

Carbon-Aware Last-Mile Delivery: Joint Scheduling and 3D Trajectory Design for ISTN-Connected AAVs via Hierarchical MARL

Huiyeon Jang¹ and Soyi Jung²,

¹Dept. Artificial Intelligence Convergence Network, Ajou University, Suwon, 16499, South Korea

²Dept. Electrical and Computer Engineering, Ajou University, Suwon, 16499, South Korea
timd0801@ajou.ac.kr, sjung@ajou.ac.kr

Abstract—Autonomous aerial vehicles (AAVs) in low Earth orbit (LEO) satellite-assisted systems face dynamic communication uncertainties and nonlinear, payload-dependent energy consumption. To address these challenges, we propose a hierarchical multi-agent reinforcement learning (HMARL) framework for mission assignment and flight control. Evaluated in a custom simulation integrating a first-order kinematic flight surrogate, a kinetic battery model, and a payload-dependent rotary-wing power model, our framework utilizes a centralized double deep Q-network (DDQN) for coarse-scale mission assignments and per-AAV multi-agent twin-delayed deep deterministic policy gradient (MATD3) controllers for continuous flight commands. Results confirm that this approach effectively optimizes task-level decision-making and trajectory control while adapting to battery dynamics and communication quality.

Index Terms—AAV cargo delivery, hierarchical multi-agent reinforcement learning, carbon-aware optimization

I. INTRODUCTION

AAV-based last-mile delivery has emerged as a transformative logistics solution with the potential to substantially reduce operational cost, shorten customer delivery latency, and mitigate environmental impact through optimized energy utilization [1]. Since AAVs are not constrained by road-network conditions or ground traffic congestion, they can provide flexible last-mile delivery services in suburban, peri-urban, and hard-to-access areas where conventional ground vehicles may suffer from limited accessibility or long travel times [2]. Stable network connectivity is essential for the efficient scheduling of AAV-based logistics delivery. Since an AAV must exchange flight-control commands and status information with a central controller in real time, mission reliability cannot be ensured without seamless communication. However, the high flight altitude of AAVs degrades the link quality to ground base stations. To overcome this limitation, the introduction of an integrated satellite-terrestrial network (ISTN), which combines satellite and terrestrial networks, is required.

As AAV logistics systems scale up, their aggregate energy consumption and associated carbon emissions also increase. AAV propulsion energy depends on the payload mass, flight distance, speed, and vertical motion. Therefore, mission assignment and flight trajectory should be jointly optimized rather than designed independently. Hence, sustainable logis-

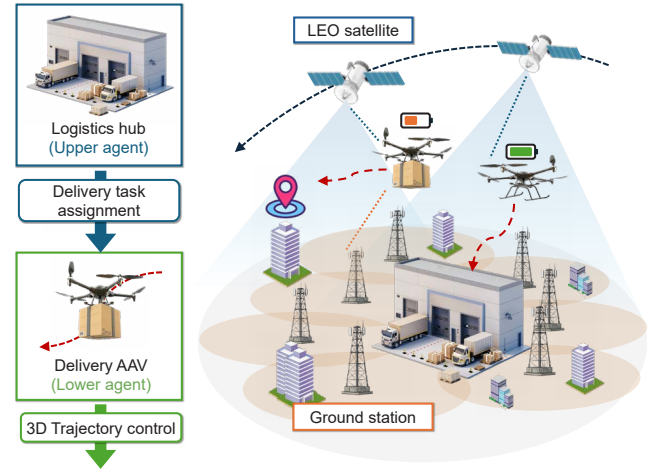


Fig. 1: Overview of the rotary-wing AAV cargo delivery system.

tics operations call for optimization that explicitly accounts for carbon emissions.

Therefore, this paper proposes a hierarchical multi-agent reinforcement learning (HMARL)-based scheme for the joint optimization of AAV logistics delivery scheduling and three-dimensional (3D) trajectory design. In the proposed scheme, the upper layer at the hub determines which AAV carries which package, whereas the lower layer optimizes the 3D trajectory along which each AAV transports its assigned package to the destination while maintaining connectivity through the ISTN.

The rest of this paper is organized as follows. Section II describes the system model, and Section III formulates the carbon-aware delivery problem. Section IV details the proposed HMARL framework, Section V presents the performance evaluation, and Section VI concludes the paper.

II. SYSTEM MODEL

We consider the AAV cargo delivery system illustrated in Fig. 1, comprising a single cargo hub and a fleet of M rotary-wing AAVs $\mathcal{M} = \{1, \dots, M\}$. The hub occupies a fixed position $\mathbf{q}_0 \in \Omega$, where $\Omega \subseteq \mathbb{R}^3$ is the operational region, and time is discretized into control steps $\mathcal{T} \in \{0, \dots, T-1\}$ of

duration Δt . To focus learning on mission allocation, collision avoidance, communication-aware routing, and energy-efficient trajectory generation, AAV motion is represented by a first-order kinematic surrogate rather than by rotor thrust, attitude, and rigid-body dynamics.

The operation follows a two-tier architecture. An upper-layer allocator, co-located with the hub, issues high-level missions (package pickup, charging, return-to-hub) to each AAV, while a lower layer of per-AAV policies generates short-horizon flight commands toward the assigned destinations. Each AAV can connect to both a low Earth orbit (LEO) satellite and a terrestrial base station, but uses exactly one link at any time. New packages arrive during step t as $A(t) \sim \text{Poisson}(\lambda \Delta t)$, independent across steps, where λ is the per-second arrival rate.

A. Rotary-Wing Energy and Carbon Emission Model

The rotary-wing AAV flight power consists of the horizontal propulsion power P_h that overcomes aerodynamic drag and the vertical-climb power P_v for the rate of change of potential energy. When the AAV flies horizontally at airspeed $V_i(t)$, the propulsion power P_h is expressed as the sum of blade profile power, induced power, and parasite drag power [3]:

$$P_h(V_i(t)) = P_0 \left(1 + \frac{3V_i(t)^2}{U_{\text{tip}}^2} \right) + P_I \left(\sqrt{1 + \frac{V_i(t)^4}{4v_0^4}} - \frac{V_i(t)^2}{2v_0^2} \right)^{1/2} + \frac{1}{2} d_0 \rho s A_r V_i(t)^3, \quad (1)$$

where P_0 is the blade profile power in hover, U_{tip} is the rotor blade tip speed, P_I is the induced-power coefficient, v_0 is the mean rotor-induced velocity in hover, d_0 is the fuselage drag ratio, ρ is the air density, s is the rotor solidity, and A_r is the rotor disc area.

We model each package as a weight-only payload. The payload affects the propulsion consumption and battery discharge through the total weight supported by the AAV, expressed as

$$W_{\text{total},i}(t) = (m_{\text{drone}} + m_i^{\text{pl}}(t)) g, \quad (2)$$

where $m_i^{\text{pl}}(t)$ denotes the payload mass carried by AAV i at time t , m_{drone} is the mass of the AAV without the payload, and g denotes the gravitational acceleration. The total weight $W_{\text{total},i}(t)$ enters the induced-power coefficient P_I and the mean induced velocity v_0 as

$$P_I = (1 + k) \frac{W_{\text{total},i}(t)^{3/2}}{\sqrt{2\rho A_r}}, \quad (3)$$

$$v_0 = \sqrt{\frac{W_{\text{total},i}(t)}{2\rho A_r}}, \quad (4)$$

where k is the induced-power correction factor.

The vertical-flight power is proportional to the altitude rate $\dot{H}_i(t) = dH_i(t)/dt$ and is consumed only during ascent. That

is,

$$P_v(\dot{H}_i(t)) = \begin{cases} W_{\text{total},i}(t) \dot{H}_i(t), & \text{if } \dot{H}_i(t) > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

During descent, the AAV is assumed to operate in autorotation and to consume no additional energy.

Consequently, the instantaneous flight power of AAV i at time t is

$$P_{\text{total},i}(t) = P_h(V_i(t)) + P_v(\dot{H}_i(t)). \quad (6)$$

Given $P_{\text{total},i}(t)$, its cumulative flight energy consumption over time is represented as

$$E_{\text{flight},i}(t) = \sum_{\tau=0}^t P_{\text{total},i}(\tau) \Delta t. \quad (7)$$

Let $\eta_{\text{ch}} \in (0, 1]$ denote the battery charging efficiency and ξ the grid carbon-intensity factor, the grid energy required to replenish $E_{\text{flight},i}(t)$ is $E_{\text{flight},i}/\eta_{\text{ch}}$, and the cumulative operational carbon emission of AAV i is therefore

$$C_i(t) = \xi \frac{E_{\text{flight},i}(t)}{\eta_{\text{ch}}}, \quad (8)$$

where ξ is the carbon intensity of the electricity grid. In this work, the carbon intensity is set to the U.S. national average $\xi = 0.357 \text{ kg CO}_2/\text{kWh}$ [4], and the battery charge efficiency is set to $\eta_{\text{ch}} = 0.81$ [5].

B. Communication Model

LEO satellites and terrestrial base stations (BSs) jointly serve the AAVs. Each LEO satellite generates eight spot beams, each treated as a separate cell, whereas each BS forms a single ground cell. We denote the sets of LEO and terrestrial cells by $\mathcal{C}_S$ and $\mathcal{C}_T$, respectively, and let $\mathcal{C} = \mathcal{C}_S \cup \mathcal{C}_T$. At each instant, every AAV associates with exactly one cell according to a maximum signal-to-interference-plus-noise ratio (SINR) criterion, while intra-cell access follows time-division multiple access (TDMA). Let the binary variable $c_{i,j}(t)$ indicate the association of AAV i with cell j at time t , so that

$$c_{i,j}(t) \in \{0, 1\}, \quad \sum_{j \in \mathcal{C}} c_{i,j}(t) = 1, \quad \forall i, t. \quad (9)$$

If AAV i associates with a LEO cell $j \in \mathcal{C}_S$, its received power obeys the free-space path-loss model

$$P_{r,i,j}^S(t) = P_j^S(t) G_S G_{\text{AAV}} \left(\frac{c}{4\pi f_S d_{i,j}(t)} \right)^2, \quad (10)$$

where $P_j^S(t)$ is the satellite-cell transmit power, G_S and G_{AAV} are the satellite and AAV antenna gains, c is the speed of light, f_S is the satellite carrier frequency, and $d_{i,j}(t)$ is the slant range. For a terrestrial cell $j \in \mathcal{C}_T$, the received power additionally incorporates small-scale fading and an environment-dependent path-loss exponent, yielding

$$P_{r,i,j}^T(t) = P_j^T(t) G_T G_{\text{AAV}} \times |h_{i,j}(t)|^2 \left(\frac{c}{4\pi f_T} \right)^2 d_{i,j}(t)^{-\alpha}. \quad (11)$$

where $P_j^T(t)$ is the BS transmit power, G_T the BS antenna gain, f_T the terrestrial carrier frequency, $d_{i,j}(t)$ the three-dimensional (3D) distance between AAV i and BS j , $\alpha \geq 2$ the path-loss exponent, and $|h_{i,j}(t)|^2$ the small-scale fading power. We write $P_{r,i,j}(t)$ for the received power from cell j , equal to $P_{r,i,j}^S(t)$ for $j \in \mathcal{C}_S$ and $P_{r,i,j}^T(t)$ for $j \in \mathcal{C}_T$.

For a candidate cell j , the SINR perceived by AAV i is

$$\phi_{i,j}(t) = \frac{P_{r,i,j}(t)}{N_0 + I_{i,j}(t)}, \quad (12)$$

where N_0 is the noise power and $I_{i,j}(t)$ is the aggregate co-channel interference from the set $\mathcal{I}_j$ of cells reusing the same resource,

$$I_{i,j}(t) = \sum_{j' \in \mathcal{I}_j} P_{r,i,j'}(t). \quad (13)$$

Under the max-SINR policy, AAV i connects to the cell with the highest SINR,

$$j_i^*(t) = \arg \max_{j \in \mathcal{C}} \phi_{i,j}(t), \quad (14)$$

and $c_{i,j}(t) = 1$ in (9) if and only if $j = j_i^*(t)$. AAV i is in per-step outage when its serving SINR fails to exceed the required threshold ϕ_{th} , i.e., $\max_{j \in \mathcal{C}} \phi_{i,j}(t) \leq \phi_{th}$.

III. PROBLEM FORMULATION

Building on the system model in Section II, we formulate carbon-aware delivery as a finite-horizon online multi-objective hybrid optimization problem under stochastic request arrivals. Specifically, delivery requests are modeled as a Poisson arrival process, while mission-allocation and flight-control decisions are updated sequentially based on the currently observed system state, without requiring prior knowledge of future requests.

At each decision slot, the control action consists of two coupled components. The first is a discrete mission-allocation decision represented as

$$\mathbf{X}(t) = x_{i,k}(t), \quad i \in \mathcal{M}, k \in \mathcal{K}_p(t), \quad (15)$$

where $x_{i,k}(t) \in 0, 1$ indicates whether AAV i is assigned to pending delivery request, or package, k at slot t , and $\mathcal{K}_p(t)$ denotes the set of pending requests observed at slot t . The second is a continuous flight-control decision $\mathbf{f}_i(t)$, which specifies the three-dimensional motion command of AAV i . The decision space is therefore mixed discrete-continuous.

We adopt two key performance indicators. With $N_{\text{req}}(t)$ and $N_{\text{del}}(t)$ the cumulative numbers of requests arrived and successfully completed up to slot t , the real-time delivery success ratio is

$$\eta_{\text{del}}(t) = \frac{N_{\text{del}}(t)}{N_{\text{req}}(t)}, \quad N_{\text{req}}(t) > 0. \quad (16)$$

The average delay $\bar{D}(t)$ accounts for both completed and pending requests through the per-package delay

$$D_k(t) = \begin{cases} t_k^c - t_k^a, & k \in \mathcal{K}_{\text{del}}(t), \\ t - t_k^a, & k \in \mathcal{K}_p(t), \end{cases} \quad (17)$$

where t_k^a and t_k^c are the arrival and completion slots of request k , and $\mathcal{K}_{\text{del}}(t)$ and $\mathcal{K}_p(t) = \mathcal{K}_{\text{req}}(t) \setminus \mathcal{K}_{\text{del}}(t)$ are the completed

and pending sets at slot t . A pending request thus keeps accruing waiting-time penalty until served. The average delay is then

$$\bar{D}(t) = \frac{1}{N_{\text{req}}(t)} \sum_{k \in \mathcal{K}_{\text{req}}(t)} D_k(t), \quad N_{\text{req}}(t) > 0, \quad (18)$$

where $\mathcal{K}_{\text{req}}(t)$ denotes the set of all requests received up to slot t .

The carbon-aware delivery problem then seeks to jointly maximize the cumulative delivery success ratio $\eta_{\text{del}}(t)$ while minimizing the average delivery delay $\bar{D}(t)$ and the total carbon emission $C_{\text{CO}_2}(t)$. The optimization problem is defined as

$$(P1) \quad \max_{\substack{\mathbf{X}(t), \\ \{f_i(t)\}_{i \in \mathcal{M}}}} \mathbb{E} \left[\omega_s \eta_{\text{del}}(T) - \omega_d \frac{\bar{D}(T)}{D_{\text{ref}}} - \omega_c \frac{C_{\text{CO}_2}(T)}{C_{\text{ref}}} \right] \quad (19)$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{C}} c_{i,j}(t) = 1, \quad c_{i,j}(t) \in \{0, 1\}, \quad (20)$$

$$\phi_i^*(t) > \phi_{th}, \quad (21)$$

$$\mathbf{p}_i(t) \in \Omega, \quad (22)$$

$$\text{dist}(\mathbf{p}_i(t), \mathcal{B}) \geq d_{\text{safe}}, \quad (23)$$

$$\delta_{\min} \leq \delta_i(t) \leq 1, \quad (24)$$

$$\|\mathbf{v}_i(t)\| \leq V_{\max}, \quad |\dot{H}_i(t)| \leq V_{\max}^z, \quad (25)$$

$$Q(t+1) = Q(t) + A(t) - \sum_{i \in \mathcal{M}} b_i(t) \geq 0, \quad (26)$$

$$\forall i \in \mathcal{M}, t \in \mathcal{T},$$

$$\omega_s + \omega_d + \omega_c = 1, \quad \omega_s, \omega_d, \omega_c \geq 0 \quad (27)$$

where D_{ref} and C_{ref} normalize the delay and emission terms in (19) to comparable scales. The constraints admit the following interpretation. The association $c_{i,j}(t)$ in (20) is not an independent decision variable but an induced quantity fully determined by the max-SINR rule (14), which enforces single-link connectivity. Constraint (21) then requires the serving SINR $\phi_i^*(t)$ to exceed the threshold ϕ_{th} . Constraint (22) confines each AAV to the operational region Ω , while (23) maintains a minimum clearance d_{safe} from the union $\mathcal{B} \subset \Omega$ of static building volumes. Constraint (24) enforces the SoC reserve, keeping the battery state of charge $\delta_i(t)$ between the minimum level $\delta_{\min}$ and full charge, and (25) imposes the kinematic speed and climb-rate limits $V_{\max}$ and $V_{\max}^z$. Constraint (26) preserves the nonnegativity of the hub backlog $Q(t)$ under stochastic arrivals $A(t)$ and per-AAV service $b_i(t)$. Finally, the nonnegative weights ω_s , ω_d , and ω_c in (27) sum to one and trade off delivery success, delay, and emissions.

IV. PROPOSED FRAMEWORK

To solve (P1), we propose an HMARL framework that decomposes the coupled allocation-control problem along two time scales. The upper layer is a centralized allocator, co-located with the hub, that assigns a discrete mission to each AAV every K steps; it is a DDQN over the joint mission space, modeled as a Markov decision process (MDP) with macro-step $\Delta T = K\Delta t$. The lower layer comprises per-AAV deterministic flight policies that output 3D commands toward the assigned destination at every step Δt , trained by MATD3

with a shared centralized critic under the centralized-training, decentralized-execution (CTDE) paradigm. This temporal abstraction replaces the monolithic hybrid action space of (P1) with two tractable sub-problems, improving convergence [6].

A. Upper-Level Agent MDP

State. At each macro-step epoch $T_k = k\Delta T$, the allocator observes the global state

$$s^h(T_k) = \left\{ \{\mathbf{p}_i, \delta_i, \sigma_i, m_i^{\text{pl}}\}_{i \in \mathcal{M}}, Q \right\}_{T_k}, \quad (28)$$

which, for every AAV $i \in \mathcal{M}$, comprises its position $\mathbf{p}_i$, SoC δ_i , operating mode $\sigma_i \in \{\text{charge, idle, delivery, return}\}$, and payload m_i^{pl} , together with the hub backlog Q .

Action. At every epoch, the allocator (re)assigns a mission to each AAV from

$$a_i^h(T_k) \in \mathcal{A}^h = \{\text{CHARGE, PICKUP}(p), \text{RETURN, NOOP}\}. \quad (29)$$

Each chosen mission defines a target waypoint $\mathbf{g}_i$ that is passed to the lower layer and held fixed over the macro-step ΔT (or until it is reached), so a mission is a temporally extended action. Since the per-AAV missions are coupled—two AAVs cannot reserve the same package, and CHARGE is valid only when docked—the allocator selects a single *joint* action from $\mathcal{A}^h = \prod_i \mathcal{A}_i^h$, where a structural mask prunes infeasible or conflicting combinations before the DDQN scores the remaining ones.

Reward. Each joint action yields a single *team* reward balancing delivery utility against CO₂ emissions. Let $\Delta C_i^{(k)} = C_i(T_{k+1}) - C_i(T_k)$ denote the CO₂ emitted by AAV i during macro-step k , the increment of (8). The reward is

$$r^h(T_{k+1}) = \sum_{i \in \mathcal{M}} \left[w_u u_i^{(k)} + w_\rho \rho_i^{(k)} - w_c \Delta C_i^{(k)} - w_\tau \bar{\tau}_i^{(k)} - w_f \mathbb{1}\{\delta_i < \delta_{\min}\} - w_g z_i^{\text{fc}} \right], \quad (30)$$

where $u_i^{(k)}$ is the number of deliveries completed, $\rho_i^{(k)} \in [0, 1]$ is the normalized queueing delay, and $\bar{\tau}_i^{(k)} = (t_i^c - t_i^a)/T$ is the normalized end-to-end delay (zero if no delivery occurs), capturing both hub queueing and in-flight time. The term w_f heavily penalizes battery depletion ($\delta_i < \delta_{\min}$), while $z_i^{\text{fc}} = 1$ discourages charging a full AAV. Infeasible assignments are removed by the structural mask rather than penalized. The dominant carbon weight w_c steers the allocator toward emission-efficient delivery and charging decisions.

B. Lower-Level Agent MDP

State. Conditioned on the assigned goal $\mathbf{g}_i$, AAV i observes the local state

$$s_i^l(t) = \{\Delta \mathbf{g}_i^b(t), \mathbf{v}_i^b(t), \psi_i(t), H_i(t), \delta_i(t), m_i^{\text{pl}}(t), \Delta \phi_i(t), \mathbf{n}_i(t), \mathbf{a}_i^l(t-1)\}, \quad (31)$$

where $\Delta \mathbf{g}_i^b(t) = \mathbf{g}_i - \mathbf{p}_i(t)$ and $\mathbf{v}_i^b(t)$ are the goal displacement and velocity in the body frame; ψ_i , H_i , δ_i , and m_i^{pl} denote the heading, altitude, SoC, and payload; $\Delta \phi_i(t) =$

$\phi_i^*(t) - \phi_{\text{th}}$ is the serving-link SINR margin under (14); $\mathbf{n}_i$ is the relative position and velocity of the nearest AAV; and $\mathbf{a}_i^l(t-1)$ is the previous command. Here, m_i^{pl} exposes the payload–power coupling of Section II-A, $\Delta \phi_i$ encodes link quality, and the body-frame representation renders the policy heading-invariant.

Action. The policy outputs a bounded 3D motion command

$$a_i^l(t) = (\Delta \psi_i(t), \nu_i(t), \zeta_i(t)), \quad (32)$$

comprising a yaw increment $\Delta \psi_i \in [-\Delta \psi_{\max}, \Delta \psi_{\max}]$, a normalized forward speed $\nu_i \in [0, 1]$ (scaled to $V_{\max}$), and a normalized climb rate $\zeta_i \in [-1, 1]$ (scaled to $V_{\max}^z$).

Reward. The lower-level reward combines goal progress, mission completion, carbon emission, and a safety/communication penalty:

$$r_i^l(t) = r_i^{\text{prog}}(t) + r_i^{\text{mis}}(t) - \beta_c \bar{C}_i(t) - \mathcal{P}_i(t). \quad (33)$$

The progress term $r_i^{\text{prog}}(t) = \beta_p \widetilde{\Delta d}_i(t)$ densely shapes motion toward the target, where $\widetilde{\Delta d}_i(t) = [\Delta d_i(t)]_+ - \lambda_r [-\Delta d_i(t)]_+$ ($\lambda_r > 1$) penalizes moving away from the goal more than it rewards approach, with $\Delta d_i(t)$ the per-step reduction in AAV-to-target distance. The mission term

$$r_i^{\text{mis}}(t) = \beta_a z_i^{\text{wp}}(t) + \beta_d z_i^{\text{del}}(t) - \beta_\tau \bar{\tau}_i^{\text{flight}}(t) z_i^{\text{del}}(t) \quad (34)$$

rewards waypoint arrival and delivery while discouraging slow flights, where $\bar{\tau}_i^{\text{flight}}(t)$ is the pickup-to-delivery delay normalized by horizon T ; restricting it to the post-pickup interval keeps the controller from being charged for queueing delay owned by the upper allocator. The term $\bar{C}_i(t)$ is the step-wise CO₂ emission (8) normalized by unladen hover, and the safety penalty

$$\mathcal{P}_i(t) = \beta_\phi z_i^\phi(t) + \beta_x z_i^{\text{area}}(t) + \beta_k z_i^{\text{col}}(t) + \beta_b z_i^{\text{bat}}(t) \quad (35)$$

flags SINR violation, region departure, collision, and battery depletion. In particular, the SINR term steers trajectories away from regions failing the SINR condition, enabling joint trajectory–communication optimization.

V. PERFORMANCE EVALUATION

A. Simulation Environment

We built a simulation environment with a single hub and two rotary-wing AAVs over a 2 km $\times$ 2 km region with a 150 m maximum altitude as shown in Fig. 2. The AAVs are controlled at 2 Hz ($\Delta t = 0.5$ s) and the upper allocator re-decides every $K = 4$ steps. Packages arrive as a Poisson process and are delivered to building rooftops; the integrated network comprises a seven-site, 21-sector terrestrial deployment and a LEO tier, with each AAV associating to the max-SINR cell under ITU-R P.1410 statistical building shadowing.

B. Simulation Results and Analysis

We compare against two non-learning allocators that share a fixed, target-directed lower controller, so the comparison isolates the allocation policy: delay-greedy, which assign each AAV to the nearest available package. As shown in Fig. 3, the DSR decreases as the package arrival rate grows from

TABLE I: Training hyperparameters of the proposed HMARL framework.

Parameter	Value
Optimizer	Adam
Activation function	ReLU
Discount factor γ	0.99
Target soft-update rate τ	0.005
Control time step Δt	0.5 s
Allocator decision period K	4 steps
Total training steps	5×10^5
<i>Lower controller (MATD3)</i>	
Actor / critic learning rate	3×10^{-4}
Replay buffer size	3×10^5
Batch size	256
Hidden layers $\times$ units	3×512
Policy delay d	2
Target / exploration noise σ	0.2 / 0.1
Target-noise clip c	0.5
Learning starts	2.5×10^4
<i>Upper allocator (DDQN)</i>	
Learning rate	3×10^{-4}
Replay buffer size	1×10^5
Batch size	64
Hidden layers $\times$ units	2×256
Exploration ϵ (start $\rightarrow$ end)	$1.0 \rightarrow 0.05$

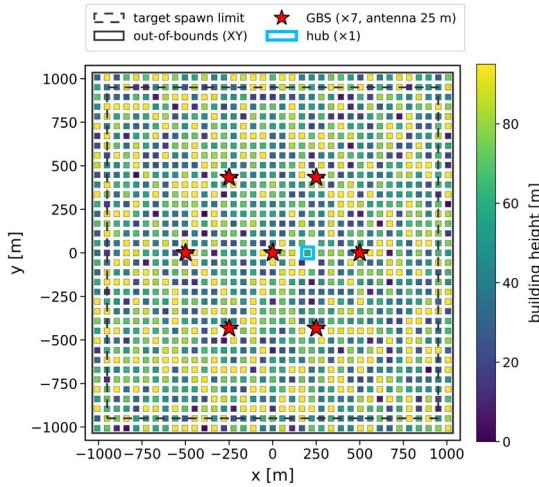


Fig. 2: AAV last-mile delivery simulation environment

TABLE II: Operational carbon emission per delivery (g CO₂/delivery) versus package arrival rate.

Arrival rate (pkg/min/hub)	Proposed HMARL	Delay-greedy	Reduction (%)
0.3	0.198	0.276	28.5
0.6	0.193	0.276	29.9
0.9	0.211	0.276	23.6
1.2	0.211	0.276	23.6

0.3 to 1.2 pkg/min/hub, because the two-AAV fleet saturates and the request backlog accumulates; the proposed allocator matches the delay-greedy DSR across the entire load range. Table II shows that it sustains this delivery performance far more efficiently, keeping the carbon emission per delivery between 0.193 and 0.211 g CO₂ against a constant 0.276 g CO₂ for delay-greedy, a 23.6–29.9% reduction across all arrival rates.

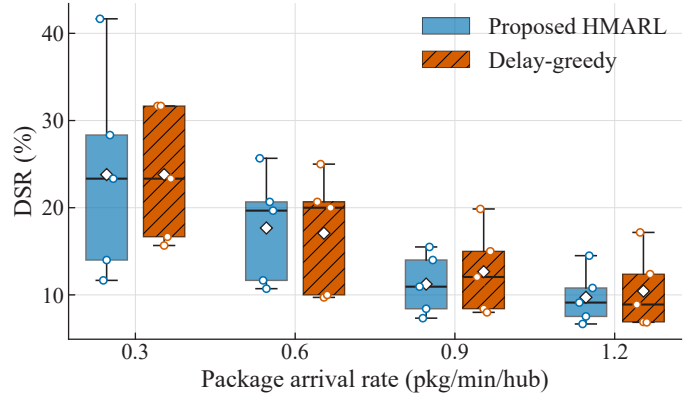


Fig. 3: Delivery success rate under varying package arrival states

VI. CONCLUSION

This paper proposes a hierarchical multi-agent reinforcement learning framework for AAV cargo last-mile delivery that jointly optimizes mission allocation and flight-trajectory control to reduce CO₂ emissions from electric power consumption. The framework integrates a centralized DDQN-based allocator for delivery sequencing and charging decisions with MATD3-based flight controllers that generate real-time 3D trajectories while considering payload-dependent energy use and communication-aware route adaptation.

ACKNOWLEDGMENT

This work was supported in part by the MSIT (Ministry of Science and ICT), South Korea, under the ITRC (Information Technology Research Center) Support Program supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP) under Grant IITP-2024-RS-2024-00436887.

REFERENCES

- [1] M. Rahmani, S. Mohammed Senouci, F. Delavernhe, and M. Berbineau, "Q-learning-based multi-objective path planning for AAV parcel delivery through public transportation vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 10, pp. 15 471–15 485, October 2025.
- [2] J. Cao, L. Yang, D. Yang, T. Zhang, L. Xiao, H. Jiang, and D. Niyato, "Trajectory optimization and pick-up and delivery sequence design for cellular-connected cargo AAVs," *IEEE Transactions on Mobile Computing*, vol. 24, no. 3, pp. 1402–1416, March 2025.
- [3] S. Jung, H. Baek, and J. Kim, "Neural myerson auction for truthful and distributed mobile charging in UAV-assisted digital-twin networks," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 11, pp. 7326–7338, November 2024.
- [4] U.S. Energy Information Administration, "U.s. electric power industry estimated emissions by state (EIA-767, EIA-906, EIA-920, and EIA-923)," Historical State Data, Final 2024 Data, November 2025, released Nov. 12, 2025.
- [5] J. K. Stolaroff, C. Samaras, E. R. O'Neill, A. Lubers, A. S. Mitchell, and D. Ceperley, "Energy use and life cycle greenhouse gas emissions of drones for commercial package delivery," *Nature Communications*, vol. 9, p. 409, February 2018.
- [6] J. Jang, I.-S. Cho, M. Shin, J. Kim, and S. Jung, "Gateway-assisted neural angular routing for hierarchical satellite networks," *IEEE Internet of Things Journal*, vol. 13, no. 6, pp. 12 322–12 339, March 2026.