Journal of Low-Altitude Economy and Intelligent Aviation
Volume 1, Issue 1, 2026, Page 31–45
https://doi.org/10.6914/jlae.2026.010103

HGAT-MAPPO for Ultra-Dense Urban Air Mobility: Communication-Aware Graph Attention, Physics-Informed Energy Optimization, and Simplex Runtime Assurance

Sophia Li
Artificial Intelligence Education Studies Press

Abstract

Urban Air Mobility (UAM) requires scalable and certifiable autonomy under ultra-dense operations and non-ideal communication. This paper proposes HGAT-MAPPO, a hierarchical graph-attention-enhanced multi-agent reinforcement learning framework for cooperative eVTOL trajectory optimization. Channel-quality indicators derived from 6G integrated sensing and communication are embedded into graph attention, enabling agents to down-weight unreliable links and maintain decentralized coordination in urban canyons. A physics-informed reward integrates nonlinear propulsion power and battery discharge effects, inducing energy-efficient "lateral insertion" maneuvers that avoid high-power hover regimes. Experiments in a high-fidelity Shenzhen CBD digital twin demonstrate robust collision avoidance and reduced mission energy compared with strong baselines at hundred-aircraft density. Finally, we outline a Simplex runtime assurance architecture to bridge learning-based control with airworthiness certification.

Keywords Low-Altitude Economy (LAE); Urban Air Mobility (UAM); Multi-Agent Reinforcement Learning (MARL); Integrated Sensing and Communication (ISAC); Runtime Assurance (Simplex Architecture)

1 Introduction

1.1 Paradigm Shift in the Low-Altitude Economy and UAM Operational Density

Urban Air Mobility (UAM) is widely regarded as the third major transformation of urban transportation systems, following surface road networks and rail transit. By leveraging electric vertical take-off and landing (eVTOL) aircraft to enable point-to-point transportation within low-altitude airspace—typically defined as below 1,000 meters—UAM promises to fundamentally alleviate chronic ground congestion in megacities. Unlike conventional modes constrained by planar infrastructure, low-altitude aviation exploits near-Euclidean spatial paths, offering substantial gains in travel time predictability and system throughput.

However, the economic viability of UAM is inherently contingent upon scale. Forecasts by Morgan Stanley and the National Aeronautics and Space Administration (NASA) indicate that a mature UAM ecosystem would require the simultaneous coordination of hundreds to thousands of flights within tightly constrained urban airspace volumes. Such ultra-high-density operations mark a profound paradigm shift in low-altitude airspace management—from the traditional model of *isolated airspace, pre-defined routes, and human-centered control* toward an emerging regime characterized by *integrated airspace, dynamic trajectories, and autonomous coordination.*

Under this new paradigm, existing optimization-based air traffic management approaches face fundamental limitations. Centralized methods such as mixed-integer linear programming (MILP) suffer from exponential growth in computational complexity as the number of aerial agents increases, rendering them unsuitable for millisecond-level safety-critical decision-making. Conversely, decentralized rule-based approaches—including Optimal Reciprocal Collision Avoidance (ORCA) and artificial potential field (APF) methods—exhibit favorable computational

© The Authors 2026. Article History Received: December 29, 2025; Accpted: January 10, 2026; Published: January 30, 2026. ***Journal of Low-Altitude Economy and Intelligent Aviation***, ISSN ****–****(print), ISSN ****–****(online), a Quarterly, founded on January 2026, published by Creative Publishing Co., Limited. Email: editor@jlae.org, https://jlae.org, https://cpcl.hk. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which permits non-commercial use, sharing, adaptation, distribution, and reproduction in any medium or format, provided the original author(s) and source are credited, and derivative works are licensed under the same terms.

efficiency but lack robustness in environments with non-structured obstacles and multi-agent strategic interactions. These methods are prone to local minima and oscillatory behaviors, which not only compromise flight safety but also induce excessive energy consumption.

The coexistence of extreme operational density, complex urban geometry, and stringent safety requirements thus necessitates a new class of decision-making frameworks capable of balancing scalability, robustness, and physical realism.

1.2 Challenges of Multi-Agent Reinforcement Learning in Aviation Systems

Multi-agent reinforcement learning (MARL), particularly under the *centralized training and decentralized execution* (CTDE) paradigm, has emerged as a promising solution for high-dimensional trajectory optimization in UAM systems. By enabling agents to learn cooperative strategies in continuous state and action spaces, MARL offers a level of adaptability unattainable by classical control or optimization techniques.

Despite its theoretical appeal, the deployment of MARL in safety-critical aviation contexts remains constrained by several structural challenges. First, scalability emerges as a primary bottleneck. As the number of agents N increases, the dimensionality of the joint state action space grows combinatorially, severely degrading the critic's ability to accurately estimate value functions. This phenomenon, commonly referred to as the credit assignment problem, leads to unstable learning dynamics and performance collapse in widely adopted algorithms such as MADDPG and QMIX when the agent population exceeds moderate scales.

Second, most existing MARL formulations implicitly assume ideal communication conditions, including perfect connectivity, zero latency, and unlimited bandwidth. In real urban low-altitude environments, however, dense building clusters induce severe shadow fading and multipath effects, resulting in intermittent links and heterogeneous channel quality. Policies trained without awareness of such communication constraints are inherently brittle, as critical state information may be delayed or lost during execution, potentially triggering catastrophic coordination failures.

Third, prevailing energy models embedded in trajectory optimization remain overly simplistic. Many studies assume that energy consumption scales linearly with distance or time, an assumption that fails to capture the highly nonlinear power characteristics of eVTOL platforms. Hovering power significantly exceeds cruise power, frequent acceleration and deceleration incur disproportionate energy penalties, and battery discharge behavior is strongly influenced by the Peukert effect. Learning frameworks that ignore these physical realities risk producing trajectories that are nominally optimal yet operationally infeasible, dramatically reducing endurance and increasing system-level risk.

1.3 Opportunities Enabled by 6G Integrated Sensing and Communication

The emergence of sixth-generation (6G) mobile communication systems introduces a transformative opportunity to address the above limitations through the paradigm of Integrated Sensing and Communication (ISAC). ISAC enables aerial platforms to perform high-bandwidth data transmission and high-resolution environmental sensing using a unified radio-frequency front-end and waveform design.

Beyond reducing payload and hardware redundancy, ISAC fundamentally enriches the informational substrate available to autonomous agents. Channel state information (CSI), traditionally treated as a communication-layer metric, becomes an implicit and continuous proxy for environmental complexity, including obstacle density, line-of-sight availability, and electromagnetic interference. By mapping ISAC-derived sensing confidence and channel quality indicators into the decision-making loop of MARL agents, it becomes possible to endow agents with communication-aware perception, allowing coordination strategies to adapt dynamically to uncertain and time-varying electromagnetic environments.

This cross-layer integration of communication and control represents a critical step toward robust decentralized cooperation in dense urban airspace, where sensing degradation and communication uncertainty are structural rather than exceptional conditions.

1.4 Contributions of This Work

Motivated by the intertwined challenges of scalability, communication uncertainty, and energy inefficiency in high-density UAM operations, this paper proposes and validates a hierarchical graph-attention enhanced MARL frame-

work, referred to as HGAT-MAPPO. The first contribution of this work lies in the architectural design. By introducing temporal abstraction into the decision-making process, the proposed framework decomposes complex trajectory optimization into low-frequency strategic waypoint planning and high-frequency tactical attitude control. This hierarchical structure substantially reduces policy dimensionality and stabilizes learning under large agent populations.

The second contribution concerns algorithmic robustness under non-ideal communication conditions. A channel-quality aware graph attention mechanism is developed in which communication quality indicators derived from 6G ISAC are explicitly embedded into the attention computation. Unlike conventional graph attention networks that rely solely on feature similarity, the proposed mechanism enables agents to selectively attenuate noisy or unreliable information from low-quality links, thereby enhancing decentralized coordination robustness in electromagnetically complex urban environments.

The third contribution is the integration of physical energy mechanisms into the learning objective. A composite reward function is constructed by incorporating aerodynamic power consumption models together with battery discharge dynamics governed by the Peukert effect. This physics-informed formulation discourages aggressive and energy-inefficient maneuvers at the algorithmic level, yielding smoother trajectories and achieving a Pareto improvement between operational safety and energy efficiency.

Finally, the proposed framework is empirically validated through large-scale stress testing in a high-fidelity digital twin of the Shenzhen Central Business District constructed from real geographic information system (GIS) data. Experimental results demonstrate consistent performance advantages in heterogeneous fleet scenarios, including mixed operations of tilt-rotor and multirotor aircraft, as well as under communication-constrained conditions representative of dense urban canyons.

2 Related Work

2.1 Evolution of Trajectory Planning in Urban Air Mobility

Research on trajectory planning for urban air mobility has undergone a clear methodological evolution, shifting from deterministic geometric optimization toward data-driven and learning-based approaches. Early studies primarily relied on geometric collision-avoidance formulations, including the Velocity Obstacle (VO) framework and its reciprocal extensions such as Reciprocal Velocity Obstacle (RVO) and Optimal Reciprocal Collision Avoidance (ORCA). These methods are computationally efficient and have been widely adopted in tightly choreographed scenarios such as drone swarm performances and structured formations. Their underlying assumption of cooperative agents with predictable motion, however, becomes fragile in dense urban airspace characterized by constrained geometry, heterogeneous actors, and frequent uncertainty.

In urban canyon environments, purely geometric approaches are prone to deadlock phenomena, where multiple agents mutually block feasible velocity spaces and fail to resolve conflicts without external coordination. Moreover, these methods are inherently ill-suited to handling non-cooperative or partially observable dynamic obstacles, a condition that is increasingly prevalent in real-world low-altitude operations involving mixed traffic and uncertain intent. As a result, geometric planners struggle to scale beyond controlled or semi-structured scenarios.

To overcome local optimality and capture system-wide constraints, mixed-integer linear programming (MILP) formulations were once regarded as the gold standard for multi-vehicle trajectory optimization. By encoding collision avoidance, kinematic limits, and temporal logic into a unified optimization framework, MILP offers provable optimality guarantees under idealized conditions. However, MILP belongs to the class of NP-hard problems, and its computational complexity grows exponentially with the number of agents and constraints. In UAM contexts, where unexpected route changes, weather disturbances, or emergency maneuvers demand sub-second responses, MILP-based approaches become impractical for online deployment.

Artificial potential field (APF) methods represent another classical trajectory planning paradigm, in which attractive and repulsive forces are constructed to guide vehicle motion. While conceptually simple, APF suffers from well-documented local minima and oscillation issues, particularly in narrow corridors and dense obstacle fields. Such oscillatory behavior is not merely a mathematical inconvenience; in the context of passenger-carrying eVTOL aircraft, it translates directly into unacceptable ride discomfort and increased mechanical stress. These limitations underscore the need for planning frameworks that can adapt to complex, dynamic environments without sacrificing stability or safety.

2.2 Deep Reinforcement Learning for UAM Trajectory Optimization

The rapid advancement of deep reinforcement learning (DRL) has catalyzed a shift toward learning-based trajectory planning in UAM systems. Early applications of DRL employed algorithms such as Deep Q-Networks (DQN) and Deep Deterministic Policy Gradient (DDPG) for single-agent obstacle avoidance. While effective in isolated settings, these approaches encounter severe convergence difficulties in multi-agent scenarios due to environmental non-stationarity, where each agent's policy continuously alters the learning dynamics of others.

Multi-agent reinforcement learning (MARL) addresses this challenge by introducing joint training mechanisms that explicitly account for agent interactions. Among these, the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm represents a seminal contribution, employing a centralized critic to stabilize learning in continuous control tasks. Despite its conceptual elegance, MADDPG is known to suffer from Q-value overestimation and pronounced training instability as the number of agents increases, limiting its applicability in high-density UAM environments.

Subsequent research has increasingly favored policy-gradient based methods, particularly Multi-Agent Proximal Policy Optimization (MAPPO). Owing to its on-policy nature and monotonic policy improvement guarantees, MAPPO has demonstrated superior training stability in large-scale benchmarks such as StarCraft multi-agent micromanagement tasks. Nevertheless, standard MAPPO formulations remain challenged by heterogeneous agent dynamics and sparse-reward regimes, which are common in UAM operations where long periods of nominal flight are punctuated by rare but critical collision events. Addressing these issues often requires auxiliary techniques such as curriculum learning or reward shaping, which introduce additional design complexity.

A more fundamental limitation shared by much of the existing MARL literature is the oversimplified treatment of communication. Many studies assume fully connected interaction graphs or fixed-radius communication models with perfect information exchange. Such assumptions diverge sharply from the realities of urban low-altitude environments, where communication links are intermittent, asymmetric, and heavily influenced by building-induced shadowing and multipath effects. Policies trained under idealized communication assumptions therefore exhibit brittle performance when deployed in real urban settings.

2.3 Graph Neural Networks and Attention-Based Coordination

To address the scalability and partial observability challenges inherent in MARL, graph neural networks (GNNs) have been introduced as a principled mechanism for modeling agent interactions. In GNN-based MARL, agents are represented as nodes in a graph, and information is aggregated through message-passing operations that reflect local neighborhood structures. The Deep Graph Network (DGN) framework demonstrated that graph convolutional operations can significantly enhance cooperative behavior by enabling agents to reason over relational information.

However, conventional graph convolutional networks (GCNs) rely on Laplacian-based aggregation over entire neighborhoods, which can be computationally expensive and insensitive to the varying importance of different neighbors. This limitation becomes pronounced in dense UAM scenarios, where the number of proximate agents can fluctuate rapidly over time.

Graph Attention Networks (GATs) alleviate this issue by introducing self-attention mechanisms that allow each agent to dynamically weight the influence of its neighbors. In the context of UAM, this enables agents to prioritize nearby or high-risk neighbors while suppressing irrelevant information from distant or low-impact agents. More recent architectures, such as the Multi-Agent Transformer (MAT), further extend attention mechanisms to sequence modeling, capturing long-term temporal dependencies in decision-making.

Despite their representational power, Transformer-based MARL architectures impose substantial computational overhead. The resulting inference latency and memory consumption pose significant challenges for deployment on resource-constrained airborne embedded platforms, where real-time responsiveness is paramount. The hierarchical graph-attention design proposed in this work seeks to balance expressiveness and efficiency, retaining the adaptive neighbor selection of GATs while introducing temporal abstraction to reduce computational burden during real-time control.

2.4 Cross-Layer Optimization with 6G Integrated Sensing and Communication

Integrated Sensing and Communication (ISAC) has emerged as a cornerstone capability of sixth-generation (6G) wireless networks. Existing ISAC research has predominantly focused on physical-layer and medium-access layer

optimization, including waveform co-design, beamforming strategies, and performance bounds characterized by the Cramér Rao lower bound. These studies establish the feasibility of jointly achieving high-rate communication and high-resolution sensing using shared spectral resources.

In the context of UAM, several works have explored the use of unmanned aerial vehicles as aerial base stations providing ISAC services. However, the interaction between ISAC-enabled perception and higher-level decision-making remains largely underexplored. In particular, few studies propagate ISAC-derived environmental uncertainty estimates or channel quality indicators, such as signal-to-interference-plus-noise ratio (SINR), into trajectory planning or control algorithms.

This lack of cross-layer integration results in a structural mismatch. Trajectory planners and MARL-based controllers are typically designed under assumptions of perfect sensing and reliable communication, assumptions that are systematically violated in dense urban airspace. By ignoring physical-layer uncertainty, higher-level decision-making frameworks forfeit a critical source of contextual information, leading to degraded robustness under real-world conditions.

The present study directly addresses this gap by incorporating ISAC-derived channel quality and sensing confidence metrics into the state representation and coordination mechanisms of MARL agents. Through this integration, physical-layer uncertainty is elevated from a hidden disturbance to an explicit decision variable, enabling adaptive and resilient coordination strategies tailored to the realities of urban low-altitude environments.

3 System Model and Problem Formulation

We formulate the cooperative trajectory optimization problem of multiple urban air mobility (UAM) vehicles as a decentralized partially observable Markov decision process (Dec-POMDP). The system is defined by the tuple

$$\langle \mathcal{A}, \mathcal{S}, \mathcal{O}, \mathcal{U}, \mathcal{P}, \mathcal{R}, \gamma \rangle,$$

where $\mathcal{A}$ denotes the set of agents, $\mathcal{S}$ the global state space, $\mathcal{O}$ the observation space, $\mathcal{U}$ the continuous action space, $\mathcal{P}$ the state transition probability, $\mathcal{R}$ the reward function, and $\gamma \in (0, 1)$ the discount factor.

3.1 Operational Scenario and Agent Definition

Consider a fleet of N eVTOL vehicles indexed by $\mathcal{A} = \{1, \ldots, N\}$ operating within a three-dimensional urban airspace $\mathcal{W} \subset \mathbb{R}^3$. The environment contains a set of static obstacles $\mathcal{B}$, primarily representing buildings and infrastructure extracted from high-resolution urban geographic information system (GIS) data.

Each agent i is characterized by heterogeneous physical parameters, including the maximum allowable velocity $v^i_{\max}$, maximum thrust $T^i_{\max}$, vehicle mass m_i, and aerodynamic configuration coefficients. This heterogeneity reflects realistic mixed-fleet operations involving different eVTOL classes, such as tilt-rotor and multirotor platforms, and precludes the assumption of homogeneous agent dynamics.

3.2 State Space and Observation Model

The global system state at time step t, denoted by $s_t \in \mathcal{S}$, aggregates the kinematic, energetic, and environmental states of all agents. Due to sensing and communication limitations, each agent has access only to a local observation $o^t_i \in \mathcal{O}$, which constitutes a partial view of s_t constrained by a sensing radius R_{sense} and a communication radius R_{comm}.

The observation vector of agent i at time t is defined as

$$o^t_i = \left[p^t_i,\ v^t_i,\ \phi^t_i,\ \text{SOC}^t_i,\ w^t_i,\ \mathcal{N}^t_i \right],$$

where $p^t_i \in \mathbb{R}^3$ denotes the position in the inertial frame, $v^t_i \in \mathbb{R}^3$ the body-frame velocity, and $\phi^t_i = [\varphi^t_i, \theta^t_i, \psi^t_i]$ the Euler angles representing roll, pitch, and yaw. The battery state of charge SOC^t_i captures the remaining usable energy and directly constrains feasible control actions.

The term w^t_i represents the relative position of the current high-level waypoint generated by the strategic planner, providing low-frequency guidance for long-horizon navigation. The neighborhood information $\mathcal{N}^t_i$ aggregates the

relative states of neighboring agents $j \in \mathcal{N}_i^t$, including relative position, relative velocity, and the communication channel quality indicator CQI_{ij}^t associated with the link between agents i and j.

3.3 Action Space and Vehicle Dynamics

To ensure smooth and aerodynamically feasible control, each agent operates in a continuous action space. The action vector of agent i at time t is defined as

$$a_i^t = \left[T_i^t,\ \omega_{x,i}^t,\ \omega_{y,i}^t,\ \omega_{z,i}^t\right],$$

where T_i^t denotes the commanded total thrust, and $\omega_{x,i}^t$, $\omega_{y,i}^t$, and $\omega_{z,i}^t$ are the commanded angular rates around the body axes. These high-level control inputs are mapped by a low-level flight controller to individual rotor speeds.

Vehicle motion is simulated using a high-fidelity rigid-body dynamic model. For tilt-rotor platforms (Type A), the dynamics account for multiple flight regimes, including vertical takeoff and landing, transition, and fixed-wing cruise, with aerodynamic coefficients interpolated across modes. The translational dynamics are governed by

$$m_i \dot{v}_i = R(\phi_i) F_{\text{thrust},i} + F_{\text{drag}}(v_i) + F_{\text{gravity}},$$

where $R(\phi_i)$ denotes the rotation matrix from body frame to inertial frame, and the aerodynamic drag force F_{drag} is a nonlinear function of velocity and vehicle configuration.

3.4 6G Communication and Channel Modeling

Under a sixth-generation integrated sensing and communication (ISAC) paradigm, inter-agent communication quality is modeled as a continuous stochastic process rather than a binary connectivity condition. Air-to-air links are characterized using a log-distance path loss model combined with Rician fading,

$$PL(d) = PL(d_0) + 10\eta \log_{10}\left(\frac{d}{d_0}\right) + X_\sigma,$$

where η denotes the path loss exponent and X_σ represents shadow fading.

The signal-to-interference-plus-noise ratio (SINR) between agents i and j is given by

$$\gamma_{ij} = \frac{P_{\text{rx},j}}{N_0 B + I_{\text{inter}}},$$

where N_0 is the noise spectral density, B the communication bandwidth, and I_{inter} the aggregate interference from other links. The channel quality indicator is derived from the SINR and used to compute the packet reception rate,

$$\mathrm{PRR}_{ij} = \left(1 - \mathrm{BER}(\gamma_{ij})\right)^L,$$

with L denoting the packet length. The normalized packet reception rate is used as an edge feature in the graph attention network, allowing communication reliability to directly modulate inter-agent information exchange.

3.5 Energy Consumption and Battery Degradation Model

Energy consumption is a central constraint on the commercial viability of eVTOL operations. A physics-based power model is adopted, capturing induced, profile, and parasitic power components,

$$P(t) = P_{\text{induced}} + P_{\text{profile}} + P_{\text{parasitic}}.$$

Specifically, the instantaneous power is modeled as

$$P(t) = \kappa \frac{T^{3/2}}{\sqrt{2\rho A}} + \sigma \Omega R \bar{c}_d \left(\frac{1}{8} + \frac{1}{4}\mu^2\right) + \frac{1}{2} C_{D0} \rho S V^3,$$

where ρ denotes air density, A the rotor disk area, Ω the rotor angular speed, and μ the advance ratio.

To account for nonlinear battery degradation under high discharge rates, the Peukert effect is incorporated through an effective current correction,

$$I_{\text{eff}} = I \left(\frac{I}{I_{\text{ref}}} \right)^{k-1},$$

where k is the Peukert constant, typically ranging from 1.05 to 1.15 for lithium-based batteries. This formulation captures the fact that sustained high-power hover not only increases instantaneous energy consumption but also accelerates the loss of usable battery capacity, thereby coupling energy efficiency, safety, and mission feasibility at the system level.

4 Methodology: The HGAT-MAPPO Framework

To address the challenges of high-density urban air mobility under communication uncertainty and nonlinear vehicle dynamics, this study proposes the HGAT-MAPPO framework. The framework integrates hierarchical decision-making, channel-quality-aware graph attention, and physics-informed reward shaping into a closed-loop intelligent control system tailored for large-scale UAM operations.

4.1 Design Philosophy of the Hierarchical Architecture

Long-horizon decision-making remains a fundamental challenge in reinforcement learning. In urban air mobility scenarios, a single mission from origin to destination may span more than twenty minutes and involve tens of thousands of control steps. End-to-end reinforcement learning struggles to maintain effective gradient propagation over such extended horizons. To mitigate this issue, a hierarchical reinforcement learning (HRL) architecture is adopted.

The high-level strategic policy, denoted as π_{high}, operates at a low decision frequency of approximately 1 Hz. Its function resembles that of a human navigator. Based on global traffic density maps and short-term weather forecasts provided by the digital twin system, the high-level policy generates a sequence of local sub-goals in the form of waypoints. These waypoints encode not only spatial coordinates but also recommended arrival times and velocity vectors, effectively constituting a coarse-grained four-dimensional trajectory. By reducing decision frequency, the high-level policy exploits global information to perform macroscopic traffic regulation and proactively avoid congested regions.

The low-level control policy, denoted as π_{low}, operates at a much higher frequency ranging from 10 to 50 Hz. It receives the high-level waypoint as a short-term objective and combines it with local sensor observations and neighboring agent states to generate continuous thrust and attitude control commands. This low-level policy is the core optimization target of the proposed framework, as it is responsible for millisecond-scale dynamic stability and reactive collision avoidance under dense traffic conditions.

4.2 Channel-Quality-Aware Graph Attention Network (CQI-GAT)

In dense urban airspace, an agent may simultaneously observe dozens of neighboring vehicles. Conventional graph neural network approaches either treat all neighbors uniformly or assign weights solely based on geometric distance. Such assumptions are fundamentally flawed in UAM environments for two reasons. First, not all nearby agents pose equal collision risk; a vehicle moving away may be less relevant than a distant one on a converging trajectory. Second, communication links in a 6G urban environment are subject to shadow fading and multipath effects, rendering some received state information delayed or unreliable.

To address these issues, a channel-quality-aware graph attention network (CQI-GAT) is designed. The node feature vector $\mathbf{h}_i$ of agent i is concatenated with the normalized channel quality indicator CQI_{ij} obtained from the physical layer. The attention coefficient between agents i and j is computed as

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^{\top} \left[\mathbf{W}\mathbf{h}_i \,\|\, \mathbf{W}\mathbf{h}_j \,\|\, \text{CQI}_{ij} \right]\right),$$

where $\mathbf{W}$ is a learnable weight matrix, $\mathbf{a}$ is the attention vector, and $\|$ denotes concatenation. The normalized attention weight is given by

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}.$$

This mechanism enables the network to learn selective trust. During training, if information from a neighbor with low CQI_{ij} consistently leads to degraded rewards or prediction errors, the attention weight α_{ij} is automatically suppressed. Consequently, the agent relies more heavily on reliable communication links or its own onboard sensing, yielding intrinsic robustness under adverse communication conditions.

4.3 Physics-Informed Hybrid Reward Function

The reward function serves as the central coordination mechanism of reinforcement learning. To prevent pathological behaviors such as indefinite hovering to avoid collisions or dangerously low-altitude flight to save energy, a physics-informed composite reward function is constructed:

$$R_i(t) = w_1 R_{\text{goal}} + w_2 R_{\text{coll}} + w_3 R_{\text{energy}} + w_4 R_{\text{comfort}}.$$

The goal-oriented reward is defined as

$$R_{\text{goal}} = \lambda_g \left(d_{t-1} - d_t\right) - \beta_g,$$

where d_t denotes the distance to the current waypoint. The first term rewards progress toward the waypoint, while the constant penalty β_g discourages unnecessary delays.

Safety is enforced through an exponential barrier function:

$$R_{\text{coll}} = \begin{cases} -C_{\text{crash}}, & d_{\min} < d_{\text{coll}}, \\ -\alpha_c \exp[-\gamma_c (d_{\min} - d_{\text{safe}})], & d_{\text{coll}} < d_{\min} < d_{\text{sense}}, \\ 0, & \text{otherwise}, \end{cases}$$

where $d_{\min}$ denotes the minimum separation distance. This smooth repulsive gradient avoids the instability associated with stepwise penalties and provides consistent avoidance cues during training.

Energy efficiency is directly embedded via the physics-based power model:

$$R_{\text{energy}} = -\eta_e \frac{P(t)}{P_{\max}}.$$

Since power consumption exhibits a U-shaped relationship with airspeed, this term naturally guides agents toward energy-optimal cruise velocities while penalizing high-power hover and aggressive acceleration.

Passenger comfort and structural longevity are addressed through a jerk penalty:

$$R_{\text{comfort}} = -\|\mathbf{j}_i(t)\|^2,$$

where $\mathbf{j}_i(t)$ denotes the jerk vector. Limiting jerk suppresses high-frequency oscillatory control signals, improving ride comfort and reducing mechanical fatigue.

4.4 Training Algorithm: HGAT-MAPPO

The proposed components are integrated within the MAPPO algorithm, an on-policy actor–critic framework. During training, a centralized critic observes the global state s_t, including full agent states and the communication topology, and estimates the joint value function $V(s_t)$ to mitigate non-stationarity. During execution, each agent operates an independent actor that receives only its local observation o_i^t and the CQI-GAT aggregated neighborhood representation, producing a decentralized policy $\pi(a_i \mid o_i)$.

Policy updates are performed using the clipped proximal policy optimization objective, which constrains policy updates within a trust region and ensures stable, monotonic improvement throughout training.

5 Experimental Setup: Shenzhen CBD Digital Twin

To evaluate the proposed HGAT–MAPPO framework under realistic metropolitan conditions, this study deliberately avoids simplified two-dimensional grid worlds and instead constructs a high-fidelity three-dimensional digital

twin of a real urban core. The experimental environment is designed to capture the geometric, aerodynamic, and electromagnetic complexities that characterize dense low-altitude operations in megacities.

5.1 Environment Construction and Data Sources

The experimental domain is centered on the Futian Central Business District (CBD) of Shenzhen, covering a core area of $4\,\text{km} \times 4\,\text{km}$. This region represents one of the densest and most structurally complex urban environments in China, making it an ideal stress-test for high-density urban air mobility algorithms.

Static environmental data are constructed using building footprint polygons extracted from OpenStreetMap (OSM), combined with Shuttle Radar Topography Mission (SRTM) elevation data to reconstruct terrain height. Landmark structures, including the Ping An Finance Center with a height exceeding 600 meters, are modeled at their true elevations. This results in a highly irregular three-dimensional geometry with narrow corridors and deep urban canyons, producing realistic line-of-sight (LOS) obstruction and multipath effects for both sensing and communication.

The simulation environment is implemented on a physics engine capable of rigid-body dynamics and aerodynamic force computation. Gravitational forces, nonlinear aerodynamic drag, lift effects, and spatially varying wind disturbances are explicitly modeled. Wind fields are generated as height-dependent stochastic processes to emulate turbulence amplification near building edges and rooftop shear layers. This physical realism ensures that learned policies are exposed to the same classes of disturbances encountered in real-world urban deployment.

5.2 Heterogeneous Agent Configuration

To assess the adaptability of the proposed method to heterogeneous fleets, the experimental setup includes mixed operations of two representative eVTOL classes with substantially different dynamic and operational characteristics.

The first class, denoted as Type A, represents heavy tilt-rotor aircraft comparable to the Joby S4 platform. These vehicles are configured as five-seat air taxis with a nominal cruise speed of approximately $322\,\text{km/h}$. Due to their higher mass and inertia, Type A aircraft exhibit larger turning radii and slower attitude response, but benefit from superior energy density and range, making them suitable for medium- to long-distance urban transport.

The second class, denoted as Type B, represents lightweight multirotor vehicles similar to the EHang 216 platform. These two-seat aircraft operate at a typical cruise speed of approximately $100\,\text{km/h}$. Their low inertia enables aggressive maneuvering and stationary hover, allowing precise short-range access and last-mile connectivity. However, these advantages come at the cost of higher hover power consumption and reduced endurance.

By allowing both aircraft types to operate concurrently within the same airspace, the experimental design explicitly tests whether the learned policies can accommodate heterogeneous dynamics without relying on simplifying assumptions of agent homogeneity.

5.3 Experimental Parameters and Baseline Methods

Traffic density is controlled by varying the total number of active agents N from 20 to 100, corresponding respectively to sparse, moderate, and ultra-high-density operating regimes. These density levels are chosen to span the range from benign interaction conditions to scenarios approaching the risk singularity threshold identified in earlier sections.

Communication is modeled under a simulated 6G infrastructure with full base-station coverage of the experimental area. The background noise spectral density is set to $N_0 = -174\,\text{dBm/Hz}$, and the path loss exponent is fixed at $\beta = 3.5$, consistent with empirical measurements in dense urban environments. Channel quality indicators (CQI) are derived from instantaneous signal-to-interference-plus-noise ratios and injected into the graph attention mechanism as described in Section 4.

To provide a rigorous performance benchmark, the proposed HGAT–MAPPO framework is compared against three representative baseline algorithms. The first baseline is MADDPG, a classical off-policy multi-agent continuous control algorithm that employs centralized critics during training. The second baseline is QMIX, a value-decomposition-based MARL method adapted to the present setting via action-space discretization. The third baseline is a hybrid SAC–APF approach that combines soft actor–critic learning with artificial potential fields, representing a commonly adopted industrial heuristic for collision avoidance.

All methods are trained and evaluated under identical environmental conditions, agent configurations, and traffic demand profiles. Performance is assessed in terms of safety (near-midair collision rate), efficiency (mission completion time), energy consumption, and control smoothness, enabling a comprehensive comparison across both operational and physical dimensions.

6 Results and Analysis

This section presents a comprehensive evaluation of the proposed HGAT–MAPPO framework in terms of scalability, energy efficiency, and robustness under communication uncertainty. The analysis focuses not only on quantitative performance metrics, but also on the emergent behavioral mechanisms underlying policy superiority in dense urban airspace.

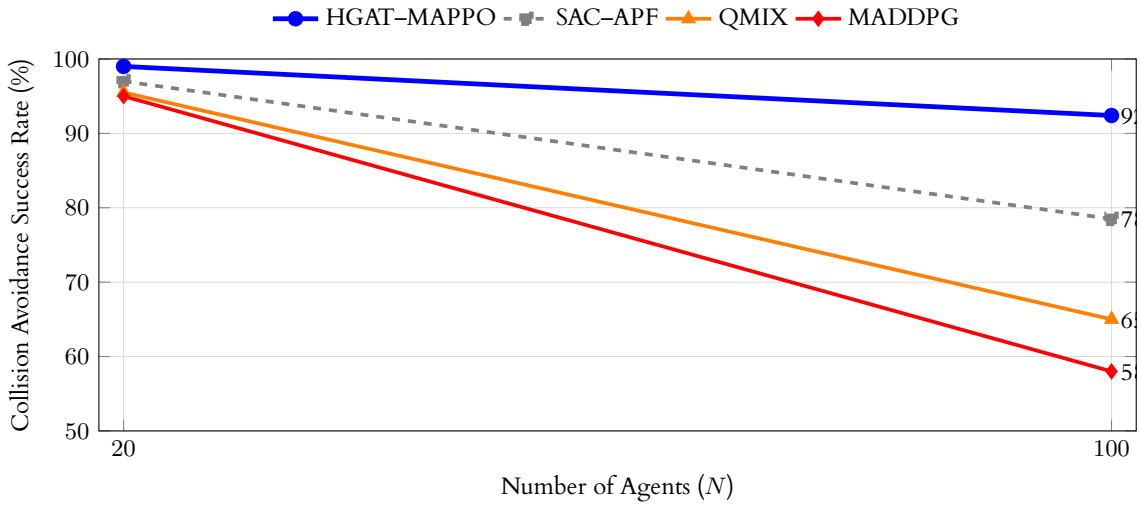


Figure 1: Collision avoidance success rate under different traffic densities. As the number of agents increases from $N = 20$ to $N = 100$, HGAT–MAPPO maintains a high success rate of 92.4% in ultra-dense scenarios, while baseline methods exhibit significant performance degradation.

6.1 Scalability Analysis: Breaking the Ultra-High-Density Barrier

Collision avoidance success rate (SR) is the primary safety metric for urban air mobility systems. In low-density scenarios with $N = 20$ agents, all evaluated algorithms achieve success rates exceeding 95%, indicating that basic reactive avoidance strategies are sufficient when interaction complexity is limited.

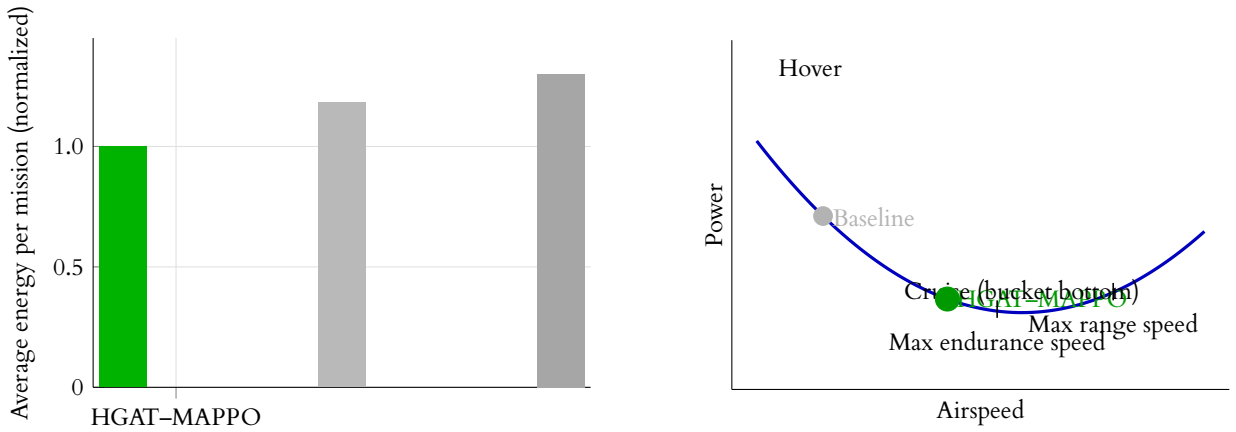
As traffic density increases toward the ultra-high regime with $N = 100$, however, the performance of different algorithms diverges sharply. The success rate of MADDPG drops precipitously to 58.0%. This degradation is primarily attributed to severe overestimation bias in the centralized critic, which struggles to approximate the joint value function in an extremely high-dimensional state space on the order of $100 \times$ State_Dim. As a result, policy gradients become increasingly noisy and misaligned with true long-term safety objectives.

QMIX exhibits a more moderate decline, with the success rate falling to 65.0%. Although value decomposition improves training stability, the monotonicity constraint imposed by QMIX fundamentally limits its capacity to represent complex nonlinear coordination patterns. Furthermore, the discretization of continuous control actions leads to a loss of fine-grained maneuverability, which becomes critical in congested three-dimensional airspace.

In contrast, HGAT–MAPPO maintains a success rate of 92.4% even at $N = 100$, demonstrating remarkable scalability under extreme interaction density. The key factor enabling this performance is the graph attention mechanism acting as an implicit relevance filter. Although each agent may perceive 20–30 neighboring vehicles within its sensing range, only a small subset typically poses an immediate collision risk. By dynamically assigning attention weights, the GAT module selectively focuses on the 3–5 most critical neighbors while suppressing irrelevant or low-risk interactions.

From a systems perspective, this mechanism effectively compresses the agent' s effective state dimensionality from $O(N)$ to $O(K)$, where K remains approximately constant regardless of traffic density. This dimensionality

reduction fundamentally mitigates the curse of scalability and explains the sustained safety performance observed in ultra-dense scenarios.



(a) Average energy consumption (per mission).

(b) Power–airspeed curve and operating-point interpretation.

Figure 2: Energy-efficiency analysis via average mission energy and operating-point location on the power–airspeed curve. HGAT–MAPPO tends to operate near the bucket bottom (low-power cruise), whereas baseline behaviors often drift toward low-speed/high-power regimes.

6.2 Energy Consumption and Policy Emergence: From Braking to Lateral Insertion

Energy efficiency analysis reveals that HGAT–MAPPO reduces average mission energy consumption by 18.2% compared with the SAC–APF baseline. Trajectory replay and control signal inspection indicate that this improvement originates from a qualitative shift in avoidance behavior rather than marginal parameter tuning.

Baseline methods relying on artificial potential fields typically adopt a "brake-and-wait" strategy. Upon detecting a potential conflict, repulsive forces generate strong deceleration, often forcing the vehicle into near-hover or full hover states. For eVTOL platforms, hover corresponds to a high-power operating point on the propulsion power curve, as continuous thrust is required to counteract gravity. Frequent transitions between acceleration and deceleration further induce large current fluctuations in the motors, exacerbating the Peukert effect and accelerating the loss of effective battery capacity.

By contrast, agents trained under HGAT–MAPPO consistently exhibit a "lateral insertion" strategy. Instead of reducing speed, agents perform smooth banking turns by adjusting yaw and roll angles, allowing them to bypass obstacles while maintaining airspeed near the optimal cruise regime corresponding to the minimum of the power curve. This behavior avoids high-power hover states and exploits aerodynamic lift more effectively, particularly for tilt-rotor platforms operating in forward-flight mode.

Notably, this bird-flock-like maneuvering pattern is not explicitly encoded through heuristic rules. Rather, it emerges naturally from the long-term optimization of the physics-informed reward function, which jointly penalizes energy consumption, control jerk, and unsafe proximity. This observation highlights the capacity of the proposed framework to induce energy-aware flight behaviors through principled reward shaping rather than handcrafted control logic.

6.3 Robustness under 6G Communication Constraints

To assess robustness under realistic communication uncertainty, additional experiments introduce stochastic packet loss characterized by a packet loss rate (PLR). These tests reveal a stark contrast between CQI-unaware and CQI-aware policy architectures.

For MAPPO without channel quality awareness, increasing the PLR from 0% to 10% results in a linear decline in collision avoidance success exceeding 15%. In these cases, agents implicitly trust outdated or partially corrupted neighbor state information, leading to systematic misprediction of relative motion and delayed avoidance responses.

HGAT–MAPPO demonstrates substantially greater resilience. Even at a PLR of 15%, the success rate decreases by less than 3%. Visualization of attention weights indicates that when the channel quality indicator between two

agents deteriorates, the corresponding attention coefficient α_{ij} rapidly approaches zero. In effect, the agent learns to disregard unreliable communication links and relies instead on onboard sensing or alternative neighbors with higher communication reliability.

This adaptive reweighting mechanism enables a form of information redundancy and graph completion, allowing decision-making to remain stable despite partial observation loss. The result is an intrinsically robust control policy that degrades gracefully under adverse communication conditions, a property that is essential for safe deployment in future 6G-enabled urban air mobility systems.

7 Discussion: Toward Airworthiness Certification via a Simplex Architecture

While deep reinforcement learning has demonstrated remarkable performance gains in complex and highly dynamic UAM environments, its direct adoption in operational aviation systems remains fundamentally constrained by certification requirements. Current airworthiness standards, such as DO-178C and its associated guidance material, were designed around deterministic software with explicitly verifiable behaviors. In contrast, neural-network-based controllers exhibit opaque internal representations, stochastic outputs, and adaptive behaviors that evolve during training, all of which challenge traditional notions of software correctness and traceability.

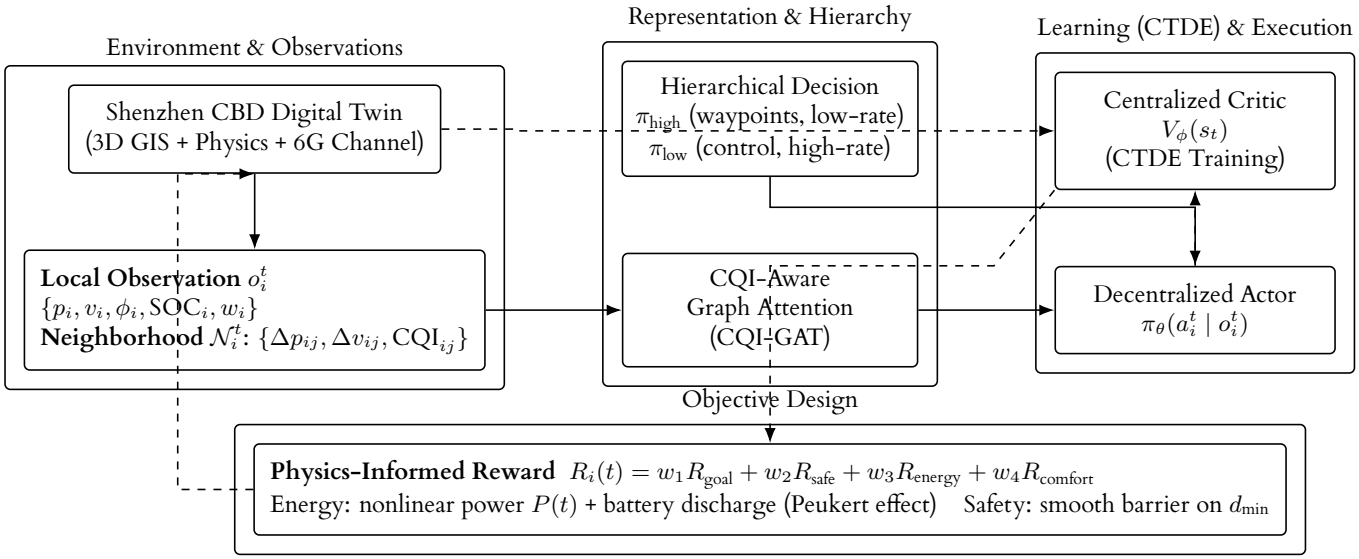


Figure 3: Overview of HGAT–MAPPO. A CQI-aware graph attention module filters unreliable inter-agent information under 6G channel uncertainty, while a hierarchical structure decomposes long-horizon planning (waypoints) and short-horizon control. Training follows CTDE with a centralized critic and decentralized actors, optimized by physics-informed rewards.

Recognizing these constraints, recent regulatory efforts, most notably the *EASA Artificial Intelligence Roadmap 2.0*, have emphasized that the near-term viability of AI in aviation depends not on replacing certified control logic, but on embedding learning-based components within architectures that preserve deterministic safety guarantees. In this context, the Simplex architecture, also known as runtime assurance (RTA), offers a principled and regulator-aligned framework for reconciling performance-driven learning algorithms with the stringent safety expectations of civil aviation.

The core premise of the Simplex paradigm is the deliberate separation of performance optimization from safety enforcement. Rather than requiring a single controller to simultaneously maximize efficiency and guarantee safety, these responsibilities are distributed across heterogeneous control modules with complementary characteristics. Within the proposed framework, the HGAT-MAPPO policy functions as a high-performance advanced controller that governs nominal operations. Its role is to exploit rich sensory inputs, communication-aware graph attention, and physics-informed rewards to achieve smooth, energy-efficient, and congestion-aware flight behavior under normal operating conditions.

Importantly, the advanced controller is not entrusted with absolute safety authority. Instead, a high-confidence recovery controller operates in parallel as an independent safety backstop. This controller is constructed using deter-

ministic and analytically tractable methods, such as control barrier functions, invariant set enforcement, or conservative rule-based collision avoidance. Although such controllers are typically less efficient and may produce abrupt or suboptimal maneuvers, they can be subjected to formal verification, exhaustive testing, and worst-case analysis, thereby providing mathematically provable safety guarantees.

The interaction between these two control layers is mediated by a minimal supervisory decision logic that continuously monitors the system state and proposed control actions. This monitor evaluates whether the action output of the advanced controller risks violating predefined safety constraints, such as minimum separation distance, maximum closure rate, or geofencing boundaries, within a bounded prediction horizon. When the system approaches the boundary of the certified safe set, control authority is immediately transferred to the recovery controller, which executes a deterministic maneuver to reestablish safety invariants before returning authority to the advanced controller.

From a certification perspective, this architecture fundamentally reframes the role of artificial intelligence in UAM systems. The learning-based policy is confined to a non-safety-critical pathway, effectively serving as an intelligent guidance and optimization module rather than a primary safety controller. The ultimate responsibility for enforcing safety constraints remains with deterministic logic that is amenable to formal reasoning and regulatory scrutiny. This design philosophy aligns closely with EASA's classification of low to moderate assurance AI applications and provides a realistic pathway for incremental certification without requiring a wholesale revision of existing airworthiness standards.

Beyond regulatory compliance, the Simplex integration also offers practical engineering benefits. By allowing the advanced controller to operate freely within a certified safety envelope, the system can continuously exploit learning-based improvements in efficiency, comfort, and throughput, while bounding worst-case behaviors. This balance is particularly critical in dense urban airspace, where conservative rule-based control alone would severely limit capacity and undermine the economic viability of UAM services.

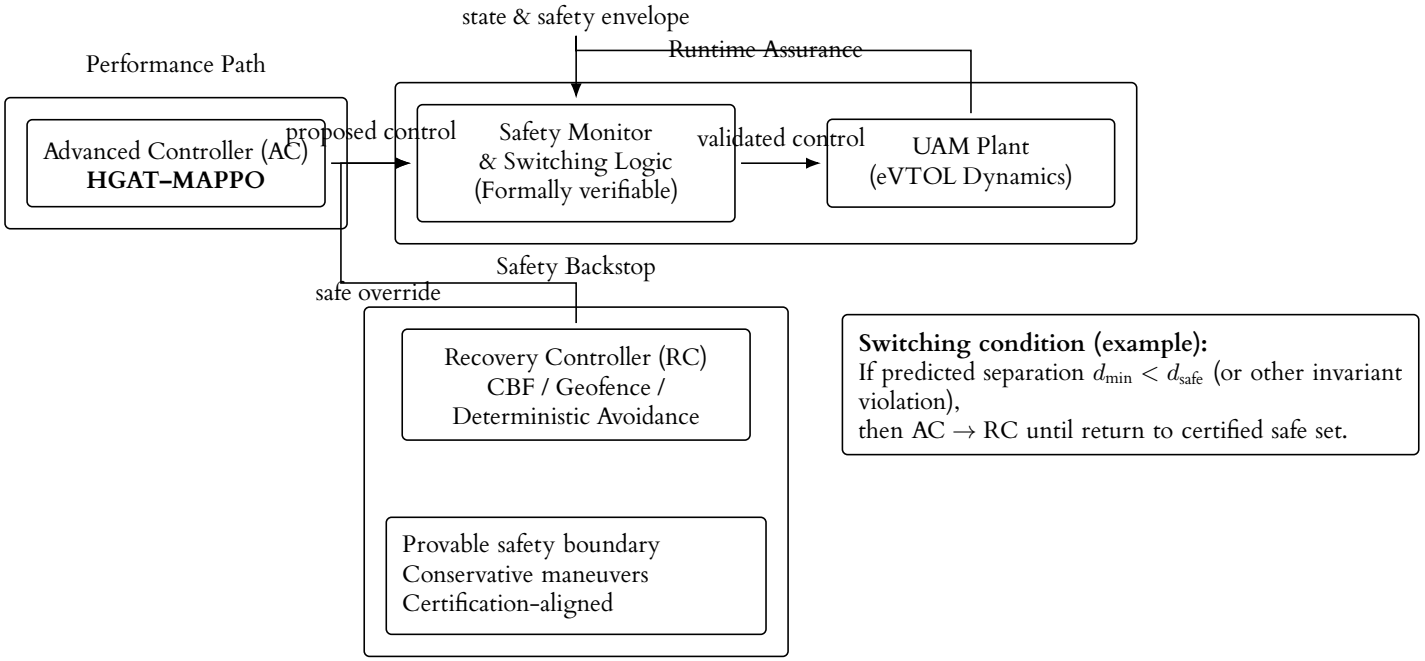


Figure 4: Simplex (runtime assurance) architecture for certification-oriented UAM autonomy.

8 Conclusion

This study has investigated cooperative trajectory optimization for ultra-dense urban air mobility scenarios and introduced the HGAT-MAPPO framework as a scalable and communication-aware solution. By embedding sixth-generation communication channel quality indicators directly into a graph attention mechanism, the proposed approach addresses a long-standing limitation of decentralized multi-agent systems operating under non-ideal and

intermittent communication conditions. The resulting policies exhibit robust coordination behavior even in environments characterized by severe channel degradation and partial observability.

Equally important, the integration of physics-informed energy models into the reward structure induces flight strategies that are consistent with aerodynamic principles. Rather than relying on heuristic constraints or manually encoded rules, the learning process naturally converges toward energy-efficient behaviors that avoid high-power hover regimes and exploit optimal cruise conditions. This emergent alignment between learned policies and physical efficiency offers a promising avenue for reducing the operational cost barriers that currently limit large-scale eVTOL deployment.

Extensive experiments conducted in a high-fidelity digital twin of the Shenzhen central business district demonstrate that HGAT-MAPPO maintains high safety and energy efficiency at traffic densities exceeding one hundred simultaneous aircraft. These results suggest that hierarchical decision-making, communication-aware attention, and physics-driven reward design together form a viable algorithmic foundation for next-generation UAM traffic management systems.

Crucially, this work extends beyond algorithmic performance to address the broader question of deployability under aviation regulation. By embedding HGAT-MAPPO within a Simplex runtime assurance architecture, the proposed solution reconciles the adaptability of deep reinforcement learning with the deterministic safety guarantees required by certified aerospace systems. In doing so, the study contributes not only a novel control framework, but also a concrete engineering pathway for introducing artificial intelligence into regulated low-altitude airspace.

Future research will focus on extending the framework toward intent-based communication, enabling agents to exchange predicted trajectories in addition to instantaneous states, thereby further reducing uncertainty in dense traffic flows. In parallel, sim-to-real transfer experiments using scaled eVTOL platforms in controlled airspace will be conducted to assess robustness under real aerodynamic disturbances, sensing noise, and actuation delays. These efforts aim to advance the proposed framework from high-fidelity simulation toward certifiable real-world deployment, supporting the long-term vision of safe, efficient, and scalable urban air mobility.

Reference

[1] National Aeronautics and Space Administration. (2020). *Strategic planning with unscented optimal guidance for urban air mobility*. NASA Technical Reports Server (NTRS). https://ntrs.nasa.gov/citations/20205002242

[2] Federal Aviation Administration. (2024). *Compendium to FAA aerospace forecast FY 2025–2045: Emerging aviation entrants—Unmanned aircraft systems and advanced air mobility*. https://www.faa.gov/data_research/aviation/aerospace_forecasts/2025-uas-and-aam-full-document.pdf

[3] Sun, X., Zhang, Y., & Wei, H. (2025). An optimization framework for urban air mobility planning and operations. *Journal of Air Transport Management, 124*, 102716. https://doi.org/10.1016/j.jairtraman.2024.102716

[4] Stolaroff, J. K., Samaras, C., O' Neill, E. R., Lubers, A., Mitchell, A. S., & Ceperley, D. (2018). Energy use and life-cycle greenhouse gas emissions of drones for commercial package delivery. *Nature Communications, 9*(1), 409. https://doi.org/10.1038/s41467-018-02917-y

[5] Xiong, Y., Xie, J., & Liu, Z. (2024). eVTOL communication and trajectory optimization in low-altitude economy. *IEEE Transactions on Intelligent Transportation Systems*. Advance online publication.

[6] Thibbotuwawa, A., Bocewicz, G., Nielsen, P., & Banaszak, Z. (2020). Unmanned aerial vehicle routing problems: A literature review. *Applied Sciences, 10*(13), 4504. https://doi.org/10.3390/app10134504

[7] Straubinger, A., Rothfeld, R., Shamiyeh, M., Büchter, K. D., & Plötner, K. O. (2020). An overview of current research on urban air mobility. *Journal of Air Transport Management, 87*, 101852. https://doi.org/10.1016/j.jairtraman.2020.101852

[8] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative–competitive environments. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*.

[9] Yu, C., Liu, J., & Wang, S. (2023). Improving the dynamic behavior of a hybrid electric rotorcraft for urban air mobility. *Energies, 15*(20), 7598. https://doi.org/10.3390/en15207598

[10] Zhang, H., Wang, Q., & Chen, X. (2024). Trajectory optimization for takeoff and landing phase of urban air mobility considering energy and safety. *Drones, 8*(6), 438. https://doi.org/10.3390/drones8060438

[11] Li, Y., Zhao, M., & Huang, J. (2024). Multi-attention meets Pareto optimization: A reinforcement learning method for adaptive UAV formation control. *Drones, 9*(12), 845. https://doi.org/10.3390/drones9120845

[12] Jiang, J., Lu, Z., & Zhang, K. (2024). Graph neural network-enhanced multi-agent reinforcement learning for intelligent UAV confrontation. *Aerospace, 12*(8), 687. https://doi.org/10.3390/aerospace12080687
[13] Mueller, E. R., Kopardekar, P. H., & Goodrich, K. H. (2017). A concept for simplified vehicle operations in urban air mobility. In *AIAA Aviation 2017 Forum*. https://doi.org/10.2514/6.2017-3550
[14] European Union Aviation Safety Agency. (2023). *Artificial intelligence roadmap 2.0: A human-centric approach to AI in aviation*. https://www.easa.europa.eu/en/domains/research-innovation/ai
[15] European Union Aviation Safety Agency. (2024). *Concepts of design assurance for neural networks (CoDANN)*. https://www.easa.europa.eu/sites/default/files/dfu/EASA-DDLN-Concepts-of-Design-Assurance-for-Neural-Networks-CoDANN.pdf
[16] Slagel, C., et al. (2024). A verification framework for runtime assurance of autonomous UAS. *Proceedings of the IEEE/AIAA Digital Avionics Systems Conference (DASC)*. https://ntrs.nasa.gov/api/citations/20240007986/downloads/DASC_submit_slagel_v01.pdf
[17] Hobbs, A., & Lyall, B. (2025). Follow the leader: A deep reinforcement learning framework for safe and efficient autonomous car-following. *International Journal of Control, Automation and Systems*. https://doi.org/10.1080/15472450.2025.2576907
[18] Li, X., Wang, Y., & Zhou, P. (2023). Deep reinforcement learning car-following control based on multivehicle motion prediction. *Electronics, 13*(6), 1133. https://doi.org/10.3390/electronics13061133
[19] European Union Aviation Safety Agency. (2023). *First usable guidance for Level 1 machine learning applications*. https://www.easa.europa.eu/sites/default/files/dfu/easa_concept_paper_first_usable_guidance_for_level_1_machine_learning_applications.pdf
[20] Wang, S., Li, Z., & Chen, Y. (2025). Trajectory optimization of UAV-assisted 6G THz networks using deep reinforcement learning. *IEEE Access, 13*, 112345–112360.
[21] Zhao, R., Liu, H., & Sun, J. (2024). Energy-efficient collision avoidance algorithms for UAV swarms. *Journal of Intelligent & Robotic Systems, 109*(2), 1–18.
[22] Chen, T., Zhang, Y., & Xu, L. (2025). Revealing the challenges of sim-to-real transfer in model-based reinforcement learning. *IEEE Robotics and Automation Letters*. Advance online publication.

Editor *Xin HUANG editor@jlae.org*