

# Multi-Agent Reinforcement Learning Guided by Signal Temporal Logic Specifications

Jiangwei Wang<sup>1</sup>, Shuo Yang<sup>2</sup>, Ziyan An<sup>3</sup>, Songyang Han<sup>1</sup>, Zhili Zhang<sup>1</sup>,  
Rahul Mangharam<sup>2</sup>, Meiyi Ma<sup>3</sup>, Fei Miao<sup>1</sup>

**Abstract**— Reward design is a key component of deep reinforcement learning (DRL), yet some tasks and designer’s objectives may be unnatural to define as a scalar cost function. Among the various techniques, formal methods integrated with DRL have garnered considerable attention due to their expressiveness and flexibility to define the reward and requirements for different states and actions of the agent. However, how to leverage Signal Temporal Logic (STL) to guide multi-agent reinforcement learning (MARL) reward design remains unexplored. Complex interactions, heterogeneous goals and critical safety requirements in multi-agent systems make this problem even more challenging. In this paper, we propose a novel STL-guided multi-agent reinforcement learning framework. The STL requirements are designed to include both task specifications according to the objective of each agent and safety specifications, and the robustness values of the STL specifications are leveraged to generate rewards. We validate the advantages of our method through empirical studies. The experimental results demonstrate significant reward performance improvements compared to MARL without STL guidance, along with a remarkable increase in the overall safety rate of the multi-agent systems.

## I. INTRODUCTION

Multi-agent reinforcement learning (MARL) have gained significant research interests in solving various sequential decision-making problems for multi-agent systems, especially with the rapid development of deep reinforcement learning (DRL). It is a key component of an MARL algorithm to define a reward function that maps each state and action to some real-valued reward [1]. However, define or encode a scalar reward function according to the desired behavior and objectives considering the dynamic interactions and typically heterogeneous goals of a multi-agent system remains challenging. Moreover, for MARL that involves both the interactions of the agents and the physical dynamic process of each individual agent, poorly designed reward functions can lead to undesired policies that are unable to accomplish the tasks, or worse, execute unsafe actions in safety-critical systems [2], [3].

A motivating example of decision-making for multi-agent system named *Traffic-jam* is shown in Fig. 1. In this case, red broken-down vehicles stopped on streets and blocked three lanes (due to an accident or other reasons), the three

blue autonomous vehicles want to drive forward and pass through the only open lane (right lane in the figure) as soon as possible while keeping a safety distance between each other and the broken-down vehicles. Additionally, the time length which autonomous vehicles remain blocked after the broken-down vehicles should be constrained below a certain threshold. To make the vehicles learn to fulfill all these requirements, designing a reward function may take lots of trials and can be computationally expensive. Regarding the safety requirements in reinforcement learning [4], [5], [6], using penalty in reward to discourage the unsafe actions, or framing the problem into a constrained optimization problem may not provide sufficient safety assurances for the selected actions. Furthermore, it is even more challenging to encode a reward function considering temporal requirements for the agents [7], [8].

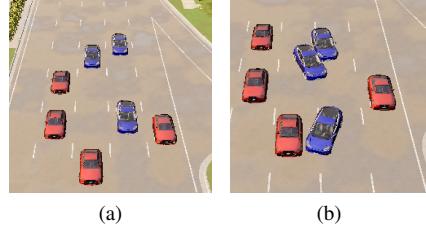


Fig. 1: *Traffic-jam* in CARLA simulator. a: Scenario initialization. b: Agents trained based on MARL with hand-engineered reward fail to cross the open lane in a timely manner and collide with other agents.

To address this challenge, we propose a safe MARL algorithm that taking full advantages of signal temporal logic (STL), which, as a formal language provides a more principled and expressive way to describe the requirements. We adopt the robustness values based on the STL requirements as rewards in our proposed STL-guided MARL algorithm. We also guarantee the satisfaction of hard safety requirements through an STL safety shield. While there are extensive works exploring the use of temporal logic in single-agent RL, very little attention has been drawn on how to leverage signal temporal logic (STL) to guide MARL [9], [10]. Our proposed algorithm shows promising results in learning a better policy for each agent to reach its objective and ensure the safety of the system. Our contributions are summarized as follows:

- We design a multi-agent reinforcement learning algorithm that is guided by STL specifications, where the STL specifications include both safety requirements and the task that each agent aims to finish. We incorporate

<sup>1</sup>Jiangwei Wang, Songyang Han, Zhili Zhang and Fei Miao are with University of Connecticut {jiangwei.wang, songyang.han, zhili.zhang, fei.miao}@uconn.edu

<sup>2</sup>Shuo Yang and Rahul Mangharam are with University of Pennsylvania {yangs1, rahulm}@seas.upenn.edu

<sup>3</sup>Ziyan An and Meiyi Ma are with Vanderbilt University {ziyan.an, meiyi.ma}@vanderbilt.edu

the STL safety shield to fulfill the safety specifications and provide safe actions for the agents.

- The proposed algorithm utilizes STL to check partial trajectories and provide robustness values as a corresponding reward during the training process.
- We validate the proposed methodology in multi-agent particle-world environment (MPE) and CARLA testbeds. We demonstrate that compared with the baseline MARL algorithms and commonly used rewards, our proposed algorithm can learn better policies with both larger rewards and higher safety rates.

## II. RELATED WORKS

### A. Multi-Agent Reinforcement Learning

There has been growing interest in the study of MARL since many real-world problems involve the interactions of multiple agents [11]. MARL approaches have been proposed for various multi-agent systems, such as unmanned aerial vehicles [12], [13], complex traffic networks [14], autonomous driving [15], and so on. Despite these successful applications, one remaining challenge in MARL is how to design good reward functions for complex tasks. Poorly-designed reward functions might lead to undesired behavior and be detrimental to safety-critical systems [16]. The complex interactions among agents and their diverse objectives make the reward design hard.

### B. Temporal Logic for Reinforcement Learning

RL reward function usually relies on hand-engineered design or approaches like reward shaping [17]. In recent years, temporal logic specifications have been used extensively as training guidance in the context of single agent reinforcement learning for its power of expressiveness. In one direction, finite state automaton (FSA) is constructed to reward the agent (e.g., [18], [19], [20]) with the benefit of easy reward-generating automaton and high interpretability. In another direction, the quantitative semantics of temporal logic formulas are captured to guide the policy training [9], [21]. Simply applying the STL-guided single agent reinforcement learning in multi-agent setting is not a good solution because they don't consider the complex interactions between the agents and their safety requirements, which is usually the case in real world systems.

In the context of MARL, very few works have been done to satisfy temporal logic specifications [10], [22], [23]. For example, [10] proposes the first MARL algorithm for temporal logic specifications with correctness and convergence guarantees. However, it uses the LTL specifically designed to satisfy the non-Markovian, infinite-horizon specifications, which may not be applicable in the real world. Also, it has not been empirically verified in MARL environments. [23] proposes an extended Markov Games as a general mathematical model that allows multiple RL agents to concurrently learn various LTL specifications. In our work, we consider STL specifications and use their robustness values as the rewards instead. Compared with LTL, STL preserves quantitative semantics that can be used to establish a robust

satisfaction value to quantify how well a trajectory fulfills a specification. This can further allow us to quantify the rewards more precisely and less sparser compared with LTL based approaches.

## III. PRELIMINARY AND PROBLEM FORMULATION

### A. Signal Temporal Logic

In this section, we introduce the syntax, semantics, and robustness metric of STL, which is a powerful formal symbolism for specifying temporal logical requirements [24]. We first define a signal  $\omega = s^0 s^1 \dots s^t$  as the state trajectory from starting state to time point  $t$ .

*Definition 3.1:* The syntax of STL is defined by:

$$\varphi ::= \mu \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \Diamond_I \varphi \mid \Box_I \varphi \mid \varphi_1 \mathcal{U}_I \varphi_2$$

where  $I = [a, b]$ ,  $a, b \in \mathbb{R}^{>0}$  denotes a bounded time interval. The atomic predicate  $\mu$  represents the underlying function  $\mu(\omega^t) \geq 0$ , where  $\omega^t$  is the signal value at time  $t$ . We use the symbols  $\Box$ ,  $\Diamond$ , and  $\mathcal{U}$  to denote temporal operators *always*, *eventually*, and *until*. The satisfaction relation  $(\omega, t) \models \varphi$  evaluates to true ( $\top$ ) if the specification  $\varphi$  is satisfied by  $\omega$  starting from  $t$  and false ( $\perp$ ) otherwise. In addition to its Boolean semantics, STL also owns the quantitative semantics (i.e. *robust satisfaction values*), which quantify the degree of satisfaction [24], [25], [26]. The quantitative semantics assign real-valued measurements to the satisfaction (positive values) or violation (negative values) of the STL formula. In the evaluation section of this work, we utilize the robustness values to measure specification satisfactions.

*Definition 3.2:* STL quantitative semantics are defined in Table I.

TABLE I: STL quantitative semantics

Formula	Semantics
$\rho(x \sim c, \varphi, t)$	$f(\omega[t]) - c$
$\rho(\neg\varphi, \omega, t)$	$-\rho(\varphi, \omega, t)$
$\rho(\varphi_1 \wedge \varphi_2, \omega, t)$	$\min\{\rho(\varphi_1, \omega, t), \rho(\varphi_2, \omega, t)\}$
$\rho(\Box_I \varphi, \omega, t)$	$\min_{t' \in (t, t+I)} \rho(\varphi, \omega, t')$
$\rho(\Diamond_I \varphi, \omega, t)$	$\max_{t' \in (t, t+I)} \rho(\varphi, \omega, t')$
$\rho(\varphi_1 \mathcal{U}_I \varphi_2, \omega, t)$	$\sup_{t' \in (t+I) \cap \mathbb{T}} (\min\{\rho(\varphi_2, \omega, t'), \inf_{t'' \in [t, t']} (\rho(\varphi_1, \omega, t''))\})$

### B. Problem Formulation of STL-guided MARL

**STL-guided MARL:** In this work, we define an STL-guided MARL for multi-agent decision-making problem such as the example shown in Fig. 1, to address the challenge of designing a reward function that utilizes the strength of STL. In particular, we define a tuple  $G = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \{r_i\}, \gamma, \{\varphi_i\}, \omega^{t-L+1:t})$ , where  $\mathcal{S}$  is the joint state space,  $\mathcal{A}$  is the joint action space,  $o_i = \mathcal{O}(s; i)$  is the local observation for agent  $i$  at global state  $s$ .  $\mathcal{T}$  corresponds to the state transition function as defined for Markov games in the literature [27]. The key new components of the tuple definition include:  $\omega^{t-L+1:t} = s^{t-L+1} \dots s^t$  being the partial state trajectory of length  $L$ , and  $\varphi_i$  being the

<sup>1</sup>To increase readability, we will omit truncation time indices  $(t-L+1 : t)$ , i.e., we will use  $\omega$  instead of  $\omega^{t-L+1:t}$  to denote the partial trajectory with a slight notation abuse.

STL formula for agent  $i$ . Take the *Traffic-jam* scenario as an example, the STL formula  $\varphi_i$  is the aggregation of several STL requirements, including reaching the destination, keeping safe distance to other agents, and waiting no more than  $T_{max}$  time steps after the broken-down vehicles. The detail of the STL formula for different tasks will be introduced in Section V. In the tuple  $G$ , reward  $r_i = \rho(\varphi_i, \omega, t)$  represents the STL robustness value, and each agent  $i$  aims to maximize its own total expected return  $R_i = \sum_{t=0}^T \gamma^t r_i^t$  where  $\gamma$  is a discount factor and  $T$  is the time horizon.

#### IV. METHODOLOGY

In this section, we first introduce the algorithm structure of STL-guided MARL algorithm. Our framework is shown in Fig. 2.

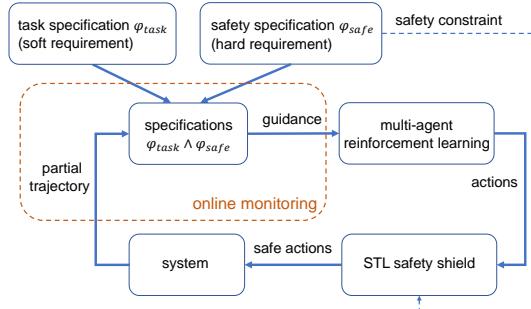


Fig. 2: Methodology overview. The user-provided task specification and safety specification are expressed in STL formula  $\varphi_{task}$  and  $\varphi_{safe}$ , respectively. The robustness values of the (partial) trajectory/signal w.r.t.  $\varphi_{task}$  and  $\varphi_{safe}$  are used to generate reward and guide the MARL policy learning. The STL safety shield, which is constructed based on safety specification  $\varphi_{safe}$ , is involved to safeguard the decisions made by MARL.

##### A. Reward Function

To address the challenge of defining a reward function for a multi-agent system that considers the objective of each individual agent and the complex interactions among agents, in this work, we check the partial trace based on the STL specifications, and provide the robustness value as the reward. It's notable that our method is generalizable to different MARL algorithms, e.g., MADDPG [28] and MAPPO [29].

At each time step of the training process, the trajectory contains the states of all the agents. Given the partial trajectory and STL specification  $\varphi_i$ , which might consist of both safety specification  $\varphi_{i,safe}$  and task specification  $\varphi_{i,task}$ , the reward for agent  $i$  can be defined as

$$r_i^t = \rho(\varphi_i, \omega, t), \forall t \in [\tau, \tau + L - 1], \quad (1)$$

where  $\rho$  is the robustness value. Given the partial trajectory, the more it satisfies the STL specifications, the larger reward it obtains. One reason we use partial trajectory robustness value rather full trajectory is that the former better discriminates different states and actions, while it's hard for latter to assign the credit to specific actions, and estimating the value function also becomes more challenging with longer future horizons. From MARL algorithm design perspective, n-step methods allow for credit assignment over a longer time

horizon such as in the A3C algorithm [30]. This enables the agent to better understand the consequences of its actions. Hence, in this work, we define the reward based on the robustness value given the partial trajectory  $\omega$ . It should also be noted that the STL requirements encompass various types, including reaching the goal, safety requirements like maintaining a safe distance from other agents, and other temporal requirements. Detailed STL requirements for task specification and safety specification are illustrated by a case study in Subsection IV-C.1.

---

##### Algorithm 1: Pseudocode for STL-guided MARL

---

```

1 Orthogonal initialization for  $\theta_i$  and  $\phi_i$ , the
   parameters for policy  $\pi_i$  and critic  $V_i$ , respectively;
2 for episode = 1 to  $M$  do
3   Episode initialization: replay buffer  $\mathcal{D} \leftarrow \emptyset$ ; initial
     state  $s$ ; step  $t \leftarrow 1$ ; rollout step number  $L$ ;
4   while  $t \leq t_{max}$  do
5      $t_{start} = t$ ;
6     Initialize an empty trajectory  $\omega$ ;
7     for  $t = t_{start}$  to  $t_{start} + L - 1$  do
8       For each agent  $i$ , select action
          $a_{i_q}^t \leftarrow \pi_{\theta_i}(o_i^t)$ , send  $a_{i_q}^t$  into STL safety
         shield, and send back  $a_i^t$ ;
9       Execute actions  $a^t = (a_1^t, \dots, a_N^t)$  chosen
         based on STL safety shield, append  $s^t$  to
          $\omega$ , observe reward  $r_i^t = \rho(\varphi_i, \omega, t)$  for
         each agent  $i$  given agent  $i$ 's STL formula
          $\varphi_i$ , observe new state  $s^{t+1}$ ;
10       $s^t \leftarrow s^{t+1}$ ;  $t \leftarrow t + 1$ ;
11    end
12  end
13  Store in  $\mathcal{D}$ 
14   $\{(s^\tau, a^\tau, r^\tau, s^{\tau+1}) \mid \tau \in [t_{start}, \dots, t_{start}+L-1]\}$ ;
15  for agent  $i = 1$  to  $N$  do
16    Randomly sample mini-batch from  $\mathcal{D}$ 
     $\{(s^\tau, a^\tau, r^\tau, s^{\tau+1}) \mid \tau \in [t, \dots, t+L-1]\}$ ;
17    Update critic by minimizing  $\mathcal{L}^{VF}(\phi_i)$ ;
18    Update actor by maximizing  $\mathcal{L}(\theta_i)$ ;
19  end
20 end

```

---

##### B. Algorithm Structure

In our work, we adopt the centralized training and decentralized execution paradigm. As depicted in Alg. 1, during training, in rollout time steps  $t$ , each agent selects potentially unsafe action  $a_{i_q}^t$  based on its local observations  $o_i^t$ . This action is then passed to the STL safety shield layer, and safe action  $a_i^t$  is returned and deployed such that the satisfaction of the STL safety specifications is guaranteed. The details of the STL safety shield layer will be presented in the next section. The global state  $s^t = (o_1^t, \dots, o_N^t)$ , representing the aggregation of the local observations, is appended to the previous states to form the partial trajectory  $\omega$ . By evaluating the partial trajectory against the STL requirements, each agent obtains its reward  $r_i^t$ .

The actor network is trained to maximize the following

objective:

$$\begin{aligned}\mathcal{L}(\theta_i) &= \frac{1}{B} \sum_{k=1}^B \min(r_{\theta_i,k} A_{i,k}, \text{clip}(r_{\theta,k}, 1 - \epsilon, 1 + \epsilon) A_{i,k}) \\ &+ \sigma \frac{1}{B} \sum_{k=1}^B S[\pi_{\theta_i}(o_{i,k})],\end{aligned}\quad (2)$$

where  $B$  is the batch size,  $r_{\theta_i,k} = \frac{\pi_{\theta_i}(a_{i,k}|o_{i,k})}{\pi_{\theta_{i,old}}(a_{i,k}|o_{i,k})}$  denotes the ratio of the probability under the new and old policies respectively,  $A_{i,k}$  is the advantage computed by GAE method [31],  $S$  is the policy entropy, and  $\sigma$  is the entropy coefficient hyperparameter. For the critic network, the loss function is:

$$\begin{aligned}\mathcal{L}^{\text{VF}}(\phi_i) &= \frac{1}{B} \sum_{k=1}^B \max[(V_{\phi_i}(s_{i,k}) - \hat{R}_{i,k})^2, \\ &\text{clip}(V_{\phi_i}(s_{i,k}), V_{\phi_{i,old}}(s_{i,k}) - \epsilon, V_{\phi_{i,old}}(s_{i,k}) + \epsilon) - \hat{R}_{i,k})^2],\end{aligned}\quad (3)$$

where  $\hat{R}_{i,k}$  is the discounted reward-to-go. For actor and critic network, we use the recurrent neural network (RNN) to take the input to enable the agents to effectively model and reason about sequential information in their interactions with the environment. By maintaining hidden states and updating them at each time step, RNN can capture the temporal dynamics and dependencies across multiple time steps.

### C. STL Safety Shield

In multi-agent systems with complex dynamics, such as the *Traffic-jam* scenario depicted in Fig. 1, ensuring system safety specification becomes paramount. One critical aspect of safety is maintaining a safe distance between agents. Consequently, we incorporate safety requirements into the Signal Temporal Logic (STL) formulas  $\varphi_{safe}$ , which will also be elaborated in Section V. To satisfy these safety requirements specified in the STL formulas, such as “the distance between agents should always be greater than a threshold”, we first convert them to CBFs. Then, we employ the quadratic programming (CBF-QP) to ensure safety. By leveraging CBF-QP, we assess whether each discrete action guarantees system safety and filter out any unsafe actions accordingly.

1) *Case Study*: We use *Traffic-jam* scenario to explain the details of STL safety shield design and the STL specifications. As shown in Fig. 1, in the *Traffic-jam* scenario, red broken-down vehicles block three lanes, a group of autonomous vehicles aim to cross the narrow road and arrive their destination as soon as possible, while maintaining the safety of the whole system. Agent  $i$ ’s observation at time  $t$  include (1) its own locations, velocities, accelerations, orientation ( $\mathbf{p}_i^t, \mathbf{v}_i^t, \mathbf{a}_i^t, \psi_i^t$ ); (2) other agent  $j$ ’s shared information ( $\mathbf{p}_j^t, \mathbf{v}_j^t, \mathbf{a}_j^t, \psi_j^t$ ),  $\forall j \in N$ , (3) its destination  $\mathbf{p}_{i,\text{dest}}$ . Agent  $i$ ’s discrete action space include:  $a_{i,1}$ : keep speed and keep in current lane;  $a_{i,2}$ : change to left lane;  $a_{i,3}$ : change to right lane;  $a_{i,4}$ : brake;  $a_{i,5} \sim a_{i,4+l}$ :  $l$  different throttle values, representing  $l$  levels of acceleration and deceleration in the current lane.

We have the following requirements for each agent  $i$ : (1) (safety) distance between itself and the leading vehicle

in its current lane and neighboring lanes should be always greater than a safe distance; (2) (task) eventually reach its destination; (3) (task) it should stop in front of the narrow road location  $\mathbf{p}_{road}$ ; (4) (task) its blocked duration  $t_{wait}$  in front of narrow road location  $\mathbf{p}_{road}$  should be less than  $T_{\max}$ . These specifications can be easily converted to STL formulas accordingly:

$$\begin{cases} \varphi_{i_1} = \square_{[0,T-1]} \|\mathbf{p}_i^t - \mathbf{p}_j^t\| - \frac{(v_j^t - v_i^t)^2}{2a_{i_1}} \geq \epsilon_1, \forall j \in N, j \neq i \\ \varphi_{i_2} = \diamondsuit_{[0,T-1]} \|\mathbf{p}_i^t - \mathbf{p}_{i,\text{dest}}\| \leq \epsilon_2 \\ \varphi_{i_3} = \square_{[0,T-1]} (\neg(\|\mathbf{p}_i^t - \mathbf{p}_{road}\| \leq L) \vee (\diamondsuit_{[0,\tau]} v_i^t \leq 0)) \\ \varphi_{i_4} = \square_{[0,T-1]} (\neg(\|\mathbf{p}_i^t - \mathbf{p}_{road}\| \leq L) \vee (t_{\text{wait}} < T_{\max})) \end{cases} \quad (4)$$

*CBF for safety*: Here we show how to define the barrier functions to fulfill the safety requirements in STL specifications. Notably, there is exiting work summarizes how to design CBF given different STL predicates generally [32]. We model the low-level agent dynamic as a nonlinear control affine system:  $\dot{\mathbf{x}}^{t+1} = f(\mathbf{x}^t) + g(\mathbf{x}^t)\mathbf{u}^t$ , where  $\mathbf{x}^t \in \mathbb{R}^n$ ,  $\mathbf{u}^t \in \mathcal{U} \subseteq \mathbb{R}^m$  denoting the set of permissible control inputs,  $f$  is the nominal unactuated dynamics, and  $g$  is the nominal actuated dynamics. We adopt the widely used kinematic bicycle model for its simplicity while still considering the non-holonomic vehicle behaviors [33].

As shown in Fig. 3, during the lane keeping mode, the ego vehicle should keep a safe distance to the vehicle in the front in its current lane; while it’s changing the lane, it should keep a safe distance to both the vehicles in its front and back. We use  $fv$  and  $bv$  to denote the front and back vehicles in the target lane respectively. The safe distance to the front vehicle can be expressed as  $D_{fv} = (1 + \epsilon)v^t + \frac{(v^t - v_{fv}^t)^2}{2a_l}$ , and the safe distance to the back vehicle is  $D_{bv} = (1 + \epsilon)v^t + \frac{(v_{bv}^t - v^t)^2}{2a_l}$ . Note that  $a_l$  and  $v^t$  are the ego vehicle’s acceleration limit and current speed respectively, and  $v_{fv}^t, v_{bv}^t$  denote the velocity of vehicle in the front and back respectively. Finally, the CBFs can be expressed as:  $h_{fv}(\mathbf{x}^t, t) = (x_{fv}^t - x^t) - D_{fv}(v^t, v_{fv}^t)$  and  $h_{bv}(\mathbf{x}^t, t) = (x^t - x_{bv}^t) - D_{bv}(v^t, v_{bv}^t)$ .

For each discrete action  $a_{i_q}^t$ , after being mapped to the corresponding continuous control input  $\mathbf{u}_{i_q}^t$  [34], [35], then the following CBF-QP is solved to return a safe control input:

$$\begin{aligned}\min \quad & \|\mathbf{u}^t - \mathbf{u}_{i_q}^t\|_2^2 \\ \text{s.t. sup } & \mathbf{u}^t \in \mathcal{U} [h(f(\mathbf{x}^t) + g(\mathbf{x}^t)\mathbf{u}^t, t+1) - h(\mathbf{x}^t, t)] \geq -\gamma h(\mathbf{x}^t, t).\end{aligned}\quad (5)$$

Then, we can have the safety guarantees [36], [37].

*Proposition 1*: Assume  $h(\mathbf{x}^t, t)$  is a valid time-varying CBF on  $\mathcal{C}$ . Then any controller  $\mathbf{u}^t(\mathbf{x}^t)$  from (5) for all  $\mathbf{x}^t \in \mathcal{C}$  will render the set  $\mathcal{C}$  forward invariant, i.e., the system is safe.

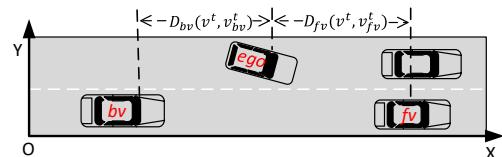


Fig. 3: Lane change in *Traffic-jam* scenario.

## V. EXPERIMENTS AND CASE STUDIES

### A. Testbeds and Common Experiment Setup

**Testbed environment:** We evaluate the performance of STL-guided MARL algorithm on two benchmarks: the multi-agent particle-world environment (MPE) [28] and CARLA [38]. For MPE, we develop two new scenarios, namely “Simple Coordination II” and “Simple Spread II”, which feature an increased number of stages for the agents to reach. These new scenarios introduce additional temporal requirements compared to the existing tasks. Within CARLA, we consider the *Traffic-jam* and *Traffic-jam Expansion* scenario, which are characterized by intricate interactions, demanding stringent safety requirements, and imposing higher temporal constraints on the agents’ decision-making processes.

**Common experimental setup:** We conduct a comparative analysis of the STL-guided MARL algorithm and the MARL algorithm with the original task reward in both testbeds. In STL-guided MARL, the STL reward are the weighted sum of the robustness values of all the STL specifications. We denote the STL reward as  $r_i^t = \sum_j c_j \rho(\varphi_{ij}, \omega, t) + b$ , where  $c_j$  are weights and  $b$  is a constant. To ensure a fair comparison, the original reward function are based on widely adopted designs found in the existing literature. We evaluate the performance of both methods by examining the episode return. We compute the return using the STL reward in both algorithms for consistency. This approach allows us to quantify the extent to which the agents learn to fulfill the designer’s goals. For our experiments, we utilize a server equipped with Intel Core i9-10900X processors and four NVIDIA RTX2080Ti GPUs. The experiments are conducted using Python 3.6.0, PyTorch 1.6.0, and CUDA 11.0.

### B. MPE Testbed

a) *Environment:* In MPE, we design two new tasks, simple coordination II and simple spread II to evaluate our algorithm. In both tasks, the observation of agent  $i$  include the relative positions to other agents and the landmarks, the discrete actions space are: stay, left, right, up and down.

**Baselines:** In both tasks,  $N$  agents need to first cover  $N$  landmarks in the first stage, and then another  $N$  landmarks in the second stage with least collisions. The difference is: in simple coordination II, agent and landmark are paired so agent only targets at its own corresponding landmark; for simple spread II, agents learn to infer the landmark they must cover, and move there while avoiding other agents. The reward function for simple coordination II is:  $r = -c_1 \sum_{i=1}^N (|\mathbf{p}_i - \mathbf{p}_{i,\text{landmark, goal}}|) + c_2 \sum_{i=1}^N (|\mathbf{p}_i - \mathbf{p}_{i,\text{landmark, others}}|)$  where  $\mathbf{p}_i$  is location of agent  $i$ ,  $\mathbf{p}_{i,\text{landmark, goal}}$  is location of the current goal landmark of agent  $i$ ,  $\mathbf{p}_{i,\text{landmark, others}}$  is location of the other landmark. The reward function for simple spread II is  $r = -c_1 \sum_{i=1}^N (\min_{j \in N} |\mathbf{p}_j - \mathbf{p}_{i,\text{landmark, goal}}|) + c_2 \sum_{i=1}^N (\min_{j \in N} |\mathbf{p}_j - \mathbf{p}_{i,\text{landmark, others}}|)$  where  $\mathbf{p}_j$  is location of agent  $j$ ,  $\mathbf{p}_{i,\text{landmark, goal}}$  is location of the landmark  $i$  in current goal stage,  $\mathbf{p}_{i,\text{landmark, others}}$  is location of the other stage’s landmark  $i$ . In both tasks, there will be a  $-1$  penalty

added on the current reward for a collision. The reward function is based on the reward designed in the existing works[28], [21]. We use MADDPG[28] as our baseline algorithm.

**STL-guided MARL:** For both tasks, given a whole trajectory of length  $T$  of agent  $i$ , the requirements include: (1) All first part of  $N$  landmarks are eventually visited by their corresponding agent; (2) All second part of  $N$  landmarks are eventually covered by their corresponding agent; (3) no collision between agents nearby (the distance is always greater than the safety threshold); (4) The first three landmarks should be visited at least once before the second three landmarks are visited.

Therefore, we write the specifications for simple coordination as:

$$\begin{aligned}\varphi_{i_1} &= \Diamond_{[0,T]} \bigwedge_{1,2,\dots} |\mathbf{p}_i^t - \mathbf{p}_{i,\text{landmark, first}}| \leq \epsilon_1, \\ \varphi_{i_2} &= \Diamond \Box_{[0,T]} \bigwedge_{1,2,\dots} |\mathbf{p}_i^t - \mathbf{p}_{i,\text{landmark, second}}| \leq \epsilon_2.\end{aligned}\quad (6)$$

Similarly, the STL specifications for simple spread II are:

$$\begin{aligned}\varphi_{i_1} &= \Diamond_{[0,T]} \bigwedge_{1,2,\dots} \min_{j \in N} |\mathbf{p}_j^t - \mathbf{p}_{i,\text{landmark, first}}| \leq \epsilon_1, \\ \varphi_{i_2} &= \Diamond \Box_{[0,T]} \bigwedge_{1,2,\dots} \min_{j \in N} |\mathbf{p}_j^t - \mathbf{p}_{i,\text{landmark, second}}| \leq \epsilon_2.\end{aligned}\quad (7)$$

The common safety requirement is:  $\varphi_{i_3} = \Box_{[0,T]} \bigwedge_{1,2,\dots} |\mathbf{p}_i^t - \mathbf{p}_j^t| \geq D_{\text{safe}}, \forall j \in N$ .

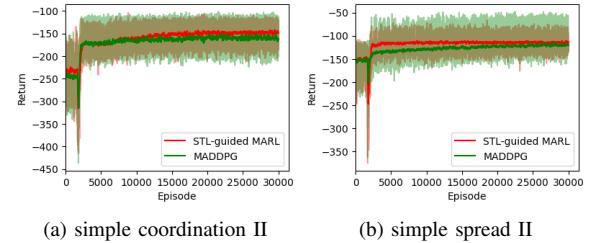


Fig. 4: Training results in MPE.

TABLE II: Mean episode return in MPE. STL-guided MARL shows higher mean episode return in both tasks compared with baseline algorithm, demonstrating the advantages of our method in helping agent learn a better policy to fulfill the designer’s intentions.

Methods	Simple coordination II	Simple spread II
STL-guided MARL	<b><math>-162.01 \pm 29.93</math></b>	<b><math>-120.47 \pm 18.13</math></b>
MADDPG	$-169.17 \pm 33.64$	$-128.36 \pm 17.90$

b) *Experiment Results:* We train the algorithms for 30000 episodes, with episode length of 25 to evaluate their performance. Fig. 4 provides insights into the mean and variance of the average returns for the two tasks. Notably, the STL-guided MARL algorithm demonstrates superior performance compared to baseline algorithm in terms of episode return. Specifically, as shown in Fig. 5, the agents trained with the STL-guided MARL approach exhibit a remarkable ability to cover the second stage landmarks after visiting the first stage. On the other hand, the agents trained with baseline algorithm MADDPG struggle to cover the second stage landmarks and tend to hover around the first stage.

TABLE III: Mean episode return and safety rate in *Traffic-jam* scenario.

Methods	Mean episode return			Safety rate
	Agent 1	Agent 2	Agent 3	
STL-guided MARL	<b>1469.98</b> ± 43.87	<b>3577.53</b> ± 580.04	<b>3699.26</b> ± 588.62	<b>97%</b>
MAPPO	1244.12 ± 174.18	1838.96 ± 392.22	2245.07 ± 150.75	74%
MAA2C	1131.51 ± 389.62	1645.80 ± 693.49	2713.19 ± 465.45	88%
MAPPO w/o STL safety shield	853.30 ± 421.46	1241.91 ± 223.38	2102.25 ± 713.09	65%

This disparity in performance highlights the advantage of the STL-guided approach in facilitating the policy learning. By incorporating STL specifications, the agents are encouraged to adhere to specific behavioral patterns that result in more successful navigation and completion of the tasks. In contrast, the agents trained with a comparison reward, without the STL-guided framework, lack the guidance necessary to achieve optimal performance and struggle to exhibit the desired behavior.

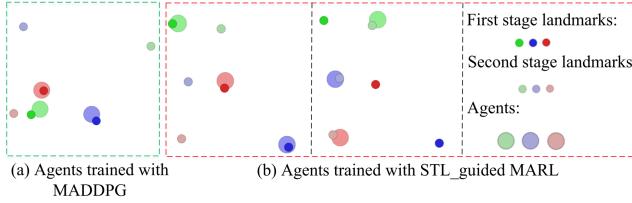


Fig. 5: Simple coordination II in MPE. Agents trained with MADDPG hover around the first stage landmarks and fail to cover the second stage as shown in (a). Agents trained with STL-guided MARL visit the first stage landmarks and then covers the second stage landmarks as shown in (b).

### C. CARLA Testbed

The *Traffic-jam* scenario settings and STL specifications are illustrated in Section IV-C.1. To further validate our algorithm, we add 3 more autonomous vehicles (agents) in the *Traffic-jam* scenario, we name it as *Traffic-jam Expansion*.

a) *Baselines*: We adopt the reward that are widely used in the existing literature for lane merging case[39], [40]. The reward for agent  $i$  are defined as follow:  $r_i = w_1 r_i^{\text{speed}} + w_2 r_i^{\text{collision}} + w_3 r_i^{\text{dest}}$ , where  $w_1, w_2, w_3 \in \mathbb{R}$  are the weights,  $r_i^{\text{speed}} = \frac{|v_i|}{v_{\max}}$ ,  $r_i^{\text{collision}} = -I_{\text{col}}$  with  $I_{\text{col}}$  being the collision intensity collected by collision sensor,  $r_i^{\text{dest}} = -|\mathbf{p}_i - \mathbf{p}_{\text{dest}}| + c$  with  $c$  being a constant. we use MAPPO algorithm [29], MAA2C algorithm[41], and MAPPO algorithm without STL safety shield as our baseline algorithms.

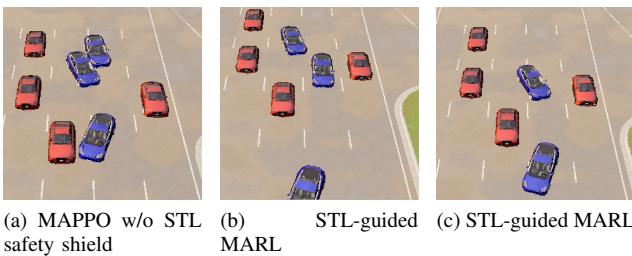


Fig. 6: *Traffic-jam* scenario testing process in CARLA. (a): Agents trained with the baseline algorithm fail to bypass the broken-down vehicle and collide. (b) and (c): Agents trained with STL-guided MARL successfully pass through the only open lane without collision in a timely manner.

b) *Experiment Results*: All the algorithms are trained 100 episodes with the episode length of 150 steps. For *Traffic-jam* scenario, training results are shown in Table III

TABLE IV: Total mean episode return and safety rate in *Traffic-jam Expansion* Scenario. "Total Mean episode return" represents the aggregated total of mean episode returns across all agents.

Methods	Total mean episode return	Safety rate
STL-guided MARL	<b>31828.34</b> ± 2772.05	<b>90%</b>
MAPPO	27617.27 ± 9504.04	64%
MAA2C	24169.17 ± 4133.64	68%
MAPPO w/o STL safety shield	21409.98 ± 2848.80	20%

and the testing visualization is shown in Fig. 6. For *Traffic-jam Expansion* scenario, training results are shown in Table IV. Here safety rate is defined as the proportion of episodes with no collisions relative to the total number of episodes. It can be observed that: (1) STL-guided MARL with STL safety shield largely outperforms the algorithm without it in terms of safety rate, therefore, showing the effectiveness of our proposed STL safety shield in ensuring the safety of the system. (2) While the STL specifications for each agent remain the same, the mode of interaction during an episode can vary between competitive and collaborative. Our proposed method consistently outperforms the baseline methods in terms of the total mean episode return, as demonstrated in Table IV. This demonstrates our algorithm can work in mixed cooperative-competitive environments. (3) The STL-guided MARL algorithm consistently outperforms the baseline algorithms in mean episode return. The superior performance of the STL-guided MARL algorithm can be attributed to its expressiveness and ability to capture the designer's goal. By leveraging STL as a guidance framework, the algorithm is able to incorporate high-level specifications and constraints into the learning process. This enables the agent to learn a policy that aligns more closely with the desired behavior outlined by the designer. Thus, the STL-guided MARL algorithm demonstrates its effectiveness in improving the learning process and enabling the agent to achieve better performance.

## VI. CONCLUSION

We propose a multi-agent reinforcement learning algorithm that leverages signal temporal logic (STL) specifications to guide the learning process and ensure the satisfaction of safety requirements and task objectives for each agent. By incorporating STL safety shields, our algorithm provides additional safety guarantees in the system. Through case studies, we demonstrate that our approach outperforms traditional MARL methods with hand-engineered rewards, as it learns better policies with higher average rewards and ensures the system safety. Our work highlights the potential of using temporal logic and formal languages in MARL to address the challenges of reward design and safety in complex multi-agent systems.

## REFERENCES

- REFERENCES

  - [1] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.
  - [2] Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Başar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8767–8775, 2021.
  - [3] Zhili Zhang, Songyang Han, Jiangwei Wang, and Fei Miao. Spatial-temporal-aware safe multi-agent reinforcement learning of connected autonomous vehicles in challenging scenarios. *arXiv preprint arXiv:2210.02300*, 2022.
  - [4] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.
  - [5] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.
  - [6] Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei, and Changliu Liu. State-wise safe reinforcement learning: A survey. *arXiv preprint arXiv:2302.03122*, 2023.
  - [7] Jan Corazza, Ivan Gavran, and Daniel Neider. Reinforcement learning with stochastic reward machines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6429–6436, 2022.
  - [8] Yiannis Kantaros. Accelerated reinforcement learning for temporal logic control objectives. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5077–5082. IEEE, 2022.
  - [9] Anand Balakrishnan and Jyotirmoy V Deshmukh. Structured reward shaping using signal temporal logic specifications. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3481–3486. IEEE, 2019.
  - [10] Lewis Hammond, Alessandro Abate, Julian Gutierrez, and Michael Wooldridge. Multi-agent reinforcement learning with temporal logic specifications. *arXiv preprint arXiv:2102.00582*, 2021.
  - [11] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.
  - [12] Jingjing Cui, Yuanwei Liu, and Arumugam Nallanathan. Multi-agent reinforcement learning-based resource allocation for uav networks. *IEEE Transactions on Wireless Communications*, 19(2):729–743, 2019.
  - [13] Han Qie, Dianxi Shi, Tianlong Shen, Xinhai Xu, Yuan Li, and Liujing Wang. Joint optimization of multi-uav target assignment and path planning based on multi-agent reinforcement learning. *IEEE access*, 7:146264–146272, 2019.
  - [14] Tianshu Chu, Jie Wang, Lara Codicà, and Zhaojian Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1086–1095, 2019.
  - [15] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.
  - [16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
  - [17] Scott Proper and Kagan Tumer. Modeling difference rewards for multiagent learning. In *AAMAS*, pages 1397–1398, 2012.
  - [18] Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, pages 2107–2116. PMLR, 2018.
  - [19] Xiao Li, Zachary Serlin, Guang Yang, and Calin Belta. A formal methods approach to interpretable reinforcement learning for robotic planning. *Science Robotics*, 4(37):eaay6276, 2019.
  - [20] Mingyu Cai, Shaoping Xiao, Junchao Li, and Zhen Kan. Safe reinforcement learning under temporal logic with reward design and quantum action selection. *Scientific reports*, 13(1):1925, 2023.
  - [21] Xiao Li, Cristian-Ioan Vasile, and Calin Belta. Reinforcement learning with temporal logic rewards. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3834–3839. IEEE, 2017.
  - [22] Ingy ElSayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. *arXiv preprint arXiv:2101.11196*, 2021.
  - [23] Borja G León and Francesco Belardinelli. Extended markov games to learn multiple tasks in multi-agent reinforcement learning. *arXiv preprint arXiv:2002.06000*, 2020.
  - [24] Oded Maler and Dejan Nickovic. Monitoring temporal properties of continuous signals. In *Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems*, pages 152–166. Springer, 2004.
  - [25] Georgios E Fainekos and George J Pappas. Robustness of temporal logic specifications for continuous-time signals. *Theoretical Computer Science*, 410(42):4262–4291, 2009.
  - [26] Jyotirmoy V Deshmukh, Alexandre Donzé, Shromona Ghosh, Xiaoaqing Jin, Garvit Juniwal, and Sanjit A Seshia. Robust online monitoring of signal temporal logic. *Formal Methods in System Design*, 51(1):5–30, 2017.
  - [27] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.
  - [28] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.
  - [29] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.
  - [30] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.
  - [31] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.
  - [32] Lars Lindemann and Dimos V Dimarogonas. Control barrier functions for signal temporal logic tasks. *IEEE control systems letters*, 3(1):96–101, 2018.
  - [33] J Kong, M Pfeiffer, G Schildbach, and F Borrelli. Autonomous driving using model predictive control and a kinematic bicycle vehicle model. In *Intelligent Vehicles Symposium, Seoul, Korea*, 2015.
  - [34] Shilpi Dixit, Saber Fallah, Umberto Montanaro, Mehrdad Dianati, Alan Stevens, Francis McCullough, and Alexandros Mouzakitis. Trajectory planning and tracking for autonomous overtaking: State-of-the-art and future prospects. *Annual Reviews in Control*, 45:76–86, 2018.
  - [35] Gianluca Cesari, Georg Schildbach, Ashwin Carvalho, and Francesco Borrelli. Scenario model predictive control for lane change assistance and autonomous driving on highways. *IEEE Intelligent transportation systems magazine*, 9(3):23–35, 2017.
  - [36] Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2016.
  - [37] Jun Zeng, Bike Zhang, and Koushil Sreenath. Safety-critical model predictive control with discrete-time control barrier function. In *2021 American Control Conference (ACC)*, pages 3882–3889. IEEE, 2021.
  - [38] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.
  - [39] Sai Krishna Sumanth Nakka, Behdad Chalaki, and Andreas A Malikopoulos. A multi-agent deep reinforcement learning coordination framework for connected and automated vehicles at merging roadways. In *2022 American Control Conference (ACC)*, pages 3297–3302. IEEE, 2022.
  - [40] Dong Chen, Mohammad Hajidavalloo, Zhaojian Li, Kaian Chen, Yongqiang Wang, Longsheng Jiang, and Yue Wang. Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic. *arXiv preprint arXiv:2105.05701*, 2021.
  - [41] Georgios Papoudakis, Filippos Christanos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869*, 2020.