electronics

MDPI

Article

AoI- and DS-Enhanced Cooperative Search for Multi-UAV Systems Under Spatially Structured Communication Constraints

Lingtao Xue [1,2], Xuewen Dong [1], Xinyu Hu [2,3], Lingxiao Yang [4] and Gang Xiao [2,*]

[1] School of Computer Science & Technology, Xidian University, Xi'an 710071, China; lingtaoxue@stu.xidian.edu.cn (L.X.); xwdong@xidian.edu.cn (X.D.)
[2] National Key Laboratory for Complex Systems Simulation, Haidian District, Beijing 100080, China; huxy53@mail2.sysu.edu.cn
[3] School of System Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China
[4] School of Artificial Intelligence and Computer Science, Shaanxi Normal University, Xi'an 710062, China; lxyang@snnu.edu.cn
* Correspondence: searchware@foxmail.com

Abstract

Multi-UAV cooperative search is important for applications such as target reconnaissance, environmental monitoring, and emergency response. In practice, communication is often spatially heterogeneous due to terrain occlusion and environmental interference, which may delay information sharing and weaken coordination efficiency when UAVs traverse communication-blocked areas. To address this issue, we propose an Age of Information (AoI)- and Dempster–Shafer (DS)-enhanced cooperative search framework for multi-UAV systems under spatially structured communication constraints. Specifically, a DS belief map is introduced to fuse uncertain observations, while AoI is used to characterize the freshness of delayed information. An AoI-aware update mechanism further integrates buffered observations into the global belief map after communication recovery. The search process is then formulated as a communication-aware multi-agent sequential decision-making problem and solved using reinforcement learning. To demonstrate the generality of the proposed framework, we instantiate it with Proximal Policy Optimization (PPO), Multi-Agent Proximal Policy Optimization (MAPPO), and Q-value Mixing Network (QMIX). Experimental results show that the proposed framework consistently outperforms the baseline methods under heterogeneous environments and different communication conditions. Among all variants, AoI-DS-MAPPO achieves the best overall performance, improving average reward, success rate, and the number of detected targets by 26.13%, 24.32%, and 3.65%, respectively, while reducing episode length by 31.96% relative to the strongest baseline.

Keywords: multi-UAV cooperative search; Age of Information; Dempster–Shafer evidence theory; reinforcement learning

Check for updates

Academic Editor: Reza K. Amineh

Received: 8 April 2026
Revised: 23 April 2026
Accepted: 24 April 2026
Published: 29 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Driven by the rapid evolution of UAV technology, coordinated multi-UAV systems are being adopted in an expanding range of applications, including target exploration, environmental sensing, and emergency rescue [1–4]. In practical deployments, UAVs often face challenges including unstable communication links, local communication outages, and delayed information feedback during search missions, which lead to the accumulation of uncertainty in global situational awareness and the degradation of information timeliness [5,6]. Meanwhile, differences in UAV positions, sensing ranges, and task progress

result in pronounced spatio-temporal heterogeneity in the distribution of information within the swarm, further increasing the complexity of collaborative decision-making [7]. How to achieve efficient information acquisition and cooperative decision-making under communication constraints and dynamically changing environments has thus become a key challenge for the intelligent development of UAV swarms [8–11], as illustrated in Figure 1.

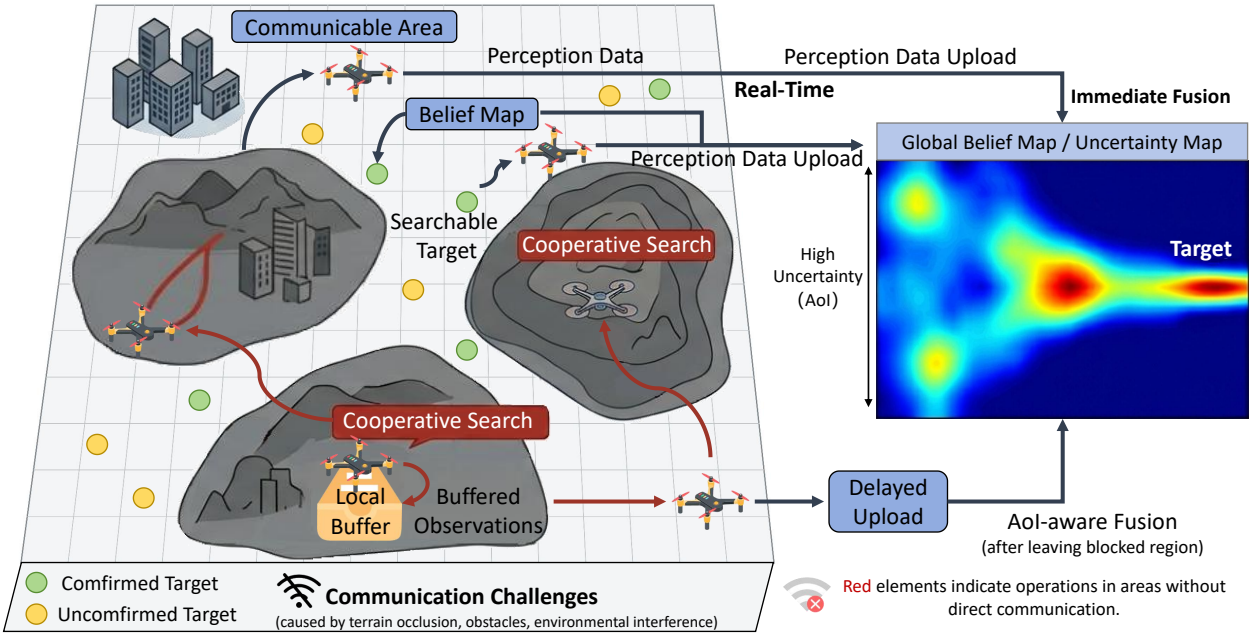


Figure 1. Illustration of the studied cooperative target search scenario for multiple UAVs, in which communication capability is environment-dependent and stable blocked zones may result in delayed data delivery.

Motivation. The motivation of this work is threefold. First, from a modeling perspective, while some existing studies have considered structured communication interruptions, the explicit coupling of location-dependent communication blackouts with AoI-weighted delayed belief synchronization and uncertainty-aware evidence fusion within a unified MARL framework remains insufficiently explored. Second, in practical UAV swarm operations, communication disruptions are often induced by the physical environment, such as terrain blockage, dense urban structures, interference-prone regions, or electromagnetic blind spots, rather than arising purely at random. As a result, agents may experience location-dependent information delays and uneven information freshness. Third, from a decision-making perspective, such delayed and asymmetric information propagation can lead to lagged global belief updates, uncertainty accumulation, distorted situational awareness, and reduced coordination efficiency. These issues motivate the development of a communication-aware cooperative framework that explicitly captures region-dependent communication constraints and their influence on shared information freshness [12,13].

Challenges. Explicitly modeling spatially structured communication constraints introduces several fundamental challenges to cooperative UAV decision-making. (1) Heterogeneous and delayed information states: Spatial variation in communication connectivity causes agents to maintain information with different update times and validity levels, making individual decision-making and coordination intrinsically non-Markovian. (2) Strong coupling between communication and exploration: Communication-induced information delays are tightly coupled with exploration dynamics, since effective sensing depends not only on area coverage but also on timely information uploading and fusion. This creates a feedback loop between mobility and communication accessibility. (3) High-dimensional and history-dependent decision-making: Persistent communication blackout regions cause

spatially uneven information aging and structured uncertainty accumulation, which cannot be captured by models assuming independent packet loss or delay. As a result, the resulting optimization problem becomes high-dimensional, history-dependent, and strongly coupled across agents, jointly involving mobility control, information freshness, and communication availability. These challenges call for adaptive, communication-aware learning mechanisms that can dynamically adjust cooperative strategies under delayed and heterogeneous information states.

Contributions. To cope with the challenges described above, this work makes the following contributions:

- **Environment-Aware Communication Formulation:** We design a communication formulation that incorporates spatial structure and explicitly describes environment-dependent communication access. Unlike conventional approaches that treat packet loss or delay as independent stochastic events, our model captures communication blockage as a persistent spatial structure tightly coupled with UAV mobility. This coupling introduces fundamentally different decision-making challenges, including non-Markovian information states and structured uncertainty accumulation, which cannot be addressed by existing random-delay or distance-attenuation models.
- **Communication-Aware Cooperative Decision Framework:** We formulate cooperative UAV decision-making under spatial communication constraints as a joint mobility–information–communication optimization problem, explicitly accounting for information freshness and heterogeneous communication accessibility across agents.
- **Adaptive MARL-Based Coordination Mechanism with Non-Trivial AoI–DS Coupling:** We develop a communication-aware multi-agent reinforcement learning (MARL) framework that integrates spatial communication constraints, information freshness (AoI), and belief evolution into the decision-making process. The proposed mechanism enables agents to adaptively balance exploration, information uploading, and coordination under location-dependent communication availability and delayed information states. Crucially, rather than treating AoI merely as an auxiliary input feature, we embed information freshness directly into the DS evidence combination process via a freshness-weighted synchronization mechanism. Specifically, stale buffered observations are down-weighted exponentially prior to belief map integration via the decay factor, principally coupling information timeliness with uncertainty quantification. To our knowledge, this AoI-weighted evidential fusion design has not been previously proposed in the context of multi-UAV cooperative search.
- **Comprehensive Evaluation:** Extensive experiments including baseline comparisons under heterogeneous communication conditions, parameter sensitivity analysis, convergence analysis, and ablation studies demonstrate the effectiveness, robustness, and generalizability of the proposed framework across different reinforcement learning paradigms.

2. Related Work

Existing work [14–26] mainly falls into three research categories for multi-UAV cooperative search according to different communication assumptions: (i) Global communication-based methods. This line of work assumes that UAVs can access globally shared information through centralized coordination or reliable network-wide communication. The goal is to improve search efficiency and coordination quality by maintaining synchronized situational representations, such as global maps, shared belief states, or centrally aggregated task information. (ii) Local communication-based methods. These methods restrict information exchange to neighboring agents within limited communication ranges or predefined topologies. They aim to support decentralized cooperation through local interaction, dis-

tributed information propagation, and topology-constrained coordination. (iii) Imperfect communication-aware methods. This category considers communication impairments such as packet loss, asynchronous transmission, or stochastic delay, and attempts to improve the robustness of cooperative search under unreliable communication conditions. We summarize representative approaches from these three categories and provide a detailed comparison in Table 1.

Table 1. Comparative summary of multi-UAV cooperative search approaches.

Reference	Comm. Scope	Spatial Communication	CDM	AoI	BM	MC	Coordination Strategy
Hollinger [14]	Global	×	×	×	✓	×	Opt
Wang [15]	Global	×	×	×	×	×	MARL
Mao [16]	Global	×	×	×	×	×	Heuristic
Abdulghafoor [17]	Local	×	×	×	×	×	Opt
Kouzeghar [18]	Local	×	×	×	×	×	MARL
Ajina [19]	Local	×	Random	×	×	×	Opt
Hou [20]	Local	×	×	×	✓	×	MARL
Fei [21]	Local	×	×	×	×	✓	Heuristic
Zhang [22]	Local	×	×	×	×	✓	Opt
Sun [23]	Global	✓	Str	×	✓	✓	Heuristic
Senthilnath [24]	None	×	×	×	×	×	Heuristic
Zhang [25]	Local	×	Random	×	×	✓	Heuristic
Ours	Global	✓	Str	✓	✓	✓	MARL

Note: ✓indicates that the corresponding feature is included, whereas × indicates that it is not included. CDM denotes communication-delay characterization; Str refers to spatially persistent delay patterns; BM indicates belief-map fusion; MC represents mobility–communication interaction; Opt stands for optimization-based methods; MARL stands for multi-agent reinforcement learning.

Global communication-based methods: Hollinger et al. studied informative trajectory planning from an optimization perspective, using a globally shared belief representation to guide information gathering under budget constraints [14]. For persistent visual coverage, Wang et al. formulated cooperative motion planning as a multi-agent reinforcement learning problem, in which agents optimize their trajectories with access to global task information [15]. Mao et al. designed a heuristic hierarchical collaboration framework for dynamic environments with complex constraints, where global spatio-temporal guidance helps improve cooperative planning efficiency [16].

Local communication-based methods: Abdulghafoor et al. developed a two-level coverage control framework in which dynamic multi-agent coverage is achieved by coupling high-level distribution control with low-level deployment optimization [17]. Kouzeghar et al. formulated multi-target pursuit in complex environments as a decentralized MARL problem and adopted a role-based MADDPG architecture to enhance both tracking and exploration performance [18]. Focusing on large-scale unknown environments, Hou et al. introduced a distributed cooperative target search approach based on MARL and convolutional neural networks to improve search efficiency [20]. Fei et al. addressed cooperative search under uncertain conditions by constructing an autonomous search model supported by a heuristic optimization framework with limited communication [21]. Unlike methods requiring richer coordination signals, Zhang et al. considered cooperative multi-UAV target tracking with only local information exchange and achieved improved convergence and collision avoidance through a distributed time-varying optimization scheme [22].

Imperfect communication-aware methods: Ajina et al. addressed multi-agent coverage control under local communication constraints by developing an asynchronous

distributed coordination scheme with event-triggered broadcasting to reduce the effect of random communication delays [19]. For multi-UAV cooperative search with unreliable communication links, Zhang et al. combined consensus-based information updates with ant colony optimization in a distributed search framework to preserve search performance under random communication disruptions [25]. Targeting communication-denied target search, Sun et al. developed a scheduling framework for heterogeneous UAV swarms that integrates slot-based communication modeling, position prediction, and heuristic region co-evolution optimization to improve search effectiveness under structured communication interruptions [23].

Comparison: In contrast to the above methods, our approach explicitly models communication blockage as a persistent spatial structure that is tightly coupled with UAV mobility rather than treating it as a random disturbance. This distinction introduces fundamentally different decision-making challenges: agents experience non-Markovian information states and structured uncertainty accumulation that existing approaches are not designed to handle. Furthermore, unlike methods that use AoI merely as an input signal, we embed information freshness directly into the DS evidence combination process via freshness-weighted synchronization, achieving a principled coupling between timeliness and uncertainty quantification. These design choices collectively address the limitations identified above and enable more effective coordination under spatially structured communication constraints.

3. System Model and Problem Formulation

This paper addresses a scenario in which multiple unmanned aerial vehicles (UAVs) operate cooperatively within an area containing several signal-blocked zones, with the objective of locating and verifying a prescribed number of targets as rapidly as possible. The key mathematical symbols and variables used throughout this paper are summarized in Table 2 for reference.

Table 2. Summary of main notation.

Symbol	Description	Symbol	Description
Ω	Operational domain	L	Side length of the area
K	Number of UAVs	H	Number of targets
D	Targets threshold	T_f	Max. steps per episode
$\Phi(u,v)$	Binary blockage map	$\mathbf{r}_t^k$	Position of UAV k at time t
$\mathbf{s}^h$	Fixed location of target h	ρ	Sensor detection radius
δ	Sensing range distance	θ_t^k	Freshness index of UAV k
θ_f	Freshness normalization constant	$\mathcal{B}_t^k$	Onboard evidence buffer
t_s	Time stamp of buffered evidence	$\psi(t,t_s)$	Exponential decay factor
η	Decay rate of freshness weight	Ψ	Overall optimization objective
ζ	Reward discount factor	c_t	Newly verified targets at step t
ω_c	Verification bonus weight	r_t^e	Evidence-gain reward
r_t^θ	Freshness-regularization penalty	ϖ_ω^k	Policy of UAV k (params. ω)
V_ψ	Value function (params. ψ)	$\tilde{Q}_k$	Individual action-utility (QMIX)
χ	QMIX utility network params.	Q_{jt}	Global team action-value (QMIX)
Q_{jt}^-	Target network (QMIX)	$\hat{G}_t$	Bootstrapped return estimate
$\hat{q}_t$	TD regression target (QMIX)	$\hat{A}_t$	Generalized advantage estimate
λ	GAE smoothing parameter	Γ_t^k	Importance-sampling ratio
ϵ_{clip}	PPO clipping coefficient	$\varkappa$	Entropy bonus coefficient

3.1. System Model

3.1.1. Operational Area and Target Description

We formulate the operational area as a discrete two-dimensional domain $\Omega = \{0, \ldots, L-1\}^2 \subset \mathbb{Z}^2$ of size $L \times L$. A fleet of K UAVs, indexed by $\mathcal{K} = \{1, \ldots, K\}$ with $|\mathcal{K}| = K$, is tasked with locating H targets whose positions are not known a priori. The mission succeeds once at least D targets are verified within a finite time horizon T_f. To capture spatial communication constraints, signal-blocked zones are characterized by a binary blockage map $\Phi : \Omega \to \{0, 1\}$. For any cell $(u, v) \in \Omega$, $\Phi(u, v) = 1$ indicates that the cell falls within a signal-blocked zone, while $\Phi(u, v) = 0$ denotes a communication-accessible cell. Each blocked zone is constructed as a connected neighborhood around a randomly sampled center point with a given radius defined under the Manhattan metric. The map Φ remains static throughout an episode and is accessible to all UAVs.

We denote the spatial position of UAV k at discrete time step t by $\mathbf{r}_t^k = (u_t^k, v_t^k) \in \Omega$, and the fixed location of target h by $\mathbf{s}^h \in \Omega$. Each target carries an active flag $f_t^h \in \{0, 1\}$, where $f_t^h = 1$ means target h has not yet been verified and $f_t^h = 0$ means it has been confirmed and excluded from subsequent processing. The episode ends once the total number of verified targets reaches D, or once the time index surpasses T_f.

3.1.2. Detection and Observation Formulation

Every UAV is equipped with an onboard sensor whose effective detection radius is ρ. At time step t, UAV k positioned at $\mathbf{r}_t^k = (u_t^k, v_t^k)$ is able to collect measurements only within its coverage zone, defined via the ℓ_1-ball:

$$\mathcal{F}_t^k \triangleq \Big\{ (u, v) \in \Omega \;\big|\; \|(u, v) - \mathbf{r}_t^k\|_1 \le \rho \Big\}. \tag{1}$$

For any cell $(u, v) \in \mathcal{F}_t^k$, let $g_t(u, v) \in \{0, 1\}$ denote the ground-truth indicator of target presence at that cell. The raw sensor reading is modeled as a binary variable $y_t^k(u, v) \in \{0, 1\}$ governed by range-dependent hit and spurious-alarm rates.

Specifically, for a cell (u, v) at range $\delta = \|(u, v) - \mathbf{r}_t^k\|_1$, the probabilities are defined as

$$\alpha(\delta) = \alpha_0 \exp\!\left(-\frac{\delta}{\rho}\right), \quad \beta(\delta) = \beta_0 \exp\!\left(-\frac{\delta}{\rho}\right), \tag{2}$$

where $\alpha_0 \in (0, 1]$ and $\beta_0 \in [0, 1)$ are the peak hit rate and peak spurious-alarm rate at $\delta = 0$, respectively. The measurement is then drawn as

$$y_t^k(u, v) \sim \begin{cases} \text{Bernoulli}\big(\alpha(\delta)\big), & g_t(u, v) = 1, \\ \text{Bernoulli}\big(\beta(\delta)\big), & g_t(u, v) = 0, \end{cases} \tag{3}$$

and $y_t^k(u, v) = 0$ is enforced for $(u, v) \notin \mathcal{F}_t^k$.

To support learning-based control, a multi-channel input tensor is constructed for each UAV. The observation of agent k at time t is denoted by $\mathbf{x}_t^k \in \mathbb{R}^{6 \times L \times L}$, comprising: (i) the global evidential belief channels $\{\pi_1, \pi_0, \pi_\phi\}$, (ii) the blockage map Φ, (iii) a one-hot location map encoding $\mathbf{r}_t^k$, and (iv) a freshness channel whose value is propagated to all cells and normalized as $\min\{1, \theta_t^k / \theta_f\}$ with a constant θ_f.

3.1.3. Evidence Accumulation and Belief Update

To characterize spatial uncertainty and aggregate measurements from multiple sources, we maintain a global evidential belief map over the domain based on Dempster–Shafer (DS) theory. For each cell $(u, v) \in \Omega$, the DS state is described by three evidential weights

$$\pi_1(u,v) \in [0,1], \quad \pi_0(u,v) \in [0,1], \quad \pi_\phi(u,v) \quad = 1 - \pi_1(u,v) - \pi_0(u,v). \tag{4}$$

where π_1, π_0, and π_ϕ represent the committed mass toward the hypotheses *target present*, *target absent*, and *undetermined*, respectively.

Local Measurement Evidence

At each time step t, UAV k at location $\mathbf{r}_t^k$ generates local evidence within its coverage zone. We denote the evidence at cell (u, v) by

$$\varepsilon_1^{k,t}(u,v),\ \varepsilon_0^{k,t}(u,v),\ \varepsilon_\phi^{k,t}(u,v) = 1 - \varepsilon_1^{k,t}(u,v) - \varepsilon_0^{k,t}(u,v), \tag{5}$$

where $\varepsilon_\phi^{k,t}$ captures uncommitted mass arising from measurement noise.

DS Update in Accessible Areas

When UAV k occupies a communication-accessible cell, i.e., $\Phi(\mathbf{r}_t^k) = 0$, its local evidence is immediately integrated into the global belief map through the DS combination operator. Let (π_1, π_0) denote the current global evidential weights and $(\varepsilon_1, \varepsilon_0)$ the incoming evidence at a given cell (subscripts omitted for brevity). The inter-source conflict factor is

$$\kappa = \pi_1\varepsilon_0 + \pi_0\varepsilon_1, \tag{6}$$

and the fused evidential weights are

$$\pi_1^+ = \frac{\pi_1\varepsilon_1 + \pi_1\varepsilon_\phi + \pi_\phi\varepsilon_1}{1-\kappa}, \quad \pi_0^+ \quad = \frac{\pi_0\varepsilon_0 + \pi_0\varepsilon_\phi + \pi_\phi\varepsilon_0}{1-\kappa}, \tag{7}$$

with $\pi_\phi^+ = 1 - \pi_1^+ - \pi_0^+$. This update is applied to every cell within the agent's coverage zone.

Remark 1. *The DS combination operator $\oplus$ is both commutative ($m_1 \oplus m_2 = m_2 \oplus m_1$) and associative ($(m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3)$), as established in [27,28]. These properties guarantee that the order in which buffered observations are fused upon zone exit does not affect the resulting belief state, ensuring consistency of updates under asynchronous evidence arrival. The freshness-weighted scaling $\psi(t, t_s)$ modifies the magnitude of evidence but preserves this algebraic structure, so commutativity and associativity are inherited by the weighted combination.*

Blocked Zones, Information Freshness, and Onboard Buffering

When UAV k enters a signal-blocked cell, i.e., $\Phi(\mathbf{r}_t^k) = 1$, it is unable to upload new measurements to the shared map. Instead, local evidence is stored in a private onboard buffer $\mathcal{B}_t^k$ together with the corresponding time stamp. Meanwhile, the information freshness index θ_t^k increments to reflect the growing staleness of unshared observations:

$$\theta_{t+1}^k = \begin{cases} \theta_t^k + 1, & \Phi(\mathbf{r}_t^k) = 1, \\ 0, & \Phi(\mathbf{r}_t^k) = 0. \end{cases} \tag{8}$$

When the UAV exits the blocked zone (i.e., $\Phi(\mathbf{r}_t^k) = 0$ and $\Phi(\mathbf{r}_{t-1}^k) = 1$), a *synchronization* operation is triggered, uploading all buffered evidence into the shared belief map.

Freshness-Weighted Synchronization with Temporal Discounting

Consider a buffered record at cell (u, v) stored at time t_s with evidential weights $(\hat{\pi}_1, \hat{\pi}_0)$. At synchronization time t, stale evidence is down-weighted via an exponential decay factor

$$\psi(t, t_s) = \exp\big(-\eta\,(t - t_s)\big), \tag{9}$$

and the scaled evidence $(\psi(t, t_s)\hat{\pi}_1,\ \psi(t, t_s)\hat{\pi}_0)$ is uploaded. The exponential form of $\psi(t, t_s)$ is not merely a heuristic choice but can be theoretically motivated from two complementary perspectives. From an information-theoretic standpoint, if the information aging process is modeled as a memoryless Markov process—meaning the rate of information loss per unit time is constant and independent of the elapsed time—then the exponential function is the *unique* decay form consistent with this assumption, following directly from the characterization of the exponential distribution as the only continuous memoryless distribution, in analogy with the survival function of a continuous-time Markov chain. From a Bayesian perspective, $\psi(t, t_s)$ can be interpreted as a time-dependent likelihood discount applied to delayed observations within the DS combination framework, which is formally equivalent to the temporal discounting factor in Bayesian forgetting models [29,30], where the reliability of buffered evidence is treated as decaying exponentially with its age. This design choice is further supported by prior AoI literature, in which exponential penalty functions have been adopted to characterize the monotonically decreasing value of aged information [12,13].

Remark 2. *The freshness-weighted synchronization mechanism establishes a theoretically traceable, end-to-end causal pathway from AoI to the policy gradient. Specifically, the exponential decay factor $\psi(t, t_s)$ directly modulates the magnitude of buffered evidence prior to DS combination. Since the DS entropy reduction upon zone exit—captured by the evidence-gain reward r_t^e—is a monotonically increasing function of ψ, staler evidence (larger θ_t^k) yields a smaller entropy reduction and hence a lower evidence-gain reward. This creates the following end-to-end causal chain: a higher freshness index leads to a smaller decay weight, which reduces the DS entropy gain upon synchronization, lowers the evidence-gain reward, decreases the estimated advantage for remaining in the blocked zone, and ultimately drives the policy toward earlier zone-exit behaviors. This three-way coupling among information freshness, uncertainty quantification, and policy optimization is the principal theoretical mechanism distinguishing the proposed framework from approaches that treat AoI merely as an auxiliary input feature.*

The weighted evidence is subsequently merged into the global belief map using the DS combination operator described above. Moreover, to model long-term uncertainty growth, a global evidence decay is applied after every step:

$$\pi_1(u, v) \leftarrow \gamma\,\pi_1(u, v), \quad \pi_0(u, v) \leftarrow \gamma\,\pi_0(u, v), \quad \pi_\phi(u, v) \leftarrow 1 - \pi_1(u, v) - \pi_0(u, v), \tag{10}$$

where $\gamma \in (0, 1]$ is an evidence-retention factor that controls the rate of belief decay. Note that γ is distinct from the RL discount factor $\zeta \in (0, 1]$ defined in (13): γ governs long-term uncertainty growth in the belief map, whereas ζ discounts future rewards in the optimization objective.

Figure 2 provides a qualitative illustration of the measurement collection and belief update process at three representative time steps. The scene row depicts the blocked zones, active targets, verified targets, and UAV positions; the belief row visualizes the target-presence mass π_1; and the undetermined-mass row visualizes π_ϕ. As the UAVs explore the domain and progressively synchronize local measurements, the global DS belief map concentrates toward likely target locations while the undetermined mass decreases

in well-covered areas. Once a target is verified with high confidence, it is withdrawn from the active set and marked accordingly. This illustration highlights the interplay among onboard sensing, DS-based evidence integration, and belief evolution within the proposed framework.

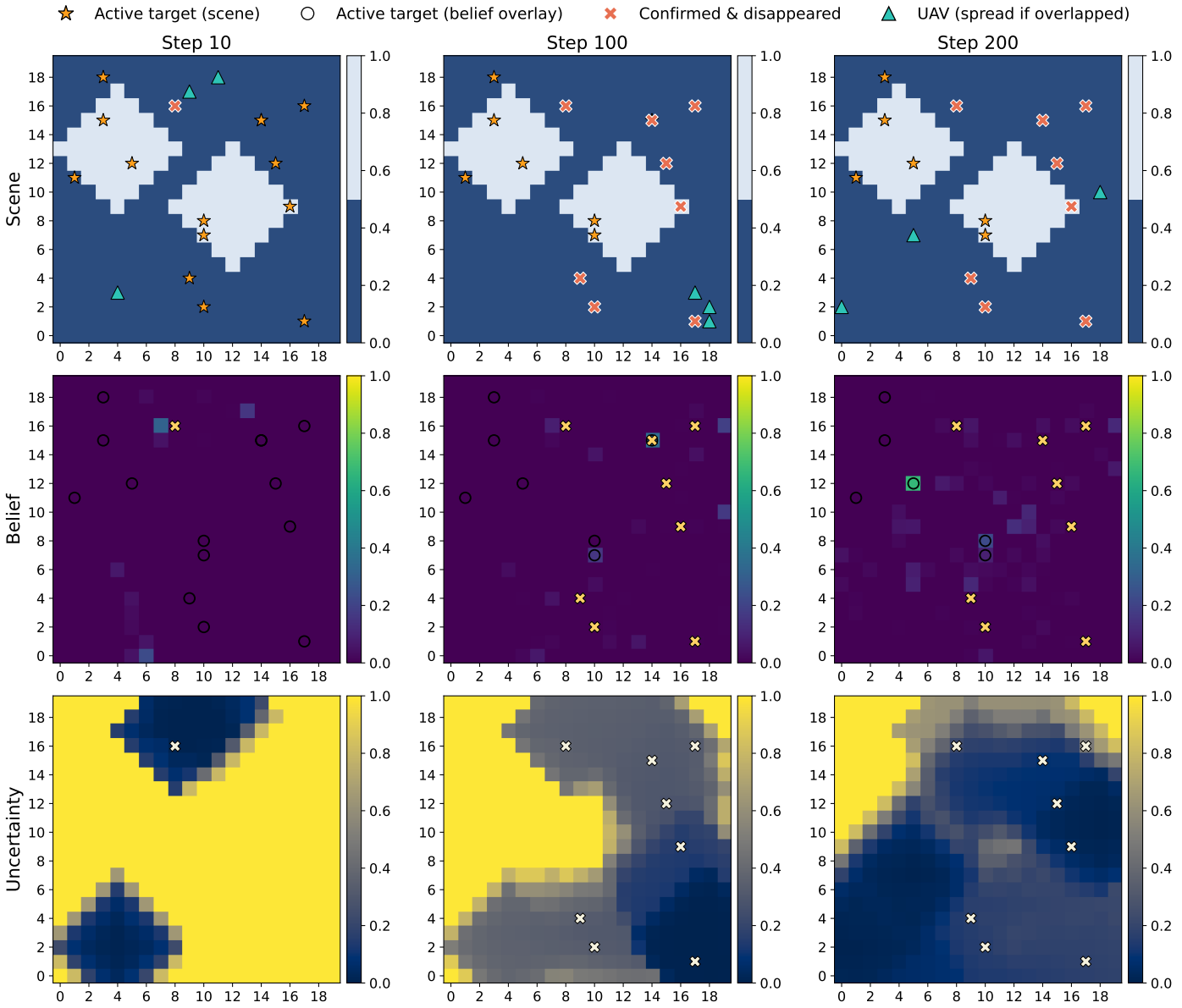


Figure 2. Visualization of the sensing and information update process at different time steps. The first row shows the scene map, including blocked regions, active targets, confirmed-and-disappeared targets, and UAV positions; the second row shows the DS belief map π_1; and the third row shows the undetermined mass map π_ϕ. From left to right, the three columns correspond to Step 10, Step 100, and Step 200, respectively. The figure shows that, as UAVs continue exploring and uploading information, the target belief distribution gradually becomes more concentrated, the undetermined mass in repeatedly observed regions decreases over time, and some targets are successfully confirmed and removed from the active target set.

3.1.4. Target Verification and Removal Procedure

A target is declared *verified* only when the global DS belief at its cell attains a sufficiently high confidence level. Let $\mathbf{s}^h = (u^h, v^h) \in \Omega$ denote the (unknown) location of target h, and let $f_t^h \in \{0,1\}$ be its active flag at time t ($f_t^h = 1$ for unverified and $f_t^h = 0$ for verified/removed targets).

Verification Precondition

Owing to signal-blocked zones, a UAV is not permitted to declare target verification while occupying a blocked cell. Specifically, UAV k at time t may initiate verification only when $\Phi(\mathbf{r}_t^k) = 0$ (i.e., it resides in a communication-accessible cell). This constraint ensures that all verifications are grounded in the up-to-date shared belief map, including any freshness-weighted synchronization updates following zone exit.

Confidence-Based Thresholding

When UAV k satisfies the verification precondition and a target cell lies within its coverage zone, the target is verified if the DS evidential weights satisfy

$$\pi_1(u^h, v^h) \geq \sigma_1, \qquad \pi_\phi(u^h, v^h) \leq \sigma_\phi, \tag{11}$$

where $\sigma_1 \in (0,1)$ and $\sigma_\phi \in (0,1)$ denote the presence-belief threshold and the undetermined-mass ceiling, respectively. This criterion demands both a strong commitment to target existence and a sufficiently small residual undetermined mass.

The thresholds σ_1 and σ_ϕ have a clear physical interpretation that justifies their use as fixed constants across all experiments. The DS belief mass π_1 at a given cell is jointly determined by both the number of times that cell falls within a UAV's sensor coverage zone and the frequency of those visits: a higher visitation count contributes more accumulated evidence, while more frequent and recent visits are less affected by the global belief decay, resulting in higher retained mass. A higher threshold σ_1 therefore corresponds to requiring sufficient cumulative evidence from adequately frequent observations before verification is triggered, while a lower undetermined-mass ceiling σ_ϕ demands that residual uncertainty be sufficiently reduced. In this sense, the thresholds jointly encode a minimum requirement on both observational coverage and visitation recency rather than constituting an arbitrary numerical choice. Since the sensor model parameters, belief decay factor, grid size, and task structure are held constant across all experimental conditions, the fixed values $\sigma_1 = 0.80$ and $\sigma_\phi = 0.20$ correspond to a consistent and comparable verification standard across all experiments.

Removal and Map Reset

Upon verifying target h, we set $f_t^h \leftarrow 0$ and exclude it from further search and reward evaluation. To prevent duplicate verifications and to flag the cell as cleared, the belief map at (u^h, v^h) is reset by assigning

$$\pi_1(u^h, v^h) \leftarrow 0, \quad \pi_0(u^h, v^h) \leftarrow 1, \quad \pi_\phi(u^h, v^h) \leftarrow 0. \tag{12}$$

The episode terminates when the cumulative number of verified targets reaches the mission target D, or when the time horizon T_f is exhausted.

3.2. Problem Formulation

Based on the foregoing system description, we cast cooperative multi-UAV target search under signal-blocked zones as an optimization problem. Mission success requires verifying at least D targets within the horizon T_f. The design goal is to achieve fast and reliable target verification, encourage informative spatial coverage, and counteract information staleness induced by communication blockage.

3.2.1. Optimization Objective

The mission objective is scalarized as a weighted combination of verification reward, evidence gain, freshness regularization, and step penalty. Let $\mathbf{a}_t = (a_t^1, \ldots, a_t^K) \in \mathcal{A}^K$ denote the joint action of all UAVs at time step t. Specifically, we seek to minimize

$$\min_{\{\mathbf{a}_t\}_{t=0}^{T_f-1}} \Psi \triangleq -\sum_{t=0}^{T_f-1} \zeta^t \Big(\omega_c \, c_t + \mu_e \, r_t^e - \mu_\theta \, r_t^\theta - \mu_s \Big). \tag{13}$$

where $\zeta \in (0, 1]$ is a discount factor, c_t denotes the number of newly verified targets at step t, $\omega_c > 0$ is the verification bonus, and $\mu_e, \mu_\theta, \mu_s \geq 0$ are scalar weighting coefficients.

The evidence-gain term is defined as the reduction in global uncertainty quantified by the DS entropy $\mathcal{E}_t$ of the belief map:

$$r_t^e \triangleq \max\{0, \, \mathcal{E}_{t-1} - \mathcal{E}_t\}, \tag{14}$$

which also captures the entropy reduction induced by freshness-weighted synchronization events upon exiting blocked zones. The freshness-regularization term penalizes prolonged operation inside blocked zones:

$$r_t^\theta \triangleq \frac{1}{K} \sum_{k \in \mathcal{K}} \frac{\theta_t^k}{\theta_f}, \tag{15}$$

where θ_t^k is the freshness index of UAV k and θ_f is a normalization constant.

Remark 3. *The freshness-regularization term r_t^θ introduces a direct and quantifiable influence on policy optimization beyond its role as a scalar penalty. Since θ_t^k increases monotonically while UAV k resides in a blocked zone and resets to zero upon exit, the expected gradient of the cumulative freshness penalty with respect to the policy parameters is nonzero and directional: it consistently penalizes trajectories with prolonged in-zone dwell time, thereby inducing a structured gradient bias toward timely zone-exit behaviors. Furthermore, because the freshness index is also embedded as an explicit channel in the observation tensor, the value function learned during training is conditioned on θ_t^k. Through Generalized Advantage Estimation (GAE), a higher freshness index leads to a lower estimated advantage for remaining inside the blocked zone, reinforcing zone-exit actions during the policy update step. These two pathways—direct reward penalty and value-conditioned advantage—constitute complementary and theoretically traceable mechanisms through which AoI quantitatively shapes the learned policy.*

3.2.2. Constraints and Termination Conditions

The admissible decision sequence must respect the following constraints:

$$\text{C1: } \mathbf{r}_t^k \in \Omega, \quad \forall k \in \mathcal{K}, \, \forall t, \tag{16}$$

$$\text{C2: } a_t^k \in \mathcal{A}, \quad \forall k \in \mathcal{K}, \, \forall t, \tag{17}$$

$$\text{C3: verification is disallowed when } \Phi(\mathbf{r}_t^k) = 1, \quad \forall k \in \mathcal{K}, \, \forall t, \tag{18}$$

$$\text{C4: } \pi_1(u, v) + \pi_0(u, v) + \pi_\phi(u, v) = 1, \quad \forall (u, v) \in \Omega, \, \forall t. \tag{19}$$

The episode terminates once the cumulative number of verified targets reaches D or when t exceeds T_f.

4. MARL-Based Search Optimization

In this section, we present a multi-agent reinforcement learning (MARL) approach to address the optimization problem stated in Section 3.2 [31–33]. The objective Ψ is highly non-convex and tightly coupled owing to (i) inter-agent dependencies, (ii) stochastic

sensing outcomes, and (iii) the DS/freshness-driven belief-update dynamics. A coordinated policy is therefore learned through repeated interaction with the environment.

4.1. Cooperative Markov Game and CTDE Paradigm

We formulate the cooperative target search as a fully cooperative Markov game $\langle \mathcal{K}, \mathcal{S}, \{\mathcal{A}^k\}, \mathcal{T}, \varrho, \zeta \rangle$, where $\varrho : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the team reward function, ϱ_t denotes its realization at step t, and all agents share the same team reward. Here, the discount factor ζ has the same meaning as that defined in Section 3.2. At time step t, each UAV $k \in \mathcal{K}$ receives a local observation $\mathbf{x}_t^k$ and selects an action $a_t^k \in \mathcal{A}^k$ according to a stochastic policy $\varpi_\omega^k(a_t^k \mid \mathbf{x}_t^k)$. The joint action is $\mathbf{a}_t = (a_t^1, \ldots, a_t^K)$ and the environment evolves as

$$\mathbf{s}_{t+1} \sim \mathcal{T}(\cdot \mid \mathbf{s}_t, \mathbf{a}_t),$$

yielding a shared reward ϱ_t.

We adopt the centralized-training decentralized-execution (CTDE) paradigm. A centralized critic accesses richer joint information during training to stabilize multi-agent learning, while each UAV acts autonomously using only its private observation at execution time.

Specifically, during centralized training, the critic network receives the concatenation of all agents' local observations as its input, enabling it to capture inter-agent dependencies and produce more accurate value estimates for policy optimization. During decentralized execution, each UAV selects actions based solely on its own local observation $\mathbf{x}_t^k$, without requiring communication with other agents or access to global state information. This design ensures that the learned policies remain deployable under partial observability and communication constraints at execution time. For PPO, each agent maintains an independent actor–critic pair trained on local observations, following an independent learning paradigm. For MAPPO, a shared centralized critic conditioned on the joint observation is used during training, while each agent's actor operates independently at execution time. For QMIX, individual Q-networks are trained in a decentralized manner, while a mixing network combines individual Q-values into a joint Q-value during training to enforce monotonic value decomposition [34].

4.2. Observation and Action Representation

4.2.1. Observation

The observation of UAV k at time t is the $6 \times L \times L$ tensor

$$\mathbf{x}_t^k = \mathrm{stack}\Big(\pi_1(\cdot),\ \pi_0(\cdot),\ \pi_\phi(\cdot),\ \Phi(\cdot),\ \mathrm{Loc}(\mathbf{r}_t^k),\ \mathrm{Stale}(\theta_t^k)\Big), \tag{20}$$

where $\mathrm{Loc}(\mathbf{r}_t^k)$ is a one-hot position map and $\mathrm{Stale}(\theta_t^k)$ is a constant-valued map whose entries equal $\min\{1, \theta_t^k/\theta_f\}$.

The inclusion of the freshness channel in the observation tensor ensures that both the policy and value networks are explicitly conditioned on the agent's current AoI state: the value head learns to associate higher staleness with lower expected return, which propagates through GAE to produce lower advantage estimates for in-zone actions, thereby strengthening the gradient signal toward zone-exit behaviors during policy updates.

4.2.2. Action

All UAVs share the same discrete action set $\mathcal{A} = \mathcal{A}^k$ for all $k \in \mathcal{K}$, defined as

$$\mathcal{A} = \{\texttt{Up}, \texttt{Down}, \texttt{Left}, \texttt{Right}, \texttt{Stay}\}, \qquad \forall k \in \mathcal{K}. \tag{21}$$

4.3. Learning-Based Instantiations with CNN Encoders

To demonstrate the generality of the proposed framework, we instantiate it with three representative deep reinforcement learning algorithms: PPO, MAPPO, and QMIX. Although these methods differ in their learning paradigms, all share CNN encoders that process map-structured observations and extract spatial features relevant to target search, blocked-zone awareness, and belief updating.

4.3.1. PPO Instantiation

For PPO, decentralized action policies are trained from local observations, where a CNN encoder distills spatial features for action selection and state-value estimation:

$$\mathbf{e}_t^k = \varphi_{\text{cnn}}(\mathbf{x}_t^k), \tag{22}$$

from which the decentralized policy and individual value function are derived as

$$\omega_\omega^k(\cdot \mid \mathbf{x}_t^k) = \text{Cat}\big(\text{softmax}(W_\omega \mathbf{e}_t^k + b_\omega)\big), \quad V_{\psi^k}(\mathbf{x}_t^k) = W_V \mathbf{e}_t^k + b_V, \tag{23}$$

where ψ^k denotes the value network parameters specific to UAV k, distinct from the shared critic parameters ψ used in MAPPO.

4.3.2. MAPPO Instantiation

For MAPPO, each UAV maintains a local actor conditioned on its own observation, while a centralized critic evaluates the aggregated team state for more stable coordination learning. The actor network is defined as

$$\mathbf{e}_t^k = \varphi_{\text{cnn}}(\mathbf{x}_t^k), \quad \omega_\omega^k(\cdot \mid \mathbf{x}_t^k) = \text{Cat}\big(\text{softmax}(W_\omega \mathbf{e}_t^k + b_\omega)\big). \tag{24}$$

Actor parameters are shared across all UAVs to improve sample efficiency and cross-agent generalization. The centralized critic aggregates observations from the entire fleet:

$$\mathbf{z}_t = \text{concat}\big(\mathbf{x}_t^1, \ldots, \mathbf{x}_t^K\big), \quad V_\psi(\mathbf{z}_t) = f_\psi(\mathbf{z}_t), \tag{25}$$

where $f_\psi(\cdot)$ is implemented via a CNN backbone followed by fully connected layers.

4.3.3. QMIX Instantiation

For QMIX, each UAV maintains an individual action-utility function derived from its local observation, and a monotonic mixing network aggregates these individual utilities into a global team action-value. The per-agent utility network also employs a CNN encoder:

$$\mathbf{e}_t^k = \varphi_{\text{cnn}}(\mathbf{x}_t^k), \quad \tilde{Q}_k(\mathbf{x}_t^k, a_t^k) = q_\chi(\mathbf{e}_t^k, a_t^k). \tag{26}$$

The global team action-value is then obtained through a monotonic aggregation:

$$Q_{\text{jt}}(\boldsymbol{\tau}_t, \mathbf{a}_t, \mathbf{z}_t) = \text{Mixer}\big(\tilde{Q}_1, \ldots, \tilde{Q}_K; \mathbf{z}_t\big), \tag{27}$$

where $\boldsymbol{\tau}_t$ denotes the agents' action-observation histories, and the mixing network is conditioned on the same global state representation $\mathbf{z}_t = \text{concat}(\mathbf{x}_t^1, \ldots, \mathbf{x}_t^K)$ as defined in the MAPPO critic.

Overall, the proposed framework is algorithm-agnostic and naturally compatible with independent actor–critic, CTDE, and value-decomposition reinforcement learning methods. This algorithm-agnostic nature follows from two key structural properties. First, the AoI-DS mechanism—comprising the DS belief map, freshness index, onboard buffering,

and synchronization—is implemented entirely within the environment dynamics and is decoupled from the internal update rules of any specific RL algorithm. Any algorithm that consumes the augmented observation $\mathbf{x}_t^k$ and optimizes the team reward ϱ_t will therefore interact with the AoI-DS mechanism transparently, without requiring modifications to its learning procedure. Second, the observation tensor defined in Equation (20) is structurally compatible with any policy network that accepts spatial feature maps as input, including recurrent, attention-based, or graph-neural-network architectures. These two properties together imply that the framework can, in principle, be instantiated with any MARL algorithm—value-based, policy-gradient, or actor–critic—operating under partial observability. The three backbones evaluated in this work, PPO, MAPPO, and QMIX, are deliberately selected to span three structurally distinct MARL paradigms (independent actor–critic, centralized-critic policy gradient, and monotonic value decomposition), providing empirical support for this claim of generality across representative algorithm families.

4.4. Training Objective for PPO and MAPPO

We collect on-policy rollouts of length T_r and optimize the clipped surrogate objective. Let $\hat{G}_t$ be the bootstrapped return and $\hat{A}_t$ be the generalized advantage estimate (GAE):

$$e_t^{\mathrm{td}} = \varrho_t + \zeta V_\psi(\mathbf{z}_{t+1}) - V_\psi(\mathbf{z}_t), \qquad \hat{A}_t = \sum_{l=0}^{T_r-1} (\zeta\lambda)^l e_{t+l}^{\mathrm{td}}, \tag{28}$$

where $\lambda \in [0,1]$ is the GAE smoothing parameter. For MAPPO with shared reward, the same $\hat{A}_t$ is used across agents, or per-agent advantages are derived from the shared critic depending on the implementation.

Define the importance-sampling ratio for UAV k:

$$\Gamma_t^k(\omega) = \frac{\varpi_\omega^k(a_t^k \mid \mathbf{x}_t^k)}{\varpi_{\omega_{\mathrm{old}}}^k(a_t^k \mid \mathbf{x}_t^k)}. \tag{29}$$

The actor is updated by maximizing the clipped surrogate objective (summed over agents):

$$\mathcal{L}^{\mathrm{surr}}(\omega) = \mathbb{E}\left[\frac{1}{K}\sum_{k\in\mathcal{K}} \min\left(\Gamma_t^k(\omega)\hat{A}_t,\ \mathrm{clip}(\Gamma_t^k(\omega), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})\hat{A}_t\right)\right], \tag{30}$$

where $\epsilon_{\mathrm{clip}} > 0$ is the clipping coefficient used in PPO-style policy updates. The shared critic is trained by minimizing the squared prediction error:

$$\mathcal{L}^{\mathrm{val}}(\psi) = \mathbb{E}\Big[\big(V_\psi(\mathbf{z}_t) - \hat{G}_t\big)^2\Big]. \tag{31}$$

An entropy bonus $\mathcal{L}^{\mathrm{ent}}(\omega)$ is optionally appended to promote exploration, yielding the composite objective

$$\mathcal{L}(\omega) = \mathcal{L}^{\mathrm{surr}}(\omega) + \varkappa\,\mathcal{L}^{\mathrm{ent}}(\omega)$$

with coefficient $\varkappa \geq 0$.

4.5. Training Objective for QMIX

Unlike PPO and MAPPO, QMIX is trained in an off-policy manner via temporal-difference (TD) learning and uses a TD target $\hat{q}_t$ in place of the bootstrapped return $\hat{G}_t$ used in PPO/MAPPO. Each agent estimates an individual action-utility function from its local history, and the mixing network fuses these into the global team action-value Q_{jt}.

For a sampled experience tuple $(\boldsymbol{\tau}_t, \mathbf{a}_t, \varrho_t, \mathbf{z}_t, \boldsymbol{\tau}_{t+1}, \mathbf{z}_{t+1})$, the regression target is constructed as

$$\hat{q}_t = \varrho_t + \zeta\, Q_{\mathrm{jt}}^{-}(\boldsymbol{\tau}_{t+1}, \mathbf{a}_{t+1}^{*}, \mathbf{z}_{t+1}), \tag{32}$$

where Q_{jt}^{-} is a periodically synchronized target network, and $\mathbf{a}_{t+1}^{*}$ is the joint greedy action derived from the individual target utilities. The network parameters are refined by minimizing the squared deviation:

$$\mathcal{L}^{\mathrm{jt}}(\chi) = \mathbb{E}\Big[\big(Q_{\mathrm{jt}}(\boldsymbol{\tau}_t, \mathbf{a}_t, \mathbf{z}_t) - \hat{q}_t\big)^2\Big]. \tag{33}$$

Minimizing $\mathcal{L}^{\mathrm{jt}}$ drives the agents toward globally coordinated action values while respecting the monotonic decomposability constraint that enables each UAV to act independently during deployment.

4.6. Execution

After training, the learned policy operates in a fully decentralized manner. At each time step, each UAV selects its action based solely on its private observation $\mathbf{x}_t^k$, using either the actor network (PPO and MAPPO) or the individual utility network (QMIX), without requiring explicit inter-agent communication. The DS evidence fusion, freshness-index accumulation, and synchronization operations remain part of the environment dynamics, shaping agent behavior through both observation evolution and reward feedback.

4.7. Generalizability of the Proposed Framework

Although the experiments in this paper are conducted in a discrete grid-world setting, the core components of the proposed framework are not inherently restricted to discrete state spaces and can be generalized to continuous environments and dynamic communication conditions.

Extension to continuous environments. The DS belief map operates over spatial cells and can be naturally extended to continuous domains through finer spatial discretization or through Gaussian process-based continuous belief representations, in which the evidential masses π_1, π_0, and π_ϕ are replaced by spatially continuous functions. The freshness-weighted synchronization mechanism depends only on the temporal structure of communication blockage, i.e., the elapsed time since the last successful upload, and is therefore agnostic to whether the underlying state space is discrete or continuous. The MARL backbone can be replaced by continuous-action variants such as Soft Actor–Critic (SAC) or continuous-action PPO to accommodate continuous mobility control, while the CNN-based observation encoder can be substituted by a more expressive architecture such as a vision transformer to handle higher-resolution spatial inputs.

Extension to dynamic communication conditions. The current framework models blocked zones as static spatial structures, which captures terrain-induced or infrastructure-related communication constraints common in practical UAV deployments. The framework can be extended to time-varying communication topologies by allowing the blockage map Φ to evolve over time. In this case, the AoI index θ_t^k and the synchronization mechanism remain structurally unchanged, as they depend only on whether communication is available at each time step rather than on the specific spatial configuration of blocked zones. To handle the resulting history-dependent belief over communication availability, recurrent policy architectures such as LSTM-based actors can be incorporated to maintain temporal context across time steps.

The above two extensions—to continuous-state environments and to time-varying communication topologies—represent promising directions for future work that will build upon the framework proposed in this paper.

5. Simulation and Results

This section evaluates the proposed framework from four complementary perspectives: overall performance against baselines, sensitivity to critical environmental parameters, convergence behavior during training, and ablation analysis of key design components. Under a shared experimental protocol, the framework is paired with PPO, MAPPO, and QMIX to assess its generality across different reinforcement learning paradigms. Robustness is further examined by varying the fleet size, the extent of signal-blocked coverage, and the onboard sensor reach. In addition, training dynamics are analyzed to characterize the learning stability and convergence properties of each variant, while ablation experiments are conducted to quantify the contribution of the major components of the proposed framework.

5.1. Simulation Environment Setup

We construct a virtual grid-world environment in which a fleet of UAVs executes coordinated reconnaissance under the DS-belief and freshness-index-driven information management scheme. The workspace is partitioned into uniform square cells, and the UAVs operate collectively to locate and verify hidden targets. At the start of each episode, UAV departure positions, target locations, and blocked-zone placements are drawn at random; thereafter, the blocked zones remain stationary. Each UAV may take up to $T_f = 300$ movement steps per episode. Default environment and training hyperparameters are listed in Table 3 unless a parameter is explicitly varied in sensitivity experiments. All algorithms are implemented in Python 3.10 with PyTorch 2.1.0 and trained on a server with an NVIDIA GeForce RTX 5090 GPU and 32 GB RAM.

Table 3. Simulation environment and parameter settings.

Parameter	Value	Parameter	Value
Map Size	20×20	Episode Horizon	$T_f = 300$
Sensor Range	4	Number of UAVs	3
Number of Targets	15	Belief Decay Factor	0.995
Number of Blocked Zones	2	Block Range	4
Learning Rate	2×10^{-4}	Batch Size	64

5.2. Baseline Methods and Evaluation Metrics

To assess the effectiveness and generality of the proposed framework, we embed it into three representative reinforcement learning backbones and contrast the augmented variants against their unmodified counterparts.

5.2.1. Baseline Methods

The comparison involves **PPO**, **MAPPO**, and **QMIX** as unmodified reference methods, alongside their belief-and-freshness-augmented counterparts **AoI-DS-PPO**, **AoI-DS-MAPPO**, and **AoI-DS-QMIX**. A rule-based reference agent, **Least-Visited Greedy**, is also included; it steers toward under-explored cells and promptly exits blocked zones when trapped. To further evaluate the contribution of the proposed framework, we additionally introduce the communication-aware baseline **MARL-DSC** [20], following the work of Hou et al., which operates under spatially structured communication constraints and employs Bayesian updates to maintain a probabilistic sensor observation map. Together, these baselines span heuristic, conventional RL, communication-aware, and framework-augmented RL categories, enabling a systematic and comprehensive evaluation of the proposed framework.

The pairwise comparison between each base algorithm and its augmented counterpart, together with the communication-aware MARL-DSC baseline, reveals whether the pro-

posed framework yields consistent and non-redundant gains irrespective of the underlying learning paradigm.

5.2.2. Evaluation Metrics

Four indicators are adopted to assess the compared methods: **average episode reward**, **success rate**, **episode length**, and **number of targets found**. These collectively reflect policy quality, mission accomplishment, search efficiency, and uncertainty reduction capability.

Experiments are structured around three objectives. First, quantitative comparisons with all baselines establish the overall advantage and task-level superiority of the proposed approach. Second, a systematic sensitivity study probes the influence of critical design variables on model behavior, substantiating the rationality of algorithmic choices and robustness to parameter shifts. Third, a training-curve analysis examines reward evolution and task-metric trajectories to characterize learning stability, sample efficiency, and convergence regularity of each method.

5.3. Performance Comparison

This subsection presents a head-to-head performance comparison between the proposed framework and the competing baselines. Quantitative results are reported in Table 4; the clear and consistent advantage of the augmented methods can be read directly from the table.

Table 4. Quantitative comparison of all methods on the multi-UAV cooperative target-search task.

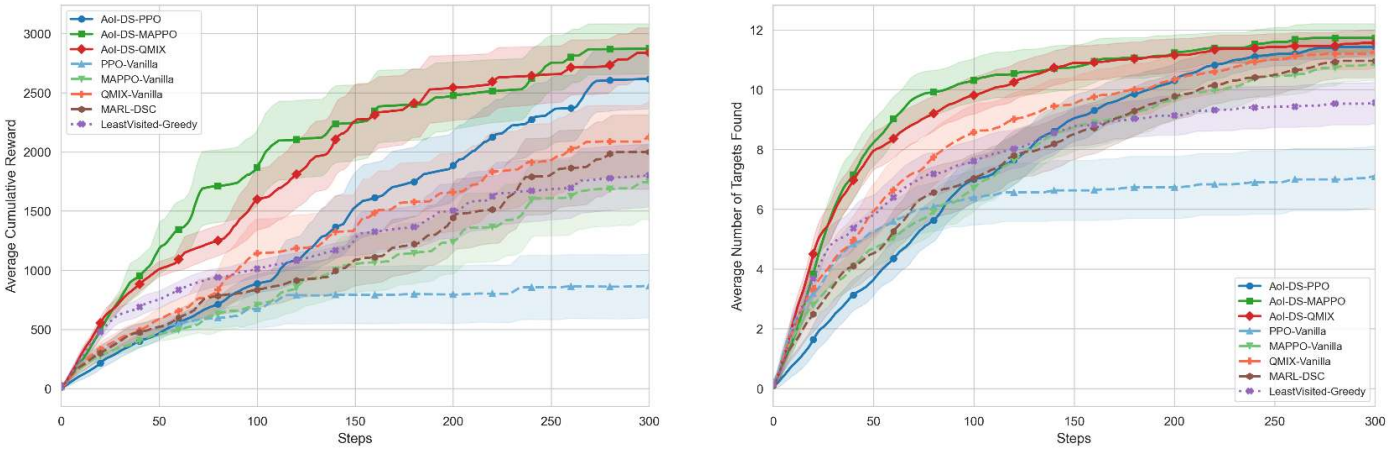
Method	Avg. Reward ↑	Success Rate ↑	Targets Found ↑	Episode Length ↓
Greedy Algorithm	2059.25 ± 813.40	0.36	10.14 ± 1.97	249.42 ± 76.62
MARL-DSC	2363.11 ± 740.76	0.74	11.52 ± 1.00	221.04 ± 64.32
PPO	1640.28 ± 1013.77	0.48	9.56 ± 3.09	238.16 ± 70.89
QMIX	2029.66 ± 822.17	0.62	11.36 ± 0.95	222.30 ± 76.86
MAPPO	1967.74 ± 863.84	0.60	11.04 ± 1.46	241.76 ± 60.01
AoI-DS-PPO (Ours)	2697.62 ± 1021.38	0.78	10.66 ± 3.06	159.56 ± 87.97
AoI-DS-MAPPO (Ours)	**2980.68 ± 401.57**	**0.92**	**11.94 ± 0.42**	**150.40 ± 83.70**
AoI-DS-QMIX (Ours)	2973.61 ± 471.48	0.90	11.80 ± 0.69	165.66 ± 80.02

Note: ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better. Bold values indicate the best performance among the compared methods, and underlined values indicate the second-best performance.

Across all four metrics, the AoI-DS-augmented variants decisively outperform their corresponding baselines. AoI-DS-MAPPO attains the highest overall scores: average reward 2980.68, success rate 0.92, targets found 11.94, and episode length 150.40, reflecting both superior search quality and faster task completion. AoI-DS-QMIX ranks closely behind, recording average reward 2973.61, success rate 0.90, targets found 11.80, and episode length 165.66, confirming its strong competitiveness. The unaugmented PPO, MAPPO, and QMIX deliver acceptable individual-metric results but fall uniformly short of their augmented counterparts in composite performance. The gains attributable to the AoI-DS mechanism are especially pronounced in success rate and execution speed. The rule-based Greedy agent achieves moderate reward and target-discovery numbers but records a success rate of only 0.36 and a comparatively long episode length of 249.42, confirming that purely heuristic exploration struggles to balance reliable completion and time efficiency in this complex cooperative setting. MARL-DSC, which operates under spatially structured communication constraints and employs Bayesian updates to maintain a probabilistic sensor observation map but excludes the AoI-weighted synchronization mechanism, achieves an intermediate performance of average reward 2363.11, success rate 0.74, targets found 11.52, and episode length 221.04. While it surpasses all unaugmented RL baselines, it falls consistently short

of the full AoI-DS variants across all four metrics, confirming that the AoI-weighted synchronization mechanism provides additional and non-redundant performance gains beyond Bayesian-based belief updating alone. Taken together, these results validate that the AoI-DS mechanism substantially lifts both mission effectiveness and execution efficiency regardless of the underlying RL backbone, with the most pronounced gains observed for MAPPO and QMIX.

Figure 3 traces how performance evolves as the available step budget grows. Regarding cumulative reward, AoI-DS-MAPPO and AoI-DS-QMIX lead the field across nearly the full step range, with AoI-DS-MAPPO pulling ahead slightly in the high-step regime. AoI-DS-PPO improves at a measured pace and eventually eclipses all unaugmented baselines, though its gains materialize more slowly in the initial phase. MARL-DSC achieves a moderate cumulative reward trajectory, consistently outperforming the three unaugmented RL methods and the Greedy agent throughout the full step range but remaining below all AoI-DS variants, further substantiating that the AoI-weighted and DS-fusion synchronization mechanism contributes meaningfully beyond Bayesian-based belief updating alone. The three unaugmented RL methods lag noticeably, with vanilla PPO saturating at a substantially lower reward plateau. The Greedy agent fares well in the low-to-mid step range but plateaus quickly, revealing limited capacity for long-horizon strategic reasoning relative to the learned cooperative policies.



(**a**) Average cumulative reward versus steps. (**b**) Average number of targets found versus steps.

Figure 3. Step-budget performance of all methods. Shaded bands denote 95% confidence intervals.

Regarding target discovery, the same ordering holds. With a larger step allowance, every method locates more targets, yet the AoI-DS variants sustain a persistent lead throughout. AoI-DS-MAPPO and AoI-DS-QMIX reach high discovery counts well before the other methods and hold those levels to the end, reflecting superior coverage efficiency and inter-agent coordination under tight time budgets. AoI-DS-PPO lags initially but progressively catches up, finishing at a level comparable to the top performers. MARL-DSC occupies an intermediate position, discovering more targets than the unaugmented RL methods and the Greedy agent but consistently falling short of the full AoI-DS variants. Vanilla PPO improves slowly and converges to a lower asymptote. Collectively, these curves confirm that integrating the AoI-DS mechanism meaningfully amplifies both search breadth and cumulative return, and the margin widens as more steps become available.

5.4. Parameter Sensitivity Analysis

Selecting appropriate settings for critical environmental variables is essential to understanding the behavioral envelope of the proposed framework. We examine three such

variables: fleet size, degree of blocked-zone coverage, and per-agent sensor reach. Every reported figure is averaged over fifty independent runs with distinct random seeds, with all other settings held at the defaults in Table 2.

5.4.1. Effect of the Number of UAVs

As shown in Figure 4a–d, scaling the fleet from two to four UAVs yields performance gains across all three algorithms, though the magnitude varies by method. AoI-DS-MAPPO benefits most consistently: average reward climbs from 2548.77 to 3514.35, success rate from 0.92 to 0.98, episode length contracts from 185.68 to 108.66, and targets found edges up from 11.77 to 11.98, demonstrating that additional UAVs translate effectively into improved coverage and cooperation.

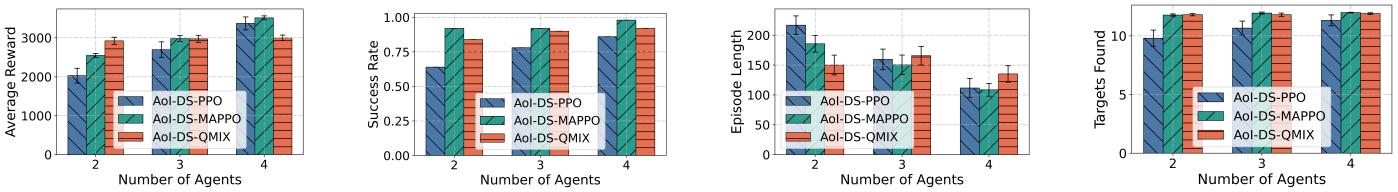


Figure 4. Sensitivity analysis with respect to the number of UAVs: (**a**) average reward, (**b**) success rate, (**c**) episode length, and (**d**) number of targets found.

AoI-DS-PPO also responds positively: average reward rises from 2030.52 to 3370.84, success rate from 0.64 to 0.86, episode length falls from 216.80 to 111.58, and targets found from 9.80 to 11.32. Its larger variance relative to AoI-DS-MAPPO likely reflects the absence of an explicit centralized coordination signal during training, making it more susceptible to team-size-induced dynamics.

AoI-DS-QMIX responds more modestly: average reward increases from 2923.56 to 2998.89, success rate from 0.84 to 0.92, episode length from 150.18 to 135.24, and targets found from 11.80 to 11.90. Overall, larger fleets consistently improve collaborative search, with AoI-DS-MAPPO gaining the most and AoI-DS-QMIX exhibiting the most stable trend.

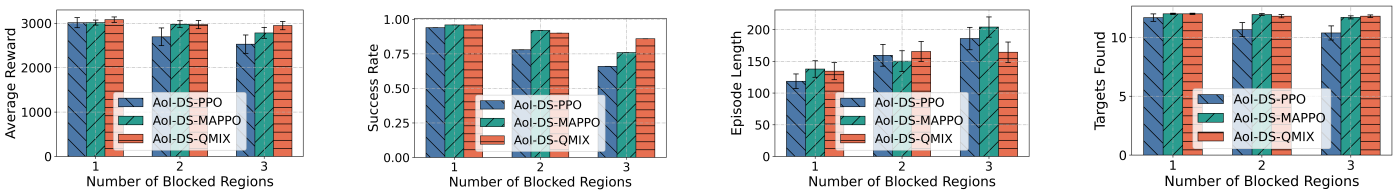
5.4.2. Effect of the Number of Blocked Zones

As shown in Figure 5a–d, increasing the number of blocked zones from 1 to 3 generally leads to performance degradation across all methods, reflecting the additional difficulty introduced by higher communication disruption density. AoI-DS-MAPPO exhibits notable performance decline: average reward drops from 3016.56 to 2785.40, success rate falls from 0.96 to 0.76, episode length increases from 137.92 to 204.12, and the number of detected targets decreases from 12.00 to 11.70, indicating that even the strongest variant is not fully immune to heavily obstructed environments. It is worth noting that when no blocked zones are present, neither the AoI-weighted synchronization mechanism nor the information cache fusion upload mechanism is triggered, making it impossible to reflect the performance benefits of our proposed framework. Therefore, the no-blocked-zone environment is excluded from consideration.

AoI-DS-PPO is the most affected: average reward falls from 3016.10 to 2531.37, success rate from 0.94 to 0.66, episode length grows from 118.66 to 186.04, and targets found decreases from 11.68 to 10.38. Its decentralized policy learning appears less adept at sustaining coordination under repeated communication interruptions.

AoI-DS-QMIX exhibits the greatest resilience: average reward declines only moderately from 3083.20 to 2950.05, success rate shifts from 0.96 to 0.86, episode length from 134.62 to 164.52, and targets found remains relatively stable from 12.00 to 11.80. In summary,

denser blocked zones harm all methods, but AoI-DS-QMIX degrades most gracefully while AoI-DS-PPO is most vulnerable to disrupted information flow.

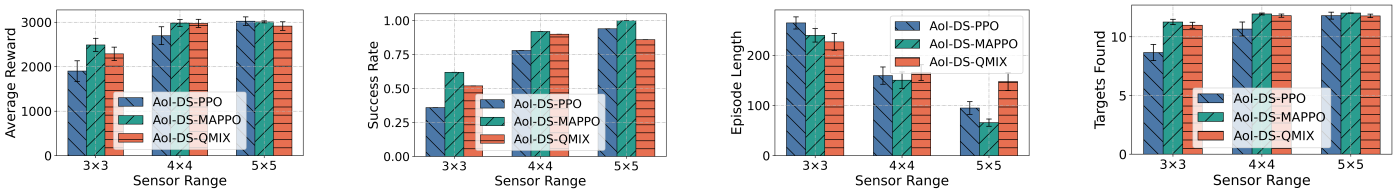


(**a**) (**b**) (**c**) (**d**)

Figure 5. Sensitivity analysis with respect to the number of blocked zones: (**a**) average reward, (**b**) success rate, (**c**) episode length, and (**d**) number of targets found.

5.4.3. Effect of the Sensing Range

As shown in Figure 6a–d, broadening the sensor detection radius from 3×3 to 5×5 cells produces clear gains in all methods, reflected by higher rewards and success rates, shorter episodes, and improved target-finding performance. AoI-DS-MAPPO achieves the strongest absolute gains: average reward rises from 2490.24 to 3007.09, success rate from 0.62 to 1.00, episode length drops sharply from 239.74 to 65.82, and targets found increases from 11.26 to 12.02, indicating that wider local perception substantially boosts both search efficiency and task reliability.



(**a**) (**b**) (**c**) (**d**)

Figure 6. Sensitivity analysis with respect to the sensor range: (**a**) average reward, (**b**) success rate, (**c**) episode length, and (**d**) number of targets found.

AoI-DS-PPO records the steepest relative improvement: average reward from 1901.88 to 3024.02, success rate from 0.36 to 0.94, episode length from 264.52 to 95.02, and targets found from 8.66 to 11.80, confirming that an extended sensor footprint effectively offsets the limitations of purely decentralized local observation.

AoI-DS-QMIX also improves steadily: average reward from 2289.31 to 2913.29, success rate from 0.52 to 0.86, episode length from 226.80 to 147.64, and targets found from 10.98 to 11.78. Sensor reach thus stands as one of the most influential task variables; enlarging local perception markedly amplifies the collaborative search capability of all AoI-DS-enhanced agents.

5.5. Convergence Analysis

To obtain a thorough picture of how the three variants learn, we examine their training dynamics from two vantage points: within-method reproducibility across random seeds, and cross-method comparison of convergence speed, performance ceiling, and training smoothness. This dual-perspective analysis goes beyond final test metrics and reveals how each method learns over the full training trajectory in the challenging cooperative search scenario.

5.5.1. Training Stability Analysis

Figure 7 presents the mean training trajectories with confidence bands across four independent random seeds for each variant. All three methods—AoI-DS-PPO, AoI-DS-

MAPPO, and AoI-DS-QMIX—display well-defined improvement trends and satisfactory cross-seed reproducibility. As training proceeds, the mean reward rises, the mission success fraction climbs, and the episode duration shrinks across all methods, signaling that the learned policies gain increasing proficiency in locating targets and finishing the cooperative search task.

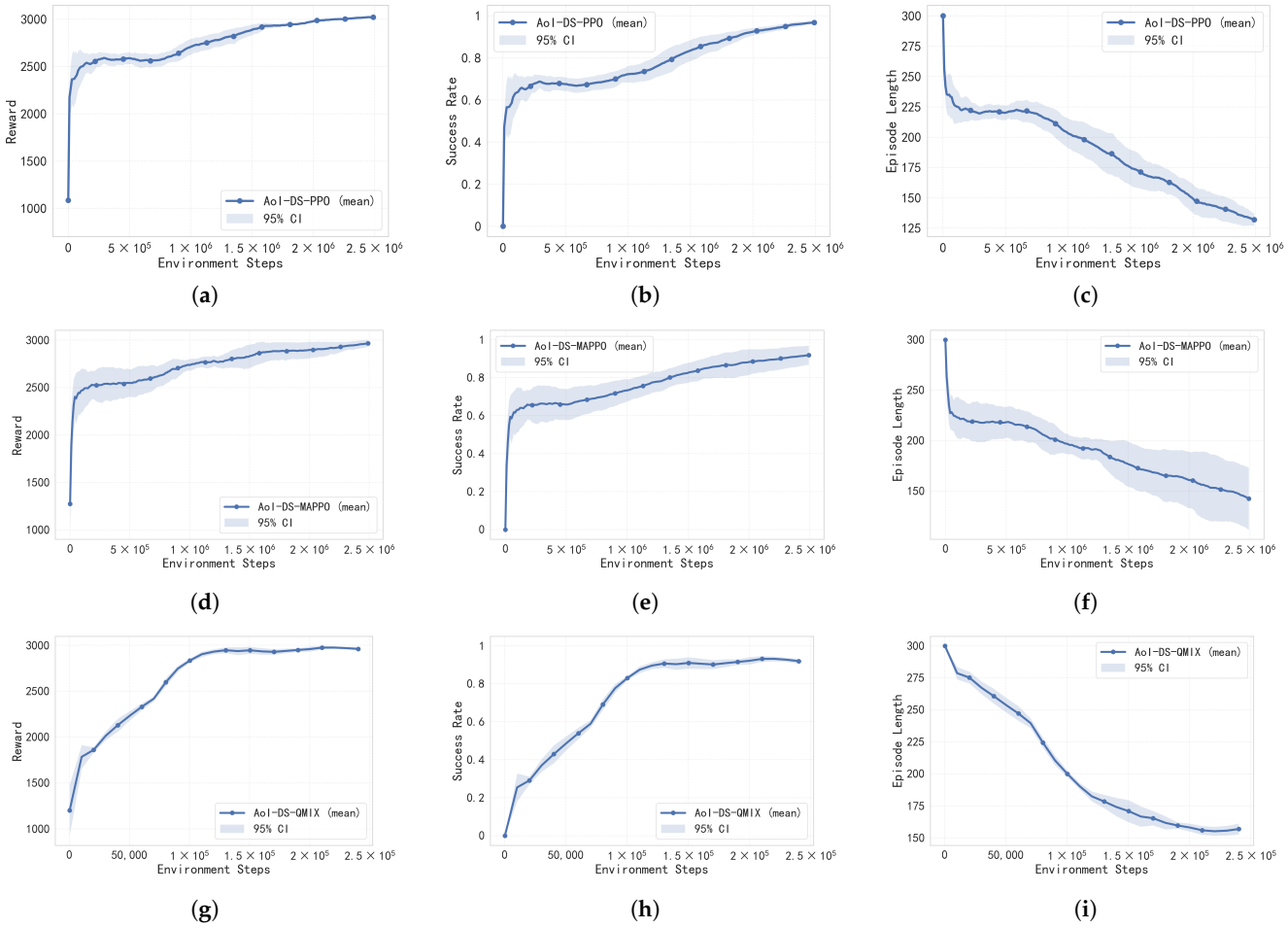


Figure 7. Cross-seed training stability of the AoI-DS-augmented methods. (**a**–**c**) AoI-DS-PPO; (**d**–**f**) AoI-DS-MAPPO; (**g**–**i**) AoI-DS-QMIX. Solid lines denote the mean trajectory across four independent seeds and shaded bands represent one standard deviation.

As shown in Table 5, AoI-DS-PPO achieves the lowest IQR (23.72), indicating the highest convergence stability among the three variants. AoI-DS-QMIX and AoI-DS-MAPPO exhibit somewhat larger IQR values (40.68 and 47.93, respectively), reflecting greater seed-to-seed variation in the converged reward distribution. In terms of AUC, AoI-DS-PPO leads (2798.37), followed by AoI-DS-MAPPO (2769.94) and AoI-DS-QMIX (2638.67), indicating that AoI-DS-PPO achieves the highest overall training efficiency. AoI-DS-MAPPO, despite showing the largest training variance, achieves the best overall performance in evaluation, which is consistent with the known characteristic of MAPPO that its centralized critic facilitates stronger policy generalization at the cost of higher training variance.

On the whole, the multi-seed results confirm that the proposed AoI-DS framework fosters reliable convergence across diverse reinforcement learning backbones and exhibits satisfactory robustness to training randomness in the considered cooperative search setting.

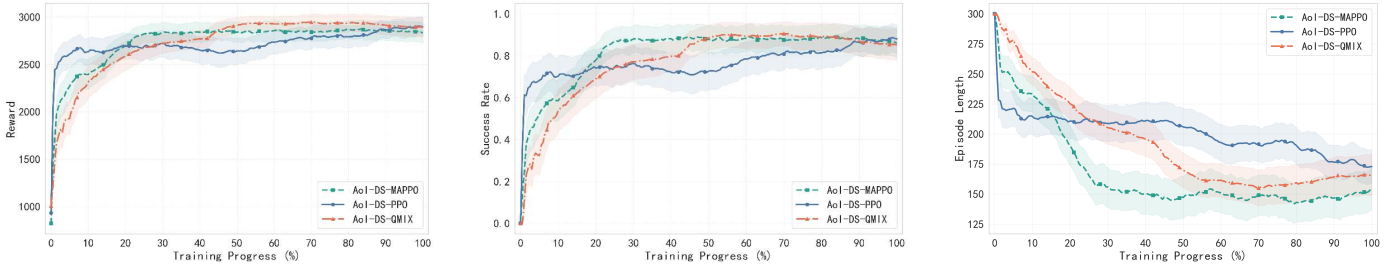
Table 5. Quantitative convergence stability metrics of the three AoI-DS variants. IQR is computed over the final 10% of training steps across all seeds; AUC is the area under the mean reward curve normalized by the total training steps of each method.

Method	IQR ↓	AUC ↑
AoI-DS-PPO	23.72	2798.37
AoI-DS-MAPPO	47.93	2769.94
AoI-DS-QMIX	40.68	2638.67

Note: ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better.

5.5.2. Comparative Analysis of Convergence Behavior

Figure 8 presents the training trajectories of all three AoI-DS variants on a shared axis. Each method exhibits a consistent improvement pattern: cumulative reward and success rate both increase while episode duration contracts, attesting to progressively more effective cooperative search strategies throughout training.



(**a**) Average reward (**b**) Success rate (**c**) Episode length

Figure 8. Inter-algorithm training comparison of AoI-DS-PPO, AoI-DS-MAPPO, and AoI-DS-QMIX across (**a**) average reward, (**b**) success rate, and (**c**) episode length. All three variants exhibit upward reward and success-rate trends alongside a shrinking episode length, confirming progressive acquisition of effective cooperative search strategies.

AoI-DS-PPO displays the steepest early-phase ascent. Reward and success rate climb sharply from the outset and episode length contracts rapidly within the first portion of training. However, the rate of improvement plateaus in later phases, and the method's final reward, success rate, and episode length remain modestly below the other two variants, suggesting that rapid initial learning comes at the cost of a somewhat lower performance ceiling.

AoI-DS-MAPPO progresses at a slightly more measured pace at the start but maintains a more regular improvement trajectory and reaches a high-performance operating region after roughly 20% of total training. It subsequently holds that level stably and records the shortest episode length among the three, reflecting superior search efficiency and the highest overall final performance.

AoI-DS-QMIX follows a more gradual ascent throughout training. Although its early gains are modest, it continues to improve steadily and closes the gap with AoI-DS-MAPPO in the mid-to-late training phases. Its terminal reward and success rate closely match those of AoI-DS-MAPPO, demonstrating strong task effectiveness; however, its episode length remains marginally longer, implying slightly lower execution efficiency.

Across all three variants, the training evidence collectively validates the effectiveness and versatility of the proposed AoI-DS-enhanced cooperative search framework.

5.6. Ablation Study

Since AoI-DS-MAPPO achieves the best overall results in the main comparison, we select MAPPO as the representative backbone for ablation analysis. Specifically, we evaluate

several reduced variants by removing or replacing key components, in order to examine their individual contributions to the overall performance.

As shown in Table 6, the full AoI-DS-MAPPO achieves the best performance on all four metrics, demonstrating the effectiveness of jointly incorporating belief-guided cooperative search and AoI-aware information updating. Removing the AoI-related design leads to a moderate performance drop, with the average reward decreasing from 2980.68 to 2948.89 and the success rate decreasing from 0.92 to 0.88, which indicates that AoI-aware updating helps improve the timeliness and quality of delayed information fusion. In contrast, removing belief guidance causes a much larger degradation, reducing the average reward to 2184.33 and increasing the episode length to 224.48, which highlights the critical role of belief-map-based global guidance in cooperative search. When both components are removed, the performance further drops to the vanilla MAPPO level, yielding the worst overall results. These findings reveal an asymmetric yet complementary contribution between the two components. Belief guidance provides the dominant performance gain, as it directly shapes the global spatial uncertainty representation that guides cooperative exploration. The AoI component, while contributing a smaller absolute gain, addresses a distinct function that belief guidance alone cannot provide: ensuring that delayed buffered evidence is integrated with appropriate temporal discounting upon zone exit. The two components are therefore complementary by design, and their combined effect underlies the overall superiority of the full AoI-DS-MAPPO framework.

Table 6. Ablation results of AoI-DS-MAPPO under communication-blocked conditions.

Method	Avg. Reward ↑	Success Rate ↑	Targets Found ↑	Episode Length ↓
AoI-DS-MAPPO (Full)	**2980.68 ± 401.57**	**0.92**	**11.94 ± 0.42**	**150.40 ± 83.70**
w/o AoI	2948.89 ± 452.92	0.88	11.84 ± 0.50	172.00 ± 78.94
w/o Belief Guidance	2184.33 ± 782.37	0.72	11.38 ± 1.16	224.48 ± 70.71
w/o AoI & Belief Guidance	1967.74 ± 863.84	0.60	11.04 ± 1.46	241.76 ± 60.01

Note: ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better. Bold values indicate the best performance among the compared methods, and underlined values indicate the second-best performance.

6. Conclusions

This paper addressed the multi-UAV cooperative target-search problem under spatially structured signal-blocked zones and introduced an AoI-DS-enhanced cooperative search framework to overcome the twin challenges of delayed belief uploading and degraded inter-agent coordination caused by communication blockage. The framework couples a Dempster–Shafer belief map for uncertainty-aware evidence fusion with a freshness-index mechanism that quantifies and compensates for observation staleness, enabling more principled belief updating and improved cooperative decision-making upon zone exit. Ablation studies confirm that belief-map guidance provides the dominant performance contribution, while the AoI-weighted synchronization mechanism offers a complementary and non-redundant gain by explicitly addressing information timeliness upon zone exit. Systematic experiments demonstrated that integrating the proposed mechanism into PPO, MAPPO, and QMIX yields consistent performance improvements: up to 26.13% higher average reward, 24.32% higher success rate, 3.65% more targets found, and 31.96% reduction in episode length relative to the strongest baseline. These outcomes substantiate the effectiveness and broad applicability of the framework for communication-constrained multi-UAV cooperative search, and future work will extend the approach to continuous-state environments and time-varying communication topologies. Additionally, topology-aware communication scheduling strategies inspired by WSN protocols such as EMO-PEGASIS,

which prioritize information upload ordering upon zone exit, represent a promising direction for further enhancing the proposed synchronization mechanism.

Author Contributions: Conceptualization, L.X., L.Y. and G.X.; Methodology, L.X. and X.H.; Software, L.X.; Validation, L.X. and X.H.; Formal analysis, L.X. and X.H.; Investigation, L.X. and L.Y.; Data curation, L.X.; Writing—original draft, L.X.; Writing—review & editing, L.X., X.D., X.H., L.Y. and G.X.; Supervision, X.D.; Funding acquisition, G.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities under QTZX25119. The APC was funded by institutional support.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
AoI	Age of Information
DS	Dempster–Shafer
MARL	Multi-Agent Reinforcement Learning
PPO	Proximal Policy Optimization
MAPPO	Multi-Agent Proximal Policy Optimization
QMIX	Q-value Mixing Network
CTDE	Centralized Training and Decentralized Execution
CNN	Convolutional Neural Network
GAE	Generalized Advantage Estimation
CDM	Communication-Delay Characterization
BM	Belief Map
MC	Mobility–Communication Interaction
Opt	Optimization-based Method

References

1. Stodola, P.; Nohel, J.; Horák, L. Dynamic reconnaissance operations with UAV swarms: Adapting to environmental changes. *Sci. Rep.* **2025**, *15*, 15092. [CrossRef] [PubMed]
2. Fan, J.; Lei, L.; Cai, S.; Shen, G.; Cao, P.; Zhang, L. Area surveillance with low detection probability using UAV swarms. *IEEE Trans. Veh. Technol.* **2023**, *73*, 1736–1752. [CrossRef]
3. Fan, X.; Wu, P.; Xia, M. Air-to-ground communications beyond 5G: UAV swarm formation control and tracking. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 8029–8043. [CrossRef]
4. Ekechi, C.C.; Elfouly, T.; Alouani, A.; Khattab, T. A survey on UAV control with multi-agent reinforcement learning. *Drones* **2025**, *9*, 484. [CrossRef]
5. García-Gil, S.; Murillo, J.M.; Galán-Jiménez, J. Enabling ultra reliable low latency communications in rural areas using UAV swarms. *Ad Hoc Netw.* **2024**, *163*, 103603. [CrossRef]
6. Shabanighazikelayeh, M.; Koyuncu, E. Optimal placement of UAVs for minimum outage probability. *IEEE Trans. Veh. Technol.* **2022**, *71*, 9558–9570. [CrossRef]
7. Du, T.; Gui, X.; Dai, H. An Attention-Driven Heterogeneous Multi-Agent Framework for UAV and Satellite-Assisted Task Offloading in Hybrid Ground Device Networks. *IEEE Internet Things J.* **2025**, *12*, 54672–54689. [CrossRef]
8. Alqefari, S.; Menai, M.E.B. Multi-UAV task assignment in dynamic environments: Current trends and future directions. *Drones* **2025**, *9*, 75. [CrossRef]
9. Zhang, W.; Liang, X.; Deng, Q.; Shu, F.; Zhang, Z.; Nie, L.; Yan, S. Joint trajectory and beamforming optimization for IRS-assisted multi-antenna UAV covert communications with a finite blocklength. *IEEE Trans. Green Commun. Netw.* **2026** , *10*, 426–439. [CrossRef]

10. Gu, J.; Wang, Y.; Ji, W.; Wei, Z.; Wang, J. LLM-Based Dynamic Event-Triggered Communication for Multi-UAV Formation Control in Urban Environments. *IEEE Trans. Cogn. Commun. Netw.* **2025**, *12*, 4825–4838. [CrossRef]
11. Yin, R.; Peng, J.; Cai, Y.; Wu, C.; Champagne, B.; Al-Dhahir, N. Radar-assisted predictive beamforming for UAV-aided networks: A deep-learning solution. *IEEE Trans. Veh. Technol.* **2025**, *74*, 16079–16093. [CrossRef]
12. Zhan, C.; Hu, H.; Wang, J.; Liu, Z.; Mao, S. Tradeoff between age of information and operation time for UAV sensing over multi-cell cellular networks. *IEEE Trans. Mob. Comput.* **2023**, *23*, 2976–2991. [CrossRef]
13. Emami, Y.; Gao, H.; Li, K.; Almeida, L.; Tovar, E.; Han, Z. Age of information minimization using multi-agent UAVs based on AI-enhanced mean field resource allocation. *IEEE Trans. Veh. Technol.* **2024**, *73*, 13368–13380. [CrossRef]
14. Hollinger, G.A.; Sukhatme, G.S. Sampling-based robotic information gathering algorithms. *Int. J. Robot. Res.* **2014**, *33*, 1271–1287. [CrossRef]
15. Wang, H.; Song, S.; Guo, Q.; Xu, D.; Zhang, X.; Wang, P. Cooperative motion planning for persistent 3d visual coverage with multiple quadrotor UAVs. *IEEE Trans. Autom. Sci. Eng.* **2023**, *21*, 3374–3383. [CrossRef]
16. Mao, Z.; Hou, M.; Li, H.; Yang, Y.; Song, W. Multi-UAV cooperative motion planning under global spatio-temporal path inspiration in constraint-rich dynamic environments. *IEEE Trans. Intell. Veh.* **2025**, *10*, 1030–1042.
17. Abdulghafoor, A.Z.; Bakolas, E. Two-level control of multiagent networks for dynamic coverage problems. *IEEE Trans. Cybern.* **2021**, *53*, 4067–4078. [CrossRef]
18. Kouzeghar, M.; Song, Y.; Meghjani, M.; Bouffanais, R. Multi-target pursuit by a decentralized heterogeneous UAV swarm using deep multi-agent reinforcement learning. *arXiv* **2023**, arXiv:2303.01799.
19. Ajina, M.; Tabatabai, D.; Nowzari, C. Asynchronous distributed event-triggered coordination for multiagent coverage control. *IEEE Trans. Cybern.* **2020**, *51*, 5941–5953. [CrossRef] [PubMed]
20. 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*, 568–578. [CrossRef]
21. Fei, B.; Bao, W.; Zhu, X.; Liu, D.; Men, T.; Xiao, Z. Autonomous cooperative search model for multi-UAV with limited communication network. *IEEE Internet Things J.* **2022**, *9*, 19346–19361. [CrossRef]
22. Zhang, B.; Hou, Y.; Yin, H.; Lv, M.; Yang, A.; Wu, L. Cooperative Dynamic Target Tracking: Distributed Time-Varying Optimization for Multi-UAV System. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 12245–12257. [CrossRef]
23. Sun, L.; Wang, J.; Wan, L.; Li, K.; Wang, X.; Lin, Y. Human-UAV interaction assisted heterogeneous UAV swarm scheduling for target searching in communication denial environment. *IEEE Trans. Autom. Sci. Eng.* **2024**, *22*, 4457–4472. [CrossRef]
24. Senthilnath, J.; Harikumar, K.; Sundaram, S. Metacognitive decision-making framework for multi-UAV target search without communication. *IEEE Trans. Syst. Man Cybern. Syst.* **2024**, *54*, 3195–3206. [CrossRef]
25. Zhang, H.; Ma, H.; Mersha, B.W.; Zhang, X.; Jin, Y. Distributed cooperative search method for multi-UAV with unstable communications. *Appl. Soft Comput.* **2023**, *148*, 110592. [CrossRef]
26. Dou, W.; Yang, P.; Zhang, Z.; Wang, Z. Cooperative Multi-UAV Search for Prioritized Targets Under Constrained Communications. *Drones* **2025**, *9*, 855. [CrossRef]
27. Shafer, G. *A Mathematical Theory of Evidence*; Princeton University Press: Princeton, NJ, USA, 1976.
28. Sentz, K.; Ferson, S. *Combination of Evidence in Dempster-Shafer Theory*; Technical Report; Sandia National Laboratories: Albuquerque, NM, USA, 2002.
29. Sun, Y.; Kadota, I.; Talak, R.; Modiano, E. *Age of Information: A New Metric for Information Freshness*; Springer: Cham, Switzerland, 2020. [CrossRef]
30. Yates, R.D.; Sun, Y.; Brown, D.R.; Kaul, S.K.; Modiano, E.; Ulukus, S. Age of information: An introduction and survey. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 1183–1210. [CrossRef]
31. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347. [CrossRef]
32. Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24611–24624.
33. Rashid, T.; Samvelyan, M.; De Witt, C.S.; Farquhar, G.; Foerster, J.; Whiteson, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *J. Mach. Learn. Res.* **2020**, *21*, 7234–7284.
34. Miuccio, L.; Riolo, S.; Samarakoon, S.; Bennis, M.; Panno, D. On learning generalized wireless MAC communication protocols via a feasible multi-agent reinforcement learning framework. *IEEE Trans. Mach. Learn. Commun. Netw.* **2024**, *2*, 298–317. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.