Journal of Zhejiang University-SCIENCE A in press
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.A2600084

Strategic flight planning and conflict management for urban air mobility operations: a mission preference-constrained MARL approach

Yan LI[1,2], Xuejun ZHANG[1,2]✉, Chuxi WANG[2,3], Yan SHEN[1,2], Chenglong LI[1,4]

[1]*School of Electronic Information Engineering, Beihang University, Beijing 100191, China*
[2]*State Key Laboratory of CNS/ATM, Beihang University, Beijing 100191, China*
[3]*National Superior College for Engineers, Beihang University, Beijing 100191, China*
[4]*Flight Technology College, Civil Aviation Flight University of China, Chengdu 641419, China*

Abstract: As an emerging low-altitude transportation paradigm, urban air mobility (UAM) is envisioned to support high-density and demand-driven operations involving diverse and flexible mission requests. However, the imbalance between limited urban airspace resources and growing operational demands inevitably causes frequent flight conflicts, posing significant challenges to safe and efficient operations. To address this issue, this paper proposes a multiagent reinforcement learning (MARL) approach to achieve strategic four-dimensional trajectory (4DT) flight planning and conflict management during the preflight window. First, a collaborative optimization framework is established, in which the deconfliction problem is formulated as a multiagent Markov decision process (MAMDP) to enable coordinated decision-making. Then, a mission preference-constrained MARL method is developed by integrating two specialized mechanisms into the multiagent deep deterministic policy gradient (MADDPG) algorithm to address UAM operational characteristics. Specifically, an action masking for mission preference (AMMP) mechanism is implemented to ensure execution compliance, and a hierarchical prioritized experience replay (HPER) mechanism is designed to improve learning efficiency. Simulation results demonstrate that the proposed AMMP-HPER-MADDPG (AH-MADDPG) method achieves an average conflict resolution rate exceeding 96% and a preference awareness rate of 100% in scenarios involving 100 flight plans, significantly outperforming other methods. The proposed approach provides an effective and adaptive solution for ensuring operational safety, mission preference, and flight efficiency in future UAM operations.

Key words: Urban air mobility (UAM); Multiagent reinforcement learning (MARL); Four-dimensional trajectory (4DT); Flight plan; Conflict management; Mission preference

1 Introduction

Urban air mobility (UAM), as an emerging yet increasingly significant component of low-altitude transportation, is progressively becoming an innovative solution for reshaping urban transport networks and promoting the development of smart cities (Cohen et al., 2021; Li et al., 2025b). UAM aims to provide a diverse range of on-demand air transport services through the deployment of advanced vehicles such as unmanned aerial vehicles (UAVs), covering areas such as logistics delivery, emergency response, and security inspection (Mu et al., 2023; Wang et al., 2025a). It is foreseeable that future UAM systems will feature high-density and demand-driven operations within limited low-altitude airspace infrastructure, further complicated by dynamic flight intents and flexible mission requests (Chen et al., 2025b; Li et al., 2025a). Consequently, the key challenge for the development and widespread application of UAM lies in reconciling potential high-density flight demands with scarce airspace resources to enable the safe and efficient management of urban airspace.

In this context, establishing an advanced urban air traffic flow management framework serves as a

✉ Xuejun ZHANG, zhxj@buaa.edu.cn
Xuejun ZHANG, https://orcid.org/0000-0003-2711-5628

Received Feb. 5, 2026; Revision accepted June 16, 2026;
Crosschecked

critical prerequisite for supporting future UAM operations (Du et al., 2025). Drawing on conventional air traffic flow management, a centralized approach can be adopted for the strategic scheduling of four-dimensional trajectory (4DT) flight plans to achieve demand-capacity balancing and conflict management (Guo et al., 2024; Simorgh et al., 2024). However, distinct from the stable long-term advance scheduling of commercial airlines, UAM operations involve diverse and unscheduled flight intents from multiple operators, each characterized by heterogeneous priorities and specific mission preferences. This necessitates short-term strategic coordination within a preflight window to resolve potential conflicts among high-density missions and generate globally deconflicted 4DT flight plans for safe and efficient operations (Wu et al., 2022). Accordingly, this paper focuses on the collaborative 4DT flight planning problem from the perspective of the UAM service supplier (USS), aiming to achieve strategic deconfliction and mission preference compliance for high-density and demand-driven operations.

In recent years, the above problem has attracted increasing research attention, with most studies formulating it as a large-scale combinatorial optimization task. Existing methods typically adopt three widely recognized strategies, namely, takeoff time adjustment, flight speed adjustment, and local rerouting, using heuristic algorithms to solve the problem (Ho et al., 2022; Xie et al., 2024). Subsequent studies have further embedded such algorithms into double-layer or two-stage planning frameworks to improve strategy coordination and scheduling performance (Pang et al., 2022; Zhong et al., 2025a). However, despite these extensions, such optimization-based studies essentially rely on online iterative searches to address NP-hard planning problems, making them computationally expensive, susceptible to high-dimensional local optima under strong spatiotemporal coupling, and difficult to scale to unscheduled mission arrivals. More critically, existing studies primarily aim to optimize operational risk and time cost while giving limited attention to varying transport demands and heterogeneous mission preferences in future UAM operations. Although recent efforts (Wu et al., 2021; Du et al., 2025) have begun to incorporate mission priority and preference constraints into strategic planning, they generally treat these mission-related factors as soft weights in the objective function, which remains inadequate for practical operations where strict compliance may be needed. Therefore, existing studies still face challenges in fully supporting short-term strategic coordination for future UAM, where decision-making efficiency and operational compliance are paramount.

Deep reinforcement learning (DRL) provides a promising learning-based perspective for addressing the aforementioned challenges and has been widely applied in robot control (Zhang et al., 2025), autonomous driving (Feng et al., 2023; Hu et al., 2025), and traffic flow control (Wang et al., 2025b). Rather than providing a one-time solution for specific scenario instances, DRL learns generalizable and executable decision policies through interaction with the environment (Yang et al., 2024; Li et al., 2025c), offering adaptive policy learning and high-dimensional decision-making capabilities for complex and flexible UAM operations. Since the strategic flight planning problem considered in this paper involves multiple interacting UAVs, heterogeneous mission preferences, and shared airspace resource constraints, multiagent reinforcement learning (MARL) provides a suitable paradigm for modeling such cooperative decision-making (Brittain and Wei, 2022; Zhong et al., 2025b). As a representative actor-critic-based MARL algorithm, the multiagent deep deterministic policy gradient (MADDPG) (Lowe et al., 2017) is well aligned with this paradigm. Under the centralized training and decentralized execution (CTDE) architecture, MADDPG exploits global state information during training while allowing each agent to make decisions based on local observations during execution, thereby helping alleviate nonstationarity in multiagent environments and supporting continuous action spaces (Cai et al., 2024; Xing et al., 2026). Despite this potential, existing approaches remain limited for the considered problem in two main aspects. First, training stability and efficiency pose ongoing challenges, as recent prioritized experience replay works (Sun et al., 2022; Gök, 2024) have attempted to enhance sample utilization but have not fully considered reward-source heterogeneity in multiobjective tasks. Second, most formulations

encode mission requirements as soft reward terms rather than hard operational constraints, making it difficult to guarantee strict compliance in practical UAM applications. These limitations motivate the development of a MARL-based method that integrates preference-constrained decision-making with enhanced experience utilization.

Motivated by the above problems, this paper develops a mission preference-constrained MARL approach for short-term strategic flight planning to support high-density and demand-driven UAM operations. Specifically, we establish a USS-coordinated optimization framework that integrates urban airspace modeling, heterogeneous mission constraints, and spatiotemporal conflict management. Building on this framework, our proposed method employs an MADDPG algorithm as the foundational architecture and achieves targeted enhancements through two designed mechanisms, including action masking for mission preference (AMMP) and hierarchical prioritized experience replay (HPER). The proposed AMMP-HPER-MADDPG (AH-MADDPG) method provides an adaptive optimization solution that jointly considers operational safety, mission preference, and flight efficiency. The main contributions are summarized as follows:

(1) A strategic flight planning framework is developed to support the integrated management process from submitted flight intents to optimized 4DT flight plans. Within this framework, the core coordination process is formulated as a multiagent Markov decision process (MAMDP), providing the decision-making basis for cooperative policy learning in preflight spatiotemporal deconfliction.

(2) A tailored MADDPG-based algorithm is proposed to account for UAM operational characteristics by integrating two specialized mechanisms. The AMMP mechanism enforces mission preferences as hard operational constraints via action masking, ensuring compliance with the diverse requirements of multiple operators. The HPER mechanism optimizes the critic training process through hierarchical experience storage and prioritized sampling, improving convergence in multiobjective environments.

(3) Multiple simulation experiments are conducted to systematically evaluate the AH-MADDPG method. The results demonstrate that our method significantly outperforms baseline algorithms in conflict resolution and preference awareness while also exhibiting superior scalability and applicability.

The remainder of this paper is organized as follows. Section 2 formulates the strategic flight planning problem. Section 3 provides a detailed description of the proposed AH-MADDPG method. Section 4 presents simulation experiments and discusses the results. Finally, Section 5 concludes the paper.

2 Problem formulation

2.1 Optimization framework and assumptions

This section presents a USS-coordinated 4DT flight planning framework for strategic conflict management in UAM operations. As shown in Fig. 1, the optimization framework consists of two connected stages. In the conflict formulation and detection stage, the urban airspace is represented as a grid-based environment with static operational attributes, heterogeneous mission intents are transformed into initial flight plans, and potential conflicts are identified according to airspace capacity and operational feasibility constraints. This stage provides the learning environment and initial state required for the subsequent MARL-based optimization. In the conflict resolution and optimization stage, each UAV is treated as an agent, and the problem is formulated as an MAMDP with tailored state, action, and reward designs. The proposed AH-MADDPG algorithm then learns coordinated adjustment policies to resolve conflicts while respecting mission preferences, ultimately generating optimized 4DT flight plans.

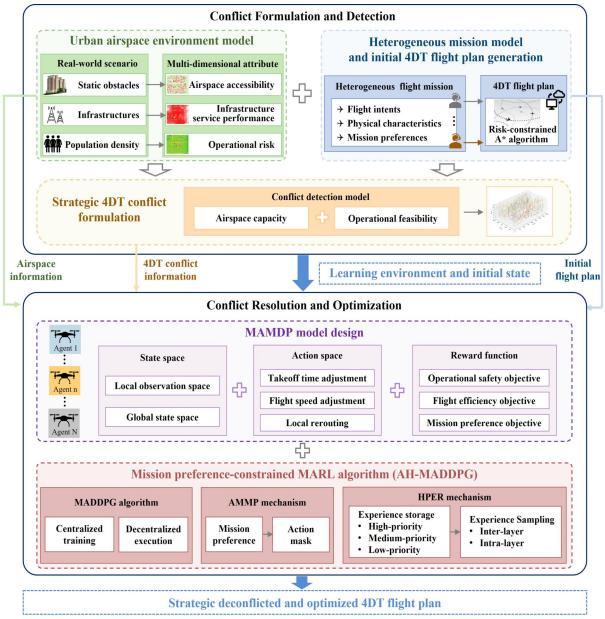


Fig. 1 Strategic 4DT flight planning framework

To further define the problem scope and support the following formulation, the main assumptions are summarized as follows:

(1) USS-based centralized coordination: The UAM-level system provides regulatory constraints and coordination rules, while the USS performs strategic flight planning and conflict management based on airspace environment data and all submitted mission information. Collaborative protocols between the USS and operators are assumed to specify mission preference attributes and derive initial trajectories, thereby transforming submitted intents into standardized planning inputs.

(2) Spatiotemporal resource allocation: A 4DT flight plan is represented as an ordered sequence of three-dimensional (3D) airspace grids with corresponding occupancy times. This paper focuses on macrolevel spatiotemporal resource allocation rather than detailed vehicle dynamics within each airspace grid, while operational feasibility is considered through simplified performance parameters.

(3) Operational uncertainty limitation: Since this paper focuses primarily on the preflight stage, operational uncertainties caused by meteorological disturbances or execution deviations are not considered, which are typically handled during tactical phases.

2.2 Urban airspace environment model

Digital airspace modeling provides the spatial foundation for flight planning in complex urban environments (Bauranov and Rakas, 2021; Chen et al., 2025b). Building on this, we discretize continuous 3D urban airspace into a finite set of uniform airspace grid cells, defined as

$$G=\left\{g_{x,y,z}\,\middle|\, x\in[1,N_X],y\in[1,N_Y],z\in[1,N_Z]\right\}, \quad (1)$$

where $g_{x,y,z}$ denotes the grid cell indexed by (x,y,z), and N_X, N_Y and N_Z are the number of grids along the three axes. On this basis, each grid is further assigned a multidimensional attribute parameter tuple $\mathcal{B}(g)=\left(A_{\text{acce}}(g),I_{\text{cns}}(g),R_{\text{risk}}(g)\right)$ to characterize its inherent operational conditions. Taking a typical operational scenario in Shenzhen, China, as an example, Fig. 2 illustrates the spatial distribution of these grid attributes.

Specifically, the airspace accessibility attribute $A_{\text{acce}}(g)$ indicates whether the grid cell is available for UAV operations and affects airspace capacity evaluation, with its value obtained using the airspace topology method based on the Alpha-Shape algorithm (Chen et al., 2026a). The infrastructure service performance attribute $I_{\text{cns}}(g)$ characterizes the local communication, navigation, and surveillance (CNS) service level and supports operational feasibility matching and is treated as a spatially distributed static parameter during strategic planning. The operational risk attribute $R_{\text{risk}}(g)$ represents the ground risk associated with population exposure and public safety and is determined from a grid-based ground risk map following the modeling approach in our earlier work (Zhu et al., 2024). Together, these static grid attributes provide environmental constraints for subsequent planning and optimization.

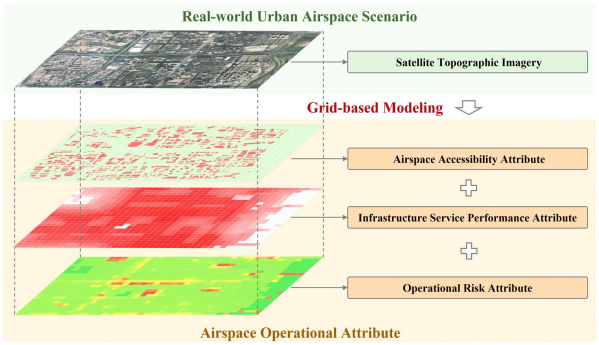


Fig. 2 Grid-based urban airspace representation

2.3 Heterogeneous mission model and initial 4DT flight plan generation

During the strategic stage, initial heterogeneous flight missions for all UAVs are submitted via their respective operators, typically comprising the flight intents, the UAV's inherent physical characteristics, and the mission preference details. The submitted missions are denoted by $\mathcal{M}$, and each mission $M_n \in \mathcal{M}$ is defined as a tuple

$$M_n = \left(U_n, \left(g^n_{\text{start}}, t^n_{\text{start}}\right), \left(g^n_{\text{goal}}, t^n_{\text{goal}}\right), V_n, L^n_{\text{cns}}, L^n_{\text{safe}}, P_n\right). \tag{2}$$

Here, U_n denotes the mission number of UAV n, while $\left(g^n_{\text{start}}, t^n_{\text{start}}\right)$ and $\left(g^n_{\text{goal}}, t^n_{\text{goal}}\right)$ represent the 4DT grid information of the starting point and goal point, respectively. $V_n = \left[v^n_{\min}, v^n_{\max}\right]$ specifies the UAV flight speed range. L^n_{cns} stands for onboard CNS service performance level inherent to the UAV itself, and L^n_{safe} expresses the airborne safety capability. These two parameters collectively reflect the core technical capability of the UAV to ensure its own autonomous and safe flight. The most crucial parameter P_n is the mission preference attribute, embodying the core heterogeneity constraint in this paper. Combining the three common scheduling strategies shown in Fig. 3, namely, (a) takeoff time adjustment, (b) flight speed adjustment, and (c) local rerouting, P_n categorizes mission preferences according to the disallowed strategies. It is defined as

$$P_n = \begin{cases} P_1, & \text{if strategy (c) or (a) is not allowed} \\ P_2, & \text{if strategy (c) is not allowed} \\ P_3, & \text{if strategy (a) is not allowed} \\ P_4, & \text{else} \end{cases}. \tag{3}$$

These mission preference categories reflect operator-specific requirements, improve the acceptability of USS-coordinated scheduling solutions in practical UAM operations, and later constrain the feasible action set of each UAV agent. For example, law enforcement, inspection, logistics delivery, and normal missions may correspond to different prohibited adjustment strategies.

Furthermore, under collaborative protocols with operators, the USS transforms the submitted mission intent $\mathcal{M}$ into a complete set of initial 4DT flight plans $\mathcal{F}_{\text{init}} = \left\{FP_n\right\}_{n=1}^{N}$, where FP_n denotes the initial plan for M_n, and N is the number of submitted missions. Each plan contains an ordered sequence of 4DT waypoints $Path_n = \left\{\left(g^n_a, t^n_a\right)\right\}_{a=0}^{N^n_P}$, where t^n_a indicates the estimated arrival time of M_n at the 3D

waypoint g_a^n, and N_P^n is the total number of waypoints. In this paper, the initial 4DT waypoints $Path_n$ are generated using the risk-constrained A* (RC-A*) algorithm as the baseline trajectory derivation method. The specific cost modeling and path planning formulation follow our previous work (Zhu et al., 2025), which jointly considers UAV performance, operational risk, and urban flight constraints.

2.4 Strategic 4DT conflict formulation

As outlined in the preceding section, an initial set of 4DT flight plans $\mathcal{F}_{\text{init}}$ has been generated for all submitted missions. Although these plans are individually feasible, limited airspace resources may still lead to global conflicts. Therefore, the USS must conduct a global spatiotemporal analysis of the initial plans and identify all potential conflicts to enable subsequent coordination. A conflict at the spatiotemporal point (g,t) is represented by a Boolean indicator $Conf(g,t)$, which is set to 1 if the airspace grid state violates the prescribed operational rules. This condition is expressed as

$$\begin{aligned} Conf(g,t) \equiv & \left(|O(g,t)| > C_{\text{cap}}(g) \right) \\ & \vee \left(\exists n \in O(g,t) \text{ s.t. Match}(U_n, g) = \text{False} \right). \end{aligned} \tag{4}$$

The first term represents an airspace capacity conflict, which occurs when $|O(g,t)|$, the number of UAVs occupying a specific airspace grid g at time t, exceeds its capacity $C_{\text{cap}}(g)$. The capacity is interpreted as the maximum number of UAVs allowed to operate simultaneously within the grid cell and is calculated as

$$C_{\text{cap}}(g) = \left[\frac{V_g A_{\text{acce}}(g)}{V_{\text{uav}}} \right] I_{\text{cns}}(g), \tag{5}$$

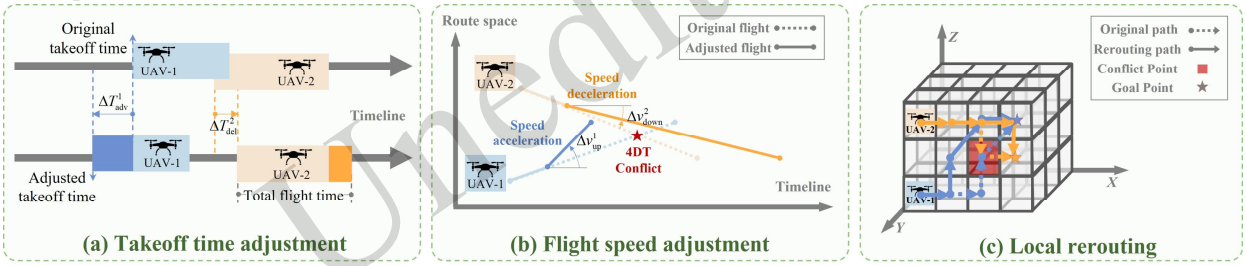


Fig. 3 Three common optimization scheduling strategies

where V_g is the physical volume of grid g, and V_{uav} denotes the safety envelope volume required for an individual UAV, derived from its ellipsoid collision zones together with the prescribed horizontal and vertical safety separation minima between UAVs (Zou et al., 2021).

The second term in Eq. (4) describes an operational feasibility conflict, which occurs when a UAV entering a grid cell fails to meet the minimum capability level prescribed for that grid. Let $O(g,t) = \{U_n \mid (g,t) \in Path_n\}$ denote the set of UAVs. Then, an operational feasibility violation exists when

$$\begin{aligned} & \text{Match}(U_n, g) = \text{False}, \\ & \quad \text{if } \left(L_{\text{cns}}^n < L_{\text{cns}}(g) \right) \text{or } \left(L_{\text{safe}}^n < L_{\text{safe}}(g) \right). \end{aligned} \tag{6}$$

Here, $L_{\text{cns}}(g)$ and $L_{\text{safe}}(g)$ represent the minimum onboard CNS and safety capability levels required for any UAV entering the grid cell, as imposed by that grid (Chen et al., 2026b).

2.5 Optimization problem formulation

Based on the grid-based airspace representation, heterogeneous mission preferences, and strategic conflict detection defined above, our optimization objective is to transform the initial flight plan set $\mathcal{F}_{\text{init}}$ into an optimized plan set $\mathcal{F}_{\text{final}}$. This set of flight plans must satisfy all operational constraints while minimizing the global comprehensive cost $J(\mathcal{F})$ of multi-UAV flight planning, which comprises operational safety cost C_{safe}, flight efficiency cost C_{effic}, and mission preference cost C_{pref}. Therefore, the strategic 4DT conflict resolution problem is formulated as

$$\min_{\mathcal{F}_{\text{final}}} J(\mathcal{F}) = f\left(C_{\text{safe}}, C_{\text{effic}}, C_{\text{pref}}\right), \text{ s.t. } \mathcal{F}_{\text{final}} \in \Omega_{\text{c}}. \quad (7)$$

Here, Ω_{c} denotes the constraint space to be satisfied, including UAV performance limitations, airspace operational conditions, and mission requirements.

The resulting problem is a multiobjective combinatorial optimization problem involving sequential interactions among multiple UAVs sharing limited airspace resources. The adjustment action of one UAV changes the spatiotemporal occupancy of grid cells and may affect the feasible decisions and future states of other UAVs. In addition, heterogeneous mission preferences impose different action constraints on UAV agents during conflict resolution. Therefore, the problem is naturally formulated as an MAMDP and is then addressed by learning a coordinated policy using the mission preference-constrained MARL method, as detailed in the next section.

3 Algorithm implementation

3.1 MADDPG algorithm

Owing to its inherent advantages in multiagent coordination, this paper adopts MADDPG as the foundational MARL algorithm under the CTDE framework. In our setting, each UAV agent corresponds to a submitted flight mission, and the learned actor policy generates strategic flight plan adjustment actions. At the planning stage, the centralized critic uses complete global information at the USS-coordinated strategic planning layer, while the actor policies are applied to individual missions in a decentralized manner during plan adjustment. In the basic learning process, each actor outputs a deterministic action with exploration noise, and the joint action is executed in the environment to obtain the global reward and the next state. The resulting transition is then stored in the replay buffer and used to update the actor-critic networks. The standard MADDPG update rules and CTDE schematic are given in Section S1 and Fig. S1 of the electronic supplementary materials (ESM). This CTDE setting enables scalable policy execution without the computational burden of fully centralized online optimization, thereby making it suitable for high-density and demand-driven UAM operations.

3.2 MAMDP model

3.2.1 State space

Under the USS-coordinated strategic planning setting, the local observation $\boldsymbol{o}_n^t \in \mathcal{O}_n$ of agent n at decision step t is constructed by the centralized planning environment based on all submitted or updated 4DT flight plans so that neighboring information is consistently derived from the shared environment. Specifically, the local observation is composed of four fixed-dimensional vectors or tensors, denoted as

$$\boldsymbol{o}_n^t = \left[\boldsymbol{s}_{\text{self}}^{n,t}, \boldsymbol{s}_{\text{mission}}^{n,t}, \boldsymbol{s}_{\text{conf}}^{n,t}, \text{Flatten}\left(\mathbf{S}_{\text{grid}}^{n,t}\right)\right]. \quad (8)$$

(1) Self-estimated state: This term $\boldsymbol{s}_{\text{self}}^{n,t}$ describes the estimated operational status of each agent at the current decision step t. It includes the estimated 3D grid $g_{\text{curr}}^n(t)$, goal grid g_{goal}^n, current flight speed $v_{\text{curr}}^{n,t}$, prescribed speed range V_n, temporal mapping information $\left(T^t, T_{\text{init}}^n\left(g_{\text{curr}}^n\right)\right)$, and onboard capability and safety levels $\left(L_{\text{cns}}^n, L_{\text{safe}}^n\right)$. These variables characterize the flight mission from spatial, temporal, kinematic, and operational perspectives.

(2) Mission preference state: This state is expressed as $\boldsymbol{s}_{\text{mission}}^{n,t} = P_n$, which encodes the hard constraints on the mission preferences that the agent must comply with, as specified in Eq. (3).

(3) Conflict awareness state: We explicitly provide a conflict-aware feature describing the nearest predicted conflict along the current plan, denoted as $\boldsymbol{s}_{\text{conf}}^{n,t} = \left[I_{\text{conf}}^n, t_{\text{toconf}}, d_{\text{toconf}}, \eta_{\text{conf}}\right]$, where the four elements represent the conflict indicator, temporal proximity, spatial proximity, and congestion severity factor.

(4) Neighboring grid perception state: This term describes the local airspace environment around the current estimated grid of each agent. It captures capacity, occupancy, and constraint-related information of nearby grid cells and is represented as a feature tensor $\mathbf{S}_{\text{grid}}^{n,t} \in \mathbb{R}^{3\times3\times3\times C}$, where C denotes the

number of feature channels. This tensor representation preserves the spatial topology of the local airspace, thereby allowing the actor to perceive local traffic density and operational constraints.

Then, the global state space $\mathcal{S}=\{\mathcal{O}_1,\cdots,\mathcal{O}_n,\cdots,\mathcal{O}_N\}$ incorporates the local observation information from all the agents, forming an omniscient view of the environment.

3.2.2 Action space

The action $\boldsymbol{a}_n^t \in \mathcal{A}$ of agent n is defined as a continuous vector associated with the three core adjustment strategies illustrated in Fig. 3, expressed as

$$\boldsymbol{a}_n^t=\left[a_n^{\text{time}},a_n^{\text{speed}},\boldsymbol{a}_n^{\text{reroute}}\right]. \tag{9}$$

Specifically, the takeoff time adjustment action a_n^{time} represents the overall takeoff time shift applied to the initial 4DT flight plan. The flight speed adjustment action a_n^{speed} modifies the current speed within the prescribed physical bounds. The local rerouting action $\boldsymbol{a}_n^{\text{reroute}}$ generates a bounded 3D displacement for strategic trajectory adjustment rather than low-level flight control. This design is more suitable for multirotor platforms with flexible maneuverability, while fixed-wing UAVs require a more constrained action design.

In the algorithm implementation, agents with heterogeneous mission preferences employ different subsets of available strategies during actual execution. Nevertheless, a unified and fixed-dimensional continuous action space is adopted for all agents to maintain a consistent actor network structure and improve parameter reusability. Accordingly, at each decision step, the actor network first outputs a raw action vector $\boldsymbol{a}_{\text{raw},n}^t$, which represents preliminary adjustment intents. Due to mission preference constraints, these intents must be filtered by the AMMP mechanism (detailed in Section 3.3.1) and subsequently mapped to a valid executable action $\boldsymbol{a}_n^t$. The detailed formulation of the executable action components is given in Section S2 of the ESM, including raw action normalization and the specific scaling process for each component.

3.2.3 Reward function

In accordance with the global optimization objectives in Eq. (7), we design a shared global reward function r^t consisting of five components:

(1) Conflict penalty reward: This term is designed to minimize the number of 4DT conflicts $N_{\text{conf}}\left(\boldsymbol{s}_t'\right)$ in the next state, defined as

$$r_{\text{conf}}^t=-N_{\text{conf}}\left(\boldsymbol{s}_t'\right)=-N\left[Conf\left(g,t'\right)\right]. \tag{10}$$

(2) Safety risk reward: This component encourages UAVs to avoid high-risk areas through the accumulation of the ground risk values of grid cells traversed by the updated flight plans, computed as

$$r_{\text{risk}}^t=-Risk_{PATH}=-\sum_{n=1}^{N}\sum_{(g_a,t_a)\in Path_n^{t'}}R_{\text{risk}}\left(g_a\right). \tag{11}$$

(3) Delay cost reward: The penalty is imposed when the actual flight plans deviate from the original takeoff or arrival timestamps to minimize deviations introduced by schedule adjustments, formulated as

$$r_{\text{delay}}^t=-T_{\text{delay}}=-\sum_{n=1}^{N}\left(\left|t_{\text{acst}}^n-t_{\text{start}}^n\right|+\left|t_{\text{acgo}}^n-t_{\text{goal}}^n\right|\right), \tag{12}$$

where T_{delay} is the total delay time, while t_{acst}^n and t_{acgo}^n denote the takeoff and arrival timestamps under the optimized actual 4DT flight plan, respectively.

(4) Efficiency cost reward: This reward improves airspace utilization by encouraging all flight plans to be completed with minimal total flight time T_{flight}, denoted as

$$r_{\text{effic}}^t=-T_{\text{flight}}=-\sum_{n=1}^{N}\left(t_{\text{acgo}}^n-t_{\text{acst}}^n\right). \tag{13}$$

(5) Mission preference violation penalty reward: This term penalizes raw actions that violate the assigned mission preferences. Let E_{time} and E_{reroute} denote violation events triggered by disallowed takeoff time adjustment and local rerouting actions, respectively. The individual penalty is defined as

$$r_{\text{pref},n}^{t}=\begin{cases}\varphi_{\text{pref}\,2}, & \text{if } E_{\text{time}} \text{ and } E_{\text{reroute}}\\ \varphi_{\text{pref}\,1}, & \text{if } E_{\text{time}} \text{ or } E_{\text{reroute}}\\ 0, & \text{else}\end{cases}, \tag{14}$$

where $\varphi_{\text{pref}\,2}>\varphi_{\text{pref}\,1}>0$ are positive penalty values. The total preference violation penalties of all flight plans can be obtained by $r_{\text{pref}}^{t}=-\sum_{n=1}^{N} r_{\text{pref},n}^{t}$.

Finally, the global reward function is derived as

$$r^{t}=\omega_{\text{c}} r_{\text{conf}}^{t}+\omega_{\text{r}} r_{\text{risk}}^{t}+\omega_{\text{d}} r_{\text{delay}}^{t}+\omega_{\text{e}} r_{\text{effic}}^{t}+\omega_{\text{p}} r_{\text{pref}}^{t}, \tag{15}$$

where ω_{c}, ω_{r}, ω_{d}, ω_{e}, and ω_{p} are nonnegative weight coefficients. In this paper, these coefficients are empirically calibrated to balance reward scales and prioritize operational objectives. In particular, r_{conf}^{t} and r_{pref}^{t} are assigned relatively dominant weights to emphasize strategic deconfliction and preference compliance, while the remaining terms encourage lower ground risk and higher flight efficiency. The robustness of this calibrated setting is further examined through a brief sensitivity analysis in Section S6.1 of the ESM.

3.3 Algorithm mechanism

3.3.1 AMMP mechanism

The above penalty r_{pref}^{t} provides soft guidance for preference compliance during learning. However, relying only on such trial-and-error learning driven by negative reward may still lead to inefficient exploration and preference-violating raw actions. To solve this issue, we further introduce a hard constraint mechanism AMMP, which imposes constraints on the policy prior to action execution. It filters the raw actor output according to the mission preference of each agent, ensuring that only preference-compliant adjustment strategies can be executed.

Specifically, the AMMP module generates an action mask vector $\boldsymbol{AM}_{n}^{t}$ based on P_n, denoted as

$$\boldsymbol{AM}_{n}^{t}=\left[m^{\text{time}}, m^{\text{speed}}, \boldsymbol{m}^{\text{reroute}}\right], \tag{16}$$

where m^{time}, m^{speed}, and $\boldsymbol{m}^{\text{reroute}}$ correspond to the masks for takeoff time adjustment, flight speed adjustment, and local rerouting, respectively. A mask value of 1 indicates that the corresponding strategy is allowed, while a value of 0 means that it is prohibited. Then, the masked action vector $\boldsymbol{a}_{\text{mraw},n}^{t}$ is obtained through the Hadamard product of the original vector and the mask vector, given by

$$\boldsymbol{a}_{\text{mraw},n}^{t}=\boldsymbol{a}_{\text{raw},n}^{t}\odot\boldsymbol{AM}_{n}^{t}. \tag{17}$$

The resulting masked action vector serves as the intermediate input for the subsequent scaling process to obtain the final executable action $\boldsymbol{a}_{n}^{t}$.

Through this mechanism, mission preference constraints are enforced before interaction with the environment, while the reward function still provides value guidance during policy learning. This dual design improves preference compliance and helps reduce ineffective exploration.

3.3.2 HPER mechanism

The standard experience replay mechanism used in MADDPG samples transitions uniformly and therefore ignores the varying importance of experience samples for learning. To address this limitation, we develop a more efficient HPER mechanism based on our previous work (Shen et al., 2026) and further refine it here for the collaborative multiobjective flight planning problem. HPER categorizes transitions $e^{j}=\left(\boldsymbol{s}^{j},\boldsymbol{a}^{j},r^{j},\boldsymbol{s}'^{j}\right)$ into three independent experience replay buffers with hierarchical priority levels based on their contribution to strategy optimization. By organizing and sampling experiences across different layers, HPER allows the learning process to focus more on critical and informative transitions while preserving exposure to regular experiences, thereby improving training efficiency and policy performance. The resulting three-layer replay buffer structure is denoted as

$$\mathcal{R}=\left\{\mathcal{R}_{\text{H}},\mathcal{R}_{\text{M}},\mathcal{R}_{\text{L}}\right\}. \tag{18}$$

The high-priority layer $\mathcal{R}_{\text{H}}$ stores the most critical transitions, mainly unresolved spatiotemporal conflicts and mission preference violations. The medium-priority layer $\mathcal{R}_{\text{M}}$ stores informative heuristic experiences, including successful terminal transitions and subcritical samples related to safety risk, delay, or

efficiency degradation. The low-priority layer $\mathcal{R}_L$ stores the remaining regular transitions to preserve sample diversity and support the learning of basic policies. Each layer has a fixed capacity and follows the first-in-first-out replacement rule when full.

During training, HPER performs two-stage sampling. First, interlayer sampling determines the number of transitions B_{layer} drawn from each of the three independent layers. Then, intralayer prioritized sampling selects transitions within each layer according to their temporal-difference errors, with importance sampling used to correct the bias introduced by nonuniform transition selection.

The detailed formulation of the HPER mechanism can be found in Section S3 of the ESM, which includes layer assignment and priority-based sampling rules.

3.4 Overall AH-MADDPG algorithm architecture

Based on the above sections, the mission preference-constrained MARL algorithm AH-MADDPG is proposed for collaborative 4DT flight planning and strategic deconfliction. The overall architecture is shown in Fig. 4, and the complete pseudocode is provided in Table S1 of the ESM. The training process enables preference-aware action execution through the AMMP mechanism, priority-guided experience utilization via the HPER mechanism, and stable multiagent learning under the CTDE framework. Overall, AH-MADDPG is designed to accelerate convergence and strengthen policy compliance in multiconstraint and multiobjective environments.

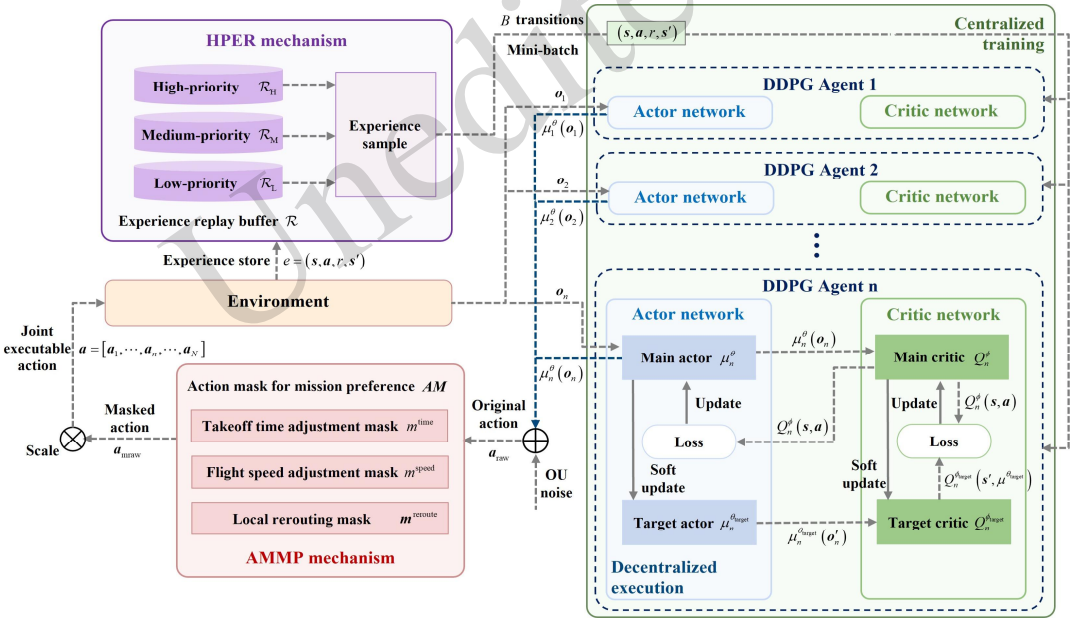


Fig. 4 Architecture and data flow of the AH-MADDPG algorithm

4 Simulation results and discussion

4.1 Environment setup

The simulation experiments are conducted in a representative urban airspace extracted from Nanshan District, Shenzhen, China, with a size of 4.0 km × 2.5 km × 120 m. Shenzhen is selected because it is one of the first pilot cities for low-altitude economic development in China and provides a realistic urban environment for evaluating the proposed method. Based on the grid-based modeling method described in Section 2.2, the selected airspace is discretized into standard 3D grid cells of 100 m × 100 m × 20 m for attribute mapping. The airspace capacity of each grid cell is calculated using Eq. (5), and the resulting capacity distribution is shown in Fig. 5. All algorithms are trained and evaluated under identical simulation settings to ensure fairness and reproducibility. The main parameter settings, including UAV performance, reward coefficients, training hyperparameters, and hardware configuration, are provided in Section S5 and Table

S2 of the ESM.

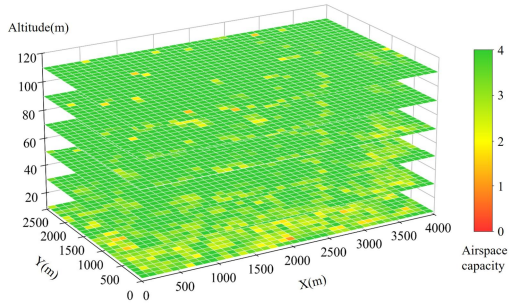


Fig. 5 Capacity distribution of the simulated airspace

4.2 Training performance analysis

To validate learning stability in uncertain conditions, we conduct training under highly randomized mixed simulation environments. A scenario library is constructed in advance, with each scenario designed to simulate 100 initial 4DT flight missions randomly scheduled for takeoff between 09:00 and 10:00. The initial flight plans are generated using the RC-A* method, which produces trajectories that are individually feasible and near-optimal for each UAV but may still exhibit conflicts from the USS perspective. In each episode of the training phase, a scenario is randomly selected from this library, and the mission preference proportion is also randomly set. Based on this setup, we establish a comprehensive benchmarking framework involving six distinct algorithms to evaluate the proposed method against both internal ablation variants and external MARL baselines. Specifically, the comparative variants include AMMP-MADDPG and HPER-MADDPG, while the representative MARL baselines consist of the multiagent proximal policy optimization (MAPPO), the multiagent twin delayed deep deterministic policy gradient (MATD3), and the traditional MADDPG. All algorithms are trained under identical environmental settings and unified neural network hyperparameters to ensure fairness.

Fig. 6 illustrates the resulting training process, where solid lines depict average rewards, and shaded areas represent variability across multiple independent runs. MADDPG exhibits the slowest reward growth, the most severe oscillations, and the lowest final reward, indicating limited learning efficiency in complex multiconstraint environments. MAPPO and MATD3 improve training stability compared with MADDPG, but their final rewards remain lower than those of the proposed method and its variants, mainly because soft penalty-based constraint handling still permits residual preference violations and limits the global reward. For the ablation variants, AMMP-MADDPG achieves higher reward levels during the early training stages by filtering infeasible actions, whereas HPER-MADDPG accelerates convergence in the middle and later stages by improving the utilization of critical samples. By integrating both mechanisms, AH-MADDPG demonstrates superior stability and minimal fluctuation throughout the training process, while converging to the highest cumulative reward of approximately -310 after approximately 19,200 episodes.

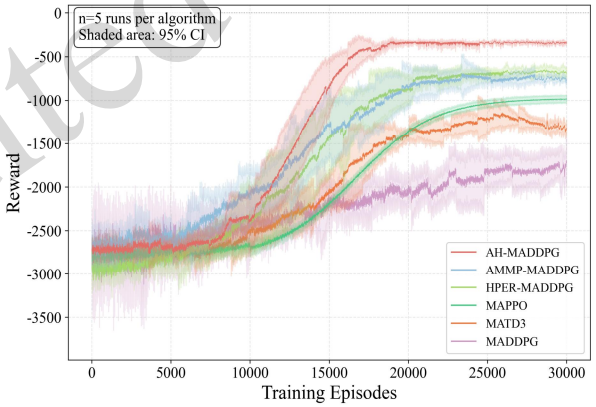


Fig. 6 Comparison of reward curves across six algorithms

4.3 Results and mechanism analysis

4.3.1 Conflict optimization result

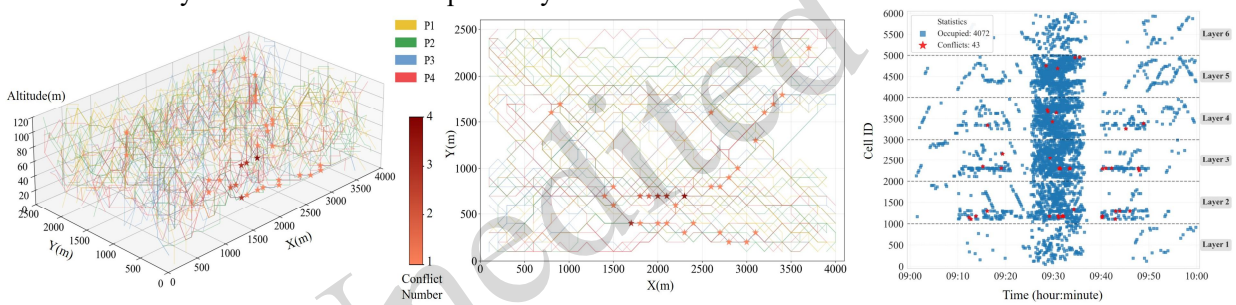
To evaluate the algorithm's effectiveness under demand-driven operations and heterogeneous mission requirements, five representative experimental scenarios are constructed. Each contains 100 randomly generated initial flight sets, while the proportions of mission preferences are systematically varied. The specific preference constraints and initial number of conflicts are provided in Table 1. These scenarios cover different levels of constraint coupling and optimization difficulty, ranging from balanced preferences in Scenario 1 to strongly constrained or relatively flexible settings in other scenarios.

Table 1 Number of missions and conflicts in five scenarios

Scenario No.	P1	P2	P3	P4	Conflicts
No.1 (balance)	25	25	25	25	43

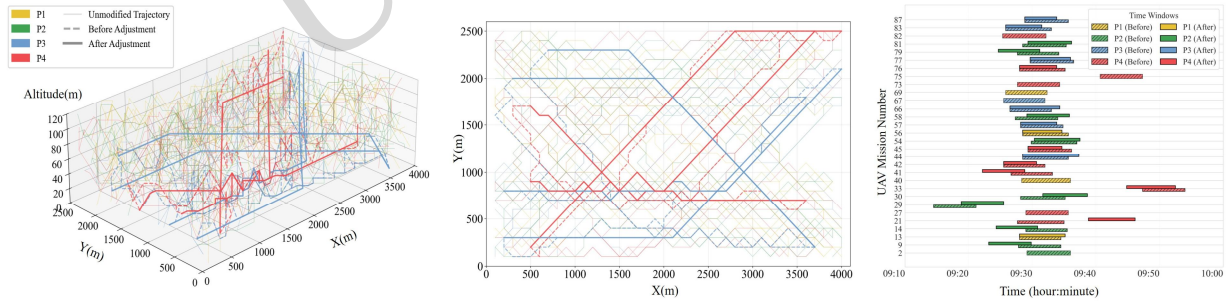
No.2 (P4 dominated)	10	10	10	70	41
No.3 (P3 dominated)	10	10	70	10	46
No.4 (P2 dominated)	10	70	10	10	42
No.5 (P1 dominated)	70	10	10	10	54

Taking Scenario 1 as an example, the initial conflict detection results are shown in Fig. 7. The conflict waypoints are distributed across the airspace but are mainly clustered around route intersections and high-density sectors. From the spatiotemporal perspective, the conflicts are concentrated primarily between approximately 09:25 and 09:35, when dense trajectory occupancy leads to overlapping traffic peaks. In addition, most conflicts occur between the 20 m and 60 m altitude ranges, corresponding to the most intensively utilized urban airspace layers. Similar conflict patterns are observed in Scenarios 2–5, as shown in Fig. S2 of the ESM. Since mission preference attributes are not incorporated during initial trajectory generation, these results indicate that the initial conflicts are mainly caused by structural airspace limitations and stochastic trajectory interactions, thereby ensuring the comparability of the five scenarios. After applying AH-MADDPG, Fig. 8 shows the optimization results for Scenario 1. The comparison indicates that all 43 initial conflicts are eliminated through coordinated spatiotemporal modifications without violating mission preference constraints. This demonstrates the capability of the proposed method to generate strategic deconflicted and optimized 4DT flight plans.



(a) Spatial distribution (3D view and top view) (b) Spatiotemporal occupancy

Fig. 7 Initial 4DT flight conflict detection results in Scenario 1 (43 conflicts)



(a) Flight path changes (3D view and top view) (b) Flight time changes

Fig. 8 Illustration results of conflict resolution in Scenario 1 (0 conflict)

Fig. 9 further quantifies the distribution of adopted strategies, revealing the adaptive optimization capability of our method under heterogeneous mission constraints. Flight speed adjustment maintains a relatively high utilization frequency in all scenarios, which is consistent with its role as a generally available strategy. Scenarios 3 and 4 exhibit clear spatiotemporal complementarity, where local rerouting or takeoff time adjustment is preferentially selected to address dominant constraint dimensions. In Scenario 5, where both takeoff time and rerouting strategies are severely restricted, the optimization burden is mainly shifted to speed adjustment and the coordinated actions of less constrained missions. In contrast, Scenarios 1 and 2 show more balanced strategy distributions, reflecting cost-oriented global optimization under evenly distributed or relatively relaxed constraints. These results indicate that the AMMP mechanism can guide the policy to select feasible adjustment strategies according to heterogeneous mission preference constraints.

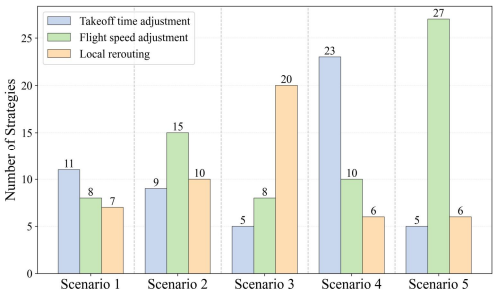


Fig. 9 Distribution of strategies adopted in five scenarios

4.3.2 Ablation study

In this section, we conduct an ablation study to quantitatively evaluate the final optimization quality and to validate the contributions of each mechanism. To this end, several metrics are introduced to assess the model across three objectives: a) For operational safety, the number of 4DT conflicts (NOC) and the conflict resolution rate (CRR) jointly measure conflict occurrences and the algorithm's effectiveness in resolving them; b) For mission preference, the number of preference violations (NOPV) quantifies the preference-violation behavior performed by the final executed actions, and the preference awareness rate (PAR) indicates the overall ratio of preference compliance; c) For flight efficiency, the number of delays (NOD) counts all the occurrences where the incurred delay exceeds the acceptable threshold, whereas the total delay time (TDT) and total flight time (TFT) provide measures of overall delay magnitude and planning efficiency, respectively. The detailed results across five scenarios are reported in Table 2.

In terms of operational safety, AH-MADDPG achieves the best overall robustness, resolving all conflicts in Scenarios 1–4 and maintaining a CRR of 96.3%, even in the most constrained Scenario 5. By comparison, the two variants show lower conflict resolution performance, indicating that relying on either AMMP or HPER alone is insufficient for stable conflict management. HPER-MADDPG generally achieves a higher CRR than AMMP-MADDPG, suggesting that priority-guided sample utilization is beneficial for learning conflict resolution policies.

Table 2 Comparison of algorithms in five scenarios

Scenario No.	Algorithm	Operational safety		Mission preference		Flight efficiency		
		NOC	CRR	NOPV	PAR	NOD	TDT(s)	TFT(s)
Scenario 1	AH-MADDPG	0	**100%**	0	**100%**	1	3230	40300
	AMMP-MADDPG	4	90.7%	0	100%	5	11540	40820
	HPER-MADDPG	2	95.3%	8	73.3%	2	6430	41100
Scenario 2	AH-MADDPG	0	**100%**	0	**100%**	0	2810	40220
	AMMP-MADDPG	2	95.1%	0	100%	1	3800	40450
	HPER-MADDPG	1	97.6%	7	82.1%	1	3720	40680
Scenario 3	AH-MADDPG	0	**100%**	0	**100%**	0	1520	40280
	AMMP-MADDPG	6	87.0%	0	100%	2	4810	40680
	HPER-MADDPG	4	91.3%	11	69.4%	0	3090	40520
Scenario 4	AH-MADDPG	0	**100%**	0	**100%**	5	9890	41220
	AMMP-MADDPG	5	88.1%	0	100%	12	22460	42380
	HPER-MADDPG	3	92.9%	13	71.1%	3	9760	41680
Scenario 5	AH-MADDPG	2	**96.3%**	0	**100%**	3	6190	41610
	AMMP-MADDPG	10	81.5%	0	100%	10	19550	42150
	HPER-MADDPG	6	88.9%	15	65.9%	3	9510	42020

For mission preference, the results highlight the irreplaceable role of AMMP. Without action masking, HPER-MADDPG exhibits unstable preference compliance, with PAR decreasing to 65.9% in Scenario 5. Conversely, the other two algorithms consistently maintain a deterministic PAR of 100%, confirming AMMP as an effective hard constraint enforcement mechanism.

However, applying hard constraints without efficient exploration often comes at the cost of flight efficiency, leading to overly conservative strategies. This limitation is evident in AMMP-MADDPG, which incurs excessive delays in Scenario 4, resulting in a NOD of 12. This indicates that agents are forced to substantially adjust their takeoff time to avoid preference violations. HPER-MADDPG shows a lower delay in this case, even achieving an NOD of 3, which slightly outperforms our method. Nevertheless,

this represents a shortcut achieved by sacrificing preference constraints, rather than a truly effective solution. Overall, these results support that AH-MADDPG can better balance conflict resolution, preference awareness, and flight efficiency by integrating AMMP and HPER under complex conditions.

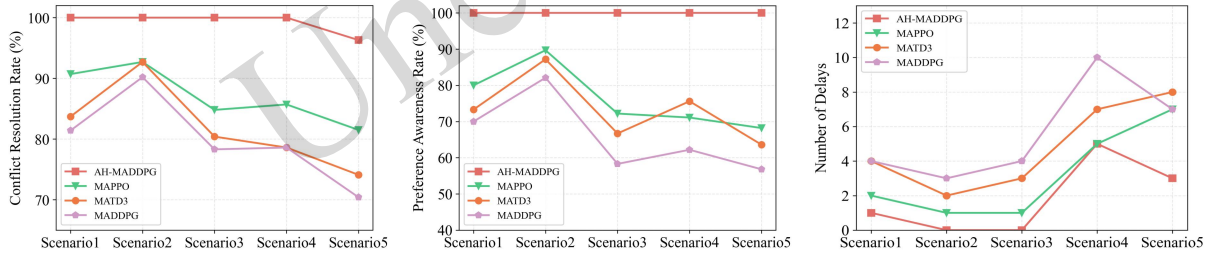
4.4 Method comparison

4.4.1 Comparison with MARL methods

We further compare AH-MADDPG with three representative MARL baselines, and the optimization results are shown in Fig. 10. AH-MADDPG demonstrates superior performance in terms of solution quality and robustness, effectively achieving conflict resolution and preference awareness across all scenarios while maintaining competitive flight efficiency. In contrast, the baseline MADDPG algorithm shows the lowest stability and optimization quality, with CRR decreasing to 70.4% and PAR dropping to 56.8% in Scenario 5. MAPPO and MATD3 achieve intermediate performance, performing well in simpler scenarios but showing residual conflicts and preference violations under complex constraints. These results indicate that soft penalty-based MARL methods are insufficient for this problem, further confirming the necessity of explicit hard constraint handling and efficient policy exploration.

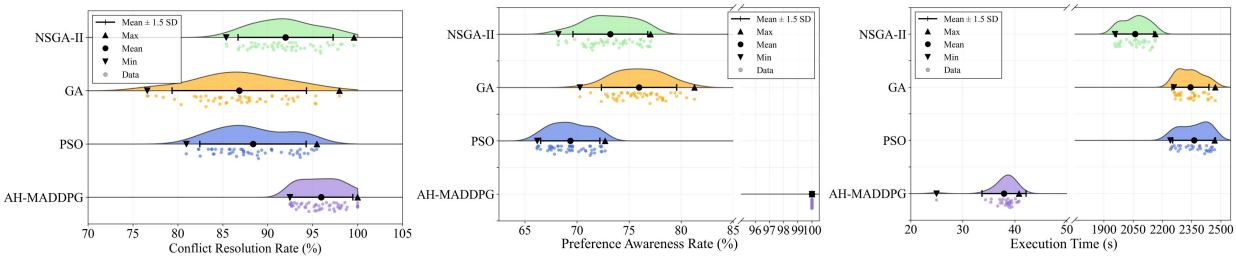
4.4.2 Comparison with heuristic methods

This section evaluates AH-MADDPG by extending the comparison to three heuristic algorithms, namely, the particle swarm optimization (PSO) algorithm, genetic algorithm (GA), and nondominated sorting genetic algorithm II (NSGA-II). The evaluation is conducted across 50 randomly generated heterogeneous flight scenarios, each comprising 100 initial flight missions with randomized preference distributions. The statistical results of the three key metrics are visualized in Fig. 11 using raincloud plots.



(a) Conflict resolution (b) Preference awareness (c) Delay result

Fig. 10 Comparison results with MARL algorithms



(a) Conflict resolution (b) Preference awareness (c) Execution time

Fig. 11 Comparison results with heuristic methods

As shown in Fig. 11a, AH-MADDPG achieves the highest median CRR of nearly 96% with the most concentrated distribution, indicating superior conflict resolution effectiveness and robustness. In contrast, all heuristic algorithms display larger performance variance and lower average CRR. Although NSGA-II performs relatively well due to its multiobjective optimization capability, it still shows a noticeable gap compared to our method. The results suggest that the proposed method, benefiting from centralized training and learned cooperative policies, enables more effective global coordination and multiobjective

balancing in strongly coupled and high-dimensional optimization problems.

In addition, considerable differences can also be found regarding preference awareness in Fig. 11b. AH-MADDPG maintains a deterministic PAR of 100% across all test scenarios, whereas the heuristic algorithms frequently fluctuate between 65% and 80%. This is mainly because standard heuristic methods typically rely on soft penalty-based objective functions, which may converge to pseudofeasible solutions that sacrifice preference compliance for conflict resolution. Strictly enforcing hard preference constraints in heuristic search would require complex initialization, repair, or decoding mechanisms, making this process difficult under coupled spatiotemporal and heterogeneous mission constraints.

Fig. 11c further compares the execution time, defined as the online solution time required to complete strategic 4DT flight planning. The results show an order of magnitude difference between the two categories of methods. This is because heuristic algorithms rely on computationally intensive online iterative searches, requiring a full-population fitness re-evaluation to accommodate even minor environmental or mission variations. By shifting most of the heavy computational burden to the offline training phase, our method supports rapid forward inference during online coordination. This capability provides practical support for USS-coordinated preflight planning by accommodating short-notice requests and reducing the required lead time for flight plan submission, thereby enhancing the flexibility and responsiveness of future on-demand UAM operations.

Additional analyses are given in the ESM. Section S6.1 reports a reward weight sensitivity analysis, and Section S6.2 performs a scalability test under varying flight densities to further evaluate the applicability of AH-MADDPG in high-density operations.

5 Conclusions

In this paper, we propose a mission preference-constrained MARL approach to effectively address strategic deconfliction through collaborative 4DT flight planning. We first introduce a comprehensive optimization framework for the USS-coordinated strategic planning layer, which integrates the complex urban airspace model, heterogeneous mission constraints, and spatiotemporal conflict management to enable centralized coordination of diverse mission intents during the preflight window. The deconfliction problem is then reasonably formulated as a cooperative MAMDP, and the AH-MADDPG method is designed for UAM operational characteristics to ensure stable and compliant decision-making in multiobjective and multiconstraint environments. Furthermore, a series of comparative experiments demonstrate that the integration of AMMP and HPER mechanisms enables the proposed method to achieve more effective conflict resolution and stricter preference awareness across various scenarios, outperforming other MARL baselines in solution quality while exhibiting superior adaptability compared with heuristic methods. These findings suggest that our AH-MADDPG method can significantly reduce the short-term planning burden while simultaneously optimizing operational safety, mission preference, and flight efficiency. Consequently, this approach offers a robust and scalable solution to support the broader deployment of future high-density and demand-driven UAM operations.

Despite these promising results, the current work still has several limitations that naturally outline our future research directions. First, as noted in the assumptions, this study focuses on preflight strategic coordination and macrolevel spatiotemporal resource allocation, while the practical transition from strategic planning to tactical execution has not yet been fully addressed. To bridge this gap, future work will incorporate UAV dynamics, flight profile constraints, and operational uncertainties to improve the executability of generated flight plans in tactical operations (Chen et al., 2025a). Second, the current MARL framework is mainly tailored for multirotor UAVs in urban low-altitude scenarios, and its applicability to more complex flight missions and heterogeneous UAV platforms remains limited. Therefore, we will refine its formulation through graph-based state representations (Zhang et al., 2023) and vehicle-specific action design (Yue et al., 2025)

to better capture complex interaction patterns and maneuverability constraints. Third, building on the current focus on low-altitude airspace planning, future work will extend our framework toward surface-air coordination scenarios (Du et al., 2024; Chai et al., 2024), where ground transportation networks and low-altitude airspace resources are jointly managed to support efficient interconnection and integrated multimodal operations. Ultimately, these extensions are expected to further enhance the practical applicability of our management framework, thereby contributing to the development of future smart cities.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62541101).

Author contributions

Yan LI designed the research, developed the methodology, and drafted the paper. Xuejun ZHANG provided supervision and funding acquisition. Chuxi WANG contributed to manuscript revision and validation. Yan SHEN discussed and revised the algorithms. Chenglong LI helped to organize the manuscript. All authors have read and approved the final version.

Conflict of interest

Yan LI, Xuejun ZHANG, Chuxi WANG, Yan SHEN, and Chenglong LI declare that they have no conflict of interest.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Bauranov A, Rakas J, 2021. Designing airspace for urban air mobility: a review of concepts and approaches. *Progress in Aerospace Sciences*, 125:100726. https://doi.org/10.1016/j.paerosci.2021.100726

Brittain M, Wei P, 2022. Scalable autonomous separation assurance with heterogeneous multi-agent reinforcement learning. *IEEE Transactions on Automation Science and Engineering*, 19(4):2837-2848. https://doi.org/10.1109/TASE.2022.3151607

Cai KQ, Li ZQ, Guo T, et al., 2024. Multiairport departure scheduling via multiagent reinforcement learning. *IEEE Intelligent Transportation Systems Magazine*, 16(2):102-116. https://doi.org/10.1109/MITS.2023.3307130

Chai RQ, Guo YL, Zuo ZY, et al., 2024. Cooperative motion planning and control for aerial-ground autonomous systems: methods and applications. *Progress in Aerospace Sciences*, 146:101005. https://doi.org/10.1016/j.paerosci.2024.101005

Chen SD, Zhang XJ, Zhang WD, 2026a. The construction and capacity evaluation of urban low-altitude reachable airspace. *Computer Engineering and Applications*, 62(4):335-343 (in Chinese). https://doi.org/10.3778/j.issn.1002-8331.2503-0046

Chen SD, Zhang XJ, Zhang ZY, et al., 2026b. Optimizing low-altitude urban airspace: identifying free route airspace and control points through spatial analysis. *International Journal of Aeronautical and Space Sciences*, 27:1606-1623. https://doi.org/10.1007/s42405-025-01053-y

Chen YT, Xu Y, Yang L, et al., 2025a. In-flight fast conflict-free trajectory re-planning considering UAV position uncertainty and energy consumption. *Transportation Research Part C: Emerging Technologies*, 171:104988. https://doi.org/10.1016/j.trc.2024.104988

Chen ZJ, Shum HY, Cao XB, et al., 2025b. Engineering and technology for low-altitude economy infrastructure. *Frontiers of Information Technology & Electronic Engineering*, 26(12):2393-2396. https://doi.org/10.1631/fitee.2530000

Cohen AP, Shaheen SA, Farrar EM, 2021. Urban air mobility: history, ecosystem, market potential, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(9):6074-6087. https://doi.org/10.1109/TITS.2021.3082767

Du RJ, Chen SK, Dong JQ, et al., 2024. Dynamic urban traffic rerouting with fog-cloud reinforcement learning. *Computer-Aided Civil and Infrastructure Engineering*, 39(6):793-813. https://doi.org/10.1111/mice.13115

Du S, Zhong G, Wang F, et al., 2025. A framework for collaborative UAM traffic flow optimization with mission preferences: incorporating customized strategy synergy into strategic conflict management. *Transportation Research Part E: Logistics and Transportation Review*, 202:104326. https://doi.org/10.1016/j.tre.2025.104326

Feng S, Sun HW, Yan XT, et al., 2023. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953):620-627. https://doi.org/10.1038/s41586-023-05732-2

Gök M, 2024. Dynamic path planning via dueling double deep Q-network (D3QN) with prioritized experience replay. *Applied Soft Computing*, 158:111503. https://doi.org/10.1016/j.asoc.2024.111503

Guo T, Mei Y, Tang K, et al., 2024. A knee-guided evolutionary algorithm for multi-objective air traffic flow management. *IEEE Transactions on Evolutionary Computation*, 28(4):994-1008. https://doi.org/10.1109/TEVC.2023.3281810

Ho F, Gonçalves A, Rigault B, et al., 2022. Multi-agent path finding in unmanned aircraft system traffic management with scheduling and speed variation. *IEEE Intelligent Transportation Systems Magazine*, 14(5):8-21. https://doi.org/10.1109/MITS.2021.3100062

Hu FQ, Fu XW, Huang HL, 2025. Safe reinforcement learning for event-triggered control of automated vehicles with uncertainty. *IEEE Transactions on Intelligent Transportation Systems*, 26(9):14039-14052. https://doi.org/10.1109/TITS.2025.3533578

Li SH, Zhang TF, Xiao YY, et al., 2025a. On-demand ridesharing based on dynamic scheduling in urban air mobility. *Transportation Research Part C: Emerging Technologies*, 175:105111. https://doi.org/10.1016/j.trc.2025.105111

Li YM, Guo T, Chen J, et al., 2025b. Urban air mobility: a review and challenges. *IEEE Intelligent Transportation Systems Magazine*, 17(3):67-87. https://doi.org/10.1109/MITS.2024.3496480

Li YM, Li JQ, Wang JJ, et al., 2025c. Multi-scale graph enhanced reinforcement learning for conflict resolution in dense UAV networks. *IEEE Internet of Things Journal*, 12(21):44290-44303. https://doi.org/10.1109/JIOT.2025.3608865

Lowe R, Wu Y, Tamar A, et al., 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. 31st Annual Conference on Neural Information Processing Systems (NeurIPS), p.6382-6393. https://doi.org/10.48550/arXiv.1706.02275

Mu ZH, Qin Y, Yu CC, et al., 2023. Adaptive cropping shallow attention network for defect detection of bridge girder steel using unmanned aerial vehicle images. *Journal of Zhejiang University-SCIENCE A*, 24(3):243-256. https://doi.org/10.1631/jzus.A2200175

Pang BZ, Low KH, Lv C, 2022. Adaptive conflict resolution for multi-UAV 4D routes optimization using stochastic fractal search algorithm. *Transportation Research Part C: Emerging Technologies*, 139:103666. https://doi.org/10.1016/j.trc.2022.103666

Shen Y, Zhang XJ, Li Y, et al., 2026. Deep reinforcement learning-based adaptive collision avoidance method for UAV in joint operational airspace. *Defence Technology*, 56:142-159. https://doi.org/10.1016/j.dt.2025.08.011

Simorgh A, Soler M, Dietmüller S, et al., 2024. Robust 4D climate-optimal aircraft trajectory planning under weather-induced uncertainties: free-routing airspace. *Transportation Research Part D: Transport and Environment*, 131:104196. https://doi.org/10.1016/j.trd.2024.104196

Sun XY, Chen JC, Du CL, et al., 2022. Multi-agent deep deterministic policy gradient algorithm based on classification experience replay. IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), p.988-992. https://doi.org/10.1109/IAEAC54830.2022.9929494

Wang F, Zhang HH, Du S, et al., 2025a. C-SPPO: a deep reinforcement learning framework for large-scale dynamic logistics UAV routing problem. *Chinese Journal of Aeronautics*, 38(5):103229. https://doi.org/10.1016/j.cja.2024.09.005

Wang LH, Jiang ZY, Qi ZY, et al., 2025b. Proactive urban expressway guidance: a hybrid approach using reinforcement learning and traffic prediction models. *IEEE Internet of Things Journal*, 12(18):37590-37603. https://doi.org/10.1109/JIOT.2025.3583833

Wu PC, Yang XX, Wei P, et al., 2022. Safety assured online guidance with airborne separation for urban air mobility operations in uncertain environments. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19413-19427. https://doi.org/10.1109/TITS.2022.3163657

Wu Y, Low KH, Pang BZ, et al., 2021. Swarm-based 4D path planning for drone operations in urban environments. *IEEE Transactions on Vehicular Technology*, 70(8):7464-7479. https://doi.org/10.1109/TVT.2021.3093318

Xie H, Han ST, Yin JN, et al., 2024. Cooperative deduction and optimal allocation method for urban low-altitude UAV flight plan. *Acta Aeronautica et Astronautica Sinica*, 45(19):330018 (in Chinese). https://doi.org/10.7527/S1000-6893.2024.30018

Xing XJ, Ma YQ, Lei YC, et al., 2026. Multi-UAV rendezvous trajectory planning based on improved MADDPG algorithm in complex dynamic obstacle environments. *IEEE Transactions on Vehicular Technology*, 75(4):5580-5591. https://doi.org/10.1109/TVT.2025.3624052

Yang YF, Huang T, Wang TQ, et al., 2024. Sampling-efficient path planning and improved actor-critic-based obstacle avoidance for autonomous robots. *Science China Information Sciences*, 67(5):152204. https://doi.org/10.1007/s11432-022-3904-9

Yue SY, Zheng D, Wei MJ, et al., 2025. Behavior-based cooperative control method for fixed-wing UAV swarm through a virtual tube considering safety constraints. *Chinese Journal of Aeronautics*, 38(11):103445. https://doi.org/10.1016/j.cja.2025.103445

Zhang JH, Ji PY, Fang LZ, et al., 2025. Stable and continuous vertical jumping control of hydraulic legged robots through reinforcement learning. *Journal of Zhejiang University-SCIENCE A*, 26:1163-1178. https://doi.org/10.1631/jzus.A2500142

Zhang XC, Zhao HT, Wei JB, et al., 2023. Cooperative trajectory design of multiple UAV base stations with heterogeneous graph neural networks. *IEEE Transactions*

on Wireless Communications, 22(3):1495-1509. https://doi.org/10.1109/TWC.2022.3204794

Zhong G, Hua JM, Du S, et al., 2025a. Urban low-altitude flight plan optimal scheduling based on complex network. *Acta Aeronautica et Astronautica Sinica*, 46(11):531479 (in Chinese). https://doi.org/10.7527/S1000-6893.2025.31479

Zhong G, Liu YP, Du S, et al., 2025b. 3D RVO-enhanced multi-agent deep reinforcement learning for collision avoidance in urban structured airspace. *Aerospace Science and Technology*, 164:110378. https://doi.org/10.1016/j.ast.2025.110378

Zhu YJ, Li Y, Zhang XJ, et al., 2025. Risk-constrained safe path planning for unmanned aerial vehicles in urban airspace. *Journal of Beijing University of Aeronautics and Astronautics*, 1-16 (in Chinese). https://doi.org/10.13700/j.bh.1001-5965.2024.0843

Zhu YJ, Zhang XJ, Li Y, et al., 2024. Grid matrix-based ground risk map generation for unmanned aerial vehicles in urban environments. *Drones*, 8(11):678. https://doi.org/10.3390/drones8110678

Zou YY, Zhang HH, Zhong G, et al., 2021. Collision probability estimation for small unmanned aircraft systems. *Reliability Engineering & System Safety*, 213:107619. https://doi.org/10.1016/j.ress.2021.107619

Electronic supplementary materials

Sections S1–S6, Tables S1–S2, Figs. S1–S4, Eqs. (S1)–(S21)

中文概要

题　目： 面向城市空中交通运行的战略飞行规划与冲突管理：一种任务偏好约束的多智能体强化学习方法

作　者： 李妍[1,2]，张学军[1,2]，王楚茜[2,3]，申炎[1,2]，李诚龙[1,4]

机　构： [1]北京航空航天大学，电子信息工程学院，中国北京，100191；[2]北京航空航天大学，空地一体新航行系统技术全国重点实验室，中国北京，100191；[3]北京航空航天大学，国家卓越工程师学院，中国北京，100191；[4]中国民用航空飞行学院，飞行技术学院，中国成都，641419

目　的： 面向未来城市空中交通（UAM）高密度、需求驱动的低空运行场景，密集飞行任务在共享空域内易产生时空资源竞争并引发四维航迹冲突，且不同任务具有异构偏好约束。本文旨在提出一种任务偏好约束下的多智能体强化学习方法，用于预飞行窗口内的战略四维飞行计划优化与冲突管理，以兼顾运行安全、任务偏好和飞行效率。

创新点： 1. 构建面向 UAM 战略运行需求的四维飞行计划优化框架，实现飞行意图转化、空域约束表达、冲突检测与计划调控的协同管理；2. 建立战略冲突消解问题的多智能体马尔可夫决策模型，提出 AH-MADDPG 方法，实现多飞行任务间的协同调整策略学习；3. 设计任务偏好动作掩码（AMMP）机制和分层优先经验回放（HPER）机制，分别提升任务偏好满足度和策略训练效率。

方　法： 1. 基于城市低空空域网格化建模和异构任务意图表征，利用风险约束 A*算法生成多无人机初始四维飞行计划，构建四维航迹冲突探测模型并识别全局时空冲突；2. 结合 UAM 战略冲突消解需求，设计面向任务偏好约束与多目标优化的状态空间、动作空间和奖励函数，并融合 AMMP 和 HPER 机制改进 MADDPG 算法；3. 通过仿真实验，与消融变体、典型 MARL 基线算法和启发式算法进行对比，验证所提方法在冲突消解、任务偏好满足和飞行效率优化方面的有效性。

结　论： 1. 所提 AH-MADDPG 方法能够有效消解多无人机初始四维飞行计划中的全局时空冲突，在代表性场景中平均冲突解决率超过 96%；2. AMMP 机制在所有仿真场景下实现 100%的任务偏好满足率，HPER 机制有效提升策略学习稳定性和效率，二者结合使算法在冲突解决、任务偏好和飞行效率之间取得更优平衡；3. 所提框架和方法为未来高密度、需求驱动的 UAM 运行提供了一种可扩展的战略冲突管理与飞行计划优化方案。

关键词： 城市空中交通；多智能体强化学习；四维航迹；飞行计划；冲突管理；任务偏好