Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079214

ARTICLE

A3TD: A Deep Reinforcement Learning Algorithm for Joint Resource Allocation in RIS-Aided CNOMA-D2D Networks

Zongchuan Li, Chen Sun* and Jian Shu

Software School, Nanchang Hangkong University, Nanchang, China

*Corresponding Author: Chen Sun. Email: sunchen@nchu.edu.cn

Received: 16 January 2026; Accepted: 13 March 2026; Published: 08 May 2026

ABSTRACT: This paper investigates the joint resource allocation problem in Reconfigurable Intelligent Surface (RIS)-assisted cooperative non-orthogonal multiple access device-to-device (CNOMA-D2D) cellular networks. To tackle the high-dimensional non-convex joint optimization of power control, RIS phase configuration and channel assignment, we propose an integrated user pairing strategy, PIP-UP, quantifying utility through factors, phase alignment, interference suppression and power difference, neglected in existing methods. Furthermore, we develop a hybrid deep reinforcement learning algorithm, A3TD, combining the parallel exploration capability of Asynchronous Advantage Actor-Critic (A3C) with the stable continuous optimization of Twin Delayed Deep Deterministic Policy Gradient (TD3). This integration enables efficient and robust joint optimization of D2D channel allocation, transmit power, and RIS phase shifts. Simulation results demonstrate that the proposed A3TD algorithm significantly outperforms baseline algorithms, Actor-Critic (AC), Deep Deterministic Policy Gradient (DDPG) and TD3, in terms of sum rate and convergence speed, validating its effectiveness for resource management in complex RIS-assisted CNOMA-D2D networks.

KEYWORDS: Reconfigurable intelligent surface; cooperative non-orthogonal multiple access; device-to-device communication; resource allocation; deep reinforcement learning; user pairing

1 Introduction

1.1 Background

The relentless densification of 5G and beyond networks necessitates efficient resource allocation techniques to meet soaring capacity demands. Device-to-device (D2D) communication, leveraging under-utilized macro-cell resources, presents a promising solution to enhance spectral efficiency and energy sustainability. However, this paradigm introduces severe intra-cell interference due to co-channel D2D transmissions, fundamentally undermining resource allocation efficacy in dense deployments [1]. While cooperative non-orthogonal multiple access (CNOMA) mitigates this challenge via superposition coding and successive interference cancellation (SIC), its performance hinges critically on user pairing strategies that simultaneously manage channel heterogeneity and interference dynamics [2]. Reconfigurable intelligent surfaces (RIS) introduces a transformative degree of freedom for wireless environment control [3]. By intelligently adjusting signal reflection phases, RIS can enhance desired signals, suppress interference, and reshape channel correlations. Its integration with CNOMA-D2D networks unlocks new potentials but also compounds the resource allocation challenge.

Predominant research on user pairing in NOMA/CNOMA systems relies heavily on heuristic methods centered on channel gains. For instance, pairing schemes often maximize inter-user channel gain differences [4] or apply combinatorial optimization like the Hungarian algorithm for spectral efficiency [5,6]. Strategies assuming fixed user roles [7] or those focusing solely on channel strength ratios [8,9] further lack adaptability in dynamic environments. Even recent advancements employing game theory [10] or multi-agent reinforcement learning (RL) [11,12] for pairing and resource allocation often incur high computational complexity and fail to jointly optimize with RIS. These approaches largely neglect the phase characteristics of user channels with RIS, which are crucial for constructive/destructive signal superposition in SIC, and the dynamic power constraints necessary to maintain the decoding order in the power domain of D2D pairs.

Existing studies in RIS-aided CNOMA systems often treat user pairing and RIS beamforming separately. Some optimize pairing based on conventional channel state information [13,14], while others focus on joint power and phase optimization without deeply coupling with user pairing dynamics [15,16]. Consequently, the synergistic impact of RIS-induced phase alignment on SIC decoding reliability within paired users remains inadequately explored. Moreover, existing optimization methods, ranging from block coordinate descent [17,18], meta-RL [19] to deep RL [20,21] and hybrid deep RL [22], struggle with the coupled, high-dimensional action space, often suffering from slow convergence, training instability, or limited generalization.

1.2 Motivations

Existing user pairing strategies in NOMA/CNOMA systems predominantly rely on channel gain disparities [4,5,11], often neglecting the phase characteristics introduced by RIS-a critical factor for SIC performance in reflected environments. Similarly, prior DRL-based resource allocation methods either suffer from high variance (e.g., A3C in continuous domains) or slow convergence (e.g., TD3 in high-dimensional spaces) when applied in isolation [19–22]. This gap motivates an integrated design philosophy: pairing must account for phase coherence and power dynamics, while optimization must balance exploration efficiency with update stability.

To include critical resource management factors-phase alignment and power control in RIS-aided CNOMA-D2D networks, We propose a Phase matching, Interference suppression, and Power control-based User Pairing (PIP-UP) mechanism. Inspired by the complementary strengths of A3C in parallelized exploration [23] and TD3 in deterministic policy refinement [24], we propose A3TD—a hybrid framework that concurrently addresses the combinatorial and continuous aspects of RIS-aided resource allocation. This approach is not only tailored to CNOMA-D2D networks but also embodies a generalizable methodology for joint optimization in other RIS-enhanced or multi-agent wireless systems where phase alignment and interference coordination are pivotal. Comparison of existing works and the proposed method on resource allocation in RIS-aided NOMA-D2D networks is shown in Table 1.

Table 1: Comparison of existing works on resource allocation in RIS-aided NOMA-D2D networks.

Work	RIS-Aided	User Pairing Metric	Joint Optimization	Solution Method	Limitation
[4]	×	Channel gain difference	power + channel	DE algorithm	No RIS; phase ignored
[5]	×	Channel gain difference	power + channel	DDPG + POPS	No RIS; phase ignored

(Continued)

Table 1 (continued)

Work	RIS-Aided	User Pairing Metric	Joint Optimization	Solution Method	Limitation
[7]	✓	Fixed strong/weak	power + phase	Alternating optimization	Pairing decoupled from phase; no channel assignment
[8]	✓	–	Pairing (power + phase)	Alternating optimization	No NOMA; heuristic pairing; no channel assignment
[10]	×	Strong/weak user grouping	Pairing (power + channel)	Matching theory + heuristic	No RIS; phase ignored
[11]	×	Channel gain difference	Pairing (power + channel)	D3PG	No RIS; phase ignored
[13]	✓	–	power + phase	Alternating optimization	No NOMA; no user pairing
[15]	✓	–	power + phase	DDQN	No NOMA; no D2D
[17]	✓	Channel quality ranking	power + phase + beamforming	BCD + SCA + SDR	No D2D; dynamic noise of active RIS
[19]	✓	–	power + phase + beamforming	MRL	No NOMA; no D2D; interference trade-off
[20]	×	Channel state	Pairing (power)	Prioritized Dueling DQN-DDPG	No RIS; no D2D; no CNOMA; static users
[22]	✓	–	power + phase + channel + position	D3QN + DDQN	No NOMA; no user pairing; discrete RIS position
This work	✓	PAD + ID + PDF	Pairing (power + phase + channel)	A3TD (A3C + TD3)	Fills all gaps

1.3 Contributions

To bridge these gaps, this paper proposes a holistic framework for joint resource optimization in RIS-aided CNOMA-D2D networks. Our work makes the following key contributions:

- **PIP-UP Strategy** quantifies pairing suitability through three metrics: (1) Phase Alignment Degree (PAD), leveraging RIS to enhance channel coherence for robust SIC; (2) Interference Degree (ID), measuring and suppressing mutual interference; (3) Power Difference Factor (PDF), ensuring adherence to NOMA's power-domain decoding constraints. PIP-UP moves beyond gain-only metrics, explicitly incorporating phase and power dynamics to improve pairing reliability.
- **A3TD Algorithm** harnesses A3C's multi-threaded parallel exploration for broad and efficient sampling of the state-action space, while employing TD3's dual-critic architecture and delayed policy updates to ensure stable and precise optimization in continuous action domains. This enables the simultaneous optimization of D2D channel allocation, transmit power, and RIS phase shifts.

- **Comprehensive Performance Validation:** Through extensive simulations, we demonstrate that the proposed A3TD algorithm, coupled with the PIP-UP strategy, significantly outperforms state-of-the-art baseline algorithms (including AC, DDPG, and TD3) in terms of system sum rate and convergence speed. The results validate the effectiveness of our framework in RIS-assisted CNOMA-D2D networks.

The remainder of this paper is organized as follows: Section 2 presents the system model and problem formulation. Section 3 details the proposed PIP-UP strategy and the A3TD algorithm. Section 4 discusses simulation results and performance analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2 System Model

As illustrated in Fig. 1, we consider a downlink RIS-aided CNOMA-D2D communication system comprising a single-antenna base station (BS), an RIS unit equipped with an $N \times N$ programmable reflecting array, M D2D user pairs denoted by $\mathbb{D} = \{D_1, D_2, \ldots, D_m, \ldots, D_M\}$, and K cellular users (CUs) denoted by $\mathbb{C} = \{C_1, C_2, \ldots, C_k, \ldots, C_K\}$, where $M > K$. The BS provides downlink service to the CUs, while each D2D pair communicates directly by reusing the downlink spectrum resources allocated to the CUs.

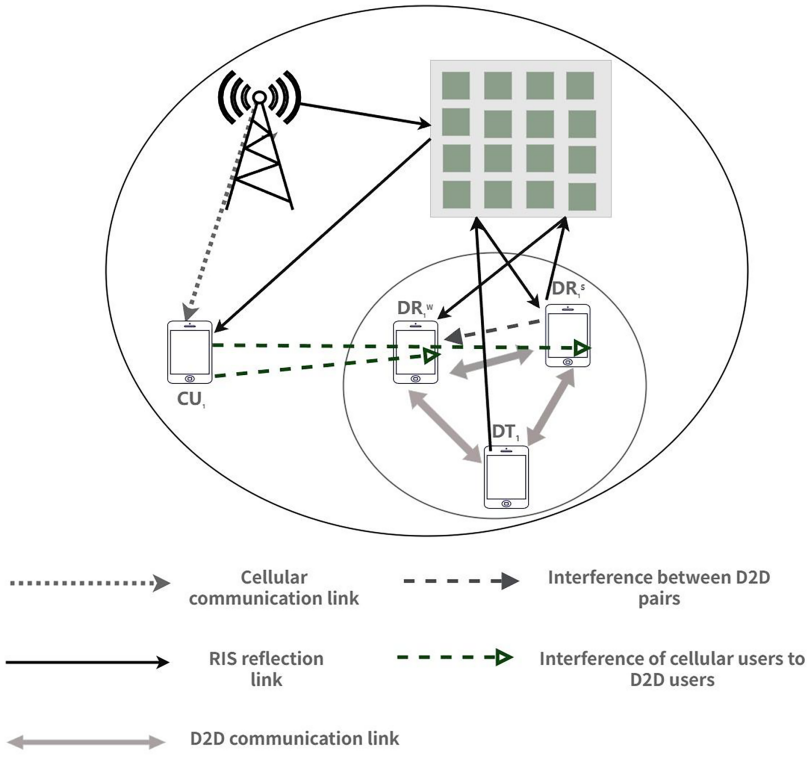


Figure 1: Illustration of an RIS-aided CNOMA-D2D cell.

Each CNOMA-D2D pair comprises a D2D transmitter (Tx) and two receivers (a strong user Rx_{strong} and a weak user Rx_{weak}). The strong user, with full-duplex capability, assists in forwarding the signal of the weak user via the D2D link after decoding its own signal, forming a cooperative NOMA transmission mechanism. The interference in the system mainly comprises three types: (1) Cellular-to-D2D interference Occurs when D2D users reuse cellular spectrum resources, causing signals from the Base Station (BS) to cellular users to interfere with D2D receivers; (2) Intra-pair D2D interference denotes the mutual interference

between the paired users; (3) Inter-pair D2D interference exists between different D2D pairs sharing the same cellular user's resources. The key notations used in this section are listed in Table 2.

Table 2: Key notations.

Symbol	Description
R_{total}	System sum rate (objective)
$\eta_{m,k}$	Binary channel allocation variable
P_m, P_m^{cp}	D2D transmit powers (continuous)
P_k^c	BS transmit power (continuous)
α_m, β_m	Power allocation coefficients
Φ	RIS phase shift matrix
θ_n	Phase shift of n-th RIS element $\in [0, 2\pi)$
γ_k^c	SINR at CU k
$H_{B,k}$	Composite channel gain from BS to CU k (direct + RIS)
I_m^s, I_m^w	Inter-pair D2D interference at strong/weak user of pair m
$\mathbb{D}$	Set of D2D pairs, $\mathcal{D} = \{D_1, \ldots, D_M\}$
$\mathbb{C}$	Set of cellular users (CUs), $\mathcal{C} = \{C_1, \ldots, C_K\}$
M	Number of D2D pairs
K	Number of cellular users
N	Number of RIS reflecting elements ($N \times N$ array)

2.1 Channel Model

The channel gain of the direct link from the BS to cellular user k is denoted as $h_{B,k}$, the channel vector from the BS to the RIS as $G \in \mathbb{C}^{N\times 1}$, and the channel vector from the RIS to cellular user k as $\mathbf{h}_{R,k} \in \mathbb{C}^{N\times 1}$. The total composite channel gain from the BS to cellular user k, denoted as $H_{B,k}$, is given by:

$$H_{B,k} = h_{B,k} + h_{R,k}^H \Phi G \tag{1}$$

where Φ is the reflection matrix of the RIS which is defined as $\Phi = \text{diag}(\phi_1, \phi_2, ..., \phi_N)$, $\phi_n = e^{j\theta_n}$, $\theta_n \in [0, 2\pi)$.

The signal transmitted by the BS to cellular user k is x_k, with transmit power P_k^c. The received signal at user k can be expressed as:

$$y_k^c = H_{B,k}\sqrt{P_k^c}x_k^c + \sum_{m=1}^{M} \eta_{m,k} H_{m,k} s_m + n_k^c \tag{2}$$

where $H_{m,k}$ is the composite channel from D2D transmitter DT_m to cellular user CU_m, s_m is the transmitted signal for DT_m, $\eta_{i,k}$ is the channel reuse coefficient, and n_k is Additive White Gaussian Noise (AWGN).

The signal-to-interference-plus-noise ratio (SINR) for user k can be calculated as:

$$\gamma_k^c = \frac{P_k^c |H_{B,k}|^2}{\sum_{m=1}^{M} \eta_{m,k} P_m |H_{m,k}|^2 + \sigma^2} \tag{3}$$

and the achievable rate of CU_k can be computed using the Shannon formula: $R_k^c = \log_2\left(1 + \gamma_k^c\right)$.

CNOMA transmission Phase I (Direct Transmission): D_m^t transmits a superimposed signal which can be expressed as:

$$s_m = \sqrt{\alpha_m P_m} x_m^s + \sqrt{\beta_m P_m} x_m^w \tag{4}$$

where P_m is the transmit power of DT_m, and α_m and β_m are power allocation coefficients for the information to the strong user and the weak user, respectively, which are satisfying $\alpha_m + \beta_m = 1$ and $\alpha_m > \beta_m = 1$.

Thus, the received signal at the strong user $DR_{s,m}$ and the weak user $DR_{w,m}$ can be calculated as:

$$y_m^s = H_{m,s} s_m + \sum_{j \neq m} \eta_{j,k} H_{j,s} s_j + H_{B,s} \sqrt{P_k^c x_k^c} + n_m^s \tag{5}$$

$$y_m^w = H_{m,w} s_m + \sum_{j \neq m} \eta_{j,k} H_{j,w} s_j + H_{B,w} \sqrt{P_k^c x_k^c} + n_m^w \tag{6}$$

CNOMA transmission Phase II (Cooperative Relaying): After successfully decoding the signal of the weak user $DR_{w,m}$, the strong user $DR_{s,m}$ acts as a decode-and-forward relay and retransmits the signal to $DR_{w,m}$. The SINR by decoding its own signal can be calculated as:

$$\gamma_m^s = \frac{\alpha_m P_m |H_{m,s}|^2}{I_m^s + I_B^s + \sigma^2} \tag{7}$$

The transmitted signal from the strong user to the weak user $DR_{w,m}$ is:

$$s_m^{cp} = \sqrt{P_m^{cp}} x_m^w \tag{8}$$

where P_m^{cp} is the cooperative transmit power of the strong user.

Hence, the received signal at the weak user during the second phase can be expressed as:

$$y_m^{cp} = H_{m,s \to w} s_m^{cp} + \sum_{j \neq m} \eta_{j,k} H_{j,w}^{cp} s_j^{cp} + n_m^{cp} \tag{9}$$

where $H_{m,s \to w}$ is the D2D cooperative channel gain from the strong user to the weak user, which includes both direct and reflected links.

For the weak user $DR_{m,w}$, maximal ratio combining (MRC) is applied to combine the signals received in both phases. The combined effective SINR can be expressed as:

$$\gamma_m^w = \frac{\beta_m P_m \left|H_{m,w}\right|^2}{\alpha_m P_m \left|H_{m,w}\right|^2 + I_m^w + I_B^w + \sigma^2} + \frac{P_m^{cp} \left|H_{m,sw}\right|^2}{I_m^{cp} + \sigma^2} \tag{10}$$

The expressions for inter-pair D2D interference I_m^s and I_m^w, interference from cellular users to D2D users I_B^s and I_B^w, and interference during the cooperative phase I_m^{cp} are:

$$I_m^s = \sum_{j \neq m} \eta_{j,k} P_j \left|H_{j,s}\right|^2, I_m^w = \sum_{j \neq m} \eta_{j,k} P_j \left|H_{j,w}\right|^2 \tag{11}$$

$$I_B^s = P_k^c |H_{B,s}|^2, I_B^w = P_k^c |H_{B,w}|^2 \tag{12}$$

$$I_m^{cp} = \sum_{j \neq m} \eta_{j,k} P_j^{cp} \left|H_{j,w}^{cp}\right|^2 \tag{13}$$

The achievable rate for the D2D user pair m is calculated as:

$$R_m = \log_2(1 + \gamma_m^s) + \log_2(1 + \gamma_m^w) \tag{14}$$

2.2 Problem Formulation

The optimization object of the RIS-aided CNOMA-D2D network is to maximize the sum rate of all D2D links, which is formulated as P1:

$$\text{P1:} \quad \max_{\{\eta_{m,k}, P_m, P_m^d, P_k^c, \Phi_m, \alpha_m, \beta_m, k^c\}} \sum_{m=1}^{M} R_m \tag{15}$$

$$\begin{aligned}
\text{s.t.} \quad & \text{C1: } \gamma_k^c \geq \gamma_{th}^c, \forall k \in C \\
& \text{C2: } \gamma_m^s \geq \gamma_{th}^d, \gamma_m^w \geq \gamma_{th}^d, \forall m \in D \\
& \text{C3: } \eta_{m,k} \in \{0,1\}, \sum_{k=1}^{k} \eta_{m,k} = 1, \forall k \in C, m \in D \\
& \text{C4: } 0 < P_m < P_{\max}^d, 0 < P_m^d < P_{\max}^d, \forall m \in D \\
& \text{C5: } 0 < P_k^c < P_{\max}^c, \forall k \in C \\
& \text{C6: } \alpha_m + \beta_m = 1, \alpha_m > \beta_m > 0, \forall m \in D \\
& \text{C7: } |\phi_n| = 1, \theta_n \in [0, 2\pi), 1 \leq n \leq N^2
\end{aligned}$$

where:

- C1 enforces a minimum SINR requirement for each cellular user to guarantee its quality of service (QoS);
- C2 ensures minimum SINR requirements for both the strong and weak users in each D2D pair to maintain D2D link reliability;
- C3 requires that each D2D pair reuses the channel of exactly one cellular user;
- C4 imposes transmit power constraints on each D2D transmitter during both the direct transmission and cooperative relaying phases;
- C5 limits the transmit power from BS to each cellular user;
- C6 mandates that the power allocation coefficient for the strong user exceeds that for the weak user ($a_m > b_m$) to enable successful SIC;
- C7 specifies a continuous phase-shift model for the RIS reflection coefficients.

Underlay spectrum sharing between CUs and D2D pairs boosts spectral efficiency via spatial reuse but introduces mutual co-channel interference. This trade-off is managed by QoS constraints C1 (CUs) and C2 (D2D users), which maximize D2D sum rate while ensuring all links meet SINR thresholds for balanced coexistence.

Constraint C3 (each D2D pair reuses exactly one CU channel) is motivated by three factors: (i) Complexity control—multi-channel reuse adds binary variables, making the non-convex problem intractable for conventional solvers and DRL algorithms; (ii) Interference localization—single-channel reuse confines interference to one CU and co-channel D2D pairs, simplifying coordination and enhancing RIS phase alignment; (iii) CNOMA compatibility—power-domain NOMA and SIC are designed for a single shared channel; multi-channel operation contradicts NOMA principles and increases transceiver complexity. This widely adopted assumption [1,4,7] ensures a tractable yet realistic evaluation framework.

Compared to discrete phase shift models, the constraints C7 enables finer and more precise phase configuration, maximizing the optimization space for phase alignment. Consequently, the system can accurately match the channel phase characteristics among users, thereby enhancing the reliability of SIC decoding and thus supporting the construction of a continuous action space. The continuous phase-shift model is a well-established assumption in the vast majority of RIS-aided communication literature [12,14,16,22].

3 Methods

3.1 User Pairing Strategy

3.1.1 Unified Pairing Weight (UPW)

In an RIS-aided CNOMA-D2D network, user pairing is critical for achieving efficient resource allocation. The proposed PIP-UP strategy integrates three key factors including phase alignment, interference, and power difference, to ultimately derive a joint pairing weight based on these metrics.

(1) Phase Alignment Degree (PAD) quantifies the similarity in phase between the equivalent channels of two users:

$$PAD_{i,j} = \frac{1}{N}\left|\sum_{n=1}^{N} e^{j\left(\theta_{i,n}-\theta_{j,n}\right)}\right| \tag{16}$$

where $\theta_{i,n}$ and $\theta_{j,n}$ denote the channel phases of user i and user j, respectively, at the n-th reflecting element, and N is the number of RIS reflecting elements.

(2) Interference Degree (ID) quantifies the channel correlation and mutual interference level between two users:

$$ID_{i,j} = \frac{\left|H_{m,i}^{H} H_{m,j}\right|}{\left\|H_{m,i}\right\| \cdot \left\|H_{m,j}\right\|} \tag{17}$$

where $H_{m,i}$ and $H_{m,j}$ are the effective equivalent channel vectors of user i and user j, respectively, which incorporate both the direct and RIS-reflected links between the transmitter and the corresponding receivers. Then, the IDs are normalized by converting it into an interference suppression capability metric $ID\prime$, where a higher value indicates lower interference.

$$ID'_{i,j} = 1 - \frac{ID_{i,j} - \min(ID)}{\max(ID) - \min(ID)} \tag{18}$$

(3) Power Difference Factor (PDF) quantifies the disparity in power requirements between two users:

$$PDF_{i,j} = \left|\frac{p_i - p_j}{p_i + p_j}\right| \tag{19}$$

where P_i and P_j denote the power demands of user i and user j, respectively. In cooperative NOMA, a larger power difference facilitates clearer signal separation and reduces interference during SIC. Hence, a higher PDF value indicates greater suitability for pairing.

UPW is proposed to combine the three key metrics—PAD, ID, and PDF—into a single measure of pairing suitability between two users in an RIS-aided CNOMA-D2D network. Based on the normalized versions of these metrics, the UPW is defined as:

$$UPW_{i,j} = PAD_{i,j} \cdot ID'_{i,j} \cdot PDF_{i,j} \tag{20}$$

where $UPW_{i,j}$ represents the pairing compatibility weight between user i and user j. The product of the three normalized indicators is deliberately chosen to reflect the conjunctive necessity of the underlying physical conditions: successful SIC-based cooperative NOMA requires simultaneous phase alignment (high PAD), low mutual interference (high ID′), and sufficient power disparity (high PDF). If any one of these conditions is severely unsatisfactory, the corresponding metric approaches zero, driving the entire UPW to a negligible value. Thus, the product acts as a soft logical AND gate, ensuring that only pairs exhibiting balanced excellence across all three dimensions are considered highly suitable.

3.1.2 PIP-UP Algorithm Description and Complexity Analysis

Based on UPW, a greedy-based user pairing algorithm is proposed to iteratively select the best user pair under the current conditions to obtain a locally optimal solution. As is illustrated in Algorithm 1, after the composite channel is computed for each receiving user i, a candidate pairing matrix $Pairs$ is constructed to record all feasible user pairs (i, j) with the initialization of $\varnothing$. For each candidate user pair (i, j), its PAD, ID and PDF are calculated and the UPW is subsequently computed. From all remaining candidate pairs, the pair (i^*, j^*) with the maximum UPW is selected:

$$(i^*, j^*) = \arg\max_{(i,j)} UPW_{i,j} \tag{21}$$

Users i^* and j^* are removed from the candidate pairing matrix $\mathcal{A}$, and the pair (i^*, j^*) is added to the pairing matrix $Pairs \leftarrow Pairs \cup \{(i^*, j^*)\}$. Then, the UPW values are recomputed for all remaining user pairs. This process is repeated until all users are paired or no further valid pairs can be formed.

Algorithm 1: PIP-UP algorithm

Input: The composite channel information of M users in user set U: power, phase and channel interference
Output: The user pairing result set *Pairs*

Initialize the candidate pairing matrix $\mathcal{A}$;
Put all possible user pairs *(i, j)* from U in $\mathcal{A}$;
while *The number of remaining users in candidate pairing matrix* $\dim(\mathcal{A}) \geq 2$ **do**
  for $i = 1$ *to* $\dim(\mathcal{A}_{:,j})$ **do**
    for $j = 1$ *to* $\dim(\mathcal{A}_{i,:})$ **do**
      Calculate the phase alignment degree $PAD_{i,j}$ according to the Formula (16);
      Calculate the interference degree $ID_{i,j}$, and normalize it to $ID'_{i,j}$ according to the Formulas (17) and (18);
      Calculate the power difference factor $PDF_{i,j}$ according to the Formula (19);
      Calculate the joint pairing weight $UPW_{i,j}$ of this user pair;
    end
  end
  Traverse the candidate pairing matrix $\mathcal{A}$ select the (i^*, j^*) that satisfies the condition (21);
  Add the optimal user pair (i^*, j^*) to the pairing result set *Pairs*;
  Remove users i^* and j^* from candidate pairing matrix $\mathcal{A}$;
end

The time complexity of the PIP-UP user pairing strategy is determined by the cumulative computational load of three sequential steps: candidate pairing matrix construction, three-dimensional core metric calculation with joint pairing weight derivation, and greedy iterative optimization. For M D2D pairs, the candidate

pairing matrix ought to be traversed M^2 times. Since the calculation of UPW is a linear operation for each candidate user pair within the matrix, M iterations is required to achieve full pairwise pairing. Thus, the time complexity of PIP-UP user pairing strategy is $O(M^3)$.

3.2 Reinforcement Learning Based Resource Allocation Algorithm

Markov Decision Process

In an RIS-aided CNOMA-D2D network, resource allocation requires joint optimization across multiple dimensions. Conventional optimization methods face significant limitations in such high-dimensional and complex environments, suffering from high computational complexity and slow convergence. To address this challenge efficiently, the resource allocation problem is formulated as a decision-making process of deep reinforcement learning (DRL). Specifically, it is modeled as a multi-agent Markov Decision Process (MDP), leveraging DRL's capabilities in parallel exploration and policy optimization to achieve efficient resource allocation.

The reward function is defined as the instantaneous total throughput at time t, incentivizing the agent to improve transmission rates through effective resource allocation.

$$R_{total}(t) = \sum_{i \in \mathcal{U}} R_i(P_t, \boldsymbol{\phi}_t, C_t) \tag{22}$$

where R_i denotes the achievable data rate of user i, which is determined by the power allocation P_t, RIS phase configuration $\boldsymbol{\phi}_t$, and channel assignment C_t at time t; and $\mathcal{U}$ represents the set of all users. Specifically, if the SINR of a user i falls below the preset threshold γ_{th}, its achievable rate is truncated to $R_i = 0$, thereby imposing an implicit penalty on QoS constraint violations.

The state s_t describes the current network environment:

$$s_t = \left\{G_{i,j},\ \boldsymbol{\phi},\ P_i,\ C_i\right\} \tag{23}$$

where $G_{i,j}$ denotes the channel gain matrix between user i and j, $\boldsymbol{\phi}$ represents the current RIS reflection phase configuration, P_i is the power allocation for user i, and C_i indicates the channel assignment of user i.

The action space a_t is defined as:

$$a_t = \{\Delta P,\ \Delta\boldsymbol{\phi},\ \Delta C\} \tag{24}$$

where ΔP denotes the adjustment in power allocation, $\Delta\boldsymbol{\phi}$ represents the update to the RIS phase shifts, and ΔC corresponds to the modification in channel assignment. If the the power, RIS phase and channel assignment violate the constraints C3–C7 after the adjustment from the action, the action will be discarded.

State space-channel gains $G_{i,j}$, RIS phases ϕ, transmit power P_t, and channel assignment C_t-determine per-link SINR via Formulas (3), (7) and (10), and thus define the achievable sum-rate in Formula (14). Action space $(\Delta P, \Delta\phi, \Delta C)$ directly modify these quantities: ΔP adjusts signal and interference power in SINR; $\Delta\phi$ reconfigures RIS coefficients, reshaping composite channels; ΔC reassigns D2D pairs to different CU channels, fundamentally altering the interference topology. A3TD is specifically designed to handle such coupled, high-dimensional action spaces, and its superior capability in learning this complex mapping is validated by the simulation results in Section 4.

The channel gain matrix $G_{i,j}(t)$ evolves according to an exogenous stochastic process; in our simulations we adopt a block fading model where $G_{i,j}$ remains constant within each episode and is re-sampled independently at the start of each new episode. This formulation conforms to the standard MDP framework and enables the agent to learn a policy that optimally responds to both its own previous decisions and the current channel conditions.

3.3 A3TD Deep Reinforcement Learning

To jointly optimize power allocation, phase configuration, and channel assignment in complex channel environments and thereby maximize spectral efficiency, this paper proposes an A3TD deep reinforcement learning algorithm, which integrates the strengths of both A3C and TD3. It leverages A3C's capability of multi-threaded parallel exploration to efficiently explore complex environments, while exploiting TD3's optimization in continuous action space to enable rapid iteration and stable convergence in resource allocation. The architecture of the A3TD algorithm is illustrated in Fig. 2.

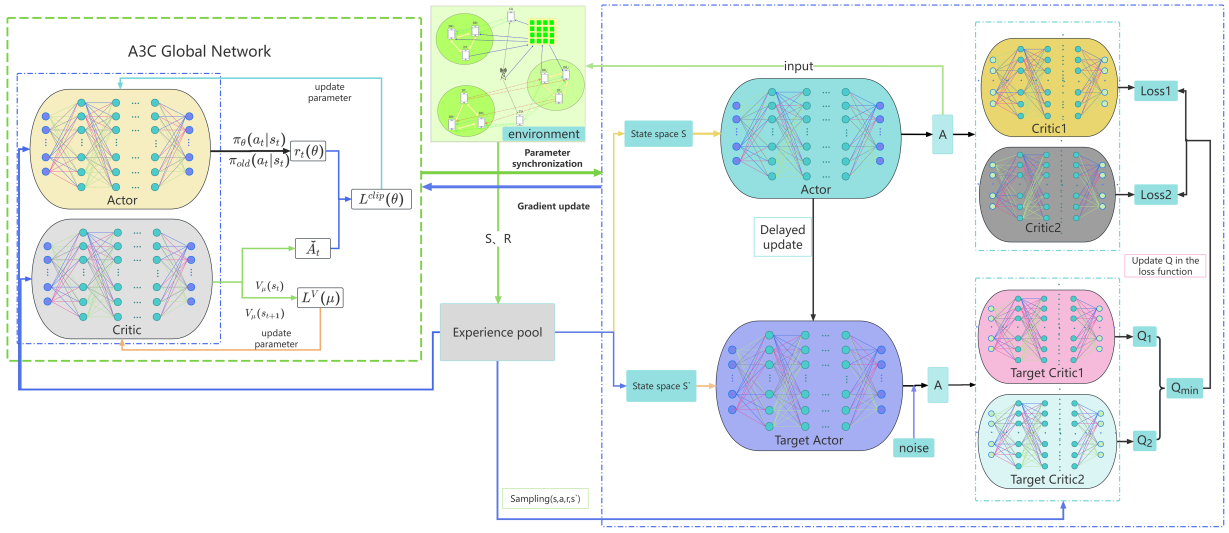


Figure 2: Architecture of the proposed A3TD hybrid deep reinforcement learning framework. Left: A3C module employs multiple parallel actors to explore the environment asynchronously, generating diverse experience trajectories. Right: TD3 module refines the policy using dual critics and delayed updates for stable and precise optimization in continuous action spaces. Center: a shared replay buffer stores transitions collected by A3C and feeds them to TD3, decoupling exploration from learning.

The A3C module accelerates agents' learning through multi-threaded parallel training. Multiple threads interact independently with the environment to generate experience samples; after updating their local AC network parameters, they asynchronously synchronize gradients with the A3C global network.

The TD3 module utilizes the samples generated by A3C: the Critic networks evaluate action values to generate Q-values, while the Actor network determines actions to guide resource allocation and maximize the system sum rate.

During training, the two modules operate in a synergistic and complementary manner: A3C provides diverse samples and exploration direction, while TD3 refines the policy updates through accurate action evaluation, leading to more efficient network performance and faster convergence.

3.3.1 Multi-Threaded A3C-Based Exploration Module

This module is responsible for generating diverse experience data. Its policy network optimizes the probability distribution over actions by maximizing the following objective function:

$$L_{\text{Actor-A3C}} = -\log \pi\left(a_t|s_t;\theta_{\text{Actor}}\right) A(s_t, a_t), \tag{25}$$

where $A(s_t, a_t) = R_{\text{total}}(t) + \gamma V(s_{t+1}) - V(s_t)$ is the advantage function, which measures the improvement of the selected action a_t compared to the average action under the current policy. Here, $\pi(a_t|s_t;\theta_{\text{Actor}})$ denotes the probability of selecting action a_t in state s_t under the policy parameterized by θ_{Actor}. The policy network updates its parameters θ_{Actor} to increase the likelihood of actions with higher advantage values:

$$\theta_{\text{Actor}} \leftarrow \theta_{\text{Actor}} + \alpha \nabla_{\theta} L_{\text{Actor-A3C}}, \tag{26}$$

where α is the learning rate.

Meanwhile, the Critic network estimates the state-value function $V(s_t)$ by minimizing the mean squared error:

$$L_{\text{Critic-A3C}} = \frac{1}{2}\left[R_{\text{total}}(t) + \gamma V(s_{t+1}) - V(s_t)\right]^2, \tag{27}$$

where $V(s_t)$ and $V(s_{t+1})$ are the outputs of the same critic network for the current and next states, respectively. A3C does not employ a separate target network for value estimation [23]. $\gamma \in [0,1)$ is the discount factor that balances current and future rewards.

The parameter update formula for the A3C Critic network is:

$$\theta_{\text{Critic}} \leftarrow \theta_{\text{Critic}} - \beta \nabla_{\theta_{\text{Critic}}} L_{\text{Critic-A3C}} \tag{28}$$

where β is the learning rate of the A3C Critic network. Multiple threads asynchronously update the global network parameters, enhancing the model's generalization ability through diverse samples.

The experience collected by A3C, (s_t, a_t, r_t, s_{t+1}), is stored in the experience replay buffer B of TD3:

$$B = B \cup (s_t, a_t, r_t, s_{t+1}) \tag{29}$$

where a batch of data u is randomly sampled as:

$$u = \left\{\left(s_t^i, a_t^i, r_t^i, s_{t+1}^i\right)\right\}_{i=1}^{N} \tag{30}$$

3.3.2 TD3-Based Policy Refinement Module

This module employs dual Critic networks and delayed policy update mechanism to improve optimization stability and efficiency, which is suitable for resource allocation problems involving high-dimensional continuous action space.

The Critic network estimates action values and updates the target values:

$$y = R_{\text{total}}(t) + \gamma \min_{i=1,2} Q_{\text{target}}^i\left(s_{t+1}, a_{t+1}\right) \tag{31}$$

where $Q_{\text{target}}^i\left(s_{t+1}, a_{t+1}\right)$ denotes the estimated value for the next state-action pair (s_{t+1}, a_{t+1}) by the i-th Critic network. y is the target value, combining the current reward and discounted future rewards. a_{t+1} is the next action generated by the target Actor network with added noise ε:

$$a_{t+1} = \pi_{\text{target}}\left(s_{t+1}\right) + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right) \tag{32}$$

The loss function for the Critic network is calculated as:

$$L_{\text{Critic-TD3}} = \frac{1}{N} \sum_{(s_t, a_t, r_t, s_{t+1}) \in u} \left[Q^i\left(s_t, a_t; \theta_{Q^i}\right) - y\right]^2 \tag{33}$$

The training objective of the Critic network is to minimize the error between estimated and target values. The parameters θ_Q^i of the Critic network are updated using gradient descent:

$$\theta_Q^i \leftarrow \theta_Q^i - \eta \nabla_{\theta_Q^i} L_{\text{Critic}} \tag{34}$$

The Actor network directly generates action a_t, with the objective of maximizing the Q-value of the Critic network:

$$L_{\text{Actor-TD3}} = -\frac{1}{N} \sum_{s_t \in u} Q\left(s_t, \pi\left(s_t; \theta_{\text{Actor-TD3}}\right)\right) \tag{35}$$

where $\pi\left(s_t; \theta_{\text{Actor-TD3}}\right)$ is the action generated by the Actor network in state s_t, and $Q\left(s_t, \pi\left(s_t; \theta_{\text{Actor-TD3}}\right)\right)$ is the Q-value of the Actor's action evaluated by the Critic network. After the training, the Actor network can generate optimal actions a_t to maximize the system's total spectral efficiency by optimizing $L_{\text{Actor-TD3}}$. The parameters $\theta_{\text{Actor-TD3}}$ are updated using gradient ascent:

$$\theta_{\text{Actor-TD3}} \leftarrow \theta_{\text{Actor-TD3}} + \alpha \nabla_{\theta_{\text{Actor-TD3}}} L_{\text{Actor-TD3}} \tag{36}$$

The TD3 enhances training stability through delayed updates of the Actor network and target networks. The Critic network is updated at every step, while the Actor network is updated every two steps. The target Actor network is updated as:

$$\theta_{\text{Actor-local}} \leftarrow \tau \theta_{\text{Actor-local}} + (1-\tau)\theta_{\text{Actor-TD3}} \tag{37}$$

The target Critic networks are updated as:

$$\theta_{\text{Q-local}}^i \leftarrow \tau \theta_{\text{Q-local}}^i + (1-\tau)\theta_Q^i \tag{38}$$

where τ is the soft update factor that controls the update rate of the target networks.

3.3.3 A3TD Algorithm Description and Complexity Analysis

As illustrated in Algorithm 2, the A3TD algorithm combines A3C's asynchronous parallelism with TD3's stability through a hybrid framework. It initializes separate networks: A3C's global Actor (θ) and Critic (ω), and TD3's primary/target Actor-Critic pairs (ϕ, ϕ'; $\psi_1, \psi_2, \psi_1', \psi_2'$), along with a shared replay buffer $\mathbb{B}$. Multiple threads interact with the environment, store transitions in $\mathbb{B}$, and asynchronously update A3C parameters (θ, ω) using local gradients. When the size of $\mathbb{B}$ exceeds threshold u, TD3 updates its twin Critics (ψ_1, ψ_2) and delays Actor (ϕ) updates every 2 steps. Periodically, TD3's Actor (ϕ) is synchronized to A3C's global Actor (θ) to propagate policy improvements. This design integrates TD3's robust value estimation with A3C's distributed exploration, enabling stable and efficient optimization via asynchronous updates and cross-module parameter sharing.

Algorithm 2: A3TD algorithm

1 Initialize global Actor network parameters θ, Critic network parameters ω of A3C module;
2 Initialize TD3 primary Actor network parameters ϕ, twin Critic network parameters ψ_1, ψ_2;
3 Initialize TD3 target Actor network $\phi' \leftarrow \phi$, target twin Critic networks $\psi_1' \leftarrow \psi_1$, $\psi_2' \leftarrow \psi_2$;
4 Initialize shared replay buffer $\mathbb{B}$;
5 Set hyperparameter: A3C thread number S, batch sampling threshold u, parameter synchronization episodes T;

(Continued)

Algorithm 2 (continued)

for $t = 0$ *to* T **do**
  for $i = 0$ *to* S **do**
    Thread i synchronizes A3C global network parameters θ and ω, store experience tuple $(s_t^i, a_t^i, r_t^i, s_{t+1}^i)$ in $\mathbb{B}$;
    if $|\mathbb{B}| > u$ **then**
      Sample batch experiences from $\mathbb{B}$;
      Update TD3 twin Critic networks ψ_1, ψ_2 then update TD3 primary Actor network ϕ and target networks ϕ', ψ_1', ψ_2' every 2 steps;
      Compute local gradient of Thread i to asynchronously update A3C global parameters θ, ω;
    end
    if $t \mod T = 0$ **then**
      Synchronize TD3 primary Actor parameters φ to A3C global Actor parameters θ;
    end
  end
end

The state space is determined by the channel, power, and phase information of M D2D pairs, K cellular users, and $N \times N$ RIS reflecting elements, resulting in the dimensionality of state space $O(M \times K \times N^2)$. The computation of one forward/backward pass of an Actor network for A3C and TD3 actors, and that of one forward/backward pass of a Critic network for A3C critic and each TD3 critic are related to the dimensionality of state space and the depth of the networks D. Besides, S and B denote parallel A3C threads and TD3 mini-batch size, respectively. Thus, the time complexity of A3TD algorithm is $O(M \times K \times N^2 \times D \times (S + B))$.

4 Simulation Results and Analysis

4.1 Simulation Settings

This section presents a comprehensive performance evaluation of the proposed A3TD algorithm within an RIS-assisted CNOMA-D2D network through systematic simulation experiments. The simulation platform is built using Python and the PyTorch deep learning framework for model construction and DRL algorithm training. To benchmark our approach, the proposed A3TD algorithm is compared against three state-of-the-art deep reinforcement learning baseline algorithms:

- AC algorithm [25] implements an Actor-Critic architecture with independent agent learning via an online experience buffer.
- TD3 algorithm [26] extends DDPG with dual Critic networks and delayed policy updates to mitigate value overestimation and enhance training stability.
- DDPG algorithm [27] incorporates experience replay and target network mechanisms with soft parameter updates, optimized for continuous action space resource allocation.

The evaluation employs two key metrics: (i) the system sum rate, representing the total achievable data rate calculated according to Formula (16), and (ii) the convergence time, defined as the number of training episodes required for an algorithm to achieve stable performance. An algorithm is considered converged when the moving average of its sum rate remains within 97% of its peak observed performance for 200 consecutive episodes.

D2D user positions are randomly generated within a single cell with the RIS deployed in the center area, which is a circular area centered at the base station with a 100-m radius—a setup commonly adopted in related works such as [13]. The SINR calculation accounts for both strong and weak user components in the NOMA scheme.

As illustrated in Table 3, the simulation parameters align with both the practical application scenarios of RIS-aided CNOMA-D2D communication systems and the standard simulation specification in the field. The hyperparameters of A3TD is given in Table 4.

Table 3: Simulation parameters.

Parameters	Values
Cell radius	250 m [24]
Carrier frequency	2 GHz [28]
Bandwidth	1 MHz [1]
BS transmit power	43 dBm [28]
D2D transmit power	0.5–2 W [29]
CU/D2D SINR threshold	3 dB/5 dB [2]
Noise power density	−174 dBm/Hz [24]
Shadowing factor	8.0 dB [29]
Pathloss model	Urban macro (UMa) with LoS/NLoS probabilities [24]

Table 4: Hyperparameters in A3TD.

Parameters	Values
Learning rate γ	3×10^{-4}
Discount factor	0.99
Replay buffer capacity	1×10^5 samples
Mini-batch size	128
A3C threads	4
Synchronization interval T	10 steps
Policy delay	2 steps
Hidden layers/neurons	2/[256, 256]

4.2 Simulation Results

Fig. 3 illustrates the relationship between convergence time and the number of D2D pairs. As network density increases, the convergence episodes for all algorithms rise due to heightened environmental complexity. The proposed A3TD algorithm consistently achieves the shortest convergence time, with its advantage becoming more pronounced in denser scenarios. For instance, in a network with 20 D2D pairs, A3TD reduces convergence episodes by approximately 30% to 45% compared to the baselines. In dense scenarios characterized by an increasing number of D2D pairs, the significantly faster convergence exhibited by A3TD is fundamentally attributed to the A3C component's multi-threaded parallel sampling capability. This architecture enables concurrent exploration from diverse environmental states, drastically accelerating the collection of experiences within the high-dimensional state-action space. Consequently, it breaks the sample temporal correlation bottleneck inherent in traditional serial sampling, allowing the algorithm to learn effective resource allocation patterns much more rapidly.

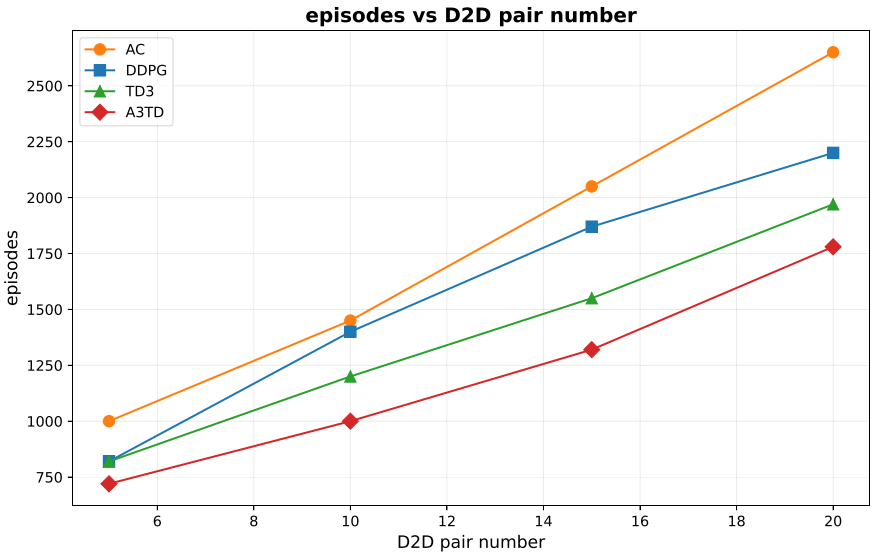
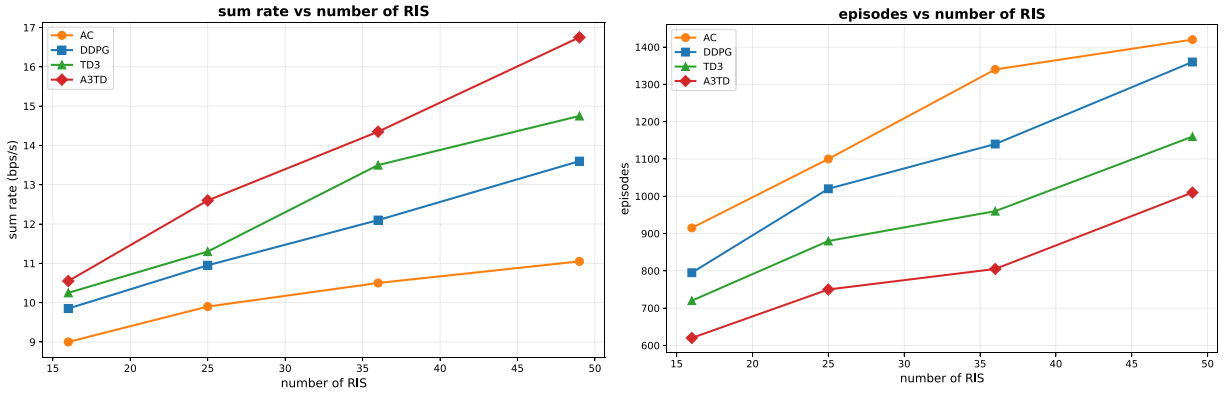


Figure 3: Relationship between D2D pair number and convergence episode.

Fig. 4a depicts the sum rate performance as the number of RIS reflecting elements increases from 16 to 49. All algorithms benefit significantly from a larger RIS, as it provides greater flexibility for phase adjustments, enabling precise signal path optimization and the creation of virtual line-of-sight links.



(**a**) Sum rate vs. Number of RIS units (**b**) Convergence episode vs. Number of RIS units

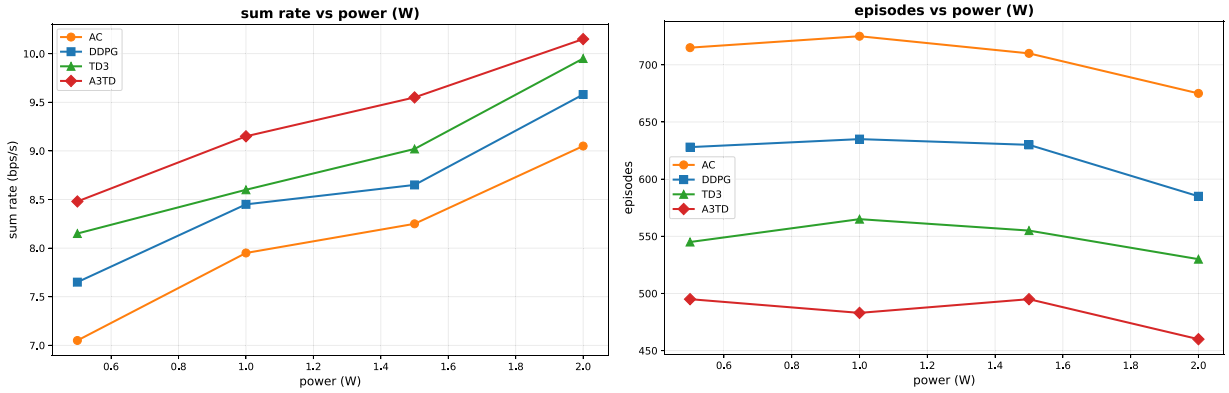
Figure 4: Sum rate, energy efficiency, and convergence episode with different numbers of RIS units.

Fig. 4b shows the corresponding convergence behavior. As the RIS dimension grows, posing a higher-dimensional non-convex optimization challenge, the convergence episodes for all baselines increase markedly. A3TD maintains the fastest convergence, achieving a 23% to 42% reduction in episodes with 49 RIS elements.

As the number of RIS reflecting elements increases—causing the dimensionality of the optimization problem to escalate sharply—the superior convergence and sum-rate performance of A3TD can be attributed to the synergistic effect within its architecture. The extensive exploration driven by A3C ensures that the algorithm avoids getting trapped in the proliferating local optima caused by dimensional expansion. Meanwhile, the TD3 component performs robust policy refinement along the explored directions in this

high-dimensional continuous space, leveraging conservative value estimation (by taking the minimum of dual critic networks) and delayed policy updates, thereby achieving steady performance improvement.

The effect of maximum transmit power is analyzed in Fig. 5. As shown in Fig. 5a, increasing power from 0.5 to 2.0 W improves the sum rate for all algorithms, following the trend predicted by the Shannon capacity formula. However, the marginal gain diminishes at higher power levels due to the concurrent increase in co-channel interference within the multi-user D2D network.



(**a**) Sum rate vs. Maximum transmit power(W)

(**b**) Convergence episode vs. Maximum transmit power(W)

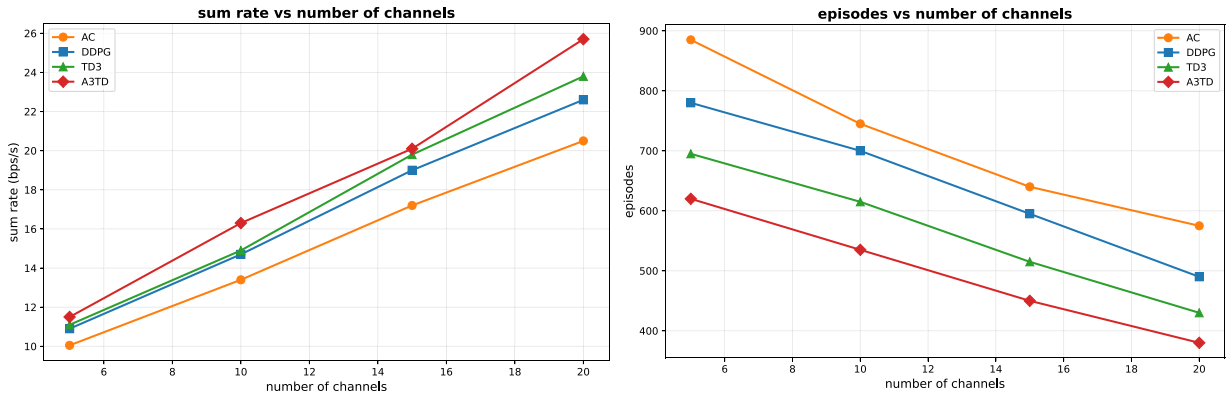
Figure 5: Sum rate and convergence episode with different maximum transmission power.

Fig. 5b reveals that transmit power variations have a relatively minor impact on convergence time compared to network density or RIS size. This is because power scaling primarily affects the reward magnitude without altering the dimensionality or fundamental complexity of the state/action space. A3TD consistently delivers the highest sum rate across all power levels, and its performance advantage remains stable, verifying the robustness of its hybrid architecture under different power constraints.

Fig. 6 investigates performance under varying numbers of available channels. Fig. 6a confirms that ample channel resources (increasing from 5 to 20) significantly alleviate co-channel interference, leading to substantial sum rate improvements for all algorithms. In resource-constrained scenarios (5 channels), algorithmic performance is similar. However, as resources become abundant, the superior optimization capability of A3TD becomes more evident, widening the performance gap.

Correspondingly, Fig. 6b shows that convergence is faster in resource-rich environments. A constrained action space forces agents into repeated iterations to resolve allocation conflicts, while a broader action space allows algorithms like A3TD to more readily discover high-reward strategies, accelerating convergence.

The evolutions of spectral efficiency during training using different algorithms are compared in Fig. 7. The learning curves distinctly highlight the differences in convergence speed, stability, and final performance. The A3TD algorithm demonstrates the most rapid convergence and the highest stability. The AC algorithm exhibits high variance and policy drift. While DDPG and TD3 improve stability through experience replay and target networks, and TD3 achieves higher final performance than DDPG via its twin critics, A3TD surpasses them all.



(**a**) Sum rate vs. number of channels (**b**) Convergence episode vs. number of channels

Figure 6: Sum rate and convergence episode with different channel numbers.

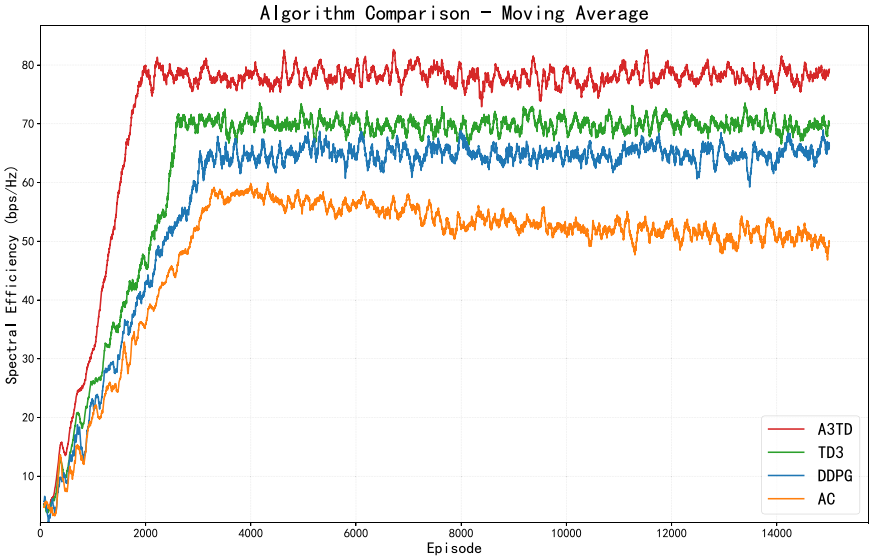


Figure 7: Sum spectral efficiency in training episodes (Moving average).

The rapid convergence and high stability of A3TD demonstrated in the learning curves are a direct manifestation of its hybrid design advantages. The A3C component provides diverse, decorrelated training samples, accelerating the initial learning phase. Concurrently, the TD3 component leverages these samples for stable policy optimization, thereby mitigating the inherent instability often associated with AC algorithms, as well as avoiding the insufficient exploration or slow convergence issues that may plague DDPG/TD3 in complex environments. In contrast, the AC framework suffers from inefficient exploration and policy instability; DDPG exhibits shortcomings in both exploration and high-dimensional optimization; while although TD3 improves stability, its exploration efficiency remains limited.

4.3 Design Guideline and Key Trade-Off

When the action space contains both discrete and continuous dimensions, a single DRL algorithm is often suboptimal. Our A3TD framework use a discrete-parallel-exploration-oriented A3C algorithm to efficiently sample diverse state-action combinations and break temporal correlations, and employ a

continuous-control-oriented TD3 algorithm with double critics and delayed updates to achieve stable and precise policy learning. This "explorer + refiner" synergy is transferable to any problem where exploration breadth and optimization precision are both critical.

A3C's multi-threaded asynchronous exploration significantly improves sample diversity but can introduce delayed and stale gradients, potentially destabilizing the shared global model. Our implementation mitigates this via experience replay buffering between A3C and TD3, but the fundamental tension remains: more exploration often comes at the cost of less stable learning. Researchers adopting this hybrid approach should carefully balance the number of parallel threads, the frequency of global updates, and the replay buffer size.

5 Conclusions

This paper has presented a deep reinforcement learning (DRL) framework for the joint optimization of multi-dimensional resources in RIS-aided cooperative NOMA device-to-device (CNOMA-D2D) networks, with the objective of maximizing the system sum rate. The core of this framework is a novel hybrid algorithm, A3TD, which seamlessly integrates A3C and TD3 paradigms. The A3TD algorithm effectively addresses the high-dimensional, non-convex challenge of simultaneous D2D channel assignment, transmit power control, and RIS phase-shift configuration by reformulating it as a MDP. This approach leverages the strong function approximation capability of deep neural networks alongside the adaptive, experience-driven optimization of reinforcement learning, thereby circumventing the prohibitive computational complexity of traditional methods. Extensive simulation results conclusively validate the superior efficacy of the proposed framework, demonstrating significant outperformance over state-of-the-art baselines in terms of spectral efficiency, energy efficiency, and convergence speed.

Looking forward, while the proposed scheme shows great promise, its practical deployment becomes more realistic. For examples, in smart factories, metallic infrastructure causes severe signal blockage, making direct links unreliable. Deploying low-cost RIS panels on ceilings or production lines creates virtual LoS paths to shadowed devices. D2D communication among AGVs and robots enables direct cooperative relaying. Our PIP-UP strategy pairs devices with aligned RIS phases and sufficient power disparity, ensuring robust SIC in fast-changing industrial channels. The A3TD algorithm then jointly optimizes transmit power, RIS phase shifts, and channel assignment without manual reconfiguration.

Future research will focus on enhancing the algorithm's robustness and applicability by incorporating critical real-world constraints. Key directions include: (i) investigating the impact of user mobility on time-varying channel states and dynamic RIS reconfiguration; (ii) developing distributed optimization strategies for scenarios involving multiple cooperating RISs [30]; (iii) accounting for practical impairments such as imperfect channel state information (CSI) and hardware distortions in transceivers and RIS elements; and (iv) exploring joint network planning problems to determine the optimal number and strategic placement of RIS panels within the cellular infrastructure.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the National Natural Science Foundation of China, grant number 62362052.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Zongchuan Li and Chen Sun; methodology, Zongchuan Li and Chen Sun; software, Zongchuan Li; validation, Zongchuan Li; formal analysis, Zongchuan Li and Chen Sun; investigation, Zongchuan Li and Chen Sun; resources, Chen Sun and Jian Shu; data curation, Zongchuan Li; writing—original draft preparation, Zongchuan Li and Chen Sun; writing—review and

editing, Chen Sun; visualization, Zongchuan Li; supervision, Chen Sun and Jian Shu; project administration, Jian Shu; funding acquisition, Jian Shu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Chen Sun, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhao X, Liu F, Zhang YJ, Chen SC, Gan J. Energy-efficient power allocation for full-duplex device-to-device underlaying cellular networks with NOMA. Electronics. 2023;12(16):3433–47. doi:10.3390/electronics12163433.
2. Nair RB, Kirthiga S. Ant colony optimization based user pairing, power allocation and relaying in a cooperative NOMA system. Wirel Pers Commun. 2025;140(3–4):905–43. doi:10.1007/s11277-025-11752-0.
3. Chen YL, Ai B, Zhang HL, Niu Y, Song LY, Han Z, et al. Reconfigurable intelligent surface assisted device-to-device communications. IEEE Trans Wirel Commun. 2021;20(5):2792–804.
4. Jia J, Tian QZ, Du A, Chen J, Wang XW. DE-based resource allocation for D2D-assisted NOMA systems. Soft Comput. 2024;28(4):3071–82. doi:10.1007/s00500-023-09266-7.
5. Khan MAA, Kaidi HM, Ahmad N, Ur Rehman M. Sum throughput maximization scheme for NOMA-enabled D2D groups using deep reinforcement learning in 5G and beyond networks. IEEE Sens J. 2023;23(13):15046–57. doi:10.1109/jsen.2023.3276799.
6. Dinh P, Arfaoui MA, Sharafeddine S, Assi C, Ghrayeb A. Joint user pairing and power control for C-NOMA with full-duplex device-to-device relaying. IEEE Trans Wirel Commun. 2023;22(5):3103–15. doi:10.1109/globecom38437.2019.9013180.
7. Gu XH, Zhang GA, Zhuo BT, Duan W, Wang J, Wen MW, et al. On the performance of cooperative NOMA downlink: a RIS-aided D2D perspective. IEEE Trans Cogn Commun Netw. 2023;9(6):1612–24.
8. Yang G, Liao YT, Liang YC, Tirkkonen O. Reconfigurable intelligent surface empowered underlaying device-to-device communication. In: Proceedings of the 2021 IEEE Wireless Communications and Networking Conference; 2021 Mar 29–Apr 1; Nanjing, China. Piscataway, NJ, USA: IEEE Press; 2021. p. 1–6.
9. Yang G, Liao YT, Liang YC, Tirkkonen O, Wang GP, Zhu X. Reconfigurable intelligent surface empowered device-to-device communication underlaying cellular networks. IEEE Trans Commun. 2021;69(11):7797–805.
10. Amer A, Hoteit S, Ben Othman J. Throughput maximization in multi-slice cooperative NOMA-based system with underlay D2D communications. Comput Commun. 2024;217(4):134–51. doi:10.1016/j.comcom.2024.01.030.
11. Vishnoi V, Budhiraja I, Gupta S, Kumar N. A deep reinforcement learning scheme for sum rate and fairness maximization among D2D pairs underlaying cellular network with NOMA. IEEE Trans Veh Technol. 2023;72(10):13506–22. doi:10.1109/tvt.2023.3276647.
12. Chandra KR, Borugadda S. Multi agent deep reinforcement learning with deep Q-network based energy efficiency and resource allocation in NOMA wireless systems. In: Proceedings of the 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT); 2023 Apr 5–7; Trichirappalli, India. p. 1–8.
13. Ao SY, Niu Y, Han Z, Zhong ZD, Ai B, Wang N, et al. Resource allocation for RIS-assisted device-to-device communications in heterogeneous cellular networks. IEEE Trans Veh Technol. 2023;72(9):11748–55. doi:10.1109/tvt.2023.3267032.
14. Sultana A, Moniruzzaman M. Spectrum efficiency maximization of reconfigurable intelligent surface assisted device-to-device networks: an actor-critic approach. In: Proceedings of the 2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC); 2023 Sep 5–8; Toronto, ON, Canada. p. 1–7.

15. Liu Y, Xu K, Xia XC, Xie W, Ma N, Xu JH. Joint power control and passive beamforming optimization in RIS-assisted anti-jamming communication. Front Inf Technol Electron Eng. 2024;25(4):537–50. doi:10.1631/fitee.2200646.
16. Wang YZ, Sun MY, Cui QM, Chen KC, Liao YX. RIS-aided proactive mobile network downlink interference suppression: a deep reinforcement learning approach. Sensors. 2023;23(14):6550. doi:10.3390/s23146550.
17. Dong GQ, Yang Z, Feng YH, Lyu B. Exploiting RIS-aided cooperative non-orthogonal multiple access with full-duplex relaying. IEICE Trans Fundam Electron Commun Comput Sci. 2023;106(7):1014–8. doi:10.1587/transfun.2022eal2067.
18. Liu YK, Chen W, Tang HY, Wang KL. Resource allocation in the RIS assisted SCMA cellular network coexisting with D2D communications. IEEE Access. 2023;11:39978–89. doi:10.1109/access.2023.3269284.
19. Zhai Q, Dong LM, Liu CX, Li Y, Cheng W. Resource management for active RIS aided multi-cluster SWIPT cooperative NOMA networks. IEEE Trans Netw Serv Manag. 2024;21(4):4421–33. doi:10.1109/tnsm.2024.3395298.
20. Saikia P, Singh K, Taghizadeh O, Huang WJ, Biswas S. Meta reinforcement learning-based spectrum sharing between RIS-assisted cellular communications and MIMO radar. IEEE Trans Cogn Commun Netw. 2024;10(1):168–79. doi:10.1109/tccn.2023.3319543.
21. Liu Y, Li Y, Li L, He M. NOMA resource allocation method based on prioritized dueling DQN-DDPG network. Symmetry. 2023;15(6):1170. doi:10.3390/sym15061170.
22. Ji Z, Qin Z, Parini CG. Reconfigurable intelligent surface aided cellular networks with device-to-device users. IEEE Trans Commun. 2022;70(3):1808–19. doi:10.1109/tcomm.2022.3145570.
23. Mnih V, Puigdomenech Badia A, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning; 2016 Jun 19–24; New York, NY, USA. p. 1928–37.
24. 3GPP. TR 36.901: study on channel model for frequencies from 0.5 to 100 GHz (Release 17). [cited 2025 Dec 1]. Available from: https://www.3gpp.org/ftp/Specs/archive/38_series/38.901/38901-h10.zip.
25. Wang X, Zhang H, Long K. Power control based on DRL algorithm for D2D-enabled networks. In: Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM); 2021 Dec 7–11; Madrid, Spain. p. 1–5.
26. Liu XY, Xu JX, Zheng KC, Zhang GL, Liu J, Shiratori N. Throughput maximization with an AoI constraint in energy harvesting D2D-enabled cellular networks: an MSRA-TD3 approach. IEEE Trans Wirel Commun. 2025;24(2):1448–66. doi:10.1109/twc.2024.3509475.
27. Guo L, Jia J, Chen J, Du A, Wang X. Deep reinforcement learning empowered joint mode selection and resource allocation for RIS-aided D2D communications. Neural Comput Appl. 2023;35(25 Suppl):18231–49. doi:10.1007/s00521-023-08745-0.
28. 3GPP. TS 38.104: NR; base station radio transmission and reception. [cited 2026 Jan 1]. Available from: https://www.3gpp.org/ftp/Specs/archive/38_series/38.104/38104-j30.zip.
29. 3GPP. TS 38.101: NR; user equipment (UE) radio transmission and reception; part 1: range 1 standalone. [cited 2025 Dec 1]. Available from: https://www.etsi.org/deliver/etsi_ts/138100_138199/13810101/19.03.01_60/ts_13810101v190301p.pdf.
30. Zhang S, Tong X, Chi K, Gao W, Chen X, Shi Z. Stackelberg game-based multi-agent algorithm for resource allocation and task offloading in MEC-enabled C-ITS. IEEE Trans Intell Transp Syst. 2025;26(10):17940–51. doi:10.1109/tits.2025.3553487.