Harmonizing Team and Individual Goals in Decentralized Multi-Robot Fire Suppression through Curriculum Reinforcement Learning

Williard Joshua Jose, Yuhao Li, Hao Zhang

***Abstract*— Coordinating multi-robot systems is essential for addressing complex real-world scenarios such as autonomous fire suppression. To effectively combat fires that may be geographically dispersed, robots must possess independent autonomous capabilities while also being able to form collaborative teams to handle larger, more intense fires. A key challenge is the inherent conflict between the overall team goal of suppressing all fires as quickly as possible and the individual robot goals, which aim to maximize the amount of fire each robot suppresses. To explicitly address this, we propose *Objective Harmonization Curriculum for Fire Suppression* (OHCF), a method that integrates curriculum learning with multi-agent reinforcement learning (MARL). We design a novel three-stage curriculum that guides the learning process by initially prioritizing individual objectives to build foundational skills, then gradually shifting the reward structure to emphasize the collective goal, and finally refining the policy to optimize efficient team coordination. Through dynamically balancing individual and team-based rewards during training, our approach explicitly harmonizes these conflicting objectives, enabling robots to learn not only when to act alone but also when to collaborate. Our method also facilitates adaptive subteaming and supports fully decentralized execution in physical multi-robot systems. Experimental results show that our approach effectively harmonizes team and individual goals for multi-robot fire suppression and outperforms baseline methods.**

More details are provided on the (anonymous) project website: `icra-ohcf.github.io`.

I. INTRODUCTION

Multi-robot systems have been widely studied over the past decades to address a broad range of real-world applications, such as logistics and delivery [1], search and rescue [2], and environmental monitoring [3]. A particularly impactful application is autonomous fire suppression [4], which aims to improve the speed and effectiveness of responses to increasingly more common wildfires and urban fires while reducing the risk posed to human firefighters [5]. Multi-robot teams are well-suited for the challenge of fire suppression, as illustrated in Fig. 1, since fires are often spatially distributed across large areas, requiring robots to disperse and address multiple incidents in parallel. In addition, large or intense fires may demand close collaboration among multiple robots to achieve successful suppression. Therefore, an effective multi-robot fire suppression strategy must enable robots to operate independently when advantageous, while still coordinating seamlessly when collaboration is required.

Given the importance of multi-robot coordination, a variety of methods have been implemented. Classical learning-free techniques, such as planning or optimization-based techniques, can find optimal paths but typically struggle with scalability and adapting to dynamic environments in real-time [6]–[8]. Multi-agent reinforcement learning (MARL) has emerged as a promising paradigm for developing complex, decentralized control policies. However, existing MARL techniques often face a critical dilemma when there are conflicting team and individual goals (Fig. 1). Purely cooperative methods can lead to inefficient "lazy agents" and unfair workload distribution [9], [10], whereas approaches focused solely on individual rewards for independent agents often fail to develop the complex coordination required for team-based tasks [11], [12]. Recent works on mixed-motive MARL have explored this dilemma from a theoretical perspective by defining specialized rewards [13], [14] or using improved gradient-based optimization methods [15], [16]. However, these approaches have largely been applied only to simple matrix-based games and have not yet been extended to more dynamic robotics environments.

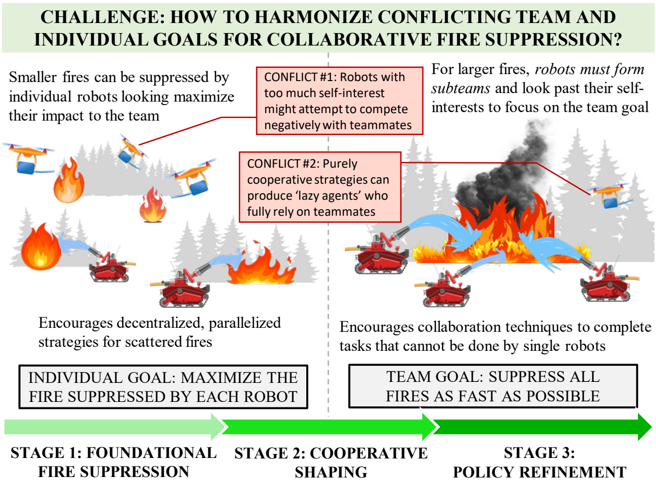
CHALLENGE: HOW TO HARMONIZE CONFLICTING TEAM AND INDIVIDUAL GOALS FOR COLLABORATIVE FIRE SUPPRESSION?
Smaller fires can be suppressed by individual robots looking maximize their impact to the team
CONFLICT #1: Robots with too much self-interest might attempt to compete negatively with teammates
For larger fires, *robots must form subteams* and look past their self-interests to focus on the team goal
CONFLICT #2: Purely cooperative strategies can produce 'lazy agents' who fully rely on teammates
Encourages decentralized, parallelized strategies for scattered fires
Encourages collaboration techniques to complete tasks that cannot be done by single robots
INDIVIDUAL GOAL: MAXIMIZE THE FIRE SUPPRESSED BY EACH ROBOT
TEAM GOAL: SUPPRESS ALL FIRES AS FAST AS POSSIBLE
STAGE 1: FOUNDATIONAL FIRE SUPPRESSION
STAGE 2: COOPERATIVE SHAPING
STAGE 3: POLICY REFINEMENT

Fig. 1. A motivating scenario involves decentralized multi-robot fire suppression, where robots must balance conflicts between team-level and individual objectives. Our curriculum-based multi-robot learning approach explicitly harmonizes these goals, which enables effective coordination while preserving individual efficiency.

In this paper, we introduce a learning-based approach called *Objective Harmonization Curriculum for Fire Suppression* (OHCF), which enables a team of robots to learn coordinated behaviors and adaptively form subteams to suppress fires of varying intensities. While the team's goal is to extinguish all fires as quickly as possible, each robot's individual goal is to maximize the number of fires it suppresses, which may conflict with the team goal. OHCF is formulated as a MARL problem, and we introduce a curriculum to guide the learning process to achieve harmonization through aligning individual

The authors are with the Human-Centered Robotics Laboratory, University of Massachusetts Amherst, Amherst, MA 01002, USA.

robot goals with the team objective. Our curriculum learning consists of three stages: (1) Foundational Fire Suppression Learning initially prioritizes the individual objective, allowing each robot to learn effective strategies for suppressing smaller fires; (2) Cooperative Shaping gradually shifts the reward structure to emphasize the collective objective; and (3) Policy Refinement incentivizes efficient team cooperation through sustained optimization of the collective team goal. OHCF supports fully decentralized execution on a physical multi-robot system in real-world environments and has demonstrated superior performance compared to previous methods.

The paper's main contribution is the introduction of a novel curriculum-based multi-agent learning method designed to specifically address objective harmonization in multi-robot fire suppression. OHCF's novelties include:

- We propose a new curriculum-based multi-agent reinforcement learning approach that dynamically shifts the reward balance from individual to team objectives during training, which harmonizes and resolves conflicts between individual robot goals and overall team objectives.
- We significantly advance multi-robot collaboration for fire suppression, with the new ability to harmonize individual and team objectives, dynamically form subteams, execute fully decentralized, and operate in real time on physical robot teams.

II. RELATED WORK

A. *Multi-Robot Fire Monitoring and Suppression*

Recently, interest has been growing in applying multi-robot systems to various real-world situations that are hazardous to humans, such as fire suppression. Multi-robot teams are well-suited to fire suppression since physically distributed robots can tackle multiple fires simultaneously, but they can also cooperate and converge to fight larger fires that are beyond the capacity of a single robot. Several works have explored using a swarm of drones for wildfire monitoring to give human commanders an understanding of the wildfire's location and dynamics [17]–[19] There have also been works exploring the action space with active fire suppression [4] and human-robot teaming [20], [21]. However, fire suppression has not yet been formulated as a problem that considers conflicts between team goals and the individual goals that are inherent in distributed systems.

B. *Learning-Free Techniques for Multi-Robot Coordination*

Historically, multi-robot coordination has been addressed using a variety of classical, learning-free approaches. Heuristic and behavior-based methods provide computationally efficient, decentralized control by employing custom-designed rules to achieve coordinated behaviors, such as formation control [22], [23]. However, these methods are typically reactive and tend to yield only locally optimal solutions. Planning-based methods such as multi-agent path finding determine optimal, collision-free paths for all robots by efficiently managing and resolving conflicts [24], [25]. However, they can have issues with scalability for dynamic environments due to exponential time complexity and also require a centralized solver. Optimization-based methods formulate multi-robot coordination as an optimization problem with robot actions as variables and environment dynamics as constraints [26], [27]. Although successful for smaller robot teams, they may also have high computation complexity with increasing number of robots and constraints, and require recomputation when unexpected changes to the environment occur.

C. *Multi-Agent Reinforcement Learning (MARL)*

1) Cooperative MARL: Recently, various multi-agent reinforcement learning (MARL) techniques were implemented to address this multi-robot coordination challenge. Classical methods focus on purely cooperative settings, where a single team objective guides the learning process for all agents [28]–[30]. However, these techniques often struggle with credit assignment, which can lead to suboptimal policies involving "lazy agents" that contribute little, and result in an unbalanced and unfair utilization of the robot team [9], [10]. To address these challenges from an individual robot's perspective, other methods formulate the problem with individual goals, rewarding each agent for its own performance [11], [12]. However, optimizing purely for individual goals places the burden of discovering cooperation entirely on the agents. In situations where a fire cannot be suppressed by a single robot, an agent may need to exhibit altruistic behavior—choosing to help other robots at the potential cost of its immediate individual reward—a complex strategy that is difficult to learn without explicit guidance.

2) Mixed-Motive MARL: To bridge this gap, mixed-motive MARL formulations have been proposed that aim to jointly optimize both team (collective) goals and individual goals [31]–[33]. However, there still exist several limitations to the state-of-the-art mixed-motive algorithms. Many methods either rely on heavily engineered reward functions that only work for specific problems [34]–[36] or that require estimations of counterfactuals (i.e., what would have happened if an agent chose a different action) that are difficult to compute in complex, real-world scenarios [13], [14]. Other approaches use gradient-based methods to combine collective and individual rewards but require differentiable objective functions to implement [15], [16]. The interplay between learning fundamental individual skills and learning high-level collaborative strategies has not yet been fully solved by existing learning-based methods, especially for deployment on physical systems.

3) Curriculum Learning for MARL: Curriculum learning is an extensively studied technique in single-agent reinforcement learning where a sequence of tasks of varying difficulty is curated to reduce the exploration time needed by an agent. However, only recently has it been explored in the multi-agent setting. A common strategy is to do population-based curriculum learning [37], [38] to improve generalization and scalability, where the number of agents is small at the start of training and is progressively increased either through a fixed or automatically determined schedule based on learning progress [39], [40]. Curriculum learning has also been applied in robotics for multi-agent path finding [41],

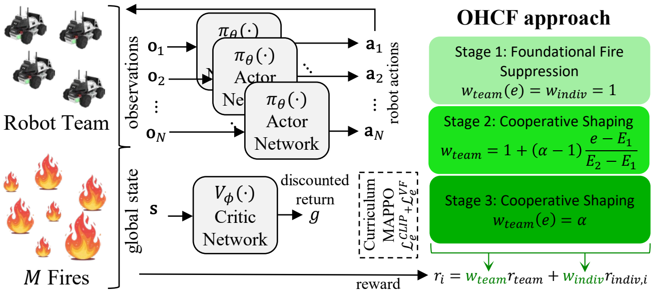

Fig. 2. Overview of our OHCF that unifies MARL with curriculum learning.

[42] and collision avoidance [43], where learning is achieved in different stages. Although progress in curriculum learning for MARL is promising, curriculum learning has not yet been applied with the objective of resolving conflicts between team goals and individual agent goals, such as in mixed-motive learning.

III. APPROACH

Notation. We represent vector spaces as calligraphic letters (e.g., $\mathcal{A}$), matrices as boldface uppercase letters (e.g., $\mathbf{B} \in \mathbb{R}^{n\times m}$, a matrix of size $n \times m$), vectors as boldface lowercase letters (e.g., $\mathbf{c} = \{c_i\} \in \mathbb{R}^d$, a d-dimensional vector whose i-th element is c_i), scalars as lowercase letters (e.g., e), and integer constants as uppercase letters (e.g., F).

A. Classic Problem Formulation

We consider the problem of multi-robot fire suppression in a large field environment, where a team of N robots must coordinate and collaborate to suppress M fires distributed across a continuous space. Each robot is represented as a circle with diameter d_i. The i-th robot's state includes its position, denoted by $\mathbf{x}_{r,i} \in \mathbb{R}^2$, and its velocity, denoted by $\mathbf{v}_i \in \mathbb{R}^2$. Each of the robots is controlled via a continuous acceleration command, $\mathbf{a}_i \in \mathbb{R}^2$. The robots' positions, velocities, and accelerations are related using standard kinematics: $\dot{\mathbf{x}}_{r,i} = \mathbf{v}_i$ and $\dot{\mathbf{v}}_i = \mathbf{a}_i$. To prevent collisions, we define a collision constraint $||\mathbf{x}_{r,i}(t) - \mathbf{x}_{r,k}(t)||_2 \geq d_i$, which must be satisfied for all robots satisfying $i \neq k$ at all times t. Each robot also has a fire suppression capacity c_i.

We denote the j-th fire's position as $\mathbf{x}_{f,j} \in \mathbb{R}^2$, with a radius r_j. Each fire also has a fire strength f_j. At time t, we denote the state of the fire as q_j^t, where $q_j^t = 1$ if the j-th fire is active or $q_j^t = 0$ if it has been suppressed. Initially, all fires are active; i.e., $q_j^0 = 1 \ \forall \ j$. An active fire becomes suppressed if the sum of the fire suppression capacities of all robots within the fire's radius exceeds the fire's strength. Mathematically, this can be expressed as $\sum_{i \in \mathcal{N}_j(t)} c_i \geq f_j(t)$, where $\mathcal{N}_j^t = \{i \mid ||\mathbf{x}_{r,i}^t - \mathbf{x}_{f,j}^t||_2 \leq r_j\}$ is the set of robots within the radius of fire j at time t.

We first define the *collective team's goal* as minimizing the number of active fires after T time steps through a sequence of coordinated robot team actions:

$$\min_{\mathbf{A}^1,\ldots,\mathbf{A}^T} \left[\sum_{j=1}^{M} f_j(T) \cdot q_j^T \right] \tag{1}$$

where $\mathbf{A}^t = (\mathbf{a}_1^t, \ldots, \mathbf{a}_N^t)$ denotes the joint action of the N robots at time t. At the same time, from the perspective of individual robots, we define each *individual robot's goal* as maximizing the total fire strength suppressed by robot i over T time steps through the sequence of its individual actions:

$$\max_{\mathbf{a}_i^1,\ldots,\mathbf{a}_i^T} \left[\sum_{t=1}^{T} \sum_{j=1}^{M} \frac{c_i}{\sum_{k \in \mathcal{N}_j^t} c_k} \cdot q_j^{t-1}(1 - q_j^t) \cdot f_j^t \right] \tag{2}$$

where $\frac{c_i}{\sum_{k \in \mathcal{N}_j^t} c_k}$ is the proportional attribution factor based on the robots that suppressed the fire.

The team's goal and the individual robot's goal may often conflict with each other, particularly when addressing fire suppression. Mathematically, solving Eqs. (1)-(2) simultaneously can lead to conflicts leading to a multi-robot optimization dilemma. In particular, the fully coordinated optimal actions that solve Eq. (1) and minimize the number of active fires are not necessarily optimal actions for solving the self-interested individual objective in Eq. (2). For example, some robots might be more heavily utilized than others, while some may remain idle. Conversely, optimal individual robot actions for maximizing the objective in Eq. (2) may not necessarily result in minimizing active fires for Eq. (1). For instance, two robots might try to competitively suppress a nearby small fire to optimize their individual goals without realizing the redundancy.

B. Fire Suppression as Multi-Agent Reinforcement Learning

We reformulate multi-robot fire suppression as a decentralized partially observable Markov decision process (Dec-POMDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{O}, \gamma \rangle$, where $\mathcal{S}$ is the state space of robots and fires, $\mathcal{A}$ is the joint action space of all robot actions $\mathbf{A}$, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the state transition function, $\mathcal{R}$ is the joint reward function of the robot team, $\mathcal{O}$ is the joint observation space of the robots, and γ is the discount factor $\gamma \in [0, 1)$ is used to value future rewards.

Observation. Each robot i perceives the environment state with respect to its own reference frame. The observation $\mathbf{o}_i$ includes its own state (i.e., position $\mathbf{x}_{r,i}$, velocity $\mathbf{v}_i$, and capacity c_i), the relative positions of other robots (i.e., $\mathbf{x}_{r,i} - \mathbf{x}_{r,k} \ \forall i \neq k$), the relative positions of all fires (i.e., $\mathbf{x}_{r,i} - \mathbf{x}_{f,j} \ \forall j$), and the net fire strength of all fires. The joint observation of the robot team is $\mathbf{O}^t = (\mathbf{o}_1^t, \ldots, \mathbf{o}_N^t) \in \mathcal{O}$.

Action. The action $\mathbf{a}_i$ for robot i is the continuous acceleration command $\mathbf{a}_i \in \mathbb{R}^2$. The joint action of the whole robot team is $\mathbf{A}^t = (\mathbf{a}_1^t, \ldots, \mathbf{a}_N^t) \in \mathcal{A}$.

Reward Function. To harmonize the team and individual objectives and address their conflicts, we design a composite reward function. A shared *collective reward*, r_{team}, is provided to all robots to promote the team objective from Eq. (1). This reward is granted whenever any fire is suppressed:

$$r_{team} = \sum_{j=1}^{M} f_j \cdot q_j^{t-1}(1 - q_j^t) + p_{step} \tag{3}$$

where the term $q_j^{t-1}(1-q_j^t)=1$ only if j was active at time $t-1$ and is suppressed at time t, and p_{step} is a small penalty applied at each timestep to encourage faster fire suppression. We also define the *individual reward*, $r_{indv,i}$, for robot i to promote the individual objective from Eq. (2), encouraging efficient and fair participation while avoiding collisions:

$$r_{indv,i} = \sum_{j=1}^{M} \frac{c_i}{\sum_{k\in\mathcal{N}_j^t} c_k} \cdot q_j^{t-1}(1-q_j^t)\cdot f_j + p_{coll} \tag{4}$$

where p_{coll} is a penalty applied to discourage robot actions that lead to collisions. Our goal is to maximize the expected discounted return by $g=\sum_{t=0}^{T}\gamma^t\left(r_{team}^t+\sum_{i=1}^{N}r_{indv}^t\right)$, which considers both r_{team} and r_{indv}. Jointly optimizing both r_{team} and r_{indv} is critical in order to train a policy that considers both team and individual objectives, especially due to potential conflicts that may arise. However, naively combining the two rewards can lead to suboptimal policies that excels at neither objective.

Decentralized Actor Networks. Each robot i is equipped with a decentralized actor network, $\pi_\theta(\mathbf{a}_i|\mathbf{o}_i)$, which maps the robot's local observation $\mathbf{o}_i$ to a probability distribution over its continuous action space $\mathbf{a}_i$. We implement the actor network with multilayer perceptron networks (MLPs) whose parameters θ are shared across all robots for sample efficiency.

Centralized Critic Network. In order to stabilize training and efficiently coordinate learning as a team, we leverage a single centralized critic network, V_ϕ, that has visibility of all robots' observations and environment states. V_ϕ takes the global state $\mathbf{s}$ as input and outputs a single scalar value estimating the expected discounted return g. The critic provides a shared baseline for all robots, enabling more informed policy updates by evaluating actions from a team-wide perspective. We implement the critic as an MLP with parameters ϕ. In this work, we define the global state as the joint observation $\mathbf{O}^t=(\mathbf{o}_1^t,\dots,\mathbf{o}_N^t)$ represented as a vector.

C. Objective Harmonization Curriculum

We discuss our novel *Objective Harmonization Curriculum for Fire Suppression* (OHCF) approach for learning a multi-robot fire suppression policy that simultaneously addresses conflicts between team objectives and individual robot objectives. The goal of OHCF is to adaptively balance learning coordinated behaviors for collaborative fire suppression with promoting the development of individual robot actions. An overview of our approach is shown in Fig. 2.

1) Curriculum Learning for Objective Harmonization: We propose a new curriculum-based learning strategy that dynamically adjusts the reward signal to guide the learning process from simple, individualistic behaviors to complex, cooperative strategies. We define the total reward for robot i at time t as a weighted sum of the components:

$$r_i = w_{team}r_{team} + w_{indv}r_{indv,i} \tag{5}$$

Then, the curriculum is designed to progress through three distinct stages by adjusting the weights w_{team} and w_{indv} over the course of training. These stages are defined in terms of the current *episode* e, with the lengths of each stage denoted as E_1, E_2, and E_3, respectively.

Stage 1: Foundational Fire Suppression Learning. In the initial phase of training (episodes $e\in[0,E_1]$), we set the weights to be balanced: $w_{team}=w_{indv}=1.0$. This reward structure encourages each robot to learn the fundamental task-related skills of navigating towards fires and participating in their suppression, driven by an equal consideration of both team success and personal contribution.

Stage 2: Cooperative Shaping. In the second stage (episodes $e\in[E_1,E_2]$), we gradually shift the policy's focus towards the collective team objective. We maintain the individual weight at $w_{indv}=1.0$ while linearly increasing the team reward weight according to the schedule:

$$w_{team}(e) = 1+(\alpha-1)\frac{e-E_1}{E_2-E_1} \tag{6}$$

where $\alpha>1$ denotes the target weight for the team reward. This curriculum smoothly steers the learned policies towards discovering more sophisticated cooperative actions that are advantageous for the team. Meanwhile, keeping w_{indv} in the objective allows us to build upon the policy developed in Stage 1 and prevents the agents from forgetting learned individual strategies.

Stage 3: Policy Refinement. During the final training stage (episodes $e>E_2$), the weights are held constant at $w_{team}=\alpha$ and $w_{indv}=1.0$. This enables the decentralized policies to stabilize and converge, fine-tuning their behavior under a harmonized objective that strongly prioritizes collective success while still accounting for individual efficiency and fairness.

2) Centralized Curriculum Training and Decentralized Execution: We address our curriculum multi-robot learning problem in the paradigm of Centralized Training for Decentralized Execution (CTDE). In this paradigm, robots leverage global information during a centralized training phase to learn coordinated policies, while executing them in a decentralized manner using only their local observations.

MAPPO Optimization. We employ Multi-Agent Proximal Policy Optimization (MAPPO) [29], an on-policy actor-critic algorithm for multi-agent settings, within our OHCF approach. MAPPO extends single-agent proximal policy optimization (PPO) [44] to deal with multiple independent robots. MAPPO's objective is to maximize a clipped surrogate objective function which encourages limited updates to the policy preventing sudden changes from destabilizing policy learning.

The per-agent actor loss is formulated as a clipped objective $\mathcal{L}^{CLIP}=\mathbb{E}\left[\min\left(\rho_t(\theta_i),\text{clip}(\rho_t(\theta_i),1-\epsilon,1+\epsilon)\right)\hat{A}_i^t\right]$, where $\hat{A}_i^t$ denotes the generalized advantage estimation [45] for each robot i at time step t, $\rho^t(\theta_i)=\frac{\pi_{\theta_i}(\mathbf{a}_i^t|\mathbf{o}_i^t)}{\pi_{\theta_{i,\text{old}}}(\mathbf{a}_i^t|\mathbf{o}_i^t)}$ denotes the probability ratio between the new and old policies, and ϵ is a hyperparameter that defines a clipping range, which constrains the magnitude of policy updates. Meanwhile, the centralized critic is updated by minimizing the mean-squared error between its predictions and the observed returns by

$\mathcal{L}^{VF}(\phi) = \mathbb{E}\left[(V_\phi(\mathbf{s}) - r_{target})^2\right]$, where r_{target} denotes the observed return. During training, we collect batches of trajectories, compute the advantages, and perform multiple epochs of stochastic gradient ascent on the actor objective alongside gradient descent on the critic loss.

Unified Learning for Objective Harmonization. We integrate curriculum learning directly into the MAPPO optimization process by conditioning the dynamic weighting of team and individual rewards on the current stage of the curriculum. The composite reward for robot i at time t during episode e is explicitly defined by the curriculum stage:

$$r_i(e) = w_{team}(e) r_{team} + w_{indv}(e) r_{indv,i} \tag{7}$$

where $w_{team}(e)$ and $w_{indv}(e)$ are determined by the schedules in Stages 1, 2, or 3. This curriculum-dependent reward directly influences the calculation of the advantage function. We modify the advantage for robot i into $\hat{A}_i^t(e)$ which is now also a function of the current episode e. Similarly, we also modify the actor's learning objective by conditioning on the curriculum and reflecting the current curriculum stage:

$$\mathcal{L}_e^{CLIP} = \mathbb{E}\left[\min\left(\rho^t(\theta_i), \text{clip}(\rho^t(\theta_i), 1-\epsilon, 1+\epsilon)\right)\hat{A}_i^t(e)\right] \tag{8}$$

The critic must similarly learn a non-stationary value function that tracks the shifting rewards during curriculum learning. Its loss also depends on the curriculum stage through the target value calculation:

$$\mathcal{L}_e^{VF} = \mathbb{E}\left[(V_\phi(\mathbf{s}) - r_{\text{target}}(e))^2\right] \tag{9}$$

where the target return $r_{\text{target}}(e)$ is computed as the discounted sum of future rewards weighted according to the curriculum. This unified learning process ensures that both the actor and critic networks adapt in concert with the curriculum, resulting in a stable and effective harmonization of the conflicting team and individual objectives.

During decentralized execution, each robot i samples its action $\mathbf{a}_i^t \sim \pi_\theta(\mathbf{a}_i^t|\mathbf{o}_i^t)$ based on its individual observation, which includes its own position and velocity, as well as the relative positions of other robots and fires. This observation information is assumed to be shared within the multi-robot team, e.g., through broadcasting. The critic $V_\phi(\mathbf{s})$ is used exclusively during training and not during execution.

IV. Experiments

To implement OHCF, the curriculum learning component is programmed with TorchRL [46] and BenchMARL [47] as a new algorithm. We build our Firefighting environment using VMAS [48] and represent robots as circular agents and fires as circular targets. We physically model the environment as a rectangular space of size 0.9×1.6. Each iteration in the simulation has a duration of 0.1s. Each robot has a radius of 0.05 while each fire has a radius of 0.25. For our experiments, we use four robots each with a suppression capacity of $c_i = 1$. We define ten fires where there are five with fire strength $f_j = 1$ rendered as yellow circles, three with fire strength $f_j = 2$ rendered as orange circles, and two with fire strength $f_j = 3$ rendered as red circles. The actor and critic networks are represented as 2-layer MLPs with 256 neurons each, with the actor network having a hyperbolic tangent activation function to limit acceleration actions to ± 1.0.

TABLE I

Quantitative results of OHCF in simulation.

Method	Success Rate ↑	Mean Steps ↓	Distance (m) ↓	Energy (J) ↓
Heuristic-Nearest	41.0%	276.79	**5.56**	**14.99**
Heuristic-Random	60.0%	223.73	9.86	38.68
Collective-only	94.5%	195.09	27.70	127.91
Individual-only	53.0%	250.93	50.48	299.52
Static-weights	97.0%	128.24	25.70	154.54
OHCF (ours)	**99.5%**	**107.71**	21.80	135.63

To train OHCF, we spawn 1,500 parallel environments and run for 100 iterations with each episode truncated at 400 timesteps for a total of 60,000,000 training timesteps. At the start of each training episode, the robots and fires are all placed randomly in the environment as our domain randomization procedure. Model training and execution are performed on 8-core Intel machines with 32GB RAM and an NVIDIA RTX 2080Ti graphics card. A single training run of OHCF while using the GPU takes approximately 3.5 hours, while execution of the policy on CPU can run at around 10 Hz.

We evaluate OHCF both in a 2D vectorized simulation environment and on a real-robot deployment running Robot Operating System (ROS). We compare OHCF against five baseline approaches. The first two are heuristic methods: (1) **Heuristic-Nearest**, where each robot attempts to suppress the nearest fire, (2) **Heuristic-Random**, where each robot stochastically chooses unsuppressed fire to approach. The next three are MARL baselines using MAPPO to illustrate the effectivity of our curriculum-based MARL formulation: (3) **Collective-only**, which exclusively optimizes the team reward r_{team}, (4) **Individual-only**, which exclusively optimizes the individual reward r_{indv}, and (5) **Static-weights**, which combines the two rewards additively as in typical MARL.

We evaluate OHCF quantitatively and compare against the baselines using four metrics: (1) **success rate**, the proportion of all scenarios where all fires were completely suppressed, (2) **mean steps**, the average length of evaluation episodes, (3) **distance**, the total distance traveled by all robots while performing fire suppression, and (4) **energy**, the total simulated energy expended by all robots computed using an energy model [49], [50] that considers mass, acceleration, and displacement (i.e., higher forces and accelerations with the same distance traveled lead to higher energy usage). Both total distance and total energy are computed in SI units by assuming each robot's physical properties are the same as an AgileX LIMO robot [51].

A. Quantitative Results of OHCF in Simulation

The quantitative results are shown in Table I. Our method achieves the highest success rate at 99.5% and also completes the fire suppression with the fewest mean steps at 107.71, which demonstrates the advantage of OHCF's curriculum

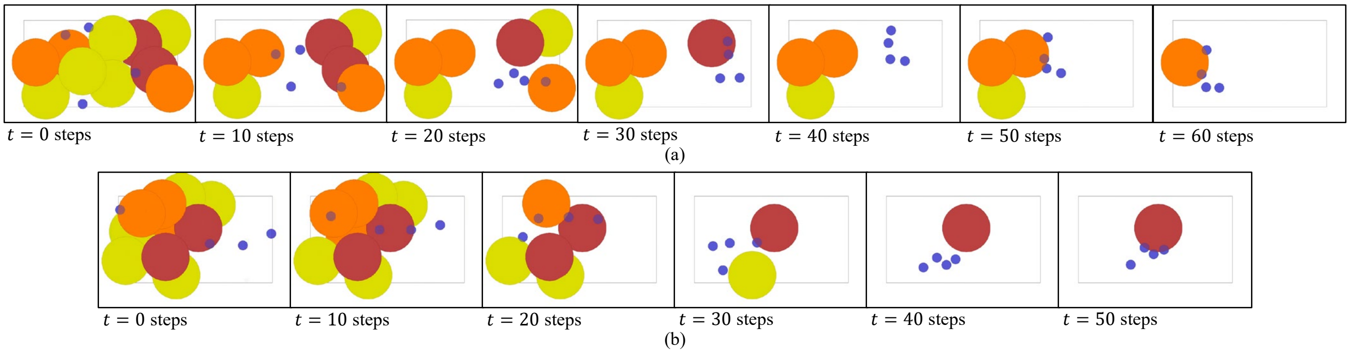

Fig. 3. Qualitative results of OHCF showing how a team of robots (represented as blue circles) collaboratively suppresses fires (represented as the larger colored circles) with strengths of 1 (yellow), 2 (orange), and 3 (red). Multiple robots initially begin suppressing low-strength fires on their own (individual goal) before forming subteams to collaboratively suppress higher-strength fires (team goal). [Best viewed in color]

learning even compared to Static-weights which combines both individual and team rewards. The heuristic-based solutions exhibit low success rates primarily due to the inability of the robots to coordinate when the need to collaborate on suppressing fires with larger strengths. This motivates the use of MARL to learn the coordination and subteaming capability. Because the Individual-only method does not have explicit coordination capability, it is unable to complete suppressing half of the scenarios and thus has a larger mean steps value. The Collective-only method performs better than Individual-only due to having a team goal of fully suppressing all fires, allowing them to collaborate when suppressing larger fires. . However, it still completes scenarios slower compared to both Static-weights and OHCF which suggests that the individual goal is important for improving fire suppression skills for the robots.

Next, we analyze the total distance traveled and energy used by the different methods. The heuristic methods' distance and energy used are low due to the robots getting stuck in fires that they cannot suppress on their own. Comparing against the MARL methods, our approach still has the lowest total distance traveled at 25.70. The Collective-only approach has slightly lower total energy use than OHCF even with longer distance traveled, which suggests that the robots travel slower with less acceleration at the expense of execution speed. On the other hand, the Individual-only approach has higher distance traveled due to problems with coordination (similar to the heuristics) and have conflicts on the larger fires that need coordination.

To validate the trained policy from OHCF, we collect steps during evaluation of OHCF and render two sample rollouts. The first rollout is shown in Fig. 3a, with the four robots represented as the small blue circles and the fires represented as the yellow, orange, and red circles. At t=10 steps, three of the $f_j = 1$ fires are suppressed by one robot each, demonstrating the ability of the robots to independently fight smaller fires. The fourth robot has started move right and wait until it is joined by the other robots to suppress one fire with $f_j = 3$ at t=20 steps and another fire with $f_j = 2$ at t=30 steps in succession. At t=40 steps, the robots move upwards to suppress the other fire with $f_j = 3$ and continue to the left side of the field. The robot team then completes suppressing a $f_j = 2$ fire and a at t=50 steps before suppressing the final fire at t=60 steps.

In the second rollout shown in Fig. 3b, the fires are spawned more concentrated towards the left side of the field. At t=10 steps, one $f_j = 1$ fire is suppressed by the leftmost robot while the other robots start moving right, At t=20 steps, two $f_j = 1$ fires have been suppressed by two robots, and a pair of robots suppressed one $f_j = 2$ fire. Since the robots are close together near the center of the field, they jointly suppress a $f_j = 3$ fire at t=30 steps and two other fires $f_j = 1$ and $f_j = 2$. At t=40 steps, they suppress a fire with $f_j = 1$ before the final $f_j = 3$ fire at t=50.

B. Real-Robot Case Study of OHCF

We employ four AgileX LIMO ROS2 mobile robots to validate OHCF in a real-world scenario. The LIMO robots are equipped with mecanum wheels for omnidirectional motion. They are velocity-controlled using the `agx_sdk` via ROS connected through the wireless network. We visualize the firefighting scenario by creating a mixed-reality setup using an overhead projected mounted to the ceiling for real-time programmable visualization, and use an OptiTrack Motion Capture system to retrieve robot pose information in order to perform closed-loop control. The projected firefighting environment has physical dimensions of 2.429×4.316m.

Our demonstration scenario is illustrated in Fig. 4. At time 0.00s, we see the four LIMO robots at their random starting positions and ten fires randomly spawned in the experiment field. They start to move according to the policy and suppress three fires of strength $f_j = 1$ by time 7.72s. At time 9.91s, the two robots on the right proceed leftwards and suppress another fire with $f_j = 1$ while the two robots on the left move upwards and also suppress another fire with $f_j = 1$. Next, the left robots collaboratively suppress the first fire with $f_j = 2$ at time 12.83s while the right robots start moving leftwards and upwards towards the center of the field. The left robots proceed to suppress the next two fires with $f_j = 2$ by first moving upward at time 15.60s then rightward at time 20.60s. At time 23.43, the two left robots are joined by one other robot and proceed rightwards to suppress the first fire with strength $f_j = 3$. Finally, the four robots all continue rightward to suppress the last fire with strength $f_j = 3$ at

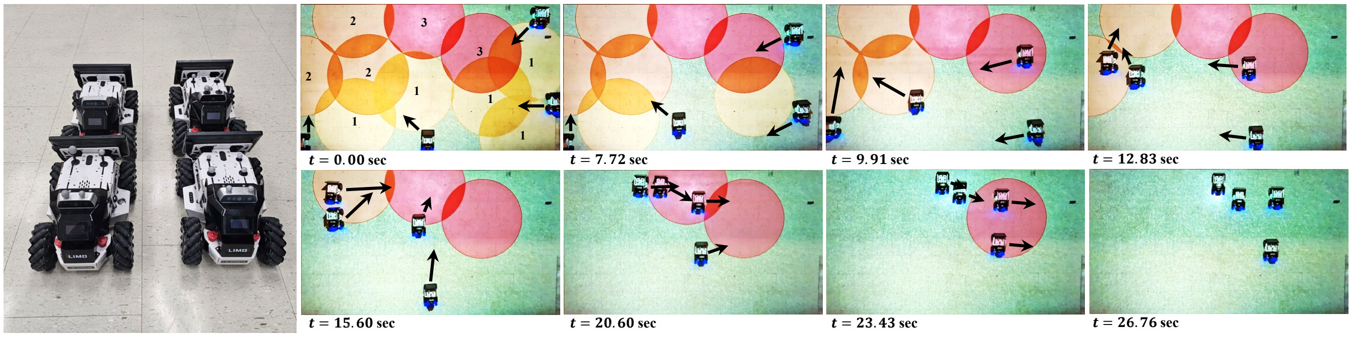

Fig. 4. Experimental validation of OHCF in a case study involving a team of four physical LIMO robots collaboratively suppressing simulated fires in real-world mixed-reality experiments.

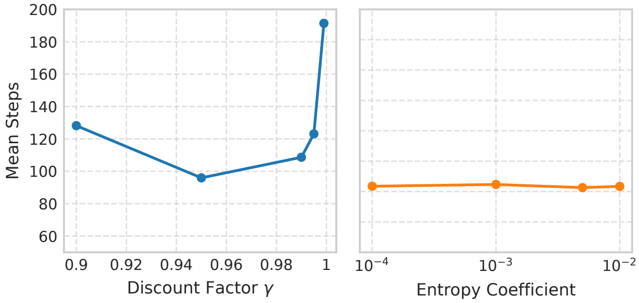

Fig. 5. Hyperparameter sensitivity analysis of OHCF with respect to the discount factor γ and the entropy coefficient.

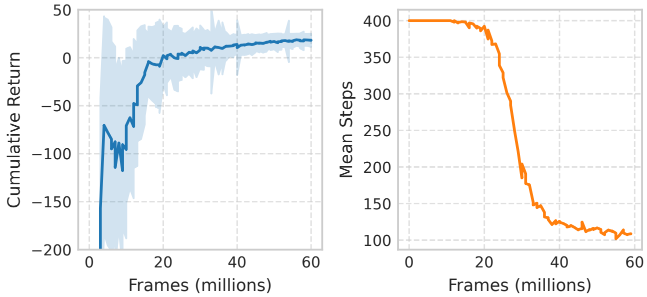

Fig. 6. Cumulative return and mean episode length of OHCF during training.

time 26.76, when the scenario is completed.

C. Discussion

1) Hyperparameter Analysis: We perform hyperparameter analysis on OHCF across the discount factor γ and entropy coefficient used to balance exploration and exploitation (Fig. 5). We see that the optimal value for the discount factor is $\gamma = 0.95$; a lower discount factor incentivizes short-term rewards (which may be useful for smaller fires) while higher discount factor slows down learning convergence. For the entropy coefficient, we do not see a significant difference across the performance, which means OHCF is not heavily dependent on exploration to successfully encounter diverse states. This may be because of the domain randomization we incorporate during training.

2) Convergence: We analyze the convergence of OHCF in Fig. 6 showing the cumulative return curve during training. OHCF's cumulative reward converges to a value around 20, unlike the baseline MAPPO methods (especially Individual-only and Collective-only) which struggles to converge well due to conflicts between the team and individual goals. We also see the mean number steps start from 400 at the start of training (i.e., all episodes are truncated while the model is still untrained) and start to go down to converge close to 100 by the end of training.

V. CONCLUSION

In this work, we propose a novel curriculum-based MARL method for multi-robot fire suppression, formulated to resolve the conflict between team-level and individual robot objectives. We introduce a three-stage curriculum that dynamically balances rewards, guiding the robots from prioritizing individual goals to optimizing for the team goal. OHCF enables a team of robots to autonomously suppress small fires and adaptively form subteams to tackle fires of higher intensity. We evaluated our OHCF approach through extensive simulations and validated its effectiveness on a decentralized physical multi-robot system. Experimental results have shown that our approach successfully addresses the objective harmonization dilemma and enables effective multi-robot fire suppression, outperforming baseline methods across a range of fire suppression scenarios.

REFERENCES

[1] A. Camisa, A. Testa, and G. Notarstefano, "Multi-Robot Pickup and Delivery via Distributed Resource Allocation," *IEEE Transactions on Robotics*, vol. 39, no. 2, pp. 1106–1118, Apr. 2023.

[2] A. Romero, C. Delgado, L. Zanzi, R. Suárez, and X. Costa-Pérez, "Cellular-enabled Collaborative Robots Planning and Operations for Search-and-Rescue Scenarios," in *IEEE International Conference on Robotics and Automation*, 2024.

[3] G. Notomista, C. Pacchierotti, and P. R. Giordano, "Multi-Robot Persistent Environmental Monitoring Based on Constraint-Driven Execution of Learned Robot Tasks," in *International Conference on Robotics and Automation*, 2022.

[4] R. N. Haksar and M. Schwager, "Distributed Deep Reinforcement Learning for Fighting Forest Fires with a Network of Aerial Robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018.

[5] J. Hwang, N.-S. Chong, M. Zhang, R. J. Agnew, C. Xu, Z. Li, and X. Xu, "Face-to-face with scorching wildfire: potential toxicant exposure and the health risks of smoke for wildland firefighters at the wildland-urban interface," *The Lancet Regional Health - Americas*, vol. 21, p. 100482, May 2023.

[6] D. Le and E. Plaku, "Multi-Robot Motion Planning With Dynamics via Coordinated Sampling-Based Expansion Guided by Multi-Agent Search," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, Apr. 2019.

[7] G. Neville, S. Chernova, and H. Ravichandar, “D-ITAGS: A Dynamic Interleaved Approach to Resilient Task Allocation, Scheduling, and Motion Planning,” *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1037–1044, Feb. 2023.

[8] R. Shome, K. Solovey, A. Dobson, D. Halperin, and K. E. Bekris, “dRRT*: Scalable and informed asymptotically-optimal multi-robot motion planning,” *Autonomous Robots*, vol. 44, no. 3, pp. 443–467, Mar. 2020.

[9] S. V. Albrecht, F. Christianos, and L. Schäfer, *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024.

[10] B. Liu, Z. Pu, Y. Pan, J. Yi, Y. Liang, and D. Zhang, “Lazy Agents: A New Perspective on Solving Sparse Reward Problem in Multi-agent Reinforcement Learning,” in *International Conference on Machine Learning*, 2023.

[11] B. Guresti, A. Vanlioglu, and N. K. Ure, “IQ-Flow: Mechanism Design for Inducing Cooperative Behavior to Self-Interested Agents in Sequential Social Dilemmas,” in *International Conference on Autonomous Agents and Multiagent Systems*, 2023.

[12] D. Ivanov, “Mediated Multi-Agent Reinforcement Learning,” in *International Conference on Autonomous Agents and Multiagent Systems*, 2023.

[13] J. L. Zhou, W. Hong, and J. C. Kao, “Reciprocal Reward Influence Encourages Cooperation From Self-Interested Agents,” in *Neural Information Processing Systems*, 2024.

[14] F. Kong, Y. Huang, S.-C. Zhu, S. Qi, and X. Feng, “Learning to Balance Altruism and Self-interest Based on Empathy in Mixed-Motive Games,” in *Neural Information Processing Systems*, 2024.

[15] Y. Li, W. Zhang, J. Wang, and S. Zhang, “Aligning Individual and Collective Objectives in Multi-Agent Cooperation,” in *Neural Information Processing Systems*, 2024.

[16] W. Kim and K. Sycara, “Fair Cooperation in Mixed-Motive Games via Conflict-Aware Gradient Adjustment,” 2025, arXiv:2508.17696 [cs].

[17] E. Seraj and M. Gombolay, “Coordinated Control of UAVs for Human-Centered Active Sensing of Wildfires,” in *American Control Conference*, 2020.

[18] E. Seraj, L. Chen, and M. C. Gombolay, “A Hierarchical Coordination Framework for Joint Perception-Action Tasks in Composite Robot Teams,” *IEEE Transactions on Robotics*, vol. 38, no. 1, Feb. 2022.

[19] E. Seraj, “Learning Efficient Diverse Communication for Cooperative Heterogeneous Teaming,” in *International Conference on Autonomous Agents and Multi-Agent Systems*, 2022.

[20] E. Seraj, A. Silva, and M. Gombolay, “Safe Coordination of Human-Robot Firefighting Teams,” 2019, arXiv:1903.06847 [cs].

[21] J. Hyun, N. R. Waytowich, and B. Chen, “CREW-WILDFIRE: Benchmarking Agentic Multi-Agent Collaborations at Scale,” 2025, arXiv:2507.05178 [cs].

[22] A. Das, R. Fierro, V. Kumar, J. Ostrowski, J. Spletzer, and C. Taylor, “A vision-based formation control framework,” *IEEE Transactions on Robotics and Automation*, vol. 18, no. 5, pp. 813–825, Oct. 2002.

[23] S. Zhao, “Affine Formation Maneuver Control of Multiagent Systems,” *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4140–4155, Dec. 2018.

[24] K. Okumura, “LaCAM: Search-Based Algorithm for Quick Multi-Agent Pathfinding,” in *AAAI Conference on Artificial Intelligence*, 2023.

[25] J. Li, A. Felner, E. Boyarski, H. Ma, and S. Koenig, “Improved Heuristics for Multi-Agent Path Finding with Conflict-Based Search,” in *International Joint Conference on Artificial Intelligence*, 2019.

[26] J. Alonso-Mora, S. Baker, and D. Rus, “Multi-robot formation control and object transport in dynamic environments via constrained optimization,” *The International Journal of Robotics Research*, vol. 36, no. 9, pp. 1000–1021, Aug. 2017.

[27] S. Park and S.-M. Lee, “Formation Reconfiguration Control With Collision Avoidance of Nonholonomic Mobile Robots,” *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 7905–7912, Dec. 2023.

[28] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, “QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning,” in *International Conference on Machine Learning*, 2018.

[29] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, “The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games,” in *Neural Information Processing Systems*, 2022.

[30] R. Lowe, Y. WU, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments,” in *Neural Information Processing Systems*, 2017.

[31] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. Dueñez-Guzman, A. García Castañeda, I. Dunning, T. Zhu, K. McKee, R. Koster, H. Roff, and T. Graepel, “Inequity aversion improves cooperation in intertemporal social dilemmas,” in *Neural Information Processing Systems*, 2018.

[32] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, and N. D. Freitas, “Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning,” in *International Conference on Machine Learning*, 2019.

[33] E. Hughes, T. W. Anthony, T. Eccles, J. Z. Leibo, D. Balduzzi, and Y. Bachrach, “Learning to Resolve Alliance Dilemmas in Many-Player Zero-Sum Games,” in *International Conference on Autonomous Agents and Multiagent Systems*, 2020.

[34] N. Anastassacos, J. García, S. Hailes, and M. Musolesi, “Cooperation and Reputation Dynamics with Reinforcement Learning,” in *International Conference on Autonomous Agents and Multiagent Systems*, 2021.

[35] E. Vinitsky, R. Köster, J. P. Agapiou, E. A. Duéñez-Guzmán, A. S. Vezhnevets, and J. Z. Leibo, “A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings,” *Collective Intelligence*, vol. 2, no. 2, p. 263391372311620, Apr. 2023.

[36] A. Lupu and D. Precup, “Gifting in Multi-Agent Reinforcement Learning,” in *International Conference on Autonomous Agents and Multiagent Systems*, 2020.

[37] Q. Long, Z. Zhou, and A. Gupta, “Evolutionary Population Curriculum for Scaling Multi-Agent Reinforcement Learning,” in *International Conference on Learning Representations*, 2020.

[38] W. Wang, T. Yang, Y. Liu, J. Hao, X. Hao, Y. Hu, Y. Chen, C. Fan, and Y. Gao, “From Few to More: Large-Scale Dynamic Multiagent Curriculum Learning,” in *AAAI Conference on Artificial Intelligence*, 2020.

[39] J. Chen, Y. Zhang, Y. Xu, H. Ma, H. Yang, J. Song, Y. Wang, and Y. Wu, “Variational Automatic Curriculum Learning for Sparse-Reward Cooperative Multi-Agent Problems,” in *Neural Information Processing Systems*, 2021.

[40] W. Zhao, Z. Li, and J. Pajarinen, “Learning Progress Driven Multi-Agent Curriculum,” in *International Conference on Machine Learning*, 2025.

[41] T. Phan, “Confidence-Based Curriculum Learning for Multi-Agent Path Finding,” in *International Conference on Autonomous Agents and Multiagent Systems*, 2024.

[42] C. Zhao, L. Zhuang, Y. Huang, and H. Liu, “Curriculum Learning Based Multi-Agent Path Finding for Complex Environments,” in *International Joint Conference on Neural Networks*, 2023.

[43] M. M. R. Komol, B. Tidd, W. Browne, F. Maire, J. Williams, and D. Howard, “Learning Behaviours for Decentralised Multi-Robot Collision Avoidance in Constrained Pathways Using Curriculum Reinforcement Learning,” *IEEE Robotics and Automation Letters*, vol. 10, no. 8, pp. 8538–8545, Aug. 2025.

[44] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” 2017, arXiv:1707.06347 [cs].

[45] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-Dimensional Continuous Control Using Generalized Advantage Estimation,” in *International Conference on Learning Representations*, 2016.

[46] A. Bou, M. Bettini, S. Dittert, V. Kumar, S. Sodhani, and X. Yang, “TorchRL: A Data-Driven Decision-Making Library for PyTorch,” in *International Conference on Learning Representations*, 2024.

[47] M. Bettini, A. Prorok, and V. Moens, “BenchMARL: Benchmarking Multi-Agent Reinforcement Learning,” *Journal of Machine Learning Research*, vol. 25, no. 217, pp. 1–10, 2024.

[48] M. Bettini, R. Kortvelesy, J. Blumenkamp, and A. Prorok, “VMAS: A Vectorized Multi-Agent Simulator for Collective Robot Learning,” in *International Symposium on Distributed Autonomous Robotic Systems*, 2022.

[49] S. Liu and D. Sun, “Minimizing Energy Consumption of Wheeled Mobile Robots via Optimal Motion Planning,” *IEEE/ASME Transactions on Mechatronics*, vol. 19, no. 2, pp. 401–411, Apr. 2014.

[50] Y. Mei, Y.-H. Lu, Y. Hu, and C. Lee, “Deployment of mobile robots with energy and timing constraints,” *IEEE Transactions on Robotics*, vol. 22, no. 3, pp. 507–522, June 2006.

[51] “LIMO ROS2.” [Online]. Available: https://global.agilex.ai/products/limo-ros2