2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Exploration-Enhanced Dueling Double Deep Q-Network with Random Network Distillation for Satellite Beam Selection

Zijing Cheng[1,2], Chuxiong Sun[1,2], Shuaijun Liu[1,2], and Lixiang Liu[1,2](✉)

[1] Institute of Software Chinese Academy of Sciences,100190, Beijing, China
[2] University of Chinese Academy of Sciences, 100049, Beijing, China
{chengzijing2024, chuxiong2016,shuaijun,lixiang}@iscas.ac.cn

Abstract. With the evolution of satellite communication systems towards achieving low-latency and high-throughput performance, dynamic beam resource scheduling emerges as a challenging sequential decision-making task that can be effectively tackled using deep reinforcement learning (DRL). However, owing to the sparse channel characteristics and complex multi-user interference in satellite communications, traditional DRL methods struggle to obtain effective learning signals during exploration, resulting in suboptimal resource allocation efficiency. To address this challenge, in this work, we propose Beam selection with Integrated RND (BIRD), a novel framework that combines the Dueling Double Deep Q-Network (DQN) architecture with Random Network Distillation (RND) to enhance exploration capabilities in sparse state spaces. Our main innovations include the design of an enhanced solution framework that integrates Dueling DQN-based value evaluation architecture with RND mechanism to improve exploration efficiency through intrinsic rewards. Additionally, we develop a novel Markov Decision Process (MDP) model for formalizing the beam selection as a sequential decision problem. Simulation results demonstrate that BIRD achieves a significant 24.1% improvement in system sum rate compared to traditional beam selection methods.

Keywords: Multi-Beam Satellite, Deep Reinforcement Learning, Random Network Distillation, Beam Selection.

1 Introduction

In recent years, the rapid development of non-geostationary orbit (NGSO) constellations such as Oneweb, Starlink, and Telesat has established satellite communication systems as a pivotal component of next-generation wireless networks [11]. However, the traditional single-beam satellite system uses a wide beam to provide coverage for ground users, which makes it difficult to meet the huge business demands in current satellite communications due to resource limitations such as channel capacity and power.

To overcome this challenge, multi-beam satellite (MBS) systems have emerged as a fundamental technological solution, as illustrated in Fig. 1. These satellites are capable of generating multiple beams concurrently and dynamically allocating beam resources,

significantly enhancing system efficiency [22]. Furthermore, with the advancement of satellite payload technology, the increased number of available beams has made the beam selection problem increasingly complex. This complexity is further compounded by the unique characteristics of satellite communication channels. Unlike terrestrial systems, satellite channels are inherently sparse due to the high-altitude deployment and line-of-sight (LoS) dominated propagation environment, resulting in limited effective propagation paths between the satellite and ground users [12]. This channel sparsity manifests in the channel matrix structure, where only a small number of elements have significant values.

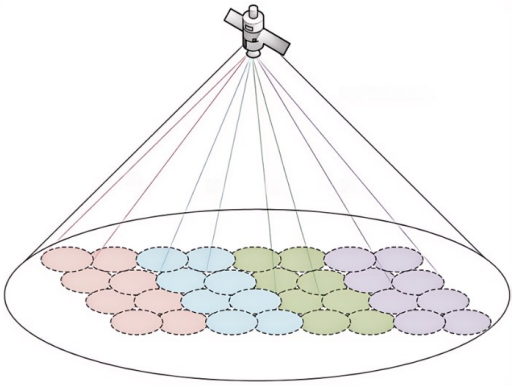

Fig. 1. Satellite coverage users

Traditional beam selection approaches, which often rely on exhaustive search or greedy algorithms, struggle to efficiently navigate this sparse solution space. These methods either become computationally intractable as the number of beams increases or fail to capture the intricate relationships between beam configurations and channel characteristics. Moreover, the dynamic nature of satellite communications, including varying user demands and channel conditions, necessitates an adaptive approach that can learn and optimize beam selection strategies over time.

Recent advances in deep reinforcement learning (DRL) have shown promising results in addressing complex decision-making problems in wireless communications. In the context of beam selection, DRL agents can learn to make sequential decisions by interacting with the environment, gradually improving their selection strategies through experience. However, existing DRL-based approaches often struggle with efficient exploration in sparse state spaces, where meaningful feedback signals are rare. This challenge is particularly pronounced in satellite beam selection scenarios, where the sparsity of channel matrices makes it difficult for conventional exploration strategies to discover optimal beam configurations.

As a result, we propose a novel framework that integrates the Dueling Double Deep Q-Network (Dueling DDQN) architecture with Random Network Distillation (RND) to enhance exploration capabilities in sparse state spaces, which we call Beam selection with Integrated RND (BIRD). The primary contributions of this paper are threefold:

- We rigorously formulate the challenging and high-dimensional satellite beam selection task, characterized by its complex dynamics and sparse state space, as a Markov Decision Process (MDP), thereby establishing a formal foundation for applying advanced DRL methods.
- As a core component of our proposed BIRD framework, we develop a novel Dueling DDQN architecture that effectively segregates state values and action advantages, leading to significantly improved value estimation accuracy in the beam selection scenarios.
- To overcome the critical exploration challenge posed by sparse reward signals inherent in satellite communication environments, we integrate a RND mechanism within the BIRD framework, substantially boosting exploration efficiency via dynamically generated intrinsic rewards.

The remainder of this paper is structured as follows. Sect. 2 reviews related work in beam selection methods and DRL. Sect. 3 presents the system model and problem formulation. Sect. 4 details the proposed BIRD framework. Sect. 5 provides experimental results and analysis. Finally, Sect. 6 concludes the paper. For convenience, the detailed notations and definitions used in this paper are summarized in Table 1.

Table 1. Notations and Definitions

Notations	Definitions
M_s	Number of satellite beams
K	Number of ground users
N_{RF}	Number of RF chains
$\mathbf{H}$	Channel matrix between satellite and users
$\mathbf{F}$	Beam selection matrix
$\mathbf{P}$	Digital precoding matrix
R_k	Achievable rate for user k
$SINR_k$	Signal-to-interference-plus-noise ratio for user k

2 Related Work

2.1 Traditional Beam Selection Methods.

Conventional beam selection methods can be broadly categorized into optimization-based and heuristic approaches [17]. For optimization-based methods, Hong et al. proposed a Lagrangian-based solution to handle the complex coupling constraints in beam allocation [8]. However, these methods often suffer from high computational complexity and poor scalability in dynamic scenarios.

Greedy-based methods have also been widely used due to their simplicity and computational efficiency. For instance, several works proposed a maximum magnitude (MM) based beam selection scheme that iteratively selects the beam with the strongest channel gain [1]. Another work developed an iterative weighted MMSE approach that greedily optimizes the sum-rate performance [16]. While these greedy algorithms are

computationally efficient, they often lead to suboptimal solutions due to their myopic nature.

To address these limitations, various heuristic algorithms have been proposed. Aravanis et al. developed a hybrid approach combining genetic algorithm with simulated annealing (GA-SA) for beam power allocation [3], while Cola et al. enhanced the simulated annealing method for joint optimization of bandwidth and power [6]. Additionally, Particle Swarm Optimization (PSO) based methods have been widely applied to beam resource scheduling problems [7]. However, as satellite communication systems become increasingly dynamic and complex, the beam selection problem gradually evolves into a complex sequential decision-making problem.

2.2 Exploration in DRL.

DRL has demonstrated remarkable performance in solving complex sequential decision-making problems and has been widely applied to resource scheduling in wireless communication systems [21]. In the context of satellite beam selection, DQN-based approaches have shown promising results [10,23]. However, a fundamental challenge in DRL is the exploration-exploitation dilemma, where agents must balance between exploiting known good strategies and exploring potentially better alternatives. Sun et al. investigated the introduction of random noise in the reward space rather than the action space, an exploration strategy that can more efficiently explore the state space while maintaining policy stability [26].This challenge becomes particularly acute in environments with sparse rewards or large state spaces, where meaningful feedback signals are rare and difficult to obtain.

Traditional exploration strategies in DRL primarily fall into uncertainty-based and probability-based approaches. The ϵ-greedy strategy, being the most fundamental, selects the current optimal action with probability $1-\epsilon$ and explores randomly with probability ϵ [2]. Boltzmann exploration introduces a temperature parameter to control exploration, assigning higher selection probabilities to actions with higher estimated values [5]. Sun revisited communication efficiency in multi-agent learning from a dimensional analysis perspective, pointing out that optimizing message timing and content solely at the receiving end is insufficient and requires a more comprehensive dimensional analysis perspective [29].Additionally, noise-based exploration strategies, such as parameter space noise [14], achieve exploration by adding perturbations to network parameters or action spaces. Li analyzed the mutual exclusion relationship between generalizability and discriminability in representational learning from an evolutionary game theory perspective, a theory that can also be applied to understanding the trade-offs in the exploration-exploitation dilemma [30]. Nevertheless, these methods often perform poorly in sparse reward environments, as their random exploration nature may require extensive time to discover valuable states.

To address the limitations of traditional exploration strategies, researchers have proposed exploration methods based on intrinsic motivation. Sun proposed an Intrinsic Motivated Multi-Agent Communication mechanism (IMMAC) based on the principle of "communicate what surprises you," which improves cooperative decision-making through intrinsic assessment of observed information importance, and similar ideas can

be applied to single-agent exploration [24].These approaches generate additional intrinsic reward signals to guide more effective exploration [13]. Among them, RND has gained significant attention due to its simplicity and effectiveness [4]. Sun employed masked state modeling and intention inference techniques in the M2I2 framework, an approach that can help agents process information more effectively in partially observable environments, sharing conceptual similarities with RND's novelty exploration principle [28]. RND generates intrinsic rewards by measuring the prediction error between a fixed random network and a trained predictor network, effectively quantifying state novelty without requiring prior knowledge of the environment. Sun proposed the T2MAC framework, which achieves more targeted and trusted multi-agent communication through selective engagement and evidence-driven integration, and this trustworthiness assessment mechanism could potentially be applied to evaluate intrinsic rewards generated by RND [25].While this approach has achieved remarkable success in challenging scenarios where traditional exploration strategies fail, its potential in wireless communication systems remains largely unexplored, particularly in satellite beam selection scenarios characterized by sparse feedback signals. This motivates our investigation into integrating RND with DQN for more effective exploration in satellite beam selection problems.

3 System Model and Problem Formulation

3.1 Task Description

This work investigates a MBS system operating in low Earth orbit (LEO) that implements the DVB-S2X standard [15]. The system architecture, depicted in Fig. 1, employs a satellite platform delivering services to K ground-based terminals through multiple beams. To facilitate efficient multi-beam transmission under hardware constraints, the satellite incorporates a hybrid beamforming architecture integrating an M_s-element discrete lens array (DLA) with N_{RF} radio frequency (RF) chains, where $N_{RF} \ll M_s$.

The DLA-based architecture enables signal transformation from the spatial domain to the beamspace domain via a lens antenna array that functions as a discrete Fourier transformer. This configuration facilitates the generation of M_s orthogonal beams, though only N_{RF} beams can be simultaneously activated due to RF chain limitations. This approach achieves significant hardware complexity reduction compared to fully-digital beamforming while preserving optimal spatial multiplexing capabilities.

The beam selection optimization is formulated as a MDP, characterized by the quintuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \beta)$. The state space $\mathcal{S}$ encompasses the system state s_t at time instant t, comprising the channel state matrix $\mathbf{H}_t$. The action space $\mathcal{A}$ constitutes the set of feasible beam selection matrices $\mathbf{F}_t$, where each action a_t represents a viable beam configuration adhering to RF chain constraints. The reward function $\mathcal{R}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ quantifies the achievable sum rate for the state-action pair (s_t, a_t), while the discount factor $\beta \in (0,1)$ optimizes the trade-off between immediate and future performance.

Signal Model In this system, we can consider a satellite communication system serving K ground users, where the signal transmission encompasses three essential stages: digital precoding, beam selection, and channel propagation. At any transmission instant, the received signal vector $\mathbf{y} \in \mathbb{C}^{K\times 1}$ at ground users can be formulated as:

$$\mathbf{y} = \mathbf{H}^H\mathbf{FPs} + \mathbf{n} = \sum_{k=1}^{K} \mathbf{h}_k^H \mathbf{F}\mathbf{p}_k s_k + \mathbf{n} \tag{1}$$

where the system parameters are defined as follows:

- The transmit symbol vector $\mathbf{s} \in \mathbb{C}^{K\times 1}$ satisfies the power normalization condition $\mathbb{E}[\mathbf{ss}^H] = \mathbf{I}_K$
- The digital precoding matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K] \in \mathbb{C}^{N_{RF}\times K}$ modifies the amplitude and phase of the input signals
- The beam selection matrix $\mathbf{F} \in \{0,1\}^{M_s\times N_{RF}}$ establishes one-to-one connections between RF chains and beam ports, with the constraint $\| \mathbf{F}(:, j) \|_0 = 1$ for each $j \in \{1, \ldots, N_{RF}\}$
- The beamspace channel matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K] \in \mathbb{C}^{M_s\times K}$ characterizes the signal propagation characteristics
- The receiver noise vector $\mathbf{n}$ follows complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_K)$

Beamspace Channel Model The satellite-to-ground channel exhibits unique characteristics due to the orbital environment [19]. These include a predominant LoS component owing to the elevated satellite position, significant free-space path loss coupled with atmospheric attenuation, and rapid channel variations caused by the high orbital velocity ($\approx$ 7.8 km/s).

Incorporating these physical characteristics, the beamspace channel matrix $\mathbf{H}$ can be expressed as:

$$\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] = \mathbf{U}\left[\beta_1\mathbf{g}_1 e^{j\phi_1}, \ldots, \beta_K\mathbf{g}_K e^{j\phi_K}\right] \tag{2}$$

where $\mathbf{U} \in \mathbb{C}^{M_s\times M_s}$ represents the DFT transformation matrix of the lens array, and $\mathbf{g}_k \in \mathbb{C}^{M_s\times 1}$ denotes the spatial channel vector for the k-th user. The composite attenuation factor β_k accounts for both free-space path loss and atmospheric attenuation, while ϕ_k represents the Doppler phase shift induced by satellite motion.

3.2 Problem Formulation.

The beam allocation problem is essentially an optimization task under limited resource constraints. By mapping this problem to the maximum weighted matching problem in bipartite graphs, its computational complexity can be proven to be NP-hard:

$$\max_{\mathbf{x}} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\, x_{ij} \leq_p \max_{\mathbf{F}} \sum_{k=1}^{K} R_k \tag{3}$$

where $\leq_p$ denotes polynomial-time reduction, and w_{ij} represents the matching weight between the i-th beam and the j-th RF chain. In LEO satellite communication systems, beam resource allocation faces unique challenges. The high-speed satellite motion not

only leads to dynamic coverage variations but also introduces significant Doppler effects. Moreover, the inter-beam coupling interference significantly impacts system performance. For user k, the signal-to-interference-plus-noise ratio can be expressed as:

$$SINR_k = \frac{\left|\mathbf{h}_k^H \mathbf{F}\mathbf{p}_k\right|^2}{\sigma^2 + \sum_{j \neq k} \left|\mathbf{h}_k^H \mathbf{F}\mathbf{p}_j\right|^2} \tag{4}$$

where $\mathbf{p}_k$ is the k-th column of precoding matrix $\mathbf{P}$, the numerator represents the effective signal power, and in the denominator, σ^2 denotes the noise power while the summation term represents co-frequency interference. Based on the DVB-S2X standard, the achievable rate for user k is given by:

$$R_k = \log_2(1 + SINR_k) \tag{5}$$

Accordingly, we formulate the sum-rate maximization problem as:

$$\begin{aligned} \max_{\mathbf{F}} \quad & \sum_{k=1}^{K} R_k \\ \text{s.t.} \quad & f_{i,j} \in \{0,1\}, \quad \forall i,j \\ & \sum_{i=1}^{M_s} f_{i,j} = 1, \quad \forall j \\ & \sum_{j=1}^{N_{RF}} f_{i,j} \leq 1, \quad \forall i \end{aligned} \tag{6}$$

The constraints reflect physical limitations: the binary constraint $f_{i,j} \in \{0,1\}$ represents the discrete nature of beam selection; $\sum_{i=1}^{M_s} f_{i,j} = 1$ ensures each RF chain selects exactly one beam; and $\sum_{j=1}^{N_{RF}} f_{i,j} \leq 1$ guarantees each beam is selected by at most one RF chain.

In the following sections, we will investigate how to solve this discrete optimization problem by constructing a MDP model to meet the practical requirements of LEO satellite systems.

4 Exploration-Enhanced DRL for Satellite Beam Selection

4.1 MDP Modeling

The beam selection problem in satellite communications presents significant computational challenges due to its NP-hard nature. Specifically, the problem exhibits the following characteristics: discrete decision variables with coupling constraints between beam selections, complex trade-offs between system sum-rate and user fairness, and high-dimensional state space due to the large number of potential beam configurations. Moreover, the sparse nature of the satellite channel matrix further complicates the learning process, as meaningful rewards become infrequent and scattered across the state space.

To address these challenges, we formulate the beam selection problem as a MDP, which can be mathematically described as follows:

State Space. The state space $\mathcal{S}$ comprises the beamspace channel matrix $\mathbf{H} \in \mathbb{C}^{M_S \times K}$ and an indicator tensor, where:

- The channel matrix $\mathbf{H}$ consists of two parts: the real and imaginary components, representing the complex channel gains between the satellite beams and ground users. the real part characterizes the in-phase component of the signal propagation, which primarily reflects the amplitude attenuation due to path loss and atmospheric effects, while the imaginary part represents the quadrature component that captures the phase shifts caused by signal propagation delay and Doppler effects due to satellite movement.
- The indicator tensor with dimension $1 \times M_s \times K$ indicates whether a beam is selected, initialized as 1 and set to 0 after selection
- M_s denotes the number of satellite beams
- K represents the number of ground users

Action Space. In the MDP framework, the action space $\mathcal{A}$ is defined as:

$$\mathcal{A} = \{a | a \in \mathbb{Z}^+, 1 \leq a \leq M_s\} \tag{7}$$

where each action a_t represents the beam index selected at time step t. The selection process has the following characteristics:

- The decision process contains N_{RF} time steps
- Only one beam is selected at each time step
- The action set updates dynamically to prevent duplicate selections

Reward Function.

In the beam selection problem, the system sum rate can only be evaluated after all RF chains have been assigned their beams, as the inter-beam interference pattern depends on the complete beam allocation. This creates a sparse reward situation where meaningful feedback is only available at the end of each episode, making it difficult for the agent to learn from intermediate steps. To address the sparse reward challenge in satellite beam selection, we design a comprehensive reward function that combines both extrinsic and intrinsic rewards

$$r_{extrinsic}(s_t, a_t) = r_{energy} + r_{rate} + r_{fairness} - r_{penalty} \tag{8}$$

where:

- r_{energy} evaluates beam energy efficiency:

$$r_{energy} = 20 \cdot \frac{\|\mathbf{f}_k\|_2}{\|\mathbf{f}_{max}\|_2} \tag{9}$$

where $\| \mathbf{f}_k \|_2$ represents the energy of the selected beam, and $\| \mathbf{f}_{max} \|_2$ is the maximum energy among currently available beams. The scaling factor 20 was determined empirically to balance with other reward components.

- r_{rate} calculates the system sum-rate, computed only after all beam selections are completed:

$$r_{rate} = \begin{cases} \text{WMMSE}(\mathbf{F}, P_{max}), & \text{if } t = N_{RF} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

  where the WMMSE algorithm considers inter-beam interference to compute the achievable rate under maximum transmit power constraint P_{max}.
- $r_{fairness}$ promotes balanced energy distribution among users:

$$\eta_k = \frac{|\mathbf{f}_k|^2}{\sum_{i=1}^{t}|\mathbf{f}_{i,k}|^2 + \epsilon} \tag{11}$$

$$r_{fairness} = 6\sum_{k=1}^{K}\eta_k \cdot \mathbb{1}_{\{\eta_k > 0.5 \text{ and } \sum_{i=1}^{t-1}|\mathbf{f}_{i,k}|^2 < 0.2\}} \tag{12}$$

- $r_{penalty}$ imposes penalties for constraint violations:

$$r_{penalty} = \begin{cases} 50, & \text{if beam already selected} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

  Additionally, to better guide exploration, we introduce an intrinsic reward based on RND:

$$r_{intrinsic} = \mathbb{E}\left[\left(\phi_{target}(s_t) - \phi_{predictor}(s_t)\right)^2\right] \tag{14}$$

where ϕ_{target} and $\phi_{predictor}$ are the target and predictor networks respectively.

This dual reward structure is particularly effective in the sparse satellite channel environment, as the intrinsic reward provides continuous learning signals even when meaningful extrinsic rewards are rare, thereby facilitating more efficient exploration of the beam selection space.

4.2 BIRD: A Dueling Double DQN Architecture with RND Enhancement.

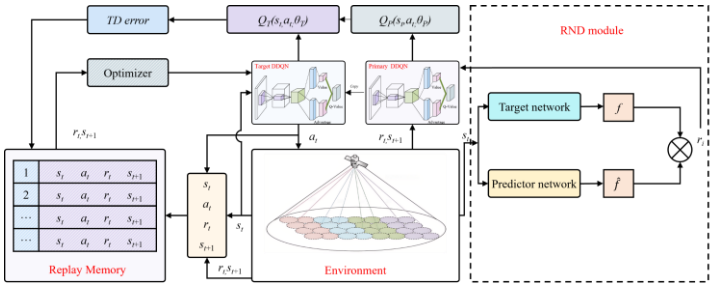

Fig. 2. Architecture of the proposed BIRD framework.

Based on our MDP formulation for satellite beam selection, where the state space comprises the complex channel matrix $\mathbf{H} \in \mathbb{C}^{M_s \times K}$ and the action space represents

beam selections, as illustrated in Fig. 2, we propose BIRD architecture to effectively handle the sparse channel characteristics and complex beam selection decisions.

BIRD Framework. Fig. 2 illustrates the BIRD framework architecture. In this framework, the agent interacts with the satellite beam environment, generating transition tuples (s_t, a_t, r_t, s_{t+1}). The agent's learning system consists of two components: a value network based on Dueling Double DQN and an exploration module based on RND. The value network adopts a dual network architecture, comprising a predictor network Q_p and a target network Q_t, with a replay buffer storing experience samples. The exploration module generates intrinsic rewards $\| \hat{f}(s_t) - f(s_t) \|^2$ through a target-predictor network structure. The agent optimizes a combined loss $\mathcal{L}_{total} = \mathcal{L}_{DQN} + \lambda \mathcal{L}_{RND}$ (where $\mathcal{L}_{DQN}$ represents the Dueling DQN loss), achieving exploration-exploitation trade-off in the beam selection task.

Dueling DQN Network Architecture. Based on our MDP formulation, where each state s_t encodes both channel information and beam selection history, BIRD employs a dueling architecture to better handle the high-dimensional state space and large-scale beam selection decisions. This architecture decomposes the Q-function into two streams (Z. Wang et al. 2016):

$$Q(s_t, a_t) = V(s_t) + \left(A(s_t, a_t) - \frac{1}{M_s} \sum_{a'} A(s_t, a') \right) \tag{15}$$

where:

- $V(s_t)$ estimates the overall value of the current channel state and beam selection history, capturing the inherent quality of the state regardless of beam choices
- $A(s_t, a_t)$ evaluates the advantage of selecting beam a_t in state s_t, reflecting the relative merit of each action
- M_s is the total number of available beams as defined in our action space

At each step, the network extracts high-dimensional channel features from state s_t, processes them through the value and advantage streams, and outputs Q-values for all possible beam selections. Combined with ϵ-greedy exploration and our designed reward function, this architecture effectively learns to make sequential beam selection decisions that maximize long-term system performance.

RND-Enhanced Exploration Mechanism. To address the exploration challenges in satellite beam selection, particularly the sparse reward nature and vast state space, we propose an RND-based exploration mechanism.

The RND component consists of two parallel networks:

- Target Network $f: \mathcal{S} \to \mathbb{R}^d$:

- Randomly initialized and fixed during training
- Maps states to a lower-dimensional embedding space

— Predictor Network $\hat{f}: \mathcal{S} \to \mathbb{R}^d$:

- Learns to predict the target network's output
- Shares the same architecture as the target network
- Updated through training to minimize prediction error

The predictor network is trained using:

$$\mathcal{L}_{RND}(\phi) = \mathbb{E}_{s_t \sim \mathcal{B}}\left[\| \hat{f}(s_t;\phi) - f(s_t) \|^2\right] \tag{16}$$

where:

- ϕ represents the predictor network parameters
- $\mathcal{B}$ is a mini-batch sampled from the experience buffer
- Gradient updates are normalized to prevent training instability

Algorithm 1 BIRD Training Algorithm

Input: Batch size B; Learning rate η; Replay period Q; Memory capacity D; Episodes J; RF chains N_{RF}; Beams M_s; Users K; Intrinsic reward weight β; Target update rate τ
Output: Trained Dueling DQN and RND predictor networks

Initialize replay memory $\mathcal{D} \leftarrow \emptyset$
Initialize Dueling DQN $Q(s,a;\theta)$ and target $Q'(s,a;\theta')$
Initialize RND target $f(s)$ and predictor $\hat{f}(s;\phi)$
for episode = 1 to J **do**
  Observe initial state s_0
  for $t = 1$ to N_{RF} **do**
    With probability ϵ, select random beam a_t; otherwise $a_t = \arg\max Q(s_t,a;\theta)$
    Execute a_t, update available beams
    Calculate extrinsic reward $r_{extrinsic}$
    if $t = N_{RF}$ **then**
      Calculate final r_{rate} and add to $r_{extrinsic}$
    end if
    Calculate intrinsic reward $r_{intrinsic} = \|\hat{f}(s_t;\phi) - f(s_t)\|^2$
    $r_{total} = r_{extrinsic} + \beta \cdot r_{intrinsic}$
    Store $(s_t, a_t, r_{total}, s_{t+1})$ in $\mathcal{D}$
    $s_t \leftarrow s_{t+1}$
  end for
  if episode mod $Q = 0$ and $|\mathcal{D}| \geq B$ **then**
    Sample mini-batch from $\mathcal{D}$
    Update RND predictor by minimizing $\mathcal{L}_{RND}$
    Calculate DDQN targets and update Dueling DQN
    Update target network: $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$
  end if
  Anneal ϵ
end for

In the next section, we will present detailed algorithm pseudocode and hyperparameter settings, demonstrating BIRD's performance in practical satellite channel scenarios. Specifically, we will experimentally validate how the synergy between Dueling DQN and RND enhances the efficiency and robustness of beam selection.

5 Simulation Results

To comprehensively evaluate BIRD's performance, we conduct experiments in a simulated MBS communication environment. Algorithm. [alg: BIRD_training] describes the overall process of the proposed BIRD scheme in detail in the form of pseudo-

code.The system performance is evaluated using the SINR defined in (4) and sum rate in (5).

Table 2. The Parameter Settings of Simulation

Parameter	Symbol	Value
Satellite orbit altitude	H	550Km
Signal-to-noise ratio	SNR	10dB
Carrier frequency	f_c	12.4 GHz
Maximum transmit antenna gain	G_m	35.9 dBi
System bandwidth	B	100MHz
Training episodes	J	10000
Mini-batch size	$\mathcal{B}$	32
Replay memory	$\mathcal{D}$	2048
Discount factor	γ	0.99
Target network update rate	τ	0.001
RND embedding dimension	d	128
Initial exploration rate	ϵ_{max}	1.0
Final exploration rate	ϵ_{min}	0.01
Exploration decay rate	λ	0.001

As shown in Table 2, we set up the simulation environment with the following key parameters: number of antennas $M_s = 64$, number of users $K = 16$, number of RF chains $N_{RF} = 8$. In our experiments, we first analyze the impact of reward weighting, where the total reward r_t for each action combines extrinsic reward r_e and intrinsic reward r_i, weighted by parameter β:

$$r_t = r_e + \beta \cdot r_i \tag{17}$$

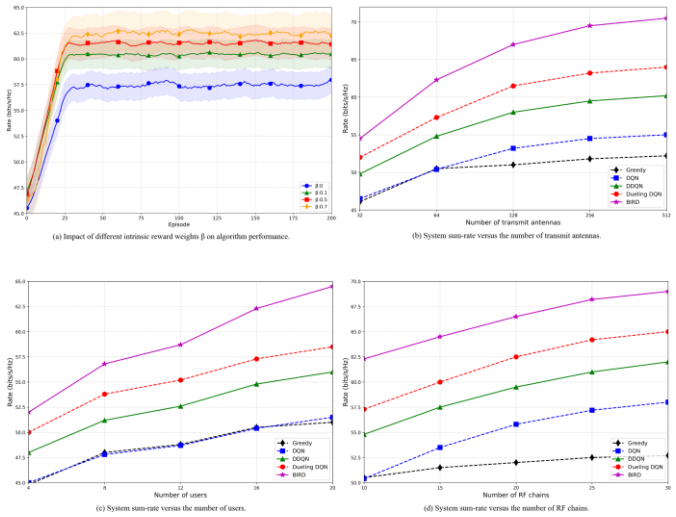

Fig. 3. Performance evaluation of beam selection algorithms under various system parameters.

The experimental results reveal that higher weight values generally lead to better performance, with β =0.7 achieving the highest sum rate of approximately 62.3 bits/s/Hz, while β=0.1 yields the lowest performance at 60.0 bits/s/Hz. This indicates that stronger intrinsic motivation helps the agent discover better policies. All configurations demonstrate rapid learning characteristics within the first 25 episodes before stabilizing, suggesting that the intrinsic reward weight primarily affects final performance rather than learning speed. Notably, while β=0.7 achieves the best performance, it also exhibits relatively larger fluctuations, as shown by the shaded areas, indicating that higher intrinsic reward weights may introduce some instability while improving performance. Considering both performance and stability, β=0.5 appears to be a reasonable compromise, maintaining good performance with reduced variance. These findings show the importance of properly balancing intrinsic and extrinsic rewards in our proposed method, as the weight directly influences both the exploration capability and stability of the learning process.

5.1 Performance Versus Different Senarios

Fig.3(b) shows the system performance as the number of transmit antennas increases from 32 to 512. The results demonstrate that BIRD consistently outperforms other methods across all antenna configurations. As the number of antennas increases, the performance gap between BIRD and conventional methods becomes more pronounced, indicating superior scalability of our proposed method. The greedy algorithm shows particularly poor scaling behavior, with its performance gap widening significantly at higher antenna numbers due to its inability to handle the exponentially growing solution space effectively. Moreover, with 512 antennas, BIRD achieves a maximum sum rate of 70.2 bits/s/Hz, showing a significant improvement of 10% over Dueling DQN and

27.4% over DQN. This performance advantage can be attributed to BIRD's enhanced exploration capability in the expanded state space provided by larger antenna arrays.

Fig.3(c) demonstrates the impact of user scaling on system performance. As the number of users increases from 6 to 20, all methods show an upward trend in sum-rate performance, but with different growth patterns. BIRD exhibits the strongest scaling capability, maintaining its performance advantage throughout the range and showing an increasingly wider gap over baseline methods as user numbers grow. The greedy algorithm, while performing reasonably well with fewer users, shows rapid performance degradation as user numbers increase due to its inability to handle the growing inter-user interference effectively. This trend is particularly evident in the enlarged view of the 8-10 user region, where BIRD achieves approximately 64.8 bits/s/Hz with 20 users, outperforming Dueling DQN by 15% and conventional DQN by 25%. The near-linear growth in system sum-rate with increasing users suggests that BIRD effectively manages the growing complexity of beam selection in multi-user scenarios, while conventional methods show signs of performance saturation at higher user loads.

Fig. 3(d) illustrates the impact of RF chain numbers on system performance. As the number of RF chains increases from 10 to 30, all methods demonstrate an upward trend in sum-rate performance. The greedy algorithm shows the poorest scaling behavior, with only marginal improvements as RF chains increase, suggesting its inability to effectively utilize additional RF resources due to its myopic decision-making process. With fewer RF chains (15-20), the performance improvement is rapid for DRL-based methods, as each additional RF chain significantly expands the available beam combinations. As the number of RF chains continues to increase (20-25), the growth rate begins to slow down, primarily because the number of effective beam combinations gradually approaches saturation. Beyond 25 RF chains, the performance improvement becomes more gradual, indicating that the system is approaching its theoretical capacity limit. With 30 RF chains, BIRD achieves approximately 69 bits/s/Hz, outperforming Dueling DQN by 12%, conventional DQN by 20%, and the greedy algorithm by 31%. This performance advantage stems from BIRD's RND mechanism, which better handles the combinatorial explosion problem caused by increased RF chains, effectively identifying and selecting optimal beam configurations in a larger action space. In contrast, traditional DQN and Dueling DQN often get trapped in local optima under large-scale RF chain configurations, unable to fully utilize the additional RF resources.

5.2 Performance Comparison

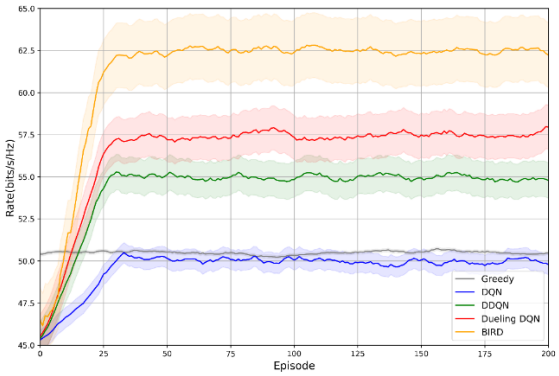

Fig. 4. Sum rate comparison among different algorithms.

This subsection compares the overall sum rate performance of the algorithms under the baseline system configuration, specifically using $M_s = 64$ satellite beams, $K = 16$ ground users, and $N_{RF} = 8$ RF chains,β=0.7.The experimental results Fig. 4 clearly demonstrate the superior performance of our proposed BIRD method, achieving a maximum sum rate of approximately 62.3 bits/s/Hz, compared to 57.3 bits/s/Hz for Dueling DQN, 54.8 bits/s/Hz for DDQN, and 50.4 bits/s/Hz for basic DQN. The greedy algorithm, while quickly reaching a performance level of about 50.2 bits/s/Hz in the early stages, shows no further improvement due to its inherent limitations. This result indicates that the myopic strategy of selecting immediate optimal choices is insufficient for handling complex beam selection problems. In contrast, although DRL-based methods require a training process, they can achieve superior performance through continuous learning and exploration. Notably, our BIRD method demonstrates rapid learning capabilities by reaching stability after only 25 episodes while achieving the best performance.

Table 3. Performance Comparison of Different Algorithms

Method	Sum Rate (bits/s/Hz)	Stability	Relative Improvement (%)
Greedy	50.2	$\pm$0.3	100.0
DQN	50.4	$\pm$2.2	100.4
DDQN	54.8	$\pm$1.8	109.2
Dueling DQN	57.3	$\pm$1.5	114.1
BIRD	**62.3**	**$\pm$1.2**	124.1

The stability analysis, as shown in Table 3, reveals that BIRD maintains minimal performance fluctuation throughout the training process, while the basic DQN exhibits the largest variance. The greedy algorithm, due to its deterministic nature, shows minimal performance variation, but this stability comes at the cost of exploring potentially better solutions. The performance improvements can be attributed to three key aspects: the Dueling structure, which contributes approximately 2.5 bits/s/Hz improvement over

DDQN, the RND mechanism, which provides an additional 5 bits/s/Hz gain, and most importantly, the ability to surpass the performance ceiling of traditional greedy methods through intelligent exploration and long-term optimization strategies.

6 Conclusion

In our work, we proposed a novel beam selection framework BIRD for satellite communication systems. By introducing the intrinsic reward to guide exploration and leveraging the advantage of Dueling architecture in value estimation, our method effectively addresses the exploration challenges in the high-dimensional beam selection space. Extensive experimental results demonstrate that our approach achieves significant improvements over baseline methods, with a 24.1% increase in sum rate performance compared to the greedy baseline. Future work could focus on extending the framework to more complex scenarios such as dynamic user distributions and multi-satellite coordination.

References

1. Amadori, P. V., and C. Masouros. 2015. "Low RF-Complexity Millimeter-Wave Beamspa ce-MIMO Systems by Beam Selection." *IEEE Trans. Commun.* 63 (6): 2212–23.
2. Amin, Susan, Maziar Gomrokchi, Harsh Satija, Herke Van Hoof, and Doina Precup. 2021. "A Survey of Exploration Methods in Reinforcement Learning." *arXiv Preprint arXiv:210 9.00157*.
3. Aravanis, A. I., B. Shankar M. R., P.-D. Arapoglou, G. Danoy, P. G. Cottis, and B. Otterste n. 2015. "Power Allocation in Multibeam Satellite Systems: A Two-Stage Multi-Objective Optimization." *IEEE Trans. Wireless Commun.*, 3171–82. https://doi.org/10.1109/twc.201 5.2402682.
4. Burda, Y., H. Edwards, A. Storkey, and O. Klimov. 2018. "Exploration by Random Netwo rk Distillation." *arXiv Preprint arXiv:1810.12894*.
5. Cesa-Bianchi, Nicolò, Claudio Gentile, Gábor Lugosi, and Gergely Neu. 2017. "Boltzman n Exploration Done Right." *Advances in Neural Information Processing Systems* 30.
6. Cocco, G., T. de Cola, M. Angelone, Z. Katona, and S. Erl. 2018. "Radio Resource Manag ement Optimization of Flexible Satellite Payloads for DVB-S2 Systems." *IEEE Trans. Bro adcast.*, 266–80. https://doi.org/10.1109/tbc.2017.2755263.
7. Durand, F., and T. Abrão. 2017. "Power Allocation in Multibeam Satellites Based on Parti cle Swarm Optimization." *AEU Int. J. Electron. Commun.* 78: 124–33. https://doi.org/10.1 016/J.AEUE.2017.05.012.
8. Hong, Y., A. Srinivasan, B. Cheng, L. Hartman, and P. Andreadis. 2008. "Optimal Power Allocation for Multiple Beam Satellite Systems." In *2008 IEEE Radio and Wireless Sympo sium*, 823–26. IEEE.
9. Hu, Xin, Shuaijun Liu, Rong Chen, Weidong Wang, and Chunting Wang. 2018. "A Deep Reinforcement Learning-Based Framework for Dynamic Resource Allocation in Multibea m Satellite Systems." *IEEE Communications Letters*, August, 1612–15. https://doi.org/10.1 109/lcomm.2018.2844243.

10. Lei, L., E. Lagunas, Y. Yuan, M. G. Kibria, S. Chatzinotas, and B. Ottersten. 2020. "Beam Illumination Pattern Design in Satellite Networks: Learning and Optimization for Efficient Beam Hopping." *IEEE Access*, 136655–67. https://doi.org/10.1109/access.2020.3011746.
11. Liu, J., Y. Shi, Z. M. Fadlullah, and N. Kato. 2018. "Space-Air-Ground Integrated Networ k: A Survey." *IEEE Commun. Surveys Tuts.*, 2714–41.
12. Liu, J., Y. Shi, L. Zhao, Y. Cao, W. Sun, and N. Kato. 2018. "Joint Placement of Controlle rs and Gateways in SDN-Enabled 5G-Satellite Integrated Network." *IEEE J. Sel. Areas Co mmun.*, 221–32. https://doi.org/10.1109/jsac.2018.2804019.
13. Mikhaylova, E., and I. Makarov. 2022. "Curiosity-Driven Exploration in VizDoom." *Proc. IEEE Int. Symp. Intell. Syst. Inform.*, 65–70. https://doi.org/10.1109/SISY56759.2022.100 36273.
14. Plappert, Matthias, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. 2017. "Parameter Space Noise for Exploration." *arXiv Preprint arXiv:1706.01905*.
15. Reimers, U. 1998. "Digital Video Broadcasting." *IEEE Communications Magazine* 36 (6): 104–10. https://doi.org/10.1109/35.685371.
16. Shi, Qingjiang, Meisam Razaviyayn, Zhi-Quan Luo, and Chen He. 2011. "An Iteratively W eighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interferin g Broadcast Channel." *IEEE Transactions on Signal Processing* 59 (9): 4331–40.
17. Shi, Shengchao, Guangxia Li, Zhiqiang Li, Hongpeng Zhu, and Bin Gao. 2017. "Joint Pow er and Bandwidth Allocation for Beam-Hopping User Downlinks in Smart Gateway Multib eam Satellite Systems." *International Journal of Distributed Sensor Networks*, May, 15501 4771770946. https://doi.org/10.1177/1550147717709461.
18. Wang, H., A. Liu, X. Pan, and J. Yang. 2014. "Optimization of Power Allocation for Multi users in Multi-Spot-Beam Satellite Communication Systems." *Math. Probl. Eng.* 2014 (1): 780823.
19. Wang, Ziqiang, Lei Xie, and Q. Wan. 2023. "Beamspace Joint Azimuth, Elevation, and De lay Estimation for Large-Scale MIMO-OFDM System." *IEEE Transactions on Instrument ation and Measurement* 72: 1–12. https://doi.org/10.1109/TIM.2023.3270972.
20. Wang, Z., T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas. 2016. "Dueling Ne twork Architectures for Deep Reinforcement Learning." In *International Conference on M achine Learning*, 1995–2003. PMLR.
21. Xu, G., F. Tan, Y. Ran, Y. Zhao, and J. Luo. 2023. "Joint Beam-Hopping Scheduling and Coverage Control in Multibeam Satellite Systems." *IEEE Wireless Communications Letter s* 12 (2): 267–71. https://doi.org/10.1109/LWC.2022.3194975.
22. Zeng, G., Y. Zhan, H. Xie, and C. Jiang. 2022. "Resource Allocation for Networked Telem etry System of Mega LEO Satellite Constellations." *IEEE Trans. Commun.* 70 (12): 8215–28.
23. Zhao, N., Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang. 2019. "Deep Reinforcemen t Learning for User Association and Resource Allocation in Heterogeneous Cellular Netwo rks." *IEEE Trans. Wireless Commun.* 18 (11): 5141–52.
24. Sun C, Wu B, Wang R, et al. Intrinsic motivated multi-agent communication[C]//Proceedin gs of the 20th International Conference on Autonomous Agents and MultiAgent Systems. 2 021: 1668-1670.
25. Sun C, Zang Z, Li J, et al. T2mac: Targeted and trusted multi-agent communication throug h selective engagement and evidence-driven integration[C]//Proceedings of the AAAI Con ference on Artificial Intelligence. 2024, 38(13): 15154-15163.

26. Sun C, Wang R, Li Q, et al. Reward space noise for exploration in deep reinforcement lear ning[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2021, 35(1 0): 2152013.
27. Wu B, Yang X, Sun C, et al. Learning effective value function factorization via attentional communication[C]//2020 IEEE International Conference on Systems, Man, and Cybernetic s (SMC). IEEE, 2020: 629-634.
28. Sun C, He P, Ji Q, et al. M2I2: Learning Efficient Multi-Agent Communication via Maske d State Modeling and Intention Inference[J]. arXiv preprint arXiv:2501.00312, 2024.
29. Sun C, He P, Wang R, et al. Revisiting Communication Efficiency in Multi-Agent Reinfor cement Learning from the Dimensional Analysis Perspective[J]. arXiv preprint arXiv:2501. 02888, 2025.
30. Li J, Zang Z, Ji Q, et al. Rethinking Generalizability and Discriminability of Self-Supervise d Learning from Evolutionary Game Theory Perspective[J]. International Journal of Comp uter Vision, 2025: 1-26.