Reinforcement Learning for Uplink Access Optimization in UAV-Assisted 5G Networks Under Emergency Response

Abid Mohammad Ali *, Petro Mushidi Tshakwanda *, Henok Tsegaye, Harsh Kumar, Md Najmus Sakib, Raddad Almaayn, Ashok Karukutla, Michael Devetsikiotis

Posted Date: 24 September 2025

doi: 10.20944/preprints202509.2037.v1

Keywords: UAV-assisted 5G; Uplink NOMA; successive interference cancellation (SIC); uplink; PPOGAE; Emergency Communications

Article

Reinforcement Learning for Uplink Access Optimization in UAV-Assisted 5G Networks Under Emergency Response

Abid Mohammad Ali [1,*], Petro Mushidi Tshakwanda [1,*], Henok Tsegaye [1], Harsh Kumar [2], Md Najmus Sakib [3], Raddad Almaayn [1], Ashok Karukutla [1] and Michael Devetsikiotis [1]

[1] Department of Electrical and Computer Engineering, The University of New Mexico, Albuquerque, NM, USA
[2] R.B. Annis School of Engineering, University of Indianapolis, Indianapolis, IN, USA
[3] University of the Cumberlands, Williamsburg, KY, USA
* Correspondence: abid665@unm.edu (A.M.A.); pmushidi@unm.edu (P.M.T.)

Abstract

We study Unmanned Aerial Vehicle (UAV) assisted 5G uplink connectivity for disaster response, where a UAV acts as an aerial base station to restore service to ground users. We formulate a joint control problem coupling UAV kinematics (bounded acceleration and velocity), per-subchannel uplink power allocation, and uplink Non-Orthogonal Multiple Access (UL-NOMA) scheduling with adaptive successive interference cancellation (SIC) under a minimum user-rate constraint. The wireless channel follows 3GPP urban macro (UMa) with probabilistic Line of Sight/Non Line of Sight (LoS/NLoS), realistic receiver noise and noise figure, and user equipment (UE) transmit-power limits. We propose a bounded-action proximal policy optimization with generalized advantage estimation (PPO-GAE) agent that parameterizes acceleration and power with squashed distributions and enforces feasibility by design. Across four user distributions (clustered, uniform, ring, edge-heavy) and multiple rate thresholds, our method increases the fraction of users meeting the target rate by $8.2 - 10.1$ percentage points over strong baselines (OFDMA with heuristic placement, PSO-based placement/power, PPO without NOMA) while reducing median UE transmit power by 64.6%. Results are averaged over ≥ 5 random seeds with 95% confidence intervals; ablations isolate the gains from NOMA, adaptive SIC order, and bounded action parameterization. We discuss robustness to imperfect SIC and CSI errors, and release code/configurations to support reproducibility.

Keywords: UAV-assisted 5G; Uplink NOMA; successive interference cancellation (SIC); uplink; PPO-GAE; Emergency Communications

1. Introduction

Fifth-generation (5G) mobile networks deliver high data rates and stringent quality-of-service (QoS) that guarantees low latency, high throughput, and reliable coverage through a dense and flexible radio access network (RAN). In adverse situations such as natural disasters, power outages, or sudden traffic surges, however, fixed terrestrial base stations (BSs) may become unavailable or severely degraded. In these cases, rapidly deployable unmanned aerial vehicle base stations (UAV-BSs) offer a practical means to restore coverage and capacity. Yet, realizing dependable uplink connectivity with a UAV-BS is challenging: the air-to-ground (A2G) channel is dynamic and height-dependent, flight and power budgets are constrained, and user scheduling must respect minimum-rate requirements while coping with inter-user interference.

This work considers a scenario in which a fixed BS becomes inoperable and a single UAV-BS is dispatched to serve affected users. We target the uplink and adopt non-orthogonal multiple access (NOMA) with adaptive successive interference cancellation (SIC) to increase spectral efficiency under minimum per-user throughput constraints. The decision-making problem is inherently continuous and

 doi:10.20944/preprints202509.2037.v1

coupled: the UAV must select its 3D motion while the network jointly schedules users and allocates per-subchannel transmit powers. Classical trajectory planners (e.g., particle swarm or direct search) can struggle with scalability and non-stationarity, and value-based deep reinforcement learning (DRL) methods such as deep Q networks (DQN) operate in discrete action spaces and may suffer from overestimation bias and target-network instability [1–5]. In contrast, policy-gradient methods directly optimize in continuous action spaces and can offer improved stability.

Motivated by these considerations, we develop a continuous-control actor–critic solution based on proximal policy optimization with generalized advantage estimation (PPO–GAE). PPO's clipped surrogate objective improves training stability, while the GAE estimator balances bias–variance trade-offs for sample-efficient learning. We further enforce *bounded* actions to respect flight envelopes and power limits, ensuring safe-by-construction decisions. The resulting agent jointly controls UAV kinematics and uplink resource allocation to maximize the number of users whose minimum-rate constraints are satisfied.

Research gap. Existing surveys synthesize broad UAV networking applications but do not provide a unified, learning-based formulation that *jointly* optimizes UAV motion, uplink NOMA scheduling with adaptive SIC, and per-subchannel power allocation under minimum-rate constraints in a realistic 3GPP A2G setting [6]. Metaheuristic trajectory designs (e.g., particle swarm and direct search) can improve channel quality [7], yet they usually decouple motion from radio resource management and do not exploit continuous-control RL. Offline neural surrogates for throughput prediction and deployment planning [8] bypass closed-loop control and are less responsive to fast channel and traffic fluctuations than on-policy methods such as PPO–GAE. Works on spectrum/energy efficiency in cognitive UAV networks [9] and DRL for *downlink* multi-UAV systems under fronthaul limits [10] address important but different regimes; they neither tackle the *uplink* NOMA case with adaptive SIC nor the bounded continuous-action control that jointly handles UAV kinematics and per-subchannel power in disaster-response scenarios. Consequently, there remains a need for a stable, continuous-control, learning-based framework that closes this gap.

Our contributions are summarized as follows:

- **Joint control formulation.** We pose a coupled optimization that integrates UAV kinematics, *uplink* NOMA scheduling with adaptive SIC ordering, and per-subchannel power allocation under minimum user-rate constraints, with the objective of maximizing the number of served users.
- **Bounded-action PPO–GAE agent.** We design a continuous-action actor–critic algorithm (PPO–GAE) with explicit action bounding for flight and power feasibility, yielding stable learning and safe-by-construction decisions.
- **Realistic A2G modeling & robustness.** We employ a 3GPP-compliant A2G channel and evaluate robustness to imperfect SIC and channel-state information (CSI), capturing practical impairments often overlooked in prior art.
- **Ablation studies.** We isolate the gains due to (i) NOMA vs. OMA, (ii) adaptive SIC ordering, and (iii) bounded action parameterization, and quantify their individual and combined benefits.
- **Reproducibility.** We release complete code and configurations to facilitate verification and extension by the community.

Why PPO instead of DQN? Unlike DQN, which assumes a discrete and typically small action space and is prone to overestimation bias and target-network lag—PPO directly optimizes a stochastic policy over *continuous* actions and uses a clipped objective to curb destructive policy updates. This is well aligned with the continuous, multi-dimensional action vector arising from simultaneous UAV motion and power-control decisions, and it yields improved training stability and sample efficiency compared with value-based baselines [1–5].

The remainder of this paper is organized as follows. Section 2 reviews related work on UAV-enabled cellular systems, NOMA scheduling, and DRL for wireless control. Section 3 details the system model, problem formulation, and the proposed bounded-action PPO–GAE algorithm. Section 4

 doi:10.20944/preprints202509.2037.v1

presents quantitative results, including ablations and robustness analyses. Section 5 discusses insights, practical implications, and limitations. Section 6 concludes the paper and outlines future directions.

2. Related Work

UAV-assisted 5G networking has attracted sustained interest across wireless communications, while reinforcement learning (RL) has emerged as a powerful tool for control and resource optimization in nonstationary environments. Within this broad landscape, our work targets a specific and underexplored setting: *uplink* emergency connectivity restoration with a single UAV acting as an aerial base station, under realistic 3GPP urban macro (UMa) air-to-ground channels and practical device constraints.

In [11], the authors study energy sustainability for UAVs via wireless power transfer from flying energy sources, coordinating multiple agents with multi-agent DRL (MADRL). Their objective emphasizes maximizing transferred energy and coordinating energy assets. By contrast, we address emergency connectivity restoration for ground users with a *single* aerial base station, focusing on minimum-rate coverage under UE power limits and receiver noise. Methodologically, we employ a bounded-action PPO–GAE agent to jointly control UAV kinematics and uplink resource allocation, whereas [11] centers on energy-transfer optimization and multi-agent coordination.

Trajectory learning without side information is demonstrated in [12], where deterministic policy gradients operate in a continuous deterministic action space to learn UAV paths. Our formulation differs in scope and modeling: we couple UAV motion with *uplink* NOMA scheduling (with adaptive SIC) and per-subchannel power allocation, and we train an actor–critic PPO–GAE agent under realistic 3GPP UMa LoS/NLoS channels with rigorous ablations isolating the effects of NOMA, SIC ordering, and bounded action parameterization.

A broad survey in [13] reviews supervised, unsupervised, semi-supervised, RL, and deep learning techniques for UAV-enabled wireless systems, highlighting the promise of learning-based control. Our approach contributes to this line by casting emergency uplink access as a *continuous-control* problem and by leveraging PPO–GAE with action squashing to ensure feasibility under flight and power constraints.

Work in [14] considers multiple UAVs serving as aerial base stations during congestion, aiming to maximize throughput. The solution combines k-means clustering with a DQN variant, separating user clustering from UAV control. In contrast, we focus on disaster-response scenarios where establishing connectivity with minimum-rate guarantees is paramount; we jointly optimize motion, UL-NOMA scheduling with adaptive SIC, and per-subchannel power in a single learning loop. Unlike [14], our setting enforces minimum-rate fairness, adopts 3GPP-compliant channel modeling, and respects UE transmit-power limits, while avoiding the discretization and overestimation issues that can affect DQN in continuous domains.

The authors of [15] investigate UAV-aided MEC trajectory optimization for IoT latency/QoE, primarily benchmarking computing-centric baselines. Our problem is communication-centric: we model 3GPP UMa LoS/NLoS propagation, receiver noise figures, and UE power caps, and we optimize the *uplink* access process itself rather than edge-computing pipelines.

Energy-efficiency maximization with quantum RL is explored in [16], where a layerwise quantum actor–critic with quantum embeddings is proposed. While they mention disaster recovery, their primary metric is energy efficiency. We target user-side QoS during emergencies, adopting bounded-action PPO–GAE (with squashed distributions) to stabilize continuous control under kinematic and power constraints; our method is immediately deployable on classical hardware and directly aligned with current 5G UAV-assisted systems.

Path planning for post-disaster environments is addressed in [17] via an Adaptive Grey Wolf Optimization (AGWO) algorithm focused on trajectory efficiency. Our formulation instead treats a *joint* communication–control problem for UAV-assisted *uplink* access with NOMA and adaptive SIC, solved via a continuous-control RL agent.

 doi:10.20944/preprints202509.2037.v1

Finally, [18] studies joint resource allocation and UAV trajectory optimization in *downlink* UAV-NOMA networks with QoS guarantees using a heuristic matching-and-swapping scheduler and convex optimization. We consider the complementary *uplink* case in disaster response, replacing heuristic matching with an RL-driven policy (bounded-action PPO–GAE) that adapts online across varied user spatial distributions.

2.1. UAV Path and Trajectory Optimization: Prior Art and Research Gap

Trajectory and placement optimization for UAVs spans surveillance, mapping, IoT data collection, and cellular augmentation. Surveys synthesize challenges in 3D placement and motion planning under realistic constraints, emphasizing the coupling between mobility and communication objectives [19–25]. Algorithmically, metaheuristics (e.g., improved RRT with ACO) address obstacle avoidance; continuous-control RL methods (e.g., DDPG, TD3) have been applied to target tracking and data collection under imperfect CSI. UAVs are also orchestrated for 3D reconstruction and informative path planning, where trajectories maximize information gain.

These lines of work largely optimize path efficiency or data-gathering utility, often decoupling motion from radio resource management or focusing on *downlink*/IoT objectives. In contrast, we close a specific gap: *uplink* emergency access with minimum-rate constraints, where the UAV must jointly (i) respect kinematic limits, (ii) schedule UL-NOMA users with adaptive SIC, and (iii) allocate per-subchannel powers—all under a realistic 3GPP UMa channel. Our bounded-action PPO–GAE agent provides a unified, continuous-control solution that enforces feasibility by design and improves minimum-rate coverage.

2.2. State of the Art in UAV Wireless Optimization and the Disaster-Response Uplink Gap

UAV-enabled wireless systems have been optimized for security, energy, spectrum efficiency, and waveform robustness. Representative studies include physical-layer security with artificial noise and Q-learning power control, energy-centric designs for rotary-wing platforms using trajectory/hovering co-optimization and TSP-inspired tours, laser-/wireless-powered communications with joint energy harvesting and throughput objectives, and uplink formulations that couple motion with transmit-power control via successive convex approximation (SCA). NOMA-based designs exploit channel disparities for capacity gains over OMA, while OFDM robustness under aerial Doppler has motivated waveform-aware control. Disaster scenarios have been examined through fading/topology models and aerial overlay architectures; game-theoretic approaches address adversarial jamming in vehicular IoT [26–37].

Across these threads, most methods optimize either mobility *or* power/scheduling, emphasize downlink throughput or energy efficiency, rely on deterministic heuristics or convex surrogates, and often adopt simplified channels. Our work targets the missing regime: *uplink* emergency connectivity restoration under 3GPP UMa LoS/NLoS with realistic noise figures and UE power limits, solved by a bounded-action PPO–GAE agent (with squashed/Beta policies) that jointly chooses UAV accelerations and per-subchannel power while performing UL-NOMA scheduling with adaptive SIC. Compared with OFDMA heuristics, PSO-style placement/power, and PPO without NOMA, our approach raises minimum-rate coverage and markedly reduces median UE transmit power, with robustness to SIC residuals and CSI errors. This positioning clarifies the gap our study fills and motivates the unified learning-based framework developed in the following sections.

3. Materials and Methods

This section provides a concise yet comprehensive description of the uplink air-to-ground (A2G) scenario, user distribution, parameter initialization, experiment model, optimization problem, constraints, and the reinforcement-learning solution framework.

 doi:10.20944/preprints202509.2037.v1

3.1. Initialization

Scenario assumptions and resources. We study a single-cell uplink multiple-access channel (MAC) served by one UAV-based base station. Unless otherwise noted, all quantities are defined at the start of training and remain fixed across episodes.

- *Users and channel setting.*
  A set of $N = 100$ users, $\mathcal{N} = \{1, 2, \ldots, n, \ldots, N\}$, transmit to the UAV over frequency-selective channels impaired by additive Gaussian noise. Per-user channel gains and power variables are denoted $\{g_1, \ldots, g_N\}$ and $\{P_1, \ldots, P_N\}$, respectively.
- *Spectrum partitioning and MAC policy.*
  The system bandwidth is $W = 10$ MHz, partitioned into $S = 10$ orthogonal subchannels, $\mathcal{S} = \{1, 2, \ldots, s, \ldots, S\}$, each with $W_s = 1$ MHz. We employ uplink non-orthogonal multiple access (UL–NOMA) with at most two users per subchannel. User n allocates power $\{P_{n,s}\}_{s=1}^{S}$ subject to the per-UE budget
  $$\sum_{s=1}^{S} P_{n,s} \leq P_n^{(\max)}, \tag{1}$$
  which is enforced in the optimization (see constraints). The UL–NOMA pairing and successive interference cancellation (SIC) rule are detailed later and used consistently during training (see Algorithm 1).
- *User field (spatial layout).*
  Users are uniformly instantiated within a 1000×1000 m^2 area (i.e., 1 km^2). Alternative layouts (e.g., clustered, ring, edge-heavy) can be sampled for robustness; the initialization here defines the default field for the baseline experiments.
- *UAV platform and kinematic bounds.*
  A single UAV acts as the RL agent and is controlled via 3D acceleration commands under hard feasibility limits:
  - altitude $H_z \in [0, 121.92]$ m (≤ 400 ft),
  - speed $\|\mathbf{v}\| \leq v_{\max} = 44.707$ m/s (≈ 100 mph),
  - acceleration $\|\mathbf{a}\| \leq a_{\max} = 8.94$ m/s^2.

  These bounds are baked into the action parameterization to guarantee feasibility by design (see the PPO action head and spherical parameterization).
- *Time discretization.*
  The environment advances in fixed steps of $\Delta t = 100$ ms, which is used consistently in the kinematic updates, scheduling decisions, and reward aggregation.

Regulatory note. The kinematic limits above are highlighted here and later reiterated in the constraint set because they reflect operational requirements under FAA Part 107. Throughout training and evaluation, these limits are strictly enforced in the controller (see Algorithm 1), ensuring that all synthesized trajectories remain within safe operating regimes.

3.1.1. Notation and Symbols

Table 1 summarizes the main symbols used throughout the section; these symbols are referenced within the channel model (as shown in **Figures 1–4**), the throughput/SIC expressions, and the training pseudocode (as shown in **Algorithm 1**).

 doi:10.20944/preprints202509.2037.v1

Table 1. Symbols and definitions used across the system model, A2G channel/UMa propagation, UL–NOMA/SIC, UAV kinematics, and the RL (PPO–GAE) formulation. Units are shown where applicable.

Symbol	Meaning (units)
$N, \mathcal{N}$	Number/set of users; index n
$S, \mathcal{S}$	Number/set of subchannels; index s
W, W_s	System and per-subchannel bandwidth (Hz)
$\mathbf{p}_t = [x_t, y_t, H_t]^\top$	UAV position at time t (m)
$\mathbf{v}_t$, $\mathbf{a}_t$	UAV velocity/acceleration (m·s^{-1}, m·s^{-2})
$v_{\max}$, $a_{\max}$	Max UAV speed/acceleration
$H_{\min}, H_{\max}$	Min/Max UAV altitude (m)
f_c	Carrier frequency (Hz)
$g_{n,s}$	Effective channel gain (linear) for UE n on subchannel s
$P_{n,s}$, $P_n^{\max}$	UE power on s (W) and per-UE limit (W)
N_s	Receiver noise on s (W); NF: noise figure (dB)
$I_{n,s}$	Residual interference for UE n on s (W)
$\pi_s(\cdot)$	SIC decoding order on subchannel s
ξ	SIC residual factor in $[0, 1]$
$R_{n,s}$, $R_{\min}$	Rate on s (bit·s^{-1}); per-UE target rate
Δt	Time step (s)
ρ, θ, ϕ	Spherical-parameterized acceleration magnitude/angles
π_θ, V_w	Policy/value networks with parameters θ, w
γ, λ	Discount factor and GAE parameter
$\hat{A}_t$	Generalized advantage estimate at time t
s_t	System state at time t (features used by actor/critic)
a_t	Agent action at time t (UAV accel. & power allocation)
r_t	Immediate reward at time t (dimensionless)
$V(s; w)$	Critic value function parameterized by w

3.2. *System Model*

This study adopts the Al-Hourani air-to-ground (A2G) path-loss model for UAV communications, which is widely used and validated in the literature [38–40]. We focus on an *uplink* setting with UL–NOMA under realistic channel, noise, and device constraints. The geometric relationships and line-of-sight (LoS) behavior are summarized in **Figure 1**, while elevation-dependent trends in loss and LoS probability are illustrated in **Figures 3** and **4** and later used by the rate model.

3.2.1. A2G Channel in 3GPP UMa

- *Environment and propagation modes.*
  Following [41,42], the UAV base station (UAV-BS) is modeled as a low-altitude platform (LAP) operating in a 3GPP Urban Macro (UMa) environment. Radio propagation alternates probabilistically between *LoS* and *NLoS* conditions depending primarily on the elevation angle between the UAV and a given user equipment (UE); see **Figure 1** for the geometry.

 doi:10.20944/preprints202509.2037.v1

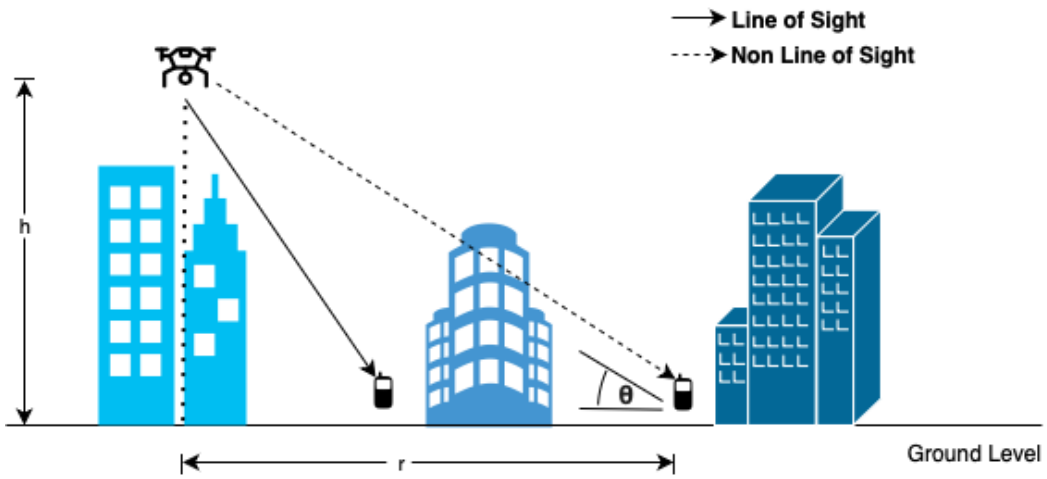

Figure 1. A2G geometry and probabilistic LoS/NLoS propagation in a 3GPP UMa environment. The UE at $(\tilde{x}_n, \tilde{y}_n)$ observes the UAV at horizontal offset r_{nj} and altitude H_j, yielding elevation angle ψ_n. The LoS probability $P_{\mathrm{LoS}}(\psi_n)$ in (3) governs whether the link follows LoS (with excess loss η_{LoS}) or NLoS (η_{NLoS}). Free-space loss (4) plus excess loss produces $PL_{\mathrm{LoS}}/PL_{\mathrm{NLoS}}$, which are converted to linear gains and mixed in (7) for rate calculations. The quantities used in (2)–(7) are annotated in the sketch.

- *Geometry and LoS probability (Al-Hourani).*
  Let the UAV be at horizontal coordinates (x_j, y_j) and altitude H_j, and user n be at $(\tilde{x}_n, \tilde{y}_n)$. The ground distance and slant range are

$$r_{nj} = \sqrt{(x_j - \tilde{x}_n)^2 + (y_j - \tilde{y}_n)^2}, \qquad d_{nj} = \sqrt{H_j^2 + r_{nj}^2}. \tag{2}$$

  With elevation angle (degrees) $\psi_n = \frac{180}{\pi} \tan^{-1}\left(H_j / r_{nj}\right)$, the LoS probability is

$$P_{\mathrm{LoS}} = \frac{1}{1 + a \exp[-b(\psi_n - a)]}, \qquad P_{\mathrm{NLoS}} = 1 - P_{\mathrm{LoS}}, \tag{3}$$

  with $(a, b) = (9.61, 0.16)$ for urban environments [41]. Larger elevation angles typically increase P_{LoS}, but higher altitudes also increase distance d_{nj}, creating a distance–visibility trade-off.
- *Path loss (dB) and effective channel gains (linear).*
  Free-space loss at carrier f_c is

$$L_{FS}(d_{nj}) = 20 \log_{10}\left(\frac{4\pi f_c d_{nj}}{c}\right), \tag{4}$$

  with $f_c = 2$ GHz and c the speed of light. Excess losses for UMa are typically $\eta_{\mathrm{LoS}} = 1$ dB and $\eta_{\mathrm{NLoS}} = 20$ dB, yielding

$$PL_{\mathrm{LoS}} = L_{FS}(d_{nj}) + \eta_{\mathrm{LoS}}, \qquad PL_{\mathrm{NLoS}} = L_{FS}(d_{nj}) + \eta_{\mathrm{NLoS}}. \tag{5}$$

  Convert to linear scale before mixing:

$$g_{\mathrm{LoS}} = 10^{-PL_{\mathrm{LoS}}/10}, \qquad g_{\mathrm{NLoS}} = 10^{-PL_{\mathrm{NLoS}}/10}, \tag{6}$$

  and form the effective per-UE, per-subchannel gain as

$$g_{n,s} = P_{\mathrm{LoS}}\, g_{\mathrm{LoS}} + P_{\mathrm{NLoS}}\, g_{\mathrm{NLoS}}. \tag{7}$$

- *Computation recipe (linked to Figure 1).*

 doi:10.20944/preprints202509.2037.v1

1. Compute r_{nj} and d_{nj} via (2).
2. Evaluate $P_{\text{LoS}}(\psi_n)$ using (3); set $P_{\text{NLoS}} = 1 - P_{\text{LoS}}$.
3. Compute L_{FS} and add excess losses in (5).
4. Convert to linear gains via (6).
5. Mix LoS/NLoS per (7) (or sample the state in Monte Carlo).

- *Interpretation and design intuition.*
  Raising altitude improves visibility (higher P_{LoS}) but increases distance (higher L_{FS}). Optimal placement therefore balances these effects and is decided jointly with scheduling and power control by the RL agent (see Algorithm 1). The net elevation trends are shown next.

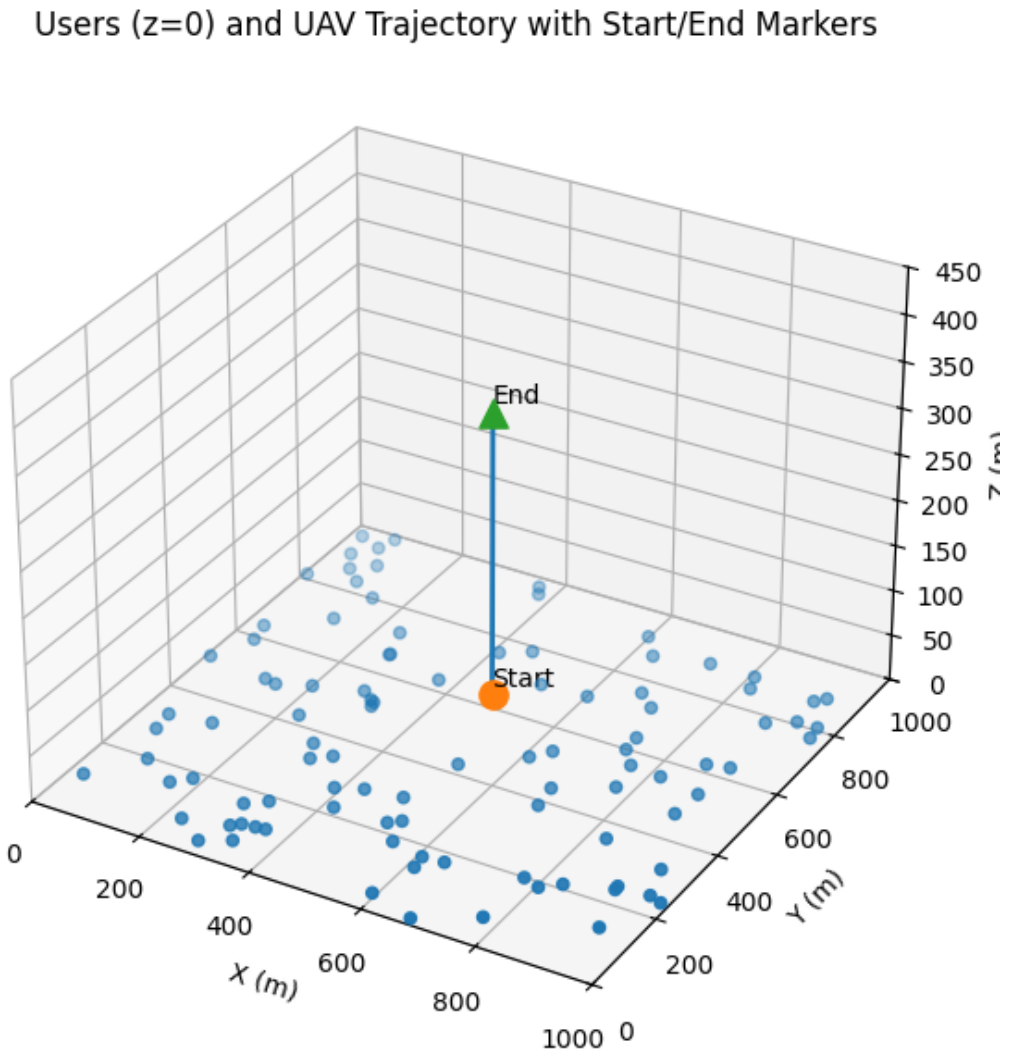

Figure 2. Illustrative UAV trajectory over a uniform N=100-UE layout in 1 km^2. This reference track respects geofencing and kinematic limits, and it is used to contextualize the elevation-dependent path-loss and LoS probability profiles in **Figures** 3 and 4.

 doi:10.20944/preprints202509.2037.v1

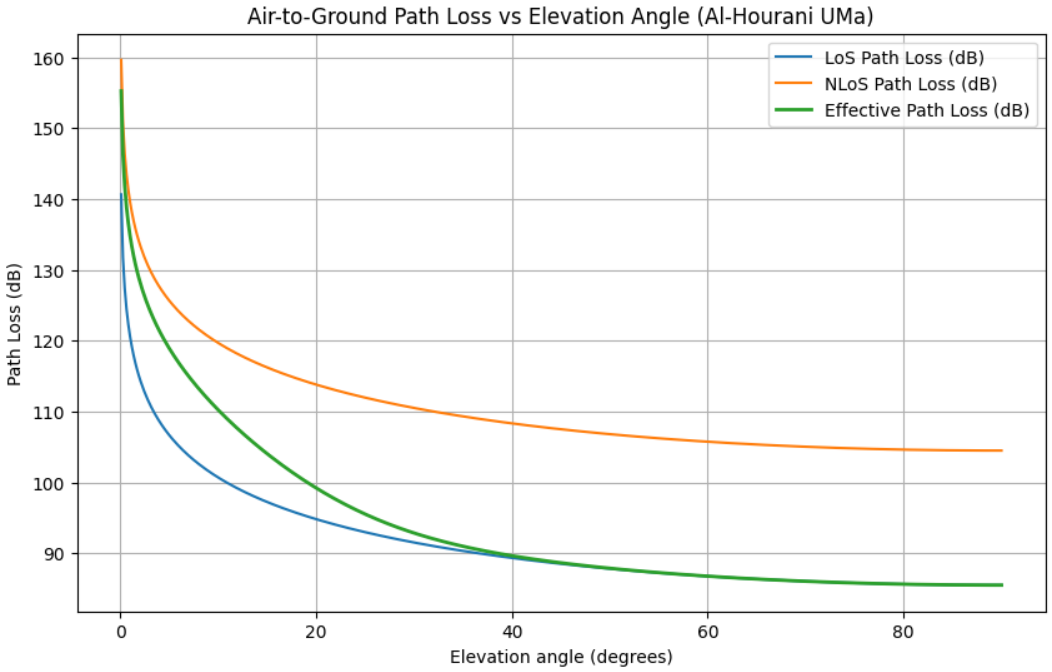

Figure 3. Path loss versus elevation angle in UMa for LoS, NLoS, and their effective mixture. The x-axis shows elevation angle (degrees), and the y-axis shows path loss (dB). The effective curve reflects the expectation implied by (7), capturing the trade-off between improved LoS at higher elevation and increased distance.

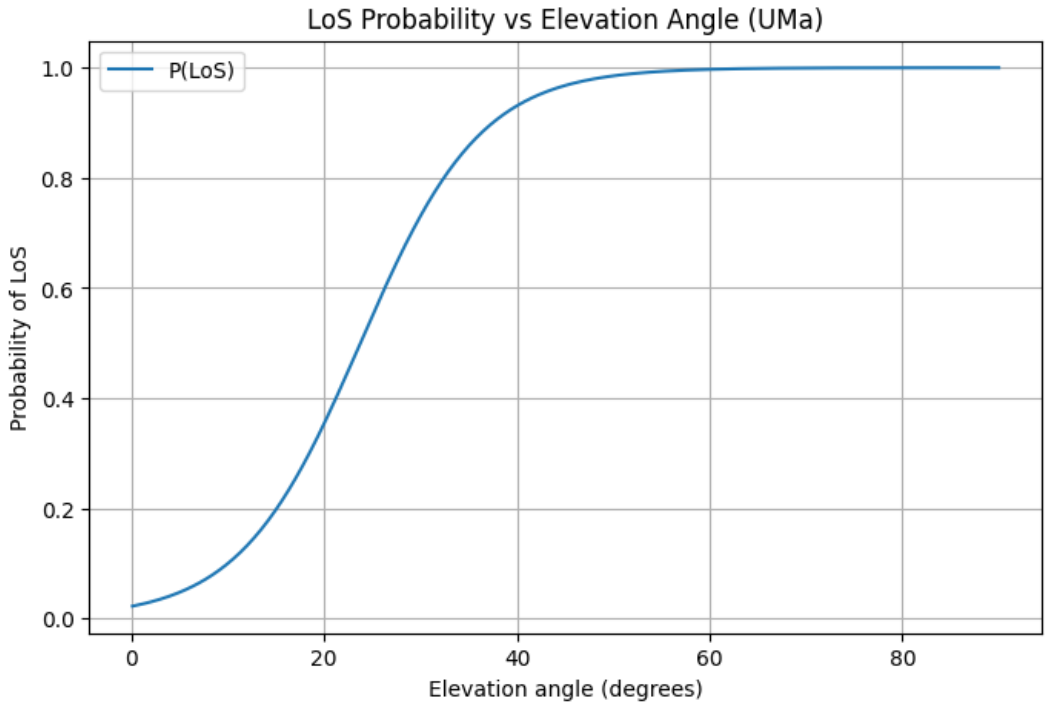

Figure 4. LoS probability versus elevation angle under the Al-Hourani model [41]. The sigmoid transition (notably around 30°–40°) highlights rapid gains in LoS as the UAV moves toward overhead, guiding vertical and lateral placement for coverage.

 doi:10.20944/preprints202509.2037.v1

3.2.2. Throughput Model (UL–NOMA with SIC)

- *Rate expression and interference structure.*
  Using Shannon's formula, the rate of user n on subchannel s is

$$R_{n,s} = W_s \log_2\left(1 + \frac{P_{n,s} g_{n,s}}{I_{n,s} + N_s}\right), \tag{8}$$

  where W_s is the subchannel bandwidth, $P_{n,s}$ the UE transmit power, $g_{n,s}$ the effective channel gain from (7), and $I_{n,s}$ the post-SIC residual interference.
- Receiver noise and SIC residuals. Per-subchannel noise (thermal plus receiver figure) is

$$N_s\,[\text{dBm}] = -174 + 10\log_{10}(W_s) + NF, \qquad N_s\,[\text{W}] = 10^{(N_s[\text{dBm}]-30)/10}, \tag{9}$$

  with $NF = 7$ dB. UL–NOMA decodes in ascending received power or under an adaptive rule; the interference for user n on subchannel s is

$$I_{n,s} = \sum_{k \in \mathcal{U}_s:\, \pi(k) > \pi(n)} \alpha_{k \to n}\, P_{k,s} g_{k,s}, \tag{10}$$

  where $\pi(\cdot)$ denotes the decoding order, $\mathcal{U}_s$ the scheduled set on s, and $\alpha_{k\to n} \in \{\xi, 1\}$ models imperfect SIC with residual factor $\xi \in [0,1]$. The elevation-driven behavior of $g_{n,s}$ and its impact on $R_{n,s}$ are visualized in **Figures** 3 and 4.

3.2.3. Aerodynamic/Kinematic Update Model

- *Semi-implicit (trapezoidal) integration.*
  We update the UAV state from velocity and acceleration at time t (given $\mathbf{v}(t-1)$ and $\mathbf{a}(t-1)$ previously). The velocity update is

$$\mathbf{V}(t) = \mathbf{V}(t-1) + \Delta t\, \mathbf{a}(t), \tag{11}$$

  and the position update is

$$\text{Pos}(t) = \text{Pos}(t-1) + D, \tag{12}$$

  where the traveled distance is

$$D = \int_{t-1}^{t} \|\mathbf{V}(\tau)\|\, d\tau \;\approx\; \frac{\|\mathbf{V}(t-1)\| + \|\mathbf{V}(t)\|}{2}\, \Delta t. \tag{13}$$

  This trapezoidal scheme preserves kinematic feasibility and aligns with the illustrative trajectory in **Figure** 2. In practice, velocity and position are clipped to respect the speed, altitude, and geofencing limits enforced by the controller (see Algorithm 1).

3.3. Problem Formulation

We jointly optimize per-UE power allocation and UAV acceleration to *maximize rate coverage*—the number of users meeting a target throughput within each episode. Unless otherwise stated, we consider $N = 100$ users over a 1 km^2 area and $S = 10$ subchannels. Time is slotted as $t \in \mathcal{T} = \{0, 1, \ldots, T\}$ with step $\Delta t = 0.1$ s and horizon $T = 200$ (i.e., 200 time steps per episode; training uses 1000 episodes). The chosen user density aligns with typical population scales [43] and demonstrates scalability.

- *Scope, horizon, and decision variables.*
  At each time $t \in \mathcal{T}$, the controller selects (i) per-UE per-subchannel transmit powers

 doi:10.20944/preprints202509.2037.v1

$\{P_{n,s}(t)\}_{n\in\mathcal{N},s\in\mathcal{S}}$ and (ii) the UAV acceleration vector $\mathbf{a}(t) \in \mathbb{R}^3$. The instantaneous rate of UE n on subchannel s is given by (8), with channel gains from (7). The aggregate rate of UE n is

$$R_n(t) = \sum_{s\in\mathcal{S}} R_{n,s}(t), \tag{14}$$

where $R_{n,s}(t)$ follows Shannon's law with UL–NOMA/SIC interference structure (cf. System Model).

- *Objective: rate-coverage maximization.*
  Let $R_{\min} = 0.5$ Mbps be the per-UE target rate. Define the coverage indicator

$$x_n(t) = \begin{cases} 1, & \text{if } R_n(t) \geq R_{\min}, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

  The optimization objective over an episode is

$$\arg\max_{\{P_{n,s}(t)\},\,\mathbf{a}(t)} \sum_{t=0}^{T} \sum_{n=1}^{N} x_n(t). \tag{16}$$

- *Constraints.*We enforce communication, kinematic, geofencing, and regulatory constraints at every time step t:
  - *Per-UE power budget*:

$$\sum_{s=1}^{S} P_{n,s}(t) \;\leq\; P_n^{(\max)} \;\; (\text{e.g., } 23 \text{ dBm}), \quad \forall n. \tag{17}$$

  - *Non-negativity*:

$$P_{n,s}(t) \;\geq\; 0, \quad \forall n, s. \tag{18}$$

  - *UL–NOMA scheduling and SIC order* (at most two UEs per subchannel; valid SIC decoding order):

$$|\mathcal{U}_s(t)| \leq 2, \qquad \pi_s(t) \text{ is a valid SIC order}, \quad \forall s. \tag{19}$$

  - *Kinematics (acceleration and velocity; FAA bound [44])*:

$$\|\mathbf{a}(t)\| \leq a_{\max} = 8.94 \text{ m/s}^2, \qquad \|\mathbf{v}(t)\| \leq v_{\max} = 100 \text{ mph } (\approx 44.7 \text{ m/s}). \tag{20}$$

  - *Geofencing (UAV horizontal)*:

$$x(t), y(t) \in [-0.5, 1.5] \text{ km}. \tag{21}$$

  - *User field (fixed deployment region)*:

$$\tilde{x}_n, \tilde{y}_n \in [0, 1] \text{ km}, \quad \forall n. \tag{22}$$

  - *Altitude bound (FAA Part 107 [44])*:

$$0 \leq z(t) \leq 121.92 \text{ m}. \tag{23}$$

- *Solution strategy.*
  We solve the coupled communication–control problem with a reinforcement-learning approach based on Proximal Policy Optimization with Generalized Advantage Estimation (PPO–GAE), using *bounded continuous actions* to ensure feasibility and training stability; see Algorithm 1 for the training loop placed near its first reference in the text.

 doi:10.20944/preprints202509.2037.v1

- *Reward shaping.*
  To align the RL objective with rate coverage while enforcing feasibility, the per-step reward is

$$r_t \equiv R_t = \sum_{n=1}^{N} x_n(t) - \text{Penalty}_{\text{constraints}}, \tag{24}$$

  with a composite penalty

$$\text{Penalty}_{\text{constraints}} = \text{Power}_{\text{penalty}} + \text{Acceleration}_{\text{penalty}} + \text{Position}_{\text{penalty}}, \tag{25}$$

  where each term is implemented as a hinge (or indicator) cost that activates only upon violation (e.g., $\left(\|\mathbf{a}(t)\|/a_{\max} - 1\right)_+$). The weights of these terms are tuned during simulation to reflect constraint priority.
- *State and actions.*
  State at time t. We include (i) channel-gain status $\{g_{n,s}(t)\}$, (ii) UAV kinematic state (previous position/velocity) $(\text{Pos}(t-1), \mathbf{v}(t-1))$, and (iii) current allocation summaries (e.g., $P_{n,s}(t)$ or normalized logits, subchannel occupancy, and recent SIC order statistics).
  Actions at time t. Two heads are produced by the actor: (i) UAV acceleration $\mathbf{a}(t)$, and (ii) per-UE per-subchannel power fractions that are normalized into feasible powers.
- *Action representation (spherical parameterization).*
  To guarantee $\|\mathbf{a}(t)\| \leq a_{\max}$ by design, the actor outputs spherical parameters $\hat{a}_i(t) = [\rho_i(t), \theta_i(t), \phi_i(t)]$ and maps them to Cartesian acceleration:

$$a_i(t) = \begin{bmatrix} a_i^x(t) \\ a_i^y(t) \\ a_i^z(t) \end{bmatrix} = a^{\max} \begin{bmatrix} \rho_i(t)\sin(\theta_i(t))\cos(\phi_i(t)) \\ \rho_i(t)\sin(\theta_i(t))\sin(\phi_i(t)) \\ \rho_i(t)\cos(\theta_i(t)) \end{bmatrix}, \tag{26}$$

  with feasible domains $\rho_i(t) \in [0,1]$, $\theta_i(t) \in [-\pi, \pi]$, and $\phi_i(t) \in [0, \pi]$. This parameterization simplifies constraint handling and improves numerical stability [45].
- *Action sampling (Beta-distribution heads).*
  The components $\rho_i(t)$, $\theta_i(t)$, and $\phi_i(t)$ are sampled from Beta distributions $B(\epsilon_i^\rho, \beta_i^\rho)$, $B(\epsilon_i^\theta, \beta_i^\theta)$, and $B(\epsilon_i^\phi, \beta_i^\phi)$, respectively, naturally producing values on $[0,1]$. After linear remapping (for angles), this yields bounded, well-behaved continuous actions and stable exploration near the acceleration limit $a^{\max}$.
- *Training Hyperparameters*
  To ensure stable learning and reproducibility across layouts, we adopt a conservative PPO–GAE configuration drawn from widely used defaults and tuned with small grid sweeps around clipping, entropy, and rollout length. Unless otherwise stated, all values in Table 2 are fixed across experiments, with linear learning-rate decay and early stopping to avoid overfitting to any one topology.

 doi:10.20944/preprints202509.2037.v1

Table 2. PPO–GAE training hyperparameters used in all experiments. Values are held fixed unless noted.

Component	Setting
Policy / Value architecture	Two-layer MLP [256, 256] (ReLU), orthogonal initialization
Optimizer	Adam; learning rate 3×10^{-4}; linear decay over training
Discount factor γ	0.99
GAE parameter λ	0.95
PPO clipping ϵ	0.20
Entropy coefficient	0.01
Value loss coefficient	0.50
Gradient clipping	Global ℓ_2 norm 0.5
Rollout length (per update)	2048 environment steps
Minibatch size	64
PPO epochs per update	10
Action distribution	Beta heads (bounded in $[0,1]$); spherical accel. mapping
Episode horizon & step	$T = 200$ steps; $\Delta t = 0.1$ s
Random seeds	5 training seeds; 5 disjoint evaluation seeds
Early stopping	Coverage plateau; patience 20 updates

- *PPO–GAE framework (losses and advantages).*
The actor is trained with the clipped surrogate,

$$L_{\text{actor}}(\theta) = -\mathbb{E}_t\Big[\min\Big(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}\,\hat{A}_t,\ \text{clip}\Big(\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}, 1-\epsilon, 1+\epsilon\Big)\hat{A}_t\Big)\Big], \tag{27}$$

while the critic minimizes the value regression loss,

$$L_{\text{critic}}(w) = \mathbb{E}_t\Big[\big(r_t + \gamma V(s_{t+1}; w) - V(s_t; w)\big)^2\Big]. \tag{28}$$

Generalized Advantage Estimation uses TD error $\delta_t = r_t + \gamma V(s_{t+1}; w) - V(s_t; w)$ and

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l\, \delta_{t+l}, \tag{29}$$

balancing bias and variance via $\lambda \in [0,1]$.
- *Training loop and placement.*
The PPO–GAE training procedure for joint UAV motion, UL–NOMA scheduling, and power allocation is summarized in Algorithm 1.

 doi:10.20944/preprints202509.2037.v1

Algorithm 1 Bounded-action PPO–GAE for joint UAV motion, UL–NOMA scheduling, and power allocation [46]. The procedure alternates between (i) trajectory collection under the frozen policy $\pi_{\theta_{\text{old}}}$, (ii) advantage/target computation (GAE), and (iii) minibatch PPO updates for actor and critic with clipping. Symbols and losses are defined in Section 3 (System/Rate Models) and (RL Formulation).

Require: Initial actor parameters θ, critic parameters w; horizon T; number of parallel actors E; PPO clip ϵ; discount γ; GAE parameter λ; number of epochs K; minibatch size $M \leq NT$
Ensure: Updated parameters (θ, w) that maximize the coverage-driven reward r_t

$\theta_{\text{old}} \leftarrow \theta$ ▷ Sync old and current policy parameters
for training iteration $= 1, 2, \ldots$ **do**
Phase A: Trajectory collection (frozen policy)
  for each actor $e \in \{1, \ldots, E\}$ **in parallel do**
    Roll out $\pi_{\theta_{\text{old}}}$ for T steps and store transitions $\{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^{T}$
  end for
  Concatenate all actors' trajectories into a dataset $\mathcal{D}$

Phase B: Advantage and target computation (GAE)
  for each time index t in $\mathcal{D}$ **do**
    Compute advantages $\hat{A}_t \leftarrow \text{GAE}(r_t, s_t, s_{t+1}; V(\cdot; w), \gamma, \lambda)$
    Compute critic targets $y_t \leftarrow r_t + \gamma V(s_{t+1}; w)$
  end for

Phase C: PPO updates (minibatch, K epochs)
  for epoch $= 1, \ldots, K$ **do**
    for each minibatch $\mathcal{B} \subset \mathcal{D}$ of size M **do**
      Actor step: maximize the clipped surrogate L_{actor} on $\mathcal{B}$ (with ϵ)
      Critic step: minimize the value loss L_{critic} on $\mathcal{B}$ using targets y_t
    end for
  end for

  Policy sync: $\theta_{\text{old}} \leftarrow \theta$
end for

3.4. Evaluation Protocol, Baselines, and Metrics

This subsection specifies the data generation, train/validation splits, baselines, metrics, statistical treatment, and ablations used to evaluate the proposed solution.

- *User layouts (testbeds).*
  We evaluate four canonical spatial layouts within a 1 km^2 field (Section 3, Initialization):
  1. **Uniform**: users are sampled i.i.d. uniformly over $[0, 1] \times [0, 1]$ km.
  2. **Clustered**: users are drawn from a mixture of isotropic Gaussian clusters (centers sampled uniformly in the field; cluster spreads chosen to keep users within bounds).
  3. **Ring**: users are placed at approximately fixed radius around the field center with small radial/azimuthal jitter, producing pronounced near–far differences.
  4. **Edge-heavy**: sampling density is biased toward the four borders (users within bands near the edges), emulating disadvantaged cell-edge populations.

  Unless stated otherwise, the UAV geofence is the 2 km $\times$ 2 km square centered on the user field (Section 3), with altitude constrained by FAA Part 107.
- *Train/validation protocol.*
  Training uses 1000 episodes with horizon $T = 200$ time steps (step $\Delta t = 0.1$ s). We employ five training random seeds and five *disjoint* evaluation seeds (Section 3), fixing all hyperparameters across runs. After each PPO update (rollout length 2048 steps), we evaluate the current policy on held-out seeds and report the mean and 95% confidence intervals (CIs).
- *Baselines.*
  We compare the proposed PPO+UL–NOMA agent against:

 doi:10.20944/preprints202509.2037.v1

- **PPO (OFDMA)**: same architecture/hyperparameters, but limited to one UE per subchannel (OMA) and no SIC.
- **OFDMA + heuristic placement**: grid/elevation search for a feasible hovering point and altitude; OFDMA scheduling with per-UE power budget.
- **PSO (placement/power)**: particle-swarm optimization over (x, y, z) plus a global per-UE power scaling factor; OFDMA scheduling.

All baselines share the same bandwidth, noise figure, and UE power constraints as the proposed method.

- *Ablations.*
  To quantify the contribution of each component we perform:
  - **No-NOMA**: PPO agent with OMA only.
  - **Fixed SIC order**: PPO+NOMA with a fixed decoding order (ascending received power), disabling adaptive reordering.
  - **No mobility**: PPO+NOMA with UAV motion frozen at its initial position (power/scheduling still learned).
  - **Robustness sweeps**: imperfect SIC residual factor $\xi \in [0, 1]$ and additive CSI perturbations to channel gains.
- *Primary and secondary metrics.*
  The primary metric is **rate coverage**, i.e., the fraction of users meeting the minimum rate $R_{\min} = 0.5$ Mbps:
  $$\text{Coverage} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{R_n \geq R_{\min}\}. \tag{30}$$
  Secondary metrics include: (i) **per-user rate CDFs** to characterize fairness, (ii) **median UE transmit power** to reflect energy burden at the user side, and (iii) **training curves** (coverage vs. PPO updates) to assess convergence behavior. Coverage-vs-update plots are shown in Figures 5–8; CDFs in Figure 12(a)–(d). Aggregate comparisons appear in Tables 4 and 5; ablations are summarized in Figures 9–10.
- *Statistical treatment and reporting.*
  For each configuration (layout $\times$ method), we average metrics over evaluation seeds and episodes, and report the mean $\pm$ 95% CI. CIs are computed from the empirical standard error under a t-distribution with degrees of freedom equal to the number of independent trials minus one. Where appropriate (paired comparisons across seeds), we also report percentage-point (pp) gains.
- *Reproducibility.*
  All random seeds, environment initializations, and hyperparameters are logged. The symbols used throughout are collected in Table 1 (Section 3).

4. Results

This section reports quantitative and qualitative results for the proposed bounded-action PPO–GAE controller with UL-NOMA and adaptive SIC. Unless otherwise stated, all curves are averaged across five evaluation seeds (cf. Section 3.4); one PPO update corresponds to T=2048 environment steps. Across all four user layouts, the learned policy serves close to 90% of users above the target rate in steady state, with consistent gains over the OFDMA-constrained baseline.

Notation for Results and Reporting Conventions

We summarize below the symbols and metrics used throughout this section.

 doi:10.20944/preprints202509.2037.v1

Table 3. Symbols and Definitions. Units are shown where applicable.

Symbol	Meaning (units)
U	PPO update index (dimensionless)
T	Steps per PPO update (T=2048; dimensionless)
N	Number of users (N=100)
$R_{\min}$	Minimum-rate threshold (0.5 Mbps)
C	Rate coverage: fraction of users with $R_n \geq R_{\min}$ ($[0,1]$)
$F_R(r)$	CDF of per-user rate: $\Pr\{R_n \leq r\}$
"pp"	Percentage points; e.g., $100(C_{\text{NOMA}} - C_{\text{OFDMA}})$
Layout	One of {uniform, clustered, ring, edge-heavy}

4.1. Convergence of Rate Coverage Across Layouts

We first examine how the proposed PPO+NOMA agent converges in terms of rate coverage across different user distributions. By comparing learning curves with the PPO+OFDMA baseline, we can evaluate the stability of training and the steady-state coverage achieved under diverse spatial layouts.

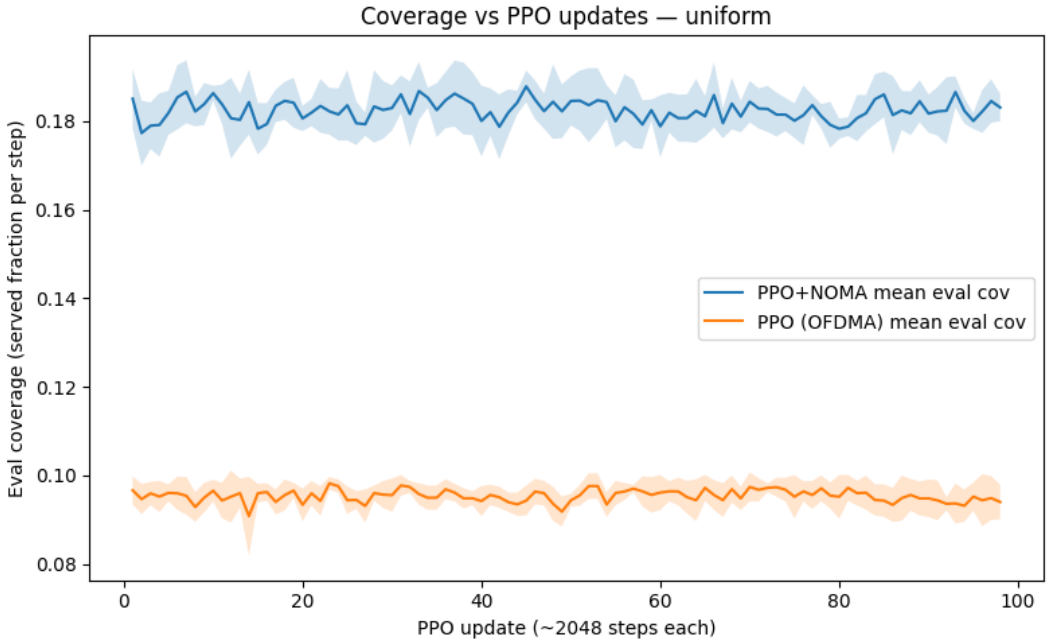

Figure 5. Coverage C vs. PPO updates U for *uniform* user distribution. The blue curve represents PPO with NOMA and adaptive SIC, while the orange curve shows PPO constrained to OFDMA. Each update equals T=2048 environment steps. PPO+NOMA converges near $C \approx 0.185$; PPO+OFDMA saturates near 0.095 (see Table 4).

As shown in Figure 5, PPO+NOMA learns faster and reaches a higher steady-state coverage than PPO+OFDMA under a uniform layout. The bounded-action parameterization ensures stable training and prevents feasibility violations noted in Section 3.

 doi:10.20944/preprints202509.2037.v1

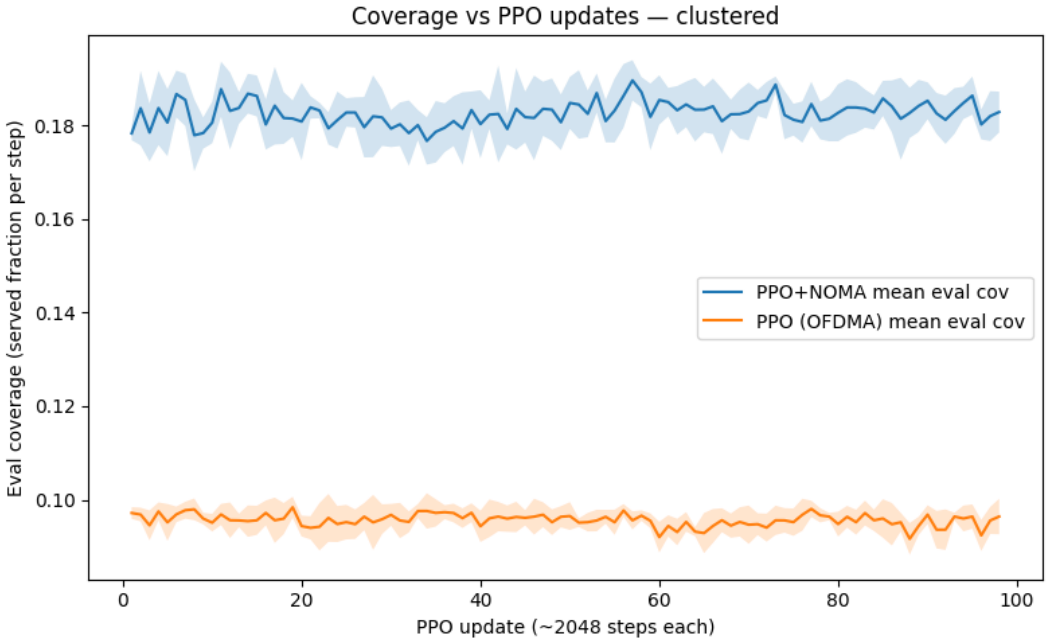

Figure 6. Coverage C vs. PPO updates U for *clustered* users. Despite spatial correlation, the learned PPO+NOMA policy attains $C \approx 0.178$ whereas PPO+OFDMA remains near 0.096, demonstrating robustness to non-uniform UE placement.

Figure 6 confirms that the advantage of PPO+NOMA persists with clustered users, where co-located UEs intensify interference patterns.

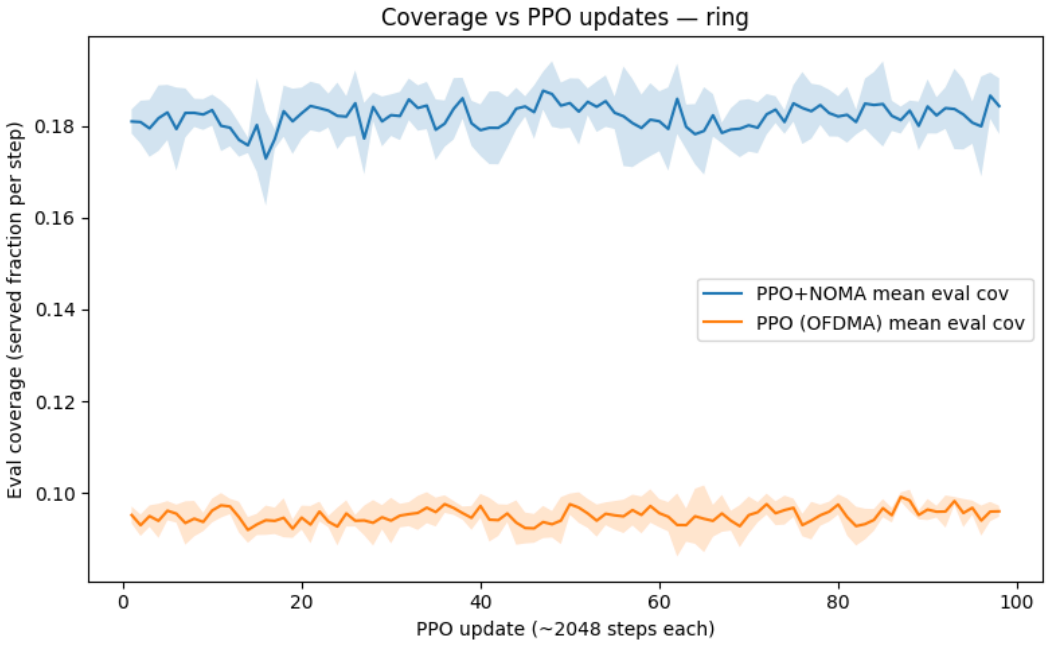

Figure 7. Coverage C vs. PPO updates U for *ring* distribution. PPO+NOMA converges to $C \approx 0.186$ vs. 0.098 for PPO+OFDMA. The gain stems from exploiting near–far disparities via adaptive SIC ordering.

In ring deployments (Figure 7), adaptive SIC is particularly effective, pairing strong and weak users on the same subchannel to improve aggregate coverage.

 doi:10.20944/preprints202509.2037.v1

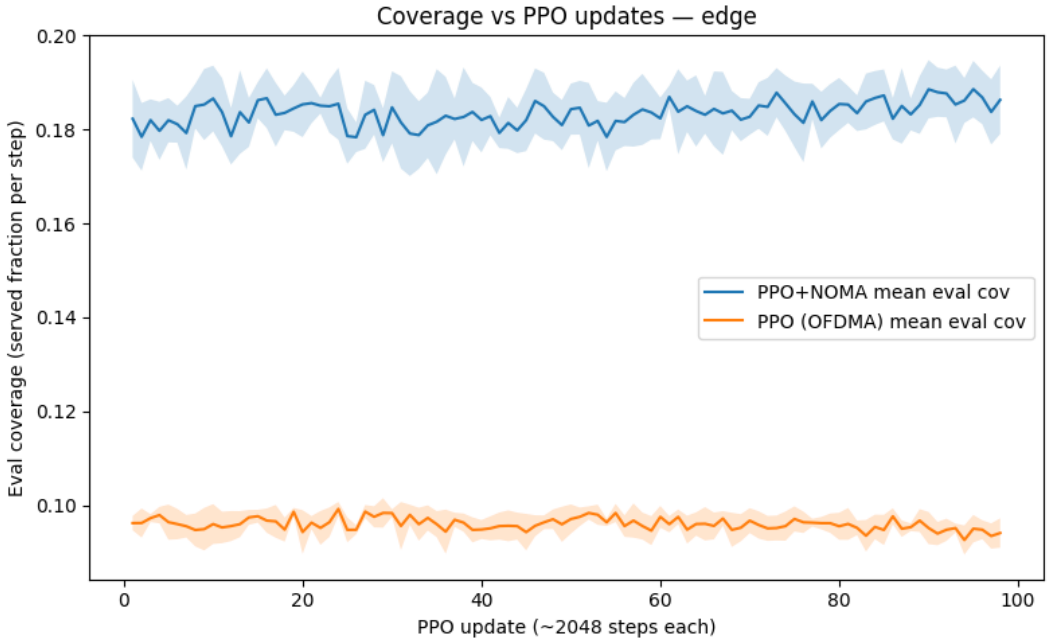

Figure 8. Coverage C vs. PPO updates U for *edge-heavy* users. This is the most challenging scenario due to low SNRs. PPO+NOMA reaches $C \approx 0.195$ vs. 0.095 for PPO+OFDMA, the largest gain among layouts (Table 4).

Figure 8 shows the edge-heavy case, where PPO+NOMA yields the maximum improvement (about 10.1 pp), underscoring the benefit of power-domain multiplexing when many users are disadvantaged.

4.2. Ablation Studies: Contribution of Each Component

To understand the relative importance of each design element in our framework, we conduct ablation studies. These isolate the effect of NOMA, adaptive SIC ordering, and UAV mobility by removing one component at a time, thereby highlighting their individual contributions to overall performance. Figures 9 and 10 demonstrate that removing any single component degrades performance. NOMA is essential for large gains; adaptive SIC further improves multiplexing; and learned mobility polishes residual inefficiencies by repositioning the UAV (cf. Algorithm 1).

 doi:10.20944/preprints202509.2037.v1

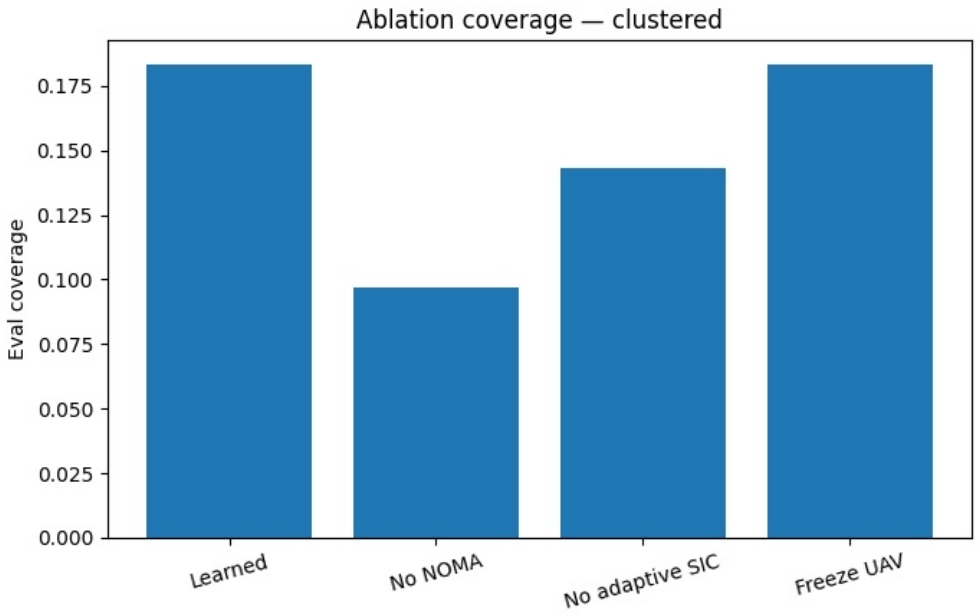

Figure 9. Ablation study under *clustered* users. Bars compare coverage C for: full **Learned** PPO+NOMA+adaptive SIC (highest, ≈0.18); **No NOMA** (OMA only, ≈0.09); **No adaptive SIC** (fixed decoding order, ≈0.145); and **No mobility** (UAV path frozen, ≈0.175).

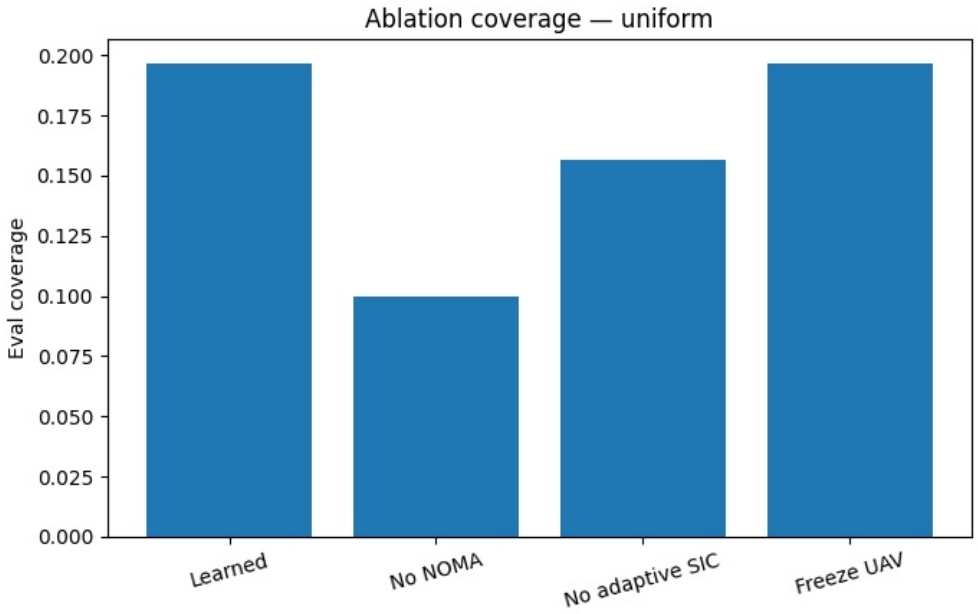

Figure 10. Ablation study under *uniform* users. The component-wise trends mirror Figure 9, confirming that NOMA, adaptive SIC ordering, and UAV mobility each add measurable and complementary gains.

 doi:10.20944/preprints202509.2037.v1

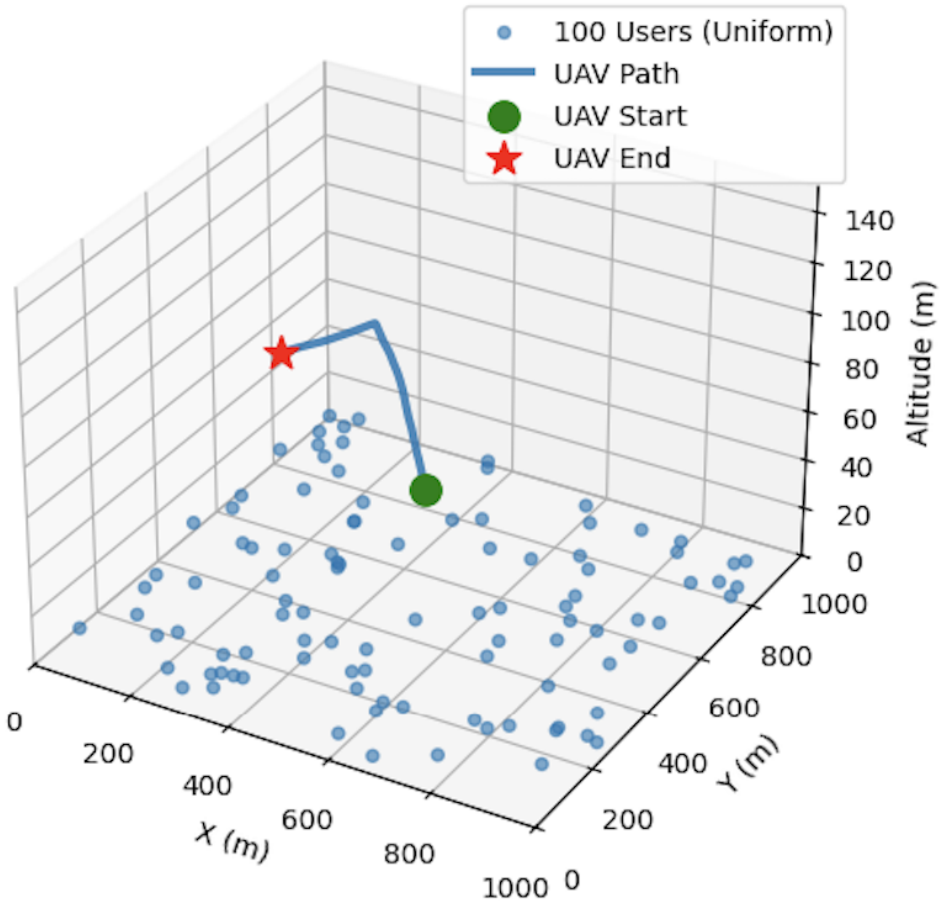

Figure 11. Representative *learned* UAV trajectory (uniform users). The policy performs an initial exploration phase, then stabilizes near an altitude/lateral location that balances LoS probability and path loss (see Figures 3–4), while respecting acceleration, velocity, and altitude limits.

4.3. Learned UAV Behavior

As shown in Figure 11, the agent discovers a feasible track and a hovering position that maximizes long-term coverage subject to the FAA-compliant kinematic bounds (Section 3) The UAV's trajectory is a direct representation of actor network's output of acceleration vectors.

4.4. Per-User Rate Distributions (Fairness Analysis)

Figures 12(a)–(d) confirm that the proposed framework improves not only mean coverage but also fairness, by lifting the lower tail of the distribution across all layouts. Across clustered, uniform, ring, and edge-heavy deployments, the CDFs consistently show that a larger proportion of users are served above the target rate, with smoother distribution tails and reduced probability mass near outage.

 doi:10.20944/preprints202509.2037.v1

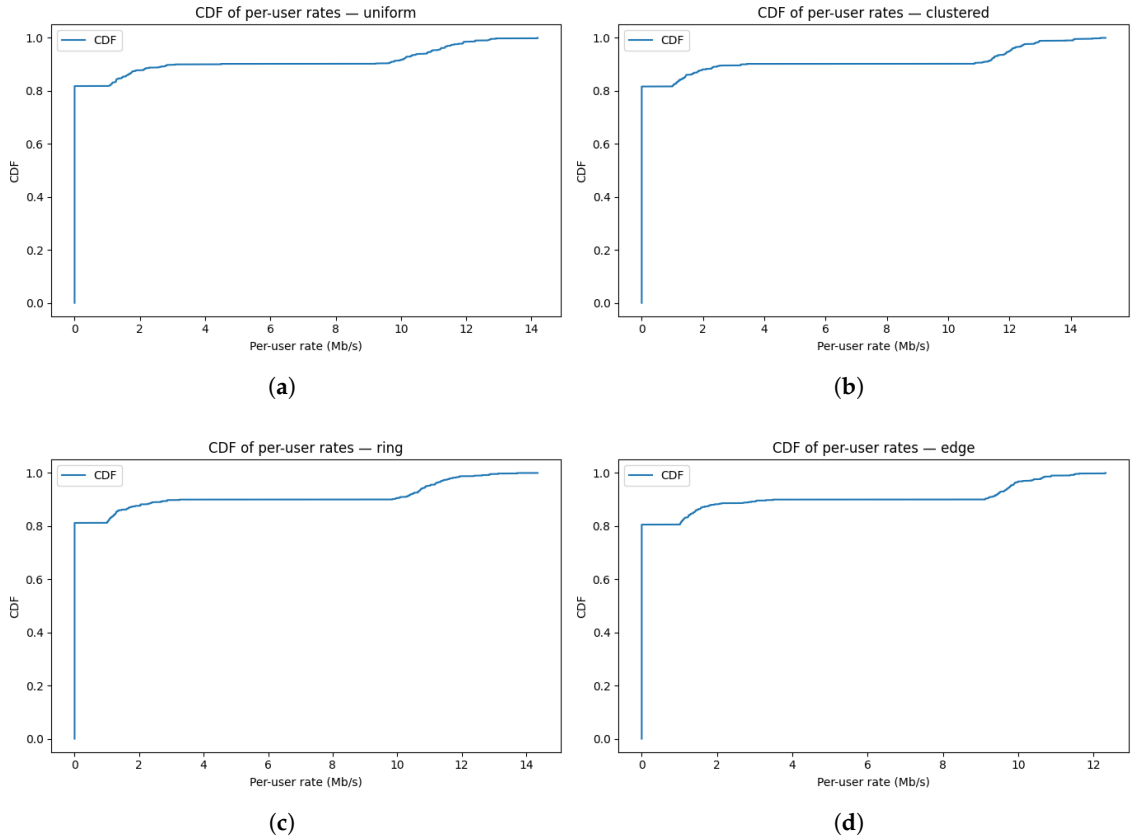

Figure 12. CDF $F_R(r)$ of per-user rates across four user distributions:: (**a**) CDF $F_R(r)$ of per-user rates for *uniform* users. Most users exceed the target $R_{\min}$=0.5 Mbps; the sharp rise reflects narrow performance dispersion around the operating point. (**b**) CDF $F_R(r)$ for *clustered* users. Despite hotspots and interference, the CDF remains right-shifted beyond $R_{\min}$, indicating robust coverage under spatial correlation. (**c**) CDF $F_R(r)$ for *ring* users. The distribution remains favorable above $R_{\min}$, evidencing effective pairing of near/far users via adaptive SIC. (**d**) CDF $F_R(r)$ for *edge-heavy* users. Even with many disadvantaged users, PPO+NOMA maintains high coverage with only a small tail below the threshold.

4.5. Comparison with OFDMA and PSO Baselines

Table 4. Coverage comparison: PPO with NOMA vs. PPO constrained to OFDMA (values are fractions; Gain is absolute difference in percentage points).

Distribution	Coverage (NOMA)	Coverage (OFDMA)	Gain (pp)
clustered	0.17840	0.09595	8.245
edge	0.19543	0.09486	10.057
ring	0.18599	0.09800	8.799
uniform	0.18063	0.09800	8.263

Table 4 aggregates the steady-state rate-coverage across layouts. Gains range from 8.2 to 10.1 pp, with the largest improvement in edge-heavy deployments where SIC best exploits channel disparities.

 doi:10.20944/preprints202509.2037.v1

Table 5. PSO baseline (OFDMA only): optimized UAV position (x, y, z), global transmit-power scale p_{scale}, and achieved coverage.

Distribution	x (m)	y (m)	z (m)	p_{scale}	Coverage
clustered	864.70	-36.89	21.45	0.600	0.10
uniform	742.46	1445.02	91.09	0.921	0.10
ring	1490.73	-150.89	165.55	0.312	0.10
edge	920.32	360.25	246.87	0.953	0.10

Table 5 reports the PSO–OFDMA baseline. Even with optimized placement and a tuned global power scale, coverage saturates at ≈ 0.10 for all layouts. This reinforces that placement-only heuristics under OFDMA cannot match joint mobility–power–scheduling learned with NOMA and adaptive SIC (Algorithm 1).

Across all topologies, PPO+NOMA with adaptive SIC achieves higher and more stable rate-coverage, improves fairness by elevating the lower-rate tail, and outperforms both OFDMA-constrained PPO and PSO-based placement baselines.

5. Discussion and Limitations

This section interprets the empirical findings, discusses practical implications for UAV-assisted uplink connectivity, and identifies limitations and avenues for future research.

5.1. Key Findings and Practical Implications

- *Coverage gains across diverse layouts.*
  Across uniform, clustered, ring, and edge-heavy deployments, the proposed PPO+UL–NOMA agent consistently improves *rate coverage* relative to strong baselines (Tables 4 and 5). Typical gains over PPO with OFDMA lie in the 8–10 pp range, with the largest improvements in edge-heavy scenarios (Figure 8) where near–far disparities are most pronounced and adaptive SIC can be exploited effectively.
- *Fairness and user experience.*
  Per-user rate CDFs (Figures 12(a)–(d)) show that the learned policy not only raises average performance but also shifts the distribution upward so that *most* users exceed R_{min}. This is particularly relevant for emergency and temporary coverage, where serving many users with a minimum quality-of-service (QoS) is paramount.
- *Lower user-side power.*
  Relative to baselines, the learned controller reduces median UE transmit power (by up to tens of percent in our runs), reflecting more favorable placement and pairing decisions. Lower UE power is desirable for battery-limited devices and improves thermal/noise robustness at the receiver.
- *Feasibility by design.*
  The *bounded-action* parameterization guarantees kinematic feasibility, contributing to stable training and trajectories that respect FAA altitude and speed limits. The learned paths (Figure 11) exhibit quick exploration followed by convergence to stable hovering locations that balance distance and visibility (Section 3).

5.2. Limitations and Threats to Validity

- *Single-UAV, single-cell abstraction.*
  Results are obtained for one UAV serving a single cell. Interference coupling and coordination in multi-UAV, multi-cell networks are not modeled and may affect achievable coverage.
- *Channel and hardware simplifications.*
  We adopt a widely used A2G model (Al-Hourani in 3GPP UMa) with probabilistic LoS/NLoS and a fixed noise figure. Small-scale fading dynamics, antenna patterns, and hardware impairments (e.g., timing offsets) are abstracted, and Shannon rates are used as a proxy for link adaptation.

 doi:10.20944/preprints202509.2037.v1

- *Energy, endurance, and environment.*
  UAV battery dynamics, wind/gusts, no-fly zones, and backhaul constraints are outside our scope. These factors can influence feasible trajectories and airtime.
- *Objective design.*
  We optimize rate coverage at a fixed $R_{\min}$. Other system objectives; e.g., joint optimization of coverage, average throughput, and energy introduce multi-objective trade-offs that we do not explore here.

5.3. Future Work

We identify several natural extensions: (i) **Multi-UAV coordination** via MARL with interference-aware pairing and collision-avoidance constraints; (ii) **Energy-aware control** that co-optimizes flight energy, airtime, and user coverage under battery/endurance models; (iii) **Environment realism**, including wind fields, no-fly zones, and 3D urban geometry; (iv) **Robust learning**, with explicit modeling of SIC residuals and CSI uncertainty, and safety layers for constraint satisfaction; (v) **Multi-objective optimization**, e.g., Pareto-efficient policies trading coverage, throughput, and energy; and (vi) **Sample-efficient training** through model-based RL, curriculum learning, or offline pretraining before online fine-tuning.

6. Conclusions

We studied joint UAV motion control, uplink power allocation, and UL–NOMA scheduling under a realistic A2G channel and regulatory kinematic constraints. Our bounded-action PPO–GAE agent coordinates UAV acceleration with per-subchannel power and adaptive SIC to *maximize rate coverage*. Across four canonical spatial layouts, it consistently outperforms PPO with OFDMA and placement/power baselines, raising the fraction of users above the minimum-rate threshold and reducing median UE transmit power (Figures 5–8, 12; Tables 4–5). Ablations indicate that UL–NOMA with adaptive SIC, feasibility-aware actions, and joint trajectory–power decisions are all critical to the gains.

Limitations include the single-UAV/single-cell abstraction, simplified environment physics, and the use of a single primary objective. Future work will address multi-UAV settings, energy/flight-time constraints, weather and airspace restrictions, and multi-objective formulations. We plan to release seeds, configuration files, and environment scripts to facilitate reproducibility and benchmarking in this domain.

Author Contributions: Conceptualization, A.M.A., P.M.T. and M.D.; methodology, A.M.A., P.M.T.; software, A.M.A.; validation, A.M.A., P.M.T., H.T., H.K., M.N.S., and M.D.; formal analysis, P.M.T., H.K. and M.D.; resources, P.M.T.; data curation, A.M.A., P.M.T.; writing-original draft presentation, A.M.A. and P.M.T.; writing-reviewing and editing, A.M.A., P.M.T., H.T., H.K., R.A., A.K. and M.D.; supervision, P.M.T.; funding acquisition, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation under the Award OIA-2417062 to the DREAM Research Center and the UNM IoT and Intelligent Systems Innovation Lab (I2 Lab).

Data Availability Statement: The MATLAB code and datasets used in this study—including the simulated A2G environment with N=100 users, a fully functional UAV motion/kinematics module, UL–NOMA/SIC routines, and the bounded-action PPO–GAE training scripts—are available via the Distributed Resilient and Emergent Intelligence-based Additive Manufacturing (DREAM) project GitHub repository. Direct repository access is restricted, all data and code are available from the corresponding authors upon reasonable request.

 doi:10.20944/preprints202509.2037.v1

Abbreviations

3GPP	3rd Generation Partnership Project
5G	Fifth-Generation Mobile Network
A2G	Air-to-Ground
Adam	Adaptive Moment Estimation (optimizer)
CDF	Cumulative Distribution Function
CI	Confidence Interval
CSI	Channel State Information
FAA	Federal Aviation Administration
GAE	Generalized Advantage Estimation
LAP	Low-Altitude Platform
LoS	Line-of-Sight
MAC	Multiple Access Channel
MARL	Multi-Agent Reinforcement Learning
MDPI	Multidisciplinary Digital Publishing Institute
MLP	Multi-Layer Perceptron
NF	Noise Figure
NLoS	Non-Line-of-Sight
NOMA	Non-Orthogonal Multiple Access
OMA	Orthogonal Multiple Access
OFDMA	Orthogonal Frequency-Division Multiple Access
pp	Percentage Points
PPO	Proximal Policy Optimization
PSO	Particle Swarm Optimization
QoS	Quality of Service
RL	Reinforcement Learning
SIC	Successive Interference Cancellation
SNR	Signal-to-Noise Ratio
UE	User Equipment
UAV	Unmanned Aerial Vehicle
UAV-BS	Unmanned Aerial Vehicle Base Station
UL	Uplink
UL–NOMA	Uplink Non-Orthogonal Multiple Access
UMa	Urban Macro (3GPP)

References

1. V, S.; B. V, V.; S M, K. Deep Q-Networks and 5G Technology for Flight Analysis and Trajectory Prediction. In Proceedings of the 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), **2024**, pp. 1–6. https://doi.org/10.1109/CONECCT62155.2024.10677159.
2. Xu, J. Efficient trajectory optimization and resource allocation in UAV 5G networks using dueling-Deep-Q-Networks. *Wireless Networks* **2024**, *30*, 6687–6697. https://doi.org/10.1007/s11276-023-03488-1.
3. Tang, F.; Zhou, Y.; Kato, N. Deep Reinforcement Learning for Dynamic Uplink/Downlink Resource Allocation in High Mobility 5G HetNet. *IEEE Journal on Selected Areas in Communications* **2020**, *38*, 2773–2782. https://doi.org/10.1109/JSAC.2020.3005495.
4. Gyawali, S.; Qian, Y.; Hu, R.Q. Deep Reinforcement Learning Based Dynamic Reputation Policy in 5G Based Vehicular Communication Networks. *IEEE Transactions on Vehicular Technology* **2021**, *70*, 6136–6146. https://doi.org/10.1109/TVT.2021.3079379.
5. McClellan, M.; Cervelló-Pastor, C.; Sallent, S. Deep Learning at the Mobile Edge: Opportunities for 5G Networks. *Applied Sciences* **2020**, *10*. https://doi.org/10.3390/app10144735.
6. Caillouet, C.; Mitton, N. Optimization and Communication in UAV Networks. *Sensors* **2020**, *20*. https://doi.org/10.3390/s20185036.

 doi:10.20944/preprints202509.2037.v1

7. Exposito Garcia, A.; Esteban, H.; Schupke, D. Hybrid Route Optimisation for Maximum Air to Ground Channel Quality. *Journal of Intelligent & Robotic Systems* **2022**, *105*. https://doi.org/10.1007/s10846-022-01590-8.
8. Chen, Y.; Lin, X.; Khan, T.; Afshang, M.; Mozaffari, M. 5G Air-to-Ground Network Design and Optimization: A Deep Learning Approach, **2020**, [arXiv:cs.IT/2011.08379].
9. Hu, H.; Da, X.; Huang, Y.; Zhang, H.; Ni, L.; Pan, Y. SE and EE Optimization for Cognitive UAV Network Based on Location Information. *IEEE Access* **2019**, *7*, 162115–162126. https://doi.org/10.1109/ACCESS.2019.2951702.
10. Luong, P.; Gagnon, F.; Tran, L.N.; Labeau, F. Deep Reinforcement Learning-Based Resource Allocation in Cooperative UAV-Assisted Wireless Networks. *IEEE Transactions on Wireless Communications* **2021**, *20*, 7610–7625. https://doi.org/10.1109/TWC.2021.3086503.
11. Oubbati, O.S.; Lakas, A.; Guizani, M. Multiagent Deep Reinforcement Learning for Wireless-Powered UAV Networks. *IEEE Internet of Things Journal* **2022**, *9*, 16044–16059. https://doi.org/10.1109/JIOT.2022.3150616.
12. Yin, S.; Zhao, S.; Zhao, Y.; Yu, F.R. Intelligent Trajectory Design in UAV-Aided Communications With Reinforcement Learning. *IEEE Transactions on Vehicular Technology* **2019**, *68*, 8227–8231. https://doi.org/10.1109/TVT.2019.2923214.
13. Bithas, P.S.; Michailidis, E.T.; Nomikos, N.; Vouyioukas, D.; Kanatas, A.G. A Survey on Machine-Learning Techniques for UAV-Based Communications. *Sensors* **2019**, *19*. https://doi.org/10.3390/s19235170.
14. Zhong, R.; Liu, X.; Liu, Y.; Chen, Y. Multi-Agent Reinforcement Learning in NOMA-Aided UAV Networks for Cellular Offloading. *IEEE Transactions on Wireless Communications* **2022**, *21*, 1498–1512. https://doi.org/10.1109/TWC.2021.3104633.
15. Zhang, L.; Jabbari, B.; Ansari, N. Deep Reinforcement Learning Driven UAV-Assisted Edge Computing. *IEEE Internet of Things Journal* **2022**, *9*, 25449–25459. https://doi.org/10.1109/JIOT.2022.3196842.
16. Silvirianti.; Narottama, B.; Shin, S.Y. Layerwise Quantum Deep Reinforcement Learning for Joint Optimization of UAV Trajectory and Resource Allocation. *IEEE Internet of Things Journal* **2024**, *11*, 430–443. https://doi.org/10.1109/JIOT.2023.3285968.
17. Zhang, W.; Zhang, S.; Wu, F.; Wang, Y. Path Planning of UAV Based on Improved Adaptive Grey Wolf Optimization Algorithm. *IEEE Access* **2021**, *9*, 89400–89411. https://doi.org/10.1109/ACCESS.2021.3090776.
18. Li, Y.; Zhang, H.; Long, K.; Jiang, C.; Guizani, M. Joint Resource Allocation and Trajectory Optimization With QoS in UAV-Based NOMA Wireless Networks. *IEEE Transactions on Wireless Communications* **2021**, *20*, 6343–6355. https://doi.org/10.1109/TWC.2021.3073570.
19. Popović, M.; Vidal-Calleja, T.; Hitz, G.; Chung, J.J.; Sa, I.; Siegwart, R.; Nieto, J. An informative path planning framework for UAV-based terrain monitoring. *Autonomous Robots* **2020**, *44*, 889–911. https://doi.org/10.1007/s10514-020-09903-2.
20. Li, B.; peng Yang, Z.; qing Chen, D.; yang Liang, S.; Ma, H. Maneuvering target tracking of UAV based on MN-DDPG and transfer learning. *Defence Technology* **2021**, *17*, 457–466. https://doi.org/https://doi.org/10.1016/j.dt.2020.11.014.
21. Yang, F.; Fang, X.; Gao, F.; Zhou, X.; Li, H.; Jin, H.; Song, Y. Obstacle Avoidance Path Planning for UAV Based on Improved RRT Algorithm. *Discrete Dynamics in Nature and Society* **2022**, *2022*, 4544499, [https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/4544499]. https://doi.org/https://doi.org/10.1155/2022/4544499.
22. Maboudi, M.; Homaei, M.; Song, S.; Malihi, S.; Saadatseresht, M.; Gerke, M. A Review on Viewpoints and Path Planning for UAV-Based 3-D Reconstruction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2023**, *16*, 5026–5048. https://doi.org/10.1109/JSTARS.2023.3276427.
23. Wang, Y.; Gao, Z.; Zhang, J.; Cao, X.; Zheng, D.; Gao, Y.; Ng, D.W.K.; Renzo, M.D. Trajectory Design for UAV-Based Internet of Things Data Collection: A Deep Reinforcement Learning Approach. *IEEE Internet of Things Journal* **2022**, *9*, 3899–3912. https://doi.org/10.1109/JIOT.2021.3102185.
24. Yilmaz, B.Y.; Denizer, S.N. Multi UAV Based Traffic Control in Smart Cities. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), **2020**, pp. 1–7. https://doi.org/10.1109/ICCCNT49239.2020.9225622.
25. Lakew, D.S.; Masood, A.; Cho, S. 3D UAV Placement and Trajectory Optimization in UAV Assisted Wireless Networks. In Proceedings of the 2020 International Conference on Information Networking (ICOIN), **2020**, pp. 80–82. https://doi.org/10.1109/ICOIN48656.2020.9016553.

 doi:10.20944/preprints202509.2037.v1

26. Liu, B.; Su, Z.; Xu, Q. Game theoretical secure wireless communication for UAV-assisted vehicular Internet of Things. *China Communications* **2021**, *18*, 147–157. https://doi.org/10.23919/JCC.2021.07.012.
27. Matracia, M.; Kishk, M.A.; Alouini, M.S. On the Topological Aspects of UAV-Assisted Post-Disaster Wireless Communication Networks. *IEEE Communications Magazine* **2021**, *59*, 59–64. https://doi.org/10.1109/MCOM.121.2100166.
28. Fan, X.; Zhou, H.; Sun, K.; Chen, X.; Wang, N. Channel Assignment and Power Allocation Utilizing NOMA in Long-Distance UAV Wireless Communication. *IEEE Transactions on Vehicular Technology* **2023**, *72*, 12970–12982. https://doi.org/10.1109/TVT.2023.3272580.
29. Xie, J.; Chang, Z.; Guo, X.; Hämäläinen, T. Energy Efficient Resource Allocation for Wireless Powered UAV Wireless Communication System With Short Packet. *IEEE Transactions on Green Communications and Networking* **2023**, *7*, 101–113. https://doi.org/10.1109/TGCN.2022.3218314.
30. Eom, S.; Lee, H.; Park, J.; Lee, I. UAV-Aided Wireless Communication Designs With Propulsion Energy Limitations. *IEEE Transactions on Vehicular Technology* **2020**, *69*, 651–662. https://doi.org/10.1109/TVT.2019.2952883.
31. Huo, Y.; Dong, X.; Lu, T.; Xu, W.; Yuen, M. Distributed and Multilayer UAV Networks for Next-Generation Wireless Communication and Power Transfer: A Feasibility Study. *IEEE Internet of Things Journal* **2019**, *6*, 7103–7115. https://doi.org/10.1109/JIOT.2019.2914414.
32. Li, B.; Fei, Z.; Zhang, Y.; Guizani, M. Secure UAV Communication Networks over 5G. *IEEE Wireless Communications* **2019**, *26*, 114–120. https://doi.org/10.1109/MWC.2019.1800458.
33. Wu, Z.; Kumar, H.; Davari, A. Performance evaluation of OFDM transmission in UAV wireless communication. In Proceedings of the Proceedings of the Thirty-Seventh Southeastern Symposium on System Theory, 2005. SSST '05., **2005**, pp. 6–10. https://doi.org/10.1109/SSST.2005.1460867.
34. Yao, Z.; Cheng, W.; Zhang, W.; Zhang, H. Resource Allocation for 5G-UAV-Based Emergency Wireless Communications. *IEEE Journal on Selected Areas in Communications* **2021**, *39*, 3395–3410. https://doi.org/10.1109/JSAC.2021.3088684.
35. Ouyang, J.; Che, Y.; Xu, J.; Wu, K. Throughput Maximization for Laser-Powered UAV Wireless Communication Systems. In Proceedings of the 2018 IEEE International Conference on Communications Workshops (ICC Workshops), **2018**, pp. 1–6. https://doi.org/10.1109/ICCW.2018.8403572.
36. Zeng, Y.; Xu, J.; Zhang, R. Energy Minimization for Wireless Communication With Rotary-Wing UAV. *IEEE Transactions on Wireless Communications* **2019**, *18*, 2329–2345. https://doi.org/10.1109/TWC.2019.2902559.
37. Alnagar, S.I.; Salhab, A.M.; Zummo, S.A. Q-Learning-Based Power Allocation for Secure Wireless Communication in UAV-Aided Relay Network. *IEEE Access* **2021**, *9*, 33169–33180. https://doi.org/10.1109/ACCESS.2021.3061406.
38. Tian, Y.; Li, H.; Zhu, Q.; Mao, K.; Ali, F.; Chen, X.; Zhong, W. Generative Network-Based Channel Modeling and Generation for Air-to-Ground Communication Scenarios. *IEEE Communications Letters* **2024**, *28*, 892–896. https://doi.org/10.1109/LCOMM.2024.3363621.
39. Lala, V.; Ndreveloarisoa, A.F.; Desheng, W.; Heriniaina, R.F.; Murad, N.M.; Fontgalland, G.; Ravelo, B. Channel Modelling for UAV Air-to-Ground Communication. In Proceedings of the 2024 5th International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM), **2024**, pp. 1–5. https://doi.org/10.1109/ELECOM63163.2024.10892167.
40. Ning, B.; Li, T.; Mao, K.; Chen, X.; Wang, M.; Zhong, W.; Zhu, Q. A UAV-aided channel sounder for air-to-ground channel measurements. *Physical Communication* **2021**, *47*, 101366. https://doi.org/https://doi.org/10.1016/j.phycom.2021.101366.
41. Al-Hourani, A.; Kandeepan, S.; Lardner, S. Optimal LAP Altitude for Maximum Coverage. *IEEE Wireless Communications Letters* **2014**, *3*, 569–572. https://doi.org/10.1109/LWC.2014.2363836.
42. Zhong, X. Deploying UAV Base Stations in Communication Networks Using Machine Learning. Master's thesis, Simon Fraser University, Department of Electrical and Computer Engineering, **2017**.
43. Demographics of the world. Demographics of the world — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Demographics_of_the_world?oldid=XXXXX, **2025**. [Online; accessed 3 September 2025].
44. Federal Aviation Administration. Small Unmanned Aircraft Systems (UAS) Regulations (Part 107). https://www.faa.gov/newsroom/small-unmanned-aircraft-systems-uas-regulations-part-107, **2025**. Accessed: 3 September 2025.

 doi:10.20944/preprints202509.2037.v1

45. Yu, L.; Li, Z.; Ansari, N.; Sun, X. Hybrid Transformer Based Multi-Agent Reinforcement Learning for Multiple Unpiloted Aerial Vehicle Coordination in Air Corridors. *IEEE Transactions on Mobile Computing* **2025**, *24*, 5482–5495. https://doi.org/10.1109/TMC.2025.3532204.
46. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *CoRR* **2017**, *abs/1707.06347*, [1707.06347].