Adaptive Communication for Joint Trajectory and RRM in MADRL-Based UAV Networks

Danila Ferretti*†, Leonardo Spampinato‡, Enrico Testi‡, Chiara Buratti‡, Riccardo Marini*
*WiLab, CNIT, Bologna, Italy email: {danila.ferretti, riccardo.marini}@wilab.cnit.it
†La Sapienza, University of Rome, Rome, Italy, email: danila.ferretti@uniroma1.it
‡DEI, University of Bologna / WiLab, CNIT, Bologna, Italy email: {leonardo.spampinato, enrico.testi, c.buratti}@unibo.it

***Abstract*—This paper addresses the joint design of Unmanned Aerial Vehicles (UAVs) trajectory and radio resource management (RRM) in dynamic wireless environments by leveraging a multi-agent deep reinforcement learning (MADRL) framework. In contrast to prior works that either assume constant synchronization between agents and the controller or overlook the communication cost, we explicitly model the interaction between UAVs and the central controller. We propose an adaptive synchronization strategy that selectively transmits model parameters and experience data based on their relevance, enabling a resource-aware RRM algorithm that optimally balances learning performance and communication overhead. The MADRL agents optimize their trajectories based on rewards that incorporate priorities derived from the RRM layer, which jointly manages both uplink and downlink communications. Simulation results demonstrate that our event-driven synchronization strategy outperforms periodic baselines in both convergence speed and communication overhead, towards scalable deployment in realistic urban environments.**

***Index Terms*—UAV, Trajectory Design, Radio Resource Management, Multi-Agent Reinforcement Learning**

I. Introduction

The application of MADRL techniques to UAV-based communication systems has become increasingly relevant, particularly in scenarios where trajectory design and RRM must be jointly optimized. This convergence addresses the dynamic nature of wireless networks and the need for intelligent coordination among UAVs. Recent literature has explored the integration of MADRL with UAV-enabled communication systems, particularly for trajectory optimization and RRM [1]–[4]. Several works adopt MADRL to enable cooperation among UAVs acting as unmanned aerial base station (UABS), optimizing service coverage or offloading performance in vehicular or Internet of Things (IoT) networks. For instance, [2] leverages graph neural networks to coordinate UAV trajectories, while [1] focuses on MADRL in NOMA-based offloading scenarios. Federated and distributed learning paradigms have also gained traction to reduce centralized overhead: [4], [5] propose asynchronous updates and hierarchical control to adapt learning processes to heterogeneous UAV systems with limited connectivity. However, most of these solutions rely on fixed or periodic synchronization between agents and the training server. Only [5] and [6] introduce mechanisms for asynchronous or event-triggered updates. In particular, [6] develops an event-triggered communication model for general MADRL systems under limited-bandwidth constraints, laying the theoretical foundation for adaptive synchronization. Yet, these approaches do not integrate the synchronization strategy within a UAV communication system that jointly optimizes trajectory, RRM, and learning coordination. In contrast, this work proposes a unified MADRL-based framework that jointly optimizes UAV trajectories and RRM. By introducing an event-driven mechanism to exchange model parameters and agent experiences based on communication relevance, we significantly reduce signaling overhead without degrading system performance. To the best of the authors' knowledge, this is the first work to tightly couple learning and communication layers through a priority-based RRM policy that accounts for both user traffic and overhead control data. The paper is organized as follows: Sec. II introduces the system model; Sec. III and Sec. IV present the proposed MADRL algorithm and its RRM integration; Sec. V reports simulation results and Sec. VI concludes the work.

This work has been carried out in the framework of the CNIT National Laboratory WiLab and the WiLab-Huawei Joint Innovation Center. We would like to thank Aman Jassal, Chan Zhou and Malte Schellmann for the very fruitful discussion on this paper.

II. System Model

We consider a scenario of dimension $X \times Y$, depicted in Fig. 1, set in Via Saragozza, Bologna. A set of macro base stations (MBSs) $m \in \mathcal{M}$ at (x_m, y_m, h_m) operate at carrier frequency f_c in the mmWave band, assisted by a set of UABSs $u \in \mathcal{U}$ to optimize resource allocation for ground user equipments (GUEs) $g \in \mathcal{G}$ via an RRM algorithm. For simplicity, in this work we only consider a single MBS, i.e., $|\mathcal{M}| = 1$. User mobility is simulated using SUMO [7] for realistic movement, with GUEs moving with an average speed v_g, through an urban environment with multiple traffic lights, generating dynamic hotspots and realistic vehicular behavior. Time is discretized as $t = 0, 1, \ldots, T$, where each timestep corresponds to a scheduling interval of duration Δt. The mission unfolds as a sequence of UAVs positions, where the position of each UABS at time t is given by $(x_u^{(t)}, y_u^{(t)}, h_u)$, and initial position $(x_u^{(0)}, y_u^{(0)}, h_u)$. GUEs establish vehicle-to-anything (V2X) communication by connecting to either the MBS or a UABS, and periodically transmit one uplink (UL) and receive one downlink (DL) packet every δt, with $\Delta t = 10\delta t$. UL and DL transmissions occur independently, as two parallel uncorrelated traffic streams. Although traffic

979-8-3503-9281-4/25/$31.00 ©2025 IEEE

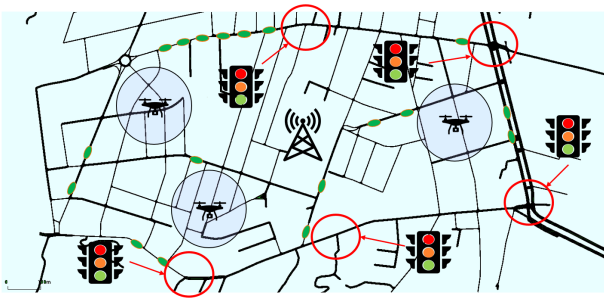

Fig. 1. An illustration of the reference scenario, modeled after the district of Via Saragozza, Bologna, Italy.

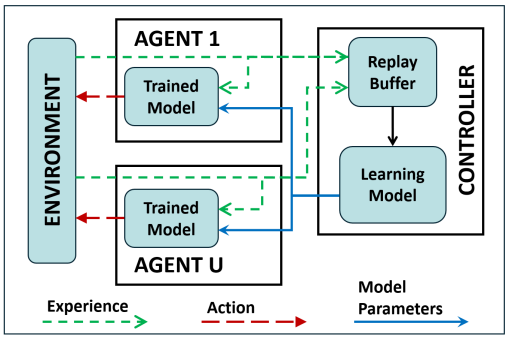

Fig. 2. Schematic representation of the multi-agent scheme showing the interaction between agents and controller.

is periodic, continuous service is essential to ensure a stable Quality of Experience (QoE). Accordingly, a GUE is considered *satisfied* if it is *served* (i.e., meets demand D_g in UL or DL) for at least $\hat{N}_\mathrm{s}$ time intervals within a service window $T_s = N_s \Delta t$. A GUE is served in UL if $\psi_g^{(ul)(t)} = 1$, and in DL if $\psi_g^{(dl)(t)} = 1$. A priority term $p_g^{(t)}$ tracks served intervals within T_s:

$$p_g^{(t)} = \begin{cases} 1, & \text{for } t = 1 \\ p_g^{(t-1)} + 1, & \text{if } \psi_g^{(t)} = 1 \\ p_g^{(t-1)}, & \text{if } \psi_g^{(t)} = 0 \end{cases} \tag{1}$$

A. Channel Model and Resource Definition

We adopt the Urban Macro (UMa) channel and resource model from 3GPP TR 38.901 [8]–[10], accounting for Line of Sight (LoS)/Non-LoS (NLoS) and modeling shadowing as log-normal. The Signal-to-Noise ratio (SNR) and Signal-to-Interference-plus-Noise ratio (SINR) in dB are given by $SNR = P_\mathrm{rx} - P_\mathrm{noise}$ and $SINR = P_\mathrm{rx} - (P_\mathrm{noise} + \sum_{i=1}^{N_{int}} P_{\mathrm{rx},i})$, with rates derived from Shannon capacity. The allocation of resources follows the 5G standard, where the number of resource units (RUs) per RRM period is $W = \frac{B_\mathrm{sys}}{12\Delta f} \cdot \frac{\Delta t}{T_\mathrm{slot}}$, with B_sys, Δf and T_slot denoting the bandwidth of the system, the subcarrier spacing, and the duration of the slots. Beamforming at the UAV is modeled as a 3x3 directional grid and MBSs and UABSs share the same resource blocks (RBs) set, W, maximizing spectral efficiency but introducing inter-cell interference across UL/DL GUE links, which remain orthogonal. The RRM algorithm schedules RUs based on priorities shaped by communication demands and synchronization needs, enabling interference prediction.

III. Multi-Agent Deep Reinforcement Learning

We address the problem of UABSs trajectory planning by modeling it as a Markov Decision Process (MDP). To jointly optimize trajectory and resource allocation, we adopt a MADRL framework based on the Double Dueling Deep Q Network (3DQN) algorithm, in which each agent's policy is modeled by a deep neural network (NN) trained to approximate the action-value function. Each UABS u flies at a speed v_u and maintains a backhaul link with the MBS for coordination and data exchange. The goal of each UABS is to follow an optimal trajectory to improve network performance while cooperating with the MBS. At each timestep t, agent u observes a tuple $o_u^{(t)}$ of elements $(x_u^{(t)}, y_u^{(t)}, t, \mathbf{L}_t, \mathbf{b}_u^{(t)})$, where $\mathbf{L}_t$ is a matrix containing all the current locations of agents in the fleet, and $\mathbf{b}_u^{(t)}$ is the *per beam information*, i.e., a vector of length $|\mathcal{B}|$ whose elements $b_{j_u}^{(t)}$ correspond to the sum of UL priority $p_g^{(ul)(t)}$ and DL priority $p_g^{(dl)(t)}$ of GUEs under the j_u-th beam of the u-th agent. Each agent stores the experience tuple $\{o_u^{(t)}, a_u^{(t)}, r_u^{(t)}, o_u^{(t+1)}\}$ in its local buffer Z_u. Over time, experiences are shared with a centralized controller maintaining a global buffer $\mathcal{Z}$ of size $|\mathcal{Z}|$. Moreover, the central controller is in charge of training the global NN model parameters θ_t and distributing the updated policy to the agents. Based on the current observation $o_u^{(t)}$ and the policy π, modeled by the shared NN parameters θ_s, each agent selects an action $a_u^{(t)}$ from the discrete action space $\mathcal{A} = \{\leftarrow, \uparrow, \rightarrow, \downarrow, \nwarrow, \nearrow, \searrow, \swarrow, \emptyset\}$, where $\emptyset$ denotes hovering. Upon executing $a_u^{(t)}$, the agent receives a reward $r_u^{(t)}$ that reflects the effectiveness of its action, promoting a balanced and cooperative behavior among all UABSs, and accounting for both UL and DL transmissions:

$$r_u^{(t)} = \frac{\sum_{g \in \mathcal{Y}^{(ul)(t)}} p_g^{(ul)(t)} + \sum_{g \in \mathcal{Y}^{(dl)(t)}} p_g^{(dl)(t)}}{|\mathcal{U}|}, \tag{2}$$

which corresponds to the sum of the priorities of GUEs in $\mathcal{Y}^{(ul)(t)}$, i.e., the subset of GUEs served by all UABSs in UL, and the sum of the priorities of GUEs in $\mathcal{Y}^{(dl)(t)}$, i.e., the subset of GUEs served by all UABSs in DL. Both UL and DL have been carefully accounted for in the reward calculation, ensuring that the performance of the reinforcement learning (RL) algorithm, and consequently the trajectories followed by the UABSs, are optimized following the resource allocation requirements for both UL and DL transmissions.

A. Model and Experiences Synchronization

Fig. 2 illustrates the MADRL framework, highlighting the bidirectional communication between controller and agents: blue arrows indicate model updates, while green arrows represent experience sharing. To make this exchange more

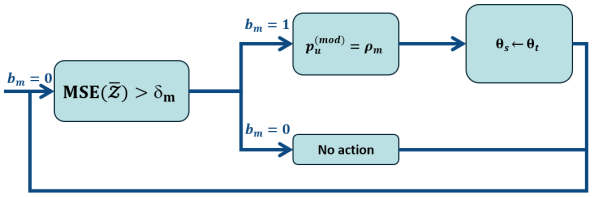

Fig. 3. Schematic representation of the model synchronization process.

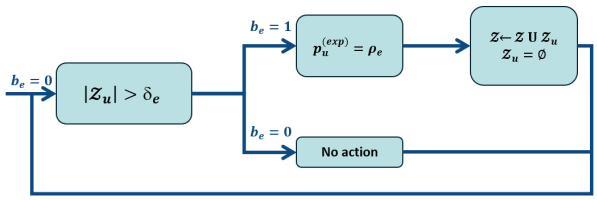

Fig. 4. Schematic representation of the experience synchronization process.

efficient, we introduce an adaptive synchronization strategy that reduces communication overhead without compromising performance. Unlike fixed-interval updates, synchronization is event-driven—UABSs transmit experiences and receive model weights only when necessary. Unlike traditional MADRL schemes that sync at every $t = 0, \ldots, T$, our approach uses two binary flags, b_m and b_e, to trigger *model and experience synchronization*, respectively. These flags, initialized to 0, are set to 1 upon meeting specific conditions. Additionally, two priority values are introduced to manage RU assignment for transmitting overhead data within the RRM process.

1) Model Synchronization

Let θ_s denote the latest model shared with the agents and θ_t the current model being trained at the controller. Let $\bar{\mathcal{Z}}$ be a batch of experiences sampled from $\mathcal{Z}$. As shown in Fig. 3, model synchronization is triggered when the Mean Squared Error (MSE) between Q-values $Q(s, a|\theta_s)$ and $Q(s, a|\theta_t)$, estimated using NN with weights θ_s and θ_t, over $\bar{\mathcal{Z}}$ exceeds a threshold δ_m. We refer to the difference between the estimated Q-values as an error, assuming that $Q(s, a|\theta_t)$ represent the ground truth. This condition can be expressed as:

$$b_m = \begin{cases} 1, & \text{if MSE}(\bar{\mathcal{Z}}) \geq \delta_m \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\text{MSE}(\bar{\mathcal{Z}}) = \frac{1}{|\bar{\mathcal{Z}}|} \sum_{(s,a,r,s') \in \bar{\mathcal{Z}}} \left(Q(s, a|\theta_s) - Q(s, a|\theta_t)\right)^2$. The following operations are performed when $b_m = 1$:

- The priority term assigned to the transmission of the overhead data $p_u^{(oh)}$ is set to $p_u^{(mod)} = \rho_m$, with ρ_m being a high enough value to ensure the model synchronization.
- The updated policy, represented by the neural network weights θ_t, is sent from the controller to the UABSs, thus $\theta_s \leftarrow \theta_t$.
- Once synchronization is completed, b_m is reset to 0.

This event-driven approach ensures that policy updates are communicated only when significant changes occur, balancing communication efficiency with system performance.

2) Experiences Synchronization

For experience synchronization, shown in Fig. 4, the trigger condition is:

$$b_e = \begin{cases} 1, & \text{if } |\mathcal{Z}_u| \geq \delta_e \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $|\mathcal{Z}_u|$ is the size of the local buffer and δ_e is the experience threshold. When $b_e = 1$, the following steps occur:

- The priority term assigned to the transmission of the overhead data $p_u^{(oh)}$ is set to $p_u^{(exp)} = \rho_e$, with ρ_e being a high enough value to ensure the experience synchronization.
- The experiences stored in $\mathcal{Z}_u$ by all UABSs $u \in \mathcal{U}$ since the last synchronization are transmitted to the controller and stored into $\mathcal{Z}$, thus $\mathcal{Z} \leftarrow \mathcal{Z} \cup \mathcal{Z}_u$.
- The local replay buffers $\mathcal{Z}_u$ are flushed, i.e. $|\mathcal{Z}_u| = 0$.
- Once synchronization is completed, b_e is reset to 0.

This adaptive mechanism transmits experiences only when enough data is available, reducing overhead. It enables an efficient integration of learning and communication, optimizing exchange frequency and minimizing data transmission between agents and controller while preserving performance.

IV. Radio Resource Management

In the following, we formulate two parallel Integer Linear Program (ILP) problems to optimize UL and DL resource allocation in an Frequency Division Duplexing (FDD) scenario. While the formulations are structurally identical, with time indexing omitted for brevity, they differ in the direction of data transmission, hence the interference dynamics and the overhead data. Unlike prior work relying on alternating allocation, our framework enables simultaneous execution of both ILPs, guided by distinct priority terms $p_g^{(ul)}$ and $p_g^{(dl)}$ to capture each GUE's full communication needs. To further improve efficiency, overhead is explicitly modeled via $p_u^{(oh)}$ in the objective, balancing learning-related data (e.g., *model and experience synchronization*) with actual network traffic. Sec. IV-A and Sec. IV-B focus on UL and DL, respectively.

A. Uplink ILP

We formulate an ILP for UL communication, solved at intervals of duration Δt to optimize RRM strategies. The objective is to maximize the number of *served* GUEs via joint MBS and UABS operations, while ensuring overhead transmission. The problem $\mathcal{P}_u$ is defined in (5a)–(5o). Here, kg, j_u indicates whether vehicle g is covered by beam j_u of UABS u. $I_{g,m,u}$ and $I_{g,u,m}$ capture potential interference between MBS–GUE and UABS–GUE links. For example, $I_{g,m,u} = 1$ implies g is in range of both m and u, so its transmission to m may be interfered by transmissions to u. The ILP produces the following outputs: $\lambda_{g,m}, \lambda_{g,u} \in 0, 1$ representing RUs assignment to GUE g by MBS m or UABS u; similarly, $\lambda_{u,m} = 1$ denotes the assignment of RUs from MBS m to

$$\mathcal{P}_u : \max \left(\sum_{g \in \mathcal{G}} \left(\psi_g^{(ul)} \cdot p_g^{(ul)} \right) + \sum_{u \in \mathcal{U}} \psi_u^{(oh)} \cdot p_u^{(oh)} \right) \tag{5a}$$

$$\textbf{s.t.:} \sum_{m \in \mathcal{M}} w_{g,m} r_{g,m} \Delta t + \sum_{u \in \mathcal{U}} \sum_{j_u \in \mathcal{K}_u} k_{g,j_u} w_{g,u} r_{g,u} \Delta t \geq \psi_g^{(ul)} D_g, \, \forall g \in \mathcal{G} \tag{5b}$$

$$\sum_{m \in \mathcal{M}} w_{g,m} r^{\mathrm{I}}_{g,m,u} \Delta t + \sum_{u \in \mathcal{U}} \sum_{j_u \in \mathcal{K}_u} k_{g,j_u} w_{g,u} r^{\mathrm{I}}_{g,u,m} \Delta t \geq \left(\sum_{m \in \mathcal{M}} \iota_{g,m,u} + \sum_{u \in \mathcal{U}} \iota_{g,u,m} \right) D_g, \, \forall g \in \mathcal{G}, \forall u \in \mathcal{U}, \forall m \in \mathcal{M} \tag{5c}$$

$$w_{u,m}^{(oh)} r_{u,m} \Delta t \geq \psi_u^{(oh)} D_u^{(oh)}, \, \forall m \in \mathcal{M}, \, \forall u \in \mathcal{U} \tag{5d}$$

$$\sum_{g \in \mathcal{G}} w_{g,m} + \sum_{u \in \mathcal{U}} \left(w_{u,m} + w_{u,m}^{(oh)} \right) \leq W, \, \forall m \in \mathcal{M} \tag{5e}$$

$$\sum_{g \in \mathcal{G}} k_{g,j_u} w_{g,u} + \sum_{m \in \mathcal{M}} \left(w_{u,m} + w_{u,m}^{(oh)} \right) \leq e_{j_u} W, \, \forall u \in \mathcal{U}, \, \forall j_u \in \mathcal{K}_u \tag{5f}$$

$$\sum_{g \in \mathcal{G}} \sum_{j_u \in \mathcal{K}_u} w_{g,u} k_{g,j_u} r_{g,u} \leq \sum_{m \in \mathcal{M}} r_{u,m} w_{u,m}, \, \forall u \in \mathcal{U} \tag{5g}$$

$$\sum_{j_u \in \mathcal{K}_u} e_{j_u} \leq N_{\text{beam}}, \, \forall u \in \mathcal{U} \tag{5h}$$

$$\sum_{m \in \mathcal{M}} \lambda_{g,m} + \sum_{u \in \mathcal{U}} \lambda_{g,u} \leq 1, \, \forall g \in \mathcal{G} \tag{5i}$$

$$w_{g,m} \leq \lambda_{g,m} W, \, \forall m \in \mathcal{M}, \forall g \in \mathcal{G} \tag{5j}$$

$$w_{g,u} \leq \lambda_{g,u} W, \, \forall u \in \mathcal{U}, \, \forall g \in \mathcal{G} \tag{5k}$$

$$\iota_{m,u} \geq \sum_{g \in \mathcal{G}} \frac{I_{g,u,m} \lambda_{g,u}}{I_{g,u,m}}, \, \forall u \in \mathcal{U}, \forall m \in \mathcal{M} \tag{5l}$$

$$\iota_{u,m} \geq \sum_{g \in \mathcal{G}} \frac{I_{g,m,u} \lambda_{g,m}}{I_{g,m,u}}, \, \forall m \in \mathcal{M}, \forall u \in \mathcal{U} \tag{5m}$$

$$\iota_{g,m,u} \geq \lambda_{g,m} + \iota_{m,u} - 1, \, \forall g \in \mathcal{G}, \, \forall u \in \mathcal{U}, \forall m \in \mathcal{M} \tag{5n}$$

$$\iota_{g,u,m} \geq \lambda_{g,u} + \iota_{u,m} - 1, \, \forall g \in \mathcal{G}, \, \forall m \in \mathcal{M}, \forall u \in \mathcal{U} \tag{5o}$$

$$\mathcal{P}_d : \max \left(\sum_{g \in \mathcal{G}} \left(\psi_g^{(dl)} \cdot p_g^{(dl)} \right) + \sum_{u \in \mathcal{U}} \psi_u^{(oh)} \cdot p_u^{(oh)} \right) \tag{6a}$$

$$\textbf{s.t.:} \ \text{(5b) - (5k)}$$

$$\iota_{m,u} \geq \sum_{g \in \mathcal{G}} \frac{I_{u,g,m} \lambda_{g,u}}{I_{u,g,m}}, \, \forall u \in \mathcal{U}, \forall m \in \mathcal{M} \tag{6b}$$

$$\iota_{u,m} \geq \sum_{g \in \mathcal{G}} \frac{I_{m,g,u} \lambda_{g,m}}{I_{m,g,u}}, \, \forall m \in \mathcal{M}, \forall u \in \mathcal{U} \tag{6c}$$

$$\iota_{m,g,u} \geq \lambda_{g,m} + \iota_{m,u} - 1, \, \forall g \in \mathcal{G}, \, \forall u \in \mathcal{U}, \forall m \in \mathcal{M} \tag{6d}$$

$$\iota_{u,g,m} \geq \lambda_{g,u} + \iota_{u,m} - 1, \, \forall g \in \mathcal{G}, \, \forall m \in \mathcal{M}, \forall u \in \mathcal{U} \tag{6e}$$

UABS u. Beam activation is modeled by $e_{j_u} = 1$ when beam $j_u \in \mathcal{K}_u$ is active on UABS u. Interference is represented by the output variables $\iota_{g,m,u} \in \{0,1\}$, $\iota_{g,u,m} \in \{0,1\}$, $\iota_{m,u} \in \{0,1\}$, and $\iota_{u,m} \in \{0,1\}$, which take the value 1 if interference occurs. For instance, $\iota_{g,m,u} = 1$ if transmission from g to m is interfered by any GUE communicating with UABS u. Moreover, $\iota_{m,u} = 1$ if for any $\iota_{g,m,u} = 1$, $\forall g \in \mathcal{G}$, and similarly for $\iota_{u,m}$. Resource allocations include $w_{g,m} \in \{0, W\}$ and $w_{g,u} \in \{0, W\}$: RUs allocated to GUEs by MBSs and UABSs, respectively and while $w_{u,m} \in \{0, W\}$ and $w_{u,m}^{(oh)} \in \{0, W\}$: RUs allocated between the MBS and UABSs for backhauling and overhead transmissions. These allocations aim to maximize the number of served GUEs, with $\psi_g^{(ul)} = 1$ and successful overhead transmission, with $\psi_u^{(oh)} = 1$. Constraints (5b) and (5c) ensure that transmission demand D_g is met, considering the rate of a unitary RU and the number of assigned RUs, based on noise and interference, respectively. The same applies to constraint (5d), where the demand $D_u^{(oh)} = D_u^{(exp)}$ is the data needed to transmit the experiences from agents to the controller, considering the rate of a unitary RU in the overhead link and the number of assigned RUs. Constraints (5e) and (5f) limit the number of RUs assigned, accounting for backhaul and overhead RUs as well. Constraint (5g) respects the maximum capacity of UL transmission for MBSs and UABSs, while ensuring sufficient backhaul capacity for UABS vehicular traffic. Constraint (5h) restricts the number of activated beams at each UABS u to N_{beam}. Constraints (5i) to (5k) ensure that each vehicle is served by only one base station (BS) at a time. Then, the constraints (5l) and (5m) verify if there is at least one effective interferer on the MBS that is connected to a UABSs u or vice versa, respectively, and constraints (5n) and (5o) verify if a UABS u is interfering the link $g - m$ given it is established or if the MBS m is interfering the link $g - u$, $u \in \mathcal{U}$ given it is established, respectively.

B. Downlink ILP

The DL ILP formulation $\mathcal{P}_d$, detailed in (6a)–(6e), mirrors the UL model with key differences arising from the direction of transmissions and resulting interference in both access and backhaul links. While UL interference stems from GUEs transmitting to MBSs and UABSs, in DL it originates from BSs transmitting to GUEs. Accordingly, interference variables $I_{g,m,u}$ and $I_{g,u,m}$ become $I_{m,g,u}$ and $I_{u,g,m}$, capturing interference from any BS to a GUE served by another node. The resource allocation process $\lambda_{g,m}$, $\lambda_{g,u}$, $\lambda_{u,m}$ and beam activation variables e_{j_u} remain unchanged, but now govern DL resource allocation. Backhaul constraints between MBSs and UABSs still apply, but the focus is on ensuring sufficient capacity for DL transmissions. Demand constraints on D_g ensure each GUE receives the required data, considering DL interference and noise. Overhead transmission in this context refers to model synchronization, with $D_u^{(oh)} = D_u^{(mod)}$.

V. Results

In this section, the results obtained considering the joint design of trajectory and RRM are presented, focusing on a scenario with limited coverage. The MBS is positioned in such a way that a significant portion of the map remains inadequately covered, resulting in a high outage probability. Consequently, the GUEs' QoE cannot be fully supported by the terrestrial network alone, necessitating the critical role of UABSs in maintaining connectivity. The key parameters of

TABLE I
SIMULATION PARAMETERS AND 3DQN HYPERPARAMETERS

T	270 s	T_s	20 s	N_s	20
X	1500 m	Y	700 m	v_g	12 m/s
ϕ_u	100°	$\lvert\mathcal{K}_u\rvert$	9	v_u	20 m/s
h_u	50 m	h_m	30 m	f_c	30 GHz
B_{sys}	200 MHz	N_{sub}	12	Δf	120 KHz
P_{noise}	-106 dBm	δt	0.1 s	T_{slot}	0.125 ms
$\lvert G\rvert$	90	$\lvert\mathcal{M}\rvert$	1	$\lvert\mathcal{U}\rvert$	3
D_g	100 kbit	N_t	1000	N_{eval}	50
n_t	1000	n_o	1	Z	50000
z	128	α	0.0001	γ	0.99
ϵ_{ratio}	0.6	$\epsilon_{\max}$	1	$\epsilon_{\min}$	0.05
δ_m	1000	δ_e	100	ρ_m, ρ_e	1000

the simulation are listed in Tab. I, including the full reuse of resources among the MBS and all UABSs, and the use of an ϵ-greedy policy during training to balance exploration and exploitation. The UABSs are deployed from initial position $(0,0)$, $(0,Y)$, $(X,0)$, (X,Y), $\left(0,\frac{Y}{2}\right)$, $\left(X,\frac{Y}{2}\right)$, $\left(\frac{X}{2},\frac{Y}{2}\right)$, and are trained for N_t training episodes across N_{epoch} epochs, with each consisting of N_e training episodes characterized by different randomly selected *traces* derived from real urban traffic conditions. Evaluation episodes are conducted N_{eval} times during training to assess the UABSs' ability to dynamically design trajectories based on real-time environmental observations, using a fixed evaluation trace. To evaluate the impact of the proposed synchronization strategy, we fixed the threshold values for triggering model and experience exchange to δ_m and δ_e, respectively (see Tab. I). We compare four configurations: in the two *periodic* settings, both model and experience transmissions occur at every timestep, with each agent sending either 1 Mbit or 100 Mbit of experience data per synchronization event. In the two *event-driven* configurations, synchronization is triggered only when the thresholds δ_m or δ_e are exceeded; although each event still carries 1 Mbit or 100 Mbit of experience data per agent, the actual transmission occurs less frequently and thus reflects experiences accumulated over multiple intervals. The performance metrics used in this study are categorized into machine learning-related and network-related. Results are averaged over five random seeds, and the performance is evaluated in terms of the cumulative reward, now based on the UL and DL reward function:

$$R = \sum_{u\in\mathcal{U}} \sum_{t=0}^{T} r_t^{(u)} \tag{7}$$

where all UABSs share the same reward, determined by the average priority of the GUEs served in UL and in DL by the entire fleet of UABSs. To better observe performance trends, the cumulative reward is smoothed using a sliding window average over 20 time intervals.

For network performance, we evaluate the number of GUEs

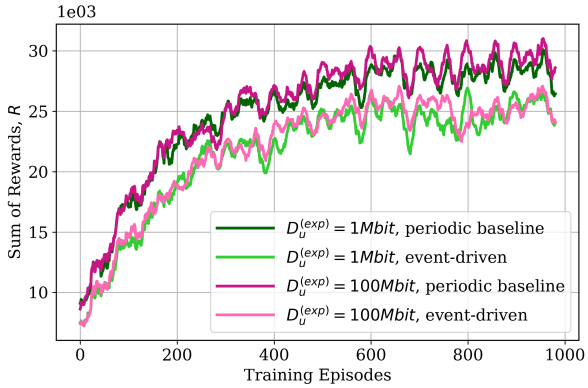

Fig. 5. Sum of Rewards for $D_u^{(exp)}$ sent periodically or event-driven.

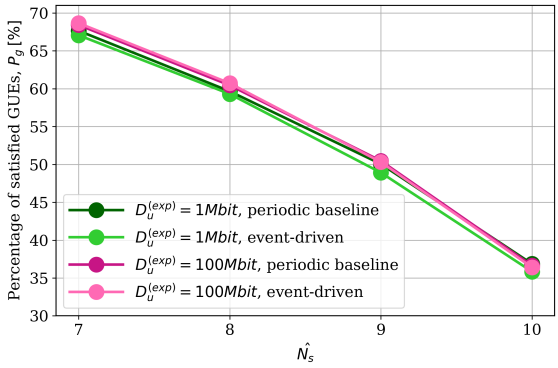

Fig. 6. Pg for $D_u^{(exp)}$ sent periodically or event-driven.

served in UL and in DL at each time step, respectively $\Psi_t^{(ul)} = \sum_{g\in G} \psi_{g,t}^{(ul)}$ and $\Psi_t^{(dl)} = \sum_{g\in G} \psi_{g,t}^{(dl)}$, and the percentage of satisfied users $P_g^{(ul)}$ and $P_g^{(dl)}$. For simplicity, we drop the (ul) and (dl), since the formulations are equal. The definition of P_g is retained as:

$$P_g = \frac{1}{|G|} \sum_{g\in G} \frac{N_g^{(sat)}}{N_g}, \tag{8}$$

where $N_g^{(sat)}$ is the number of satisfied service windows for GUE g and N_g is the total amount of services windows, which accounts for the GUE travel duration. Moreover, we analyze how the network throughput S fluctuates through the N_{epochs} epochs. By defining $S_{i,e}$ as the throughput obtained during the i-th training episode in an epoch e, we obtain $S_{i,e} = \frac{1}{T}\sum_{g\in\mathcal{G}} D_g N_g^{(sat)}$. The average return over the N_e episodes in the e-th epoch can be written as:

$$S_{e,\text{avg}} = \frac{1}{N_e} \sum_{i=0}^{N_e} S_{i,e}. \tag{9}$$

The parameters for the *online network* and *target network* updates, as well as the hyper-parameters for the 3DQN, are also unchanged and are referenced in Table I. Fig. 5 illustrates the impact of different experience data transmission strategies

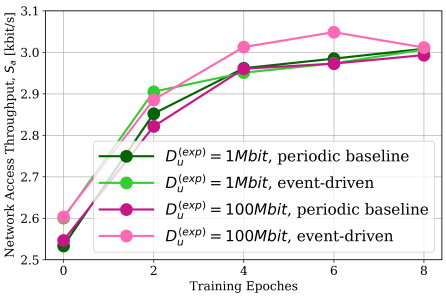

Fig. 7. Network access throughput for $D_u^{(exp)}$ sent periodically or event-driven.

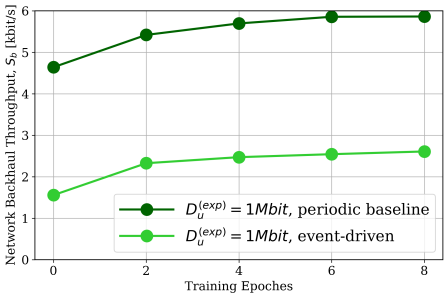

Fig. 8. Network backhaul throughput for $D_u^{(exp)}$= 1 Mbit sent periodically or event-driven.

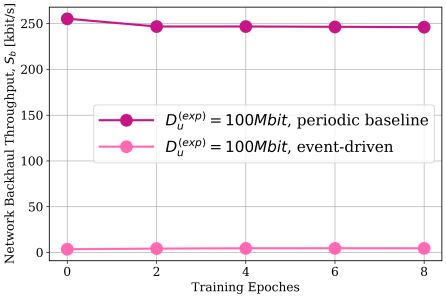

Fig. 9. Network backhaul throughput for $D_u^{(exp)}$= 100 Mbit sent periodically or event-driven.

on the overall reward. Whereas using an event-driven approach slightly reduces the cumulative reward, it significantly improves network efficiency, as explained below. Moreover, Fig. 6 confirms that the event-driven synchronization strategy does not negatively affect GUEs satisfaction. When $D_u^{(exp)} =$ 1 Mbit, satisfaction levels remain stable with only a minor 1.45% decrease, while in the $D_u^{(exp)} = 100$ Mbit case, they are practically unchanged (0.06% increase). These findings indicate that the event-driven approach maintains user satisfaction while adjusting experience synchronization dynamics. To further assess its impact on network performance, we examine how it influences throughput and signaling overhead, as presented in the following. Specifically, Fig. 7 shows that network access throughput remains stable and even improves when using the event-driven approach. This indicates that vehicles can transmit the same amount of data or more, while benefiting from a significant reduction in signaling overhead. At the same time, Fig. 8 and 9 demonstrate that network backhaul throughput decreases drastically, effectively reducing unnecessary signalling overhead. Specifically, when $D_u^{(exp)} =$ 1 Mbit, as shown in Fig. 8, the network backhaul throughput decreases from 6 kbit/s to 2.5 kbit/s, resulting in a 58% reduction in signalling overhead, while when $D_u^{(exp)} = 100$ Mbit, as presented in Fig. 9, the network backhaul throughput decreases from 245 kbit/s to 5 kbit/s, leading to a 98% reduction in signalling overhead. These results underline the substantial efficiency gains achieved through the event-driven synchronization approach, significantly minimizing unnecessary signaling and optimizing network resource usage. This balance is crucial as it ensures that network resources are used more efficiently, leading to improved overall performance.

VI. Conclusion

This paper has presented an enhanced MADRL framework for the joint optimization of UAV trajectory planning and RRM in communication networks with both UL and DL traffic. The proposed architecture introduces an event-driven synchronization mechanism for model updates and experience sharing, enabling dynamic signaling between agents and controller. Unlike traditional approaches that assume continuous or periodic communication, our method adapts to the learning relevance and network conditions, thereby reducing unnecessary overhead. The framework integrates a reward function shaped by service priorities derived from the RRM process, aligning local learning objectives with global communication goals. In parallel, the RRM incorporates communication overhead into the decision process, achieving a realistic trade-off between user traffic and control signaling. Results confirm that the event-driven synchronization strategy substantially improves communication efficiency. Notably, while access throughput improves, the backhaul signaling load is significantly reduced—by up to 98% in high-volume scenarios, without compromising user satisfaction. The results demonstrate that adaptive, experience-driven communication can significantly reduce backhaul load without compromising performance, especially under constrained backhaul capacity.

References

[1] R. Zhong et al., "Multi-Agent Reinforcement Learning in NOMA-Aided UAV Networks for Cellular Offloading," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1498–1512, 2022.

[2] X. C. Zhang et al., "Cooperative Trajectory Design of Multiple UAV Base Stations with Heterogeneous Graph Neural Networks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 1495–1509, 2023.

[3] J. Huang et al., "Joint Data Caching and Computation Offloading in UAV-Assisted Internet of Vehicles via Federated Deep Reinforcement Learning," *IEEE Transactions on Vehicular Technology*, vol. 73, pp. 17644–17656, 2024.

[4] F. Li et al., "Multi-UAV Hierarchical Intelligent Traffic Offloading Network Optimization Based on Deep Federated Learning," *IEEE Internet of Things Journal*, vol. 11, pp. 21312–21324, 2024.

[5] S. Shen et al., "Asynchronous Federated Deep Reinforcement Learning-Based Dependency Task Offloading for UAV-Assisted Vehicular Networks," *IEEE Internet of Things Journal*, vol. 11, pp. 31561–31574, 2024.

[6] G. Hu et al., "Event-Triggered Communication Network With Limited-Bandwidth Constraint for Multi-Agent Reinforcement Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 8, pp. 3966-3978, Aug. 2023, doi: 10.1109/TNNLS.2021.3121546.

[7] P. Álvarez López et al., "Microscopic Traffic Simulation using SUMO," *Proc. IEEE 21st Int. Conf. Intelligent Transportation Systems (ITSC)*, pp. 2575–2582, 2018, doi: 10.1109/ITSC.2018.8569938.

[8] L. Spampinato et al., "Joint Trajectory Design and Radio Resource Management for UAV-aided Vehicular Networks," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 2, pp. 1–14, 2024.

[9] D. Ferretti et al., "QoE and Cost-Aware Resource and Interference Management in Aerial-Terrestrial Networks for Vehicular Applications," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 8, pp. 11249–11261, 2024.

[10] 3GPP, "Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz," *3GPP TR 38.901, v16.1.0, Release 16*, Dec. 2019.