Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2026.073202

ARTICLE

# Segment-Conditioned Latent-Intent Framework for Cooperative Multi-UAV Search

**Gang Hou[1,#], Aifeng Liu[1,#], Tao Zhao[1], Wenyuan Wei[2], Bo Li[1], Jiancheng Liu[3,*] and Siwen Wei[4,5,*]**

[1]Northwest Institute of Mechanical and Electrical Engineering, Xianyang, 712099, China
[2]Department of Railway Transportation Operations Management, Baotou Railway Vocational & Technical College, Baotou, 014060, China
[3]School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China
[4]Shaanxi Key Laboratory of Antenna and Control Technology, Xi'an, 710076, China
[5]39th Research Institute of China Electronics Technology Group Corporation, Xi'an, 710076, China

*Corresponding Authors: Jiancheng Liu. Email: jianchengliu@njust.edu.cn; Siwen Wei. Email: siwen_wei@stu.xidian.edu.cn

[#]These authors contributed equally to this work

Received: 12 September 2025; Accepted: 24 December 2025; Published: 10 February 2026

**ABSTRACT:** Cooperative multi-UAV search requires jointly optimizing wide-area coverage, rapid target discovery, and endurance under sensing and motion constraints. Resolving this coupling enables scalable coordination with high data efficiency and mission reliability. We formulate this problem as a discounted Markov decision process on an occupancy grid with a cellwise Bayesian belief update, yielding a Markov state that couples agent poses with a probabilistic target field. On this belief–MDP we introduce a segment-conditioned latent-intent framework, in which a discrete intent head selects a latent skill every $K$ steps and an intra-segment GRU policy generates per-step control conditioned on the fixed intent; both components are trained end-to-end with proximal updates under a centralized critic. On the $50 \times 50$ grid, coverage and discovery convergence times are reduced by up to 48% and 40% relative to a flat actor-critic benchmark, and the aggregated convergence metric improves by about 12% compared with a state-of-the-art hierarchical method. Qualitative analyses further reveal stable spatial sectorization, low path overlap, and fuel-aware patrolling, indicating that segment-conditioned latent intents provide an effective and scalable mechanism for coordinated multi-UAV search.

**KEYWORDS:** Multi-agent reinforcement learning; Markov decision process; multi-UAV cooperative search

## 1 Introduction

Cooperative search with multiple unmanned aerial vehicles (UAVs) enables rapid, wide-area situational awareness for surveillance, humanitarian response, and defense operations, where parallel sensing and timely localization of mission-relevant targets are critical [1,2]. These demands are further amplified by the proliferation of IoT devices and the emergence of next-generation (6G-ready) network architectures, in which UAV-assisted infrastructures are increasingly deployed as agile aerial nodes for real-time sensing, surveillance, and data collection in dense IoT environments [3–5]. In such settings, decision policies must simultaneously promote broad spatial coverage, high-probability target discovery, and fuel-aware persistence under partial observability and dynamic interaction among agents, which renders joint planning inherently high dimensional and nonstationary.

Conventional approaches to multi-UAV search span graph-based planning, swarm heuristics, game-theoretic coordination, and evolutionary/metaheuristic optimizers [6–8]. While these methods provide valuable baselines, they typically assume static or fully observable environments, rely on strong prior modeling or hand-crafted heuristics, and offer limited support for online adaptation, multi-agent credit assignment, and principled trade-offs between coverage, discovery, and endurance. Deep reinforcement learning has recently advanced UAV navigation by learning directly from interaction and coping with uncertainty [9–11]. In multi-agent settings, centralized training with decentralized execution (CTDE) alleviates nonstationarity [12–14], yet flat (single-level) DRL still struggles with long-horizon exploration, joint action-space blowup, and ambiguous credit assignment as team size grows [15].

Hierarchical reinforcement learning (HRL) provides temporal abstraction and subgoal structure through manager–worker decompositions and value-function factorizations [16,17]. Canonical architectures such as FeUdal Networks and the Option-Critic framework instantiate these principles via goal-conditioned managers and learnable options [18], but typically require delicate intrinsic-reward design or learned termination functions that are prone to option collapse and training instability. Recent multi-agent extensions further exploit hierarchical organization to facilitate cooperative decision-making [19,20], yet often rely on hand-crafted subtask taxonomies, predefined communication patterns, or centralized coordinators that do not transfer seamlessly to fully cooperative, partially observed settings. Surveys underscore the potential of HRL for scalable aerial coordination [21,22]; nevertheless, these designs can incur additional variance and computational overhead when applied to belief-based multi-UAV search with tightly coupled objectives. Maximum-entropy regularization has been shown to enhance exploration and coordination [23], yet systematically aligning exploratory behavior with mission-level coverage objectives and energy constraints remains a central open challenge.

This work introduces a segment-conditioned latent-intent framework for cooperative multi-UAV search (SCLI–CMUS) that unifies temporal abstraction, coordinated exploration, and endurance awareness within a single CTDE policy. The environment is modeled as a discounted Markov decision process on a discretized workspace endowed with a Bayesian cellwise update of the occupancy field. Within one end-to-end differentiable policy, a discrete *intent head* selects a latent skill every $K$ steps to guide medium-horizon behavior, while an *action head* driven by an intra-segment GRU issues per-step yaw increments conditioned on the fixed intent and local features. Compared with FeUdal-style and Option-Critic architectures, this fixed-horizon, discrete-intent design retains sufficient temporal expressiveness while avoiding termination-related instabilities and keeping per-step computation close to that of a recurrent actor-critic with a single additional categorical head. To reconcile heterogeneous signal scales and stabilize training, we employ a three-parameter, scale-calibrated saturated reward that jointly accounts for information gain, coverage efficiency, and energy–time cost. The principal contributions of this work are summarized as follows:

1. We propose Segment–Conditioned Latent–Intent for Cooperative Multi–UAV Search (SCLI–CMUS), a CTDE framework that couples a discrete intent selector—updated at fixed segment boundaries—with an intra-segment recurrent controller in a single end-to-end differentiable policy.
2. We develop a three-coefficient, scale-calibrated saturated reward that jointly balances information gain, coverage efficiency, and energy–time costs.
3. Comprehensive experiments demonstrating faster learning, improved overage/discovery convergence times, and robust qualitative behaviors across representative UAV team sizes.

The remainder of this paper is structured as follows. Section 2 reviews related work on cooperative multi-UAV search, planning-based coordination, and (hierarchical) multi-agent reinforcement learning. Section 3 presents the belief-based MDP formulation and the proposed segment-conditioned

latent-intent framework, while Section 4 details the experimental setup, benchmarks, and ablation studies. Finally, Section 5 summarizes the findings and outlines directions for future research.

## 2 Related Work

Conventional planning approaches for UAV search largely build on graph-search and shortest-path heuristics, which perform well in static and fully observable environments [6,24]. However, such methods do not naturally accommodate multi-agent credit assignment, online replanning under sensing uncertainty, or principled division of labor among multiple vehicles. Swarm-style schemes based on artificial potential fields, flocking rules, or pheromone deposition provide lightweight, decentralized coordination with low communication burden [7,25], and learned heuristics can amortize local perception-to-action mappings [26]. These approaches, though attractive for their simplicity, are susceptible to local minima, lack mechanisms to globally optimize coupled coverage–discovery–endurance trade-offs, and often require extensive manual retuning when environment statistics change.

Game-theoretic and metaheuristic frameworks offer alternative tools for cooperative search and routing. Potential games and market-based task allocation furnish equilibrium concepts and scalable assignment rules [27,28], while differential games capture adversarial or pursuit–evasion interactions in continuous time [28]. Evolutionary and metaheuristic optimizers traverse nonconvex search spaces and can handle multiple objective criteria in path planning and routing [29,30], with recent advances in motion-encoded multi-parent crossovers improving solution diversity [8]. Nonetheless, their reliance on strong prior modeling, offline optimization, and substantial computational budgets limits their ability to adapt online in uncertain, time-critical environments.

Deep reinforcement learning (DRL) has achieved notable success in UAV navigation by learning policies directly from interaction, thereby coping with uncertainty and enabling online adaptation [9]. Single-vehicle studies have demonstrated obstacle-aware, threat-aware maneuvering and memory-augmented exploration in complex environments [10,11,31,32]. In multi-agent settings, centralized training with decentralized execution (CTDE) has become a standard paradigm: decentralized actors are trained against a centralized critic to mitigate nonstationarity and stabilize learning [12], and recent work emphasizes resilience and meta-adaptation under distribution shift [13,14]. However, flat (single-level) DRL typically struggles to align long-horizon exploration with local motion control, suffers from joint action-space blowup as team size grows, and faces ambiguous multi-agent credit assignment [15]. Maximum-entropy and entropy-regularized formulations can improve exploration and coordination [23], but designing reward structures that explicitly couple information gain, coverage efficiency, and energy–time costs remains challenging in belief-based multi-UAV search.

Hierarchical reinforcement learning (HRL) introduces temporal abstraction and subgoal structure through manager–worker decompositions and value-function factorizations [16,17]. Canonical architectures such as FeUdal Networks and the Option-Critic framework realize these principles via goal-conditioned managers and learnable options [18,33], but typically require carefully designed intrinsic rewards or learned termination functions that are prone to option collapse, premature termination, and training instability. Against this backdrop, the segment-conditioned latent-intent framework developed in this work is intended to preserve the advantages of temporal abstraction while mitigating practical difficulties associated with option termination and hand-crafted subtask hierarchies.

## 3 Method

Fig. 1 provides an overview of the proposed framework. The environment yields a probabilistic occupancy map $b(t)$, local observations $o_t^u$, and agent coordinates $\mathbf{p}_t^u$. These signals are fused by an encoder

(GRU) into global information features that condition two heads within a single policy: a discrete skill head $\pi_\phi$ that selects a segment intent $z^u_{t_k}$ every $K$ steps, and an action head $\pi_\theta$ that issues per-step yaw-increment commands conditioned on the fixed intent and an autoregressive hidden state. A centralized critic $V_\psi$ evaluates behaviour on the belief state and supplies advantages for step-level and segment-level updates; experience tuples are stored in a replay buffer.

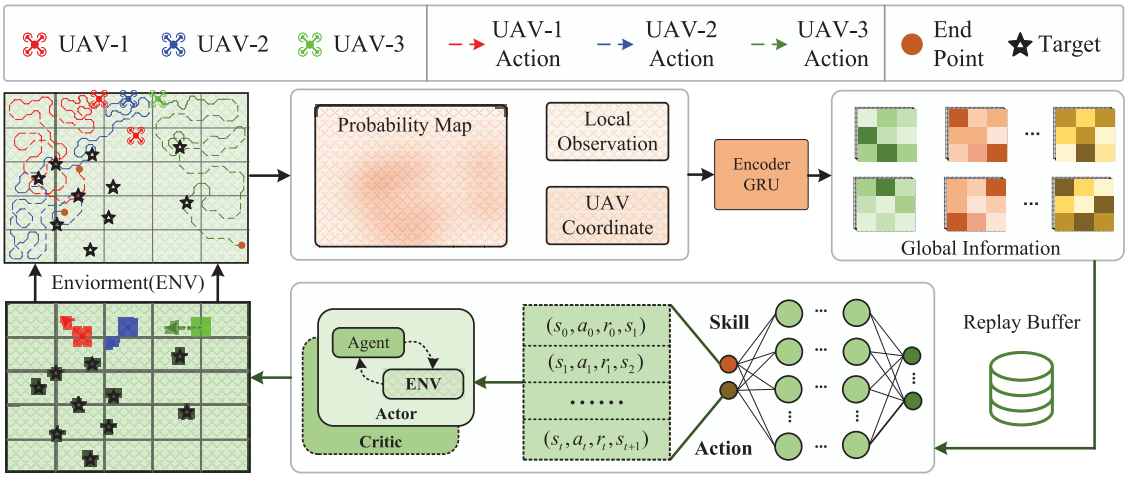


**Figure 1:** Overview of the segment-conditioned latent-intent framework for cooperative multi-UAV search (SCLI-CMUS). Left: environment and trajectories of multiple UAVs (colours identify agents; stars indicate targets; orange dots denote end points). Middle: inputs to the encoder comprising the probability map $b(t)$, local observation $o^u_t$, and UAV coordinates $\mathbf{p}^u_t$; the encoder (GRU) produces global information features. Right: policy heads and learning signals. The skill head $\pi_\phi$ selects a discrete intent every $K$ steps; the action head $\pi_\theta$ outputs per-step yaw increments conditioned on the intent and an autoregressive state

The main symbols used in the formulation and policy parameterization are summarized in Table 1.

**Table 1:** Core notation used in the problem formulation and policy

| Category | Symbol | Description |
|---|---|---|
| Environment & State | $\mathcal{D},\ D_X, D_Y$ | Discretized grid workspace and its size. |
| | $\mathcal{U},\ N$ | UAV index set and team size. |
| | $S_t$ | MDP state (UAV poses and belief field). |
| Sensing & Coverage | $R_{\text{sen}},\ \text{Cov}_u(x, y; t)$ | Sensing radius and per-UAV cell coverage indicator. |
| | $\mathcal{V}_t,\ \text{CovRate}(t)$ | Visited cell set and instantaneous coverage ratio. |
| Belief & Reward | $b_{x,y}(t),\ b(t)$ | Cellwise occupancy belief and full belief field. |
| | $R_t,\ \widehat{\mathcal{I}}_t, \widehat{\mathcal{C}}_t, \widehat{\mathcal{E}}_t,$ | Team reward, normalized components, reward weights $\lambda_{\mathcal{I}}, \lambda_{\mathcal{C}}, \lambda_{\mathcal{E}}$, discount factor $\gamma$, and discounted return $J$. |
| | $\lambda_{\mathcal{I}}, \lambda_{\mathcal{C}}, \lambda_{\mathcal{E}},\ \gamma,\ J$ | |

(Continued)

**Table 1 (continued)**

| Category | Symbol | Description |
|---|---|---|
| Latent Intents & Segments | $K,\ t_k$ | Segment length and segment-start time index. |
| | $z^u_{t_k},\ M$ | Latent intent of UAV $u$ at $t_k$ and number of discrete intents. |
| Actions & Policy | $\mathcal{A},\ a^u_t,\ \alpha$ | Discrete yaw–increment set, action of UAV $u$, and increment magnitude. |
| | $\pi^u_\phi,\ \pi^u_\theta,\ V_\psi$ | Skill head, action head, and centralized critic. |
| Features & Hidden state | $o^u_t,\ g^{\text{map}}_t, g^{\text{loc},u}_t, G^u_t$ | Local observation, belief-map feature, local feature, and aggregated summary. |
| | $h^u_t,\ \text{GRU}_\omega$ | Recurrent hidden state and GRU encoder. |

### *3.1 MDP Formulation for Cooperative Multi-UAV Search*

We model cooperative multi-UAV search as a discounted Markov decision process on a discretized workspace. Let the two-dimensional workspace be discretized as

$$\mathcal{D} = \{(x, y) \mid x = 1, \ldots, D_X,\ y = 1, \ldots, D_Y\}, \tag{1}$$

where each cell $(x, y)$ either contains a target or is empty, and $|\mathcal{D}| = D_X D_Y$ denotes the total number of grid cells. Time is discrete $t = 0, 1, 2, \ldots$. The cellwise occupancy posterior is

$$b_{x,y}(t) = \mathbb{P}\big(\text{cell } (x, y) \text{ contains a target} \,\big|\, \mathcal{Z}_{1:t}\big), \tag{2}$$

with $\mathcal{Z}_{1:t}$ the $\sigma$-algebra generated by all measurements up to $t$. The initial prior $b_{x,y}(0)$ is specified from domain knowledge (uniform). Denote by $b(t) = \{b_{x,y}(t)\}_{(x,y)\in\mathcal{D}}$ the full occupancy field.

We now specify the agent set, action space, kinematics, joint control, and state representation. Let $\mathcal{U} = \{1, \ldots, N\}$ index the UAVs. UAV $u$ has planar position $\mathbf{p}^u_t \in \mathbb{R}^2$ and heading $\psi^u_t \in (-\pi, \pi]$. Each UAV selects a discrete yaw–increment action

$$a^u_t \in \mathcal{A} = \{-\alpha,\ 0,\ +\alpha\} \ \text{(degrees)}, \tag{3}$$

where $\alpha = 45$, and the action evolves under constant–speed kinematics with sampling step $\Delta t > 0$ and speed $v > 0$.

Given a chosen yaw–increment action $a^u_t$ and constant-speed motion, the heading and planar position of UAV $u$ evolve according to

$$\begin{cases} \psi^u_{t+1} = \psi^u_t + \dfrac{\pi}{180}\, a^u_t, \\ \mathbf{p}^u_{t+1} = \mathbf{p}^u_t + v\Delta t \begin{bmatrix} \cos\psi^u_{t+1} \\ \sin\psi^u_{t+1} \end{bmatrix}, \end{cases} \tag{4}$$

where $v$ is the constant forward speed and $\Delta t$ is the sampling interval.

Collecting all per-agent yaw–increment actions yields the joint action vector

$$\mathbf{a}_t = (a_t^1, \ldots, a_t^N) \in \mathcal{A}^N. \tag{5}$$

The MDP state collects agent poses and the occupancy field, namely

$$S_t = \big(\{\mathbf{p}_t^u, \psi_t^u\}_{u\in\mathcal{U}},\ b(t)\big). \tag{6}$$

Each UAV carries an omnidirectional sensing disc of radius $R_{\text{sen}} > 0$. A grid cell $(x, y)$ is classified as fully observed by UAV $u$ at time $t$ if all four vertices $\partial\mathcal{D}(x, y)$ lie within the disc centered at $\mathbf{p}_t^u$:

$$\mathrm{Cov}_u(x, y; t) = \begin{cases} 1, & \max\limits_{(x',y')\in\partial\mathcal{D}(x,y)} \left\|\mathbf{p}_t^u - (x', y')\right\| \le R_{\text{sen}}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The instantaneous covered set is defined by

$$\mathcal{V}_t = \left\{(x, y) \in \mathcal{D} :\ \sum_{u=1}^{N} \mathrm{Cov}_u(x, y; t) \ge 1\right\}, \tag{8}$$

where the coverage ratio at time $t$ is given by $\mathrm{CovRate}(t) = \frac{|\mathcal{V}_t|}{|\mathcal{D}|}$. The cumulative exploration status of a cell is recorded by

$$\mathrm{Visited}(x, y; t) = \mathbb{I}\left\{\exists\, \tau \le t :\ \sum_{u=1}^{N} \mathrm{Cov}_u(x, y; \tau) \ge 1\right\}. \tag{9}$$

Let $Z_t^u$ denote the measurement field collected by UAV $u$ at time $t$. For any covered cell $(x, y)$ with $\mathrm{Cov}_u(x, y; t) = 1$, the binary detector on UAV $u$ obeys

$$\mathbb{P}\big(Z_{x,y}^u(t) = 1 \,\big|\, T_{x,y}\big) = \begin{cases} P_D^u, & \text{if } T_{x,y} = 1, \\ P_{FA}^u, & \text{if } T_{x,y} = 0, \end{cases} \tag{10}$$

where $Z_{x,y}^u(t) \in \{0, 1\}$ is the binary measurement at cell $(x, y)$ and time $t$, $T_{x,y} \in \{0, 1\}$ denotes the hidden occupancy of cell $(x, y)$, and $P_D^u$, $P_{FA}^u$ are the detection and false–alarm probabilities of UAV $u$, respectively. The set of UAVs that cover $(x, y)$ at time $t$ is

$$\mathcal{U}_{x,y}(t) = \big\{u \in \mathcal{U} :\ \mathrm{Cov}_u(x, y; t) = 1\big\}. \tag{11}$$

Under conditional independence given $T_{x,y}$, the joint likelihoods for the measurements on $(x, y)$ at time $t$ are

$$\begin{cases} \mathcal{L}_1 = \prod\limits_{u\in\mathcal{U}_{x,y}(t)} \big(P_D^u\big)^{Z_{x,y}^u(t)} \big(1 - P_D^u\big)^{1-Z_{x,y}^u(t)}, \\ \mathcal{L}_0 = \prod\limits_{u\in\mathcal{U}_{x,y}(t)} \big(P_{FA}^u\big)^{Z_{x,y}^u(t)} \big(1 - P_{FA}^u\big)^{1-Z_{x,y}^u(t)}. \end{cases} \tag{12}$$

The cellwise Bayesian update is then

$$b_{x,y}(t+1) = \begin{cases} \dfrac{b_{x,y}(t)\,\mathcal{L}_1}{b_{x,y}(t)\,\mathcal{L}_1 + \big(1 - b_{x,y}(t)\big)\,\mathcal{L}_0}, & \mathcal{U}_{x,y}(t) \neq \varnothing, \\ b_{x,y}(t), & \mathcal{U}_{x,y}(t) = \varnothing. \end{cases} \tag{13}$$

The transition kernel induced by the deterministic kinematics and the cellwise Bayesian update is

$$\mathbb{P}(S_{t+1} \mid S_t, \mathbf{a}_t) = \prod_{u=1}^{N} \delta\left(\mathbf{p}_{t+1}^{u} - \mathbf{p}_t^{u} - v\Delta t[\cos\psi_{t+1}^{u}, \sin\psi_{t+1}^{u}]^{\top}\right) \prod_{(x,y)\in\mathcal{D}} \delta\left(b_{x,y}(t+1) - \mathrm{Bayes}(b_{x,y}(t), Z_{x,y}(t))\right). \tag{14}$$

The team reward adopts a scale-calibrated saturated form, first define analytic normalizers

$$U_{\max} = N\pi R_{\mathrm{sen}}^2, \qquad E_{\max} = c_0 N\Delta t + c_\psi \frac{\pi}{180} N\alpha, \tag{15}$$

with fixed coefficients $c_0 = 1.0 \times 10^{-3}$ and $c_\psi = 1.0 \times 10^{-2}$, and construct dimensionless components

$$\widehat{\mathcal{I}}_t = \frac{H(b(t)) - H(b(t+1))}{U_{\max} \ln 2}, \qquad \widehat{\mathcal{C}}_t = C_t, \qquad \widehat{\mathcal{E}}_t = \frac{\mathcal{E}_t}{E_{\max}}. \tag{16}$$

The saturated team reward is then

$$R_t = \lambda_{\mathcal{I}} \tanh\big(\widehat{\mathcal{I}}_t\big) + \lambda_C \tanh\big(\widehat{\mathcal{C}}_t\big) - \lambda_{\mathcal{E}} \tanh\big(\widehat{\mathcal{E}}_t\big), \tag{17}$$

where $\lambda_{\mathcal{I}}$, $\lambda_C$, and $\lambda_{\mathcal{E}}$ are positive weighting coefficients that control the relative importance of information gain, coverage efficiency, and energy–time cost, respectively.

The reward components are defined as follows. The entropy of the belief field is

$$H(b(t)) = \sum_{(x,y)\in\mathcal{D}} \Big[ - b_{x,y}(t) \ln b_{x,y}(t) - \big(1 - b_{x,y}(t)\big) \ln\big(1 - b_{x,y}(t)\big)\Big], \tag{18}$$

the coverage efficiency term is

$$C_t = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} b_{x,y}(t)\,\Delta V_{x,y}(t) \cdot \frac{\sum_{(x,y)\in\mathcal{D}} \mathbb{I}\big\{\sum_{u=1}^{N} \mathrm{Cov}_u(x,y;t) \geq 1\big\}\; b_{x,y}(t)}{\max\Big\{1,\; \sum_{(x,y)\in\mathcal{D}} \sum_{u=1}^{N} \mathrm{Cov}_u(x,y;t)\, b_{x,y}(t)\Big\}}, \tag{19}$$

where the energy–time cost term is defined as

$$\mathcal{E}_t = c_0\, N\, \Delta t \;+\; c_\psi\, \frac{\pi}{180} \sum_{u=1}^{N} |a_t^u|, \tag{20}$$

with $\Delta V_{x,y}(t) = \mathrm{Visited}(x,y;t) - \mathrm{Visited}(x,y;t-1)$ denoting the incremental visitation indicator at cell $(x,y)$, and $\mathbb{I}\{\cdot\}$ the indicator function. The coefficients $c_0 > 0$ and $c_\psi > 0$ control the time-related and turning-related energy costs, respectively.

The planning objective is the discounted return

$$J = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right], \qquad \gamma \in (0,1), \tag{21}$$

which fully specifies the discounted MDP on the state $S_t$ and underpins the two-timescale latent–skill policy in Section 3.2.

### *3.2 Segment-Conditioned Latent-Intent Policy Parameterization*

On the discounted MDP of Section 3.1, each agent is equipped with a slow latent skill that governs behaviour over a fixed segment of $K$ steps, while fast per-step actions are generated conditionally on the current skill and an autoregressive hidden state.

Let segment boundaries be $t_k = kK$ with $K \in \mathbb{N}$. At each segment start $t_k$, agent $u$ samples a categorical latent skill from a skill head conditioned on a summary of global and local information,

$$z^u_{t_k} \sim \pi^u_\phi\big(z \mid G^u_{t_k}\big), \qquad z^u_t \equiv z^u_{t_k}, \tag{22}$$

where $z \in \{1, \ldots, M\}$ denotes the discrete intent index, and the skill remains fixed within the current segment $t \in [t_k,\ t_k + K - 1]$. The summary $G^u_t$ aggregates a spatial feature of the belief field with local encodings and team statistics,

$$G^u_t = \mathrm{Agg}\Big(g^{\mathrm{map}}_t,\ g^{\mathrm{loc},u}_t,\ \{p^v_t\}_{v \in \mathcal{U}}\Big), \tag{23}$$

where $g^{\mathrm{map}}_t = \mathrm{CNN}\big(b(t)\big)$ encodes the global belief map, and $g^{\mathrm{loc},u}_t = \mathrm{enc}(o^u_t)$ encodes the local observation of UAV $u$. $\mathrm{Agg}(\cdot)$ is a permutation-invariant aggregation function over team states, and $\mathrm{Emb}(z) \in \mathbb{R}^{d_z}$ denotes a learnable embedding associated with intent $z$.

For each agent $u$, a hidden state evolves within the segment via a GRU driven by local features and the current skill embedding,

$$h^u_t = \mathrm{GRU}_\omega\Big(h^u_{t-1},\ \big[g^{\mathrm{loc},u}_t \ \oplus\ \mathrm{Emb}(z^u_t)\big]\Big), \tag{24}$$

with parameters $\omega$ and concatenation $\oplus$. The resulting state, coupled with the belief feature, parameterizes the action head:

$$\ell^u_t = \mathrm{MLP}_\theta\Big(\big[h^u_t \ \oplus\ g^{\mathrm{map}}_t\big]\Big) \in \mathbb{R}^3, \qquad \pi^u_\theta\big(a^u_t = k \mid h^u_t, \mathrm{Emb}(z^u_t), g^{\mathrm{map}}_t\big) = \frac{\exp(\ell^u_{t,k})}{\sum_{k' \in \{-\alpha, 0, +\alpha\}} \exp(\ell^u_{t,k'})}, \tag{25}$$

with parameters $\theta$.

Collecting parameters $\Theta = (\phi, \theta, \omega)$, the joint policy over a segment factorizes into skill selections at $t_k$ and per-step action selections conditioned on the fixed skill and the evolving hidden state,

$$\Pi_\Theta\big(\mathbf{z}_{t_k}, \mathbf{a}_{t_k:t_k+K-1} \mid S_{t_k:t_k+K-1}\big) = \prod_{u=1}^{N} \pi^u_\phi\big(z^u_{t_k} \mid G^u_{t_k}\big) \prod_{t=t_k}^{t_k+K-1} \prod_{u=1}^{N} \pi^u_\theta\big(a^u_t \mid h^u_t, \mathrm{Emb}(z^u_{t_k}), g^{\mathrm{map}}_t\big), \tag{26}$$

where $h^u_t$ evolves according to (24).

Learning objectives are matched to these timescales. The segment return starting at $t_k$ is

$$G^{(K)}_{t_k} = \sum_{\tau=0}^{K-1} \gamma^\tau R_{t_k+\tau}, \tag{27}$$

where $\gamma \in (0,1)$ is the discount factor, and the stepwise (infinite-horizon) return is given by $G_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} R_{t+\tau}$. A centralized value function on the belief state augmented by latent and hidden summaries is introduced as

$$\Xi_t = \left(S_t,\ \{z_{t_k}^u\}_{u\in\mathcal{U}},\ \{h_t^u\}_{u\in\mathcal{U}}\right), \qquad V_\psi(\Xi_t) \approx \mathbb{E}\left[G_t \mid \Xi_t\right], \tag{28}$$

which supports low-variance advantage estimates at both levels. With temporal-difference residuals and generalized advantage estimation,

$$\delta_t = R_t + \gamma V_\psi(\Xi_{t+1}) - V_\psi(\Xi_t), \qquad \hat{A}_t = \sum_{\ell=0}^{\infty} (\gamma\lambda)^{\ell}\, \delta_{t+\ell}, \quad \lambda \in [0,1], \tag{29}$$

the segment-level advantage at $t_k$ is

$$\hat{A}_{t_k}^{\text{skill}} = G_{t_k}^{(K)} - \bar{b}_\varphi(\Upsilon_{t_k}), \qquad \Upsilon_{t_k} = \left(S_{t_k}, \{G_{t_k}^u\}_{u\in\mathcal{U}}\right), \tag{30}$$

where $\bar{b}_\varphi$ is a $K$-step baseline with parameters $\varphi$.

Optimization is carried out by proximal policy updates at both timescales. For the action head, define the probability ratio

$$r_t^u(\theta) = \frac{\pi_\theta^u(a_t^u \mid h_t^u, \text{Emb}(z_{t_k}^u), g_t^{\text{map}})}{\pi_{\theta_{\text{old}}}^u(a_t^u \mid h_t^u, \text{Emb}(z_{t_k}^u), g_t^{\text{map}})}, \tag{31}$$

and minimize the clipped surrogate aggregated over agents,

$$\mathcal{L}_{\text{PPO-act}}(\theta,\psi) = -\sum_{u=1}^{N} \mathbb{E}\left[\min\left(r_t^u \hat{A}_t,\ \text{clip}(r_t^u, 1\pm\varepsilon)\hat{A}_t\right)\right] + c_v\, \mathbb{E}\left[(V_\psi(\Xi_t) - \hat{V}_t)^2\right] - c_{\text{ent}} \sum_{u=1}^{N} \mathbb{E}\left[\mathcal{H}(\pi_\theta^u)\right], \tag{32}$$

with clip parameter $\varepsilon > 0$, weights $c_v, c_{\text{ent}} > 0$, and bootstrap target $\hat{V}_t$. For the skill head, the segment–start ratio

$$r_{t_k}^{(z),u}(\phi) = \frac{\pi_\phi^u(z_{t_k}^u \mid G_{t_k}^u)}{\pi_{\phi_{\text{old}}}^u(z_{t_k}^u \mid G_{t_k}^u)} \tag{33}$$

leads to the segment–level surrogate

$$\mathcal{L}_{\text{PPO-skill}}(\phi) = -\sum_{u=1}^{N} \mathbb{E}\left[\min\left(r_{t_k}^{(z),u}\, \hat{A}_{t_k}^{\text{skill}},\ \text{clip}(r_{t_k}^{(z),u}, 1\pm\varepsilon)\, \hat{A}_{t_k}^{\text{skill}}\right)\right] - c_{\text{ent-z}} \sum_{u=1}^{N} \mathbb{E}\left[\mathcal{H}(\pi_\phi^u)\right], \tag{34}$$

with $c_{\text{ent-z}} > 0$. Combining both timescales yields the overall objective

$$\min_{\theta,\phi,\psi}\ \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PPO-act}}(\theta,\psi) + \mathcal{L}_{\text{PPO-skill}}(\phi). \tag{35}$$

## 4 Experiments

### *4.1 Parameter Setting*

All experiments strictly follow the discounted MDP and sensing specification described above, and adopt the conventional benchmark setting used in prior cooperative search studies [28,34], with stationary targets and ideal, latency-free inter-UAV communication. The basic setup consists of a $50 \times 50$ grid with three UAVs and ten stationary targets. The sensor footprint is set to $R_{\text{sen}} = 0.8$ (in grid-cell units), which determines

per-step observable area and thereby the rate of uncertainty reduction. Optimization proceeds with PPO using $\gamma = 0.99$, $\lambda = 0.95$, and clipping coefficient $\varepsilon = 0.20$; actor and critic learning rates are both $3 \times 10^{-4}$. Reward shaping follows the three-parameter saturated design in Eq. (17), with $(\lambda_I, \lambda_C, \lambda_{\mathcal{E}}) = (1.0, 1.0, 0.1)$ to balance information gain and coverage against energy–time penalization on a common scale.

All baseline networks are configured with comparable capacity and are trained using similar batch sizes. For the PPO-based SCLI–CMUS, we use on-policy trajectory batches without long-term replay, whereas the MADDPG-based methods rely on a replay buffer of fixed capacity with minibatch sampling.

### *4.2 Performance Benchmark*

1. **DQN method [35]:** Each UAV runs an independent Deep Q-Network on its local observation to estimate action values for the discrete yaw increments. The absence of explicit coordination limits information sharing and typically degrades scalability in larger or cluttered maps.
2. **ACO method [36]:** Ant Colony Optimization governs motion through a pheromone field over the grid, where each UAV behaves as an "ant" that deposits and follows trails. The induced three-dimensional pheromone tensor encodes heading preferences per cell and per agent.
3. **MADDPG [28]:** A canonical CTDE actor–critic baseline on the same Markov decision formulation. It employs decentralized actors with a centralized critic and experience replay. Using identical observation and reward interfaces enables a direct assessment of the gains attributable to hierarchical intent mechanisms and difference-reward shaping.
4. **Maximum-Entropy RL (ME-RL) [23]:** An entropy-regularized extension of MADDPG that incorporates spatial entropy and fuzzy logic to encourage exploration and coordination under communication and energy considerations.
5. **DTH–MADDPG [34]:** A hierarchical reinforcement-learning framework with a slow strategic controller and a set of fast decentralized executors. The strategic layer updates intermittently to assign high-level intents (region/waypoint directives) to the team, while the executor layer implements per-UAV control via MADDPG under CTDE with replay.

### *4.3 Evaluation Metrics*

We report task performance through spatial coverage and target discovery, and we quantify search efficiency via convergence times to fixed performance levels. Let $\mathcal{D}$ denote the grid, $\mathcal{V}_t \subseteq \mathcal{D}$ the set of cells visited at least once by time $t$, $N_\star$ the total number of targets, and $N_{\mathrm{det}}(t)$ the number detected by time $t$. Define the instantaneous fractions

$$\kappa(t) = \frac{|\mathcal{V}_t|}{|\mathcal{D}|}, \qquad \delta(t) = \frac{N_{\mathrm{det}}(t)}{N_\star}. \tag{36}$$

To capture the speed at which operational effectiveness is achieved, introduce coverage and discovery convergence times as first hitting times of prescribed thresholds $\rho_{\mathrm{cov}}, \rho_{\mathrm{det}} \in (0, 1]$:

$$\tau_{\mathrm{cov}} = \inf\{t \in \mathbb{N}_0 : \kappa(t) \geq \rho_{\mathrm{cov}}\}, \qquad \tau_{\mathrm{det}} = \inf\{t \in \mathbb{N}_0 : \delta(t) \geq \rho_{\mathrm{det}}\}. \tag{37}$$

In all experiments we set $\rho_{\mathrm{cov}} = \rho_{\mathrm{det}} = 0.85$. Smaller values of $\tau_{\mathrm{cov}}$ and $\tau_{\mathrm{det}}$ indicate faster attainment of wide-area exploration and target acquisition, respectively, and correlate with reduced flight time and energy expenditure.

### *4.4 Performance Evaluation*

This section quantifies learning efficiency and asymptotic performance of the proposed method relative to a strong baseline. Fig. 2 reports episode–wise learning curves, where the horizontal axis denotes the training episode index and the vertical axis denotes the total episode reward computed under the reward design in Section 3.1. Curves correspond to the mean over repeated runs, and shaded bands depict variability across runs.

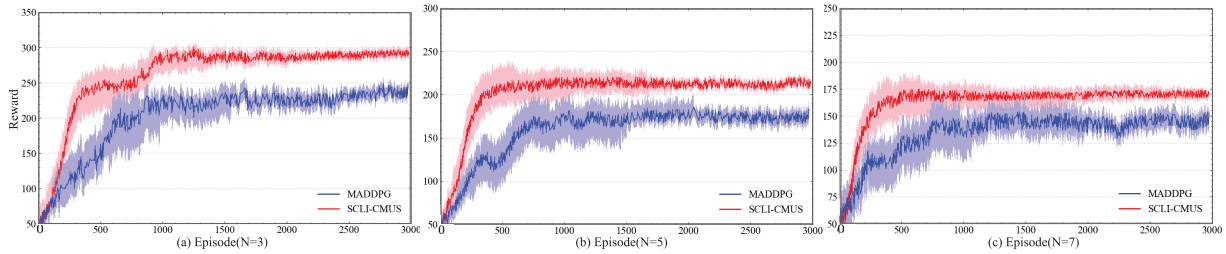


**Figure 2:** Episode reward vs. training episode for MADDPG (blue) and SCLI–CMUS (red). The horizontal axis denotes episode index; the vertical axis denotes total reward per episode; shaded regions indicate variability across runs. (**a**) $N = 3$ UAVs, (**b**) $N = 5$ UAVs, (**c**) $N = 7$ UAVs

A consistent pattern emerges across all subplots. The proposed SCLI–CMUS (red) rises sharply at early episodes and reaches a high plateau with markedly reduced dispersion, whereas MADDPG (blue) exhibits slower ascent, a lower steady level, and wider fluctuations. This behaviour is most pronounced in Fig. 2a with three agents, where SCLI–CMUS achieves a visibly higher steady reward and converges in substantially fewer episodes. The gap persists in Fig. 2b, indicating that the advantage is robust when scaling to five agents. The reduced variance of SCLI–CMUS is consistent with the segment–conditioned intent mechanism and the saturated, scale–balanced reward, which together suppress redundant exploration and stabilize gradient updates.

The scaling trend with agent count is also informative. Moving from three to seven agents, both methods display a gradual reduction in asymptotic reward, which is consistent with fixed–horizon evaluation: faster attainment of high coverage leaves a longer terminal phase dominated by energy–time penalization. Despite this shift in absolute level, SCLI–CMUS maintains a persistent margin and tighter confidence bands in Fig. 2c, indicating improved coordination under higher platform density.

In Table 2 (Search Area = 50 × 50), the proposed SCLI–CMUS achieves the best coverage and discovery convergence across all team sizes. For $N = 3$, SCLI–CMUS reduces $\tau_{\text{cov}}$ to 1102 and $\tau_{\text{det}}$ to 1232, yielding improvements of approximately 36% and 29% relative to MADDPG (1718/1723) and 32% and 23% relative to ME–RL (1611/1598). Against the strongest hierarchical baseline (DTH–MADDPG), SCLI–CMUS still provides 21% faster coverage and 9% faster discovery (1397/1348 vs. 1102/1232). As the team scales, the margins persist. For $N = 5$, $\tau_{\text{cov}}$ and $\tau_{\text{det}}$ fall to 860/1137, improving over MADDPG by 43%/20% and over DTH–MADDPG by 13%/10%. At $N = 7$, SCLI–CMUS attains 710/997, exceeding MADDPG by 48%/40% and DTH–MADDPG by 11%/6%. The aggregate "Total" column confirms the trend: 6038 for SCLI–CMUS vs. 6853 for DTH–MADDPG (≈12% gain) and 9400 for MADDPG (≈36% gain). These gains are attributable to segment-conditioned intent selection and scale-calibrated reward saturation, which jointly suppress redundant footprint overlap, prioritize high-entropy regions, and stabilize critic estimates under CTDE.

**Table 2:** Coverage convergence time $\tau_{\text{cov}}$ and discovery convergence time $\tau_{\text{det}}$ for 15–target scenarios under varying team size $N \in \{3, 5, 7\}$. Lower is better; the best value in each column is typeset in **bold** and marked with ↓

| Search area | Method | N = 3 | | N = 5 | | N = 7 | | Total |
|---|---|---|---|---|---|---|---|---|
| | | $\tau_{\text{cov}}$ | $\tau_{\text{det}}$ | $\tau_{\text{cov}}$ | $\tau_{\text{det}}$ | $\tau_{\text{cov}}$ | $\tau_{\text{det}}$ | |
| 50 × 50 | **SCLI–CMUS (Ours)** | **1102↓** | **1232↓** | **860↓** | **1137↓** | **710↓** | **997↓** | **6038↓** |
| | DTH-MADDPG | 1397 | 1348 | 987 | 1266 | 798 | 1057 | 6853 |
| | ME–RL | 1611 | 1598 | 1377 | 1308 | 1241 | 1537 | 8673 |
| | MADDPG | 1718 | 1723 | 1510 | 1424 | 1357 | 1668 | 9400 |
| | DQN | 2256 | 1784 | 1998 | 1601 | 1657 | 1828 | 11124 |
| | ACO | 3304 | 3202 | 2160 | 2148 | 1882 | 1964 | 14662 |
| 60 × 60 | **SCLI–CMUS** | **1214↓** | **1371↓** | **970↓** | **1315↓** | **820↓** | **1107↓** | **6797↓** |
| | DTH-MADDPG | 1561 | 1577 | 1065 | 1410 | 897 | 1188 | 7698 |
| 70 × 70 | **SCLI–CMUS** | **1377↓** | **1456↓** | **1084↓** | **1470↓** | **991↓** | **1317↓** | **7695↓** |
| | DTH-MADDPG | 1419 | 1571 | 1139 | 1554 | 1010 | 1399 | 8092 |

The scaling behavior with $N$ is also consistent and informative. All methods exhibit decreasing $\tau_{\text{cov}}$ and $\tau_{\text{det}}$ as the team grows, reflecting the intrinsic parallelism of multi-UAV coverage. However, SCLI–CMUS shows the steepest decline, indicating that additional agents are efficiently utilized rather than inducing interference. In particular, the improvement from $N = 3$ to $N = 7$ is 36% for coverage (1102→710) and 19% for discovery (1232→997), whereas MADDPG improves by 21% and 3% over the same range. The hierarchical baseline DTH–MADDPG narrows the gap relative to flat actor–critic learners, yet it remains consistently behind SCLI–CMUS, suggesting that segment-consistent skill conditioning and the three-parameter saturated reward yield more effective division of labor and faster attainment of operational performance.

To further probe this behaviour, we extended the comparison between SCLI–CMUS and the strongest hierarchical baseline (DTH–MADDPG) from the 50 × 50 workspace to larger search areas of 60 × 60 and 70 × 70. The 50 × 50 case already shows that DTH–MADDPG is the closest competitor in terms of coverage and discovery convergence, so these larger maps provide a more stringent test of scalability. As the search area grows to 60 × 60, the advantage of SCLI–CMUS becomes most pronounced: the total convergence-score gap between the two methods increases to 6797 vs. 7698, i.e., an absolute difference of 901, larger than the corresponding gap on the 50 × 50 grid (6038 vs. 6853). This indicates that on moderately larger workspaces, segment-conditioned intents and belief-based reward shaping yield more efficient spatial partitioning and reduce redundant coverage more effectively than the dual-timescale controller in DTH–MADDPG.

### *4.5 Sensitivity Analysis*

We next examine the sensitivity of SCLI–CMUS to the three reward weights $\lambda_I$, $\lambda_C$, and $\lambda_{\mathcal{E}}$ in (17). Since information gain is the primary driver of target discovery in belief-based search, we fix $\lambda_I = 1$ throughout and vary $\lambda_C$ and $\lambda_{\mathcal{E}}$ around the nominal setting $(\lambda_C, \lambda_{\mathcal{E}}) = (1.0, 0.1)$ used in Section 4.1.

Representative results for $N = 5$ are summarized in Table 3. The coverage weight $\lambda_C$ controls how strongly the policy prioritizes expanding the visited set: a low value ($\lambda_C = 0.5$) yields markedly larger $\tau_{\text{cov}}$ and $\tau_{\text{det}}$ (up to 1150/1450), whereas a high value ($\lambda_C = 1.5$) achieves the fastest coverage (851 steps at $\lambda_{\mathcal{E}} = 0.05$) but consistently slower target discovery ($\tau_{\text{det}} \approx 1230$–$1280$). The energy–time coefficient $\lambda_{\mathcal{E}}$ regulates motion

aggressiveness: with $\lambda_C = 1.0$, the setting $(\lambda_C, \lambda_{\mathcal{E}}) = (1.0, 0.10)$ attains the best overall trade-off, with $\tau_{\text{cov}} = 860$ and the globally minimal $\tau_{\text{det}} = \mathbf{1137}$, while both smaller and larger $\lambda_{\mathcal{E}}$ slightly degrade either coverage or discovery.

**Table 3:** Sensitivity of coverage and discovery convergence times to reward weights for $N = 5$ UAVs

| $\lambda_C$ | $\lambda_{\mathcal{E}}$ | $\tau_{\text{cov}}/\tau_{\text{det}}$ | $\lambda_C$ | $\lambda_{\mathcal{E}}$ | $\tau_{\text{cov}}/\tau_{\text{det}}$ | $\lambda_C$ | $\lambda_{\mathcal{E}}$ | $\tau_{\text{cov}}/\tau_{\text{det}}$ |
|---|---|---|---|---|---|---|---|---|
| 0.5 | 0.05 | 1040/1320 | 1.0 | 0.05 | 869/1180 | 1.5 | 0.05 | **851**↓/1250 |
| | 0.10 | 1075/1360 | | 0.10 | 860/**1137**↓ | | 0.10 | 867/1247 |
| | 0.15 | 1110/1405 | | 0.15 | 878/1210 | | 0.15 | 863/1230 |
| | 0.20 | 1150/1450 | | 0.20 | 875/1260 | | 0.20 | 889/1280 |

The best value in each column is typeset in **bold** and marked with (↓).

### *4.6 Case Study*

This case study provides a qualitative examination of cooperative behaviour under the proposed policy in a $50 \times 50$ workspace with three UAVs and ten fixed targets. Fig. 3 depicts colour–coded trajectories at three representative time stamps. At $t = 300$ (Fig. 3a), the team has already established a clear spatial allocation: trajectories exhibit strong inter-agent separation and limited crossovers. Large uncovered areas are partitioned implicitly, and each UAV conducts frontier-seeking sweeps within its assigned sector. The resulting footprints cover disjoint corridors with small overlap, which accelerates global coverage while preventing early concentration around the same cells.

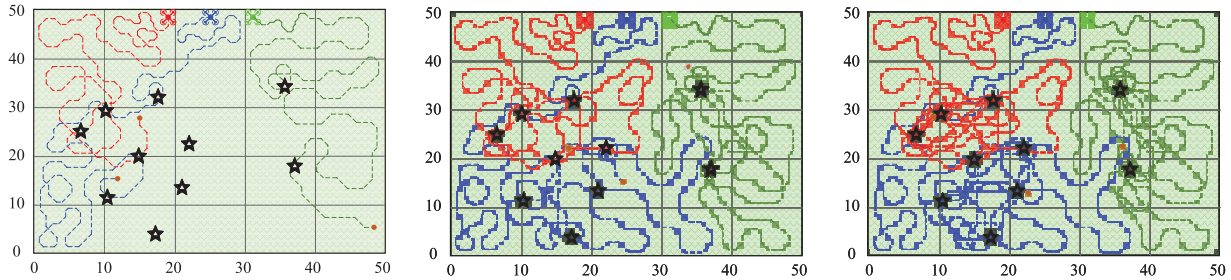


**Figure 3:** Trajectories of three UAVs in a $50 \times 50$ workspace with ten fixed targets at representative time stamps. Dashed paths are colour coded by agent (red, blue, green); black stars mark target locations. Panels: (**a**) $t = 300$, early exploration with clear sector separation; (**b**) $t = 1000$, intensified sampling around informative regions with limited boundary crossings; (**c**) $t = 1500$, steady patrolling within sectors with low redundant coverage

At $t = 1000$ (Fig. 3b), the belief map has concentrated around multiple target locations, and paths become denser in those neighbourhoods. Agents maintain sector integrity while adapting their local loops to repeatedly interrogate high-probability cells. Boundary incursions are rare and occur only where adjacent sectors meet, indicating stable intent selection and limited handover cost. The joint pattern reflects a balanced exploration–exploitation regime: residual unexplored pockets are swept, and detected vicinities receive increased sampling frequency.

At $t = 1500$ (Fig. 3c), the team enters a persistent monitoring phase. Each UAV continues to patrol its sector with short, recurrent loops centred on previously informative regions. The path overlap remains low and the blank areas show no redundant revisits, which is consistent with the energy–time penalization in the reward and the segment-consistent action generation.

## 5 Conclusion

This work has presented a segment-conditioned latent-intent framework for cooperative multi-UAV search that formulates the problem as a discounted MDP on an occupancy grid with a cellwise Bayesian belief update and parameterizes decision making by a single end-to-end policy combining a discrete intent head, updated every $K$ steps, with an intra-segment GRU action head trained under a centralized critic, together with a three-coefficient, scale-calibrated saturated reward balancing information gain, coverage efficiency, and energy–time cost. Across grids of size $50 \times 50$, $60 \times 60$, and $70 \times 70$, the proposed method consistently outperforms strong flat and hierarchical reinforcement-learning baselines: on the $50 \times 50$ workspace, coverage and discovery convergence times are reduced by up to 48% and 40% relative to a flat actor–critic method, and the aggregated convergence metric improves by about 12% compared with a state-of-the-art hierarchical baseline, with the largest total improvement observed on the $60 \times 60$ grid. Future work will extend the framework to adaptive intent durations and heterogeneous platforms, incorporate bandwidth–limited communication and collision–avoidance constraints, model moving targets and three–dimensional kinematics, and pursue field deployment with sim–to–real transfer and formal performance guarantees.

**Acknowledgement:** None.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** Gang Hou, Aifeng Liu, Tao Zhao: Investigation, Data Curation, Writing—Original Draft. Siwen Wei, Wenyuan Wei, Bo Li: Review and Editing, Visualization. Jiancheng Liu, Siwen Wei: Writing—Review and Editing, Supervision. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** Not applicable.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Yanmaz E. Joint or decoupled optimization: multi-UAV path planning for search and rescue. Ad Hoc Netw. 2023;138(3):103018. doi:10.1016/j.adhoc.2022.103018.
2. Liu X, Su Y, Wu Y, Guo Y. Multi-conflict-based optimal algorithm for multi-UAV cooperative path planning. Drones. 2023;7(3):217. doi:10.3390/drones7030217.
3. Chen Z, Zhang T, Hong T. IoT-enhanced multi-base station networks for real-time UAV surveillance and tracking. Drones. 2025;9(8):558. doi:10.3390/drones9080558.
4. Andreou A, Mavromoustakis CX, Markakis E, Bourdena A, Mastorakis G. UAV-asisted IoT network framework with hybrid deep reinforcement and federated learning. Sci Rep. 2025;15(1):37107. doi:10.1038/s41598-025-21014-5.
5. Nguyen DC, Ding M, Pathirana PN, Seneviratne A, Li J, Niyato D, et al. 6G Internet of Things: a comprehensive survey. IEEE Internet Things J. 2021;9(1):359–83. doi:10.1109/jiot.2021.3103320.
6. Qu L, Fan J. Unmanned combat aerial vehicle path planning in complex environment using multi-strategy sparrow search algorithm with double-layer coding. J King Saud Univ—Comput Inf Sci. 2024;36(10):102255. doi:10.1016/j.jksuci.2024.102255.
7. Elmokadem T, Savkin AV. Computationally-efficient distributed algorithms of navigation of teams of autonomous UAVs for 3D coverage and flocking. Drones. 2021;5(4):124. doi:10.3390/drones5040124.
8. Alanezi MA, Bouchekara HR, Apalara TAA, Shahriar MS, Sha'aban YA, Javaid MS, et al. Dynamic target search using multi-UAVs based on motion-encoded genetic algorithm with multiple parents. IEEE Access. 2022;10:77922–39. doi:10.1109/access.2022.3190395.

9. Wang F, Zhu XP, Zhou Z, Tang Y. Deep-reinforcement-learning-based UAV autonomous navigation and collision avoidance in unknown environments. Chin J Aeronaut. 2024;37(3):237–57. doi:10.1016/j.cja.2023.09.033.
10. Chen J, Liu L, Zhang Y, Chen W, Zhang L, Lin Z. Research on UAV path planning based on particle swarm optimization and soft actor-critic. In: Proceedings of the 2024 China Automation Congress (CAC); 2024 Nov 1–3; Qingdao, China. p. 6166–71.
11. Xue D, Lin Y, Wei S, Zhang Z, Qi W, Liu J, et al. Leveraging hierarchical temporal importance sampling and adaptive noise modulation to enhance resilience in multi-agent task execution systems. Neurocomputing. 2025;637(1–2):130134. doi:10.1016/j.neucom.2025.130134.
12. Lee J, Friderikos V. Interference-aware path planning optimization for multiple UAVs in beyond 5G networks. J Commun Netw. 2022;24(2):125–38. doi:10.23919/jcn.2022.000006.
13. Wang K, Gou Y, Xue D, Liu J, Qi W, Hou G, et al. Resilience augmentation in unmanned weapon systems via multi-layer attention graph convolutional neural networks. Comput Mater Contin. 2024;80(2):2941–62. doi:10.32604/cmc.2024.052893.
14. Wang K, Xue D, Gou Y, Qi W, Li B, Liu J, et al. Meta-path-guided causal inference for hierarchical feature alignment and policy optimization in enhancing resilience of UWSoS. J Supercomput. 2025;81(2):358. doi:10.1007/s11227-024-06848-6.
15. Wang N, Li Z, Liang X, Li Y, Zhao F. Cooperative target search of UAV swarm with communication distance constraint. Math Probl Eng. 2021;2021(1):3794329. doi:10.1155/2021/3794329.
16. Gou Y, Wei S, Xu K, Liu J, Li K, Li B, et al. Hierarchical reinforcement learning with kill chain-informed multi-objective optimization to enhance resilience in autonomous unmanned swarm. Neural Netw. 2025;195(2):108255. doi:10.1016/j.neunet.2025.108255.
17. Huh D, Mohapatra P. Multi-agent reinforcement learning: a comprehensive survey. arXiv:2312.10256. 2023.
18. Vezhnevets AS, Osindero S, Schaul T, Heess N, Jaderberg M, Silver D, et al. Feudal networks for hierarchical reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia, pp. 3540–9.
19. Ahilan S, Dayan P. Feudal multi-agent hierarchies for cooperative reinforcement learning. arXiv:1901.08492. 2019.
20. Tang H, Hao J, Lv T, Chen Y, Zhang Z, Jia H, et al. Hierarchical deep multiagent reinforcement learning with temporal abstraction. arXiv:1809.09332. 2018.
21. Lyu M, Zhao Y, Huang C, Huang H. Unmanned aerial vehicles for search and rescue: a survey. Remote Sens. 2023;15(13):3266. doi:10.3390/rs15133266.
22. Rahman M, Sarkar NI, Lutui R. A survey on multi-UAV path planning: classification, algorithms, open research problems, and future directions. Drones. 2025;9(4):263. doi:10.3390/drones9040263.
23. Zhao L, Gao Z, Hawbani A, Zhao W, Mao C, Lin N. Fuzzy-MADDPG based multi-UAV cooperative search in network-limited environments. In: Proceedings of the 2024 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM); 2024 Nov 19–21; Setif, Algeria, p. 1–7.
24. Kelner JM, Burzynski W, Stecz W. Modeling UAV swarm flight trajectories using rapidly-exploring random tree algorithm. J King Saud Univ—Comput Inf Sci. 2024;36(1):101909. doi:10.1016/j.jksuci.2023.101909.
25. Zhang X, Ali M. A bean optimization-based cooperation method for target searching by swarm UAVs in unknown environments. IEEE Access. 2020;8:43850–62. doi:10.1109/access.2020.2977499.
26. Chaves AN, Cugnasca PS, Jose J. Adaptive search control applied to search and rescue operations using unmanned aerial vehicles (UAVs). IEEE Latin Am Trans. 2014;12(7):1278–83. doi:10.1109/tla.2014.6948863.
27. Qamar RA, Sarfraz M, Rahman A, Ghauri SA. Multi-criterion multi-UAV task allocation under dynamic conditions. J King Saud Univ—Comput Inf Sci. 2023;35(9):101734. doi:10.1016/j.jksuci.2023.101734.
28. Hou Y, Zhao J, Zhang R, Cheng X, Yang L. UAV swarm cooperative target search: a multi-agent reinforcement learning approach. IEEE Trans Intell Veh. 2023;9(1):568–78. doi:10.1109/tiv.2023.3316196.
29. Hou K, Yang Y, Yang X, Lai J. Distributed cooperative search algorithm with task assignment and receding horizon predictive control for multiple unmanned aerial vehicles. IEEE Access. 2021;9:6122–36. doi:10.1109/access.2020.3048974.

30. Phung MD, Ha QP. Safety-enhanced UAV path planning with spherical vector-based particle swarm optimization. Appl Soft Comput. 2021;107(2):107376. doi:10.1016/j.asoc.2021.107376.
31. Tang J, Liang Y, Li K. Dynamic scene path planning of uavs based on deep reinforcement learning. Drones. 2024;8(2):60. doi:10.3390/drones8020060.
32. Sabzekar S, Samadzad M, Mehditabrizi A, Tak AN. A deep reinforcement learning approach for UAV path planning incorporating vehicle dynamics with acceleration control. Unmanned Syst. 2024;12(03):477–98. doi:10.1142/s2301385024420044.
33. Bacon PL, Harb J, Precup D. The option-critic architecture. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; 2017 Feb 4–9; San Francisco, CA, USA. p. 1726–34.
34. Liu J, Wei S, Li B, Wang T, Qi W, Han X, et al. Dual-timescale hierarchical MADDPG for Multi-UAV cooperative search. J King Saud Univ Comput Inf Sci. 2025;37(6):1–17. doi:10.1007/s44443-025-00156-6.
35. Harikumar K, Senthilnath J, Sundaram S. Multi-UAV oxyrrhis marina-inspired search and dynamic formation control for forest firefighting. IEEE Trans Autom Sci Eng. 2018;16(2):863–73. doi:10.1109/tase.2018.2867614.
36. Perez-Carabaza S, Besada-Portas E, Lopez-Orozco JA, de la Cruz JM. Ant colony optimization for multi-UAV minimum time search in uncertain domains. Appl Soft Comput. 2018;62(4):789–806. doi:10.1016/j.asoc.2017.09.009.