Learning Coordinated Maneuver in Adversarial Environments

Zechen Hu, Manshi Limbu, Daigo Shishika, Xuesu Xiao, and Xuan Wang

***Abstract*— This paper aims to solve the coordination of a team of robots traversing a route in the presence of adversaries with random positions. Our goal is to minimize the overall cost of the team, which is determined by (i) the accumulated risk when robots stay in adversary-impacted zones and (ii) the mission completion time. During traversal, robots can reduce their speed and act as a 'guard' (the slower, the better), which will decrease the risks certain adversary incurs. This leads to a trade-off between the robots' guarding behaviors and their travel speeds. The formulated problem is highly non-convex and cannot be efficiently solved by existing algorithms. Our approach includes a theoretical analysis of the robots' behaviors for the single-adversary case. As the scale of the problem expands, solving the optimal solution using optimization approaches is challenging, therefore, we employ reinforcement learning techniques by developing new encoding and policy-generating methods. Simulations demonstrate that our learning methods can efficiently produce team coordination behaviors. We discuss the reasoning behind these behaviors and explain why they reduce the overall team cost.**

I. Introduction

Coordination of multi-robot systems has been studied under various contexts [1], including cooperative path planning [2], resource sharing and task allocation [3], and geometric formation maintenance [4]. Complementary to the challenges addressed in these works, this paper introduces a new problem centered around generating coordinated team behaviors to reduce risks caused by adversaries. Considering a graph-based representation, team coordination problems have been studied in [5]–[7]. In this work, as shown in Fig. 1, we consider a route-based version of the problem, fine-grinding the movements of robots in continuous space. The environment features adversaries whose positions are randomly initialized from a set. Robots accumulate 'risks' when traveling through adversary-controlled zones, and such risks can be reduced by robots if they slow down and 'guard' a certain adversary. We define the team cost as a combination of mission completion time and accumulated risk. Therefore, minimizing this cost requires robots' coordination, trading off between their speed and the adoption of guarding behaviors. This adds a novel dimension of complexity and strategic decision-making, which is unattended in conventional multi-robot coordination tasks.

The described scenario has direct applications in many domains. For instance, when multiple rescue robots need to pass a fire-engulfed corridor [8], some robots might deploy countermeasures as *guard* to quell flames, ensuring safer passage for their peers; On battlefields, when multiple vehicles need to traverse enemy-controlled territories [9], suppressive fire from allies can guard and mitigate threats posed by the enemy. The formulated multi-robot coordination problem is challenging due to the combinatorial nature of robots' states, their hybrid actions (speed and guard), and multiple constraints embedded through robot-adversary interactions.

George Mason University. {zhu3, klimbu2, dshishik, xiao, xwang64}@gmu.edu.

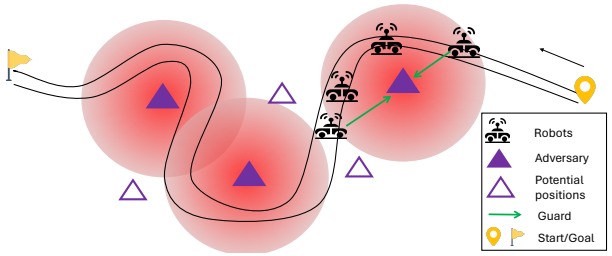


Fig. 1: A team of robots traversing an environment with risk. Adversary positions are randomly initialized from a set.

To solve this problem, one feasible approach from the optimization literature is Mixed Integer Programming (MIP). However, since MIP for multi-robot coordination generally provides an *open-loop* solution for the full trajectories of all robots [10], it faces scalability challenges as the number of robots increases, and requires repeated re-planning if the adversary position or any environmental features change. Motivated by these challenges, in this paper, we seek *closed-loop* policies using a Reinforcement Learning (RL)-based approach. The *main contributions* are as follows: (i) We rigorously formulate a new multi-robot coordination problem that incorporates guard behaviors among team members to mitigate risks from adversaries. (ii) We investigate the conversion of the problem into Markov Decision Processes (MDPs) with hybrid move and guard actions. (iii) We introduce a Hybrid Proximal Policy Optimization algorithm tailored to our problem, featuring special treatment of reward reshaping and a unique multi-weighted hot encoding for representing robots' states and adversary positions. (iv) We perform extensive simulated experiments to validate the effectiveness of the proposed method and compare it with MIP methods. We also elucidate the rationale behind observed behaviors and their effectiveness in reducing the collective team cost.

II. Literature Review

We review related work on MIP and RL for solving multi-robot coordination problems.

A. Mixed Integer Programming

MIP has been applied to various multi-robot coordination problems, including task allocation [11], multi-robot path planning [7], and environmental coverage and exploration [12]. Solving MIP problems is NP-complete, making

traditional MIP solvers sensitive to the number of variables [13] and primarily applicable to small-scale problems. However, for our problem of interest, the coordination of robots occurs at every time step, depending on their coupled actions and states. Such constraints and task specifications can be encoded through logic formulation and piece-wise nonlinear functions [14]. Yet, considering the entire trajectories of all robots as variables can quickly render MINLP (Mixed Integral NonLinear Programming) with coupled constraints intractable. To address this issue, the advances in MIP such as interior point [15], branch and bound [5], [16] methods, and heuristic approaches [17], can drastically improve computation scalability by leveraging convexity properties or the decomposability of the problem. While these properties may not inherently exist for general applications, one can introduce special cost function design and relaxation [5] to promote convexity; or by simplifying explicit collaboration between robots [17] to enhance problem decomposability.

In this paper, these techniques [5], [17] are not directly applicable since explicit coordination is critical to formulating robots' coordination and impact significantly on the optimal behavior of the team[1]. On the other hand, given the hybrid action spaces of the robots, even simple linear risk functions can lead to highly non-smooth non-convex objectives and constraints, which can make MINLP solvers numerically unstable, or converge to sub-optimal solutions [18]. Finally, our problem requires fast deployment and quick adaptation to adversaries with random positions, which favors closed-loop solutions over the open-loop solutions provided by MINLP.

B. Multi-Agent Reinforcement Learning

While optimization techniques suffer from computational complexity, Reinforcement Learning (RL) methods enable robots to employ trial and error to efficiently find empirical solutions for complex problems. Moreover, the learned policy is closed-loop and can adapt in real-time to new adversary positions. For centralized problem solving, Deep Q-Networks (DQN) [19], a value-based RL method suitable for discrete actions, has been applied to learn team formations in battle games [20]. For complex tasks with continuous actions such as traffic optimization [21], Advantage Actor-Critic (A2C) methods [22] can achieve faster convergence and better exploration by utilizing a policy-based model as an actor. Building on A2C, there are generalizations such as Proximal Policy Optimization (PPO) [23] and Deep Deterministic Policy Gradient (DDPG) [24] with improved stability or data efficiency. For systems with hybrid action space, there exist hybrid-PPO [25] methods that can output discrete actions simultaneously with continuous actions. In addition, considering the agent-based nature of our problem, the mentioned algorithms also have multi-agent variants such as Deep Coordination Graph [26], Multi-agent PPO [27], and Multi-agent DDPG [28], allowing for decentralized execution, where each robot learns a local model and determines actions according to local observation. The key advantage of MARL is to improve algorithmic scalability. However, the solution may be sub-optimal due to partial information.

In our setup, although we follow an idea similar to existing centralized hybrid-PPO methods, we face challenges in training efficiency. This has motivated us to apply special treatment to the state encoding of robots and adversaries, as well as reward reshaping, to address these issues.

III. Problem Formulation

In this section, we will first formulate the *task* of the multi-robot team, then introduce the notions of *risk* and *guard*. Based on these, we quantify the *team cost* and describe our *problem of interest*. Throughout the following definitions, adversaries are considered to be heterogeneous while robots are homogeneous.

Robots: Consider a number of n robots traversing a route of length L. The position of the i-th robot along the route at time t is represented by $s_i^t \in [0, L]$. Let $v_i^t \in [0, v_{\max}]$ denote the speed of the i-th traveling robot at time t. Thus, the position update for each traveling robot is given as

$$s_i^{t+1} = s_i^t + v_i^t \Delta t. \tag{1}$$

where Δt is the time interval. Before robot i arrives at the destination, each time step will produce a time penalty, denoted by P_i^t. We define

$$P_i^t = \begin{cases} p & \text{if } s_i^t < L, \\ 0 & \text{if } s_i^t = L. \end{cases} \tag{2}$$

Adversaries and Risk: Let m denote the number of adversaries. The position of adversary j is represented as

$$z_j \in \mathcal{D}_j \subset (0, L), \tag{3}$$

which is randomly chosen from a set $\mathcal{D}_j$ for each trial of experiment. Around z_j is the impact zone of the adversary, denoted by $\mathcal{M}_j$. If a robot is in this region, i.e., $s_i^t \in \mathcal{M}_j$, a cost $r_{i,j}^t$ will be incurred,

$$r_{i,j}^t = \begin{cases} f_j(s_i^t, z_j) \geq 0 & \text{if } s_i^t \in \mathcal{M}_j, \\ 0 & \text{if } s_i^t \notin \mathcal{M}_j, \end{cases} \tag{4}$$

which depends on the relative positions of s_i^t and z_j^t.

Guard: During traversal, if $s_k^t \in \mathcal{M}_j$, robot $k \in \{1, \cdots, n\}$ can reduce its speed and counteract adversary j as a 'guard'. Specifically, let $g_k^t \in \{1, 2, \cdots, m\}$ denote the index of the adversary that robot k is guarding against at time t. Then, the risks that adversary j incurs to robots i, $\forall s_i^t \in \mathcal{M}_j$ are discounted to $\alpha_{k,j}^t r_{i,j}^t$, where

$$\alpha_{k,j}^t = \begin{cases} 1 - \beta \dfrac{|v_{\max} - v_k^t|}{v_{\max}} & \text{if } g_k^t = j, \\ 1 & \text{otherwise}, \end{cases} \tag{5}$$

is the discount factor with $\beta \in (0, 1)$. When $v_k^t = 0$, robot k achieves best guarding performance $\alpha_{k,j}^t = 1 - \beta$, while as $v_k^t \to v_{\max}$, the guarding effect vanishes. Furthermore, we assume the guarding effects stack with each other, thus,

[1] In simulated experiments, a naive baseline algorithm with simplified robot coordination will be given to reflect this gap

considering all robots in the system guarding an adversary j, the risk it incurs to robot i is discounted by all guards as $\prod_{k=1}^{n} \alpha_{k,j}^t r_{i,j}^t$.

Team Cost: Let T be the total time for all robots in the team to traverse the route. Based on the above definitions, the team cost considers the risks and time penalties accumulated by all robots,

$$\boldsymbol{J} = \sum_{i=1}^{n} (\boldsymbol{R}_i + \boldsymbol{P}_i), \tag{6}$$

where $\boldsymbol{R}_i = \sum_{t=1}^{T} R_i^t \Delta t$ and $\boldsymbol{P}_i = \sum_{t=1}^{T} P_i^t \Delta t$, with

$$R_i^t = \sum_{j=1}^{m} \prod_{k=1}^{n} \alpha_{k,j}^t r_{i,j}^t. \tag{7}$$

taking into account the guarding effect.

Problem of Interest: In each time step, robot i's action is composed of traveling speed v_i^t and guard target g_i^t. Assume the robots can observe adversaries' positions z_j but do not know $\mathcal{D}_j$. To strategically design all robots' behaviors, let $\boldsymbol{v}^t = \{v_1^t, \ldots, v_n^t\}$, and $\boldsymbol{g}^t = \{g_1^t, \ldots, g_n^t\}$. The problem is to minimize the team cost, i.e., $t \in \{1, \cdots, T\}$

$$\min_{\{\boldsymbol{v}^t, \boldsymbol{g}^t\}} \boldsymbol{J}. \tag{8}$$

Note that the moving and guarding behaviors of robots lead to team coordination, as one robot can decrease its speed to benefit all traveling robots within the influence region of the guarded adversary.

IV. METHOD

In this section, we present methods for solving the formulated problem. We perform a simple analysis for the case of single-robot single-adversary. For more complicated cases, we introduce proximal policy optimization (PPO) based RL algorithms with a special multi-weighted hot state encoding mechanism and reward reshaping to improve training efficiency.

A. An observation for the single adversary case

We start by considering a single robot and a single adversary case that makes the robot's behavior analyzable. Although this analysis does not yield directly applicable robot coordination strategies, its conclusions align with several real-world observations and can offer insights into the multiple robots and adversaries cases.

Given that the robot's velocity affects its guarding performance $\alpha_{i,j}^t$ as shown in (5), the single robot faces a trade-off: slowing down to incur a discount factor $\alpha_{i,j}^t$ to the risk it takes or speeding up to decrease the time it stays in the adversary controlled area.

Proposition 1. *Suppose there's only one robot and one adversity. Consider the guard discount defined in* (5)*. If* $\Delta t \to 0$*, the cost* $\boldsymbol{J}$ *is minimized if the robot always maintains maximum speed, i.e.,* $\forall t$*,* $v = v_{\max}$.

Proof. For ease of performing analysis, given $\Delta t \to 0$, we convert the risk accumulation to a continuous form, which reads[2]

$$\begin{aligned} \boldsymbol{R} &= \int_0^T (1 - \beta + \frac{\beta v}{v_{\max}}) f(s,z) dt \\ &= \int_0^L (1 - \beta + \beta \frac{v}{v_{\max}}) f(s,z) \frac{ds}{v} \\ &= \int_0^L (\frac{1-\beta}{v} + \frac{\beta}{v_{\max}}) f(s,z) ds \end{aligned} \tag{9}$$

where $v > 0$ and $\int_0^T s \; dt = L$. In the second line of the equation, we substitute the integration variable using the property $\frac{ds}{dt} = v$. From (9), it's clear that $\boldsymbol{R}$ is minimized if $\forall t, v = v_{\max}$. Furthermore, given the definition of $\boldsymbol{P}$ as the time penalty, it is also minimized by maintaining $v = v_{\max}$. This completes the proof. □

Proposition 1 only analyzes a single-robot single-adversary case. However, in the multi-robot case, if there exist spots $s_i^t \in \mathcal{M}_j$ such that $f_j(s_i^t, z_j^t) = 0$, indicating that some robots can guard moving robots with 0 risk, then they should stop to guard, and the moving robots should adopt the strategy of Proposition 1 to reduce $\sum_{i=1}^{n} \boldsymbol{R}_i$. This leads to a 'bang-bang behavior' [29], where the robot either remains in a 'safe spot' to guard others or moves at full speed when under the protection of other robots. This moving pattern, known as 'bounding overwatch', is well-justified in the military domain. However, this strategy also incurs significant time costs in terms of $\sum_{i=1}^{n} \boldsymbol{P}_i$, as it necessitates some robots to fully stop when guarding others.

When the number of robots grows large and the time penalty $\sum_{i=1}^{n} \boldsymbol{P}_i$ begins to dominate the overall cost, the strategy might change into two possible variations: (i) the robots take special scheduling to stop or move, and when moving, they move at full speed; (ii) several robots move at intermediate speeds to perform move and guard simultaneously. Both strategies are observed later in our simulations section, depending on the environment setup. Nevertheless, these behaviors consider the movements of multiple robots in a coupled manner, which makes theoretical analysis intractable. In addition, as will be discussed in the simulation, due to the piece-wise and condition-dependent non-linear cost structure and coupled constraints, the MINLP formulation of the problem is very difficult to solve. Motivated by these, we seek to use reinforcement learning to solve the problem with a closed-loop solution.

B. MDP Formulation and RL methods

The Markov Decision Process (MDP) formulation of our problem is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, R)$, including state, action, state transition, discount factor, and reward. Here, we propose a centralized MDP to address the multi-robot coordination problem formulated in Sec. III. Specifically, let $\boldsymbol{s}^t = \{s_1^t, \ldots, s_n^t, z_1, \ldots, z_m\} \in \mathcal{S}$ be the state set

[2]The subscripts i, j are omitted since only considering one robot and one adversary. For continuous representation, the notation for time t is also omitted.

at time t, where $s_i^t, z_j \in [0, L]$ are associated with robots positions and adversaries positions, respectively. Since adversaries may appear at random positions, their information needs to be encoded into the system state so that the learned policy can generate different strategies corresponding to the adversary positions. Following this, the state space is defined as:

$$\mathcal{S} \coloneqq [0, L]^n \times [0, L]^m.$$

For the actions of robots, we have $a_i^t = (v_i^t, g_i^t)$, which is a hybrid combination of continuous speed $v_i^t \in [-v_{\max}, v_{\max}]$ and discrete guard behavior $g_i^t \in \{1, 2, \ldots, m\}$. The action space for all robots is

$$\mathcal{A} \coloneqq \left([-v_{\max}, v_{\max}] \times \{1, 2, \ldots, m\}\right)^n. \tag{10}$$

Based on $\mathcal{S}$ and $\mathcal{A}$, the state transition is $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. The robot states follow motion dynamics (1), which is deterministic, and the adversary positions are static and do not depend on $\mathcal{A}$. The $R(\boldsymbol{s}^t, \boldsymbol{a}^t)$ is the immediate reward of action $\boldsymbol{a}^t \in \mathcal{A}$ with state $\boldsymbol{s}^t \in \mathcal{S}$, defined as the negative team cost

$$R(\boldsymbol{s}^t, \boldsymbol{a}^t) \coloneqq -\sum_{i=1}^{n} (R_i^t + P_i^t). \tag{11}$$

We complete our MDP formulation by choosing a discounter factor $\gamma = 0.995$. The goal of RL is to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ to maximize the expected cumulative reward for the whole team over the task horizon T, i.e.,

$$\max_{\pi} \quad \mathop{\mathbb{E}}_{\boldsymbol{a}^t \sim \pi(\cdot|\boldsymbol{s}^t)} \left[\sum_{t=0}^{T} (\gamma)^t R^t \right]. \tag{12}$$

Multi-weighted hot state encoding. To employ RL methods to solve our MDP problem, we can directly feed a scalar representation of each robot's and adversary's state (position) to the model. However, we observe that the dimension of our state space is much smaller than that of the action space due to the joint speed and guard behaviors. This discrepancy hinders the neural network's reasoning capabilities, which cannot efficiently learn parameters [30]. Inspired by the one-hot encoding, we seek to expand the state space using a similar mechanism. However, one-hot encoding is typically suited for discrete variables, whereas our state space is continuous. To address this, for robot positions, we introduce a new *weighted hot encoding* mechanism, which represents a continuous variable as the weighted average of two neighboring one-hot vectors. Specifically, let $h(s) \in [0, 1]^{L+1}$ denote the weighted hot encoding for a state $s \in [0, L]$, and $s = s_{\text{int}} + s_{\text{dec}}$, which has both integer and decimal parts. Let $h(s)[k], k \in \mathbb{Z}$ denote the kth element of vector $h(s)$. Then $h(s)$ is a vector with two nonzero entries:

$$\begin{cases} h(s)[s_{\text{int}} + 1] = 1 - s_{\text{dec}} \\ h(s)[s_{\text{int}} + 2] = s_{\text{dec}} \end{cases} \tag{13}$$

As a simple example, if $L = 4$ and $s = 3.2$. Since $3.2 = 0.8 \times 3 + 0.2 \times 4$, one has, $h(s) = [0 \;\; 0 \;\; 0 \;\; 0.8 \;\; 0.2]^\top$. If $s = 0.7 = 0 \times 0.3 + 1 \times 0.7$, one has $h(s) = [0.3 \;\; 0.7 \;\; 0 \;\; 0 \;\; 0]$.

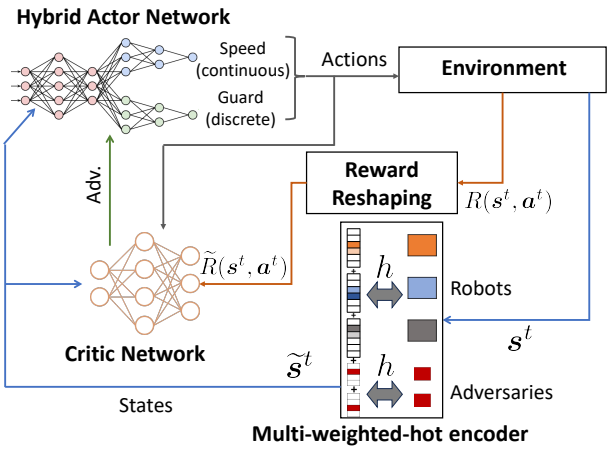


Fig. 2: RL Implementation: H-PPO with multi-weighted hot encoding and reward reshaping

Since we consider a centralized MDP, we vector stack each robot's encoding and adversaries encoding through $h(\cdot)$ to obtain a Multi-weighted hot state encoding as follows:

$$\widetilde{\boldsymbol{s}}^t = \text{vec}\{h(s_1^t), \cdots, h(s_n^t), h(z_1), \cdots, h(z_m)\},$$

which has a dimension of $[0, 1]^{(n+1)(L+1)}$. Ths allows us to individually encode robots' positions and adversary positions in a way that is easily understood by the neural network. It is worth **remarking** that one-hot encoding is typically used for categorical variables. In our problem, the positions of robots, whether inside or outside of adversary-impacted zones, correspond to completely different properties and different feasible guard actions. This distinction favors one-hot encoding, as all input entries are orthogonal to each other. This fact further justifies why multi-weighted hot state encoding can enhance our learning efficiency.

Reward reshaping. Since our task requires all robots to move to the terminal position, it is common to introduce a one-time constant reward $Q(\boldsymbol{s}^t) = q$, if $s_i^t = L, \forall i$; $Q(\boldsymbol{s}^t) = 0$, otherwise. However, this terminal reward is so sparse that it provides limited guidance to robots for state-space exploration and policy updates. Due to the greedy nature of the action section process, robots are reluctant to enter adversary zones. To address this, we further introduce a reshaping reward

$$F(\boldsymbol{s}^t, \boldsymbol{s}^{t+1}) = c \sum_{i=1}^{n} \left(\gamma s_i^{t+1} - s_i^t\right), \tag{14}$$

which incites robots to move forward and speeds up the learning process [31]. The final reshaped reward reads

$$\widetilde{R}(\boldsymbol{s}^t, \boldsymbol{a}^t) = R(\boldsymbol{s}^t, \boldsymbol{a}^t) + Q(\boldsymbol{s}^t) + F(\boldsymbol{s}^t, \boldsymbol{s}^{t+1}) \tag{15}$$

where $\boldsymbol{s}^{t+1}$ is determined by $\boldsymbol{s}^t, \boldsymbol{a}^t$ through $\mathcal{T}$. We **note** that the reshaped reward does not change the optimal solution compared to the original formulation. This is guaranteed by [32], as both $Q(\boldsymbol{s}^t)$ and $F(\boldsymbol{s}^t, \boldsymbol{s}^{t+1})$ can be rewritten into a potential-based function: $\gamma\Phi(\boldsymbol{s}^{t+1}) - \Phi(\boldsymbol{s}^t)$, where $\Phi(\cdot)$ is a real-valued function of state and γ is the discount factor.

RL Implementation. Combining the formulated MPD with multi-weighted hot state encoding and reward reshaping, we use two proximal policy optimization (PPO) based RL algorithms to solve the multi-robot path traveling problem. The key difference lies in the way we handle the hybrid action space. First, for simplicity, we consider pure discrete action space, assuming robots only take integral speeds. This has led to a standard PPO with discrete actions (D-PPO) [33], and the proposed multi-weighted hot state encoding degrades to multi-one hot encoding. Second, we consider PPO with hybrid action space (H-PPO), and let the actor-network simultaneously output continuous speed actions and discrete guard actions. The policy losses (log probabilities) of the two actions are combined and used for training the network parameters. A conceptual diagram of the RL implementation is shown in Fig. 2, with a centralized structure to handle all robots' rewards and actions. Leaving as our future work, a possible decentralized implementation of the RL paradigm is to let each robot possess a local model of Fig. 2. Then, leveraging our multi-weighted-hot encoder, if the robot cannot observe certain robot's states, the corresponding weighted-hot vector has all entries being zero.

V. Simulated Experiments

In this section, we present simulated experiment results to validate the analytical statements in Sec. IV-A and the RL implementation for team coordination behaviors in Sec. IV-B. For complex cases, we discuss the reasoning behind these behaviors and why they reduce team costs. We also provide comparisons with baselines and numerical methods based on MINLP using Surrogate optimization [34] and BONMIN Solver [35].

A. Simulation environment

We consider three different environments shown in Fig. 3, including 1-3 adversaries with potential overlaps over their impact zones. The route in Fig. 1 is abstracted as a linear distance from the starting point to the target. In M1, the adversary's position z_j is randomly chosen from a discrete integer set $\mathcal{D}_j$, with width 10 centered at the middle of the environment; In M2, the two adversary's positions are chosen from two discrete sets $\mathcal{D}_j$, each has a width 5; the risk zones may overlap with each other. For both M1 and M2, when risk functions $f_j(s_i^t, z_j^t)$ are homogeneous and depends linearly on the distance between the robot and an adversary

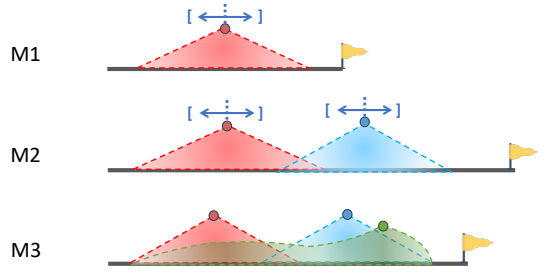


Fig. 3: Experiment environments with different adversary configurations. The height represents the unit risk each adversary generates at different locations.

(visualized by the height of the shades in the figure). In M3, we consider stationary adversary positions, but the risk functions are heterogeneous and nonlinear, and with more complex overlaps.

We choose the following environment parameters: time interval $\Delta t = 1$, max robot speed $v_{\max} = 3$, risk coefficient $\eta = 1$, time penalty $p = 1$, guard discount coefficient $\beta = 0.6$. All rewards, before being sent to D-PPO, H-PPO models, are re-scaled for normalization purposes. For each environment (M1, M2, M3), the learning model is individually trained. Within each environment, since adversary positions are part of the state space, a single model is trained with all possible adversary positions. During execution, this single model can adapt to adversary positions that are randomly generated for each test. This reflects the generalization capability of our model.

B. Team coordination and model generalization with homogeneous adversaries.

The results in Fig. 4a-c consider the M1 scenario with two and three robots, respectively. The D-PPO and H-PPO methods generate almost the same results except for slight differences in velocity when robots leave or approach the boundaries of adversary-impacted zones, which leads to minor changes to the final reward. For conciseness, we only visualize the results from D-PPO. In Fig. 4a, the behaviors of the two robots follow a 'bounding overwatch' pattern described in Sec. IV-A, i.e., one robot guards at the boundary of the adversary-impacted zone with minimal risk until the other robot moves across the zone at full speed. Then, the two robots switch roles to cooperatively accomplish the task with minimum cost. In Fig. 4b, as the number of robots increases, we observe (from the vertical dashes) a change in robots' coordination such that the guarding robots start moving 2 seconds before the traveling robots arrive at the other end. This adjustment is due to the increase in the number of robots; the time penalty encourages the robots to arrive at the destination more quickly. In Fig. 4c, we change the adversary's position in the middle of the execution. Although our algorithm is not designed to account for the dynamic behaviors of adversaries, the closed-loop nature of the policy and the fact that all possible adversary positions have been learned during the training process allow robots to quickly make adaptations. At the behavior level, robots maintain their position at the boundary of the zone when guarding. This demonstrates the model's generalization capability. If using optimization-based approaches, such as mixed-integer programming, then replanning is necessary.

We employ D-PPO and H-PPO methods to solve optimal coordination under the M2 environment. Since the positions of the two adversaries are randomly initialized, it is possible that the two adversary-impacted zones may or may not overlap with each other, which will then impact the coordination patterns of the robots. Here, we choose two representative cases: without overlap and with overlap. In Fig.4d, without overlap, the three robots simply reproduce coordination behaviors in Fig. 4b over the two zones, respectively. In the case

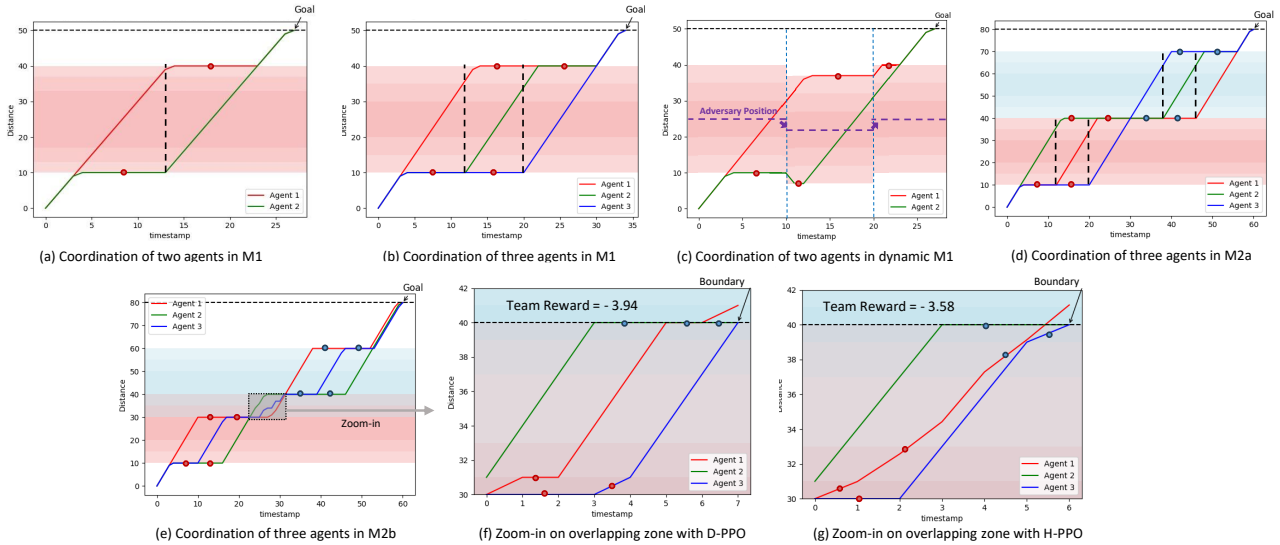


Fig. 4: Using D-PPO and H-PPO to solve team coordination problem with first two environments in Fig. 3. The x-axis represents time and the y-axis represents the length the robot traveled in the environment. The slope represents the robot's speed. The color dots on the trajectories represent the adversary the robot is currently guarding against. In (c) the purple dash represents the current adversary position. The shades are the adversary-impacted zones with darker colors in the middle to represent higher risk, corresponding to Fig. 3.

of Fig.4e, where two adversaries have an overlapped area (c.f. M2 in Fig. 3), the results of D-PPO and H-PPO show relatively consistent strategies for the $[10, 30]$ and $[40, 60]$ zones, as shown in Fig. 4e, but exhibit variant behaviors in the overlapping zone $[30, 40]$. To investigate this, we zoom into the overlapping zone and observe the difference between H-PPO and D-PPO results shown in Fig. 4f-g. Note that robots start at $\{30, 31, 30\}$ instead of $\{30, 30, 30\}$ because when entering the zone, robot 2 moves with $v_{\max} = 3$ from $s_2 = 28$ and directly arrives at $s = 31$. For a similar reason, the terminal states are $\{41, 40, 40\}$. In both Fig. 4f and g, at least one robot takes intermediate speeds that perform move and guard behaviors simultaneously. This aligns with the hypothesis-(ii) at the end of Sec. IV-A. When robots are close to the boundaries of the overlapping zone, their risks are dominated by red and blue adversaries, respectively. As observed in plots, the stationary robot always guards the adversary which causes more risk, while the robot with intermediate speed will guard the other adversary. Moreover, a comparison of the results reveals that H-PPO, with its ability to handle continuous speeds, can refine strategies more effectively, resulting in better rewards. The asymmetry in robot's behavior may be due to time discretization, where the cost is computed at the end of each time step.

C. Validation of Reward Reshaping and Weighted-hot State Encoding

Using the same environment setup as in M1 and M2, we validate the effectiveness of the proposed reward reshaping and weighted-hot state encoding techniques, and compare them with the results when these techniques are not used.

1) Necessity of Reward Reshaping: Using environment M1, the effectiveness of reward reshaping is demonstrated in Fig. 5 (a). It is observed that without reward reshaping, the reward curve in the PPO training shows a trend of downward

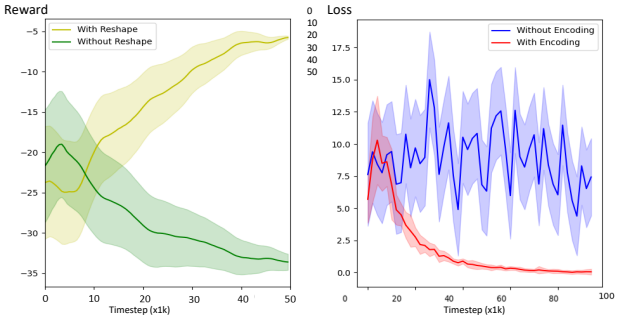


Fig. 5: Validations of Reward Reshaping and Weighted-hot State Encoding Techniques.

convergence. When executing the learned policy in the gym environment, it was noticed that all robots halted before entering the adversary zone. We believe this is due to the original problem formulation, where entering the adversary zone triggers significant costs immediately, causing unfavorable exploration because robots tend to avoid such heavily penalized actions. Furthermore, since the environment has a maximum simulation times step of 100[3], robots will simply wait until the simulation stops and learn a bad behavior with poor rewards. On the other hand, with reward reshaping applied, the training could stably converge toward the optimal solution's reward values. As we execute the learned policy in the gym environment, robots can generate meaningful coordination behaviors. This verifies the effectiveness of our reward reshaping as discussed in Sec. IV-B.

2) Effectiveness of Weighted-hot State Encoding.: We use the M2 environment to verify the effectiveness of our

[3]We note that in the ideal case, 100 steps are more than enough for all robots to reach the destination.

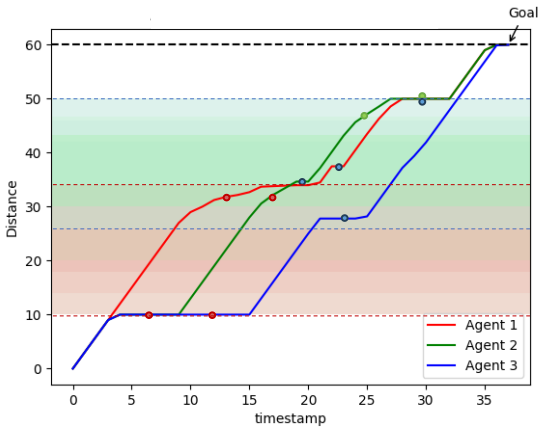
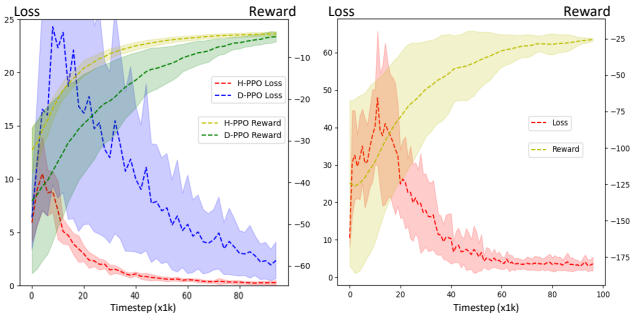


Fig. 6: Coordination of three robots in M3 using H-PPO. The figure is read the same way as Fig. 4.

weighted-hot state encoding mechanism. From Fig. 5 (b), the training loss curve indicates that directly using the scalar values of robots' and adversaries' positions as neural network inputs, i.e., without weighted-hot encoding results in continuous oscillations in the PPO training process. This implies that the neural network faces difficulties in understanding the state description. Instead, when weighted-hot encoding is used and all other model parameters (latent layers, learning rate, exploration rate) are kept the same, the training stability is significantly improved and the loss converges to zero quickly. This verifies the effectiveness and necessity of our weighted-hot state encoding mechanism.

D. Behavior analysis of team coordination in complex environments with heterogeneous adversaries.

For the case of three adversaries with multiple overlaps and heterogeneous nonlinear risk functions (M3), we deploy three robots and obtain coordination results as shown in Fig. 6. Here, the adversaries are stationary. We added colored dashes to better visualize the boundaries of red and blue adversaries. The environment's complexity makes it difficult to judge the optimality of the obtained coordination. In the following, we only discuss the reasoning behind the obtained result and explain why it reduces the overall team cost. First, according to Fig. 3, the M3 environment adds M2 with an extra green adversary. This green adversary generates an unsymmetric and nonlinear risk zone, which poses little risk early in the path but grows large as robots proceed. Consequently, in Fig. 6, robots first follow a pattern very similar to that of Fig. 4e. However, midway through the path, as risks associated with the green adversary become large, the predecessor robot 1 does not fully stop to guard others. Instead, it takes an intermediate speed to guard the red adversary. This lasts until robot 1 meets robot 2 and both leave the boundary of the red adversary. On the other hand, we observe robot 3 stops in the middle of the path to perform guard. This happens because, at the moment, the other two robots are suffering huge risks from both blue and green adversaries. By stopping and thereby increasing its own risk, robot 3 contributes to the cost-saving of the whole team.



(a) M2 environment with overlapping zone using H-PPO and D-PPO.

(b) M3 environment using H-PPO.

Fig. 7: Convergence results for training loss and rewards.

Finally, when robot 1 and 2 both arrives at the boundary, they help robot 3 by guarding red and blue adversaries, respectively. We note that the coordination presented in Fig. 6 may not be the global optimal, and its optimality gap is difficult to quantify. However, as discussed above, the observed robot coordination does exhibit rational behaviors, with the goal of reducing the overall team cost.

E. Reward Comparison with Baseline and MINLP Solvers

To visualize the training process, in Fig. 7, we use M2 and M3 environments as representative results to show training losses and rewards. We observe that H-PPO performs better than D-PPO and we note that D-PPO struggles to converge for the M3 environment.

For comparison purposes, we introduce a naive baseline strategy where all robots' actions are decoupled, and each robot individually uses greedy choices for move and guard. Additionally, we write the MINLP formulation of the problem and solve it using the BONMIN Solver [35] in Python and the Surrogate algorithm in MATLAB. We note that the piece-wise and condition-dependent non-linear functions (2)-(5) require large numbers of auxiliary variables to formulate, which leads to low computational efficiency. When run indefinitely, BONMIN takes hours trying to close the optimality gap. Therefore, we stop BONMIN in 10 minutes if it does not converge and record the result. For all environments in Fig. 3, a fixed set of adversary positions is chosen, and three robots are deployed to test all approaches. Specifically for M2, there is an overlap between the two adversaries. As shown in Table I, the developed H-PPO generally outperforms or provides comparable performance to other methods. The Surrogate does not exhibit meaningful convergence on rewards for the two complex cases. As a final remark, while BONMIN is stopped after 10 minutes, H-PPO obviously requires significantly more training time. However,

TABLE I: Comparison with Baseline and MINLP Solvers

Env	D-PPO	**H-PPO**	Baseline	BONMIN	Surrogate
M1	-5.80 ± 0.2	$\mathbf{-5.78 \pm 0.0}$	-9.94	$\mathbf{-5.78}$	-6.11
M2	-8.63 ± 0.7	-8.22 ± 0.4	-20.18	$\mathbf{-8.03}$	N/A
M3	-40.13 ± 8.5	$\mathbf{-26.87 \pm 2.0}$	-44.70	-30.32	N/A

H-PPO can be trained offline and provides a general closed-loop policy for all possible adversary positions. In contrast, for BONMIN, each adversary position needs to be solved individually for an open-loop solution. Thus, for real-time applications, the proposed H-PPO is more desirable.

VI. Conclusion and Future Work

We have formulated a coordination problem considering a team of robots traversing a route with adversaries. The cost of the team was determined by the time penalty and the risk robots accumulated when crossing adversary-impacted zones, which can be reduced by robots' guard behavior when moving at a lower speed. We analyzed the optimal coordination strategy for a single adversary scenario. For complex environments, we proposed and implemented an H-PPO method with reward reshaping and a new multi-weighted hot state encoding mechanism. Simulated experiments are performed under different environments and compared with alternative approaches to validate our analysis and the effectiveness of the H-PPO method. Based on the simulation results, we discussed the reasoning behind these behaviors in terms of reducing the overall team cost. Future work will consider developing a decentralized learning paradigm to achieve scalable coordination with a larger number of robots and more complicated environments. We also seek to expand the proposed formulation to non-route-based environments considering the geometries of risk zones and terrain, and the decentralization of the proposed algorithm.

References

[1] A. Dorri, S. S. Kanhere, and R. Jurdak, "Multi-agent systems: A survey," *Ieee Access*, vol. 6, pp. 28573–28593, 2018.

[2] A. Torreno, E. Onaindia, A. Komenda, and M. Štolba, "Cooperative multi-agent planning: A survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–32, 2017.

[3] M. Afrin, J. Jin, A. Rahman, A. Rahman, J. Wan, and E. Hossain, "Resource allocation and service provisioning in multi-agent cloud robotics: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 842–870, 2021.

[4] K.-K. Oh, M.-C. Park, and H.-S. Ahn, "A survey of multi-agent formation control," *Automatica*, vol. 53, pp. 424–440, 2015.

[5] C. A. Dimmig, K. C. Wolfe, and J. Moore, "Multi-robot planning on dynamic topological graphs using mixed-integer programming," *arXiv preprint arXiv:2303.11966*, 2023.

[6] M. Limbu, Z. Hu, S. Oughourli, X. Wang, X. Xiao, and D. Shishika, "Team coordination on graphs with state-dependent edge cost," in *20230 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023.

[7] C. A. Dimmig, K. C. Wolfe, M. Kobilarov, and J. Moore, "Uncertainty-aware planning for heterogeneous robot teams using dynamic topological graphs and mixed-integer programming," *arXiv preprint arXiv:2310.08396*, 2023.

[8] R. Bogue, "The role of robots in firefighting," *Industrial Robot: the international journal of robotics research and application*, vol. 48, no. 2, pp. 174–178, 2021.

[9] P. Mátyás and N. Máté, "Brief history of uav development," *Repüléstudományi Közlemények*, vol. 31, no. 1, pp. 155–166, 2019.

[10] A. Cauligi, P. Culbertson, B. Stellato, D. Bertsimas, M. Schwager, and M. Pavone, "Learning mixed-integer convex optimization strategies for robot planning and control," in *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 1698–1705, IEEE, 2020.

[11] C. Zhang and J. A. Shah, "Co-optimizating multi-agent placement with task assignment and scheduling.," in *IJCAI*, pp. 3308–3314, 2016.

[12] F. Gul, A. Mir, I. Mir, S. Mir, T. U. Islaam, L. Abualigah, and A. Forestiero, "A centralized strategy for multi-agent exploration," *IEEE Access*, vol. 10, pp. 126871–126884, 2022.

[13] D. Ioan, I. Prodan, S. Olaru, F. Stoican, and S.-I. Niculescu, "Mixed-integer programming in motion planning," *Annual Reviews in Control*, vol. 51, pp. 65–87, 2021.

[14] R. Calegari, G. Ciatto, V. Mascardi, and A. Omicini, "Logic-based technologies for multi-agent systems: a systematic literature review," *Autonomous Agents and Multi-Agent Systems*, vol. 35, no. 1, p. 1, 2021.

[15] F. A. Potra and S. J. Wright, "Interior-point methods," *Journal of computational and applied mathematics*, vol. 124, no. 1-2, pp. 281–302, 2000.

[16] L. D. Beal, D. C. Hill, R. A. Martin, and J. D. Hedengren, "Gekko optimization suite," *Processes*, vol. 6, no. 8, p. 106, 2018.

[17] J. Yu and S. M. LaValle, "Optimal multirobot path planning on graphs: Complete algorithms and effective heuristics," *IEEE Transactions on Robotics*, vol. 32, no. 5, pp. 1163–1177, 2016.

[18] E. Klotz, "Identification, assessment, and correction of ill-conditioning and numerical instability in linear and integer programs," in *Bridging Data and Decisions*, pp. 54–108, INFORMS, 2014.

[19] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," *Advances in neural information processing systems*, vol. 29, 2016.

[20] E. A. O. Diallo and T. Sugawara, "Learning strategic group formation for coordinated behavior in adversarial multi-agent with double dqn," in *PRIMA 2018: Principles and Practice of Multi-Agent Systems: 21st International Conference*, pp. 458–466, Springer, 2018.

[21] H. Zhang, W. Chen, Z. Huang, M. Li, Y. Yang, W. Zhang, and J. Wang, "Bi-level actor-critic for multi-agent coordination," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 7325–7332, 2020.

[22] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *Advances in neural information processing systems*, vol. 12, 1999.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[24] S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, and S. Russell, "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 4213–4220, 2019.

[25] Z. Fan, R. Su, W. Zhang, and Y. Yu, "Hybrid actor-critic reinforcement learning in parameterized action space," *arXiv preprint arXiv:1903.01344*, 2019.

[26] W. Böhmer, V. Kurin, and S. Whiteson, "Deep coordination graphs," in *International Conference on Machine Learning*, pp. 980–991, PMLR, 2020.

[27] J. Wang and L. Sun, "Dynamic holding control to avoid bus bunching: A multi-agent deep reinforcement learning framework," *Transportation Research Part C: Emerging Technologies*, vol. 116, p. 102661, 2020.

[28] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.

[29] R. Bellman, I. Glicksberg, and O. Gross, "On the "bang-bang" control problem," *Quarterly of Applied Mathematics*, vol. 14, no. 1, pp. 11–18, 1956.

[30] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis, "Reinforcement learning, fast and slow," *Trends in cognitive sciences*, vol. 23, no. 5, pp. 408–422, 2019.

[31] S. M. Devlin and D. Kudenko, "Dynamic potential-based reward shaping," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), pp. 433–440, IFAAMAS, 2012.

[32] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proceedings of the 16th International Conference on Machine Learning*, pp. 278–287, 1999.

[33] C. C.-Y. Hsu, C. Mendler-Dünner, and M. Hardt, "Revisiting design choices in proximal policy optimization," *arXiv preprint arXiv:2009.10897*, 2020.

[34] N. V. Queipo, R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. K. Tucker, "Surrogate-based analysis and optimization," *Progress in aerospace sciences*, vol. 41, no. 1, pp. 1–28, 2005.

[35] P. Bonami and J. Lee, "Bonmin user's manual," *Numer Math*, vol. 4, pp. 1–32, 2007.