text{rec}}, \ldots, e_{t_c}^{\text{rec}}, e_{t_c+1}^{\text{sim}}, \ldots, e_T^{\text{sim}}], \tag{7}$$

因此，由从数据库回放的转移 $\mathbf{e}_t^{\text{rec}} = \left([\mathbf{s}_t^{\text{r}}, \mathbf{s}^{\text{v}^{\text{rec}}}\mathbf{t}], \mathbf{a}_t, \mathbf{r}([\mathbf{s}_t^{\text{r}}, \mathbf{s}^{\text{v}^{\text{rec}}}\mathbf{t}], \mathbf{a}_t), [\mathbf{s}_{t+1}^{\text{r}}, \mathbf{s}^{\text{v}^{\text{rec}}}\mathbf{t}_{+1}]\right)$ 以及模拟得到的转移 $\mathbf{e}_t^{\text{sim}} = \left([\mathbf{s}_t^{\text{r}}, \mathbf{s}^{\text{v}^{\text{sim}}}\mathbf{t}], \mathbf{a}_t, \mathbf{r}([\mathbf{s}_t^{\text{r}}, \mathbf{s}^{\text{v}^{\text{sim}}}\mathbf{t}], \mathbf{a}_t), [\mathbf{s}_{t+1}^{\text{r}}, \mathbf{s}^{\text{v}^{\text{sim}}}\mathbf{t}_{+1}]\right)$组成，假设接触发生在时间 $\mathbf{t}_c$。训练过程中，策略根据式（4）中的完整状态采样动作，其中不包含虚拟部分是回放还是模拟得到的信息。

HySR 特别适用于以下应用场景：(i) 虚拟部分与真实部分的动力学复杂度不平衡；(ii) 使用虚拟部分的真实实例进行训练较为方便或更安全；(iii) 分离部分之间的接触发生在每个滚动过程相对靠后的阶段，从而能够更好地利用记录的数据；(iv) 虚拟部分对真实部分的影响可忽略不计，而反向影响已知且能够良好地迁移至现实世界。以上各点均适用于许多球类运动，例如足球、板球、棒球或篮球，在这些运动中，物体必须通过与运动员的交互到达目标状态。尽管迄今为止 HySR 的有效性仅在机器人乒乓球中得到验证，但在 HySR 设置下进行训练并不限于球类运动。例如，对于滑动推物或投掷物体等任务（与拾取物体相对），我们可以将不同重量或材质的物体与单一机械臂动作相结合。(i) 至 (iv) 点同样适用于避开移动障碍物等任务，因为首先就应避免接触。例如，机器人在杂乱且动态的环境中行走，或自动驾驶汽车在拥挤的城市中行驶。

C. 事后状态

HySR 还支持生成额外数据，而事后状态 (HiS)是我们提出的、利用这些数据进行高效训练的方法。HiS 实现了在 HySR 设置中生成额外虚拟数据这一想法。具体而言，对于每条 HySR 轨迹 $\tau^{\text{HySR}} = [e1, \ldots, e_T]$，它会生成轨迹 $\tau^{\text{HiS}} = [e_1^m, \ldots, e_T^m]$，其中 $m = 1, \ldots, M$。对于这 M 条轨迹中的每一条，我们都会采样一组不同的虚拟状态 $\mathbf{s}1^{\text{v}_m}{:}T$ $\sim \mathcal{D}_{\text{rec}}$。我们对 HySR 轨迹进行扩充，以反映虚拟状态为 $\mathbf{s}1^{\text{v}_m}{:}T$ 时可能发生的情况。具体而言，我们生成事后转移

$$e_t^m = \left([\mathbf{s}_t^{\text{r}}, \mathbf{s}_t^{\text{v}_m}], \mathbf{a}_t, r([\mathbf{s}_t^{\text{r}}, \mathbf{s}_t^{\text{v}_m}], \mathbf{a}_t), [\mathbf{s}_{t+1}^{\text{r}}, \mathbf{s}_{t+1}^{\text{v}_m}]\right) \tag{8}$$

对于HySR生成的每个转移 $\mathbf{e}_t = \left([\mathbf{s}_t^{\text{r}}, \mathbf{s}_t^{\text{v}}], \mathbf{a}_t, \mathbf{r}([\mathbf{s}_t^{\text{r}}, \mathbf{s}_t^{\text{v}}], \mathbf{a}_t), [\mathbf{s}_{t+1}^{\text{r}}, \mathbf{s}_{t+1}^{\text{v}}]\right)$，真实状态$\mathbf{s}_t^{\text{r}}$、$\mathbf{s}_{t+1}^{\text{r}}$ 以及动作$\mathbf{a}_t$保持与HySR转移中相同，但虚拟状态会通过事后方式重新标记为虚拟状态 $\mathbf{s}^{\text{v}_m}\mathbf{t}$ 、$\mathbf{s}^{\text{v}_m}\mathbf{t}_{+1}$的不同实例，这些实例取自预先记录的数据或仿真数据。随后，根据重新标记后的状态计算事后转移的奖励。请注意，以这种方式，HiS生成了真实部分可能如何与虚拟物体的各个实例交互的丰富多样情形。因此，HiS能够就这一特定动作序列在不同情境下的质量提供有价值的反馈。在图1c所示的乒乓球示例中，球拍击中了一整团具有不同速度和轨迹的球；因此，所得球会到达截然不同的落点。

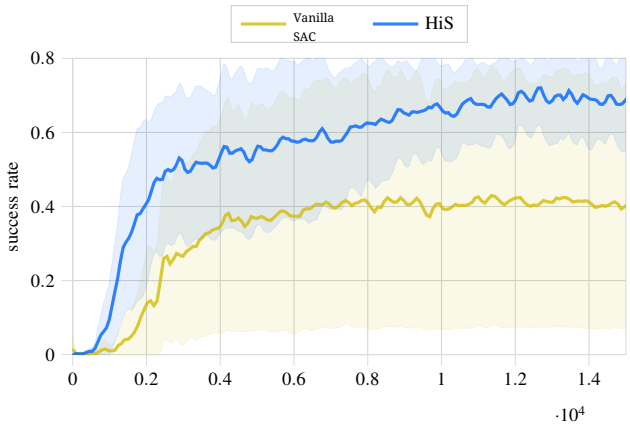
a）选择策略：将所有HiS轨迹加入回放缓冲区并由离策略算法进行采样的朴素方法，可能会损害学习过程。影响强化学习算法样本复杂度的一个关键因素是回放缓冲区中转移的离策略程度 [11]。离策略数据具有挑战性，因为它们是在多个不同策略下收集的，而式（1）中回报的期望值是相对于由当前策略诱导的分布 π计算的。这种分布偏移会给回报估计引入偏差和方差，并导致对Q 值的错误估计。通常，缓冲区的大小决定了缓冲区的离策略程度，因为较早的转移更可能源自较早、因而不同的策略 [11]。额外的HiS轨迹会增加缓冲区的离策略程度，因为HiS轨迹中的动作是在以原始HySR状态为条件的情况下生成的，而非以重新标记后的状态为条件。

因此，HiS 训练包含一种选择策略，用于挑选最能加速学习过程的转移。我们提出根据奖励$r(\mathbf{s},\mathbf{a})$或式（2）中的时序差分误差 $\delta(\mathbf{s},\mathbf{a},r,\mathbf{s}')$对 HiS 转移进行排序。优先选择奖励较高的转移，其依据是尽早获得高奖励，这在稀疏奖励设置下尤其有用。时序差分误差对应于对Q函数更新的影响程度，因此是对转移进行排序的合理依据。选择过程基于时序差分误差的绝对值 $|\delta(\mathbf{s},\mathbf{a},r,\mathbf{s}')|$，因为使用高值和低值转移进行更新，都有助于使动作价值函数更接近其最优形式Q^*。如何将这些度量（时序差分误差或奖励）应用于 HiS 轨迹中的转移至关重要。我们研究了两种方法：（i）对每个转移分别应用度量；（ii）对轨迹中的所有转移整体应用度量，即对整条轨迹的奖励或时序差分误差求和。第一种实现是直接明了的方法；第二种则用于检验这样的假设：度量值较高的轨迹中的所有转移都具有重要意义。后一种思路曾在优先回放的背景下得到探讨，结果表明，时间上相关的转移具有相关的时序差分误差 [4]。

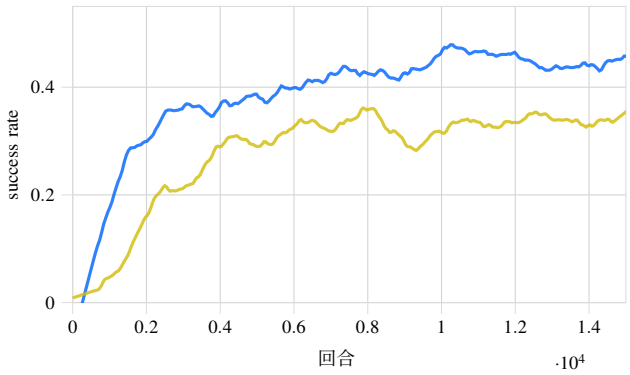
在定义了对 HiS 转移进行排序的判据后，我们现在介绍确定要添加到回放缓冲区中的 HiS 转移数量的方法。要纳入缓冲区，判据 c 必须同时满足：(i) 超过阈值 ψ_c，且 (ii) 按照该判据排在 k_c 个最佳转移之列。条件 (i) 可防止添加那些虽在本轮同类转移中排名靠前、但值过低的转移。条件 (ii) 控制回放缓冲区中离策略数据的最大数量。算法 1详细概述了 HiS。

b) 与 *HER* 的关系：HiS 与 HER [1] 有一些共同之处。HER 训练目标条件策略，并在事后将期望目标重新标记为实际达成的目标。HER 和 HiS 都利用事后数据提高采样效率，在稀疏奖励场景下尤其有用。与之不同的是，HiS 重新标记的是虚拟状态，而非目标。

图2展示了标准马尔可夫决策过程、



(a) 仿真机器人



(b) 真实机器人

图3：乒乓球任务的结果。HiS 提高了采样效率和渐近性能（以机器人步数衡量）。

HySR、HER和HiS在其相应图模型方面的差异。请注意，HER设置还允许每个转移中包含多个目标，并且 [9] 将动力学引入HER的目标空间。如图2所示，目标和动态目标 [9] 可被视为虚拟状态的特殊情形。因此，我们的方法也可应用于目标条件设置。与HER应用于该设置时的区别在于，HER会采样目标，使其对应于特定的已达成目标（例如，事后采样的目标是机器人夹爪的最终位置）。然而，HiS会从初始目标分布中采样，然后使用优先级策略从采样目标集合中进行选择。在采样足够多的目标并采用相应的优先级策略后，HiS可能会收敛到与HER相同的行为（例如，HiS采样足够多的目标，使得优先级策略能够选取与HER策略本会选择的目标相似的目标）。在这种设置下，与HER相比，HiS增加的复杂性似乎没有必要。然而，在一般的HySR设置中，HER无法应用于虚拟状态，而HiS则可以应用。

对于同时具有目标状态和虚拟状态的任务，可以在 HiS 的基础上应用 HER。

c) 结合 *HiS* 和 *HER*：HER 和 HiS 从两个不同的方向解决学习过程中的稀疏性问题。HER 会生成额外的经验，这些经验能够解决目标并获得较高奖励，因此有益于学习过程；但由于 HER 只会重新标注目标，因此并不能直接帮助探索状态空间。另一方面，HiS 通过创建环境的额外虚拟状态，帮助探索状态空间。与 HER 不同，HiS 创建的经验可能无法到达目标，也可能无法获得高奖励。为了在以目标为导向的 HySR 设置中获得 HER 的部分益处，HiS 可以采样大量事后轨迹，直到其中一些轨迹也能到达目标。然而，对于具有挑战性且奖励稀疏的任务，这种策略在计算上可能代价高昂，甚至无法实施。

相反，对于此类任务，结合HER和HiS可能更有益，并能发挥两者的优势。实际上，可以按照算法1实现这种组合，并将完整的HiS回合添加到HER所使用的回放缓冲区中。HER会重新标记HiS轨迹的目标，随后即可对带有重新标记的虚拟状态和重新标记目标的轨迹进行采样，以优化策略。

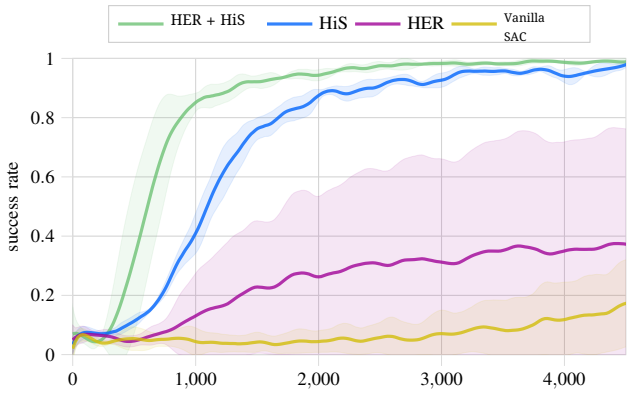
IV. 实验

我们提出的新方法HiS有望在HySR设定下提高样本效率。我们在两种场景下开展HiS评估：1）在若干适用HySR假设的仿真任务上进行广泛实验；2）在具有挑战性的真实机器人乒乓球任务上进行实验。此外，我们还研究了HiS在多大程度上适用于操作任务，并展示了HER和HiS组合的效率。

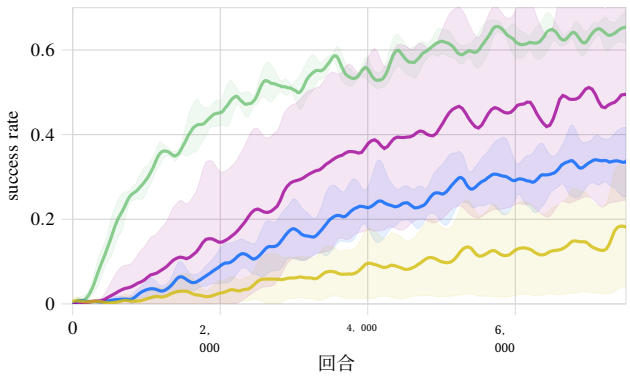
A. 将*HiS*应用于原始*HySR*乒乓球任务

最初提出HySR的工作 [5] 使用肌肉机器人学会了打乒乓球，并自然地将难以控制的柔性肌肉机器人 [7, 6] 与相对简单的球模型区分开来。不同于原始任务，我们不再使用手工设计的稠密奖励函数，而是采用简单的稀疏奖励：当球落在半径为40cm的圆内时返回1，否则返回0。为了应用HiS，我们从100条记录的球轨迹中采样虚拟球，并回放20个并行实例。我们在SAC [12] 之上应用HiS，并以原始SAC作为基线进行比较。在轨迹选择方面，我们采用阈值为 $\psi_c = 0.5$的奖励标准。附录详细介绍了实验细节以及每项实验的参数，其中还包含对不同选择标准的评估。

图3所示的实验表明，无论是在完全仿真环境还是HySR 设置下，HiS 都能以更少的机器人时间步学会这项任务。图3a中的仿真实验展示了使用不同随机种子进行的十次运行的均值和方差。HiS 的平均性能最高可达到约 70% 的成功率，而原始 SAC 约为 40%。HiS 仅用 SAC 达到渐近性能所需样本的约 29%，便达到了与 SAC 相当的渐近性能



（a）FetchPush



(b) FetchSlide

图4：在仿真 Gym 机器人任务 FetchPush 和 FetchSlide 上的结果。就机器人步数而言，HER 和 HiS 的学习速度都快于 SAC。结合 HER 和 HiS 可获得最佳结果。

（分别为 2k 和 7k 回合）。另一个观察结果是，与 SAC 相比，不同运行之间的方差更小。

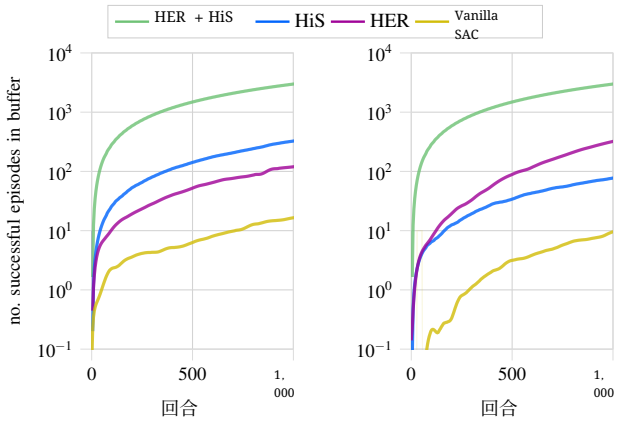
使用真实机器人时，HiS 也更高效。图3b 展示了原始 SAC 和 HiS 的一次训练轮次。HiS 的成功率约为 45%，而 SAC 仅约为 35%。与仿真实验类似，HiS 仅用真实机器人收集约 38% 的样本数量（3k 和 8k 回合），便达到了 SAC 的最高成功率。

B. HySR 和 *HiS* 在操作任务中的表现

FetchPush 和 FetchSlide [1] 是用于目标条件强化学习的两项基准操作任务。这两项任务基于现有机器人硬件，并使用 MuJoCo [35]进行仿真。在推物任务中，目标是将箱子移动到目标位置。机器人的夹爪被锁定，以防止抓取。滑动任务略有不同：其目标是将冰球移动到机器人够不到的目标位置，因此需要采用滑动策略。这两项任务都满足 HySR假设，因此，我们的方法可用于在真实机器人上学习策略。在这两项任务中，被操作物体遵循简单的动力学规律。机械臂

是刚性的，与物体接触时产生的微小扰动会由位置控制器迅速补偿。因此，从物体传递到机器人的力影响很小。

然而，这两项任务都使用动力学相对简单、能够精确仿真的机器人。仿真到现实训练已被证明能够成功解决这些任务 [1]，因此更适合这类系统。然而，当使用更复杂的机器人执行类似任务时，例如配备柔性部件或使用精度较低的位置控制器的机器人，仿真与现实之间更大的差异很可能会降低迁移后策略的性能。HySR 非常适合在这类机器人上学习策略。我们的实验表明，除了 HySR 带来的实际优势外，HiS 还能显著提高这些任务的样本效率，尤其是在与 HER 结合时。我们将在下一小节详细分析这些结果。此外，我们还将展示，在以仿真到现实方式学习策略时，HiS 仍能在仿真阶段加快学习速度。



(a) FetchPush (b) FetchSlide

图 5：在 FetchPush 和 FetchSlide 任务的学习过程中，前 1000 个回合内添加到强化学习缓冲区的成功轨迹数量。HER和HiS产生的成功轨迹数量，比单独使用 HER 或 HiS 多一个数量级。

C. 结合 *HER* 和 *HiS*

在第III-C节中，我们讨论了 HER 与 HiS 之间的异同。由于 FetchPush 和 FetchSlide 任务具有目标条件特性，因此适合用于实验研究 HER 与 HiS 之间的关系。图4展示了推和滑动任务的结果。为了应用 HiS，我们对虚拟物体进行了100个并行实例的仿真。对于虚拟轨迹选择，我们使用机器人移动物体的轨迹。与乒乓球任务类似，我们在 SAC 之上应用 HiS 和/或 HER，SAC 也作为基线。

在这两项任务中，HER 和 HiS 的学习速度都快于原始SAC。在推物任务中，HiS 的学习速度明显快于 HER，仅用 4k 回合便几乎完美地解决了任务；而在滑动任务中，HER 的表现则超过了 HiS。不过，在这两项任务中，HER 与 HiS 结合使用取得了最佳性能。与乒乓球任务类似，我们还发现，单独使用 HiS 或将其与 HER 结合使用，都能显著降低不同运行之间的方差。

事后方法似乎能够很好地发挥作用，一个关键原因在于它们会生成额外的正向标记经验。为说明这一现象，我们考察一个简单指标：在训练的前1000个回合中放入强化学习缓冲区的成功轨迹数量。这些轨迹要么通过同策略收集，要么在涉及HiS和/或HER时通过事后方式生成。我们过滤掉那些虽被标记为成功、但物体位置并未发生变化的无意义回合。这类回合所包含的有用信息很少。我们注意到，图5中被标记为成功的轨迹数量与图4中的结果之间存在很强的相关性。对于两个任务而言，它们显示出所比较算法之间相同的排序关系。这一发现很好地说明了HER和HiS如何相互补充。对于所研究的两个任务，在训练的初始阶段，HER和HiS的组合所生成的成功轨迹数量比单独使用HER或HiS多一个数量级。其背后的直觉是，HiS会生成大量事后轨迹，而HER会将这些轨迹标记为成功。

D. HiS 在完全仿真任务中的效率

HiS 最初旨在加速仿真与现实混合场景中的学习，因为真实数据获取成本高，而仿真数据成本低。因此，迄今为止的实验——包括仅在仿真中运行的实验——都是按照真实机器人本应执行的步数进行评估的。但在纯仿真任务中应用 HiS 是否有意义，例如在仿真到真实环境设置的仿真阶段？在这种场景下，目标通常不是尽量减少机器人步数，而是缩短总实际运行时间并减少计算量。与基线强化学习算法相比，HiS 会增加计算量，因为它会模拟额外的虚拟状态，依据选择标准对这些数据进行排序，然后将其送入算法的回放缓冲区。不过，如前文所示，HiS 也能提高学习过程中的样本效率。

表 I 展示了对 HiS 在仿真任务中效率的评估。我们将 HiS 与原始 SAC 进行比较，并将 HER 和 HiS 的组合与仅使用 HER 进行比较。为了比较总计算成本，我们将学习限制在单个 CPU 上（Intel Xeon W-2145，时钟频率为 3.70GHz）。计算量增加所带来的影响取决于任务以及额外虚拟物体的数量。对于相同步数，使用 HiS 时的墙钟时间在乒乓球任务中仅增加约 2%，而在操作任务中则增加约 60%。为了量化样本效率的提升，我们考察 HiS 达到原始SAC渐近性能或最终性能所需的机器人时间步数；同样，也考察 HER 和 HiS 达到 HER 渐近性能或最终性能所需的机器人时间步数。正如我们在图3和

图4所示，使用 HiS 能显著提升性能。因此，综合这两方面的效果，即使以墙钟时间而非机器人时间步数进行评估，HiS 仍然表现出色。总之，我们已经证明，对于只希望在仿真中优化策略的场景，HiS 会有所帮助。

因此，HiS 的应用范围不仅限于满足第 III-B 节所述所有 HySR 假设的任务。在仿真环境中，无需考虑任务不同部分的动力学复杂度差异。不过，HiS 还假设虚拟状态不会影响真实状态。对于纯仿真任务，可以对这些影响机器人状态的效应进行建模，但仍存在一个问题：多个彼此不影响的物体可能同时对机器人施加力。为克服这些差异，对于接触稀疏的任务，一种潜在的解决方案是将训练分为两个阶段：首先，使用 HiS 进行训练，并忽略这些差异。这个阶段有助于机器人与值得关注的物体发生接触。随后，再不使用 HiS 对这些接触进行微调。

V. 结论与未来工作

在这项工作中，我们提出了后见状态（HiS），这是一种适用于混合仿真与真实环境设置的样本高效训练方法，其中任务的虚拟部分包含多个实例，并与单个真实部分配对。HiS 通过有选择地挑选训练数据来利用这些额外数据。我们在多种仿真任务上评估了 HiS，并证明它能够提高样本效率，应用于一项复杂的真实世界肌肉机器人任务时亦是如此。我们还证明，结合 HER 和 HiS 比单独应用任一方法都能带来更好的性能。

我们工作的一项局限性是要求虚拟部分不会影响真实部分。在许多任务中，这种影响不能被忽略，尤其是在处理较重物体时。一种解决方法是将仿真中的力映射回真实机器人的关节力矩。

另一种策略是将学习分为多个阶段。对于满足该条件的阶段，可以使用 HiS 进行学习；否则，则完全采用真实训练。例如，对于抓取任务，机器人学习移动到夹爪接触物体之前的位置这一阶段，可以使用 HiS 进行学习。

对于需要长时交互的复杂任务，在整个回合中持续采样虚拟物体（而不是仅在开始时采样）可能会产生更好的结果。这种方法的优势在于，对机器人附近的物体进行重新采样，由于增加了接触，生成更重要经验的概率更高。

本工作的重点是在HySR设置下设计一种样本效率更高的训练方案。然而，如第IV-D节所示，HiS对纯仿真任务也可能有益。探索将HiS应用于仿真任务的最佳策略，尤其是应用于不满足HySR条件的任务，具有很大潜力。

表I：对HiS在纯仿真任务中效率的评估。这里存在两种影响。首先，单独使用HiS，或将HiS与HER结合使用，会增加完成相同步数所需的总实际运行时间。然而，其次，平均而言，HiS只需更少的时间步即可达到匹配性能（此处指实验期间评估的SAC的渐近性能或最终性能）；HER和HiS也只需更少的时间步即可达到匹配性能（此处指实验期间评估的HER的渐近性能或最终性能）。对于本工作研究的任务，第二种影响更大，因此即使完全在仿真中进行学习，达到相同性能所需的平均墙钟时间也更短。

	Task		
	推	滑动	乒乓球
HiS 与 SAC 的比较			
总实际运行时间	163.9 %	156.3 %	102.3 %
机器人的时间步数匹配性能	16.7 %	50.8 %	30.7 %
性能所需的墙钟时间匹配性能	27.4 %	79.4 %	31.4 %
HER+HiS 与 HER 的比较			
总实际运行时间	165.6 %	162.3 %	
机器人时间步，用于匹配性能	11.7 %	42.6 %	
墙钟时间，用于匹配性能	19.4 %	69.1 %	

未来的工作还可以探索将 HiS 与课程学习相结合，并使用数据增强。对虚拟物体进行变换和扰动等数据增强，有助于提升泛化能力，而且虚拟物体本身也可以符合物理规律。此外，还可以调整虚拟部分的物理定律，例如重力或时间流速，以改变整个任务的难度。通过这种方式，可以加入课程学习，例如在训练期间放慢或加快时间流速，使机器人更快到达高奖励区域。随后，在训练接近尾声时，逐步将时间流速调整至真实速度，可以提升性能向现实世界迁移的效果。

参考文献

[1]Marcin Andrychowicz、Filip Wolski、Alex Ray、Jonas Schneider、Rachel Fong、Peter Welinder、Bob Mc-Grew、Josh Tobin、OpenAI 的 Pieter Abbeel 和 Wojciech Zaremba。事后经验回放。神经信息处理系统进展，30，2017年。[2] 陈佳白、Peng Liu、Wei Zhao 和 Xianglong Tang。面向事后多目标强化学习的引导式目标生成。*Neurocomputing*，359：353–367，2019年。[3] Sofanit Wubeshet Beyene 和 Ji-Hyeong Han。用于元强化学习的带双缓冲区的优先事后回放。*Electronics*，11(24)：4192，2022年。

[4] Marc Brittain、Josh Bertram、Xuxi Yang 和 Peng Wei。优先序列经验回放。*arXiv* 预印本 *arXiv:1905.12726*，2019年。[5] Dieter B¨uchler、Simon Guist、Roberto Calandra、Vincent Berenz、Bernhard Sch¨olkopf 和 Jan Peters。使用肌肉机器人从零开始学习打乒乓球。*IEEE*机器人学汇刊，38(6)：3850–3860，2022年。[6] Dieter B¨uchler、Roberto Calandra 和 Jan Peters。学习控制肌肉机器人上的高速弹道运动。机器人与自主系统，159：104230，2023年。[7] Dieter B¨uchler、Heiko Ott 和 Jan Peters。用于机器人学习的轻量级气动肌肉机械臂。载于国际机器人与自动化会议（*ICRA*），斯德哥尔摩，2016年5月。doi：10.1109/icra.2016.7487599。[8] Shirli Di-Castro、Dotan Di Castro 和 Shie Mannor。仿真与现实：协同增效。神经信息处理系统进展，34：6868–6880，2021年。[9] Meng Fang、Cheng Zhou、Bei Shi、Boqing Gong、Jia Xu 和 Tong Zhang。DHER：面向动态目标的事后经验回放。载于国际学习表征会议，2018年9月。[10]Meng Fang、Tianyi Zhou、Yali Du、Lei Han 和 Zhengyou Zhang。课程引导的事后经验回放。神经信息处理系统进展，32，2019年。[11] William Fedus、Prajit Ramachandran、Rishabh Agarwal、Yoshua Bengio、Hugo Larochelle、Mark Rowland 和 Will Dabney。重新审视经验回放的基础。载于国际机器学习会议，第3061–3071页。PMLR，2020年。[12]Tuo mas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。软演员-评论家：带随机演员的离策略最大熵深度强化学习。载于国际机器学习会议，第1861–1870页。PMLR，2018年。[13] D. F. B. Haeufle、M. G¨unther、A. Bayer 和 S. Schmitt。带串联阻尼和偏心力–速度关系的 Hill 型肌肉模型。生物力学杂志，47(6)：1531–1536，2014年4月。ISSN 0021-9290。doi：10.1016/j.jbiomech.2014.02.009。[14]Irina Higgins、Arka Pal、Andrei Rusu、Loic Matthey、Christopher Burgess、Alexander Pritzel、Matthew Botvinick、Charles Blundell 和 Alexander Lerchner。Darla：改进强化学习中的零样本迁移。载于国际机器学习会议，第1480–1490页。PMLR，2017年。[15]Stephen James、Paul Wohlhart、Mrinal Kalakrishnan、Dmitry Kalashnikov、Alex Irpan、Julian Ibarz、Sergey Levine、Raia Hadsell 和 Konstantinos Bousmalis。通过仿真到仿真实现仿真到现实：利用随机到规范适应网络实现数据高效的机器人抓取。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12627–12637页，2019年。

[16]Katie Kang、Suneel Belkhale、Gregory Kahn、Pieter Abbeel 和 Sergey Levine。通过仿真实现泛化：将仿真数据与真实数据融入基于视觉的自主飞行深度强化学习。发表于 国际机器人与自动化会议（*ICRA*），第 6008–6014 页。IEEE，2019年。[17] Aviral Kumar、Justin Fu、Matthew Soh、George Tucker 和 Sergey Levine。通过自举误差降低稳定离策略 Q 学习。神经信息处理系统进展，第 32 卷，2019 年。[18] Aviral Kumar 、 Aurick Zhou 、George Tucker 和 Sergey Levine。面向离线强化学习的保守 Q 学习。神经信息处理系统进展，第 33 卷：1179–1191，2020年。[19]Sascha Lange、Thomas Gabel 和 Martin Riedmiller。批量强化学习。强化学习：最新进展，第 45–73 页，2012年。[20] Sergey Levine、Aviral Kumar、George Tucker 和 Justin Fu。离线强化学习：关于开放问题的教程、综述与观点。*arXiv:2005.01643 [cs,stat]*，2020年11月。网址：http://arxiv.org/abs/2005. 01643。arXiv：2005.01643。[21] Long-Ji Lin。基于强化学习、规划和教学的自我改进反应式智能体。机器学习，8(3)：293–321，1992年。出版方：Springer。[22] Peng Liu、陈佳白、Yingnan Zhao、Chenyao Bai、Wei Zhao 和 Xianglong Tang。为优先级后见之明强化学习生成注意力目标。知识型系统，203：106140，2020年。[23]Jeffrey Mahler、Matthew Matl、Vishal Satish、Michael Danielczuk、Bill DeRose、Stephen McKinley 和 Ken Goldberg。学习双手灵活的机器人抓取策略。*ScienceRobotics*，4(26)，2019年。[24]Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Alex Graves、Ioannis Antonoglou、Daan Wierstra 和 Martin Riedmiller。使用深度强化学习玩 Atari 游戏。*arXiv* 预印本 *arXiv:1312.5602*，2013年。网址：http://arxiv.org/abs/1312.5602。[25] Andrew W. Moore 和 Christopher G. Atkeson。优先级扫描：以更少数据和更短时间进行强化学习。机器学习，13(1)：103–130，1993年10月。国际标准连续出版物编号 1573-0565。doi：10.1007/BF00993104。网址：https://doi.org/10.1007/BF00993104。[26] Fabio Muratore、Fabio Ramos、Greg Turk、Wenhao Yu、Michael Gienger 和 Jan Peters。从随机化仿真中进行机器人学习：综述。*Frontiers in Roboticsand AI*，第 31 页，2022年。[27] Jing Peng 和 Ronald J. Williams。在 Dyna 框架中进行高效学习与规划。*AdaptiveBehavior*，1(4)：437–454，1993年3月。国际标准连续出版物编号 1059-7123。doi：10.1177/105971239300100403。出版方：SAGE Publications Ltd STM。[28]X. B. Peng、M. Andrychowicz、W. Zaremba 和 P. Abbeel。通过动力学随机化实现机器人控制的仿真到现实迁移。发表于 *2018*年 *IEEE* 国际机器人与自动化会议（*ICRA*）

1–8，2018年5月。doi：10.1109/ICRA.2018.8460528.[29]Rafael Rafailov、Tianhe Yu、Aravind Rajeswaran 和 Chelsea Finn。利用潜在空间模型从图像进行离线强化学习。载于动力学与控制学习，第1154–1168页。PMLR，2021年。[30]Antonin Raffin。RL Baselines3 Zoo。github.com/DLR-RM/rl-baselines3-zoo，2020年。[31] Paulo Rauber、Avinash Ummadisingu、Filipe Mutz 和 J¨urgen Schmidhuber。事后策略梯度。*arXiv:1711.06006 [cs]*,2019年2月。URL http://arxiv.org/abs/1711.06006。arXiv：1711.06006。[32] Zhizhou Ren、Kefan Dong、Yuan Zhou、Qiang Liu 和 Jian Peng。通过事后目标生成进行探索。神经信息处理系统进展，第32卷，2019年。[33] Tom Schaul、John Quan、Ioannis Antonoglou 和 David Silver。优先经验回放。*arXiv*预印本 *arXiv:1511.05952*，2015年。[34]Noah Y Siegel、Jost Tobias Springenberg、Felix Berkenkamp、Abbas Abdolmaleki、Michael Neunert、Thomas Lampe、Roland Hafner、Nicolas Heess 和 Martin Riedmiller。坚持有效的方法：离线强化学习的行为建模先验。*arXiv*预印本 *arXiv:2002.08396*，2020年。[35] Emanuel Todorov、Tom Erez 和 Yuval Tassa。MuJoCo：用于基于模型的控制的物理引擎。载于智能机器人与系统（*IROS*），*2012*年*IEEE/RSJ*国际会议，第5026–5033页。IEEE，2012年。[36] Yifan Wu、George Tucker 和 Ofir Nachum。行为正则化离线强化学习。*arXiv*预印本 *arXiv:1911.11361*，2019年。[37] Tianhe Yu、Garrett Thomas、Lantao Yu、Stefano Ermon、James Zou、Sergey Levine、Chelsea Finn 和 Tengyu Ma。MOPO：基于模型的离线策略优化。*arXiv:2005.13239 [cs, stat]*, 2020年5月。URL http://arxiv.org/abs/2005.13239。arXiv：2005.13239。[38] Tianhe Yu、Aviral Kumar、Rafael Rafailov、Aravind Ra- jeswaran、Sergey Levine 和 Chelsea Finn。COMBO：保守离线基于模型的策略优化。神经信息处理系统进展，第34卷：28954–28967，2021年。[39] Jingwei Zhang、Lei Tai、Peng Yun、Yufeng Xiong、Ming Liu、Joschka Boedecker 和 Wolfram Burgard。面向机器人的VR护目镜：用于视觉控制的真实到仿真的领域自适应。*IEEE*机器人与自动化快报，4（2）：1148–1155，2019年。[40] Rui Zhao 和 Volker Tresp。基于能量的事后经验优先级排序。载于机器人学习会议，第113–122页。PMLR，2018年。[41] Wei Zhu、Xian Guo、Dai Owaki、Kyo Kutsuzawa 和 Mitsuhiro Hayashibe。应用于仿生机器人的强化学习仿真到现实迁移技术综述。*IEEE*神经网络与学习系统汇刊，第1–16页，2021年。国际标准连续出版物编号2162-2388。doi：10.1109/TNNLS.2021.3112718。

VI. 附录

A. 实验细节

本节将详细介绍用于评估我们方法的任务。

1) 模拟乒乓球任务：我们在自定义的 MuJoCo 环境中进行实验，并使用 Hill 型肌肉模型模拟肌肉 [13]。与 B¨uchler 等人 [5],类似，我们会重放球轨迹直到发生接触，并在此后进行模拟。根据策略的不同，回合平均持续 37–39 个时间步。当球触碰球拍或超出球拍的可触及范围时，回合结束。

对于 HiS，主回合结束后，并行回合仍会继续。问题在于，我们采样的策略分布以包含主虚拟物体的状态为条件。在这些情况下，我们会使用其他虚拟物体之一来采样策略，直到所有虚拟物体对应的回合都结束。

2) 真实机器人乒乓球任务：我们保留了大部分任务细节，使其与 B¨uchler 等人 [5]的原始实验相似，并列出我们所做的调整。

- 我们将策略应用于新一代机器人硬件。
- 初始化过程有所不同：我们不再使用位置控制器，而是发送固定的目标压力，使机器人到达初始位置。
- 我们以 25 Hz 而不是 100Hz 的频率向机器人发送动作，因为我们发现这样使用 SAC 能显著改善结果。

在真实机器人上采集 15k 个回合大约需要 24 小时。

3) FetchPush 和 *FetchSlide*：每个回合的固定长度为 50 个时间步。为了应用 HiS，我们根据与采样主要物体相同的分布，额外采样虚拟物体。

B. 超参数

三个任务的超参数见表 II、III 和 IV。乒乓球任务的 SAC 超参数是在不使用 HiS 的仿真任务上调优的。Gym 机器人环境实验采用了在 [30] 中找到的超参数。

C. 消融研究：选择标准

使用奖励标准（主要选择成功轨迹）对 HiS 在我们的所有任务上都效果很好。由于 HER 会将轨迹重新标记为成功，因此在同时应用 HER 和 HiS 时，选择成功的 HiS 轨迹就不那么重要了。因此，在这种情况下，选择虚拟物体发生移动的轨迹效果更好。在这项工作中，我们着重研究了这些标准，以展示我们的方法。实验表明，时序差分误差标准有助于学习进展，但效果不如另外两个标准。表 V 展示了对不同选择标准的评估。需要注意的是，真实乒乓球任务仅评估了一轮训练。模拟乒乓球任务的结果是对采用不同随机种子的十轮训练取平均值，而滑动任务的结果则是对五轮训练取平均值。乒乓球任务（真实和

表 II：乒乓球超参数

	超参数	值
SAC	伽马	0.9999
	熵系数	0
	学习率	0.0003
	批量大小	256
	策略网络	MLP
	层数	1
	隐藏单元数	200
	梯度步数	500
	训练频率	1
	训练频率单位	回合
	缓冲区大小	5000000
	学习开始步数	10000
HiS	判据	每回合奖励
	k_c	3
	ψ_c	0.5

表 III：FetchPush 超参数

	超参数	值
SAC	伽马	0.95
	熵系数	auto
	学习率	0.001
	批量大小	256
	策略网络	MLP
	层数	2
	隐藏单元数	64
	梯度步数	1
	训练频率	1
	训练频率单位	step
	缓冲区大小	5000000
	学习开始步数	1000
HER	目标选择策略	未来
	采样目标数	4
HiS	判据	Δx 虚拟物体
	k_c	3
	ψ_c	0.02

表 IV：FetchSlide 超参数

	超参数	值
SAC	伽马	0.95
	熵系数	auto
	学习率	0.001
	批量大小	2048
	策略网络	MLP
	层数	3
	隐藏单元数	512
	梯度步数	1
	训练频率	1
	训练频率单位	step
	缓冲区大小	5000000
	学习开始步数	1000
HER	目标选择策略	未来
	采样目标数	4
HiS	判据	Δx 虚拟物体
	k_c	3
	ψ_c	0.02

模拟）在 15000 回合后进行评估，滑动任务则在 5000 回合后进行评估。

表 V：评估 HiS 和 HER+HiS 的不同选择标准对最终成功率的影响。

	Task		
	模拟乒乓球	真实乒乓球	滑动
Vanilla SAC	41.2 %	34.5 %	9.9 %
HiS奖励	70.6 %	44.7 %	33.7 %
HiS Δx 虚拟物体			25.5 %
HiS 时序差分误差	50.8 %	34.2 %	13.2 %
HER+HiS奖励			53.8 %
HER+HiS Δx 虚拟物体			64.5 %