arXiv:2602.06227v1 [cs.AI] 5 Feb 2026

为事后经验回放而做：强化学习中的一阶时序逻辑奖励规范（扩展版）

Pierriccardo Olivieri米兰理工大学 DEIB，意大利米兰
pierriccardo.olivieri@polimi.it

Fausto Lasca
米兰理工大学 DEIB，意大利米兰 fausto.lasca@mail.polimi.it

Alessandro Gianola
INESC-ID／里斯本高等技术学院，里斯本大学，葡萄牙里斯本
alessandro.gianola@tecnico.ulisboa.pt

Matteo Papini
米兰大学计算机科学系，意大利米兰
matteo.papini@unimi.it

摘要

本文提出了一种新颖的框架，用于在状态空间较大的马尔可夫决策过程（MDP）中，对非马尔可夫性奖励进行逻辑规范。我们的方法采用有限轨迹上的理论模线性时序逻辑（LTLfMT）。与经典时序逻辑相比，LTLfMT 的表达能力更强，其中谓词是一阶理论中任意的一阶公式，而非简单的布尔变量。增强的表达能力使其能够规范非结构化、异构数据领域中的复杂任务，从而推动形成统一且可复用的框架，无需手动编码谓词。然而，与标准 LTLf 规范相比，LTLfMT 更强的表达能力也带来了额外的理论和计算挑战。针对这些挑战，我们从理论角度出发，确定了 LTLf MT 的一个片段，该片段既易于处理，又足以在无限状态空间情境下进行奖励规范。从实践角度来看，我们提出了一种基于奖励机和事后经验回放（HER）的方法，用于转换一阶逻辑规范并应对奖励稀疏性。我们在一个采用非线性算术理论的连续控制场景中评估了该方法，结果表明，该方法能够自然地规范复杂任务。实验结果表明，针对具体任务定制的 HER 实现对于解决目标复杂的任务至关重要。

1 引言

在强化学习（RL）中，智能体与环境交互，并接收标量奖励信号作为反馈，利用该信号随时间改进其决策能力。原则上，根据奖励假设 [Sutton and Barto， 2018],任何任务都可以通过指定适当的奖励函数来建模。然而，奖励工程是一项艰巨的任务，通常具有启发式性质，并且依赖领域知识。此外，许多我们感兴趣的任务本质上是非马尔可夫的，这使得常用的马尔可夫奖励函数对学习的效果较差。通过逻辑规范为奖励引入结构，有助于应对这些挑战，使奖励工程更贴近人类语言，并提高可解释性。在逻辑奖励规范中，智能体在每个时间步接收的奖励取决于基于该时间步的

预印本。

真实命题符号集合是否满足某个公式。这种方法可以使用时序逻辑表达复杂任务。为了更直观地理解这一点，可以考虑一个在仓库中运行的机器人，它必须将一个箱子运送到特定位置。该任务包含两个按顺序执行的子目标：(*A*)在给定位置取起箱子（x_o， y_o），然后（B）将其运送到指定位置（x_g， y_g）。可以通过编码事件 A 与 B之间时序关系的逻辑规范来约束这一行为，例如"先执行 A，然后执行 B"。

目前，这一方向的大多数研究都采用基于有限轨迹线性时序逻辑（LTLf）的逻辑规范 [De Giacomo and Vardi, 2013, Brafman et al., 2018]。LTLf 源自其对应的无限轨迹版本（LTL [Pnueli, 1977a]），后者是形式化验证领域中最常用于描述系统属性的语言。过去十年间，LTLf 在人工智能领域也日益流行，因为对于规划任务等场景中的有限执行进行推理更加直观 [Giacomo and Rubin, 2018]。LTLf 对许多离散领域而言具有足够的表达能力，并且拥有成熟的工具链，可将此类规范转换为自动机或"奖励机" [Icarte et al., 2018]。然而，LTLf 的表达能力仅限于布尔谓词，这限制了奖励规范的复杂度。要判断这些谓词的真假，通常需要手动定义编码函数 。对于离散领域或简单的奖励函数而言，这一限制通常并不构成问题。然而，在许多现实场景中，例如业务流程管理 [Gianola et al., 2024] 或机器人学 [Li et al., 2017], ，更丰富的奖励规范语言能够以更少的人工干预为学习器提供同等乃至更多的信息，并显著加快为同一环境制定新任务规范的速度。当存在异构数据类型时，富有表达力的规范语言的重要性就更加明显，下面的扩展示例将对此进行说明。

仓库机器人示例。 机器人必须到达物体的位置，识别具有特定标识符的物体，并满足重量约束。由此得到的谓词可能是"$A = (x - x_o)^2 + (y - y_o)^2 < r^2 \wedge i = \text{H123} \wedge weight < 10$"。机器人的状态是异构的，由以下变量描述：$x$， y， $w \in \mathbb{R}^+$，分别表示位置和物体重量（$w = 0$ 如果没有携带物体）以及$i \in \{\text{H123}, \text{S456}, \ldots\}$ 是给定的标识符集合。

通常，评估此类谓词需要一些手工设计的编码器：一个用于判断智能体的位置（x， y）与物体位置（x_o， y_o）之间的欧几里得距离是否低于阈值 r，另一个用于检查标识符并在库存数据库中查找物体的重量。即使在这个相对简单的示例中，实现规范也需要（i）手动编码每个条件，而且很可能（ii）生成的规范无法在不同任务之间复用。若没有用户定义的此类编码，就无法在 LTLf 中处理连续或异构状态，因为 LTLf 只在命题层面运作。

为克服这些局限性，我们建议使用有限轨迹上的理论模线性时序逻辑（LTLfMT） [Geatti 等人, 2022年、2023年，Faella 和 Parlato，2024年，Rodríguez 和 Sánchez，2024年、2023年，Rodríguez 等人， 2025], 它是 LTLf 的一种扩展，其中原子命题被替换为在（组合的）任意一阶理论上解释的一阶公式。LTLfMT 能够利用多样的一阶理论对奖励规范进行推理，例如实数算术和/或整数算术、未解释函数和关系（例如用于表示关系数据库 [Ghilardi 等人， 2023],[Gianola， 2023]）、复杂数据类型（例如列表、数组），乃至自定义的领域专用谓词，从而提供自然编码更复杂规范所需的表达能力。从实践角度看，我们不再为每个条件手动定义单独的编码器，而是提出一个统一框架，该框架仅需要任务的一个公式以及对常量的赋值。该公式可以使用不同的理论来表达，并通过通用的理论模可满足性（SMT）求解器 [Barrett 等人， 2021], 进行求值，而无需将其显式编码为命题谓词。继续以配送机器人为例，前文讨论的谓词 A 可以在 LTLfMT 中使用实数上的（非线性）算术理论来表达和求值，并通过标准 SMT 求解器进行求值，而无需手部制作的、依赖领域的编码。通过将复杂性转移到逻辑规范层，该框架能够定义比 LTLf 更具表达力的行为，同时借助现成的 SMT求解器 [Barbosa 等人，2022年，de Moura 和 Bjørner， 2008], 在不同任务和领域之间保持通用性与可复用性。从理论上讲，LTLfMT 框架可被视为一种用于奖励规范的可组合方法，其中每种理论都充当一个构建模块，添加更多理论（或其组合）会增强该框架的表达能力。如果除布尔谓词之外不包含任何额外理论，我们便恢复标准的 LTLf 语义。为获得更强的表达能力，我们可以纳入并

组合诸如算术和数据库理论等任意理论。然而，添加到 LTLfMT 中的理论选择（某些理论可能缺乏可用的求解器）以及完整逻辑的一些高级特征（如量词和变量前瞻算子）都会影响可判定性和求解器支持，进而影响该框架的效率与实际可用性。因此，重要的是识别出可靠且具有足够表达力的逻辑片段，以及能够处理所关注领域特有数据类型的、得到良好支持的理论。

贡献与论文结构。 在第 2 节讨论相关工作并在第 3 节介绍必要的技术背景后，我们提出两项主要贡献——在**第 4 节**中，我们通过将经典的 LTLf 逻辑扩展为表达能力更强的 LTLfMT，定义了一个奖励规范理论框架。在该框架中，原子命题被任意理论上的一阶公式所取代。随后，我们确定了 LTLfMT 的一个片段，该片段（i）除理论本身引入的问题外，不存在其他可判定性问题；（ii）支持将公式轻松转换为自动机；同时（iii）具有足够的表达能力，能够刻画大多数实际感兴趣的目标——在 **第 5 节** 中，当集成理论为非线性实数算术理论（NRA）时，我们应用所提出的框架，以处理连续控制场景中的奖励规范。在这一领域，我们利用一阶框架中目标内在的“结构化”特性，解决逻辑规范中典型的奖励稀疏性问题。我们以一种并非简单的方式结合奖励机的反事实经验（CRM）和事后经验回放（HER） [Andrychowicz 等人， 2017] 提出了一种解决方案。在具有复杂目标的连续控制任务上进行的实证结果表明，与仅使用 CRM 相比，该方法有所改进（第 6 节）。代码可从以下地址获取：https://bit.ly/4i20C4Z。

2 相关工作

我们围绕本文所应对的两个关键挑战讨论相关工作：处理非布尔数据和处理奖励稀疏性。

配备标记函数的命题语言。鉴于基于 LTL 的奖励具有内在的“离散”特性，大多数工作都聚焦于具有有限状态的马尔可夫决策过程 [Bacchus 等，1996，Thiébaux 等，2002，Gretton 等，2003，2011，Gretton，2014，Brafman 等，2018，Bozkurt 等， 2020]。但也有一些值得注意的例外：Cai 等人 [2021] 考虑连续马尔可夫决策过程状态，但逻辑公式是用与状态空间中“区域”等高级概念相对应的命题字母定义的。这种定义属于逻辑之外的范畴，留由用户在实现中处理。“约束螺栓”框架 [De Giacomo 等，2019，2020a,b]则更明确地采用了这种方法，其中的“特征”将“世界”（不一定能被智能体观测到）的属性映射为命题原子。当这种映射直接应用于马尔可夫决策过程的状态时，通常称为标记函数 [Lacerda 等，2015，Hasanbeig 等，2020，2023，Jiang 等， 2021]。

奖励机（RM） [Icarte 等人， 2018, 2022] 是一种逻辑规范的程序化替代方案，用户可以直接构建有限状态自动机。事实上，我们可以将由 LTLf 公式生成的自动机视为一种称为简单奖励机的 RM，它具有二元奖励函数。与上文讨论的许多逻辑相关工作一样，MDP 状态与命题字母之间的映射必须单独实现，在此语境下也称为标记函数。
借助标记函数，逻辑指定奖励和奖励机均得以应用于具有连续状态（和动作）的 MDP，从而为深度强化学习实验打开了大门 [例如，Hasanbeig 等，2020年，Li 等人， 2024]。然而，在上述所有示例中，标记函数（或其等效形式）都被视为黑箱。这种方法使任务间的泛化更加困难，容易出错，也削弱了逻辑公式的可解释性。一个值得注意的例外是 Li 等人 [2017] 为机器人控制提出的时序逻辑，它允许使用如下形式的一阶项 $f(s) < c$，其中 s 是 MDP 的连续状态， c 是常量， f 是函数。我们也将论证，这类不等式项足以编码连续控制场景中大多数值得关注的目标。不过，Li 等人 [2017] 的一阶项只是命题原子，其标记函数的形式预先设定。我们将说明 LTLfMT 在何种意义上更加灵活。

稀疏奖励的现有解决方法。 根据定义，通过逻辑公式指定的目标往往会产生稀疏 （即罕见的）奖励。这类奖励在强化学习中出了名地难以优化，因此促成了

通用奖励塑形策略的早期发展 [Dorigo 和 Colombetti，1994年，Ng 等人， 1999]——即增加额外反馈，引导智能体朝目标前进。在近期的深度强化学习 文献中，尤其是在机器人学应用方面，事后经验回放（HER） [Andrychowicz 等人， 2017] 已成为基于目标的任务的一种强大替代方案，甚至可谓标准方法。然而，HER 的工作原理是利用神经网络的状态泛化能力，并将其应用于定义为状态或状态变量函数的目标。基于命题逻辑或有限状态自动机的规范，或是封装在标记函数等黑箱组件中的规范，通常不具备目标与状态之间这种共同结构。奖励机文献明确探讨了稀疏性问题，Icarte 等人 [已经提出了若干特设方法 2022], ，即奖励机反事实经验（CRM）、层次化方法（HRM）和自动奖励塑形（ARS）。他们还建议探索奖励机与 HER 之间的协同作用，但据我们所知，这一想法之后并未得到进一步探讨。以时间抽象形式呈现的层次化方法也见于线性时序逻辑文献 [Araki 等人， 2021]。值得注意的是，Jothimurugan 等人 [2019] 共同提出了一种（命题）奖励规范语言和一种奖励塑形方法。

3 预备知识

本节回顾强化学习以及基于时序逻辑的奖励规范的主要概念。

强化学习。强化学习问题 [Sutton and Barto，2018] 通常被定义为求解一个马尔可夫决策过程（MDP） [Puterman，1990]。一个 MDP 由元组 $\langle S, A, R, P, \gamma \rangle$ 构成，其中 S 是状态集合， A 是动作集合， R： $S \times A \to \mathbb{R}$ 是奖励函数。环境的动态由转移函数 P： $S \times A \times S \to [0, 1]$ 描述，它表示在状态 $s \in S$ 下采取动作 $a \in A$ 后转移到状态 s' 的概率。折扣因子 $\gamma \in [0, 1)$ 用于权衡未来奖励与即时奖励的重要性。智能体根据策略 π： $S \to \Delta(A)$ 行动，该策略根据观测到的状态 $s \in S$ 为动作分配一个概率分布。最优策略是能够获得最大回报的策略，即从初始状态出发并遵循该策略所收集奖励的期望折扣和。基于时序逻辑或奖励机的任务规范会自然地导出非马尔可夫奖励函数。这种表达能力更强的框架通常称为非马尔可夫奖励决策过程（NMRDP）。形式上，一个 NMRDP 是一个元组 $\langle S, A, R, P, \gamma \rangle$ ，与 MDP 等价，但其奖励函数 function $R : (S \times A)^* \to \mathbb{R}$ 是非马尔可夫的，并且取决于完整的状态-动作对历史 [Bacchus et al.,1996]。经典强化学习方法无法直接应用于 NMRDP。

有限轨迹上的线性时序逻辑。LTLf逻辑 [De Giacomo和Vardi， 2013] 是定义在有限轨迹上的LTLf [Pnueli，1977b] 的一种变体。以LTLf给出的逻辑公式，依据字母表 $\mathcal{P}$ 中定义的命题布尔符号进行求值。有限轨迹 $\sigma = \langle \sigma_0, \sigma_1, \ldots \sigma_n \rangle$ 是一个状态序列（不要与马尔可夫决策过程的状态相混淆）。每个状态 $\sigma_i \subseteq \mathcal{P}$ 是在第 i 个时间步为真的命题符号集合，其中有限整数 $n \in \mathbb{N}$ 指定最后一个时间步。在LTLf中， 公式 φ 由文法 $\varphi ::= p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid X\varphi \mid \varphi_1 \mathcal{U} \varphi_2$ 定义，其中 $p \in \mathcal{P}$， X 是下一步算子， $\mathcal{U}$ 是直到算子。其他算子可通过标准缩写来定义，例如 $\varphi_1 \vee \varphi_2 ::= \neg(\neg\varphi_1 \wedge \neg\varphi_2)$ 、弱下一时刻算子 $X\varphi ::= \neg X \neg \varphi$ 以及最终算子 $F\varphi ::= \top \mathcal{U} \varphi$。公式 φ 在轨迹 σ 的时间步 t 上求值，形式化记为 $(\sigma, t) \models \varphi$。我们采用LTLf的标准语义，并通过对公式 [De Giacomo和Vardi， 2013]进行归纳来定义。在命题逻辑语义之上，其关键方面是： $(\sigma, t) \models X\varphi$ 当且仅当 $(\sigma, t+1) \models \varphi$，并且 $(\sigma, t) \models \varphi_1 \mathcal{U} \varphi_2$ 当且仅当 $(\sigma, j) \models \varphi_2$ 对某个 $j \geq t$ 成立，且 $(\sigma, k) \models \varphi_1$ 对所有 $t \leq k < j$ 成立。在基于时序逻辑的奖励规范中，我们将 $\mathcal{P}$ 中的命题符号解释为世界特征，即智能体在每个时间步 t 接收到的特征。在每个时间步 t，智能体获得的（二元）奖励，由在当前逻辑状态 σ_t 上对 φ 求值来决定。通常， $\mathcal{P}$ 中的命题符号或流畅项表示关于马尔可夫决策过程状态的真值。例如，De Giacomo等人 [2019] 将这种编码称为特征函数 f_i： $W \to D_i$ ，它把世界状态 w 从世界状态集合 W 映射到有限域 D ，例如{真、假}。轨迹的状态 $\sigma_t = \langle f_1(w_t) \quad f_2(w_t) \ldots \rangle$ 通过对特征 f_i 的提取得到，这些特征取自当前世界状态 w_t 、时刻为 t。该奖励信号明显的非马尔可夫性质，使该决策问题成为非马尔可夫奖励决策过程。一种被广泛采用的解决方案 [例如，De Giacomo等人，2020b]是将 φ 转换为有限自动机 $\mathcal{A}_\varphi = \langle Q, 2^{\mathcal{P}}, \delta, q_0, F \rangle$，其中 Q 是自动机状态的集合，

$2^{\mathcal{P}}$ 是输入字母表， δ 是转移函数， $q_0 \in Q$ 是初始状态， $F \subseteq Q$ 是接受（终止）状态的集合。自动机接受轨迹 $\sigma = \langle \sigma_0, \sigma_1, \ldots, \sigma_n \rangle$，如果它从 q_0出发并依次消耗来自 σ的输入，从而到达接受状态 $q \in F$。通过将非马尔可夫奖励决策过程 $\mathcal{M}$ 与自动机 $\mathcal{A}_\varphi$相结合，我们构造乘积马尔可夫决策过程 $\mathcal{M}'$: $= \langle S', A', R', P', \gamma \rangle$，其中状态空间 $S' = S \times Q$ 同时跟踪环境状态和自动机状态。乘积马尔可夫决策过程中的转移与奖励具有马尔可夫性，因为自动机状态捕获了关于历史的必要信息。因此， $\mathcal{M}'$ 等价于原始的非马尔可夫奖励决策过程，但可以使用标准强化学习技术求解。

4 通过 LTLfMT 定义奖励规范

LTLfMT [Geatti 等人 2022] 通过允许 $\mathcal{P}$ 中的谓词为任意理论上的一阶公式，扩展了时序逻辑的表达能力。谓词 $p \in \mathcal{P}$ 不再是布尔符号，而是可以表示一个根据所选背景理论进行解释的一阶公式。本节将正式定义 LTLfMT 的语法，说明如何将不同理论集成到这一框架中。随后，我们提出 LTLfMT 的一个可命题化片段，在保留足够表达能力的同时，可用于强化学习中的奖励规范。最后，我们说明如何将该片段中的公式转换为确定性有限自动机（DFA），从而用于定义强化学习的乘积马尔可夫决策过程。

4.1 有限轨迹上的模理论线性时序逻辑

为了将谓词表示为一阶公式，在 LTLfMT 中，字母表 $\mathcal{P}$ 被多排序一阶签名 $\Sigma = \mathcal{S} \cup \mathcal{P} \cup \mathcal{C} \cup \mathcal{F} \cup \mathcal{V} \cup \mathcal{W}$ 取代，其中 $\mathcal{S}$ 是类型符号（即数据类型）的集合， $\mathcal{C}$ 是常量集合， $\mathcal{P}$ 是谓词集合，[1] $\mathcal{F}$ 是函数集合， $\mathcal{V}$ 和 $\mathcal{W}$ 分别是变量集合和量化变量集合。 Σ-项 t 是构建 LTLfMT 原子公式的构建模块，其生成规则如下:

$$t := v \mid w \mid c \mid f(t_1, \ldots, t_k) \mid \bigcirc v \mid \tilde{\bigcirc} v \tag{1}$$

其中， $c \in \mathcal{C}$、 $v \in \mathcal{V}$、 $w \in \mathcal{W}$ 和 $f \in \mathcal{F}$ 是元数为 k的函数符号。算子 $\bigcirc$ 和 $\tilde{\bigcirc}$分别表示前瞻算子和弱前瞻算子，类似于 LTLf 中的 next X 和 weak˜next X，但作用于 Σ-变量；它们用于关联变量在某一时刻的取值与其在下一时刻的取值。LTLfMT 的完整文法定义如下:

$$\begin{aligned} &\alpha := p(t_1, \ldots, t_k), \\ &\lambda := \alpha \mid \neg\alpha \mid \lambda_1 \vee \lambda_2 \mid \lambda_1 \wedge \lambda_2 \mid \exists v \lambda \mid \forall v \lambda, \\ &\varphi := \top \mid \lambda \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 \wedge \varphi_2 \mid X\varphi \mid \tilde{X}\varphi \mid \varphi_1 \mathcal{U} \varphi_2. \end{aligned} \tag{2}$$

与 LTLf 相比，该文法引入了两个额外层次 α 和 λ。直观而言， α 定义项上的原子谓词，而 λ 允许对这些原子进行一阶组合。顶层公式 φ 是这些一阶表达式上的标准时序公式，与 LTLf 类似。理论通过对 Σ-项的解释融入这一框架。理论 $\mathcal{T}$可定义为一组 Σ-结构，即解释 Σ 中所有符号（因此也包括 Σ-项）的集合，这些结构称为理论模型。直观而言， Σ-签名定义了可用符号（例如，1、2、 $+$、 $>$，.....）， Σ-结构则提供具体解释（例如， $+$(1, 2) 被解释为整数加法）。理论 $\mathcal{T}$为 Σ中的每个符号提供解释。例如，线性整数算术（LIA）理论在 $\mathcal{S}$中引入整数数据类型的一个排序，在集合 $\mathcal{F}$中引入诸如 $+$ 及其逆运算 $-$ 之类的算术运算，通过 $\mathcal{C}$、 $\mathcal{V}$和$\mathcal{W}$引入取值于 $\mathbb{Z}$ 的常量和变量，并在 $\mathcal{P}$中引入比较谓词（例如， $=$、 $<$）。可以组合多个理论，包括其 Σ-签名，其解释由组合的 Σ-结构给出。形式上，一个$\mathcal{T}$-状态 $s = (M, \mu)$包含两个元素: $M \in \mathcal{T}$（即一个 $\mathcal{T}$-模型）是一阶 Σ-结构，它解释 Σ 中的所有符号；而 μ： $\mathcal{V} \to \text{dom}(M)$ 是一个将非量化变量映射到其参照域的函数。量化变量由单独的环境函数ξ： $\mathcal{W} \to \text{dom}(M)$解释。在 LTLfMT 中，轨迹（或词）是由一列有限的 $\mathcal{T}$-状态构成的，跨越多个时间步： $\sigma = \langle (M, \mu_0)$ ....., $(M, \mu_n) \rangle$，其中 $M \in \mathcal{T}$在整个轨迹中固定不变，而 μ_i可能不同，反映变量随时间的演变。项 t 在轨迹 σ 的时间 i 处、给定变量

[1]请注意，LTLfMT 中没有字母表。因此，从现在起， $\mathcal{P}$ 表示一阶签名 Σ中的谓词符号。

环境 ξ 使用函数 μ_i 和 ξ 定义：对于 $\mathcal{V}$ 和 $\mathcal{W}$中的变量，分别在基本情形下定义；对于常量和函数项，则按通常方式归纳定义。对轨迹 σ上时间步 i 处、环境函数为 ξ 的 LTLfMT 公式 φ 的满足关系记作$\sigma, \xi, i \models \varphi$，其定义遵循通常的一阶语义。此满足关系用于定义 LTLfMT 公式 φ 在轨迹 σ 的时间点（时间步）i处的模满足 $\mathcal{T}$，记作 $\sigma, i \models \varphi$；该关系按标准方式归纳定义，以处理时序算子（语义的形式化定义见 [Geatti 等人， 2022] ）。

4.2 强化学习中的奖励规范片段

LTLfMT 框架允许非常一般的表达式，但总体而言，它是不可判定的 [Geatti 等人， 2022]。我们可以定义一个半可判定的 LTLfMT 片段。形式上，对于底层一阶理论可判定的情况，Geatti 等人提出的半判定过程证明了 LTLfMT 是半可判定的（即，对于可满足的实例，总能得到肯定答案；而对于不可满足的实例，推理过程可能无法终止），该过程也已在 BLACK求解器中实现。[2022], [2] 直观地说，要确定 LTLfMT 的一个可判定片段，底层理论 $\mathcal{T}$必须是可判定的，但通常这还不够：在 LTLf 的一阶扩展中，时序算子与量词的相互作用仍可能破坏可判定性；对于 LTLfMT 而言，这种情况是由前瞻算子的存在导致的。我们通过从 Σ 层移除（弱）前瞻算子 $\bigcirc$（$\widetilde{\bigcirc}$），提出了一个适用于强化学习中奖励规范的易处理 LTLfMT 片段。虽然时序逻辑层中的（弱）下一步算子 $X^{\tilde{}}(X)$ 对于指定智能体可能需要完成的“子任务”之间的时序约束至关重要，但在 Σ 项上添加（弱）前瞻算子 $\bigcirc$（$\widetilde{\bigcirc}$）并不自然，也无益于奖励规范，因为它作用于变量。Σ层可视为数据收集接口，用于捕获非结构化且异质的环境信息。我们认为，对数据收集定义时序约束对奖励规范并无用处，Σ层应仅用于捕获智能体可获取的即时信息。请注意，我们的 LTLfMT 片段仍支持任意一阶理论。

$$
\begin{aligned}
t &:= v \mid w \mid c \mid f(t_1, \ldots, t_k) \\
\alpha &:= p(t_1, \ldots, t_k) \\
\lambda &:= \alpha \mid \neg\alpha \mid \lambda_1 \vee \lambda_2 \mid \lambda_1 \wedge \lambda_2 \mid \exists v\lambda \mid \forall v\lambda \\
\varphi &:= \top \mid \lambda \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 \wedge \varphi_2 \mid X\varphi \mid \tilde{X}\varphi \mid \varphi_1 \mathcal{U} \varphi_2
\end{aligned} \tag{3}
$$

在此片段中，不可判定性的来源变成了理论本身（以及量词的使用）。假设理论本身是可判定的，我们便可以利用任何实现该理论判定过程的工具来处理一阶维度：当该理论的无量词片段是可判定的，并且量词不会影响可判定性时，情况便是如此，例如存在量词消去过程（如实数算术的情形 [塔斯基， 1951]）。在此类场景中，我们可以使用先进的量词消去模块来消除量词 [Collins，1974], ，然后使用诸如 SMT 求解器之类的通用工具 [Barrett 等人， 2021] 来评估所得到的无量词公式。根据所采用的理论，如果量词的使用会导致不可判定性，我们还可以通过移除量词进一步限制我们的片段。这样，我们便得到一个类似于 Gianola等人 [2024]提出的“data-LTLf”的 LTLfMT 片段。

4.3 转换为确定性有限自动机和乘积马尔可夫决策过程

Geatti 等人 [2022] 给出了一个非构造性证明：任何可由无前瞻LTLfMT定义的语言，其布尔抽象也可以由 LTLf 定义，反之亦然。我们采取更务实的视角，并按以下方式对我们的片段进行命题化：(i) 将无前瞻LTLfMT公式句法变换为 LTLf 公式；(ii) 将 LTLfMT迹变换为 LTLf迹。这足以定义一个自动机和一个乘积马尔可夫决策过程，只需使用针对所选理论的理论求解器（例如，当该理论无量词时使用 SMT 求解器），便可高效地进行模拟。请参阅式 (3) 中我们这一片段的文法。给定我们语言中的一个公式φ，我们首先将每个 α（其仅包含来自 $\mathcal{V}$ 的自由变量）替换为一个不同的命题字母。然后，将每个剩余的 λ 替换为一个不同的命题字母。结果得到一个 LTLf 公式 φ'。现在令 $\mathcal{P}$ 表示所有已使用命题字母的集合——也就是我们的

[2]https://www.black-sat.org/en/stable/

新字母表。通过在每个时间步为 $\mathcal{P}$ 中的每个命题字母赋予一个真值，将一条 LTLfMT迹 $\sigma = (s_0, \ldots, s_t)$ 映射为一条 LTLf迹 $\sigma' = (\sigma'_1, \ldots, \sigma'_n)$。例如，可以借助定理证明器 [Kovács 和 Voronkov，2013]，或针对所选理论的判定过程（若该理论无量词，则可使用 SMT 求解器，详见第 4.2 节）来完成这一操作。关键在于，在给定时间步交由求解器处理的问题纯属一阶问题，不含任何时序算子。例如，假设理论是实数算术，如果将 ϕ 中的"$\forall y \exists z(y * z > 0 \implies z < y)$"替换为 ϕ'中的字母 P，那么求解器会判定 $P \in \sigma'_t$ 当且仅当 $s_t \models \forall y \exists z(y * z > 0 \implies z < y)$，并首先进行高效的量词消去。在运行时，会在 $LTLf$迹 σ'上评估命题抽象 φ'。SMT 求解器充当极其通用的标记函数（参见第 2 节）。为了便于强化学习，我们可以将 φ'转换为确定性有限自动机（DFA）$\mathcal{A}_{\varphi'} = \langle Q, 2^{\mathcal{P}}, \delta, q_0, F \rangle$，并用它定义一个乘积马尔可夫决策过程 $M' := \langle S', A', R', P', \gamma \rangle$，其扩展状态空间为 $S' = S \times Q$，与 LTLf 中的做法完全相同（参见第 3 节）。我们可以使用 Zhu 等人 [2019][3] 定义的转换技术自动完成这一过程。在训练时，DFA 中的状态转移由 SMT 求解器确定。这是一种特殊类型的奖励机（具体而言，是具有布尔值奖励函数的简单奖励机，参见 [Icarte 等人，2018]），其中 SMT 求解器再次充当标记函数。

5 基于 LTLfMT 奖励的强化学习

在本节中，我们针对使用所提出的规范语言定义的任务给出一种强化学习解决方案，重点关注连续控制领域。

我们首先回顾奖励规范流程，并强调与其他方法相比，它能显著减少人工干预。该规范的最终目标是构建一个乘积马尔可夫决策过程，以便应用标准强化学习算法（以及受奖励机启发的更复杂技术）。首先，根据所关注的领域（例如连续控制），我们选择一种理论（例如非线性实数算术）及其关联的 SMT 求解器（并遵循第 4.2 节中的全部注意事项）。在所关注的领域内可以指定若干不同的任务。对于定义任务的用户，即奖励工程师，我们唯一需要的输入是以第 4.2 节所定义的无前瞻 LTLfMT 片段书写的公式 φ。与现有的奖励规范框架（参见第 2 节）相比，用户无需手动实现标记函数或任何类似内容。至多，为了启用多目标强化学习 （例如使用事后经验回放，后文将详述），用户需要提供公式中出现的一阶常量到数值的字典赋值。这样一来，任务便由（一部分）常量"参数化"。接下来，我们将聚焦于连续控制，既为我们的奖励规范流程提供一个直观、贴近实际的示例，也研究事后经验回放在处理由逻辑规范生成的稀疏奖励方面的有效性。

连续控制案例研究 连续控制中的常见任务包括控制一个变量达到某个目标值或范围，或控制一组变量构成的向量达到某个目标点。例如，机器人在空间中的目标位置、暖通空调系统的理想温度，以及巡航控制的速度。这些"子任务"可以用状态向量之间的距离来描述 L^p（例如，欧几里得距离、盒约束）。这可以通过实数算术实现，并且可以使用时序算子将子任务组合成更复杂的任务。因此，对于连续控制，我们可以选择非线性实数算术（NRA）作为无前瞻LTLfMT的理论基础，并获得第 4.2 节中讨论的所有计算优势。例如，考虑一辆在停车区域内二维空间中行驶的自动驾驶汽车。一个状态 $s \in \mathbb{R}^2$ 表示空间坐标。我们指定一项（参数化）任务，要求智能体先到达空间中的某个点 A，再到达另一个点B，且始终不得经过某个"危险"区域。该任务可以用我们的无前瞻LTLfMT片段表示为

$$\varphi = \neg F \neg\ x \geq x_{min} \wedge x \leq x_{max}\ \wedge F\ \ x - x_a\ ^2 +\ y - y_a\ ^2 < (\ )\ ((\ )\ (\ )\ r_a^2 \wedge F\ \ x - x_b\ ^2 +\ y - y_b\ ^2 < r_b^2$$

((\ )\ (\ )\ )). 在这种情况下，我们使用适用于线性实数算术（LRA）和非线性实数算术（NRA）的SMT 求解器，将 φ 命题化并得到$\varphi' := \neg F \neg \alpha_1 \wedge F(\alpha_2 \wedge F(\alpha_3)$。

我们使用这个示例来强调我们的奖励规范框架的优势。首先，我们要求用户用于指定任务的输入非常简单。它由表示规范公式 φ 的简单字符串，以及可选的将常量赋值为数值以参数化任务的设定组成。[4] 相比之下，现有方法通常需要复杂的输入。例如，

[3]在我们的实现中，我们使用 LTLf2DFA 库 (https://github.com/whitemech/LTLf2DFA) [4]这可以简单到只需一个 JSON 配置文件，其中指定常量的名称和值。

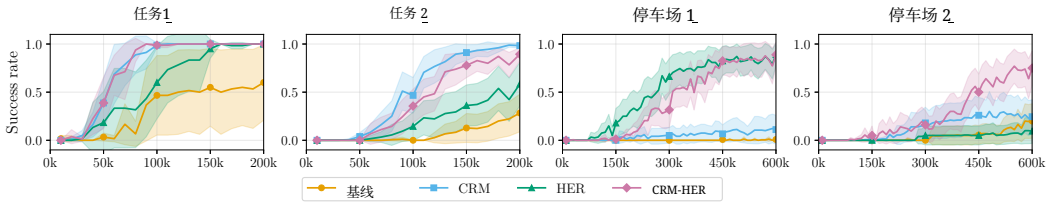

图 1：停车环境中各基线方法在 4 个不同任务上的成功率，误差范围为 20 次独立实验得到的 95% 自助法置信区间。横轴表示 DDPG 的训练步数。每个图上方标出了任务名称，并按任务复杂度从左到右排列。

Icarte 等人 [提出的奖励机 2018]要求对自动机进行显式规范。他们还要求实现标记函数，这与第 2 节讨论的若干基于逻辑的方法一样。此外，在我们的框架中，为同一领域定义新任务可以非常快速。在我们概述的连续控制领域中，非线性实数算术允许通过仅仅编写新的公式来定义许多新任务。对领域进行更实质性的修改只需改变理论（例如，从非线性实数算术改为线性算术）；具体而言，这几乎可以毫不费力，取决于是否有高效的求解器可用。最后，在连续控制的情况下，如下一节所示，我们的框架借助事后经验回放可以轻松处理稀疏奖励。

5.1 缓解乘积马尔可夫决策过程中的稀疏性

基于逻辑的奖励规范方法的一个众所周知的问题是奖励稀疏性。对于连续控制领域，我们提出将奖励机中的反事实推理（CRM）思路 [Icarte 等人， 2022] 与事后经验回放（HER）结合起来。这两种技术都作用于数据收集层面，通过生成额外的人工经验来改进学习。二者都可应用于乘积马尔可夫决策过程生成的轨迹 $\tau = \{< s_0, q_0 >, a_0, r_1, < s_1, q_1 >, \ldots, <s_T, q_T >\}$。反事实推理利用底层的奖励自动机 $\mathcal{A}_\varphi$，通过将确定性有限自动机的状态替换为其他可能状态来生成额外的“伪”经验。由于确定性有限自动机由智能体自身模拟，而马尔可夫决策过程并非如此，因此始终可以进行这种替换。也就是说，给定单个经验 $\{\langle s_t, q_t\rangle, a_t, r_{t+1}, \langle s_{t+1}, q_{t+1}\rangle\}$，我们可以创建 $|Q|$ 人工转移：将q_t 替换为任意其他 $q \in Q$ ，并重新计算奖励。事后经验回放专为具有稀疏二元奖励反馈的基于目标的场景设计。它通过从未成功的轨迹中学习，利用神经网络的状态泛化能力来提高样本效率。事后经验回放要求定义一个“基于目标的问题”，其中包括目标空间 $\mathcal{G}$、谓词奖励函数f_g： $S \to \{0,1\}$ ，用于检查目标 $g \in \mathcal{G}$ 是否在状态 $s \in S$下达成；以及映射m： $S \to \mathcal{G}$ ，使得 $\forall s \in S, f_{m(s)}(s) = 1$ ，用于生成虚拟目标 $g' = m(s)$。基于目标的转移是一个元组 $((s_t, g)\ a_t, r_{t+1}, (s_{t+1}, g))$。首先，我们展示如何将 CRM 和 HER 适配到我们的框架中，然后将二者结合起来，利用其协同效应缓解奖励稀疏性。

将 HER 应用于乘积马尔可夫决策过程。 HER 需要一个目标规范，但在乘积马尔可夫决策过程的状态上定义目标规范并不直观，而且通常由用户提供。为了自动化 HER，我们将乘积马尔可夫决策过程的“真实”目标定义为状态 $\langle s, q\rangle \in S'$，其中 q 是自动机 $\mathcal{A}_{\varphi'}$ 上接受状态 $q_f \in F$的前驱，并且 s 满足 $q_f = \delta(q, L(s, \cdot, \cdot))$，也就是说，该状态会触发向 q_f的转移。然后，基于目标的规范可定义如下：$\mathcal{G} = S', f_g(s) = \mathbb{I}[m(s) = g]$，以及 $m(s) = s$，其中 $s \in S'$ 是乘积马尔可夫决策过程的一个状态。

结合 CRM 与事后经验回放。给定一条轨迹 τ，我们将通过复制 τ 并应用上述反事实推理所获得的 $|Q|$ 人工经验添加到回放缓冲区。然后，对每条（真实或反事实）轨迹应用事后经验回放，生成总计 $2|Q|$ 条人工经验，并将它们全部添加到回放缓冲区。事后经验回放的中间目标根据每条轨迹中最后访问的状态来定义 τ [Andrychowicz 等人， 2017]。在下一节中，我们将比较这两种单独使用和组合使用的方法。

6 实验

在本节中，我们使用深度确定性策略梯度（DDPG） [Lillicrap等人， 2016] 从为停车环境构建的乘积马尔可夫决策过程中学习策略。我们以第 5.1 节中介绍的、在回放缓冲区 层面运行的不同方法作为基线。**基线**：直接使用普通 DDPG 从乘积马尔可夫决策过程中学习任务。**CRM**：将 CRM 应用于用于指定任务的确定性有限自动机，为 DDPG 生成虚拟经验。**事后经验回放**：直接将事后经验回放应用于乘积马尔可夫决策过程的增广状态，而不使用 CRM。**CRM-HER**：我们提出的主要方法，将事后经验回放应用于真实经验以及使用 CRM 生成的反事实经验。我们的目标是评估事后经验回放和 CRM-HER 在不同复杂度任务中的表现与先前方法相比如何，以及 CRM 与事后经验回放所带来的预期协同效应能否转化为实际优势。

环境与任务。 我们在两个不同的连续控制环境中定义的一系列任务（即不同的 φ 公式）上进行了实验：停车 和*Reacher*。停车环境来自*HighwayEnv* 集合 [Leurent， 2018],其中智能体在空停车场中控制一辆汽车。我们修改了该环境，以支持由任意LTLfMT-NRA公式定义的任务规范。在这两个底层环境中，我们使用第 5 节详述的规约流水线定义了多个复杂度递增的任务。这些任务的复杂度主要来自两个因素：用于定义任务的公式的规模与结构（体现在所得确定性有限自动机/乘积马尔可夫决策过程的复杂度中），以及发现目标的难度——由于存在稀疏奖励，这一点尤其具有挑战性。关于任务规范（包括LTLfMT公式）以及*Reacher* 结果的完整细节见附录。

结果分析。 图 1 展示了各方法在多次训练运行上对性能取平均后得到的结果。性能以训练过程中定期进行的评估回合的成功率 来衡量。若智能体在回合时限内达到目标，则该回合被视为成功。我们的结果表明，在大多数任务中，基线方法难以稳定地学习到能够达成目标的策略，落后于其他方法。这一趋势在所有难度级别上都成立，但在更复杂的任务中尤为明显，此时它完全无法发现成功策略（例如，参见任务停车_1、停车_2的结果）。在大多数情况下，CRM更能有效应对确定性有限自动机复杂度的增加。这在停车环境的任务_1 和任务_2 中尤为明显，其性能与表现最佳的方法相当。然而，当目标难以发现时，仅使用CRM收益甚微，其性能落后于基于事后经验回放的方法（尤其参见停车_1）。RM-HER的表现参差不齐，结果因任务性质而异。在主要挑战在于目标发现的任务中，它表现良好，清楚地体现了事后经验回放的益处（例如，参见任务_2、停车_2）。然而在其他任务中，相对于基线的改进不太明显，有时几乎难以察觉。CRM-HER在所有任务中都表现稳定，要么取得最佳结果，要么接近表现最好的方法。结果表明，CRM-HER有效结合了CRM与事后经验回放的优势，从而能够应对不断上升的任务复杂度。性能优势在更具挑战性的任务中尤为明显——尤其是在停车_2任务中，它是唯一能够学习到成功策略的方法。代码与实验结果见：https://bit.ly/4i20C4Z。

7 结论

我们提出了一种基于 LTLfMT 逻辑片段的强化学习奖励规范新框架。与 LTLf 相比，我们的语言表达能力更强，因为它允许谓词是任意理论上的一阶逻辑公式。尽管逻辑增强 MDP 的定义遵循 LTLf 的标准流程，但语言表达能力的提升给负责指定奖励的用户带来了两个重要影响：(i) 用户无需再承担编写 标记函数的负担，这些函数由现成的 SMT 求解器取代；用户只需编写一个逻辑公式； (ii) 在连续控制领域，用户只需付出极少努力，就能使用公式中的同一组变量，为 HER 提供目标规范。我们利用这一第二项优势，将奖励机的 CRM 技术与 HER 相结合，以应对这种基于逻辑的规范所导致的奖励不可避免的稀疏性。在基准连续控制领域中，对采用无前瞻 LTLf-modulo-NRA 规范的任务所进行的实验显示出令人鼓舞的结果。

未来工作应探索将这些技术与分层强化学习方法相结合 [Icarte 等人， 2022]。最后，我们将一阶项纳入 规范语言，为未来进一步结合符号方面与连续空间 强化学习铺平了道路，例如规范安全约束、连续奖励和多任务问题，并最终实现对连续领域中学习过程的形式化验证。

致谢

本研究部分由“OptiGov”项目资助，项目编号为 2024.07385.IACDC（DOI：10.54499/2024.07385.IACDC）。该项目由“复苏与韧性计划”（PRR）通过“RE-C05-i08 - 更数字化科学”投资项目（措施“RE-C05-i08.m04”）全额资助，属于“葡萄牙复苏任务结构”（EMRP）与科学与技术基金会（FCT）签署的融资协议框架，后者为中间受益方。本研究还部分获得葡萄牙国家资金支持，由科学与技术基金会（FCT）通过 UID/50021/2025 和 UID/PRR/50021/2025 项目提供。

参考文献

Marcin Andrychowicz、Dwight Crow、Alex Ray、Jonas Schneider、Rachel Fong、Peter Welinder、Bob McGrew、Josh Tobin、Pieter Abbeel 和 Wojciech Zaremba。《事后经验回放》。载于*NIPS*，第 5048–5058 页，2017年。

Brandon Araki、Xiao Li、Kiran Vodrahalli、Jonathan A. DeCastro、Micah J. Fry 和 Daniela Rus。逻辑选项框架。载于*ICML*，第 139 卷《style id='3'>机器学习研究论文集，第 307–317 页。PMLR，2021年。

Fahiem Bacchus、Craig Boutilier 和 Adam Grove。奖励行为。载于全国人工智能大会，第 1160–1167 页，1996年。

Haniel Barbosa、Clark W. Barrett、Martin Brain、Gereon Kremer、Hanna Lachnitt、Makai Mann、Abdalrhman Mohamed、Mudathir Mohamed、Aina Niemetz、Andres Nötzli、Alex Ozdemir、Mathias Preiner、Andrew Reynolds、Ying Sheng、Cesare Tinelli 和 Yoni Zohar。cvc5：一款功能全面、适用于工业级应用的 SMT 求解器。载于 Dana Fisman 和 Grigore Rosu 编，系统构造与分析工具和算法——第 *28* 届国际会议 *TACAS 2022*，作为欧洲软件理论与实践联合会议 *ETAPS 2022* 的一部分于德国慕尼黑举行，*2022*年*4*月*2*日至*7*日，论文集，第一部分，《计算机科学讲义》第 13243 卷，第 415–442 页。施普林格，2022年。doi：10.1007/978-3-030-99524-9_24。URL：https://doi.org/10.1007/978-3-030-99524-9_24。

Clark W. Barrett、Roberto Sebastiani、Sanjit A. Seshia 和 Cesare Tinelli。理论模可满足性。载于 Armin Biere、Marijn Heule、Hans van Maaren 和 Toby Walsh 编，《可满足性手册 - 第二版》，人工智能与应用前沿第 336 卷，第 1267–1329 页。IOS 出版社，2018 年。doi：10.3233/FAIA201017。URL：https://doi.org/10. 3233/FAIA201017。

Alper Kamil Bozkurt、Yu Wang、Michael M. Zavlanos 和 Miroslav Pajic。利用无模型强化学习，根据线性时序逻辑规范进行控制综合。载于《国际机器人与自动化会议》，第 10349–10355 页。IEEE，2020年。

Ronen I. Brafman、Giuseppe De Giacomo 和 Fabio Patrizi。LTLf/LDLf 非马尔可夫奖励。载于《AAAI》，第 1771–1778 页。AAAI出版社，2018年。

Mingyu Cai、Mohammadhosein Hasanbeig、Shaoping Xiao、Alessandro Abate 和 Zhen Kan。基于时序逻辑的连续运动规划模块化深度强化学习。*IEEERobotics Autom.Lett.*，6(4):7973–7980，2021年。

George E. Collins。通过柱形代数分解对实闭域进行量词消去——初步报告。*SIGSAM Bull.*，8(3):80–90，1974年。doi: 10.1145/1086837.1086852。URL https://doi.org/10.1145/1086837.1086852。

Giuseppe De Giacomo 和 Moshe Y Vardi。有限轨迹上的线性时序逻辑和线性动态逻辑。载于 *IJCAI*，第 13 卷，第 854–860 页，2013年。

Giuseppe De Giacomo、Luca Iocchi、Marco Favorito 和 Fabio Patrizi。约束螺栓的基础：采用 LTLf/LDLf约束规范的强化学习。载于自动规划与调度国际会议论文集，第 29 卷，第 128–136 页，2019 年。

Giuseppe De Giacomo、Marco Favorito、Luca Iocchi、Fabio Patrizi 和 Alessandro Ronca。通过转换器实现时序逻辑监控奖励。载于*KR*，第 860–870 页，2020a。

Giuseppe De Giacomo、Luca Iocchi、Marco Favorito 和 Fabio Patrizi。面向强化学习智能体的约束螺栓。载于*AAAI*，第 13659–13662 页。AAAI出版社，2020b。

Leonardo Mendonça de Moura 和 Nikolaj S. Bjørner。Z3：一种高效的 SMT 求解器。载于 C. R. Ramakrishnan 和 Jakob Rehof 编，*系统构造与分析工具和算法，第 14 届国际会议 TACAS 2008，作为软件理论与实践联合欧洲会议 ETAPS 2008 的一部分举办，匈牙利布达佩斯，2008 年 3 月 29 日至 4 月 6 日*。论文集，第 4963 卷，收录于《计算机科学讲义》，第 337–340 页。施普林格，2008 年。doi: 10.1007/978-3-540-78800-3_24。网址 https://doi.org/10.1007/978-3-540-78800-3_24。

Marco Dorigo 和 Marco Colombetti。《机器人塑造：通过学习开发自主智能体》。《人工智能》，71(2):321–370，1994年。

Marco Faella 和 Gennaro Parlato。模理论LTLf的统一自动机理论方法。载于 Ulle Endriss、Francisco S. Melo、Kerstin Bach、Alberto José Bugarín Diz、Jose Maria Alonso-Moral、Senén Barro 和 Fredrik Heintz 编，*ECAI 2024——第27届欧洲人工智能会议，2024年10月19日至24日，西班牙圣地亚哥-德孔波斯特拉——暨第13届智能系统杰出应用会议（PAIS 2024）*，人工智能与应用前沿第392卷，第1254–1261页。IOS出版社，2024年。doi：10.3233/FAIA240622。URL：https://doi.org/10.3233/FAIA240622。

Luca Geatti、Alessandro Gianola 和 Nicola Gigante。有限轨迹上的理论模线性时序逻辑。载于 *IJCAI*，第22卷，第2641–2647页，2022年。

Luca Geatti、Alessandro Gianola、Nicola Gigante 和 Sarah Winkler。LTLf 模理论的可判定片段。载于 Kobi Gal、Ann Nowé、Grzegorz J. Nalepa、Roy Fairstein 和 Roxana Radulescu 编，*ECAI 2023——第26届欧洲人工智能会议，2023年9月30日至10月4日，克拉科夫，波兰——暨第12届智能系统杰出应用会议（PAIS 2023）*，人工智能与应用前沿第372卷，第811–818页。IOS出版社，2023年。doi：10.3233/FAIA230348。URL https://doi.org/10.3233/ FAIA230348。

Silvio Ghilardi、Alessandro Gianola、Marco Montali 和 Andrey Rivkin。关系动作基的安全验证与全称不变式。载于第三十二届国际人工智能联合会议论文集，*IJCAI 2023，2023年8月19日至25日*，中国澳门特别行政区，第 3248–3257 页。ijcai.org，2023年。doi：10.24963/IJCAI.2023/362。URL：https://doi. org/10.24963/ijcai.2023/362。

Giuseppe De Giacomo 和 Sasha Rubin。面向 LTLf 和线性动态逻辑有限迹目标的 FOND规划自动机理论基础。载于 Jérôme Lang 编第二十七届国际人工智能联合会议论文集，*IJCAI 2018，2018年7月13日至19日*，瑞典斯德哥尔摩，第 4729–4735 页。ijcai.org，2018年。doi：10.24963/IJCAI.2018/657。URL：https://doi.org/10. 24963/ijcai.2018/657。

Alessandro Gianola。 通过理论模可满足性验证数据感知流程，第 470 卷，收录于商业信息处理讲义。施普林格，2023年。doi：10.1007/ 978-3-031-42746-6。网址：https://doi.org/10.1007/978-3-031-42746-6。

Alessandro Gianola、Marco Montali 和 Sarah Winkler。基于覆盖与自动机对数据感知流程进行模理论线性时间验证。载于*AAAI*，第 10525–10534 页。AAAI出版社，2024年。

Charles Gretton。一种表达能力更强的决策理论规划行为逻辑。载于*PRICAI*，《计算机科学讲义》第 8862 卷，第 13–25 页。施普林格，2014年。Charles Gretton、David Price 和 Sylvie Thiébaux。具有非马尔可夫奖励的决策过程求解方法的实现与比较。载于*UAI*，第 289–296 页。Morgan Kaufmann，2003年。Charles Gretton、Froduald Kabanza、David Price、John K. Slaney 和 Sylvie Thiébaux。具有非马尔可夫奖励的决策理论规划。*CoRR*，abs/1109.2355，2011年。Hosein Hasanbeig、Daniel Kroening 和 Alessandro Abate。逻辑引导的认证强化学习。人工智能，322：103949，2023年。Mohammadhosein Hasanbeig、Daniel Kroening 和 Alessandro Abate。结合时序逻辑的深度强化学习。载于*FORMATS*，《计算机科学讲义》第 12288 卷，第 1–22 页。施普林格，2020年。Rodrigo Toro Icarte、Toryn Klassen、Richard Valenzano 和 Sheila McIlraith。在强化学习中使用奖励机进行高层任务规范与分解。载于国际机器学习会议，第 2107–2116 页。PMLR，2018年。Rodrigo Toro Icarte、Toryn Q Klassen、Richard Valenzano 和 Sheila 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》，73：173–208，2022年。Yuqian Jiang、Suda Bharadwaj、Bo Wu、Rishi Shah、Ufuk Topcu 和 Peter Stone。面向持续强化学习任务的时序逻辑奖励塑形。载于*AAAI*，第 7995–8003 页。AAAI出版社，2021年。Kishor Jothimurugan、Rajeev Alur 和 Osbert Bastani。一种用于强化学习任务的可组合规范语言。载于*NeurIPS*，第 13021–13030 页，2019年。Laura Kovács 和 Andrei Voronkov。一阶定理证明与 Vampire。载于计算机辅助验证——第*25*届国际会议，*CAV 2013*，俄罗斯圣彼得堡，*2013*年*7*月*13*日至*19*日。论文集，《计算机科学讲义》第 8044 卷，第 1–35 页。施普林格，2013年。doi：10.1007/978-3-642-39799-8_1。URL：https://doi.org/10.1007/978-3-642-39799-8_1。Bruno Lacerda、David Parker 和 Nick Hawes。针对部分可满足余安全线性时序逻辑规范的最优策略生成。载于*IJCAI*，第 1587–1593 页。AAAI出版社，2015年。Edouard Leurent。一个用于自动驾驶决策的环境。https://github. com/eleurent/highway-env，2018年。Andrew C. Li、Zizhao Chen、Toryn Q. Klassen、Pashootan Vaezipoor、Rodrigo Toro Icarte 和 Sheila A. McIlraith。用于嘈杂和不确定环境中深度强化学习的奖励机。载于*NeurIPS*，2024年。Xiao Li、Cristian Ioan Vasile 和 Calin Belta。带有时序逻辑奖励的强化学习。载于*IROS*，第 3834–3839 页。IEEE，2017年。Timothy P. Lillicrap、Jonathan J. Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。结合深度强化学习的连续控制。载于*ICLR*（海报），2016。Andrew Y. Ng、Daishi Harada 和 Stuart Russell。奖励变换下的策略不变性：理论及其在奖励塑形中的应用。载于*ICML*，第 278–287 页。Morgan Kaufmann，1999年。Amir Pnueli。程序的时序逻辑。载于第*18*届计算机科学基础年会，普罗维登斯，罗得岛州，美国，*1977*年*10*月*31*日至*11*月*1*日，第 46–57 页。IEEE计算机学会，1977a。doi：10.1109/SFCS.1977.32。URL：https://doi.org/10. 1109/SFCS.1977.32。Amir Pnueli。程序的时序逻辑。载于第*18*届计算机科学基础年会（*SFCS 1977*），第 46–57 页。IEEE，1977b。

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

Antonin Raffin、Ashley Hill、Adam Gleave、Anssi Kanervisto、Maximilian Ernestus 和 Noah Dormann。Stable-baselines3：可靠的强化学习实现。《机器学习研究杂志》，22(268)：1–8，2021年。URL http://jmlr.org/papers/v22/20-1364.html。

Andoni Rodríguez 和 César Sánchez。用于模理论可实现性的布尔抽象。载于 Constantin Enea 和 Akash Lal（编），计算机辅助验证——第*35*届国际会议，*CAV 2023*，巴黎，法国，*2023*年*7*月*17*日至*22*日，论文集，第三部分，《计算机科学讲义》第13966卷，第305–328页。施普林格，2023年。doi: 10.1007/978-3-031-37709-9_15。URL https://doi.org/10.1007/978-3-031-37709-9_15。

Andoni Rodríguez 和 César Sánchez。针对线性时序逻辑和模理论LTLf的自适应反应式综合。载于 Michael J. Wooldridge 、 Jennifer G. Dy 和 Sriraam Natarajan 编，第三十八届 *AAAI* 人工智能大会、*AAAI 2024*、第三十六届人工智能创新应用大会、*IAAI 2024*、第十四届人工智能教育进展研讨会、*EAAI 2014*，*2024*年*2*月*20*日至*27*日，加拿大温哥华，第 10679–10686 页。AAAI出版社，2024年。doi: 10.1609/AAAI.V38I9.28939。URL https://doi.org/10.1609/aaai.v38i9.28939。

Andoni Rodríguez、Guy Amir、Davide Corsi、César Sánchez 和 Guy Katz。针对模理论LTLf的盾牌综合。载于 Toby Walsh、Julie Shah 和 Zico Kolter 编，*AAAI-25*，由人工智能促进协会赞助，*2025*年*2*月*25*日至*3*月*4*日，美国宾夕法尼亚州费城，第 15134–15142 页。AAAI出版社，2025年。doi: 10.1609/AAAI.V39I14. 33660。URL https://doi.org/10.1609/aaai.v39i14.33660。

Richard S. Sutton 和 Andrew G. Barto。《强化学习：导论》（第 2 版）。麻省理工学院出版社，2018年。

Alfred 塔斯基。初等代数与几何学的判定方法。《加利福尼亚大学 *a* *P*出版社》，1951年。

Sylvie Thiébaux、Froduald Kabanza 和 John K. Slaney。用于处理具有非马尔可夫奖励的决策过程的任意时刻状态式求解方法。载于《UAI》，第 501–510 页。Morgan Kaufmann，2002年。

Shufang Zhu、Geguang Pu 和 Moshe Y. Vardi。LTLf 到自动机转换的一阶与二阶编码，2019年。网址：https://arxiv.org/abs/1901.06108。

任务描述

本节将进一步介绍实验设置。我们会描述所用的环境，以及为实验定义的任务（包括公式和文字描述）。某些常量的数值赋值体现了我们对任务的手动配置（目标分配），同时也体现了哪些常量可由事后经验回放操控，以定义虚构目标。

A.1 停车环境

停车环境 [Leurent, 2018] 表示一个二维停车场，其中智能体控制一辆汽车，由连续向量 $s \in \mathbb{R}^6$表示。状态包含空间坐标 x， y、速度 v_x， v_y和朝向 $\alpha = \sin\ \theta$， $\beta = \cos\theta$，其中 θ是偏航角。动作空间为连续的二维空间（油门和转向角）。接下来，我们列出为该环境定义的任务。公式中出现的变量只有x， y， α， β，而 a， b， c， d则为常量。

任务1 在任务_1 中，智能体需要先前往检查点位置 $A = (-0.2, -0.08)$，然后在停车场内驶向目标位置 $G = (a, b)$，并遵循这一时间顺序。描述该任务的公式由汽车空间坐标变量 (x, y) 上的距离约束构成，容差为 $r = 0.03$:

$$\varphi_1 = F(\ (x+0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge XF\ (x-a)^2 + (y-b)^2 < 0.03^2\ ).$$

目标分配： $\{a = 0.2, b = 0.08\}$。

任务 2 在任务_2中，我们引入了一个额外的目标位置 $B = (e,\ f)$。智能体需要以任意顺序通过两个检查点 $A,\ B$，然后移动到目标 G。下面给出相应的公式：

$$\begin{aligned}\varphi_2 = & F(\ (x+0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge X \\ & F(\ (x-0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge XF\ (x-a)^2 + (y-b)^2 < 0.03^2\ )\ ) \\ & \vee \\ & F(\ (x-0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge X \\ & F(\ (x+0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge XF\ (x-a)^2 + (y-b)^2 < 0.03^2\ )\ ).\end{aligned}$$

目标分配： $\{a = 0.2, b = 0.08\}$。

任务 3 此任务在图 1 中表示为停车_1，针对目标G引入了更严格的规范。在通过检查点 A 后，智能体不再只是到达停车场中的某个位置，而是需要将汽车正确停放在特定位置，这意味着汽车还必须满足方向约束。这通过对空间变量和方向变量 $x,\ y,\ \alpha,\ \beta$施加框约束（加权 L_1 范数）来形式化，具体如下：

$$\begin{aligned}\varphi_3 = & F\ (\ (x+0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge X \\ & F\ (|x-a| + 0.2|y-b| + 0.02|\alpha - c| + 0.02|\beta - d| < 0.0144\ )\ ).\end{aligned}$$

where $r = 0.0144$是容差阈值。目标分配： $\{a = 0.18, b = 0.14, c = 0, d = 1\}$.

任务 4 这项称为停车_2的任务是我们测试中最难的，因为它综合了此前各项任务的复杂性。智能体需要依次通过两个检查点位置 $A,\ B$（顺序不限），然后将汽车准确停在指定位置。

$$\begin{aligned}\varphi_4 = & F(\ (x+0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge XF(\ (x-0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge X \\ & F\ (|x-a| + 0.2|y-b| + 0.02|\alpha - c| + 0.02|\beta - d| < 0.0144\ )\ ) \\ & \vee \\ & F(\ (x-0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge XF(\ (x+0.2)^2 + (y+0.08)^2 < 0.03^2\ \wedge X \\ & F\ (|x-a| + 0.2|y-b| + 0.02|\alpha - c| + 0.02|\beta - d| < 0.0144\ )\ ).\end{aligned}$$

目标分配： $\{a = 0.18, b = 0.14, c = 0, d = 1\}$。

停车结果分析 停车的更多结果如图3所示，其中针对每项任务，我们报告了 20 次运行中的累积遗憾和累积奖励。累积遗憾是相对于始终能够成功的理想策略来衡量的。这些补充结果体现了第 6 节中的分析。

A.2 Reacher 环境

Reacher 是一条在二维空间中运动的双关节机械臂。我们使用 Reacher-v5[5] 的一个变体，并修改了状态空间。我们的状态 $s \in \mathbb{R}^{11}$ 编码了每个手臂角度的正弦和余弦 $\alpha_1,\ \beta_1,\ \alpha_2,\ \beta_2$、每个手臂的角速度 $\omega_1,\ \omega_2$、目标的空间位置（$x_t,\ y_t$）以及机械臂末端的坐标（$x,\ y,\ z = 0$）。动作空间是连续的二维空间（两个铰链的扭矩）。下面介绍我们为 Reacher 定义的任务。公式中出现的变量只有 x 和 y，而 a 和 b是常量。

[5]https://gym nasium.farama.org/environments/mujoco/rea cher/

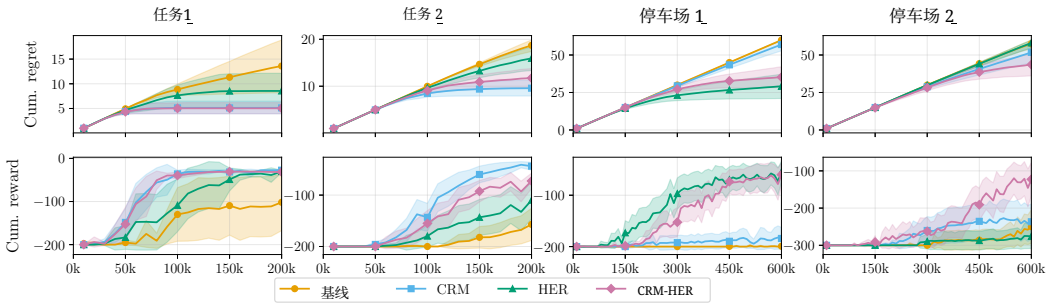

图2：停车环境中各任务的遗憾和累积奖励图。

任务1 在图3中名为任务_1 的任务里，智能体必须将手部（x，y）向左移动，也就是移动到框约束 $x < -0.19$所描述的任意位置，然后在 $r_g = 0.01$ 容差半径内将其移动到指定的目标位置 $G =$ （a，b）。公式如下:

$$\varphi_1 = F(\, x \leq -0.19 \, \wedge XF \, (x-a)^2 + (y-b)^2 < 0.01^2 \,).$$

目标分配： $\{a = 0.1, b = 0.1\}$。

任务 2 任务 task_2 需要额外的一步。将手臂向左移动后，智能体必须将手部移动到靠近其基座 $S =$ (0，0)的位置，容差距离为 $r_s = 0.03$，然后才能到达目标位置 G。公式如下:

$$\varphi_2 = F(\, x \leq -0.19 \, \wedge XF \, (\, x^2 + y^2 < 0.03^2 \, \wedge XF \, (x-a)^2 + (y-b)^2 < 0.01^2 \,)).$$

目标分配： $\{a = 0.1, b = 0.1\}$。

任务 3 任务 task_3 与任务 task_2类似，但到达目标位置所需的容差 $r_g = 0.005$ 更严格，容差距离从 0.01 降至 0.005，使任务复杂度大幅提升，可接受目标区域也显著缩小。该公式与前一个任务中的公式几乎相同:

$$\varphi_3 = F(\, x \leq -0.19 \, \wedge XF \, (\, x^2 + y^2 < 0.03^2 \, \wedge XF \, (x-a)^2 + (y-b)^2 < 0.005^2 \,))$$

目标分配： $\{a = 0.1, b = 0.1\}$。

Reacher 的结果分析 Reacher 环境的结果如图3所示。与之前一样，我们比较了 20 次运行中的成功率、累积遗憾和累积奖励。尽管没有哪一种技术在成功率或累积奖励方面明显占优，但 CRM-HER 的表现始终不逊于其他方法，并且在任务 task_1 和 task_2中展现出更快的收敛速度。分析累积遗憾时，这一优势更加明显：与其他方法相比，CRM-HER 始终是更优的选择。

B 超参数

表 1 和表 2 展示了各环境和任务的学习过程中选用的超参数。各参数说明如下:

- “MLP (400,300)”表示一个包含两个隐藏层的多层感知机，两个隐藏层分别由 400 和 300 个单元组成。
- α 表示 ADAM优化器进行梯度更新时所使用的学习率。
- τ 表示对网络进行 Polyak平均时，相对于相应目标网络所使用的软更新系数。
- γ 是折扣因子。

大多数实验采用 *StableBaselines 3* [Raffin 等人 2021] 提供的 DDPG 实现中的默认超参数。唯一的例外是停车环境中的任务 4，此任务增大了批次大小，以便算法能够学习到有意义的策略。

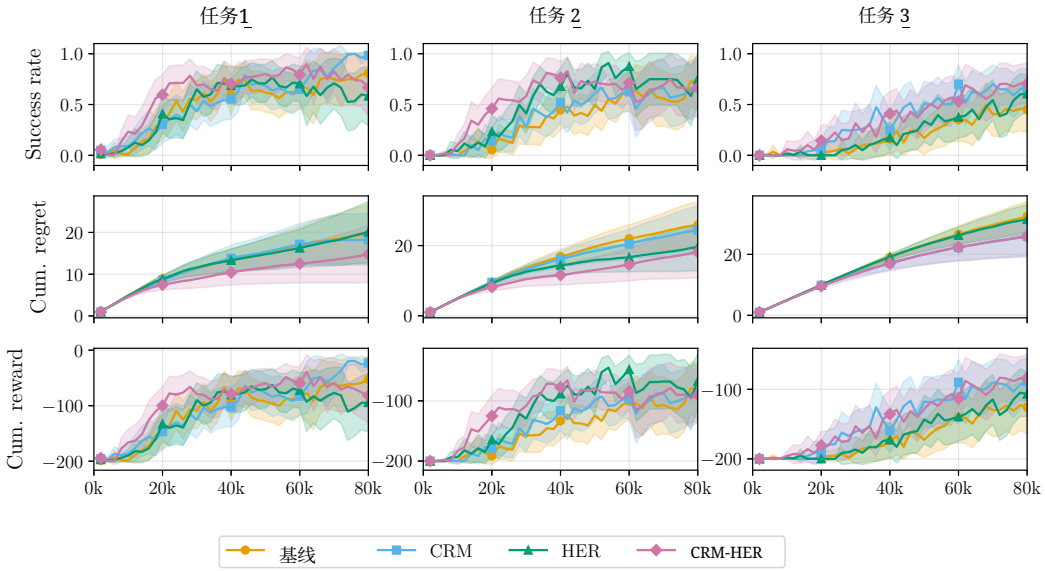

图3：Reacher 环境中各任务的成功率、累积遗憾和累积奖励。

停车环境超参数

	Task 1	Task 2	Task 3	任务 4
回合长度	200	200	200	300
缓冲区大小	10^6	10^6	10^6	10^6
批次大小	256	256	256	1024
演员网络	多层感知机（400，300）	多层感知机（400，300）	多层感知机（400，300）	多层感知机（400，300）
评论家网络	多层感知机 (400，300)	多层感知机 (400，300)	多层感知机 (400，300)	多层感知机 (400，300)
α	0.001	0.001	0.001	0.001
τ	0.005	0.005	0.005	0.005
γ	0.99	0.99	0.99	0.99

表1：表格展示了我们在停车环境中的实验里各任务所使用的超参数。

Reacher 环境超参数

	Task 1	Task 2	任务 3
回合长度	200	200	200
缓冲区大小	10^6	10^6	10^6
批次大小	256	256	256
演员网络	多层感知机（400，300）	多层感知机（400，300）	多层感知机（400，300）
评论家网络	多层感知机（400，300）	多层感知机（400，300）	多层感知机 (400,300)
α	0.001	0.001	0.001
τ	0.005	0.005	0.005
γ	0.99	0.99	0.99

表 2：该表展示了我们在 Reacher 环境实验中每个任务所使用的超参数。

C 计算基础设施

所有实验均使用 Python3.10 实现，并在以下硬件配置的机器上运行：

- **中央处理器**：AMD Ryzen 9 7950X
- **图形处理器**：NVIDIA RTX 4080 Super
- **内存**：32GB DDR5 6000MHz