Hybridized MA-DRL for Serving xURLLC with Cognizable RIS and UAV Integration

Anal Paul, *Member, IEEE*, Raviteja Allu, *Student Member, IEEE*, Keshav Singh, *Member, IEEE*, Chih-Peng Li, *Fellow, IEEE*, and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—This work proposes a new model of reconfigurable intelligent surface (RIS) called cognizable RIS (CRIS) that is specifically designed to meet the unique demands of users who require extreme-ultra-reliable and low-latency Communication (xURLLC) in the sixth generation (6G) wireless networks. The programmable elements in the proposed CRIS unit can adapt to different modes of operation to provide significant performance gain. To improve reliability at the receiver, we integrate unmanned aerial vehicles with the CRIS module, which enhances network performance through beamforming and mobility. Our study focuses on maximizing the sum throughput in a multiple-input multiple-output scenario using the rate-splitting multiple access communication system. To achieve this, we introduce a novel hybridized multi-agent-based deep reinforcement learning (DRL) algorithm for optimal resource allocation that maximizes the sum throughput. We incorporate long-short-term memory (LSTM) networks into our proposed DRL to address the temporal dependencies due to stochastic channel conditions. By utilizing the proposed LSTM-based multi-agent DRL (MA-DRL) algorithm, we achieve notable gains of 11.7% and 26.9% in sum throughput over widely recognized DRL benchmark algorithms, all while adhering to xURLLC's stringent maximum packet error probability constraint of 10^{-9}.**

***Index Terms*—Rate-splitting multiple access, reconfigurable intelligent surface, extreme ultra-reliable low-latency communication, long short-term memory, deep reinforcement learning.**

I. Introduction

EXtreme ultra-reliable and low-latency communication (xURLLC) is a service category in the sixth generation (6G) wireless communication networks designed to meet the stringent quality-of-service (QoS) requirements of latency-sensitive applications, e.g., self-driving, telemedicine, industrial automation, and other real-time communication services [1]. The primary objective of xURLLC is to ensure that the communication link should be available with very high reliability (e.g., 99.999999%) and provide end-to-end latency as low as 0.1 milliseconds (ms) [2]. These challenging requirements require novel communication frameworks, network architectures, and resource allocation schemes that can adapt to the dynamic nature of wireless channels and efficiently handle interference [3].

Several advanced technologies such as reconfigurable intelligent surfaces (RISs) [4]–[6], THz communication [4], unmanned aerial vehicle (UAV) [5], [7], massive multiple-input multiple-output (MIMO) [8], deep reinforcement learning (DRL) [9]–[11], along with non-orthogonal multiple access (NOMA) [12], [13], space division multiple access (SDMA), or rate splitting multiple access (RSMA) schemes [14] are reported for conventional ultra-reliable and low-latency communication (URLLC) services. However, the existing literature rarely contains significant substantive work exploring the potential of the aforementioned technologies or new network paradigms that have met the requirement for xURLLC services up to now [1]–[3].

A. Motivations of the Present Work

RISs are recognized for their energy efficiency and environmental adaptability [4], [5]. In contrast, UAVs are noted for their deployment flexibility and capability to establish direct line-of-sight (LoS) links [5], [7]. This system utilizes the additional degrees of freedom from RISs, in conjunction with UAVs, to optimize transmission quality among terrestrial nodes. The integration of UAVs and RISs has the potential to enhance data delivery in stringent wireless scenarios. The study in [15] evaluated an integrated RIS-UAV relay system. The approach of mounting RISs on UAVs has demonstrated substantial performance improvements in NOMA and orthogonal frequency division multiple access (OFDMA) communication systems [16].

In the literature, several works studied the efficiency of the UAV-mounted RIS systems [17]–[20]. Zhai *et al.* introduced an innovative UAV-mounted RIS (U-RIS) system that combines the benefits of UAVs and RIS for enhanced task offloading, where user signals are efficiently reflected towards a ground mobile edge computing server via the U-RIS [17]. The study in [18] further employed UAV-mounted RIS technology to reduce hardware requirements and signal processing complexities on the UAV, thereby elevating the network's energy efficiency and coverage. Xiao *et al.* [19] investigated a solar-powered UAV-mounted RIS system, aiming to broaden network access by optimizing beamforming, UAV trajectory, and addressing system constraints. Meanwhile, the research presented in

The work of K. Singh and C.-P. Li was supported in part by Qualcomm through a Taiwan University Research Collaboration Project and also supported in part by the National Science and Technology Council of Taiwan under Grants NSTC 112-2221-E-110-038-MY3 and NSTC 112-2811-E-110-018-MY3. The work of T. Q. Duong was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109. (*Corresponding author: Keshav Singh.*)

A. Paul, R. Allu, K. Singh, and C.-P. Li are with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (Email: apaul@ieee.org, d093070005@nsysu.edu.tw, keshav.singh@mail.nsysu.edu.tw, cpli@faculty.nsysu.edu.tw).

T. Q. Duong is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada, and with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN Belfast, U.K. (Email: tduong@mun.ca).

[20] utilized UAV-mounted RIS to maximize the data rate across multiple users and reduce the UAV's flight energy consumption, employing a trained online learning strategy. Collectively, these works highlight the adaptable utility of UAV-mounted RIS in revolutionizing network performance, energy efficiency, and user connectivity in modern wireless communication systems.

It is worth mentioning that our extensive literature study indicates that RISs predominantly operate in a static mode, such as passive, active, or simultaneously transmitting and reflecting (STAR) [21]. This time-invariant mode of operation of RISs might not be effective for supporting xURLLC services, especially in dynamically varying channel conditions and when the channel state information (CSI) is not fully available. Consequently, we are motivated to re-evaluate the operational principles of RISs, allowing each programmable element (PE) (i.e., made by meta-materials [4]) to operate independently in a designated mode.

The independent trajectory control of UAV-assisted RISs and optimal resource distribution to the base station, UAVs, and RISs become imperative in handling the dynamic nature of xURLLC networks. Given their inherent characteristics, DRL algorithms represent a promising solution for embedded optimization and real-time decision-making in wireless contexts [22]–[24]. Utilizing multi-agent-based DRL can enhance the service to latency-sensitive users (LSUs) amidst diverse network constraints [7]. Within xURLLC service domains, adopting a multi-agent DRL (MA-DRL) paradigm for overseeing UAV-assisted RIS systems can substantially optimize UAV trajectories and positioning strategies. This ensures robust communication link establishment between the base station and LSUs while minimizing latency and other potential bottlenecks [25].

The cooperative multi-agent reinforcement learning (MARL) problem, handled by [26] through the innovative value-decomposition network (VDN), aimed to resolve the "lazy agent" issue and partial observability challenges by proposing a single joint reward signal. Advancing from VDN, the work in [27] introduced QMIX, an MA-DRL strategy that outperforms by allowing a richer representation of joint action-value functions. This enhancement is primarily due to a monotonicity constraint, enabling a more sophisticated handling of agent interactions [27]. While both VDN and QMIX focus on discrete action spaces and promote centralized training with an aim on decentralized execution, the work in [28] extended the MARL scope to continuous actions with the multi-agent deep deterministic policy gradient (MADDPG) algorithm. Notably, MADDPG introduced training with policy ensembles to increase robustness and adaptability, a feature not emphasized by VDN or QMIX. The authors in [29] employed the MADDPG algorithm to maximize the total data rate and simultaneously minimize the total communication power in UAV networks. However, MADDPG [28], being an actor-critic method that uses deterministic policy gradient, is computationally intensive and sometimes harder to scale to large numbers of agents. The multi-agent proximal policy optimization (MAPPO) [30] then builds on and refines the approaches of VDN, QMIX, and MADDPG by utilizing proximal policy optimization to make stable policy updates, offering an effective solution for dynamic and unpredictable environments. The authors in [31] used MAPPO to activate base stations in a heterogeneous network dynamically, optimizing energy efficiency and service quality by adapting to real-time environmental conditions.

B. Contributions of this Work

This study proposes a cognizable RIS (CRIS) model in which individual RIS element (i.e., PE) can change their operating modes to provide time-variant, ultra-reliable QoS requirements to assist the xURLLC users. We assume that each CRIS module has N independent PEs. The wireless signal incident on the n-th PE of a CRIS is denoted by s_n, where $n \in \mathcal{N} \triangleq \{1, 2, \ldots, n, \ldots N\}$. The n-th PE transmits and reflects the following signals:

$$\mathtt{t}_n = \phi_n^{\mathtt{t}} s_n, \text{ and } \mathtt{r}_n = \phi_n^{\mathtt{r}} s_n, \tag{1}$$

where, $\phi_n^d = \left(\sqrt{\eta_n^d} e^{j\theta_n^d}\right)$, $\forall d \in \mathcal{D} \triangleq \{\mathtt{r}, \mathtt{t}\}$. Here, $d = \mathtt{r}$ represents the reflection, and $d = \mathtt{t}$ represents the transmission regions of CRIS. $\eta_n^{\mathtt{t}} \in [0\ \eta_n^{\max}]$, $\theta_n^{\mathtt{t}} \in [0\ 2\pi]$ and $\eta_n^{\mathtt{r}} \in [0\ \eta_n^{\max}]$, $\theta_n^{\mathtt{r}} \in [0\ 2\pi]$ denote the amplitude and phase shift response of the n-th PE's transmission and reflection coefficients, respectively. From Fig. 1, the types of operating modes of each PE in a CRIS include a) passively transmitting (PT), b) actively transmitting (AT), c) passively reflecting (PR), d) passively simultaneously transmitting and reflecting (PSTAR), e) passive reflecting actively transmitting (PRAT), f) actively reflecting (AR), g) actively reflecting passively transmitting (ARPT), and h) active STAR (ASTAR). At a specific time instant (t), the individual PE in CRIS can operate in any one mode by controlling ${\eta_n^{\mathtt{t}}}^{\max}$ and ${\eta_n^{\mathtt{r}}}^{\max}$ as shown in Table I. We deploy these CRIS units using independent UAVs in a cell area to serve xURLLC users.

TABLE I: CRIS's element parameters in different operating modes

Mode	${\eta_n^{\mathtt{r}}}^{\max}$	${\eta_n^{\mathtt{t}}}^{\max}$	Mode	${\eta_n^{\mathtt{r}}}^{\max}$	${\eta_n^{\mathtt{t}}}^{\max}$
a) PT	0	1	e) PRAT	1	> 1
b) AT	0	> 1	f) AR	> 1	0
c) PR	1	0	g) ARPT	> 1	1
d) PSTAR	1	1	h) ASTAR	> 1	> 1

With each PE of the RIS having multiple modes of operation, the configuration space is vast. Finding the optimal configuration becomes imperative to harness the full potential of RIS and ensure seamless communication. The problem of determining the best configuration for a given RIS scenario can be viewed as a combinatorial optimization problem. Given n programmable elements and each element with 8 possible modes, the total number of combinations is 8^n. The problem of finding the best configuration for a CRIS with a vast configuration space is likely NP-hard. As n grows, even a modest increase in the number of elements can result in exponential growth in potential configurations, making traditional optimization techniques computationally intensive or even infeasible. Using traditional methods such as an exhaustive search like Brute-force, heuristic search techniques like particle swarm optimization (PSO) and genetic algorithms (GA),

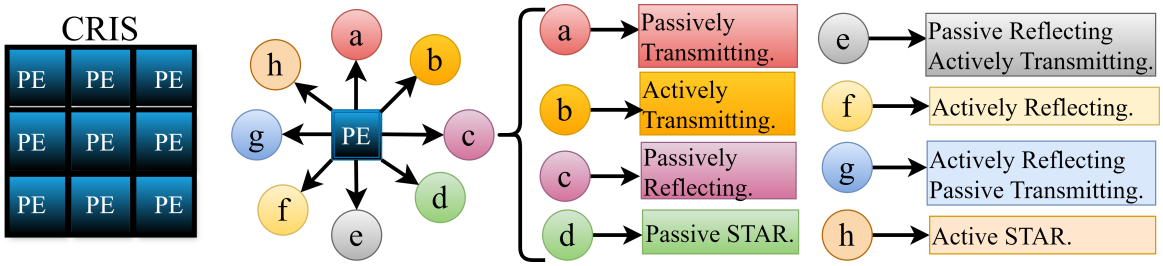


Fig. 1: Each PE of the CRIS module can function independently in any one mode at time instant t.

or gradient-based optimization are not entirely convincing to address the present RIS configuration problem [32], [33]. PSOs and GAs, or gradient-based optimizations, are faster than Brute-force policy; however, they cannot make adaptive decisions in a time-variant environment, which is crucial given the dynamic nature of wireless communication scenarios. [32], [33]. The promising DRL emerges as a compelling alternative as DRL can handle large state and action spaces and continuously learns and adapts its strategies based on interactions with the environment [22], [23], [34], [35]. DRL strikes an effective balance between exploration and exploitation, ensuring that the CRIS configuration does not stagnate at local optima but consistently seeks improved solutions. With its compatibility with modern computational hardware, DRL further strengthens its position as a robust and scalable solution for the complex challenge of RIS optimization [11].

In a time-variant wireless channel, we propose an MA-DRL algorithm scheme to adapt the dynamic operating mode, UAVs' position, and optimal resource allocation to maximize the sum throughput of all xURLLC users. Each UAV-mounted CRIS module employs a dedicated agent for determining all the PEs' operating configurations, optimizing the decision parameters and its UAV's positions. Using common control channels, all spatially distributed agents keep updating the BS with various information, such as current position, moving direction, and channel state estimation.

Our proposed MA-DRL framework significantly advances the MARL by incorporating a sophisticated multi-dimensional action and state space, explicitly capturing both spatial decisions and temporal dynamics to adjust UAV and CRIS parameters precisely. It includes a carefully designed reward function that balances throughput maximization, latency minimization, and power efficiency, facilitating sophisticated policy learning appropriate to the complexities of UAV-mounted CRIS communications. Unlike existing models like VDN [26], QMIX [27], MADDPG [28], [29], and MAPPO [30], [31], our framework emphasizes adaptive and dynamic interaction with the environment and integrates an advanced coordinated multi-agent strategy. Whereas in VDN and QMIX, the focus is predominantly on joint action-value functions without detailing the operational collaboration between different types of agents. Our architecture employs a multi-agent coordination mechanism that enables cooperative interactions among UAV-CRIS units and the base station, enhancing overall system performance through cooperative decision-making. This approach enables effective collaboration among agents while optimizing system-wide objectives, significantly improving over previous methods. Our proposed MA-DRL framework enhances the capabilities of MAPPO [30], [31] by incorporating both Kullback-Leibler (KL) divergence [36] and generalized advantage estimation (GAE). Using KL divergence in our framework ensures a more controlled and stable policy update mechanism, effectively managing the exploration-exploitation trade-off and preventing drastic policy deviations. The integration of GAE further refines this approach by providing a more accurate advantage estimation, which optimizes the policy gradient updates for a balanced and effective learning progression.

Although our proposed DRL technique offers an efficient and stable policy optimization algorithm, after going through extensive analysis, we find that it struggles to capture long-term temporal dependencies. To avoid such cases, the agent must rely on past observations/historical patterns or actions to make optimal decisions, which is challenging for standard feed-forward on-policy-based neural networks used in our DRL algorithm. Therefore, our proposed DRL algorithm integrates the long-short-term memory technique (LSTM) [37] as a key component. The research reported in [4], [38], [39], and [40] mainly explores the use of LSTM networks in DRL for tasks like making predictions [4], [38] or to help manage network resources within certain frameworks [39]. These studies showed how LSTMs can understand patterns over time, but they often focus on improving one part of a system or making better forecasts. Unlike these approaches, our work takes a significant step forward by applying LSTMs to make decisions in complex and dynamic environments, particularly for networks using UAV-mounted CRIS units and aiming for extremely reliable and fast communications, as might be seen in upcoming 6G technologies. In the proposed MA-DRL, the agent learns to take actions in the time-variant environment to maximize a reward signal. LSTM networks are used to model sequential state representations, which help capture long-term dependencies in the environment. This modification enhances the agent's ability to handle problems with complex time dependencies and partial observability. The simulation results find the efficacy of the proposed distributed multi-agent LSTM-DRL-based UAV-mounted CRIS over the conventional DRL approaches regarding convergence speed and data rate to the xURLLC users.

In brief, the main contributions of this study are:

1) For the very first time in the literature, we present a novel RIS model, CRIS, in which individual RIS elements (i.e., PE) can dynamically adjust their operating modes to ensure ultra-reliable QoS for xURLLC users. This flexibility offers a richer configuration space, allowing fine-tuned optimization to meet specific communication

requirements. Moreover, we propose to deploy CRIS units using UAVs to serve xURLLC users. A combination of UAV mobility and CRIS adaptability can greatly enhance wireless network performance, particularly in challenging contexts.

2) Recognizing the complexity and dynamic nature of the problem, we propose an MA-DRL algorithm. Each UAV-mounted CRIS has a dedicated agent responsible for determining the configuration of its PEs, optimizing decision parameters, and adjusting UAV positions. This decentralized approach allows for more scalable and efficient network optimization. We incorporate LSTMs into our DRL framework to better capture long-term dependencies in the environment with complex temporal dynamics.
3) Our comprehensive simulations confirm that our multi-agent LSTM-based DRL outperforms traditional DRL techniques in rapid convergence, spectrum efficiency, and higher data rate in stringent reliability constraints for xURLLC users. Additionally, the UAV-mounted CRIS design demonstrates a notable advantage of exploring the scope of RSMA over SDMA and NOMA transmission schemes for xURLLC users.

The rest of the paper is organized as follows: Section II provides a detailed system model description. Section III discusses the problem formulation, while Section IV introduces a multi-agent LSTM-based DRL algorithm that addresses the present problem. Section V presents extensive numerical results and analysis demonstrating our proposed model's efficacy. Finally, we conclude our work in Section VI.

Notational conventions: In this work, vectors are denoted by bold lowercase letters, e.g., $\mathbf{a}$, while matrices are represented by bold uppercase letters, e.g., $\mathbf{A}$. Scalars and sets are signified by a and $\mathcal{A}$, respectively. The symbols $(\cdot)^{\mathrm{T}}$ and $(\cdot)h$ represent the transpose and conjugate transpose operations, respectively. The absolute value is given by $|\cdot|$, and $||\cdot||$ indicates the Frobenius norm. The Gaussian distribution with mean μ and variance σ is represented as $\mathcal{CN}(\mu,\sigma)$. Lastly, $diag(\mathbf{a})$ refers to a diagonal matrix with vector $\mathbf{a}$ as its diagonal elements.

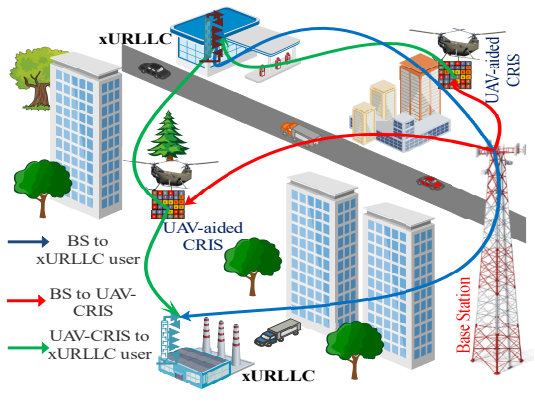


Fig. 2: UAV-CRIS-enabled MIMO communication for xURLLC.

II. System Model

Fig. 2 depicts the proposed system model that consists of a base station (BS) with $\mathcal{M} \triangleq \{1,\ldots,m,\ldots,M\}$ antennas, $\mathcal{K} \triangleq \{1,\ldots,k,\ldots,K\}$ number of xURLLC users with $\mathcal{L} \triangleq \{1,\ldots,l,\ldots,L\}$ antennas each, and $\mathcal{Q} \triangleq \{1,\ldots,q,\ldots,Q\}$ number of UAV-mounted CRIS units where each CRIS contains $\mathcal{N} \triangleq \{1,\ldots,n,\ldots,N\}$ number of PEs. The reflection and transmission coefficient matrix of the q-th CRIS is given by $\mathbf{\Phi}_q^d = diag\{\phi^d_{(q,1)},\ldots,\phi^d_{(q,N)}\}$, $\forall d \in \mathcal{D} \triangleq \{\mathtt{r},\mathtt{t}\}$. $\phi^d_{(q,n)} = \left(\sqrt{\eta^d_{(q,n)}}e^{j\theta^d_{(q,n)}}\right)$, $\eta^t_{(q,n)} \in [0\ \eta^{\mathtt{t}^{\max}}_{(q,n)}]$, $\theta^t_{(q,n)} \in [0\ 2\pi]$ and $\eta^r_{(q,n)} \in [0\ \eta^{\mathtt{r}^{\max}}_{(q,n)}]$, $\theta^r_{(q,n)} \in [0\ 2\pi]$ denote the amplitude and phase shift response of the n-th PE's transmission and reflection coefficients of q-th CRIS, respectively. According to the law of conservation of energy, the sum of the energies of the reflected and transmitted signals are constrained by the power amplifier, i.e., $\phi^{\mathtt{t}}_{(q,n)} + \phi^{\mathtt{r}}_{(q,n)} \leqslant \eta^{\max}_{(q,n)}, \forall n \in \mathcal{N}, q \in \mathcal{Q}$. We assume that xURLLC users and UAV-CRISs are spatially distributed geographically. Depending on the position of the users, each user is located either in the reflection or transmission space of a UAV-mounted CRIS unit. Let $\mathbf{U}(k,q) = \{\mathtt{r},\mathtt{t}\}$ denote the region of k-th user w.r.t. q-th CRIS, i.e., $\mathbf{U}(k,q) = \mathtt{r}$ represents that k-th user is in the reflection region of the CRIS, whereas $\mathbf{U}(k,q) = \mathtt{t}$ represents that k-th user is in the transmission region of the CRIS.

A. RSMA-Based Communication Strategy for xURLLC Users

In this work, under the RSMA-based transmission scheme [41], each xURLLC user receives L_k messages such that $L_k = \min\{M,L\}, \forall k \in \mathcal{K}$. The collection of messages for the k-th user is defined as $\mathbf{w_k} = \{W^k_1,\ldots,W^k_{L_k}\}$. The i-th message designated for the k-th user is divided into two components: the common and private segments, denoted as $W^{\mathrm{c},k}_i$ and $W^{\mathrm{p},k}_i$, respectively. The collection of common and private messages of the k-th user are denoted as $\mathbf{w}_{c,k} = \{W^{\mathrm{c},k}_1,\ldots,W^{\mathrm{c},k}_{L_k}\}$ and $\mathbf{w}_{\mathrm{p},k} = \{W^{\mathrm{p},k}_1,\ldots,W^{\mathrm{p},k}_{L_k}\}$, respectively. The common messages of all the users are combined into $L_c \in \{1,\ldots,\min(M,L)\}$ messages denoted by $\mathbf{w}_c \in \mathbb{C}^{L_c\times 1}$, and encoded together into a common stream vector of $\mathbf{s}_c = \left[s^c_1,\ldots,s^c_{L_c}\right]^{\mathrm{T}}$. This $\mathbf{s}_c$ is decoded by all xURLLC users. The private parts of the k-th user are independently encoded in a private stream vector $\mathbf{s}_k = \left[s^{\mathrm{p},k}_1,\ldots,s^{\mathrm{p},k}_{L_k}\right]^{\mathrm{T}}$ decoded by the k-th user. Therefore, the general vector of the data stream to be transmitted is expressed as $\mathbf{s} = [\mathbf{s}_{\mathrm{c}},\mathbf{s}_1,\ldots,\mathbf{s}_K]^{\mathrm{T}}$ that satisfies $\mathbb{E}\{\mathbf{s}\mathbf{s}^h\} = \mathbf{I}$. Linear precoders $\mathbf{P} = [\mathbf{P}_c,\mathbf{P}_1,\ldots,\mathbf{P}_K]$ precode data streams, where $\mathbf{P}_c \in \mathbb{C}^{M\times L_c}$ is the precoder for the common stream vector and $\mathbf{P}_k \in \mathbb{C}^{M\times L_k}$ is the precoder for the private stream vector of the user k-th user.

B. Received Signal Modeling for xURLLC User

The signal transmitted by the BS at a time instant t is:

$$\mathbf{x}(t) = \mathbf{P}_{\mathrm{c}}(t)\mathbf{s}_{\mathrm{c}}(t) + \sum_{k=1}^{K}\mathbf{P}_k(t)\mathbf{s}_k(t). \tag{2}$$

The k-th user, $\forall k \in \mathcal{K}$ of xURLLC, receives the transmitted signal $\mathbf{x}(t)$ from the BS and R the number of UAV-mounted CRIS units. The signal received by the k-th user is given by

$$\mathbf{y}_k(t) = \mathbf{A}_k(t)\mathbf{x}(t) + \mathbf{b}_k(t) + \mathbf{z}_{1,k}(t), \tag{3}$$

where,

$$\mathbf{A}_k(t) = \mathbf{h}_k(t) + \sum\nolimits_{q=1}^{Q} \mathbf{G}_k^{(\mathbf{U}(k,q),q)(t)}(t)\mathbf{\Phi}_q^{\mathbf{U}(k,q)(t)}(t)\mathbf{F}_q(t), \tag{4}$$

$$\mathbf{b}_k(t) = \sum\nolimits_{q=1}^{Q} \mathbf{G}_k^{(\mathbf{U}(k,q),q)(t)}(t)\mathbf{\Phi}_q^{\mathbf{U}(k,q)(t)}(t)\mathbf{z}_{2,q}(t). \tag{5}$$

Here $\mathbf{z}_{1,k} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_{1,k}^2\mathbf{I}_L\right)$ and $\mathbf{z}_{2,q} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_{2,q}^2\mathbf{I}_N\right)$ are the additive white Gaussian noise (AWGN) vectors at xURLLC user and CRIS, respectively. Furthermore, $\mathbf{h}_k \in \mathbb{C}^{L\times M}$, $\mathbf{F}_q \in \mathbb{C}^{N\times M}$ and $\mathbf{G}_k^{(\mathbf{U}(k,q),q)(t)} \in \mathbb{C}^{L\times N}$ represent the channels from BS to k-th xURLLC user, BS to q-th CRIS and q-th CRIS to k-th xURLLC user, which is in $\mathbf{U}(k,q)$ region of CRIS. The achievable data rate for MIMO communication is derived using [42]. The common rate of the k-th user is given by

$$R_k^c(t) = \log_2(|\mathbf{I} + \mathbf{A}_k(t)\mathbf{P}_c(t)\mathbf{A}_k^h(t)\mathbf{P}_c^h(t)\mathbf{T}_k^{-1}(t)|), \tag{6}$$

where $\mathbf{T}_k(t) = \sum_{i=1}^{K} \mathbf{A}_k(t)\ \mathbf{P}_i(t)\ \mathbf{P}_i^h(t)\ \mathbf{A}_k^h(t) + \mathbf{b}_k(t)\ \mathbf{B}_k^h(t)\ \sigma_{2,k}^2(t) + \sigma_{1,k}^2(t)\mathbf{I}_L$. The private rate of the k-th user is

$$R_k^p(t) = \log_2(|\mathbf{I} + \mathbf{A}_k(t)\mathbf{P}_k(t)\mathbf{P}_k^h(t)\mathbf{A}_k^h(t)\mathbf{J}_k^{-1}(t)|), \tag{7}$$

where $\mathbf{J}_k(t) = \sum_{i=1,i\neq k}^{K} \mathbf{A}_k(t)\ \mathbf{P}_i(t)\mathbf{P}_i^h(t)\ \mathbf{A}_k^h(t) + \mathbf{b}_k(t)\ \mathbf{B}_k^h(t)\ \sigma_{2,k}^2(t) + \sigma_{1,k}^2(t)\mathbf{I}_L$. The power budget at the q-th CRIS is

$$P_q^{\text{CRIS}} = \left(\left\|\mathbf{\Phi}_q^{\mathtt{r}}(t)\mathbf{F}_q(t)\mathbf{P}(t)\right\|_2^2\right) + \sigma_{2,q}^2\left(\left\|\left(\mathbf{\Phi}_q^{\mathtt{r}}(t)\right)\right\|_2^2\right) + \left(\left\|\mathbf{\Phi}_q^{\mathtt{t}}(t)\mathbf{F}_q(t)\mathbf{P}(t)\right\|_2^2\right) + \sigma_{2,q}^2\left(\left\|\left(\mathbf{\Phi}_q^{\mathtt{t}}(t)\right)\right\|_2^2\right), \forall q \in \mathcal{Q}. \tag{8}$$

In wireless communication systems, when the l_k is finite block length (FBL) and the error probability ξ_k is non-zero, there is a fundamental limit on the achievable data rates R_k^c and R_k^p [41]. This is because, as the block length decreases or the error probability increases, the system is more prone to errors, which can limit the information rate. To quantify this loss, the achievable rate for a given FBL $l_k^c(t)$ and $l_k^p(t)$ in common and private data rate in RSMA along with decoding error probability ξ_k is expressed as [43]

$$\overline{R_k^c}(t) = R_k^c(t) - \sqrt{\frac{V_k^c(t)}{l_k^c(t)}}\frac{\mathrm{Q}^{-1}(\xi_k)}{\log_e 2}, \forall k, \tag{9}$$

$$\overline{R_k^p}(t) = R_k^p(t) - \sqrt{\frac{V_k^p(t)}{l_k^p(t)}}\frac{\mathrm{Q}^{-1}(\xi_k)}{\log_e 2}, \forall k, \tag{10}$$

where $V_k^c(t) = 1-(1-\Gamma_k^c(t))$ and $V_k^p(t) = 1-(1-\Gamma_k^p(t))$ are the channel dispersion values for the common and private channels, respectively, where $\Gamma_k^c(t) = ||\mathbf{A}_k(t)\mathbf{P}_c(t)||_F^2/||\mathbf{T}_k(t)||_F^2$ and $\Gamma_k^p(t) = ||\mathbf{A}_k(t)\mathbf{P}_k(t)||_F^2/||\mathbf{J}_k(t)||_F^2$, in which $||\mathbf{A}||_F$ denotes the Frobenius norm of $\mathbf{A}$. The inverse error function is represented as $Q^{-1}(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(\frac{-t^2}{2}\right)dt$. To ensure the successful decoding of the common message at each user, the total common transmission rate is defined as $\overline{R_c}(t) \triangleq \sum_{k\in\mathcal{K}} c_k(t)$, to satisfy the minimum required data rate constraint such that $\overline{R_c}(t) \leqslant \min(\overline{R_k^c}(t))$. The symbol $c_k(t)$ represents the portion of the common rate allocated in time-instant t to the k-th user.

C. Trajectories of UAV-mounted CRIS Units

In this scenario, we consider a three-dimensional (3D) surface for accommodating K stationary xURLLC users. Q non-stationary UAV-mounted CRIS units serve these users. The mobility of independent users is beyond our control, but we can control the placement of UAVs to meet the service requirements of xURLLC users. Consequently, finding the optimal locations of UAVs that enable users' mobility freedom is paramount within the current system model. Let the q-th UAV-mounted CRIS be located at the 3D coordinate $(\mathrm{X}_t^q, \mathrm{Y}_t^q, \mathrm{Z}_t^q)$ during time instant t, exhibiting a uniform velocity of $V_t^q \in \mathbf{U} \sim \left[0, v_{\max}^q\right]$. Similarly, at time t, the k-th user is at $(\mathrm{X}_t^k, \mathrm{Y}_t^k, \mathrm{Z}_t^k)$. Here, we assume that the orientation of the RIS is fixed, i.e., all the RISs are oriented in the xz plane. Therefore, the region at which the k-th user lies w.r.t the q-th UAV-mounted CRIS can be decided using the following condition. When $\mathrm{Y}_t^k \leqslant \mathrm{Y}_t^q$, the k-th user lies in the reflection region of the q-th UAV-mounted CRIS, otherwise the k-th user lies in the transmission region of q-th UAV-mounted CRIS.

$$d = \begin{cases} \mathtt{r}, & \text{if } \mathrm{Y}_t^k \leqslant \mathrm{Y}_t^q \\ \mathtt{t}, & \text{otherwise.} \end{cases} \tag{11}$$

However, in practical scenarios, a single UAV-CRIS unit cannot simultaneously serve all the xURLLC users at time t. We assume that the present BS is responsible for serving any xURLLC user in $\frac{4}{3}\pi(\mathtt{R}_{\mathtt{bs}})^3$ unit of spherical volume, where $\mathtt{R}_{\mathtt{bs}}$ is the radius of the sphere. To enable the q-th UAV-CRIS-assisted service provision for the k-th user, that user must fall within the transmission coverage radius ($\mathtt{R}_{\mathtt{q}}$) of the q-th UAV-CRIS unit. The distance D_q^k between the q-th UAV and k-th xURLLC user at time instant t in 3D space is calculated using the Euclidean distance as

$$D_q^k(t) = \sqrt{(\mathrm{X}_t^q - \mathrm{X}_t^k)^2 + (\mathrm{Y}_t^q - \mathrm{Y}_t^k)^2 + (\mathrm{Z}_t^q - \mathrm{Z}_t^k)^2}. \tag{12}$$

D. Channel Modeling in the Presence of Imperfect CSI

Due to the mobility of the UAVs, all the communication links related to the UAVs suffer from a serious Doppler spread effect. The Doppler frequency for the q-th UAV is $f_q(t) = v_q(t)\cos(\varrho_q(t))\cos(\varphi_q(t))/\lambda(t)$, where $v_q(t)$ is the speed of the q-th UAV in meters/second (m/s), $\lambda(t)$ is the wavelength of the signal, $\varrho_q(t) \in [0, \pi/2]$ and $\varphi_q(t) \in [0, 2\pi)$ are the elevation and azimuth angles of arrival (AoAs) at the q-th UAV, respectively. Further, we assume the BS maintains LoS with UAVs. Hence, all the above channels follow the Rician fading distribution [42] with the Doppler effect. For example, the channels from the BS to the q-th UAV-CRIS are modeled as

$$\mathbf{F}_q(t) = e^{j2\pi f_q(t-1)T}\alpha\sqrt{\frac{\beta_{\mathrm{F}_q}}{\beta_{\mathrm{F}_q}+1}}\mathbf{F}_q^{\text{LoS}}(t) + \sqrt{\frac{1}{\beta_{\mathrm{F}_q}+1}}\mathbf{F}_q^{\text{NLoS}}(t), \tag{13}$$

where β_{F_q} is the Rician factor, and α is the propagation constant. $\mathbf{F}_q^{\text{LoS}}$ and $\mathbf{F}_q^{\text{NLoS}}$ are the LoS (deterministic) and non line-of-sight NLoS (Rayleigh) components. The deterministic LoS component $\mathbf{F}_q^{\text{LoS}}$ is modeled as $\mathbf{F}_q^{\text{LoS}} =$

$\mathbf{a}_N^q\left(\vartheta_q^{AoA}(t)\right)(\mathbf{a}_M^q)h\left(\vartheta_q^{AoD}(t)\right)$, where,

$$\mathbf{a}_N^q\left(\vartheta_q^{AoA}(t)\right)=\Big[1,e^{j\frac{2\pi\zeta}{\lambda}\sin\vartheta_q^{AoA}(t)},\ldots,e^{j\frac{2\pi\zeta}{\lambda}(N-1)\sin\vartheta_q^{AoA}(t)}\Big]^{\mathrm{T}},\quad(14)$$

$$\mathbf{a}_M^q\left(\vartheta_q^{AoD}(t)\right)=\Big[1,e^{j\frac{2\pi\zeta}{\lambda}\sin\vartheta_q^{AoD}(t)},\ldots,e^{j\frac{2\pi\zeta}{\lambda}(M-1)\sin\vartheta_q^{AoD}(t)}\Big]^{\mathrm{T}},\quad(15)$$

where ζ is defined as antenna separation distance. λ is the wavelength, and we set ζ/λ =1/2. The angle of departure $\vartheta_q^{AoD}(t)$ and the angle of arrival $\vartheta_q^{AoA}(t)$ are assumed to be randomly distributed within $[0,2\pi)$.

The channels from the q-th UAV-CRIS to the k-th user is modeled as:

$$\mathbf{G}_k^{(d,q)}(t)=e^{j2\pi f_q(t-1)T}\alpha\sqrt{\frac{\beta_{\mathbf{G}_k^{(d,q)}}}{\beta_{\mathbf{G}_k^{(d,q)}}+1}}\left(\mathbf{G}_k^{(d,q)}\right)^{\mathrm{LoS}}+\sqrt{\frac{1}{\beta_{\mathbf{G}_k^{(d,q)}}+1}}\left(\mathbf{G}_k^{(d,q)}\right)^{\mathrm{NLoS}}(t),\quad(16)$$

where $\beta_{\mathbf{G}_k^{(d,q)}}$ is the Rician factor, f_q denotes the carrier frequency, α is the propagation constant. $\left(\mathbf{G}_k^{(d,q)}\right)^{\mathrm{LoS}}$ and $\left(\mathbf{G}_k^{(d,q)}\right)^{\mathrm{NLoS}}$ are the LoS (deterministic) and NLoS (Rayleigh) components. The deterministic LoS component $\left(\mathbf{G}_k^{(d,q)}\right)^{\mathrm{LoS}}$ is modeled as $\left(\mathbf{G}_k^{(d,q)}\right)^{\mathrm{LoS}}=\mathbf{a}_L^k\left(\vartheta_{(k,q)}^{AoA}(t)\right)(\mathbf{a}_N^{(q,k)})h\left(\vartheta_{(q,k)}^{AoD}(t)\right)$, where $\mathbf{a}_N\left(\vartheta_{(k,q)}^{AoA}(t)\right)$ is defined as:

$$\mathbf{a}_L^{(k,q)}\left(\vartheta_{(k,q)}^{AoA}(t)\right)=\Big[1,e^{j\frac{2\pi\zeta}{\lambda}\sin\vartheta_{(k,q)}^{AoA}(t)},\ldots,e^{j\frac{2\pi\zeta}{\lambda}(L-1)\sin\vartheta_{(k,q)}^{AoA}(t)}\Big]^{\mathrm{T}},\quad(17)$$

$$\mathbf{a}_N^{(q,k)}\left(\vartheta_{(q,k)}^{AoD}(t)\right)=\Big[1,e^{j\frac{2\pi\zeta}{\lambda}\sin\vartheta_{(q,k)}^{AoD}(t)},\ldots,e^{j\frac{2\pi\zeta}{\lambda}(N-1)\sin\vartheta_{(q,k)}^{AoD}(t)}\Big]^{\mathrm{T}}.\quad(18)$$

The angle of departure $\vartheta_{(q,k)}^{AoD}(t)$ and angle of arrival $\vartheta_{(k,q)}^{AoA}(t)$ are assumed to be randomly distributed within $[0,2\pi]$.

The channel from the BS to the k-th user is modeled as Rayleigh distribution. Further to this, the large-scale path-loss (in dB) is $P_{L_i}=P_{L_0}(dist_i/dist_0)^{-\alpha_i}$, where P_{L_0} in dB denotes the path-loss at the reference distance of $dist_0$, and α_i where $\forall i\in\{\mathbf{h}_k,\mathbf{F}_q,\mathbf{G}_k^{(d,q)}\}$, represents the path-loss exponent between the BS to k-th xURLLC user, BS to q-th CRIS and q-th CRIS to k-th xURLLC user, respectively. Moreover, $dist_i$ denotes the distance of the i-th link.

In practice, due to several unwanted obstacles, such as hardware impairments and fading in the channel, the CSI may suffer an estimation error. For example, the CSI from the BS to the k-th user is expressed as $\mathbf{h}_k(t)=\tilde{\mathbf{h}}_k(t)+\Delta\mathbf{h}_k(t)$ and the CSI from the BS to the q-th UAV-CRIS is expressed as $\mathbf{F}_q(t)=\tilde{\mathbf{F}}_q(t)+\Delta\mathbf{F}_q(t)$ and q-th UAV-CRIS to k-th user is expressed as $\mathbf{G}_k^{(d,q)}(t)=\tilde{\mathbf{G}}_k^{(d,q)}(t)+\Delta\mathbf{G}_k^{(d,q)}(t)$, where $\tilde{\mathbf{h}}_k$,$\tilde{\mathbf{F}}_q$, $\tilde{\mathbf{G}}_k^{(d,q)}$ are estimated CSI and $\Delta\mathbf{h}_k(t)$, $\Delta\mathbf{F}_q(t)$ and $\Delta\mathbf{G}_k^{(d,q)}(t)$, indicate error matrix. In this work, these imperfections are modeled as the norm-bounded error model [44] given by

$$\|\Delta\mathbf{h}_k\|_2\leqslant\varrho_k,\ \left\|\Delta\mathbf{F}_q\right\|_2\leqslant\varepsilon_q,\ \left\|\Delta\mathbf{G}_k^{(d,q)}\right\|_2\leqslant\varphi_k^q,\quad(19)$$

where ϱ_d and ε_u denote the downlink and uplink channel's error bound, respectively. Considering these uncertainties, the imperfect channels lie in the bounded region ($\mathcal{B}$) defined as

$$\mathbf{h}_k(t)\in\mathcal{B}_1=\left\{\tilde{\mathbf{h}}_k(t)+\Delta\mathbf{h}_k:\|\Delta\mathbf{h}_k\|_2\leqslant\varrho_k\right\},\quad(20)$$

$$\mathbf{F}_q(t)\in\mathcal{B}_2=\left\{\tilde{\mathbf{F}}_q(t)+\Delta\mathbf{F}_q:\left\|\Delta\mathbf{F}_q\right\|_2\leqslant\varepsilon_q\right\},\quad(21)$$

$$\mathbf{G}_k^{(d,q)}(t)\in\mathcal{B}_3=\left\{\tilde{\mathbf{G}}_k^{(d,q)}(t)+\Delta\mathbf{G}_k^{(d,q)}:\left\|\Delta\mathbf{G}_k^{(d,q)}\right\|_2\leqslant\varphi_k^q\right\},\quad(22)$$

III. Problem formulation

This section focuses on mathematical problem formulation for RSMA-based sum rate maximization of the xRULLC users under various constraints. Latency and reliability in the xURLLC play a crucial role. In this work, we focus on the end-to-end (E2E) latency, which is approximately proportional to the block length for point-to-point (P2P) communication [45]. The sum-rate maximization for all the xURLLC users under the assistance of URV-aided CRIS units is formulated as follows:

$$\max_{\mathbf{a}(t)}f(\mathbf{a}(t))=\left[\sum\nolimits_{k\in\mathcal{K}}(c_k(t)+\overline{R_k^p}(t))\right],\quad(23)$$

$$\text{s.t. }(C.1):\sum\nolimits_{k\in\mathcal{K}}|\mathbf{P}_k(t)|^2+|\mathbf{P}_c(t)|^2\leqslant p_{max}^{\mathrm{bs}}(t),$$
$$(C.2):\sum\nolimits_{i\in\mathcal{K}}c_i(t)\leqslant\overline{R_c}(t),\forall i\in\mathcal{K},$$
$$(C.3):c_k(t)\geqslant 0,\ \forall k\in\mathcal{K},$$
$$(C.4):c_k(t)+\overline{R_k^p}(t)\geqslant R_k^{\min}(t),\forall k\in\mathcal{K},$$
$$(C.5):P_q^{\mathrm{CRIS}}(t)\leqslant p_{q,max}^{\mathrm{RS}}(t),\forall q\in\mathcal{Q},$$
$$(C.6):\phi_{(q,n)}^{\mathrm{t}}+\phi_{(q,n)}^{\mathrm{r}}\leqslant\eta_{(q,n)}^{\max},\forall n\in\mathcal{N},q\in\mathcal{Q},$$
$$(C.7):|\phi_{(q,n)}^d(t)|\leqslant(\eta_{(q,n)}^d)^{\max},\forall n\in\mathcal{N},q\in\mathcal{Q},d\in\mathcal{D},$$
$$(C.8):-(\mathtt{R_{bs}}-\mathtt{R_q})\leqslant\{\mathrm{X}_t^q,\mathrm{Y}_t^q,\mathrm{Z}_t^q\}\leqslant(\mathtt{R_{bs}}-\mathtt{R_q}),$$
$$(C.9):D_q^k(t)\leqslant\mathtt{R_q},\ \forall k\in\mathcal{K},\ \forall q\in\mathcal{Q},$$
$$(C.10):0\leqslant v_q(t)\leqslant v_q^{\max}(t),\ \forall q\in\mathcal{Q},$$
$$(C.11):0\leqslant\xi_k(t)\leqslant\xi_k^{\max}(t),\ \forall k\in\mathcal{K},$$
$$(C.12):l_c^k(t)+l_p^k(t)\leqslant l_{\max}(t),\ \forall k\in\mathcal{K},$$

where $\mathbf{a}(t)=\{\mathbf{P}_c(t),\mathbf{P}_k(t),l_c^k(t),l_p^k(t),P_q^{\mathrm{CRIS}}(t),\phi_{(q,n)}^{\mathrm{t}},\phi_{(q,n)}^{\mathrm{r}},\mathrm{X}_t^q,\mathrm{Y}_t^q,\mathrm{Z}_t^q,v_q(t)\}$. In the following constraints, $(C.1)$ sets a maximum power budget ($p_{\max}^{\mathrm{bs}}$) at the BS. $(C.2)$ ensures the successful decoding of the common message at each user, while $(C.3)$ guarantees that the common rate remains positive. Constraint $(C.4)$ enforces a minimum data rate ($R_{k,\min}$) for each user in the network. $(C.5)$ represents the power constraint at the CRIS, where $P_{q,\max}^{\mathrm{RS}}$ denotes the maximum power budget assigned to the q-th CRIS. $(C.6)$ indicates the CRIS's reflection and transmission coefficient constraint. $(C.7)$ denotes the maximum amplification of the signal $(\eta_{(q,n)}^d)^{\max}$ at the n-th element of the q-th CRIS. Assuming the BS is located at the center point of a 3D sphere, i.e., $(0,0,0)$. There are strict restrictions on inter-cell interference; each q-th UAV-CRIS, at time instant t, must

satisfy the positional constraint ($C.8$). These constraints ensure that the positional coordinates must lie within the range. For successful data transmission from the q-th UAV-CRIS unit to the k-th xURLLC user, a distance constraint is imposed as given in ($C.9$). Additionally, a maximum velocity constraint $v_q^{\max}(t)$ is enforced on the velocity $v_q(t)$ of the q-th UAV-mounted CRIS unit in ($C.10$). To ensure reliability, a maximum decoding error probability constraint is defined in ($C.11$). The latency constraint is addressed in ($C.12$), which states that the maximum fixed blocklength should not exceed $l_{\max}$. Finally, for E2E delay/latency, the point-to-point communication is approximately proportional to the block length [45].

A. Analysis of the Objective Function and Constraints

The present objective function is non-convex. A detailed analysis shows that the rest of the constraints are convex except constraint ($C.4$). Constraint ($C.4$) is non-convex due to the presence of the logarithm function. Their linearity or bound nature determines the convexity of the other constraints. Furthermore, the formulated problem formulation includes various types of constraints, including linear inequalities ($C.1$), quadratic inequalities ($C.5$), and non-linear inequalities ($C.9$). Additionally, there are other types of constraints, such as a range constraint ($C.8$). ($C.10$), ($C.11$), and ($C.12$) involve upper and lower bounds on variables, and reliability and latency constraints, respectively.

B. Solving Complex and Time-Variant Optimization Problems

The recent literature exploits DRL techniques to overcome the limitations of conventional optimization techniques to solve such problems, as DRL combines the power of deep neural networks and reinforcement learning together without relying on explicit mathematical models [33]. DRL agents learn optimal decision-making policies through trial and error, using rewards as feedback to guide their learning process. This approach allows them to adapt and find solutions in complex and dynamic problem environments. However, training DRL models is computationally intensive and requires substantial time. Thus, selecting appropriate network architectures and training procedures is crucial to ensure convergence and good performance. To this aim, we explore many DRL-based techniques to solve the present problem and propose an effective DRL approach while comparing their relative performance. However, before applying any DRL algorithms, we must reformulate the present optimization problem as a Markov decision process (MDP) problem [35].

C. MDP Formulation of the Present Problem

MDP is a mathematical framework that models decision-making problems in a stochastic environment [23]. It is characterized by its key attributes, including states, actions, transition probabilities, and rewards. MDP provides a formal structure for decision-making problems and is particularly useful in dynamic and uncertain environments. In our optimization problem, MDP formulation is employed to capture the sequential decision-making nature of the problem. By formulating the problem as an MDP, we can exploit the power of DRL algorithms to learn an optimal policy. Our model consists of two types of agents: BS-agents, which work for the BS, and UAV-CRIS agents, which work for a UAV-mounted CRIS unit. It is important to note that if there are Q UAV-mounted CRIS units, then each unit will have its UAV-CRIS agent. The attributes of the MDP formulation in our problem are defined as follows:

- State space ($\mathcal{S}$): The continuous state space $\mathcal{S}$ represents the system's configuration, including the channel conditions and its associated relevant variables. These variables in $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_t, \ldots, \mathbf{s}_T\}$ can take on any real value within a specified range, allowing for a more detailed representation of the system's dynamics and interactions. In the context of our MA-DRL framework, we define the state representations for each of the agent. The state for the BS-agent is written as: $\mathbf{s}_t^{\text{bs}}(t) = \{\xi_k(t), l_{\max}(t), p_{\max}^{\text{bs}}(t), c_k(t), R_k^{\min}(t), \mathbf{F}_q(t), \mathbf{h}_k(t)\}$. The state of each UAV-CRIS agent is expressed as: $\mathbf{s}_t^q(t) = \{D_q^k(t), \xi_k(t), p_{q,max}^{\text{RS}}(t), \eta_{(q,n)}^{\max}, c_k(t), R_k^{\min}(t), l_k(t), \mathbf{G}_k(t)\}$. Collectively, the aggregated system state at time instant t, incorporating both the BS-agent and all UAV-CRIS agents, is represented as $\mathbf{s}_t = \{\mathbf{s}_t^{\text{bs}} \cup \mathbf{s}_t^q\}, \forall q \in Q$, where Q denotes the set of all UAV-CRIS agents, and $\mathbf{s}_t \in \mathcal{S}$.
- Action space ($\mathcal{A}$): The continuous action space $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_t, \ldots, \mathbf{a}_T\}$ corresponds to the choices that the decision-making agent can make. Actions include adjusting UAV positions, velocities, transmission power levels, CRIS parameters, and other system variables. At time instant t, the actions available to the BS agent are represented by a set as $\mathbf{a}_t^{\text{bs}} = \{\mathbf{P}_c(t), \mathbf{P}_k(t), l_c^k(t), l_p^k(t)\}$. These actions are crucial for the adaptive management of the network's power distribution and the strategic selection of communication links to optimize performance. Contrarily, for each UAV-CRIS agent, the action set is defined as $\mathbf{a}_t^q = \{P_q^{\text{CRIS}}(t), \phi_{(q,n)}^{\text{t}}, \phi_{(q,n)}^{\text{r}}, \mathrm{X}_t^q, \mathrm{Y}_t^q, \mathrm{Z}_t^q, v_q(t)\}$. Therefore, the total collaborative action space at time instant t is written as $\mathbf{a}_t = \{\mathbf{a}_t^{\text{bs}} \cup \mathbf{a}_t^q\}$, $\forall q \in Q$, and $\mathbf{a}_t \in \mathcal{A}$.
- Rewards (r_t): Rewards in our context are quantified by evaluating key performance metrics such as data rates, latency violations, and power consumption. The challenge lies in designing a reward function that maximizes the sum rate and considers various constraints. This function must align with the system's objectives, promote desired behavior, and guide the DRL agent toward optimal UAV-mounted CRIS communication system decision-making. The overall system performance depends on three main components: throughput maximization (i.e., leads to latency minimization) power efficiency, and constraint satisfaction. Therefore, we design the reward function as

$$r_t(s_t, a_t) = w_1 \sum\nolimits_{k\in\mathcal{K}} f(\mathbf{a}(t)) - w_2 \Bigg(\mathbf{P}_c(t) + \sum\nolimits_{k\in\mathcal{K}} \mathbf{P}_k(t) + \sum\nolimits_{q\in Q} P_q^{\text{CRIS}} \Bigg) - w_3 \sum\nolimits_i \lambda_i(t) \vartriangle_i (t), \tag{24}$$

where w_1, w_2, and w_3 are weights that determine the relative importance of each component as mentioned above, respectively. The symbol λ_i is also a weight parameter and $\triangle_i$ quantifies the degree of violation for each constraint i, using mean squared error (MSE). We consider $w_1 + w_2 + w_3 = 1$.

- Transition probabilities: The transition probabilities describe the likelihood of transitioning from one state to another when a particular action is taken. These probabilities capture the system's dynamics, including UAV mobility and channel variations. However, in the present work, we do not have any defined transition probabilities due to the time-variant characteristics of the system. In DRL, when transition probabilities are unavailable, the agent learns policies through trial and error.

IV. Proposed Multi-agent LSTM-based DRL

Our goal is to enhance the learning capabilities of our agents. One primary challenge is to enable our agents to comprehend the relationship between temporal events, which is crucial for effective action decisions. To address this, our DRL methodology incorporates the LSTM system [46]. LSTMs serve as an intelligent memory module for our agents, enabling them to retain and utilize essential historical information to make informed present-time decisions [47]. LSTMs assess the ongoing processes at each time instance, adjusting their memory accordingly. This memory comprises two segments: a general information store ($\mathbf{C}_t$) and a monitor for recent events ($\mathbf{h}_t$). Such a design aids agents in concentrating on relevant past data while excluding inconsequential details. We then merge this sophisticated memory with a policy network, directing our agents' action choices and a value network, evaluating the current scenario's desirability.

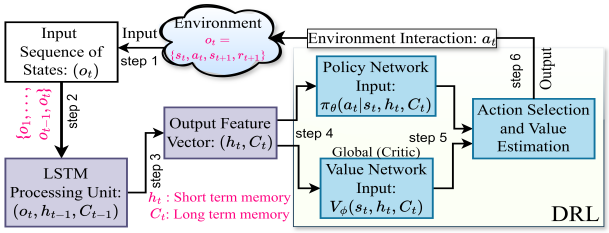


Fig. 3: Flowchart for LSTM output integration with DRL.

In the process of integrating LSTM with DRL, we outline a sequential workflow in Fig. 3 that commences with the input of a series of environmental states observed over time, denoted as $O = \{\mathbf{o}_1, \ldots, \mathbf{o}_t\}$. The observed state ($\mathbf{o}_t = \{\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, \mathbf{r}_{t+1}\}$) serve as inputs to the LSTM network, which systematically processes the current observation $\mathbf{o}_t$ with its accumulated memory from past sequences, represented by $\mathbf{h}_{t-1}$ and $\mathbf{C}_{t-1}$. Through this process, the LSTM updates its internal memory and outputs a feature vector $\{\mathbf{h}_t, \mathbf{C}_t\}$. This vector captures temporal dependencies and contextual information extracted from the sequence of inputs, effectively capturing the dynamism inherent in the environmental states $\mathbf{s}_t = \{\mathbf{s}_t^{\text{bs}} \cup \mathbf{s}_t^q\}, \forall q \in Q$. Subsequently, this feature vector is channelled into the policy network, π_θ, and the global value network, V_ϕ (i.e., critic network). The policy network is responsible for determining the next action $\mathbf{a}_t$ to be taken for the individual agents. In contrast, the value network estimates the potential returns from the current state using Algorithm 3. The action selection process for each agent (i.e., $\mathbf{a}_t = \{\mathbf{a}_t^{\text{bs}} \cup \mathbf{a}_t^q\}, \forall q \in Q$), denoted as $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)$, and the centralized state value estimation, $V_\phi(\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)$, are directly impacted by the LSTM's output.

Our LSTM network consists of various units: input gate $\mathbf{I}$, forget gate $\mathbf{F}$, memory cell $\mathbf{C}$, and output gate $\mathbf{O}$. Each of these gates has associated weights and biases, denoted as $(\mathbf{I}, \mathbf{W_I}, \mathbf{b_I})$, $(\mathbf{F}, \mathbf{W_F}, \mathbf{b_F})$, $(\mathbf{C}, \mathbf{W_C}, \mathbf{b_C})$, and $(\mathbf{O}, \mathbf{W_O}, \mathbf{b_O})$ [40]. At state t, the input gate utilizes historical sequences from $\{\mathbf{o}_1, \ldots, \mathbf{o}_{t-1}\}$ (i.e., see $\mathbf{o}_{t-1}$ in Fig. 4), while the output gate predicts $\hat{\mathbf{o}}_t$. Let ρ be the set of LSTM gates, so $\rho = \{\mathbf{I}, \mathbf{F}, \mathbf{C}, \mathbf{O}\}$. The LSTM structure using system parameters $\mathbf{o}_t$ is mathematically represented as [40], [47]:

$$\mathbf{I}_t = f_{\text{sig}}^{\text{act}}(\mathbf{W_I}\mathbf{o}_t + \mathbf{W_{IH}}\mathbf{h}_{t-1} + \mathbf{W_{IC}}\mathbf{C}_{t-1} + \mathbf{b_I}), \tag{25}$$

$$\mathbf{F}_t = f_{\text{sig}}^{\text{act}}(\mathbf{W_F}\mathbf{o}_t + \mathbf{W_{FH}}\mathbf{h}_{t-1} + \mathbf{W_{FC}}\mathbf{C}_{t-1} + \mathbf{b_F}), \tag{26}$$

$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot f_{\text{tanh}}^{\text{act}}(\mathbf{W_C}\mathbf{o}_t + \mathbf{W_{CH}}\mathbf{h}_{t-1} + \mathbf{W_{CC}}\mathbf{C}_{t-1} + \mathbf{b_C}), \tag{27}$$

$$\mathbf{O}_t = f_{\text{sig}}^{\text{act}}(\mathbf{W_O}\mathbf{o}_t + \mathbf{W_{OH}}\mathbf{h}_{t-1} + \mathbf{W_{OC}}\mathbf{C}_{t-1} + \mathbf{b_O}), \tag{28}$$

$$\mathbf{h}_t = f_{\text{peep}}^{\text{act}}(\mathbf{O}_t \odot f_{\text{tanh}}^{\text{act}}(\mathbf{W_O}\mathbf{o}_t + \mathbf{W_{OH}}\mathbf{h}_{t-1} + \mathbf{W_{OC}}\mathbf{C}_{t-1} + \mathbf{b_O})), \tag{29}$$

where $\odot$ is element-wise multiplication, and $f_{\text{sig}}^{\text{act}}$, $f_{\text{tanh}}^{\text{act}}$, and $f_{\text{peep}}^{\text{act}}$ are the sigmoid, tangent, and peephole activation functions, respectively. The loss function for the LSTM is:

$$f_{\text{Loss}} = \sum_{i=1}^{M} \left\| \mathbf{o}_t^i - \hat{\mathbf{o}}_t^i \right\|^2, \tag{30}$$

with $\hat{\mathbf{o}}_t$ and $\mathbf{o}_t$ denoting the predicted and target system parameters at time t.

Algorithm 1 Optimization of LSTM function in (32)

Initialization:

Learning rate: $\alpha = 0.001$, 1st moment vector: $\mathbf{m}_0 = \mathbf{0}$, 2nd moment vector: $\mathbf{v}_0 = \mathbf{0}$, Tolerance: ϵ, Number of epochs: N, $\{\beta_1, \beta_2\} = 0.9$, Initialization: $\boldsymbol{\iota}_{\text{LM},0} \sim \mathcal{N}(0, \frac{1}{n})$.

Optimization Procedure:

for $epoch = 1$ to N **do**
  for all mini-batches $\mathbf{B}$ **do**
    Compute gradient: $\nabla \boldsymbol{\iota}_{\text{LM}} = \frac{\partial}{\partial \boldsymbol{\iota}_{\text{LM}}} \left(\boldsymbol{\iota}_{\text{LM}}^{\text{approx}} \right)$
    Update: $\mathbf{m}_{epoch} = \beta_1 \mathbf{m}_{epoch-1} + (1 - \beta_1) \nabla \boldsymbol{\iota}_{\text{LM}}$.
    Update: $\mathbf{v}_{epoch} = \beta_2 \mathbf{v}_{epoch-1} + (1 - \beta_2)(\nabla \boldsymbol{\iota}_{\text{LM}})^2$.
    Correct bias in: $\hat{\mathbf{m}}_{epoch} = \frac{\mathbf{m}_{epoch}}{1 - \beta_1^{epoch}}$.
    Correct bias: $\hat{\mathbf{v}}_{epoch} = \frac{\mathbf{v}_{epoch}}{1 - \beta_2^{epoch}}$.
    Update parameters: $\boldsymbol{\iota}_{\text{LM}} = \boldsymbol{\iota}_{\text{LM}} - \alpha \frac{\hat{\mathbf{m}}_{epoch}}{\sqrt{\hat{\mathbf{v}}_{epoch}} + \epsilon}$.
  end for
end for
return $\boldsymbol{\iota}_{\text{LM}}$.

1) LSTM Network Optimization: The objective of the training is to optimize the parameters symbolized by $\boldsymbol{\iota}_{\text{LM}}^*$ and to develop a curve $\wp(\mathbf{o}_t^*, \boldsymbol{\iota}_{\text{LM}})$ that minimizes the divergence from the target system parameters $\mathbf{o}_t$. The optimization function for the LSTM network is given by:

$$\boldsymbol{\iota}_{\text{LM}}^* = \arg \min_{\boldsymbol{\iota}_{\text{LM}}} \text{Dev}(\boldsymbol{\iota}_{\text{LM}})$$

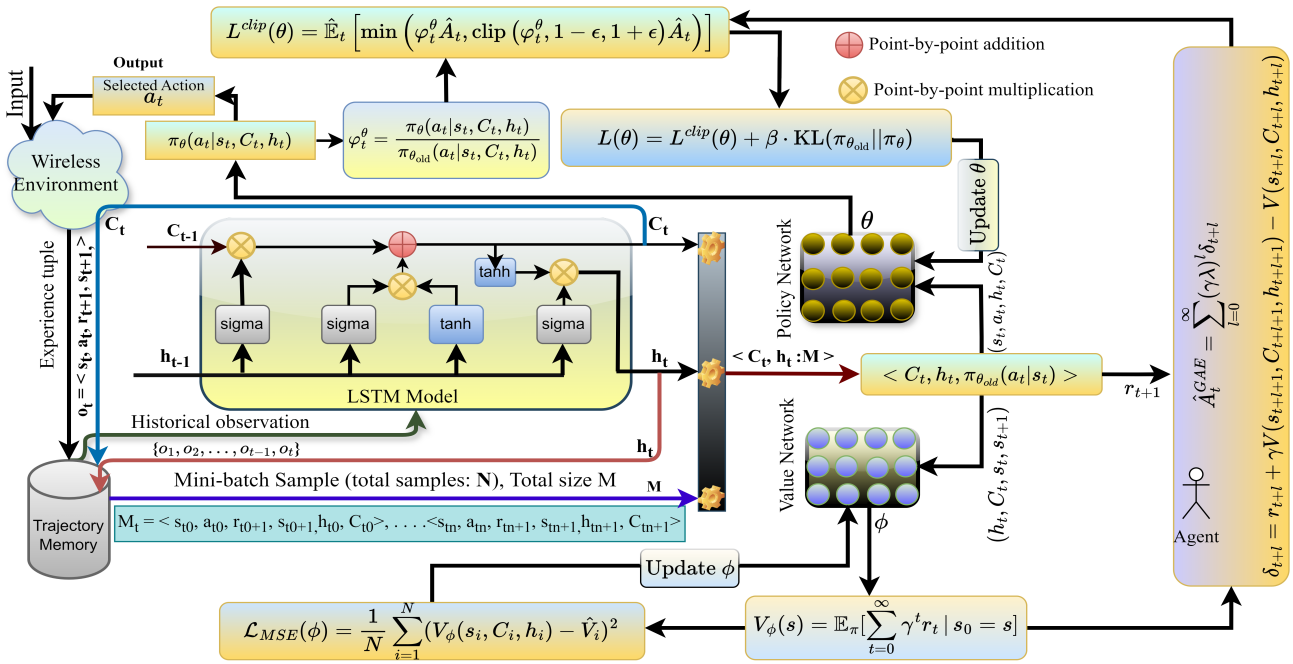


Fig. 4: Block diagram of the proposed hybridized LSTM-DRL model

$$= \arg\min_{\boldsymbol{\iota}_{\mathrm{LM}}} \sum_{i=1}^{\|\mathbf{o}\|} \left[\mathbf{o}_t^i - \wp(\hat{\mathbf{o}}_t^i, \boldsymbol{\iota}_{\mathrm{LM}})\right]^2. \tag{31}$$

Given the complexity in obtaining $\boldsymbol{\iota}^*_{\mathrm{LM}}$, an approximation is considered for simplicity:

$$\begin{aligned}\boldsymbol{\iota}_{\mathrm{LM}}^{\mathrm{approx}} &= \arg\min_{\boldsymbol{\iota}_{\mathrm{LM}}} \mathrm{Dev}(\boldsymbol{\iota}_{\mathrm{LM}}) \\ &\approx \sum_{i=1}^{\|\mathbf{o}\|} \left[\mathbf{o}_t^i - \wp(\hat{\mathbf{o}}_t^i, \boldsymbol{\iota}_{\mathrm{LM}})\right]^2 + \nu \sum_{i=1}^{\|\mathbf{o}\|} T(\mathbf{o}_t^i, \mathbf{o}_{t-1}^i) \\ &\quad + \lambda \|\boldsymbol{\iota}_{\mathrm{LM}}\|_2 - \kappa \sum_{i=1}^{\|\mathbf{o}\|} E(\wp(\hat{\mathbf{o}}_t^i, \boldsymbol{\iota}_{\mathrm{LM}})). \end{aligned} \tag{32}$$

In (32), $\sum_{i=1}^{\|\mathbf{o}\|} \left[\mathbf{o}_t^i - \wp(\hat{\mathbf{o}}_t^i, \boldsymbol{\iota}_{\mathrm{LM}})\right]^2$ quantifies the divergence between the predicted system parameters, $\hat{\mathbf{o}}_t$, and the authentic system parameters, $\mathbf{o}_t$. Additionally, the term $T(\mathbf{o}_t^i, \mathbf{o}_{t-1}^i)$, $\nu \in (0,1)$, signifies the temporal consistency criterion, striving for coherence between subsequent time steps. The temporal consistency term penalizes large deviations between subsequent model predictions to ensure smoothness over time. We use the log-cosh function for this: $T(\mathbf{o}_t^i, \mathbf{o}_{t-1}^i) = \log\left(\cosh\left(\mathbf{o}_t^i - \mathbf{o}_{t-1}^i\right)\right)$. The regularization term, represented by $\lambda \|\boldsymbol{\iota}_{\mathrm{LM}}\|_2$, aids in preventing overfitting by penalizing large model parameter values; we use $\lambda = 0.1$. The entropy term, $\kappa \sum_{i=1}^{\|\mathbf{o}\|} E(\wp(\hat{\mathbf{o}}_t^i, \boldsymbol{\iota}_{\mathrm{LM}}))$, where $\kappa \in (0,1)$, promotes prediction diversity, ensuring that the model does not become overly deterministic and captures the inherent stochasticity. To optimize the LSTM objective function, we employ the proposed Algorithm 1.

A. Ensembling of Present LSTM with Proposed DRL

Using LSTM states improves our DRL model's decision-making capabilities by considering temporal dependencies. We integrate those LSTM states with the policy and value network outputs. The policy network, denoted as $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)$, outlines the action probabilities. Meanwhile, the value network symbolized as $V_\phi(\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)$ provides estimates of expected rewards from any given state. Following a strategy similar to the policy optimization methodology presented in [24], we adopt a clipped surrogate function, $L^{clip}(\theta)$, ensuring a fair balance between exploration and exploitation as follows:

$$L^{clip}(\theta) = \mathbb{E}_t\left[\min\left(\varphi_t^\theta \hat{A}_t, \mathrm{clip}\left(\varphi_t^\theta, 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right]. \tag{33}$$

To understand the influence of shifts in policy parameters on the clipped objective, we derive $L^{clip}(\theta)$ with respect to θ. This derivation captures the sensitivity of the surrogate function to changes in the policy parameters. Recognizing small changes in φ_t^θ due to changes in θ can be challenging. However, using the Taylor series expansion, we approximate these shifts as:

$$\varphi_t^\theta \approx \varphi_t^{\theta_{\mathrm{old}}} + \frac{\partial \varphi_t^\theta}{\partial \theta}\Big|_{\theta_{\mathrm{old}}} (\theta - \theta_{\mathrm{old}}). \tag{34}$$

In this expression, the gradient describes how the function changes for small alterations in the policy parameters around the point θ_{old}. Here, the ratio φ_t^θ contrasts the new policy, π_θ, with the preceding policy, $\pi_{\theta_{\mathrm{old}}}$, at the time instance t and which is written as:

$$\varphi_t^\theta = \frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)}{\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)}. \tag{35}$$

To eliminate any sudden policy alterations, we adopt KL divergence [36] between the novel and earlier policies remains within the confines of ϵ. Therefore, by including the KL policy, the primary objective function in (33) is consolidated to:

$$L(\theta) = L^{clip}(\theta) + \beta \cdot \mathrm{KL}(\pi_{\theta_{\mathrm{old}}}||\pi_\theta). \tag{36}$$

In this equation, β steers the KL constraint. With the Taylor expansion, the divergence between the old and the current policy becomes:

$$\mathrm{KL}(\pi_{\theta_{\mathrm{old}}}||\pi_\theta) \approx \frac{1}{2}(\theta - \theta_{\mathrm{old}})^T \mathbb{E}_t\left[\nabla\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)\right.$$

Algorithm 2 MA-DRL for xURLLC Communication

Initialization:

Define states $\mathcal{S}$ and actions $\mathcal{A}$ for agents. Define the reward function r_t with weights w_1, w_2, w_3, Initialize LSTM. Initialize policy networks π_θ and value networks V_ϕ, *Parameters:* episode = K, number of steps per episode T, clip parameter ϵ, number of LSTM layers L, number of hidden units H, discount factor γ, GAE parameter λ, mini-batch size M, entire trajectory $\tau(\forall t \in T)$.

Output: Each agent optimizes resource to maximize (24).

Interaction with environment:

for Episode = 1, 2, ..., K **do**
  for time step t **do**
    Base Station Agent:
      Collect state $\mathbf{s}_t^{\text{bs}}$
      Select action $\mathbf{a}_t^{\text{bs}} = \pi_\theta^{\text{bs}}(\mathbf{s}_t^{\text{bs}}, \mathbf{h}_t^{\text{bs}}, \mathbf{C}_t^{\text{bs}})$.
    for each UAV-mounted CRIS Agent $\forall i \in Q$ **do**
      Collect states $\mathbf{s}_t^q[i]$ for agent q.
      Select action $\mathbf{a}_t^q[i] = \pi_\theta^q(\mathbf{s}_t^i, \mathbf{h}_t^i, \mathbf{C}_t^i)$.
    end for
    Execute $\mathbf{a}_t$, where, $\mathbf{a}_t = (\mathbf{a}_t^{\text{bs}} \cup \mathbf{a}_t^q)$, $\mathbf{a}_t^q[i] \in \mathbf{a}_t^q$.
    Receive reward: $r_t(\mathbf{s}_t, \mathbf{a}_t)$, $\mathbf{s}_t = (\mathbf{s}_t^{\text{bs}} \cup \mathbf{s}_t^q)$
    for BS Agent **do**
      Update BS agent's LSTM using **Algorithm 1**:
        $\mathbf{C}_t^{\text{bs}} = f_c(\mathbf{C}_{t-1}^{\text{bs}}, \mathbf{o}_t^{\text{bs}})$, $\mathbf{h}_t^{\text{bs}} = f_h(\mathbf{h}_{t-1}^{\text{bs}}, \mathbf{o}_t^{\text{bs}})$
      Add $(\mathbf{s}_t^{\text{bs}}, \mathbf{a}_t^{\text{bs}}, r_{t+1}, \mathbf{h}_t^{\text{bs}}, \mathbf{C}_t^{\text{bs}})$ to trajectory τ
      if timestep equals T **then**
        Add trajectory τ to the set D
        Reset the environment to its initial state
        Reset the LSTM hidden and cell states
        Initialize a new trajectory $\tau = \{\}$
      end if
    end for
    for each UAV-mounted CRIS agent $\forall i \in Q$ **do**
      Update i^{th} agent's LSTM using **Algorithm 1**:
        $\mathbf{C}_t^i = f_c(\mathbf{C}_{t-1}^i, \mathbf{o}_t^i)$, $\mathbf{h}_t^i = f_h(\mathbf{h}_{t-1}^i, \mathbf{o}_t^i)$
      Append $(\mathbf{s}_t^i, \mathbf{a}_t^i, r_{t+1}, \mathbf{h}_t^i, \mathbf{C}_t^i)$ to trajectory τ
      if timestep equals T **then**
        Add trajectory τ to the set D
        Reset the environment to its initial state
        Reset the LSTM hidden and cell states
        Initialize a new trajectory $\tau = \{\}$
      end if
    end for
  end for
  Share information among all agents for cooperation
  Call Training **Algorithm 3**
end for

$$\times \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)^T \Big] (\theta - \theta_{\text{old}}), \quad (37)$$

where the term $\mathbb{E}_t[\nabla\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t) \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{C}_t)^T]$ represents the Fisher information matrix [48], which captures the expected curvature of the log-likelihood of the policy trajectory. It quantifies how the distribution of the policy's output is sensitive to parameter changes θ. The quadratic form of the approximation ensures that it is more accurate when the current policy parameters θ are close to the old parameters θ_{old}.

Having established the modifications and constraints on policy updates to ensure a smooth evolution of the agent's strategy, our next focus is on maximizing the agent's cumulative rewards over time, which is encapsulated as:

$$V_\phi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \mathbf{s}_0 = \mathbf{s} \right]. \quad (38)$$

To ensure effective learning, it becomes imperative to understand how shifts in the value function parameters influence the expected return. This sensitivity of the value function, in terms of its parameters ϕ, is expressed as:

$$\frac{\partial V_\phi(s)}{\partial \phi} = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t A_t^{\text{GAE}} \frac{\partial \delta_t}{\partial \phi} \mid \mathbf{s}_0 = \mathbf{s} \right]. \quad (39)$$

By understanding the above gradient, agents can make more informed decisions, attributing the influence of ϕ on expected returns, thereby aiding in efficient credit assignment. With the aid of GAE in (39), the advantage $\hat{A}_t$ is formulated as:

$$\hat{A}_t^{\text{GAE}} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}, \quad (40)$$

where γ signifies the discount factor, λ manages the bias-variance equilibrium, and δ_{t+l} equates the disparities between authentic rewards and predicted values:

$$\delta_{t+l} = r_{t+l} + \gamma V_\phi(\mathbf{s}_{t+l+1}, \mathbf{C}_{t+l+1}, \mathbf{h}_{t+l+1}) - V_\phi(\mathbf{s}_{t+l}, \mathbf{C}_{t+l}, \mathbf{h}_{t+l}). \quad (41)$$

The $L(\theta)$ in (36) updates the policies and concurrently updates the value network using MSE loss:

$$\mathcal{L}_{MSE}(\phi) = \frac{1}{M} \sum_{i=1}^{M} (V_\phi(\mathbf{s}_i, \mathbf{h}_i, \mathbf{C}_i) - \hat{V}_i)^2. \quad (42)$$

By minimizing $\mathcal{L}_{MSE}(\phi)$, the agent refines its value estimates. Concurrently, optimizing $L(\theta)$ guides the policy towards actions that maximize expected future rewards. The delicate balance between these updates, ensured by our surrogate loss and KL divergence constraints, leads to a robust and effective learning algorithm. Our MA-DRL-LSTM-based algorithm is provided in Algorithm 2.

Algorithm 3 Training in DRL for xURLLC Communication

Initialization:

Markovian property: The environment follows the MDP framework. Policy Initialization: Initial policies are stochastic. Exploration: ϵ-greedy.

for each trajectory τ in D **do**
  for timestep = 1, 2, ..., length of τ **do**
    BS agent do
    Compute TD error:
      $\delta_{t+l}^{\text{bs}} = r_{t+l} + \gamma V(\mathbf{s}_{t+l+1}^{\text{bs}}, \mathbf{C}_{t+l+1}^{\text{bs}}, \mathbf{h}_{t+l+1}^{\text{bs}}) - V(\mathbf{s}_{t+l}^{\text{bs}}, \mathbf{C}_{t+l}^{\text{bs}}, \mathbf{h}_{t+l}^{\text{bs}})$
    for each UAV-mounted CRIS agent $\forall i \in Q$ **do**
      Compute TD error:
        $\delta_{t+l}^q[i] = r_{t+l} + \gamma V(\mathbf{s}_{t+l+1}^i, \mathbf{C}_{t+l+1}^i, \mathbf{h}_{t+l+1}^i) - V(\mathbf{s}_{t+l}^i, \mathbf{C}_{t+l}^i, \mathbf{h}_{t+l}^i)$
    end for
    Collect TD error: $\delta_{t+l} = (\delta_{t+l}^{\text{bs}} \cup \delta_{t+l}^q)$
    Compute GAE: $\hat{A}_t^{GAE(\gamma,\lambda)} = \sum_{l=0}^{T-t} (\gamma\lambda)^l \| \delta_{t+l} \|$
    Compute $r_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k+1}$
  end for
end for
for each mini-batch of size M **do**
Policy update:
  Compute stochastic policy gradient: $\nabla_\theta L^{clip}(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} \nabla_\theta \min\left(\varphi_i^\theta \hat{A}_i, \text{clip}\left(\varphi_i^\theta, 1-\epsilon, 1+\epsilon\right) \hat{A}_i\right)$
  Compute overall objective function: $L(\theta) = L^{clip}(\theta) - \beta \cdot \text{KL}(\pi_{\theta_{\text{old}}} || \pi_\theta)$
Value network update:
  Compute MSE loss gradient for V_ϕ: $\nabla_\phi \mathcal{L}_{MSE}(\phi) \approx \frac{1}{M} \sum_{i=1}^{M} \nabla_\phi (V_\phi(\mathbf{s}_i, \mathbf{h}_i, \mathbf{C}_i) - \hat{V}_i)^2$
  Update: $\phi \leftarrow \phi + \alpha \nabla_\phi \mathcal{L}_{MSE}(\phi)$
end for

B. Multi-Agent Approach

Our MA-DRL approach emphasizes coordination and cooperation among UAV-mounted CRIS agents and BS agents to enhance sum data rate maximization in xURLLC networks.

TABLE II: List of simulation parameters and their corresponding values.

Parameters	Value	Parameters	Value	Parameters	Value	Parameters	Value
BS's antennae	$M = 4$ [14]	BS transmission power	33 dBm	Noise at RIS	$\sigma^2_{2,q} = -80$dBm [6]	Maximum block-length	$l_{max} = 256$ [49]
xURLLC's antennae	$L = 3$ [12]	Minimum data rate	$R_k^{\min} = 1$bps/Hz [14]	Noise at xURLLC	$\sigma^2_{1,k} = -80$dBm [6]	Carrier frequency	$f_q = 2.0$GHz [7]
No. of xURLLC	$K = 6$ [16]	UAV max velocity	$v_q^{\max} = 10$m/s [7]	Rician factor	$\forall\beta = 10$ [24]	Decoding error probability	$\xi_k^{\max} = 10^{-9}$ [2]
Antenna separation	$\zeta = 7.5$cm	PE's max amplitude	$\eta_{(q,n)}^{dmax} = 10$ [50]	Path-loss exponent: $\mathbf{h}_k$	$\alpha_i = (3.5, 3.0)$	No. UAV-mounted CRIS unit	$Q = 4$ [16]
No. of PE in CRIS	$N = 72$	CSI error norm bound	$\{\varrho, \varepsilon, \varphi\} = 0.01$ [44]	Path-loss exponent: $\mathbf{G}_k^{(d,q)}$	$\alpha_i = (2.75, 2.50)$	Path-loss exponent: $\mathbf{F}_q$	$\alpha_i = (3.0, 2.75)$

We carefully extend the traditional actor-critic methods by integrating a joint action-learning phase where agents learn not just individually but also consider the impact of their actions on the system's objectives. This is done through collaborative reward calculation, helping to understand how actions interrelate and guide the system toward goals.

To further enhance the stability and performance of our multi-agent LSTM-based DRL framework, we introduce target networks for both the actor and the critic components. This addition aims to mitigate the rapid oscillations in value estimates and policy updates that can hamper the learning process. Target networks provide more stable target values for the TD error calculations and policy updates, thereby smoothing the training dynamics and facilitating better convergence. We employ separate target networks for the policy and value functions, denoted as $\pi_{\theta'}$ (target actor) and $V_{\phi'}$ (target critic), respectively. These networks are clones of their corresponding main networks but with their parameters (θ' for the actor and ϕ' for the critic) updated less frequently. This setup ensures that the target values against which the main networks' outputs are compared remain relatively stable over several iterations. Target networks' parameters are updated using a soft update strategy defined by the equations:

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta', \tag{43}$$

$$\phi' \leftarrow \tau\phi + (1-\tau)\phi', \tag{44}$$

where $\tau = 0.005$ controls the rate at which the target networks are updated. This integration ensures that the target values evolve smoothly, contributing to the overall stability of the training process. During the training phase, the TD error for the base station agent and the UAV-mounted CRIS agents is calculated using the target critic network. This modification to the TD error calculation utilizes the target critic for estimating the value of the next state, significantly enhancing the stability of the value updates. The adjusted TD error calculation in Algorithm 3 for the base station agent, for instance, is represented as:

$$\delta_{t+l}^{\text{bs}} = r_{t+l} + \gamma V_{\phi'}(\mathbf{s}_{t+l+1}^{\text{bs}}, \mathbf{C}_{t+l+1}^{\text{bs}}, \mathbf{h}_{t+l+1}^{\text{bs}}) - V_{\phi}(\mathbf{s}_{t+l}^{\text{bs}}, \mathbf{C}_{t+l}^{\text{bs}}, \mathbf{h}_{t+l}^{\text{bs}}). \tag{45}$$

The inclusion of target networks in our framework is anticipated to significantly reduce the volatility of policy and value estimates during training.

C. Computational Complexity Analysis of Proposed Algorithm

During interaction with the environment, the algorithm collects state-action-reward data for both the base station agent and each UAV-mounted CRIS agent, updates LSTM memory cells ($O(L)$), appends trajectory data, shares information, and conducts mini-batch training updates ($O(M)$). The overall time complexity for this interaction is approximately $O(K \cdot T \cdot (1 + L \cdot q + M))$. For the training algorithm, which involves TD errors, GAE, discounted rewards, and policy and value network updates, the time complexity is approximately $O(D \cdot T \cdot (1 + M))$. Therefore, the total time complexity of the entire algorithm is $O(K \cdot T \cdot (1 + L \cdot q + M) + D \cdot T \cdot (1 + M))$.

D. Proposed MA-DRL Convergence Analysis

Consider a multi-agent system where each agent i utilizes a policy π_{θ_i} and a critic V_{ϕ_i}, parameterized by vectors θ_i and ϕ_i, respectively. Target networks for actors and critics are denoted as $\pi_{\theta_i'}$ and $V_{\phi_i'}$, updated with coefficients τ_θ and τ_ϕ, ensuring smooth updates. According to the policy gradient theorem, the update rule for the policy parameters θ_i is:

$$\theta_i^{(k+1)} = \theta_i^{(k)} + \alpha \mathbb{E}_{\pi_{\theta_i}} \left[\nabla_{\theta_i} \log \pi_{\theta_i}(\mathbf{a}_t|\mathbf{s}_t) A^{\pi}(\mathbf{s}_t, \mathbf{a}_t) \right], \tag{46}$$

where $A^{\pi}(\mathbf{s}_t, \mathbf{a}_t) = Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t) - V^{\pi}(\mathbf{s}_t)$ represents the advantage function, and α is the learning rate. The critic parameters ϕ_i are updated using the MSE loss:

$$\phi_i^{(k+1)} = \phi_i^{(k)} - \beta \nabla_{\phi_i} \left[r_{t+1} + \gamma V_{\phi_i}(\mathbf{s}_{t+1}) - V_{\phi_i}(\mathbf{s}_t) \right]^2, \tag{47}$$

where β is the critic's learning rate. The soft update rules for the target networks are defined as:

$$\theta_i' \leftarrow (1-\tau_\theta)\theta_i' + \tau_\theta \theta_i, \quad \phi_i' \leftarrow (1-\tau_\phi)\phi_i' + \tau_\phi \phi_i, \tag{48}$$

which stabilizes learning by providing more consistent target values for policy and value updates.

Under the assumptions of Lipschitz continuity and bounded gradients [51], stochastic approximation theory guarantees convergence of $\{\theta_i^{(k)}\}$ and $\{\phi_i^{(k)}\}$ to local minimizers of the expected return and value functions. The updates for θ_i and ϕ_i reduce the KL-divergence between the policies and the target policies, thereby ensuring convergence to a policy that maximizes the expected return as follows:

$$\min_{\theta_i, \phi_i} \text{KL}(\pi_{\theta_i'} || \pi_{\theta_i}) + \lambda \mathbb{E}_{\pi_{\theta_i'}} \left[Q^{\pi_{\theta_i'}}(s, a) \right], \tag{49}$$

where λ is a penalty term moderating the rate of convergence and promoting exploration. This describes the agents' gradual convergence towards an optimal policy performance measure in a complex, stochastic multi-agent environment.

V. Numerical Results and Analysis

In this section, we extensively evaluate the performance of our proposed scheme using comprehensive simulations. As benchmarks, we compare our proposed LSTM-based MA-DRL algorithm against three popular DRL algorithms: deep-Q networks (DQN) [11], DDPG [10], and the PPO approach [24], known for its stability through the clipped objective function.

In our DRL simulations, we adhere to rigorously selected parameter configurations derived from best practices in the literature [4], [7], [9]–[11], [16], [24]. The learning rate (α) is fixed at 0.001, complemented by a discount factor (γ) of

0.99 [25]. We allocate an experience replay buffer of 10^6 samples and designate a batch size of 64 for consistent network updates. For the DQN algorithm, the exploration rate (ϵ) starts at an assertive 1.0, decaying to a minimum threshold of 0.01 at a decay rate of 0.995. The pivotal update of the target network occurs every 500 step. Our network configuration boasts four strategically crafted hidden layers (i.e., fully connected with a dropout probability of 0.2), each with 256 neurons, operationalized with the ReLU activation function. In the DDPG domain, the actor and critic networks, equipped with four hidden layers of 256 neurons, exhibit learning rates of 0.0001 and 0.001, respectively. The soft update coefficient (τ) stands at a precise 0.005, and our exploration is accentuated by the Ornstein-Uhlenbeck noise process with a θ of 0.15 and a σ of 0.2. Turning our attention to the PPO algorithm, it inherits the actor and critic learning rates of 0.0001 and 0.001, respectively. The policy's clip range remains non-negotiable at 0.2 [24], and we've instated the GAE parameter (λ) at 0.9. Both the policy and value networks are fortified with four hidden layers of 128 neurons each, operating under the ReLU activation function. For scenarios demanding recurrent architectures, our LSTM layers are carefully designed with 256 cells. Furthermore, we set the LSTM's dropout rate at 0.25 to mitigate overfitting. The sequence length is fixed to 100 time steps, ensuring adequate memory for temporal dependencies.

A. Simulation Setup and Parameters

In the subsequent subsection, we comprehensively describe the environmental setup, other simulation parameters, and relevant details. In addition to the proposed algorithm, DQN, DDPG, and PPO algorithms are implemented using Python and TensorFlow. The computational resources employed for these simulations comprise an RTX GPU with 16 GB of memory, 40 GB of RAM, and an Intel i7 CPU operating at 2.90 GHz. Our system model operates in a stochastic environment wherein the channel state undergoes continuous alterations within specified boundaries at each time step.

In our model, we define a spherical space encompassing 750m^3, within which all communication nodes coexist. Base Station (BS) is located at the central coordinate $(0, 0, 10)$ in the (x, y, z) plane, while $z \geqslant 0$. The UAV-mounted CRIS units navigate throughout the designated volume and are uniformly distributed. Additionally, xURLLC users are spatially distributed within a subset of this space, occupying 300m^3. Further essential parameters and their corresponding values are carefully tabulated in Table II. Unless we redefine the parameter with new values, the value remains fixed as given in Table II anywhere in the following text.

B. Numerical Results and Discussions

In Fig. 5, the convergence plot distinctly illustrates the superior performance of our proposed LSTM-based MA-DRL algorithm in comparison to the established DQN [11], DDPG [10], PPO [24], MADDPG [28], and MAPPO [30] benchmarks. Our method's fusion of LSTM networks allows it to effectively capture long-term dependencies in state-action sequences, thus providing a minute understanding of temporal

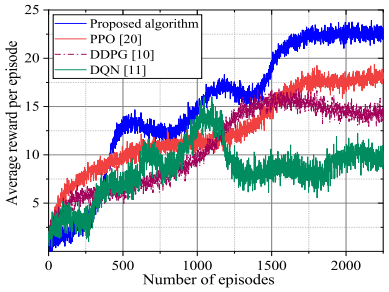


Fig. 5: Convergence plot: Comparing with existing DRL algorithms.

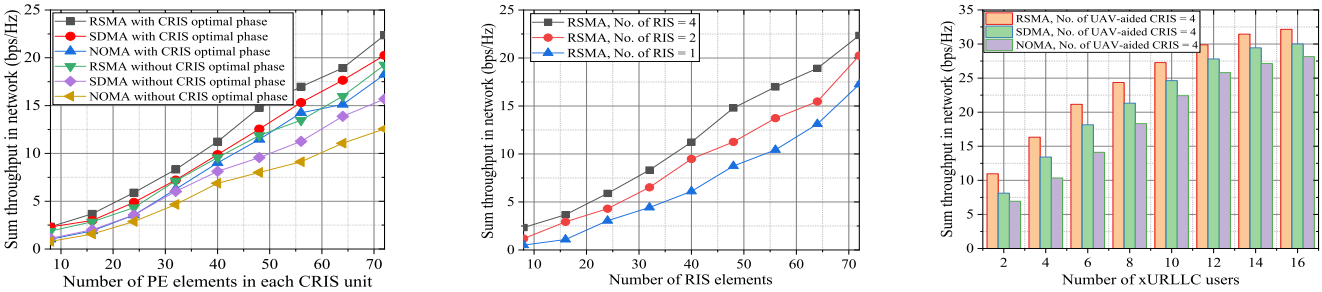
dynamics. When paired with our precisely designed functions, this inherent capability ensures optimal action selection and value approximation explicitly tailored for our problem domain. Moreover, our multi-agent framework facilitates a more comprehensive view of the environment, emphasizing cooperative and strategic decision-making. Together, these elements of temporal understanding and function design precision give our algorithm a decisive advantage in the defined environment. Around episode number 1750, it can be observed that both the proposed algorithm and PPO approach a state of convergence, indicating stable learning. In contrast, DDPG and DQN display evident oscillatory behaviours, suggesting they have not converged consistently. Notably, the proposed LSTM-based DRL algorithm's performance significantly surpasses PPO, DDPG, and DQN, registering accumulated average reward gains of 1.19 times, 1.62 times, and 2.34 times higher, respectively.

TABLE III: A comparative analysis of different RIS modes in sum rate maximization (bps/Hz), $Q = 4$, $N = 72$, $K = 6$, $\{M, L\} = 4$.

RIS	Perfect CSI	Imperfect CSI	RIS	Perfect CSI	Imperfect CSI
PT	21.587	14.259	PRAT	24.846	17.153
AT	23.417	16.287	PSTAR	27.129	18.351
PR	20.118	15.637	ARPT	27.032	18.027
AR	22.324	17.178	ASTAR	27.758	18.658
Proposed: CRIS		Perfect CSI: 31.327		Imperfect CSI: 22.654	

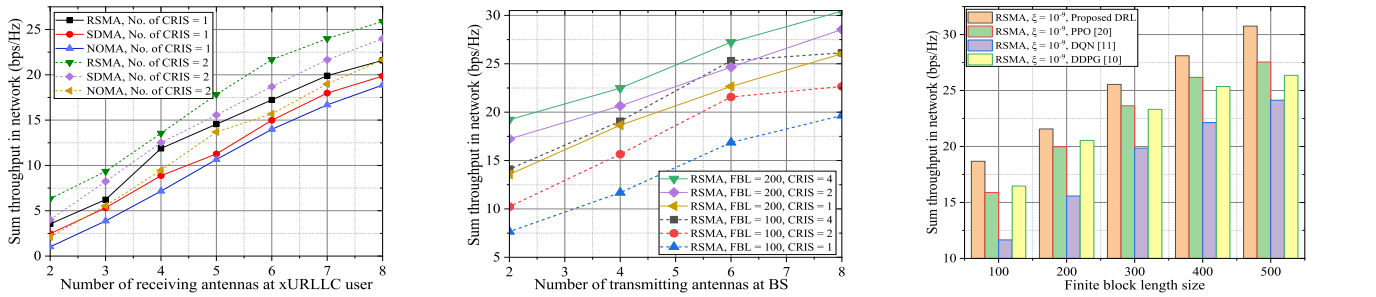
Table III showcases the superior performance of the proposed CRIS model when it integrates with our system model, and the LSTM-based DRL algorithm manages the resource allocation. The results manifest the edge the CRIS model holds over other RIS operation modes in perfect and imperfect channel scenarios. This enhanced performance is attributed to the adept self-assessment and mode selection capabilities of individual PE elements in the RIS, a distinctive feature of our model. It is pivotal to underline that traditional RIS modes also operate within our proposed system model, further emphasizing the efficacy of CRIS. For example, compared to the advanced ASTAR mode, the CRIS model shows an approximately $\sim$ 12.87% performance improvement in the Perfect CSI scenario. Similarly, CRIS exhibits superiority in the Imperfect CSI context, improving about $\sim$ 21.42% relative to the ASTAR mode.

In our subsequent in-depth performance analysis presented in Fig.6a, we compare various transmission schemes. Specifically, we investigate our chosen RSMA technique, contrasting it with SDMA and NOMA schemes. Empirical evaluations



(a) Comparison with different transmission schemes. (b) Impact of CRIS number on throughput. (c) xURLLC users' impact on total throughput.

Fig. 6: Overall analysis of transmission schemes, CRIS number impact, and xURLLC user impact on total throughput.



(a) Impact of receiving antennas. (b) FBL and CRIS unit effect on throughput. (c) Decoding error probability effect on throughput.

Fig. 7: Analysis of receiving antennas, FBL and CRIS unit effect, and decoding error probability effect on throughput.

show that the RSMA scheme, coupled with the UAV-mounted CRIS modules in a MIMO system, consistently outperforms the other methods, whether we account for optimal phase design for the PE elements in CRIS or ignore it. The salient advantage of RSMA emerges from its inherent ability to utilize both the spatial domain, through multiple antennas, and the power domain. This dual-domain utilization ensures that users experiencing a range of channel conditions receive optimal service, all while minimizing interference. Meanwhile, SDMA, which allows spatial differentiation of users in a MIMO setting, sometimes faces obstacles, particularly in high-user-density scenarios. Similarly, NOMA, known for its power domain multiplexing, occasionally exhibits sensitivity when encountering imperfect channel state information, a challenge that MIMO settings tend to amplify. For the scenario with optimal CRIS phase shift design, RSMA outperforms SDMA by approximately 10.38% and surpasses NOMA by around 22.48% when 72 PE elements are in individual CRIS units.

As depicted in Fig. 6b, a systematic examination of the sum throughput for RSMA reveals the influence of the number of PE elements within varying CRIS configurations. As the number of PE elements escalates, coupled with increased CRIS units, the resultant throughput experiences significant enhancement. For a configuration deploying 4 CRIS units and 72 PE elements, a throughput of 22.339 bps/Hz is achieved. This value is approximately 10.32% superior to the ~ 20.25 bps/Hz realized with 2 CRIS units. Furthermore, compared with the throughput of 17.247 bps/Hz from a single CRIS unit setup, the four-unit configuration demonstrates an impressive gain of nearly ~ 29.53%.

Fig. 6c demonstrates the sum throughput performances of RSMA, SDMA, and NOMA schemes as the number of xURLLC users increases, each utilizing a fixed number of UAV-mounted CRIS units. Specifically, when the number of xURLLC users is 16, RSMA manifests a 32.156 bps/Hz throughput. This throughput is approximately ~ 7.14% superior to the 30.012 bps/Hz achieved using SDMA and a commendable ~ 14.22% higher than the 28.154 bps/Hz observed with NOMA. An intriguing observation across all schemes is the consistent increase in sum rate despite the constant resources such as transmission power and the number of UAV-mounted CRIS units. This phenomenon can be justified by the multi-user diversity inherent to wireless communication systems. As the user count grows, the system exploits these users' distinct and independent channel conditions, maximizing the sum rate. In particular, schemes like RSMA proficiently exploit the spatial domain, implying that an increment in user count leads to enhanced utilization of spatial resources, culminating in an augmented sum rate.

Fig. 7a presents the sum throughput performance of RSMA, SDMA, and NOMA schemes against the increasing number of receiving antennas at the xURLLC user. Distinctly, when the number of receiving antennas is 8 and with 1 CRIS, RSMA achieves a throughput of 21.5847 bps/Hz, which is approximately ~ 8.68% higher than the 19.856 bps/Hz achieved by SDMA and about ~ 14.39% more than the 18.865 bps/Hz observed with NOMA. In the configuration with 2 CRIS, RSMA records a throughput of 25.879 bps/Hz, marking a gain of approximately ~ 7.88% over SDMA's 23.987 bps/Hz and an impressive ~ 19.47% over NOMA's 21.6547 bps/Hz. This consistent outperformance of RSMA can be attributed to its adeptness in harnessing the spatial domain. Unlike SDMA and NOMA, RSMA optimally leverages spatial multiplexing, enabling simultaneous transmission of several

data streams, thus boosting the sum rate. The addition of CRIS units further amplifies RSMA's capabilities, permitting it to shape the wireless environment favourably. As the number of receiving antennas rises, the benefits from spatial diversity and multiplexing in RSMA become even more pronounced, leading to its pronounced superiority over SDMA and NOMA.

In Fig. 7b, we investigate the interplay between the FBL, the count of CRIS units, and the number of transmitting antennas at the BS on the sum throughput within the RSMA paradigm. Focusing on configurations where the number of transmitting antennas at the BS equals 4, we find substantial fluctuations in the RSMA's efficacy contingent on the FBL and the employed CRIS elements. Specifically, with FBL = 100, escalating the CRIS units from 1 to 4 augments the throughput by an impressive $\sim 63.15\%$. In contrast, for FBL = 200, the throughput experiences a boost of $\sim 20.51\%$ when transitioning from a singular CRIS unit to a 4-unit assembly. These findings underscore the intrinsic merit of integrating more CRIS modules, particularly in settings characterized by a reduced FBL.

Fig. 7c showcases the sum throughput performance of RSMA under a fixed packet error probability rate of $\xi_k = 10^{-9}$ as facilitated by various algorithms, namely our Proposed LSTM-based DRL, PPO [24], DQN [11], and DDPG [10], as the finite block length size increases. An imperative observation from the table is the distinct supremacy of the Proposed LSTM-based DRL over the other algorithms at all finite block lengths. For instance, at a block length of 500, the LSTM-based DRL achieves a throughput of 30.757 bps/Hz. This is an impressive 11.7% improvement over the next best, which is 27.547 bps/Hz achieved by PPO for the same block length and a massive 26.9% gain over DDPG's 24.132 bps/Hz. The DQN, in this scenario, lags even further. The results suggest that our LSTM-based DRL adeptly handles the challenges posed by the intricate dynamics of RSMA, especially under the rigorous reliability constraints of xURLLC. LSTM, which inherently captures temporal dependencies, probably empowers the DRL to better predict and react to the ever-changing state of the wireless environment.

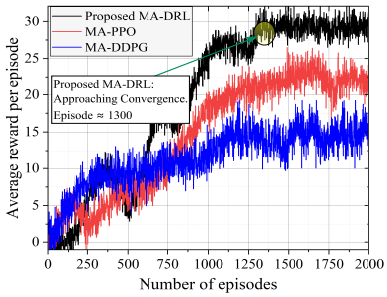


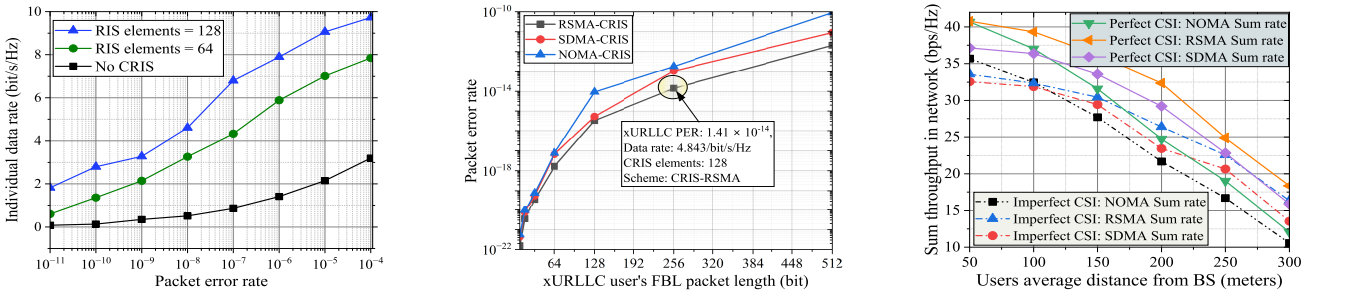
Fig. 8: Convergence plot: multi-agent multi-actor-critic networks.

Fig. 8 illustrates the convergence characteristics of a multi-agent framework encompassing our proposed MA-DRL and established benchmark DRL algorithms such as MA-DDPG [28] and MA-PPO [30]. Each agent, whether a BS or UAV-mounted CRIS, is equipped with an actor-network for making action decisions and a critic network for assessing the current state's desirability. These networks are complemented by corresponding target networks, a target actor-network ($\pi_{\theta'}$) and a target critic network ($V_{\phi'}$), which are updated less frequently to provide stable targets for temporal-difference error calculations and policy updates, thereby smoothing the training dynamics. This configuration of multi-actor-critic-based learning networks facilitates smoother convergence and achieves a marginally higher accumulated reward compared to the results shown in Fig. 5, which utilized a single actor-critic network. The superior performance of our MA-DRL can be attributed to the enhanced stability provided by the target networks, which mitigates the oscillations commonly seen with frequent updates in MA-DDPG [28] and the over-exploration issues in MA-PPO [30]. Additionally, our framework's ability to maintain a consistent learning trajectory without the abrupt policy shifts typically observed in MA-PPO further enhances its reward accumulation over time. For the proposed MA-DRL, we observe a more rapid convergence by $\approx$ 1300th episode in Fig. 8, as compared to the convergence rates depicted in Fig. 5. The integration of target networks and the refined handling of multi-agent dynamics significantly reduce the instability associated with frequent updates of Q-values and policies, offering a more stable benchmark for learning, which enhances the overall convergence properties of the algorithm.

Fig. 9a demonstrates the impact of packet error rate on individual data rate for different numbers of PE elements (i.e., RIS elements) in the proposed CRIS module. It is noted that relaxing the packet error rate increases the individual data rate. The data rate also gets enhanced as the number of PE elements is increased. At a packet error rate of 10^{-9}, which is considered the maximum for xURLLC, the CRIS module with 128 RIS elements significantly increases the data rate by approximately 9.28 times compared to the system with no CRIS module, and about 1.53 times when compared to the system with 64 RIS elements. These gains highlight the efficiency of incorporating a larger number of PE elements in the CRIS module for enhancing communication performance in critical network scenarios.

Fig. 9b demonstrates the impact of xURLLC's short packet length (i.e., FBL) on packet error rate for the present work. Increasing the packet length in xURLLC increases the packet error rate due to greater exposure to channel imperfections and error propagation risks. Importantly, the RSMA technique substantially lowers the PER over SDMA and NOMA schemes. For a packet length of 256 bits, RSMA achieves a PER of 1.41×10^{-14}, whereas SDMA and NOMA exhibit PERs of 1.07×10^{-13} and 1.821×10^{-13}, respectively. This shows that RSMA reduces the PER by approximately 7.59 times compared to SDMA and about 12.92 times compared to NOMA at this packet length. This significant reduction in PER by RSMA can be attributed to its relative robustness to imperfect CSI, which is critical in SDMA for beamforming and in NOMA for power allocation. By not strictly requiring perfect CSI, RSMA maintains lower PERs under practical channel conditions.

Fig. 9c illustrates the throughput dynamics in xURLLC networks, focusing on users' average proximity to the base station under varying CSI conditions. Users are grouped within this range at a distance of 50 meters, revealing NOMA's



(a) Packet error rate's impact on xURLLC data rate. (b) FBL's impact on xURLLC packet error rate. (c) Throughput comparison with users' proximity.

Fig. 9: Analysis of packet error rates, packet lengths, and throughput in comparison with users' proximity.

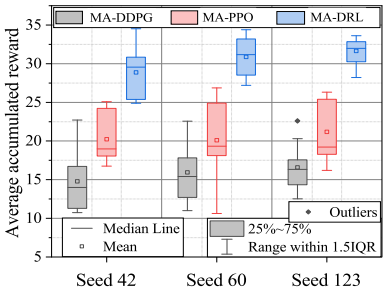


Fig. 10: Variance in training: a Box plot of accumulated rewards for MA-DRL, MA-PPO, and MA-DDPG across different random seeds.

superior performance in managing dense user clusters by efficiently leveraging power domain resources. This results in a throughput gain of 9.55% under imperfect CSI and 9.54% under perfect CSI compared to SDMA, emphasizing NOMA's effectiveness in high-density settings up to 100 meters. As the distance increases to 300 meters, RSMA begins to significantly outperform both NOMA and SDMA, with gains of 20.53% and 54.46% over SDMA and NOMA respectively, under imperfect CSI and 15.57% and 52.17% under perfect CSI. These notable improvements with RSMA at extended ranges are due to its hybrid approach, which adeptly combines power and spatial domain strategies. This dual-domain exploitation enables RSMA to adaptively manage interference and allocate resources more dynamically, thereby enhancing throughput and network efficiency in scenarios with dispersed user distributions and varied channel conditions. The robustness and flexibility of RSMA make it particularly effective for ensuring reliable and high-throughput communications across larger distances in xURLLC networks.

Fig. 10 presents the box plot for the MA-DRL, MA-PPO, and MA-DDPG algorithms to demonstrate the variance in the training phase. Each box plot captures the median, interquartile range (IQR), and outliers, providing a comprehensive view of the data distribution across different training scenarios. The median, a robust measure of central tendency, is indicated by the line within each box. The box itself, spanning from the first quartile (25th percentile) to the third quartile (75th percentile), encapsulates the middle 50% of the data and illustrates the data's dispersion. Whiskers on the plots extend up to 1.5 times the IQR from the quartiles, effectively highlighting the spread of the majority of the data. This box plot layout is useful in conveying how each algorithm performs under varying conditions influenced by the random seed selection, specifically using seeds 42, 60, and 123. Such visualization aids in understanding the sensitivity of each algorithm to initialization and other stochastic factors, thereby providing insights into their stability and robustness. The accumulated reward in episode 2000 is used as a metric to measure the effectiveness of the training process for each algorithm in the MA-DRL context.

VI. Conclusions

A key aspect of our research was the innovative design of a CRIS unit. This unit is characterized by its adaptability; each element within the CRIS is proficiently engineered to synchronize with the environment, determining and selecting the most optimal operational mode. This dynamism ensures that the CRIS consistently yields superior outcomes, optimizing communication over any conventional mode of operations by the RIS. Moreover, integrating this CRIS with UAVs endows the system with heightened mobility and expansive area coverage, surmounting conventional limitations. Our introduction and successful deployment of the multi-agent LSTM-based DRL strategy fortify the research's significance. This method, set against traditional algorithms like PPO, DQN, and DDPG, consistently displayed superior performance. It demonstrated its superiority in navigating the complex dynamics of RSMA, especially when offering gains of 11.7% and 26.9% in sum throughput compared to PPO and DDPG, both operating under the strict xURLLC reliability constraint of a packet error probability rate of 10^{-9}. Our developed system showcases innovative design and precision in performance, which ensures the highest level of xURLLC reliability. The key insights that are derived from our work can be briefly summarized as

The practical implementation of UAV-mounted CRIS technologies introduces complex challenges that span across various interdisciplinary fields, from micro-controller design specific to CRIS operations to advancements in UAV dynamics for enhanced stability and adaptability. While presently unexplored in our study, this multifaceted research domain opens a significant scope for future investigations. Recognizing these challenges does not emphasize a limitation of our current research but rather outlines a clear path for future work.

References

[1] J. Park, S. Samarakoon, H. Shiri, M. K. Abdel-Aziz, T. Nishio, A. Elgabli, and M. Bennis, "Extreme ultra-reliable and low-latency communication," *Nature Electronics*, vol. 5, no. 3, pp. 133–141, 2022.

[2] B. Han *et al.*, "Flexible and dependable manufacturing beyond xURLLC: A novel framework for communication-control co-design," in *Proc. Int. Conf. Softw. Qual., Reliab., Secur. Companion*, 2022, pp. 562–568.

[3] E. Dias *et al.*, "Sliding window network coding enables next generation URLLC millimeter-wave networks," *IEEE Netw. Lett.*, vol. 5, no. 3, pp. 159–163, 2023.

[4] H. Z. *et al.*, "Resource management for multiplexing eMBB and URLLC services over RIS-aided THz communication," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1207–1225, 2023.

[5] A. Ranjha*et al.*, "URLLC facilitated by mobile UAV relay and RIS: A joint design of passive beamforming, blocklength, and UAV positioning," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4618–4627, 2021.

[6] K. Zhi, C. Pan, H. Ren, K. K. Chai, and M. Elkashlan, "Active RIS versus passive RIS: Which is superior with the same power budget?" *IEEE Commun. Lett.*, vol. 26, no. 5, pp. 1150–1154, 2022.

[7] P. Yang *et al.*, "Proactive UAV network slicing for URLLC and mobile broadband service multiplexing," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3225–3244, 2021.

[8] M. Elwekeil, A. Zappone, and S. Buzzi, "Power control in cell-free massive MIMO networks for UAVs URLLC under the finite blocklength regime," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1126–1140, 2023.

[9] M. Alsenwi *et al.*, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 7, pp. 4585–4600, 2021.

[10] A. S. Abdalla and V. Marojevic, "Multiagent learning for secure wireless access from UAVs with limited energy resources," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 22 356–22 370, 2023.

[11] P. Chen, X. Li, M. Matthaiou, and S. Jin, "DRL-based RIS phase shift design for OFDM communication systems," *IEEE Wirel. Commun. Lett.*, vol. 12, no. 4, pp. 733–737, 2023.

[12] Q. Chen, J. Wu, J. Wang, and H. Jiang, "Coexistence of URLLC and eMBB services in MIMO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 839–851, 2023.

[13] D.-B. Ha, V.-T. Truong, and Y. Lee, "Performance analysis for rf energy harvesting mobile edge computing networks with SIMO/MISO-NOMA schemes," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 8, no. 27, Apr. 2021.

[14] X. Ou, X. Xie, H. Lu, and H. Yang, "Resource allocation in MU-MISO rate-splitting multiple access with SIC errors for URLLC services," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 229–243, 2023.

[15] T. Shafique, H. Tabassum, and E. Hossain, "Optimization of wireless relaying with flexible UAV-borne reflecting surfaces," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 309–325, 2021.

[16] Y. Li, C. Yin, T. Do-Duy, A. Masaracchia, and T. Q. Duong, "Aerial reconfigurable intelligent surface-enabled URLLC UAV systems," *IEEE Access*, vol. 9, pp. 140 248–140 257, 2021.

[17] Z. Zhai, X. Dai, B. Duo, X. Wang, and X. Yuan, "Energy-efficient UAV-mounted RIS assisted mobile edge computing," *IEEE Wirel. Commun. Lett.*, vol. 11, no. 12, pp. 2507–2511, 2022.

[18] D. Tyrovolas, P.-V. Mekikis, S. A. Tegos, P. D. Diamantoulakis, C. K. Liaskos, and G. K. Karagiannidis, "Energy-aware design of UAV-mounted RIS networks for IoT data collection," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1168–1178, 2023.

[19] Y. Xiao, D. Tyrovolas, S. A. Tegos, P. D. Diamantoulakis, Z. Ma, L. Hao, and G. K. Karagiannidis, "Solar powered UAV-mounted RIS networks," *IEEE Commun. Lett.*, vol. 27, no. 6, pp. 1565–1569, 2023.

[20] E. M. Mohamed, S. Hashima, and K. Hatano, "Energy aware multiarmed bandit for millimeter wave-based UAV mounted RIS networks," *IEEE Wirel. Commun. Lett.*, vol. 11, no. 6, pp. 1293–1297, 2022.

[21] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surv. Tutorials*, vol. 23, no. 3, pp. 1546–1577, 2021.

[22] K. K. Nguyen *et al.*, "Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 3, pp. 358–368, 2022.

[23] X. Yuan *et al.*, "Deep reinforcement learning-driven reconfigurable intelligent surface-assisted radio surveillance with a fixed-wing UAV," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 4546–4560, 2023.

[24] C. Meng, K. Xiong, W. Chen, B. Gao, P. Fan, and K. B. Letaief, "Sum-rate maximization in STAR-RIS-assisted RSMA networks: A PPO-based algorithm," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 5667–5680, 2024.

[25] R. Zhong, X. Mu, Y. Liu, Y. Chen, J. Zhang, and P. Zhang, "STAR-RISs assisted NOMA networks: A distributed learning approach," *IEEE J. Sel. Top. Signal Process.*, vol. 17, no. 1, pp. 264–278, 2023.

[26] P. Sunehag *et al.*, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 2085–2087.

[27] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4295–4304.

[28] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.

[29] M. R. Maleki, M. R. Mili, M. R. Javan, N. Mokari, and E. A. Jorswieck, "Multi-agent reinforcement learning trajectory design and two-stage resource management in CoMP UAV VLC networks," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7464–7476, 2022.

[30] O. Lohse, N. Pütz, and K. Hörmann, "Implementing an online scheduling approach for production with multi agent proximal policy optimization (MAPPO)," in *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems: IFIP WG 5.7 Int. Conf., APMS 2021, Nantes, France, September 5–9, 2021, Proceedings, Part V*. Springer, 2021, pp. 586–595.

[31] A. Lotfolahi and H.-W. Ferng, "A multi-agent proximal policy optimized joint mechanism in mmwave hetnets with CoMP toward energy efficiency maximization," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 1, pp. 265–278, 2024.

[32] B. Liu, W. Ni, R. P. Liu, Y. J. Guo, and H. Zhu, "Decentralized, privacy-preserving routing of cellular-connected unmanned aerial vehicles for joint goods delivery and sensing," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9627–9641, 2023.

[33] P. Qin, S. Wang, Z. Lu, Y. Xie, and X. Zhao, "Deep reinforcement learning-based energy minimization task offloading and resource allocation for air ground integrated heterogeneous networks," *IEEE Syst. J.*, vol. 17, no. 3, pp. 4958–4968, 2023.

[34] K. K. Nguyen, A. Masaracchia, and C. Yin, "Deep reinforcement learning for intelligent reflecting surface-assisted D2D communications," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 10, no. 1, pp. 1–8, Jan. 2023.

[35] H. Mei *et al.*, "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, 2022.

[36] E. Laz and U. Orguner, "Gaussian mixture filtering with nonlinear measurements minimizing forward kullback-leibler divergence," *Signal Processing*, vol. 208, p. 108992, 2023.

[37] F. Kavehmadavani, V.-D. Nguyen, T. X. Vu, and S. Chatzinotas, "Intelligent traffic steering in beyond 5G open RAN based on LSTM traffic prediction," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7727–7742, 2023.

[38] J. Xue, K. Yu, T. Zhang, H. Zhou, L. Zhao, and X. Shen, "Cooperative deep reinforcement learning enabled power allocation for packet duplication URLLC in multi-connectivity vehicular networks," *IEEE Trans. Mobile Comput.*, pp. 1–15, 2024.

[39] Y. Liu, X. Wang, J. Mei, G. Boudreau, H. Abou-Zeid, and A. B. Sediq, "Situation-aware resource allocation for multi-dimensional intelligent multiple access: A proactive deep learning framework," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 116–130, 2021.

[40] R. Li, C. Wang, Z. Zhao, R. Guo, and H. Zhang, "The LSTM-based advantage actor-critic learning for resource management in network slicing with user mobility," *IEEE Commun. Lett.*, vol. 24, no. 9, pp. 2005–2009, 2020.

[41] S. Pala, M. Katwe, K. Singh, B. Clerckx, and C.-P. Li, "Spectral-efficient RIS-aided RSMA URLLC: Toward mobile broadband reliable low latency communication (mBRLLC) system," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 3507–3524, 2024.

[42] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1719–1734, 2020.

[43] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[44] R. Allu *et al.*, "Robust beamformer design in active RIS-assisted multiuser MIMO cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 2, pp. 398–413, 2023.

[45] K. Singh*et al.*, "Transceiver design and power control for full-duplex ultra-reliable low-latency communication systems," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 2, pp. 1392–1406, 2022.
[46] Y. Hua, Z. Zhao, R. Li, X. Chen, Z. Liu, and H. Zhang, "Deep learning with long short-term memory for time series prediction," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 114–119, 2019.
[47] C. Yang *et al.*, "Logic synthesis optimization sequence tuning using RL-based LSTM and graph isomorphism network," *IEEE Trans. Circuits Syst. II, Express Briefs*, vol. 69, no. 8, pp. 3600–3604, 2022.
[48] Y. Takeishi, M. Iida, and J. Takeuchi, "Approximate spectral decomposition of fisher information matrix for simple ReLU networks," *Neural Networks*, vol. 164, pp. 691–706, 2023.
[49] S. He, Z. An, J. Zhu, J. Zhang, Y. Huang, and Y. Zhang, "Beamforming design for multiuser uRLLC with finite blocklength transmission," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 12, pp. 8096–8109, 2021.
[50] H. Niu, Z. Lin, K. An, J. Wang, G. Zheng, N. Al-Dhahir, and K.-K. Wong, "Active RIS assisted rate-splitting multiple access network: Spectral and energy efficiency tradeoff," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1452–1467, 2023.
[51] T. T. Doan, S. T. Maguluri, and J. Romberg, "Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach," *IEEE Trans. Autom. Control*, vol. 66, no. 10, pp. 4469–4484, 2021.

Anal Paul (Member, IEEE) received his Bachelor of Technology degree from the Government College of Engineering and Ceramic Technology, India, in 2008, and his Master of Engineering degree from Jadavpur University, India, in 2010. In 2021, he received his Ph.D. degree from the Indian Institute of Engineering Science and Technology, Shibpur. From July to December 2022, he worked as a postdoctoral researcher in the Department of Information and Communication Engineering at Yeungnam University, South Korea. Since January 2023, he has been a Postdoctoral Researcher at National Sun Yat-sen University, Taiwan, conducting research in Digital Twin and Metaverse applications for Wireless Communication Systems.

Raviteja Allu (Graduate Student Member, IEEE) received his B.Tech degree in Electrical and Electronics Engineering from GMR institute of technology affiliated to JNTU Kakinada, India in 2016, M. Tech degree in Electrical Engineering from NIT Rourkela, India in 2019. He is currently a PhD scholar at the Institute of Communication engineering, National Sun Yat-sen University (NSYSU), Taiwan. His current research interests are in the areas of cognitive radio communications, backscatter communications, reconfigurable intelligent surfaces-assisted communications, full-duplex wireless systems, integrated sensing and communications, and MIMO communications.

Keshav Singh (Member, IEEE) received the Ph.D. degree in Communication Engineering from National Central University, Taiwan, in 2015. He currently works at the Institute of Communications Engineering, National Sun Yat-sen University (NSYSU), Taiwan as an Associate Professor. Prior to this, he held the position of Research Associate from 2016 to 2019 at the Institute of Digital Communications, University of Edinburgh, U.K. From 2019 to 2020, he was associated with the University College Dublin, Ireland as a Research Fellow. He chaired workshops on conferences like IEEE GLOBECOM 2023 and IEEE WCNC, 2024. He also serves as leading guest editor of IEEE Transactions on Green Communications and Networking Special Issue on Design of Green Near-Field Wireless Communication Networks and IEEE Internet of Things Journal Special Issue on Positioning and Sensing for Near-Filed (NF)-driven Internet-of-Everything. He leads research in the areas of green communications, resource allocation, transceiver design for full-duplex radio, ultra-reliable low-latency communication, non-orthogonal multiple access, machine learning for wireless communications, integrated sensing and communications, non-terrestrial networks, and large intelligent surface-assisted communications.

Chih-Peng Li (Fellow, IEEE) received the B.S. degree in Physics from National Tsing Hua University, Hsin Chu, Taiwan, and the Ph.D. degree in Electrical Engineering from Cornell University, NY, USA. Dr. Li was a Member of the Technical Staff with Lucent Technologies. Since 2002, he has been with National Sun Yat-sen University (NSYSU), Kaohsiung, Taiwan, where he is currently a Distinguished Professor. Dr. Li has served in various positions with NSYSU, including the Chairman of the Electrical Engineering Department, the VP of General Affairs, the Dean of Engineering College, and the VP of Academic Affairs. His research interests include wireless communications, baseband signal processing, and data networks. He is now the Director General of the Engineering and Technologies Department, at the National Science and Technology Council, Taiwan.

Dr. Li is currently the Chapter Chair of the IEEE Broadcasting Technology Society Tainan Section. Dr. Li has also served as the Chapter Chair of the IEEE Communication Society Tainan Section, the President of the Taiwan Institute of Electrical and Electronics Engineering, the Editor of IEEE Transactions on Wireless Communications, the Associate Editor of IEEE Transactions on Broadcasting, and the Member of Board of Governors with IEEE Tainan Section. Dr. Li has received various awards, including the Outstanding Research Award from the Ministry of Science and Technology. Dr. Li is a Fellow of the IEEE.

Trung Q. Duong (Fellow, IEEE) is a Canada Excellence Research Chair (CERC) and a Full Professor at Memorial University, Canada. He is also the adjunct Chair Professor in Telecommunications at Queen's University Belfast, UK and a Research Chair of Royal Academy of Engineering, UK. He was a Distinguished Advisory Professor at Inje University, South Korea (2017-2019), an Adjunct Professor and the Director of Institute for AI and Big Data at Duy Tan University, Vietnam (2012-present), and a Visiting Professor (under Eminent Scholar program) at Kyung Hee University, South Korea (2023-2025). His current research interests include quantum communications, wireless communications, quantum machine learning, and quantum optimisation.

Dr. Duong has served as an Editor/Guest Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS MAGAZINES, and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He received the Best Paper Award at the IEEE VTC-Spring 2013, IEEE ICC 2014, IEEE GLOBECOM 2016, 2019, 2022, IEEE DSP 2017, IWCMC 2019, 2023, and IEEE CAMAD 2023. He has received the two prestigious awards from the Royal Academy of Engineering (RAEng): RAEng Research Chair (2021-2025) and the RAEng Research Fellow (2015-2020). He is the recipient of the prestigious Newton Prize 2017.