arXiv:2303.05038v2 [cs.AI] 4 Apr 2024

利用上下文结构生成有用的辅助任务

Benedict Quartey计算机科学系 布朗大学

Ankit Shah计算机科学系 布朗大学

George Konidaris计算机科学系 布朗大学

摘要

强化学习需要与环境交互，而这可能代价高昂到难以承受，尤其是在机器人学领域。这一限制要求我们采取能够在环境交互有限的情况下，通过最大限度地复用以往经验来解决问题的方法。我们提出一种方法，在学习解决给定任务的同时，通过生成并同步学习有用的辅助任务，最大限度地复用经验。为生成这些任务，我们构建给定任务的抽象时序逻辑表示，并利用大语言模型生成上下文感知对象嵌入，以促进对象替换。反事实推理和离策略方法使我们能够在解决给定目标任务的同时学习这些辅助任务。我们将这些见解融入一个新的多任务强化学习框架，并通过实验表明，生成的辅助任务与给定任务具有相似的底层探索需求，从而最大限度地发挥定向探索的效用。我们的方法使智能体无需额外的环境交互，就能自动学习更多有用的策略。

1 引言

强化学习（RL）是一种通用范式，用于建模智能体与环境之间的交互。它已被证明是解决序贯决策问题的一种稳健方法，近年来也取得了若干令人振奋的成果 [1, 2, 3]。然而，要学会能够完成任务的有价值行为，智能体必须通过反复与环境交互来进行探索，而这可能代价高昂，尤其是在机器人学中。因此，必须高效利用经验数据。一种方法是利用这样一个事实：智能体在探索并解决某项任务时，会获得有关环境的经验，而这些经验也可能有助于学习解决许多密切相关的任务。考虑在一所陌生的房子里学习泡茶这一任务。要完成这项任务，一个简单的目标序列可以是：前往厨房，拿一个杯子，拿一个茶包，从水龙头接水，以及从冰箱里拿牛奶。智能体在学习完成这项任务时，一旦发现了冰箱的位置，就应该能够重复利用这一经验来完成相关任务，例如从冰箱里拿一杯饮料。

以往的经验回放工作 [4, 5, 6] 介绍了重用先前经验的有效方法，而离策略学习算法 [7, 8, 9] 也支持跨任务学习；智能体可以利用由行为策略 生成的数据，学习执行不同的目标策略。然而，这些方法在指定辅助任务方面能力有限，无法充分发挥反事实经验推理和离策略学习的效益，部分原因在于它们没有对任务结构作出任何假设。

NeurIPS 2023 规划泛化研讨会（GenPlan 2023）。

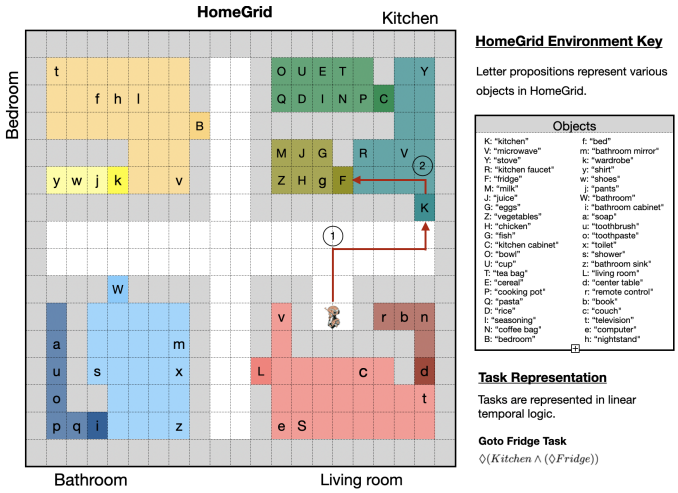


图1：HomeGrid，一个确定性离散网格世界领域。在这个世界中，智能体通过前往环境中与对象对应的网格位置来完成任务。任务使用 LTL公式进行描述，这些公式表示完成给定任务所需子目标的序列/顺序。满足这些任务需要按任务规范所确定的可接受顺序访问相关网格单元。例如，图中的编号箭头表示满足**前往冰箱**任务的策略。

我们认为，可以利用任务结构以及以对象为中心的环境中的语义和上下文结构来生成辅助任务。我们提出了一种新方法 TaskExplore：给定一个目标任务，该方法使用时序逻辑表达式生成上下文相似的辅助任务，并通过替换对象来实现替换所用的上下文感知嵌入由大语言模型生成。在这种设定下，我们可以利用反事实推理和离策略方法，在学习给定任务的同时学习这些辅助任务，而行为策略仅以给定任务为条件。我们的方法能够最大限度地发挥定向探索经验的效用，尤其适用于需要智能体限制或引导探索的复杂环境。我们通过实验证明，TaskExplore 生成的辅助任务能够最大限度地利用单任务课程中的定向经验。

总之，本文提出了两项主要贡献：

1. 我们提出一种方法，利用上下文感知对象嵌入和抽象时序逻辑任务表示，生成与给定目标任务具有相同底层探索需求的有用辅助任务。2. 我们通过实验证明，这类生成任务比随机生成的任务集能带来更好的经验迁移，并且能独特地受益于对主要任务进行定向探索。

2 背景

在这项工作中，我们聚焦于以对象为中心的环境，并利用线性时序逻辑 [10] 描述涉及对象的时间扩展任务。这些任务使用针对线性时序逻辑 [9]的离策略学习进行学习。

2.1 线性时序逻辑

线性时序逻辑（LTL） [10], 提供了一种富有表现力的语法，用于描述时序行为。LTL 表达式/公式由原子命题、逻辑连接词（否定（¬）、合取（∧）、析取（∨）、蕴含（→））以及一组时序算子（下一时刻（◯）、直到（U）、总是（□）和最终（◊））组成。LTL 公式的最简语法定义如下：

$$\varphi := p \mid \neg\varphi_1 \mid \varphi_1 \wedge \varphi_2 \mid \bigcirc \varphi \mid \varphi_1 \mathrm{U} \varphi_2 \tag{1}$$

其中， p 是一个原子命题，即用于描述环境某种属性的布尔文字； φ_1和 φ_2 是有效的 LTL 公式。在强化学习场景中，LTL 公式为描述奖励目标提供了一种替代方式，通常更具表现力且更自然 [11]。它们能够表达显式规格，用以刻画任务的成功执行。考虑一个序贯导航任务：先前往厨房，再前往冰箱，其中厨房 和冰箱是可观测的布尔原子命题。可以使用下列 LTL 公式来描述此任务：

$$\Diamond(Kitchen \wedge (\Diamond Fridge)) \tag{2}$$

给定命题的真值赋值序列 [12] ，可以推进线性时序逻辑公式，以确定截至目前所见状态已满足公式的哪些部分，以及哪些部分仍未满足。这在强化学习领域尤其有用，因为它提供了一种跟踪非马尔可夫目标的方法。事实上，几何线性时序逻辑（G-LTL） [11], 奖励机 Q 学习（Q-RM） [8] 和 LPOPL [9]等一类强化学习算法利用这一方法，使用环境状态空间和线性时序逻辑规范的自动机表示来构建并求解乘积马尔可夫决策过程。

2.2 基于线性时序逻辑的离策略学习

离策略强化学习方法处理这样一种情形：智能体在环境中交互时使用一种目标策略，同时学习所需的行为策略。线性时序逻辑离策略学习推进（LPOPL） [9] 调整 Q 学习离策略算法 [13], 以便在同一环境交互过程中同时学习多个线性时序逻辑任务的策略。

给定一组线性时序逻辑任务规范 ϕ，LPOPL 首先通过线性时序逻辑推进 [12], 从每个规范中提取子任务，即对给定的每个公式，遍历环境命题集合的所有可能真值赋值并推进公式。随后，LPOPL 反复执行一系列回合，并为每个任务及提取出的子任务（也以 LTL公式表示）分别学习一个 Q值函数。每次迭代时，从 ϕ 中选取一个任务作为该回合的目标，并根据该任务采用 ε-贪心行为策略选择动作。不过，每一步都会通过离策略更新更新所有 Q值函数，使智能体能够推进那些当前并非其目标的任务。

为帮助说明 LPOPL，考虑一个包含给定任务集合的示例： $\phi = \{\Diamond（a \wedge \Diamond b） \ \Diamond（c \wedge \Diamond b）\}$。推进这两个任务会提取出子任务 $\Diamond b$。随后，LPOPL 将初始化三个 Q值函数 $Q_{\Diamond(a \wedge \Diamond b)}$、 $Q_{\Diamond(c \wedge \Diamond b)}$ 和 $Q_{\Diamond b}$。执行一个回合时，将从 ϕ 中选取一个任务，例如 $\Diamond （ a \wedge \Diamond b ）$，然后根据 $Q_{\Diamond(a \wedge \Diamond b)}$ 采用 ε-贪心方式选择动作。回合期间，一旦命题 a 变为 真 ，行为策略就会转为根据 $Q_{\Diamond b}$ 采用 ε-贪心策略，因为在命题 a 的赋值为 True 时推进 $\Diamond（a \wedge \Diamond b）$，会将线性时序逻辑公式转换为 $\Diamond b$。这体现了 LPOPL 的一个理想特性：共享相同推进后 LTL 形式的任务规范可以共享子任务策略。

我们借鉴 LPOPL 的子任务提取方法和 Q值函数更新策略，以加速多个任务的离策略学习。然而，我们的方法有所不同，因为我们并非从一系列课程任务中学习。相反，我们使用一个以单个给定任务为条件的行为策略，并对该任务的经验进行反事实推理，以同时解决生成的辅助任务。

3 相关工作

经验回放方法研究单任务课程问题，即用于提取迁移样本的任务集合的基数为一（1），且仅包含目标任务 [14]。这些工作的重点是探索如何组织并利用从单个任务中获得的经验进行训练，以实现最优效果。优先经验回放 [5] 改进了经验回放 [4],：后者从回放记忆中均匀采样，而前者则优先选择重要的状态转移，使其被更频繁地采样。事后经验回放（HER） [6] 将探索作为一种隐式课程，并引入了从回放记忆中的经验的其他实现方式进行学习的方法：根据实际达成的目标，而不是智能体试图达成的目标，重新标记经验。HER 的反事实经验重用仅限于单一目标状态，因此无法表示具有丰富表现力、且在到达目标状态的同时

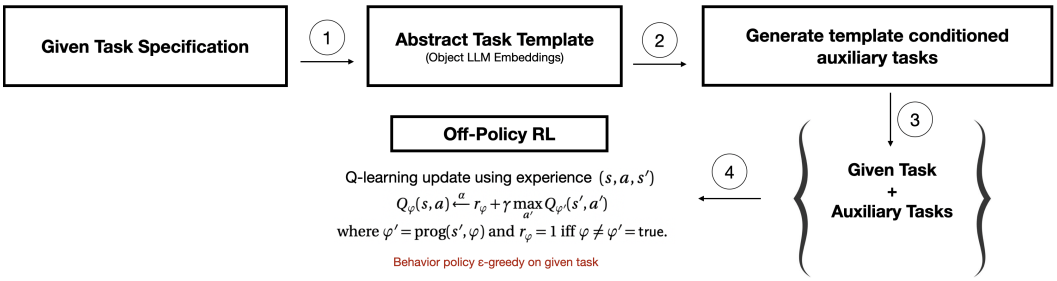


图 2：该图展示了 TaskExplore 框架。给定一个以线性时序逻辑指定的任务，我们构建一个抽象任务模板，将给定公式中的实例对象命题替换为对其描述的大语言模型嵌入，从而捕捉每个对象的各种相关属性。随后，我们使用余弦相似度度量，从环境中为抽象任务模板中的每个命题节点选择对象，以生成辅助任务。我们为给定任务和所有辅助任务初始化策略（Q值函数），并执行强化学习，其中仅针对给定任务以 ϵ-贪心方式选择动作，收集解决给定任务所需的定向经验。在每个学习步骤中，所有 Q值函数都会通过离策略 Q学习更新进行更新。

遇到特定中间状态 [7]的时序扩展行为。此外，HER 的替代目标是从目标任务中采样的中间样本，而非不同的替代任务。

其他研究，例如奖励机的反事实经验（CRM） [7], 奖励机Q学习 [8] 和 LPOPL [9] ，提出了将离策略更新应用于同时学习多个动作价值函数的算法。这些研究解决了 HER 在将反事实经验应用于时间上延展的替代任务时所面临的局限。然而，在这些研究中，能够从反事实推理中受益的替代任务被假定为已知或给定。此外，用于决定如何与环境交互、并由此生成反事实经验的行为策略，取决于由多个任务构成的课程——通常包括能够从这些合成经验中受益的替代任务。

我们的工作旨在提出一种解决方案，使智能体能够自动生成表达能力强、时间延展的辅助任务——而非中间任务——从而最大限度地利用以对象为中心的环境中单任务课程所提供的定向经验。TaskExplore 有别于课程学习文献中数量有限的任务生成研究 [15, 16], ，这些研究中的任务并非由人工设计。我们的方法并非像课程学习中的任务生成那样，旨在生成优质的中间任务以获取经验样本 [14]。相反，我们专注于生成各不相同的辅助任务，以最大限度地利用单任务课程中的定向经验；这更接近终身学习情境，在此类情境中，智能体需要从极少量或受限的数据集中学习泛化能力 [17]。

4 问题定义

为了形式化定义我们的问题，我们提出面向对象的非马尔可夫奖励决策过程（OO-NMRDP），结合了面向对象的马尔可夫决策过程（OOMDP） [18]和非马尔可夫奖励决策过程（NMRDP） [19, 20, 21] 形式体系中的思想。OO-NMRDP 是一个 8 元组 $M = <O,\ C,\ L,\ S,\ A,\ T,\ R_\varphi,\ \gamma>$，其中 O 是一组布尔命题，用于表示环境中存在的对象，并由标记函数 L 检测：$S \to 2^O$ 该函数将状态映射到这些布尔命题，指定哪些命题在何种状态下为真。C 是对象类别集合，S 是状态集合，A 是动作集合，$\gamma \in [0,1]$ 是折扣因子，而T：$S \times A \times S \to [0,1]$ 表示环境的状态转移动态。与常规马尔可夫决策过程不同，奖励函数 R_φ 定义在状态历史之上；当且仅当智能体在给定回合中经历的状态序列满足线性时序逻辑公式 φ时，智能体才会获得 1 的奖励。

$$R_\varphi(\langle s_0, ..., s_n \rangle) = \begin{cases} 1 & \text{if } \delta_{0:n-1} \nvDash \varphi \text{ and } \delta_{0:n} \models \varphi \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

其中 $\delta_{i:j} = \langle L(s_i) ...,\ L(s_j) \rangle$。学习智能体无法访问对象类别集合 C，也无法获取环境 T的状态转移动态。我们将序列 [11] 或软排序约束任务 [22] 表示为命题集合 O 上的线性时序逻辑公式。

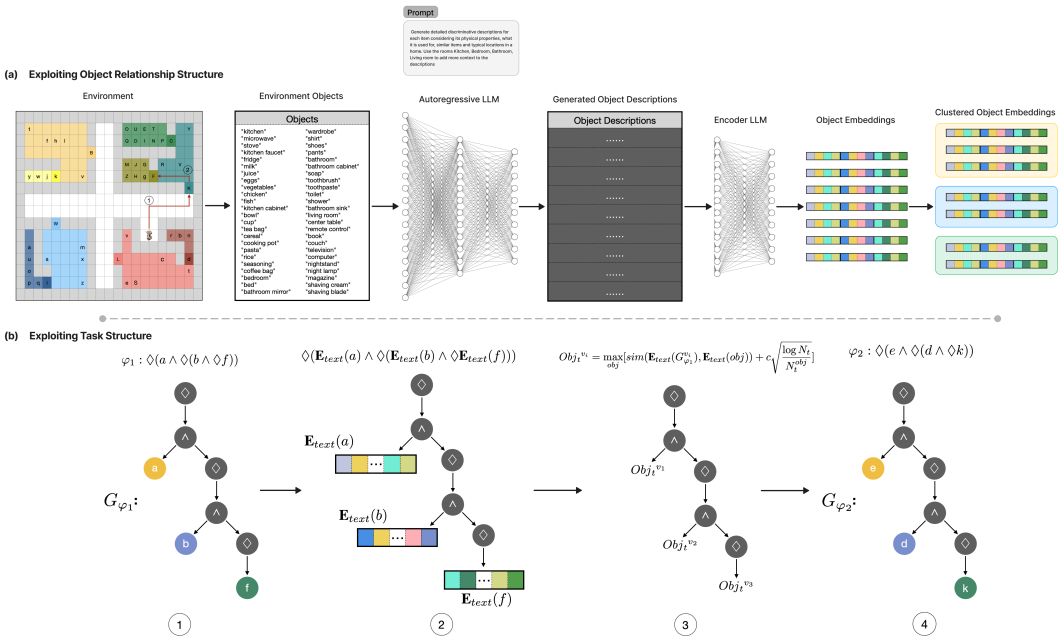


图 3：该图展示了 TaskExplore 如何构建并利用具有上下文感知能力的对象嵌入和抽象任务表示/模板。在**图 a**中，我们使用自回归大语言模型为环境中的对象列表生成详细描述，并使用编码器语言模型将这些生成的描述编码为每个对象对应的一个 768维向量。随后，我们对这些描述嵌入进行聚类，发现能够体现对象之间语义和上下文相似性的对象类别。在**图 b**中，我们通过使用相应对象的嵌入来表示给定线性时序逻辑公式的抽象语法图中的命题节点，从而构建任务模板。借助该任务模板，我们可以根据对象的余弦相似度从环境中选择对象，创建上下文相似的新任务，并在高度相关的对象与相关但尚未见过的对象之间平衡选择。

5 利用上下文结构生成辅助任务

人类能够利用以对象为中心的环境中固有的结构（包括对象及其相互关系），以及任务的组合结构。凭借这种结构，他们能够学习任务，同时思考所积累的经验还可以如何用于其他任务。这种能力使人们在学习任何一项特定事物时，都能获得多种有用技能。我们介绍如何利用任务的抽象时序逻辑表示和环境中对象的上下文感知嵌入，赋予强化学习智能体这种能力。

图 2 展示了我们方法的主要步骤。给定一个用线性时序逻辑 φ_1 指定的序列任务，我们的框架利用 LTL 的组合语法以及对象之间现实世界的上下文关系，通过对象交换构建一组辅助任务 $Aux_{\varphi 1}= \{a_1, ..., a_n\}$，使其具有与 φ_1 相似的底层探索需求。随后，我们为 φ_1 以及生成的辅助任务集 Aux_{φ_1} 中的每个任务初始化一个包含 Q值函数的策略库。我们通过类似 LPOPL [9] 的离策略更新，同时学习 φ_1 的策略和所有辅助任务的策略。智能体始终遵循相对于 φ_1 的 ϵ-贪心策略。我们的学习方法与 LPOPL 的一个关键区别在于：我们不使用目标任务的多任务课程，并且我们的行为策略仅以给定任务为条件，从而限制了在环境中的探索。

在第 5.1 节中，我们详细介绍如何从环境中构建上下文感知对象嵌入，并用这些嵌入确定适合交换的对象。第 5.2 节介绍如何利用这些对象嵌入构建抽象任务模板，并据此生成辅助任务。图 3 展示了这两个过程。最后，在第 5.3 节中，我们说明如何通过反事实推理和离策略学习，同时学习用于生成的辅助任务的策略。



图 4：该图展示了对每个环境物体的 768 维嵌入向量进行 K 均值聚类的结果，并将结果可视化于二维潜在空间中。图 (a) 中每个对象的嵌入由 Sentence-T5 模型对图中所示的物体名称进行编码而生成。相反，图 (b) 中的嵌入由 Sentence-T5 对 text-davinci-003 生成的各物体文本描述进行编码而生成。K 均值算法使用的簇数为四 (4)，这是根据 HomeGrid 中不同探索区域的数量确定的。**请注意，由大语言模型生成的物体描述所生成的嵌入改善了新兴簇边界的分离效果，并理想地增大了潜在空间中相似但上下文不同的对象之间的距离，例如厨房橱柜和浴室橱柜。**

5.1 利用对象关系中的结构

大语言模型在大型文本语料库上训练，编码了有用的常识性知识和上下文相关的人类知识，因此可作为良好的先验，用于构建以对象为中心的环境中对象之间的关系。我们的方法利用这类模型生成具有区分性的对象上下文感知嵌入。自回归语言模型（如 GPT [23]）擅长生成语言，因为它们经过训练，能够最大化给定前序词元时下一个词元的似然。另一方面，编码器-解码器和仅编码器大语言模型（如 T5 [24] 和 BERT [25]）擅长生成紧凑表示，能够有效捕捉句子上下文和语义。我们使用 InstructGPT *text-davinci-003* 模型 [26] 为环境中的对象列表生成详细描述，并使用 *Sentence-T5* 编码器模型 [27] 将生成的描述编码为每个对象对应的 768维向量。随后，我们使用 K 均值算法 [28, 29]和 K 均值++ 初始化[30]对这些描述嵌入进行聚类。

作为消融实验，我们研究 Sentence-T5 仅依据物体名称生成上下文感知嵌入以进行聚类的能力，如图 4a 所示。图 4b 则展示了使用*text-davinci-003* 模型生成的描述所带来的聚类结果改进，揭示了利用大语言模型进行上下文感知数据增强对下游任务的益处。

5.2 利用任务组合中的结构

给定一个实例线性时序逻辑公式，我们利用线性时序逻辑固有的组合性，创建任务的泛化表示或模板。借助这一任务模板，我们可以生成相关的辅助任务。与先前工作 [31] 中使用的表示格式类似，我们将给定公式 φ_1 解析为其抽象语法树，并将其表示为有向图 $G_{\varphi_1} = (V_{\varphi_1}, E_{\varphi_1})$。该图中的边 E_{φ_1} 连接父算子与其子公式，顶点 V_{φ_1} 表示算子或原子命题节点，如图 3b.1 所示。这种表示方式使我们能够高效利用给定任务的结构来构建辅助任务。

我们通过遍历 G_{φ_1} 并将实例命题节点（表示环境中的对象）替换为其描述的嵌入，来创建抽象任务模板，如图 3b.2 所示。借助这一抽象任务模板，我们通过在每个命题节点处替换环境中的相关对象，生成 x 个新的辅助任务。我们选择嵌入位于同一嵌入簇中的对象，并优先选取与当前模板嵌入节点余弦相似度最高的对象。我们提出一种基于上置信界 [32] 的简单、基于价值的对象选择指标，以平衡对高余弦相似度对象和相关但尚未见过对象的选择。

公式 4 规定了给定模板中每个嵌入节点的对象选择方式。

$$Obj_t{}^{v_i} = \max_{obj}[sim(\mathbf{E}_{text}(G_{\varphi_1}^{v_i}), \mathbf{E}_{text}(obj)) + c\sqrt{\frac{\log N_t}{N_t^{obj}}}] \tag{4}$$

其中，$Obj_t{}^{v_i}$ 是在试验 t 中为模板节点 i 选定的对象命题；$sim(\mathbf{E}_{text}(G_{\varphi_1}^{v_i}),\mathbf{E}_{text}(obj))$ 是对象 obj 的嵌入与模板 G_{φ_1}中节点 v_i 的嵌入之间的余弦相似度；**c** 是一个可调参数，用于平衡选择余弦相似度较高的对象与选择相关但尚未见过的对象；N_t 是对象选择试验的总次数；N_t^{obj} 是对象 obj 被选中的试验次数。

5.3 基于反事实经验的离策略更新

对于离策略更新，当使用 ε-贪心行为策略针对给定的线性时序逻辑公式运行一个回合时，每个辅助任务的 Q值函数也会随之更新，仿佛对应的公式就是当前目标。假设在状态*s* 中采取动作*a* 后到达新状态 s'，若要更新特定的 Q_φ，则需计算在该次状态转移中，假如智能体的目标是 φ，本应获得的奖励。为此，将 φ 在新状态 s'上推进；如果得到的公式 φ' 为 *true*，则奖励为 1，否则为 0。随后，Q_φ 按照以下规则更新：

$$Q_\varphi(s,a) \leftarrow Q_\varphi(s,a) + \alpha(r + \gamma \max_{a'} Q_{\varphi'}(s', a') - Q_\varphi(s,a)) \tag{5}$$

如 LPOPL 所示，[9] 学习在全局上是最优的，因为 $Q_\varphi(s,\ a)$ 会根据其已推进子任务 $Q_{\varphi'}(s',\ a')$ 中每个动作 a' 的最大值进行更新。

6 实验

我们在实验中使用的任务规范是一项食物准备任务：智能体必须前往**厨房橱柜**，拿取**烹饪锅**，拿取**调味料**，然后前往**冰箱**，拿取**鸡肉**，最后前往**炉灶**。在 HomeGrid 中，这项任务对应于按正确顺序访问相应的格子。下面的线性时序逻辑公式使用表示各个相关对象的原子命题来表示这项任务：

$$\Diamond(C \wedge \Diamond(P \wedge \Diamond(I \wedge \Diamond(F \wedge \Diamond(H \wedge \Diamond Y))))) \tag{6}$$

我们的环境 HomeGrid 的说明见图1。我们发现，选择最少对象簇数量的一种良好启发式方法，是根据环境中的不同探索区域数量来确定。对于 HomeGrid，这个数量是四（4），因为其中有四个不同且有用的探索区域，即“厨房”“浴室”“客厅”和“卧室”。我们在以下三种条件下评估我们的方法：

1. **我们的方法**：给定我们的食物准备任务 φ_1，我们按照所述方法生成**20 个辅助任务**。我们使用在 φ_1 上采用 ε- 贪心策略的行为策略，同时学习这些任务和 φ_1，将探索引向与 φ_1 更相关的经验。2. **基线 1**：为了证明 TaskExplore 生成的任务能够独特地利用单任务课程中的定向经验，我们重复上述方法，但将行为策略替换为随机策略，以进行更广泛的探索。

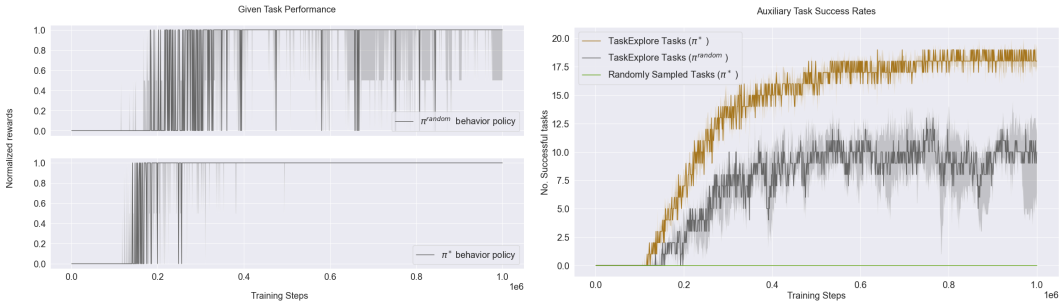


图 5：图 (a) 展示了智能体在学习解决给定任务的同时，使用随机行为策略（π^{random}）和 ε-贪心行为策略（π^*）解决 TaskExplore 生成的辅助任务时，在给定任务上获得的归一化折扣奖励。图 (b) 展示了随着学习的推进，辅助任务上的任务成功率。使用 ε-贪心（π^*）行为策略解决给定任务的同时学习 TaskExplore 生成的任务，其表现显著优于所有其他基线方法。所有结果均基于 7 次不同随机种子的运行进行归一化。

3. **基线 2**：为了说明，与一般的可能任务集合相比，TaskExplore 生成的任务能够更有效地受益于定向探索经验，我们从环境中的命题集合里随机抽取序列任务，生成 **20 个辅助任务**，这与以往研究中的常见做法一致 [31, 22]。抽取的任务所含命题数量与给定任务相同，因此难度等级也相同。我们使用在 φ_1 上采用 ε-贪心策略的行为策略，同时学习这些任务和 φ_1。

7 结果与讨论

图 5 展示了我们实验的结果，凸显了 TaskExplore 构建的辅助任务集的若干有趣特性。首先，在图 5a 中可以看到，我们的方法在给定任务上采用 ϵ-贪心行为策略，同时学习解决辅助任务，并不会对给定任务的性能产生不利影响。然而，使用随机行为策略时，给定任务上的性能会下降，这凸显了定向探索的重要性。
直观来看，使用随机行为策略进行学习会收集到更多样化的经验，相较于定向探索期间收集到的经验，这应该更有利于多个辅助任务。然而，我们在图5b中的结果表明，TaskExplore 开发的辅助任务能够最大限度地利用这些受限经验，实际表现甚至优于使用随机行为策略的情况。这是因为生成的任务在情境上与给定任务相似，并且具有相似的底层探索需求。

图6进一步揭示了这一现象。图中展示了一个样例探索热图，来自我们使用随机行为策略和 ε-贪心行为策略学习食物准备任务的一次完整实验。在第一个回合中，两种行为策略都进行了广泛探索；然而，随着学习的推进，ε-贪心策略带来了更具定向性、范围更受限的经验，并将探索重点放在厨房探索区域。另一方面，随机行为策略仍然进行广泛探索，没有特别关注其本应聚焦的厨房探索区域，因此无法在相关辅助任务上取得进展。

最后，图5b中的结果表明，从环境中可能任务的一般分布空间随机采样任务来构建课程，无法像 TaskExplore 开发的辅助任务那样有效地利用定向探索经验。

8 结论

本文提出了一种方法，使智能体能够自动生成表达力丰富的时间延展辅助任务，从而最大限度地利用以对象为中心的环境中单任务课程的定向经验。这种辅助任务生成方法尤其

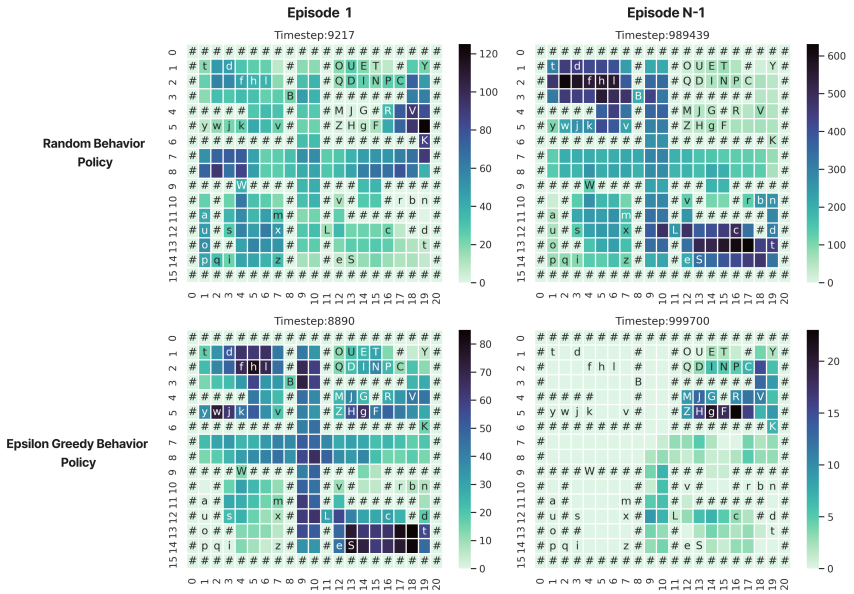


图6：该图展示了在使用 TaskExplore 生成的辅助任务进行反事实更新、学习给定食物准备任务时，随机行为策略和 ε-贪心行为策略的探索热图示例。智能体在每个回合开始时都位于单元格 (x 轴=14，y 轴=13)，颜色越深的单元格表示智能体访问该单元格的次数越多。这是一张消融图，有助于直观展示定向探索在复杂环境中的益处，以及这种定向经验如何帮助上下文相似、共享相同底层探索需求的辅助任务；相比之下，探索策略可能产生更多样化的经验。

适用于终身学习场景，因为智能体可以从受限数据集中生成并解决新任务。本着重复利用计算的精神，可以保存这些策略组成的策略库，并在未来任务中重复使用，以加速学习。未来的工作可以采用能够检测真实环境中开放词汇对象的现代视觉语言模型（VLMs），以减少 TaskExplore 对预定义对象命题集合的依赖（任务可由这些命题来表达），以及对将状态映射为命题真值的标注函数的依赖，

参考文献

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, 等，“通过深度强化学习实现人类水平的控制”，《自然》，第 518 卷，第 7540 期，第 529–533 页，2015年。[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford 和 O.Klimov，“近端策略优化算法”，*arXiv* 预印本 *arXiv:1707.06347*，2017年。[3] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, 等，“通过大规模深度强化学习玩《Dota 2》”，*arXiv* 预印本 *arXiv:1912.06680*，2019年。[4] L.-J. 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，第 8 卷，第 293–321 页，1992年。[5] T. Schaul, J. Quan, I. Antonoglou 和 D. Silver，“优先经验回放”，*arXiv* 预印本 *arXiv:1511.05952*，2015年。[6] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. To- bin, O. Pieter Abbeel 和 W. Zaremba，“ 事后经验回放 ”，《神经信息处理系统进展》，第 30 卷，2017 年。[7] R. T. Icarte, T. Q. Klassen, R. Valenzano 和 S. A. McIlraith，“奖励机：利用强化学习中的奖励函数结构”，《人工智能研究杂志》，第 73 卷，第 173–208 页，2022年。

[8] R. T. Icarte, T. Klassen, R. Valenzano 和 S. McIlraith，“在强化学习中使用奖励机进行高级任务规范和分解”，载于国际机器学习会议，第 2107–2116 页，PMLR，2018年。[9] R. Toro Icarte, T. Q. Klassen, R. Valenzano 和 S. A. McIlraith，“使用线性时序逻辑向强化学习智能体教授多个任务”，载于第 17 届国际自主智能体与多智能体系统会议论文集，AAMAS ’18，第 452–461 页，国际自主智能体与多智能体系统基金会，2018年。[10] A. Pnueli，“程序的时序逻辑”，载于第 18 届计算机科学基础年度研讨会论文集，SFCS ’77，第 46–57 页，IEEE计算机学会，1977年。[11] M. L. Littman, U. Topcu, J. Fu, C. Isbell, M. Wen 和 J. MacGlashan，“通过广义线性时序逻辑实现与环境无关的任务规范”，arXiv 预印本 arXiv:1704.04341，2017年。[12] F. Bacchus 和 F. Kabanza，“使用时序逻辑表达规划的搜索控制知识”，人工智能，第 116 卷，第 1-2 期，第 123–191 页，2000年。[13] C. J. Watkins 和 P. Dayan，“Q学习”，机器学习，第 8 卷，第 279–292 页，1992年。[14] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor 和 P. Stone，“强化学习领域的课程学习：框架与综述”，机器学习研究期刊，第 21 卷，第 1 期，第 7382–7431 页，2020年。[15] S. Narvekar, J. Sinapov, M. Leonetti 和 P. Stone，“课程学习的源任务创建”，载于 2016 年国际自主智能体与多智能体系统会议论文集，第 566–574 页，2016年。[16] F. L. D. Silva 和 A. H. R. Costa，“面向强化学习的面向对象课程生成”，载于第 17 届国际自主智能体与多智能体系统会议论文集，第 1026–1034 页，2018年。[17] S. Thrun，“终身学习算法”，《学习如何学习》，第 8 卷，第 181–209 页，1998年。[18] C. Diuk, A. Cohen 和 M. L. Littman，“用于高效强化学习的面向对象表示”，载于第 25 届国际机器学习会议论文集，第 240–247 页，2008年。[19] R. Brafman, G. De Giacomo 和 F. Patrizi，“有限迹线性时序逻辑/有限迹动态逻辑中的非马尔可夫奖励”，载于 AAAI人工智能会议论文集，第 32 卷，2018年。[20] A. Camacho, O. Chen, S. Sanner 和 S. A. McIlraith，“用线性时序逻辑表达的非马尔可夫奖励：通过奖励塑形引导搜索”，载于第十届组合搜索年度研讨会，2017年。[21] S. Thiébaux, C. Gretton, J. Slaney, D. Price 和 F. Kabanza，“具有非马尔可夫奖励的决策理论规划”，《人工智能研究杂志》，第 25 卷，第 17–74 页，2006年。[22] J. X. Liu, A. Shah, E. Rosen, G. Konidaris 和 S. Tellex，“面向时间扩展任务规范的技能迁移”，arXiv 预印本 arXiv:2206.05096，2022年。[23] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever 等，“通过生成式预训练提升语言理解能力”，2018年。[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li 和 P. J. Liu，“探索统一文本到文本转换器的迁移学习极限”，机器学习研究期刊，第 21 卷，第 1 期，第 5485–5551 页，2020年。[25] J. Devlin, M.-W. Chang, K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向转换器预训练”，arXiv 预印本 arXiv:1810.04805，2018年。[26] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray 等，“通过人类反馈训练语言模型遵循指令”，arXiv 预印本 arXiv:2203.02155，2022年。[27] J. Ni, G. H. Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer 和 Y. Yang，“Sentence-T5：基于预训练文本到文本模型的可扩展句子编码器”，arXiv 预印本 arXiv:2108.08877，2021。[28] S. Lloyd，“脉冲编码调制中的最小二乘量化”，IEEE信息论汇刊，第 28 卷，第 2 期，第 129–137 页，1982年。[29] J. MacQueen，“多变量观测分类与分析的一些方法”，载于第五届伯克利数学、统计与概率研讨会论文集，第 281 页，1965年。

[30] D. Arthur 和 S. Vassilvitskii，“k-means++：谨慎初始化的优势”，技术报告，斯坦福，2006。[31] P. Vaezipoor、A. C. Li、R. A. T. Icarte 和 S. A. Mcilraith，“Ltl2action：为多任务强化学习泛化线性时序逻辑指令”，发表于国际机器学习会议，第 10497–10508 页，PMLR，2021。[32] P. Auer，“使用置信界权衡利用与探索”，机器学习研究杂志，第 3 卷，第 11 期，第 397–422 页，2002。