IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 November 2025, accepted 26 December 2025, date of publication 30 December 2025,
date of current version 5 January 2026.

Digital Object Identifier 10.1109/ACCESS.2025.3649591

RESEARCH ARTICLE

Multi-Agent Deep Reinforcement Learning-Based RIS-Aided UAV Communications

YANG CHEN, HANIEH AHMADI, AND SABA AL-RUBAYE, (Life Senior Member, IEEE)
Faculty of Engineering and Applied Sciences, Cranfield University, MK43 0AL Milton Keynes, U.K.

Corresponding author: Yang Chen (yang.chen@cranfield.ac.uk)

This work was supported by Engineering and Physical Sciences Research Council (EPSRC) Communications Hub for Empowering Distributed Cloud Computing Applications and Research (CHEDDAR) Project under Grant EP/X040518/1 and Grant EP/Y037421/1.

ABSTRACT However, traditional model-based phase-shift optimization is highly sensitive to imperfect CSI and becomes computationally prohibitive for large UPA-based RIS, while existing model-free solutions relying on single-agent DRL struggle with the exponentially growing action space. This paper presents a scalable multi-agent deep Q-network (MADQN)–based RIS controller designed for large-scale UAV–RIS systems under realistic channel dynamics. An end-to-end channel inference architecture is first introduced to mitigate CSI imperfection and reconstruct stable channel representations under UAV mobility. A multi-objective formulation is then developed to jointly optimize sum rate, energy consumption, and control latency, which is transformed into a multi-agent Markov decision process (MMDP) compatible with quantized RIS hardware. Building on this formulation, a dual-agent RIS controller is proposed, in which row and column agents cooperatively determine the quantized phase configuration of a large UPA RIS. Extensive simulations demonstrate that the proposed framework significantly outperforms benchmark schemes, showing acceptable robustness against varying Rician factor SNRs, UAV densities, and RIS sizes. These results confirm that the proposed MADQN-based controller is a promising and practical solution for scalable RIS control in large-scale multi-UAV communication systems.

INDEX TERMS Uncrewed aerial vehicles, reconfigurable intelligent surfaces, uplink wireless communication, multi-agent reinforcement learning.

I. INTRODUCTION

The rapid growth of wireless communication demands has exposed substantial limitations of traditional terrestrial base stations (BS) [1], [2], especially as moving towards the sixth generation (6G) mobile networks [3], [4]. Owing to static properties of the beam pattern in the long term, fixed BS struggled to provide ubiquitous coverage and adapt to dynamic user distribution or service requirements. As a result, their propagation usually suffered from signal blind spots and infrastructure failure, which significantly reduced their connectivity to remote and hard-to-reach areas. In contrast, uncrewed aerial vehicles (UAVs) offer BS with unparalleled flexibility and efficiency by dynamically adjusting the positions of BS in three-dimensional (3D) space [5]. In that case, UAVs will benefit wireless communications from extending network coverage [6], eliminating dead zones [7], and adapting to evolving network traffic in real-time [8]. This adaptability enables the construction of dynamic and more accurate channel models [9], supporting robustness and resilience of network deployments. Furthermore, UAVs play a critical role in scenarios such as disaster recovery and rural communications [10], where rapid and adaptive network deployment is essential to restore multi-user communications. Thus, UAV communication has become a promising solution to addresses the inherent shortcomings of terrestrial base stations and paves the way for more adaptive, reliable, and wide-reaching communication services in the future.

Nevertheless, the classic UAV communication has to forward 6G signals from ground users with high transmit power, which leads to a great waste of limited UAV energy [11], [12]. In addition, buildings and other obstacles are prone to block the connectivity between BS and the UAV in both urban area and distant area, from which the

The associate editor coordinating the review of this manuscript and approving it for publication was Ronald Chang.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

network failure will reduce the information flow from UAV to BS [13]. Fortunately, Reconfigurable Intelligent Surface (RIS) emerges as a groundbreaking innovation to release communication potential without wasting transmit power. RIS is a predominantly passive and economical technology that modulates wireless environments by dynamically manipulating the magnitude and phase shift of RIS, thus transforming the wireless medium into a programmable entity [14]. Compared to conventional direct UAV-BS scenarios, RIS benefits beamforming from minimal power consumption and adaptable configuration [15]. Practical applications of RIS have shown considerable advantages. For example, a 2430-unit cell RIS can increase the received signal strength by 15dB through avoiding indoor obstruction in the Sub-6 GHz band, highlighting their advantage to mitigate blind spots in 5G/6G networks [16]. Advanced genetic algorithms exhibited improved signal coverage and propagation efficiency of RIS in complex urban environments, which is benefited from multi-beam and wide-beam reflection patterns [17]. In addition, RIS has demonstrated efficient and enhanced channel capacity, reduced interference, and released signal degradation from obstructions or multipath fading [18], the application of RIS has involved several communication systems, such as indoor communications, vehicle networks, and millimeter-wave communications [19]. As an essential element of 6G networks, RIS aims to tackle issues such as spectrum scarcity, energy inefficiency, and inadequate connection in remote regions [20].

The amalgamation of RIS with UAVs presents a promising strategy to increase the coverage, reliability, and resilience of the wireless networks system [21]. UAV-assisted communication systems offer a flexible framework for RIS implementations where conventional infrastructure is limited to installation and deployment, such as wireless communications in rural or disaster-affected regions [15]. Integrating UAVs with RIS enables dynamic beamforming, trajectory optimization, and improved spectral efficiency [14], [22]. UAV-mounted RIS panels have proven effective in alleviating obstructions in mmWave communications and improving downlink rates in vehicular and urban settings [19], [23]. Moreover, current research indicated that the joint optimization of UAV trajectories and RIS phase shifts can significantly enhance system performance regarding energy efficiency and achievable rates [24]. These outcomes highlighted the capability of RIS-UAV systems to facilitate next-generation applications with enhanced dependability and agility.

A. RELATED WORK

Generally, the deployment of RIS is divided into two types in UAV communications, including airborne UAV-terrestrial RIS and aerial RIS [25]. Early studies focus on the investigation of aerial RIS, where the UAV-RIS link can be ignored to simplify communication scenarios. To assist terrestrial communication, RIS in [26] was mounted on UAV platform for smart reflection, thereby providing a direct optimal controller for the connectivity of UAV-Ground communications. However, they only designed framework and architecture for initial investigation instead of detailed algorithm and theoretical analyses. To achieve line-of-sight (LoS) and full-area coverage, an energy-efficient aerial backhaul system is developed in [27], where RIS is mounted on a high-altitude aerial platform to release a sudden increase of communications demands at peak times of urban environments. Nevertheless, the optimization of RIS was so sensitive to the mobility of high-altitude aerial platform. In addition, the implementation of aerial platform also hindered the direct connection between UAV and BS, which reduced the reliability of UAV-BS link. Considering the flexibility of UAV-mounted RIS, Aung et al. [28] integrated UAV with simultaneously transmitting and reflecting (STAR)-RIS to facilitate communication between IoT devices and a mobile edge computing (MEC) server. However, the performance and efficiency of UAV-mounted RIS is highly dependent on the battery limitation of UAV, so aerial RIS is currently not suitable for long-term communication tasks.

Instead, considering the deployment of RIS on the ground, airborne UAV-terrestrial RIS can benefit both Users (UEs)-UAV and UAV-BS links [29], [30]. To reduce the power consumption of UAV jitters and release hardware constraints, Adam et al. [31] jointly optimized passive beamforming, active beamforming, and UAV trajectory with decoupled non-convex objectives. Moreover, a hybrid semi-unfolding deep neural network (HSUDNN) was developed to restore imperfect channel state information (CSI) for RIS-UAV links in real-time. However, the proposed S-procedure and convex-concave technique can only achieve suboptimal channel capacity due to the limitation of linearization to large-scale and high dimension objectives. To enable wireless power transfer (WPT) systems, RIS in [32] is deployed on the buildings to enhance the energy efficiency of UAV-enabled WPT systems. The successive convex approximation (SCA) is further developed to jointly optimize the trajectory and hovering time of UAV and reflection coefficient of RIS for energy minimization. However, SCA is driven with high computation complexity, limited to dynamic radio environments, particularly in UAV-BS link. For fair resource allocation of UAV-enabled wireless network, a hybrid active-passive RIS is developed in [33] to maximize the minimum rate among users with block coordinate ascend and SCA. Regardless, the aforementioned mode-based optimization is not strong enough to deal with high dimension system coupling between internal variables, such as UAV trajectory and reflection coefficient.

Owing to the nature of gradient descent, model-free algorithm has been proposed to solve the non-convex optimization without sacrificing precision [34], especially for RIS control. To maximize the data transfer rate of the UAV in time, the Double Deep Q-Network (DDQN) and Deep Deterministic Policy Gradient (DDPG) algorithms are adopted in [14], such that they can enable joint optimization for the 3D

space of the UAV and the phase shift of the RIS. However, DDQN and DDPG can only deal with limited-scale reflective coefficient and CSI instead of large-scale deployment of RIS or UAV. To avoid propagation block in dense urban environments, Adhikari et al. [35] constructed a RIS-assisted UAV network by joint optimization for 3D trajectory of UAV and phase shift of RIS. For online reduction of computational complexity, Twin-delayed Deep Deterministic Policy Gradient (TD3) was leveraged in RIS controller to achieve higher energy efficiency and throughput for RIS-assisted UAV network. Nevertheless, considering the high training overhead of TD3, the proposed framework was lack of sufficient adaptability. For further solving the blockage and energy-hungry, federated learning is designed by [20] along with long short-term memory to achieve fair resource allocation in RIS controller. However, as an explicit distributed algorithm, federated learning was more suitable for the optimization of large group of physical entities, including server and clients. As a result, it is so challenging to enable implicit distributed control in RIS for optimal UAV-to-Ground networks.

Beyond single-agent DRL, multi-agent reinforcement learning (MARL) has recently been explored to address the scalability and coordination challenges inherent to RIS-assisted wireless systems. In RIS-aided cell-free massive MIMO, a heterogeneous-agent TRPO framework [36] was introduced to jointly optimize precoding at access points and RIS reflection coefficients with only local CSI, enabling distributed decision-making while limiting backhaul usage. For RIS-assisted NR-U and Wi-Fi coexistence, MARL has been employed to coordinate multiple cells and multiple RIS panels to maximize system-wide sum-rate under unlicensed spectrum sharing [37]. In wireless-powered RIS networks, hierarchical MARL with attention mechanisms [38] has been proposed to jointly learn user access strategies, RIS beamforming, and BS transmission parameters to improve throughput under highly dynamic power constraints. MARL has also been applied to RIS-enabled codebook design [39], where agents cooperatively adjust BS beams and RIS phase patterns using only binary received signal strength, thereby mitigating CSI acquisition overhead in rapidly changing environments. In AAV–RIS-assisted ISAC systems [40], MARL has been used to jointly optimize AAV trajectory, RIS phase control, and beamforming to balance sensing and communication in dynamic scenarios, further demonstrating the flexibility of MARL in aerial RIS applications. Although these MARL-based approaches demonstrate promising scalability and distributed control capabilities, they mainly address network-level cooperation rather than the fine-grained, quantized, row–column phase coordination required for large UPA RIS surfaces in UAV-assisted links. This gap further motivates the need for a tailored MARL framework suitable for implicit distributed control within the RIS itself under hardware quantization and UAV-induced channel dynamics.

B. MOTIVATION AND CONTRIBUTIONS

Despite the significant progress achieved by recent studies in the domains of aerial RIS, airborne UAV–terrestrial RIS integration, and model-free RIS control, several key challenges remain unaddressed, especially under quantized UPA RIS architectures and UAV-assisted dynamic radio environments when scaling up to large-scaled RIS elements:

- Firstly, most prior works rely on model-based optimization tools such as alternating optimization (AO) and successive convex approximation (SCA), which assume accurate and tractable channel state information (CSI). However, CSI acquisition in UAV–RIS systems is highly imperfect due to mobility, rapidly varying UAV–RIS–UE channel, and the large number of RIS elements that must be estimated jointly [41]. These imperfections propagate through the coupled optimization of UAV position, beamforming vectors, and RIS phase shifts, making the resulting non-convex problem extremely sensitive to estimation errors. Linearization and convex approximation techniques are often used to address this complexity, which introduces further mismatch and leads to suboptimal performance. Moreover, even for moderate UPA RIS sizes, the computational overhead of these algorithms is too high for real-time UAV operation, where channel conditions evolve on a fast timescale.
- Secondly, most existing DRL-based RIS schemes rely on single-agent architectures, which become inadequate for UPA RIS structures where the control must be decomposed into coordinated row and column decisions. The resulting expansion of the joint state–action space leads to slower convergence, limited exploration efficiency, and degraded performance under dynamic channel conditions. Although multi-agent DRL has been applied in other wireless systems [42], these approaches do not fully address the unique constraints of RIS control, including quantized phase-shift actions, row–column coordination for UPA hardware, coherence requirements for reflection patterns, and the stringent latency constraints imposed by UAV-assisted operation. Consequently, there remains a need for a multi-agent, quantization-aware, and latency-efficient learning framework specifically designed to support cooperative control of UPA RIS in UAV-enabled communications.

Motivated by the recent advances of Multi-Agent Reinforcement Learning (MARL) in distributed wireless control, we develop a multi-agent DRL framework tailored to quantized RIS control in UAV-assisted communication systems. By decomposing the RIS configuration task into coordinated row and column adjustments, the proposed approach overcomes the curse of dimensionality and enables scalable, low-latency optimization under dynamic channel conditions. The main contributions of this paper are summarized as follows:

- An end-to-end (E2E) channel inference architecture is designed to mitigate the negative effects of CSI imperfection during UAV–RIS–BS channel dynamics.

The proposed inference model captures temporal variations and reconstructs stable channel representations, thereby significantly enhancing the robustness of RIS control under mobility-induced uncertainty.

- The system objective is formulated to jointly optimize sum rate, control latency, and energy consumption, reflecting practical constraints in UAV-assisted RIS operation. This multi-objective optimization problem is then reformulated as a multi-agent Markov decision process (MMDP), which enables decentralized yet coordinated decision-making across RIS row and column controllers while remaining compatible with quantized RIS hardware.
- A novel multi-agent deep Q-network (MADQN)–based controller is developed to adjust the phase shifts of the RIS elements. This design allows for more precise and scalable control of both the row and column elements, which is especially beneficial for large-scale RIS in UAV communication scenarios. The MADQN framework also demonstrated adaptability to the mobility inherent in UAV-assisted systems.
- Comprehensive simulations validate performance of the proposed framework. The results demonstrate that the MADQN-based method significantly outperforms single-agent DQN and conventional baseline approaches in terms of convergence behavior, sum rate performance, energy efficiency, and control latency. The proposed architecture also exhibits consistent robustness across different channel conditions, noise levels, user populations, and RIS dimensions.

The remaining sections of this work are structured as follows. Section II delineates the system model and problem formulation. Section III presents the proposed deep reinforcement learning system, detailing the training methodology and decision-making procedures for the two agents. Section IV presents the simulation results, illustrating the superiority of the dual-agent framework. Consequently, Section V draws conclusions.

C. NOTATIONS

The sets, vectors and matrices are denoted by Italic boldface characters. $\mathcal{CN}(\cdot,\cdot)$ means complex normal distribution. $\mathbb{C}^{N_{BS}\times m_r m_c}$ and $\mathbb{C}^{m_r m_c\times 1}$ mean complex matrices with dimensions of $N_{BS}\times m_r m_c$, and complex vectors with dimension of $m_r m_c \times 1$. $(\cdot)^\star$, $|\cdot|$, $\|\cdot\|_2$, $(\cdot)^T$, $(\cdot)^H$ represent desired value, absolute value, L_2-norm, transpose, and conjugate transpose of matrix respectively. $\odot$ denotes Hadamard product, and $\otimes$ denotes the Kronecker product. Table 1 summarizes the specific notation in this paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

To support UAV communications in XL-MIMO systems, the system model has been developed as below to benefit uplink multi-UAV communication system from the advanced RIS controller. In addition, considering the main objective of this paper, the channel model is further designed to adapt to the RIS-enabled uplink multi-UAV communication systems.

TABLE 1. Notation Table.

Symbol	Description
$\boldsymbol{H}_U$	Channel matrix for uplink multi-UAV communications
$\boldsymbol{H}_{UB}$	Channel matrix between UAVs and BS
$\boldsymbol{\Phi}$	Reflective pattern of RIS
$\boldsymbol{H}_{RB}$	Channel matrix between RIS and BS
$\boldsymbol{H}_{UR}$	Channel matrix between UAVs and RIS
$\boldsymbol{H}^{LoS}$	The LoS components of channel model
$\boldsymbol{H}^{NLoS}$	The NLoS components of channel model
κ	Rician factor
$L_i(\boldsymbol{\theta}_{i,t})$	Loss function from the i^{th} agent
$\boldsymbol{\mathcal{D}}_i$	Experiences of the i^{th} agent
$\boldsymbol{A}_i(\phi_i)$	ULA-enabled array response vector
$\boldsymbol{A}_R(\phi_{R_r},\phi_{R_c})$	Array response vector of UPA-enabled RIS
β_{XY}	Complex path gain between X and Y
(d_r, d_c)	The spacing of row and column element in RIS
$(\mathbf{p}_{U_i}^{(1)}, \mathbf{p}_{U_i}^{(2)})$	The departure and destination of UAV i
$\mathbf{p}_{BS}$	Position of BS
$\mathbf{p}_{\text{RIS}}$	Position of RIS
P_{U_i}	Transmitted power of the i^{th} UAV
$\boldsymbol{n}_{BS}$	AWGN at BS
$\boldsymbol{v}_{U_i}$	Combining vector at the i^{th} UAV
C_i	Achievable data rate from i^{th} UAV
$C_{i,min}$	Minimum data rate at the i^{th} UAV
$P_{i,max}$	Maximum transimitted power at the i^{th} UAV
$\phi_{_,t}$	The reflective vector for row or column element of RIS
$\phi(\varphi)$	Steering vector with phase φ
C_{ij}	The data rate from the j^{th} UAV during the i^{th} channel block
C_{th}	Threshold of data rate
$\alpha_{g,t}$	Global learning rate at the t^{th} iteration
$\boldsymbol{y}_{i,t}^{target}$	Target value of training DQN
$Q(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}; \boldsymbol{\theta}_t)$	Q function of train DQN
$Q_i(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t-1})$	Q function of target DQN

A. SYSTEM MODEL

In this work, a RIS-enabled uplink multi-UAV communication system was built up as shown in Figure. 1, consisting of BS, RIS, and N_{UAV} UAVs as $\boldsymbol{U}$ with $\boldsymbol{U} = [U_1, U_2, \cdots, U_{N_{UAV}}]$ in the uplink. The BS was facilitated with N_{BS} antennas for receiving, while each UAV is facilitated with single antenna for transmitting. Moreover, it is assumed that the direct link between UAV and BS is low enough, so that RIS is leveraged for uninterrupted UAV communications. In that case, RIS is installed on the top of a building between BS and UAVs, and located on the Y-Z plane. Taking into account the uniform planar array (UPA) model, RIS is made up of m elements aligned at the origin of the coordinate system, including m_r row elements and m_c column elements. In addition, each element of RIS can adjust amplifier and phase shifter simultaneously for amplified reflection. The initial position of the i^{th} UAV is denoted as $\mathbf{p}_{U_i}^{(1)} = (x_{U_i}^{(1)}, y_{U_i}^{(1)}, z_{U_i}^{(1)})$, and the final position of them is $\mathbf{p}_{U_i}^{(2)} = (x_{U_i}^{(2)}, y_{U_i}^{(2)}, z_{U_i}^{(2)})$. The UAV travels along a predefined trajectory, maintaining a fixed altitude z_{UAV} while its x- and y-coordinates transition linearly over time. The RIS is located at $\mathbf{p}_{\text{RIS}} = (x_{\text{RIS}}, y_{\text{RIS}}, z_{\text{RIS}})$, and the BS is fixed at

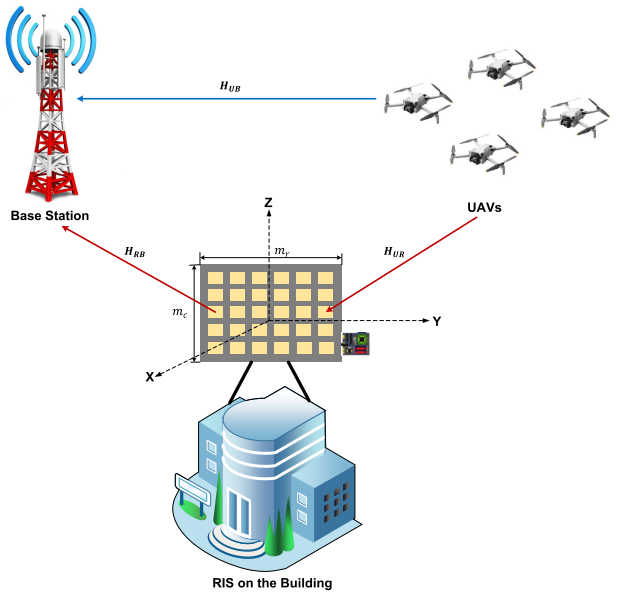


FIGURE 1. RIS-enabled uplink Multi-UAV communication system.

$\mathbf{p}_{\text{BS}} = (x_{\text{BS}}, y_{\text{BS}}, z_{\text{BS}})$. These positions are parametrized to allow flexibility in the deployment of the system and can be adjusted based on the specific urban topology. The trajectory of the i^{th} UAV was designed as (1):

$$\mathbf{p}_{U_i}(t) = (1-t)\mathbf{p}_{U_i}^{(1)} + t\mathbf{p}_{U_i}^{(2)}, \tag{1}$$

where $t \in [0, 1]$ represents the normalized time over the whole flight of UAV. The velocity of UAV is denoted as $\mathbf{v}_{\text{UAV}}$, allowing for analysis of system performance under varying speeds. The trajectory simplification focuses attention on RIS optimization while providing a controlled testbed for evaluating the robustness and efficiency of the proposed algorithm.

B. CHANNEL MODEL AND TRANSMISSION MODEL

To accurately model the communication environment, channel models between the i^{th} UAV and BS are employed as (2) for the three communication links, including UAV-to-RIS, RIS-to-BS, and UAV-to-BS.

$$\boldsymbol{h}_{U_i} = \boldsymbol{h}_{U_iB} + \boldsymbol{H}_{RB}\boldsymbol{\Phi}\boldsymbol{h}_{U_iR}, \tag{2}$$

where $\boldsymbol{h}_{U_iB} \in \mathbb{C}^{N_{BS}\times 1}$ means the channel vector between i^{th} UAV and BS, $\boldsymbol{h}_{U_iR} \in \mathbb{C}^{m_r m_c \times 1}$ means the channel vector between i^{th} UAV and RIS, and $\boldsymbol{H}_{RB} \in \mathbb{C}^{N_{BS}\times m_r m_c}$ denotes the channel vector between RIS and BS. Furthermore, $\boldsymbol{\Phi} = \text{diag}\{\boldsymbol{\phi}\}$ consists the phase shift vectors $\boldsymbol{\phi} \in \mathbb{C}^{m_r m_c \times 1}$ of RIS, where the j^{th} element of the phase shift vectors satisfied $|\boldsymbol{\phi}(j)| = 1$. Correspondingly, the channel model is defined as (3) for multi-UAV communications:

$$\boldsymbol{H}_U = \boldsymbol{H}_{UB} + \boldsymbol{H}_{RB}\boldsymbol{\Phi}\boldsymbol{H}_{UR}, \tag{3}$$

where $\boldsymbol{H}_U = [\boldsymbol{h}_{U_1}, \boldsymbol{h}_{U_2}, \ldots, \boldsymbol{h}_{U_{N_{UAV}}}] \in \mathbb{C}^{N_{BS}\times N_{UAV}}$, $\boldsymbol{H}_{UB} = [\boldsymbol{h}_{U_1B}, \boldsymbol{h}_{U_2B}, \ldots, \boldsymbol{h}_{U_{N_{UAV}}B}] \in \mathbb{C}^{N_{BS}\times N_{UAV}}$, and $\boldsymbol{H}_{UR} = [\boldsymbol{h}_{U_1R}, \boldsymbol{h}_{U_2R}, \ldots, \boldsymbol{h}_{U_{N_{UAV}}R}] \in \mathbb{C}^{m_r m_c \times N_{UAV}}$. Considering the Rician factor, the channel models $\boldsymbol{H}_{UB}$, $\boldsymbol{H}_{UR}$, and $\boldsymbol{H}_{RB}$ are defined as (4), (5), and (6).

$$\boldsymbol{H}_{UR} = \sqrt{\frac{\kappa}{\kappa+1}}\boldsymbol{H}_{UR}^{LoS} + \sqrt{\frac{1}{\kappa+1}}\boldsymbol{H}_{UR}^{NLoS}, \tag{4}$$

$$\boldsymbol{H}_{UB} = \sqrt{\frac{\kappa}{\kappa+1}}\boldsymbol{H}_{UB}^{LoS} + \sqrt{\frac{1}{\kappa+1}}\boldsymbol{H}_{UB}^{NLoS}, \tag{5}$$

$$\boldsymbol{H}_{RB} = \sqrt{\frac{\kappa}{\kappa+1}}\boldsymbol{H}_{RB}^{LoS} + \sqrt{\frac{1}{\kappa+1}}\boldsymbol{H}_{RB}^{NLoS}, \tag{6}$$

where $\boldsymbol{H}_{UR}^{LoS}$, $\boldsymbol{H}_{UB}^{LoS}$, and $\boldsymbol{H}_{RB}^{LoS}$ represent the deterministic channel models for line-of-sight (LoS) path, while $\boldsymbol{H}_{UR}^{NLoS}$, $\boldsymbol{H}_{UB}^{NLoS}$, and $\boldsymbol{H}_{RB}^{NLoS}$ represent the fast fading channel models for non-line-of-sight (NLoS) path. In addition, κ means the ratio of power between LoS path and NLoS path.

Given the UAV communications of Figure. 1 is in the far-field regime, the NLoS components of channel models are formulated by $\boldsymbol{H}_{XY}^{NLoS} \sim \mathcal{CN}(0, \sigma_{XY}^2\boldsymbol{I})$, while the LoS components of channel models are defined as below for each link $XY \in \{UR, UB, RB\}$:

$$\boldsymbol{H}_{XY}^{LoS} = \beta_{XY}\boldsymbol{A}_Y(\phi_Y)\boldsymbol{A}_X(\phi_X)^H, \tag{7}$$

where $\boldsymbol{A}_X(\phi_X)$ and $\boldsymbol{A}_Y(\phi_Y)$ mean the array response vector at transmitter and receiver with the angles of departure (AoD) ϕ_X and arrival (AoA) ϕ_Y separately, while $\beta_{XY} = \sqrt{G_X G_Y}\frac{\lambda}{4\pi d_{XY}}e^{-\frac{2\pi d_{XY}}{\lambda}}$ means the complex path gain, where d_{XY} denotes link distance between X and Y, and λ is the carrier wavelength. Moreover, G_X and G_Y mean the antenna gain of X and Y, respectively. For a uniform linear array (ULA) i with N_i antennas and spacing d_i, the array response vector of i are defined as:

$$\boldsymbol{A}_i(\phi_i) = \frac{1}{\sqrt{N_i}}\left[1, e^{j\frac{2\pi d_i}{\lambda}\sin(\phi_i)}, \cdots, e^{j\frac{2\pi d_i}{\lambda}(N_i-1)\sin(\phi_i)}\right]^T. \tag{8}$$

In particular, considering the configuration of UPA in RIS, (8) can be extended to the following to define $\boldsymbol{A}_R(\phi_{R_r}, \phi_{R_c})$:

$$\begin{aligned}\boldsymbol{A}_R(\phi_{R_r}, \phi_{R_c}) = \frac{1}{\sqrt{m_r m_c}}\Bigg[&1, e^{j\frac{2\pi d_r}{\lambda}\sin(\phi_{R_r})\cos(\phi_{R_c})}, \cdots, \\ &e^{j\frac{2\pi d_r}{\lambda}(m_r-1)\sin(\phi_{R_r})\cos(\phi_{R_c})}\Bigg]^T \\ \otimes\Bigg[&1, e^{j\frac{2\pi d_c}{\lambda}\sin(\phi_{R_r})\sin(\phi_{R_c})}, \cdots, \\ &e^{j\frac{2\pi d_c}{\lambda}(m_c-1)\sin(\phi_{R_r})\sin(\phi_{R_c})}\Bigg]^T\end{aligned} \tag{9}$$

where ϕ_{R_r} and ϕ_{R_c} represent the vertical and horizontal AoA separately. In addition, d_r means the spacing of RIS row elements, and d_c means the spacing of RIS column elements.

The total received signal at the BS is the superposition of the direct signal from the UAV and the signal reflected by the RIS, the received signal from each UAV is given by:

$$\boldsymbol{y}_{U_i} = \sqrt{P_{U_i}}\boldsymbol{h}_{U_i}^H\boldsymbol{x}_{U_i} + \sum_{l\neq i}^{N}\sqrt{P_{U_l}}\boldsymbol{h}_{U_l}^H\boldsymbol{x}_{U_l} + \boldsymbol{n}_{BS}, \tag{10}$$

where x_{U_i} is the transmitted signal from the i^{th} UAV, P_{U_i} means the transmitted power from the i^{th} UAV, and n_{BS} is the additive white Gaussian noise (AWGN) at the BS with $\boldsymbol{n_{BS}} \sim \mathcal{CN}(0, \sigma_{BS}^2 \boldsymbol{I})$.

The goal of the proposed system is to maximize this data rate by dynamically optimizing the RIS phase shifts, taking into account UAV mobility and urban obstructions. The primary performance metric is the achievable data rate from the i^{th} UAV at the BS, defined as:

$$C_i = \log_2\left(1 + \frac{P_{U_i}|\boldsymbol{v}_{U_i}^H \boldsymbol{h}_{U_i}|^2}{\sum_{l\neq i}^{N} P_{U_l}|\boldsymbol{v}_{U_i}^H \boldsymbol{h}_{U_l}|^2 + ||\boldsymbol{v}_{U_i}||^2 \sigma_{BS}^2}\right), \quad (11)$$

where $\boldsymbol{v}_{U_i} \in \mathbb{C}^{N_{BS}\times 1}$ means the combining vector of the i^{th} UAV.

C. PROBLEM FORMULATION

Considering the limitation of imperfect CSI estimation, UAV energy, and real-time control latency, the main objective of this paper is formulated to achieve trade-off between sum rate and above challenges under the proposed system model as below:

$$\begin{aligned}
3(\text{P0}) \quad & \max_{\{\boldsymbol{\Phi}_t, \boldsymbol{V}_t, \boldsymbol{P}_t\}_{t=1}^{T}} \; w_1 \sum_{i=1}^{N} \mathbb{E}_{\text{CSI}}[C_{i,t}] + w_2\, \mathbb{E}_{\text{CSI}}\left[\frac{\sum_{i=1}^{N} C_{i,t}}{P_t^{\text{tot}}}\right] \\
& \qquad - w_3\,[L_t - L_{\max}]_+ && (12a)\\
& \text{s.t. } \boldsymbol{v}_{U_i,t} \in \boldsymbol{V}_t,\ \left\|\boldsymbol{v}_{U_i,t}\right\| = 1, \quad \forall i, t \in \mathcal{N}, && (12b)\\
& \boldsymbol{P}_{i,t} \in \boldsymbol{P}_t, \boldsymbol{P}_{i,t} \le \boldsymbol{P}_{i,\max}, \quad \forall i, t \in \mathcal{N}, && (12c)\\
& \phi_{n,t} \in \mathcal{Q}_b \triangleq \left\{0, \frac{2\pi}{2^b}, \cdots, \frac{2\pi(2^b-1)}{2^b}\right\}, \\
& \Phi_{n,t} = e^{j\phi_{p,q}(t)}, \forall n, t \in \mathcal{N}, \Phi_{n,t} \in \boldsymbol{\Phi}_t && (12d)\\
& P_t^{\text{tot}} = \sum_i \frac{P_{i,t}}{\eta_i} + P_{\text{RIS}}(\boldsymbol{\Phi}_t) + P_C(\boldsymbol{\Phi}_t, \boldsymbol{\Phi}_{t-1}) \\
& \qquad + P_{BB}(\boldsymbol{V}_t) + P_{UAV},\ \forall i, t \in \mathcal{N} && (12e)\\
& E_{i,t+1} = E_{i,t} - P_{i,t}^{tot}\Delta t, E_{i,t} \ge E_{i,\min}, \forall i, t \in \mathcal{N} \\
& && (12f)\\
& \mathbb{P}\{C_{i,t} \ge C_{i,\min}\} \ge 1 - \delta_i, \forall i, t \in \mathcal{N} && (12g)\\
& \mathbb{P}\{L_t \le \Delta t \le T_c\} \ge 1 - \epsilon_L, \forall t \in \mathcal{N} && (12h)\\
& w_1 + w_2 - w_3 = 1 && (12i)
\end{aligned}$$

where $\{\boldsymbol{\Phi}_t, \boldsymbol{V}_t, \boldsymbol{P}_t\}$ denotes the RIS configuration, combining vector of UAV, and transmit power at time slot t of statistical CSI. In addition, $\boldsymbol{v}_{U_i,t}$ and $C_{i,t}$ mean the combining vector and data rate of the i^{th} UAV at time slot t. Moreover, P_t^{tot} and L_t denote the total power consumption and UAV control latency at time slot t separately, while $\mathbb{E}_{\text{CSI}}[\cdot]$ means that expected value across statistical CSI. (12b) and (12c) limit the energy consumption of the baseband and UAV, where $||\boldsymbol{v}_{U_i,t}||$ defines the unit energy consumption for decoding at the i^{th} UAV, and $P_{i,t}$ means the energy consumption of the i^{th} UAV with upper bound $P_{i,\max}$. (12d) specifies the phase configuration of each RIS $\Phi_{n,t}$ as the discrete quantization with the number of control bits b. (12e) defines the total power consumption as PA based RF transmit power $\sum_i \frac{P_{i,t}}{\eta_i}$, active element power $P_{\text{RIS}}(\boldsymbol{\Phi}_t)$, phase shift power $P_C(\boldsymbol{\Phi}_t, \boldsymbol{\Phi}_{t-1})$, baseband power $P_{BB}(\boldsymbol{V}_t)$, and UAV platform power P_{UAV}, where $P_{i,t} \in P_t$ and η_i means the energy consumption and PA efficiency of the i^{th} UAV. (12f) limits the bound and update of battery energy for each UAV. (12g) and (12h) constrain the lower bound of achievable capacity and the upper bound of control latency of the proposed system, where $C_{i,min}$ means the minimum data rate at the i^{th} UAV, Δt means time-slot duration, and T_c means channel coherence time. In that case, $C_{i,t}$, P_t^{tot}, and L_t align well with the indicator of imperfect CSI, UAV energy accounting, and real-time control latency, so (12i) is defined to achieve a balance between the optimization for channel state, energy consumption, and control latency.

Although problem (P0) captures the complete optimization of RIS phase configuration $\boldsymbol{\Phi}_t$, UAV combining vector $\boldsymbol{V}_t$, and transmit power allocation P_t, direct optimization over element-wise RIS phase matrix $\Phi_t \in \mathcal{Q}_b^N$ is still a struggle for practical implementation. In particular, the cardinality of Φ_t increases exponentially with the number of RIS elements, which leads to extremely high computational overhead for real-time control, large switching energy due to frequent per-element updates, and severe latency violations under fast UAV mobility. Moreover, the imperfections in statistical CSI estimation make per-element optimization unstable and highly sensitive to noise. To overcome these challenges, the RIS configuration is redesigned with two low-dimensional steering coefficients $\{k_{r,t}, k_{c,t}\}$, which jointly describe the entire RIS phase profile through a structured representation $\phi_{r,c,t}$. This transformation reduces the RIS control space from $O(2^{bN})$ to a compact and dimensional domain, enabling robust optimization under statistical CSI, significantly lower energy consumption for RIS reconfiguration, and improved control-latency compliance. Accordingly, by substituting Φ_t with $\{k_{r,t}, k_{c,t}\}$ to define the achievable capacity, energy consumption, and control latency, (P0) is reformulated as (P1) in (13), where the objective and constraints remain conceptually identical but are expressed in terms of the steering-based RIS model. This yields a tractable formulation that preserves the physical behavior of RIS control while ensuring computational efficiency and practical deployability.

$$\begin{aligned}
(\text{P1}) \quad & \max_{\{k_{r,t}, k_{c,t}, \boldsymbol{V}_t, \boldsymbol{P}_t\}_{t=1}^{T}} \; w_1 \sum_{i=1}^{N} \mathbb{E}_{\text{CSI}}[C_{i,t}(k_{r,t}, k_{c,t}, \boldsymbol{V}_t, \boldsymbol{P}_t)] \\
& \qquad + w_2\, \mathbb{E}_{\text{CSI}}\left[\frac{\sum_{i=1}^{N} C_{i,t}(k_{r,t}, k_{c,t}, \boldsymbol{V}_t, \boldsymbol{P}_t)}{P_t^{\text{tot}}(k_{r,t}, k_{c,t}, \boldsymbol{V}_t, \boldsymbol{P}_t)}\right] \\
& \qquad - w_3\left[L_t(k_{r,t}, k_{c,t}) - L_{\max}\right]_+ && (13a)\\
& \text{s.t. } \boldsymbol{v}_{U_i,t} \in \boldsymbol{V}_t,\ \left\|\boldsymbol{v}_{U_i,t}\right\| = 1, \quad \forall i, t \in \mathcal{N}, && (13b)\\
& \qquad \boldsymbol{P}_{i,t} \in \boldsymbol{P}_t,\ \boldsymbol{P}_{i,t} \le \boldsymbol{P}_{i,\max}, \quad \forall i, t \in \mathcal{N}, && (13c)\\
& \qquad \phi_{p,q}(t) = \mathcal{Q}_b\left[k_{r,t}x_p + k_{c,t}y_q\right], \Phi_{p,q,t} = e^{j\phi_{p,q}(t)}, \\
& \qquad \mathcal{Q}_b(\theta) = \underset{\phi \in \left\{0, \frac{2\pi}{2^b}, \ldots, \frac{2\pi(2^b-1)}{2^b}\right\}}{\arg\min} |\theta - \phi|,\ \forall p, q, t, \\
& && (13d)
\end{aligned}$$

$$P_t^{\text{tot}}(k_{r,t}, k_{c,t}, \boldsymbol{V}_t, \boldsymbol{P}_t) = \sum_i \frac{P_{i,t}}{\eta_i} + P_{\text{RIS}}\big(N_{\text{act}}(t)\big) + P_C\big(\Delta_\phi(t)\big) + P_{BB}(\boldsymbol{V}_t) + P_{UAV},\ \forall t \in \mathcal{N}, \tag{13e}$$

$$L_t(k_{r,t}, k_{c,t}) = L_{\text{base}} + \alpha_{\text{sw}}\,\Delta_\phi(t) + \alpha_{\text{act}}\,N_{\text{act}}(t) + \alpha_{\text{snr}}\,\ell(\text{SNR}_t),\ \forall t \in \mathcal{N} \tag{13f}$$

$$E_{i,t+1} = E_{i,t} - P_{i,t}^{\text{tot}}\Delta t,\ \ E_{i,t} \geq E_{i,\min},\ \forall i, t \in \mathcal{N}, \tag{13g}$$

$$\mathbb{P}\{C_{i,t}(k_{r,t}, k_{c,t}, \boldsymbol{V}_t, \boldsymbol{P}_t) \geq C_{i,\min}\} \geq 1 - \delta_i,\ \forall i, t, \tag{13h}$$

$$\mathbb{P}\{L_t(k_{r,t}, k_{c,t}) \leq \Delta t \leq T_c\} \geq 1 - \epsilon_L,\ \forall i, t \in \mathcal{N}, \tag{13i}$$

$$w_1 + w_2 - w_3 = 1. \tag{13j}$$

In particular, for the structured RIS configuration generated by $(k_{r,t}, k_{c,t})$, we further redefine the power consumption of RIS P_{RIS}, baseband P_{BB}, and phase shifter P_c with the load of active RIS elements N_{act} and the normalized phase-update magnitude $\Delta_\phi(t)$ as below:

$$N_{\text{act}}(t) = f_{\text{act}}\big(k_{r,t}, k_{c,t}\big) \approx \Big\lfloor \big(|k_{r,t}|_{\text{norm}} + |k_{c,t}|_{\text{norm}}\big)\, N_r N_c \Big\rfloor, \tag{14a}$$

$$\Delta_\phi(t) = \frac{1}{2\pi} \sum_{p=1}^{N_r} \sum_{q=1}^{N_c} \big|\phi_{p,q}(t) - \phi_{p,q}(t-1)\big|_{2\pi}, \tag{14b}$$

$$P_{\text{RIS}}\big(N_{\text{act}}(t)\big) = N_{\text{act}}(t)\, P_{\text{elem}}, \tag{14c}$$

$$P_C\big(\Delta_\phi(t)\big) = P_{C,0} + \kappa_P\,\Delta_\phi(t), \tag{14d}$$

$$P_{BB}(\boldsymbol{V}_t) = P_{BB,0} + \kappa_B\, g(\boldsymbol{V}_t), \tag{14e}$$

$$\ell(\text{SNR}_t) = \max\left\{0,\ \frac{\text{SNR}_0 - \text{SNR}_t}{\text{SNR}_0}\right\}, \tag{14f}$$

where $|\cdot|_{2\pi}$ denotes the wrapped angular distance, P_{elem} is the per-element RIS power, $P_{C,0}$ and $P_{BB,0}$ are fixed offsets, and κ_P, κ_B are hardware-dependent scaling factors. The functions $f_{\text{act}}(\cdot)$ and $g(\cdot)$ are chosen as monotone proxies that capture the load of active RIS elements and the computational complexity of baseband processing, respectively. According to (14a), (14b), and (14f), the control latency in (P1) is written as (13f), where L_{base} collects the fixed delay over control-plane, driver, computation, and networking.

III. MADQN BASED RIS CONTROLLER

In this section, multi-agent deep Q network (MADQN) is initially overviewed as a preliminary. Then, to optimize the objective in (12a), the proposed system models are transformed to markov decision process (MDP). Subsequently, a novel dual-agent RIS controller was developed under the proposed MDP to jointly improve the sum rate, energy efficiency, and control latency under imperfect CSI estimation at uplink.

A. BACKGROUND OF MULTI-AGENT DEEP Q NETWORK

Deep Q network (DQN) is a general deep reinforcement learning framework to select optimal actions by interacting with their environment. However, DQN is normally limited to the constraints of action space and computation complexity. As a result, the advanced research has extended the principles of DQN to multi-agent reinforcement learning (MARL) as MADQN, where multi-agents work together within a shared environment.

Considering the framework of decentralized training and decentralized execution (DTDE) [43], each DQN agent is driven by independent training and action execution. The multiple MDP (MMDP) is defined as experiences $(\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T})$, where $\mathcal{N}$ is the number of agents, $\mathcal{S}$ represents the state space, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_N$ is a joint action space consisting of the action space of each agent $\mathcal{A}_i$, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{N_{ex}}$ is defined as the expected reward function, and $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]^{N_{ex}}$ means the transition function, where N_{ex} means the memory storage of aforementioned experiences. The objective of each agent is to maximize the cumulated reward since $t = k$ in the future as below:

$$R_t = \sum_{t=k}^{T} \gamma^{t-k} r_t, \tag{15}$$

where r_t means the temporal reward, T means the number of episodes during training, and γ is the discount factor. Given the policy π, current state $\boldsymbol{s}$, and action of each agent $\boldsymbol{a} = \{a_1, a_2, \cdots, a_N\}$, the action-value is defined as the expected R_t by Q function:

$$Q^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}[R_t | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}], \tag{16}$$

According to the principle of Q-learning, the optimal action-value can be defined with Bellman equation as below:

$$Q^\star(\boldsymbol{s}_t, \boldsymbol{a}_t) \approx \mathbb{E}[r_t + \gamma \max_{\boldsymbol{a}_{t+1}} Q^\star(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) | \boldsymbol{s}_t, \boldsymbol{a}_t], \tag{17}$$

where $\boldsymbol{s}_{t+1}$ means the next state when the temporal action $\boldsymbol{a}_t$ is executed in state $\boldsymbol{s}_t$, and $\boldsymbol{a}_{t+1}$ means the optimal action at state $\boldsymbol{s}_{t+1}$. The aforementioned $Q^\star(\boldsymbol{s}_t, \boldsymbol{a}_t)$ can be approximated with deep Q network $Q_i(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}; \boldsymbol{\theta}_{t-1})$ at each agent, the loss function of which can be defined as below:

$$L_i(\boldsymbol{\theta}_{i,t}) = \mathbb{E}_{(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, r_t, \boldsymbol{s}_{t+1}) \sim (\boldsymbol{s}, \boldsymbol{a}_i, \boldsymbol{r}, \boldsymbol{s}')}[\boldsymbol{y}_{i,t} - Q(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t})]^2, \tag{18}$$

$$\boldsymbol{y}_{i,t} = \mathbb{E}_{(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, r_t, \boldsymbol{s}_{t+1}) \sim (\boldsymbol{s}, \boldsymbol{a}_i, \boldsymbol{r}, \boldsymbol{s}')} \times \big[r_{t+1} + \gamma \max_{\boldsymbol{a}_{i,t+1}} Q^\star(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{i,t+1}; \boldsymbol{\theta}_t) \mid \boldsymbol{s}_t, \boldsymbol{a}_{i,t}\big] \tag{19}$$

where $\boldsymbol{s}_t$ means the temporal parameters of deep Q network, and the tuple $(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1})$ is sampled from the experiences $(\boldsymbol{s}, \boldsymbol{a}_i, \boldsymbol{r}, \boldsymbol{s}')$, where $\boldsymbol{s}, \boldsymbol{a}_i, \boldsymbol{r}, \boldsymbol{s}'$ mean the storage for current state, action of the i^{th} agent, current reward, and next state,

respectively. Furthermore, the aforementioned loss function can be updated by gradient descent as below:

$$\nabla_{\boldsymbol{\theta}_{i,t}} L_i(\boldsymbol{\theta}_{i,t}) = \mathbb{E}_{(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, r_t, \boldsymbol{s}_{t+1}) \sim (\boldsymbol{s}, \boldsymbol{a}_i, \boldsymbol{r}, \boldsymbol{s}')} [\boldsymbol{y}_{i,t} - Q(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t-1})] \nabla_{\boldsymbol{\theta}_{i,t}} Q(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t}), \tag{20}$$

B. MULTI-AGENT MARKOV DECISION PROCESS FORMULATION

Generally, the optimization for RIS controller is a NP-hard problem with high computational complexity from non-linear control. In addition, it is so difficult for phase shifter to achieve the global optimization due to the non-convex nature of imperfect CSI acquisition. To solve (13) in a data-driven manner, we reformulate the RIS-control problem at uplink UAV communications as a cooperative multi-agent Markov decision process (MMDP) and learn a policy by means of a multi-agent deep Q-network (MADQN). The MMDP is characterized by the set of agents $\mathcal{N}$, the global state space $\mathcal{S}$, the joint action space $\mathcal{A}$, the reward function $\mathcal{R}$, and the transition kernel $\mathcal{T}$. The objective of the learned policy is to approximate the joint RIS steering for (P1) by maximizing a reward that balances sum-rate, energy consumption, and control latency.

1) ENVIRONMENT EVOLUTION

The evolution of signal propagation between UAVs and BS was further defined as multi-MDP (MMDP), where two agents optimized phase shifter for row element and column element of RIS separately. At each slot t, the true uplink channels $\mathbf{H}_{U,t}$ are drawn from the statistical CSI model, and noisy pilot measurements are obtained at the BS. As illustrated in Fig. 2, a neural channel front-end (NCF) denoises the pilots and outputs a latent representation of the CSI as well as the uncertainty of them. This latent feature is then processed by a recurrent belief encoder, which aggregates the long-term CSI observations and previous RIS configuration into a compact belief vector $\boldsymbol{b}_t \in \mathbb{R}^{d_b}$ as the global state s_t for RIS control.

Given the belief vector $\boldsymbol{b}_t$, the two agents simultaneously select their steering updates for the RIS row and column apertures as their optimal action $a_{i,t}$. The joint steering update is mapped to the structured RIS configuration $\Phi p, q, t$ as in (P1), from which the instantaneous data rate $C_{i,t}$, total power consumption P_t^{tot}, and control latency L_t can be derived with (11), (13e), and (13f). Then, the UAV batteries $E_{i,t}$ will be updated to $E_{i,t+1}$ in terms of $P_{i,t}$ as (13g), after which new noisy pilots should be generated through the probability of transition function $\mathcal{T}$ of MMDP, further converting to a new belief vector $\boldsymbol{b}_{t+1}$ from recurrent belief encoder as the next state s_{t+1}.

2) OBSERVATION SPACE

Considering that the row agent and column agent need to cooperate with each other for higher link capacity, both agents must share a consistent description of the wireless

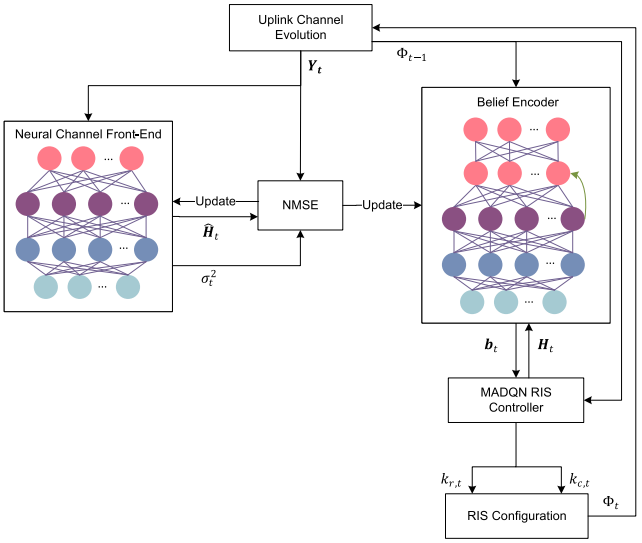


FIGURE 2. E2E channel inference architecture.

environment and the current RIS configuration. Due to imperfect CSI acquisition, the true multi-UAV channel $\boldsymbol{H}_U$ cannot be directly observed without noise. Instead, the BS receives the pilot measurement Y_t at each time slot t, which is processed by the NCF to produce a denoised CSI estimate $\hat{H}_t$ along with channel uncertainty term σ_t^2. Given the controller also retains full knowledge of the current RIS configuration Φ_t, the observation is then formed as $\boldsymbol{o}_t = \{\boldsymbol{Y}_t, \hat{\boldsymbol{H}}_t, \sigma_t^2, \boldsymbol{\Phi}_{t-1}\}$. However, the raw observation reflects only instantaneous and noisy information, which cannot represent Markovian system state directly. To construct a stable, temporally informed representation, the observation is passed through the end-to-end neural inference pipeline illustrated in Fig. 2. This pipeline consists of the NCF followed by a recurrent belief encoder, where the NCF extracts latent CSI features and the belief encoder aggregates past observations and RIS configurations to form a compact belief vector b_t. The resulting belief captures the long-term temporal evolution of the statistical CSI and provides a robust summary of the relevant channel dynamics. Since the RIS configuration is fully known at the controller, the global state at time slot t is defined as $\boldsymbol{\mathcal{S}} = \{\boldsymbol{b}_t, \boldsymbol{\Phi}_{t-1}\}$, where $\boldsymbol{b}_t$ denotes the latent, uncertainty-aware channel representation, and $\boldsymbol{\Phi}_{t-1}$ denotes the previous phase matrix of RIS. In that case, both agents operate on a shared state that jointly encodes the belief about the uplink channels and the instantaneous RIS configuration. To keep the balance between exploration and exploitation of each agent, the ϵ-greedy policy is adopted during training. Accordingly, the aforementioned state space is extended to the observation space as $\boldsymbol{\mathcal{O}} = \{\boldsymbol{b}_t, \boldsymbol{\Phi}_{t-1}, \epsilon\}$, where ϵ means the probability to choose random action.

3) ACTION SPACE

Action space limited the boundary of RIS control during UAV communications. Considering the complexity of fine-grained analog bias voltages, the adjustment of the reflection pattern

has to be performed with a discrete action space along the RIS row and column dimensions separately. The action space is defined as the incremental steering along the row and column axes as below:

$$A_i = \left\{ -\frac{4}{N_i}, -\frac{2}{N_i}, 0, \frac{2}{N_i}, \frac{4}{N_i} \right\}, i \in \{r, c\} \tag{21}$$

where N_r and N_c are the number of RIS elements per row and column. At slot t, the row agent selects an index $a_{r,t} \in \{1, \ldots, |A_r|\}$, and the column agent selects $a_{c,t} \in \{1, \ldots, |A_c|\}$. The corresponding steering increments are $\Delta\kappa_{r,t} = A_r[a_{r,t}]$ and $\Delta\kappa_{c,t} = A_c[a_{c,t}]$, which update the normalized steering parameters as $\kappa_{r,t+1} = \kappa_{r,t} + \Delta\kappa_{r,t}$ and $\kappa_{c,t+1} = \kappa_{c,t} + \Delta\kappa_{c,t}$, respectively. According to the quantized phase field in (13d), the per-element RIS configuration will be generated by spatial wavenumbers, which are defined as $k_{r,t+1} = \pi\kappa_{r,t+1}N_r$ and $k_{c,t+1} = \pi\kappa_{c,t+1}N_c$ separately. To collect the decisions of both agents, the joint action space of the MMDP is determined by $\mathcal{A} = \mathcal{A}_r \times \mathcal{A}_c$, where $\boldsymbol{\mathcal{A}}_r$ and $\boldsymbol{\mathcal{A}}_c$ are the discrete sets of row and column steering increments. This formulation is fully consistent with the discrete action sets $\boldsymbol{\mathcal{A}}_i$ for the MADQN implementation as Section. III-A and directly generates the RIS configuration in (P1).

4) REWARD FUNCTION

To align the MMDP with the optimization objective in (12), the immediate reward at slot t must reflect the trade-off between sum rate, energy consumption, and control latency, where $C_t = \sum_i C_{i,t}$ denote the instantaneous sum rate at the BS, P_t^{tot} means the total power consumption in (12e), and L_t means the control latency. In that case, the instantaneous energy efficiency is initially defined as $EE_t = \frac{C_t}{P_t^{tot}}$. Since the scales of C_t and EE_t may vary during training, on-line normalization is applied to obtain $\tilde{C}_t = \frac{C_t - C_{\min}}{C_{\max} - C_{\min}}$ and $\tilde{EE}_t = \frac{EE_t - EE_{\min}}{EE_{\max} - EE_{\min}}$, where $(C_{\min}, C_{\max})$ and $(\text{EE}_{\min}, \text{EE}_{\max})$ are running bounds updated during training. In addition, the latency violations are penalized through the positive part of the delay budget as $[L_t - L_{\max}]_+ = \max\{0, L_t - L_{\max}\}$. As a result, the per-step reward is then designed as below:

$$r_t = w_1\tilde{C}_t + w_2\tilde{\text{EE}}_t - w_3[L_t - L_{\max}]_+, \tag{22}$$

where w_1, w_2, w_3 satisfies (12i), which efficiently approximate the MMDP reward to the per-slot contribution of (P1). To further encourage exploration of high-quality RIS configurations at the early stage of learning, a simple shaping can be applied by penalizing extremely low-rate episodes as below:

$$C_t = \begin{cases} C_t, & C_t \geq C_{th} \\ C_t - \Delta_{pen}, & C_t < C_{th} \end{cases} \tag{23}$$

where C_{th} is a design threshold of sum rate and $\Delta_{\text{pen}} > 0$ is a small penalty constant. This shaping of r_t is consistent with the regularization in (23), but now embedded into a multi-objective reward that also accounts for energy and latency.

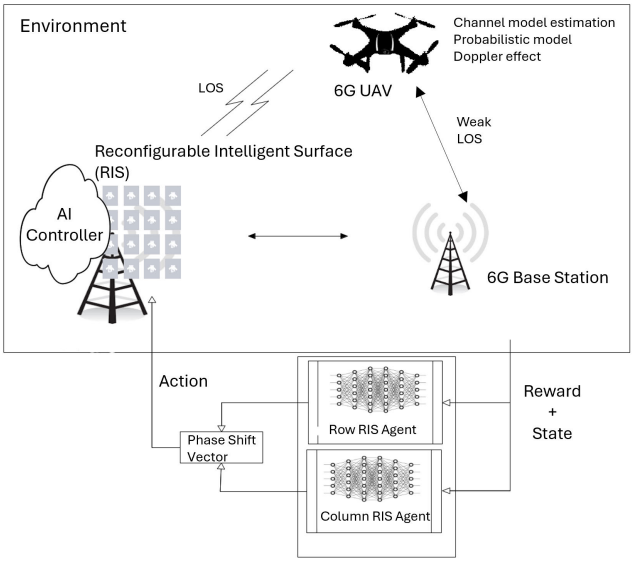


FIGURE 3. The architecture for dual-agent RIS control in uplink UAV communications.

Remark 1 (Specialization to ULA-based RIS): The aforementioned formulation of the action space is derived for a dual-agent RIS controller operating on a UPA, where independent row and column agents adjust the reflective pattern with discrete incremental phase shifts. In the simplified case of a ULA configuration, as employed in the single-agent baseline, the control dimension will be reduced to a single steering parameter κ_t and the corresponding wavenumber k_t. The RIS configuration is then generated from a single 1D phase profile, and the action space $\boldsymbol{\mathcal{A}}$ collapses to a one-dimensional discrete set of steering updates. This reduction decreases the cardinality of the action space and simplifies the learning task, which is particularly useful for initial verification experiments and benchmarking against conventional schemes.

C. DUAL-AGENT RIS CONTROLLER FRAMEWORK

To precisely manage the reflective pattern during UAV communications, a dual-agent RIS controller is developed with MADQN in Section. III-A. The proposed framework divides the phase shift control task into two dedicated agents, including a row RIS agent and a column RIS agent. Each agent independently adjusts the corresponding phase shifts from their own dimension, jointly optimizing propagation efficiency and maximizing the overall sum rate of the uplink communications between the UAVs and BS.

In Figure. 3, a comprehensive architecture has been designed for the proposed dual-agent RIS controller framework within UAV communications. The top frame of the aforementioned figure illustrated the main communication environment, where two LoS path have been built up between UAVs and BS in the uplink of 6G communications. On the one hand, each UAV was connected to BS directly with weak LoS path due to imperfect CSI, Doppler effect, and dense urban obstacles. On the other hand, the reflective

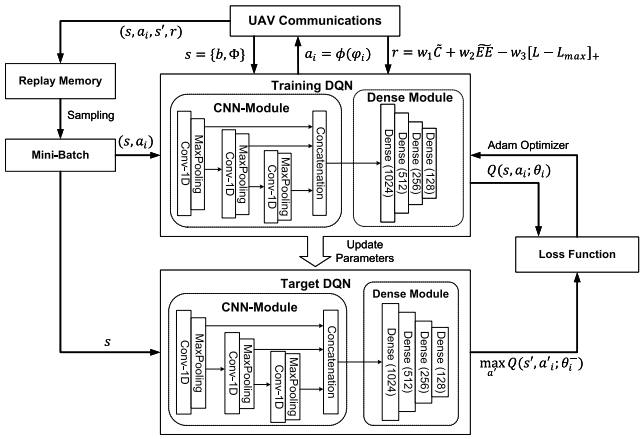


FIGURE 4. The architecture of each agent for RIS controller in Uplink UAV communications.

pattern of RIS bridged another distinct LoS path from each UAV to BS, where the AI controller of RIS can adjust wireless propagation with dynamic phase shifter. As a result, AI controller-based RIS can extend an alternative uplink LoS to compensate the direct uplink LoS within dynamic radio environments, especially around dense urban infrastructures.

Instead, the bottom architecture demonstrated the workflow of the proposed dual-agent RIS controller in detail, where the control task has been distributed to two deep neural network-based agent, including the row RIS agent and the column RIS agent. According to the proposed MMDP, both agents receive CSI and current reflective pattern as feedback from the aforementioned communication environment, so as to calculate the optimal phase shifter from the respective dimension of RIS elements. To unleash more potential of dual-agent in the proposed framework, the local action has been combined into a unified phase shift vector with (21) to reduce the computational complexity of the reward function as (22) and (23). Once obtaining the unified phase shift vector from AI controller, RIS will adapt the alternative LoS to the dynamic wireless communications system by learning how to manipulate the uplink propagation between each UAV and BS. This joint and hierarchical control architecture is particularly designed to address challenges for complexity and scalability posed by large-scale RIS deployments in dynamic UAV communication environments.

To keep the robustness of RIS control, a double deep Q network (DDQN) [44] has been applied to each agent as Figure. 4. The processing pipeline for each agent initiates with a convolutional neural network (CNN) module, which is responsible for extracting high-level spatial and contextual features from the multidimensional environment state, including the UAV-to-BS channel matrix and historical RIS phase configurations. The CNN module is made up of multiple one-dimensional convolutional layers, each of which is followed by concatenation and max-pooling operations. This layered structure enables the agent to efficiently capture both temporal and spatial dependencies inherent in CSI and reflective pattern, which are critical for precise RIS phase shift optimization.

Subsequent to feature extraction, the output of the CNN module is forwarded to a stack of fully connected dense layers. These layers will further abstract the feature representation and model complex nonlinear relationships between the observation and the desired RIS phase adjustments. The outputs of final dense layer serve as the input to mean squared error (MSE) loss function, which computes the Q-values corresponding to the possible discrete actions available to the local agent. Action selection is managed by an ϵ-greedy policy, ensuring an appropriate balance between exploration and exploitation throughout the training process.

To further enhance learning stability and convergence, each agent maintains a separate target DQN network with an identical architecture. The parameters of the target network are periodically synchronized with those of the online DQN network. Within training DQN, the loss function will be extended as below from (18):

$$L_i(\boldsymbol{\theta}_{i,t}) = \mathbb{E}_{(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, r_t, \boldsymbol{s}_{t+1}) \sim (\boldsymbol{s}, \boldsymbol{a}_i, \boldsymbol{r}, \boldsymbol{s}')}[\boldsymbol{y}_{i,t}^{target} - Q_i(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t-1})]^2, \tag{24}$$

where $\boldsymbol{y}_{i,t}^{target}$ means the target value fed to training DQN, which depends on temporal reward r_t and training DQN $Q^{\star}(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}; \boldsymbol{\theta}_{t-1})$ as (19) in general DQN algorithm. To avoid overfitting action selections during training processes, a separate target network is adopted to predict $\boldsymbol{y}_{i,t}^{target}$ as below:
where $\boldsymbol{\theta}^-$ means the neural network parameters from target DQN, which is acquired by periodic transferring from training DQN as Figure. 4. Moreover, to keep the trade-off between exploration and exploitation in training DQN, the local deep Q network $Q_i(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t-1})$ is redefined further as follows:

$$\begin{aligned} Q_i(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t-1}) = (1-\alpha) Q(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t-2}) \\ + \alpha Q(\boldsymbol{s}_t, \boldsymbol{a}_{i,t}, \boldsymbol{\theta}_{i,t-1}), \end{aligned} \tag{25}$$

where α means the critical parameters of Q-function, which will fine-tune the balance between convergence and robustness during training processes. The whole procedure of dual-agent RIS controller has been displayed in Algorithm. 1.

Remark 2 (Computational Complexity): In a single-agent DQN controlling the entire RIS aperture, the action space grows as $|\mathcal{A}| = |\mathcal{A}_r||\mathcal{A}_c|$, and both the forward pass and the Bellman update scale proportionally with this joint cardinality. Consequently, $\mathcal{O}(|\mathcal{A}_r||\mathcal{A}_c|)$ operations per training step are required for evaluating $Q(s, a)$ and computing the target values. In contrast, the proposed dual-agent MADQN factorizes the control problem into two independent Q-networks of size $|\mathcal{A}_r|$ and $|\mathcal{A}_c|$. The forward pass, action selection, and gradient updates therefore scale only as

[1]The latency penalty weight w_3 in (P0) is treated as a Lagrange multiplier in the RL formulation. Instead of fixing w_3, the learning algorithm dynamically updates this quantity via a dual ascent rule denoted as Λ_{lat}, which automatically satisfies the latency constraint during training.

Algorithm 1 Dual-Agent MADQN RIS Controller With Belief-State Representation

Input: Statistical CSI model, RIS geometry, latency budget $L_{\max}$, learning rate η, discount factor γ, number of episodes T, steps per episode K, replay buffers $\mathcal{D}_r, \mathcal{D}_c$, batch size N_{ex}, target update period N_T.

Initialization: Initialize NCF, belief encoder, row/column Q-networks Q_r, Q_c, target networks Q_r^-, Q_c^-, exploration rate ϵ, dual latency weight Λ_{lat} [1].

for $t = 1$ to T **do**
  Sample episode channel H_{true} from statistical CSI model; reset belief state and hidden state.
  for $k = 1$ to K **do**
    Observe noisy pilots Y_k; obtain $(\hat{H}_k, \sigma_k^2, z_k)$ via NCF.
    Update belief state b_k using belief encoder and previous action.
    Row/column agents select actions $a_{r,k}$, $a_{c,k}$ via ϵ-greedy over Q_r, Q_c.
    Apply joint action to update RIS phases Φ_k; compute $(R_k, P_k^{\text{tot}}, \text{EE}_k, L_k)$.
    Compute scalar reward r_k with (22), where metrics are normalized online; obtain time-aware discount γ_k.
    Process next pilots to obtain b_{k+1}; store $(b_k, a_{r,k}, r_k, b_{k+1}, \gamma_k)$ in $\mathcal{D}_r$, and corresponding tuple in $\mathcal{D}_c$.
    if $|\mathcal{D}_r| \geq N_{\text{ex}}$ **and** $|\mathcal{D}_c| \geq N_{\text{ex}}$ **then**
      Sample mini-batches; update Q_r, Q_c by minimizing Bellman loss.
      Soft-update target networks: $\theta_i^- \leftarrow \tau\theta_i + (1-\tau)\theta_i^-$.
    end if
  end for
  Update dual variable: $\Lambda_{\text{lat}} \leftarrow \left[\Lambda_{\text{lat}} + \eta_{\text{dual}}(\overline{L} - L_{\max})\right]_+$.
  Anneal ϵ and learning rate η.
end for

Output: Trained row and column Q-networks Q_r, Q_c, NCF, and belief encoder.

$\mathcal{O}(|\mathcal{A}_r| + |\mathcal{A}_c|)$ per step, rather than the multiplicative joint space. This reduces both computation and memory costs by approximately a factor of $\frac{|\mathcal{A}_r||\mathcal{A}_c|}{|\mathcal{A}_r|+|\mathcal{A}_c|}$, which grows linearly with the action resolution and becomes substantial for large RIS dimensions. The dual-agent formulation is therefore significantly more scalable for real-time RIS control.

IV. RESULTS AND ANALYSIS

In this section, numerical results are derived from simulation experiments to verify the convergence and robustness of the proposed RIS controller against various radio environments.

A. EXPERIMENTAL SETTING

The experimental settings have been summarized in Table. 2, consisting of communication scenarios setup and training parameters setup as below.

1) COMMUNICATION SCENARIOS SETUP

In the ULA configuration, the BS is placed at position $[-60, 60, 10]$ along the y-axis with the number of antennas $N_{BS} = 8$. Instead, RIS is facilitated at position $[0, 0, 10]$ with the UPA configuration in the YoZ plane, whose size is shaped by 128×128. Moreover, 10 UAVs, each with a single antenna, are scheduled to cruise from $[-100, 100, 20]$ to $[100, -100, 20]$, following the trajectory of (1). The noise variance is set as $\sigma_{BS} = 0.1$ at BS, while the noise variance at each UAV is set as $\sigma_{U_i} = 0.5$. According to the real-time location of BS, RIS, and each UAV, the LoS components of channel models are constructed with (7) and (8) separately in ULA configurations. Considering the UPA configuration of RIS, the array response vector of (7) can be extended to (9) for channel modeling of RIS. Instead, the NLoS components of channel models are subject to the Gaussian distribution with $H_{XY}^{NLoS} \sim \mathcal{CN}(0, \mathbf{I})$ for each link $XY \in \{UR, UB, RB\}$.

2) TRAINING PARAMETERS SETUP

The number of channel episodes is set as $T = 3000$, while 20 channel blocks are allocated to each channel episode. The number of experience is applied as $N_{ex} = 200$ with batch size $N_{mb} = 64$. In addition, the period of target update is determined as $N_T = 20$, while the initial learning rate is applied as $\eta = 3 \times 10^{-4}$. The memory storage $|\mathcal{M}|$ is upper-bounded by 10000. For impact factor, the discount factor is set as $\gamma = 0.95$, and the critical parameter of Q function is set as $\alpha = 0.7$. The parameters for training DQN are updated with Adam optimizer. Under the ϵ-greedy policy, the exploration probability is initialized as $\epsilon_0 = 1$.

B. CONVERGENCE PERFORMANCE ANALYSES OF PROPOSED ALGORITHM

To examine the convergence behavior and operational advantages of the proposed framework, we evaluate a broad set of reinforcement learning controllers that include DQN, DDPG, TD3, SAC, MA-DQN, and MADDPG, together with classical baselines such as alternating optimization (AO). Following Remark 1, the single-agent controller adopts a simplified ULA-based RIS with 32 elements, whereas the multi-agent controller operates over a 32×32 UPA in which row and column agents jointly optimize the two-dimensional reflective pattern. Performance assessment focuses on six key metrics, consisting of total return, average sum rate, energy efficiency, power consumption, control latency, and CSI-related NMSE.

The total return curves in Figure 5a reveal that MA-DQN achieves the most stable and highest long-term return among all learning-based methods, demonstrating fast convergence and strong robustness against channel dynamics in RIS-aided UAV communications. TD3 and SAC both converge smoothly but remain below the level achieved by DQN-based methods. MADDPG converges more slowly due to increased non-stationarity in the multi-agent setting,

TABLE 2. Parameter Settings.

Param	Definition	Value
Communication Scenarios in Testbed		
N_{BS}	Number of Antennas at BS	8
N_R	Shape of RIS Elements	128×128
N	Number of UAVs	15
N_{U_i}	Number of Antennas at Each UAV	1
σ^2_{BS}	Noise Variance at BS	0.1
σ^2_i	Noise Variance at Each UAV	0.5
$\boldsymbol{H}^{NLoS}_{-}$	NLoS Channel Matrix	$\mathcal{CN}(0, \mathbf{I})$
–	Position of BS	[-60, 60, 10]
–	Position of RIS	[0, 0, 10]
–	Departure Position of each UAV	[-100, 100, 20]
–	Destination Position of each UAV	[100, -100, 20]
Training Parameters Setup		
T	Number of Channel Episodes	3000
N_{mb}	Mini-batch Size	64
N_{ex}	Number of Experiences	200
N_T	Period of Target Update	20
$\lvert\mathcal{M}\rvert$	Size of Memory Storage	10000
–	Neural Network Optimizer	Adam
–	Types of Learning Rate Decay	Exponential Decay
ϵ_0	Initial Exploration Probability	1
$\epsilon_{\min}$	Minimum Exploration Probability	0.1
ϵ_{decay}	Decay Rate of Exploration Probability	0.995
K	Number of Channel Blocks	20
η_0	Initial Learning Rate	3×10^{-4}
η_{decay}	Decay Rate of Learning Rate	0.92
γ	Discount Factor	0.95
α	Critical Parameter of Q Function	0.7

while DDPG exhibits the lowest stability across episodes. These results indicate that DQN remains highly effective for RIS phase-shift control, even when extended to multi-agent cooperative settings.

Figure 5b shows that MA-DQN consistently achieves the highest sum rate among the learning-based methods and approaches the AO upper bound, which leverages full CSI and iterative model-based optimization. TD3 and SAC deliver moderate performance, while DDPG and MADDPG indicate noticeable gaps due to their sensitivity to noise in continuous policy updates. The multi-agent framework enhances performance relative to single-agent configurations by enabling coordinated row–column phase-shift adaptation across the UPA, which improves spatial controllability and yields a clear performance gain.

The energy efficiency is compared in Figure 5c, which further illustrates the superiority of MA-DQN under the multi-agent control architecture. MA-DQN maintains higher bit-per-Joule efficiency due to avoid unnecessary phase-shifting operations and stabilize earlier in training. SAC and TD3 both achieve moderate efficiency, while DDPG and MADDPG exhibit reduced performance due to their less consistent convergence behavior and higher action exploration overhead.

Figure 5d demonstrates that MA-DQN maintains the lowest long-term power consumption among all learning-based methods apart from AO, which is optimized with full CSI but is computationally impractical for real-time UAV scenarios. TD3 and SAC both offer relatively stable power usage, whereas DDPG and MADDPG suffer from fluctuations driven by unstable policy updates. The multi-agent structure again improves efficiency by distributing control across spatial dimensions, which reduces redundant phase adjustments.

Figure 5e shows that MA-DQN and DQN exhibit the strongest robustness to CSI estimation errors, achieving the lowest NMSE among learning-based methods. The discrete policy structure of them enables stable mapping between noisy observations and RIS phase actions. SAC and TD3 achieve moderate robustness, while DDPG and MADDPG are more sensitive to imperfect CSI due to the instability of continuous gradient-based learning under noisy observations. Regarding full and accurate CSI, AO shows the highest sensitivity to estimation imperfections.

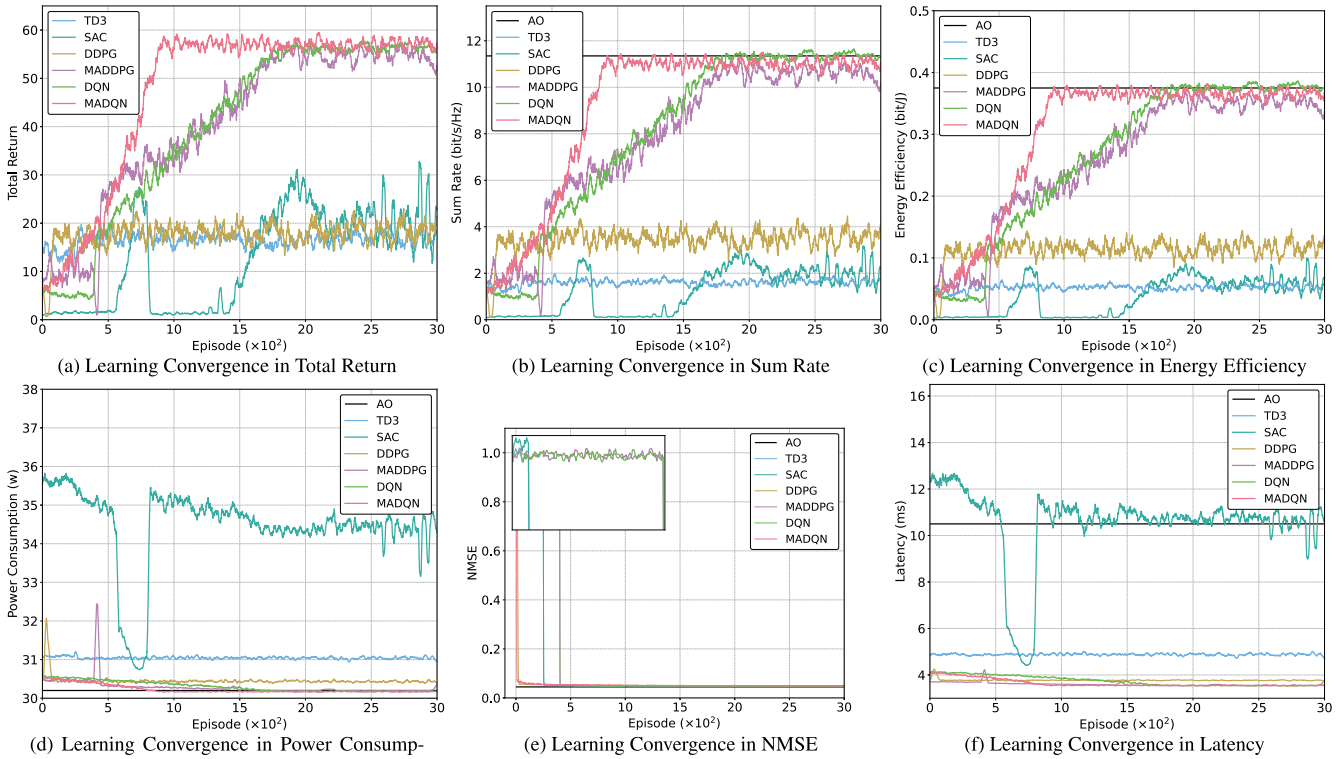
The latency results in Figure 5f highlight that MA-DQN achieves the lowest inference and control latency among all learning-based approaches because of the efficient discrete-action design and compact Q-network structure. Continuous-control algorithms such as DDPG, TD3, and SAC incur higher computational costs due to action-space projection and gradient-based updates. MADDPG experiences the highest latency among the learning-based schemes due to decentralized critic updates and the enlarged joint-action space. AO suffers from the largest latency and is unsuitable for real-time UAV-assisted RIS control.

C. ROBUSTNESS PERFORMANCE ANALYSES OF PROPOSED ALGORITHM

This subsection evaluates the robustness of the proposed MA-DQN-based RIS controller under varying propagation conditions, signal quality, network density, and RIS dimensions. The analysis considers the effects of the Rician factor, SNR, number of UAVs, and RIS size on the convergence behaviour of the total return across 3000 training episodes.

Figure 6 identifies the impact of the Rician factor on the total return. The Rician factor determines the relative strength of the deterministic line-of-sight (LoS) component compared with multipath scattering. When the Rician factor is very small, the training curves show noticeable fluctuations because the agents must operate in highly stochastic fading environments. As the Rician factor increases, the environment becomes more predictable and the return gradually stabilizes. For large Rician factors, the algorithm converges rapidly with minimal oscillations, indicating that the MA-DQN controller can exploit more reliable CSI information. These observations confirm that the proposed method remains stable even when operating under severe fading conditions.

Figure 7 presents the effect of SNR on the learning behaviour. The curves for different SNR levels are highly similar, indicating that the total return is largely insensitive to SNR variations within the tested range. The MA-DQN controller converges to nearly the same return value in all cases. This observation suggests that the adopted state representation and reward design allow the agent to maintain stable learning performance even when observation noise



(a) Learning Convergence in Total Return (b) Learning Convergence in Sum Rate (c) Learning Convergence in Energy Efficiency

(d) Learning Convergence in Power Consumption (e) Learning Convergence in NMSE (f) Learning Convergence in Latency

FIGURE 5. Performance analyses of MADQN-based RIS controller.

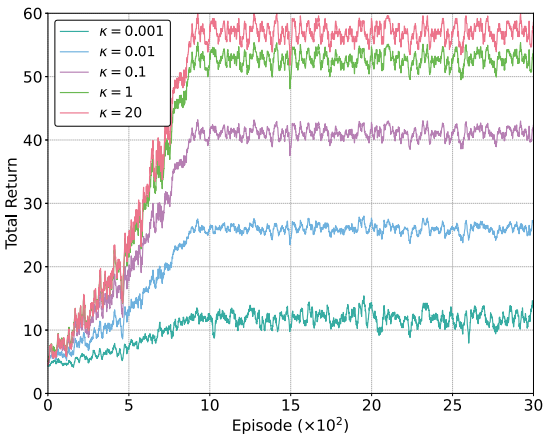


FIGURE 6. Performance comparison of MA-DQN against κ.

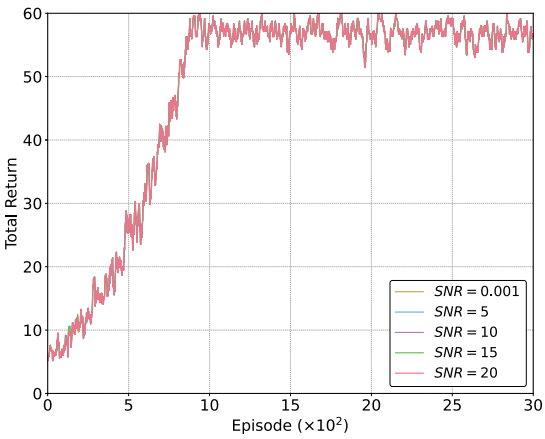


FIGURE 7. Performance comparison of MA-DQN against SNR.

varies. The algorithm is therefore robust to SNR differences and does not rely heavily on high signal quality to achieve effective policy learning.

Figure 8 presents the total return for different numbers of UAVs in the system. With a small number of UAVs, the convergence is fast but the overall return remains limited because the communication demand and achievable multi-user gains are relatively small. As the number of UAVs increases, the total return rises significantly due to richer multi-user diversity and higher aggregated throughput potential. This indicates larger numbers of UAVs introduce greater interaction complexity and lead to more variation in the early training stages, but the MA-DQN controller still converges reliably in all cases. These results show that the algorithm maintains robust performance under increasing network load.

Figure 9 investigates the effect of RIS size on the learning outcome. Smaller RIS dimensions provide limited beamforming gain, resulting in lower returns and more variability. As the RIS size increases, the number of controllable reflecting elements grows and the agents gain more spatial freedom for phase-shift optimization. This leads to higher long-term returns and smoother convergence behaviour, so the larger RIS configurations can converge

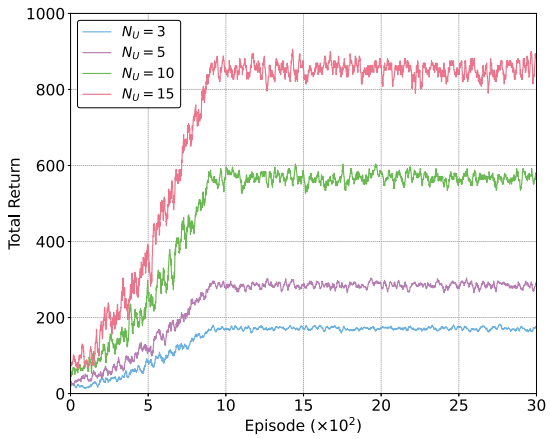


FIGURE 8. Performance comparison of MA-DQN against N_U.

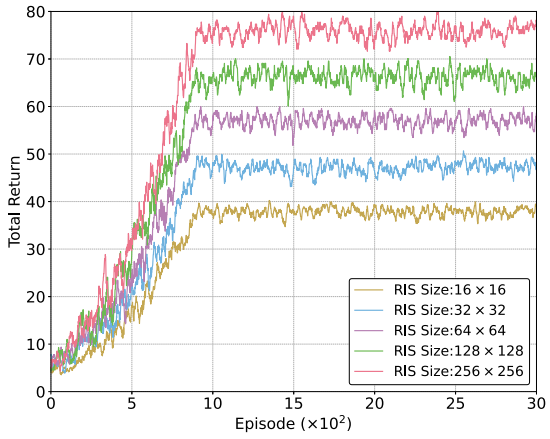


FIGURE 9. Performance comparison of MA-DQN against RIS Size.

rapidly and reach the higher return values. The performance improvement with RIS size confirms that the MA-DQN controller scales effectively with the dimensionality of the RIS and is capable of revealing the additional degrees of freedom offered by large passive arrays.

Overall, the results demonstrate that the proposed MA-DQN framework maintains stable and consistent convergence under diverse channel conditions, noise levels, user populations, and RIS dimensions. The algorithm adapts effectively to increasing system complexity and consistently approaches high-return solutions under dynamic and realistic UAV-assisted RIS deployments.

V. CONCLUSION

In this paper, we proposed an MA-DQN–based RIS controller for uplink UAV communications, where a row–column agent structure enables scalable optimization of quantized RIS phase shifts under dynamic channel conditions. By reformulating the communication objective as an MMDP and integrating a dual-agent learning strategy, the framework achieved superior convergence, achievable capacity, and robustness compared with existing baselines across varied Rician fading, SNR levels, UAV densities, and RIS sizes. However, some import limitations are still remaining from practical considerations such as RIS hardware constraints, phase-switching delays, and UAV-side computational overhead. Future work will incorporate these implementation factors and extend the controller design toward near-field beamforming and large-scale multi-RIS deployments.

REFERENCES

[1] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, J. S. Thompson, E. G. Larsson, M. D. Renzo, W. Tong, P. Zhu, X. Shen, H. V. Poor, and L. Hanzo, "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 905–974, 2nd Quart., 2023.

[2] X. Wang, Z. Zhang, D. He, K. Guan, D. Liu, J. Dou, S. Mumtaz, and S. Al–Rubaye, "A multi–task learning model for super resolution of wireless channel characteristics," in *Proc. IEEE Global Commun. Conf.*, Mar. 2022, pp. 952–957.

[3] N. Abosata, S. Al-Rubaye, and G. Inalhan, "Lightweight payload encryption-based authentication scheme for advanced metering infrastructure sensor networks," *Sensors*, vol. 22, no. 2, p. 534, Jan. 2022.

[4] S. Al-Rubaye, J. Rodriguez, L. Z. Fragonara, P. Theron, and A. Tsourdos, "Unleash narrowband technologies for industrial Internet of Things services," *IEEE Netw.*, vol. 33, no. 4, pp. 16–22, Jul. 2019.

[5] A. Warrier, S. Al-Rubaye, D. Panagiotakopoulos, G. Inalhan, and A. Tsourdos, "Interference mitigation for 5G-connected UAV using deep Q-learning framework," in *Proc. IEEE/AIAA 41st Digit. Avionics Syst. Conf. (DASC)*, Sep. 2022, pp. 1–8.

[6] H. Whitworth, S. Al–Rubaye, A. Tsourdos, J. Jiggins, N. Silverthorn, and K. Thomas, "Aircraft to operations communication analysis and architecture for the future aviation environment," in *Proc. IEEE/AIAA 40th Digit. Avionics Syst. Conf. (DASC)*, Mar. 2021, pp. 1–8.

[7] V. Towhidlou, S. Al-Rubaye, and A. Tsourdos, "LTE handover design for cellular-connected aircraft," in *Proc. IEEE/AIAA 41st Digit. Avionics Syst. Conf. (DASC)*, Sep. 2022, pp. 1–5.

[8] S. Al-Rubaye, A. Al-Dulaimi, and Q. Ni, "Power interchange analysis for reliable vehicle-to-grid connectivity," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 105–111, Aug. 2019.

[9] P. S. Bithas, V. Nikolaidis, A. G. Kanatas, and G. K. Karagiannidis, "UAV-to-ground communications: Channel modeling and UAV selection," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5135–5144, Aug. 2020.

[10] X. Tang, F. Chen, F. Wang, and Z. Jia, "Disaster-resilient emergency communication with intelligent air–ground cooperation," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 5331–5346, Feb. 2024.

[11] X. Liu, Y. Yu, F. Li, and T. S. Durrani, "Throughput maximization for RIS-UAV relaying communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19569–19574, Oct. 2022.

[12] F. Alhashmi, S. Al-Rubaye, and A. Tsourdos, "Advanced mobility flight dynamics restriction to support high availability communication systems," in *Proc. AIAA DATC/IEEE 43rd Digit. Avionics Syst. Conf. (DASC)*, Sep. 2024, pp. 1–10.

[13] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2016.

[14] H. Mei, K. Yang, Q. Liu, and K. Wang, "3D-trajectory and phase-shift design for RIS-assisted UAV systems using deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3020–3029, Mar. 2022.

[15] X. Gu, W. Duan, G. Zhang, M. Wen, J. Choi, and P.-H. Ho, "Aerial reconfigurable intelligent surface-assisted terrestrial communications," *IEEE Internet Things Mag.*, vol. 7, no. 2, pp. 54–60, Mar. 2024.

[16] A. Araghi, M. Khalily, M. Safaei, A. Bagheri, V. Singh, F. Wang, and R. Tafazolli, "Reconfigurable intelligent surface (RIS) in the sub-6 GHz band: Design, implementation, and real-world demonstration," *IEEE Access*, vol. 10, pp. 2646–2655, 2022.

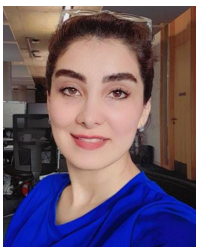
[17] A. Bagheri, M. Safaei, A. Araghi, S. M. Shahabi, F. Wang, M. Khalily, and R. Tafazolli, "Mathematical model and real-world demonstration of multi-beam and wide-beam reconfigurable intelligent surface," *IEEE Access*, vol. 11, pp. 19613–19621, 2023.

[18] W. Sun, Z. Xu, M. Zhu, L. Li, Y. Ni, and H. Zhao, "Achievable rate maximization for RIS-UAV assisted vehicular communication network," in *Proc. IEEE 24th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2023, pp. 306–310.

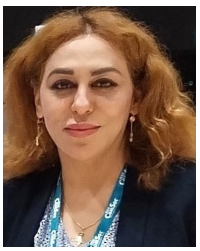
[19] G. Sun, B. Liu, J. Li, S. Liang, H. Pan, and X. Zheng, "Enabling urban mmWave communications with UAV-carried IRS via deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2024, pp. 4985–4990.
[20] Y. Chen, S. Al-Rubaye, A. Tsourdos, K.-F. Chu, Z. Wei, L. Baker, and C. Gillingham, "Federated deep reinforcement learning-based intelligent surface configuration in 6G secure airport networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 10, pp. 17304–17320, Oct. 2025.
[21] S. Subasu, S. Al-Rubaye, A. Warrier, and H. Whitworth, "Unmanned air vehicle operational framework for public safety communications," in *Proc. AIAA DATC/IEEE 43rd Digit. Avionics Syst. Conf. (DASC)*, Sep. 2024, pp. 1–9.
[22] H. Zhang, M. Huang, H. Zhou, X. Wang, N. Wang, and K. Long, "Capacity maximization in RIS-UAV networks: A DDQN-based trajectory and phase shift optimization approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2583–2591, Apr. 2023.
[23] S. Al-Rubaye, A. Al-Dulaimi, S. Mumtaz, and J. Rodriguez, "Dynamic pricing mechanism in smart grid communications is shaping up," *IEEE Commun. Lett.*, vol. 22, no. 7, pp. 1350–1353, Jul. 2018.
[24] S. Yi, F. Wang, and X. Zhang, "Hybrid multi-agent deep reinforcement learning for active-IRS-based rate maximization over 6G UAV mobile wireless networks," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2023, pp. 858–863.
[25] S. Shang, T. Zhang, D. Xu, L. Liu, C. Wu, S. Mumtaz, and C. Yuen, "Joint trajectory and beamforming optimization for UAV-RIS-empowered multiuser communication networks: A double deep Q-network approach," *IEEE Trans. Veh. Technol.*, vol. 74, no. 10, pp. 1–15, Oct. 2025.
[26] C. You, Z. Kang, Y. Zeng, and R. Zhang, "Enabling smart reflection in integrated air-ground wireless network: IRS meets UAV," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 138–144, Dec. 2021.
[27] H.-B. Jeon, S.-H. Park, J. Park, K. Huang, and C.-B. Chae, "An energy-efficient aerial backhaul system with reconfigurable intelligent surface," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6478–6494, Aug. 2022.
[28] P. S. Aung, L. X. Nguyen, Y. K. Tun, Z. Han, and C. S. Hong, "Aerial STAR-RIS empowered MEC: A DRL approach for energy minimization," *IEEE Wireless Commun. Lett.*, vol. 13, no. 5, pp. 1409–1413, May 2024.
[29] L. Yang, F. Meng, J. Zhang, M. O. Hasna, and M. D. Renzo, "On the performance of RIS-assisted dual-hop UAV communication systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10385–10390, Sep. 2020.
[30] Y. Xu, T. Zhang, Y. Liu, D. Yang, L. Xiao, and M. Tao, "Computation capacity enhancement by joint UAV and RIS design in IoT," *IEEE Internet Things J.*, vol. 9, no. 20, pp. 20590–20603, Oct. 2022.
[31] A. B. M. Adam, X. Wan, M. A. M. Elhassan, M. S. A. Muthanna, A. Muthanna, N. Kumar, and M. Guizani, "Intelligent and robust UAV-aided multiuser RIS communication technique with jittering UAV and imperfect hardware constraints," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10737–10753, Aug. 2023.
[32] H. Ren, Z. Zhang, Z. Peng, L. Li, and C. Pan, "Energy minimization in RIS-assisted UAV-enabled wireless power transfer systems," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5794–5809, Apr. 2023.
[33] N. T. Nguyen, V.-D. Nguyen, H. Van Nguyen, Q. Wu, A. Tölli, S. Chatzinotas, and M. Juntti, "Fairness enhancement of UAV systems with hybrid active-passive RIS," *IEEE Trans. Wireless Commun.*, vol. 23, no. 5, pp. 4379–4396, May 2024.
[34] P.-H. Chou, B.-R. Zheng, W.-J. Huang, W. Saad, Y. Tsao, and R. Y. Chang, "Deep reinforcement learning-based precoding for multi-RIS-aided multiuser downlink systems with practical phase shift," *IEEE Wireless Commun. Lett.*, vol. 14, no. 1, pp. 23–27, Jan. 2025.
[35] B. Adhikari, A. S. Khwaja, M. Jaseemuddin, A. Anpalagan, and A. Nallanathan, "Energy efficient RIS-assisted UAV networks using twin delayed DDPG technique," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18423–18439, Dec. 2024.
[36] Y. Zhu, J. Zhang, E. Shi, Z. Liu, C. Yuen, D. Niyato, and B. Ai, "Joint precoding and phase shift design for RIS-aided cell-free massive MIMO with heterogeneous-agent trust region policy," *IEEE Trans. Veh. Technol.*, vol. 74, no. 1, pp. 1794–1799, Jan. 2025.
[37] M. Zeng, X. Ning, W. Wang, Q. Wu, and Z. Fei, "RIS aided NR-U and Wi-Fi coexistence in single cell and multiple cell networks on unlicensed bands," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 3, pp. 1528–1541, Sep. 2023.
[38] C. Chen, S. Gong, B. Gu, L. Li, Z. Liu, N. C. Luong, and D. Niyato, "Learning RIS-assisted hybrid NOMA transmissions for capacity maximization in wireless-powered networks," *IEEE Trans. Veh. Technol.*, early access, Aug. 11, 2025, doi: 10.1109/TVT.2025.3597318.
[39] A. Abdallah, A. Celik, M. M. Mansour, and A. M. Eltawil, "Multi-agent deep reinforcement learning for beam codebook design in RIS-aided systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7983–7999, Jul. 2024.
[40] A. M. Huroon, G. B. Tarekegn, A. M. A. Abdo, A. Amjad, L.-C. Tai, and L.-C. Wang, "Multiagent deep reinforcement learning for AAV-RIS-assisted integrated sensing and communication," *IEEE Internet Things J.*, vol. 12, no. 19, pp. 40083–40097, Oct. 2025.
[41] Y. Chen, S. Al-Rubaye, A. Tsourdos, H. Li, X. Shi, Z. Wei, Z. Liu, L. Baker, and C. Gillingham, "Meta-hierarchical reinforcement learning based beamforming for near-field multi-user communications," *IEEE Trans. Wireless Commun.*, early access, 2025.
[42] Y.-C. Chuang, W.-Y. Chiu, R. Y. Chang, and Y.-C. Lai, "Deep reinforcement learning for energy efficiency maximization in cache-enabled cell-free massive MIMO networks: Single- and multi-agent approaches," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10826–10839, Aug. 2023.
[43] H. Wang, Y. Yu, and Y. Jiang, "Fully decentralized multiagent communication via causal inference," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10193–10202, Dec. 2023.
[44] H. v. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 2094–2100.



YANG CHEN received the B.Eng. degree in transportation from the Civil Aviation University of China, the M.S. degree in aerospace from Cranfield University, U.K., where he is currently pursuing the Ph.D. degree with the Advanced Connectivity and System Integration Research Group, Faculty of Engineering and Applied Sciences. His current research interests include 6G connectivity and security and machine learning.



HANIEH AHMADI is currently pursuing the Ph.D. degree in electronic engineering with the Institute for Safe Autonomy (ISA), University of York. She was a Research Assistant of 6G wireless communications with the Faculty of Engineering and Applied Sciences, Cranfield University, where she is contributed to the Communications Hub for Empowering Distributed Cloud Computing Applications (CHEDDAR) Project. Her research interests include 6G wireless and satellite communications, ORAN, reconfigurable intelligent surfaces (RIS), and antenna design.



SABA AL-RUBAYE (Life Senior Member, IEEE) is currently the Chair Professor of telecommunications and autonomous systems with the Faculty of Engineering and Applied Sciences, Cranfield University, also a member of expert at the UKTIN-UK Telecoms Innovation Network. She leads the development of 6G technologies and their applications in future communication networks, particularly in support of space and air mobility. She contributed to the development of industry standards, including IEEE P1932.1 for License/Unlicensed Interoperability and IEEE P1920.2 for Vehicle-to-Vehicle Communications for Uncrewed Aircraft Systems. She has been actively involved in various high-profile projects funded by Innovate U.K., Research England, EPSRC, the Department for Transport, the Natural Sciences and Engineering Research Council of Canada (NSERC), and U.S. Government. Her research has been widely recognized, winning best technical paper awards at the IEEE Vehicular Technology conferences in 2011 and 2015, and the best paper awards at the IEEE DASC conference in 2020, 2021, and 2024. She is registered as a Chartered Engineer (CEng), MIET.



https://dspace.org/

Multi-agent deep reinforcement learning-based RIS-aided UAV communications

Chen, Yang; Ahmadi, Hanieh; Al-Rubaye, Saba

2026
Institute of Electrical and Electronics Engineers (IEEE)
https://doi.org/10.1109/access.2025.3649591
Chen Y, Ahmadi H, Al-Rubaye S. (2026) Multi-agent deep reinforcement learning-based RIS-aided UAV communications. IEEE Access, Volume 14, 2026, pp. 1522-1536

Downloaded from a DSpace repository. https://dspace.org/