Efficient Dynamic Beamforming Activation for UAV-Enabled Vehicular Networks

Leonardo Spampinato*, Lorenzo Mario Amorosa*, Marco Skocaj*[1],
Greta Vallero†, Daniela Renga†, Chiara Buratti*

*DEI Department, University of Bologna, Bologna, Italy; †DET Department, Politecnico di Torino, Torino, Italy
Emails: {leonardo.spampinato, lorenzomario.amorosa, marco.skocaj, c.buratti}@unibo.it,
{greta.vallero, daniela.renga}@polito.it

***Abstract*—Unmanned aerial base stations (UABSs) are a promising solution for improving coverage and capacity in vehicle-to-everything (V2X) communications, particularly in dense urban areas. However, their operation is constrained by onboard energy consumption, required for both flight and communication. Beamforming, while enhancing network performance, adds to this challenge due to its energy-intensive nature. This paper proposes a sequential and hierarchical decision-making framework for UABS operations, considering trajectory planning, dynamic beamforming, and radio resource assignment (RRA). While heuristic and optimal solutions are employed for trajectory planning and RRA respectively, the beamforming model is modeled as Markov decision process (MDP) to maximize served user demand, weighted by time-varying priorities, under strict energy constraints. Leveraging a dueling double deep q-network (3DQN) algorithm that penalizes energy budget violations, an agent policy for the beamformer is then trained. Simulation results demonstrate that the proposed approach outperforms static beamforming benchmarks and closely matches an ideal step-wise oracle, achieving a balance between energy efficiency and served user demand while adapting to dynamic V2X traffic conditions.**

***Index Terms*—UAV, beamforming, energy budget, DRL**

I. Introduction

The advent of vehicle-to-everything (V2X) has enabled vehicles to communicate with their surroundings. As demand for high-reliability, low-latency V2X applications grows, introducing new communication infrastructure becomes critical, particularly in dense urban environments. Unmanned aerial base station (UABS), which is unmanned aerial vehicle (UAV) equipped with onboard base station (BS), has emerged as an effective solution to address the coverage and capacity challenges of conventional fixed terrestrial BSs. UABSs offer the unique advantage of mobility, allowing them to dynamically reposition themselves to maintain line-of-sight (LOS) wireless links with vehicles. Despite their flexibility, UABSs are constrained by small onboard batteries, which limit their available energy resources. UABS energy consumption is primarily driven by flight operations and communication systems [1], including beamforming, which is a critical component for enhancing network performance through improved signal quality and spatial multiplexing. Although effective, this technique is particularly energy-intensive, as it requires a dedicated radio frequency (RF) chain for each active beam. Efficient energy management is therefore essential to sustain UABS missions. Much of the existing literature has focused on energy optimization for UAVs through trajectory planning [2], [3]. However, for UABS-specific missions, the energy consumed by the communication system cannot be overlooked. Joint trajectory and energy optimization in aerial networks often focus on balancing downlink transmit power among users [4], [5], without considering a total energy budget for the communication system over a period of time. Moreover, they cannot be applied in uplink-dominated applications. Dynamic beamforming has been explored in terrestrial networks, such as in [6], where beam selection and transmit power were jointly optimized to improve energy efficiency while ensuring user coverage. Similarly, [7] proposed a hybrid beamforming strategy combining beam selection and precoding, though it relied on a one-to-one association between beams and users, which is impractical for UABS serving large networks. Both works, however, assume terrestrial infrastructure with unlimited energy, lacking considerations for a total energy budget. Energy-efficient beam selection under strict energy constraints for UAVs was addressed in [8], where a two-stage beamforming scheme based on a budgeted combinatorial multi-armed bandit approach was proposed to improve spectral and energy efficiency. However, it does not account for user quality of service (QoS), critical in vehicular communications.

In this paper, we propose a sequential decision making framework tailored for V2X uplink-dominated applications. The architecture decomposes the intertwined challenges of trajectory planning, dynamic beamforming, and subsequent radio resource assignment (RRA) into separate modules. We then formulate a beam activation problem to maximize served user demand while adhering to a total energy budget over the UABS's flight mission. Due to the problem's complexity in highly dynamic scenarios with limited prior knowledge, we propose a deep reinforcement learning (DRL) agent to solve it. Finally, we evaluate the trained agent's performance using heuristic and diverse trajectories, which represent the trajectory planning module, alongside an optimized RRA module. The paper is structured as follows: Sec.II introduces the system model and proposes a modular agent, in Sec.III the beamforming activation problem is formulated, Sec.IV explains solutions adopted, Sec.V presents numerical results and in Sec.VI conclusions are drawn.

[1]Marco Skocaj is now with Huawei Technologies, Munich Research Center. His work was conducted while at the University of Bologna.

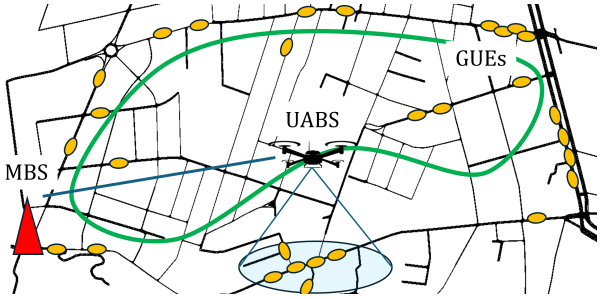

Fig. 1: UABS coordinates with a MBS to provide additional capacity and coverage to vehicular GUEs. The green line represents the trajectory for the current UABS flight mission.

II. System Model

We consider the system model depicted in Fig.1. Within a reference area, an UABS is deployed to provide additional coverage and capacity to a set $\mathcal{G}$ ($|\mathcal{G}| = G$) of vehicular ground user equipments (GUEs), in coordination with a macro base station (MBS). To do so, the UABS is assigned to a flight mission of duration $\hat{T}$, discretized into timesteps t of duration Δt, and its position (x_t, y_t) evolves over time depending on the flight trajectory, while its altitude h is kept constant. The UABS features a multi-antenna system that supports hybrid beamforming at the carrier frequency f_c. To enhance spatial diversity and antenna gain, the UABS dynamically selects a beam configuration from an available set. Independently of the specific beam configuration, the UABS always covers an area of fixed size. Moreover, each beam configuration involves activating a specific number of RF chains, impacting the energy consumption and duration of the battery dedicated to the communication system.

GUEs act as distributed agents in an extended sensing application, sharing data from local sensors and cameras towards a remote server for further processing. This generates periodic transmission of uplink packets of fixed size D every Δt and within service windows of fixed duration T_{w}. To meet continuity of service requirements, a RRA algorithm runs at the MBS, maximizing the number of served users weighted by a service priority term $p_{g,t}$. This term is initialized at the beginning of each new service window and increments whenever g is served and successfully uploads its packet at time step t. To enable efficient network operations, the UABS must dynamically change their beamforming pattern to adapt to the varying traffic demand of GUEs and their coverage while minimizing the overall energy consumption to improve battery life. In this work, we build upon our prior findings [9] by addressing the dynamic beamforming selection under an energy budget constraint.

A. Sequential Decision Making Model

Trajectory planning, beamforming, and cooperative terrestrial-aerial RRA are tightly intertwined. The optimization of one significantly impacts the others, making joint optimization particularly challenging. To this end, we introduce the decision making framework represented in Fig.2, where

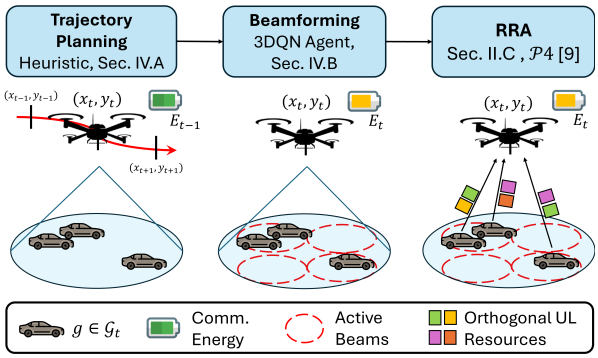

Fig. 2: Sequential decision-making framework with three modules ensuring continuous UL service to GUEs. This paper focuses on the beamforming module, which dynamically activates beams under an energy budget constraint.

multiple modules interact sequentially and hierarchically to ensure continuous service to users. In particular:

- The *trajectory planning module* determines the UABS position at each time step t. It plays a crucial role in the definition of $\mathcal{G}_t \subseteq \mathcal{G}$, which represents the subset of GUEs currently covered by the UABS.
- The *beamforming module* is responsible for adapting the beamforming configuration to the dynamic user demand while respecting an energy budget constraint over the flight mission. The number of active beams at time instant t is referred to as b_t.
- Finally, the *RRA module* perceives the decisions of previous modules and it assigns radio resources, split among MBS and UABS, to meet GUEs uplink demand.

The first two modules are physically implemented at the UABS, while RRA is executed on a mobile edge computing (MEC) server, also considering the MBS user traffic. The exchange of information and data between MEC server, UAV, and MBSs is supported via backhaul communication. This work focuses on the design and optimization of the beamforming module while assuming heuristics for the trajectory planning, as detailed in Sec.IV-A, and the RRA algorithm briefly recalled in Sec. II-C.

B. Channel Model and Beam Gain

The channel model follows the urban macro (UMa) specifications as described in 3GPP TR 38.901 [10], distinguishing between line-of-sight (LoS) and non-LoS (NLoS) conditions based on the probability ρ_L of the link being in LoS. The link signal-to-noise ratio (SNR) in dB is calculated as $\mathrm{SNR} = P_{\mathrm{tx}} + G_{\mathrm{tx}} + G_{\mathrm{rx}} - L_{\mathrm{path}} - P_{\mathrm{noise}}$, where $P_{\mathrm{tx}}, G_{\mathrm{tx}}$, and G_{rx} denote transmit power transmitter and receiver antenna gains respectively, L_{path} is the path loss and P_{noise} is the noise power. For beamforming, the UABS coverage is defined by a vertical field of view angle ϕ, giving a solid angle $\Phi \approx 2\pi\left(1 - \cos\left(\frac{\phi}{2}\right)\right)$. Assuming b_t circular and non-overlapping beams active within this overall solid angle, the receiver gain can be expressed as in [11]:

$$G_{\mathrm{rx}}(b_t) = 10\log_{10}\frac{41000}{\left(\frac{\Phi}{b_t}\frac{360}{2\pi}\right)^2} \tag{1}$$

Notably, as b_t increases, beamwidth narrows, improving receiving gain.

C. Radio Resource Assignment

The RRA module tackles the challenge of user assignment and resource allocation. Specifically, it determines whether each user should connect to the UABS or to the MBS, while allocating resources to maximize the number of served users based on their priority. The objective is to ensure continuous service for users, following the approach solving problem $\mathcal{P}4$ in [9]. At each time step t, the RRA module allocates up to W_{m} and W_{u} resources from the shared bandwidth B_{sys} to users associated to the MBS and UABS, respectively. Resources span over time and frequency, with the UABS further exploiting spatial multiplexing and operating in full-reuse over each b_t active beams, thus it holds:

$$W_{\mathrm{u}}(b_t) = W_{\mathrm{m}} \cdot b_t = \frac{B_{\mathrm{sys}}}{N_{\mathrm{sub}}\Delta f} \cdot \frac{\Delta t}{T_{\mathrm{slot}}} \cdot b_t, \tag{2}$$

where time is divided in time slots of duration T_{slot} and frequency into groups of N_{sub} consecutive sub-carriers each spaced Δf. Parameters $T_{\mathrm{slot}}, N_{\mathrm{sub}}, \Delta f$ are taken from [9] following 5G numerology. The result of the algorithm at time step t is a set of binary values $\psi_{g,t}(b_t)$, which is 1 if user g associated to the UABS successfully uploaded its packet, 0 otherwise. Each priority term $p_{g,t}$ evolves as $p_{g,t+1} = p_{g,t} + \psi_{g,t}(b_t)$.

D. Energy Consumption Model

In this work, we focus on a strategy that dynamically adjusts the number of active beams at the UABS to reduce energy consumption while adhering to user demand. For our analysis, we adopt a simplified linear energy consumption model based on the number of active RF chains. While more comprehensive models also account for factors like load-dependent components [12], our simplified model highlights the energy savings attributable to the proposed dynamic beam strategy.

Let b_t represent the number of active beams at time step t, each corresponding to a dedicated RF chain. The energy consumption e_t of the BS on board the UABS is given by:

$$e_t = b_t \cdot P_{\mathrm{RF}} \cdot \Delta t, \tag{3}$$

where P_{RF} is the power consumption of a single RF chain. While specific values can be found in [13], without loss of generality we assume $P_{\mathrm{RF}} = 1$ W. The available communication energy E_t is updated as $E_{t+1} = E_t - e_t$, with its initial value E_0 corresponding to the total energy budget available for the communication system throughout the flight mission.

III. Problem Formulation

The dynamic beams activation problem under budget constraint can be framed as an optimization problem, that is

$$\text{maximize} \quad \hat{\rho}(b_0, b_1, \ldots, b_{\hat{T}}) \tag{4}$$

$$\text{subject to} \quad \sum_{i=0}^{\hat{T}} b_i \leq B \tag{5}$$

$$b_i \geq 0 \quad \text{for all} \quad i = 0, 1, \ldots, \hat{T}\ ,$$

where $b_0, \ldots, b_{\hat{T}}$ denotes the number of beams activated at each time step, $\hat{\rho}$ is the objective function we aim at maximizing, and B denotes a budget constraint. Here, B represents a maximum number of beams that can be activated during the whole flight mission due to energy budget. The objective function is expressed as:

$$\hat{\rho}(b_0, \ldots, b_{\hat{T}}) = \sum_{t=0}^{\hat{T}} \rho_t(b_t), \tag{6}$$

with

$$\rho_t(b_t) = \underbrace{\sum_{g\in\mathcal{G}_t} \psi_{g,t}(b_t)\cdot p_{g,t}\cdot D}_{\text{served demand}} - \underbrace{\sum_{g\in\mathcal{G}_t} p_{g,t}\cdot D}_{O_t,\ \text{offered demand}} \quad . \tag{7}$$

At time step t, $\mathcal{G}_t \subseteq \mathcal{G}$ represents the set of users covered by the UABS, which depends on its current position, and the GUEs' movement. $\rho_t(b_t)$ denotes the difference between the served demand and the offered demand O_t, both weighted by the users' priority $p_{g,t}$ According to Eq.(6), $\max\hat{\rho} = 0$ and it occurs when all users covered by the UABS in a flight mission are successfully served. However, this is not always feasible due to the budget constraint. In this case, $\hat{\rho}$ is negative and it represents the total amount of weighted user demand not served. This may result from either a situation where the UABS does not activate enough beams for spatial multiplexing in traffic hotspots or an overly aggressive beam activation leading to premature communication battery depletion. As such, it is essential for the UABS to learn to resort to a higher beam count only when strictly necessary, ensuring enough energy to guarantee efficient network operations throughout the whole flight duration.

Without a priori knowledge of the vehicle's trajectory or the resulting RRA behavior, a planned optimal solution for the whole flight is intractable. We therefore adopt a temporal difference (TD) learning approach, modeling the beam activation problem as a Markov decision process (MDP) defined by a *state space*, *action space*, and *reward signal*.

A. Markov Decision Process

The state space $\mathcal{S}$ defines all possible states that the agent can observe, where $s_t = \left(\mathbf{z_t}, E_t, \hat{T} - t\right)$ denotes the state at time step t. Here, $\mathbf{z_t}$ is a vector of length Z^2, representing a a $Z \times Z$ discretized grid of the area underneath the UABS. Elements $z_{t,i}$ corresponds to the sum of the priority values $p_{g,t}$ of all GUEs g located within the i-th region. This

information reflects a coarse and local distribution of users covered by the UABS. Furthermore, E_t indicates the energy available for communication at time t, while $\hat{T}-t$ represents the remaining flight time. [2]

At each time step t the UABS can select a beamforming pattern a_t out of a discrete set $\mathcal{A} = \{0, 1, 2\text{x}2, 3\text{x}3, 4\text{x}4\}$. Action '0' corresponds to turning off the communication system to save energy, action '1' generates a single beam for the entire covered area, while action '$n$xn' generates n^2 beams arranged in a grid layout. Each action a_t determines the total number of simultaneous beams b_t and the corresponding number of active RF chains. Notably, the number of beams changes the receiver gain (Eq.(1)), enables spatial multiplexing of resources (Eq.(2)), and influences energy consumption (Eq.(3)), while the covered area remains unchanged.

The reward at time step t is defined as:

$$r_t = \rho_t(b_t) - C(b_0, b_1, \dots, b_t), \tag{8}$$

where $\rho_t(b_t)$ is expressed in Eq.(7) and $C(b_0, b_1, \dots b_t)$ is a penalty function introduced to take into account the constraint in (5). It is defined as:

$$C(b_0, b_1, \dots b_t) = \begin{cases} b_t^{\alpha} & \text{if } \sum_{i=0}^{t} b_i \leq B \quad \text{(a)} \\ (\hat{T}-t)^{\beta} & \text{otherwise .} \quad \text{(b)} \end{cases} \tag{9}$$

Here, Eq.(9a) aims to encourage the UABS to serve available user demand with the fewest active beams possible and Eq.(9b) represents the remaining flight duration at time step t once constraint in (5) are violated and the UABS service ends. Indeed, when the energy budget is not respected, the drone may need to end its route earlier to recharge. While such behavior reflects inefficient network operations, in practice, it may be justified in cases of extreme traffic congestion to maximize overall network service. These penalties can be further tuned via the hyper-parameters α and β.

IV. Algorithms

A. Trajectory Design

Following the modular approach, each state s_t perceived by the beamforming agent depends on the trajectories followed by the drone in each flight throughout training. Thus, providing trajectories that enrich the diversity of situations the beamformer can encounter is fundamental to train a well-behaving agent. To this end, three types of trajectories are defined:

- *hovering*, the UABS remains stationary throughout the flight, with its position randomly chosen from a limited set. These positions include high-density hotspots, like traffic lights, or less crowded areas;
- *loop*, the UABS moves along a predefined close loop within the considered area. The starting point along the loop and the drone speed are selected randomly, creating diverse position sequences across trajectories. An example of loop trajectory is represented in Fig.1.
- *trailing*, the UABS begins above a randomly chosen GUE and follows its movements. Once the user completes its path, the UABS switches to another nearby user, repeating this process until the flight duration $\hat{T}$ is reached.

[2]The linear model in Eq. (3) is not a limitation, as the agent relies only on the current battery level E_t, not on a specific evolution model. Alternative communication energy consumption models may be used.

B. Beamformer Training

The beamforming agent is trained using the dueling double deep Q-network (3DQN) algorithm, aiming at estimating the optimal Q-values $Q_{\pi^*}(s,a)$ provided by the optimal policy π^*. Each $Q_{\pi^*}(s,a)$ represents the expected discounted cumulative reward obtained by an optimal agent taking action a while in a state s. They can be expressed iteratively following the Bellman equality as $Q_{\pi^*}(s,a) \approx r(s,a) + \gamma \max_a Q_{\pi^*}(s',a)$, where $r(s,a)$ is the reward obtained for taking action a in state s, s' is the subsequent state after performing action a in state s and γ is the discount factor, balancing the importance between immediate and future rewards. To do so, 3DQN employs two neural networks, *online* and *target*, whose parameters are $\theta = \eta \cup \mu$ and $\theta^- = \eta^- \cup \mu^-$, respectively. The *target* is a delayed copy of *online* and it is introduced to stabilize training. A dueling architecture is used, thus *online* computes the value of $Q(s,a|\theta)$ by separately estimating the state-value $V(s|\eta)$ and the action-advantage $A(s,a|\mu)$, with $Q(s,a|\theta) = V(s|\eta) + \left(A(s,a|\mu) - \frac{1}{|A|}\sum_{a'} A(s,a'|\mu)\right)$. Similarly, the *target* computes $Q(s,a|\theta^-)$. This allows the agent to differentiate between the importance of states versus the relative value of each action.

At the i-th training step, a batch $\mathcal{K}$ of k experiences $\{s,a,s',r\}$ is randomly drawn from a buffer replay of dimension K. Then, the *online* network parameters θ_i are updated applying gradient descent on the loss function:

$$L(\theta_i) = \mathbb{E}_{\{s,a,s',r\}\sim\mathcal{K}}\left[\left(y - Q(s,a|\theta_i)\right)^2\right] \tag{10}$$

with target $y = r + \gamma Q(s', \arg\max_{a'} Q(s',a'|\theta_i)|\theta_i^-)$, where θ_i^- denotes the *target* parameters at the i-th training step. The calculation of y follows the Double Q-learning algorithm to reduce target overestimation. Every l training step, the *online* parameters are copied to the *target* network.

V. Numerical results

In this section, we analyze the numerical results obtained by training the proposed beamforming agent and comparing its performance against benchmarks. Two traffic scenarios are considered: *Low Traffic* and *High Traffic*. The *Low Traffic* scenario features a limited volume of traffic data, with fewer GUEs and lower demand D. In contrast, the *High Traffic* scenario doubles the number of users and their demand, resulting in a fourfold increase in the total data traffic potentially offered to the UABS.

The agent was trained over N episodes using a decaying ϵ-*greedy* policy to collect experiences $\{s,a,r,s'\}$. Specifically, at each interaction with the environment, the agent selects a random action $a \in \mathcal{A}$ with probability ϵ. Otherwise, it chooses the action with the highest Q-value under

TABLE I: Simulation Parameters

$\hat{T}$	270 s	Δt	1 s	B_{sys}	50 MHz
f_c	30 GHz	ϕ	100°	E_0	1350 J
h	100 m	Z	4	$\lvert\mathcal{G}\rvert$	[100,200]
P_{tx}	1 dBm	T_w	10 s	D	[1,2] Mbit
G_{tx}	1 dB	N	2000	P_{noise}	-104 dBm
K	50000	k	128	(α, β)	(0.3, 1.2)
γ	0.8	l	500	(η, μ)	(64x32x1, 64x32x5)

the current state s_t and policy parameters θ, that is $a = \arg\max_a Q(a, s|\theta)$. This method balances exploration of new strategies with the exploitation of the current policy. To encourage exploitation over time, the parameter ϵ is linearly decayed as learning progresses. At the beginning of each training episode, a UABS trajectory $[(x_0, y_0), \ldots, (x_{\hat{T}}, y_{\hat{T}})]$ is randomly generated as described in IV-A, whereas GUEs paths are generated using simulation urban mobility (SUMO), an open-source simulation software designed to handle large road networks. Simulation parameters are reported in Table I, with key differences between *Low Traffic* and *High Traffic* scenarios noted in square brackets. Notably, the total energy budget is set to $E_0 = \hat{T} \cdot 5$, so that the beamformer can complete a flight mission if the average number of active beams in an episode is below 5.

A. Benchmarks and Performance Metric

Two benchmarks are here introduced, namely:

- *Fixed n*, which is a beamformer that always activates n beams during its flight mission with $n \in \{1, 4, 9, 16\}$.
- *Energy Efficient Oracle (EEO)*, which is an ideal agent that evaluates all beam configurations and selects the one maximizing the instantaneous reward r_t. Thus, it chooses either the smallest b_t that meets user demand due to the beam penalty factor or the largest one when demand is too high. However, a different channel realization may occur when performing the action chosen. Due to the short time scale and high mobility of GUEs and the UABS, this solution cannot be adopted in a real-time system, thus it is proposed as a stepwise energy efficient oracle.

Comparison among benchmarks is studied in terms of episode return $R_n = \sum_{t=1}^{T_n} r_t$, coinciding with the sum of rewards r_t obtained in the n-th episode ending at time step T_n. As highlighted in Sec.III-A, the return accounts for the overall served traffic demand, the sequence of beams activated over time, and the fulfillment of the energy budget constraint. Moreover, we present the average number of beams activated, along with the 25-th and 75-th percentile, in each episode to show how our agent adapts to different traffic conditions.

B. Learning Performance

Fig. 3 shows R_n for the proposed agent and benchmarks, considering training over N episodes under the two traffic settings. Early in training, the 3DQN agent lacks a suitable policy, selecting the number of beams randomly within an

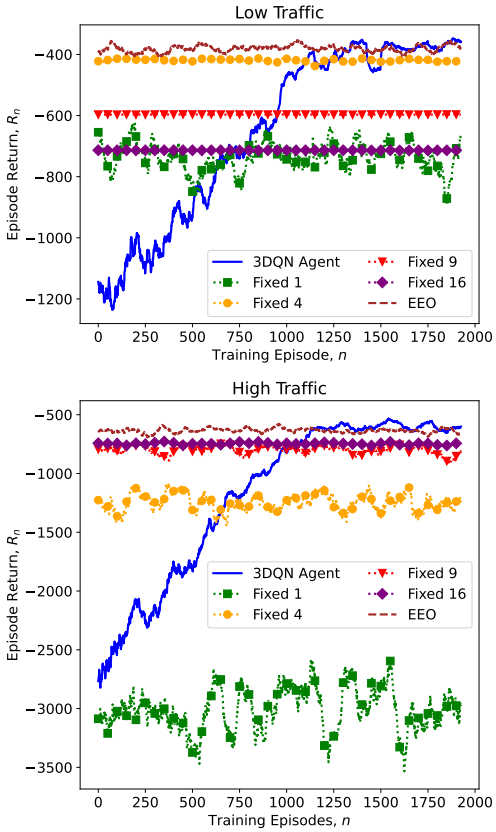

Fig. 3: Return comparison between trained agent and benchmarks.

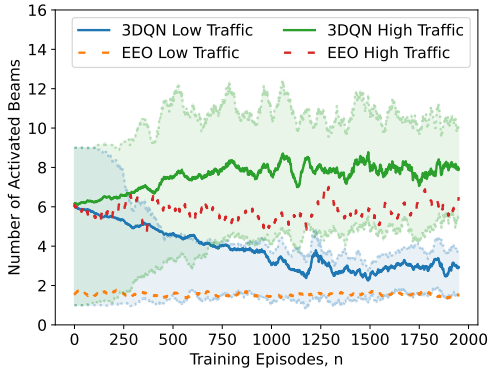

Fig. 4: Average number of activated beams per episode, with a shaded area spanning between the 25th to 75th percentiles. Dashed lines indicate the average for the EEO benchmark.

episode. This results in frequent penalties due to either insufficient beams to meet the user demand or an excessive number wasting energy and draining the communication battery prematurely. However, as training progresses the agent leans an improved policy, maximizing the obtainable return and achieving performance comparable to the ideal oracle in both traffic settings. The performance of *Fixed 1* is inadequate for both low and high traffic scenarios, indicating that the reception gain G_{rx} is insufficient when no directional beams are used. Although *Fixed 4* performs well with low traffic conditions, it is less energy-efficient than the proposed agent, which surpasses it in the second half of training. However,

in case of high traffic, the spatial resource multiplexing is insufficient, thus a drop in its performance is observed. *Fixed 9* and *Fixed 16* waste significant energy during low traffic, incurring penalties from premature battery depletion despite meeting user demand during the UABS flight. Under high traffic, their returns are comparable: while 16 beams deplete the battery faster than 9, they serve more user demand within that shorter duration. In contrast to *Fixed n* benchmarks, the proposed agent, after adequate training, provides a balanced solution that maximizes served user demand and enhances communication energy efficiency.

Fig.4 reports the evolution of the number of activated beams throughout training for the proposed agent and the average number of beams for the ideal oracle for comparison. Initially, the 3DQN policy activates an average of 6 beams in each episode regardless of traffic type. Over time, however, it adapts to the specific traffic conditions. Under low traffic conditions, the agent more likely selects a low number of beams, averaging around 3, closely aligning with the oracle's average, while rarely selecting 9 or 16 beams. On the contrary, under high traffic conditions, the number of beams selected exhibits greater variability, as indicated by a wider gap between the 25-th and 75-th percentiles. This behavior arises from substantial fluctuations in user demand and the agent's efforts to maintain energy-efficient communication. Notably, while the ideal oracle averages 6 beams, the proposed agent stabilizes at approximately 8 beams. This difference comes from the increased complexity of the optimization problem under higher demand. While the oracle can exhaustively evaluate all configurations to identify the most energy-efficient solution, the 3DQN agent must activate more beams to ensure higher reliability in the face of uncertain conditions due to channel variability and unknown future demand.

Without loss of generality, Fig.5 shows the inference of the agent learned policy for a representative episode under high traffic conditions. As weighted user demand evolves over time, the agent dynamically activates the appropriate number of beams b_t, enabling the UABS to serve users while managing its energy budget. Notably, past UABS's actions influence the offered weighted demand O_t, as serving user g at timestep $t-1$ increments its priority term $p_{g,t}$, possibly leading to highly negative rewards, as it can be seen in the figure around time step 100 and 175. At the final time step T, the UABS depletes its communication battery and incurs in penalty $C(b_0, \ldots, b_T)$ for failing to adhere to the energy budget constraint.

VI. Conclusions

In this paper, we proposed a modular and hierarchical approach to deal with the interdependent problem of trajectory planning and dynamic beamforming activation for UABSs in vehicular networks. Focusing on the latter, the dynamic beam activation problem is formulated to maximize the served user demand weighted by time-varying priorities while adhering to an energy budget constraint over the flight trajectory. However, due to problem complexity, we framed it as a MDP

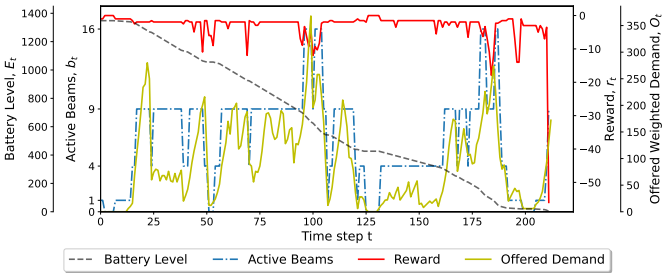

Fig. 5: Agent policy tested in an evaluation episode.

and solved it via TD learning using 3DQN algorithm. Numerical results highlight the benefit of dynamic beamforming compared to benchmarks considering fixed beam activation sequences, and achieving comparable performance to those of an ideal step-wise oracle. The development of end-to-end learning strategies to improve modules coordination between trajectory planning and beamforming - for instance, based on the use of curriculum learning - is left for future research.

VII. Acknowledgment

This work has been carried out in the framework of the CNIT WiLab-Huawei Joint Innovation Center and also partially supported by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP F83C22001690001, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

References

[1] H. Yan *et al.*, "Optimum battery weight for maximizing available energy in uav-enabled wireless communications," *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1410–1413, 2021.

[2] Y. Zeng and R. Zhang, "Energy-efficient uav communication with trajectory optimization," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, 2017.

[3] L. Ding, D. Zhao, H. Ma, H. Wang, and L. Liu, "Energy-efficient min-max planning of heterogeneous tasks with multiple uavs," in *2018 IEEE 24th Int. Conf. on Parallel and Distrib. Systems*, 2018, pp. 339–346.

[4] L. Guo, S. Zhang, S. Shi, and X. Ji, "Joint optimization for energy efficient uav relaying with multiple user pairs," in *2023 IEEE 23rd Int. Conf. on Commun. Tech.*, 2023, pp. 1620–1626.

[5] J. Wang *et al.*, "Joint resource allocation and trajectory design for energy-efficient uav assisted networks with user fairness guarantee," *IEEE Internet of Things Journal*, vol. 11, no. 13, pp. 23 835–23 849, 2024.

[6] Y. Dantas *et al.*, "Beam selection for energy-efficient mmwave network using advantage actor critic learning," in *IEEE Int. Conf. on Commun.*, 2023, pp. 5285–5290.

[7] Y. Liu *et al.*, "Joint beam selection and precoding based on differential evolution for millimeter-wave massive mimo systems," in *2022 IEEE Int. Conf. on Acoust., Speech and Signal Processing*, 2022, pp. 5318–5322.

[8] G. Wang *et al.*, "Spectral efficient two-stage beamforming for uav mimo under energy constraint: A budgeted combinatorial mab-based approach," *IEEE Commun. Lett.*, vol. 28, no. 8, pp. 1904–1908, 2024.

[9] L. Spampinato, D. Ferretti, C. Buratti, and R. Marini, "Joint trajectory design and radio resource management for uav-aided vehicular networks," *IEEE Trans. on Veh. Technology*, pp. 1–14, 2024.

[10] 3GPP, "Technical specification group radio access network; study on channel model for frequencies from 0.5 to 100 GHz," *TR 38 901 version 16.1.0*, Dec. 2019.

[11] V. Salvia, *Antenna and Wave Propagation*. Laxmi Publications, 2007.

[12] D. López-Pérez *et al.*, "A survey on 5g radio access network energy efficiency: Massive mimo, lean carrier design, sleep modes, and machine learning," *IEEE Commun. Surveys & Tutorials*, vol. 24, no. 1, pp. 653–697, 2022.

[13] G. Auer *et al.*, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, 2011.