Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Topology-Aware Multi-Agent Reinforcement Learning for Cooperative Multi-UAV Coverage Path Planning

Li Wang
University of Science and Technology of China
Wenjie Gan
Southeast University
Junhao Liang
Southeast University
Wanyuan Wang
wywang@seu.edu.cn
Southeast University
Yanyan Liu
North Information Control Research Academy Group Co., Ltd.

Research Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11009922/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Topology-Aware Multi-Agent Reinforcement Learning for Cooperative Multi-UAV Coverage Path Planning

Li Wang[1,3], Wenjie Gan[2], Junhao Liang[4], Wanyuan Wang[2*], Yanyan Liu[3*]

[1]School of Cyber Science and Technology, University of Science and Technology of China, Hefei, 230026, Anhui, China.
[2]School of Computer Science and Engineering, Southeast University, Nanjing, 211189, Jiangsu, China.
[3]North Information Control Research Academy Group Co., Ltd., Nanjing, 211153, Jiangsu, China.
[4]School of Software Engineering, Southeast University, Nanjing, 211189, Jiangsu, China.

*Corresponding author(s). E-mail(s): wywang@seu.edu.cn; 22377069@qq.com;

Abstract

Cooperative area coverage by multiple unmanned aerial vehicles (UAVs) requires efficient exploration while avoiding collisions and maintaining communication within each group. Existing coverage methods are mainly rule-based or learning-based. Rule-based methods depend heavily on handcrafted heuristics and have limited adaptability to complex environments, whereas learning-based methods often couple exploration planning with motion control in a single policy, increasing the difficulty of policy learning. As the UAV fleet size increases, group-based organization becomes increasingly important for scalable coordination. However, most existing methods still treat UAVs as a flat swarm and lack explicit mechanisms to distinguish intra-group interactions from inter-group coordination. To address these limitations, we develop a topology-aware hierarchical coverage framework tailored to fixed-group multi-UAV exploration. First, a planning–execution hierarchy decouples frontier-based target assignment from low-level motion control, thereby reducing the policy-learning burden. Second, the execution layer employs a group-aware hierarchical topology encoder. Its sequential Graphormer modules construct UAV-centric action representations

by first encoding intra-group communication dependencies and then inter-group exploration relationships, enabling complementary exploration across groups while preserving intra-group communication. We evaluate the proposed method and the baselines on 200 unseen randomly generated maps, comprising 50 Easy, 100 Medium, and 50 Hard maps. On the Hard maps, our method achieves a final coverage rate (FCR) of 91.33%, exceeding the strongest reinforcement-learning-based baseline by 14.98 percentage points, while limiting the collision rate (CR) and communication violation rate (CVR) to 1.81% and 9.32%, respectively, with the lowest CVR among all compared methods.

Keywords: Multi-UAV coverage, multi-agent reinforcement learning, hierarchical decision-making, topology-aware

1 Introduction

In tasks such as area reconnaissance, disaster search and rescue, and infrastructure inspection, rapid and comprehensive information acquisition is essential for target identification, risk detection, and subsequent operation planning (Lyu et al 2023; Nguyen et al 2024). Owing to limitations in flight speed, sensing range, and endurance, a single unmanned aerial vehicle (UAV) can hardly cover a large area within a limited time, whereas multiple UAVs operating in parallel can substantially improve information-acquisition efficiency (Song et al 2023; Rahman et al 2025). However, as the fleet size increases, coordinating all UAVs as a single swarm becomes increasingly difficult, while limited communication ranges may prevent persistent direct communication among all UAVs. The UAVs can therefore be organized into groups according to communication connectivity or task relationships (Poudel and Moh 2023), enabling close information exchange and cooperation within each group while allowing different groups to coordinate their exploration of the same task area. Consequently, efficient cooperative coverage by multiple UAV groups is an important problem in autonomous information acquisition within complex environments.

Multi-group UAV coverage presents two coupled coordination challenges. At the inter-group level, different groups should explore complementary unexplored regions to reduce redundant coverage and accelerate global exploration. At the intra-group level, UAVs should disperse sufficiently to expand the covered region while remaining close enough to preserve communication. Excessive dispersion may break intra-group communication links, whereas overly compact formations can cause local clustering and repeated coverage (Güven and Yanmaz 2024). Meanwhile, each UAV must avoid obstacles and inter-UAV collisions while adapting its movement to the evolving coverage state (Luna et al 2024; Ramezani et al 2024). Consequently, efficient multi-group coverage requires complementary exploration across groups and communication-aware motion coordination within each group.

Existing cooperative coverage methods mainly fall into two categories (Frattolillo et al 2023; Rahman et al 2025). (1) Rule-based methods, such as Boustrophedon decomposition (Gong et al 2024), frontier-based exploration (Tran et al 2022), and artificial potential fields (Wang et al 2024), generate coverage paths using predefined

geometric or control rules (Xiao et al 2023; Du et al 2023; Wang and Wang 2024). These methods primarily aim to maximize coverage but rarely explicitly maintain intra-group communication (Li et al 2024a; Ahmed et al 2024). Moreover, collision avoidance is often triggered only when a potential collision is imminent, causing UAVs to spend additional time adjusting their movements and thereby reducing coverage efficiency. (2) Reinforcement-learning-based methods have introduced task-specific mechanisms for adaptive multi-UAV coverage. An IPPO-based UAV-cluster planner combines local and global information to improve coverage efficiency (Wang et al 2025), while DDMDAC exchanges the states of connected UAVs to support decentralized coverage decisions (Aydemir and Cetin 2023). Other studies develop zero-shot coverage policies for unseen maps and changing UAV networks (Carvalho and Aguiar 2026), or use sparse dual-attention to selectively encode targets, obstacles, and neighboring UAVs (Chen et al 2026). Although these methods improve environmental adaptability and cooperative decision-making (Westheider et al 2023), they generally treat UAV interactions through unified information sharing or attention aggregation, without explicitly separating intra-group communication dependencies from inter-group exploration coordination. This may result in redundant exploration across groups or communication violations within groups.

To address these limitations, we propose a hierarchical cooperative decision framework for multi-group UAV coverage. The framework separates the decision process into a frontier-guided planning layer and a topology-aware execution layer. (1) The planning layer determines where the UAVs should explore by extracting candidate targets from the evolving coverage boundary and evaluating them according to travel distance and exploration value. This process supplies each UAV with an explicit exploration direction and reduces the burden of learning long-term target selection entirely from delayed coverage rewards. (2) The topology-aware execution layer combines the assigned target and local observation with a hierarchical topology representation to generate motion actions. Specifically, we construct two Graphormer-based topology-aware modules for intra-group and inter-group coordination. The intra-group Graphormer captures communication dependencies and member distributions within each group, while the inter-group Graphormer further models exploration relationships across different groups. Together, these modules enable the execution layer to adapt motion decisions to local environmental conditions while preserving intra-group communication and reducing redundant exploration across groups.

We evaluate the method on 200 unseen randomly generated maps, comprising 50 Easy, 100 Medium, and 50 Hard maps with progressively increasing environmental complexity. The experiments demonstrate the consistent performance of our method across these difficulty levels. On the challenging Hard maps, our method achieves a final coverage rate (FCR) of 91.33%, exceeding the strongest reinforcement-learning-based baseline by 14.98 percentage points, while maintaining a collision rate (CR) of only 1.81% and the lowest communication violation rate (CVR) of 9.32%. These results demonstrate its ability to balance coverage performance, collision avoidance, and communication maintenance. Ablation studies further confirm the distinct roles of intra-group topology modeling in communication maintenance and inter-group topology modeling in coordinated exploration.

The main contributions of this work are summarized as follows:

- We devise a two-layer cooperative decision framework in which the upper layer performs frontier-guided exploration-target assignment, while the lower layer learns target-conditioned motion control. This hierarchy reduces the policy-learning burden by separating global exploration planning from local action execution.
- We introduce a group-aware hierarchical topology encoder that explicitly differentiates intra-group communication dependencies from inter-group exploration relationships. Cascaded Graphormer modules derive UAV-centric action representations from each UAV's decision-making perspective, supporting topology-aware coordination within groups and complementary exploration across groups.
- Extensive evaluations establish the superior coverage performance and coordination safety of our method, while targeted ablation studies quantify the respective contributions of intra-group and inter-group topology modeling.

2 Related Work

2.1 Rule-Based Coverage and Exploration Methods

Rule-based multi-robot coverage methods generally construct trajectories through geometric decomposition, frontier selection, task allocation, or explicit optimization. Boustrophedon decomposition partitions the environment into manageable cells and assigns the resulting coverage paths among multiple robots. Recent work further improves its robustness through online path redistribution when robots fail or leave the team (Gong et al 2024). Energy-aware extensions jointly optimize coverage routes and flight speeds while considering the battery constraints of individual UAVs (Datsko et al 2024). For large graph-based environments, local-search methods directly optimize multiple coverage paths to reduce the makespan (Tang and Ma 2024), while informative path planning combines area partitioning with coordinated search trajectories to improve search efficiency in open fields (Kim and Kim 2025).

For unknown environments, frontier-based methods repeatedly identify unexplored boundaries and assign them as navigation targets. Frontier-led swarming distributes UAVs across different unexplored regions (Tran et al 2022), while frontier clustering and task allocation can reduce redundant exploration and recover from robot failures (Yan et al 2023). RACER combines decentralized space decomposition, workload allocation, and hierarchical trajectory planning for rapid multi-UAV exploration under limited communication (Zhou et al 2023). Sampling-based methods such as HMS-RRT improve frontier extraction and collaborative exploration using multiple RRT strategies (Ning et al 2024), whereas artificial-potential-field approaches provide computationally efficient online coverage and local collision avoidance (Wang et al 2024). These methods are interpretable and require no policy training, but their performance depends heavily on hand-crafted costs, allocation rules, and reactive motion strategies. More importantly, rule-based methods often struggle to achieve efficient exploration while simultaneously maintaining intra-group communication and operational safety.

2.2 Reinforcement-Learning-Based Coverage Methods

Reinforcement learning enables robots to acquire coverage and coordination policies through interaction with the environment. Connected-agent reinforcement learning incorporates neighboring agents into the decision process to improve dynamic area coverage (Aydemir and Cetin 2023), while communication-efficient distributed Q-learning selectively exchanges frontier-related information instead of complete maps or Q-tables (Latif and Parasuraman 2024). Parameter-sharing multi-agent reinforcement learning has been applied to multi-UAV path planning and following (Zhao et al 2024b) and multi-AUV cooperative search in unknown environments (Li et al 2024b). Attention-based policies further aggregate spatial and inter-agent information for cooperative UAV coverage (Chen et al 2026), while dynamic UAV networks improve transfer to unseen team configurations (Carvalho and Aguiar 2026). Although these methods improve adaptive coverage and multi-agent coordination, most of them organize UAV interactions through a flat coordination structure. Such a structure can be effective for relatively small teams, but its coordination complexity increases as the number of UAVs grows, because all agents are modeled under a unified interaction pattern despite their potentially different coordination roles. Group-based organization provides a more structured way to decompose large UAV teams and support scalable cooperation. However, existing methods rarely explicitly distinguish intra-group communication maintenance from inter-group exploration coordination. In contrast, our method combines frontier-guided planning with a topology-aware hierarchical encoder that separately models intra-group communication dependencies and inter-group exploration relationships.

3 Problem Formulation

3.1 Scenario Definition

We consider a cooperative coverage task in a two-dimensional grid environment. The environment is represented by an occupancy map $\mathbf{M} \in \{0,1\}^{H\times W}$, where H and W denote the map dimensions. The set of traversable cells is denoted by $\mathcal{F}$, while the remaining cells correspond to obstacles.

Let $\mathcal{U} = \{1,\dots,N\}$ denote a team of homogeneous UAVs. The team is partitioned into K groups $\mathcal{G}_1,\dots,\mathcal{G}_K$, satisfying

$$\bigcup_{k=1}^{K} \mathcal{G}_k = \mathcal{U}, \qquad \mathcal{G}_k \cap \mathcal{G}_l = \varnothing, \quad k \neq l. \tag{1}$$

The position of UAV i at time step t is denoted by $\mathbf{p}_i^t \in \mathcal{F}$. Each UAV senses the traversable cells within a local neighborhood determined by the sensing radius r_s. The sensing region of UAV i, the cumulative covered region, and the coverage ratio are

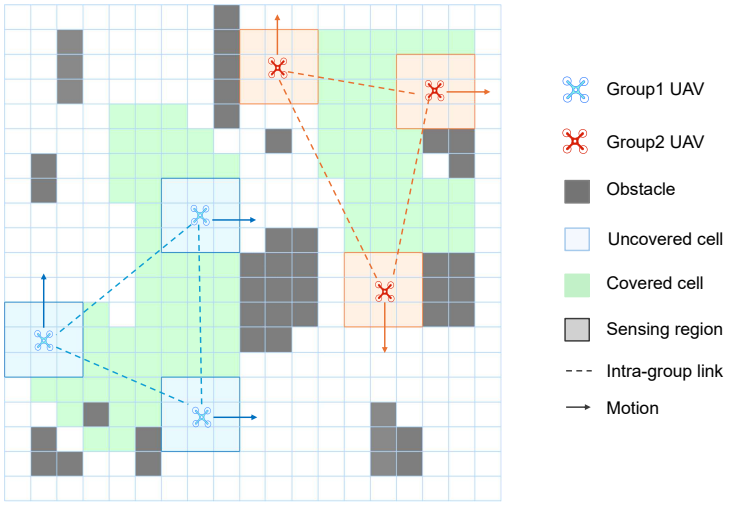


Fig. 1 Illustration of the cooperative multi-UAV coverage scenario. UAVs are partitioned into groups and move synchronously to cover traversable cells while avoiding obstacles and maintaining intra-group communication.

defined as

$$\begin{aligned} \mathcal{V}_i^t &= \left\{ \mathbf{q} \in \mathcal{F} \mid \left\| \mathbf{q} - \mathbf{p}_i^t \right\|_\infty \le r_\mathrm{s} \right\}, \\ \mathcal{C}_t &= \bigcup_{\tau=0}^{t} \bigcup_{i \in \mathcal{U}} \mathcal{V}_i^\tau, \qquad \rho_t = \frac{|\mathcal{C}_t|}{|\mathcal{F}|}. \end{aligned} \tag{2}$$

Since $\mathcal{C}_t$ is defined by set union, repeatedly observed cells are counted only once.

Based on its local observation, UAV i selects a discrete motion action $\mathbf{a}_i^t \in \mathcal{A}$. All UAVs execute their selected actions synchronously. A motion is feasible only if the resulting position remains within $\mathcal{F}$, and a collision occurs when two or more UAVs occupy the same grid cell after action execution.

The UAVs must also maintain intra-group communication. For each group $\mathcal{G}_k$, we define a communication graph $\mathcal{H}_k^t = (\mathcal{G}_k, \mathcal{E}_k^t)$, where two UAVs are connected if their Euclidean distance does not exceed the communication radius r_c. The communication requirement is expressed as

$$\begin{aligned} \mathcal{E}_k^t &= \left\{ (i,j) \mid i,j \in \mathcal{G}_k,\ i \ne j,\ \left\| \mathbf{p}_i^t - \mathbf{p}_j^t \right\|_2 \le r_\mathrm{c} \right\}, \\ \kappa\left(\mathcal{H}_k^t\right) &= 1, \qquad \forall k \in \{1, \dots, K\}, \end{aligned} \tag{3}$$

where $\kappa(\mathcal{H}_k^t)$ denotes the number of connected components in the communication graph. This constraint requires all UAVs within each group to remain connected through direct or multi-hop communication links. The overall cooperative coverage scenario is illustrated in Fig. 1.

3.2 Cooperative Coverage Objective

The UAVs coordinate their motion decisions to maximize the coverage of the traversable region. Each UAV executes an individual policy π_i, while all individual policies share the same parameters θ. Given the available information $\mathbf{o}_i^t$, the action of UAV i is selected according to

$$\mathbf{a}_i^t \sim \pi_i\left(\cdot \mid \mathbf{o}_i^t; \theta\right), \qquad \forall i \in \mathcal{U}. \tag{4}$$

Here, o_i^t denotes the local observation available to UAV i at time step t, and $\pi_i(\cdot \mid o_i^t; \theta)$ denotes the action probability distribution determined by the shared policy parameters θ. The action a_i^t is sampled from this distribution.

Let T denote the terminal time step of the coverage task. The shared policy parameters are optimized to maximize the expected coverage ratio at T:

$$\begin{aligned} \theta^* = \arg\max_{\theta} \quad & \mathbb{E}_\theta\left[\rho_T\right] \\ \text{subject to} \quad & \mathbf{p}_i^t \in \mathcal{F}, && \forall i \in \mathcal{U},\ t = 0, \ldots, T, \\ & \mathbf{p}_i^t \neq \mathbf{p}_j^t, && \forall i, j \in \mathcal{U},\ i \neq j,\ t = 0, \ldots, T, \\ & \kappa\left(\mathcal{H}_k^t\right) = 1, && \forall k \in \{1, \ldots, K\},\ t = 0, \ldots, T. \end{aligned} \tag{5}$$

The first constraint requires every UAV to remain within the traversable region $\mathcal{F}$ throughout the episode. If a candidate action would move a UAV into an obstacle or outside $\mathcal{F}$, the action is rejected, the UAV remains at its current position, and an obstacle penalty is imposed without terminating the episode. The second constraint prevents two UAVs from occupying the same cell simultaneously; conflicting actions are similarly rejected and penalized. The third constraint requires the communication graph of each UAV group to remain connected, such that any two UAVs within the same group can communicate through either a direct or multi-hop communication path. Together, the objective and constraints encourage the UAVs to expand the covered region while avoiding collisions and maintaining intra-group communication.

4 Proposed Method

4.1 Method Overview

Multi-group UAV coverage requires both long-term exploration planning and real-time motion control. A flat policy must simultaneously determine promising unexplored regions and generate feasible actions under obstacles, inter-UAV interactions, and communication constraints, which substantially increases the learning difficulty (Kong et al 2024; Ni et al 2025; Choi et al 2025). To separate these two decision responsibilities, we propose a hierarchical framework consisting of a frontier-guided planning layer and a topology-aware execution policy, as illustrated in Fig. 2.

The planning layer extracts candidate targets from the evolving coverage map and assigns exploration directions to the UAVs, thereby providing explicit global guidance. Given these targets, the execution policy generates local motion actions based

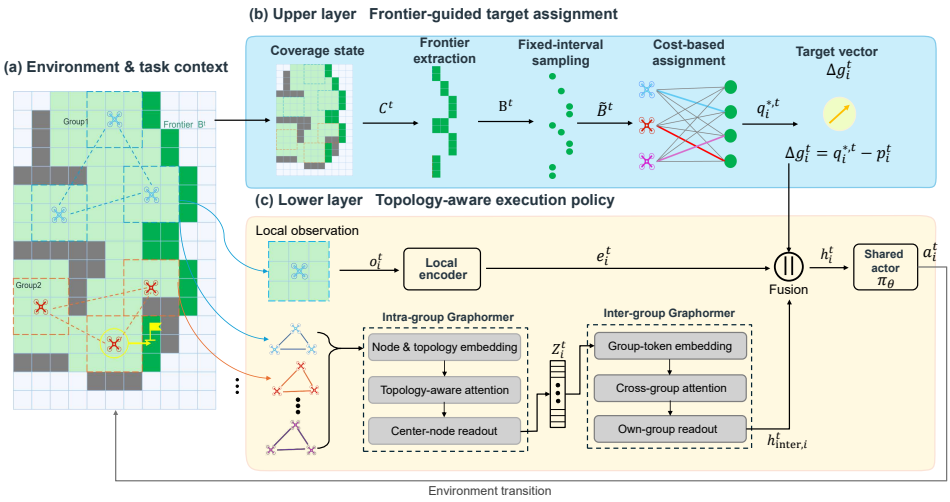


Fig. 2 Overall architecture of the proposed topology-aware hierarchical coverage framework. The planning layer assigns exploration targets to individual UAVs, while the execution layer employs hierarchical topology modeling to generate movement actions.

on environmental observations and multi-UAV interactions. Within the execution policy, intra-group and inter-group relations are modeled sequentially. The intra-group encoder captures communication and cooperation among UAVs within the same group, while the inter-group encoder coordinates exploration across different groups to reduce redundant coverage.

4.2 Frontier-Guided Planning Layer

Efficient coverage requires the UAVs to continuously identify promising unexplored regions according to the evolving global coverage state. Without such global guidance, a UAV may repeatedly select locally feasible actions without making meaningful progress toward the remaining uncovered regions. The planning layer addresses this limitation by extracting exploration targets from the global coverage state and supplying each UAV with a task-level exploration direction (Arias-Perez et al 2024; Goarin and Loianno 2024).

A frontier is defined as an uncovered traversable cell adjacent to the currently covered region. Accordingly, the frontier set at time step t is expressed as

$$\mathcal{B}^t = \{\mathbf{q} \in \mathcal{F} \setminus \mathcal{C}_t \mid \mathcal{N}(\mathbf{q}) \cap \mathcal{C}_t \neq \varnothing\}, \tag{6}$$

where $\mathcal{N}(\mathbf{q})$ denotes the neighborhood of cell $\mathbf{q}$. These frontier cells characterize the interface through which the currently covered region can expand toward unexplored space.

Because adjacent frontier cells frequently belong to the same local boundary, directly treating every frontier cell as an independent target would introduce substantial redundancy. The raw frontier set is therefore consolidated into a candidate target set:

$$\widetilde{\mathcal{B}}^t = \mathcal{Q}\left(\mathcal{B}^t\right), \tag{7}$$

where $\mathcal{Q}(\cdot)$ denotes a fixed-interval downsampling operation that retains one candidate every s frontier cells, thereby reducing redundant neighboring candidates.

Based on the downsampled candidate set, the planning layer assigns a target to each UAV by jointly considering the travel distance and exploration value (Goarin and Loianno 2024; Kong et al 2024). Specifically, for UAV i and candidate target $\mathbf{q}$, the assignment cost is defined as

$$C_{i,\mathbf{q}}^t = \lambda_d \widehat{D}\left(\mathbf{p}_i^t, \mathbf{q}\right) + \lambda_e \left(1 - \widehat{V}(\mathbf{q})\right), \tag{8}$$

where $\widehat{D}(\mathbf{p}_i^t, \mathbf{q})$ is the normalized distance cost and $\widehat{V}(\mathbf{q})$ is the normalized exploration value, measured by the unexplored area around the candidate. The assignment mechanism selects targets with lower costs, thereby favoring candidates that are both closer to the UAV and more valuable for further exploration.

4.3 Topology-Aware Execution Policy

After the planning layer assigns exploration targets, the execution policy determines how each UAV moves toward its assigned target while avoiding obstacles and inter-UAV collisions. During this process, UAVs within the same group need to maintain communication and coordinate their movements, whereas different groups need to coordinate their exploration to reduce redundant coverage. These distinct coordination requirements cannot be adequately captured by treating all UAV interactions in the same manner (Goeckner et al 2024; Zhao et al 2024a; Feng et al 2024).

Motivated by group-aware coordination modeling (Duan et al 2024), we design a hierarchical topology encoder comprising two sequential Graphormer modules (Ying et al 2021). The intra-group module first aggregates information along valid communication links within each group, producing topology-aware UAV representations. These representations are subsequently processed by the inter-group module to capture coordination across different groups. The resulting topology representation therefore integrates intra-group communication dependencies and inter-group coordination information for subsequent action decision-making.

Let x_i^t denote the initial node feature of UAV i at time step t. For each UAV group $\mathcal{G}_k$, an intra-group graph $\mathbf{G}_{\text{intra},k}^t = (\mathcal{G}_k, \mathcal{E}_{\text{intra},k}^t)$ is constructed (Cattai et al 2025). Its edge set is defined as

$$\mathcal{E}_{\text{intra},k}^t = \left\{ (j,l) \;\middle|\; j,l \in \mathcal{G}_k,\ j \neq l,\ \left\|\mathbf{p}_j^t - \mathbf{p}_l^t\right\|_2 \leq r_c \right\}, \qquad k = 1, \ldots, K. \tag{9}$$

For a given UAV i, the hierarchical topology encoder processes all UAV groups, where $c(i)$ denotes the index of the group containing UAV i. For $\mathcal{G}_{c(i)}$, UAV i is

selected as the center node so that the resulting representation describes the intra-group topology and member distribution from the perspective of UAV i. For each $\mathcal{G}_k$ with $k \neq c(i)$, no individual UAV is selected as the reference node. Instead, a virtual center node v_k is introduced and initialized by averaging the features of all UAVs in $\mathcal{G}_k$. This virtual node provides an overall representation of the external group without relying on the perspective of any particular UAV. This asymmetric design reflects the distinct information requirements of intra-group and inter-group coordination: the current group is encoded from the perspective of UAV i to preserve its relevant communication context, whereas each external group is summarized through a virtual center node to provide an objective group-level view.

The intra-group Graphormer processes the node features and communication links within each UAV group. Its output at the selected center node is retained as the representation of $\mathcal{G}_k$ from the perspective of UAV i:

$$z_{i,k}^t = \begin{cases} \text{Graphormer}_{\text{intra}} \left(\left\{ x_j^t \right\}_{j \in \mathcal{G}_k}, \mathbf{G}_{\text{intra},k}^t \right) [i], & k = c(i), \\ \text{Graphormer}_{\text{intra}} \left(\left\{ x_j^t \right\}_{j \in \mathcal{G}_k} \cup \left\{ x_{v_k}^t \right\}, \mathbf{G}_{\text{intra},k}^t \cup \{v_k\} \right) [v_k], & k \neq c(i). \end{cases} \tag{10}$$

Specifically, the output of UAV i is retained for $\mathcal{G}_{c(i)}$, whereas the output of the virtual center node v_k is retained for every other UAV group.

After processing all UAV groups, their representations are stacked as

$$Z_i^t = \left[z_{i,1}^t, \ldots, z_{i,K}^t \right]^\top. \tag{11}$$

The inter-group Graphormer then models the coordination relationships among the K UAV-group representations. Its output corresponding to $\mathcal{G}_{c(i)}$ is retained as the final topology representation of UAV i:

$$h_{\text{inter},i}^t = \text{Graphormer}_{\text{inter}} \left(Z_i^t, \mathbf{G}_{\text{inter}}^t \right) [c(i)]. \tag{12}$$

Thus, $h_{\text{inter},i}^t$ integrates the intra-group communication information contained in the K UAV-group representations and the coordination relationships among different UAV groups, while maintaining the decision-making perspective of UAV i.

In parallel with the topology encoder, the execution policy processes the local neighborhood of each UAV. The local observation o_i^t contains the information associated with the eight neighboring grid cells around UAV i. This information describes the immediate movement conditions, including nearby obstacles and the feasibility of candidate movements, and is encoded as

$$e_i^t = \phi_o \left(o_i^t \right), \tag{13}$$

where $\phi_o(\cdot)$ denotes the local neighborhood encoder.

The local neighborhood feature and the topology representation are concatenated to form the input of the actor network:

$$h_i^t = e_i^t \parallel h_{\text{inter},i}^t, \tag{14}$$

where $\parallel$ denotes feature concatenation. Based on the fused representation, UAV i selects its action according to the shared policy:

$$a_i^t \sim \pi_i\left(\cdot \mid h_i^t; \theta\right). \tag{15}$$

All UAVs share the policy parameters θ, while each UAV constructs its topology representation and action distribution from its own decision-making perspective.

4.4 Policy Optimization

All UAVs share the actor parameters and independently select actions based on their individual representations, while a centralized critic utilizes global information during training (Ekechi et al 2025; Yu et al 2022). Parameter sharing reduces the number of trainable parameters and promotes consistent coordination behavior across UAVs.

4.4.1 Reward Design

The step-wise reward jointly considers coverage progress, target guidance, obstacle avoidance, collision avoidance, and communication maintenance (Ni et al 2025; Choi et al 2025):

$$r_i^t = \lambda_{\text{cov}} r_{\text{cov}}^t + \lambda_{\text{goal}} r_{\text{goal},i}^t - \lambda_{\text{obs}} r_{\text{obs},i}^t - \lambda_{\text{uav}} r_{\text{uav},i}^t - \lambda_{\text{com}} r_{\text{com},i}^t, \tag{16}$$

where the coefficients λ_{cov}, λ_{goal}, λ_{obs}, λ_{uav}, and λ_{com} control the relative contributions of the corresponding terms.

Coverage reward. The coverage reward is determined by the proportion of newly covered cells relative to all traversable cells, encouraging the UAV team to explore previously uncovered regions.

Target-guidance reward. The target-guidance reward is calculated from the reduction in the normalized Euclidean distance between UAV i and its assigned target. Moving closer to the target produces a positive reward, whereas moving away produces a negative reward.

Obstacle-collision penalty. A binary penalty is imposed when the candidate position of UAV i lies in an obstacle or outside the traversable region (Liu et al 2025).

collision penalty. A binary penalty is imposed when the candidate position of UAV i conflicts with that of another UAV.

Communication penalty. Let $d_{\text{com},i}^{t+1} = \min_{j \neq i,\, c(j)=c(i)} \|\mathbf{p}_i^{t+1} - \mathbf{p}_j^{t+1}\|_2$ denote the distance from UAV i to its nearest same-group UAV. The soft communication penalty

is defined as

$$r_{\mathrm{com},i}^{t}=\begin{cases}0, & d_{\mathrm{com},i}^{t+1}\le r_w,\\ \dfrac{d_{\mathrm{com},i}^{t+1}-r_w}{r_c-r_w}, & r_w<d_{\mathrm{com},i}^{t+1}\le r_c,\\ 1, & d_{\mathrm{com},i}^{t+1}>r_c,\end{cases} \tag{17}$$

where r_w and r_c denote the warning threshold and communication limit, respectively, with $r_w < r_c$. The penalty is zero within r_w, increases linearly as the distance approaches r_c, and remains at its maximum beyond r_c. This anticipatory penalty encourages same-group UAVs to maintain a safe communication margin before disconnection occurs (Cattai et al 2025).

4.4.2 Optimization Objective

Let θ_{old} denote the actor parameters used to collect the current batch of trajectories. The probability ratio between the current policy and the old policy for UAV i at time step t is defined as (Schulman et al 2017)

$$r_i^t(\theta)=\frac{\pi_i\left(a_i^t \mid h_i^t;\theta\right)}{\pi_i\left(a_i^t \mid h_i^t;\theta_{\mathrm{old}}\right)}, \tag{18}$$

The clipped policy loss is defined as

$$\mathcal{L}_{\mathrm{policy}}(\theta)=-\mathbb{E}_{t,i}\left[\min\left(r_i^t(\theta)\widehat{A}_i^t,\mathrm{clip}\left(r_i^t(\theta),1-\epsilon,1+\epsilon\right)\widehat{A}_i^t\right)\right], \tag{19}$$

where $\widehat{A}_i^t$ denotes the estimated advantage of UAV i, and ϵ is the clipping coefficient.

The centralized critic is optimized using the value loss

$$\mathcal{L}_{\mathrm{value}}(\phi)=\mathbb{E}_t\left[\left(V\left(s^t;\phi\right)-\widehat{R}^t\right)^2\right], \tag{20}$$

where s^t denotes the global state, ϕ denotes the critic parameters, and $\widehat{R}^t$ is the estimated return.

The overall optimization loss is defined as

$$\mathcal{L}(\theta,\phi)=\mathcal{L}_{\mathrm{policy}}(\theta)+c_v\mathcal{L}_{\mathrm{value}}(\phi)-c_e\mathbb{E}_{t,i}\left[\mathcal{H}\left(\pi_i\left(\cdot \mid h_i^t;\theta\right)\right)\right], \tag{21}$$

where c_v and c_e are the coefficients of the value loss and entropy regularization, respectively.

After training, the planning layer continuously updates the exploration targets according to the global coverage state, while the learned execution policy generates motion actions based on the assigned targets, local observations, and topology-aware representations. The two layers jointly enable efficient coverage planning and responsive multi-UAV coordination in obstacle-rich environments.

Table 1 Configuration of the evaluation dataset.

Difficulty	Number	Obstacle ratio	Main characteristics
Easy	50	0%–5%	Open regions with sparse and isolated obstacles.
Medium	100	5%–15%	Clustered obstacles, partial barriers, and separated local regions.
Hard	50	15%–30%	Dense obstacles, narrow passages, corridors, and room-like regions.

5 Experimental Setup

5.1 Evaluation Dataset

To evaluate the methods under different environmental complexities, we construct an evaluation dataset containing 200 independently generated 50×50 grid maps. Each map consists of traversable and obstacle cells and is categorized as Easy, Medium, or Hard according to its obstacle ratio and structural complexity. Let $\mathcal{O}$ denote the set of obstacle cells. The obstacle ratio is defined as

$$\eta_{\text{obs}} = \frac{|\mathcal{O}|}{HW}. \tag{22}$$

Easy maps contain open regions with sparse or isolated obstacles. Medium maps introduce clustered obstacles and partial barriers that divide the traversable area into several local regions. Hard maps contain denser obstacles and more complex structures, such as narrow passages, corridors, and room-like regions, requiring more frequent motion adjustment and stronger coordination among UAVs. The composition of the evaluation dataset is summarized in Table 1.

5.2 Baselines

We compare our method with six baselines: three rule-based methods, namely Boustrophedon-CPP, Frontier-Swarm, and APF-CPP, and three reinforcement-learning-based methods, namely IPPO, Rainbow-DQN-SF, and SDA-MATD3.

Boustrophedon-CPP. This method decomposes the environment into traversable subregions and generates regular back-and-forth paths within each subregion (Gong et al 2024). These paths are then assigned to different UAVs for cooperative coverage.

Frontier-Swarm. This method detects frontiers between covered and uncovered regions and guides UAVs toward nearby frontiers (Tran et al 2022). Local interaction and priority rules are used to maintain formation and resolve motion conflicts.

APF-CPP. This method combines frontier exploration with artificial potential fields (Wang et al 2024). Uncovered regions attract UAVs, while obstacles and nearby UAVs generate repulsive forces. Frontier-based backtracking is used to escape locally covered regions.

IPPO. Following the IPPO-based UAV-cluster coverage path planning method of Wang et al (2025), we adapt IPPO to our grid-based coverage environment as a baseline. Each UAV independently selects movement actions from its local observation using a shared actor–critic network. The baseline does not employ frontier-guided target planning or explicit topology modeling.
Rainbow-DQN-SF. This baseline employs parameter-shared Rainbow DQN with the robustness and safety filters of the original method (Carvalho and Aguiar 2026). These filters redirect locally trapped UAVs and eliminate actions that may enter obstacles, cross map boundaries, or cause conflicts.
SDA-MATD3. This method uses sparse dual-attention modules to select relevant targets, UAVs, and obstacles (Chen et al 2026). We adapt it to the grid environment while retaining entity attention, dynamic sparse selection, and centralized training.

5.3 Evaluation Metrics

We evaluate all methods using four metrics: Final Coverage Rate (FCR), Steps to 80% Coverage (S80), Collision Rate (CR), and Communication Violation Rate (CVR). FCR, CR, and CVR are reported as percentages, whereas S80 is measured in time steps.
Final Coverage Rate (FCR). FCR measures the percentage of traversable cells visited by UAVs during an episode:

$$\mathrm{FCR} = 100 \times \rho^T = 100 \times \frac{|C^T|}{|\mathcal{F}|}. \tag{23}$$

Here, C^T denotes the set of cells cumulatively covered by all UAVs up to the terminal time step T, and $\mathcal{F}$ denotes the set of all traversable cells in the environment. Therefore, a higher FCR indicates that the UAV team covers a larger portion of the environment within the given execution budget.
Steps to 80% Coverage (S80). S80 records the first time step at which the coverage ratio reaches 80%:

$$\mathrm{S80} = \min\left\{t \in \{0, \ldots, T\} \mid \rho^t \geq 0.8\right\}. \tag{24}$$

A lower S80 indicates faster coverage.
Collision Rate (CR). CR measures the percentage of time steps in which at least one collision occurs. Let c^t denote the collision indicator at time step t:

$$c^t = \mathbb{I}\left[\exists\, i, j \in \{1, \ldots, N\},\ i < j : \mathbf{p}_i^t = \mathbf{p}_j^t\right]. \tag{25}$$

Here, $\mathbb{I}[\cdot]$ equals 1 if the enclosed condition is satisfied and 0 otherwise. CR is calculated as

$$\mathrm{CR} = \frac{100}{T} \sum_{t=1}^{T} c^t. \tag{26}$$

A lower CR indicates safer motion coordination among the UAVs.
Communication Violation Rate (CVR). CVR measures the percentage of time steps in which at least one UAV group loses communication connectivity. The

communication violation indicator at time step t is defined as

$$v^t = \mathbb{I}\left[\exists\, k \in \{1, \ldots, K\} : \kappa\left(\mathcal{H}_k^t\right) > 1\right]. \tag{27}$$

CVR is calculated as

$$\mathrm{CVR} = \frac{100}{T} \sum_{t=1}^{T} v^t. \tag{28}$$

A lower CVR indicates fewer time steps with disconnected intra-group communication graphs.

5.4 Implementation Details

The experiments involve (N=9) homogeneous UAVs partitioned into (K=3) fixed groups, with each group containing three UAVs. The sensing radius and communication radius are set to $r_s = 1$ and $r_c = 8$ grid cells, respectively, and the maximum episode length is (T=150) steps. To ensure a fair comparison, all learning-based methods are trained on the same set of training maps under an identical budget of 90,000 environment-interaction steps. During evaluation, all methods are tested on the same 200 unseen maps. For each map, the obstacle layout, UAV initial positions, and all environment settings are kept identical across methods. Detailed training hyperparameters and network configurations are provided in Appendix A.

6 Results and Discussion

6.1 Overall Performance

To comprehensively evaluate the proposed method, we compare it with three rule-based methods and three reinforcement-learning-based methods on Easy, Medium, and Hard maps. Table 2 presents the Final Coverage Ratio (FCR), the number of steps required to reach 80% coverage (S_{80}), the collision rate (CR), and the communication violation rate (CVR). A higher FCR indicates better coverage performance, while lower S_{80}, CR, and CVR indicate better exploration efficiency, collision avoidance, and communication maintenance, respectively. Based on the experimental results, the following conclusions can be drawn.

The proposed method demonstrates consistently strong coverage performance across all map difficulty levels. It achieves FCR values of 96.61%, 96.18%, and 91.33% on the Easy, Medium, and Hard maps, respectively. These results outperform the strongest RL-based baseline by 7.45–15.88 percentage points and also remain substantially higher than Frontier-Swarm and Boustrophedon-CPP. The frontier-based planning mechanism provides goal-directed exploration targets, while inter-group topology modeling supports coordination across groups. Together, these components help sustain high coverage performance as map complexity increases.

The proposed method achieves the strongest communication maintenance while keeping collision risk consistently low. It obtains the lowest CVR across all difficulty levels, with values of 6.59%, 7.75%, and 9.32% on the Easy, Medium, and Hard maps, respectively. Compared with Frontier-Swarm, the baseline with the lowest

Table 2 Evaluation results under different map difficulty levels. FCR, CR, and CVR are reported in percentages, while S_{80} is measured in steps. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Easy				Medium				Hard			
	FCR	S_{80}	CR	CVR	FCR	S_{80}	CR	CVR	FCR	S_{80}	CR	CVR
Rule-based												
APF-CPP	**97.70**	**88.70**	<u>0.10</u>	97.80	**98.30**	**92.90**	<u>0.70</u>	97.90	**93.10**	<u>109.00</u>	1.80	98.00
Frontier-Swarm	70.50	–	**0.00**	<u>12.60</u>	66.70	–	**0.00**	<u>13.90</u>	52.30	–	**0.00**	<u>23.40</u>
Boustrophedon-CPP	75.40	–	**0.00**	96.00	75.80	–	**0.00**	96.20	75.90	–	**0.00**	96.80
RL-based												
IPPO	61.97	–	24.64	33.36	58.10	–	29.38	34.78	54.18	–	33.34	33.79
SDA-MATD3	61.10	–	8.79	60.04	61.30	–	20.52	64.29	58.93	–	30.25	64.37
Rainbow-DQN-SF	89.16	108.72	**0.00**	68.80	80.30	110.97	**0.00**	69.97	76.35	–	<u>0.01</u>	64.94
Ours	<u>96.61</u>	<u>95.34</u>	0.47	**6.59**	<u>96.18</u>	<u>99.80</u>	1.17	**7.75**	<u>91.33</u>	**108.64**	1.81	**9.32**

CVR, these results represent reductions of 6.01–14.08 percentage points. Meanwhile, its CR remains below 2% across all maps, demonstrating reliable collision avoidance even as environmental complexity increases. By explicitly modeling intra-group topology, the proposed encoder enables each UAV to consider the relative states of its group members, thereby improving compliance with the intra-group communication constraint while maintaining a low collision risk.

The proposed method achieves a more favorable trade-off between exploration performance and safe coordination. On the Hard maps, it reaches 80% coverage in 108.64 steps and achieves a final FCR of 91.33%, while limiting CR and CVR to 1.81% and 9.32%, respectively. In comparison, APF-CPP reaches 80% coverage in nearly the same number of steps (109.00 versus 108.64) and improves the final FCR by only 1.77 percentage points, while its CVR reaches 98.00%, indicating that its slightly higher coverage is achieved at the expense of intra-group communication maintenance. Rainbow-DQN-SF reduces CR to 0.01%, but fails to reach 80% coverage, with a final FCR of only 76.35% and a CVR of 64.94%. The proposed topology-aware hierarchical encoder separately models inter-group exploration coordination and intra-group communication dependencies. By incorporating both relationships into UAV-centric action representations, it balances exploration performance and safe coordination in complex environments.

6.2 Ablation Study

To evaluate the contributions of the two topology modeling components, we construct three variants by removing the inter-group topology module, the intra-group topology module, or both modules from the full model. All other network components and training settings remain unchanged to ensure a fair comparison. Based on the ablation results shown in Table 3, the following conclusions regarding the inter-group and intra-group topology modeling components can be drawn.

Intra-group topology modeling is essential for maintaining communication within each group. For example, on the Medium maps, removing the intra-group

Table 3 Ablation results under different map difficulty levels. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Easy				Medium				Hard			
	FCR	S_{80}	CR	CVR	FCR	S_{80}	CR	CVR	FCR	S_{80}	CR	CVR
w/o Intra-group	<u>77.46</u>	–	26.74	19.87	<u>78.44</u>	–	17.64	24.22	<u>79.10</u>	–	11.64	17.94
w/o Inter-group	68.19	–	<u>13.66</u>	<u>8.69</u>	67.57	–	<u>14.42</u>	<u>9.06</u>	68.05	–	<u>7.54</u>	**8.65**
w/o Topology	65.08	–	18.13	17.61	66.54	–	17.98	26.74	65.52	–	12.62	18.65
Full	**96.61**	**95.34**	**0.47**	**6.59**	**96.18**	**99.80**	**1.17**	**7.75**	**91.33**	**108.64**	**1.81**	<u>9.32</u>

topology increases CVR from 7.75 to 24.22, whereas removing the inter-group topology only increases it to 9.06. A similar degradation is observed on the Easy maps when the entire topology encoder is removed, with CVR increasing from 6.59 to 17.61. By capturing the relative states and communication dependencies of group members, the intra-group topology module provides the information required to maintain reliable intra-group connections.

Inter-group topology modeling substantially improves coordinated exploration across groups. On the Medium maps, for instance, removing the inter-group topology decreases FCR from 96.18 to 67.57, a larger reduction than the decrease to 78.44 caused by removing the intra-group topology. When the entire topology encoder is removed, FCR further declines to 66.54. These comparisons show that the inter-group component accounts for a substantial portion of the topology encoder's contribution to coverage performance. By modeling exploration relationships across groups, it supports more effective coordination of their exploration decisions and thereby improves overall exploration efficiency.

6.3 Training Convergence Analysis

To evaluate the learning efficiency and training stability of the proposed method, we compare its training reward curve with those of IPPO, Rainbow-DQN-SF, and SDA-MATD3. Figure 3 shows the evolution of the mean per-UAV reward during training.

The proposed method exhibits faster reward improvement and converges to a substantially higher reward than the baseline methods. All methods begin with mean rewards of approximately (-5), indicating similarly poor initial policies. The reward of the proposed method becomes positive after approximately 17 episodes, exceeds 4 after approximately 30 episodes, and subsequently stabilizes near 7. In comparison, Rainbow-DQN-SF improves relatively quickly during the early training stage but converges to approximately 4.8. SDA-MATD3 stabilizes at approximately 3.0 after 50 episodes, whereas IPPO improves more slowly and reaches approximately 4.5 by the end of training. Although the raw episode rewards fluctuate because of changes in obstacle layouts and initial UAV positions, the smoothed curve of the proposed method remains consistently above those of the baselines throughout most of the training process. The frontier-guided planning layer reduces the burden of simultaneously learning long-term exploration and local motion control, while the sequential intra-group and inter-group Graphormer modules provide structured representations

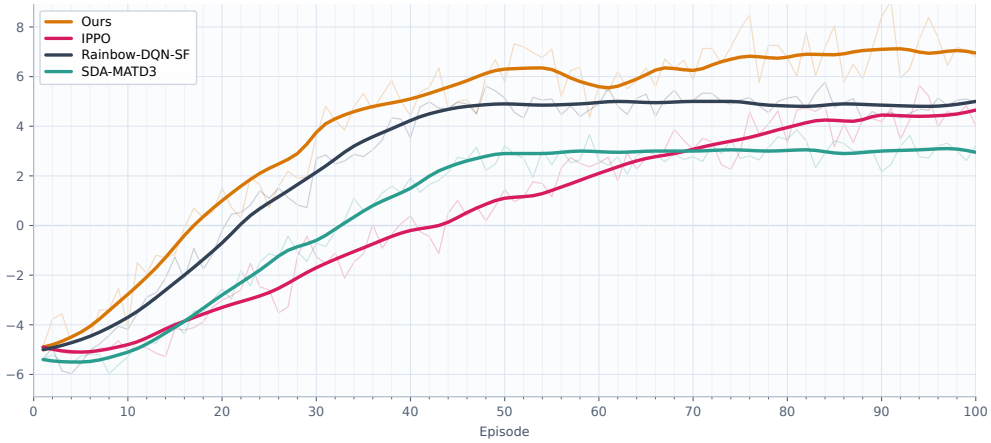


Fig. 3 Training reward curves of the proposed method and the reinforcement-learning-based baselines.

of communication and coordination relationships. These components improve the effectiveness and stability of policy learning, enabling the proposed method to achieve both faster reward improvement and a higher converged performance.

6.4 Generalization Evaluation

6.4.1 Generalization across Communication Ranges

To evaluate whether the learned policies remain effective under communication conditions different from those encountered during training, we vary the communication range as $r_c \in 4, 6, 8, 10, 12$, while fixing the number of UAVs to (N=9) and the group configuration to (K=3) with (3+3+3) UAVs. All learning-based methods are trained only under the default setting $r_c = 8$ and are then directly evaluated at each communication range without retraining or fine-tuning. The same 200 unseen maps and initial UAV positions are used for all methods. Figures 4(a) and 4(b) report the resulting FCR and CVR, respectively.

The proposed method maintains the best coverage and communication performance across all communication ranges. When the communication range decreases from the training setting $r_c = 8$ to $r_c = 4$, the proposed method still achieves an FCR of 85.72% and a CVR of 14.23%. Its FCR is 18.55 percentage points higher than that of Rainbow-DQN-SF, the strongest coverage baseline at this range, while its CVR is 25.98 percentage points lower than that of IPPO, the baseline with the lowest CVR. As r_c increases from 4 to 12, its FCR increases from 85.72% to 95.17%, while its CVR decreases from 14.23% to 3.45%. Moreover, when $r_c \geq 8$, its FCR remains between 94.21% and 95.17%, indicating stable coverage under more permissive communication conditions. The results indicate that the learned policy is not restricted to the communication range used during training and generalizes effectively to both more restrictive and more permissive communication conditions.

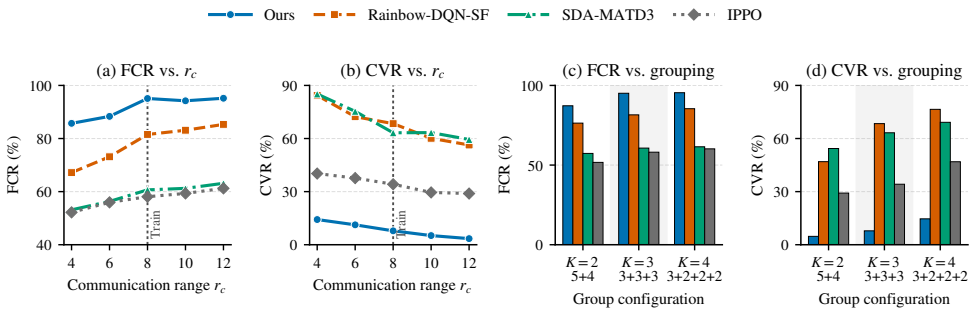


Fig. 4 Generalization performance across communication ranges and group configurations. (a) FCR and (b) CVR under different communication ranges. (c) FCR and (d) CVR under different group configurations. The vertical dashed lines and shaded regions indicate the corresponding training settings.

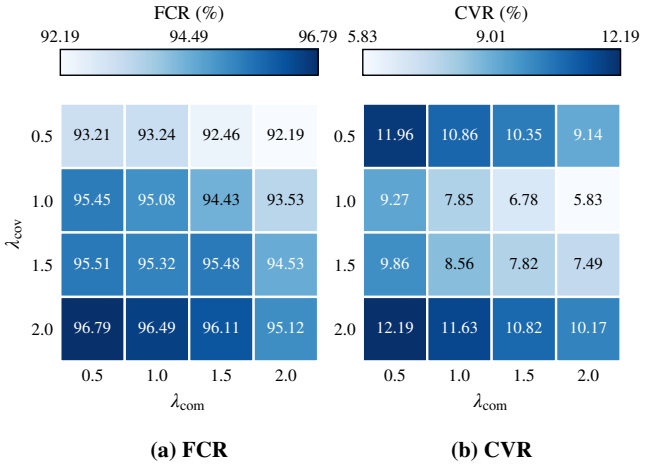
6.4.2 Generalization across Group Configurations

To examine whether the learned policies can adapt to group structures unseen during training, we fix the total number of UAVs to (N=9) and the communication range to $r_c = 8$, and evaluate three group configurations: (K=2) with (5+4) UAVs, (K=3) with (3+3+3) UAVs, and (K=4) with (3+2+2+2) UAVs. All learning-based methods are trained only under the default (K=3) configuration and are directly applied to the other configurations without retraining or fine-tuning. Figures 4(c) and 4(d) present the FCR and CVR results on the same 200 unseen maps.

The proposed method remains robust when both the number and sizes of UAV groups change. It achieves FCR values of 87.23%, 95.08%, and 95.43% under the (K=2), (K=3), and (K=4) configurations, respectively, while maintaining the lowest CVR values of 4.73%, 7.85%, and 14.63%. Across the three configurations, its FCR exceeds that of the strongest baseline by between 10.04 and 13.55 percentage points, and its CVR is lower than that of the lowest-CVR baseline by between 24.46 and 32.26 percentage points. With (K=2), the larger groups impose stronger spatial coupling among their members, limiting exploration dispersion but providing more same-group neighbors for communication maintenance. In contrast, the (K=4) configuration allows more groups to explore different regions, improving coverage, but its smaller groups provide fewer alternative communication neighbors and therefore increase CVR. By recomputing intra-group and inter-group topology relationships for each group partition, the sequential Graphormer modules adapt the UAV representations to the changed coordination structure. These results demonstrate that the proposed method is not tied to the group configuration used during training and retains a favorable coverage–communication trade-off across different group partitions.

6.5 Hyperparameter Analysis

To investigate how the reward design affects the trade-off between exploration performance and communication compliance, we vary the coverage-reward weight λ_{cov} and



(a) FCR **(b) CVR**

Fig. 5 Effects of the coverage-reward weight λ_{cov} and communication-penalty weight λ_{com} on FCR and CVR.

communication-penalty weight λ_{com} over $\{0.5, 1.0, 1.5, 2.0\}$. For each of the resulting 16 combinations, the policy is trained under the same settings and evaluated on the same 200 maps, while all other hyperparameters remain fixed. Figure 5 reports the resulting FCR and CVR.

The two reward components exert competing influences on UAV spatial behavior, resulting in a trade-off between exploration performance and communication compliance. A larger coverage-reward weight encourages UAVs to move toward distinct unexplored regions, increasing their spatial dispersion and improving coverage. However, excessive dispersion also increases the risk that same-group UAVs exceed the communication range. Accordingly, when $\lambda_{\text{com}} = 1.0$, increasing λ_{cov} from 1.0 to 2.0 raises FCR from 95.08% to 96.49%, but also increases CVR from 7.85% to 11.63%. In contrast, a larger communication-penalty weight encourages same-group UAVs to remain closer, improving communication compliance at the cost of restricting their exploration range. When $\lambda_{\text{cov}} = 1.0$, increasing λ_{com} from 0.5 to 2.0 reduces CVR from 9.27% to 5.83%, while FCR decreases from 95.45% to 93.53%. The default setting $(\lambda_{\text{cov}}, \lambda_{\text{com}}) = (1.0, 1.0)$ maintains a high FCR of 95.08% while limiting CVR to 7.85%, providing a balanced compromise between spatial dispersion for exploration and intra-group connectivity.

7 Conclusions

This paper presented a topology-aware hierarchical multi-agent reinforcement learning framework for cooperative multi-UAV coverage under collision and intra-group communication constraints. The framework separates long-term exploration planning from local action execution. The frontier-guided planning layer extracts candidate targets from the evolving coverage boundary and assigns them according to travel distance and exploration value. Given these targets, the execution policy employs sequential

intra-group and inter-group Graphormer modules to capture communication dependencies within groups and exploration relationships across groups. This design enables the UAVs to expand the covered region while maintaining intra-group connectivity and avoiding unsafe movements.

Experiments on 200 unseen maps demonstrated the effectiveness of the proposed framework. On the Hard maps, it achieved an FCR of 91.33%, outperforming the strongest reinforcement-learning-based baseline by 14.98 percentage points, while limiting CR and CVR to 1.81% and 9.32%, respectively. The ablation results confirmed that intra-group topology modeling primarily supports communication maintenance, whereas inter-group topology modeling substantially improves coordinated exploration. The proposed method also exhibited faster training convergence and generalized effectively to unseen communication ranges and group configurations. Future work will extend the framework to environments with dynamic obstacles, time-varying communication links, and adaptive group membership, and will further evaluate its performance on real multi-UAV platforms.

Supplementary information. No supplementary material is associated with this article.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62676094, in part by the National Key Research and Development Program of China under Grant 2024YFB4303805 and in part by the research start-up fund of USTC.

Statements and Declarations

Funding. This work was supported in part by the National Natural Science Foundation of China under Grant 62676094, in part by the National Key Research and Development Program of China under Grant 2024YFB4303805 and in part by the research start-up fund of USTC.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. Not applicable. This study did not involve human participants or animals.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Materials availability. Not applicable.

Code availability. The code used in this study is available from the corresponding author upon reasonable request.

Author contributions. Li Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing–original draft, and Writing–review and editing. Wenjie Gan, Junhao Liang, Wanyuan Wang, and Yanyan Liu contributed to validation and Writing–review and editing. All authors read and approved the final manuscript.

Appendix A Detailed Implementation Settings

Table A1 summarizes the training, planning, and reward settings used in our experiments.

Table A1 Detailed implementation settings of the proposed method.

Category	Parameter	Value
Training	Environment-interaction budget	90,000 steps
	Optimizer	Adam
	Actor learning rate	3×10^{-4}
	Critic learning rate	3×10^{-4}
	Discount factor γ	0.99
	GAE parameter λ	0.95
	PPO clipping coefficient ϵ	0.2
	Value-loss coefficient c_v	0.5
	Entropy coefficient c_e	0.01
	Rollout length	128 steps
	Optimization epochs per rollout	10
	Mini-batch size	64
Planning	Frontier downsampling interval s	5 cells
	Distance-cost weight λ_d	0.6
	Exploration-value weight λ_e	0.4
Reward	Coverage-reward weight λ_{cov}	1.0
	Target-guidance weight λ_{goal}	1.0
	Obstacle-penalty weight λ_{obs}	1.0
	UAV-collision penalty weight λ_{uav}	1.0
	Communication-penalty weight λ_{com}	1.0
	Communication warning threshold r_w	6 grid cells

References

Ahmed G, Sheltami TR, Mahmoud A (2024) Energy-efficient multi-UAV multi-region coverage path planning approach. Arabian Journal for Science and Engineering 49(9):13185–13202. https://doi.org/10.1007/s13369-024-09295-w

Arias-Perez P, Gautam A, Fernandez-Cortizas M, et al (2024) Exploring unstructured environments using minimal sensing on cooperative nano-drones. IEEE Robotics and Automation Letters 9(12):11202–11209. https://doi.org/10.1109/LRA.2024.3486212

Aydemir F, Cetin A (2023) Multi-agent dynamic area coverage based on reinforcement learning with connected agents. Computer Systems Science and Engineering 45(1):215–230. https://doi.org/10.32604/csse.2023.031116

Carvalho JP, Aguiar AP (2026) Multi-agent reinforcement learning for zero-shot coverage path planning with dynamic UAV networks. Robotics and Autonomous Systems 195:105163. https://doi.org/10.1016/j.robot.2025.105163

Cattai T, Frattolillo F, Lacava A, et al (2025) Multi-UAV reinforcement learning with realistic communication models: Recent advances and challenges. IEEE Open Journal of Vehicular Technology 6:2067–2081. https://doi.org/10.1109/OJVT.2025.3586774

Chen X, Yang J, Zhang L (2026) A multi-UAV cooperative coverage method based on sparse dual-attention reinforcement learning. IEEE Robotics and Automation Letters 11(4):4058–4065. https://doi.org/10.1109/LRA.2026.3664265

Choi SC, Lee Y, Cho SW (2025) Reinforcement learning-integrated evolutionary algorithm for enhanced unmanned aerial vehicle coverage path planning. Swarm and Evolutionary Computation 97:102051. https://doi.org/10.1016/j.swevo.2025.102051

Datsko D, Nekovar F, Penicka R, et al (2024) Energy-aware multi-UAV coverage mission planning with optimal speed of flight. IEEE Robotics and Automation Letters 9(3):2893–2900. https://doi.org/10.1109/LRA.2024.3358581

Du L, Fan Y, Gui M, et al (2023) A multi-regional path-planning method for rescue UAVs with priority constraints. Drones 7(12):692. https://doi.org/10.3390/drones7120692

Duan W, Lu J, Xuan J (2024) Group-aware coordination graph for multi-agent reinforcement learning. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Jeju, South Korea, pp 3926–3934, https://doi.org/10.24963/ijcai.2024/434

Ekechi CC, Elfouly T, Alouani A, et al (2025) A survey on UAV control with multi-agent reinforcement learning. Drones 9(7):484. https://doi.org/10.3390/drones9070484

Feng Z, Wu D, Huang M, et al (2024) Graph-attention-based reinforcement learning for trajectory design and resource assignment in multi-UAV-assisted communication. IEEE Internet of Things Journal 11(16):27421–27434. https://doi.org/10.1109/JIOT.2024.3397823

Frattolillo F, Brunori D, Iocchi L (2023) Scalable and cooperative deep reinforcement learning approaches for multi-UAV systems: A systematic review. Drones 7(4):236. https://doi.org/10.3390/drones7040236

Goarin M, Loianno G (2024) Graph neural network for decentralized multi-robot goal assignment. IEEE Robotics and Automation Letters 9(5):4051–4058. https://doi.org/10.1109/LRA.2024.3371254

Goeckner A, Sui Y, Martinet N, et al (2024) Graph neural network-based multi-agent reinforcement learning for resilient distributed coordination of multi-robot systems. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, Abu Dhabi, United Arab Emirates, pp 5732–5739, https://doi.org/10.1109/IROS58592.2024.10802510

Gong J, Kim H, Lee S (2024) Resilient multi-robot coverage path redistribution using Boustrophedon decomposition for environmental monitoring. Sensors 24(23):7482. https://doi.org/10.3390/s24237482

Güven İ, Yanmaz E (2024) Multi-objective path planning for multi-UAV connectivity and area coverage. Ad Hoc Networks 160:103520. https://doi.org/10.1016/j.adhoc.2024.103520

Kim K, Kim J (2025) Coordinated informative path planning for multi-robot search in open fields. Journal of Intelligent & Robotic Systems 111(2):65. https://doi.org/10.1007/s10846-025-02270-z

Kong X, Zhou Y, Li Z, et al (2024) Multi-UAV simultaneous target assignment and path planning based on deep reinforcement learning in dynamic multiple obstacles environments. Frontiers in Neurorobotics 17:1302898. https://doi.org/10.3389/fnbot.2023.1302898

Latif E, Parasuraman R (2024) Communication-efficient multi-robot exploration using coverage-biased distributed Q-learning. IEEE Robotics and Automation Letters 9(3):2622–2629. https://doi.org/10.1109/LRA.2024.3358095

Li Y, Chen W, Fu B, et al (2024a) A global coverage path planning method for multi-UAV maritime surveillance in complex obstacle environments. Drones 8(12):764. https://doi.org/10.3390/drones8120764

Li Y, Ma M, Cao J, et al (2024b) A method for multi-AUV cooperative area search in unknown environment based on reinforcement learning. Journal of Marine Science and Engineering 12(7):1194. https://doi.org/10.3390/jmse12071194

Liu H, Long X, Li Y, et al (2025) Adaptive multi-UAV cooperative path planning based on novel rotation artificial potential fields. Knowledge-Based Systems 317:113429. https://doi.org/10.1016/j.knosys.2025.113429

Luna MA, Molina M, Da-Silva-Gomez R, et al (2024) A multi-UAV system for coverage path planning applications with in-flight re-planning capabilities. Journal of Field Robotics 41(5):1480–1497. https://doi.org/10.1002/rob.22342

Lyu M, Zhao Y, Huang C, et al (2023) Unmanned aerial vehicles for search and rescue: A survey. Remote Sensing 15(13):3266. https://doi.org/10.3390/rs15133266

Nguyen LV, Le TH, Torres Herrera I, et al (2024) Intelligent path planning for civil infrastructure inspection with multi-rotor aerial vehicles. Construction Robotics 8(2):19. https://doi.org/10.1007/s41693-024-00135-9

Ni J, Ge Y, Zhao Y, et al (2025) An improved multi-UAV area coverage path planning approach based on deep Q-networks. Applied Sciences 15(20):11211. https://doi.org/10.3390/app152011211

Ning Y, Li T, Yao C, et al (2024) HMS-RRT: A novel hybrid multi-strategy rapidly-exploring random tree algorithm for multi-robot collaborative exploration in unknown environments. Expert Systems with Applications 247:123238. https://doi.org/10.1016/j.eswa.2024.123238

Poudel S, Moh S (2023) Priority-aware task assignment and path planning for efficient and load-balanced multi-UAV operation. Vehicular Communications 42:100633. https://doi.org/10.1016/j.vehcom.2023.100633

Rahman M, Sarkar NI, Lutui R (2025) A survey on multi-UAV path planning: Classification, algorithms, open research problems, and future directions. Drones 9(4):263. https://doi.org/10.3390/drones9040263

Ramezani M, Amiri Atashgah MA, Rezaee A (2024) A fault-tolerant multi-agent reinforcement learning framework for unmanned aerial vehicles–unmanned ground vehicle coverage path planning. Drones 8(10):537. https://doi.org/10.3390/drones8100537

Schulman J, Wolski F, Dhariwal P, et al (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:170706347 https://doi.org/10.48550/arXiv.1707.06347

Song J, Zhao K, Liu Y (2023) Survey on mission planning of multiple unmanned aerial vehicles. Aerospace 10(3):208. https://doi.org/10.3390/aerospace10030208

Tang J, Ma H (2024) Large-scale multi-robot coverage path planning via local search. Proceedings of the AAAI Conference on Artificial Intelligence 38(16):17567–17574. https://doi.org/10.1609/aaai.v38i16.29707

Tran VP, Garratt MA, Kasmarik K, et al (2022) Frontier-led swarming: Robust multi-robot coverage of unknown environments. Swarm and Evolutionary Computation 75:101171. https://doi.org/10.1016/j.swevo.2022.101171

Wang J, Wang R (2024) Multi-UAV area coverage track planning based on the Voronoi graph and attention mechanism. Applied Sciences 14(17):7844. https://doi.org/10.3390/app14177844

Wang Q, Zhang Y, Wang Y, et al (2025) Coverage path planning of UAV cluster based on independent PPO in complex environment. In: Proceedings of the 37th Chinese Control and Decision Conference (CCDC). IEEE, Xiamen, China, pp 1917–1922, https://doi.org/10.1109/CCDC65474.2025.11090996

Wang Z, Zhao X, Zhang J, et al (2024) APF-CPP: An artificial potential field based multi-robot online coverage path planning approach. IEEE Robotics and Automation Letters 9(11):9199–9206. https://doi.org/10.1109/LRA.2024.3432351

Westheider J, Rückin J, Popović M (2023) Multi-UAV adaptive path planning using deep reinforcement learning. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, Detroit, MI, USA, pp 649–656, https://doi.org/10.1109/IROS55552.2023.10342516

Xiao P, Li N, Xie F, et al (2023) Clustering-based multi-region coverage-path planning of heterogeneous UAVs. Drones 7(11):664. https://doi.org/10.3390/drones7110664

Yan X, Zeng Z, He K, et al (2023) Multi-robot cooperative autonomous exploration via task allocation in terrestrial environments. Frontiers in Neurorobotics 17:1179033. https://doi.org/10.3389/fnbot.2023.1179033

Ying C, Cai T, Luo S, et al (2021) Do transformers really perform badly for graph representation? In: Advances in Neural Information Processing Systems, pp 28877–28888

Yu C, Velu A, Vinitsky E, et al (2022) The surprising effectiveness of PPO in cooperative multi-agent games. In: Advances in Neural Information Processing Systems, pp 24611–24624

Zhao B, Huo M, Li Z, et al (2024a) Graph-based multi-agent reinforcement learning for large-scale UAVs swarm system control. Aerospace Science and Technology 150:109166. https://doi.org/10.1016/j.ast.2024.109166

Zhao X, Yang R, Zhong L, et al (2024b) Multi-UAV path planning and following based on multi-agent reinforcement learning. Drones 8(1):18. https://doi.org/10.

3390/drones8010018

Zhou B, Xu H, Shen S (2023) RACER: Rapid collaborative exploration with a decentralized multi-UAV system. IEEE Transactions on Robotics 39(3):1816–1835. https://doi.org/10.1109/TRO.2023.3236945