arXiv:2603.01292v1 [cs.LG] 1 Mar 2026

将线性时序逻辑约束融入近端策略优化，实现安全强化学习

Maifang Zhang[1], *Hang Yu[2], *Qian Zuo[1], *Cheng Wang[3],Vaishak Belle[1],和 Fengxiang He[1], †

摘要—本文提出带线性时序逻辑约束的近端策略优化（PPO-LTL），这是一种将以线性时序逻辑表述的安全约束融入近端策略优化、用于安全强化学习的框架。线性时序逻辑约束能够严格表示复杂的安全要求，例如机器人学中普遍存在的各类规范，从而实现对安全要求的系统化监控。通过极限确定性 B¨uchi 自动机监控违反线性时序逻辑约束的情况，再由逻辑到代价机制将其转化为惩罚信号。随后，利用拉格朗日方案通过这些信号引导策略优化。在 Zones 和 CARLA 环境中进行的大量实验表明，与最先进的方法相比，我们提出的 PPO-LTL 能够持续减少安全违规，同时保持具有竞争力的性能。代码见 https://github.com/EVIEHub/PPO-LTL。

I. 引言

强化学习（RL）已在包括机器人学在内的多个领域取得了显著成功 [1]。在强化学习中，近端策略优化（PPO）是一种被广泛采用的同策略方法，它通过裁剪代理目标约束策略更新，从而在探索与稳定性之间取得平衡 [2]。将强化学习部署于安全关键环境仍然极具挑战，其中对安全约束的违反项可能导致灾难性后果。安全强化学习已得到全面综述，并在约束优化框架下应对这一挑战：智能体在限制累积安全代价的同时寻求最大化奖励 [3]。在这一方法族中，通过拉格朗日方案实现的约束近端策略优化（PPO-Lagrangian）已成为一种主要方法 [4]。

尽管取得了这些进展，一个关键局限仍然存在：约束必须以智能体的状态和动作的解析不等式来表述。这与一大类抽象安全约束并不兼容，例如机器人学中广泛存在的法规。例如，英国《公路法典》对驾驶行为作出规定 [5]，而这些规定很难转化为上述不等式。这就要求在强化学习训练过程中采用机器可计算且有原则的安全规范。

为解决这一问题，我们提出带线性时序逻辑约束的近端策略优化（PPO-LTL），这是一种用于表示抽象约束的新方法以线性时序逻辑 [6], [7],表示，并将这些约束嵌入近端策略优化智能体。线性时序逻辑提供了一种严格且可由机器验证的工具，可将时序性质（如“始终避开不安全状态”“最终到达目标”）以及监管规则（如“红灯停车直至变绿”）编码为逻辑规范。

我们设计了一种逻辑到代价机制，可将违反时序逻辑约束的情况系统地转换为引导策略学习的代价信号。该机制可应用于广泛的环境中，作为即插即用的解决方案。每个线性时序逻辑规范首先被编译为一个 ω-自动机，通常是极限确定性布奇自动机（LDBA） [8], [9]。该自动机通过基于原子命题定义有限状态集和带标签的转换，编码时序属性的满足条件。在执行过程中，它与智能体—环境交互同步演化，实质上充当运行时监视器，用于检查智能体的轨迹是否满足规范。

发生违反时，监视器会发出代价信号，以反映违反的严重程度；该严重程度由与不同安全规则相关联的一组预定义权重决定。这些违规代价会随时间进行聚合，并整合到安全强化学习框架中。由此产生的信号随后通过拉格朗日方案与策略优化相结合，使智能体能够在确保满足安全要求的同时优化任务性能。与手工设计的惩罚不同，这种方法以一种有原则、可泛化且模块化的方式，将高级安全要求编码到学习过程中。

我们对所提出的方法进行了严格的理论分析。我们将 PPO-LTL 表述为一种由有偏随机梯度预言机驱动的不精确投影原始–对偶方法。具体而言，我们将 PPO 中的两个常见组成部分——裁剪代理目标和有限轮次的小批量更新——视为对真实拉格朗日梯度进行有偏随机近似的机制。随后，我们证明了支撑 PPO-LTL 的投影原始–对偶动力学具有遍历平稳性保证。该结果表明，尽管梯度估计存在偏差和噪声，我们的算法仍能稳定地收敛到平稳点的一个邻域。实际上，这意味着 PPO-LTL 无需依赖精确的梯度评估即可稳定减少约束违反，凸显了其在自动驾驶等具有挑战性的场景中的鲁棒性。

我们开展了全面的实验，以在多种基准环境中评估近端策略优化-线性时序逻辑，包括 ZonesEnv（具有逻辑区域的连续控制） [10]和 CARLA（自动驾驶模拟器） [11]。我们的方法

*共同贡献。

†通讯作者。电子邮箱：F.He@ed.ac.uk。[1]M. Zhang、Q. Zuo、V. Belle 和 F. He 曾就职于苏格兰爱丁堡 EH8 9AB 的爱丁堡大学信息学院。[2]H. Yu 曾就职于澳大利亚新南威尔士州达灵顿 2008 的悉尼大学工程学院计算机科学系。[3]C. Wang 曾就职于苏格兰爱丁堡 EH14 4AS 的赫瑞瓦特大学工程与物理科学学院。

与作为基线的近端策略优化 [2], 以及一系列最先进的安全强化学习方法进行了比较，包括 TIRL-PPO、TIRL-SAC [12], PPO-Mask、PPO-Shielding [13], [14],和 PPO-Lagrangian。实证结果表明，在保持有竞争力的任务性能的同时，安全违规持续减少。广泛的消融研究和敏感性分析进一步验证了我们的贡献。

II. 相关工作

安全屏蔽。安全屏蔽方法通过使用经验证的策略或在抽象模型上进行模型检测，在线预先阻止不安全动作，从而保障安全性 [13], [14]。这能实现二元安全行为，并提供强有力的形式化保证，但可能限制探索，并导致数据分布非平稳。相比之下，软集成方法通过逻辑到代价机制将违反项映射为成本信号，提供密集反馈，兼容基于梯度的优化和约束马尔可夫决策过程求解器 [4], [15]。使用线性时序逻辑监视器生成逐规则成本，还能进一步实现对多个约束的模块化处理、跨规范的组合推理，以及向大型规则集直接扩展 [7], [16]。与运行时动作过滤策略相比，PPO-LTL 将时序逻辑约束直接纳入策略优化循环，提供密集反馈信号，以支持在复杂时序要求下进行学习。

强化学习中的逻辑工具。除线性时序逻辑外，其他形式逻辑也已在强化学习中得到探索。信号时序逻辑（STL） [17] 将时序推理扩展到实值信号，而道义逻辑 [18]则表达义务和许可等规范性概念。这些方法具有更强的表达能力，但通常计算成本较高，且任务特定的设计复杂度较高，因而其实用性受到限制。概率逻辑编程（PLP） [19] 也已被用于对逻辑推理中的不确定性建模，使智能体能够编码概率约束。相比之下，线性时序逻辑实现了良好平衡：它轻量、易于编译为自动机，并且足以表达时序安全性质 [20]。近期研究 [21]通过乘积马尔可夫决策过程和线性规划，在基于模型的驾驶框架中应用线性时序逻辑以进行策略综合；而我们的方法侧重于基于学习的约束优化，而非基于模型的综合。

III. 预备知识

马尔可夫决策过程（MDP）。我们考虑一个折扣约束马尔可夫决策过程 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, c, \mu, \gamma)$，其中$\mathcal{S}$ 和 $\mathcal{A}$ 分别表示状态空间和动作空间，$P(s'|s,a)$表示在给定动作 a 的情况下从状态 s 转移到状态 s' 的转移概率，$r : (s,a) \to [0, R_{\max}]$ 是奖励函数，$c : (s,a) \to [0, C_{\max}]$ 是聚合成本 $c(s,a) = \sum_{i=1}^{K} w_i c^{(i)}(s,a)$，其中 w_i是 $i \in [K]$的固定权重，μ 是初始状态分布，$\gamma \in (0,1)$ 是折扣因子，用于控制智能体对未来奖励的重视程度。策略 $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ 是从状态空间映射到动作空间上的概率分布空间的映射，而 $\pi(a \mid s)$ 表示在状态 s下选择动作 a 的概率。令 $\theta \in \Theta \subset \mathbb{R}^p$表示随机策略 $\pi_\theta(\cdot|s)$的参数。对于给定策略 π_θ，定义在 $\mathcal{S} \times \mathcal{A}$ 上的折扣占用度量为 $d_\theta^\gamma = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbf{P}\pi_\theta(s_t = s, a_t = a|\mu)$。我们将奖励价值函数和成本价值函数定义为:

$$J_R(\theta) = \mathbb{E}_{(s,a)\sim d_\theta^\gamma}\left[r(s,a)\right], \; J_C(\theta) = \mathbb{E}_{(s,a)\sim d_\theta^\gamma}\left[c(s,a)\right].$$

约束优化目标如下，

$$\max_{\theta\in\Theta} J_R(\theta) \quad \text{s.t.} \quad J_C(\theta) \le d,$$

其中 $d = \sum_{i=1}^{K} w_i d_i$ 是聚合安全预算。相应的拉格朗日函数定义为 $\mathcal{L}(\theta, \lambda) = J_R(\theta) - \lambda(J_C(\theta) - d)$，其中 $\lambda \in [0, \Lambda]$ 是拉格朗日乘子，且 $\Lambda < \infty$。

近端策略优化（PPO）。策略梯度方法直接优化策略，但如果更新过大，则可能出现稳定性问题 [22]。近端策略优化通过裁剪更新来解决这一问题，相当于给学习过程加上一条“安全带”。裁剪代理目标 $L^{\text{PPO}}(\pi)$定义如下，

$$\mathbb{E}_t\Big[\min\left(\rho_t(\pi)A_t,\ \text{clip}(\rho_t(\pi), 1-\epsilon, 1+\epsilon)A_t\right)\Big],$$

其中 $\rho_t(\pi) = \frac{\pi(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$ 是重要性比率，A_t 是优势估计（衡量一个动作相对于平均水平有多好），ϵ是防止剧烈更新的裁剪参数。在实践中，近端策略优化通过对收集到的轨迹采样进行多轮小批量随机梯度更新来优化该目标。

线性时序逻辑（LTL）。线性时序逻辑是一种形式语言，用于描述系统状态无限序列上的时序性质。线性时序逻辑中的公式称为规范，用于定义系统期望的时序行为。线性时序逻辑公式由有限集合的原子命题$\mathcal{P}$构成，这些原子命题表示有关环境的基本事实（例如“发生了碰撞”或“已到达目标”）。每个命题都是关于系统状态的布尔语句，其值非真即假，例如车辆当前是否处于安全区域，或交通信号灯是否为绿色。这些命题通过布尔联结词（$\neg$、$\wedge$、$\vee$）以及$\mathbf{G}$ （始终）、$\mathbf{F}$ （最终）、$\mathbf{X}$ （下一时刻）和$\mathbf{U}$ （直到）等时序算子组合起来，以表达随时间变化的更复杂要求。线性时序逻辑公式的语义规定了规范被视为满足时所需的精确条件。这些语义定义在轨迹 $\sigma = s_0 s_1 s_2 \ldots$上，其中 $\sigma, t \models \varphi$ 表示规范 φ 在时间 t 成立——换言之，系统截至时间t 的行为符合 φ 所描述的性质。例如，$\mathbf{G}\neg$collision要求永远不会发生碰撞，而$\mathbf{F}$goal表示最终必须达到目标状态。

IV. PPO-LTL

A. 将约束表示为 LTL规范

在现实应用中，许多安全要求同时取决于环境状态（例如位置、与障碍物的距离）以及事件的时序结构，

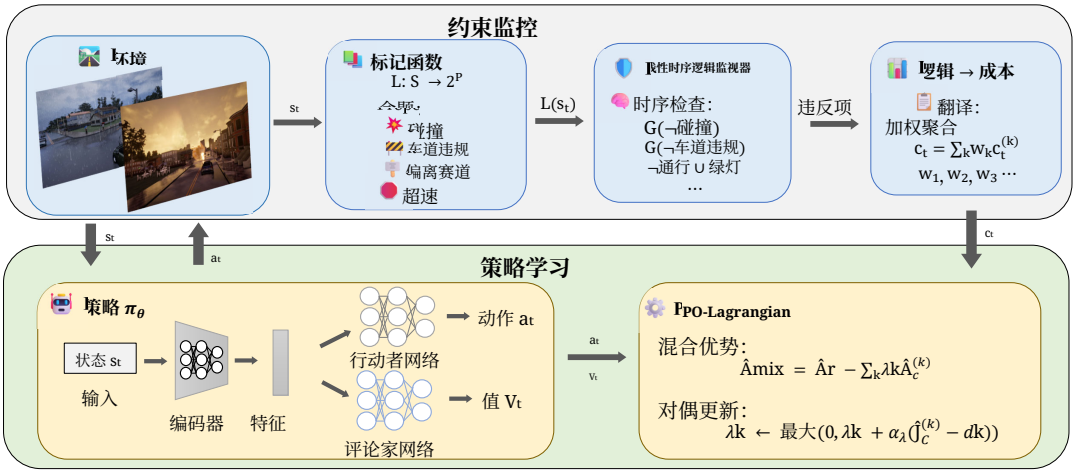


图1：PPO-LTL：环境状态以原子命题标记，由 LTL 检查器监控并生成约束代价，再将其与任务奖励结合，用于策略优化。

（例如事件的顺序和时机） [3], [4], [16], [23]。例如，“访问检查点后最终抵达目标”这类要求，无法简单地编码为标量惩罚。它们需要对状态序列进行时序推理 [24], [20]。为形式化地表示此类要求，我们将其定义为 LTL 规范。每个规范都是一个逻辑公式，用于描述系统轨迹上期望的时序行为，并作为智能体策略应满足的形式化要求。已有先导研究将交通法规表示为 LTL 规范。线性时序逻辑将时序算子与布尔联结词结合起来，以描述复杂行为，例如始终避免碰撞，同时最终抵达目的地（**G**¬碰撞 $\wedge$ **F**目的地），或要求进入交叉路口后最终亮起绿灯（**G**(交叉路口 $\rightarrow$ **F**绿灯）。这种表达能力使自动驾驶场景中的多阶段驾驶规则得以紧凑建模。

B. 逻辑到代价机制

B¨uchi 自动机与 LDBA。每个 LTL 规范都可以转换为 B¨uchi 自动机（BA），这是一种状态转移结构，用于监控智能体的事件序列是否满足时序规则。自动机读取交互轨迹；执行过程中，只有当指定的接受状态被无限次访问时，才视为满足规范。在强化学习中，我们采用一种称为极限确定性 B¨uchi 自动机（LDBA）的简化变体，它能提供更可预测的检查，并提高计算效率。LDBA 使高层时序逻辑能够在训练过程中逐步评估，从而为将符号规则转换为指导策略学习的数值信号奠定基础。

监视器。训练期间，时序规范由与环境同步演化的运行时监视器进行检查。每个监视器都会观察轨迹，并判断规范 ϕ_i 是否得到满足。检测到违规相关转移时，监视器会发出非负代价信号 $c_t^{(i)}$，其大小由反映相对重要性的规则特定权重决定。多个监视器可能同时生成代价，所有违规代价都会汇总起来，在 CMDP框架中引导策略优化：$c_t = \sum_{i=1}^{K} c_t^{(i)}$。安全关键规则会产生更大的代价，而目标导向要求则会持续产生较高代价，直到条件得到满足。

到达-避让分解。为进一步简化策略优化，每个编译自动机都可以分解为一系列到达-避让子任务 [24], [25]。每个子任务包含：(1) 到达条件：最终必须发生的状态或事件；(2) 避让条件：绝不能发生的事件。例如，**F**(目标) $\wedge$**G**¬碰撞可分解为两个子任务：“始终避开碰撞”和“最终到达目标”。这种转换在保留原有时序语义的同时，降低了策略搜索的复杂度。

逻辑到代价机制。因此，环境最终反馈为：(s_{t+1}, r_t, c_t, info)，其中 r_t表示任务性能， c_t 表示汇总后的约束成本，而 info 则提供每条规则的诊断信息。这种运行时逻辑到成本转换与领域无关，可应用于各种不同的环境，引导强化学习策略在优化性能的同时满足时序逻辑约束。

C. PPO-LTL 中的拉格朗日方案

鉴于逻辑到成本机制产生的逐步违规成本，PPO-LTL将其直接纳入策略优化过程。我们使用原始-对偶方法求解约束优化问题，其中约束信息通过混合优势信号$\hat{}$ $A_{\text{mix}} = \hat{}A_r - \sum_{k=1}^{K} \lambda_k \hat{}A_c^{(k)}$影响策略更新，其中$\hat{}$ A_r 和$\hat{}$$A_c^{(k)}$分别是奖励和成本的广义优势估计。每次近端策略优化更新后，乘子通过投影梯度上升进行更新：

$$\lambda_k \leftarrow \max\left(0,\ \lambda_k + \alpha_\lambda\big(\hat{J}_C^{(k)} - d_k\big)\right).$$

当成本超过其预定义限值时， λ_k 增大，从而加强对违反项所施加的惩罚。相反，当成本保持在可接受范围内时， λ_k 减小或保持不变，使策略能够继续提升任务性能。

V. 理论保证

本节分析 PPO-LTL 的收敛性质。我们在乘积马尔可夫决策过程框架内对学习过程进行建模：将状态空间视为增广空间 $\mathcal{S} = \mathcal{S}_{env} \times \mathcal{Q}$，其中 $\mathcal{S}_{env}$ 是环境状态，$\mathcal{Q}$是 LDBA 状态。具有时间依赖性的线性时序逻辑代价由此变为严格满足马尔可夫性质，使得代价函数可定义为 $c(s, a) = \sum_{k=1}^{K} w_k c^{(k)}(s, a)$。相应地，PPO-LTL 按如下方式生成迭代序列$\{(\theta_t, \lambda_t)\}_{t\geq 0}$：

$$\theta_{t+1} = \Pi_{\Theta}(\theta_t + \alpha \hat{g}_t), \quad \lambda_{t+1} = \Pi_{[0,\Lambda]}(\lambda_t + \beta \hat{u}_t), \quad (1)$$

其中， Π 表示欧几里得投影， α、 $\beta > 0$ 为学习率，$\hat{g}_t$ 是原始目标的随机上升方向，$\hat{u}_t$ 是约束残差 $\hat{J}_C(\theta_t) - d$ 的随机估计。我们进一步将 PPO-LTL 抽象为一个带有有偏随机梯度预言机的不精确投影原始-对偶框架：PPO 中的截断代理优化和有限轮次小批量更新被建模为对真实拉格朗日梯度的有偏随机估计。对于给定的学习率 $\alpha, \beta > 0$，原始梯度映射和对偶梯度映射定义如下：

$$\mathcal{G}(\theta, \lambda) := \frac{1}{\alpha}\left(\Pi_{\Theta}\left(\theta + \alpha \nabla_\theta \mathcal{L}(\theta, \lambda)\right) - \theta\right),$$
$$\mathcal{H}(\theta, \lambda) := \frac{1}{\beta}\left(\Pi_{[0,\Lambda]}\left(\lambda + \beta(J_C(\theta) - d)\right) - \lambda\right),$$

量化该问题在投影下的一阶平稳性。

该理论依赖于对原始域和随机梯度的两个温和假设，这些假设见第 VII 节，并遵循通常的做法 [26], [2], [27]。我们证明以下定理。

定理*1*：在假设1和假设2以及学习率 $0 < \alpha \leq 1/(4L_{\mathcal{L}})$ 和 $\beta > 0$ 的条件下，设 $\{\theta_t, \lambda_t\}_{t\geq 0}$ 由 (1) 定义。那么，对于所有 $T \geq 1$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{G}(\theta_t, \lambda_t)\|] \\
\leq &\sqrt{\frac{2(\Delta_{\mathcal{L}} + 2\Lambda U_{\max})}{\alpha T}} + O\left(\sqrt{\sigma_\theta^2 + a^2 + \alpha}\right) \\
&+ O\left(\left(G_{\max}^2 + \sigma_\theta^2 + a^2\right)^{1/4}\right), \\
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\mathcal{H}(\theta_t, \lambda_t)\|] \\
\leq &\sqrt{\frac{\Delta_{\mathcal{L}}}{\beta T}} + O\left(\sqrt{U_{\max}^2 + \sigma_\lambda^2 + b^2 + \alpha^2 + \frac{G_{\max}^2}{\beta}}\right) \\
&+ O\left(\frac{\alpha}{\sqrt{\beta}}\sqrt{G_{\max}^2 + \sigma_\theta^2 + a^2}\right),
\end{aligned}$$

表 I：ZonesEnv 中的结果。

方法	奖励	撞墙率	λ
PPO	17.95 ± 0.84	3.7 ± 2.1%	-
PPO-Mask	10.14 ± 2.82	6.0 ± 2.0%	-
PPO-Shielding	15.92 ± 0.55	12.0 ± 2.0%	-
PPO-Lagrangian	23.23 ± 1.51	6.0 ± 2.0%	0.00
PPO-LTL-A	17.86 ± 0.73	4.3 ± 3.5%	0.0048
PPO-LTL-B	18.61 ± 0.67	4.7 ± 3.1%	0.0018

其中 $\Delta_{\mathcal{L}} = \sup_{\Theta\times[0,\Lambda]} \mathcal{L}(\theta, \lambda) - \inf_{\Theta\times[0,\Lambda]} \mathcal{L}(\theta, \lambda) < \infty$ 上确界 (,) 下确界 (,)， $U_{\max} = \sup_{\theta\in\Theta} |J_C(\theta) - d|$ 上确界 ()， $G_{\max} = \sup_{(\theta,\lambda)} \|\nabla_\theta \mathcal{L}(\theta, \lambda)\|$，而 $O(\cdot)$ 表示与问题相关、但与 T 无关的常数。

详细证明见第七节。定理1为 PPO-LTL 所依据的投影原始-对偶动力学建立了遍历平稳性保证。该定理表明，尽管 PPO 固有的梯度估计存在偏差和噪声（例如，由裁剪和小批量采样导致，这些影响由方差项和偏差项 σ 和 a, b 表示），算法仍能可靠地收敛到平稳点的邻域。在实践中，这意味着 PPO-LTL 无需精确计算梯度，就能稳定地减少约束违反，从而证实其在自动驾驶等复杂环境中的鲁棒性。

六、实验

我们在 ZonesEnv 和 CARLA 上进行了大量实验，充分验证了我们的算法。

基线与比较方法。我们采用 PPO 作为无约束基线。将 TIRL-PPO 和 TIRL-SAC 纳入标准安全强化学习基线，以评估其他约束优化技术的表现。PPO-Mask 是一种启发式安全过滤器，会在不安全动作即将发生时，主动将其替换为预先定义的安全备选动作（例如紧急制动）。我们将其纳入实验，以展示纯粹依赖反应式、基于规则的干预的局限性（例如死锁和过度保守）。PPO-Shielding 是实验中的主要竞争方法。我们还纳入 PPO-Lagrangian 作为标准约束强化学习方法。为确保公平比较，所有方法均采用卷积神经网络骨干网络，由 6 个带有 ReLU激活的卷积层组成，其后接策略网络和价值网络；二者均采用包含[500, 300]个单元的 2 层多层感知机。

环境与实现。我们在两个环境中进行评估，使用 256 维的 CustomMultiIn- putExtractor [28] 和 3 个随机种子。(1)**ZonesEnv：** 一个 Safety Gymnasium [10] 网格世界，点机器人在代表原子命题的彩色区域中导航（例如，“避开蓝色区域，直到到达绿色区域”）。墙壁碰撞会产生惩罚，线性时序逻辑违反会产生成本。模型训练 200k 步。(2) **CARLA：** 一个自动驾驶模拟器 [11]（v0.9.13, Town02）。观测包含语义分割、自车状态和独热编码 LDBA状态。线性时序逻辑监视器将事件映射为确定性成本，用于近端策略优化-拉格朗日更新。模型训练 100k 步。

TAB表 II：CARLA 中的结果。箭头（↑，↓）表示数值严格优于的方向 ter.

方法	碰撞率 ↓	路线 ↑	距离	速度	长度	碰撞次数	中心偏差	Cost	λ
PPO	0.262 ± 0.115	0.013 ± 0.008	6.58	0.45	4030	8.0	0.77	-	-
TIRL-PPO	0.173 ± 0.129	0.083 ± 0.111	6.37	0.07	9460	7.3	0.87	-	-
TIRL-SAC	0.336 ± 0.063	0.027 ± 0.011	6.92	0.92	926	57.3	0.12	-	-
PPO-Mask	0.408 ± 0.058	0.010 ± 0.012	2.38	1.06	1338	38.0	0.37	-	-
PPO-Shielding	0.267 ± 0.056	0.072 ± 0.036	18.87	10.63	93	164.3	0.52	-	-
PPO-Lagrangian	0.233 ± 0.089	0.077 ± 0.051	7.78	0.96	6616	103.0	0.60	0.00	0.01
PPO-LTL-A	**0.143 ± 0.110**	0.077 ± 0.107	5.96	3.57	703	154.3	0.52	0.25	0.03
PPO-LTL-B	0.170 ± 0.148	**0.236 ± 0.037**	12.79	1.48	4859	10.0	0.73	0.08	0.00

表 III：CARLA 中的消融研究和敏感性分析。

配置	碰撞率 ↓	路线 ↑	距离	速度	长度	碰撞次数	中心偏差	Cost	λ
*LTL*组件消融									
无碰撞	**0.159 ± 0.027**	0.009 ± 0.003	2.71	0.68	4123	11.0	0.64	0.07	0.00
无驶出赛道	0.275 ± 0.129	0.032 ± 0.044	3.29	0.06	10501	5.7	0.62	0.07	0.00
无车道侵入	0.207 ± 0.046	**0.024 ± 0.009**	6.17	1.98	14323	21.7	0.57	0.06	0.01
成本限制 *(cl)* 和对偶学习率 (α_λ) 的敏感性									
cl = 0.001	0.229 ± 0.153	0.033 ± 0.035	5.19	1.10	4749	21.0	0.78	0.06	0.01
cl = 0.05	0.273 ± 0.199	0.020 ± 0.023	3.04	4.65	907	228.7	0.56	0.22	0.71
cl = 0.5	0.395 ± 0.130	0.045 ± 0.059	4.86	0.09	6614	11.7	0.91	0.01	0.00
α_λ = 0.00001	**0.160 ± 0.059**	0.062 ± 0.063	8.67	0.89	2957	11.7	0.95	0.01	0.00
α_λ = 0.0001	0.243 ± 0.081	0.053 ± 0.033	8.98	0.54	3029	9.7	0.63	0.07	0.00
α_λ = 0.01	0.233 ± 0.097	**0.099 ± 0.065**	8.28	2.05	3776	8.3	0.86	0.06	0.03
约束严格程度与混合公式									
高松弛	0.258 ± 0.067	0.042 ± 0.040	4.51	0.09	3380	13.3	0.62	0.01	0.00
中度松弛	0.205 ± 0.100	0.056 ± 0.053	3.79	1.50	2787	53.7	0.53	0.01	0.00
碰撞聚焦	0.233 ± 0.206	0.010 ± 0.008	0.07	0.22	3020	15.7	0.39	0.00	0.00
超松散	0.393 ± 0.110	0.141 ± 0.098	8.81	1.04	3996	18.0	0.77	0.00	0.00
混合轻型	0.282 ± 0.076	0.042 ± 0.056	3.37	0.64	4218	18.0	0.90	0.01	0.00
混合轻型松散	0.242 ± 0.018	0.041 ± 0.033	2.86	0.26	5728	7.0	0.48	0.01	0.00
混合介质	0.229 ± 0.131	0.044 ± 0.029	5.62	3.69	3248	50.7	0.81	0.06	0.00
不含时序线性时序逻辑的仅碰撞变体									
仅碰撞 (cl=0.5)	0.243 ± 0.079	0.033 ± 0.043	4.93	0.52	4155	11.3	1.04	0.00	0.00
仅碰撞 (cl=0.1)	0.248 ± 0.228	0.028 ± 0.030	6.10	0.40	3573	14.3	0.90	0.00	0.00
仅碰撞中档	0.346 ± 0.175	0.158 ± 0.200	9.07	2.07	1584	52.0	0.86	0.01	0.00
仅碰撞宽松	0.193 ± 0.081	**0.269 ± 0.297**	12.66	0.93	9531	9.3	0.56	0.00	0.00

评估指标。我们评估任务性能和约束满足情况。安全指标包括各类约束（碰撞、偏离赛道、侵入车道、航向、蛇形行驶、超速、转向急变）的违规率，即违规事件数与总步数之比（例如，VR $=N_{\text{vio}}/N_{\text{total}}$） [4]。任务指标包括路线完成率（RCR）、平均速度 $\bar{v}$和总行驶距离 D_{total}，其中 RCR $= d_{\text{completed}}/d_{\text{target}}$。约束动态通过平均回合成本 $\hat{J}_C = \frac{1}{N}\sum_i C_i$、最终拉格朗日乘子 λ^*以及训练期间的收敛行为来衡量。

ZonesEnv 中的比较结果。表 I 展示了 3 个随机种子下的性能。PPO-LTL-A 优先考虑稳定性，而 PPO-LTL-B 则略微放宽约束。各基线方法存在不同缺陷：启发式 PPO-Mask 严重限制了探索，奖励仅为 10.14；PPO-Shielding 难以应对连续动态，撞墙率最高，达到 12.0%。尽管 PPO-Lagrangian 的表面奖励最高，为 23.23，但这具有欺骗性；由于缺少 LTL 记忆，它忽视了时序规则，并产生了高达 56.98 的违规成本（未在图中显示）。标准 PPO 的撞墙率较低，为 3.7%，但无法执行复杂的时序规范。相比之下，两种 PPO-LTL 变体都能提供较为均衡的策略。它们在有效奖励方面显著优于 Mask 和 Shielding，同时严格遵守 LTL 约束，且撞墙率也具有竞争力，分别为 4.3% 和 4.7%。

CARLA 中的比较结果。表 II 比较了 PPO-LTL-A（严格成本上限 0.02）和 PPO-LTL-B（宽松上限 0.1）与各基线方法的表现。PPO 无法平衡安全性与活性。标准安全强化学习基线方法表现出严重的异常行为：TIRL-PPO 存在机器人冻结问题（尽管存活了 9460 步，速度仍接近于零），而 TIRL-SAC 未能安全收敛（碰撞率为 0.336）。此外，PPO-Shielding 表现出鲁莽驾驶模式：尽管速度看似很高，却很快发生碰撞（长度为 93 步，碰撞次数为 164.3），路线完成度极低（0.072）。相反，PPO-Mask 的突然停止导致保守型死锁（距离为 2.38），并且反而带来了更高的碰撞率（0.408）。PPO-Lagrangian 缺乏时序前瞻能力，限制了进展（距离为 7.78）。相比之下，PPO-LTL 平衡了

表 IV：PPO-LTL 在 ZonesEnv 中对成本限制和拉格朗日学习率 (α_λ) 的敏感性分析。

成本限制	α_λ	奖励 ↑	撞墙率 ↓	λ
0.03	0.008	18.04 ± 1.30	6.7 ± 1.2%	0.0072
0.05	0.010	17.86 ± 0.73	**4.3 ± 3.5%**	0.0048
0.07	0.015	**18.61 ± 0.67**	4.7 ± 3.1%	0.0018
0.10	0.020	18.35 ± 1.42	5.7 ± 2.1%	0.0002

主动安全与任务活性，既避免过度保守，也避免鲁莽速度。PPO-LTL-A 达到了最低碰撞率（0.143，比标准 PPO 降低 45%）。PPO-LTL-B 达到了最高路线完成度（0.236），并保持了较长且稳定的回合。

消融研究与敏感性分析。表 III 通过系统性移除约束，评估了 CARLA 环境中的多种约束配置、各项约束的贡献以及超参数的稳健性。结果验证了谨慎平衡 LTL 约束的必要性：过于简单或过度放宽的边界容易导致速度策略过于天真或驾驶行为鲁莽，而经过适当调优的时序逻辑则能确保安全、有效地执行任务。此外，移除单项约束的实验也证实，多组件 LTL 约束对于实现均衡的驾驶性能至关重要。表 IV 展示了在 ZonesEnv 中，PPO-LTL 在不同成本限制和拉格朗日学习率下的敏感性分析结果。在成本限制相差三倍的范围内，该框架仍能保持稳定行为，表明约束机制能够对策略行为提供可解释且一致的控制。

计算成本。与标准 PPO 相比，PPO-LTL 的额外开销可以忽略不计。在 3 个随机种子下，在 ZonesEnv 中训练 200k 步，PPO-LTL（相较于 PPO）耗时 235.3s ± 0.9s（相较于 226.3s± 0.5s）；在 CARLA 中训练 100k 步，耗时 2557.3s ± 97.8s（相较于 2536.3s ± 39.0s）。这证实 LTL 监控和拉格朗日对偶更新仅带来极小的计算负担，使其在实际应用中仍能保持较高的效率。

VII. 证明

假设*1*：原始域 Θ 是紧致凸集，它刻画了信赖域正则化、裁剪或式（1）中的显式投影所导致的有界迭代。价值函数 $J_R(\theta)$ 和 $J_C(\theta)$ 在 Θ上连续可微，且其梯度满足 Lipschitz 条件：

$$\|\nabla_\theta J_R(\theta) - \nabla_\theta J_R(\theta')\| \le L_R\|\theta - \theta'\|$$
$$\|\nabla_\theta J_C(\theta) - \nabla_\theta J_C(\theta')\| \le L_C\|\theta - \theta'\|.$$

因此，对于任意 $\lambda \in [0, \Lambda]$，拉格朗日函数 $\mathcal{L}(\theta, \lambda)$ 关于 θ 是 $L_{\mathcal{L}}$-光滑的，且满足 $L_{\mathcal{L}} = L_R + \Lambda L_C$。

假设*2*：令 $\mathcal{F}_t$ 表示由截至 t开始时刻的所有随机性生成的 σ-域。估计值 $\hat{}g_t$ 满足

$$\mathbb{E}[\hat{g}_t|\mathcal{F}_t] = \nabla_\theta\mathcal{L}(\theta_t, \lambda_t) + a_t,$$

其中 $\|a_t\| \le a$，并且

$$\mathbb{E}[\|\hat{g}_t - \mathbb{E}[\hat{g}_t|\mathcal{F}_t]\|^2|\mathcal{F}_t] \le \sigma_\theta^2.$$

对偶信号 $\hat{}u_t$ 满足 $\mathbb{E}[\hat{}u_t|\mathcal{F}_t] = (J_C(\theta_t) - d) + b_t$，其中 $\|b_t\| \le b$，且具有有界方差 $\mathbb{E}[(\hat{}u_t - \mathbb{E}[\hat{}u_t|\mathcal{F}_t])^2|\mathcal{F}_t] \le \sigma_\lambda^2$。偏差项 a_t 和 b_t 涵盖了近似效应，包括裁剪不匹配、小批量更新和估计误差。

~证明。对于固定的 t，我们定义真实梯度 θ_{t+1}：~$=\Pi_\Theta(\theta_t + \alpha\nabla_\theta\mathcal{L}(\theta_t, \lambda_t))$。因此，有 $\theta_{t+1} - \theta_t = \alpha\mathcal{G}(\theta_t, \lambda_t)$。根据假设1，我们有

$$\mathcal{L}(\tilde{\theta}_{t+1}, \lambda_t) \ge \mathcal{L}(\theta_t, \lambda_t) + \langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t), \tilde{\theta}_{t+1} - \theta_t\rangle - \frac{L_{\mathcal{L}}}{2}\left\|\tilde{\theta}_{t+1} - \theta_t\right\|^2.$$

~~由于对于所有 θ，$\langle\theta_{t+1} - (\theta_t + \alpha\nabla_\theta\mathcal{L}(\theta_t, \lambda_t))\ \theta_t - \theta_{t+1}\rangle \ge 0$，取 $\theta = \theta_t$ 可得

$$\langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t), \tilde{\theta}_{t+1} - \theta_t\rangle \ge \frac{1}{\alpha}\|\tilde{\theta}_{t+1} - \theta_t\|^2.$$

于是得到

$$\mathcal{L}(\tilde{\theta}_{t+1}, \lambda_t) - \mathcal{L}(\theta_t, \lambda_t) \ge \left(\frac{1}{\alpha} - \frac{L_{\mathcal{L}}}{2}\right)\|\tilde{\theta}_{t+1} - \theta_t\|^2,$$
$$= \alpha\left(1 - \frac{\alpha L_{\mathcal{L}}}{2}\right)\|\mathcal{G}(\theta_t, \lambda_t)\|^2.$$

在 $\alpha < \frac{1}{4L_{\mathcal{L}}}$下，我们有 $1 - \frac{\alpha L_{\mathcal{L}}}{2} \ge \frac{7}{8}$，因此

$$\mathcal{L}(\tilde{\theta}_{t+1}, \lambda_t) - \mathcal{L}(\theta_t, \lambda_t) \ge \frac{7}{8}\alpha\left\|\mathcal{G}(\theta_t, \lambda_t)\right\|^2. \quad (2)$$

现在令 $\delta_t := \hat{}g_t - \nabla_\theta\mathcal{L}(\theta_t, \lambda_t)$ 表示 $\hat{}g_t$ 与 $\mathcal{L}(\theta_t, \lambda_t)$ 的梯度之间的差距。因此，有 $\|\theta_{t+1} - \theta_{t+1}\| \le \alpha\|\delta_t\|$。由于对于所有 $\theta \in \Theta$，$\theta_{t+1} = \Pi_\Theta(\theta_t + \alpha(\nabla_\theta\mathcal{L}(\theta_t, \lambda_t) + \delta_t))$，我们有$\langle\theta_{t+1} - (\theta_t + \alpha(\nabla_\theta\mathcal{L}(\theta_t, \lambda_t) + \delta_t))\ \theta - \theta_{t+1}\rangle \ge 0$。取 ~$\theta = \theta_{t+1}$，则，

$$\langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t) + \delta_t, \theta_{t+1} - \tilde{\theta}_{t+1}\rangle \ge \frac{1}{\alpha}\langle\theta_{t+1} - \theta_t, \theta_{t+1} - \tilde{\theta}_{t+1}\rangle. \quad (3)$$

~同样，根据 θ_{t+1}的定义，

$$\langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t), \tilde{\theta}_{t+1} - \theta_{t+1}\rangle \ge \frac{1}{\alpha}\langle\tilde{\theta}_{t+1} - \theta_t, \tilde{\theta}_{t+1} - \theta_{t+1}\rangle. \quad (4)$$

我们将其分解为

$$\langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t), \theta_{t+1} - \theta_t\rangle = \langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t), \tilde{\theta}_{t+1} - \theta_t\rangle + \langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t), \theta_{t+1} - \tilde{\theta}_{t+1}\rangle.$$

对于第二项，重新整理式 (3) 得到$\langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t)\ \theta_{t+1} - \tilde{\theta}_{t+1}\rangle \ge \frac{1}{\alpha}\langle\theta_{t+1} - \theta_t\ \tilde{\theta}_{t+1} - \tilde{\theta}_{t+1}\rangle - \langle\delta_t, \tilde{\theta}_{t+1} - \theta_{t+1}\rangle$。合并这些项并利用几何恒等式 $\langle x, x - y\rangle = \frac{1}{2}(\|x\|^2 + \|x - y\|^2 - \|y\|^2)$，得到

$$\langle\nabla_\theta\mathcal{L}(\theta_t, \lambda_t), \theta_{t+1} - \theta_t\rangle \ge \frac{1}{2\alpha}\|\tilde{\theta}_{t+1} - \theta_t\|^2 - \langle\delta_t, \theta_{t+1} - \tilde{\theta}_{t+1}\rangle \ge \frac{\alpha}{2}\|\mathcal{G}(\theta_t, \lambda_t)\|^2 - \alpha\|\delta_t\|^2. \quad (5)$$

再次根据 $L_{\mathcal{L}}$-光滑性，对于任意 $x,\ y \in \Theta$,

$$\mathcal{L}(x, \lambda_t) \ge \mathcal{L}(y, \lambda_t) + \langle\nabla_\theta\mathcal{L}(y, \lambda_t), x - y\rangle - \frac{L_{\mathcal{L}}}{2}\|x - y\|^2.$$

取 $x=\theta_{t+1}$ 和 $y=\theta_t$，我们有

$$\mathcal{L}(\theta_{t+1},\lambda_t) \geq \mathcal{L}(\theta_t,\lambda_t) + \langle \nabla_\theta \mathcal{L}(\theta_t,\lambda_t), \theta_{t+1} - \theta_t\rangle - \frac{L_{\mathcal{L}}}{2}\|\theta_{t+1}-\theta_t\|^2.$$

使用 (5) 和 $\|\theta_{t+1}-\theta_t\| \leq \alpha(\|\nabla_\theta \mathcal{L}(\theta_t,\lambda_t)\| + \|\delta_t\|)$，我们得到

$$\mathcal{L}(\theta_{t+1},\lambda_t) - \mathcal{L}(\theta_t,\lambda_t) \geq \frac{\alpha}{2}\|\mathcal{G}(\theta_t,\lambda_t)\|^2 - \alpha\|\delta_t\|^2 - L_{\mathcal{L}}\alpha^2\left(\|\nabla_\theta \mathcal{L}(\theta_t,\lambda_t)\|^2 + \|\delta_t\|^2\right).$$

由于 $\Theta \times [0,\Lambda]$ 是紧集且 $\nabla_\theta \mathcal{L}$ 连续，因此存在 $G_{\max} < \infty$ 使得 $\|\nabla_\theta \mathcal{L}\| \leq G_{\max}$。由此可得

$$\mathcal{L}(\theta_{t+1},\lambda_t) - \mathcal{L}(\theta_t,\lambda_t) \geq \frac{\alpha}{2}\|\mathcal{G}(\theta_t,\lambda_t)\|^2 - C_1\alpha\|\delta_t\|^2 - C_2\alpha^2,$$

其中 C_1、$C_2 > 0$ 是与 T 无关的常数。重新排列，并对从 $t=0$ 到 $T-1$ 的各项求和后取期望，得到

$$\frac{\alpha}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathcal{G}(\theta_t,\lambda_t)\|^2\right] \leq \mathbb{E}\left[\sum_{t=0}^{T-1}\left(\mathcal{L}(\theta_{t+1},\lambda_t) - \mathcal{L}(\theta_t,\lambda_t)\right)\right] + C_1\alpha\sum_{t=0}^{T-1}\mathbb{E}\|\delta_t\|^2 + C_2T\alpha^2.$$

此外，

$$\sum_{t=0}^{T-1}\left(\mathcal{L}(\theta_{t+1},\lambda_t) - \mathcal{L}(\theta_t,\lambda_t)\right) = \mathcal{L}(\theta_T,\lambda_{T-1}) - \mathcal{L}(\theta_0,\lambda_0) + \sum_{t=1}^{T-1}\left(\mathcal{L}(\theta_t,\lambda_{t-1}) - \mathcal{L}(\theta_t,\lambda_t)\right).$$

我们将真实对偶上升方向定义为 $u_t := J_C(\theta_t) - d$，得到相应的投影 $\tilde{\lambda}_{t+1} := \Pi_{[0,\Lambda]}(\lambda_t + \beta u_t)$，并回顾 $\mathcal{L}(\theta,\lambda) = J_R(\theta) - \lambda(J_C(\theta) - d)$，第二项变为 $\mathcal{L}(\theta_t,\lambda_{t-1}) - \mathcal{L}(\theta_t,\lambda_t) = (\lambda_t - \lambda_{t-1})u_t$。定义 $S_T := \sum_{t=1}^{T-1}(\lambda_t - \lambda_{t-1})u_t$，由此可得

$$S_T = \lambda_{T-1}u_{T-1} - \lambda_0 u_1 - \sum_{t=1}^{T-2}\lambda_t\left(u_{t+1} - u_t\right).$$

由于 $J_C(\theta)$ 在 Θ 上有界，因此存在 $U_{\max} < \infty$，使得对所有 t 都有 $|u_t| \leq U_{\max}$。因此，$|\lambda_{T-1}u_{T-1} - \lambda_0 u_1| \leq 2\Lambda U_{\max}$。由于 Θ 是紧的，且 $\nabla_\theta J_C(\theta)$ 连续，因此存在 $G_C < \infty$，使得 $|u_{t+1} - u_t| = |J_C(\theta_{t+1}) - J_C(\theta_t)| \leq G_C\|\theta_{t+1} - \theta_t\|$。因此，由柯西-施瓦茨不等式和詹森不等式可得

$$\left|\sum_{t=1}^{T-2}\lambda_t(u_{t+1}-u_t)\right| \leq \Lambda G_C\sum_{t=1}^{T-2}\|\theta_{t+1}-\theta_t\|, \leq \sqrt{T}\Lambda G_C\left(\sum_{t=1}^{T-2}\|\theta_{t+1}-\theta_t\|^2\right)^{1/2}.$$

由于 $\|\theta_{t+1}-\theta_t\| \leq \alpha(\|\nabla_\theta \mathcal{L}(\theta_t,\lambda_t)\| + \|\delta_t\|)$，因此有 $\mathbb{E}\left[\|\theta_{t+1}-\theta_t\|^2\right] \leq 2\alpha^2\left(G_{\max}^2 + \sigma_\theta^2 + a^2\right)$。结合上述界可得

$$E[|S_T|] \leq 2\Lambda U_{\max} + \alpha T\Lambda G_C\sqrt{2(G_{\max}^2 + \sigma_\theta^2 + a^2)}.$$

Combining $\mathbb{E}[\mathcal{L}(\theta_T,\lambda_{T-1}) - \mathcal{L}(\theta_0,\lambda_0)] \leq \Delta_{\mathcal{L}}$, we obtain

$$\frac{\alpha}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathcal{G}(\theta_t,\lambda_t)\|^2] \leq \Delta_{\mathcal{L}} + C_1\alpha T(\sigma_\theta^2 + a^2) + C_2T\alpha^2 + 2\Lambda U_{\max} + \alpha T\Lambda G_C\sqrt{2(G_{\max}^2 + \sigma_\theta^2 + a^2)}.$$

因此，

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathcal{G}(\theta_t,\lambda_t)\|^2] \leq \frac{2(\Delta_{\mathcal{L}} + 2\Lambda U_{\max})}{\alpha T} + 2C_1(\sigma_\theta^2 + a^2) + 2C_2\alpha + 2\Lambda G_C\sqrt{2(G_{\max}^2 + \sigma_\theta^2 + a^2)}.$$

由詹森不等式和基本界，我们得到该结果。由于 $\lambda_{t+1} = \Pi_{[0,\Lambda]}(\lambda_t + \beta u_t)$ 且 $\Pi_{[0,\Lambda]}(\lambda_t + \beta\hat{u}_t)$，有 $\tilde{\lambda}_{t+1} - \lambda_t = \beta\mathcal{H}(\theta_t,\lambda_t)$。令 $\varepsilon_t := \hat{u}_t - u_t$ 表示对偶估计误差。对于任意固定的θ_t，有

$$\begin{aligned}&\mathcal{L}(\theta_t,\lambda_{t+1}) - \mathcal{L}(\theta_t,\lambda_t) \\ &= -(\tilde{\lambda}_{t+1} - \lambda_t)u_t - (\lambda_{t+1} - \tilde{\lambda}_{t+1})u_t.\end{aligned} \tag{6}$$

由于对所有 $\lambda \in [0,\Lambda]$，$\langle \lambda_{t+1} - (\lambda_t + \beta u_t)\ \lambda - \lambda_{t+1}\rangle \geq 0$，取 $\lambda = \lambda_t$ 可得

$$\begin{aligned}(\tilde{\lambda}_{t+1} - \lambda_t)u_t &\geq \frac{1}{\beta}(\tilde{\lambda}_{t+1} - \lambda_t)^2, \\ &\geq \beta\|\mathcal{H}(\theta_t,\lambda_t)\|^2.\end{aligned} \tag{7}$$

由于

$$\begin{aligned}&|\lambda_{t+1} - \tilde{\lambda}_{t+1}| \\ &= |\Pi_{[0,\Lambda]}(\lambda_t + \beta(u_t + \varepsilon_t)) - \Pi_{[0,\Lambda]}(\lambda_t + \beta u_t)| \leq \beta|\varepsilon_t|,\end{aligned}$$

由此得到

$$-(\lambda_{t+1} - \tilde{\lambda}_{t+1})u_t \leq \beta|\varepsilon_t||u_t|. \tag{8}$$

结合式（6）、（7）和（8），可得

$$\begin{aligned}\mathcal{L}(\theta_t,\lambda_{t+1}) - \mathcal{L}(\theta_t,\lambda_t) &\leq -\beta\|\mathcal{H}(\theta_t,\lambda_t)\|^2 + \beta|\varepsilon_t||u_t|, \\ &\leq -\beta\|\mathcal{H}(\theta_t,\lambda_t)\|^2 + \frac{\beta}{2}\varepsilon_t^2 + \frac{\beta}{2}u_t^2.\end{aligned}$$

此外，$|\lambda_{t+1} - \lambda_t| \leq \beta(|u_t| + |\varepsilon_t|)$且

$$-(\lambda_{t+1}-\lambda_t)u_{t+1} = -(\lambda_{t+1}-\lambda_t)u_t - (\lambda_{t+1}-\lambda_t)(u_{t+1}-u_t)$$

对所有 t均成立。因此，应用杨氏不等式可得

$$\begin{aligned}&|(\lambda_{t+1} - \lambda_t)(u_{t+1} - u_t)| \leq \beta(|u_t| + |\varepsilon_t|)G_C\|\theta_{t+1} - \theta_t\|, \\ &\leq \frac{\beta}{4}(u_t^2 + \varepsilon_t^2) + 2\beta G_C^2\|\theta_{t+1} - \theta_t\|^2.\end{aligned}$$

根据假设2，$\mathbb{E}[\varepsilon_t^2|\mathcal{F}_t] \leq \sigma_\lambda^2 + b^2$。将所有项合并，并使用

$$\mathcal{L}(\theta_{t+1},\lambda_{t+1}) - \mathcal{L}(\theta_{t+1},\lambda_t) = -(\lambda_{t+1} - \lambda_t)u_{t+1},$$

我们得到

$$\begin{aligned}&\mathbb{E}[\mathcal{L}(\theta_{t+1},\lambda_{t+1}) - \mathcal{L}(\theta_{t+1},\lambda_t)] \\ &\leq -\beta\mathbb{E}[\|\mathcal{H}(\theta_t,\lambda_t)\|^2] + \beta C_3(U_{\max}^2 + \sigma_\lambda^2 + b^2) + \beta C_4\alpha^2,\end{aligned}$$

其中 C_3、C_4 是与 T 无关的常数。将该分解从 $t=0$ 到 $T-1$ 求和可得

$$\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{L}(\theta_{t+1},\lambda_{t+1})-\mathcal{L}(\theta_t,\lambda_t)] = \mathbb{E}[\mathcal{L}(\theta_T,\lambda_T)-\mathcal{L}(\theta_0,\lambda_0)] \geq -\Delta_{\mathcal{L}}.$$

由于

$$\begin{aligned}&(\mathcal{L}(\theta_{t+1},\lambda_{t+1})-\mathcal{L}(\theta_t,\lambda_t))\\ =&\underbrace{(\mathcal{L}(\theta_{t+1},\lambda_t)-\mathcal{L}(\theta_t,\lambda_t))}_{:=A_t}+\underbrace{(\mathcal{L}(\theta_{t+1},\lambda_{t+1})-\mathcal{L}(\theta_{t+1},\lambda_t))}_{:=B_t},\end{aligned}$$

将其对 $t=0,\ldots,T-1$ 求和可得

$$\sum_{t=0}^{T-1}\mathbb{E}[B_t]\geq-\Delta_{\mathcal{L}}-\sum_{t=0}^{T-1}\mathbb{E}[A_t].$$

根据 $L_{\mathcal{L}}$-光滑性，我们有

$$A_t\leq G_{\max}\|\theta_{t+1}-\theta_t\|+\frac{L_{\mathcal{L}}}{2}\|\theta_{t+1}-\theta_t\|^2.$$

取期望并应用杨氏不等式可得 $\mathbb{E}[A_t]\leq \frac{G^2{}_{\max}}{2}+\frac{1+L_{\mathcal{L}}}{2}\mathbb{E}\|\theta_{t+1}-\theta_t\|^2$。因此，

$$\sum_{t=0}^{T-1}\mathbb{E}[A_t]\leq T\frac{G_{\max}^2}{2}+\frac{1+L_{\mathcal{L}}}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\theta_{t+1}-\theta_t\|^2.$$

利用 $\mathbb{E}\left[\|\theta_{t+1}-\theta_t\|^2\right]\leq 2\alpha^2\left(G_{\max}^2+\sigma_\theta^2+a^2\right)$，我们有

$$\sum_{t=0}^{T-1}\mathbb{E}[A_t]\leq\frac{G_{\max}^2}{2}T+(1+L_{\mathcal{L}})T\alpha^2\left(G_{\max}^2+\sigma_\theta^2+a^2\right).$$

合并各项并除以 βT，可得

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathcal{H}(\theta_t,\lambda_t)\|^2]\\ \leq&\frac{\Delta_{\mathcal{L}}}{\beta T}+C_3(U_{\max}^2+\sigma_\lambda^2+b^2)+C_4\alpha^2+\frac{G_{\max}^2}{2\beta}\\ &+\frac{\alpha^2}{\beta}\left((1+L_{\mathcal{L}})(G_{\max}^2+\sigma_\theta^2+a^2)\right).\end{aligned}$$

最后，应用詹森不等式和基本界限，证明得证。□

八、结论

本文提出了 PPO-LTL框架，通过拉格朗日方案将以线性时序逻辑表示的安全规范融入近端策略优化，从而为编码复杂的安全要求（如监管规则）提供了一种精确方法。本文还给出了收敛性的理论保证。在 Zones 和 CARLA 环境中进行的实验充分验证了我们的方法。

参考文献

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot 等，“利用深度神经网络和树搜索掌握围棋”，《自然》，第 529 卷，489，2016 年，第 7587 期，484 页。[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] J. Garc´ıa 和 F. Fern´andez，“安全强化学习综合综述”，《机器学习研究期刊》，第 16 卷，第 42 期，第 1437–1480 页，2015 年。[4] A. Ray, J. Achiam 和 D. Amodei，“深度强化学习中安全探索的基准测试”，载于 NeurIPS 安全机器学习研讨会，2019 年。[5] C. Tennant, C. Neels, G. Parkhurst, P. Jones, S. Mirza 和 J. Stilgoe，“代码、文化与混凝土：自动驾驶车辆与道路规则”，《可持续城市前沿》，第 3 卷，第 710478 页，2021 年。[6] A. Pnueli，“程序的时序逻辑”，第 18 届计算机科学基础年度研讨会（SFCS 1977），第 46–57 页，1977 年。[7] C. Baier 和 J.-P. Katoen，《模型检测原理》。MIT 出版社，2008 年。[8] M. Y. Vardi，“重新审视自动机理论模型检测”，载于《验证：理论与实践》。Springer，1996 年，第 137–150 页。[9] S. Sickert, J. Esparza, B. Finkbeiner 和 M. Leucker，“线性时序逻辑的极限确定性 b¨uchi 自动机”，载于《计算机辅助验证》。Springer，2016 年，第 312–332 页。[10] J. Ji, J. Zhou, B. Zhang, J. Dai, X. Pan, R. Sun, W. Huang, Y. Geng, M. Liu 和 Y. Yang，“OmniSafe：加速安全强化学习研究的基础设施”，《机器学习研究期刊》，第 25 卷，第 285 期，第 1–6 页，2024 年。[11] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez 和 V. Koltun，“CARLA：一个开放的城市驾驶模拟器”，载于《机器人学习会议》。PMLR，2017 年，第 1–16 页。[12] T. Haarnoja, A. Zhou, P. Abbeel 和 S. Levine，“软演员-评论家：使用随机演员的离策略最大熵深度强化学习”，载于国际机器学习会议。PMLR，2018 年，第 1861–1870 页。[13] M. Alshiekh, R. Bloem, R. Ehlers, B. K¨onighofer, S. Niekum 和 U. Topcu，“通过安全屏蔽实现安全强化学习”，载于 AAAI 人工智能会议论文集，第 32 卷，第 1 期，2018 年。[14] N. Jansen, B. K¨onighofer, S. Junges, A. Serban 和 R. Bloem，“使用概率屏蔽器实现安全强化学习”，载于第 31 届并发理论国际会议（CONCUR 2020）。Schloss Dagstuhl–Leibniz 信息学中心 f¨ur，2020 年，第 3–1 页。[15] E. Altman，《约束马尔可夫决策过程》。CRC 出版社，2021 年。[16] A. Camacho, R. T. Icarte, T. Q. Klassen, R. A. Valenzano 和 S. A. McIlraith，“LTL 及其他：强化学习中奖励函数规范的形式语言”，载于 IJCAI，第 19 卷，2019 年，第 6065–6073 页。[17] O. Maler 和 D. Nickovic，“连续信号时序性质监测”，载于实时与容错系统形式化技术国际研讨会。Springer，2004 年，第 152–166 页。[18] P. Y. Chan, X. Li, Y. Lu, Y. Lin 和 A. Bundy，“运用右开时序义务可废止逻辑形式化自动驾驶车辆法规”，载于创新人工智能技术与应用国际会议。Springer，2024 年，第 194–207 页。[19] L. De Raedt 和 A. Kimmig，“概率（逻辑）编程概念”，《机器学习》，第 100 卷，第 1 期，第 5–47 页，2015 年。[20] J. Chen, S. E. Li 和 M. Tomizuka，“利用潜在深度强化学习实现可解释的端到端城市自动驾驶”，《IEEE 智能交通系统汇刊》，第 23 卷，第 6 期，第 5068–5078 页，2021 年。[21] S. Qi, Z. Zhang, Z. Sun 和 S. Haesaert，“ 基于线性时序逻辑规范的风险感知自动驾驶”，载于 2025 年 IEEE/RSJ 智能机器人与系统国际会议（IROS）。IEEE，2025 年，第 14 877–14 883 页。[22] R. S. Sutton 和 A. G. Barto，《强化学习导论》，第 2 版。马萨诸塞州剑桥：MIT 出版社，2018 年。 [在线]。网址：http://incompleteideas.net/book/the-book-2nd.html [23] Y. Li, F. He, M. Xue 等，“时序逻辑引导的安全强化学习”，载于第 37 届 AAAI 人工智能会议论文集，2023 年。

[24] P. Vaezipoor, A. C. Li, R. A. T. Icarte 和 S. A. Mcilraith，“Ltl2action：面向多任务强化学习的线性时序逻辑指令泛化”，载于国际机器学习会议。PMLR，2021 年，第 10 497–10 508 页。[25] M. Jackermeier 和 A. Abate，“Deepltl：学习高效满足复杂线性时序逻辑规范，以支持多任务强化学习”，载于第十三届国际学习表征会议，2025 年。[26] J. Baxter 和 P. L. Bartlett，“无限时域策略梯度估计”，人工智能研究期刊，第 15 卷，350，2001 年，第 319 页。 – 。[27] J. Achiam, D. Held, A. Tamar 和 P. Abbeel，“约束策略优化”，载于国际机器学习会议。PMLR，2017 年，第 22–31 页。[28] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus 和 N. Dormann，“Stable-baselines3：可靠的强化学习实现”，载于机器学习研究期刊，第 22 卷，2021 年，第 1–8 页。