arXiv:2305.01521v1 [cs.LG] 2 May 2023

释放长期新颖性探索中表征的力量

Alaa Saade [*1] **Steven Kapturowski** [*1] **Daniele Calandriello** [*] [1] **Charles Blundell** [1] **Pablo Sprechmann** [1] **Leopoldo Sarra** [2] **Oliver Groth** [1] **Michal Valko** [1] **Bilal Piot** [1]

摘要

我们提出了基于聚类的在线密度估计实现稳健探索（RECODE），这是一种用于基于新颖性的探索的非参数方法，它根据状态在选定嵌入空间中的相似度，估计状态簇的访问计数。通过将经典聚类方法适配到深度强化学习的非平稳环境中，RECODE能够在数千个回合中高效地跟踪状态访问计数。我们进一步提出了一种逆动力学损失的新型泛化方法，该方法利用掩码 Transformer 架构进行多步预测；结合 RECODE，该方法在DM-HARD-8的一系列具有挑战性的 3D 探索任务中实现了新的最先进水平。RECODE 在高难度 Atari 游戏的探索任务中同样达到了新的最先进水平，并且是首个在 *Pitfall!*中到达结束画面的智能体。

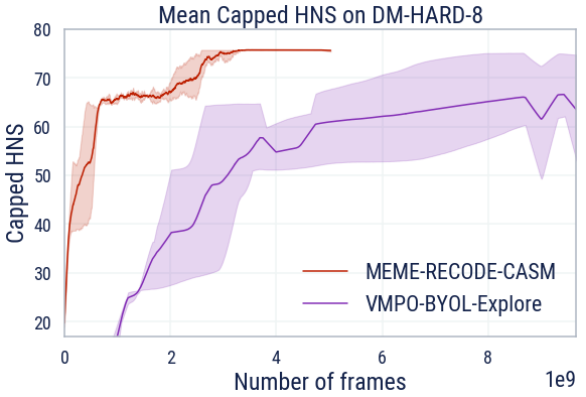

图 1： RECODE的一个关键结果是，它使我们能够利用更强大的状态表征进行长期新颖性估计；从而在具有挑战性的 3D任务套件 DM-HARD-8中实现新的最先进水平。

1. 引言

探索机制是强化学习（RL，Sutton 和 Barto，2018）智能体的关键组成部分，尤其是在稀疏奖励任务中，智能体需要执行很长的动作序列才能获得奖励。探索问题已在多臂老虎机（Lattimore 和 Szepesv´ari，2020）和马尔可夫决策过程（MDP，Puterman，1990年；Jaksch 等人，2010年）的背景下得到理论研究（Kearns 和 Singh，2002年；Azar 等人，2017年；Brafman 和 Tennenholtz，2003年；Auer 等人，2002年；Agrawal 和 Goyal，2012年；Audibert 等人，2010年；Jin 等人，2020年）。在 MDP 中实现高效探索的一种简单而理论可靠的方法，是将访问计数的递减函数用作探索奖励（Strehl 和 Littman，2008年；Azar 等人，2017年）。然而，对于大型或连续状态空间，这种方法会变得难以处理，因为智能体不太可能多次访问完全相同的状态，因此需要某种形式的状态间有意义的泛化。人们提出了多种访问计数和访问密度的近似方法与代理指标，使这种探索方式能够应用于复杂环境。在深度强化学习中，有两种取得部分成效的方法：参数化方法利用神经网络直接估计访问密度；非参数化方法则利用已访问状态的记忆来引导探索。

参数化方法要么使用密度模型显式估计访问计数（Bellemare 等人，2016年；Ostro- vski 等人，2017年），要么采用访问情况的代理指标，例如动力学模型的预测误差（Pathak 等人，2017年；Guo 等人，2022年），或者通过预测当前观测的特征来衡量，例如使用固定且随机初始化的神经网络生成的特征，如 RND （Burda 等人，2019年）。尽管这类方法在许多场景中为探索提供了强有力的基线方法（Burda 等人，2018年），但在持续学习场景中，它们容易受到深度学习常见问题的影响，尤其是适应缓慢和灾难性遗忘。通过梯度下降训练的参数化模型不适合快速适应（例如，在单个回合内），因为在探索奖励能够跟上之前，必须先更新状态表示。此外，灾难性遗忘使参数化方法容易受到

[*]贡献相同 [1]DeepMind [2]弗里德里希-亚历山大大学物理系¨，埃朗根-纽伦堡，实习于 DeepMind 期间完成。通信请联系：< alaas@deepmind.com，skapturowski@deepmind.com，dcalan- driello@deepmind.com>。

所谓“脱离”问题的影响，即算法会失去对值得探索区域的追踪（Ostrovski 等人，2017年）。非参数化方法依靠记忆来存储遇到的状态（Savinov 等人，2018年；Badia 等人，2020b）。这使方法能够响应最新经验，同时避免相互干扰地保留记忆。然而，由于计算约束，必须限制内存大小，因此需要对状态进行选择或采用聚合机制。

为了兼顾两者的优势，Never Give Up（NGU，Ba- dia 等人，2020b）将基于情节记忆的短期新颖性信号与基于 RND 的长期新颖性信号结合为单一的内在奖励。然而，同时估计两种不同的新颖性信号会增加算法的复杂性，并且需要进行仔细的参数调优。此外，正如Pathak 等人（2017年）所指出的，任何探索算法的最终效果都在很大程度上取决于所选择的状态表示。如果状态编码容易受到观测中的噪声或不可控特征的影响，就可能产生无关的新颖性信号，并阻碍在不同状态之间实现有意义的泛化。由于 NGU 依赖于 RND 进行表征，因此在存在噪声观测的情况下，它也会继承其编码缺陷，从而限制了该方法在随机或复杂环境中的适用性。

在本文中，我们通过将探索问题分解为两个相互解耦的子问题，来解决这些问题。首先，(i) **表征学习**：利用嵌入函数编码有意义的状态相似性，同时对观测中的不可控因素保持鲁棒性。其次，(ii) **计数估计**：提供长期的基于访问的探索奖励，同时保持对最新经验的响应能力。为解决 (i)，我们扩展了 Pathak 等人 (2017年) 提出的逆动力学模型，并利用掩码序列变换器（Devlin 等人，2018年）的优势，构建了一种编码器，能够在较长轨迹上生成丰富的表征，同时抑制对不可控特征的编码。我们将这种表征称为 CASM，即耦合动作-状态掩码。为实现 (ii)，我们提出了一种新颖的非参数方法，称为基于聚类的在线密度估计鲁棒探索（RECODE）。具体而言， RECODE 通过将密度估计和聚类技术应用于在线强化学习场景，在嵌入空间中估计软访问计数。我们的方法跟踪跨越数千个回合的交互历史，大幅提升了记忆容量；相比之下，现有的非参数探索方法通常只存储最近的历史，例如当前回合。在存在噪声的情况下，我们证明，该方法严格优于最先进的探索奖励，例如 NGU 或 RND。RECODE 在 Atari 上，取得了与最先进水平相当或更优的探索成果，并且是首个在*Pitfall!*中到达终幕的智能体。除二维场景外，我们的方法在三维场景中也表现出色领域中表现出色，并与 CASM结合，在具有挑战性的 DM-HARD-8 测试套件中创下新的最先进水平（图1）。

2. 背景

我们考虑智能体与其环境之间的离散时间交互（McCallum，1995年；Hutter，2004年；Hutter 等人，2009年；Daswani 等人，2013年）。在每个时间步 $t \in \mathbb{N}$，智能体接收一个观测 $o_t \in \mathcal{O}$，该观测部分反映了环境的底层状态 $s \in \mathcal{S}$，并生成一个动作 $a_t \in \mathcal{A}$。我们考虑策略 $\pi: \mathcal{O} \to \Delta_{\mathcal{A}}$，这些策略将观测映射为动作上的概率分布。最后，外在奖励函数 $r_e: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 将观测映射为标量反馈。该函数可以与内在奖励函数r_i 相结合，以鼓励探索行为，而这种行为可能无法仅通过 r_e 来诱导。

每个时间步提供给智能体的观测t 用于通过嵌入函数 f_θ 构建状态表征：$\mathcal{O} \to \mathcal{E}$，将 o_t 与向量$e_t = f_\theta(o_t)$关联起来。通常，嵌入空间 $\mathcal{E}$ 是向量空间 $\mathbb{R}^D$，其中 $D \in \mathbb{N}^*$ 为嵌入维度。学习 f_θ 的常见方法包括：对观测 o_t 使用自编码损失（Burda 等人，2018年）、使用逆向动力学损失（Pathak 等人，2017年）、在潜在层级使用多步预测损失（Guo 等人，2020年；2022年），或采用其他类似的表示学习方法。具体而言，Pathak 等人（2017年）和 Badia 等人（ 2020b ）强调了逆向动力学损失在过滤噪声或不可控特征方面的作用，例如《Pitfall!》中的屏幕死亡倒计时器。

在离散环境中，一种广受欢迎且有原则的探索方法是提供与访问计数成反比的内在奖励（Strehl 和 Littman，2008年；Azar 等人，2017年）。然而，在大型或连续空间中，同一状态可能很少被重复访问。Badia 等人(2020b)通过引入基于槽位的记忆 M解决了这一问题。该记忆存储当前回合中的所有过往嵌入，并以查询嵌入 $e_t = f_\theta(o_t)$与其在核函数 $\mathcal{K}$下的 k 个最近邻 $\text{Neigh}_k(e_t)$之间的相似度之和取代离散计数：：

$$r_t \propto \frac{1}{\sqrt{N(f_\theta(o_t))}} \approx \frac{1}{\sqrt{\sum_{m \in \text{Neigh}_k(e_t)} \mathcal{K}(e_t, m)}}. \quad (1)$$

由于在训练过程中存储完整的嵌入历史需要占用大量空间，这种基于槽位的记忆通常仅用于短期时间范围，并且在 NGU 中会在每个回合结束时重置。因此，基于槽位的记忆必须与另一种能够估计长期新颖性的独立机制结合使用，从而增加方法复杂度并带来权衡。接下来，我们将介绍一种简单且高效的基于槽位的记忆，它能够有效追踪数千个回合中的新颖性。

算法 1 RECODE

1：输入：嵌入 e**，记忆** $M = \{m_l\}_{l=1}^{|M|}$**，原子访问计数**$\{c_l\}_{i=l}^{|M|}$**，邻居数量** k**，用于判断候选新原子是否距离较远的相对容差** κ**，平方距离估计** d^2 **ema，** d^2 **ema 的衰减率** τ**，折扣因子**γ**，插入概率** η**，核函数** $\mathcal{K}$**，内在奖励常数** n_0**2：输出：更新后的记忆** $M = \{m_l\}_{l=1}^{|M|}$**，更新后的原子访问计数**$\{c_l\}_{i=l}^{|M|}$**，更新后的平方距离** d^2 **ema，内在奖励** r**3：计算** $N_{\mathcal{K}}(M, e) = \sum_{l=1}^{|M|}(1 + c_l)\ \mathcal{K}(m_l, e)$**；4：计算内在奖励** $r = \left(\sqrt{N_{\mathcal{K}}(M, e)} + n_0\right)^{-1}$

5 k e $_k$ $e = \{m_j\}_{j=1}^{k}$**：查找距**
6 d_{ema} $d^2 \leftarrow 1 - \tau\ d^2$ $-$
离该嵌入最近的原子：邻居 ( )：更新估计值：ema ( ) ema $+ {}_k^{\tau} \sum_{m \in}$**邻居**$_k(e)$ $\|m - e\|_2^2$ **7：折扣所有原子计数** $c_l \leftarrow \gamma\, c_l \quad \forall l \in \{1, \cdots, |M|\}$
8：查找最近原子 $m_\star = \textbf{arg min}_{m \in M, m \neq m_j} \|m - e\|_2$**9：在** $[0, 1]$ **中均匀采样一个实数：** $u \sim U[0, 1]$**10：如果**$\|m_\star - e\|_2^2 > \kappa\, d^2$ **ema 且** $u < \eta$**，则 11：以概率** $P(j) \propto 1/c_j^2$ **采样要移除的原子** m_j **12：查找距离** m_j **最近的原子** $m_\dagger$**：** $m_\dagger = \textbf{arg min}_{m \in M, m \neq m_j} \|m - m_j\|_2$**13：重新分配被移除原子的计数：** $c_\dagger \leftarrow c_j + c_\dagger$**14：以计数 1 将** e **插入索引** j**：** $m_j \leftarrow e, c_j \leftarrow 1$**15：否则**
16 $m_\star \leftarrow {}_{c_\star+1}^{c_\star} m_\star + {}_{c_\star+1}^{1} e$**：更新最近**
17 $c_\star \leftarrow c_\star + 1$
原子的位置：更新最近原子的计数 18：结束如果

3. RECODE

本节介绍我们的方法——基于聚类的在线密度估计鲁棒探索（RECODE），用于计算探索的内在奖励。RECODE 其灵感来自 NGU (Badia 等人，2020b）提出的奖励机制，但 NGU 会将单个嵌入后的观测存储在 M中，并通过周期性重置来限制空间复杂度；而 RECODE则通过在记忆中聚合相似观测来控制空间复杂度。这需要为记忆中的每个元素存储一个独立的计数器；新观测无需直接添加到记忆中，而通常会被分配给最近的现有元素，并使其计数器递增。由于计数器永不重置，且合并后的观测对嵌入空间有更好的覆盖范围，RECODE的记忆比简单的基于槽位的方法能持续更长时间，并在许多高难度探索环境中达到最先进的性能。它还将新颖性的估计简化为一种机制，而非像 NGU那样采用两种机制。此外， RECODE 架构具有很强的灵活性，能够轻松与各种强化学习智能体结合，尤其是与不同的表征学习方法结合。正如我们在实验中所展示的，能够更好地利用学习所得表征中的先验的方法（例如 RECODE），其性能优于那些需要直接在原始观测上估计新颖性的方法，例如 RND (以及由此衍生的 NGU）。下面我们将详细介绍实现方式，并在算法1中概述所用技术。

近似访问计数。我们的估计器基于一个有限的基于槽位的容器 $M = \{m_j\}_{j=1}^{|M|}$，其中 $|M|$ 是内存大小。我们将 $m_j \in \mathcal{E}$ 称为原子，因为它们不必像 Badia 等人(2020b;a) 中那样对应单个嵌入。我们还单独存储一个计数向量 c，使得c_i 是对 m_i的访问计数的估计。具体而言，c_i 不仅反映对 m_i 的访问次数，还捕获了之前与其足够接近的任何访问。

给定一个新的嵌入 e，我们根据相似性核估算其软访问计数 (算法 1：第 3–4 行)，即记忆中所有接近 e 的原子的加权和：

$$N_{\mathcal{K}}(M, e) = \sum\nolimits_l (1 + c_l)\mathcal{K}(m_l, e; d_{\text{ema}}). \tag{2}$$

具体来说，我们选择如下核函数：

$$\mathcal{K}(m_l, e) = \frac{1}{1 + \frac{\|e - m_l\|_2^2}{\epsilon d_{\text{ema}}^2}} \mathbb{1}_{\left\{\|e - m_l\|_2^2 < d_{\text{ema}}^2\right\}}, \tag{3}$$

其中 $\epsilon \in \mathbb{R}_+$ 是一个固定参数。此定义与 Badia 等人 (2020b) 类似，但我们将他们对 e的前k 个邻居求和，改为对距离 d_{ema} 不超过 e的所有原子求和。这一选择避免了使用 k-近邻方法计算计数时可能出现的反直觉现象。具体来说，给定嵌入的软访问计数理应在将其添加到记忆后增加。然而，向记忆中添加原子可能会改变 k-近邻列表。如果从列表中移出的原子具有较大的计数，这实际上可能会降低附近软访问计数的估算值，而不是使其增加。相反，我们的方法不受这一问题影响。

最后，我们返回内在奖励 r，如方程1所示，但在分母中加入一个小常数 n_0 以确保数值稳定性，并按照 r 中的方法，通过内在奖励标准差的运行估计值进行归一化（Burda 等人，2019年）。

构建记忆。构建记忆时，我们采用了用于估计软访问计数的同一聚合原理。具体而言，我们将原子 m_i 与观测聚类后得到的质心相对应。我们借鉴了经典聚类和密度估计方法，例如 k-均值或 DP-均值（Kulis 和 Jordan，2011年）；并对其进行调整，以应对大规模强化学习场景带来的挑战：内存大小有限，无法存储所有历史数据；观测按顺序到达，其分布并非平稳；用于嵌入观测的表征也会随时间变化。下面我们将介绍 RECODE 如何解决这些问题。

在每一步，我们都必须更新记忆 M 以反映观察到 e 对软访问计数的影响，同时保持其大小 $|M|$ 不变。直观而言，我们可以想到两种可能的方法：要么用新的嵌入替换一个已有原子，要么更新某个已有原子的位置和计数，使其更接近 e。令 $m_\star$ 为 e 在 M中最近的原子。我们采用以下规则（算法 1：L8-18）来整合新的

将嵌入存入记忆，这与 DP-means 聚类算法密切相关（Kulis 和 Jordan，2011年）：

- 如果 e 满足 $||m_\star - e||^2 < \kappa d^2_{\text{ema}}$，其中 d_{ema} 是一个自适应阈值，而 $\kappa > 0$ 是一个固定参数，则会将其“分配”给由 $m_\star$ 编码的簇，并根据现有嵌入与新嵌入的计数的凸组合，更新$m_\star$的值：

$$m_\star \longleftarrow \frac{c_\star}{c_\star + 1} m_\star + \frac{1}{c_\star + 1} e \tag{4}$$

  其权重 $c_\star$ 也会增加 1；

- 如果附近没有原子，我们会以概率 η 抛硬币，随机决定是否创建一个新原子。如果抛硬币成功，我们会将新嵌入作为一个新原子引入，同时按照下一段所述的流程移除一个现有原子。如果抛硬币失败，我们则按照公式4更新。$m_\star$

引入随机抛硬币是为了提高聚类算法对噪声的稳定性。具体而言，只有当远离记忆的嵌入平均出现 $1/\eta$ 次后，才会将其插入，从而减少偶发离群点带来的影响。同时，当一个远处的嵌入多次出现，并因此与软访问计数相关时，将其加入以提升记忆的覆盖范围的可能性很高。但为确保内存大小有限，必须移除一个现有原子。我们研究了三种基于簇数量 m_i选择要移除的原子 c_i的不同策略：(a) 以概率 $\propto \frac{1}{c_i^2}$移除；(b) 以概率 $\propto \frac{1}{c_i}$移除；(c) 移除 c_i最小的原子。附录 C.2 (图 8 和 9)中对移除策略的消融研究通过实证表明，对于我们考虑的设置，策略 (a) 效果最好，但结果总体上也对具体的策略选择相当稳健。

每当一个原子 i 被移除时，其计数 c_i 会重新分配到其最近邻原子的计数中，以保持记忆中的总计数不变。更新规则 RECODE也可以从理论角度解释为潜在概率聚类模型中的一种近似推断方案。附录C中进一步详细说明了我们的更新规则与 DP-means 算法之间的联系。

处理非平稳分布。嵌入观测之间的距离尺度可能在不同环境之间以及训练过程中发生显著变化，这是策略和嵌入函数均存在非平稳性的结果 f_θ。为处理这一问题，我们采用了 NGU(Badia 等人，2020b) 中的自适应带宽机制。具体而言，每当接收到新的嵌入 d^2_{ema} 时，我们就更新核函数参数 e，

根据新嵌入与以下对象之间的均方距离 k个最近的现有原子（算法1：L5-6）。为了让 d_{ema}能够更快速地适应，我们将 NGU 中使用的运行平均值替换为参数为 τ 的指数移动平均。

然而，我们注意到，这种机制不足以应对 f_θ 在较长时间尺度上的非平稳性。原始的 NGU 记忆受此问题的影响并不大，因为它会在每个回合后重置，因此表征几乎没有时间发生显著变化。然而，在 RECODE中，如果旧簇未能得到频繁更新，这些不断变化的表征最终可能会破坏长期记忆。具体来说，在一种表征下，某个原子可能获得较高的计数，但在另一种表征下却变得无法访问（因而毫无用处），同时仍不太可能被移除。为解决这一问题，我们加入一个衰减常数γ，在每一步按 $c_i \longleftarrow \gamma c_i$折扣记忆中所有原子的计数，其中 $\gamma < 1$ （算法1：L7）。这会逐渐降低陈旧原子的计数，并提高它们在后续插入过程中被移除的可能性：长时间没有新观测被“分配”给它们的簇最终会被替换。与此同时，相关簇的存续时间远长于以往方法。图3展示了一位已学会到达*Pitfall!*结束画面的智能体记忆中所含簇的簇龄直方图。图3 中的红线表示单个回合内可能达到的最大步数，该值由*Pitfall!*的游戏内死亡计时器限定；对于每回合都会重置记忆的方法，这也代表其最大记忆时域。可以看到，大多数簇的簇龄都远超一个回合，最早的记忆甚至可追溯到数千个回合之前。我们在附录C.2 （图10至12和14）中更详细地讨论折扣的影响。

值得注意的是，与 NGU 中每个执行器都维护各自的记忆副本不同， RECODE 在分布式智能体的所有执行器之间共享记忆，这大大提高了每个原子的更新频率，从而减少了记忆更新之间的表征漂移。

4. 表示学习方法

如第 2 节所述，嵌入函数 $f_\theta : \mathcal{O} \to \mathcal{E}$ 会对探索质量产生显著影响；许多不同的表示学习技术都已在这一背景下得到研究（Burda 等人，2018年；Guo 等人，2020年；2022年；2021年；Erraqabi 等人，2021年）。下文我们聚焦于动作预测嵌入，首先介绍标准的一步预测公式（Pathak 等人，2017年；Badia 等人，2020b;a）。嵌入函数 f_θ 由一个前馈神经网络参数化，该网络以 o_t（时刻 t的观测）作为输入。我们定义一个分类器 g_ϕ，给定以下内容的嵌入：

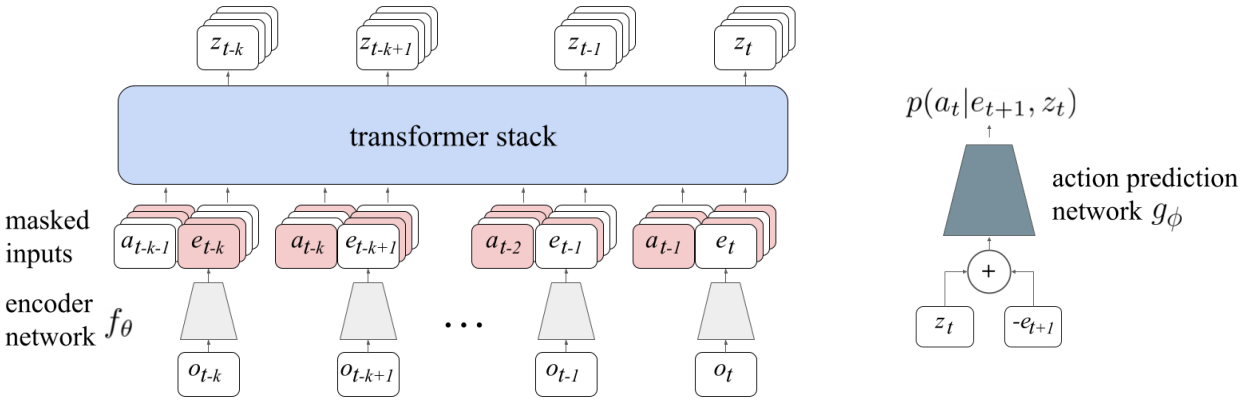

图2：用于在部分可观测环境中学习表征的耦合动作-状态掩码（CASM）架构。Transformer 将由动作 a_i 和嵌入后的观测$e_i = f_\theta(o_i)$组成、长度为 k 的掩码序列作为输入，并尝试在输出中重构缺失的嵌入。在时间步$t-1$和 t 处的重构嵌入随后用于构建一个 1 步动作预测分类器。用作RECODE 表征的嵌入函数为 f_θ。掩码输入以粉色阴影标示；训练期间会采样 $N = 4$掩码序列（图中由 a、 e 和 z 的堆叠表示）。

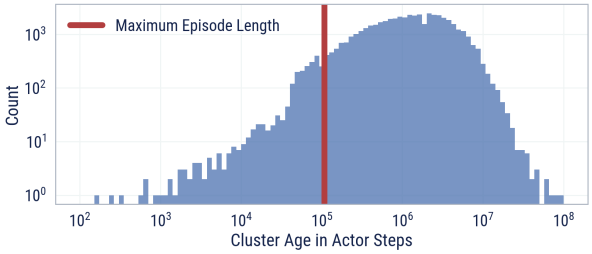

图3：智能体学习抵达*Pitfall!*的结束画面时，其记忆中的内容。

两个连续观测 $f_\theta(o_t)$ $f_\theta(o_{t+1})$，输出一个估计值 $p_{\theta,\ \phi}($ $a_t|o_t,\ o_{t+1})\ =\ g_\phi\,(f_\theta(o_t)\ f_\theta(o_{t+1}))$，表示给定两个连续观测（$o_t$，$o_{t+1}$）时采取某个动作的概率。随后通过最小化负对数似然的期望，对 f_θ 和 g_ϕ 进行联合训练：

$$\min_{\theta,\phi} \mathcal{L}(\theta,\phi)(a_t) = -\ln(p_{\theta,\phi}(a_t|o_t, o_{t+1}))\,, \qquad (5)$$

其中，a_t 是 o_t 和 o_{t+1}之间采取的真实动作。这些嵌入已被证明有助于处理观测中包含许多不可控特征的环境（Badia 等人，2020b），例如 Atari 的*Pitfall!*；在这些环境中，即使智能体静止不动，观测也包含许多虚假的新颖性来源。

虽然 RECODE可以与任意嵌入函数搭配使用，例如针对特定领域设计的嵌入函数，但选择有意义的表征也是影响最终性能的关键因素。标准的一步动作预测方法的一个主要缺点是，预测任务过于简单，通常只需学习高度局部化的低级特征（例如，单个物体在状态转移下如何移动）即可解决，而这些特征未必能反映全局环境结构。相比之下，理想的嵌入应捕捉有关环境的更高层次信息，例如智能体位置或先前观测到的地标的相对位置；这些信息可能不会同时出现在单个观测o_t 和 o_{t+1}中。要实现这一点，可能需要更长的时间步上下文。

然而，如果我们直接将完整轨迹提供给预测器，预测任务就会变得更加容易。为了解决这一局限，我们提出使用一种随机上下文 h_t，在每个时间步 $k \le t$，要么提供$f_\theta(o_k)$，要么提供 a_{k-1}。[1] 核心思路是，模型仍然可以通过在给定 (h_{t-1}, a_{t-1}) 的情况下，从 $f_\theta(o_t)$ 推断缺失信息来学习预测 a_t。这样一来，动作预测器就不会仅仅依赖 $f_\theta(o_t)$ 提供的信息，还会在 h_t中构建冗余表征。

从实现角度来看，我们首先构建一个由观测嵌入和动作组成的序列：$(f_\theta(o_0)\ a_0, f_\theta(o_1) \ldots, a_{t-1}, f_\theta(o_t))$。接着，受掩码语言模型（Devlin 等人，2018年）的启发，在每个时间步 t，我们随机将 $f_\theta(o_t)$ 或 a_t 替换为表示缺失信息的特殊标记。随后将这些掩码序列输入因果掩码 Transformer，其输出再投影到嵌入的维度（$\dim\ z_t = \dim\ f_\theta(o_t)$），并将差值

[1]我们避免同时对 $f_\theta(o_k)$ 和 a_{k-1} 进行掩蔽，因为这会增加预测任务无法确定的似然。

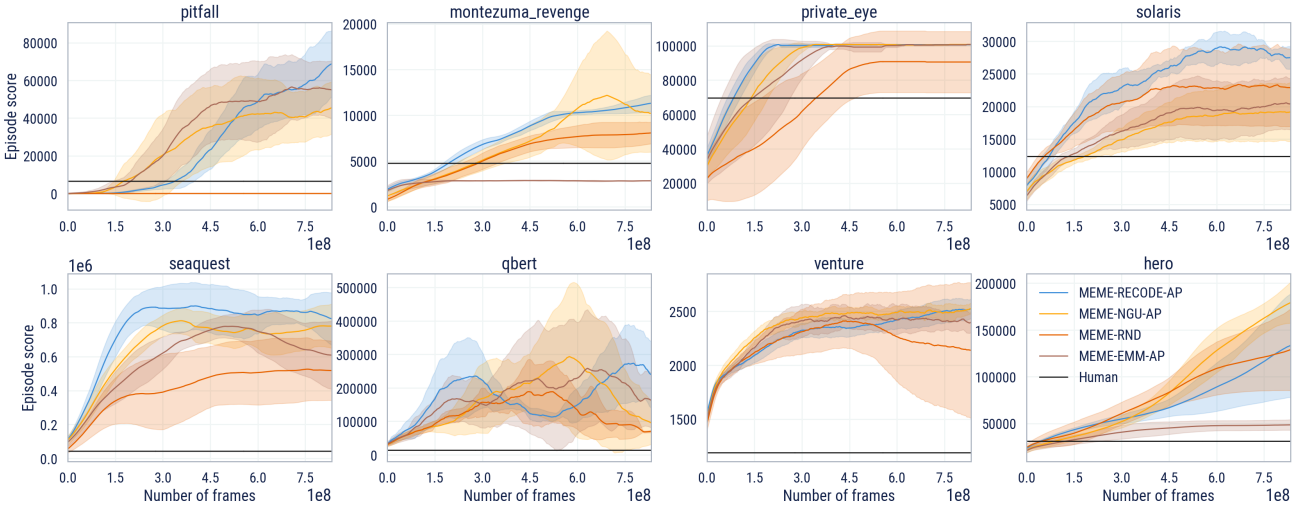

图 4：比较 RECODE 与 Atari 的高难度探索游戏中的其他探索奖励。所有智能体均基于 MEME，并使用相同的表征学习机制（AP）。请注意，*Q*bert* 中的高方差源于游戏中的一个漏洞；利用该漏洞可以获得显著更高的分数（Chrabaszcz 等人，2018年）。

输入最终的多层感知机分类器 g_ϕ。与1步动作预测一样，我们使用最大似然训练表征。下文将此方法称为耦合动作-状态掩码（CASM）。训练期间，我们从每条轨迹中随机采样多个掩码序列（$N = 4$），以帮助降低梯度方差。请注意，我们提供给 RECODE的最终嵌入是 $e_t = f_\theta(o_t)$，即 Transformer 的输入，以避免泄露有关智能体轨迹的信息。图2展示了该架构的示意图。

5. 实验

在本节中，我们通过实验验证所提出方法在两个用于探索的成熟基准上的有效性，分别针对二维和三维环境：一组 Atari 学习环境（ALE，Bellemare 等人，2013年）子集，包含八款游戏，例如 Pitfall 和 Montezuma's Revenge，这些游戏被视为困难探索问题（Bellemare 等人，2016年）；以及 DM-HARD-8 （Gulcehre 等人，2019年），一套部分可观测的三维游戏。所有游戏都带来了显著的探索挑战，例如极长的时间跨度（$\mathcal{O}$10K 步）、需要回溯、稀疏奖励、物体交互以及程序化环境生成。我们的方法在两个基准上都达到了最先进水平，还解决了此前两个尚未解决的游戏：在 Atari 的 *Pitfall!* 中，我们的方法首次到达结束画面；在 DM-HARD-8的 *Push Block* 中，我们首次实现了超人类性能。我们还进行了一系列消融实验，以进一步了解表征学习机制的影响，以及其对噪声观测的鲁棒性。

在以下实验（以及附录 K）中评估的所有候选架构均由三个主要模块组成：（1）负责收集观测、更新策略等强化学习核心任务的基础智能体；（2）负责生成探索奖励的算法；（3）负责学习有意义的观测表征的嵌入机制。我们的命名方式反映了模块的选择，即 AGENT-EXPLORATION-EMBEDDING。例如， MEME智能体在 Kapturowski 等人（2022年）中被记为MEME-NGU-AP。在所有实验中，我们都使用 MEME 智能体，但会改变探索机制和表征机制。探索方面，我们考察 EMM（纯情节记忆）、 NGU和 RECODE ；表征方面，我们尝试AP 和 CASM。附录E列出了所有智能体和基线方法的完整超参数。

5.1. Atari

Belle- mare 等人所确定的 Atari 困难探索子集（2016年）在优化时域方面提出了相当大的挑战，回合持续时间最长可达 27,000步，使用标准的四步动作重复。此外，奖励在尺度和密度方面都存在显著差异。在 Atari领域的所有实验中，我们将智能体的内存大小设为 $5 \cdot 10^4$ 个原子。我们遵循先前工作（Mnih 等人，2015年；Van Hasselt 等人，2016年）所确立的实验设置评估所有智能体：使用 30 次随机无操作，不使用"粘性动作"（Machado 等人，2018年），并对六个随机种子的性能取平均值。

我们比较了使用我们的探索奖励 RECODE 与其他方法所获得的游戏分数，同时保持智能体

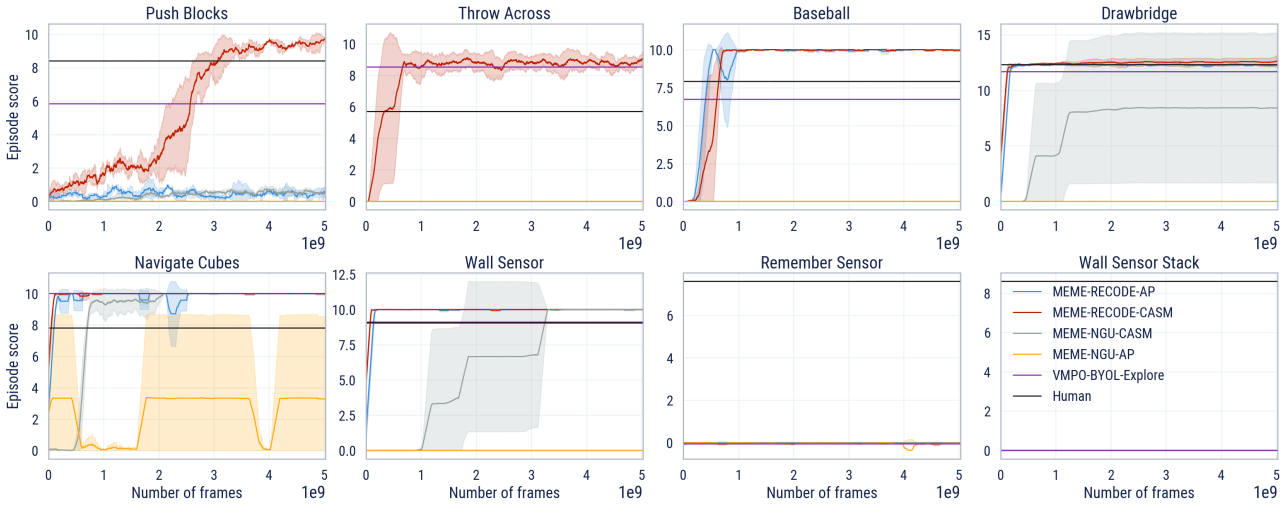

图 5：在 RECODE 上， NGU 和 BYOL-Explore 在 DM-HARD-8的单任务版本中的性能。BYOL-Explore 的结果对应于 Guo 等人 （2022年）报告的最终性能，训练共进行了 1e10 环境帧。所有结果均为 3 个随机种子的平均值。

架构和表征机制不变。图4中的结果表明，与 MEME-NGU-AP相比，我们的方法使用概念上更简单的探索奖励，在全部八款游戏中均达到了最先进水平，性能超越人类。图 MEME-EMM-AP和 MEME-RND 消融实验揭示了单独使用短期新颖性和长期新颖性各自的不足。 EMM 单独使用无法解决蒙特祖玛的复仇，因为这款游戏需要长期记忆。反之， RND 单独使用无法解决*Pitfall!*，因为观测中存在许多不可控特征，而且它无法利用 AP 嵌入。相比之下， RECODE能够利用 AP表征实现短期和长期新颖性，这是因为基于聚类的记忆能够整合长时间跨度的信息，从而使单一内在奖励即可解决这两款游戏。

5.2. DM-HARD-8

DM-HARD-8 （Gulcehre 等人，2019年）由八项探索任务组成，旨在挑战强化学习智能体：在具有部分可观测性、连续控制、稀疏奖励以及高度多变初始条件的程序生成三维世界中完成任务。每项任务都要求智能体与环境中的特定物体互动，以到达一个能提供奖励的大苹果（示例见附录图16 ）。程序生成机制会在每个回合随机改变物体的形状、颜色和位置。在我们针对该 DM-HARD-8 领域开展的所有实验中，我们将智能体的内存大小设为 $2 \cdot 10^5$ 个原子。我们还在这些实验中默认使用能力更强的 CASM 表征，而非 AP ，并在第5.3节中通过消融实验展示表征的影响。所有报告的性能均针对评估结果为三个随机种子的平均值。

我们在该领域将 RECODE 与 NGU 以及近期提出的 BYOL-Explore （Guo 等人，2022年）进行了比较。图5 中的结果表明，我们的方法在八项任务中有六项达到了超越人类的性能，在该基准测试中创下新的最先进水平；这也是人类基准首次在*Push Block*任务上被超越。为评估表征的贡献，我们还运行了 NGU 的一个版本，该版本使用能力更强的 CASM 表征，而非其默认的 AP表征。将 AP 切换为 CASM 显著提升了 NGU的性能，并凸显了在该领域的表征机制中纳入更长轨迹的信息、以应对部分可观测性这一挑战的重要性。然而，只有 RECODE 能够充分发挥 CASM所赋予的表征能力，因为它能够利用这种能力，同时获取短期和长期新颖性奖励。

5.3. 消融实验

作为实验的最后部分，我们进行了两项消融研究，以评估我们的方法对噪声观测的存在以及底层表征机制选择的敏感程度。

对观测噪声的鲁棒性。观测空间中的噪声是探索方法必须克服的最显著对抗性条件之一。要在任何实际场景中发挥作用，就必须考虑传感器总是不完美这一情况。“噪声电视问题”（Schmidhuber，2010年；Pathak 等人，2017年）是一个常见的比喻，用来描述探索方法陷入

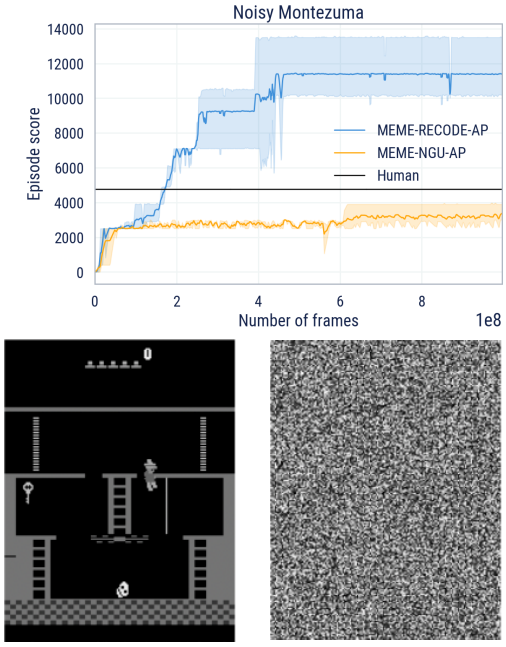

图 6：对观测噪声的鲁棒性。上： RECODE 与 NGU 在含噪蒙特祖玛上的性能比较。下：含噪蒙特祖玛中的一帧，噪声与原始帧拼接在一起。

将噪声预测为毫无意义的新颖性信号这一困境。为了评估我们的方法对观测噪声的鲁棒性，我们通过在图像高度维度上，将包含白噪声的帧（范围为$[0, 255]$）与游戏原始的 210×160 灰度观测拼接起来，构造了一个带噪声的蒙特祖玛的复仇版本。我们在此设置下使用相同的RECODE 对比 NGU，并采用相同的 AP 骨干网络，在表征层面抑制不可控噪声，并评估探索奖励对噪声的敏感性。该实验结果见图6。我们发现，存在噪声时， MEME-NGU-AP 的性能显著下降。这可能是因为 NGU 依赖 RND 来计算长期探索奖励，而在存在不可控噪声时，这会退化为随机探索（Kapturowski 等人，2018年）。这实际上将基线方法限制为在单个回合内进行短期探索。相比之下， RECODE的平均性能并未显著下降，取得的分数与图4相近，但方差更大。

利用不同的表征机制。 在 DM-HARD-8 上的实验表明，在更复杂的部分可观测环境中，采用更强大的表征学习技术至关重要。然而，尽管更丰富的表征通常能为下游任务学习带来普遍提升，但它本身无法解决探索问题。在图 7中，我们比较了 AP 和 CASM 对汇总性能的贡献

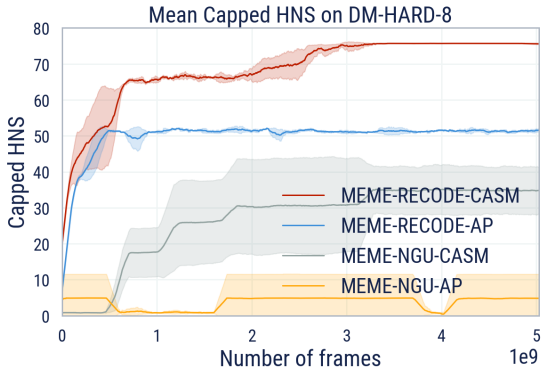

图 7：使用封顶的人类归一化分数，比较 RECODE 和 NGU 在所有 DM-HARD-8 任务中利用 AP 和CASM表征的表现。

中 NGU 和 RECODE 在 DM-HARD-8上的表现。结果始终表明， CASM 在这一领域是优于 AP 的表征，使用这两种探索方法都能显著提升性能。然而，对于两种表征，RECODE 都优于NGU，这表明，利用表征能力来提取短期和长期新颖性信号，是我们提出的方法的一项关键优势。

6. 结论

本文介绍了基于聚类的在线密度估计鲁棒探索（RECODE），这是一种原理严谨且概念简单的探索奖励，适用于深度强化学习智能体，可通过基于槽位的记忆估计访问计数，实现鲁棒探索。RECODE 通过在线聚类机制将有效记忆跨度扩大数个数量级，优于以往的非参数探索方法。我们的方法在两个成熟的探索基准上取得了任务性能的最先进水平：Atari 的困难探索子集和 DM-HARD-8。此外，它还是首个在时限内到达 *Pitfall!* 结束画面的智能体，这体现了 RECODE高效利用长期（即以往经验）和短期（即回合内）新颖性信号的能力。除了基准测试之外，RECODE的性能也不会受到噪声观测的影响——这是一种对抗性条件，会显著降低 RND 和 NGU等先前方法的性能。此外，我们还表明，该方法不受所选观测嵌入表征技术的限制，并且能够随着表征能力的增强而良好扩展，例如采用我们提出的 CASM 架构这类多步序列预测 Transformer。最后，我们得出结论： RECODE可作为一种简单而鲁棒的即插即用探索方法，与任何强化学习智能体和表征学习方法兼容，并能将表征学习的改进直接转化为探索性能的提升。

参考文献

Agrawal, S. 和 Goyal, N. 多臂老虎机问题中汤普森采样的分析。载于学习理论会议，第 39–1 页。机器学习研究期刊研讨会与会议论文集，2012年。

Audibert, J.-Y., Bubeck, S., 和 Munos, R. 多臂老虎机中的最佳臂识别。载于*COLT*，第 41–53 页。Citeseer，2010年。

Auer, P., Cesa-Bianchi, N., 和 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。

Azar, M. G., Osband, I., 和 Munos, R. 强化学习的极小极大遗憾界。载于第 *34* 届国际机器学习会议论文集——第 *70* 卷，第 263–272 页。机器学习研究期刊，2017年。

Badia, A. P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., 和 Blundell, C. Agent57：超越 Atari人类基准。载于国际机器学习会议，第 507–517 页。PMLR，2020a。

Badia, A. P., Sprechmann, P., Vitvitskyi, A., Guo, D., Piot, B., Kapturowski, S., Tieleman, O., Arjovsky, M., Pritzel, A., Bolt, A. 和 Blundell, C. Never Give Up：学习定向探索策略。发表于国际学习表征会议，2020b。

Barreto, A. M., Precup, D. 和 Pineau, J. 实用的基于核函数的强化学习。《机器学习研究杂志》，17(1):2372–2441，2016年。

Bellemare, M., Veness, J., 和 Talvitie, E. 跳跃上下文树切换。载于国际机器学习会议，第 1458–1466 页。PMLR，2014 年。

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D. 和 Munos, R. 统一基于计数的探索与内在动机。载于《神经信息处理系统进展》，第 1471–1479 页，2016年。

Bellemare, M. G., Naddaf, Y., Veness, J. 和 Bowling, M. 街机学习环境：通用智能体的评估平台。《人工智能研究杂志》，47：253–279，2013年。

Brafman, R. 和 Tennenholtz, M. R-max——一种用于近最优强化学习的通用多项式时间算法。机器学习研究杂志，3：213–231，2003年。

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T. 和 Efros, A. A. 好奇心驱动学习的大规模研究。*arXiv* 预印本*arXiv:1808.04355*，2018年。

Burda, Y., Edwards, H., Storkey, A. 和 Klimov, O. 通过随机网络蒸馏进行探索。收录于第七届国际学习表征会议，第 1–17 页，2019年。Chrabaszcz, P., Loshchilov, I. 和 Hutter, F. 回归基础：评测用于玩 Atari 的经典进化策略。*arXiv* 预印本 *arXiv:1802.08842*，2018年。

Daswani, M.、Sunehag, P. 和 Hutter, M. 基于历史的强化学习的 Q学习。在亚洲机器学习会议，第 213–228 页。PMLR，2013年。

Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本*arXiv:1810.04805*，2018年。

Domingues, O. D., Menard, P., Pirotta, M., Kaufmann, E., and Valko, M. 基于核的强化学习：有限时间分析。载于第*38*届国际机器学习会议论文集，第139卷《机器学习研究论文集》，第2783–2792页。PMLR，2021a。

Domingues, O. D., Tallec, C., Munos, R., and Valko, M. 面向无奖励探索的学习到的表示上的基于密度的奖励，应用于深度强化学习。载于*ICML 2021*年研讨会，2021b。

Erraqabi, A., Zhao, M., Machado, M. C., Bengio, Y., Sukhbaatar, S., Denoyer, L., and Lazaric, A. 强化学习中的探索驱动型表征学习。载于*ICML 2021*年无监督强化学习研讨会，2021年。

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., 等。IMPALA：采用重要性加权的 actor-学习器架构实现可扩展的分布式深度强化学习。载于国际机器学习会议，第 1407–1416 页。PMLR，2018年。

Gulcehre, C., Le Paine, T., Shahriari, B., Denil, M., Hoffman, M., Soyer, H., Tanburn, R., Kapturowski, S., Rabinowitz, N., Williams, D., 等。有效利用示范来解决困难探索问题。载于国际学习表征会议，2019年。

Guo, Z. D., Pires, B. A., Piot, B., Grill, J.-B., Altch´e, F., Munos, R. 和 Azar, M. G. 用于多任务强化学习的自举潜在预测表征。发表于国际机器学习会议，第 3875–3886 页。PMLR，2020年。

Guo, Z. D., Azar, M. G., Saade, A., Thakoor, S., Piot, B., Pires, B. A., Valko, M., Mesnard, T., Lattimore, T. 和 Munos, R. 几何熵探索。*arXiv* 预印本*arXiv:2101.02055*，2021年。

Guo, Z. D., Thakoor, S., P˘ıslar, M., Pires, B.A., Altch´e, F., Tallec, C., Saade, A., Calandriello, D., Grill, J.-B., Tang, Y., 等。Byol-explore：通过自举预测进行探索。*arXiv* 预印本*arXiv:2206.08332*，2022年。

Hazan, E., Kakade, S., Singh, K. 和 Van Soest, A. 可证明高效的最大熵探索。载于国际机器学习会议，第2681–2691 页，2019年。

Hutter, M. 通用人工智能：基于算法概率的序列决策. SpringerScience &BusinessMedia,2004.

Hutter, M. 等人。特征强化学习：第一部分。非结构化马尔可夫决策过程。De Gruyter Open，2009年。

Jaksch, T.、Ortner, R. 和 Auer, P. 强化学习的近最优遗憾界。机器学习研究期刊，11（4月）：1563–1600，2010年。

Jin, C.、Krishnamurthy, A.、Simchowitz, M. 和 Yu, T. 强化学习的无奖励探索。载于国际机器学习会议，第 4870–4879 页。PMLR，2020年。

Kaelbling, L. P.、Littman, M. L. 和 Cassandra, A. R. 部分可观测随机领域中的规划与行动。人工智能，101：99–134，1998年5月。ISSN00043702。doi：10.1016/s0004-3702(98)00023-x。

Kapturowski, S.、Ostrovski, G.、Quan, J.、Munos, R. 和 Dabney, W. 分布式强化学习中的循环经验回放。载于国际学习表征会议，2018年。

Kapturowski, S., Campos, V., Jiang,R., Raki´cevi´c, N., van Hasselt, H., Blundell, C., and Badia, A. P. 人类水平雅达利速度提升 200 倍。 *arXiv* 预印本*arXiv:2209.07550*，2022年。

Kearns, M. and Singh, S. 多项式时间内的近最优强化学习。机器学习，49(2-3)：209–232，2002年。

Kulis, B. and Jordan, M. I. 重新审视 k 均值：通过贝叶斯非参数方法实现的新算法。 *arXiv* 预印本*arXiv:1111.0352*，2011年。

Kveton, B. and Theocharous, G. 基于核函数的代表性状态强化学习。在第二十六届 *AAAI* 人工智能会议，2012年。

Lattimore, T. 和 Szepesv´ari, C. 赌博机算法。《剑桥大学出版社》，2020年。

Lee, L.、Eysenbach, B.、Parisotto, E.、Xing, E.、Levine, S. 和 Salakhutdinov, R. 通过状态边缘分布匹配实现高效探索。 *arXiv* 预印本*arXiv:1906.05274*，2019年。Liu, H. 和 Abbeel, P. 从虚无中获得行为：无监督主动预训练。《神经信息处理系统进展》，34:18459–18473，2021年。

Machado, M. C., Bellemare, M. G., Talvitie, E., Veness, J., Hausknecht, M. 和 Bowling, M. 重新审视街机学习环境：通用智能体的评估协议与开放问题。人工智能研究杂志，61：523–562，2018年。

McCallum, R. A. 用于隐藏状态强化学习的基于实例的效用区分。载于机器学习会议论文集 *1995*，第387–395页。爱思唯尔，1995年。

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. 等。通过深度强化学习实现人类水平控制。《自然》，2015年。

´Ormoneit, D. 和 Sen, S. 基于核函数的强化学习。机器学习，2002年。

Ostrovski, G., Bellemare, M. G., van den Oord, A. 和 Munos, R. 使用神经密度模型进行基于计数的探索。载于第*34*届国际机器学习会议论文集，第*70*卷，第2721–2730页。机器学习研究期刊.org，2017年。

Parzen, E. 关于概率密度函数及众数的估计。《数学统计年鉴》，33，1962年。ISSN 0003-4851。doi: 10.1214/aoms/1177704472。

Pathak, D., Agrawal, P., Efros, A. A. 和 Darrell, T. 通过自监督预测驱动的好奇心探索。载于*IEEE*计算机视觉与模式识别会议研讨会论文集，第16–17页，2017年。

Pazis, J. 和 Parr, R. 连续空间马尔可夫决策过程中的PAC最优探索。载于第二十七届*AAAI*人工智能会议，2013年。

Pong, V. H., Dalal, M., Lin, S., Nair, A., Bahl, S., and Levine, S. Skew-fit：覆盖状态的自监督强化学习。*arXiv* 预印本*arXiv:1903.03698*，2019年。

Pritzel, A., Uria, B., Srinivasan, S., Puigdom`enech, A., Vinyals, O., Hassabis, D., Wierstra, D., and Blundell, C. 神经情景控制。*ICML*，2017年。

Puterman, M. L. 马尔可夫决策过程。《运筹学与管理科学手册》，2:331–434，1990年。

Rosenblatt, M. 关于密度函数的一些非参数估计的评论。《数学统计年刊》，27，1956年。ISSN 0003-4851。doi: 10.1214/aoms/1177728190。Savinov, N., Raichuk, A., Vincent, D., Marinier, R., Pollefeys, M., Lillicrap, T. 和 Gelly, S. 通过可达性实现情节好奇心。载于国际学习表征会议，2018年。Schaul, T., Quan, J., Antonoglou, I. 和 Silver, D. 优先经验回放。*CoRR*，abs/1511.05952，2015年。Schmidhuber, J. 创造力、乐趣与内在动机的形式理论（1990–2010）。《*IEEE*自主心理发展汇刊》，2(3):230–247，2010年。Seo, Y., Chen, L., Shin, J., Lee, H., Abbeel, P. 和 Lee, K. 使用随机编码器最大化状态熵以实现高效探索。载于国际机器学习会议，第9443–9454页。PMLR，2021年。Strehl, A. L. 和 Littman, M. L. 基于模型的马尔可夫决策过程区间估计分析。《计算机与系统科学杂志》，74(8):1309–1331，2008年。Sutton, R. S. 通过时间差分方法学习预测。《机器学习》，3(1):9–44，1988年。Sutton, R. S. 和 Barto, A. G.《强化学习：导论》。麻省理工学院出版社，2018年。Tang, H., Houthooft, R., Foote, D., Stooke, A., Xi Chen, O., Duan, Y., Schulman, J., DeTurck, F. 和 Abbeel, P. #探索：深度强化学习中基于计数的探索研究。载于《神经信息处理系统进展》，2017年。URL https://proceedings.neurips.cc/paper/2017/file/3a20f62a0af1aa152670bab3c602feed-Paper.pdf。Tao, R. Y., Franc¸ois-Lavet, V. 和 Pineau, J. 面向样本高效探索的表征空间新颖性搜索。《神经信息处理系统进展》，33: 8114–8126，2020年。Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A. 等。使用PixelCNN解码器进行条件图像生成。《神经信息处理系统进展》，29，2016年。Van Hasselt, H., Guez, A. 和 Silver, D. 使用双重Q学习进行深度强化学习。载于*AAAI*人工智能会议论文集，第30卷，2016年。

A. 相关工作

本节将简要 且不求面面俱到地概述以往计算访问计数或估计密度的研究 s
在强化学习中。我们将其分 为参数化方法和非参数化方法。

参数化方法。Bellemare 等人（2016年）和 Ostrovski 等人（2017年）提出使用密度估计器，根据图像计算伪访问计数，例如使用上下文树切换（CTS，Bellemare 等人，2014年）或 PixelCNN（Van den Oord 等人，2016年）。另一方面，Tang 等人（2017年）使用局部敏感哈希将连续状态映射为离散嵌入，并据此计算显式访问计数。一些方法，例如 RND （Burda 等，2019），可以被解释为通过训练神经网络来预测一个随机初始化且未经训练的神经网络的输出，从而隐式估计观测的密度；后者以观测作为输入。Hazan 等人（2019年）、Pong 等人（2019年）、Lee 等人（2019年）和 Guo 等人（2021年）提出了搜索策略的算法，使其最大化所诱导的状态空间分布的熵。具体而言，Guo 等人（2021年）优化的损失既可以计算密度估计，又能最大化熵。最后，Domingues 等人（2021b）在学习到的表征之上进行密度估计，这些表征的设计受到无奖励有限马尔可夫决策过程中所使用奖励的启发。

非参数化方法。我们所采用的非参数密度估计可追溯至 Rosenblatt（1956年）和 Parzen（1962年）（Parzen–Rosenblatt 窗），由于对数据分布的假设非常宽松，因此广泛应用于机器学习。Kveton 与 Theocharous（2012年）以及 Barreto 等人（2016年）已将非参数化的基于核的方法用于强化学习，并证明其在规模较小的环境中具有良好的实证效果；Ormoneit 与 Sen（2002年）、Pazis 与 Parr（2013年）以及 Domingues 等人（2021a）则对其进行了理论分析。在 NGU （Badia 等人，2020b）和 MEME （Kapturowski 等人，2022年）中，Agent57（Badia 等人，2020a）采用非参数化方法，在情节层面计算短期奖励。Liu 与 Abbeel（2021年）提出了一种用于强化学习的无监督预训练方法，通过最大化在抽象表示空间中计算的非参数化熵来探索环境。作者表明，该方法提升了在 Atari 游戏和连续控制任务上的迁移性能。Seo 等人（2021年）使用随机嵌入和非参数化方法估计状态访问熵，但无法推广到并行学习的嵌入。Tao 等人（2020年）表明，基于 K-NN 的探索可以提升基于模型的强化学习中的探索效果和数据效率。非参数化方法虽适合对复杂数据建模，但其挑战在于需要存储整个数据集，并计算其密度。我们在正文第 3 节中通过提出一种方法来应对这一挑战：该方法基于长期状态历史估计访问计数，使我们的方法能够扩展到比以往工作所研究的问题规模大得多的任务；同时，它不对表征作任何假设，表征可以与探索过程并行训练，无需预先固定。

B. 一般符号

我们考虑通常的强化学习设定，其中智能体与环境交互，以最大化折扣奖励之和，折扣因子为 $\gamma \in [0,1)$，如 Sutton（1988年）所述。具体而言，环境可描述为一个部分可观测马尔可夫决策过程（POMDP）（Kaelbling 等人，1998年）。首先，我们通过一个元组 $(\mathcal{S}, \mathcal{A}, T, R)$ 定义马尔可夫决策过程（MDP），其中 $\mathcal{S}$ 是状态集合，$\mathcal{A}$ 是可能动作的集合，T 是转移函数，它将状态-动作映射为下一状态上的概率分布，而 $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 是奖励函数。具体而言，马尔可夫决策过程是智能体与其环境之间的离散时间交互过程（McCallum，1995年；Hutter，2004年；Hutter 等人，2009年；Daswani 等人，2013年）。在部分可观测 MDP 中，智能体接收的不是来自 $\mathcal{S}$ 的状态，而是观测 $o \in \Omega$，其中 $\mathcal{O}$ 是将未观测状态映射为观测上的概率分布的函数。观测 o 只能包含关于底层状态 $s \in \mathcal{S}$ 的部分信息。此函数可以与内在奖励函数 r_i 结合，以促进行为探索。环境通过执行状态转移来响应智能体的动作$a \in \mathcal{A}$，转移到状态 $s' \sim T(\cdot|s,a)$；智能体随后收到新的观测 $o' \sim \Omega(\cdot|s')$和奖励$r \sim R(s,a)$。在步骤 t，我们用 $h_t = \{o_0, a_1, o_1, \ldots, a_t, o_t\} \in \mathcal{H}_t$ 表示过去观测-动作的历史，其中 $\mathcal{H}_t = \mathcal{H}_{t-1} \times \mathcal{A} \times \mathcal{O}, \mathcal{H}_0 = \mathcal{O}$，整体历史空间为 $\mathcal{H} = \bigcup_{t\in\mathbb{N}} \mathcal{H}_t$。我们考虑策略 $\pi\colon \mathcal{H} \to \Delta_{\mathcal{A}}$，它将过去观测-动作的历史映射为动作上的概率分布。

C. RECODE 从聚类的角度来看

更新规则 RECODE在算法1 中的记忆结构，可以解释为潜在概率聚类模型中的一种近似推断方案。我们在此探讨这种联系，以便更好地理解并论证

算法 2一种流式聚类算法

. 1：参数：簇数量 $|M|$**最近簇中心的数量** k**每一步的计数折扣** γ**提议创建新簇的距离阈值** κ**接受创建新簇的概率** η**2：状态：创建新簇的阈值（即平均簇间距离）** $d = 0$**簇中心** $m_l = 0 \quad \forall l \in 1 \dots |M|$**簇计数** $c_l = 0 \quad \forall l \in 1 \dots |M|$**点的** k**个最近邻的索引** e**：邻域**$_k$**(**e**) 3：实现：4：对于所有接收到的嵌入** $e \in \{e_0, e_1, e_2, \dots\}$**，执行 5：更新平均簇间距离** $d \leftarrow$ **(**$1-\tau$**)**$d + \frac{\tau}{k} \sum_{l \in}$**邻域**$_k$**(**$e$**)** $\|m_l - e\|_2^2$**6：折扣所有簇中心的计数** $c_l \leftarrow \gamma\, c_l \quad \forall l \in 1$**，.....，** $|M|$**7：查找最近簇中心的索引** $m_\star =$ **arg min**$_{m \in M} \|m_l - e\|_2$**8：如果** $\|m_i - e\|_2^2 > \kappa\, d$ **且以概率** η **执行，则 9：按概率** P**(**j**) 抽取要移除的簇中心索引** $j \propto 1/c_j^2$**10：查找距离** m_j**最近的簇中心索引：** $m_\dagger =$ **argmin**$_{m \in M, l \neq j} \|m_l - m_j\|_2$**11：重新分配被移除簇中心的计数：** $c_\dagger \leftarrow c_j + c_\dagger$**12：将簇** j**替换为位于** e**的新簇：** $m_j \leftarrow e$**,** $c_j \leftarrow 1$**13：否则 14：更新最近簇中心** $m_\star \leftarrow \frac{c_\star}{c_\star + 1} \mu_i + \frac{1}{c_\star + 1} e$**15：更新最近簇中心的计数** $c_\star \leftarrow c_\star + 1$**16：结束如果 17：结束循环**

将所提出的算法视为密度估计器。该规则与（Kulis 和 Jordan，2011年）的 DP-means 算法密切相关，但有两个关键区别：

- 每一步都会对簇中心的计数进行折扣，使我们的方法能够应对策略和嵌入函数变化所导致的数据非平稳性，从而有效降低记忆中陈旧簇中心的权重，
- 创建新簇中心时，我们会移除一个样本数不足的簇中心，以保持记忆的大小不变。

这些调整对于适应我们设定中的额外复杂性是必要的：该设定遵循流式协议（即数据到达时必须明确地使用或存储，未存储的数据之后将无法再次访问），且具有非平稳性（即不假设数据随时间推移始终服从相同的分布）。经过这些调整后得到的聚类算法如算法 2 所示。 RECODE 实现了这样一种算法来更新记忆，同时还会为观测到的嵌入计算内在奖励 e，详见第 3 节。

C.1. 应对有限内存限制。

我们首先介绍为应对流式场景的记忆限制而引入的修改：1）每个数据点（在我们的记法中为嵌入 e_t）仅被纳入簇分布近似一次，随后即被丢弃；2）簇的总数会随机投影至簇数量上限（否则簇的数量将无限增长——尽管增长速度会逐渐放缓）。这两项修改使我们的方法能够保持常数空间

复杂度，以应对无限数据流。

该逐步算法 2 的逐步论证相对直观。在步骤 t中，对于嵌入 e_t，我们证明以下目标函数达到最小值：

$$\min_{l\in 1,\dots,|M|} \|m_l - e_t\|_2^2 \quad \text{s.t.} \quad \|m_l - e_t\|_2^2 \le \kappa d \tag{6}$$

反向推导：更新簇中心会直接降低目标函数，且不会违反约束（除非它原本就违反了约束；这一情况已在该分支的前置条件中排除）。这对应“else”分支。“if”分支恰好在 e_t处引入一个新簇中心，因此，公式6 完全达到最小值：此分支下其值为零。最后，选择最近簇中心的索引，可直接最小化公式6中相应分支的取值，而不考虑约束（约束由“if/else”确保满足）。请注意，公式6 中的硬约束取代了 DP-means 的软簇惩罚项（Kulis 和 Jordan，2011年）。

与 k 均值和 DP-means 不同，簇中心的更新是对嵌入进行指数加权移动平均，而不是利用所有数据执行全局优化步骤。因此，尤其重要的是，对6 进行评估时 e_s，其中 $s \neq t$，这一点具有重要意义，因为 k 均值和 DP-means 的目标函数考虑的是所有数据，而不是单个数据点。

我们在图 8 中测试了不同移除策略的定性表现。这项研究表明，以概率 $\propto c^{-2}$ 随机移除一个簇更加稳定，也能更好地追踪非平稳分布。从这些简单示例中得到的直觉，也得到了在 Atari 环境中开展的消融实验的证实，如图 9 所示，其中我们比较了采用不同移除规则的 `RECODE` 次运行。

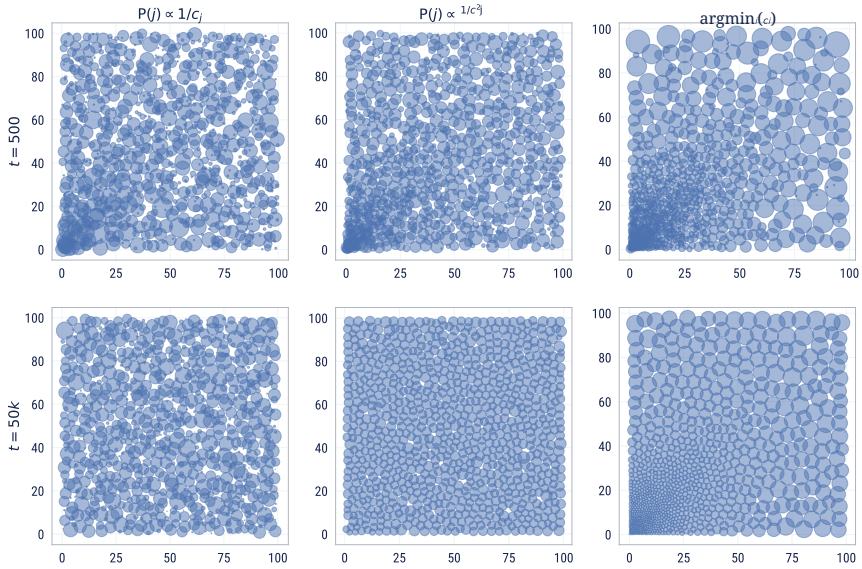

图 8：移除策略对簇中心和计数演变的影响（计数对应于标记的大小）。在每个时间步 t，我们从边长为 $\min(100, t)$ 的正方形中采样一批包含 64 个二维嵌入的数据。$t = 100$ 之后，分布保持稳定，我们希望经过足够长的时间后，簇中心和计数的分布都变得近似均匀。对于一种确定性移除策略，即移除计数最低的簇，在分布停止变化很久之后，簇中心仍可能呈现偏斜。对于两种概率移除策略，簇中心都会变得近似均匀，但只有使用 $1/c_j^2$ 移除策略时，我们观察到簇中心和计数都会变得均匀。请注意，我们使用的折扣因子为 $\gamma = 0.9999$。

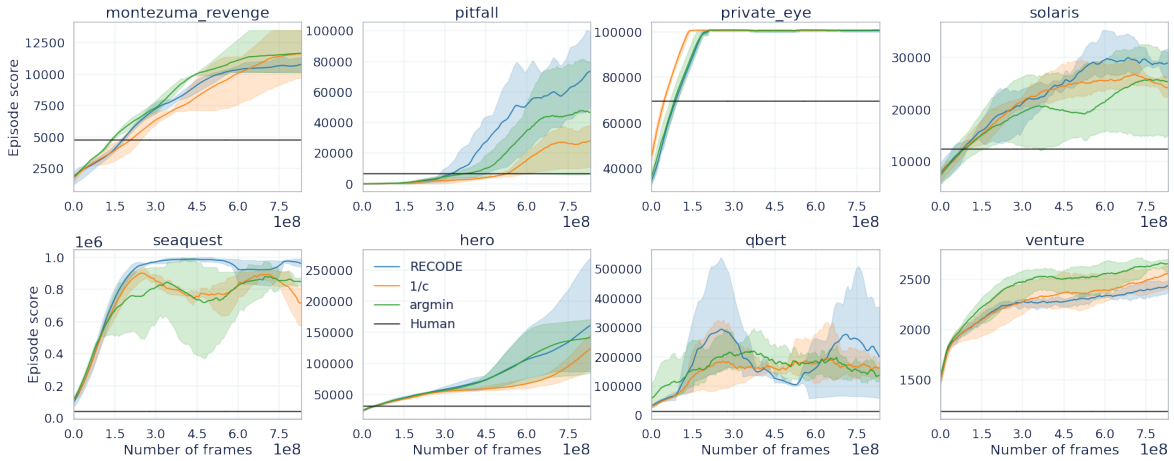

图9：移除策略对性能的影响。考虑的所有移除策略都能构成可行的算法，但在某些环境中（最明显的是*Pitfall!*），选择 $1/c^2$ 的策略似乎更稳健。

C.2. 处理非平稳嵌入分布。

现在我们来讨论如何应对嵌入分布的非平稳性。为处理数据流的非平稳性，我们进行了以下修改：

1. 簇数量会衰减；2. 可以合并两个簇以容纳一个新簇；3. 使用指数加权移动平均来更新簇中心。

在 k 均值中，所有数据都会被保留。这使得 k 均值的成本较高：每一步拟合时，都会检查整个数据集，以更新簇分配和簇均值。相反，我们对与每个簇相关的数据采用分布近似，并在根据6重新调整簇分配时，基于这一分布近似进行处理。

具体而言，每个簇都近似为一个精度为 1 的高斯分布，其均值未知，但先验均值为零、精度为 1。具体如下：

$$\mu_l \sim \mathcal{N}(0, c_0)$$
$$e_i|\mu_l \sim \mathcal{N}(\mu_l, 1)$$

其中，$\mathcal{N}(\mu, \tau)$ 表示均值为 μ、精度为 τ 的高斯（或正态）分布（精度是方差的倒数）。由于 μ_l 的先验与 e_i 的似然共轭，我们知道 m_l 的后验将具有 $\mathcal{N}(\mu_l, c_l)$ 的形式。使用单个嵌入 e_i 更新此后验的形式如下：

$$m \leftarrow \frac{c_l}{c_l + 1} m + \frac{1}{c_l + 1} e_i$$
$$c_l \leftarrow c_l + 1$$

这正是算法2中的更新。

请注意，在该模型中，计数 c_l 也是分布的精度参数，表示每个簇的扩散程度的倒数（或集中度）。在算法2的每一步中，这些计数都会衰减。实际上，这会使表示每个簇的分布的方差增大：因此在每个时间步，每个簇的集中度都会降低，对哪些数据点属于该簇也变得更加不确定。超参数 γ 刻画了所有簇以这种方式扩散的速率。每一步都会增加这种不确定性，从而形成一种“遗忘”机制，帮助算法应对不断变化的数据分布。

如上文根据公式 e_t 所述，簇重采样可确保簇的数量不超过 $|M|$。有两个细节需要考察：合并什么，以及如何合并。如 $c_j \mapsto 0$所示，高斯似然分配给簇 j 的新数据点的概率也趋近于零，因此计数最少的簇可能对未来的密度估计影响最小（因为它的扩散程度最大）。然而，当 $c_j \gg 0$时，应该移除哪个簇就不那么明确了。因此，我们以与计数平方成反比的概率随机选择要移除的簇（使用计数平方能比$1/c_j$更突出计数之间的微小差异）。该簇可以被彻底移除，但我们选择将其计数重新分配给最近的簇，因为实验发现，这种策略对超参数选择不那么敏感。

为了帮助建立对折扣因子影响的直观认识，我们在图10中通过一个具有非平稳嵌入分布的玩具示例来展示其影响。我们发现，调整折扣因子 γ 可以在短期记忆和长期记忆之间平滑过渡。



图 10：使用 RECODE 在一个玩具示例中进行非平稳密度估计。对于步骤 $t = 0, \ldots, 100$，我们从边长为 $1 + \sqrt{t}$ 的正方形中均匀采样一批 64 个二维嵌入。因此，嵌入分布的支持集会随时间扩展，以模拟一种非平稳分布，类似于强化学习智能体在探索过程中访问的状态分布。我们绘出了 RECODE 学习到的原子，其大小与计数成正比。我们发现，当折扣因子足够小时， RECODE 表现出短期记忆，能够准确近似最终分布。随着折扣因子增大， RECODE 表现出更长期的记忆，近似状态的历史密度；从左下角概率质量的集中情况可以看出这一点。

为确认簇折扣在实践中的必要性，我们在 Atari 环境中进行了额外的消融实验。如果我们不在探索过程中训练表征，而是从预训练的表征开始，可以看到 RECODE 即使不使用折扣因子，也能取得尚可的表现（见图12）。然而，如图11所示，只要我们同时训练表征，遗忘旧观测的能力就能弥补分布偏移，并显著提升性能。

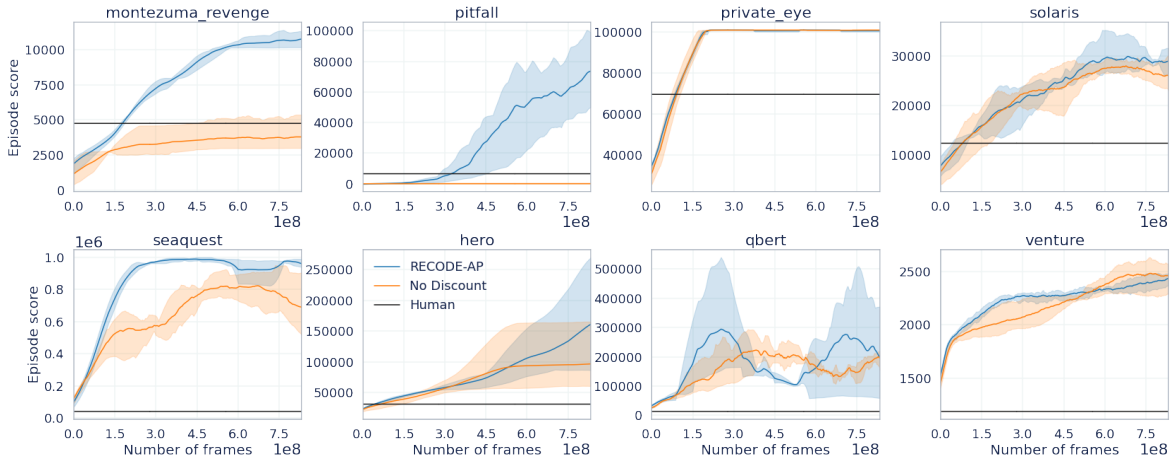

图 11：折扣因子对性能的影响。随着嵌入在整个训练过程中不断演变，旧簇可能会在当前表征下失去意义。停用折扣因子（即 $\gamma = 1$）会导致性能显著下降，尤其是在蒙特祖玛的复仇和 Pitfall! 等困难探索场景中。

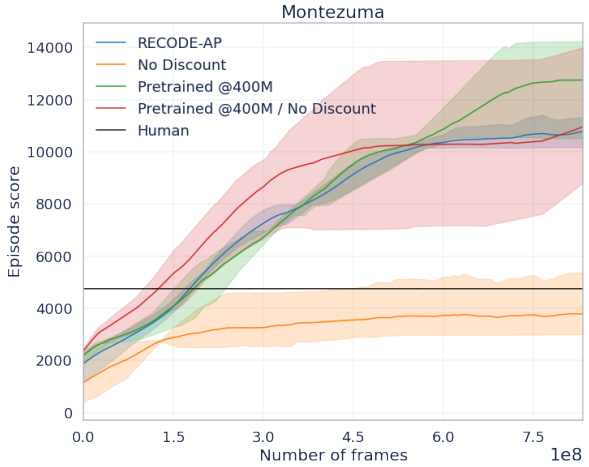

图 12：预训练嵌入与同步训练的嵌入，以及对折扣因子的敏感性。我们在 4 亿帧时对嵌入进行快照，然后使用这些冻结嵌入，让智能体从头开始重新训练。我们发现，即使使用 $\gamma = 1$（即不进行折扣），也能取得相近的性能。有趣的是，我们还观察到，智能体取得的分数远高于原智能体在拍摄快照时取得的分数。

D. 使用 RECODE分析探索

本节通过一个简单示例，展示RECODE的探索过程如何展开。我们使用*Random Disco Maze*的一种变体：这是 Badia 等人在2020b 提出的一种网格世界环境，用于说明使用可控状态表征估计探索奖励的重要性，如图13（左）所示。智能体每回合都从一个完全可观测、大小为 21×21 的迷宫中的固定位置出发。智能体可以执行四种动作 {左、右、上、下}。如果智能体撞上墙壁、到达目标状态，或达到 500 步的上限，回合就会结束。

至关重要的是，环境会在每个时间步对每个墙壁碎片的颜色进行随机变化。具体而言，每个墙壁碎片的颜色都会从五种可能的颜色中随机且独立地选取。这会给系统引入大量无关变化，从而对基于新颖性的探索奖励方法构成严峻挑战。原因在于，智能体永远不会两次看到完全相同的状态，因为墙壁碎片的颜色在每个时间步都会不同。

我们运行了 RECODE 一亿个时间步。智能体在收集约五千万至六千万个时间步后能够找到目标。在图13 （右）中展示了智能体探索该环境时簇分布的变化情况。由于智能体总是从同一位置（迷宫左下角）开始，因此分布严重偏向于过度代表这一点。我们可以看到，随着时间推移，簇中心均匀地覆盖整个迷宫。



图 13: （左）Random Disco Maze（右） RECODE 所学习的簇的分布随时间的演变，详情请参阅正文。

我们研究了 RECODE 在 蒙特祖玛的复仇中的记忆可以追溯到多远。结果如图 14所示。我们发现， RECODE 所学习簇的年龄分布（即每个原子在多少步之前被插入记忆中）在约 $2 \cdot 10^6$ 个智能体步数附近呈现众数，这相当于数百个回合；此外，还有相当数量的簇的年龄是其十倍。需要提醒的是， NGU的短期非参数新颖性最多只估计一个回合（图中的红线）。

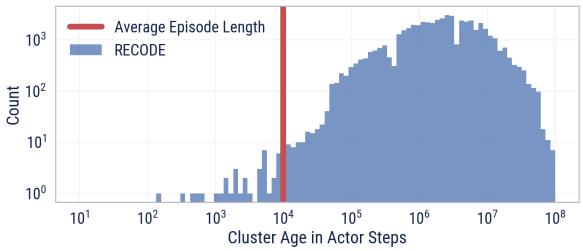

图 14: 在 蒙特祖玛的复仇中所学习簇的年龄分布。 RECODE的记忆时域远超单个回合。我们将 $\gamma = 0.999$ 设置为与图4中的实验相同。我们用红色标出了回合的平均长度，表明在这一设置下， RECODE的记忆可以追溯数千个回合。

E. 超参数

此处展示了我们实验中使用的精确超参数值，表格1 适用于 Atari，表格 2 适用于 DM-HARD-8 我们省略了与基础智能体 MEME（Kapturowski 等人，2022年）相同的超参数。

表 1：Atari 超参数。

参数	值
RECODE 内存大小	5×10^4
RECODE 折扣因子 γ	0.999
RECODE 插入概率 η	0.05
RECODE 相对容差 κ	0.2
RECODE 奖励常数 c	0.01
RECODE 衰减率 τ	0.9999
RECODE 邻居 k	20
IM奖励尺度 β_{IM}	1.0
最大折扣	0.9997
最小折扣	0.97
回放周期	80
轨迹长度	160
回放比率	6.0
回放容量	2×10^5 轨迹
批次大小	64
强化学习 Adam 学习率	3×10^{-4}
嵌入 Adam 学习率	6×10^{-4}
强化学习权重衰减	0.05
嵌入权重衰减	0.05
强化学习主体初始步幅	4
强化学习主体块数	(2，3，4，4)
强化学习主体通道数	(64，128，128，64)
强化学习主体步幅	(1，2，2，2)

表 2： DM-HARD-8 超参数。

参数	值
RECODE内存大小	2×10^5
RECODE折扣因子 γ	0.997
RECODE插入概率 η	0.2
RECODE相对容差 κ	0.2
RECODE奖励常数 c	0.01
RECODE衰减率 τ	0.9999
RECODE邻居 k	20
IM奖励尺度 β_{IM}	0.1
最大折扣	0.997
最小折扣	0.97
回放周期	40
轨迹长度	80
回放比率	2.0
回放容量	5000 条轨迹
批次大小	128
强化学习 Adam 学习率	1×10^{-4}
嵌入 Adam 学习率	3×10^{-4}
强化学习权重衰减	0.1
嵌入权重衰减	0.1
强化学习主体初始步幅	2
强化学习主体块数	(2，4，12，6)
RL躯干通道数	(64，128，128，64)
RL躯干步幅	（1，2，2，2）

表 3： CASM超参数。

参数	值
Transformer 类型	门控 TransformerXL
状态掩码率	0.8
每条轨迹的掩码数	4
动作嵌入维度	32
层数	2
注意力维度	128
注意力头数	4
多层感知机隐藏层维度	(512,)
预测器隐藏层维度	(128,)

F. 使用 RECODE的分布式智能体架构

现在我们详细说明如何将 RECODE高效集成到典型的分布式强化学习智能体（Espeholt 等人，2018年；Kapturowski 等人，2018年）中。这类智能体由多个并行运行并相互交互的进程组成，从而能够开展大规模实验。传统上，学习器会执行梯度更新步，以训练策略 π_θ 和嵌入（表征）函数f_θ，并将参数 θ 转发给推理工作器。多个相互独立的执行器会向推理工作器请求动作，在环境中执行这些动作，并将产生的状态转移发送给学习器（可选择通过一个

（优先级）回放（Mnih 等人，2015年；Schaulet 等人，2015年）。使用 RECODE时，执行器还会与一个实现算法 1的共享记忆通信：在每个步骤 t中，它们会向推理服务器查询其历史的嵌入 $f_\theta(h_t)$，并将其发送到共享记忆。共享记忆会返回一个内在奖励 r_t，该奖励随后会加到外在奖励上，用于训练学习器进程中的策略。图 15 展示了使用 RECODE 的典型分布式智能体架构概览。15。

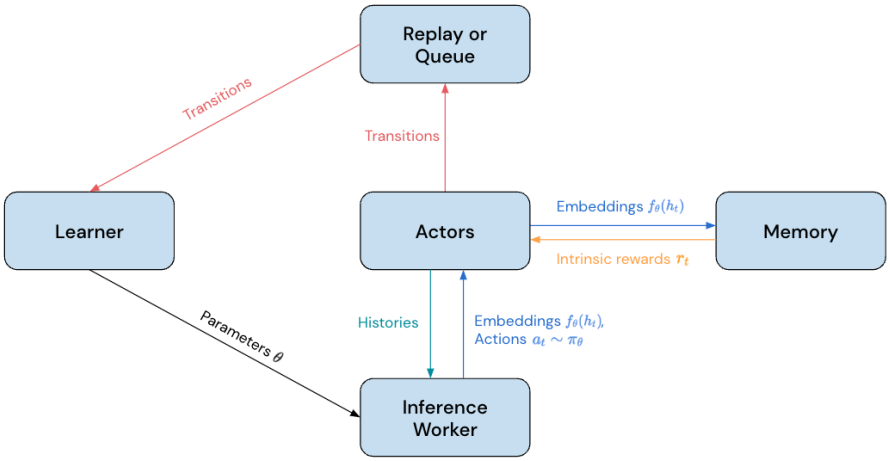

图15：使用 RECODE的分布式智能体架构概览。

G. 智能体分类

主论文中的实验和附录K 中的扩展实验所评估的所有方法均由三个主要组件构成。

- 一个负责整体强化学习过程的基础智能体（例如，执行动作、收集观测、计算调整后的回报、更新策略等）。我们重点关注 MEME (Kapturowski 等人，2022年)，这是对 Agent57（Badia 等人，2020a）的近期改进，具有高得多的样本效率，目前在 Atari 上达到最先进水平；以及一种基于 VMPO 的智能体（Guo 等人，2022年），它是目前在 DM-HARD-8上的最先进方法。

- 一种表征学习机制，用于生成供内在奖励生成器使用的观测嵌入。我们同时考虑动作预测（AP）和 CASM 嵌入。请注意，有些内在奖励模块无法有效利用表征学习模块（例如， RND），而另一些则将第二和第三个模块合并为一种方法（例如， BYOL-Explore）

- 一种用于生成内在奖励的算法。除此之外， RECODE我们还考虑了近期的 BYOL-Explore (Guo 等人，2022年)、 NGU (Badia 等人，2020b) 以及 NGU的两个构建模块： RND (Burda 等人，2019年) 和情节记忆 (EMM)(Pritzel 等人，2017年)。

例如，在我们更详细的分类体系中，（ MEME Kapturowski 等人，2022年）中描述的原始智能体被标记为 MEME-NGU-AP 基线，并与我们的新型 MEME-RECODE-AP 智能体进行比较，唯一的修改是探索奖励发生了变化。表K报告了我们实验中所有可用组合的更多细节。

表 4：实验中使用的智能体分类体系。

智能体名称		基础智能体	内在奖励	表征学习
MEME-NGU-AP	（Kapturowski 等人，2022年）	MEME	NGU	AP
MEME-RND	（消融）	MEME	RND	不适用[a]
MEME-EMM-AP	(消融)	MEME	EMM	AP
MEME-RNDonAP	(消融)	MEME	RND	AP [b]
MEME-RECODE-AP	（本文）	MEME	RECODE	AP
MEME-RECODE-CASM	（本文）	MEME	RECODE	CASM
MEME-NGU-CASM	（消融）	MEME	NGU	CASM
VMPO-BYOL-Explore	(Guo 等人，2022年)	VMPO	BYOL-Explore	BYOL-Explore [c]

(a) 与原论文一样， RND 以原始观测作为输入。(b) 为了测试 RND应对非平稳表征的能力，我们与策略同步训练一个 AP 编码器，并用它创建输入 RND 的观测嵌入（即，在 AP之上运行 RND ）。(c) BYOL-Explore 机制会在内部训练一个神经网络，用于预测观测的动态演变。这样一来，智能体便获得了奖励/新颖性信号（预测误差），以及观测的嵌入表征（可从网络最后几层提取）。

H. 在 DM-HARD-8 环境中的探索

三维任务 DM-HARD-8 的待探索状态空间可能极其庞大。例如，Baseball task，如图 16所示：智能体需要观察场景、找到球棒、将其捡起、把球扔下、捡起球，并成功拿到苹果。

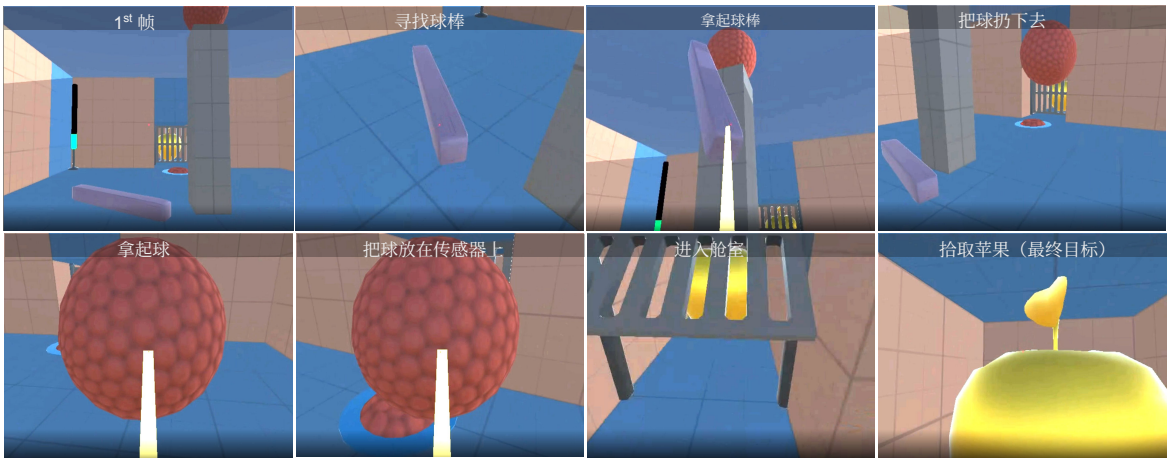

图 16：智能体完成 DM-HARD-8 Baseball 任务时的第一人称视角快照。图像按照时间顺序从左到右、从上到下排列。每张图像都展示了任务的一个特定阶段。智能体必须与环境中的特定物体互动，才能完成任务。

I. 汇总结果

在本节中，我们展示了跨不同环境的总体汇总结果，既包括 Atari 套件，也包括 DM-HARD-8。为确保不会因奖励尺度较大而使某个环境占据主导地位，我们在每个环境中使用人类归一化分数（Mnih 等人，2015年），然后在取平均值之前，将超过 100% 的分数封顶。

如图 17 所示，（左），未封顶分数可能随时间大幅波动，在此例中，这仅仅是 *Q*bert*中高方差所造成的假象。这种方差源于 *Q*bert*中的一个漏洞，利用该漏洞便可获得高得多的分数。

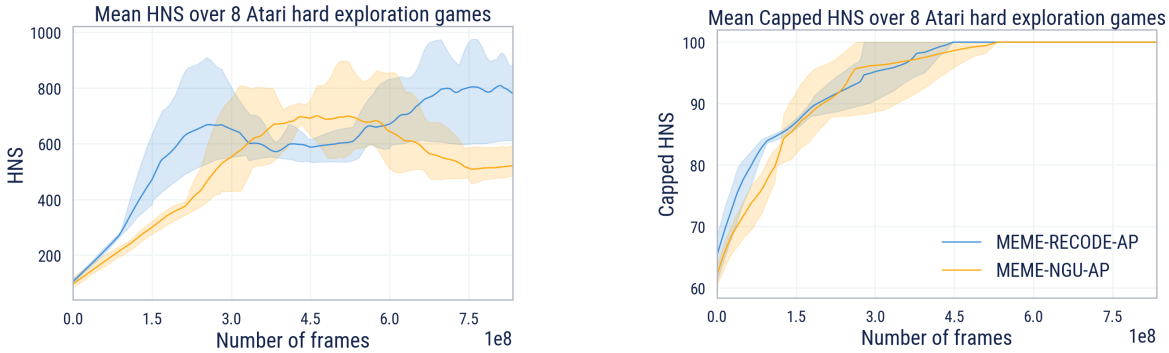

图 17：Atari 上的汇总结果。（左）： MEME-RECODE-AP 与MEME-NGU-AP在 Atari 游戏上的平均人类归一化分数。（右）：封顶的人类归一化分数。

在图 18中，我们展示了按每个任务套件中的所有环境汇总的主要结果。 DM-HARD-8 表 5总结了结果，并比较了 RECODE 新颖性奖励机制与 NGU的性能。我们着重强调了与 BYOL-Explore 等替代方案相比，我们的方法应用于 DM-HARD-8等三维环境时的有效性。

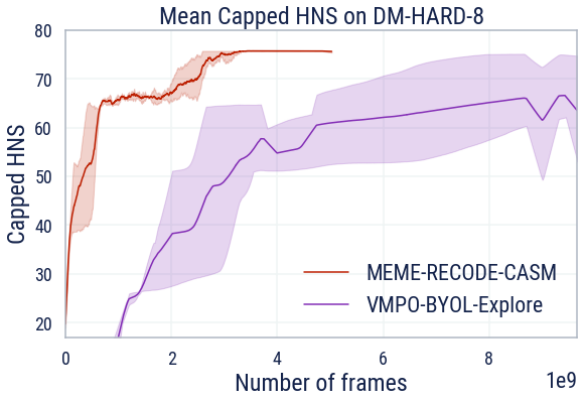

图 18： DM-HARD-8上的汇总结果。比较 MEME-RECODE-CASM 与 VMPO-BYOL-Explore的平均封顶人类归一化分数。

表 5：在 DM-HARD-8 个环境任务上的汇总结果。

Game	人类	MEME-NGU-AP	MEME-NGU-CASM	MEME-RECODE-AP	MEME-RECODE-CASM
棒球	7.90	0.00 ± 0.00	0.00 ± 0.00	**10.00 ± 0.00**	**10.00 ± 0.00**
吊桥	12.30	0.00 ± 0.00	4.15 ± 6.62	12.41 ± 0.07	**12.86 ± 0.13**
导航立方体	7.80	3.33 ± 5.33	9.76 ± 0.25	**10.00 ± 0.00**	**10.00 ± 0.00**
推方块	**8.40**	0.00 ± 0.00	0.24 ± 0.24	1.07 ± 0.62	4.06 ± 2.14
记忆传感器	**7.60**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
跨越投掷	5.70	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**9.72 ± 0.31**
墙壁传感器	9.10	0.00 ± 0.00	0.00 ± 0.00	**10.00 ± 0.00**	**10.00 ± 0.00**
墙壁传感器堆栈	**8.60**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

J. 多任务实验

我们还在 RECODE 一个基于 VMPO、与 BYOL-Explore 所使用的智能体类似的智能体中实现了该方法（Guo 等人，2022年），并在多任务设置下将我们的性能与 BYOL-Explore 进行了比较。该实验有两个不同的目的。首先，这证明了我们的探索奖励具有普适性：事实表明，它适用于迥异的强化学习智能体，无论这些智能体是基于价值还是基于策略的。其次，我们可以在多任务设置下，直接将其与最先进水平的 BYOL-Explore 智能体进行比较。不过，我们注意到，本实验所使用的表征学习技术——一步动作预测——基于一种会丢弃过去历史的前馈嵌入，因此可能并不适合在部分可观测马尔可夫决策过程（POMDP）中进行探索。尽管如此，图19 显示， RECODE的性能与 BYOL-Explore 相当，仅缺少一个等级即可达到其性能。使用更适合的表征（例如 CASM）来提升这一性能，则留待未来工作。

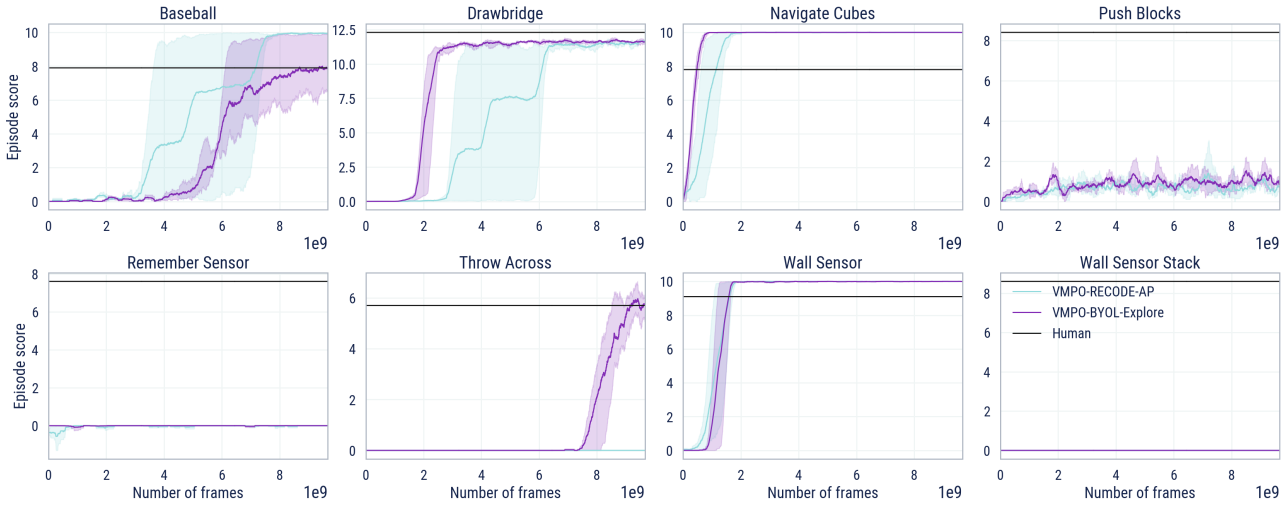

图 19： RECODE 与 BYOL-Explore 在多任务版本的 DM-HARD-8 上的性能比较。本实验中的 RECODE实现基于 VMPO，使用连续动作集。

K. 附加消融研究

K.1. CASM 掩蔽

这里我们分析在 CASM 中移除掩码策略后（见正文）我们的方法的性能。对于序列中的每个元素，我们不再只提供嵌入后的观测或动作，而是始终同时提供两者，这使得上游分类器任务过于简单。掩码策略可以提供额外的上下文（相对于动作预测而言），同时确保预测任务仍然足够困难，从而要求表征对高层特征进行编码。

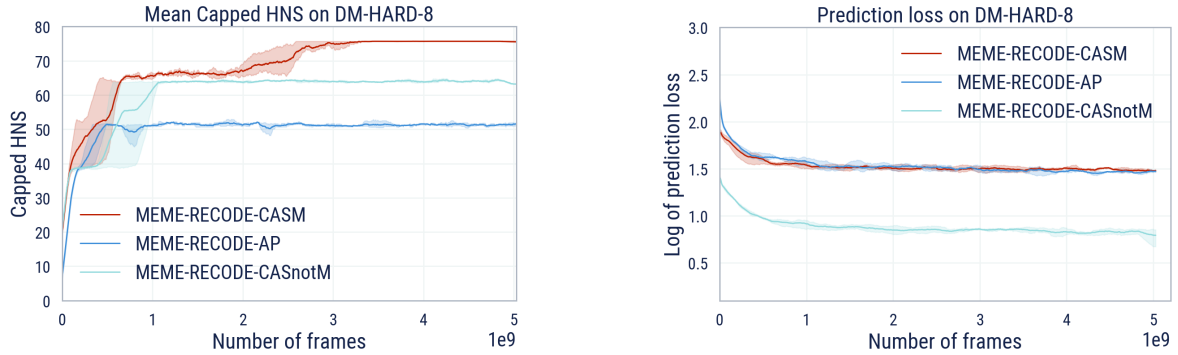

图20： CASM 掩码消融实验。（左）： CASnotM 是在未对轨迹应用任何掩码的情况下进行的。即使单独使用，额外的上下文也比 AP具有明显优势，但掩码策略对于解决 Push Block 至关重要。（右）： CASnotM 的损失远低于 AP 和 CASM，这表明在没有任何掩码的情况下增加上下文会使预测任务更容易（但会导致表征的稳健性降低）。

K.2. RND 基于动作预测嵌入 ngs

我们调整了 RND，以利用经过训练的动作预测嵌入，并将其称为 RNDonAP。为此，我们使用一个随机初始化的多层感知机（MLP）对嵌入进行随机投影，再使用另一个经过训练的 MLP 重建该随机投影。重建误差为探索提供内在奖励；我们按照（Burda 等人，2019年）的方法，使用其标准差的运行估计值对奖励进行归一化。我们发现，由此得到的智能体无法攻克一些最难的探索游戏，例如 Montezuma's Revenge 或 Pitfall。这项消融实验的结果见图 21。使用预训练嵌入进行的实验似乎表明，在这种设置下， RNDonAP 能够取得更好的性能，但由于无法同时训练嵌入，该方法的普适性受到很大限制。

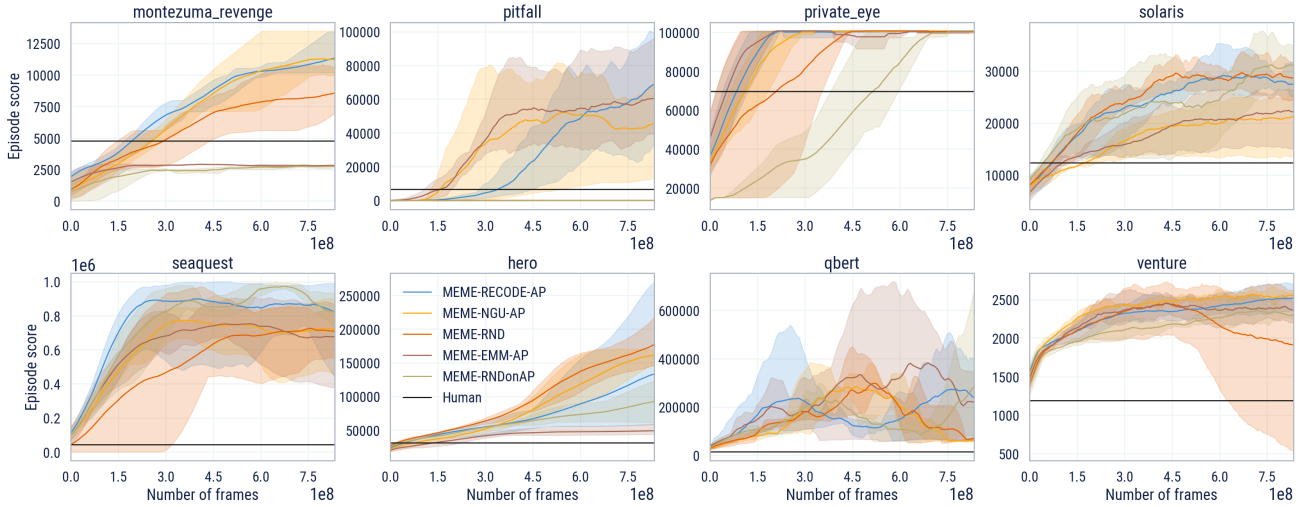

图21： RECODE 与 MEME 及其消融版本在 8 款高难度 Atari 探索游戏中的性能对比。我们发现，这种方法无法解决一些最难的游戏，例如蒙特祖玛的复仇 或 *Pitfall!*

这一失败的一种可能解释是，较大的 RND 误差可能是由观测到新状态，或已观测状态的表征发生漂移所造成的。RND 未能区分这两种影响，导致探索效果不佳。

K.3. RECODE 基于 BYOL 嵌入

为进一步展示 RECODE ro　对表征变化的稳健性，在图22中，我们比较了以下方法的性能： RECO
DE基于动作　预测表征与 BYOL 表征的比较（如Guo 等人（2022年）　).

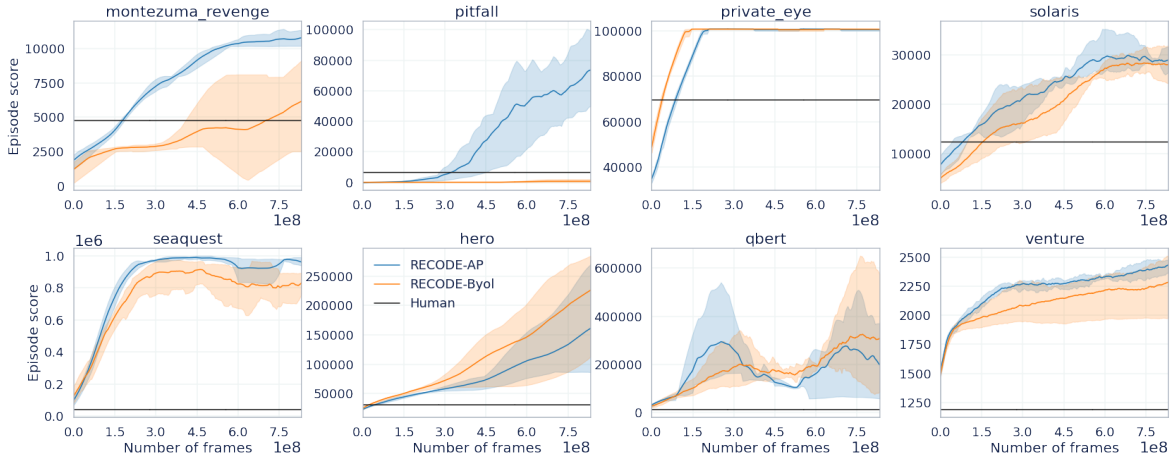

图22：比较 RECODE 使用 AP 嵌入与 BYOL 嵌入的情况。我们观察到， RECODE 能够利用 BYOL 嵌入，在蒙特祖玛的复仇中达到超人类性能，并在*Pitfall!*中取得正分数，但与 AP 嵌入相比，性能明显较差。对于 BYOL，我们对嵌入维度 $\{32, 128, 512\}$进行了扫描，并报告表现最佳的维度，即 32。