D3QN-Based Joint Power and Reflection Optimization for Energy-Efficient Anti-Jamming against Dynamic Jamming Attacks in Backscatter-Enabled 6G IoT Networks

DOI: https://doi.org/10.54654/isj.v2i28.6393

Le Hoang Hiep, Ngo Huu Huy*

Abstract— **This paper develops a Dueling Double Deep Q-Network (D3QN) framework for joint transmit-power and reflection-coefficient control in backscatter-enabled 6G Internet-of-Things networks subject to random, reactive, and time-varying jamming. The problem is formulated as a jammer-aware Markov decision process with a 20-action power–reflection space and a normalized reward that balances spectral efficiency, transmit-power consumption, and low-SINR outage. The proposed method is evaluated against DQN, DDQN, PPO, TD3, SAC, random selection, and fixed control over five independent seeds with 95% confidence intervals. Under random jamming at $P_j = 30$ dBm, D3QN improves spectral efficiency by 19.83% and reduces outage probability by 6.99% relative to DQN. Under a common dynamic-jamming trace, it provides a 2.12% spectral-efficiency gain while using an average transmit power of 0.427 W, with the maximum-power action selected in only 17.6% of the slots. The 20-action design also improves spectral efficiency by 4.21% over a 9-action grid while using 59.18% fewer outputs than a 49-action grid. Under strong reactive jamming, D3QN improves energy efficiency over DQN by 2.54%, but exhibits lower spectral efficiency and slightly higher outage, revealing a rate - energy - reliability trade -off. Overall, the proposed framework offers a practical balance among communication performance, energy use, and action-space complexity.**

Tóm tắt— **Bài báo này phát triển một khung điều khiển dựa trên mạng Dueling Double Deep Q-Network (D3QN) nhằm tối ưu đồng thời công suất phát và hệ số phản xạ trong mạng IoT 6G hỗ trợ truyền thông tán xạ ngược dưới tác động của nhiễu ngẫu nhiên, nhiễu phản ứng và nhiễu biến đổi theo thời gian. Bài toán được mô hình hóa dưới dạng quá trình quyết định Markov có nhận biết trạng thái gây nhiễu, với không gian gồm 20 hành động kết hợp công suất - phản xạ và hàm phần thưởng chuẩn hóa nhằm cân bằng hiệu suất phổ, mức tiêu thụ công suất phát và xác suất gián đoạn do SINR thấp. Phương pháp đề xuất được so sánh với DQN, DDQN, PPO, TD3, SAC, lựa chọn ngẫu nhiên và điều khiển cố định trên năm hạt giống ngẫu nhiên độc lập, với khoảng tin cậy 95%. Trong điều kiện nhiễu ngẫu nhiên tại $P_j = 30$ dBm, D3QN cải thiện hiệu suất phổ 19.83% và giảm xác suất gián đoạn 6.99% so với DQN. Dưới một chuỗi nhiễu động chung, phương pháp đạt mức tăng 2.12% về hiệu suất phổ, trong khi công suất phát trung bình chỉ là 0.427 W và hành động công suất cực đại chỉ được lựa chọn trong 17.6% số khe thời gian. Cấu hình 20 hành động cũng cải thiện hiệu suất phổ 4.21% so với lưới 9 hành động, đồng thời sử dụng ít hơn 59.18% số đầu ra so với lưới 49 hành động. Trong điều kiện nhiễu phản ứng mạnh, D3QN cải thiện hiệu suất năng lượng 2.54% so với DQN, nhưng đạt hiệu suất phổ thấp hơn và xác suất gián đoạn cao hơn một cách nhẹ, cho thấy sự đánh đổi giữa tốc độ truyền, năng lượng và độ tin cậy. Nhìn chung, khung đề xuất cung cấp sự cân bằng thực tiễn giữa hiệu năng truyền thông, mức sử dụng năng lượng và độ phức tạp của không gian hành động.**

Keywords— 6G IoT; anti-jamming; backscatter communication; D3QN; energy efficiency.

Từ khóa— IoT 6G; chống nhiễu; truyền thông; backscatter; D3QN; hiệu xuất năng lượng.

This manuscript was received on April 6, 2026. It was reviewed on July 6, 2026, revised July 13, 2026 and accepted August 18, 2026.

* Corresponding author.

I. Introduction

Sixth-generation (6G) wireless networks are expected to support massive Internet-of-Things (IoT) connectivity, ubiquitous sensing, and low-latency services while accommodating a large number of energy- and hardware-constrained devices [1 - 4]. The convergence of advanced communication paradigms and intelligent resource management is central to realizing the 6G vision, yet achieving reliable communication under stringent energy and hardware constraints remains challenging, particularly in interference-limited and adversarial wireless environments [5].

Backscatter communication has emerged as a promising low-power communication paradigm in which passive devices convey information by modulating and reflecting incident radio-frequency (RF) signals rather than generating an independent carrier [6 - 8]. By eliminating several power-demanding RF components, backscatter devices can support long-term sensing and connectivity with substantially lower energy consumption than conventional active transmitters [9, 10], Reconfigurable intelligent surfaces (RIS) further extend backscatter capabilities by enabling controllable signal reflection and beamforming, thereby improving link quality in challenging propagation environments [11 - 13]. Symbiotic radio allows active and passive transmissions to share both spectrum and RF sources, improving spectral and energy utilization [14, 15]. However, the relatively weak reflected signal and the dependence on an external RF source make backscatter links particularly vulnerable to interference and intentional jamming [16].

Jamming attacks deliberately inject interference into the wireless medium to reduce the signal-to-interference-plus-noise ratio (SINR), increase outage probability, and disrupt legitimate communications. As reviewed in [17], intelligent anti-jamming mechanisms are essential for future wireless systems in which adversaries can observe legitimate transmissions and adapt their attack strategies [18]. Unlike static interference sources, a dynamic jammer may vary its power, activation pattern, and attack mode over time [19]. Conventional frequency-hopping, spread-spectrum, and threshold-based defense mechanisms can become less effective when jammer behavior evolves faster than the corresponding sensing and adaptation procedures [20]. Game-theoretic formulations have provided a principled foundation for modeling adversarial interactions between legitimate users and jammers [21, 22]; however, computing equilibrium strategies in practice requires channel and jammer models that are typically unavailable.

Dynamic jamming creates a sequential decision problem because the quality of a control action depends on the instantaneous propagation conditions, current jammer state, and subsequent environmental evolution. Traditional optimization methods generally require an accurate analytical model, full channel knowledge, or repeated optimization whenever the environment changes [23]. These requirements limit their applicability to real-time anti-jamming control [24]. Earlier model-based approaches formulated anti-jamming defense as a Markov decision process (MDP) and derived policies using policy iteration or game-theoretic equilibria. However, such approaches become difficult to apply when channel and jammer transition models are unknown or time-varying. Deep reinforcement learning (DRL) has consequently attracted substantial attention as a model-free alternative for sequential anti-jamming decision-making [25 - 27].

In a backscatter-assisted system, anti-jamming performance depends on both the transmit power P_t of the active IoT transmitter and the reflection coefficient β of the passive device. These two variables are intrinsically coupled because they jointly determine the received power of the direct and backscattered components [28]. Optimizing only P_t may consume unnecessary energy, whereas independently selecting β may fail to account for the current direct-link quality and jamming level. Joint resource allocation in backscatter-enabled systems is therefore nonlinear and generally non-convex [29]. Under dynamic jamming, the problem becomes more challenging because the controller must repeatedly select an appropriate power–reflection pair without knowing future channel and attack realizations.

DRL provides a model-free approach for learning sequential control policies through interaction with the environment. DQN has been employed for discrete anti-jamming power and channel selection [30]; however, its use of the same value estimator for action selection and evaluation can produce overestimation bias [31]. Double DQN mitigates this effect by separating these operations, and the dueling architecture separately estimates state values and action advantages. Their combination, commonly referred to as Dueling Double DQN or D3QN, has demonstrated effectiveness for finite control sets in which several actions may have similar values under a given channel state [26]. Recent work has further explored anti-jamming policy learning under wideband and frequency-hopping scenarios [32].

Actor–critic algorithms provide alternative solutions for continuous or hybrid wireless control. Proximal policy optimization (PPO) stabilizes policy updates using a clipped objective and has been applied in multi-agent anti-jamming contexts [33]. Twin delayed deep deterministic policy gradient (TD3) uses twin critics, target-policy smoothing, and delayed actor updates to reduce value overestimation in continuous-action problems, and has been applied to wireless resource allocation and anti-jamming [34]. Soft actor–critic (SAC) combines twin critics with entropy-regularized policy learning to improve exploration and robustness. Comprehensive surveys have catalogued these and related DRL techniques for next-generation wireless resource management [35, 36]. Actor–critic architectures have also been investigated for IRS-aided anti-jamming and dynamic channel access [37, 38].

Nevertheless, DQN, PPO, TD3, and SAC are generic learning algorithms; they are not inherently aware of backscatter propagation, jammer dynamics, or the coupling between active transmit power and passive reflection. Their effectiveness therefore depends on the system state, action representation, reward function, and deployment constraints used in a particular anti-jamming formulation. Existing studies typically focus on active-radio channel selection, power control, trajectory design, or reflecting-surface configuration [39, 40]. Comparatively limited attention has been given to learning a unified power–reflection policy for a backscatter-assisted link under random and reactive jamming.

The integration of AI-empowered approaches with backscatter communication has opened promising research avenues for IoT networks [41]. While RIS-aided anti-jamming approaches have demonstrated benefits through passive beamforming optimization [37, 42], these methods typically assume active transmitters with dedicated RF chains [43]. The unique challenges of jointly optimizing an active transmitter's power alongside a passive backscatter device's reflection coefficient under adversarial conditions remain insufficiently explored. Furthermore, intelligent anti-jamming strategies for IoT systems with limited channel state information require principled approaches that do not rely on exhaustive environment models [44].

Three research gaps consequently remain. First, many resource-control frameworks optimize active transmit power and passive reflection/phase variables separately, despite their coupled impact on the received SINR. Second, existing DRL-based anti-jamming studies often omit the jammer mode, jammer–receiver channel, or previous control action from the decision state, which limits their ability to capture temporal attack behavior. Third, several studies report single-run learning curves or point-wise gains without quantifying random-seed variability, confidence intervals, or the sensitivity of the results to the selected learning algorithm.

To address these limitations, this paper develops a D3QN-based joint power and reflection control framework for backscatter-enabled 6G IoT communications under dynamic jamming. The main contributions are summarized as follows:

1) We establish a backscatter-assisted communication and dynamic jamming model that incorporates the direct transmitter–receiver link, the cascaded transmitter–backscatter–receiver link, and the jammer–receiver link. Random and reactive attack modes with time-varying jamming power are represented through a finite-state Markov process.
2) We formulate the anti-jamming problem as a jammer-aware MDP. The state includes normalized channel gains, the estimated

jammer power and mode, and the previous control action. Transmit power and reflection coefficient are selected jointly rather than through independent optimization stages.

3) We construct a coupled discrete action space $\mathcal{A} = \mathcal{P} \times \mathcal{B}$, in which each action corresponds to one feasible (P_t, β_t) pair. The adopted configuration contains four transmit-power levels and five reflection-coefficient levels, resulting in 20 joint actions. A normalized reward is designed to balance spectral efficiency, transmit-power consumption, and low-SINR outage avoidance.
4) We employ D3QN as the solution mechanism for the finite joint action space. Double Q-learning reduces value overestimation, while the dueling architecture separates the quality of a channel–jammer state from the relative advantages of alternative power–reflection pairs. Computational complexity and practical edge-assisted deployment are also analyzed.
5) We evaluate the proposed framework against DQN, DDQN, PPO, TD3, SAC, random-action, and fixed-control baselines under common channel and jammer traces. The experiments use multiple independent random seeds and report means, standard deviations, and 95% confidence intervals. Ablation tests further isolate the effects of Double Q-learning, the dueling structure, and reflection adaptation.

The novelty of the proposed work is methodological rather than a modification of the generic D3QN architecture. Specifically, the contribution lies in integrating a jammer-aware state representation, a coupled hardware-feasible power–reflection action space, and an energy- and outage-aware reward into a unified backscatter anti-jamming MDP. D3QN is selected because its finite output layer maps directly to the available RF power and impedance states, whereas continuous-action baselines require clipping and projection to the nearest feasible control pair during deployment.

Table I compares the learning mechanisms and action implementations used in the unified experimental setting. All methods are evaluated with the same backscatter channel model, dynamic-jammer observations, reward definition, interaction budget, and statistical protocol. Therefore, the table compares the algorithmic differences among the implemented baselines rather than implying that each generic algorithm is inherently backscatter- or jammer-aware.

The remainder of this paper is organized as follows. Section II presents the system model, dynamic jammer, MDP formulation, D3QN solution, computational complexity, and statistical evaluation protocol. Section III reports and discusses the simulation results, including baseline and ablation comparisons. Section IV concludes the paper and outlines future research directions.

II. Research Methodology

A. System and Signal Model

We consider a backscatter-enabled 6G Internet-of-Things (IoT) communication system consisting of an active IoT transmitter, a passive backscatter device, a legitimate receiver, and a dynamic jammer, as illustrated in Figure 1. The active transmitter sends an RF signal to the receiver, while the backscatter device conveys its information by modulating and reflecting the incident waveform rather than generating an independent carrier. This architecture is suitable for energy-constrained IoT deployments because the backscatter device requires only a small amount of power for impedance switching and baseband processing.

Let $h_{tr,t}$, $h_{tb,t}$, $h_{br,t}$, and $h_{jr,t}$ denote the channel coefficients of the transmitter–receiver, transmitter–backscatter, backscatter–receiver, and jammer–receiver links at time slot t, respectively. Each channel is represented as:

$$h_{xy,t} = \sqrt{L_0 \left(\frac{d_{xy}}{d_0} \right)^{-\nu}} \widetilde{h}_{xy,t}, \qquad (1)$$

where L_0 is the reference path gain at distance d_0, d_{xy} is the corresponding link distance, ν is the path-loss exponent, and $\widetilde{h}_{xy,t} \sim \mathcal{CN}(0,1)$ models Rayleigh fading. The associated channel power gain is denoted by

TABLE I. ALGORITHMIC COMPARISON OF DRL METHODS UNDER THE UNIFIED ANTI-JAMMING EVALUATION SETTING

Method	Native Action Domain	Value/Policy Stabilization	Joint P_t–β Realization	Executed Action in This Study	Jammer-Aware State	Multi-Seed Evaluation
DQN [32]	Discrete	Target network; single-Q estimation	Cartesian-product action indexing	20 discrete pairs	Yes	Yes
DDQN [38]	Discrete	Double Q-learning	Cartesian-product action indexing	20 discrete pairs	Yes	Yes
PPO [45]	Discrete	Clipped actor–critic objective	Categorical policy over joint actions	20 discrete pairs	Yes	Yes
TD3 [39]	Continuous	Twin critics; delayed actor; target-policy smoothing	Two-dimensional continuous actor	Clipped and mapped to nearest feasible pair	Yes	Yes
SAC [46]	Continuous	Twin critics and entropy regularization	Two-dimensional stochastic actor	Clipped and mapped to nearest feasible pair	Yes	Yes
Proposed D3QN	**Coupled discrete**	**Double Q-learning and dueling value decomposition**	**Direct Cartesian-product action indexing**	**20 discrete pairs**	**Yes**	**Yes**

Note: All learning methods use the same backscatter-enabled system model, dynamic-jammer information, reward definition, training interaction budget, and seed-specific channel and jammer traces. For TD3 and SAC, the two continuous outputs are clipped to the feasible ranges and mapped to the nearest available power–reflection pair before execution. Multi-seed evaluation denotes reporting over independent runs using the mean, standard deviation, and 95% confidence interval.

$$g_{xy,t} = |h_{xy,t}|^2 . \tag{2}$$

Let x_t and z_t denote the unit-power signals transmitted by the legitimate transmitter and jammer, respectively. The backscatter symbol is denoted by b_t, with $\mathbb{E}[|b_t|^2] = 1$. The received signal can be expressed as:

$$\begin{aligned} y_t = \sqrt{P_t}\, h_{tr,t} x_t + \sqrt{\beta_t P_t}\, h_{tb,t} h_{br,t} x_t b_t \\ + \sqrt{P_{j,t}}\, h_{jr,t} z_t + n_t, \end{aligned} \tag{3}$$

where P_t is the transmit power, $P_{j,t}$ is the jamming power, $\beta_t \in [0,1]$ is the power reflection coefficient, and $n_t \sim \mathcal{CN}(0, \sigma^2)$ is additive white Gaussian noise.

Under independent phase variations and symbol averaging, the direct and backscattered components are combined in the power domain. The resulting signal-to-interference-plus-noise ratio (SINR) is:

$$\Gamma_t = \frac{P_t g_{tr,t} + \beta_t P_t g_{tb,t} g_{br,t}}{P_{j,t} g_{jr,t} + \sigma^2}. \tag{4}$$

The instantaneous spectral efficiency and throughput are respectively defined as:

$$R_t = \log_2(1 + \Gamma_t) \quad [\mathrm{bit/s/Hz}], \tag{5}$$

$$C_t = B \log_2(1 + \Gamma_t) \quad [\mathrm{bit/s}], \tag{6}$$

where B is the system bandwidth. Throughout the simulation section, results reported in $\mathrm{bit/s/Hz}$ are referred to as spectral efficiency rather than absolute throughput.

The energy-efficiency metric is computed as:

$$\eta_{\mathrm{EE},t} = \frac{R_t}{P_t} \quad [\mathrm{bit/J/Hz}], \tag{7}$$

or equivalently $B\eta_{\mathrm{EE},t}$ in $\mathrm{bit/J}$. This distinction is maintained consistently in the revised figures and tables.

B. Dynamic Jamming Model

The jammer is allowed to vary both its transmission power and attack mode over time. Its operating mode belongs to:

$$\mathcal{M}_j = \{\text{random}, \text{reactive}\}, \tag{8}$$

while its transmit power is selected from a finite set:

$$\mathcal{P}_j = \left\{ P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(N_j)} \right\}. \tag{9}$$

The jammer state at slot t is therefore represented by

$$s_t^{(j)} = \left(m_t^{(j)}, P_{j,t} \right) \in \mathcal{M}_j \times \mathcal{P}_j. \tag{10}$$

Its temporal evolution follows a discrete-time Markov process:

$$\Pr\left(s_{t+1}^{(j)} = q \,\middle|\, s_t^{(j)} = p \right) = \mathbf{T}_{p,q}^{(j)}, \tag{11}$$

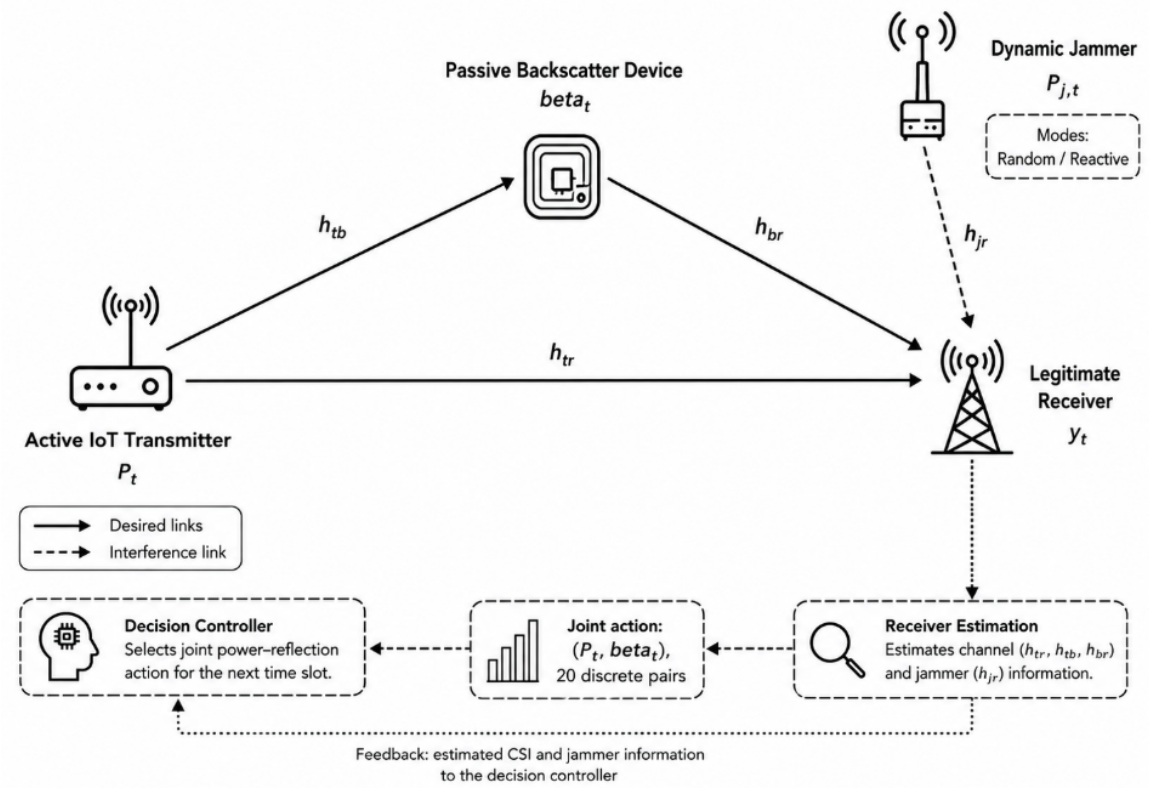


Figure 1. Backscatter-enabled 6G IoT system with joint power–reflection control under dynamic random and reactive jamming

where $\mathbf{T}^{(j)}$ is the jammer-state transition matrix. Random jamming changes its power independently of the legitimate transmission decision, whereas reactive jamming increases or activates its power after detecting legitimate RF activity. The main jammer settings are summarized in Table II.

At the beginning of each slot, the receiver estimates the channel power gains, jamming power, and attack mode using pilot measurements and spectrum sensing. These estimates are fed back to the decision controller. Under this assumption, the augmented control state can be modeled as an MDP. If the jammer mode or channel state cannot be reliably estimated, the same formulation becomes a partially observable MDP, which is outside the scope of this work.

C. Optimization Objective and Discrete Action Construction

The physical control variables satisfy

$$0 < P_t \le P_{\max}, \tag{12}$$

$$0 < \beta_t \le 1. \tag{13}$$

Although these constraints define continuous feasible intervals, D3QN requires a finite action space. Therefore, the transmit-power and reflection-coefficient sets are discretized as:

$$\mathcal{P} = \{0.2,\ 0.5,\ 0.8,\ 1.0\}\ \mathrm{W}, \tag{14}$$

$$\mathcal{B} = \{0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0\}. \tag{15}$$

The joint action space is their Cartesian product:

$$\mathcal{A} = \mathcal{P} \times \mathcal{B}, \qquad |\mathcal{A}| = N_P N_\beta = 4 \times 5 = 20. \tag{16}$$

Each D3QN output consequently corresponds to one feasible transmit-power/reflection-coefficient pair. The reflection coefficients are uniformly quantized, whereas the power levels form a compact nonuniform grid spanning low-, medium-, and high-power operating regions. This construction limits the network output dimension while retaining sufficient control resolution.

A finer action grid may reduce quantization loss but increases the number of output actions,

exploration time, and sample complexity. Conversely, an excessively coarse grid reduces training complexity but may prevent the agent from identifying an effective operating point. The adopted 20-action configuration therefore provides a balance between control resolution and computational cost.

The objective is to determine a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected discounted return:

$$\max_{\pi} J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1} \rho^t r_t\right], \tag{17}$$

where T is the episode horizon, r_t is the instantaneous reward, and $\rho \in [0, 1)$ is the RL discount factor. The notation ρ is used for temporal discounting to avoid confusion with the SINR Γ_t.

The resulting problem is a stochastic sequential decision problem with nonlinear coupling between P_t, β_t, the direct link, and the backscattered link. Therefore, a closed-form slot-independent solution is generally unavailable under time-varying channels and dynamic jamming. No NP-hardness claim is made without a formal complexity reduction.

D. MDP Formulation

The anti-jamming control problem is formulated as the MDP

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}_s, r, \rho), \tag{18}$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the 20-element joint action space, $\mathcal{P}_s$ is the state-transition kernel, r is the reward function, and ρ is the discount factor.

1. State Space

The state observed by the controller at slot t is

$$\begin{aligned} s_t = \big[&\widetilde{g}_{tr,t}, \widetilde{g}_{tb,t}, \widetilde{g}_{br,t}, \widetilde{g}_{jr,t}, \widetilde{P}_{j,t}, \\ &\text{onehot}\left(\widehat{m}_t^{(j)}\right), \widetilde{P}_{t-1}, \beta_{t-1}\big], \end{aligned} \tag{19}$$

where the tilde denotes min–max normalization, and $\widehat{m}_t^{(j)}$ is the estimated jammer mode. In particular,

$$\widetilde{P}_{j,t} = \frac{P_{j,t}}{P_{j,\max}}, \qquad \widetilde{P}_{t-1} = \frac{P_{t-1}}{P_{\max}}. \tag{20}$$

Unlike formulations based only on instantaneous desired-link CSI, (19) contains the jammer–receiver gain, the estimated jamming power and mode, and the previous control action. These components provide information about adversarial dynamics and short-term decision dependence.

2. Action Space

At slot t, the agent selects

$$a_t = (P_t, \beta_t) \in \mathcal{A}. \tag{21}$$

The action is executed jointly rather than optimizing P_t and β_t in separate stages. This construction captures their coupled influence in (4) and avoids a sequential solution in which one variable is optimized while the other is held fixed.

3. Reward Function

To combine spectral efficiency, transmit-power consumption, and outage avoidance on a dimensionless scale, the reward is defined as:

$$r_t = w_R \frac{R_t}{R_{\text{ref}}} - w_P \frac{P_t}{P_{\max}} - w_O \mathbf{1}\left\{\Gamma_t < \Gamma_{\text{th}}\right\}, \tag{22}$$

where w_R, w_P, and w_O are nonnegative weights, R_{ref} is a reference spectral efficiency, Γ_{th} is the minimum acceptable SINR, and $\mathbf{1}\{\cdot\}$ is the indicator function.

The first term rewards reliable high-rate communication. The second term discourages unnecessarily high transmission power. The third term penalizes slots in which the SINR falls below the required threshold. Normalization prevents the reward from directly combining quantities with incompatible units and substantially different numerical scales. The reward weights and their selection rationale are provided in Table II.

4. State Transition

The augmented state evolves according to:

$$s_{t+1} \sim \mathcal{P}_s\left(s_{t+1} \mid s_t, a_t\right), \tag{23}$$

where the transition is determined by channel evolution, jammer-state transitions, and the current control action. Although the observed interference sequence appears non-stationary when the jammer dynamics are omitted, augmenting the state with the estimated jammer power and mode allows the interaction to be approximated by a stationary Markov model.

5. Methodological Distinction

The methodological contribution does not lie in modifying the standard D3QN architecture itself. Instead, it lies in constructing a jammer-aware MDP with a coupled discrete power–reflection action space for backscatter-assisted anti-jamming communication. In contrast to single-variable DQN-based power control, the proposed policy directly maps each observed channel and jammer state to a feasible (P_t, β_t) pair. The reward additionally accounts for energy use and low-SINR events. These distinctions are summarized in Table I.

E. D3QN-Based Solution

A Dueling Double Deep Q-Network is employed to estimate the optimal action-value function over the finite joint action space. D3QN combines Double Q-learning with a dueling network structure to reduce overestimation bias and improve state-value estimation.

1. Double Q-Learning

Let Q_θ and Q_{θ^-} denote the online and target networks, respectively. For a transition $(s_i, a_i, r_i, s_i', d_i)$, the Double-Q target is

$$y_i = r_i + \rho(1 - d_i) Q_{\theta^-}\left(s_i', \arg\max_{a \in \mathcal{A}} Q_\theta(s_i', a)\right), \tag{24}$$

where $d_i = 1$ for a terminal transition and $d_i = 0$ otherwise. The online network selects the next action, whereas the target network evaluates it. This separation reduces the positive estimation bias that can occur when a single network performs both operations.

The online network is optimized using the mean-squared temporal difference loss

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left[y_i - Q_\theta(s_i, a_i)\right]^2, \tag{25}$$

where $\mathcal{M}$ denotes a sampled mini-batch.

2. Dueling Architecture

The action-value function is decomposed into a state-value stream $V_\theta(s)$ and an advantage stream $A_\theta(s, a)$:

$$Q_\theta(s, a) = V_\theta(s) + A_\theta(s, a) - \frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} A_\theta(s, a'). \tag{26}$$

The value stream estimates the quality of the current channel and jamming state independently of a specific action. The advantage stream then distinguishes among the 20 power–reflection combinations. This decomposition is useful when several actions have similar performance under the same propagation and interference conditions.

3. Rationale for Selecting D3QN

PPO can operate with either categorical or continuous policies, whereas TD3 and SAC naturally generate continuous actions. However, practical RF transmitters and backscatter devices commonly provide a finite number of power and impedance states. D3QN therefore maps directly to the considered hardware-constrained action set without requiring continuous-action projection during deployment.

The D3QN architecture is not claimed to be universally superior to actor–critic methods. PPO, TD3, and SAC are included as additional baselines under identical state information, jammer traces, reward definition, interaction budget, and statistical protocol. For fair hardware-constrained evaluation, continuous outputs generated by TD3 and SAC are clipped and mapped to the nearest feasible (P_t, β_t) pair before execution.

F. Training Procedure

Algorithm 1 summarizes the training procedure. An experience-replay buffer stores state transitions and breaks temporal correlation

among consecutive samples. A target network is periodically synchronized with the online network. Exploration follows an exponentially decaying ϵ-greedy strategy with a nonzero lower bound.

The target network is updated every C interaction steps:

$$\theta^{-} \leftarrow \theta. \tag{27}$$

The exploration probability is:

$$\epsilon_u = \max\left\{\epsilon_{\min}, \epsilon_0 \exp(-ku)\right\}, \tag{28}$$

where u is the global environment-interaction index. Using a global counter prevents the exploration schedule from being unintentionally reset at the beginning of each episode.

G. Computational Complexity and Deployment

Let d_s be the state dimension, $N_A = N_P N_\beta$ the number of joint actions, and $n_1, \ldots, n_L$ the numbers of neurons in the hidden layers. The approximate computational complexity of one D3QN forward pass is:

$$\mathcal{O}\left(d_s n_1 + \sum_{\ell=2}^{L} n_{\ell-1} n_\ell + n_L(N_A + 1)\right), \tag{29}$$

where the N_A outputs correspond to the advantage stream and the additional output corresponds to the value stream. Therefore, D3QN has the same asymptotic inference order as a conventional DQN, with only a small additional value-stream overhead.

For a mini-batch of size N_B, one training update has complexity

$$\mathcal{O}\left(N_B C_{\text{net}}\right), \tag{30}$$

where C_{net} denotes the forward/backward complexity of the neural network. The memory requirement is approximately

$$\mathcal{O}\left(|\theta| + N_{\mathcal{D}} d_{\text{tr}}\right), \tag{31}$$

where $|\theta|$ is the number of trainable parameters and d_{tr} is the storage dimension of one transition.

Training can be performed offline at a cloud or mobile-edge server. During deployment, only a forward pass and an $\arg\max$ operation are required at the active transmitter, gateway, or edge controller. The passive backscatter device does not execute the neural network; it only applies the selected impedance state associated with β_t.

Practical deployment requires channel estimation, jammer-power sensing, feedback of the selected action, and synchronization within each control slot. The finite action design is compatible with transmitters providing discrete RF power levels and backscatter tags providing a finite set of impedance states.

Representative future applications in Vietnam include low-power environmental monitoring in remote mountainous regions, smart agriculture and aquaculture, industrial-zone sensing, port logistics, and interference-resilient IoT monitoring. These are prospective 6G IoT scenarios rather than claims of current commercial 6G deployment.

H. Statistical Evaluation Protocol

To reduce dependence on random initialization, channel realizations, and exploration trajectories, each algorithm is evaluated over $N_s = 5$ independent random seeds. All competing methods use the same seed-specific channel and jammer traces, episode horizon, interaction budget, state information, and reward definition.

For a metric X, the sample mean and standard deviation are:

$$\overline{X} = \frac{1}{N_s} \sum_{i=1}^{N_s} X_i, \tag{32}$$

$$s_X = \sqrt{\frac{1}{N_s - 1} \sum_{i=1}^{N_s} \left(X_i - \overline{X}\right)^2}. \tag{33}$$

The 95% confidence interval is reported using the Student-t distribution:

$$\overline{X} \pm t_{0.975, N_s - 1} \frac{s_X}{\sqrt{N_s}}. \tag{34}$$

Algorithm 1: D3QN-Based Joint Power and Reflection Control

Input: Power set $\mathcal{P}$, reflection set $\mathcal{B}$, replay capacity $N_{\mathcal{D}}$, mini-batch size N_B, discount factor ρ, target-update interval C, and exploration parameters $\epsilon_0, \epsilon_{\min}, k$

Output: Learned policy $\pi^*(s) = \arg\max_{a\in\mathcal{A}} Q_\theta(s,a)$

Construct $\mathcal{A} = \mathcal{P} \times \mathcal{B}$ with $|\mathcal{A}| = 20$;
Initialize replay buffer $\mathcal{D}$, online network Q_θ, and target network Q_{θ^-} with $\theta^- \leftarrow \theta$;
Set global interaction counter $u \leftarrow 0$;
for *episode* $e = 1, \ldots, E$ **do**
  Initialize the channel and jammer states;
  Observe the normalized initial state s_0;
  for *time slot* $t = 0, \ldots, T-1$ **do**
    Set $\epsilon_u = \max\{\epsilon_{\min}, \epsilon_0 \exp(-ku)\}$;
    if $\mathrm{rand}() < \epsilon_u$ **then**
      Select a_t uniformly from $\mathcal{A}$;
    else
      Select $a_t = \arg\max_{a\in\mathcal{A}} Q_\theta(s_t, a)$;
    Execute $a_t = (P_t, \beta_t)$, and observe r_t, s_{t+1}, and terminal flag d_t;
    Store $(s_t, a_t, r_t, s_{t+1}, d_t)$ in $\mathcal{D}$;
    if $|\mathcal{D}| \geq N_B$ **then**
      Sample mini-batch $\mathcal{M} = \{(s_i, a_i, r_i, s'_i, d_i)\}_{i=1}^{N_B}$;
      Compute $a_i^* = \arg\max_{a\in\mathcal{A}} Q_\theta(s'_i, a)$;
      Compute $y_i = r_i + \rho(1-d_i) Q_{\theta^-}(s'_i, a_i^*)$;
      Update θ by minimizing $\frac{1}{N_B}\sum_{i=1}^{N_B} [y_i - Q_\theta(s_i, a_i)]^2$;
    if $\mathrm{mod}(u, C) = 0$ **then**
      Update target network: $\theta^- \leftarrow \theta$;
    Set $s_t \leftarrow s_{t+1}$ and $u \leftarrow u + 1$;
    if $d_t = 1$ **then**
      break;

Performance improvements are calculated as:

$$G_X = \frac{X_{\mathrm{D3QN}} - X_{\mathrm{baseline}}}{X_{\mathrm{baseline}}} \times 100\%. \qquad (35)$$

Each reported gain explicitly identifies the comparison baseline, jamming condition, and whether the value is point-wise or averaged over multiple jamming levels. The revised convergence, spectral efficiency, energy-efficiency, robustness, and ablation results are therefore presented as mean values with standard deviations or 95% confidence intervals.

III. Performance and Evaluation

A. Simulation Setup

The proposed framework is evaluated in a backscatter-enabled IoT communication environment subject to dynamic random and reactive jamming. The simulation configuration follows the system, jammer, and MDP models presented in Section II. All physical-layer, learning, action-space, implementation, and statistical-evaluation parameters are summarized in Table II.

Channel, Jammer, and Action Settings: All desired and interfering links follow independent distance-dependent Rayleigh fading, and the receiver noise is modeled as complex Gaussian noise. Spectral efficiency and energy efficiency are evaluated using the definitions in Section II. The jammer evolves according to a finite-state Markov model and switches between random and reactive modes with time-varying power. Perfect estimation of the channel gains, jammer power, and attack mode is assumed to isolate the control

TABLE II. SIMULATION AND LEARNING PARAMETERS

Parameter	Value	Parameter	Value
Channel model	Rayleigh fading	Noise power, σ^2	-90 dBm
Bandwidth, B	1 MHz	Path-loss exponent, ν	2.5
Maximum power, $P_{\max}$	1 W	Jamming power, P_j	0–30 dBm
Jammer modes	Random/reactive	Jammer evolution	Finite-state Markov process
Power set, $\mathcal{P}$	$\{0.2, 0.5, 0.8, 1.0\}$ W	Reflection set, $\mathcal{B}$	$\{0.2, 0.4, 0.6, 0.8, 1.0\}$
Joint action space	$\lvert\mathcal{A}\rvert = 20$	SINR threshold, Γ_{th}	5 dB
Reward weights	$w_R = 1,\ w_P = 0.01,\ w_O = 1$	Discount factor, ρ	0.95
Learning rate	10^{-3}	Replay-buffer size	10^5
Mini-batch size	64	Target-update interval	100 steps
Training episodes	500	Exploration parameters	$\epsilon_0 = 1,\ k = 0.001$
Baselines	DQN, DDQN, PPO, TD3, SAC, Random, Fixed control	Independent seeds	5
Reported statistics	Mean, SD, and 95% CI	Confidence interval	Student-t

Note: All learning methods use the same state, reward, interaction budget, channel realizations, jammer trajectories, and feasible power–reflection actions. TD3 and SAC outputs are mapped to the nearest feasible action pair.

performance. The controller selects one of 20 feasible transmit-power/reflection-coefficient pairs listed in Table II. Continuous actions produced by TD3 and SAC are clipped and mapped to the nearest feasible pair.

Learning and Evaluation Protocol: All learning methods use the same normalized reward, observable state, episode horizon, interaction budget, channel realizations, and jammer trajectories. D3QN is trained using experience replay, a target network, and decaying ϵ-greedy exploration. Convergence is determined using a common moving-average reward criterion rather than visual inspection. Each experiment is repeated over five independent paired seeds, and the results are reported as the mean, standard deviation, and 95% Student-t confidence interval. Every percentage gain explicitly states its baseline, jamming condition, and whether it is point-wise or averaged. Implementation and reproducibility settings are summarized in Table II.

B. Baseline Schemes

The proposed D3QN is compared with the following methods under identical state, reward, action, training-budget, and evaluation conditions:

- **DQN:** Standard Q-learning over the same 20 joint actions, without Double-Q estimation or dueling decomposition.
- **DDQN:** Double Q-learning over the same action space, without separate value and advantage streams.
- **PPO:** A categorical actor–critic policy over the 20 feasible joint actions.
- **TD3:** A continuous actor with twin critics, delayed policy updates, and target-policy smoothing.
- **SAC:** An entropy-regularized stochastic actor with twin critics.
- **Random:** Uniform selection from the feasible joint action set.
- **Fixed Control:** A constant transmit-power and reflection-coefficient pair independent of channel and jammer conditions.

Ablation variants are additionally used to quantify the individual effects of Double Q-learning, the dueling architecture, reflection adaptation, and joint power–reflection control.

C. Results and Analysis

The simulations were implemented in Python using PyTorch and Gymnasium, with a custom backscatter anti-jamming environment, Stable-Baselines3 baselines, and multi-seed statistical evaluation using NumPy, SciPy, and Matplotlib.

1. Learning Convergence and Stability

Figure 2 compares the 20-episode moving-average reward of DQN, DDQN, PPO, TD3, SAC, and D3QN over five independent seeds. Under the predefined convergence criterion with a 20-episode window, a reward-change tolerance of 0.01, and a 20-episode persistence requirement, none of the considered methods satisfies the criterion before the 500-episode training horizon. Accordingly, the recorded convergence episode is 500 for D3QN, DQN, DDQN, PPO, TD3, and SAC.

Over the final 20 training episodes, D3QN achieves an average reward of 248.395, corresponding to improvements of 0.576% and 0.666% over DQN and DDQN, whose corresponding rewards are 246.972 and 246.752, respectively. The 95% confidence interval of D3QN is ±17.304, which is narrower than that of DQN (±25.402) but wider than that of DDQN (±14.792).

SAC obtains the highest final average reward of 266.826, whereas D3QN provides a modest reward improvement over the value-based DQN and DDQN baselines.

2. Performance under Random Jamming

As shown in Figure 3a, the spectral efficiency decreases as the jamming power increases. At $P_j = 30$ dBm, D3QN achieves 1.748 ± 0.182 bit/s/Hz, compared with 1.458 ± 0.447 bit/s/Hz for DQN and

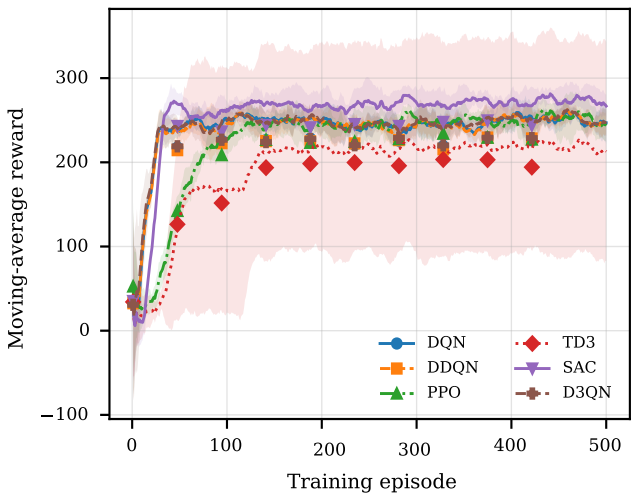


Figure 2. Convergence of the considered DRL schemes

1.510 ± 0.163 bit/s/Hz for DDQN, which is the strongest competing learning-based baseline under this condition. These results correspond to spectral-efficiency gains of 19.83% over DQN and 15.70% over DDQN.

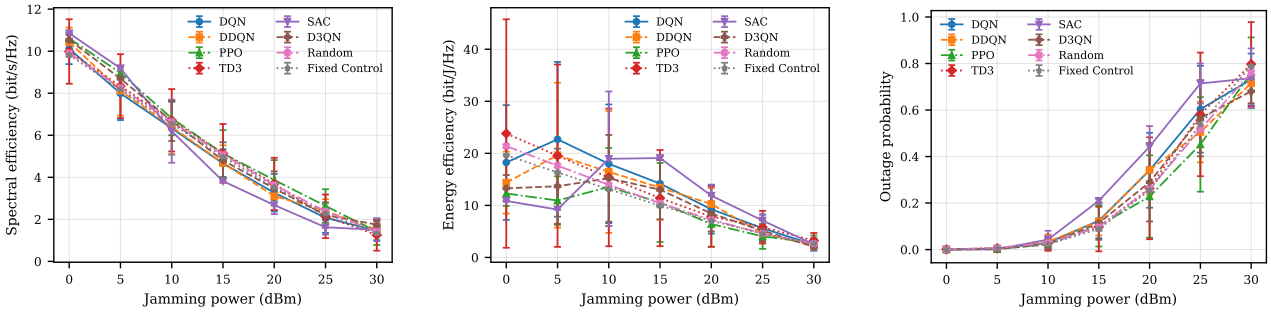
Figure 3b reveals a rate–energy trade-off. D3QN obtains an energy efficiency of 2.043 ± 0.127 bit/J/Hz at $P_j = 30$ dBm, whereas DQN and TD3 achieve 2.447 ± 1.109 bit/J/Hz and 3.217 ± 1.487 bit/J/Hz, respectively. Consequently, D3QN is 16.52% lower than DQN and 36.49% lower than TD3, the strongest actor–critic baseline in terms of energy efficiency. This result indicates that, under strong random jamming, the learned D3QN policy prioritizes spectral-efficiency preservation over energy-efficiency maximization.

As reported in Figure 3c, D3QN achieves an outage probability of 0.680 ± 0.051 at 30 dBm, compared with 0.732 ± 0.111 for DQN and 0.715 ± 0.034 for DDQN. The corresponding outage reductions are 6.99% relative to DQN and 4.88% relative to DDQN. These results are consistent with the spectral-efficiency advantage observed in Figure 3a.

3. Performance under Reactive Jamming

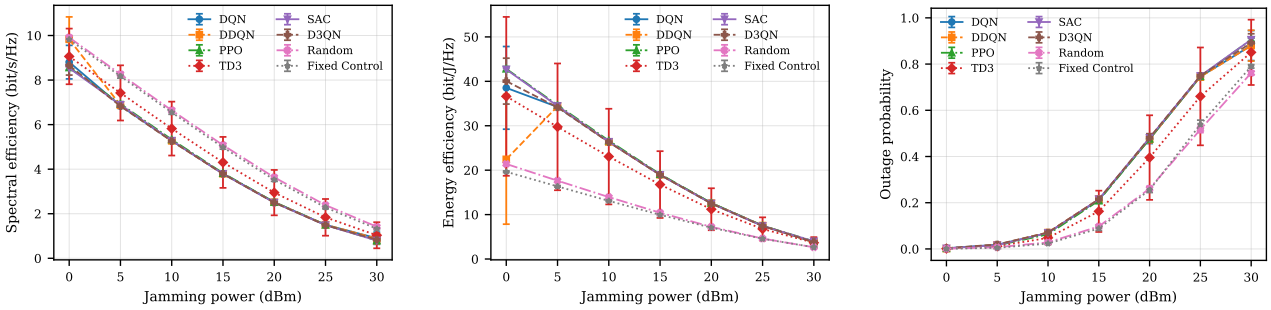
Reactive jamming causes a substantial degradation in the D3QN spectral efficiency. As shown in Figure 4a, D3QN achieves 0.836 ± 0.116 bit/s/Hz at $P_j = 30$ dBm, whereas DQN and TD3 obtain 0.900 ± 0.274 bit/s/Hz and 1.034 ± 0.585 bit/s/Hz, respectively. Thus, the D3QN spectral efficiency is 7.03% lower than DQN and 19.07% lower than TD3, which is the strongest learning-based baseline under this condition. Random and fixed control also achieve 1.417 ± 0.055 bit/s/Hz and 1.326 ± 0.046 bit/s/Hz, respectively, indicating that the current reactive-mode policy requires further reward or state-representation refinement.

In Figure 4b, D3QN achieves 3.833 ± 0.379 bit/J/Hz, compared with 3.739 ± 0.802 bit/J/Hz for DQN. This represents a 2.54% improvement over DQN. However, PPO obtains the highest actor–critic energy efficiency of 3.985 ± 0.103 bit/J/Hz; therefore, D3QN remains 3.81% below PPO. The



(a) Spectral efficiency (b) Energy efficiency (c) Outage probability

Figure 3. Performance under random jamming. Error bars denote the 95% confidence intervals over five independent seeds



(a) Spectral efficiency (b) Energy efficiency (c) Outage probability

Figure 4. Performance under reactive jamming. Error bars denote the 95% confidence intervals over five independent seeds

higher energy efficiency of D3QN relative to DQN is associated with its more conservative transmit-power selection under reactive jamming.

Figure 4c shows that D3QN yields an outage probability of 0.892 ± 0.039, while DQN and TD3 obtain 0.880 ± 0.065 and 0.851 ± 0.141, respectively. Hence, D3QN increases outage by 1.32% relative to DQN and by 4.81% relative to TD3. The reactive-jamming results therefore demonstrate an energy–reliability trade-off: the conservative D3QN policy slightly improves energy efficiency over DQN but sacrifices spectral efficiency and outage performance.

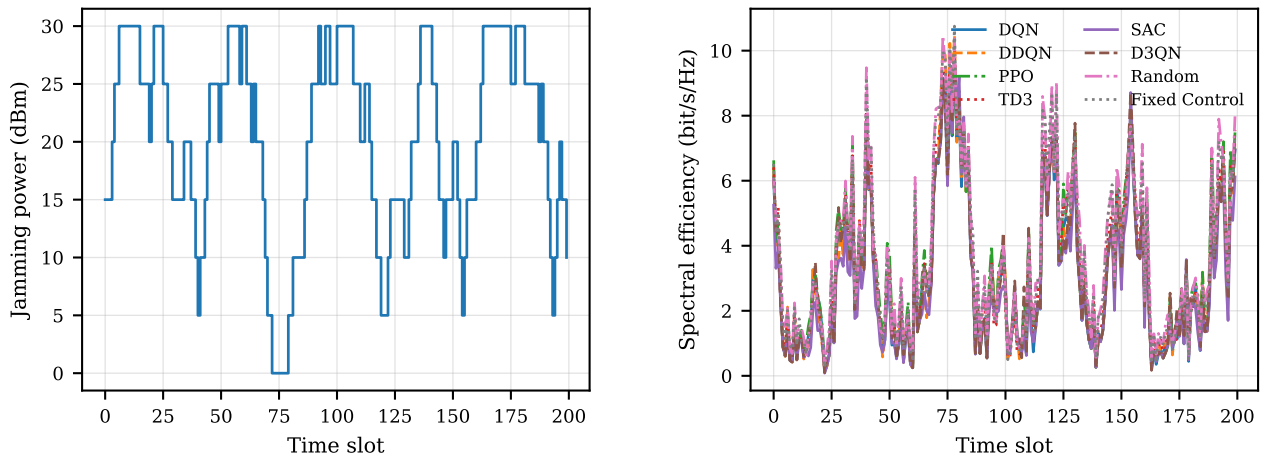
4. Robustness under Dynamic Jamming

Figure 5a presents the common jammer sequence applied to all schemes. The evaluation spans 200 time slots, during which the jamming power varies between 0 and 30 dBm. The trace contains 34 transitions between random and reactive jamming modes, including 109 random-jamming slots and 91 reactive-jamming slots.

Under the same jammer sequence, Figure 5b shows that D3QN achieves an average spectral efficiency of 3.212 ± 0.287 $\mathrm{bit/s/Hz}$, compared with 3.145 ± 0.268 $\mathrm{bit/s/Hz}$ for DQN, where the intervals represent the 95% confidence intervals over five independent seeds. This corresponds to a 2.12% improvement over DQN.

The temporal standard deviations of the seed-averaged spectral-efficiency traces are 2.268 for D3QN and 2.274 for DQN. Hence, D3QN reduces temporal variability by only 0.26%. The two schemes therefore exhibit comparable temporal fluctuations, while D3QN provides a modest improvement in average spectral efficiency under the considered dynamic-jamming sequence.

Table III summarizes the quantitative results under the strongest random and reactive jamming conditions. The table highlights that D3QN provides clear spectral-efficiency and outage advantages under random jamming, while exhibiting an energy–reliability trade-off under reactive jamming.



(a) Dynamic-jammer power trace

(b) Spectral-efficiency response

Figure 5. Robustness under a common dynamic-jamming sequence.

TABLE III. QUANTITATIVE PERFORMANCE UNDER STRONG JAMMING AT $P_j = 30$ dBm

Method	Random Jamming			Reactive Jamming		
	SE	EE	Outage	SE	EE	Outage
DQN	1.458 ± 0.447	2.447 ± 1.109	0.732 ± 0.111	0.900 ± 0.274	3.739 ± 0.802	0.880 ± 0.065
DDQN	1.510 ± 0.163	2.292 ± 0.435	0.715 ± 0.034	0.882 ± 0.231	3.741 ± 0.591	0.880 ± 0.067
PPO	1.409 ± 0.634	2.824 ± 1.390	0.760 ± 0.152	0.797 ± 0.021	$\mathbf{3.985 \pm 0.103}$	0.906 ± 0.006
TD3	1.255 ± 0.745	$\mathbf{3.217 \pm 1.487}$	0.797 ± 0.181	1.034 ± 0.585	3.668 ± 1.254	0.851 ± 0.141
SAC	1.503 ± 0.542	2.512 ± 1.215	0.737 ± 0.128	0.796 ± 0.042	3.982 ± 0.212	0.906 ± 0.009
Random	1.417 ± 0.055	2.638 ± 0.089	0.761 ± 0.009	$\mathbf{1.417 \pm 0.055}$	2.638 ± 0.089	$\mathbf{0.761 \pm 0.009}$
Fixed Control	1.326 ± 0.046	2.653 ± 0.091	0.790 ± 0.008	1.326 ± 0.046	2.653 ± 0.091	0.790 ± 0.008
D3QN	$\mathbf{1.748 \pm 0.182}$	2.043 ± 0.127	$\mathbf{0.680 \pm 0.051}$	0.836 ± 0.116	3.833 ± 0.379	0.892 ± 0.039

SE: spectral efficiency in bit/s/Hz; EE: energy efficiency in bit/J/Hz. Values are reported as the mean ± 95% Student-t confidence interval over five independent seeds. Bold values indicate the best result in each column; lower outage probability is preferred.

5. *Learned Joint Power–Reflection Adaptation*

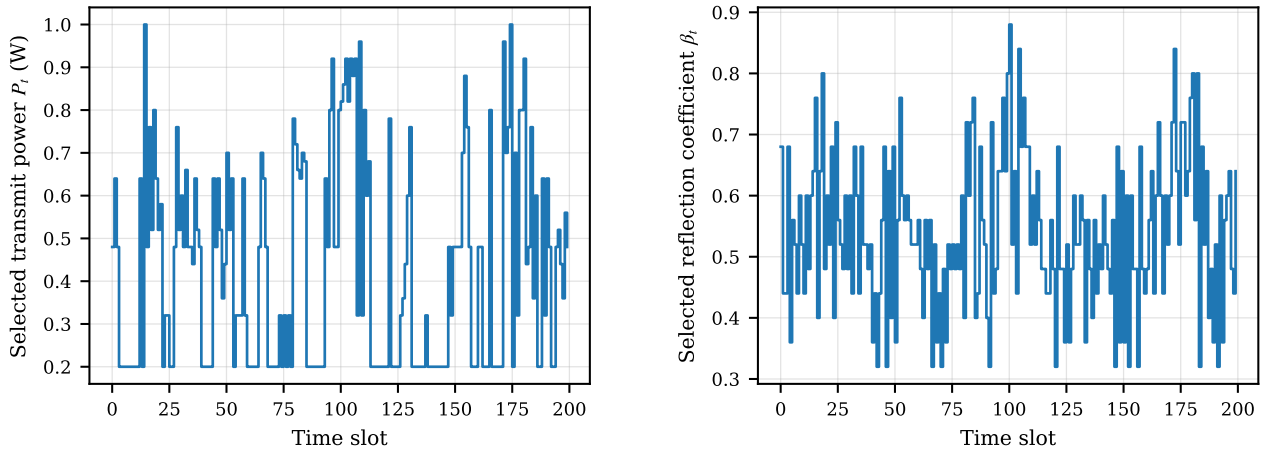
Figure 6a shows that D3QN selects different transmit-power levels instead of continuously operating at $P_{\max} = 1$ W. Across the five seeds and 200 time slots, the average selected transmit power is 0.427 W. The maximum-power action is selected in only 17.6% of the evaluated slots, whereas the lowest power level of 0.2 W is selected in 67.2% of the slots. This result indicates that the learned policy generally avoids persistent maximum-power transmission.

As shown in Figure 6b, D3QN also adapts the reflection coefficient across the complete feasible set. The average selected reflection coefficient is 0.541, while high-reflection actions satisfying $\beta_t \geq 0.8$ account for 31.2% of the evaluated slots. In particular, $\beta_t = 0.8$ and $\beta_t = 1.0$ are selected in 13.3% and 17.9% of the slots, respectively.

The combined results in Figs. 6a and 6b show that the learned D3QN policy does not collapse to a fixed-power or fixed-reflection strategy. Instead, it performs joint adaptation of P_t and β_t in response to the time-varying channel and jamming states.

6. *Ablation Analysis*

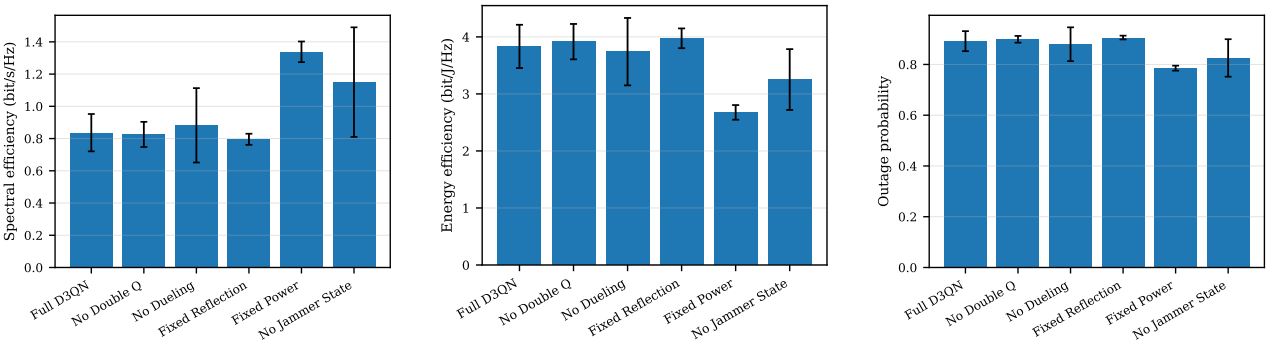
Figure 7a shows that Full D3QN achieves a spectral efficiency of 0.836 ± 0.116 bit/s/Hz. Removing Double Q-learning reduces this value



(a) Selected transmit power

(b) Selected reflection coefficient

Figure 6. Joint actions learned by D3QN



(a) Spectral efficiency

(b) Energy efficiency

(c) Outage probability

Figure 7. Ablation analysis under reactive jamming at $P_j = 30$ dBm. Error bars denote the 95% confidence intervals over five independent seeds

to 0.826 ± 0.078 bit/s/Hz, corresponding to a 1.30% degradation. Fixing the reflection coefficient yields 0.796 ± 0.034 bit/s/Hz, which is 4.89% lower than Full D3QN. In contrast, the No Dueling, Fixed Power, and No Jammer State variants obtain 0.882 ± 0.231, 1.338 ± 0.064, and 1.150 ± 0.340 bit/s/Hz, respectively. These values are 5.47%, 59.97%, and 37.49% higher than Full D3QN, indicating that the complete policy does not maximize spectral efficiency alone.

As shown in Figure 7b, Full D3QN achieves an energy efficiency of 3.833 ± 0.379 bit/J/Hz. The No Dueling, Fixed Power, and No Jammer State variants reduce this value to 3.741 ± 0.591, 2.676 ± 0.128, and 3.252 ± 0.533 bit/J/Hz, corresponding to reductions of 2.40%, 30.18%, and 15.15%, respectively. By contrast, the No Double Q and Fixed Reflection variants achieve 3.917 ± 0.310 and 3.975 ± 0.173 bit/J/Hz, which are 2.18% and 3.70% higher than Full D3QN. These results reveal that the proposed architecture balances rate, reliability, and power consumption rather than optimizing a single metric independently.

Figure 7c reports an outage probability of 0.892 ± 0.039 for Full D3QN. Removing Double Q-learning and fixing the reflection coefficient increase outage to 0.899 ± 0.013 and 0.906 ± 0.007, corresponding to relative increases of 0.82% and 1.60%, respectively. In contrast, the No Dueling, Fixed Power, and No Jammer State variants reduce outage to 0.880 ± 0.067, 0.785 ± 0.010, and 0.826 ± 0.074, representing reductions of 1.36%, 11.93%, and 7.42%, respectively.

Overall, the ablation results expose a clear

rate–energy–reliability trade-off. Fixed Power and No Jammer State improve spectral efficiency and outage performance, but reduce energy efficiency substantially. Conversely, Full D3QN maintains a more conservative power-control policy and achieves a more balanced operating point across the three metrics.

Table IV summarizes the ablation results under reactive jamming at $P_j = 30$ dBm. Removing Double Q-learning or fixing the reflection coefficient degrades spectral efficiency and increases outage, whereas the Fixed Power and No Jammer State variants improve rate and reliability at the expense of energy efficiency. These results confirm that the complete D3QN framework provides a balanced rate–energy–reliability operating point rather than maximizing a single metric.

7. Action-Space Granularity

Figure 8a compares the coarse, proposed, and fine action grids containing 9, 20, and 49 joint power–reflection actions, respectively. The coarse grid achieves a spectral efficiency of 0.803 ± 0.047 bit/s/Hz, whereas the proposed 20-action grid obtains 0.836 ± 0.116 bit/s/Hz. This corresponds to a 4.21% improvement over the coarse configuration.

The fine 49-action grid achieves the highest mean spectral efficiency of 0.964 ± 0.185 bit/s/Hz. The proposed 20-action grid is therefore 13.25% below the fine-grid result, or equivalently retains approximately 86.75% of its mean spectral efficiency. However, the fine-grid result also exhibits the widest confidence interval, indicating greater seed-dependent variability.

As shown in Figure 8b, none of the three action-space configurations satisfies the predefined convergence criterion before the 500-episode training horizon. Consequently, the recorded convergence episode is 500 for the 9-, 20-, and 49-action grids across all five seeds. The current results therefore do not support a claim that the proposed 20-action grid converges faster than the finer configuration.

Overall, the 20-action design provides a moderate spectral-efficiency improvement over the coarse grid while using 59.18% fewer output actions than the 49-action configuration. Nevertheless, the fine grid achieves the highest mean spectral efficiency, and additional training or a less restrictive convergence criterion is required to determine whether the 20-action grid offers a measurable convergence advantage.

8. Discussion

The results demonstrate that D3QN provides an effective joint power–reflection control mechanism, particularly under random and time-varying jamming. At $P_j = 30$ dBm, it improves spectral efficiency by 19.83% and reduces outage by 6.99% relative to DQN under random jamming. Under the common dynamic trace, D3QN achieves a 2.12% spectral-efficiency gain while using an average transmit power of only 0.427 W; the maximum-power action is selected in 17.6% of the slots. The 20-action design also improves spectral efficiency by 4.21% over the 9-action grid while using 59.18% fewer outputs than the 49-action grid. Nevertheless, reactive-jamming results reveal an energy–reliability trade-off: D3QN improves energy efficiency over DQN by 2.54%, but exhibits lower spectral efficiency and slightly higher outage. Overall, the proposed scheme offers balanced rate, energy, and implementation complexity rather than optimizing a single metric.

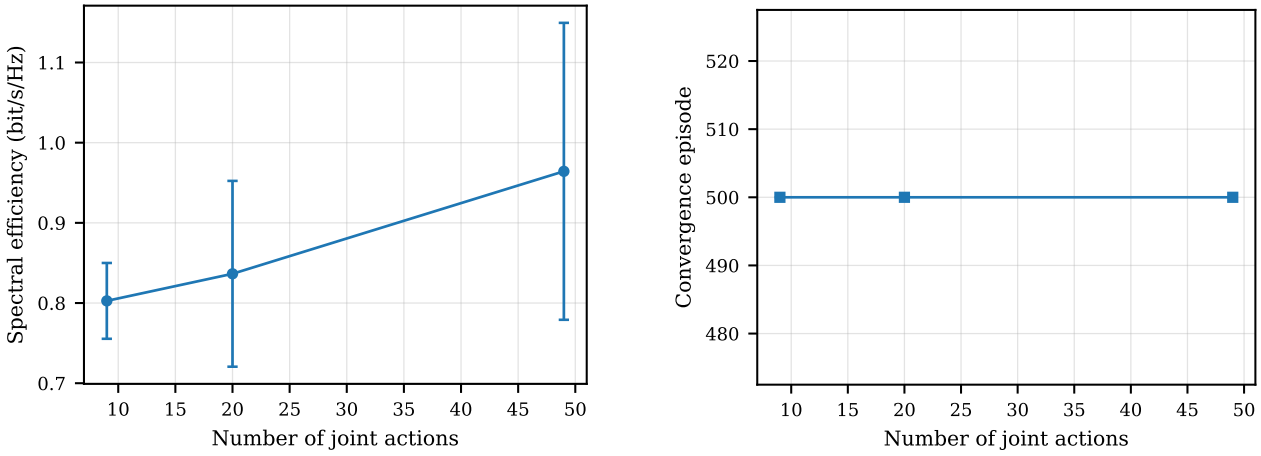
IV. Conclusion

This paper developed a D3QN-based framework for joint transmit-power and reflection-coefficient control in backscatter-enabled 6G IoT networks subject to random and reactive jamming. The anti-jamming problem was formulated as a jammer-aware MDP with a 20-action power–reflection space and a reward that jointly accounts for spectral efficiency, power consumption, and low-SINR outages.

The results show that the proposed method is most effective under random and time-varying jamming. At $P_j = 30$ dBm, D3QN improves spectral efficiency by 19.83% and reduces outage by 6.99% relative to DQN under random jamming. Under the common dynamic-jammer trace, it provides a 2.12% spectral-efficiency gain over DQN while using an average transmit power of 0.427 W; the maximum-power action is

TABLE IV. ABLATION RESULTS UNDER REACTIVE JAMMING AT $P_j = 30$ dBm

Variant	SE	EE	Outage
Full D3QN	0.836 ± 0.116	3.833 ± 0.379	0.892 ± 0.039
No Double Q	0.826 ± 0.078	3.917 ± 0.310	0.899 ± 0.013
No Dueling	0.882 ± 0.231	3.741 ± 0.591	0.880 ± 0.067
Fixed Reflection	0.796 ± 0.034	3.975 ± 0.173	0.906 ± 0.007
Fixed Power	1.338 ± 0.064	2.676 ± 0.128	0.785 ± 0.010
No Jammer State	1.150 ± 0.340	3.252 ± 0.533	0.826 ± 0.074



(a) Spectral efficiency

(b) Convergence episode

Figure 8. Effect of action-space granularity under reactive jamming at $P_j = 30$ dBm. Error bars denote the 95%confidence intervals over five independent seeds

selected in only 17.6% of the slots. The 20-action design also improves spectral efficiency by 4.21% over the 9-action grid while using 59.18% fewer outputs than the 49-action configuration.

The results also reveal an important limitation. Under strong reactive jamming, D3QN improves energy efficiency over DQN by 2.54%, but does not outperform all baselines in spectral efficiency or outage. Accordingly, the proposed framework should be viewed as a balanced rate–energy–complexity solution rather than a universally dominant policy. Future work will refine the reward and jammer-state representation, incorporate imperfect sensing and partial observability, and investigate adaptive action discretization.

References

[1] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, and H. V. Poor, "6G Internet of Things: A comprehensive survey," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 359–383, 2022.

[2] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Enabling massive IoT toward 6G: A comprehensive survey," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 11 891–11 915, 2021.

[3] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road toward 6G," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1117–1174, 2022.

[4] L. H. Hiep and H. H. Ngo, "Adaptive multi-mode DRL framework for intelligent RIS-assisted anti-jamming in 6G wireless networks," *Journal of Communications Software and Systems*, vol. 22, no. 3, pp. 295–307, 2026.

[5] S. Basharat, S. A. Hassan, H. Pervaiz, A. Mahmood, Z. Ding, and M. Gidlund, "Exploring reconfigurable intelligent surfaces for 6G: State-of-the-art and the road ahead," *IET Communications*, vol. 16, no. 12, pp. 1353–1372, 2022.

[6] T. Jiang, Y. Zhang, W. Ma, M. Peng, Y. Peng, M. Feng, and G. Liu, "Backscatter communication meets practical battery-free Internet of Things: A survey and outlook," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 3, pp. 2021–2051, 2023.

[7] W. Wu, X. Wang, A. Hawbani, L. Yuan, and W. Gong, "A survey on ambient backscatter communications:

Principles, systems, applications, and challenges," *Computer Networks*, vol. 216, p. 109235, 2022.

[8] L. H. Hiep and H. H. Ngo, "AI-driven jamming detection and DRL-based anti-jamming for ambient backscatter communications in 6G networks," *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 42, no. 2, pp. 98–114, 2026.

[9] U. S. Toro, K. Wu, and V. C. M. Leung, "Backscatter wireless communications and sensing in green Internet of Things," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 1, pp. 37–55, 2022.

[10] L. H. Hiep, "Probabilistic analysis of anti-jamming frequency hopping in backscatter IoT networks," *Journal of Science & Technology on Information and Communications*, vol. 11, no. 1, pp. 19–33, 2026.

[11] Y.-C. Liang, Q. Zhang, J. Wang, R. Long, H. Zhou, and G. Yang, "Backscatter communication assisted by reconfigurable intelligent surfaces," *Proceedings of the IEEE*, vol. 110, no. 9, pp. 1339–1357, 2022.

[12] M. Sayem, M. Z. Chowdhury, S. R. Hasan, and Y. M. Jang, "Reconfigurable intelligent surface assisted BackCom: An overview, analysis, and future research directions," *ICT Express*, vol. 9, no. 5, pp. 927–940, 2023.

[13] L. H. Hiep and H. H. Ngo, "RIS-assisted anti-jamming backscatter communication using deep reinforcement learning," *Journal of Science and Technology on Information Security*, vol. 1, no. 27, pp. 70–86, 2026.

[14] M. Hua, Q. Wu, L. Yang, R. Schober, and H. V. Poor, "A novel wireless communication paradigm for intelligent reflecting surface based symbiotic radio systems," *IEEE Transactions on Signal Processing*, vol. 70, pp. 550–565, 2022.

[15] L. H. Hiep and L. X. Hieu, "An energy-efficient deep reinforcement learning-based anti-jamming strategy for next-generation wireless networks," *TNU Journal of Science and Technology*, vol. 231, no. 7, pp. 35–43, 2026.

[16] L. H. Hiep and H. H. Ngo, "D3QN-based joint power control and reflection adaptation for anti-jamming ambient backscatter communications with mobile UAV jammers," *International Journal of Robotics and Control Systems*, vol. 6, no. 5, 2026.

[17] L. Jia, N. Qi, Z. Su, F. Chu, S. Fang, K.-K. Wong, and C.-B. Chae, "Game theory and reinforcement learning for anti-jamming defense in wireless communications: Current research, challenges, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 3, pp. 1798–1838, 2025.

[18] L. T. H. Van, N. Q. Minh, P. V. Huong, and N. H. Minh, "A novel approach for 1D-CNN hyperparameter optimization in IoT attack detection using particle swarm optimization," *Journal of Science and Technology on Information Security*, vol. 1, no. 24, pp. 53–71, 2025.

[19] H. Pirayesh and H. Zeng, "Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 767–809, 2022.

[20] W. Li, J. Chen, X. Liu, X. Wang, Y. Li, D. Liu, and Y. Xu, "Intelligent dynamic spectrum anti-jamming communications: A deep reinforcement learning perspective," *IEEE Wireless Communications*, vol. 29, no. 3, pp. 60–67, 2022.

[21] L. Jia, N. Qi, F. Chu, S. Fang, X. Wang, S. Ma, and S. Feng, "Game-theoretic learning anti-jamming approaches in wireless networks," *IEEE Communications Magazine*, vol. 60, no. 5, pp. 60–66, 2022.

[22] A. Gouissem, K. Abualsaud, E. Yaacoub, T. Khattab, and M. Guizani, "Accelerated IoT anti-jamming: A game theoretic power allocation strategy," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 607–10 620, 2022.

[23] L. H. Hiep, D. M. Tran, and N. N. Duong, "Adaptive reflection control for anti-jamming backscatter communication under dynamic interference," *Vinh University Journal of Science*, vol. 55, no. 2A, pp. 115–126, 2026.

[24] A. Pourranjbar, G. Kaddoum, A. Ferdowsi, and W. Saad, "Reinforcement learning for deceiving reactive jammers in wireless networks," *IEEE Transactions on Communications*, vol. 69, no. 6, pp. 3682–3697, 2021.

[25] X. Li, J. Chen, X. Ling, and T. Wu, "Deep reinforcement learning-based anti-jamming algorithm using dual action network," *IEEE Transactions on Wireless Communications*, vol. 22, no. 7, pp. 4625–4637, 2023.

[26] N. Van Huynh, D. T. Hoang, D. N. Nguyen, and E. Dutkiewicz, "DeepFake: Deep dueling-based deception strategy to defeat reactive jammers," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6898–6914, 2021.

[27] C. Liu, Y. Zhang, G. Niu, L. Jia, L. Xiao, and J. Luan, "Towards reinforcement learning in UAV relay for anti-jamming maritime communications," *Digital Communications and Networks*, vol. 9, no. 6, pp. 1477–1485, 2023.

[28] Q. Zhang, Y.-C. Liang, and H. V. Poor, "Reconfigurable intelligent surface assisted MIMO symbiotic radio networks," *IEEE Transactions on Communications*, vol. 69, no. 7, pp. 4832–4846, 2021.

[29] H. Yang, Y. Ye, K. Liang, and X. Chu, "Energy efficiency maximization for symbiotic radio networks with multiple backscatter devices," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1431–1444, 2021.

[30] L. Xiao, Y. Ding, J. Huang, S. Liu, Y. Tang, and H. Dai, "UAV anti-jamming video transmissions with QoE guarantee: A reinforcement learning-based approach," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5933–5947, 2021.

[31] L. H. Hiep, V. T. Hoang, and H. H. Ngo, "RLIoT: Reinforcement learning based network resource optimization using IoT sensor data," *International Journal of Computer Networks & Communications*, vol. 18, no. 4, pp. 1–16, 2026.

[32] J. Qi, H. Zhang, X. Qi, and M. Peng, "Deep reinforcement learning based hopping strategy for wideband anti-jamming wireless communications," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 3, pp. 3568–3579, 2024.

[33] Z. Lv, L. Xiao, Y. Du, G. Niu, C. Xing, and W. Xu, "Multi-agent reinforcement learning based UAV swarm communications against jamming," *IEEE Transactions on Wireless Communications*, vol. 22, no. 12, pp. 9063–9075, 2023.

[34] Z. Shao, H. Yang, L. Xiao, W. Su, Y. Chen, and Z. Xiong, "Deep reinforcement learning-based resource management for UAV-assisted mobile edge computing against jamming," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 13 358–13 374, 2024.

[35] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1226–1252, 2021.

[36] H. Ke, H. Wang, and H. Sun, "Deep reinforcement learning-based power control and bandwidth allocation policy for weighted cost minimization in wireless networks," *Applied Intelligence*, vol. 53, pp. 26 885–26 906, 2023.

[37] H. Yang, Z. Xiong, J. Zhao, D. Niyato, Q. Wu, H. V. Poor, and M. Tornatore, "Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1963–1974, 2021.

[38] Y. Liu, K. Xu, X. Xia, W. Xie, N. Ma, and J. Xu, "Joint power control and passive beamforming optimization in RIS-assisted anti-jamming communication," *Frontiers of Information Technology & Electronic Engineering*, vol. 24, no. 12, pp. 1791–1802, 2023.

[39] F. Yao, B. Zhao, J. Zhao, F. Shu, Y. Wu, and X. Cheng, "Anti-jamming technique for IRS aided JRC system in mobile vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 12 550–12 560, 2024.

[40] Y. Sun, K. An, J. Luo, Y. Zhu, G. Zheng, and S. Chatzinotas, "Outage constrained robust beamforming optimization for multiuser IRS-assisted anti-jamming communications with incomplete information," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 13 298–13 314, 2022.

[41] F. Xu, T. Hussain, M. Ahmed, K. Ali, M. A. Mirza, W. U. Khan, A. Ihsan, and Z. Han, "The state of AI-empowered backscatter communications: A comprehensive survey," *IEEE Internet of Things Journal*, vol. 10, no. 24, pp. 21 763–21 786, 2023.

[42] D. Balasubramaniam, S. Anbalagan, K. S. Thivya, and S. Suresh, "Advancements in IoT system security: A reconfigurable intelligent surfaces and backscatter communication approach," *The Journal of Supercomputing*, vol. 81, p. 183, 2025.

[43] L. H. Hiep, M. V. Ho, and D. M. Tran, "RIS-assisted deep reinforcement learning for intelligent anti-jamming communications in 6G networks," *TNU Journal of Science and Technology*, vol. 231, no. 6, pp. 419–428, 2026.

[44] Q. Zhou, Y. Niu, P. Xiang, and B. Wan, "Intelligent anti-jamming decision algorithm for wireless communication under limited channel state information conditions," *Scientific Reports*, vol. 15, p. 6271, 2025.

[45] H. Ding, Y. Niu, Q. Zhou, and X. Peng, "A novel intelligent anti-jamming communication algorithm based on proximal policy optimization," *Physical Communication*, vol. 65, p. 102366, 2024.

[46] K. Jia, D. Chen, and X. Sun, "Soft actor-critic based power control algorithm for anti-jamming in D2D communication," in *IEEE Int. Conf. Control, Electron. Comput. Technol. (ICCECT)*, Jilin, China, 2023, pp. 1–5.

About the Authors

Le Hoang Hiep

Workplace: Thai Nguyen University of Information and Communication Technology, Vietnam

Email: lhhiep@ictu.edu.vn

Education: Hiep Le-Hoang received the B.E. degree in Information Technology in 2009 from Thai Nguyen University of Information and Communication Technology, Vietnam, and the M.S. degree in 2013 from Manuel S. Enverga University Foundation, Philippines. He is currently pursuing the Ph.D. degree at Thai Nguyen University of Information and Communication Technology, Vietnam.

Recent research direction: His research interests include Wireless network security, Deep reinforcement learning, Internet of Things technologies, and UAV–satellite communications.

Tên tác giả: **Lê Hoàng Hiệp**

Cơ quan công tác: Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên.

Email: lhhiep@ictu.edu.vn

Quá trình đào tạo: Nhận bằng Kỹ sư tại Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên, Việt Nam năm 2009, và nhận bằng Thạc sĩ tại Trường Đại học Manuel S. Enverga, Philippines năm 2013. Hiện nay là nghiên cứu sinh tại Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên, Việt Nam.

Hướng nghiên cứu hiện nay: An toàn mạng không dây, học sâu tăng cường, công nghệ IoT và truyền thông vệ tinh – UAV.

Ngo Huu Huy

Workplace: Thai Nguyen University of Information and Communication Technology, Vietnam.

Email: nhhuy@ictu.edu.vn

Education: He received his B.S. and M.S. degrees from Thai Nguyen University of Technology, Vietnam, in 2010 and 2012, respectively, and Ph.D. degrees in Information Engineering and Computer Science from Feng Chia University, Taiwan, in 2021. Currently, he is a lecturer at the Thai Nguyen

University of Information and Communication Technology, Vietnam.
Recent research direction: Computer vision, deep learning, deep reinforcement learning, Internet of Things (IoT), and neural networks.

Tên tác giả: **Ngô Hữu Huy**

Cơ quan công tác: Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên, Việt Nam.

Email: nhhuy@ictu.edu.vn

Quá trình đào tạo: Nhận bằng Đại học và Thạc sĩ tại Trường Đại học Công nghệ Thông tin và Truyền thông –Đại học Thái Nguyên, Việt Nam lần lượt vào các năm 2010 và 2012; sau đó nhận bằng Tiến sĩ chuyên ngành Kỹ thuật thông tin và Khoa học máy tính tại Đại học Feng Chia, Đài Loan năm 2021. Hiện nay là giảng viên tại Trường Đại học Công nghệ Thông tin và Truyền thông –Đại học Thái Nguyên, Việt Nam.

Hướng nghiên cứu hiện nay: Thị giác máy tính, học sâu, học tăng cường sâu, Internet vạn vật (IoT) và mạng nơ-ron.