EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Reinforcement Learning Method for Collaborative Multi-Objective Path Planning of IIoT Drones and Unmanned Vehicles in Random Demand and Dynamic Delivery Environments

Li Qiao[1,*], Changkai Xu[1] and Bin Ni[1]

[1]Air Force Logistics Academy, Xuzhou 221000, China

Abstract

INTRODUCTION: Dynamic Industrial Internet of Things (IIoT) delivery is characterized by stochastic order arrivals, heterogeneous UAV–UGV capabilities, and real-time demand updates, making conventional static routing methods insufficient for collaborative delivery planning. Existing UAV trajectory-planning and truck–drone routing studies usually focus on either single-platform trajectory control or offline vehicle routing, and they rarely integrate stochastic IIoT demand updating, UAV–UGV rendezvous coordination, operational constraints, and multi-objective decision-making into a unified reinforcement-learning framework.

OBJECTIVES: To address this limitation, this study proposes a guidance-factor-enhanced MADDPG framework for collaborative multi-objective path planning of UAVs and UGVs in random-demand and dynamic delivery environments.

METHODS: Order arrivals are modeled by a Poisson process, while order location, payload, priority, and time-window attributes are generated through multivariate random distributions and updated by IIoT terminals every 5 s. Payload, endurance, speed, rendezvous, and task-sequencing constraints are incorporated into a normalized multi-objective function considering delivery time, delivery cost, demand satisfaction, and UAV endurance loss. A distance–demand–priority guidance factor and a reconstructed reward function are further introduced to alleviate sparse-reward effects and guide agents toward high-value demand regions.

RESULTS: Simulation and ablation results show that the reconstructed reward is essential for stable learning, while the guidance factor significantly accelerates convergence. Under 40 IIoT nodes, the proposed method converges after approximately 2,000 episodes, whereas the model without guidance information requires about 5,000 episodes. Comparative experiments further show that the proposed method improves system throughput and collaborative delivery efficiency while maintaining feasible UAV trajectories.

CONCLUSION: The proposed framework provides an adaptive reinforcement-learning solution for UAV–UGV cooperative logistics in dynamic IIoT environments.

Keywords: Industrial Internet of Things, Drone-Unmanned Vehicle Collaboration, Stochastic Demand, Multi-Objective Path Planning, MADDPG, Guidance Factor

Received on 24 June 2026, accepted on 27 July 2026, published on 19 August 2026

doi: 10.4108/eetsis.13700

*Corresponding author. Email: qiaoli8.7@163.com

1. Introduction

The Industrial Internet of Things (IIoT) enables real-time sensing, data acquisition, and collaborative control among heterogeneous industrial devices, providing an important foundation for intelligent logistics and autonomous delivery systems [1-2]. In IIoT-oriented delivery scenarios, order demand may arrive randomly, node information is updated continuously, and unmanned platforms must make decisions under limited load, energy, speed, and communication resources [3]. These characteristics make static routing methods insufficient for real-time collaborative delivery.

Existing studies have investigated IIoT logistics, UAV trajectory optimization, truck–drone routing, and multi-agent reinforcement learning [4-5]. UAV trajectory-planning methods are effective for continuous flight control and communication-aware optimization, but they usually focus on UAV-only systems [6-7]. Truck–drone routing models consider vehicle–drone synchronization, but many of them are offline optimization models and are less suitable for real-time stochastic IIoT demand [8-11]. MADDPG-based UAV coordination studies provide useful tools for multi-agent decision-making, but heterogeneous UAV–UGV constraints, rendezvous coordination, demand satisfaction, and endurance-safety loss are not always integrated into a unified multi-objective framework. In addition, multi-objective reinforcement learning and uncertainty-aware multi-agent reinforcement learning provide important theoretical references for dynamic IIoT delivery[12-14]. Multi-objective sequential decision-making studies have shown that

scalarization is suitable when the decision preference is predefined, whereas Pareto-frontier learning is more appropriate when a set of trade-off policies is required [15–16]. Robust multi-agent reinforcement learning studies further indicate that model uncertainty and state uncertainty may significantly affect policy robustness in practical multi-agent systems [17-18]. Compared with these studies, the present work adopts a weighted scalarization strategy because the delivery objectives are predefined as delivery time, delivery cost, demand satisfaction, and UAV endurance safety. However, Pareto-frontier learning and uncertainty-aware MARL remain important future directions for improving adaptive preference selection and deployment robustness in dynamic IIoT logistics environments.

Existing studies can be broadly divided into three categories. UAV trajectory-optimization studies mainly focus on continuous flight control, communication efficiency, or edge-computing performance, but usually do not consider UGV-assisted rendezvous delivery and stochastic order updating. Truck–drone routing studies emphasize vehicle–drone synchronization and routing cost, but most of them are formulated as offline or scenario-based optimization problems and are less suitable for real-time IIoT demand updates. MADDPG-based UAV coordination studies provide useful multi-agent decision-making tools, but they are commonly designed for homogeneous UAV teams and do not explicitly incorporate UAV–UGV heterogeneity, rendezvous constraints, delivery-demand satisfaction, and endurance-safety loss in a unified multi-objective framework. Therefore, the novelty of this study lies in the integration of stochastic IIoT demand modeling, heterogeneous UAV–UGV collaboration, guidance-factor-enhanced reward reconstruction, and multi-objective path planning within one MADDPG-based framework.

Recent MAPPO and QMIX studies provide strong baselines for cooperative multi-agent reinforcement learning. MAPPO has shown competitive performance in several cooperative benchmarks, while QMIX improves decentralized value-based coordination through monotonic value-function factorization. However, the present problem involves continuous UAV–UGV motion decisions, rendezvous coordination, and multi-objective delivery constraints. Therefore, this paper proposes a MADDPG-based collaborative multi-objective path planning method for IIoT drones and unmanned vehicles. The proposed

method constructs a stochastic demand model, establishes UAV – UGV collaborative constraints, introduces a guidance factor, and reconstructs the reward function to improve convergence and collaborative decision-making.

2. Algorithms and Performance Analysis for Multi-Agent Reinforcement Learning

MADDPG is adopted to address continuous-control collaborative decision-making among UAV – UGV agents under dynamic IIoT delivery environments. The MADDPG framework extends actor – critic learning to multi-agent scenarios through centralized training and decentralized execution [19], while the deterministic policy gradient mechanism provides a basis for continuous action control [20-21].For the i-th agent, the actor objective and critic loss are expressed in Equations (1) and (2), respectively:

$$J^{\pi_i} = \mathrm{E}\left[Q_i(S, a_1, \ldots, a_n) \middle| \; s_i, \pi_i \right] \quad (1)$$

$$J^{Q} = \mathrm{E}\left[\left(r_i + \gamma Q_{i'}(S', a_{1'}, \ldots, a_{n'}) \right) - Q_i(S, a_1, \ldots, a_n) \right] \quad (2)$$

where $J^{\pi i}$ and J^{Qi} are the actor objective and critic loss of agent i; Q_i and Q'_i are the current and target value functions; Si and S' are the current and next states; r_i is the immediate reward; γ is the discount factor; and $\{a_1, \ldots, a_n\}$ and $\{a'_1, \ldots, a'_n\}$ are the current and next joint action sets. By integrating reinforcement learning, the "weighted method" is employed to transform the multi-objective problem into a single-objective one, as expressed in Equation (3):

$$\min F(x) = \sum_{i=1}^{m} \omega_i f_i(x) \quad (3)$$

where F(x) is the integrated objective; f_1, f_2, f_3, and f_4 denote delivery time, delivery cost, demand satisfaction rate, and UAV endurance loss rate, respectively; and w_1-w_4 are weight coefficients with $w_1 + w_2 + w_3 + w_4 = 1$.

The original sparse task reward is insufficient for the UAV–UGV collaborative delivery problem because agents may receive delayed feedback only after order completion. Therefore, reward shaping is introduced to provide denser learning signals. Following the idea of potential-based reward shaping, the reconstructed reward consists of three parts: the negative normalized multi-objective cost, the positive guidance reward obtained from high-value completed or selected orders, and the penalty for infeasible decisions. The detailed guidance factor, reconstructed reward, and penalty function are defined in Section 3.3 and Section 3.4. This design does not replace the MADDPG policy-gradient update; instead, it improves the reward signal used by the critic during policy evaluation and actor updating.

3. MADDPG Model Construction

Figure 1 illustrates the overall architecture of the proposed MADDPG-based UAV – UGV collaborative path-planning framework. The framework consists of five main modules: stochastic demand generation, IIoT sensing and information updating, UAV – UGV collaborative constraints, multi-objective optimization, and reinforcement-learning-based policy training. First, random delivery orders are generated according to arrival time, location, payload demand, and priority attributes, and the unfulfilled order set is dynamically updated through IIoT terminals. Then, UAV and UGV status information, including position, load, speed, endurance, and task execution state, is incorporated into the collaborative decision model. Based on these inputs, the guidance factor and reconstructed reward function are introduced to enhance state representation and improve training efficiency. Finally, the MADDPG framework outputs collaborative routing and task-assignment decisions, enabling UAVs and UGVs to jointly optimize delivery time, cost, demand satisfaction, and endurance safety in dynamic delivery environments.

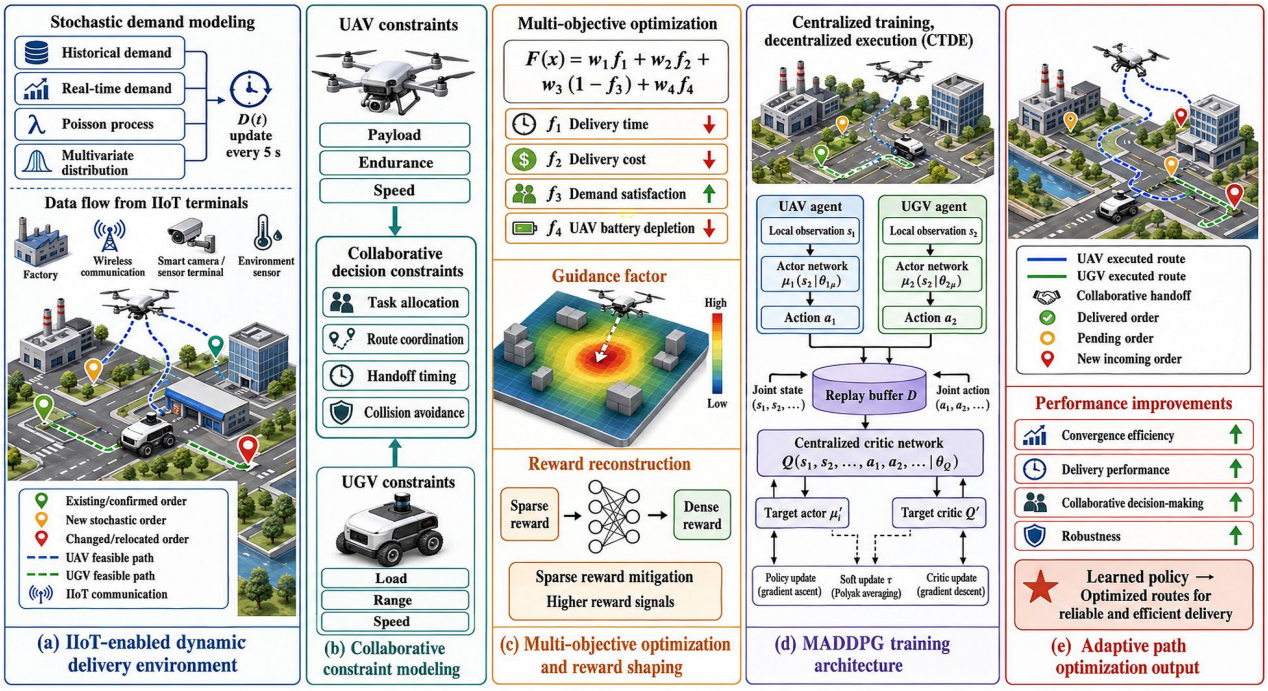

Figure 1. The architecture of the MADDPG algorithm

3.1 Environment and Requirements Model

The stochastic nature of delivery orders manifests across five dimensions—including arrival time and delivery location—and is modeled using a "Poisson process + multivariate random distribution" approach, integrated with historical and real-time demand data collected via IIoT terminals. The arrival of orders per unit of time follows a Poisson distribution with parameter λ, as shown in Equation (4):

$$P(N(t)=k)=\frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{4}$$

where N(t) is the number of new orders in time interval t; λ is the average arrival rate; and k is the number of arriving orders. The quantification of a single order is shown in Formula (5):

$$d_i=\left(x_i, y_i, t_i^{\text{start}}, t_i^{\text{end}}, q_i, \tau_i\right) \tag{5}$$

where $O_i = (x_i, y_i, q_i, T_i, t_i)$ denotes order i; xi and yi are node coordinates; qi is payload demand; Ti is the priority coefficient; and ti is the order-arrival time.

3.2 Collaborative Constraint Model

Based on the hardware characteristics of UAVs and UGVs, as well as the IIoT-enabled collaborative delivery process, these constraints are categorized into two types: equipment constraints and collaborative decision constraints. Equipment constraints are formulated based on equipment status data acquired through IIoT sensing. UAVs are required to satisfy constraints regarding payload capacity, endurance, and speed—as presented in Equations (6) through (8), respectively—while UGV constraints are outlined in Equations (11) through (11):

$$\sum_{d_i \in D_u(k)} q_i \le 0.9 Q_u^{\max} \tag{6}$$

$$L_u(k) \le L_u^{\max} \cdot \frac{1}{1+0.1 v_w(t)+0.1(\zeta(t)-1)} \tag{7}$$

$$\left[v_u^{\min}, v_u^{\max} \cdot \frac{1}{1+0.05 v_w(t)}\right] \tag{8}$$

$$\sum_{d_i \in D_g(m)} q_i \le Q_g^{\max} \tag{9}$$

$$L_g(m) \le L_g^{\max} \tag{10}$$

$$v_g^{\min} \le v_g(m,t,x,y) \le \frac{v_g^{\max}}{\eta(t,x,y)} \tag{11}$$

In Equations (6) – (11), L_u and L_g denote the actual payloads of the UAV and UGV, respectively; L_u^{max} and L_g^{max} are their maximum allowable payload capacities; $E_u(t)$ represents the remaining UAV battery energy at time step t; E^u_{min} denotes the minimum safety energy threshold required for return or rendezvous; v_u and v_g are the operating speeds of the UAV and UGV; v_u^{max} and v_g^{max} represent their maximum allowable speeds; d_{ij} denotes the travel distance between nodes i and j; and t_{ij} represents the corresponding travel time. These constraints jointly ensure that the generated path satisfies payload capacity, endurance safety, speed limitation, and UAV – UGV collaborative operation requirements.

3.3 Mathematical Model for Multi-Objective Path Planning

The UAV–UGV collaborative delivery environment is modeled as a directed graph G=(V,E), where V={0} ∪ D(t) ∪ R. Node 0 denotes the depot, D(t) is the set of unfulfilled IIoT orders at time t, and R represents optional rendezvous nodes. Let $x^g_{mij(t)}$ and $x^u_{kij(t)}$ be binary decision variables. $x^g_{mij(t)} = 1$ indicates that UGV m moves from node i to node j, while $x^u_{kij(t)} = 1$ indicates that UAV k flies from node i to node j. $z_{i(t)} = 1$ means that order i is completed.

The multi-objective path-planning problem is transformed into a normalized weighted minimization problem:

$$F(t) = w_1 f_{1'}(t) + w_2 f_{2'}(t) + w_3 \cdot \left[1 - f_{3'}(t)\right] + w_4 \cdot f_{4'}(t) + P(t) \tag{12}$$

$$w_1 + w_2 + w_3 + w_4 = 1 \tag{13}$$

where f1'(t), f2'(t), f3'(t), and f4'(t) denote the normalized delivery time, delivery cost, demand satisfaction rate, and UAV endurance loss rate, respectively. P(t) is the penalty term for infeasible decisions. The normalization is expressed as:

$$f_{r'}(t) = \frac{f_r(t) - f_{r,\min}}{f_{r,\max} - f_{r,\min} + \varepsilon} \tag{14}$$

The delivery time is defined as the maximum completion time of all UAVs and UGVs:

$$f_1(t) = \max\left\{\max_{m\in T} g_m(t), \max_{k\in T} u_k(t)\right\} \tag{15}$$

The demand satisfaction rate and UAV endurance loss rate are calculated as:

$$f_3(t) = \sum_{i\in D(t)} q_i \cdot \frac{z_i(t)}{\sum_{i\in D(t)} q_i} \tag{16}$$

$$f_4(t) = \frac{1}{|K|}\sum_{k\in K} \frac{E_{k,\max} - E_{k,\text{remain}}(t)}{E_{k,\max} - E_{k,\min}} \tag{17}$$

The guidance factor is designed according to the marginal utility of selecting an unfulfilled order under dynamic IIoT delivery demand. In collaborative delivery, an order with larger payload demand and higher priority generally has higher service value, while an order located farther away imposes higher flight time, energy consumption, and rendezvous cost. Therefore, the proposed guidance factor combines normalized demand, distance-decay utility, and priority amplification. The exponential distance-decay term is used because the marginal attractiveness of a demand node decreases nonlinearly with travel distance and energy consumption. This design is not intended to prove global optimality as a deterministic routing heuristic; rather, it serves as a potential-like reward-shaping signal that improves early-stage exploration while the final policy is still learned through the MADDPG actor–critic update.

To improve exploration efficiency, a guidance factor is introduced for each unfulfilled order:

$$G_i(t) = \frac{q_i}{q_{\max}} \exp\left[-\frac{d_i(t)}{d_0}\right]\left(1 + \chi T_i\right) \tag{18}$$

where qi is the demand of order i, di(t) is the distance to order i, Ti is the priority coefficient, and d0 and χ are scaling parameters. The reconstructed reward is defined as:

$$R(t) = -F(t) + \alpha \sum_{i\in D(t)} z_i(t) G_i(t) - P(t) \tag{19}$$

The proposed optimization process can be summarized as follows: first, IIoT terminals update the unfulfilled order set D(t) every 5 s; second, the guidance factor Gi(t) is calculated according to demand, distance, and priority; third, MADDPG outputs UAV–UGV collaborative actions; finally,

the environment updates vehicle states, order-completion states, penalties, and rewards. Through repeated interaction and policy updating, the agents learn a feasible collaborative routing strategy that balances delivery efficiency, cost, demand satisfaction, and UAV endurance safety.

3.4 Reinforcement-Learning Implementation and Constraint Handling

To improve reproducibility, the reinforcement-learning implementation is further specified in this subsection. The agent set consists of K UAV agents and M UGV agents. During centralized training, the critic receives the global state, including the unfulfilled order set D(t), order attributes, UAV and UGV positions, remaining loads, remaining UAV energy, speed states, task-completion flags, and rendezvous information. During decentralized execution, each agent makes decisions according to its own observation, including its current position, residual load, residual energy or operating state, nearby candidate orders, local guidance-factor values, and the nearest feasible rendezvous nodes.

The state vector at time step t is defined as:

$$S(t) = \{D(t),\ P_u(t),\ P_g(t),\ L_u(t),\ L_g(t),\ E_u(t),\ V_u(t),\ V_g(t),\ z(t),\ G(t)\} \quad (20)$$

where $P_{u(t)}$ and $P_{g(t)}$ are the UAV and UGV position sets; $L_{u(t)}$ and $L_{g(t)}$ are the residual load states; $E_{u(t)}$ is the remaining UAV energy; $V_{u(t)}$ and $V_{g(t)}$ are the speed states; Z(t) denotes the order-completion state; and G(t) is the guidance-factor vector. The action of each UAV includes the selection score of the next service order, the target flying direction or target node, and the operating speed. The action of each UGV includes the next movement target, speed, and optional rendezvous decision. Continuous actions output by the actor network are mapped to the nearest feasible order or rendezvous node according to the current constraint set.

Infeasible actions are handled by a two-stage mechanism. First, action projection is applied to map the selected action to the nearest feasible candidate if the violation is mild. Second, a penalty is added when the action violates payload, endurance, speed, rendezvous, or sequencing constraints. The total penalty is defined as:

$$P(t) = \beta_L C_L(t) + \beta_E C_E(t) + \beta_V C_V(t) + \beta_R C_R(t) + \beta_S C_S(t) \quad (21)$$

where $C_L(t)$, $C_E(t)$, $C_V(t)$, $C_R(t)$, and $C_S(t)$ denote violations of load capacity, energy safety, speed limit, rendezvous coordination, and task sequencing, respectively; β_L, β_E, β_V, β_R, and β_S are non-negative penalty coefficients. The rendezvous constraint requires that a UAV returns to the depot or reaches the assigned UGV rendezvous node before its remaining energy falls below the safety threshold. If the spatial distance between the UAV return position and the UGV rendezvous position exceeds the tolerance εr, or if the arrival-time difference exceeds Δτmax, a rendezvous penalty is imposed. In the baseline setting, the multi-objective weights are set as $w_1=w_2=w_3=w_4=0.25$ to avoid subjective preference toward a single objective.

4. Simulation Experiments and Results Analysis

4.1 Ablation Study

Ablation experiments were conducted to validate the contribution of the guidance factor and the reconstructed reward function, following the common evaluation strategy of removing or retaining key modules to identify their individual effects on convergence and performance [22-24]. Regarding the impact of reward shaping—as observed in Figure 2—under a scenario involving 40 IIoT nodes, the TD3 algorithm failed to function at all without the reconstructed reward function. Conversely, both the MADDPG algorithm (enhanced with information augmentation) and the TD3 algorithm (equipped with the reconstructed reward function)—represented by the blue and red curves, respectively—successfully completed the data collection task and converged to higher, more stable reward values; this demonstrates that the reconstructed reward function effectively guarantees the convergence of the algorithm. Furthermore, in the absence of guidance factor information, the convergence of cumulative rewards was noticeably slower, requiring approximately 5,000 episodes to ensure convergence. In contrast, the MADDPG algorithm required only about 2,000 episodes to achieve convergence.

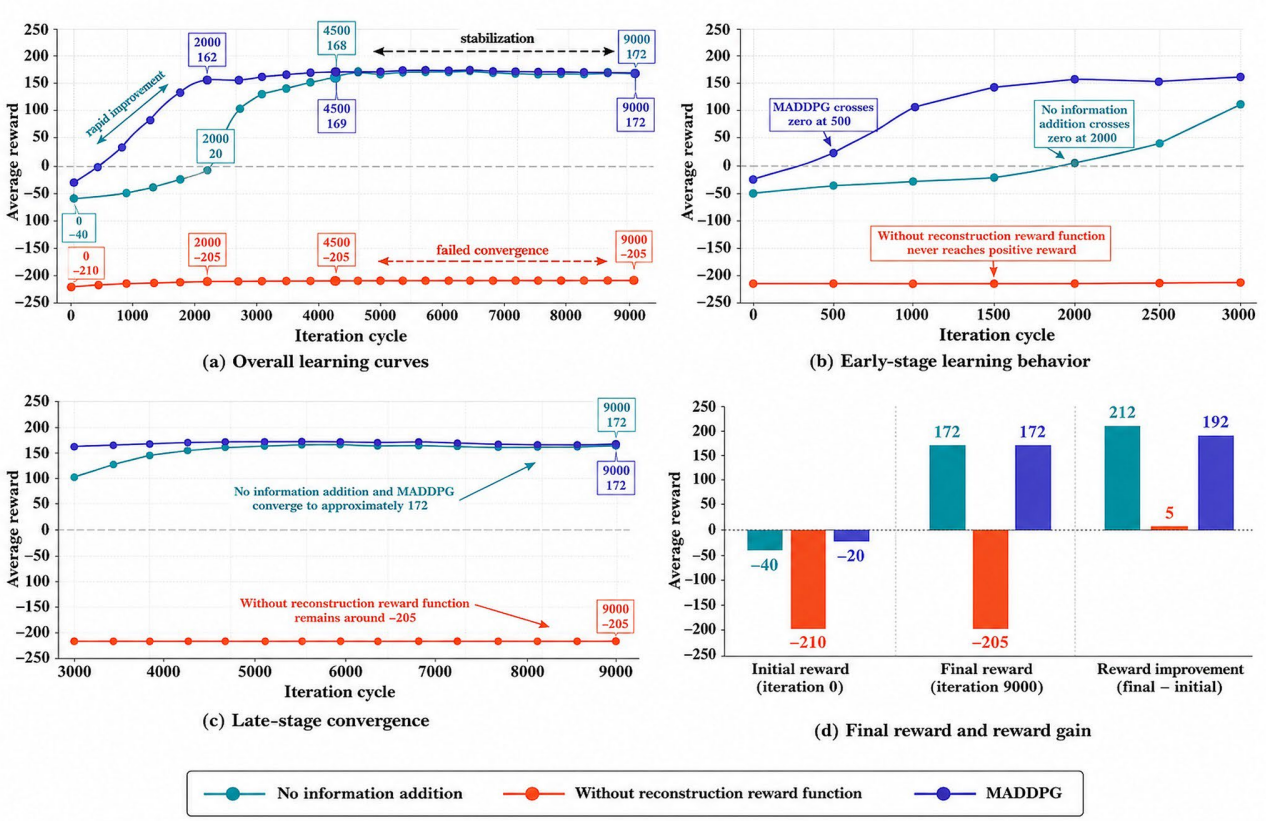

Figure 2. The Impact of Reconstructing the Multidimensional Reward Function and Introducing a Guidance Factor on Rewards

Figure 2 shows the ablation results of the reconstructed reward function and the guidance factor under 40 IIoT nodes. Without the reconstructed reward function, the learning curve remains at a low reward level and fails to form an effective delivery policy, indicating that the original sparse reward is insufficient for UAV–UGV collaborative path planning. After introducing the reconstructed reward function, the agents obtain denser feedback from task completion, constraint satisfaction, and movement toward high-value demand regions, which improves training stability. In addition, the guidance factor mainly accelerates early-stage exploration. The proposed MADDPG model reaches a stable reward level after approximately 2,000 episodes, whereas the model without guidance information requires about 5,000 episodes. These results demonstrate that the reconstructed reward ensures stable learning, while the guidance factor reduces exploration cost and improves convergence efficiency.

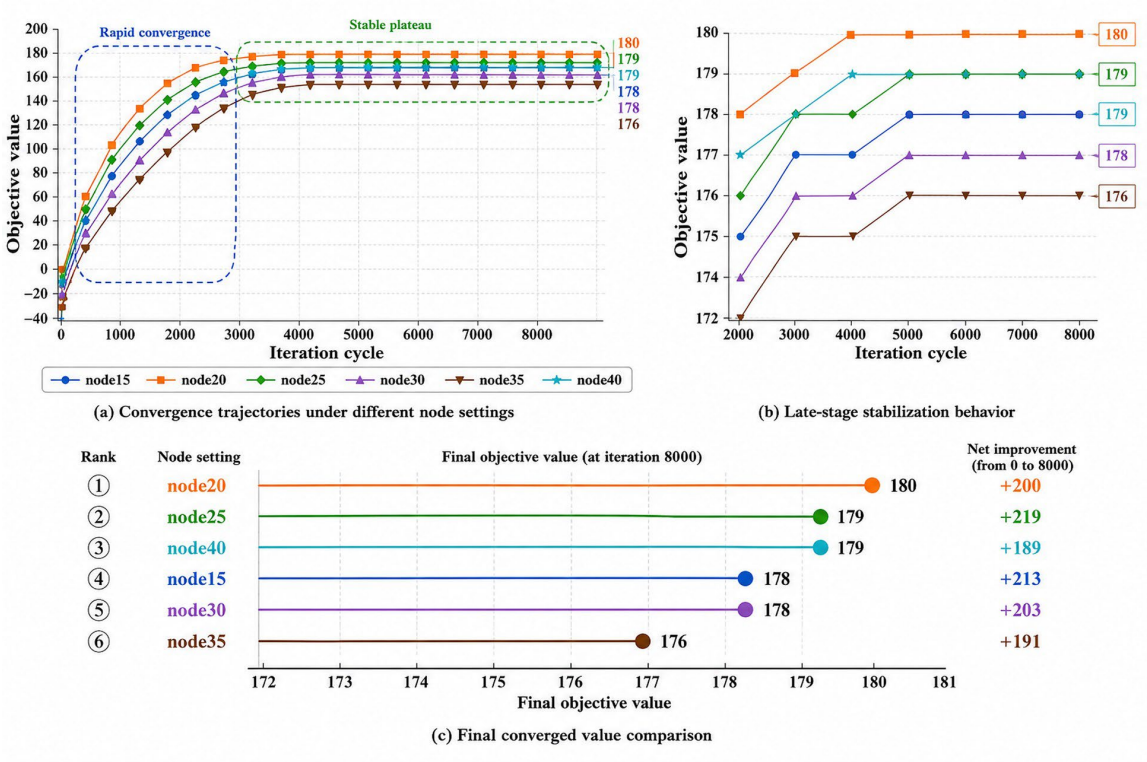

Figure 3. Total Cumulative Reward Per Cycle for Different Numbers of IIoT Nodes

Figure 3 presents the cumulative reward curves of the proposed MADDPG-based method under different numbers of IIoT nodes. The rewards increase rapidly in the early training stage and gradually converge, indicating that the agents can continuously improve their collaborative delivery policies through interaction with the environment. When the number of IIoT nodes is small, the task-assignment space is limited and the reward curve is more stable. As the number of nodes increases, the delivery environment becomes more complex because of more random orders, dispersed demand points, route conflicts, and energy constraints, leading to stronger reward fluctuations during training. Nevertheless, all curves eventually reach stable levels, demonstrating that the proposed method maintains convergence capability and robustness under different node densities.

4.2 Comparative Experiments of Different Algorithms

To further clarify the advantage of the proposed method, the benchmark algorithms are interpreted from the perspective of collaborative delivery. Traditional greedy or nearest-priority heuristics can quickly select nearby high-priority nodes, but they lack long-term coordination and may lead to local route redundancy when orders arrive dynamically. Single-agent or weakly coordinated reinforcement-learning algorithms can improve local path selection, but they cannot fully capture the coupling between UAV service decisions, UGV rendezvous movement, endurance safety, and demand satisfaction. In contrast, the proposed MADDPG-based framework learns coordinated policies through centralized training and decentralized execution, allowing UAVs and UGVs to balance short-term delivery efficiency and long-term collaborative feasibility.

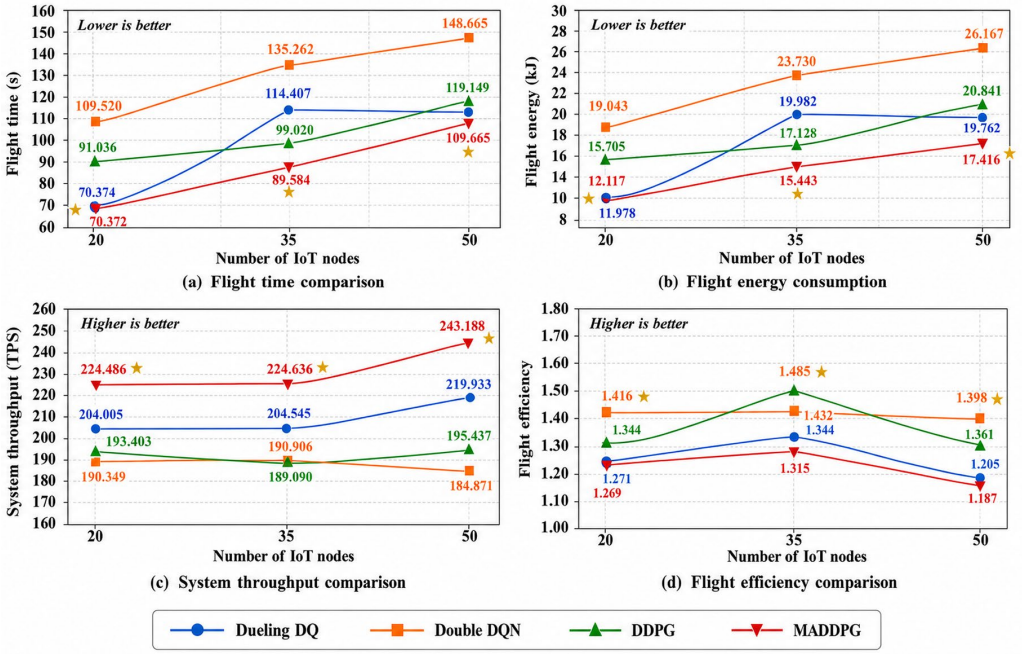

Figure 4. Comparison of Experimental Metrics for UAVs Employing Different Algorithms at a Fixed Number of Nodes

Figure 4 compares the performance of different reinforcement learning algorithms in terms of flight time, flight energy consumption, system throughput, and flight efficiency. As the number of IIoT nodes increases, both flight time and energy consumption generally increase because more delivery tasks and longer routing distances are introduced. Compared with the benchmark algorithms, the proposed MADDPG-based method achieves lower flight time and energy consumption under most node scales, indicating that the learned policy can reduce redundant detours and better consider UAV endurance constraints. Meanwhile, the proposed method obtains the highest system throughput, especially when the number of nodes reaches 50, showing stronger task-completion capability in complex delivery environments. Overall, Figure 4 verifies that the proposed method achieves better comprehensive performance by balancing delivery efficiency, energy consumption, throughput, and resource utilization.

The computational and deployment characteristics of the proposed method are also considered. The main computational cost occurs during offline centralized training, where the critic uses global state information to evaluate joint UAV–UGV actions. After training, decentralized execution only requires each actor network to output an action according to local observation and periodically updated IIoT information, which reduces the online decision burden. For larger fleets or denser node configurations, the state and action dimensions increase, and the training cost may rise accordingly. This issue can be mitigated by limiting the candidate order set to nearby high-value nodes, using parameter sharing among homogeneous UAVs, and updating global demand information through edge servers. Although the guidance factor is computed from global IIoT demand information during training, practical deployment can approximate it using periodically broadcast demand maps or local sub-region demand information.

4.3 Trajectory Optimization

Figures 5 and 6 show the optimized UAV trajectories under two different IIoT node distributions. In both scenarios, the two UAVs start from the same initial point but gradually form differentiated service regions after training, indicating that the agents can learn cooperative behaviors and avoid highly overlapped routes. In Figure 5, UAV 1 mainly serves the right-side sensor cluster, while UAV 2 covers the left-side and lower sensor clusters. The trajectories are spatially

separated and relatively smooth, suggesting that the learned policy can reduce unnecessary detours and repeated coverage. In Figure 6, although the node distribution changes, the UAVs still generate feasible and coordinated trajectories. UAV 1 mainly covers the right-side demand region, while UAV 2 adjusts its route to serve the left-side and lower-left nodes. These results indicate that the proposed method is not limited to a fixed demand scenario and can adaptively generate stable, smooth, and spatially coordinated trajectories under different IIoT node distributions.

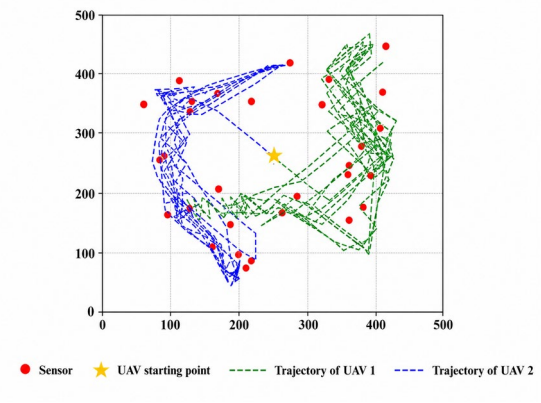

Figure 5. Optimized trajectory of UAV in Scenario 1

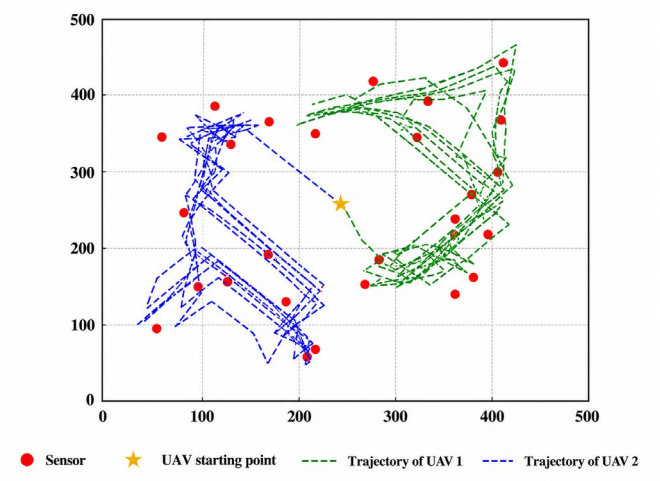

Figure 6. Optimized trajectory of UAV in Scenario 2

The two trajectory scenarios also indicate the generalization potential of the learned policy. Although the spatial distribution of IIoT nodes changes between Scenario 1 and Scenario 2, the UAVs still form differentiated service regions and avoid highly overlapped routes. This suggests that the proposed reward reconstruction and guidance factor can support adaptive path planning under different node configurations.

The current simulation does not fully include communication delay, packet loss, wind disturbance, localization error, dynamic obstacles, or temporary no-fly zones. These factors may affect real-time decision-making, UAV energy consumption, route feasibility, and UAV–UGV rendezvous reliability. Therefore, the present results should be regarded as simulation-based validation of the proposed decision-making framework. Future work will introduce communication-delay models, wind-affected energy consumption, obstacle-avoidance constraints, and no-fly-zone restrictions into a digital-twin-based IIoT logistics platform.

5. Conclusion

This study investigated UAV–UGV collaborative multi-objective path planning for random-demand and dynamic IIoT delivery environments. The proposed framework addresses a practical problem in intelligent industrial logistics, where delivery orders arrive stochastically, demand information is updated in real time, and heterogeneous unmanned platforms must coordinate under payload, endurance, speed, and rendezvous constraints. The main conclusions are as follows.

1. A guidance-factor-enhanced MADDPG framework was established by integrating stochastic IIoT demand modeling, UAV–UGV collaborative constraints, and normalized multi-objective optimization. Compared with conventional static routing or single-platform trajectory-planning methods, the proposed framework can dynamically update the unfulfilled order set every 5 s and support collaborative decision-making under random demand.

2. The reconstructed reward function and the guidance factor improved the learning process. The reconstructed reward provided continuous feedback from task completion, constraint satisfaction, and multi-objective performance, while the distance–demand – priority guidance factor directed agents toward high-value demand regions. The ablation results showed that the proposed method converged after approximately 2,000 episodes under 40 IIoT nodes, whereas the model without guidance information required approximately 5,000 episodes.

3. The comparative and trajectory experiments verified the effectiveness of the proposed method. Compared with the

benchmark algorithms, the proposed MADDPG-based method achieved better overall performance in terms of flight time, energy consumption, system throughput, and flight efficiency. The optimized trajectories further showed that UAVs could form differentiated service regions and generate feasible, smooth, and spatially coordinated paths under different node distributions.

Future work will further evaluate the proposed framework under broader and more challenging delivery scenarios, including uneven spatial demand distributions, burst order arrivals, partial UAV or UGV failures, communication interruption, and temporary node unavailability. In addition, the current guidance factor is manually designed according to demand, distance, and priority. Although the ablation study verifies its effectiveness, an adaptive or learnable guidance mechanism may provide a more flexible solution when demand patterns change significantly. Therefore, future research will explore meta-learning or attention-based mechanisms to learn guidance weights automatically from historical IIoT delivery data. Due to project-related configuration and platform-interface restrictions, the complete code and simulation environment are not publicly released at this stage. To improve reproducibility, the revised manuscript provides detailed state definitions, action definitions, reward components, penalty functions, constraint-handling rules, and scenario-generation settings.

References

[1] SISINNI E, SAIFULLAH A, HAN S, et al. Industrial Internet of Things: Challenges, opportunities, and directions[J]. IEEE Transactions on Industrial Informatics, 2018, 14(11): 4724–4734.

[2] XU X, LU Y, VOGEL-HEUSER B, et al. Industry 4.0 and Industry 5.0—Inception, conception and perception[J]. Journal of Manufacturing Systems, 2021, 61: 530–535.

[3] MADDIKUNTA P K R, PHAM Q V, PRABADEVI B, et al. Industry 5.0: A survey on enabling technologies and potential applications[J]. Journal of Industrial Information Integration, 2022, 26: 100257.

[4] ZHAO G, WANG Y, MU T, et al. Reinforcement-learning-assisted multi-UAV task allocation and path planning for IIoT[J]. IEEE Internet of Things Journal, 2024, 11(16): 26766–26777.

[5] HAN J, LIU Y, LI Y. Vehicle routing problem with drones considering time windows and dynamic demand[J]. Applied Sciences, 2023, 13(24): 13086.

[6] CUI H, LI K, JIA S, et al. Dynamic collaborative truck-drone delivery with en-route synchronization and random requests[J]. Transportation Research Part E: Logistics and Transportation Review, 2024, 192: 103802.

[7] WANG F, LI H, XIONG H. Truck-drone routing problem with stochastic demand[J]. European Journal of Operational Research, 2025, 322(3): 854–869.

[8] NING Z, YANG Y, WANG X, et al. Multi-agent deep reinforcement learning based UAV trajectory optimization for differentiated services[J]. IEEE Transactions on Mobile Computing, 2024, 23(5): 5818–5834.

[9] FAN C, XU H, WANG Q. Multi-agent deep reinforcement learning for trajectory planning in UAVs-assisted mobile edge computing with heterogeneous requirements[J]. Computer Networks, 2024, 248: 110469.

[10] JU T, LI L, LIU S, et al. A multi-UAV assisted task offloading and path optimization for mobile edge computing via multi-agent deep reinforcement learning[J]. Journal of Network and Computer Applications, 2024, 229: 103919.

[11] KANG H, CHANG X, MIŠIĆ J, et al. Cooperative UAV resource allocation and task offloading in hierarchical aerial computing systems: A MAPPO-based approach[J]. IEEE Internet of Things Journal, 2023, 10(12): 10497–10509.

[12] SUN H, ZHOU Y, ZHANG H, et al. Joint optimization of caching, computing and trajectory planning in aerial mobile edge computing networks: An MADDPG approach[J]. IEEE Internet of Things Journal, 2024, 11(24): 40996–41007.

[13] HWANG S, LEE H, KIM M, et al. Multi-agent deep reinforcement learning for decentralized multi-UAV mobile edge computing networks[J]. IEEE Internet of Things Journal, 2025, 12(10): 14484–14497.

[14] XU S, LIU Q, GONG C, et al. Energy-efficient multi-agent deep reinforcement learning task offloading and resource allocation for UAV edge computing[J]. Sensors, 2025, 25(11): 3403.

[15] Roijers, Diederik M., et al. "A survey of multi-objective sequential decision-making." Journal of Artificial Intelligence Research 48. 2013: 67-113.

[16] Van Moffaert, Kristof, and Ann Nowé. "Multi-objective reinforcement learning using sets of pareto dominating policies." The Journal of Machine Learning Research 15.1. 2014: 3483-3512.

[17] Pirotta, Matteo, Simone Parisi, and Marcello Restelli. "Multi-objective reinforcement learning with continuous pareto frontier

approximation." Proceedings of the AAAI conference on artificial intelligence. Vol. 29. No. 1. 2015.

[18] Zhang, Kaiqing, et al. "Robust multi-agent reinforcement learning with model uncertainty." Advances in neural information processing systems 33. 2020: 10571-10583.

[19] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Advances in Neural Information Processing Systems. 2017: 6379–6390.

[20] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. arXiv:1509.02971, 2015.

[21] NG A Y, HARADA D, RUSSELL S. Policy invariance under reward transformations: Theory and application to reward shaping[C]//Proceedings of the 16th International Conference on Machine Learning. 1999: 278–287.

[22] Saini, Hemant Kumar. "Connecting the 6G Autonomous Worlds with Real Time Edge Intelligence (Autonomous Vehicle)." Edge Intelligence for 6G-Enabled Industrial Internet of Things (2026): 163-180.

[23] Rakhra, Manik, Deepak Prashar, and Sudan Jha. "A Graph-based AI Generated Visual Manipulation Detector for UAV-based Industrial Systems." Cognitive Security for Industrial IoT. CRC Press, 2026. 110-124.

[24] Rashmi, K. B., et al. "Intrusion Detection for Unmanned Aerial Vehicles (UAVs) via Lightweight Federated Continuous Learning." 2026 5th International Conference on Sentiment Analysis and Deep Learning (ICSADL). IEEE, 20