Intelligent Spectral Efficiency Maximization for OTFS-RSMA Communications in High-Mobility IoT Networks

Kiet Nguyen Tuan Tran, Thanh Phung Truong, *Graduate Student Member, IEEE,* Tung Son Do *Graduate Student Member, IEEE,* Manh Cuong Ho *Graduate Student Member, IEEE,* Nhu-Ngoc Dao, *Senior Member, IEEE,* and Sungrae Cho, *Senior Member, IEEE*

***Abstract*—Ensuring high spectral efficiency (SE) for mobile Internet of Things (IoT) devices in high-mobility environments poses challenges due to severe Doppler effects, rapid channel variations, and multiuser interference. This work addresses these problems by proposing a downlink orthogonal time-frequency space (OTFS) and rate-splitting multiple access (RSMA) framework that exploits the Doppler-robust delay–Doppler domain modulation of OTFS and the flexible interference management capability of RSMA to increase SE under high-mobility conditions. In this framework, common and private precoding vectors are jointly designed to maximize the system SE. The resulting nonconvex design problem is formulated as a continuous control task and is solved using a deep reinforcement learning approach based on a normalization-integrated deep deterministic policy gradient agent. The agent outputs precoders for common and private streams, and a closed-form normalization layer ensures feasibility by projecting the actions onto the power-constrained set without penalty tuning. Numerical results over high-mobility OTFS channels demonstrate stable convergence of the learned policy and reveal that the proposed framework consistently outperforms OTFS-wo-RSMA and OTFS-RSMA baselines across a broad signal-to-noise ratio range. These results confirm that learning-based precoding optimization for OTFS-RSMA consolidates Doppler-resilient modulation with robust interference management, enabling spectrally efficient, robust mobile IoT communications in high-mobility scenarios.**

***Index Terms*—Deep reinforcement learning, orthogonal time-frequency space, rate-splitting multiple access.**

I. Introduction

Mobile Internet of Things (IoT) deployments require reliable, high-throughput connectivity for numerous heterogeneous devices operating in mobile environments, with intermittent traffic, and under tight power budgets [1]. Conventional time-frequency (TF) waveforms severely degrade under large Doppler spreads, making them unsuitable for mobile and vehicular IoT scenarios [2], [3]. To address this problem, the orthogonal TF space (OTFS) maps information symbols to 2D basis functions in the delay-Doppler (DD) domain rather than the TF grid [4]. Bits are placed on the DD grid, and the inverse symplectic finite Fourier transform is applied to obtain TF samples. The Heisenberg transform converts the samples to the time domain for transmission. At the receiver, the inverse operations recover the DD symbols [5]. By modulating information in the DD domain, the OTFS transforms a rapidly time-varying multipath channel into a 2D quasi-time-invariant representation over the duration of a data frame. This approach ensures that every symbol is exposed to all channel paths, achieving full channel diversity and nearly constant channel gain. This robustness to high-Doppler effects makes OTFS an attractive and reliable solution for mobile, vehicular, and other mission-critical IoT links where terminals are moving [1].

Although OTFS yields a nearly time-invariant effective channel in the DD domain by exploiting full TF diversity, dense IoT access still requires robust multidevice interference management. Rate-splitting multiple access (RSMA) has emerged as a versatile strategy for this purpose [6]. In the downlink, RSMA partitions each user's message into a common stream, decoded by all devices, and a private stream, decoded individually after interference mitigation. These linearly precoded streams are transmitted nonorthogonally, enabling RSMA to interpolate flexibly between treating interference as noise and performing full interference decoding, scaling to the dense IoT connectivity anticipated in systems beyond the fifth and sixth generations [7], [8].

Integration of OTFS with RSMA is considered on the premise that this combination can improve spectral efficiency (SE) under high-mobility conditions. In large-scale IoT deployments, OTFS provides robustness to severe Doppler and rapid channel variations, whereas RSMA manages multiuser interference. Combined, this integration enables robust, spectrally efficient multidevice communication that meets the reliability and scalability requirements of dense IoT networks.

Beyond waveform and multiple-access design, deep reinforcement learning (DRL) has been adopted across a broad range of adjacent IoT resource-allocation problems. For instance, [9] employs quantum federated reinforcement learning for joint mode selection and resource allocation in STAR-RIS-aided vehicle–road cooperation systems, and [10] uses quantized federated reinforcement learning, including an adaptive-quantized DDPG (AQ-DDPG) for energy-efficiency optimization and an adaptive-quantized DDQN (AQ-DDQN) for latency minimization, to navigate an energy–latency tradeoff in STAR-IRS-assisted carbon-aware IIoT. These works demonstrate the flexibility of DRL formulations across diverse IoT settings and motivate its application to the distinct waveform and multiple-access codesign problem addressed here.

A. Related Work

A critical advantage of OTFS, particularly in high-mobility scenarios, is its ability to transform doubly selective (time-

K.N.T. Tran, T.P. Truong, T.S. Do, M.C. Ho and S.Cho are with Chung-Ang University, School of Computer Science and Engineering, Seoul, Republic of Korea (email: knttran, tptruong, tsdo, hmcuong@uclab.re.kr, srcho@cau.ac.kr)

N.N. Dao is with Research Institute for Data Science and Artificial Intelligence, Hallym University, Seoul, Republic of Korea (email: nndao@hallym.ac.kr).

Corresponding author: Sungrae Cho

TABLE I: Related Work Comparison

Ref	Channel/Network Conditions	Multiaccess	Multiantenna	OTFS	Design / Optimization Method
[5]	Sparse delay–Doppler channel, perfect CSI at receiver	✗	✗	✓	Message-passing iterative detector
[11]	Integer/fractional Doppler, embedded pilots	✗	✗	✓	Embedded pilot-aided channel estimation
[12]	High mobility with imperfect CSI	✗	✗	✓	Model-driven deep-learning detector (OAMP)
[13]	Downlink SISO/MISO RSMA with SIC decodability constraints	RSMA	✓	✗	Joint rate allocation + power control via SCA
[14]	UAV-base station downlink (LoS), joint placement and precoding	RSMA	✓	✗	Analytical performance study with optimized precoding
[15]	Multibeam satellite, inter-beam interference	RSMA	✓	✗	Energy-efficient power/rate allocation (DC programming + SCA)
[16]	Indoor MISO visible light communication	RSMA	✓	✗	Analytical RSMA framework unifying NOMA/SDMA
[17]	Uplink OTFS-RSMA, heterogeneous mobility groups	RSMA	✗	✓	User grouping + cross-domain decoding-order design
[18]	Downlink, high- Doppler, path-selective receiver	RSMA	✗	✓	SCA-based power allocation
Proposed	Downlink OTFS-RSMA with precoding	RSMA	✓	✓	DRL-based joint common+private precoder (DDPG + closed-form power-normalization layer)

and frequency-varying) channels into a quasi-static representation in the DD domain. This transformation yields channel coefficients that remain nearly invariant over the transmission frame, significantly enhancing the communication robustness and reliability of high-velocity IoT devices (IoTDs). For example, a study [5] introduced a structured interference cancellation scheme that employs soft-output, message-passing detection to counteract high-Doppler distortions in OTFS. Similarly, another study [11] proposed an embedded-pilot scheme for OTFS that includes pilot and guard symbols in every frame. This method enables a threshold-based message-passing algorithm to estimate the DD channel and recover data jointly. The study presents a model-driven deep learning-based orthogonal approximate message-passing detector [12], that applies the orthogonal approximate message-passing algorithm in a lightweight neural network to mitigate OTFS channel estimation errors.

Many research efforts have focused on improving the SE in wireless networks using RSMA. A study [13] investigated maximizing the total data rate and proposed a nearly optimal algorithm based on successive convex approximation. In addition, RSMA has been explored for integration into aerial systems [14], [19], [20], positioning it as a valuable technology for non-terrestrial networks. Researchers [15] have focused on achieving energy-efficient RSMA systems using diverse techniques, including multigroup multicast and multibeam satellite communication. The combining of RSMA and another advanced hardware such as reconfigurable intelligent surface has been conducted by [21], which investigates the trade-off between energy efficiency and SE in a multiuser RSMA-assisted active reconfigurable intelligent surface. Further, RSMA has been implemented in visible light communication systems to minimize interference and improve overall communication [16]. However, the RSMA design becomes more challenging under rapid channel variations due to rate-splitting and successive interference cancellation timing constraints.

Building on the robustness of OTFS, researchers have combined it with advanced multiple-access techniques. Studies have explored the synergy between OTFS and RSMA (OTFS-RSMA), reporting significant performance gains. For instance, an uplink cross-domain transmission scheme employing heterogeneous mobile device grouping demonstrated that OTFS-RSMA significantly improves outage performance and device fairness compared with OTFS nonorthogonal multiple access (NOMA), with fairness quantified using a modified Atkinson index [17]. In parallel research, optimizing SE in OTFS-RSMA systems using a path-selective one-tap receiver with power allocation yields substantial average sum-rate improvements in high-mobility scenarios, surpassing benchmarks (e.g., OTFS-NOMA and conventional orthogonal frequency-division multiplexing) [18]. These findings underscore the significant potential of appropriately designed OTFS-RSMA systems to enhance reliability, throughput, and fairness in next-generation high-mobility wireless networks. Table I compares critical related studies, providing an overview of the existing literature. The comparison is based on the considered channel conditions, the multiple-access scheme employed, the use of multiantenna systems, and OTFS modulation.

B. Motivations and Contributions

Previous OTFS-RSMA studies have demonstrated that combining these two methods can enhance outage performance, fairness, and the sum-rate compared with the OTFS-NOMA and orthogonal frequency-division multiplexing (OFDM) baselines. However, these studies have often neglected joint precoding optimization for multiantenna downlink systems and overlook the algorithmic complexity of solving the resulting nonconvex problems under stringent power constraints. Although convex optimization approaches (e.g., weighted minimum mean square error) are standard for RSMA [6], they incur high computational overhead, which is prohibitive in high-mobility OTFS scenarios where channel coherence times are short, necessitating a low-latency, learning-based approach that can map channel states directly to precoding actions. Motivated by these limitations, this work proposes a learning-based OTFS-RSMA design that applies the robustness of

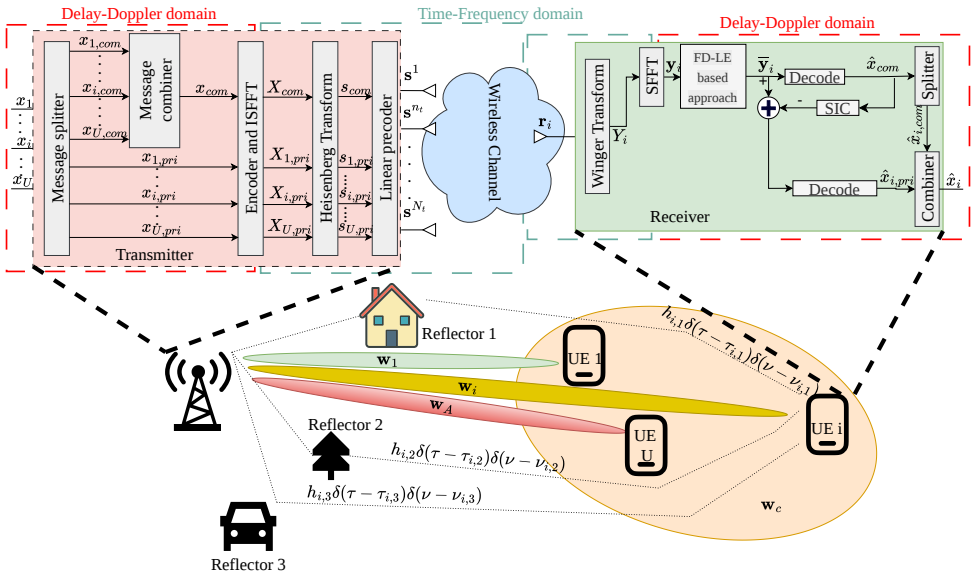


Fig. 1: Downlink OTFS-RSMA system model.

OTFS to channel fluctuations caused by Doppler effects and the flexible interference management RSMA to boost SE in realistic high-mobility, power-constrained mobile IoT scenarios. This work considers a multidevice downlink system in which the base station (BS) employs RSMA to enhance communication efficiency for high-mobility IoTDs, optimizing the precoding of common and private streams to maximize SE. The primary contributions of this work are summarized below.

- To the best of our knowledge, this work is among the earliest studies that consider the precoding optimization in OTFS-RSMA systems. In such systems, a downlink system SE maximization problem is formulated by jointly optimizing the common and private precoding vectors at the BS under a total transmission power constraint, consistent with the frequency-domain linear equalizer (FD-LE) detection chain.
- This work proposes a novel deep reinforcement learning (DRL) framework based on the deep deterministic policy gradient (DDPG) algorithm to learn continuous precoding policies. A closed-form normalization of the DDPG algorithm is integrated to ensure the feasibility of the optimization variables at every step. This algorithm projects the output of the actor network onto the power-constrained set, avoiding environment violation while preserving learning stability.
- The performance of the proposed framework is evaluated under various simulation scenarios. The results reveal fast and stable convergence and consistent sum-rate gains over OTFS-wo-RSMA and OTFS-RSMA baselines across a broad signal-to-noise-ratio (SNR) range. The framework scales favorably with the number of devices, confirming its robustness under heterogeneous mobility and various conditions.

The remainder of this paper is organized as follows. Section II outlines the system model and formal problem statement. Next, Section III describes the proposed framework. Then, Section IV discusses the extensive simulation results. Finally, Section V presents the conclusions.

Notation: The notation $(\cdot)^H$ denotes the conjugate transpose, and $\otimes$ represents the Kronecker product. The trace operator is denoted by $Tr(\cdot)$, and $diag(\cdot)$ and $circ(\cdot)$ construct diagonal and circulant matrices, respectively. In addition, $\mathbf{F}_N$ and $\mathbf{F}_M$ indicate the discrete Fourier transform matrices of orders N and M, respectively.

II. Problem Statement

Fig. 1 illustrates the considered system model, where a BS with N_t antennas serves a group of U high-mobility IoTDs in an OTFS system with N symbols and M subcarriers. Let $\mathcal{A} \triangleq \{1, \ldots, N_t\}$ and $\mathcal{U} \triangleq \{1, ..., U\}$ be the sets of BS antennas and devices, respectively.

A. Transmit Signal Model

At the transmitter, the signal $x_i[k,l]$ of the ith device is placed in the DD domain, where $k \in \{0, \ldots, N-1\}$ and $l \in \{0, \ldots, M-1\}$ denote the Doppler and delay bins, respectively [5]. Next, this approach adopts one-layer rate-splitting, the fundamental architecture of existing RSMA schemes, to process the signal [6]. Based on the RSMA principle, a message splitter divides each transmit signal $x_i[k,l]$ into a common stream $x_{i,\text{com}}[k,l]$ and a private stream $x_{i,\text{pri}}[k,l]$. A message combiner combines the common streams from all IoTDs into a single signal $x_{\text{com}}[k,l]$. Then, the inverse symplectic finite Fourier transform maps each $[k,l]$ symbol to the TF domain [5]. Therefore, the mapped common and private streams are respectively expressed as follows:

$$X_{\text{com}}[n,m] = \frac{1}{NM} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} x_{\text{com}}[k,l]\, e^{j2\pi\left(\frac{nk}{N} - \frac{ml}{M}\right)}, \quad (1)$$

$$X_{i,\text{pri}}[n,m] = \frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1} x_{i,\text{pri}}[k,l]\, e^{j2\pi\left(\frac{nk}{N}-\frac{ml}{M}\right)}, \quad (2)$$

where $n \in \{0,\ldots,N-1\}$ and $m \in \{0,\ldots,M-1\}$. The Heisenberg transform [4] is applied to the TF domain signals $X_{\text{com}}[n,m]$ and $X_{i,\text{pri}}[n,m]$ to obtain the corresponding time-domain signals. Further, $\mathbf{s}_{\text{com}} \in \mathbb{C}^{NM\times 1}$ and $\mathbf{s}_{i,\text{pri}} \in \mathbb{C}^{NM\times 1}$ denote the complete time-domain signal vectors for the common and private streams , respectively, of the ith device. These streams are linearly precoded using a precoding matrix $\mathbf{W} = [\mathbf{w}_{\text{com}}, \mathbf{w}_1, \ldots, \mathbf{w}_U] \in \mathbb{C}^{N_t\times(U+1)}$, where $\mathbf{w}_{\text{com}}, \mathbf{w}_i \in \mathbb{C}^{N_t\times 1}$ are the precoding vectors. Finally, the transmit signal at the BS is expressed as follows:

$$\mathbf{s}^{n_t} = w_{\text{com}}^{n_t}\,\mathbf{s}_{\text{com}} + \sum_{i\in\mathcal{U}} w_i^{n_t}\,\mathbf{s}_{i,\text{pri}}. \quad (3)$$

The terms $w_{com}^{n_t}$ and $w_i^{n_t}$ denote the n_tth entries of the precoding vectors $\mathbf{w}_{\text{com}}$ and $\mathbf{w}_i$, respectively. The transmitted signal $\mathbf{s}^{n_t}$ propagates through a channel characterized by the DD domain response $h_i^{n_t}(\tau,\nu)$, representing the channel between the n_tth antenna at the BS and the ith device, where τ and ν denote the delay and Doppler variables, respectively. This work considers a multipath channel model comprising P propagation paths in the DD domain, where each path is associated with a specific delay and Doppler shift, yielding the following channel impulse response [22]:

$$h_i^{n_t}(\tau,\nu) = \sum_{p=0}^{P-1} h_{i,p}^{n_t}\delta(\tau-\tau_{i,p})\delta(\nu-\nu_{i,p}). \quad (4)$$

for the pth path of the ith device, $h_{i,p}^{n_t}$ represents the complex channel gain, $\tau_{i,p}$ denotes the time delay, and $\nu_{i,p}$ indicates the Doppler shift.

B. Signal Detection Process

At the receiver, the time-domain vector $\mathbf{r}_i \in \mathbb{C}^{NM\times 1}$ is the superposition of signals transmitted from all N_t antennas, convolved with their respective time-domain channels, plus noise:

$$\mathbf{r}_i = \sum_{n_t\in\mathcal{A}} \mathbf{G}_i^{n_t}\left(w_{\text{com}}^{n_t}\mathbf{s}_{\text{com}} + \sum_{q\in\mathcal{U}} w_q^{n_t}\mathbf{s}_{q,\text{pri}}\right) + \mathbf{z}_i, \quad (5)$$

where $\mathbf{G}_i^{n_t} \in \mathbb{C}^{NM\times NM}$ denotes the time-domain channel matrix between the n_tth antenna at the BS and the ith device, and $\mathbf{z}_i \sim CN(0, \mathbf{I}_{NM}\sigma^2)$ represents the additive white Gaussian noise vector. The Wigner transform [5] is applied to $\mathbf{r}_i$. By assuming perfect orthogonality between the transmitted and received pulses, the signal Y_i at the TF domain is expressed as follows:

$$Y_i[n,m] = \sum_{n_t\in\mathcal{A}} H_i^{n_t}[n,m]\left(X_{TF}[n,m]\right)^{n_t} + Z_i[n,m], \quad (6)$$

where $\left(X_{TF}[n,m]\right)^{n_t} \triangleq X_{\text{com}}[n,m]w_{\text{com}}^{n_t} + \sum_{q\in\mathcal{U}} X_{q,\text{pri}}[n,m]w_q^{n_t}$ denotes the total signal transmitted from the n_t antenna at the BS. The detection phase transforms the signal Y_i back to the DD domain via the symplectic finite Fourier transform. The matrix form of the signal in the DD domain is based on [22]:

$$\mathbf{y}_i = \sum_{n_t\in\mathcal{A}} \mathbf{H}_i^{n_t} w_{\text{com}}^{n_t}\mathbf{x}_{\text{com}} + \sum_{q\in\mathcal{U}}\sum_{n_t\in\mathcal{A}} \mathbf{H}_i^{n_t} w_q^{n_t}\mathbf{x}_{q,\text{pri}} + \mathbf{z}_i, \quad (7)$$

where $\mathbf{y}_i \in \mathbb{C}^{NM\times 1}$ represents the received signal vector at the DD domain. The vectors $\mathbf{x}_{\text{com}} \in \mathbb{C}^{NM\times 1}$ and $\mathbf{x}_{i,\text{pri}} \in \mathbb{C}^{NM\times 1}$ collect all symbols $x_{\text{com}}[k,l]$ and $x_{q,\text{pri}}[k,l]$, respectively, in the DD domain. The matrix $\mathbf{H}_i^{n_t} \in \mathbb{C}^{NM\times NM}$ represents the equivalent channel from the n_tth BS antenna to the ith device in the DD domain.

For the detection process, this method employs an FD-LE [23] due to its low computational complexity and effectiveness in diagonalizing the block-circulant OTFS channel, which transforms the complex block-circulant channel matrix $\mathbf{H}_i^{n_t}$ into a diagonal form to facilitate efficient per-subcarrier detection . During this stage, the private signals are treated as interference to facilitate the detection $\mathbf{x}_{\text{com}}$. This process yields the signal expressed in Proposition 1.

Proposition 1: Applying the result from [23], the received signal is written as follows:

$$\bar{\mathbf{y}}_i = \mathbf{x}_{\text{com}} + \sum_{q\in\mathcal{U}} (\mathbf{F}_N\otimes\mathbf{F}_M^H)^{-1}\mathbf{D}^q(\mathbf{F}_N\otimes\mathbf{F}_M^H)\mathbf{x}_{q,\text{pri}} + \tilde{\mathbf{z}}, \quad (8)$$

where $\mathbf{D}^q = (\mathbf{D}_i^{\text{com}})^{-1}\mathbf{D}_{i,q}^{\text{eff}}$, and the terms $\mathbf{D}_i^{\text{com}}$ and $\mathbf{D}_{i,q}^{\text{eff}}$ represent effective diagonal channel matrices for the common and private streams. The matrices are defined as $\mathbf{D}_i^{\text{com}} = \sum_{n_t\in\mathcal{A}} \mathbf{D}_i^{n_t} w_{\text{com}}^{n_t}$ and $\mathbf{D}_{i,q}^{\text{eff}} = \sum_{n_t\in\mathcal{A}} \mathbf{D}_i^{n_t} w_q^{n_t}$. Each antenna diagonal matrix $\mathbf{D}_i^{n_t}$ has its $(kM+l+1)$th main diagonal element given by

$$(D_i^{n_t})^{k,l} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} d_{i,n}^{m,1,n_t}\, e^{j2\pi\left(\frac{ml}{M}-\frac{nk}{N}\right)}, \quad (9)$$

where $d_{i,n}^{m,1,n_t}$ corresponds to the $(nM+m+1)$th row and the first column of $\mathbf{H}_i^{n_t}$. Finally, the noise term is expressed as

$$\tilde{\mathbf{z}} = (\mathbf{F}_N\otimes\mathbf{F}_M^H)^{-1}(\mathbf{D}_i^{\text{com}})^{-1}(\mathbf{F}_N\otimes\mathbf{F}_M^H)\mathbf{z}. \quad (10)$$

Proof: Please see Appendix A.

According to Proposition 1 and assuming normalized noise power, the signal-to-interference-plus-noise ratio (SINR) of the decoded common signal ($\text{SINR}_{i,\text{com}}$) is calculated as follows:

$$\text{SINR}_{i,\text{com}} = \frac{1}{\sum_{q\in\mathcal{U}} Q^q + N_i}, \quad (11)$$

where $Q^q \triangleq \frac{1}{NM}\text{Tr}\{\mathbf{D}^q(\mathbf{D}^q)^H\}$ represents the interference from private streams, and $N_i = \frac{1}{NM}\text{Tr}\{(\mathbf{D}_i^{\text{com}})^{-1}(\mathbf{D}_i^{\text{com}})^{-H}\}$ represents the noise-enhancement expression for the FD-LE process. Following the decoding process, the estimated common signal $\hat{x}_{\text{com}}$ is subtracted from the received signal $\bar{\mathbf{y}}$ using successive interference cancellation. From the decoded signal $\hat{x}_{\text{com}}$, message splitter is used to obtain $\hat{x}_{i,\text{com}}$. The resulting residual signal is then used to detect the private signal $\hat{x}_{i,\text{pri}}$ of the ith device, while treating the private signals of other IoTDs as interference. Thus, the SINR for decoding the private signal is given by

$$\text{SINR}_{i,\text{pri}} = \frac{Q^i}{\sum_{q\in\mathcal{U}\setminus\{i\}} Q^q + N_i}. \quad (12)$$

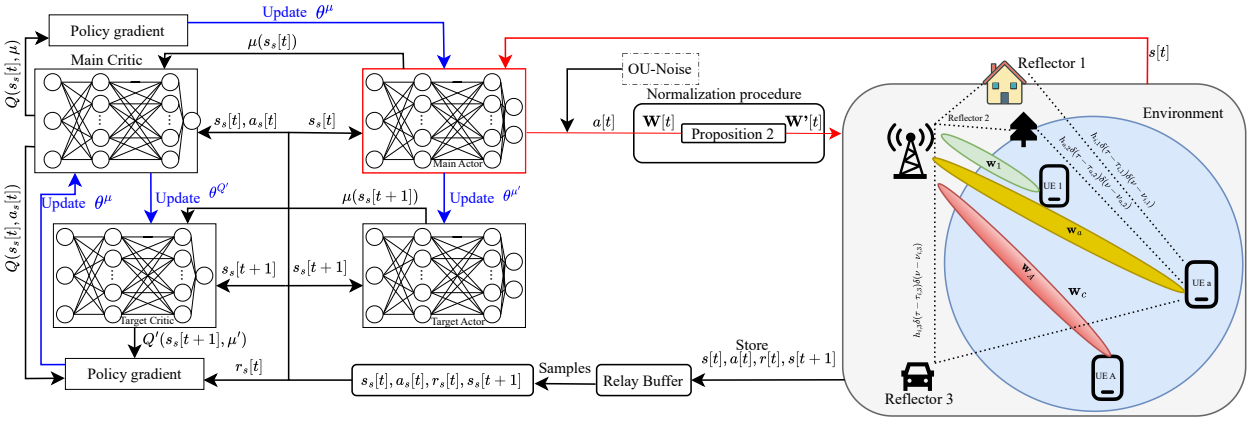


Fig. 2: Proposed framework

The estimated common signal $\hat{x}_{i,\text{com}}$ is combined with the decoded private signal $\hat{x}_{i,\text{pri}}$ to reconstruct the original message $\hat{x}_i$ for device U_i. Based on (11) and (12), the common and private rates are calculated as follows [24]:

$$R_{i,\text{com}} = \log_2(1 + \text{SINR}_{i,\text{com}}), \tag{13}$$

$$R_{i,\text{pri}} = \log_2(1 + \text{SINR}_{i,\text{pri}}). \tag{14}$$

To ensure all IoTDs can decode the common stream, the common rate is determined by the minimum of the common rate of all IoTDs $R_{\text{com}} = \min\{R_{1,\text{com}}, \ldots, R_{i,\text{com}}, \ldots, R_{U,\text{com}}\}$. Then, the selected common rate is continuously distributed among all IoTDs as follows:

$$R_{\text{com}} = \sum_{i \in \mathcal{U}} C_{i,\text{com}}, \tag{15}$$

where $C_{i,\text{com}}$ represents the portion of rate R_{com} distributed to device U_i. The relation in the preceding equation fixes only the aggregate $\sum_{i \in \mathcal{U}} C_{i,\text{com}}$ and therefore leaves the individual shares unspecified. Since the unweighted sum spectral efficiency is invariant to the particular allocation, the equal split $C_{i,\text{com}} = R_{\text{com}}/U$ is adopted throughout for exposition. Thus, the SE for device U_i is calculated as

$$R_{i,\text{total}} = C_{i,\text{com}} + R_{i,\text{pri}}. \tag{16}$$

C. Problem Formulation

The problem focuses on maximizing the system SE by optimizing the precoding matrix subject to the BS transmission power constraints, formulated as follows:

$$\max_{\mathbf{W}} \quad \sum_{i \in \mathcal{U}} R_{i,\text{total}} \tag{17a}$$

$$\text{s.t.} \quad \text{Tr}(\mathbf{W}\mathbf{W}^H) \leq P_t, \tag{17b}$$

where $\mathbf{W}$ denotes the precoding matrix, and the total BS transmission power is P_t. The constraint (17b) ensures that the transmission power does not exceed the power of the BS.

The significant computational complexity of iterative convex approximation algorithms (e.g., successive convex approximation) poses a considerable challenge for solving the nonconvex SE maximization problem in real time [25]–[27]. To address this complexity, this work employs a DRL framework that operates directly on the high-dimensional continuous action space, capturing interactions between precoding, interference, and SE to approximate the optimal precoding matrices.

III. DDPG-based SE Maximization Framework

To address the challenge of maximizing downlink SE in OTFS-RSMA systems operating under high-mobility and resource-constrained conditions, this work employs a DRL framework based on the DDPG algorithm [28]. By formulating the precoding design problem as a Markov decision process , the proposed framework enables efficient exploration of the high-dimensional continuous action space, overcoming the limitations of conventional optimization techniques in dynamic environments with large action sets. As an off-policy, model-free method, the DDPG algorithm is suitable for real-time precoding optimization in rapidly varying channels, where traditional approaches often fail due to imperfect channel state information and complex system dynamics. Furthermore, DRL methods, including the DDPG algorithm, perform well in other dynamic, resource-constrained domains (e.g., aerial communications, adaptive video streaming, and underwater networks) [29]–[32].

A. Problem Formulation as a Markov Decision Process

This work formulates the precoding optimization problem in OTFS-RSMA systems as a discrete-time Markov decision process, where the state space, action space, and reward function are defined as follows.

- **State space:** Each state captures the current communication environment. This work defines the state as the OTFS domain channel gain vector for IoTDs. These states are represented as flattened real and imaginary components of the complex-valued $\mathbf{h}_i^{n_t} \in \mathbb{C}^{P\times 1}$. Formally, the state is

$$s[t] = \{\mathbf{h}_i^{n_t}[t], \ldots, \mathbf{h}_U^{n_t}[t]\}, \quad \forall i \in \mathcal{U}, \forall n_t \in \mathcal{A}. \tag{18}$$

- **Action space:** The action vector comprises continuous variables representing the RSMA precoding matrix for common and private streams:

$$a[t] = \left\{ \ \mathbf{w}_{\text{com}}[t], \mathbf{w}_1[t], \ldots, \mathbf{w}_i[t], \ldots, \mathbf{w}_U[t] \ \right\}, \forall i \in \mathcal{U}. \tag{19}$$

  This vector is passed through a scaling function to ensure that the BS power constraints are met.
- **Reward function:** The reward at each time step is defined as the aggregate of the total achievable rate across all IoTDs:

$$r[t] = \sum_{i \in \mathcal{U}} \mathrm{R}_{i,\text{total}}. \tag{20}$$

Remark 1: The channel state evolves through device mobility and multipath propagation and is statistically independent of the precoding action and is drawn independently across time slots, so the transition kernel does not depend on $a[t]$. The formulation is therefore a contextual bandit, i.e., a degenerate Markov decision process whose transitions are action independent, rather than a genuinely sequential problem. Because $s[t+1]$ is independent of both $s[t]$ and $a[t]$, the action-value function separates as $Q^{\mu}(s,a) = r(s,a) + \gamma\, \mathbb{E}_{s'}[V^{\mu}(s')]$, in which the second term depends on neither the state nor the action. Hence, $\nabla_a Q^{\mu} = \nabla_a r$ and $\arg\max_a Q^{\mu}(s,a) = \arg\max_a r(s,a)$, so the discounted term only adds a constant that cannot change the maximizing precoder. The MDP scaffolding is retained because it is the standard formalism for the actor-critic machinery used here.

B. DDPG-Based SE Maximization

In light of Remark 1, the DDPG agent is employed as a continuous-action solver that amortizes per-realization precoder optimization into a single forward pass, rather than as a mechanism for temporal credit assignment. As illustrated in Fig. 2, the continuous precoding optimization problem in the OTFS-RSMA downlink is solved by the DDPG framework [28], which couples a deterministic policy network (actor) with a value-function network (critic) in an off-policy, model-free setting. Particularly, discrete-action value-based methods such as deep Q-networks are unsuitable here because the joint common and private precoder action $\mathbf{W} \in \mathbb{C}^{N_t \times (U+1)}$ is continuous and high-dimensional under the hard transmit-power constraint in (17b), which would demand an intractable discretization of the action space. Inspired by this observation, DDPG parameterizes the actor as a deterministic mapping $\mu_\theta(s)$ and updates it by the deterministic policy gradient theorem of [33],

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho^{\mu}}\left[\nabla_a Q^{\mu}(s,a)\big|_{a=\mu_\theta(s)} \nabla_\theta \mu_\theta(s)\right], \tag{21}$$

which propagates the gradient directly through the learned action via the chain rule, achieving lower variance than score-function-based stochastic policy gradients in continuous, high-dimensional action spaces, with exploration restored by an additive Ornstein–Uhlenbeck (OU) noise process . To satisfy (17b) at every step while preserving this differentiability chain, the actor output is passed through the closed-form power-normalization layer of Proposition 2. Hence, viewing $\mu(s \mid \theta^{\mu})$ as the complete normalized actor mapping, the update rule in (24) is the standard

Algorithm 1 DDPG-based SE Maximization Algorithm

```
Set up the environment.
Initialize DDPG neural networks: μ(s | θ^μ), Q(s, a | θ^Q).
Initialize model parameters.
Set the number of episodes E and steps per episode T.
for e = 1, 2, ..., e_max do
    Get initial state s[1]
    for t = 1, 2, ..., T do
        Observe current environment state s[t]
        Select action a[t].
        Normalize precoding matrix as Proposition 2 to satisfy power constraint (17b).
        Get next state s[t+1], reward r[t] .
        Store (s[t], a[t], r[t], s[t+1]) in replay buffer.
        Set s[t] = s[t+1].
        Update the actor and critic networks.
        Update the target actor and critic network as (26).
    end for
end for
```

DDPG sample-based approximation of (21), with the on-policy expectation $\mathbb{E}_{s\sim\rho^{\mu}}$ replaced by minibatch states drawn from the replay buffer and the true Q^{μ} replaced by the learned critic $Q(\cdot \mid \theta^{Q})$.

At each time step, the actor network $\mu\big(s \mid \theta^{\mu}\big)$ selects an action a based on the current state s according to its policy, whereas the critic network $Q\big(s, a \mid \theta^{Q}\big)$ evaluates how well that action performs, where θ^{μ} and θ^{Q} denote the weight parameters of the actor and critic networks, respectively.

The weight parameters of the critic network are optimized by minimizing its loss function as follows:

$$L\big(\theta^{Q}\big) = \left[\big(Q(s[t], a[t] \,|\, \theta^{Q}) - y[t]\big)^2\right] \tag{22}$$

where $Q\big(s[t], a[t] \mid \theta^{Q}\big)$ denotes the estimate of the value of taking action $a[t]$ in state $s[t]$ in the critic network, and $y[t]$ represents the target value applied to update the critic network. The computation of $y(t)$ is given by

$$y[t] = r[t] + \gamma\, Q\big(s[t+1],\, \mu(s[t+1]) \mid \theta^{Q}\big) \tag{23}$$

where $\gamma \in [0,1]$ denotes the discount factor. The parameters of the actor network are adjusted using the evaluations of the critic network, by applying the policy gradient update as follows:

$$\nabla_{\theta^{\mu}} J \approx \nabla_{\theta^{\mu}} Q\big(s[t], \mu(s[t] \mid \theta^{\mu}) \mid \theta^{Q}\big). \tag{24}$$

The DDPG algorithm separates the target into the actor and critic networks, denoted by $\mu'(s \mid \theta^{\mu'})$ and $Q'(s, a \mid \theta^{Q'})$, respectively, to enhance training stability. Using these target networks, the update target $y(t)$ from (23) is formulated as follows:

$$y[t] = r[t] + \gamma Q'\big(s[t+1],\, \mu'(s[t+1] \mid \theta^{\mu'}) \mid \theta^{Q'}\big). \tag{25}$$

Target networks are softly updated at each iteration according to

$$\theta^{Q'} \leftarrow \epsilon\, \theta^{Q} + (1-\epsilon)\, \theta^{Q'}, \quad \theta^{\mu'} \leftarrow \epsilon\, \theta^{\mu} + (1-\epsilon)\, \theta^{\mu'} \tag{26}$$

where $\epsilon \in (0,1)$ denotes the soft-update rate for target networks with $\epsilon \ll 1$. Adding noise to the action ensures exploration in the training process. In the DDPG algorithm, noise is generated based on the OU noise process [34]. Let $\mathcal{N}[t]$ denote the noise at

time slot t. Then, the action determined by the DDPG algorithm during training is given by

$$a[t] \;=\; \mu\big(s[t] \mid \theta^{\mu}\big) \;+\; \mathcal{N}[t]. \tag{27}$$

Particularly, the proposed framework realizes the state-to-action mapping through an actor-critic architecture of two deep neural networks trained jointly, which has been regarded as an effective approach for optimizing complex systems in recent works [35], [36]. At each time step t, the state vector $s[t]$, which stacks the real and imaginary parts of the OTFS delay–Doppler channel coefficients $\mathbf{h}_i^{n_t}[t]$ for all users $i \in \mathcal{U}$ and antennas $n_t \in \mathcal{A}$, is fed into the actor network $\mu(s[t] \mid \theta^{\mu})$. The actor output is reshaped into the complex precoder candidates $\{\mathbf{w}_{\text{com}}, \mathbf{w}_1, \ldots, \mathbf{w}_U\}$ and projected onto the feasible set $\{\mathbf{W} : \|\mathbf{W}\|_F^2 \le P_t\}$ by the closed-form power-normalization layer defined in Proposition 2, yielding the executed action $a[t]$. The critic network $Q(s, a \mid \theta^{Q})$ then estimates the action value $Q^{\mu}(s[t], a[t])$, and its parameters are optimized by minimizing the temporal-difference loss built from the target $y[t]$ in (23). The actor parameters are updated by the deterministic policy gradient in (24), which backpropagates $\nabla_a Q$ through μ and shifts the network toward actions of higher expected return. Iterating this coupled actor-critic update across episodes drives $\mu(\cdot \mid \theta^{\mu})$ toward the policy that maximizes the discounted SE objective $\mathbb{E}[\sum_t \gamma^t r[t]]$, which, by Remark 1, is equivalent to maximizing the instantaneous SE $r[t]$ for each channel realization.

Proposition 2: Let $\mathbf{W'} = [\mathbf{w'}_1, \mathbf{w'}_2, \ldots, \mathbf{w'}_{U+1}] \in \mathbb{C}^{N_t \times (U+1)}$ be the precoding matrix after normalization that satisfies the constraint in (17b) with $\mathbf{w'}_b \in \mathbb{C}^{N_t \times 1}$, $b \in \mathcal{B} \triangleq \{1, ..., U+1\}$. The value of $\mathbf{w'}_b$ is calculated as follows:

$$\mathbf{w'}_b = \Theta \mathbf{w}_b, \quad \forall b \in \mathcal{B}, \tag{28}$$

where $\Theta \triangleq \frac{\sqrt{P_t}}{||\mathbf{W}||_F}$ denotes the scaling factor, $\mathbf{w}_b$ represents the precoding vector selected by the actor network, and $||\mathbf{W}||_F$ indicates the Frobenius norm of the matrix.

Proof: The precoding matrix before normalization is $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{U+1}] \in \mathbb{C}^{N_t \times (U+1)}$. The total power transmission is written as follows:

$$\mathrm{Tr}(\mathbf{W}\mathbf{W}^H) = ||\mathbf{W}||_F^2 = \sum_{b=1}^{U+1} ||\mathbf{w}_b||_2^2. \tag{29}$$

The total transmission power after normalization is

$$\mathrm{Tr}(\mathbf{W'}\mathbf{W'}^H) = \sum_{b=1}^{U+1} ||\mathbf{w'}_b||_2^2 \overset{(a)}{=} \Theta^2 \sum_{b=1}^{U+1} ||\mathbf{w}_b||_2^2, \tag{30}$$

where (a) is obtained due to (28). Replacing $\Theta = \frac{\sqrt{P_t}}{||\mathbf{W}||_F}$ with (30) yields

$$\Theta^2 \sum_{b=1}^{U+1} ||\mathbf{w}_b||_2^2 = \frac{P_t}{||\mathbf{W}||_F^2} ||\mathbf{W}||_F^2 = P_t. \tag{31}$$

Therefore, the normalized precoder $\mathbf{W'}$ satisfies the power constraint (17b), which completes the proof.

TABLE II: Delay-Doppler Profile for Device Channels

Propagation path index (P)	0	1	2	3
Delay (τ_p) μs	8.33	25	41.67	58.33
Delay tap index (l_{τ_p})	1	3	5	7
Doppler (ν_p) Hz	0	0	468.8	468.8
Doppler tap index (κ_{ν_p})	0	0	1	1

C. Framework Formulation

The proposed method, detailed in **Algorithm 1** and illustrated in Fig. 2 , outlines the DDPG-based algorithm for maximizing SE in OTFS-RSMA. The training occurs over $e_{\max}$ episodes, each consisting of T time steps. Within each time step t, the agent observes the current state $s[t]$ and decides on an action $a[t]$, which is to add OUnoise to promote exploration. The precoding matrix $\mathbf{W}[t]$ induced by $a[t]$ is passed through the normalization block (Proposition 2) to produce $\mathbf{W'}[t]$, which satisfies the power constraint (17b). The environment executes $\mathbf{W'}[t]$, returning the reward $r[t]$, corresponding to the system SE, calculated in (16). This experience, a sample comprising $(s[t], a[t], r[t], s[t+1])$, is stored in a replay buffer for training. Finally, the state advances to the next time step $s[t+1]$ to begin the next iteration. In Fig. 2, black arrows depict data flow, blue arrows indicate parameter update paths for the main and target networks, and red arrows highlight the executed action and its normalization before interacting with the environment.

D. Complexity Analysis

In this subsection, we analyze the computational complexity of the proposed DDPG-based approach and compare it with a representative optimization-theoretic baseline for the same optimization problem.

1) Complexity of the Proposed DRL-Based Approach: The training process of the DRL approach is performed off-line in advance; thus , this work evaluates the computational complexity of the proposed DRL framework during its interaction with the environment based solely on the actor network. Following the formulation in [35], this complexity is expressed as follows:

$$O\left(\sum_{e=1}^{E} \zeta_{e-1}\zeta_e\right), \tag{32}$$

where ζ_e denotes the number of nodes in layer e. The input layer ($e = 0$) and output layer ($e = E$) correspond to the dimensions of the state $s[t]$ and action $a[t]$ spaces, respectively, such that

$$\begin{aligned} \zeta_0 &= U \times P \times N_t \times 2, \\ \zeta_E &= (U+1) \times 2 \times N_t. \end{aligned} \tag{33}$$

The overall computational complexity of the algorithm simplifies to

$$O(UPN_t). \tag{34}$$

With a fixed-size actor network, the per-step computational cost is $O(UPN_t)$; hence, it linearly increases with each environmental parameter. Increasing U, P, or N_t by a factor of ϱ increases the complexity by the same factor, whereas a simultaneous increment of multiple parameters multiplies their effects. Various parameters, such as grid sizes (N, M), do not affect the complexity in (34) unless they are explicitly included in the state representation, where they increase the input dimension ζ_0 and cost. This linear scaling, combined with a constant depth network, makes latency predictable and the proposed algorithm feasible for real-time operation as the environment scales.

2) Complexity of Other Approaches: For the same optimization problem, we adopt the successive convex approximation (SCA) method as a representative optimization-theoretic complexity reference, which addresses the non-convex precoding objective by linearizing it around the current iterate and solving each resulting convex sub-problem with an interior-point solver. Besides the analytical complexity derived below, the SCA is also implemented and evaluated numerically in Section IV-B3, where its SE and measured runtime are

TABLE III: Online computational complexity comparison.

Algorithm	Per-iteration cost	Total online cost	Offline training
SCA [35]	$O((UN_t)^{3.5}\log_2(\epsilon^{-1}))$	$O(I_{\text{SCA}}\cdot(UN_t)^{3.5}\log_2(\epsilon^{-1}))$	No
Proposed DDPG	Single feed forward pass	$O(UPN_t)$ (linear)	Yes

compared with those of the proposed scheme in Table V. Particularly, the interior-point analysis in [35] applies the standard self-concordant-barrier result to show that the cost to solve each convex sub-problem to ϵ-accuracy scales as $O\big((UN_t)^{3.5}\log_2(\epsilon^{-1})\big)$, where ϵ is the convergence tolerance and the $(UN_t)^{3.5}$ factor combines the $O\big((UN_t)^3\big)$ cost of the linear system solved at each Newton step with the $O\big(\sqrt{UN_t}\big)$ self-concordant-barrier iteration bound, while the $\log_2(\epsilon^{-1})$ factor accounts for the precision required to reach an ϵ-accurate solution. Specifically, because the linearization is local, the convex sub-problem must be reconstructed and re-solved for I_{SCA} outer iterations per channel realization to recover a stationary point of the original non-convex objective, yielding a total online cost of $O\big(I_{\text{SCA}}\cdot(UN_t)^{3.5}\log_2(\epsilon^{-1})\big)$. Hence, the entire iterative procedure is re-executed from scratch whenever the channel changes, with no inherent mechanism to amortize past computation across independent channel realizations.

3) Complexity Comparison: The preceding analyses expose a fundamental structural gap between the two approaches that becomes decisive in dynamic high-mobility OTFS-RSMA IoT scenarios. Particularly, the SCA reference incurs a polynomial per-iteration cost in the environment parameters and multiplies that cost by I_{SCA} outer iterations per channel realization, yet admits no offline amortization phase, so its full iterative procedure must be re-executed within the channel coherence interval for every new channel, which makes practical application in a dynamic environment problematic. Inspired by this observation, the proposed DDPG-based approach confines all iterative cost to a one-off offline training phase and reduces online inference to a single feed-forward pass of the actor network with the linear scaling $O(UPN_t)$ in (34). Specifically, the resulting reduction is from polynomial to linear in the joint user-antenna dimension, and the gap widens rapidly with problem size: doubling UN_t multiplies the SCA per-iteration cost by roughly $2^{3.5}\approx 11.3\times$ while only doubling the proposed online cost. Hence, the proposed scheme is the only one of the two whose online inference stays within the per-frame budget imposed by the OTFS coherence interval, and Table III summarizes this comparison.

The linear scaling $O(UPN_t)$ in (34) has a direct implication for the algorithmic decision latency (ADL) of the proposed framework. Because the online inference cost grows linearly, rather than polynomially, in the three environment parameters (U, P, N_t), a fixed device count U yields an ADL that is bounded by a small multiple of a single actor forward pass, is independent of the training horizon, and does not require iterative per-decision optimization loops. This bound makes the framework deployable under the short coherence times typical of high-mobility IoT channels, whereas the iterative SCA reference of III-D2 accumulates decision latency polynomially in the problem size. An extended evaluation covering energy efficiency, transmission latency, and reliability under varying channel coherence times is identified as an important direction for our future work.

IV. Numerical Results

Consistent with the primary objective of this paper, which is to introduce a DRL framework for SE maximization in OTFS-RSMA rather than to analyze channel estimation or interference cancellation impairments, we adopt the widely used simplifying assumptions of perfect CSI at the BS and perfect SIC at the receivers [23], [24], [37].

A. Simulation Settings

This work evaluates the proposed DDPG-based SE maximization framework under an OTFS-RSMA downlink system, where $N_t = 4$ for the BS, $U = 10$ IoTDs are served, and the OTFS grid is configured with $N = M = 16$. This work assumes all IoTDs have $P = 4$ propagation paths. To ensure channel energy normalization, the path gains follow a uniform power delay profile modeled as complex Gaussian random variables, $h_{i,p}^{n_t} \sim \mathcal{CN}(0, 1/P)$, which implying a normalized expected power of $\sum_{p=0}^{P-1}\mathcal{E}\{|h_{i,p}^{n_t}|^2\} = 1$. Consistent with the parameters in Tables II and IV, the simulation applies a maximum IoTD speed of 126.6 km/h, corresponding to a peak Doppler shift of 468.8 Hz.

TABLE IV: Simulation Parameters

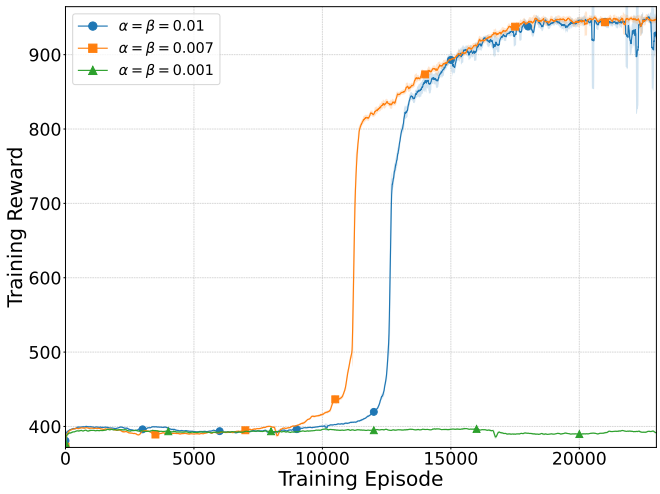
Parameter	Value
M	16 [22], [23]
N	16 [22], [23]
Users	6–10
No. of propagations	4 [22], [23]
Δf	7.5 kHz
f_c	4 GHz
σ^2	1
Hidden layers	512 nodes and 256 nodes
Learning rate	0.007
Buffer capacity	20 000
Training episodes	23 000
Steps per episode	300

B. Performance Evaluation

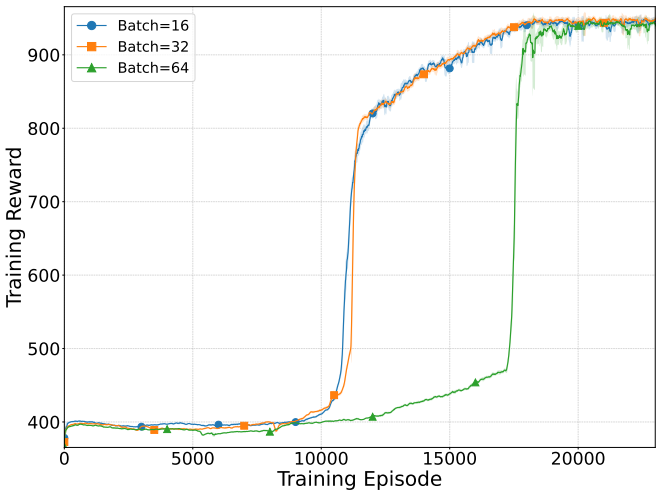
This section presents an evaluation of the proposed framework to analyze the convergence behavior and parameter sensitivity of the DRL agent, quantify the SE performance gains relative to the established baselines, and assess the framework scalability with respect to critical system parameters. For this comparative analysis, three baseline schemes were established.

- **OTFS-wo-RSMA [38]:** This scheme uses the same OTFS modulation and DRL-based optimization framework as the primary proposed method but omits the RSMA component. This crucial benchmark enables the isolation and precise measurement of the performance benefits that RSMA contributes, particularly in managing multiuser interference.
- **OTFS-RSMA:** This scheme retains the complete OTFS-RSMA system architecture but replaces the DDPG learning policy with a heuristic greedy algorithm. Instead of learning a continuous policy, the algorithm initializes with a random vector and iteratively improves it by optimizing one precoding parameter at a time over a fixed discrete grid to maximize the immediate reward. This comparison highlights the advantages of the proposed DRL method in handling complex, high-dimensional continuous spaces where standard iterative search techniques often become stuck in local optima.
- **OFDM-RSMA [39]:** This baseline adopts the traditional OFDM modulation scheme integrated with the complete RSMA component while maintaining the same DRL-based optimization framework as the primary proposed method. This benchmark isolates performance advantages attributed to the OTFS modulation waveform in high-mobility environments. By subjecting both modulation schemes to identical intelligent resource allocation and multiple access strategies, this baseline underscores the superior robustness of OTFS against Doppler-induced inter-carrier interference compared with the significant performance degradation observed in OFDM.

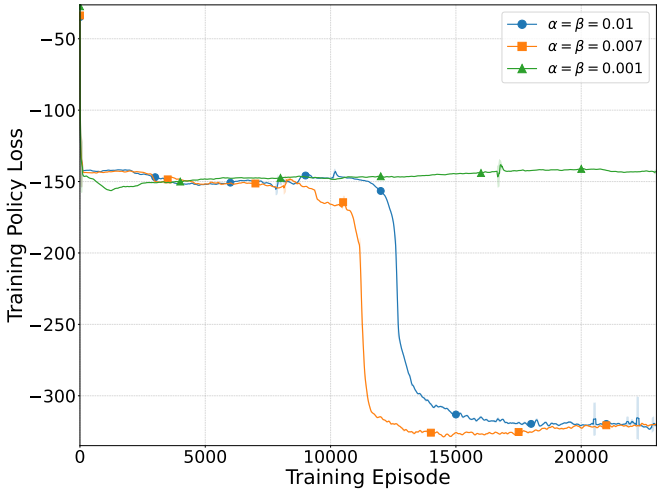
1) Convergence Analysis: Finding appropriate hyperparameters for the model is a hurdle in deep reinforcement learning. To this end, we conduct a sensitivity study by varying three critical hyperparameters of the proposed framework, the actor and critic learning rates, the minibatch size, and the OU exploration noise schedule to reveal the model's best-fit operating point. The chosen operating point is listed in Table IV, actor and critic learning rates $\alpha = \beta = 0.007$, soft-update coefficient $\tau = 0.1$, discount $\gamma = 0.99$, replay buffer of 2×10^4 transitions, minibatch size 32, and 2.3×10^4 training episodes of 300 steps each.



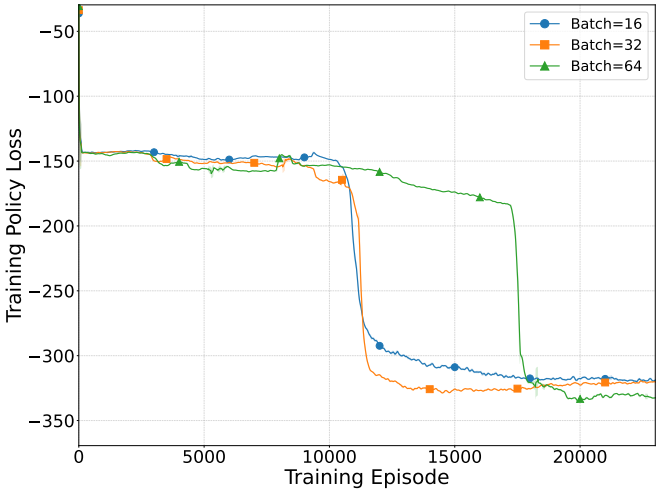
(a) Different learning rate training reward



(b) Different batch size training reward

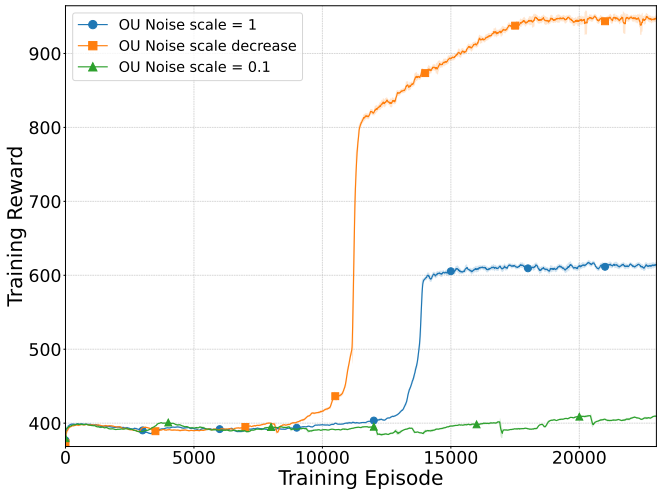


(c) Different learning rate policy loss

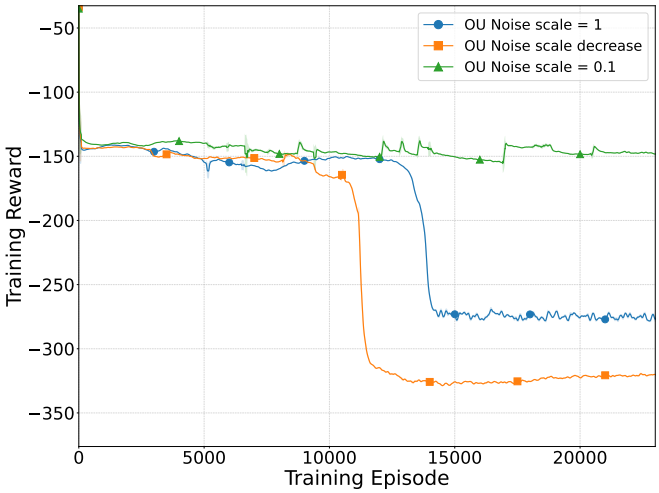


(d) Different batch size policy loss

Fig. 3: Convergence of the proposed scheme.



(a) Training reward



(b) Training policy loss

Fig. 4: OU exploration noise scale impact.

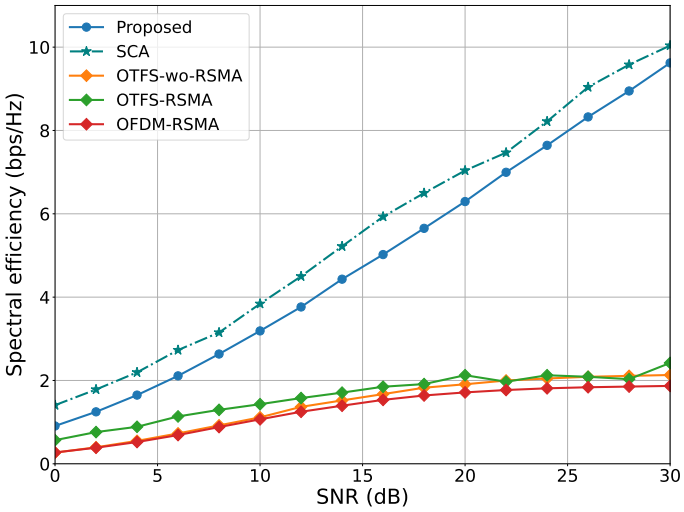


Fig. 5: Spectral efficiency versus SNR.

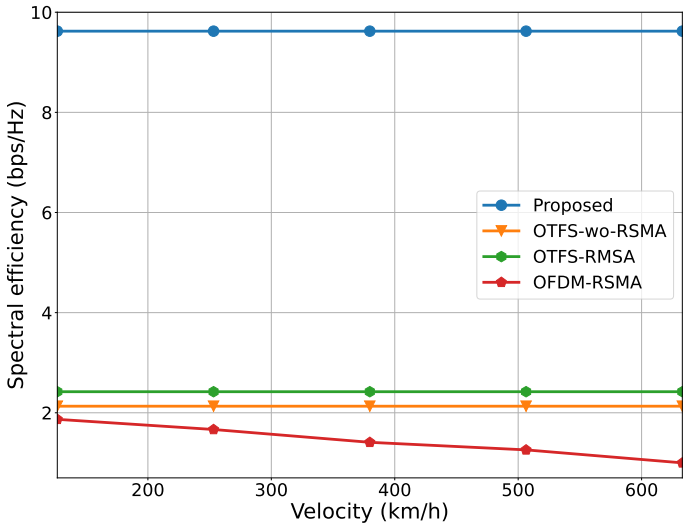


Fig. 6: Spectral efficiency vs. velocity under normalized path energy and perfect delay–Doppler CSI.

Firstly, we tune the learning rate of the model by selecting three cases of actor learning rate α and critic learning rate β across an order-of-magnitude range bracketing the chosen $\alpha = \beta = 0.007$, with the training reward and policy loss reported in Fig. 3a and Fig. 3c, respectively. The learning rate determines the size of each gradient step in the actor and critic updates and therefore directly shapes both the time at which the policy escapes the initial low-reward region and the stability of the policy after that escape. Because the too small learning rates result in too small update steps, which affects the model's learning speed and stability, the case $\alpha = \beta = 0.001$ fails to learn — the reward stagnates near 395 over the entire training horizon, and the policy loss remains flat near -145. Furthermore, with a high learning rate, the model may oscillate around but cannot move into the optimal, manifesting in the case $\alpha = \beta = 0.01$ as a delayed policy-loss drop until around episode 12 500 and visible reward dips during the late training horizon. Hence, the case $\alpha = \beta = 0.007$ gives the best performance, where the agent is the first to leave the low-reward region near episode 11 000, reaches the highest reward plateau of approximately 945 by episode 14 000, and drives the policy loss to its lowest value of approximately -325 along the smoothest trajectory of the three cases — about 139% better than the failed case $\alpha = \beta = 0.001$.

Next, to evaluate the efficiency of the minibatch size to the training, we train the model with three values as $B \in \{16, 32, 64\}$ and get the result in Fig. 3b and Fig. 3d. The minibatch size determines the number of replay-buffer samples used in each gradient update and therefore controls the variance of each policy-gradient estimate and, equivalently, the diversity of update directions seen per episode. In the case $B = 16$, the small sample set yields noisy policy-gradient estimates that produce visibly higher reward fluctuations during and after the transition and leave the policy loss near -315 at the end of training, even though the agent does transition out of the low-reward region near episode 11 000. In the case $B = 64$, each update averages over a larger sample set drawn from a replay buffer of fixed size 2×10^4, so the consecutive gradient estimates become highly correlated and over-smoothed; the resulting update direction is too consistent to explore the action space efficiently within the early episodes, and the model starts to converge after about 17 500 episodes, which is very slow compared to the others. Hence, the case $B = 32$ gives the best performance, where the reward and policy loss reach approximately 945 and -325, respectively, with the smoothest trajectory among the three cases. We note that $B = 64$ does eventually attain a slightly lower policy loss of approximately -335 near the end of training, but only after its substantially delayed transition.

At the chosen training setting, the policy is trained for $N_e N_s = 2.3\times10^4\times300 = 6.9\times10^6$ environment transitions, with one minibatch update applied at each transition as listed in Algorithm 1. Particularly, this large training effort happens only once, before the model is deployed. After training, generating a precoder for a new channel takes only a single forward pass through the actor, with linear complexity $O(U P N_t)$ from Section III-D. Hence, the one-time training effort is reused for every later deployment, and the per-decision online cost grows only linearly with the number of users, paths, and antennas.

We further examine how the OU exploration noise schedule affects the convergence of the model. Particularly, the OU process generates the exploration noise $\mathcal{N}[t]$ via the discrete-time recursion

$$\mathcal{N}[t+1] = \mathcal{N}[t] + \kappa\big(\mu_{\mathrm{OU}} - \mathcal{N}[t]\big) + \sigma_0\,\sigma\,\boldsymbol{\xi}[t], \quad \boldsymbol{\xi}[t] \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{35}$$

with mean-reversion rate κ, long-run mean μ_{OU} (set to zero), intrinsic Gaussian-driver scale σ, and outer noise-scale parameter σ_0 that rescales the per-step perturbation. Following the standard DDPG configuration of [28], we fix $\kappa = 0.15$ and $\sigma = 0.2$, and treat σ_0 as the schedule-tunable parameter swept in this study, so that the effective per-step kick standard deviation is $\sigma_0\,\sigma$. To evaluate this trade-off, we test three schedules over the 2.3×10^4-episode training horizon: a constant scale $\sigma_0 = 1.0$, a constant scale $\sigma_0 = 0.1$, and a linearly decaying schedule that starts at 1.0 and decreases to 0.1 over the first 18 000 episodes and is held at 0.1 thereafter, with the corresponding training reward and policy loss illustrated in Fig. 4a and Fig. 4b, respectively. With σ_0 held constant at 1.0, the persistent large perturbation dominates the actor output at every step, so the agent fails to learn the best reward which make the reward remains flat after achieving 615 across the entire training horizon while the policy loss never leaves its initial neighborhood near -275. With σ_0 held constant at 0.1, the agent fail to learn which make the reward fluctuate around 400 because the narrow perturbation neighborhood prevents the policy from discovering the high-reward region of the action space, and the policy loss also keep around -150. Hence, the decaying schedule gives the best performance, where the agent benefits from broad early exploration that surfaces high-reward policies and from progressively narrower late-phase perturbation that exploits the policy already learned: the reward rises gradually begin at the 7 000 episodes, climbs steadily through the decay phase, and reaches a high-reward plateau of approximately 950 by episode 17 500.

To assess training stability across random initializations, the proposed DDPG agent and the Twin Delayed Deep Deterministic Policy Gradient (TD3) [40] algorithm are trained with three random seeds selected from $\{1, \ldots, 5\}$ under the hyperparameters of Table IV with the training horizon extended to 2.5×10^4 episodes, where each seed fixes both network initialization and channel realizations, exposing both algorithms to identical channels. As shown in Fig. 7, all six runs initially

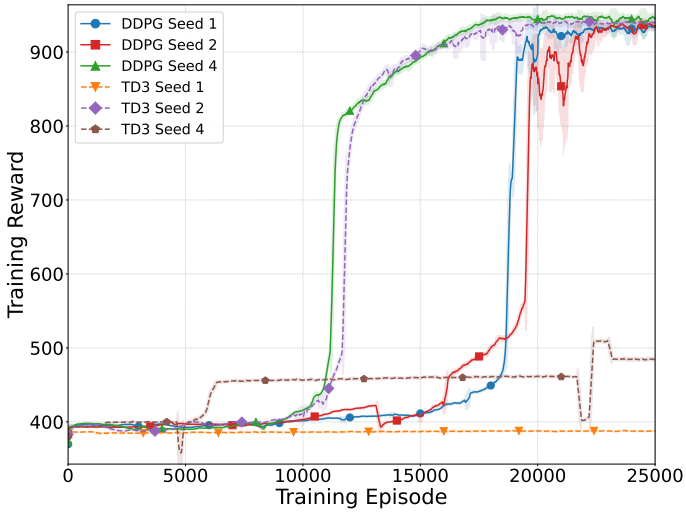


Fig. 7: Training rewards of DDPG and TD3 across random seeds.

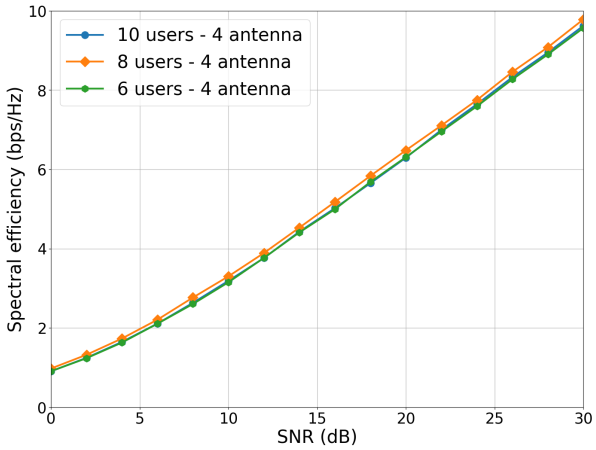


Fig. 8: Spectral efficiency vs. SNR by number of devices.

remain on a plateau at a training reward of approximately 400, on which actor output saturation renders the policy gradient vanishingly small. DDPG runs escape this plateau after approximately from range 11,000 to 18,000 episodes and converge to 931.3 and 945.9, whereas one TD3 run escapes after approximately 11,700 episodes and converges to 939.7. The escaped runs reach the same level within 1% irrespective of the algorithm, indicating that the clipped double-Q learning and delayed policy updates of TD3 do not improve the attainable precoding SE. Moreover, the outcome depends strongly on the seed, as the second seed escapes under both algorithms while the third seed fails under both, which is consistent with the sensitivity of DRL training to random initialization [41], [42]. Notably, under the hyperparameter setting of Table IV, the proposed DDPG scheme exhibits more stable training than TD3 across random initializations. In particular, DDPG attains a plateau escape rate of 3/3, whereas TD3 achieves only 1/3, demonstrating that DDPG is less susceptible to actor saturation in the considered environment.

2) Spectral Efficiency Performance: Fig. 5 comprehensively compares the SE performance of the proposed scheme with the that of the baselines across a practical SNR range of 0 to 30 dB. The superiority of the OTFS is evident in the massive performance gap between the proposed method and OFDM-RSMA baseline. The proposed method reaches a SE of about 9.62 bps/Hz at an SNR of 30 dB while the OFDM-based approach is around 1.87 bps/Hz, highlighting the inability of the conventional OFDM to cope with severe intercarrier interference in high-mobility environments and confirming the robustness of OTFS in the DD domain.

Moreover, the necessity of intelligent resource management is demonstrated by a comparison with the OTFS-RSMA baseline. Despite employing the same advanced architecture, the OTFS-RSMA without the proposed strategy stagnates at about 2.42 bps/Hz at 30 dB, proving that the structural advantages of OTFS-RSMA cannot be realized without the sophisticated DRL-based optimization policy to manage precoding and power allocation dynamically. Finally, the specific gain from the multiple access layer is isolated by the OTFS-wo-RSMA baseline. This non-RSMA variant attains only about 2.13 bps/Hz at 30 dB, making its lags behind the proposed scheme by about 7.49 bps/Hz, confirming that the flexible interference suppression provided by RSMA through its splitting of messages into common and private streams is critical for maximizing SE in multi-user scenarios. Overall, these results confirm that the superior performance of the proposed system is not attributable to any single factor but to integrating the robust OTFS, adaptive RSMA protocol, and DRL optimization framework.

Fig. 6 illustrates the divergence in performance, as the OFDM-RSMA scheme significantly degrades at higher velocities due to severe inter carrier interference. All OTFS-based curves remain flat under these conditions, maintaining flat SE from 120 to over 600 km/h. This stability affirms the robustness of OTFS in overcoming high-Doppler spreads, a benefit obtained by modulating signals in the delay-Doppler (DD) domain. Such a transformation converts a rapidly time-varying fading channel into a sparse, quasi-time-invariant representation. In the simulation framework, this resilience is further consolidated using FD-LE with perfect channel state information . Because the channel gains are normalized, an increase in velocity and the resulting Doppler shift alters the tap location indices on the DD grid without diminishing the total received signal energy. With accurate channel state information, the FD-LE receiver precisely tracks these location shifts and compensates for Doppler-induced phase rotations, rendering the achievable SE invariant to speed. This invariance is therefore an exact consequence of the idealized conditions adopted in this study rather than a general property of the waveform. Accordingly, Fig. 6 is to be read as a waveform-versus-waveform comparison that isolates the structural robustness of the delay–Doppler representation by factoring out channel estimation. In a practical deployment, fractional Doppler leakage and channel estimation error or aging would introduce a moderate degradation at extreme velocities, so the reported flatness should be interpreted as an upper bound.

3) Framework Scalability: Fig. 8 evaluates the scalability of the proposed framework as the number of devices varies over $U \in \{6, 8, 10\}$ while the antenna budget is fixed at $N_t = 4$. The three SE-vs-SNR curves are tightly clustered across the entire SNR sweep, indicating that the framework remains robust even as the system enters the overloaded regime where the user count exceeds the number of antennas. Particularly, the eight-device configuration consistently lies slightly above the six- and ten-device cases, reaching approximately 9.78 bps/Hz at 30 dB, which is roughly 2.3% and 1.7% better than the ten- and six-device cases respectively. With a small user count $U = 6$, the framework cannot fully exploit multiuser diversity and the common stream is multiplexed over a smaller user pool. With a large user count $U = 10$, the multicast bottleneck R_{com} is dragged down by the worst user rate, and the additional private streams aggravate the residual inter-user interference that the four-antenna array cannot null. Hence, the eight-device case emerges as the sweet spot where the multiuser-diversity gain is maximized before the multicast bottleneck takes over. Inspired by this observation, we note that the high-SNR slope of all three curves approaches $\log_2(10) \approx 3.32$ bps/Hz per 10 dB increment, signifying that the DRL agent has converged to a common-stream-dominant strategy that secures the one-DoF multicast floor regardless of the precise user count. Therefore, the proposed framework is structurally insensitive to the number of devices in the overloaded regime, a property that makes it well suited to dynamic IoT

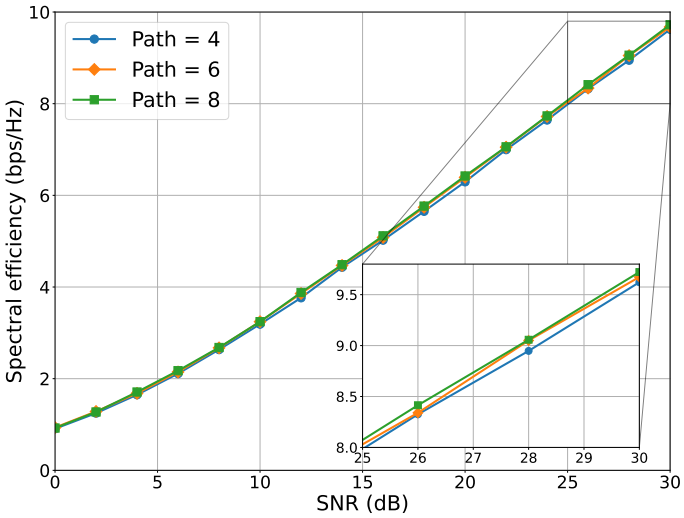


Fig. 9: Spectral efficiency vs. SNR by number of paths.

deployments in which the device count is not known a priori and may fluctuate during operation.

Fig. 9 reports the spectral efficiency of the proposed OTFS-RSMA-DDPG scheme under three propagation-path settings $P \in \{4, 6, 8\}$ across an SNR range from 0 to 30 dB. Specifically, at 0 dB the three settings differ by only 0.026 bps/Hz because the noise floor dominates and the diversity gain from extra paths cannot yet be exploited, whereas at the high-SNR end (30 dB) the curves separate into 9.72, 9.67, and 9.62 bps/Hz for $P = 8$, $P = 6$, and $P = 4$, which is approximately 1.0% and 0.5% better than $P = 4$, respectively. In the 25–30 dB range, the ordering $P = 8 > P = 6 > P = 4$ becomes consistent and reflects the additional DD diversity carried by richer multipath: once the SNR is high enough for the actor to discriminate among DD bins, more resolvable paths broaden the channel support in the DD grid, and the precoders align with stronger channel directions. At lower SNRs the three settings cross occasionally because the residual exploration noise and finite-rate convergence of the trained policy mask the small diversity gap.

Table V compares the spectral efficiency and online execution time of the proposed DDPG scheme against the benchmark SCA algorithm outlined in Section III-D2 under various combinations of U and P. For a fair comparison, both schemes are evaluated over the same 200 channel realizations at SNR = 10 dB on a single CPU thread. In particular, because the SCA iteratively refines both the common and private precoders by solving convex subproblems, it achieves a higher SE than the learning-based approach. Nevertheless, the proposed DDPG scheme attains approximately 76%, 84%, 83%, and 86% of the SCA benchmark across the four configurations, respectively. However, the optimization complexity of each convex subproblem increases significantly with the number of devices, causing the average SCA runtime to grow by roughly 4.5× as U scales from 6 to 10. In contrast, the online inference of the proposed DDPG scheme requires only a single forward pass over the actor network with the linear complexity formulated in (34). Consequently, its execution latency remains nearly constant at 0.062–0.069 ms, operating 3.8×10^4 to 1.6×10^5 times faster than the SCA. To put this into practical perspective, considering the OTFS frame duration $NT = N/\Delta f \approx 2.133$ ms, the SCA execution delay exceeds the frame duration by more than three orders of magnitude, rendering its calculated precoders severely outdated before deployment in fast-fading channels. Hence, the proposed DDPG scheme effectively trades a moderate SE gap for an ultra-low decision latency approximately 30× shorter than a single OTFS frame, ensuring robust real-time feasibility in dynamic high-mobility environments.

TABLE V: Spectral Efficiency and Computational Time Comparison at SNR = 10 dB

U	P	SCA (bps/Hz)	DDPG (bps/Hz)	SCA Runtime (s)	DDPG Runtime (ms)
6	4	4.085	3.121	2.34	0.062
8	4	4.010	3.363	4.55	0.063
10	4	4.024	3.326	10.42	0.065
10	6	3.812	3.294	8.49	0.069

V. Conclusion

This paper presents a DRL-based framework for SE maximization in OTFS-RSMA downlink systems under high-mobility conditions. By formulating the precoding design as a continuous optimization problem, the proposed method addresses interference management, power constraints, and dynamic channel conditions in the DD domain. The simulation results demonstrated that the proposed scheme outperforms the OTFS-wo-RSMA and OTFS-RSMA baselines in SE, with performance gains increasing at higher SNR levels. Furthermore, scalability tests with varying grid sizes and antenna counts confirmed that the framework remains robust and benefits from array and resolution gains. These results indicate the potential of integrating DRL with OTFS-RSMA to enable high-throughput, reliable communication in next-generation high-mobility wireless networks.

Several extensions of the proposed framework merit further research. Particularly, the perfect CSI and perfect SIC assumptions adopted here isolate the SE-maximization role of the DRL agent, whereas practical IoT deployments inevitably contend with channel estimation errors and residual SIC interference; explicitly modeling these impairments and re-tuning the actor under imperfect CSI and SIC is the most direct extension. Furthermore, a dedicated velocity study that removes these idealizations by accounting for fractional Doppler leakage, which arises when Doppler shifts do not fall on the delay–Doppler grid and produces inter-Doppler interference that grows with velocity, as well as CSI aging at high mobility, remains an important topic for future work. Specifically, validating the trained policy on real-world datasets, beyond the calibrated delay–Doppler benchmark used in this paper, also deserves to be investigated to expose the framework to channel statistics that synthetic benchmarks do not reproduce. Besides, energy efficiency and transmission latency are also important performance indicators for practical IoT systems, and evaluating the proposed framework along these axes would give a fuller picture of how the scheme behaves once deployed, beyond the SE results reported in this paper. Hence, more heterogeneous IoT scenarios, where devices simultaneously differ in path counts, mobility profiles, and quality-of-service requirements, open an intriguing research opportunity to stress-test the generalization of the DRL agent.

Appendix
Proof of Proposition 1

The following proof reveals how the diagonalized channel matrix in (8) is related to the original block-circulant matrix in (7). According to the lemma in [43], any block circulant matrix $\mathbf{Q} = \text{circ}[\mathbf{Q}_0, \mathbf{Q}_1, \ldots, \mathbf{Q}_{N-1}] \in \mathbb{C}^{NM \times NM}$ with $\mathbf{Q}_n \in \mathbb{C}^{M \times M}$ for $0 \le n \le N-1$ can be block-diagonalized in the following forms:

$$\begin{aligned} \mathbf{Q} &= (\mathbf{F}_N^H \otimes \mathbf{I}_M)\text{diag}(\Theta_i)(\mathbf{F}_N \otimes \mathbf{I}_M) \\ &= (\mathbf{F}_N \otimes \mathbf{I}_M)\text{diag}(\Theta_i')(\mathbf{F}_N^H \otimes \mathbf{I}_M), \end{aligned} \tag{36}$$

with $0 \le i \le N-1$, $\Theta = \sum_{n=0}^{N-1} \mathbf{Q}_n e^{\frac{-j2\pi ni}{N}}$, and $\Theta' = \sum_{n=0}^{N-1} = \mathbf{Q}_n e^{\frac{j2\pi ni}{N}}$.

Based on this, considering a block-circulant matrix as presented in (37)

$$\mathbf{H}_i^{n_t} = \begin{bmatrix} \mathbf{A}_{i,0}^{n_t} & \mathbf{A}_{i,N-1}^{n_t} & \cdots & \mathbf{A}_{i,2}^{n_t} & \mathbf{A}_{i,1}^{n_t} \\ \mathbf{A}_{i,1}^{n_t} & \mathbf{A}_{i,0}^{n_t} & \ddots & \mathbf{A}_{i,3}^{n_t} & \mathbf{A}_{i,2}^{n_t} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{A}_{i,N-2}^{n_t} & \mathbf{A}_{i,N-3}^{n_t} & \ddots & \mathbf{A}_{i,0}^{n_t} & \mathbf{A}_{i,N-1}^{n_t} \\ \mathbf{A}_{i,N-1}^{n_t} & \mathbf{A}_{i,N-2}^{n_t} & \ddots & \mathbf{A}_{i,1}^{n_t} & \mathbf{A}_{i,0}^{n_t} \end{bmatrix} \tag{37}$$

we apply $(\mathbf{F}_N \otimes \mathbf{I}_M)$ to (7), yielding the following:

$$\begin{aligned}
&(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{y}_i \\
&= \sum_{n_t \in \mathcal{A}} (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{H}_i^{n_t} w_{\text{com}}^{n_t} \mathbf{x}_{\text{com}} + \sum_{q \in \mathcal{U}} \sum_{n_t \in \mathcal{A}} (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{H}_i^{n_t} w_q^{n_t} \mathbf{x}_{q,\text{pri}} \\
&+ \mathbf{z} \\
&= \sum_{n_t \in \mathcal{A}} (\mathbf{F}_N \otimes \mathbf{I}_M)(\mathbf{F}_N^H \otimes \mathbf{I}_M)\text{diag}(\mathbf{B}_l^{n_t}, 0 \le l \le N-1)(\mathbf{F}_N \otimes \mathbf{I}_M) w_{\text{com}}^{n_t} \mathbf{x}_{\text{com}} \\
&+ \sum_{q \in \mathcal{U}} \sum_{n_t \in \mathcal{A}} (\mathbf{F}_N \otimes \mathbf{I}_M)(\mathbf{F}_N^H \otimes \mathbf{I}_M)\text{diag}(\mathbf{B}_l^{n_t}, 0 \le l \le N-1)(\mathbf{F}_N \otimes \mathbf{I}_M) w_q^{n_t} \\
&\times \mathbf{x}_{q,\text{pri}} + (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z} \\
&= \sum_{n_t \in \mathcal{A}} \text{diag}(\mathbf{B}_l^{n_t}, 0 \le l \le N-1) w_{\text{com}}^{n_t} (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{x}_{\text{com}} + \sum_{q \in \mathcal{U}} \sum_{n_t \in \mathcal{A}} \text{diag}(\mathbf{B}_l^{n_t} \\
&, 0 \le l \le N-1) w_q^{n_t} (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{x}_{q,\text{pri}} + (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z},
\end{aligned} \tag{38}$$

where $\text{diag}(\mathbf{B}_l^{n_t}, 0 \le l \le N-1)$ denotes a block-diagonal matrix with $\mathbf{B}_l^{n_t} = \sum_{n=0}^{N-1} \mathbf{A}_{i,n}^{n_t} e^{\frac{-j2\pi ln}{N}}, 0 \le l \le N-1$, and $\mathbf{A}_{a,n}^{n_t}$ denotes the first-block column size $M \times M$ of $\mathbf{H}_i^{n_t}$. Because each $\mathbf{A}_{i,n}^{n_t}$ is circulant, the matrix can be further diagonalized by $\mathbf{F}_M$

$$\mathbf{A}_{i,n}^{n_t} = \mathbf{F}_M \Lambda_{i,n}^{n_t} \mathbf{F}_M^H \tag{39}$$

where $\Lambda_{i,n}^{n_t} = \text{diag}(\sum_{m=0}^{M-1} o_{i,n}^{n_t,m,1} e^{\frac{j2\pi rm}{M}}, 0 \le r \le M-1)$ with $o_{a,n}^{n_t,m,1}$ represents the element in the mth row and first column of $\mathbf{A}_{i,n}^{n_t}$. Thus, we apply $(\mathbf{I}_N \otimes \mathbf{F}_M^H)$ to (38), yielding the following :

$$\begin{aligned}
&(\mathbf{I}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{y}_i \\
&= \sum_{n_t \in \mathcal{A}} \text{diag}(\hat{\mathbf{B}}_l^{n_t}, 0 \le l \le N-1) w_{\text{com}}^{n_t} (\mathbf{F}_N \otimes \mathbf{I}_M)(\mathbf{I}_N \otimes \mathbf{F}_M^H) \\
&\times \mathbf{x}_{\text{com}} + \sum_{q \in \mathcal{U}} \sum_{n_t \in \mathcal{A}} \text{diag}(\hat{\mathbf{B}}_l^{n_t}, 0 \le l \le N-1) w_q^{n_t} (\mathbf{F}_N \otimes \mathbf{I}_M) \\
&\times (\mathbf{I}_N \otimes \mathbf{F}_M^H)\mathbf{x}_{q,\text{pri}} + (\mathbf{I}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z} \\
&= \sum_{n_t \in \mathcal{A}} \text{diag}(\hat{\mathbf{B}}_l^{n_t}, 0 \le l \le N-1) w_{\text{com}}^{n_t} (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{x}_{\text{com}} \\
&+ \sum_{q \in \mathcal{U}} \sum_{n_t \in \mathcal{A}} \text{diag}(\hat{\mathbf{B}}_l^{n_t}, 0 \le l \le N-1) w_q^{n_t} (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{x}_{q,\text{pri}} \\
&+ (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z},
\end{aligned} \tag{40}$$

where $(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z} \sim CN(0, \mathbf{I}_{NM}\sigma^2)$ and $\hat{\mathbf{B}}_l^{n_t} = \sum_{n=0}^{N-1} \Lambda_{i,n}^{n_t} e^{\frac{-j2\pi ln}{N}}, 0 \le l \le N-1$. In addition, $(\mathbf{I}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{I}_M) = (\mathbf{F}_N \otimes \mathbf{I}_M)(\mathbf{I}_N \otimes \mathbf{F}_M^H) = (\mathbf{F}_N \otimes \mathbf{F}_M^H)$ because of the result of the Kronecker product property, where $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$.

By defining $\mathbf{D}_i^{n_t} \triangleq \text{diag}(\hat{\mathbf{B}}_l^{n_t}, 0 \le l \le N-1)$, the following per-antenna diagonal matrix becomes

$$\begin{aligned}
&(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{y}_i \\
&= \sum_{n_t \in \mathcal{A}} \mathbf{D}_i^{n_t} w_{\text{com}}^{n_t} (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{x}_{\text{com}} + \sum_{q \in \mathcal{U}} \sum_{n_t \in \mathcal{A}} \mathbf{D}_i^{n_t} w_q^{n_t} (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{x}_{q,\text{pri}} \\
&+ (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}.
\end{aligned} \tag{41}$$

The common and private precoding contributions are grouped in the diagonal matrix $\mathbf{D}_i^{n_t}$ as $\mathbf{D}_i^{\text{com}} = \sum_{n_t \in \mathcal{A}} \mathbf{D}_i^{n_t} w_{com}^{n_t}$ and $\mathbf{D}_{i,q}^{\text{eff}} = \sum_{n_t \in \mathcal{A}} \mathbf{D}_i^{n_t} w_q^{n_t}$ to obtain the received signal:

$$\begin{aligned}
&(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{y}_i \\
&= \mathbf{D}_i^{\text{com}} (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{x}_{\text{com}} + \sum_{q \in \mathcal{U}} \mathbf{D}_{i,q}^{eff} (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{x}_{q,\text{pri}} + (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}.
\end{aligned} \tag{42}$$

The next step is to apply $(\mathbf{F}_N \otimes \mathbf{F}_M^H)^{-1}(\mathbf{D}_i^{\text{com}})^{-1}$ to (42) to detect $\mathbf{x}_{\text{com}}$; thus, the complete signal is given by

$$\overline{Y}_i = \mathbf{x}_{\text{com}} + \sum_{q \in \mathcal{U}} (\mathbf{F}_N \otimes \mathbf{F}_M^H)^{-1} \mathbf{D}_0^q (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{x}_{q,\text{pri}} + \tilde{\mathbf{z}}, \tag{43}$$

where $\mathbf{D}^q = (\mathbf{D}_i^{\text{com}})^{-1}\mathbf{D}_{i,q}^{\text{eff}}$, and $\tilde{\mathbf{z}} = (\mathbf{F}_N \otimes \mathbf{F}_M^H)^{-1}(\mathbf{D}_i^{\text{com}})^{-1}(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}$, which completes the proof.

References

[1] L. Xiao, S. Li, Y. Qian, D. Chen, and T. Jiang, "An overview of otfs for internet of things: Concepts, benefits, and challenges," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7596–7618, 2022.

[2] P. Yang, Y. Xiao, M. Xiao, and S. Li, "6g wireless communications: Vision and potential techniques," *IEEE Network*, vol. 33, no. 4, pp. 70–75, 2019.

[3] F.-M. Han and X.-D. Zhang, "Wireless multicarrier digital transmission via weyl-heisenberg frames over time-frequency dispersive channels," *IEEE Transactions on Communications*, vol. 57, no. 6, pp. 1721–1733, 2009.

[4] R. Hadani, S. Rakib, M. Tsatsanis, A. Monk, A. J. Goldsmith, A. F. Molisch, and R. Calderbank, "Orthogonal time frequency space modulation," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6.

[5] P. Raviteja, K. T. Phan, Y. Hong, and E. Viterbo, "Interference cancellation and iterative detection for orthogonal time frequency space modulation," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6501–6515, 2018.

[6] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2073–2126, 2022.

[7] O. Dizdar, Y. Mao, and B. Clerckx, "Rate-splitting multiple access to mitigate the curse of mobility in (massive) mimo networks," *IEEE Transactions on Communications*, vol. 69, no. 10, pp. 6765–6780, 2021.

[8] O. Dizdar, Y. Mao, Y. Xu, P. Zhu, and B. Clerckx, "Rate-splitting multiple access for enhanced urllc and embb in 6g: Invited paper," in *2021 17th International Symposium on Wireless Communication Systems (ISWCS)*, 2021, pp. 1–6.

[9] S. Chaudhary, I. Budhiraja, R. Chaudhary, N. Kumar, D. Garg, and A. M. Almuhaideb, "Quantum federated reinforcement-learning-based joint mode selection and resource allocation for star-ris-aided vrcs," *IEEE Internet of Things Journal*, vol. 11, no. 22, pp. 36 242–36 256, 2024.

[10] S. Chaudhary, I. Budhiraja, and R. Chaudhary, "Energy and latency tradeoff for star-irs-assisted vehicle road cooperative system in carbon intelligent iiot leveraging quantized federated reinforcement learning," *IEEE Internet of Things Journal*, vol. 12, no. 23, pp. 50 759–50 770, 2025.

[11] P. Raviteja, K. T. Phan, and Y. Hong, "Embedded pilot-aided channel estimation for otfs in delay–doppler channels," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4906–4917, 2019.

[12] Y. Yue, J. Shi, Z. Li, J. Hu, and Z. Tie, "Model-driven deep learning assisted detector for otfs with channel estimation error," *IEEE Communications Letters*, vol. 28, no. 4, pp. 842–846, 2024.

[13] Z. Yang, M. Chen, W. Saad, and M. Shikh-Bahaei, "Optimization of rate allocation and power control for rate splitting multiple access (rsma)," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5988–6002, 2021.

[14] W. Jaafar, S. Naser, S. Muhaidat, P. C. Sofotasios, and H. Yanikomeroglu, "On the downlink performance of rsma-based uav communications," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 258–16 263, 2020.

[15] H. Cui, L. Zhu, Z. Xiao, B. Clerckx, and R. Zhang, "Energy-efficient rsma for multigroup multicast and multibeam satellite communications," *IEEE Wireless Communications Letters*, vol. 12, no. 5, pp. 838–842, 2023.

[16] S. Naser, P. C. Sofotasios, L. Bariah, W. Jaafar, S. Muhaidat, M. Al-Qutayri, and O. A. Dobre, "Rate-splitting multiple access: Unifying noma and sdma in miso vlc channels," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 393–413, 2020.

[17] Q. Huai, W. Yuan, Y. Wu, and P. Fan, "Cross-domain transmission scheme for uplink otfs-rsma based on heterogeneous mobile user grouping," *IEEE Communications Letters*, vol. 28, no. 5, pp. 1181–1185, 2024.

[18] J. Liu, Y. Liang Guan, Y. Ge, and B. Clerckx, "Power allocation for high mobility otfs-rsma system with path-selective one-tap receiver," *IEEE Wireless Communications Letters*, vol. 14, no. 3, pp. 661–665, 2025.

[19] W. Jaafar, S. Naser, S. Muhaidat, P. C. Sofotasios, and H. Yanikomeroglu, "Multiple access in aerial networks: From orthogonal and non-orthogonal to rate-splitting," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 372–392, 2020.

[20] Z. Lin, M. Lin, T. de Cola, J.-B. Wang, W.-P. Zhu, and J. Cheng, "Supporting iot with rate-splitting multiple access in satellite and aerial-integrated networks," *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11 123–11 134, 2021.

[21] 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 Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1452–1467, 2023.

[22] S. K. Mohammed, R. Hadani, and A. Chockalingam, *OTFS Modulation: Theory and Applications*. John Wiley & Sons, 2024.

[23] Z. Ding, R. Schober, P. Fan, and H. Vincent Poor, "Otfs-noma: An efficient approach for exploiting heterogenous user mobility profiles," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7950–7965, 2019.

[24] Q. Huai, W. Yuan, Y. Wu, and P. Fan, "Downlink otfs-rsma cross-domain transmission scheme and sum-rate maximization," *IEEE Communications Letters*, vol. 29, no. 3, pp. 600–604, 2025.

[25] V. Kumar, R. Zhang, M. D. Renzo, and L.-N. Tran, "A novel sca-based method for beamforming optimization in irs/ris-assisted mu-miso downlink," *IEEE Wireless Communications Letters*, vol. 12, no. 2, pp. 297–301, 2023.

[26] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface assisted uav communication: Joint trajectory design and passive beamforming," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 716–720, 2020.

[27] S. Singh, K. Deka, S. Sharma, Y.-P. Lin, and N. Rajamohan, "Beamforming optimization and admm-based detection in irs-aided otfs systems," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 1948–1964, 2025.

[28] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[29] W. J. Yun, S. Park, J. Kim, M. Shin, S. Jung, D. A. Mohaisen, and J.-H. Kim, "Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-uav control," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7086–7096, 2022.

[30] W. J. Yun, D. Kwon, M. Choi, J. Kim, G. Caire, and A. F. Molisch, "Quality-aware deep reinforcement learning for streaming in infrastructure-assisted connected vehicles," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2002–2017, 2022.

[31] D. Kwon, J. Jeon, S. Park, J. Kim, and S. Cho, "Multiagent ddpg-based deep learning for smart ocean federated learning iot networks," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9895–9903, 2020.

[32] Y. He, B. Sheng, H. Yin, D. Yan, and Y. Zhang, "Multi-objective deep reinforcement learning based time-frequency resource allocation for multi-beam satellite communications," *China Communications*, vol. 19, no. 1, pp. 77–91, 2022.

[33] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Machine Learning (ICML)*, 2014, pp. 387–395.

[34] G. E. Uhlenbeck and L. Ornstein, *On the theory of the Brownian motion*. Physical Review, 1930.

[35] T. Phung Truong, T. My Tuyen Nguyen, T. Vi Nguyen, N.-N. Dao, and S. Cho, "Energy efficiency in rsma-enhanced active ris-aided quantized downlink systems," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 3, pp. 834–850, 2025.

[36] T. P. Truong, V. D. Tuong, N.-N. Dao, and S. Cho, "Flyreflect: Joint flying irs trajectory and phase shift design using deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4605–4620, 2023.

[37] J. Huang, W.-B. Sun, Z. Zhang, L. Wang, D. Turlykozhayeva, Z. Yin, and W.-X. Meng, "Joint beamforming and power allocation for ris-assisted otfs-noma systems," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2026.

[38] G. Zhou, Y. Mao, and B. Clerckx, "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff," *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 4816–4828, 2022.

[39] A. K. Kowshik, S. Gurugopinath, and S. Muhaidat, "Deep learning-based detection for rsma with orthogonal time frequency space modulation," *IEEE Communications Letters*, vol. 29, no. 4, pp. 669–673, 2025.

[40] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1587–1596.

[41] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 3207–3214.

[42] C. Colas, O. Sigaud, and P.-Y. Oudeyer, "How many random seeds? statistical power analysis in deep reinforcement learning experiments," 2018. [Online]. Available: https://arxiv.org/abs/1806.08295

[43] P. Raviteja, Y. Hong, E. Viterbo, and E. Biglieri, "Practical pulse-shaping waveforms for reduced-cyclic-prefix otfs," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, pp. 957–961, 2019.