基于选项框架、采用多模式探索的自主非单体式智能体

JaeYoon Kim*[1]、Junyu Xuan†[1]、Christy Liang†[2] 和 Farookh Hussain†[1]*电子邮件：JaeYoon.Kim@student.uts.edu.au
†电子邮件：{Junyu.Xuan,Jie.Liang, Farookh.Hussain}@uts.edu.au[1]澳大利亚人工智能研究院（AAII），澳大利亚悉尼科技大学[2]悉尼科技大学可视化研究院，澳大利亚

arXiv:2305.01322v3 [cs.AI] 4 May 2024

摘要——**大多数强化学习（RL）探索研究关注“探索方式”，即“如何探索”。另一类探索研究关注“何时探索”，但这并非强化学习探索研究的主要焦点。在通常的强化学习探索行为中，单一探索的“何时”问题将探索动作与智能体的利用动作绑定在一起。近年来，非单体式探索研究开始考察人类和动物的模式切换探索行为。我们的研究最终旨在使智能体能够自主决定何时探索或利用。本文介绍了在选项框架下，对非单体式行为的自主多模式探索开展的初步研究。通过对比实验结果，我们证明了该方法优于现有的非单体式探索方法。**

索引词—**非单体式探索、自主多模式探索、选项框架**

I. 引言

探索是强化学习（RL）算法的关键部分，因为它使智能体能够选择发现未知状态。已有许多从不同视角研究强化学习探索的工作，例如内在奖励 [1], [2], [3], [4],技能发现 [5], [6], [7], 基于记忆 [8], [9], [10], [11],和基于Q值 [12]。尽管探索研究不断发展，但其关注重点一直是“如何探索”，也就是智能体如何选择探索动作。然而，对于“何时探索”的研究尚未得到充分开展。

关于“何时进行探索”有两种方法，分别是单一探索和非单体式探索。基于噪声的单一探索是一种典型的单一探索：在将最终动作输入环境之前，将从随机分布中采样的噪声添加到行为策略的原始动作中。策略的原始动作和所添加的噪声分别发挥利用和探索的作用。因此，采用单一探索的行为策略属于时间齐次行为策略。

本研究得到澳大利亚研究委员会通过探索早期职业研究员奖DE200100245和联结项目LP210301046的支持。

策略。然而，在非单体式探索中，行为策略的原始动作不会添加噪声。它们会在单独的步骤中为各自的目的而行动。因此，采用非单体式探索的行为策略属于异质模式切换策略（图1）。

我们对非单体式探索的初步研究 [13] 进行了探讨。作为一项初步工作，该研究仍存在若干局限。首先，只有一种探索策略（我们称之为单模式探索）。智能体可能需要更多探索模式选择，而更高的探索模式熵意味着探索模式数量多于一种。其次，可控的探索时长不应固定，而应具有可变性。第三，该研究采用了一种简单的阈值超参数函数，称为“稳态”，用于调整探索与利用之间切换触发信号的尺度。然而，应通过策略本身实现自然的切换机制。研究还提出了其他有信息量的触发器，包括基于动作不匹配的触发器和基于方差的触发器。

本文提出一种基于选项框架的自主非单体式智能体，采用多模式探索来解决上述问题。具体而言，我们采用分层强化学习（HRL）作为选项框架，将一系列探索模式与利用串联起来，以实现幕内时间尺度上的切换行为。因此，我们可以实现具有不同熵的多模式探索。为实现探索策略与利用策略之间的自主切换，并使切换基于内在信号，我们通过修改各切换模式的奖励来调整引导式探索。此外，我们还研究了鲁棒最优策略，以保持潜在性能。

与此同时，本研究应回答以下 5 个问题。如何采用选项框架，以利用分层强化学习的上下文进行探索模式选择和利用？智能体如何灵活调整探索周期？智能体如何获得更多熵，以便选择探索模式？智能体如何在不使用任何辅助函数的情况下，自主确定非单体多模式探索的切换？

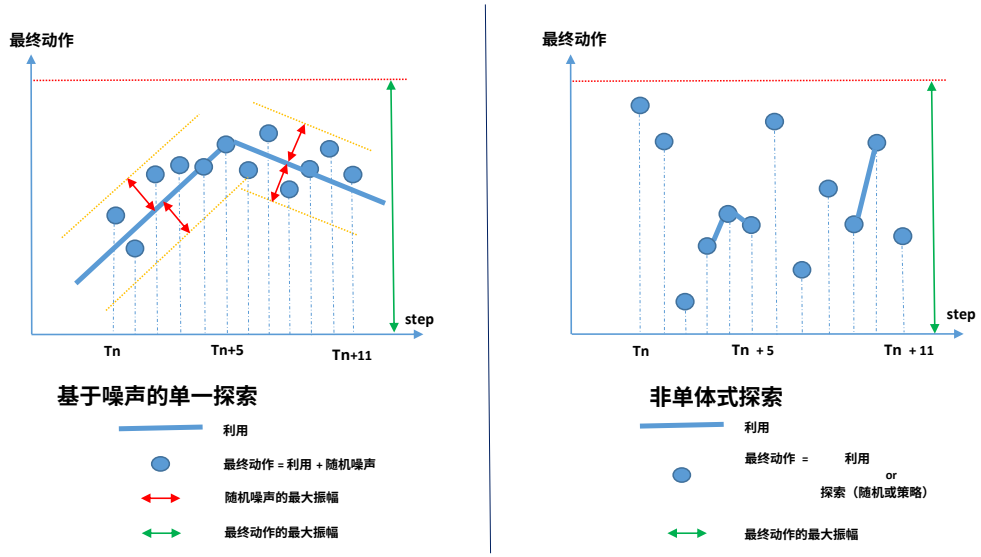

图 1：基于噪声的单一探索（左）和非单体式探索（右）的示例。为了便于理解，本例中的最终动作是一个标量，表示智能体在每一步执行的动作，智能体以实心圆表示。实线表示利用，即行为策略的原始动作。基于噪声的单一探索中的实心圆表示最终动作，它结合了行为策略的原始动作和每一步采样得到的有界噪声。然而，非单体式探索中的实心圆则根据每一步的模式定义，即利用（行为策略的原始动作）或探索（随机噪声或策略）。

例如“稳态”？智能体如何避免策略固有的扰动，同时拥有鲁棒最优策略？

值得一提的是，文献中没有类似的研究，因此参考方法部分基于 [13] 提出的方法，尽管他们的工作并非基于选项框架。最终，我们的探索方法表现更好。

贡献总结如下。

- 支持自主非单体式多模式探索的选项框架模型开发：针对前 3 个研究问题，我们提出了一种新颖的分层强化学习模型架构，以支持自主非单体式多模式探索。

- 利用策略的内在特征开发非单体式探索的切换方法：针对第四个研究问题，我们的模型通过奖励修改采用了引导式探索。

- 提升策略的鲁棒性：针对最后一个研究问题，通过利用评估过程，可以确保获得鲁棒的最优策略。

下面将对本研究的其余内容作如下说明。第II节综述与本研究相关的探索和分层强化学习研究。第III 节介绍我们提出的模型。第IV节描述了用于性能测量的实验，将我们的模型与非单体模型进行比较， [13],作为参考模型和单体探索方法 HIRO。我们在实验中讨论了若干已获认可的问题第V节。最后，我们在第VI节介绍当前研究的结论以及对未来工作的建议。

II. 相关工作

A. 选项框架

马尔可夫决策过程中的选项集合是动作这一广义概念的体现，由一个半马尔可夫决策过程（SMDP）构成。半马尔可夫决策过程旨在根据可变时长处理抽象动作的不同层级。分层强化学习是强化学习的一种代表性泛化形式，其中环境被建模为半马尔可夫决策过程 [14]。

采用多模式探索的非单体式探索模式中，每个动作在不同时间段产生的效果各不相同。因此，动作序列通过利用多模式探索的选项框架来定义。

B. 探索

无论各种触发器事件是否源自不确定性，都已对其进行了考虑 [15], [16],[17]。

[18] 的实验展示了机器人通过自主探索以及在人类伙伴支持下进行的社会引导探索来学习行为的有效性。[19]提出了关于贝叶斯框架的主张，该框架支持在线适应变化动力学，并通过使用方差奖励揭示逆境转移程度来避免保守倾向。 [20]提出了战术乐观与悲观（TOP）估计

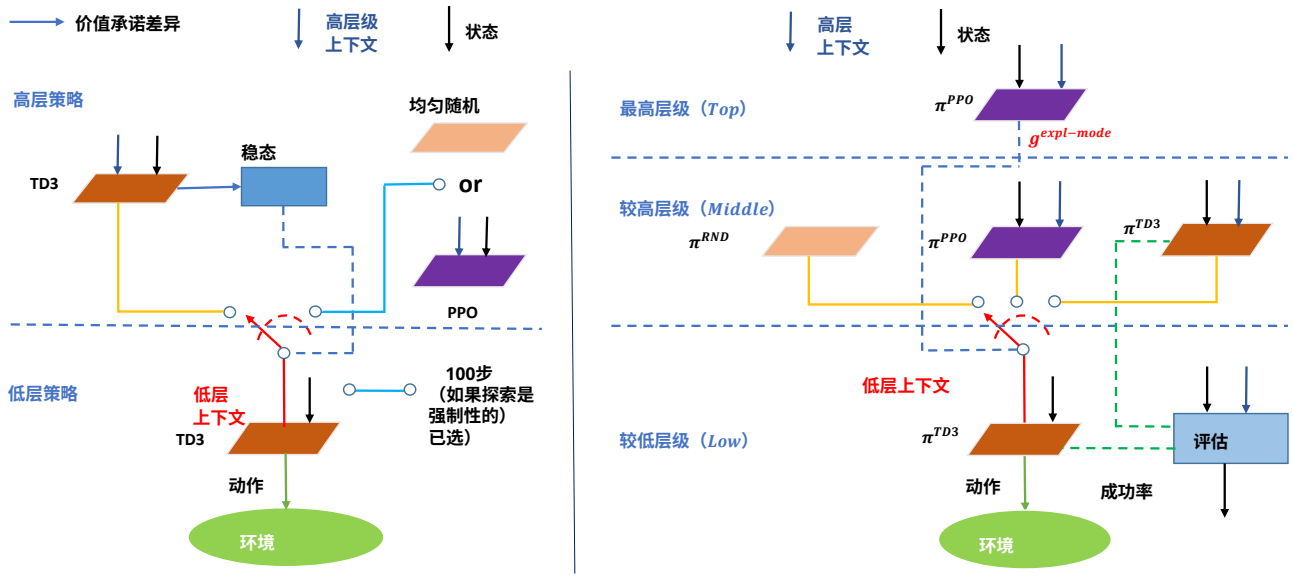

图 2：我们提出的模型（右）的架构与参考论文中使用稳态的模型架构（左）对比 [13]

，用于通过分位数近似 [21]实现乐观与悲观在线价值估计的策略。因此，信念分布按照分位数估计构建：

$$q_{Z^{\pi}(s,a)} = q_{\bar{Z}(s,a)} + \beta q_{\sigma(s,a)} \quad (1)$$

其中，$q_{\bar{Z}(s,\ a)}$和 $q_{\sigma(s,a)}$分别是分位数估计的均值和标准差，Z(s,a) 是一个回报随机变量。当 $\beta \geq 0$ 和 $\beta < 0$ 时，信念分布分别是乐观的和悲观的。

[13]指出，相较于单一探索，非单一探索十分重要。其代表性的非单一探索方法利用「稳态」，该方法基于相隔 k 步的价值函数差异，称为以下「价值承诺差异」：

$$D_{promise}(t-k,t) := \left|V(s_{t-k}) - \sum_{i=0}^{k-1} \gamma^i R_{t-i} - \gamma^k V(s_t)\right| \quad (2)$$

其中，$V(s)$ 是智能体在状态 s 下的价值估计，γ 是折扣因子，R 是奖励。

上述关于引导探索框架、稳健马尔可夫决策过程研究以及自适应乐观的研究，启发我们基于 [13] 开展研究。

C. 分层强化学习

半马尔可夫决策过程（SMDP）利用选项来定义领域知识，并能够重用子目标的解决方案 [14]。 [22] 声称，自适应技能、自适应划分框架支持学习近最优技能，这些技能通过技能划分自动并行组合而成，并对初始的错误指定模型进行修正。 [23] 提出了一种解决稀疏奖励任务的算法，研究表明，通过构建可最小化覆盖时间的选项，该算法能够加速探索。 [24] 也针对稀疏奖励环境展开研究。因此，它正式提出了快慢好奇心的概念，以促进长时间范围探索。选项评论家架构包含内部策略，该策略遵循策略从选项中选出的选项，直至满足选项终止条件 [25]。[26]声称，分层强化学习中的每个策略都利用流式深度生成模型来保持完整的表达能力，并通过潜变量以自底向上的逐层方法进行训练。HIRO提出了一种同步分层强化学习相邻层级的方法，以高效训练高层策略。

我们的模型在探索研究中采用 HIRO，因为它是一种传统的目标条件分层强化学习方法。

III. 我们的模型

鉴于 [13],中出现的价值承诺差异问题日益突出，我们的研究重点关注一种自主多模式非单体探索模型，其中智能体会自主决定何时启动和退出探索模式。此外，预期模型会利用其内在特性来采取动作。为此，我们的研究采用了选项框架。

表 I：关键符号说明。

符号	含义
t	动作步
$state$	当前状态
$next_state$	下一状态
Top	最高层
$Middle$	较高层
Low	较低层
$Action$	低层的动作低层 （ π_L^{TD3}的动作）
$target_pos$	顶层顶层 和中层的上下文
$goal$	三个子项的当前低层上下文 中层策略（当前目标为 π_L^{TD3}）
$next_goal$	三个子策略的下一个低层上下文 中层策略（ π_L^{TD3}的下一个目标）
R	从环境中获得的奖励（ 在 OpenAI 的 Ant领域中，R 的符号为负 Gym)
π_T^{PPO}	顶层的策略（同策略）
π_M^{TD3}	中层的策略（离策略）中间
π_M^{PPO}	中层的策略（同策略）
π_M^{RND}	策略（随机策略中的均匀随机。 icy) of中间 级
π_L^{TD3}	低层级的策略（离策略）
$g^{\text{expl-mode}}$	π_T^{PPO}的顶部层动作
$\alpha_{g^{\text{expl-mode}}}$	根据 $g^{\text{expl-mode}}$ 的 α 预设值
$S_O_{g^{\text{expl-mode}}}$	根据 $g^{\text{expl-mode}}$ 得出的 $Success_ratio$ 的参考值
$loss$	π_T^{PPO}的损失（顶层）
S_E	评估函数的 $Success\ rate$ _ π^{TD3} 的中级 层
$Done_m$	完成在顶层时间范围内的次数
R_m	R 在顶层级别时域内的总和
$Count_m$	顶层级在规划视野内的计数
S_O_m	关于 $g^{\text{expl-mode}}$ of 的成功次数比例 π_T^{PPO} 在顶层的时间范围内
ρ	目标速率的预设值，即平均值 参考模型的切换次数

选项框架尤其适用于目标条件分层强化学习，可通过完全依赖状态的分层策略来控制多模式探索。针对第I节提出的第一个研究问题，我们的模型包含三个层级的分层强化学习，如图 2所示，并结合了 [13]的实现模型。我们的模型根据层级的高度为各层命名：Top、 $Middle$和 Low。各层的策略分别为： π_T^{PPO}对应顶层， π_M^{TD3}、 π_M^{PPO}和 π_M^{RND}对应中层，以及 π_L^{TD3}对应底层。

分层控制过程易于系统地构建多模式探索， $g^{\text{expl-mode}}$，将其作为针对稳态等函数控制的选项。探索模式策略 π_T^{PPO} 可从以下三个中层策略中进行选择：中层

$$g^{\text{expl-mode}} \sim \pi_T^{PPO}. \tag{3}$$

因此， $g^{\text{expl-mode}}$ 的值表示两种探索模式之一，即均匀随机或PPO，或者表示一种利用方式，即TD3。此外，选项框架的固有特性也为探索提供了若干控制优势。

为了实现我们的研究目的，我们的选项框架模型包含四个要素：策略本身固有-的切换模式决策、赋予探索更多的熵维度、引导探索模式，以及采用评估过程来确保鲁棒性。

A. 策略本身固有的切换模式决策

由于我们的模型采用了 π_T^{PPO}固有的训练方法，因此可以根据中层 选项 $g^{\text{expl-mode}}$从 π_T^{PPO}中选择一个策略。 π_T^{PPO}是在没有辅助机制的情况下，将所有模式下策略的奖励最大化整合到其自身策略中。这意味着，探索和利用的持续时间都由智能体的固有特性控制。最终，非单体式探索模式策略的所有特性都可以整合到策略的奖励最大化中，以回答第I节中的第二个研究问题。我们可以根据各探索模式的次数验证切换模式的选择，如第IV节所示。

B. 为探索赋予更多熵选择

我们的模型根据探索模式的熵程度，将其视为乐观程度，针对探索模式策略进行多模式探索。我们的模型有两种探索模式，分别是 π_M^{RND}和 π_M^{PPO}，以及一个利用策略 π_M^{TD3}，位于第I节第三个研究问题的中层级。因此，在训练智能体的过程中，我们假设三个策略各自的熵程度如下：

$$H\left(\pi_M^{RND}\right) > H\left(\pi_M^{PPO}\right) > H\left(\pi_M^{TD3}\right) \tag{4}$$

其中，$H\left(\pi_M^{\bullet}\right)$表示策略的总体熵 $\pi_M^{\bullet}$。

我们的模型仅将 PPO 用于一种探索模式，因此在训练结束时会将其丢弃。我们的模型只将离策略 TD3 作为最终目标策略。同时，只要由于中层级的三个子策略之一而产生数据，PPO 和 TD3 就会共同进行训练。如果 PPO 得到了一定程度的训练，我们的模型预计，关于探索结果，PPO 的性能将高于均匀随机的性能。

C. 引导式探索

在智能体训练过程中，潜在奖励进展分为两个阶段。因此，为了维持第一阶段，我们的模型将对智能体进行引导式探索纳入考虑。由于我们的模型在目标条件分层强化学习中采用选项框架，探索模式策略可以遵循奖励最大化策略，从而使用预设参数$\alpha_{g^{\text{expl-mode}}}$ 对来自环境的奖励 R 进行修正，具体为

$$R_{final} = R + \alpha_{g^{\text{expl-mode}}} * R \tag{5}$$

其中， R_{final}表示根据预设参数$\alpha_{g^{\text{expl-mode}}}$ 和环境奖励 R得到的修正奖励。

根据 $g^{\text{expl-mode}}$ 的类型， $\alpha_{g^{\text{expl-mode}}}$ 的值可能预先设定为不同的值，有时也可能相同

$$\alpha_{\text{uniform random}} > \alpha_{\text{ppo}} > or\ equal\ to\ \alpha_{\text{td3}} \tag{6}$$

算法1 基于选项框架的多探索模式

1: **初始化：**
根据 $g^{expl\text{-}mode}$ **的** π_T^{PPO} **设置** $\alpha_{g^{expl\text{-}mode}}$ **的值；根据** $g^{expl\text{-}mode}$ π_T^{PPO} **的** $S_O_{g^{expl\text{-}mode}}$ **设置** $S_O_{g^{expl\text{-}mode}}$
2
的值：过程 Evaluate_π_M^{TD3}**(...,** π_M^{TD3}, π_L^{TD3}**, ...) 3：根据** π_M^{TD3} *and* π_L^{TD3}**，计算 S_E 4：结束过程 5：过程 Train**$_T$**(...,** S_E**,** $g^{\text{expl-mode}}$**, ...) 6：如果** $g^{\text{expl-mode}}$ **是均匀随机或 PPO，则 7：**$loss_{final} = loss + S_E * loss$ **8：否则 9：**$loss_{final} = loss - S_E * loss$**10：结束条件判断 11：结束过程 12：对于** $t = 0, \ldots, T-1$**，执行 13：**$action \leftarrow Clamp_Max(\pi_L^{TD3}(state, goal) + Noise)$ **14：执行动作，观察** R **和** $next_state$**15：**$Done \leftarrow Judge_success(state, target_pos)$ **16：如果 Done 为 True，则将** $Done_m$ **增加 1 17：将** R_m**增加** R**18：将** $Count_m$ **增加 1 19：如果达到顶层的时间范围，则 20：**$S_O_m \leftarrow Done_m \,|\, Count_m$**； 21：**$R_{final} \leftarrow R_m + \alpha_{g^{\text{expl-mode}}} * R_m$**22：如果**
$S_O_m >= S_O_{g^{\text{expl-mode}}}$**，则 23：**$Done\ m \leftarrow True$**_24：结束条件判断 25：如果达到顶层的训练时间，则 26：Train**$_T$**(...,** S_E**,** $g^{\text{expl-mode}}$**, ...) 27：结束条件判断 28：如果处于起始模式，则 29：**$g^{\text{expl-mode}} \leftarrow Random\ policy$ **30：否则 31：**$g^{\text{expl-mode}} \sim \pi_T^{PPO}$**32：结束条件判断 33：**$R_m, Count_m, Done_m \leftarrow 0$**34：结束条件判断 35：如果达到中层的时间范围，则**
36 $g^{expl\text{-}mode}$ $next_goal \sim \pi_M^{RND}$ π_M^{PPO} π_M^{TD3}**:**
37

通过 、 、 或 ：如果达到 π_M^{PPO} **的训练步数，则 38：Train_**π_M^{PPO}**(...) 39：结束条件判断 40：结束条件判断 41：**$state \leftarrow next_state$**；** $goal \leftarrow next_goal$**42：如果达到** π_M^{TD3} **的训练步数，则 43：Train_**π_M^{TD3}**(...) 44：结束条件判断 45：如果达到中层的评估步数，则 46：Evaluate_**π_M^{TD3}**(...,** π_M^{TD3}**,** π_M^{TD3}**, ...) 47：结束条件判断 48：结束循环**

其中，$\alpha_{\text{uniform random}}$，$\alpha_{\text{ppo}}$， *and* α_{td3} 表示为$\alpha_{g^{\text{expl-mode}}}$ π_M^{RND}、 π_M^{PPO}和 π_M^{TD3}预设的参数，分别对应*Middle*。

最后，由于 R_{final}的值用于探索模式策略的训练，因此，针对第 I 节中的第四个研究问题，探索模式策略会根据$\alpha_{g^{\text{expl-mode}}}$的值选择奖励最大化的选项。随着 $\alpha_{g^{\text{expl-mode}}}$ 的值增大，其探索模式的发生概率会降低。

D. 鲁棒性评估

对于智能体潜在奖励进展的第二阶段，我们的模型采用在线评估过程，以维持鲁棒最优策略。在线评估过程中成功率的出现表明，本研究中的智能体性能进入了奖励进展的第二阶段。从第二阶段开始，我们的智能体需要通过在线评估过程获得鲁棒最优策略。在线过程每隔预设步数评估离策略 π_M^{TD3}，然后根据 $g^{\text{expl-mode}}$的类型输出成功率 S_E。因此，本研究计算了第 I 节中第五个研究问题所对应的探索模式策略 $loss_{final}$ π_T^{PPO}，计算方式如下：

$$loss_{final} = loss + S_E * loss \tag{7}$$

其中， $loss_{final}$是根据 S_E修改后的损失。

在我们的模型中，随着 S_E 的在线值增加，均匀随机模式下探索模式策略和在线策略的损失会变得大于其原始在线损失。

IV. 实验

我们的模型作为自主非单体式智能体的多模式探索控制，通过探索模式和利用的计数来体现，因为这些计数对于分析我们的模型至关重要。每个计数都描述了策略在所有模式上的当前奖励最大化情况。通过分析，我们旨在回答以下关键问题。我们的模型能否表现得优于参考论文中的代表性模型、*[13],* 以及基于噪声的单一探索策略？我们在 OpenAI Gym 的 Ant领域中的两个任务 Ant Push 和 Ant Fall 上，对我们的模型及其进行评估。用于比较的参考模型是 [13],XU-intra(100, informed， p^*， X)和 XI-intra(100, informed， p^*， X)，在我们的参考模型中分别称为“Ref:Uniform random”和“Ref:PPO”[1]。PPO 被用于参考模型的内在探索模式。基于噪声的单一探索策略是 HIRO，它在每个层级都由 TD3 组成。我们的模型和两个参考模型同样基于 HIRO 实现。为了评估三个模型中的最佳性能，我们设置了以下四个分析项目：

1）整个训练过程中分配给每种策略的次数是多少?

[1]请阅读 [13]的第 3.1 节，了解实验详情

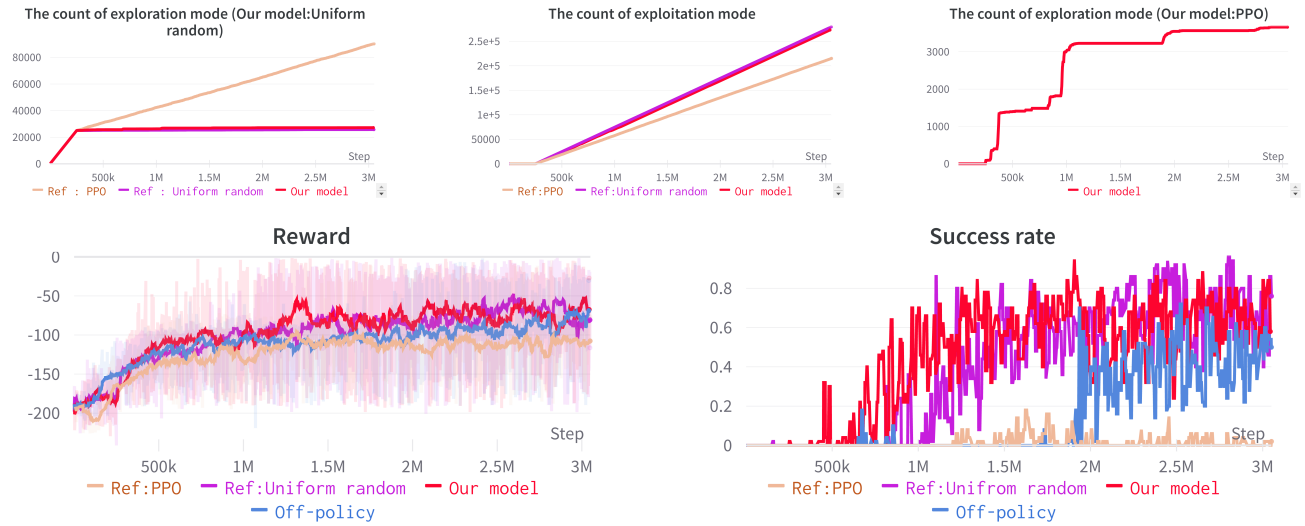

图 3：在 Ant Push 任务中，我们的模型、参考论文：均匀随机、参考论文：PPO 和 HIRO 的探索模式与利用次数，以及高层策略的奖励和成功率



图 4：在 Ant Fall 任务中，我们的模型、参考论文：均匀随机、参考论文：PPO 和 HIRO 的探索模式与利用次数，以及高层策略的奖励和成功率

2）与参考模型被迫从探索转为利用的过程相比，我们的模型如何在探索模式与利用之间进行转移？3）基于引导式探索策略时，均匀随机与同策略作为我们模型探索策略的差异有多大？4）评估过程对第二奖励阶段的表现有多大影响?

AntPush 和 AntFall 的结果分别见图 3 和图 4 。此外，我们的源代码和实现细节已在线公开[2]。算法 1 展示了基于参考代码实现的算法主体部分。

A. 与参考论文和纯离策略方法的比较

1)Ant Push: 我们的模型在几乎所有训练步数中都优于其他模型。我们的模型和两个参考模型在大多数训练步数中都进行利用。我们的模型和两个参考模型的探索模式所占比重均低于利用。HIRO 在早期表现最佳，但随着其他模型利用多样化的探索模式，其潜力很快便在整个训练过程中逐渐丧失。“Ref:均匀随机”的表现优于“Ref:PPO”。

[2]https://github.com/jangikim2/An-Autonomous-Non-monolithic-Agent-with-Multi-mode-Exploration-based-on-Options-Framework



图5：针对 Ant Push 中正常模型的三种消融研究

我们的模型的均匀随机探索模式和“Ref:Uniform random”并不会持续很长时间，而是短暂且逐渐地进行。同时，根据预设目标率 ρ，“Ref:PPO”的探索程度高于“Ref:Uniform random”，并且持续探索会在起始模式之后发生。

在算法中的起始模式之后 1，我们模型的 PPO 探索模式约有 3600 步，多于模型的均匀随机探索模式（约 2100 步）。我们模型的 PPO 探索模式大多发生在 1M 步之前。我们模型的两种探索模式与利用的总步数对比如下

$$Total_Step\left(\pi_M^{TD3}\right) >> Total_Step\left(\pi_M^{PPO}\right) > Total_Step\left(\pi_M^{RND}\right) \quad (8)$$

总_步数$\left(\pi_M^{\cdot}\right)$表示 $\pi_M^{\cdot}$ 的策略已执行的总步数。

引导式探索策略根据奖励的修改（公式（5））以及成功次数的比率 S_O_m，生成我们模型的探索模式。

Ant PUSH 任务的第二阶段从奖励超过 -100 时对应的步数开始，因为评估过程的成功率 S_E约从 500K 步开始出现，并在 1M 步时超过 0.6。由于损失修改（公式（7））依赖于评估过程，因此奖励崩溃的情况不会长时间发生。

2) 蚂蚁坠落： 经过 3M 步数后，我们的模型相较于所有其他模型都表现出具有竞争力的性能。蚂蚁坠落的奖励修改预设与蚂蚁推挤的不同，这意味着 $\alpha_{\text{on-policy}}$ 等于 $\alpha_{\text{off-policy}}$。我们的模型的利用阶段比两个参考模型少持续了 1M 步数，这与蚂蚁推挤中的情况不同。HIRO 的性能保持稳定，并在后期有所下降。与蚂蚁推挤不同，“Ref:PPO”的性能优于“Ref:均匀随机”。

在蚂蚁坠落中，“Ref:均匀随机”和我们的模型的探索模式持续时间，比蚂蚁推挤中“Ref:均匀随机”和我们的模型的探索模式更长。与此同时，“Ref:PPO”和“Ref:均匀随机”的探索模式几乎相同，因为“Ref:均匀随机”和“Ref:均匀随机”的预设目标率 ρ 相同。

虽然 $\alpha_{\text{on-policy}}$ 等于 $\alpha_{\text{off-policy}}$，但由于第二层各个动作的成功次数比例也经过了修改，因此在起始模式之后，我们模型中两种探索模式与利用模式的总步数比较贯穿整个训练步数，如下所示

$$Total_Step\left(\pi_M^{TD3}\right) > Total_Step\left(\pi_M^{PPO}\right) >> Total_Step\left(\pi_M^{RND}\right). \quad (9)$$

与 Ant PUSH 任务不同，AntFall 任务的第二阶段在 4M 至 4.5 M 步数之间出现奖励下降。在此期间，评估过程的成功率维持在 0.5 至 0.6 之间。与 IV-B2 相比，由于成功率的作用，奖励很快恢复。

B. 消融研究

我们研究了不采用奖励修改、不采用损失修改以及两者均不采用的模型。图 5 展示了在 Ant Push 任务中，与我们的正常模型进行对比的实验结果。因此，为了分别对各情况进行实验，我们移除了与 Ant Push 中正常模型相关的部分。

1) 不采用奖励修改：利用模式的步数少于我们的正常模型，而均匀随机和 PPO 的探索步数则多于我们的正常模型。奖励表现和成功率缓慢提升。

2) 不进行损失修改：同样，这两种探索模式的步数都比我们的普通模型更多，而利用模式的步数则比我们的普通模型更少。由于，奖励在 220 万步至 300 万步之间有所下降。

PPO 探索的增加。尽管这段时间的成功率高于我们的普通模型，但其表现不如我们的普通模型。

3) 不进行奖励修改和损失修改：探索过多而利用不足，导致性能最差。

V. 讨论

A. 同策略对探索的影响

在探索初期采用同策略时，同策略 π_M^{PPO} 的表现并不具竞争力。然而，经过自身或其他策略的一定程度训练后，同策略的表现优于随机策略。同时，在实践中，π_T^{PPO} 不太可能陷入局部极小值，因为有三种中级 策略。

B. 奖励修改的影响

在 Ant Fall 任务中，我们的模型在前期步数（约 2M 步以内）的表现落后于所有其他模型。原因是，在此之前同策略运行了很长时间，因为 $\alpha_{\text{on-policy}}$ 等于 $\alpha_{\text{off-policy}}$。引导式探索的奖励修改利用了 $\alpha_{g^{\text{expl-mode}}}$ 的固定值，这并非一种自适应策略。

C. 损失修改的影响

在 Ant Fall 任务中，400 万至 450 万步之间出现的同策略策略和随机策略，导致智能体的性能下降。具体而言，可以考虑对反映成功率的 S_E 进行建模。S_E 越高，不确定性就越低。因此，S_E 与不确定性相关。

VI. 结论

为克服非单体式探索的问题，本文提出了一种基于选项框架、采用多模式探索的自主非单体式智能体。我们揭示了该模型遵循人类和动物行为模式的潜力。我们的模型借助引导探索框架，利用各探索策略熵的差异。通过评估过程，有望获得鲁棒最优策略。未来需要研究自主非单体式智能体多模式探索的自适应策略的引导式探索。为获得鲁棒最优策略，还需要进一步研究智能体中 S_E 的建模。

参考文献

[1] J. Fu、J. Co-Reyes 和 S. Levine，“Ex2：利用样例模型进行深度强化学习探索”，神经信息处理系统进展，第 30 卷，2017年。[2] Y. Burda、H. Edwards、A. Storkey 和 O. Klimov，“通过随机网络蒸馏进行探索”，*arXiv* 预印本 *arXiv:1810.12894*，2018年。[3] H. Tang、R. Houthooft、D. Foote、A. Stooke、O. Xi Chen、Y. Duan、J. Schulman、F. DeTurck 和 P. Abbeel，“#探索：深度强化学习中基于计数的探索研究”，神经信息处理系统进展，第 30 卷，2017年。

[4] Y. Burda、H. Edwards、D. Pathak、A. Storkey、T. Darrell 和 A. A. Efros，“大规模好奇心驱动学习研究”，arXiv 预印本 arXiv:1808.04355，2018年。[5] B. Eysenbach、A. Gupta、J. Ibarz 和 S. Levine，“多样性就是你所需要的一切：无奖励函数学习技能”，arXiv 预印本 arXiv:1802.06070，2018年。[6] K. Gregor、D. J. Rezende 和 D. Wierstra，“变分内在控制”，arXiv 预印本 arXiv:1611.07507，2016年。[7] J. Achiam、H. Edwards、D. Amodei 和 P. Abbeel，“变分选项发现算法”，arXiv 预印本 arXiv:1807.10299，2018年。[8] A. P. Badia、P. Sprechmann、A. Vitvitskyi、D. Guo、B. Piot、S. Kapturowski、O. Tieleman、M. Arjovsky、A. Pritzel 和 A. Bolt，“永不放弃：学习定向探索策略”，arXiv 预印本 arXiv:2002.06038，2020年。[9] N. Savinov、A. Raichuk、R. Marinier、D. Vincent、M. Pollefeys、T. Lillicrap 和 S. Gelly，“通过可达性实现情节好奇心”，arXiv 预印本 arXiv:1810.02274，2018年。[10] A. Ecoffet、J. Huizinga、J. Lehman、K. O. Stanley 和 J. Clune，“Go-explore：高难度探索问题的一种新方法”，arXiv 预印本 arXiv:1901.10995，2019年。[11] Y. Guo、J. Choi、M. Moczulski、S. Feng、S. Bengio、M. Norouzi 和 H. Lee，“基于记忆的轨迹条件策略：从稀疏奖励中学习”，神经信息处理系统进展，第 33 卷，第 4333–4345 页，2020年。[12] I. Osband、C. Blundell、A. Pritzel 和 B. Van Roy，“通过自助法 DQN 进行深度探索”，神经信息处理系统进展，第 29 卷，2016年。[13] M. Pislar、D. Szepesvari、G. Ostrovski、D. Borsa 和 T. Schaul，“智能体何时应该探索？”，arXiv 预印本 arXiv:2108.11811，2021年。[14] R. S. Sutton、D. Precup 和 S. Singh，“介于马尔可夫决策过程与半马尔可夫决策过程之间：强化学习中时间抽象的框架”，人工智能，第 112 卷，第 1–2 期，第 181–211 页，1999年。[15] M. A. Wiering 和 H. Van Hasselt，“强化学习中的集成算法”，IEEE 系统、人类与控制论汇刊 B 部分（控制论），第 38 卷，第 4 期，第 930–936 页，2008年。[16] J. Buckman、D. Hafner、G. Tucker、E. Brevdo 和 H. Lee，“通过随机集成价值扩展实现样本高效的强化学习”，神经信息处理系统进展，第 31 卷，2018年。[17] S. Flennerhag、J. X. Wang、P. Sprechmann、F. Visin、A. Galashov、S. Kapturowski、D. L. Borsa、N. Heess、A. Barreto 和 R. Pascanu，“将时序差分不确定性作为探索信号”，arXiv 预印本 arXiv:2010.02255，2020年。[18] A. L. Thomaz 和 C. Breazeal，“社会引导探索实验：构建能在有人类教师指导或无指导的情况下学习的机器人的经验教训”，Connection Science，第 20 卷，第 2–3 期，第 91–110 页，2008年。[19] E. Derman、D. Mankowitz、T. Mann 和 S. Mannor，“鲁棒强化学习的贝叶斯方法”，载于《人工智能中的不确定性》。PMLR，2020年，会议论文集，第 648–658 页。[20] T. Moskovitz、J. Parker-Holder、A. Pacchiano、M. Arbel 和 M. Jordan，“深度强化学习中的战术性乐观与悲观”，神经信息处理系统进展，第 34 卷，第 12 849–12 863 页，2021年。[21] W. Dabney、M. Rowland、M. Bellemare 和 R. Munos，“采用分位数回归的分布式强化学习”，载于《AAAI 人工智能大会论文集》，第 32 卷，2018年，会议论文集。[22] D. J. Mankowitz、T. A. Mann 和 S. Mannor，“自适应技能自适应划分（ASAP）”，神经信息处理系统进展，第 29 卷，2016年。[23] Y. Jinnai、J. W. Park、D. Abel 和 G. Konidaris，“通过最小化覆盖时间发现用于探索的选项”，载于《国际机器学习大会》。PMLR，2019年，会议论文集，第 3130–3139 页。[24] N. Bougie 和 R. Ichise，“强化学习中用于高层探索的快慢好奇心”，Applied Intelligence，第 51 卷，第 2 期，第 1086–1107 页，2021年。[25] P.-L. Bacon、J. Harb 和 D. Precup，“选项评论家架构”，载于《AAAI 人工智能大会论文集》，第 31 卷，第 1 期，2017年。[26] T. Haarnoja、K. Hartikainen、P. Abbeel 和 S. Levine，“用于分层强化学习的潜在空间策略”，载于《国际机器学习大会》。PMLR，2018年，第 1851–1860 页。