Transportation Research Part C 185 (2026) 105553

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Deep reinforcement learning-based multi-agent negotiation framework for conflict-free U-Space scenarios

Zhiqiang Liu, Juan José Ramos Gonzalez, Ender Çetin, Jose Luis Munoz-Gamarra *

Technical Innovation Cluster on Aeronautical Management, Universitat Autonoma de Barcelona, Sabadell, Barcelona, Spain

ARTICLE INFO

Keywords:
U-Space
Deep reinforcement learning
Negotiation framework
Multi-agent system
Proximal policy optimization
Airspace capacity optimization
Strategic deconfliction

ABSTRACT

Very-Low Level urban airspace is expected to accommodate large numbers of simultaneous drone missions in the near future, posing acute challenges for the strategic planning and deconfliction services mandated by U-space regulations. The current strategic planner algorithm relies on the first-come-first-served or batch policy that rejects nearly half of the submitted flight plans in large-scale scenarios. This paper presents a novel study that integrates a Deep Reinforcement Learning-based Multi-Agent Negotiation Framework into the U-space strategic planning and deconfliction service. The framework reallocates rejected flight plans through an iterative sealed-bid auction mechanism in a fully competitive auction environment. Four role-specific drone operator agents were constructed: Achiever, Optimizer, Economizer and Normalizer, to learn bidding policies with the Proximal Policy optimization algorithm. Each round's reward combines individual profit, bidding cost and a global utilisation signal, guiding agents towards system-efficient yet self-interested behaviour. The simulation experiments show that the Deep Reinforcement Learning-based Multi-Agent Negotiation Framework lifts the mission re-accommodation ratio to a notable level while converging to a final allocation within a short strategic planning time. Analysis of reward distribution and bidding patterns confirms that agents adapt their strategies to heterogeneous operational objectives without explicit coordination. These findings indicate that learning-enabled auctions can reconcile operator competition with network-level efficiency, offering a scalable path towards conflict-free, high-density U-space operations.

1. Introduction

The rapid integration of Unmanned Aerial Vehicles (UAVs) into Very-Low Level (VLL) airspace presents significant challenges for airspace management (EASA, 2020), particularly within the framework of U-space (SESAR Joint Undertaking, 2017), which aims to provide a harmonised regulatory and technological ecosystem for autonomous drone operations. With the anticipated growth in demand for drone-enabled services, such as logistics, surveillance, and infrastructure inspection, the airspace is expected to become increasingly congested, underscoring the need for robust, scalable, and efficient mechanisms for strategic planning and deconfliction (European Union Aviation Safety Agency, 2024). Ensuring safe and efficient operations requires optimized resource allocation, particularly in high-density environments where multiple Drone Operators (DOs) compete for limited airspace resources. Evidence from

* Corresponding author.
E-mail addresses: Zhiqiang.Liu@uab.cat (Z. Liu), JuanJose.Ramos@uab.cat (J.J. Ramos Gonzalez), Ender.Cetin@uab.cat (E. Çetin), JoseLuis.Munoz.Gamarra@uab.cat (J.L. Munoz-Gamarra).

https://doi.org/10.1016/j.trc.2026.105553
Received 23 June 2025; Received in revised form 6 November 2025; Accepted 31 January 2026
Available online 13 February 2026

CORUS-XUAM (CORUS-XUAM, 2020; CORUS-XUAM Consortium, 2023) project and previous innovative methodologies (Schefers et al., 2018; Liu et al., 2024b; Ramos González et al., 2025; Munoz-Gamarra et al., 2024; Munoz-Gamarra et al., 2024; Munoz-Gamarra et al., n.d.) highlights these challenges.

The First-Come-First-Served (FCFS) protocol is the most elementary, granting airspace access in the chronological order of Flight Plans (FPs) submission (Pongsakornsathien et al., 2025; Dai et al., 2021; Liu et al., 2023; Munoz-Gamarra et al., 2023). While its simplicity and transparency are advantageous, FCFS is notoriously inefficient in dense traffic scenarios, often leading to a high rate of FP rejections and suboptimal use of available VLL airspace capacity (Liu et al., 2024b). In response to these limitations, alternative methods such as Batch processing has been introduced. By evaluating and approving FPs in groups rather than sequentially, Batch processing can achieve a higher density of accepted missions, as demonstrated in studies (Ramos González et al., 2025), which showed an increase in the average acceptance ratio clearly. Despite this improvement, a significant portion, nearly half of all submitted FPs are still rejected in large-scale scenarios, underscoring a critical need for more sophisticated and adaptive allocation mechanisms.

The core of the strategic deconfliction problem lies in the efficient allocation of a finite resource airspace among competing users. This challenge is not unique to U-space and has been a central theme in Air Traffic Management (ATM) for decades (Hamissi et al., 2025; Pellegrini et al., 2020; Schefers et al., 2020, 2018). Early research focused on centralized optimization models, often using mathematical programming techniques to minimize delays and resolve conflicts (Liu et al., 2023; Pongsakornsathien et al., 2025; Schefers et al., 2020). While effective on a smaller scale, these centralized approaches struggle with the computational complexity and scalability required for high-density, dynamic environments. This led to the exploration of distributed approaches, with a significant body of work focusing on Multi-Agent Systems (MAS) (Chen et al., 2024).

MAS offer a decentralized paradigm where autonomous agents, representing individual aircraft or operators, interact and negotiate to achieve both individual and system-level goals. The application of MAS to air traffic management has been explored in various contexts, from en-route conflict resolution to airport ground movement. Weigang et al. (2010), Schefers et al. (2020) and Chen et al. (2009) have demonstrated the potential of multi-agent frameworks to optimize air traffic flow, while others have applied similar concepts to urban traffic control and autonomous vehicle coordination (Chen et al., 2009; Gao et al., 2021; Dubey et al., 2025; Wang et al., 2025; Xie et al., 2025), highlighting the versatility of agent-based negotiation. These systems often employ auction mechanisms as a means of resource allocation. Auctions provide a structured and competitive framework for agents to express their preferences and valuations for scarce resources, leading to efficient and often optimal outcomes.

The advent of Deep Reinforcement Learning (DRL) has opened new frontiers for creating intelligent, adaptive agents capable of learning complex strategies in dynamic and uncertain environments. DRL combines the perceptual power of deep neural networks with the goal-oriented learning capabilities of reinforcement learning, enabling agents to master high-dimensional decision-making problems without explicit programming (Li et al., 2019; Ribeiro et al., 2022). Its success in complex games like Go, as demonstrated by Silver et al. (2016) has inspired a wave of applications in diverse fields, including robotics, finance, and transportation (El-Sayed et al., 2021; Zhang et al., 2021; Li et al., 2022). In the context of airspace management, DRL offers a powerful tool for training agents to navigate the intricate trade-offs of strategic deconfliction. DRL algorithms, particularly policy gradient methods (Bohn et al., 2019; Koch et al., 2019) like Proximal Policy Optimization (PPO) (Schulman et al., 2017), are well-suited for the continuous and competitive nature of airspace allocation. PPO's actor-critic architecture (Konda and Tsitsiklis, 2000) allows agents to learn both a policy (the actor) for taking actions and a value function (the critic) for evaluating states, while its clipped surrogate objective function ensures stable and reliable policy updates, preventing the drastic performance swings that can plague other algorithms.

This paper builds upon these converging streams of research to propose a novel Deep Reinforcement Learning-based Multi-Agent Negotiation Framework (DRL-MANF) specifically designed for the strategic deconfliction challenge in U-space. This framework introduces a optimized mechanism for FPs that are initially rejected by conventional FCFS or batch processing systems. Instead of being discarded, these Rejected Flight Plans (RFPs) are reintroduced into a competitive, multi-round auction where DO agents can bid for alternative launch windows. This approach is predicated on the observation that many drone missions, such as agricultural monitoring or infrastructure inspection, possess a degree of temporal flexibility, making them amenable to rescheduling without compromising their objectives.

A key innovation of this work is the introduction of heterogeneous agent roles. The study moves beyond the assumption of homogenous operators and constructs four distinct agent archetypes, each with a unique reward structure reflecting different business objectives and risk appetites. This heterogeneity fosters a more realistic and complex competitive landscape, allowing us to study how diverse strategic motivations interact and influence overall system performance. The agents learn their bidding policies using the PPO algorithm, guided by a composite reward signal that balances individual profit, the cost of bidding, and a global signal reflecting the overall efficiency of VLL airspace utilization. The negotiation itself is structured as an iterative, sealed-bid auction, mediated by a virtual currency called "P-coins" that regulates competition and encourages strategic resource management.

Through extensive simulation experiments, this work demonstrates that the DRL-MANF significantly increases the mission re-accommodation ratio, pushing it to levels unattainable by traditional planning strategies alone. The framework converges to a final, efficient allocation within a short timeframe, making it suitable for the dynamic nature of strategic planning. By analyzing the emergent bidding patterns and reward distributions, this study confirms that the agents learn to adapt their strategies according to their predefined roles without any explicit coordination, reconciling individual, self-interested behavior with network-level efficiency. These findings suggest that learning-enabled auctions (Liu et al., 2024a) represent a promising and scalable solution for managing the high-density, conflict-free U-space operations of the future, offering a pathway to significantly enhance airspace capacity and utilization. This study, therefore, makes a significant contribution by integrating advanced DRL techniques with a sophisticated multi-agent auction mechanism to address a critical and pressing challenge in the evolution of autonomous aviation.

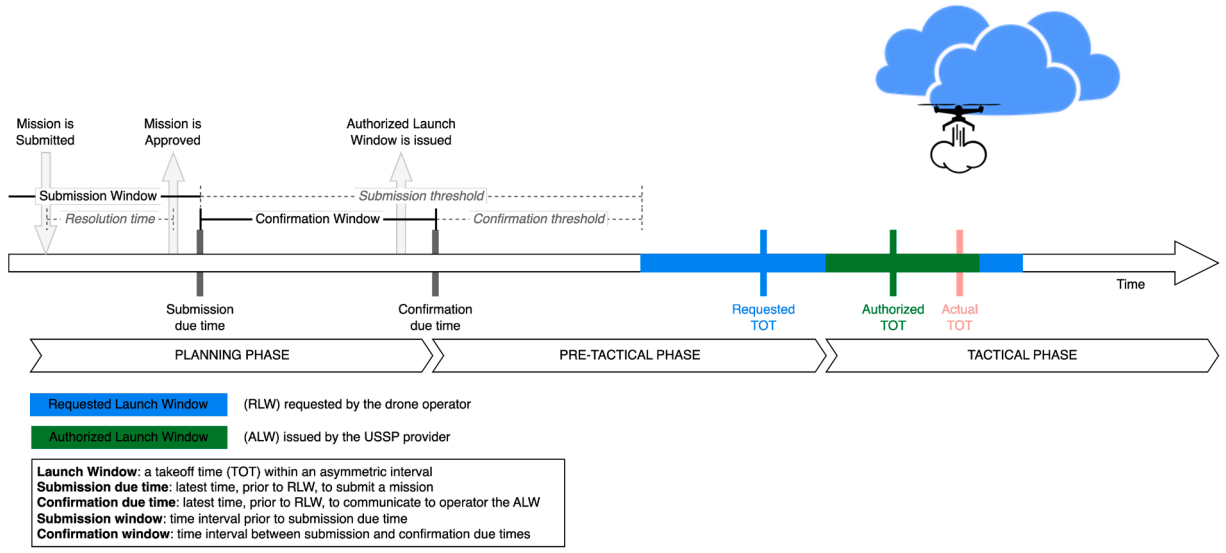

Fig. 1. Strategic deconflicting service.

The remainder of the paper unfolds in four further sections. Section 2 formalises the strategic re-allocation problem, and details the DRL-MANF architecture covering state vectors, neural-network topologies, auction protocol and reward shaping. Section 3 describes the simulation configuration such as agent roles, learning algorithm and training protocol and Section 4 presents quantitative results: convergence diagnostics, acceptance ratios, bidding behaviour and sensitivity to demand intensity and P-coin budgets. Section 5 discusses policy implications, scalability to larger markets and concludes by outlining future research directions.

2. Problem statements and formulations

2.1. Problem background

As shown in Fig. 1, to ensure structured airspace allocation, DOs must submit their FPs before the Submission Window (SW, the time interval before take-off during which mission submission is permitted) deadline, including temporal and spatial details such as the Requested Launch Window (RLW, the time interval of given length around the requested take-off time) and flight trajectories. This requirement applies to all airspace users to ensure orderly scheduling. The Airspace Manager (AM) is the operational entity, designated or authorised by the competent authority that created a U-space airspace (European Union, 2021). AM leveraging common information services supplied by the Common Information Service Provider (CISP), executes strategic planning and deconfliction algorithm, evaluates FP conflicts in both temporal and spatial conditions and assigns at least one Authorised Launch Window (ALW, the time interval during which a mission's takeoff can safely occur without conflicting with any previously approved missions) within the RLW. Each FP receives a single ALW approval, guaranteeing a conflict-free take-off. This process is completed before the Confirmation Window (CW, the time interval between the SW and the RLW, which is used to determine the ALW while accommodating arriving FPs before the interval's end) deadline. FPs without ALW authorisation are generally rejected due to conflicts.

Although the AM determines ALW feasibility in advance, this information is not immediately disclosed. Instead, the interim confirmation period (when a final take-off value is assigned) allows for further optimization, reassessing ALW availability for both incoming and existing FPs to enhance airspace utilization.

The FCFS policy ensures fairness and transparency in U-space airspace management by approving flight plans in the order they are received. However, this rigid approach limits airspace utilisation efficiency, as it does not account for potential optimizations in scheduling and resource allocation. To address this problem, a Batch processing method was introduced (Ramos González et al., 2025), allowing the AM to evaluate multiple flight plans simultaneously using the strategic planning and deconfliction algorithm. The Batch processing approach increases the number of ALWs assigned within a set of given RLWs, improving overall airspace capacity. As shown in Fig. 2 (Ramos González et al., 2025), while Batch processing demonstrates improvements over FCFS, experimental results indicate that further optimization opportunities remain.

Airspace utilisation in this context is defined as the percentage of successfully accepted FPs relative to the total submitted FPs. The acceptance ratio serves as a key metric for evaluating the efficiency of different allocation strategies. Compared to the 41.5 % average acceptance ratio of FCFS, the Batch processing increases the acceptance ratio to 49.8 %, demonstrating its effectiveness in improving airspace utilisation. However, despite this improvement, nearly 50 % of submitted FPs are still rejected due to airspace constraints and conflicts (Ramos González et al., 2025). It means that while Batch processing improves allocation efficiency, it does not fully resolve the issue of FP rejections, leaving a significant number of FPs unaccommodated.

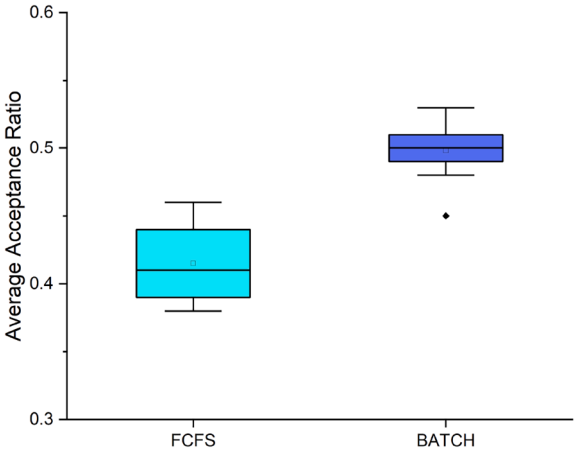

Fig. 2. Comparison of FCFS and BATCH.

This raises a pivotal question: To improve VLL airspace efficiency, can RFPs be reconsidered and re-accommodated through a more adaptive mechanism? With the question in mind, the research group investigated the airspace condition after the strategic planning and deconfliction algorithm had been executed. Observations of airspace utilisation and the characterisation of FPs inspired the optimization approach proposed in this study. The remainder of this work provides a thorough answer to the raised question.

After a detailed analysis of airspace occupancy, it was observed that the airspace remaining after strategic planning and deconfliction is not always fully utilised. In fact, even within certain simulation periods, a substantial portion of the airspace remains available. FPs that were previously rejected due to temporal and spatial conflicts could be re-accommodated within the existing airspace resources, provided their DOs are willing to sacrifice their originally scheduled take-off times. It is essential to note that the feasibility of re-accommodation depends on the AM's ability to identify an Alternative Requested Launch Window (ARLW, period of time when the submited mission would be free of conflict) for previously RFPs within a predefined temporal buffer and on the DO's willingness to accept that ARLW. Importantly, unlike in traditional ATM, some of the FPs in a U-space environment frequently correspond to non-time-critical missions. For example, consider a routine aerial photogrammetry mission conducted to update high-resolution orthomosaics of an agricultural field. Such a mission does not require take-off at an exact minute; rather, departing within a ±50-min interval around the originally planned launch time will not compromise data quality or operational objectives. In these cases, allowing the FP to re-accommodate into available airspace resources, rather than having it be outright rejected and indefinitely delayed, yields a clear advantage in terms of both mission continuity and overall airspace utilisation.

In light of these observations, to complement FCFS and Batch processing, a DRL-MANF is initiated as an additional optimization mechanism, providing DOs a second chance to re-enter the airspace, which allows RFPs to be reconsidered through a multi-agent negotiation framework. Instead of permanently discarding these FPs, they are reintroduced into the allocation process via an auction mechanism, where DOs can compete for several solutions, each of which contains ARLWs to the original FPs information provided by the AM. Through a multi-agent negotiation process, potentially viable FPs can be strategically reintegrated, further enhancing FP acceptance rates and overall airspace efficiency.

2.2. Global solution generation

The global solution, dynamically generated by the AM, serves as the foundation for the MANF process, providing alternative launch opportunities for previously rejected FPs. Unlike an immediate update after each rejection, the AM, which functions as the environment within the DRL framework, periodically constructs the global solution at regular intervals. This approach allows the AM to accumulate enough RFPs before searching for feasible alternatives, ensuring a more comprehensive and optimized reassignment process.

AM searches for ARLWs within a predefined time range around the originally requested RLW. Specifically, for each RFP, the AM explores available time slots within a ±50-min window from the original RLW. This ±50-min range was empirically selected as a balanced compromise between DOs' operational flexibility, solver efficiency, and the AM's capability to identify feasible conflict-free alternatives. Tests under high-density scenarios indicated that the original 20-min RLW used in the strategic deconflicting service was often insufficient, while extending the search range beyond ±50 min produced marginal benefits with significantly higher computational cost and operator delay.

- If feasible ARLWs exist, they are provided with the corresponding DOs as potential rescheduling options.
- If no available time slots are found within this range, the FP remains unallocated and cannot be reintegrated.
- Some FPs may have multiple ARLWs, offering DOs different alternative solutions. In such cases, each ARLW represents a distinct rescheduling opportunity, creating multiple possible solutions for the DO to evaluate.

Table 1
Example of global solution table.

	DO-A		DO-B	
RFPs	RFP1	RFP2	RFP3	RFP4
Solution 1	ARLW1,1	–	ARLW1,3	ARLW1,4
Solution 2	–	ARLW2,2	–	–
Solution 3	ARLW3,1	ARLW3,2	–	ARLW3,4
Solution 4	–	–	ARLW4,3	–

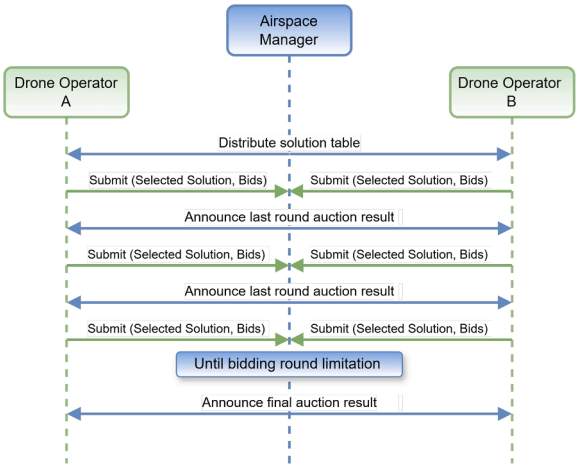

Fig. 3. The process of negotiation framework.

The global solution consists of multiple solutions for all DOs, where each solution corresponds to all RFPs along with their associated ARLWs. These solutions could consider various airspace constraints, such as traffic density and operational priorities, but the RLW time slot is the most focused information in the current stage. Once generated, the solution table is distributed to all DOs, initiating the multi-agent negotiation phase, where DOs compete for available allocations based on their learned strategies. The example of the global solution table is shown in Table 1. In Table 1, $ARLW_{i,r}$ represents the ARLW assigned to RFP r in candidate Solution i generated by the AM. Rows correspond to candidate solutions, while columns correspond to individual RFPs. A dash (-) indicates that no feasible ARLW exists for that (i, r) pair.

The observation can be made from Table 1, DO-B prefers solution 1 as it reassigns both RFPs-RFP3 and RFP4, while DO-A favours solution 3 since it accommodates two RFPs instead of one in solution 1. This creates strong competition, making winning the negotiation crucial for each DO.

When transmitting the solution table to each DO, the AM ensures that each DO receives only the information relevant to its operations. For example, if DO-A is assigned four possible solutions, each solution will contain only the ARLWs corresponding to its associated RFPs (RFP1 and RFP2). This restricted information access is essential because FP data is considered commercially sensitive. Given that DOs operate in a fully competitive environment, FP details are not shared between competing entities to maintain data confidentiality and fair competition.

2.3. Deep reinforcement learning based multi-agent negotiation framework

DRL provides a data-driven optimization method where agents continuously learn from interactions with the environment. Unlike rule-based heuristics, DRL allows agents to make decisions based on learned policies, adapting to changing airspace constraints, competition, and mission information (Bus¸oniu et al., 2008; Wang et al., 2022).

In this study, DRL-MANF is designed to dynamically reallocate airspace resources by enabling DOs to strategically compete for multiple alternative solutions, each solution consisting of different alternative requested launch windows. Within this framework, each DO operate as an autonomous reinforcement learning agent, whose objective is to maximise its mission acceptance rate while optimising its bidding strategy in a complete competitive negotiation environment.

By deploying DRL, the framework ensures adaptive decision-making through iterative learning, allowing DOs to refine their negotiation tactics over time. This continuous optimization process enhances airspace utilisation while maintaining flexibility in mission scheduling and resource allocation.

Once the solution table is distributed, the multi-agent negotiation process begins. Each DO receive an individual solution table from the global solution. And they analyse the solutions received and prepare for the first round of bidding, where they must decide which solution to bid on and how much to bid. Priority Coins (P-coins) are the virtual currency units allocated to each DO agent

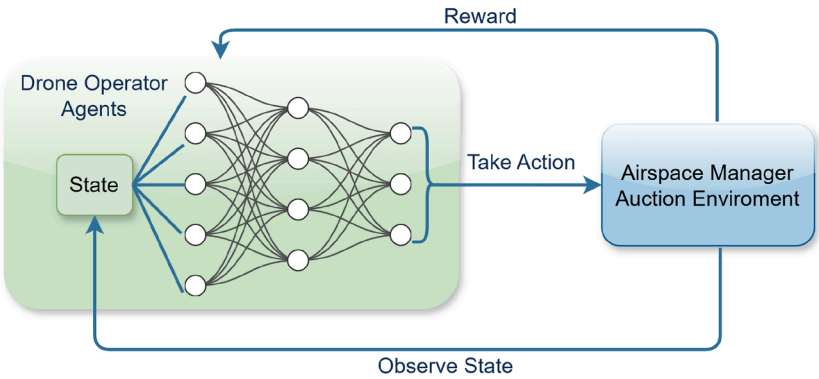

Fig. 4. The structure of deep reinforcement learning.

at the start of the negotiation process. It is important to note that P-coins are not related to the real currency, it is a mechanism to control the bidding process. These coins serve as a bidding resource that regulates competition and prevents irrational or excessively high bidding behaviours. Each DO must strategically allocate its limited P-coins across multiple negotiation rounds to optimize the bidding strategy.

After collecting all bids, the AM evaluates the submissions and determines the winning bid for the current round. The DO that offers the highest bid for the selected solution secures the execution rights for the selected solution. If multiple DOs bid the same amount for the same solution, a random selection is applied. Once the AM finalises the auction results, it announces the outcome of the previous round to all DOs. This feedback enables DOs to adjust their bidding strategies dynamically in response to competition. As Fig. 3 shows the ontology of the negotiation process. Following the auction resolution, a new negotiation round begins, allowing DOs to re-evaluate their strategies based on past performance. In each round, DOs can adjust their bid amounts based on the holding P-coins and competition trends or change their selected solution if previous bids were unsuccessful. This iterative process continues until the predefined bidding round limit is reached. It is also possible for the negotiation to end before reaching the predefined negotiation round limit if only one agent remains bidding for the selected solution, meaning that all other competitors have withdrawn from the bidding process. At the conclusion of all negotiation rounds, the AM announces the final auction results, officially determining the final execution of the solution. The DOs that successfully secured their selected solution can proceed with their preferences, while others may have to accept the assigned arrangement for their RFPs within the executed solution.

Through this structured, multi-round auction mechanism, the DRL-MANF ensures that DOs continuously learn from past behaviours to improve their bidding strategies. It is important to note that P-coins are not deducted during the negotiation process but only at the end, when the winner is determined. This ensures that DOs can strategically adjust their bids throughout the negotiation rounds without immediate financial constraints, promoting a more adaptive decision-making process. Ideally, this negotiation framework operates in parallel with the strategic deconfliction service, running at fixed time intervals. During each interval, the system collects information on RFPs and searches for ARLWs to be used in subsequent negotiations. The entire negotiation process is designed to be highly efficient, with the expectation that the final execution solution is determined within one minute under optimal conditions. This rapid resolution ensures minimal delays and optimizes airspace utilisation efficiency.

Fig. 4 illustrates the closed-loop interaction between DO agents and the AM-auction environment. In each bidding round, every agent compresses the instantaneous auction status into a state vector, passes it through its policy network, and outputs an action (a bid tied to a selected solution). The AM then evaluates the joint action profile, updates the auction book, and returns to each agent a scalar reward-round reward after each bidding round along with the updated state vector and global plus individual reward at episode completion. This observe-act-feedback cycle iterates until the prescribed bidding rounds concludes.

This framework is characterised by three key features:

- **Role Specific Agents** Each drone operator agent is instantiated as an independent autonomous agent with a bespoke negotiation profile that reflects its business objectives and risk appetite. Each agent maintains its own neural-network policy; they share only the auction environment and exchange neither parameters nor gradients. This design deliberately avoids the coordination assumptions of conventional multi-agent reinforcement learning, framing the problem instead as multiple independent agents interacting with the same state-transition dynamics. Four agent archetypes are introduced to reflect real-world heterogeneity: depending on mission urgency, commercial constraints, or tolerance for a delayed take-off, an operator may bid aggressively for any viable solution, bid conservatively, or elect not to reschedule at all. By encoding these divergent attitudes into distinct reward weightings, the framework reproduces the full spectrum of strategic behaviours that emerge when operators with different priorities compete for ARLW solutions.
- **Competitive Auction-Based Environment** All DOs operate independently, making strategic decisions without any collaboration, as the auction environment is fully competitive. The AM hosts a sealed bid, single winner auction that is repeated over multiple discrete rounds. Only one agent (i.e., one DO) can ultimately win the negotiation for a given solution and secure the final execution of the selected solution. Agents have no direct communication channel and must infer rivals' intentions only from historical bid

outcomes, fostering an intrinsically competitive landscape. For each candidate solution, the highest bid wins; ties are broken by a deterministic priority rule to guarantee uniqueness. P-coins are debited only when an agent wins, mitigating the risk of early budget exhaustion and supporting exploration.

- **Multi-Solution Bidding Mechanism** The AM provides multiple feasible solutions to each DO rather than a single re-accommodating option. Each solution specifies a distinct set of ARLWs for the RFPs, and these sets need not overlap: for example, as Table 1 shows, Solution 1 may include ARLWs that permit the re-accommodate of RFP 3 and RFP 4, whereas Solution 2 might omit RFP 3 and RFP 4 but add RFP 2. By exposing every feasible ARLW combination, the AM ensures that all DOs have equitable access to the full universe of re-accommodation options. During the auction, each DO evaluates these candidate solutions, comparing projected mission acceptance, temporal trade-offs, remaining P-coin budgets, and submits bids on those that best align with its operational goals. This design underpins systemwide fairness: no operator is constrained to a single re-accommodating option, and every DO receives the entire spectrum of conflict-free ARLW configurations uncovered by the AM's search.

2.4. Formalization of the DRL-MANF

The DRL-MANF is structured as a Markov Decision Process (MDP) (Sutton et al., 1999; Puterman, 2008), where each DO is modelled as an independent reinforcement learning agent. The primary objective of each DO is to maximise its reward, which varies based on its assigned role. Since there are four distinct agent types: Achiever, Optimizer, Economizer, and Normalizer, each follows a different reward function, leading to distinct negotiation strategies and learning objectives.

2.4.1. State representation

Each DO i at negotiation round r observes a state vector (Eq. (1)):

$$S_r^i = \{\, ID,\ P\text{-}coins,\ Win,\ SolutionFeatures \,\}, \tag{1}$$

where:

- i: Represents the index of a specific DO in the multi-agent system.
- r: Represents the negotiation round.
- ID: A normalised value representing the agent's unique identifier.
- $P\text{-}coins$: The number of P-coins the DO has available for bidding, normalised to $[0, 1]$.
- Win: A binary feature indicating whether the DO has won the selected solution in previous rounds.
- $SolutionFeatures$ (for each solution j):
  - Solution Index - Normalised index of the solution in the list of available solutions.
  - Missions Could Fly $M_{fly}(j)$ Number of RFPs could be re-accommodate in this solution.
  - Max Fly Missions $M_{\max}(j)$– Number of RFPs could be re-accommodate in the global solution.
  - Rejected Missions $M_{reject}(j)$– Number of RFPs for the DO in the current global solution.
  - Last Bid: The last bid placed by the DO on this solution.
  - Is Win: A binary feature indicating whether this solution is the winning solution for last round.
  - Last Round Highest Bid: The highest bid recorded in the last round for this solution.
  - Current Round: Normalised value representing the current round in the negotiation process.

2.4.2. Action representation

At each negotiation round r, each DO i selects an action A_r^i which consists of choosing a solution and placing a bid (Eq. (2)):

$$A_r^i = \{\, SelectedSolution,\ BidValue \,\}, \tag{2}$$

where:

- Selected Solution j: The index of the solution from the solution table.
- Bid Value B_r^i: The amount of P-coins allocated for the selected solution, determined by the policy network.

The Actor Network outputs two probability distributions:

1. Solution Selection Distribution $\pi_\theta(S_r)$: A categorical probability distribution over all available solutions.
2. Bid Selection Distribution $\beta_\theta(S_r, A_r^{\text{solution}})$: A categorical probability distribution over bid values, conditioned on the selected solution.

The final bid is constrained by the DO's available P-coins, ensuring that no DO bids more than its current balance (Eq. (3)):

$$B_r^i = \min\left(\beta_\theta\left(S_r,\ A_r^{\text{solution}}\right),\ AvailableP\text{-}coins\right), \tag{3}$$

where this action representation ensures that DOs make strategic bidding decisions while adhering to their resource constraints.

2.4.3. Reward function

Each DO i at negotiation round r receives reward R_r^i from three distinct sources (Eq. (4)):

$$R_r^i = R_{\text{round}}^i + R_{\text{individual}}^i + R_{\text{global}}^i, \tag{4}$$

where R_{round}^i is the immediate reward issued in each round r (Eq. (5)), $R_{\text{individual}}^i$ is the role-weighted individual component (Eq. (6)), and R_{global}^i is the system-level component (Eq. (7)).

R_{round}^i (Round Reward, RR): Immediate reward based on chosen solution in each round (Eq. (5)).

$$R_{\text{round}}^i(j) = \begin{cases} 0.01 + \frac{M_{\text{fly}}(j)}{M_{\text{reject}}(j)} * 0.01 & \text{if } M_{\max}(j) = M_{\text{max\_global}}, \\ 0.005 + \frac{M_{\text{fly}}(j)}{M_{\text{reject}}(j)} * 0.005 & \text{if } M_{\max}(j) \geq 0.8 * M_{\text{max\_global}}, \\ -0.01 & \text{otherwise.} \end{cases} \tag{5}$$

Where:

- $M_{\text{fly}}(j)$ number of RFPs that could be re-accommodate under the selected solution.
- $M_{\text{reject}}(j)$ number of RFPs in the selected solution.
- $M_{\max}(j)$ maximum RFPs that could be re-accommodate in this solution.
- $M_{\text{max\_global}}$: maximum possible re-accommodate RFPs across all solutions.

This function rewards DOs if they contribute to optimal mission allocation, while penalising poor allocation. RR applies for all roles of agents.

$R_{\text{individual}}^i$ (Individual Reward, IR): Agent-specific reward based on its role at the end of the negotiation process (Eq. (6)).

$$R_{\text{individual}}^i = \left(w_1 R_{\text{cost\_penalty}} + w_2 R_{\text{bonus}} + w_3 R_{\text{lose\_penalty}}\right)/10, \tag{6}$$

where:

- $R_{\text{cost\_penalty}}$ for excessive spending of P-coins. Calculate as -(final bid/5) if DO is the winner.
- R_{bonus} winning bonus for securing solution. Calculate as $+10$ if DO wins the auction.
- $R_{\text{lose\_penalty}}$: if DO lose the auction will receive penalty. Calculate as -10 if DO lose the auction.

R_{global}^i (Global Reward, GR): The global reward (Eq. (7)), which encourages fair competition and efficient resource allocation, will be given at the end of the negotiation.

$$R_{\text{global}}^i = \begin{cases} \frac{M_{\text{total}}}{M_{\text{max\_global}}}, & \text{if } M_{\text{total}} \geq 0.8 * M_{\text{max\_global}}, \\ -(1 - \frac{M_{\text{total}}}{M_{\text{max\_global}}}), & \text{otherwise.} \end{cases} \tag{7}$$

Where:

- M_{total} the total number of successfully re-accommodated RFPs across all DOs.
- $M_{\text{max\_global}}$: the global maximum possible RFPs that could be re-accommodate.
- $0.8 * M_{\text{max\_global}}$: a threshold ensuring that at least 80 % of the total RFPs capacity is utilized before positive rewards are granted.

2.4.4. Transition representation

The state transition function governs how the system updates the state of each DO at the end of each negotiation round. Since the framework follows a multi-round auction mechanism, transitions occur discretely between negotiation rounds rather than continuously over time. At each negotiation round r, the DO i observes a state S_r^i, selects an action A_r^i, receives a reward R_r^i, and transitions to the next round's state S_{r+1}^i. The transition function is formally represented as (Eq. (8)):

$$S_{r+1}^i = P\left(S_r^i, A_r^i\right), \tag{8}$$

where $P(\cdot)$ defines how the state is updated based on the previous round's decisions and auction outcomes.

In the process of negotiation, the DO updates its state based on the four key factors belonging to Solution Features. The information update is determined either by the DO's actions or by the information received from the AM:

- Last bid: The DO's most recent bid amount is stored.
- Is win: A binary indicator (0 or 1) is updated based on whether the current solution is the winning solution in the last round.
- Last round highest bid: The highest bid recorded in the previous round for the current solution.

- Current round: The negotiation round counter increments by 1.

At the end of the negotiation, the final bid winner is determined, and the P-coins deduction occurs (Eq. (9)):

$$S_{\text{final}} = \begin{cases} P_{\text{final\_coins}} = P_{\text{initial\_coins}} - B_{\text{final\_win}} & \text{if DO wins,} \\ P_{\text{final\_coins}} = P_{\text{initial\_coins}} & \text{if DO loses.} \end{cases} \tag{9}$$

Winning DO's P-coins are reduced by the final bid amount. Other DOs retain their full P-coin balance. The winner DO's state vector Win will be updated from 0 to 1 to indicate the final winner.

2.5. Formal problem formulation

The negotiation and learning process can be interpreted as a constrained optimization problem supervised by the AM. Let D denote the set of DOs, and S the set of feasible global solutions generated by the AM. Each solution $s \in S$ specifies a set of conflict-free ARLWs for all RFPs, satisfying the allowable time-shift bound $\pm\Delta_t$ (50-min in this work). $M_i(s)$ denotes the number of RFPs of DO i that are successfully re-accommodated under solution s, which is determined by the specific ARLW assignments contained in s, and $M_{\text{total}}(s) = \sum_{i \in D} M_i(s)$ represents the total number of re-accommodated RFPs across all DOs. The AM seeks to maximize the overall re-accommodation capacity:

$$\max_{s \in S} M_{\text{total}}(s) = \sum_{i \in D} M_i(s), \tag{10}$$

subject to:

Feasibility: s must be conflict-free and satisfy $|\text{start}_{\text{ARLW}} - \text{start}_{\text{RLW}}| \le \Delta_t$;

Budget: $0 \le b_i^r \le \text{AvailableP-coins}_i^r$ for each round r, with $P_i^{\text{init}} = 100$;

Process: $r \le R_{\max}\ (= 30)$, bids are sealed and agents cannot communicate;

Information: each DO observes only its own RFPs realted solutions.

The system operationally realises (10) through a repeated sealed-bid auction over at most $R_{\max}$ rounds. In each round r, DO i selects an action $A_i^r = (j_i^r, b_i^r)$, where j_i^r is the index of the targeted global solution and b_i^r the bid level, both sampled from policy π_i. The winning global solution is determined by the highest valid bid when the negotiation converges.

Each agent aims to maximize its expected cumulative return over the episode, composed of three reward components previously defined in Eqs. (5)–(7):

$$\max_{\pi_i} \mathbb{E}_{\pi_i}\left[\sum_{r=1}^{R_{\max}} R_i^{\text{round}}(j_i^r) + R_i^{\text{ind}} + R^{\text{glob}}\right], \tag{11}$$

where the round reward $R_i^{\text{round}}(j_i^r)$ depends on the quality of the selected solution in round r, R_i^{ind} captures the role-weighted individual outcome (cost penalty, bonus, or lose penalty), and R^{glob} represents the system-level performance shared by all DOs. The coefficients (w_{1i}, w_{2i}, w_{3i}) determine the role type (Achiever, Optimizer, Economizer, or Normalizer).

This formulation explicitly connects the decentralized learning process to an underlying social optimization objective, clarifying the decision variables, agent roles, and operational constraints.

Having established this mathematical formulation, the next subsection translates it into an operational procedure. Specifically, it presents the algorithmic implementation that governs the multi-round negotiation and the interactions between the DO agents and the AM.

2.6. Algorithmic implementation of the DRL-MANF

Algorithm 1 summarizes the end-to-end process of the proposed DRL-MANF framework. The AM first generates and distributes feasible solution tables for each DO agent, after which the agents iteratively submit sealed bids under the P-coin feasibility constraint. A round reward is issued at every bidding round, while the individual and global rewards are granted once the negotiation terminates. As the PPO routine follows the standard implementation, it is omitted here for clarity.

3. Methodology

The proposed DRL-MANF offers distinct advancements specifically designed to tackle strategic deconfliction challenges in U-space scenarios. One significant innovation lies in the fine-grained differentiation of DO roles. While existing frameworks typically treat drone operators homogeneously, DRL-MANF introduces explicitly defined, heterogeneous agent roles, each reflecting nuanced operational motivations and resource management strategies. This differentiation not only enrich the agents' adaptive capability but also ensures robust learning and strategic diversity across negotiations.

Algorithm 1 DRL-MANF for one episode (learning routine omitted).

Require: RFP; AM search window Δt (e.g., ± 50 min); DO agents $\mathcal{A} = \{1, \ldots, N\}$ with role weights (w_1, w_2, w_3).

Fixed episode settings (as used in experiments)

$R_{\max} \leftarrow 30$ ▷ Maximum negotiation rounds

for each DO $i \in \mathcal{A}$ **do**

  $P_i^0 \leftarrow 100$

end for ▷ Initial P-coin budget

(A) AM: global solutions and private distribution

$G \leftarrow \varnothing$ ▷ Global solution table, partitioned by DO

for each DO $i \in \mathcal{A}$ **do**

  $C_i \leftarrow$ feasible ARLWs for i's RFPs within Δt; compose up to K solutions

  $G \leftarrow G \cup \{(i, C_i)\}$

end for

for each DO $i \in \mathcal{A}$ **do**

  $T_i \leftarrow \textsc{ExtractIndividualSolutionTable}(G, i)$ ▷ Privacy-preserving slice

  $S_i^{(1)} \leftarrow \textsc{BuildState}(T_i, \mathrm{ID} = i, P_i = P_i^0, \mathrm{Win} = 0)$ ▷ Eq. (1)

  $\tau_i \leftarrow \varnothing$ ▷ Per-agent log (for analysis)

end for

(B) Sealed-bid auction: multi-round rollout

for $r = 1$ **to** $R_{\max}$ **do**

  $A^{(r)} \leftarrow \varnothing$ ▷ Joint actions at round r

  for each $i \in \mathcal{A}$ **do**

    $j_i \leftarrow \textsc{SelectSolution}(S_i^{(r)}, T_i)$ ▷ Agent policy; Eq. (2)

    $b_i^{\mathrm{raw}} \leftarrow \textsc{BidValue}(S_i^{(r)}, j_i)$ ▷ Agent policy; Eq. (2)

    $b_i \leftarrow \min\{b_i^{\mathrm{raw}}, \textsc{AvailablePCoins}(i)\}$ ▷ Per-round feasibility; Eq. (3)

    $A^{(r)} \leftarrow A^{(r)} \cup \{(i, j_i, b_i)\}$

  end for

  $\mathcal{B}_{\mathrm{active}} \leftarrow \{\, i \in \mathcal{A} \;:\; b_i > 0 \,\}$ ▷ Agents that actually bid this round

  if $|\mathcal{B}_{\mathrm{active}}| = 1$ **then**

    winner $\leftarrow$ the unique $i \in \mathcal{B}_{\mathrm{active}}$; $j^\star \leftarrow j_i$; $b^\star \leftarrow b_i$

    for each $k \in \mathcal{A}$ **do**

      $\mathrm{RR}_k^{(r)} \leftarrow \textsc{RoundReward}(j_k \mid T_k)$ ▷ Issued every round, Eq. (5)

      $S_k^{(r+1)} \leftarrow \textsc{Transition}\big(S_k^{(r)}, (j_k, b_k), (\mathrm{winner}, j^\star, b^\star)\big)$ ▷ Eq. (8)

      Append $\big(S_k^{(r)}, (j_k, b_k), \mathrm{RR}_k^{(r)}\big)$ to τ_k

    end for

    break

  end if

  $(\mathrm{winner}, j^\star, b^\star) \leftarrow \textsc{ResolveSealedBids}(A^{(r)})$ ▷ Highest bid

  for each $i \in \mathcal{A}$ **do**

    $\mathrm{RR}_i^{(r)} \leftarrow \textsc{RoundReward}(j_i \mid T_i)$ ▷ Issued every round, Eq. (5)

    $S_i^{(r+1)} \leftarrow \textsc{Transition}\big(S_i^{(r)}, (j_i, b_i), (\mathrm{winner}, j^\star, b^\star)\big)$ ▷ Eq. (8)

    Append $\big(S_i^{(r)}, (j_i, b_i), \mathrm{RR}_i^{(r)}\big)$ to τ_i

  end for

end for

(C) Episode close: allocation, end-of-episode rewards, budgets

$\textsc{ApplyFinalAllocation}(\mathrm{winner}, j^\star)$; $\textsc{DeductPCoins}(\mathrm{winner}, b^\star)$ ▷ Eq. (9)

$M_{\mathrm{total}} \leftarrow \textsc{CountReaccommodatedMissions}(G, j^\star)$; $M_{\max}^{\mathrm{global}} \leftarrow \textsc{GlobalMaxPossibleMissions}(G)$

for each $i \in \mathcal{A}$ **do**

  Set IR_i per Eq. (6) and GR_i per Eq. (7).

  Add $(\mathrm{IR}_i + \mathrm{GR}_i)$ to the last reward entry of τ_i

end for

return $(\mathrm{winner}, j^\star)$ and per-agent logs $\{\tau_i\}$.

Table 2
Reward adjustments.

Agent role	Cost weight (w_1)	Bonus weight (w_2)	Lose penalty weight (w_3)
Optimizer	0.5	1.0	1.0
Achiever	0.0	1.5	1.0
Economizer	1.5	0.5	1.0
Normalizer	0.5	0.5	0.5

Additionally, DRL-MANF's auction mechanism incorporates a dynamic, iterative bidding structure enhanced by a virtual currency termed P-coins. Contrary to conventional auction methods, P-coins in DRL-MANF are deducted exclusively upon winning a bid, allowing continuous strategic recalibration across negotiation rounds without immediate budgetary constraints. This innovation promotes nuanced strategic behaviors, enabling agents to adaptively refine their bidding patterns based on real-time competitive feedback, thus achieving an unprecedented balance between individual objectives and collective airspace efficiency.

3.1. Agent roles

Before the negotiation process begins, each DO is assigned a specific role that determines its strategic bidding behaviour and decision-making logic within the DRL-MANF. Four distinct roles have been defined to reflect realistic variations in operator priorities and attitudes towards rescheduling: Achiever, Optimizer, Economizer, and Normalizer. In DRL, the reward mechanism plays a crucial role in shaping agent behaviour. Therefore, in this framework, each of these roles is associated with a unique reward structure, influencing how DOs evaluate and adjust their bidding strategies.

The ideal strategic positioning of the four agent types is summarised below:

- **Achiever** The Achiever agent prioritises mission success above all else, aggressively competing for airspace slots with minimal concern for resource expenditure. Its behaviour reflects scenarios in which mission completion is critical or highly valuable, and the associated operational or economic costs of rescheduling are negligible compared to the importance of successfully executing the planned mission. Consequently, Achievers tend to frequently win auctions and achieve a high acceptance rate, but at significantly elevated costs. This aggressive, risk-tolerant bidding behaviour ensures maximum flexibility in mission scheduling, but it can lead to rapid depletion of allocated bidding resources (P-coins).
- **Economizer** The Economizer agent emphasises cost-effective decision-making, prioritising conservative bidding to minimise resource expenditure. This role models operators who assign higher weight to P-coins economy considerations or have limited willingness to adjust mission schedules. Economizers actively avoid high-cost bidding scenarios unless the potential reward justifies the expenditure, resulting in fewer auction wins but greater preservation of resources. Consequently, they participate selectively and strategically, bidding only in situations where the return on investment clearly favours mission rescheduling.
- **Optimizer** The Optimizer agent adopts a balanced bidding approach, carefully weighing mission acceptance probabilities against the cost implications of rescheduling. This role represents operators with moderate flexibility in mission timing who aim for a strategic compromise between securing airspace slots and conserving bidding resources. Optimizers dynamically evaluate auction contexts, balancing short-term gains against long-term strategic positioning. This measured behaviour enables the Optimizer to maintain stable performance over time, consistently securing beneficial solutions without excessive resource depletion.
- **Normalizer** The Normalizer agent serves primarily as a baseline or control reference within the negotiation framework. It therefore values winning, cost-efficiency, and mission fulfilment equally, without the extreme biases exhibited by the Achiever, Economizer, or Optimizer. This measured behaviour provides a realistic reference point: it mimics a typical operator who neither overpays to secure every bid nor disengages to save coins.This agent adheres strictly to predefined bidding patterns without actively pursuing competitive advantage, reflecting a neutral or indifferent operator profile that lacks incentives for strategic bidding. The Normalizer's passive behaviour provides a stable benchmark against which the performance and adaptive learning of the other agents can be objectively assessed. Its inclusion ensures that improvements or changes observed in other agent roles can be confidently attributed to their distinct reward structures and strategic adaptations.

These agents collectively represent the heterogeneous priorities and behaviours likely to be encountered in real-world U-space environments, allowing the DRL-MANF framework to realistically capture diverse operational scenarios and effectively guide operators toward strategies that improve overall airspace utilization. According to predefined agent roles, the role-specific reward adjustments shown in Table 2.

To operationalise heterogeneous operator objectives, each agent is assigned a triple of scalar weights (w_1, w_2, w_3) that respectively modulate bidding cost, mission bonus and loss penalty (Table 2). The Optimizer adopts a balanced profile (0.5, 1.0, 1.0): it values mission acceptance but still internalises half of the bidding expenditure and residual risk. The Achiever (0.0, 1.5, 1.0) neglects cost altogether, reflecting missions that willingly overspend to guarantee execution yet remain sensitive to failure. In contrast, the Economizer (1.5, 0.5, 1.0) minimises expenditure, tolerates moderate opportunity gains, and is indifferent to forfeiture: representative of low-margin logistics. Finally, the Normalizer (0.5, 0.5, 0.5) receives only half of all the reward types and therefore serves as a behavioural control. By injecting these priors directly into the reward function, the framework steers policy search towards role-consistent bidding strategies while preserving inter-agent competition.

Table 3
PPO hyperparameters.

Parameter	Value
Learning rate	1.0×10^{-4}
Clip parameter ϵ	0.2
Epochs per update	5
Initial entropy coefficient	0.09
Mini-batch size	10
Discount factor γ	0.99
GAE smoothing λ	0.95
Maximum gradient norm	0.5
Actor hidden layers	[256, 128, 64]
Critic hidden layers	[256, 128, 64]
Optimiser	ADAM

3.2. Learning algorithm

The DRL-MANF trains the four role-specific DO agents with a PPO actor-critic implemented in PYTORCH (Paszke et al., 2019). The key design decisions are grouped into three themes-algorithmic motivation, software framework, and network realization-and are summarised below:

- Algorithmic motivation: PPO is adopted because its clipped surrogate objective effectively suppresses the policy oscillations that hampered preliminary runs with naïve policy-gradient updates, while its actor-critic formulation enhanced with Generalised Advantage Estimation(GAE) offers the favourable bias variance trade-off required in the non-stationary, multi-agent auction. Moreover, PPO accommodates the factorised action space of each negotiation round: the joint log-probability of a chosen solution and its corresponding discrete bid is computed as the sum of two categorical heads and can therefore be clipped in a single, coherent operation, yielding a stable yet responsive learning signal for all four role-specific agents.
- Software framework: The learning pipeline is implemented in PyTorch whose eager, dynamic computation graphs handle the variable-length control flow that arises when auctions terminate as soon as only one bidder remains active. Given that the networks comprise three modest hidden layers, inference saturates CPU, whereas PyTorch's vectorised kernels exploit multi-threading efficiently. Tight coupling with the Mesa discrete-event simulator is also simplified: tensors can be exchanged with native Python objects at every time-step without the tracing barriers or graph/eager conversions, resulting in a lean software stack that accelerates experimentation.
- Network realisation and learning signal: The actor and critic share fully connected layers of 256, 128 and 64 ReLU units before bifurcating into their respective heads; weights are Xavier (Glorot and Bengio, 2010) initialised and updated with Adam (Kingma and Ba, 2017). At each round the actor outputs a categorical distribution over candidate solutions and, conditional on that choice, a second categorical distribution over bid levels; their log probabilities are aggregated for the PPO ratio. Advantages are standardised within every mini-batch, entropy regularisation is geometrically annealed across episodes to balance exploration and exploitation, and gradient norms are clipped to ensure numerical stability. Together, these design choices enable heterogeneous reward definitions to translate into role-consistent bidding strategies while preserving convergence of the overall auction dynamics.

3.3. Training protocol

Building on the learning algorithm in Section 3.2, the present section details the numerical schedule that drives policy updates and the auction environment in which the four DO agents interact. Table 3 collects the hyperparameters that determine how frequently and how aggressively the actor-critic is updated. Together, they balance exploration against stability, ensure bounded parameter shifts, and regulate gradient magnitudes.

The learning rate and clip parameter jointly set a trust region that tempers policy shifts; four epochs per update recycle every sample exactly four times, striking a bias variance compromise. Mini-batch size, gradient clipping, and entropy decay keep optimisation numerically stable, while (γ, λ) control the temporal credit assignment in GAE.

The quantitative settings above follow the "standard landing zone" for PPO; nevertheless, each number carries a predictable behavioural trade-off:

- **Learning rate** (10^{-4}). Raising it accelerates convergence but can trigger divergence; lowering it slows learning yet improves stability.
- **Clip parameter** ($\epsilon = 0.2$). A larger value lets the policy jump farther between updates (fast but risky); a smaller value yields conservative, slower progress. The chosen 0.2 is the conventional PPO default.
- **Epochs per update** (5). Re-using each mini-batch more than four times risks over-fitting; fewer passes waste data. Values in the 3-5 range are typical, and 5 strikes the bias variance compromise.
- **Mini-batch size** (10). Bigger batches reduce gradient noise and improve stability at the expense of compute; smaller batches speed each step but may introduce high variance.

Table 4
Negotiation-environment configuration.

Parameter	Value
Name of agents	JUNO, UTAH, SWORD, OMAHA
Agent types	Achiever, Optimizer, Economizer, Normalizer
Training episodes	300
Rounds per episode (max.)	30
Initial P-coin budget	100

- **Max gradient norm** (0.5). Increasing the clip limit admits larger updates yet can destabilise training; tighter clipping guards against spikes but may slow learning. The 0.5 threshold is a common default in PyTorch RL baselines.
- **Initial entropy coefficient** (0.09, geometrically decayed). A higher coefficient promotes broader exploration early on; a lower one forces early exploitation and risks premature convergence.
- **Discount factor** γ (0.99). Values closer to 1 emphasise distant rewards and lengthen convergence; smaller values bias the agent toward immediate gains. The 0.99 setting is standard for continuing control tasks.
- **GAE smoothing** λ (0.95). Increasing λ lowers variance but introduces estimation bias; decreasing it does the opposite. The $(\gamma, \lambda) = (0.99, 0.95)$ pairing is the canonical PPO choice.

These guidelines reveal that most numbers already occupy well-tested “safe zones”; in practice, only the learning rate or clip parameter are tuned first when convergence is either sluggish (raise the rate) or unstable (lower the rate or ϵ).

Table 4 specifies the boundary conditions of every negotiation episode: agent composition, horizon length, and budget constraints. These settings create a finite-horizon, budget-constrained game that mirrors practical U-space operations.

At the start of each episode the four heterogeneous agents receive 100 P-coins and bid in a sealed auction for at most 30 rounds. This uniform allocation ensures comparability among agent roles by isolating the effects of learned bidding strategies. In real-world deployment, however, P-coin endowments could be dynamically adjusted according to mission priority or operator performance - for example, rewarding DOs that consistently execute their planned trajectories and thereby reduce operational uncertainty. Episodes terminate prematurely if only a single bidder remains. Over 300 episodes, the recorded trajectories supply sufficient diversity for the entropy-annealed PPO learner to converge the global mission-acceptance ratio without overfitting to short-term negotiation patterns. Together, the hyperparameter schedule and environment configuration realize a training loop that is stable (bounded updates, clipped gradients), efficient (small batches, CPU-only), and realistic (budgeted, finite-round auctions), thus bridging the methodological foundations of Section 3.2 with the empirical results that follow.

4. Experiments

This section reports the empirical evaluation of DRL-MANF. First, it presents the training result analysis, covering agent performance and reward distribution. Learning curves and reward traces for the four agent archetypes oscillate during the early exploratory phase but converge steadily thereafter. Second, it provides the validation analysis including global solution decomposition, agent performance, reward distribution, and behavioural patterns. Overall, the learning-enabled auction mechanism yields a marked improvement in airspace utilization.

4.1. Training result analysis

The training results are analysed from multiple perspectives to evaluate the effectiveness of the DRL-MANF. Through these analyses, the aim is to validate whether the negotiation framework effectively optimizes airspace utilization while maintaining strategic competition among DOs. The results also provide evidence of how different reward structures shape the learning behaviour of each agent type. The analysis focuses on the following key aspects:

4.1.1. Agents performance

Examining the number of times each DO successfully wins the auction, provides insights into how different agent roles perform under competitive bidding conditions. The Fig. 5 illustrates the total number of wins achieved by each DO type throughout the 300 training episodes. The empirical negotiation framework training outcomes mirror the behavioural assumptions embedded in the four agent archetypes:

- Achiever secures the greatest share of successful wins 49.3 %, as expected, due to its unrestricted spending of P-coins. This agent is designed to prioritise winning over cost efficiency, leading to frequent success in securing airspace allocations.
- Optimizer ranked second, 24.3 %, it balances assertive bidding with prudent expenditure. Its intermediate performance reflects a strategic trade-off between mission acceptance and budget preservation.
- Economizer wins with a lowest frequency 12.0 %, consistent with its cost-minimizing behaviour, which results in more conservative bidding. thus the bidding strategy limits competitive intensity. Although this restraint reduces costs, it correspondingly diminishes the probability of winning negotiations.

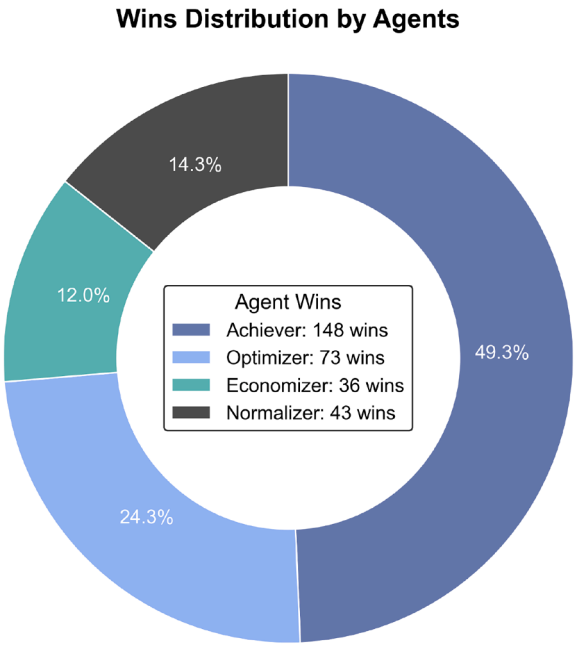

Fig. 5. Training result: wins distribution by agents.

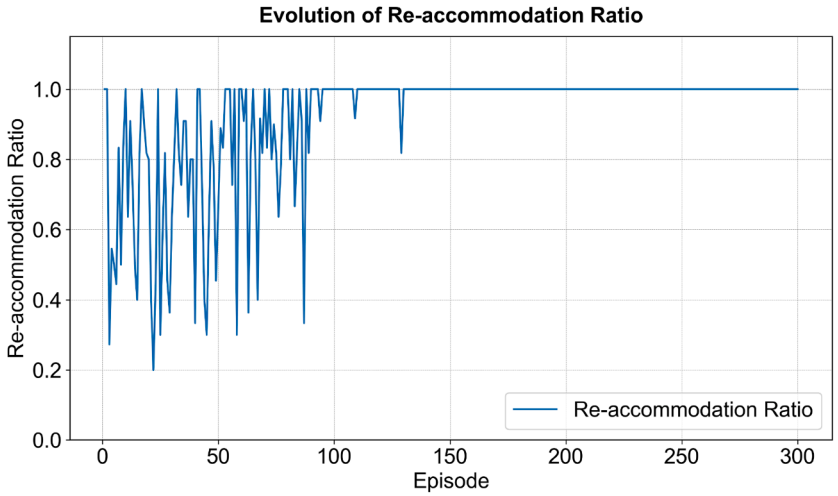

Fig. 6. Training result: evolution of re-accommodation ratio.

- Normalizer, functioning solely as a reference baseline, wins slightly more than the Economizer 14.3 % due to its special reward structure, it acts more reasonably, neither aggressively nor conservatively, yielding a measured, economically rational bidding profile that approximates a prudent real-world operator.

Evaluating the proportion of successfully reallocated flight plans serves as a metric for overall airspace optimization. Fig. 6 depicts the evolution of the Re-Accomodation Ratio (RAR) of the final executed solution over training episodes. This metric directly measures the efficacy of the negotiation framework in optimizing airspace usage. Notably, during the initial episodes, the RAR fluctuated markedly, indicating that agents were still refining their bidding strategies. After approximately 150 episodes, the RAR consistently exceeds 0.8 reach 1.0 (meaning every previously rejected mission was re-accommodated), demonstrating that the reward structure effectively guides agents toward solutions that maximize airspace utilization. Moreover, the sustained high RAR across different winning decision agents (DOs) suggest that, irrespective of which agent prevails, the resulting solution converges toward a more optimized airspace allocation.

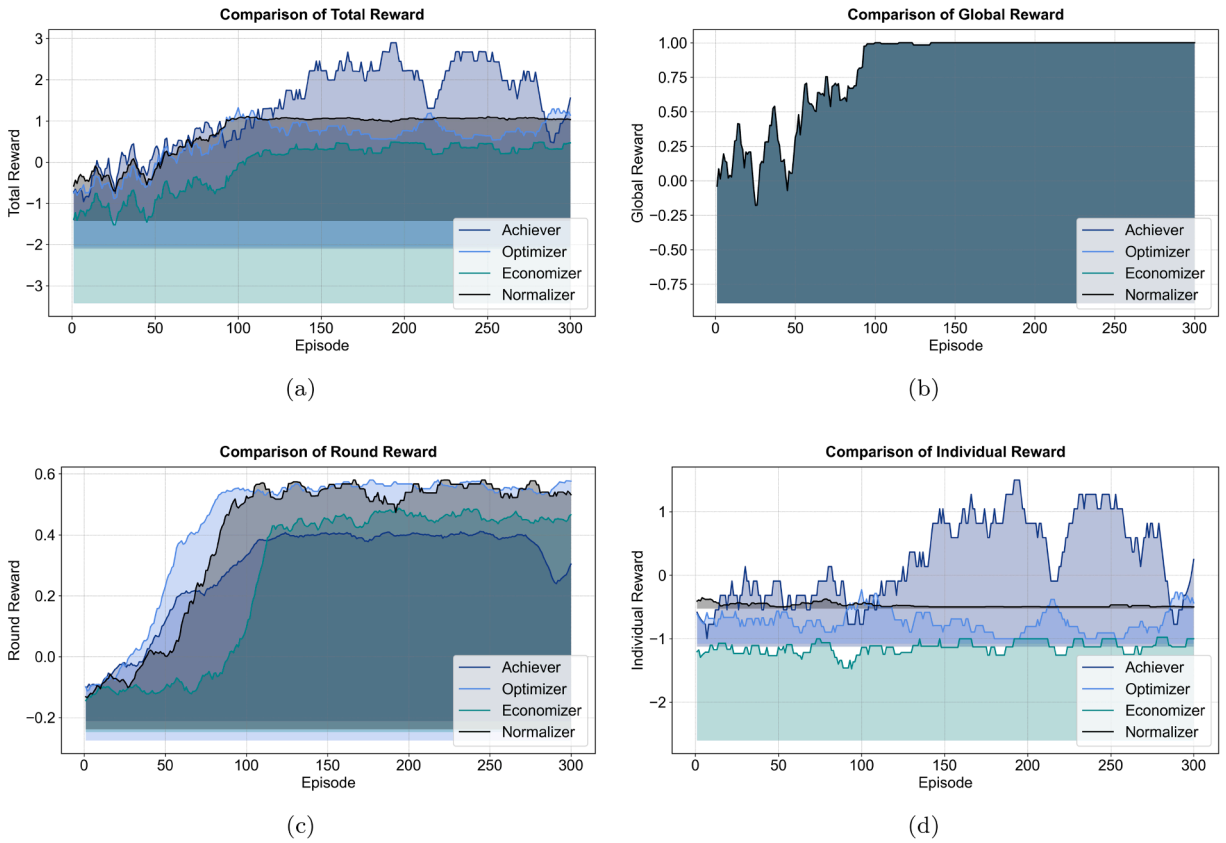

Fig. 7. Training result: rewards.

4.1.2. Reward distribution

Analysing the cumulative rewards obtained by each DO over the course of 300 training episodes, revealing differences in learned bidding behaviours. The reward dynamics offer a direct reflection of the effectiveness of the DRL framework, as well as the impact of different reward structures on each DO's performance. Fig. 7 plotted after a centred 10 episodes moving average to attenuate stochastic episode-to-episode noise, reduces visual clutter so that multiple agents and reward categories can be compared on a common axis.

Fig. 7a presents the reward progression for four different DO types. Each DO's rewards fluctuate over time as they learn from negotiation outcomes, refine their bidding policies, and adapt to competitive dynamics. The early training episodes (0-100) exhibit high reward fluctuations across all DO types. This variability is a result of exploration, where agents experiment with different bidding strategies to optimize their outcomes. Around episode 150, a clearer trend emerges as agents start to stabilise their strategies, leading to a reduction in extreme reward swings. By episode 250, all agents have converged to more structured reward patterns, reflecting learned negotiation behaviours. Between episodes 200 and 300, the residual oscillations primarily reflect the zero-sum nature of the bidding protocol, wherein only one agent prevails per episode. For example, near episode 220, the dip in the Achiever's curve corresponds to its loss, concurrently mirrored by a spike in the Optimizer's series indicating its win. Following these reciprocal fluctuations, both trajectories rapidly return to their pre-perturbation levels, signifying that the agents' policies have effectively converged. Nonetheless, the strict competition inherent to the negotiation mechanism still permits occasional outliers, manifesting as the observed transient deviations.

The Achiever exhibits the highest reward trajectory throughout the training process. This behaviour aligns with its aggressive bidding strategy, where it prioritises winning negotiations over cost efficiency. However, its reward pattern is also highly volatile, characterised by frequent spikes and dips. This instability suggests that while the Achiever successfully secures many negotiations, it also experiences instances where it fails to win. When the Achiever loses a negotiation, it receives a negative reward, further contributing to its fluctuating reward trend. The overall trajectory indicates that the Achiever has learned to dominate the auction system in most cases, though it still encounters occasional failures.

The Optimizer demonstrates a steady upward trend in rewards, showing an increasingly stable performance over time. In the early training episodes, the rewards fluctuate significantly, suggesting that the agent is initially exploring different bidding strategies. As training progresses, the Optimizer learns to balance cost efficiency with competitive bidding, leading to smoother reward progression. By episode 150, it exhibits a more controlled learning pattern, indicating that it has refined its decision-making process to maximize long-term gains while maintaining a sustainable bidding approach.

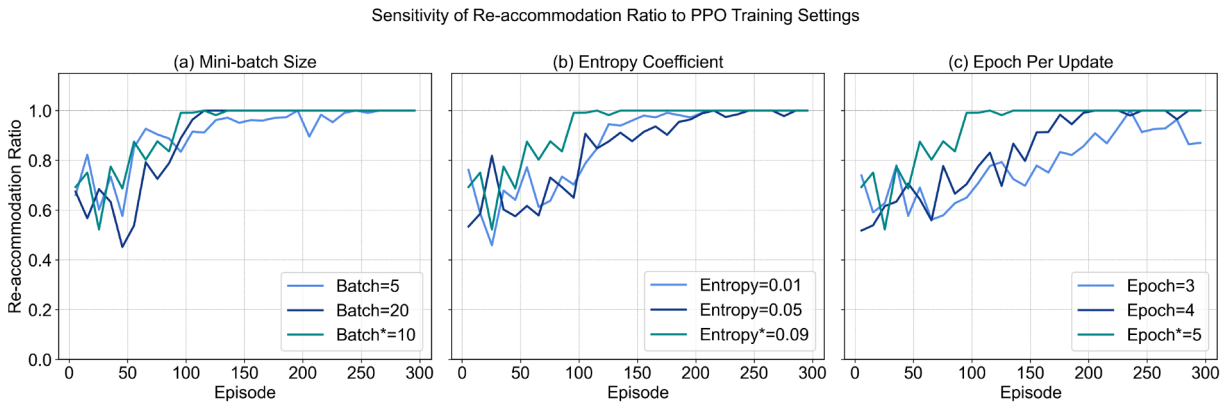

Fig. 8. Re-accommodation Ratio under PPO training sweeps.

The Economizer follows a trajectory similar to the Optimizer in the early episodes, where reward fluctuations are relatively high. However, as training advances, it stabilises at a lower reward level compared to the Achiever, the Optimizer and the Normalizer. This outcome reflects the Economizer's cost-conscious approach, where it aims to win negotiations while minimising expenditure. The learned strategy suggests that this agent selectively participates in bidding, avoiding unnecessary competition. The gradual stabilisation of rewards indicates that it effectively adapts to the negotiation environment, optimising its bidding behaviour while still ensuring a reasonable number of wins.

The Normalizer's total reward rises from negative values to zero during the initial phase and, from episode 100 onward, remains tightly clustered around 1. By assigning a uniform coefficient of 0.5 to every reward component, it adopts a neutral bidding posture–neither aggressively pursuing wins nor excessively conserving resources. This symmetric reward schema minimizes variance and solidifies the Normalizer's role as a calm, rational baseline against which the competitive behaviors of the other agents can be benchmarked.

Further decomposing rewards into Round Reward, Individual Reward, and Global Reward to understand how different reward components influence agent decision-making. The breakdown of RR, IR, and GR provides further insights into how different reward components contribute to each DO's learning process and final performance. It is important to note that the GR is determined by the final executed solution, meaning that all DOs receive the same GR per episode. Therefore, the variations in total reward trends primarily stem from differences in RR and IR, which are specific to each agent's negotiation strategy.

The remainder of Fig. 7 presents the three reward components-GR (Fig. 7b), RR (Fig. 7c), and IR (Fig. 7d)-for each decision-operator archetype over 300 training episodes. The GR rises rapidly from near zero to unity by episode 90 and remains saturated thereafter, indicating that the solution selection strategy has stabilized. However, the stability of both the bidding strategy and the individual goal strategy must be assessed via the RR and IR curves.

Fig. 7c traces the RR trajectories of the four agents archetypes. During the first 50 episodes all agents hover near, or just below, zero while their policies are still exploratory. Thereafter the Optimizer and Normalizer surge to the front, exceeding 0.55 by episode 90 and maintaining that level for the remainder of training evidence that both can reliably identify the locally optimised solution in each auction. The Achiever rises more gradually, plateaus near 0.40, indicating the selection of solutions in each auction round are not always the best one, and drifts downward after episode 260 as its win first strategy occasionally misfires. The Economizer lags initially but converges toward 0.45 after episode 120, consistent with its selective, cost conscious participation in bidding rounds.

Fig. 7d depicts the IR curves. The Achiever shows the greatest amplitude: after starting below zero it surges to sustained peaks above $+1.0$ between episodes 140–240, then plunges whenever consecutive losses occur-an unmistakable fingerprint of its aggressive bidding doctrine. The Optimizer oscillates around -0.7 early on, drifts up to roughly -0.5, and then stabilises, indicating that it has learned to temper spending while still extracting value. The Economizer remains consistently negative (≈ -1.1) because any victory incurs P-coin expenditure that outweighs the gain, whereas non-winning episodes yield zero IR. The Normalizer stays almost flat near -0.45 across all 300 episodes, reaffirming its role as a neutral baseline under the symmetric 0.5 coefficient reward scheme.

4.2. *Sensitivity to PPO training settings*

Robustness of the DRL-MANF to PPO training settings (Table 3) is evaluated by sweeping three factors: mini-batch size, entropy coefficient, and epochs per update. Fig. 8 reports the RAR under these controlled sweeps. For visual clarity, the episode-wise trajectories are shown after applying a 10-episode moving-average smoothing; the unsmoothed curves display similar qualitative trends - same plateau and ordering - but with higher variance.

Fig. 8-a varies the mini-batch size $\{5, 10, 20\}$. All curves rise quickly and settle on the same plateau, indicating that batch size has negligible effect on the asymptotic RAR. Fig. 8-b varies the entropy coefficient $\{0.01, 0.05, 0.09\}$. Larger entropy mildly accelerates early exploration (slightly faster initial rise), yet all settings converge to the same limit. Fig. 8-c varies the epochs per update $\{3, 4, 5\}$. More epochs per update shorten the transient learning phase but do not alter the terminal performance. Across all panels, RAR exceeds

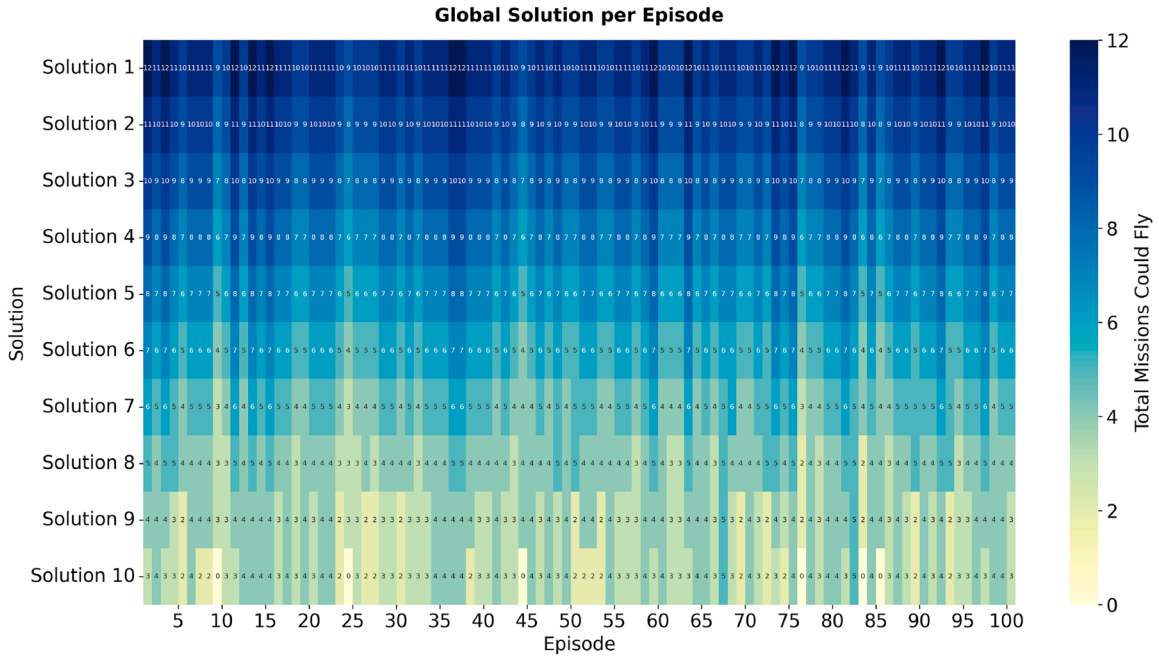

Fig. 9. Global solution.

0.8 early in training and stabilizes at ≈ 1.0 within about 150-200 episodes, with the baseline configuration (Batch = 10, Entropy = 0.09, Epochs = 5; starred in the legends) lying well within this plateau.

These results indicate that the PPO hyperparameters considered here influence the speed of convergence but not the asymptotic efficiency of the system as measured by RAR; re-accommodation performance remains effectively unchanged once training stabilizes. The observed stability is consistent with clipped PPO and advantage normalization (Section 3.2), which promote stable policy updates in a nonstationary, competitive environment.

4.3. Model validation analysis

4.3.1. Global solutions

Fig. 9 illustrates the re-accommodation capacity in the 100 global solutions produced by the AM search specifically for model validation. It is important to stress that this validation set is entirely disjoint from the solutions used during training. Each column represents one global solution generated by the AM, and the ten rows within each column correspond to the candidate solutions (Solution 1-Solution 10) that constitute that global solution. The color in each cell indicates the number of RFPs that can be re-accommodated under that solution, with the overlaid integer providing an exact count. From Fig. 9, observation can be made that Solution 1 consistently offers the highest capacity, from 9 to 12, while Solution 10 is usually the sparsest, from 0 to 5. In the reward evaluation system, Solution 1 is always the best solution.

To complement the global view provided in Fig. 9, the Fig. 10 disaggregates the negotiation outcome and analyses, for each DO, the internal composition of its candidate solutions. The title of the subfigures contains the name and role of the DO, indicating that the DO was trained as an agent of this role during the training phase. The subfigires, Fig. 10a–d represent the global solution information assigned to Achiever, Optimizer, Economizer and Normalizer respectively. It can be seen from the Fig. 10 that it has a similar trend to the Fig. 9, that is, Solution 1 usually contains the most capacity of the re-accommodation missions, and then decreases one by one until Solution 10, which contains the least capacity of the re-accommodation missions.

4.3.2. Agents performance

Fig. 11 depicts the relative success rates achieved by the four agents across the 100 competitive validaton auctions. The Achiever class secures 51 victories (51.0 %), thus capturing almost half of the available first place outcomes. The Optimizer agent follows with 35 wins (35.0 %), indicating that its price-sensitive bidding strategy remains highly effective, albeit marginally less so than the reward-maximising posture of the Achiever. In contrast, the Normalizer and Economizer policies convert their more conservative stances into just 9 (9.0 %) and 5 (15.0 %) wins, respectively.

The polarity between the two leading agents and the two budget-oriented agents corroborates the earlier mission-yield analysis: aggressive coin expenditure (Achiever, Optimizer) consistently translates into higher clearing scores, whereas risk averse bidding (Economizer, Normalizer) rarely attains the top rank. Nevertheless, the combined 15.0 % share claimed by the latter pair confirms

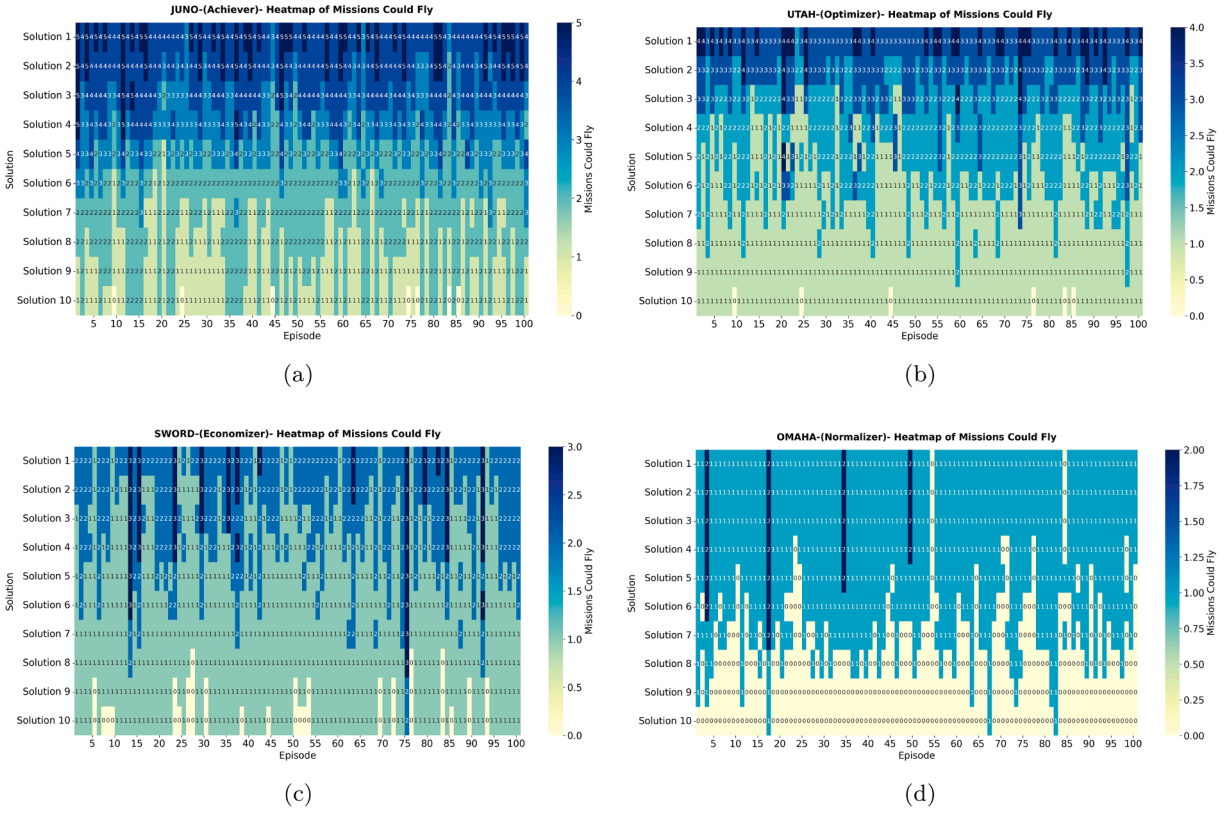

Fig. 10. Global solution analysis per agent.

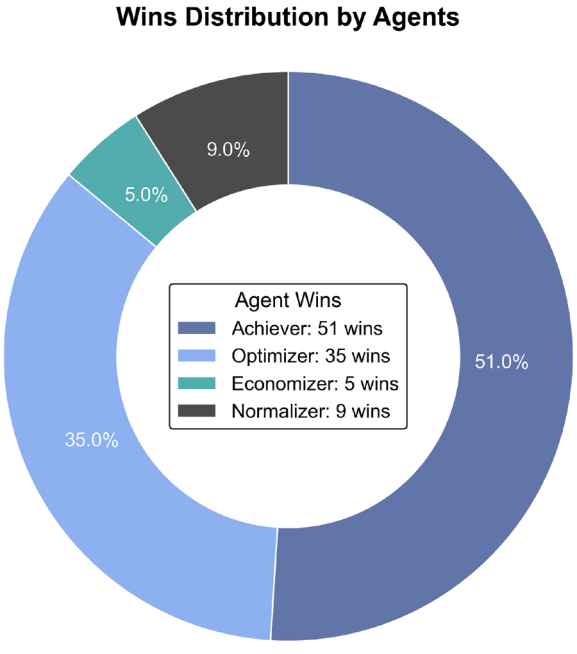

Fig. 11. Validation result: wins distribution.

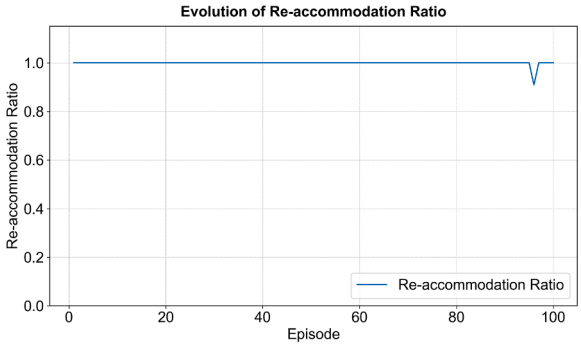

Fig. 12. Validation result: evolution of re-accommodation ratio.

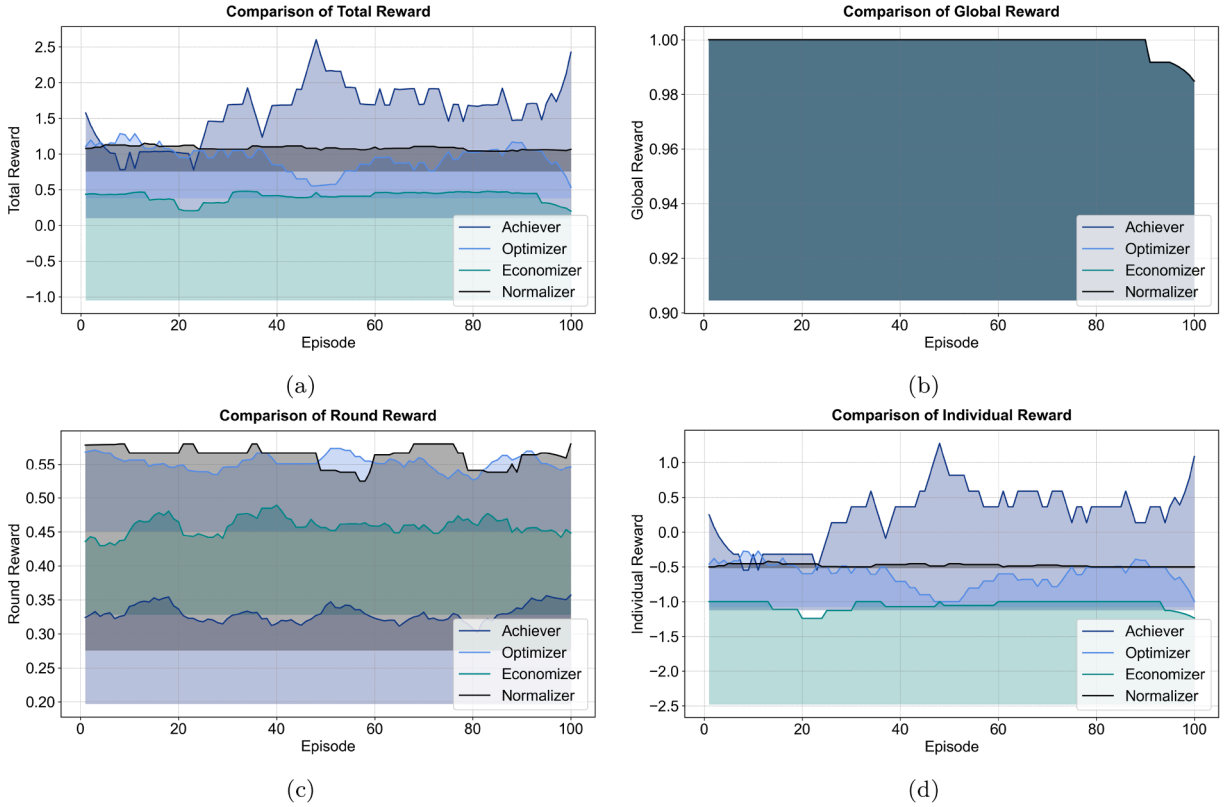

Fig. 13. Validation result: reward analysis.

that the adaptive mechanism does not preclude low-budget strategies from prevailing when market congestion or price spikes deter their high-spending rivals.

Fig. 12 tracks the episode-by-episode trajectory of the RAR, i.e. the share of initially rejected mission requests that are ultimately reassigned to an ARLW. Throughout the 100 episodes the curve remains pinned to RAR = 1.00, evidencing perfect recovery, with the exception of one transient dips (approximatly episodes 96) where the ratio briefly falls to 0.90 before rebounding in the next episode. Consistent with Fig. 2, the Batch processing rejects roughly 50 % of the submitted FPs, whereas the MANF lifts the RAR to approximately 1.0, meaning that almost every previously rejected mission is successfully re-accommodated.

4.3.3. Reward distribution

Fig. 13a and d reveals a clear total reward spectrum among the four agent archetypes: the Achiever ultimately secures the highest aggregate surplus (rising above 2.5 units), whereas the Normalizer delivers a steady, moderate gain ($\approx 1.0 \pm 0.05$) with minimal variance. The Economizer maintains a low but stable reward (≈ 0.45), and the Optimizer oscillates between aggressive and conservative regimes suffering early over bidding losses (according to Fig. 13d) before partially recovering, thus tracing a dynamic path between the Achiever's peak seeking and the Normalizer's fiscal discipline. Together, these patterns underscore that learning to calibrate

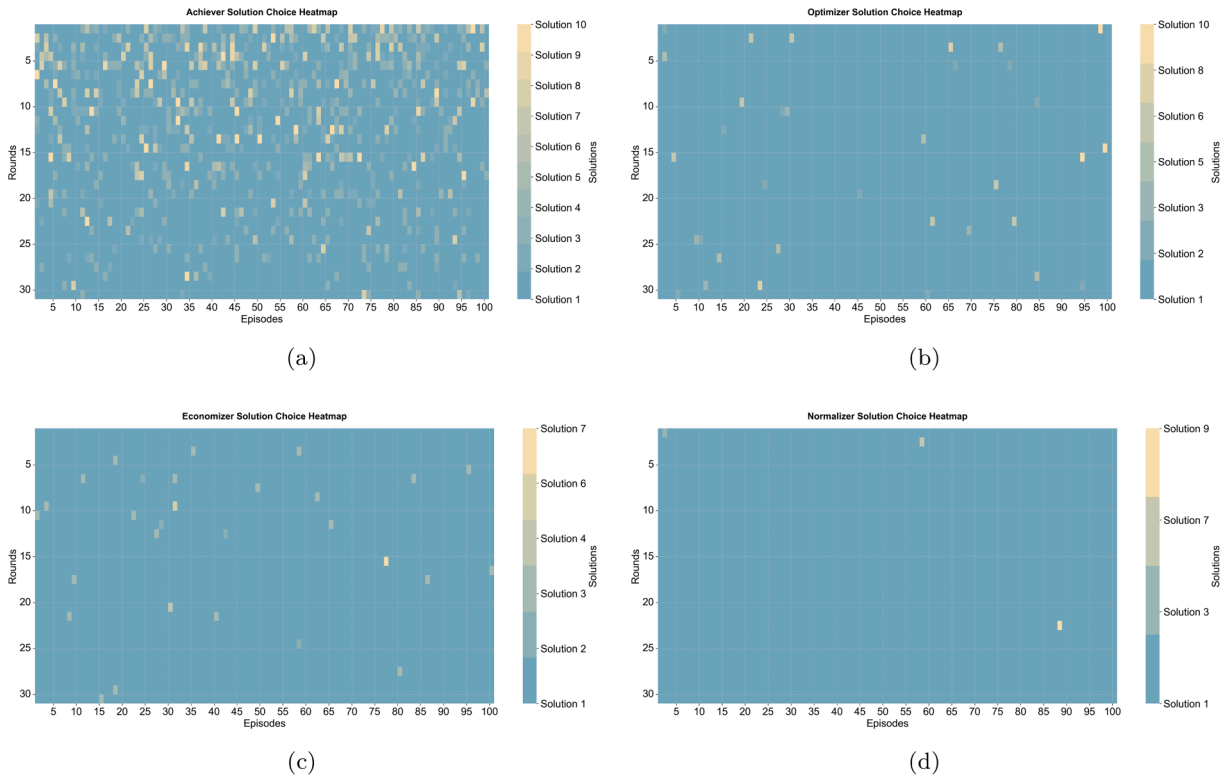

Fig. 14. Validation result: agents solution selection.

bidding strategy yields superior long run pay-off, but that DO can trade off win rate against budgetary stability by choosing the appropriate strategic archetype.

Fig. 13b juxtaposes the smoothed global reward trajectories accrued by the four agents over 100 independent validation episodes. Because the global reward is normalised to 1, the vertical scale accentuates even marginal efficiency losses. Near perfect system efficiency. For the first 95 episodes the reward curve is flat at 1.00, implying that irrespective of which agent wins a given auction, the adaptive mechanism consistently converts the full demand set into executable missions without incurring any penalty or slack capacity. In other words, strategic heterogeneity among bidders does not erode system level performance. Late episode micro degradation, around episode 90-100 (due to the 10 episode smooth process) a very small but systematic drop emerges, with the reward settling at ≈ 0.985 by episode 100. Because the four coloured traces are fully superposed, the decline is common to all agent types rather than the consequence of an individual policy failure.

Fig. 13c illustrates the round reward trajectories for all four agent archetypes over 100 episodes, revealing a persistent rank ordering and shared dynamic pattern. The Normalizer consistently achieves the highest per-round utility (≈ 0.56–0.58) with minimal fluctuation, while the Optimizer tracks just below (≈ 0.54–0.56), indicating that solution selection sensitive nearly matches the Normalizer's schedule stability gains. The Economizer occupies an intermediate, flatter band around (≈ 0.45–0.48), reflecting its risk-averse strategy's trade-off between steadiness and solution selection. Finally, the Achiever begins with the lowest reward (≈ 0.32), gradually climbs to (≈ 0.35) by episode 100, and exhibits a mild mid-horizon trough, signifying that aggressive capacity-seeking requires extended learning before converting high mission counts into commensurate per-round utility.

4.3.4. Agents behaviour

This section presents a detailed examination of agent behavior over the 100 validation episodes, with the aim of elucidating how each archetype responds once the global solution is released. Structure of the analysis along two complementary dimensions:

- Solution selection dynamics: tracking which candidate solution each agent chooses at every auction round, and how those choices evolve as their learned policies converge.
- Bidding value patterns: characterizing the temporal and structural variation in the P-coin expenditure profiles that underlie those solution selections.

By combining these perspectives, a comprehensive portrait of post solution conduct is obtained, revealing both the strategic preferences and fiscal discipline that each policy brings to bear under the adaptive mechanism.

Solution selection dynamics

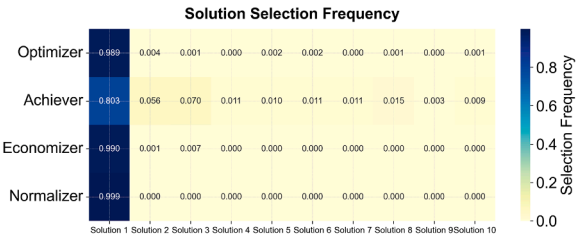

Fig. 15. Validation result: agents solution selection distribution.

Achiever: The Fig. 14a exhibits a dense mosaic of colours, signalling that the Achiever samples almost the entire ladder of bundles. Although Solution 1 remains the modal choice, higher-indexed sets (Solutions 2-8) appear with high frequency across all 30 rounds and throughout the 100 validation episodes. This pronounced volatility confirms that the Achiever continually re-optimises its allocation in response to local price signals, willingly shifting to mid and high tier solutions whenever an incremental mission slot justifies the additional coin outlay. The pattern is consistent with an exploitative policy that values capacity above budget conservation.

Optimizer: In stark contrast, the Optimizer's Fig. 14b is almost uniformly filled with the hue of Solution 1, punctuated by isolated pixels corresponding to Solutions 2, 3, 5, 6, 8 and 10. These sparse deviations cluster in rounds where bidding congestion is anticipated (e.g., early-episode first rounds and late-episode penultimate rounds), suggesting a disciplined rule: hold the base solution unless a specific market cue indicates that a short-lived switch will secure additional capacity at modest extra cost. The resulting behaviour aligns with the Optimizer's price-sensitive ethos-aggressive only when the marginal utility of an upgrade exceeds a self-imposed expenditure threshold.

Economizer: The Economizer selects the conservative Solution 1 in the majority of cases but intersperses a moderate share of Solution 3 and, less often, Solutions 2, 4 and 6 in the Fig. 14c. The upgrades occur roughly once every 5-6 episodes and are scattered across rounds, reflecting opportunistic exploitation of temporary price troughs. Absence of any selections beyond Solution 6 indicates that the agent deliberately avoids the high-tier solutons, preserving its coin inventory for sustained participation rather than maximal throughput. The pattern substantiates the Economizer's design objective: minimise spending while retaining a fallback option to marginally boost capacity when competition momentarily abates.

Normalizer: The Normalizer's Fig. 14d is an almost unbroken field of the Solution 1 colour, interrupted only three times (a single instance each of Solutions 2, 3 and 9). Such near-deterministic selection underscores an overriding preference for schedule stability: by locking into the lowest-risk solutions the agent guarantees one executable mission per auction and avoids the need for real-time recalibration. The vanishingly rare excursions to alternative solutions coincide with rounds where rival bids were anomalously weak, confirming that the Normalizer will deviate from its baseline only when it can do so without compromising its conservative coin budget.

Fig. 15 summarizes all 3,000 auction decisions (30 rounds × 100 episodes) in a single selection-frequency matrix, thereby illuminating the long-run bundle preferences of each agent archetype. Three principal patterns are evident:

1. **Near deterministic baselines.** The Normalizer (99.9 %) and Economizer (99.0 %) nearly always select Solution 1, thereby validating their design goals of schedule stability and coin conservation. Deviations are statistically negligible and never extend beyond the third rung of the capacity ladder.
2. **Price sensitive restraint.** The Optimizer likewise favours Solution 1 in 98.9 % of auctions, yet exhibits systematic-albeit minute-excursions to mid-tier bundles: at most 0.4 % for Solutions 2-3 and 0.2 % for Solutions 5-8 and 10. These rare upgrades align with the low-congestion windows identified in the heat-map analysis, indicating a calibrated willingness to pay a modest premium when a marginal capacity increase is almost certain.
3. **Diverse exploration.** The Achiever exploits the full ten-step ladder, selecting Solution 1 in 80.3 % of cases and distributing the remaining 19.7 % across higher tiers-most notably Solutions 2 (5.6 %), 3 (7.0 %), 8 (1.5 %) and 10 (0.9 %). This broad distribution corroborates its capacity-maximizing ethos: whenever additional missions can be secured, the Achiever is predisposed to escalate its bid and switch bundles.

Collectively, the frequency matrix reinforces the round-level observations: post-convergence behaviour remains heterogeneous, with the Achiever traversing the entire solution spectrum, the Optimizer exercising selective flexibility, and the Normalizer,Economizer dyad adhering to a single low risk baseline.

Bidding value patterns

Achiever: After round 25, the Fig. 16a is dominated by deep-blue cells punctuated by occasional white gaps, indicating that the Achiever almost always places very high bids (≈ 70–100) P-coins as each negotiation round draws to a close–especially in the final round. In the earlier rounds, it typically bids at medium-to-high levels (≈ 20–70) P-coins. This pattern suggests that the Achiever has learned during training how to steer the negotiation process end-to-end and has internalized the critical importance of deploying its highest bids in the closing round.

Optimizer: Predominantly dark-blue shading (≈ 70–100) P-coins characterises the Fig. 16b, yet the colour intensity exhibits discernible cyclical waves rather than the Achiever's random bursts. This structured high-bidding behaviour reflects a price sensitive calibrator: the agent remains aggressive but modulates its bid ceiling in response to perceived market saturation, leading to rhythmic

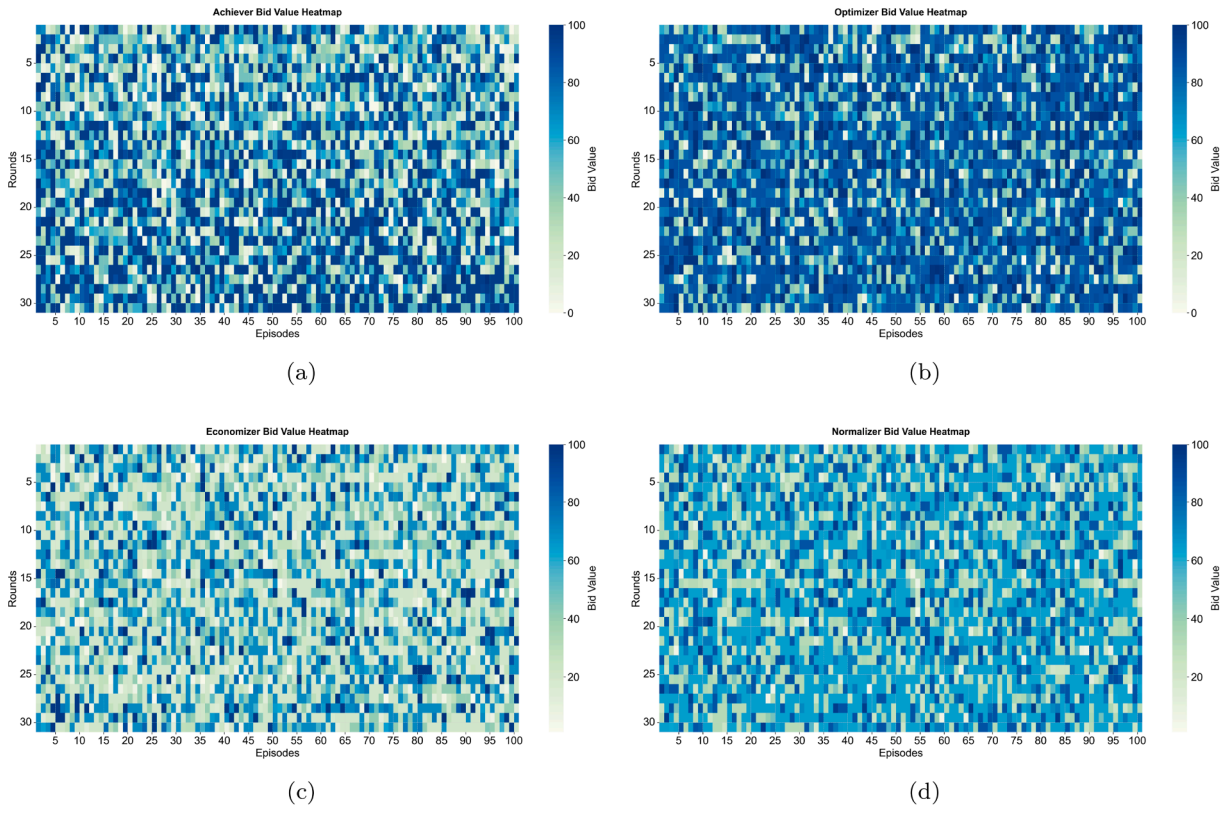

Fig. 16. Validation result: agents bid value.

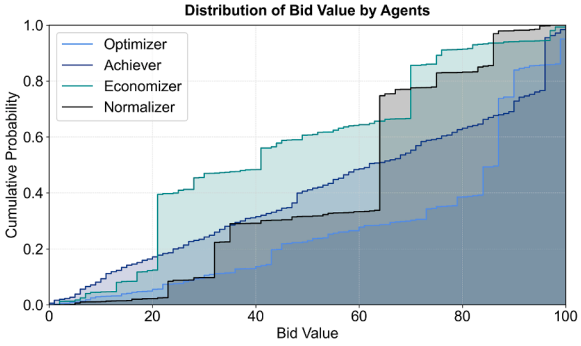

Fig. 17. Validation result: agents bid value distribution.

alternation between very high and moderately high values. The pattern substantiates the Optimizer's intermediate standing capturing substantial capacity while avoiding the extreme volatility of the Achiever.

Economizer: The Fig. 16c is dominated by pale greens (≈ 20–50) P-coins with isolated dark spikes reaching the maximum scale. Such a pattern signifies a thrift-oriented strategy: the agent generally submits minimal bids to preserve coins and only launches high bids opportunistically with a small presentage probably cause the solution selected was the highest utility. The resulting mosaic of low-level bids interlaced with sharp peaks aligns with the Economizer's lower average reward but occasional upset victories.

Normalizer: A uniform teal palette centred on the mid-range (≈ 50–60) P-coins spans the entire Fig. 16d, with only mild excursions toward either extreme. This consistency reflects the Normalizer's objective of schedule continuity: by maintaining a stable, moderate bid level it secures at least one mission per round while conserving sufficient liquidity to operate throughout all episodes. The absence of abrupt colour shifts corroborates the agent's low variance reward profile observed in the performance metrics.

Fig. 17 demonstrates pronounced heterogeneity in bidding intensity across archetypes. The Optimizer occupies the upper tail of the price spectrum: its cumulative probability remains well below 0.4 until roughly 60 P-coins and then rises steeply, implying that more than 60 % of its submissions fall in the 80-100 P-coin band. By contrast, the Achiever exhibits a broad, quasi-linear profile with

mass distributed over the entire 0-100 range; only at 90 P-coins does its cumulative probability exceed 0.75, reflecting a willingness to scale bids progressively rather than concentrate at the maximum. The Economizer ascends fastest in the lower half of the support: approximately 55 % of its bids are below 40 P-coins and 90 % below 80 P-coins, confirming its thrift-oriented posture. Finally, the Normalizer shows a two-step pattern–little activity below 50 P-coins, followed by an abrupt surge such that nearly 50 % of bids lie in the 60-80 interval–consistent with a strategy that locks into a narrow, mid-range price corridor to safeguard schedule stability while containing expenditure.

5. Conclusion and future work

This study introduces a DRL-MANF to enhance airspace resource allocation in a competitive auction-based environment. The proposed framework builds upon the strategic deconfliction service, which initially allocates airspace based on predefined rules, such as FCFS or Batch processing, to optimize the allocation of RLWs. However, despite this initial optimization, a significant number of FPs are still rejected due to airspace constraints. The AM then collects these RFPs and searches for ARLWs to provide additional opportunities for rescheduling. The DRL-MANF serves as a secondary optimization layer, allowing DOs to engage in a competitive, multi-round auction process to secure available ARLWs. By employing PPO, each DO learn role-specific bidding strategies, ensuring adaptive decision-making in a dynamic airspace environment. This approach effectively extends the initial strategic deconfliction service-based airspace optimization, providing a flexible and autonomous decision-making mechanism that improves overall mission acceptance rates while maintaining strategic competition.

Experimental results show that the agents progressively refine their bidding strategies, thereby increasing airspace utilization efficiency, all missions that were previously rejected are subsequently accommodated, yielding a 100 % re-accommodation ratio. The Achiever, Optimizer, Economizer and Normalizer agents exhibit distinct learning behaviours aligned with their predefined roles, demonstrating how different reward-driven strategies influence competitive negotiation outcomes. The high re-accommodation ratio of final executed solutions further validates that regardless of which DO wins a negotiation, the final selected solution always moves toward a more optimized airspace configuration. Across the 100 validation episodes every rejected flight was re-accommodated in at most the next round, save for one short-lived events where the ratio dipped to 0.90 and rebounded immediately (Fig. 12). The near-unity outcome is secured even though the four archetypes pursue antagonistic objectives aggressive bididng (Achiever), price sensitivity (Optimizer), budget preservation (Economizer) and schedule stability (Normalizer). Hence strategic heterogeneity, far from impeding system performance, provides a self-balancing mechanism. The cumulative distribution functions of bid values (Fig. 17) show a monotone shift from left to right, and the interval probabilities quantify those qualitative differences. The Achiever's heavy mass in the 80-100 P-coin band mirrors empirical evidence from high-stakes auctions, whereas the Normalizer's strategy resembles fixed-offer policies. The learned strategies are therefore not black boxes; they map onto recognisable real-world behaviours. These findings highlight the effectiveness of reinforcement learning in enhancing post-deconfliction airspace allocation, offering an additional layer of flexibility beyond traditional allocation mechanisms.

While the DRL-MANF framework introduces a novel approach to airspace reallocation, certain limitations remain:

1. Dependence on predefined ARLWs: The negotiation process is limited to the ARLWs provided by the AM, meaning that if no feasible ARLWs exist within the predefined search window, the framework cannot recover those RFPs. Future improvements should explore more flexible time adjustments.
2. Constraints on DO's willingness to accept ARLWs: The success of the negotiation process assumes that DOs are willing to adjust their original launch time preferences to bid for an available ARLW. However, if a DO insists on maintaining its original RLW and refuses to accept any alternative time slots, then the negotiation mechanism cannot provide a feasible solution for that DO. This highlights a fundamental limitation where negotiation can only be effective when DOs exhibit scheduling flexibility. This constraint also suggests a complementary layer in which DOs trade already authorised slots, implementing such post-allocation exchanges would require a dedicated combinatorial matching algorithm that searches for mutually beneficial swaps while maintaining trajectory separation.
3. Static role taxonomy: Each operator is irrevocably assigned to one of four archetypes with fixed reward weights. The framework cannot emulate carriers that change strategic posture over time or blend multiple objectives within a single policy.
4. Scalability constraints: The study is conducted in a four-agent environment, and further validation is required to assess performance in larger-scale airspace allocation scenarios with increased agent diversity and mission complexity.

To further enhance the DRL-MANF and address its current limitations, several key research directions are proposed:

- Flexible ARLW search and dynamic rescheduling. The current framework relies on a fixed time window when searching ARLWs. This approach may not always yield feasible rescheduling opportunities, especially in highly constrained airspace environments. Future research should explore dynamic ARLW generation mechanisms, considering additional variables such as traffic flow patterns, airspace congestion levels, and mission urgency. By integrating adaptive rescheduling policies, DOs could receive customized time slot recommendations, improving the likelihood of successful mission reassignment.
- Comparative evaluation with alternative algorithms. While PPO is the current choice for training agents, evaluating alternative reinforcement learning algorithms, such as Deep Q-Networks (DQN), Deep Deterministic Policy Gradient (DDPG), could provide deeper insights into performance trade-offs. Future work should compare these methods in terms of convergence speed, stability, and overall negotiation efficiency to determine the most effective approach for optimising DRL-MANF.

- Dynamic P-coin economy. Introduce a more flexible and comprehensive P-coin mechanism. To enhance long-term viability, P-coins can be redesigned as a cumulative, roll-over credit system. After each negotiation, operators automatically earn a set number of P-coins for every successful launch and landing, while delays or cancellations subtract credits. Any unused balance carries forward into the next auction, enabling firms to save tokens and then spend them in bulk when a high-value solution appears. Under this scheme, each operator must continually balance spending now against saving for later, ensuring regular participation in every round while still allowing a well-timed push for victory in future, more competitive negotiations.
- Scalability testing in high-density airspace. The current framework has been tested with four DOs, meaning that its performance in larger-scale, high-density airspace scenarios remains unverified. As urban air mobility (UAM) and autonomous drone operations expand, it is essential to assess whether DRL-MANF can effectively manage negotiations among a significantly larger number of DOs. Future studies should conduct scalability tests in multi-agent simulations, exploring factors such as negotiation convergence time, bid competition dynamics, and airspace capacity utilization under increasing agent populations.

These future research directions aim to further refine the DRL-MANF framework, ensuring its adaptability, efficiency, and scalability in increasingly complex airspace management systems.

CRediT authorship contribution statement

Zhiqiang Liu: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization; **Juan José Ramos Gonzalez:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization; **Ender Çetin:** Writing – review & editing, Supervision; **Jose Luis Munoz-Gamarra:** Writing – review & editing, Supervision, Project administration.

Data availability

Data will be made available on request.

Acknowledgments

This research is part of the National Spanish Project: “A Multi-Agent negotiation framework for planning conflict-free U-space scenarios” (Grant PID2020-116377RB-C22 funded by MCIN/AEI/10.13039/501100011033). Opinions expressed in this article reflect only the author’s views.

References

Bohn, E., Coates, E.M., Moe, S., Johansen, T.A., 2019. Deep reinforcement learning attitude control of fixed-wing UAVs using proximal policy optimization. In: 2019 International Conference on Unmanned Aircraft Systems, ICUAS 2019. https://doi.org/10.1109/ICUAS.2019.8798254

Bus¸oniu, L.B., Babuška, R., De Schutter, B., Bus¸oniu, L.B., Babuška, R., De Schutter, B., 2008. A Comprehensive Survey of Multi-Agent Reinforcement Learning. Technical Report.

Chen, B., Cheng, H.H., Palen, J., 2009. Integrating mobile agent technology with multi-agent systems for distributed traffic detection and management systems. Transp. Res. Part C Emerging Technol. 17 (1), 1–10. https://doi.org/10.1016/j.trc.2008.04.003

Chen, S., Evans, A.D., Brittain, M., Wei, P., 2024. Integrated conflict management for UAM with strategic demand capacity balancing and learning-based tactical deconfliction. IEEE Trans. Intell. Transp. Syst. 25 (8), 10049–10061. https://doi.org/10.1109/TITS.2024.3351049

CORUS-XUAM, 2020. Concept of operations for European u-space services - extension for urban air mobility https://doi.org/10.3030/101017682.

CORUS-XUAM Consortium, 2023. U-space concept of operations (ConOps) 4th edition. Project Deliverable D4.2. SESAR 3 Joint Undertaking. Brussels, Belgium. Horizon 2020 “CORUS-XUAM” Very Large-Scale Demonstration. https://www.sesarju.eu/sites/default/files/documents/reports/U-space%20CONOPS%204th%20edition.pdf.

Dai, W., Pang, B., Low, K.H., 2021. Conflict-free four-dimensional path planning for urban air mobility considering airspace occupancy. Aerosp. Sci. Technol. 119, 107154. https://doi.org/10.1016/j.ast.2021.107154

Dubey, R.K., Dailisan, D., Argota Sánchez-Vaquerizo, J., Helbing, D., 2025. Fairlane: a multi-agent approach to priority lane management in diverse traffic composition. Transp. Res. Part C Emerging Technol. 171, 104919. https://doi.org/10.1016/j.trc.2024.104919

EASA, 2020. Easa opinion no 01/2020: high-level regulatory framework for the u-space. https://www.easa.europa.eu/en/documentlibrary/opinions/opinion-012020.

El-Sayed, I., Khan, K., Arboleya, P., 2021. Realtime framework for EV charging coordination using multi-agent systems. In: 2021 IEEE Vehicle Power and Propulsion Conference, VPPC 2021 - ProceedingS. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/VPPC53923.2021.9699306

European Union, 2021. Commission implementing regulation (EU) 2021/664 of 22 April 2021 on a regulatory framework for the u-space. https://eur-lex.europa.eu/eli/reg_impl/2021/664/oj. Official Journal of the European Union, L 139, 23 Apr 2021, pp. 161–183.

European Union Aviation Safety Agency, 2024. Easy access rules for unmanned aircraft systems (regulations (EU) 2019/947 and 2019/945). https://www.easa.europa.eu/en/document-library/easy-access-rules/easy-access-rules-unmanned-aircraft-systems-regulations-eu.

Gao, J., Wong, T., Wang, C., 2021. Social welfare maximizing fleet charging scheduling through voting-based negotiation. Transp. Res. Part C Emerging Technol. 130, 103304. https://doi.org/10.1016/j.trc.2021.103304

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. PMLR, Chia Laguna Resort, Sardinia, Italy, pp. 249–256. https://proceedings.mlr.press/v9/glorot10a.html.

Hamissi, A., Dhraief, A., Sliman, L., 2025. A comprehensive survey on conflict detection and resolution in unmanned aircraft system traffic management. IEEE Trans. Intell. Transp. Syst. 26 (2), 1395–1418. https://doi.org/10.1109/TITS.2024.3509339

Kingma, D.P., Ba, J., 2017. Adam: a method for stochastic optimization. 1412.6980. https://arxiv.org/abs/1412.6980.

Koch, W., Mancuso, R., West, R., Bestavros, A., 2019. Reinforcement learning for UAV attitude control. ACM Trans. Cyber-Phys. Syst. 3. https://doi.org/10.1145/3301273

Konda, V.R., Tsitsiklis, J.N., 2000. Actor-critic algorithms. In: Advances in Neural Information Processing Systems.

Li, K., Niu, L., Yang, Y., Wang, Y., 2022. An approach to optimize lab-seat allocation problem based on multi-agent negotiation. In: IEIR 2022 - IEEE International Conference on Intelligent Education and Intelligent Research. Institute of Electrical and Electronics Engineers Inc., pp. 247–250. https://doi.org/10.1109/IEIR56323.2022.10050056

Li, S., Egorov, M., Kochenderfer, M.J., 2019. Optimizing collision avoidance in dense airspace using deep reinforcement learning. In: 13th USA/Europe Air Traffic Management Research and Development Seminar 2019.

Liu, Y., Zhou, Z., Naqvi, W., Chen, J., 2023. Strategic deconfliction of unmanned aircraft based on hexagonal tessellation and integer programming. J. Guid. Contr. Dyn. 46 (12), 2362–2372. https://doi.org/10.2514/1.G007459

Liu, Z., Gonzalez, J. J.R., Luis, J., Gamarra, M., 2024a. A multi-agent negotiation framework for planning conflict-free u-space scenarios. In: ICRAT.

Liu, Z., Munoz-Gamarra, J.L., Ramos Gonzalez, J.J., 2024b. Characterization of strategic deconflicting service impact on very low-level airspace capacity. Drones 8 (9). https://doi.org/10.3390/drones8090426

Munoz-Gamarra, J., Ramos, J., Liu, Z., Sobejano, A., 2023. Impact of USSPs Performance in Shared U-Space Volumes. Vol. 2023-November of *SESAR Innovation Days*. SESAR Joint Undertaking.

Munoz-Gamarra, J., Ramos Gonzalez, J., Liu, Z., 2024. U-space Contingency Management Service: Enhancing U-space Volumes Safety.

Munoz-Gamarra, J.L., Ramos, J.J., Liu, Z., UTM contingency management analysis and mitigation based on strategic vertiportsl assignment. https://doi.org/10.2514/6.2025-1807

Munoz-Gamarra, J.L., Ramos, J.J., Liu, Z., 2024. U-Space contingency management based on enhanced mission description. Aerospace 11 (11). https://doi.org/10.3390/aerospace11110876

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. 1912.01703. https://arxiv.org/abs/1912.01703.

Pellegrini, A., Sanzo, P.D., Bevilacqua, B., Duca, G., Pascarella, D., Palumbo, R., Ramos, J.J., Piera, M.A., Gigante, G., 2020. Simulation-based evolutionary optimization of air traffic management. IEEE Access 8, 161551–161570. https://doi.org/10.1109/ACCESS.2020.3021192

Pongsakornsathien, N., Safwat, N. E.-D., Xie, Y., Gardi, A., Sabatini, R., 2025. Advances in low-altitude airspace management for uncrewed aircraft and advanced air mobility. Prog. Aerosp. Sci. 154, 101085. https://doi.org/10.1016/j.paerosci.2025.101085

Puterman, M.L., 2008. Markov Decision Processes: Discrete Stochastic Dynamic Programming. https://doi.org/10.1002/9780470316887

Ramos González, J.J., Liu, Z., Muñoz-Gamarra, J.L., Pastor, E., Barrado, C., 2025. Impact assessment of the strategic planning performance in shared u-space volumes. Eur. J. Transp. Infrastruct. Res. 25 (1). https://doi.org/10.59490/ejtir.2025.25.1.7472

Ribeiro, M., Ellerbroek, J., Hoekstra, J., 2022. Distributed conflict resolution at high traffic densities with reinforcement learning. Aerospace 9. https://doi.org/10.3390/aerospace9090472

Schefers, N., Amaro Carmona, M.A., Ramos González, J.J., Saez Nieto, F., Folch, P., Munoz-Gamarra, J.L., 2020. Stam-based methodology to prevent concurrence events in a multi-airport system (mas). Transp. Res. Part C Emerging Technol. 110, 186–208. https://doi.org/10.1016/j.trc.2019.11.012

Schefers, N., González, J. J.R., Folch, P., Munoz-Gamarra, J.L., 2018. A constraint programming model with time uncertainty for cooperative flight departures. Transp. Res. Part C Emerging Technol. 96. https://doi.org/10.1016/j.trc.2018.09.013

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms http://arxiv.org/abs/1707.06347.

SESAR Joint Undertaking, 2017. U-space blueprint. https://www.sesarju.eu/u-space-blueprint.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Driessche, G. V.D., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of go with deep neural networks and tree search. Nature 529, 484–489. https://doi.org/10.1038/nature16961

Sutton, R.S., Precup, D., Singh, S., 1999. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. Artif. Intell. 112. https://doi.org/10.1016/S0004-3702(99)00052-1

Wang, T., Ma, M., Liang, S., Yang, J., Wang, Y., 2025. Robust lane change decision for autonomous vehicles in mixed traffic: a safety-aware multi-agent adversarial reinforcement learning approach. Transp. Res. Part C Emerging Technol. 172, 105005. https://doi.org/10.1016/j.trc.2025.105005

Wang, Z., Pan, W., Li, H., Wang, X., Zuo, Q., 2022. Review of deep reinforcement learning approaches for conflict resolution in air traffic control. Aerospace 9. https://doi.org/10.3390/aerospace9060294

Weigang, L., Dib, M. V.P., Alves, D.P., Crespo, A. M.F., 2010. Intelligent computing methods in air traffic flow management. Transp. Res. Part C Emerging Technol. 18 (5), 781–793. Applications of Advanced Technologies in Transportation: Selected papers from the 10th AATT Conference. https://doi.org/10.1016/j.trc.2009.06.004

Xie, N., Chen, Y., Tang, W., Chen, X.M., 2025. Multi-agent reinforcement learning with causal communication for ride-sourcing pricing in mixed autonomy mobility. Transp. Res. Part C Emerging Technol. 176, 105164. https://doi.org/10.1016/j.trc.2025.105164

Zhang, W., Liu, H., Wang, F., Xu, T., Xin, H., Dou, D., Xiong, H., 2021. Intelligent electric vehicle charging recommendation based on multi-agent reinforcement learning. In: The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021. Association for Computing Machinery, Inc, pp. 1856–1867. https://doi.org/10.1145/3442381.3449934