Journal of Advanced Digital Communications
https://www.sciltp.com/journals/jadc

Article

Deep Reinforcement Learning—Based Beamforming for RIS-Aided 6G Systems

John Feng

School of Computer Science, The University of Sydney, Camperdown, NSW 2050, Australia; yfen7825@uni.sydney.edu.au

How To Cite: Feng, J. Deep Reinforcement Learning—Based Beamforming for RIS-Aided 6G Systems. *Journal of Advanced Digital Communications* **2025**, *2*(1), 3. https://doi.org/10.53941/jadc.2025.100003

Received: 20 November 2025
Revised: 18 December 2025
Accepted: 25 December 2025
Published: 26 December 2025

Abstract: Reconfigurable Intelligent Surfaces (RIS) have emerged as a key enabling technology for beyond-5G and 6G wireless networks, offering programmable control over the radio propagation environment with extremely low power consumption. However, jointly optimizing the base station (BS) beamforming vector and the high-dimensional RIS phase configuration remains a fundamentally challenging task due to non-convex coupling, hardware constraints, imperfect channel knowledge, and fast-varying user mobility patterns. Traditional optimization-based approaches, such as alternating optimization and convex relaxations, struggle to scale with large RIS arrays and are unable to adapt efficiently to rapidly changing channel conditions. To address these limitations, this work proposes a deep reinforcement learning (DRL) framework that learns an adaptive control policy through direct interaction with the wireless environment, without requiring explicit channel models or handcrafted optimization procedures. The proposed actor–critic architecture simultaneously outputs continuous beamforming and RIS phase-shift actions and incorporates domain-specific reward shaping to balance spectral efficiency, energy consumption, and phase-switching smoothness. Comprehensive experiments across diverse propagation scenarios—including shadowing variations, multipath sparsity levels, mobile users, and hardware ablation settings—demonstrate that the proposed method achieves significantly higher rate, energy efficiency, and robustness than conventional baselines, while maintaining efficient online inference suitable for real-time 6G deployments. The results confirm that DRL-driven beamforming provides a scalable and model-agnostic solution for next-generation intelligent wireless environments.

Keywords: reconfigurable intelligent surface (RIS); 6G wireless networks; deep reinforcement learning (DRL); actor-critic networks; energy-efficient communications; adaptive control; multi-path propagation

1. Introduction

The evolution toward sixth–generation (6G) wireless networks is driven by the demand for immersive communications, large-scale connectivity, and ultra-reliable high-capacity services, motivating the adoption of transformative physical-layer technologies beyond the limits of 5G systems [1,2]. Among various emerging paradigms, reconfigurable intelligent surfaces (RIS) have rapidly become a key enabler due to their capability to shape the wireless propagation environment in an energy-efficient and cost-effective manner [3]. By leveraging programmable metasurfaces, RISs can reconstruct wavefronts and provide controllable reflections, enabling near-real-time channel reconfiguration in future 6G scenarios [4]. Meanwhile, the integration of RIS into millimeter-wave and terahertz communication further enhances coverage and reliability, especially under severe blockage and non-line-of-sight (NLoS) conditions [5,6]. As system scale grows, particularly in extremely large RIS deployments, optimizing the joint beamforming between base stations (BS) and RIS becomes increasingly critical [7,8].

However, efficient control of RIS-assisted systems remains challenging due to their high dimensionality, fast-varying wireless conditions, and the need for low-latency real-time optimization. Existing resource allocation and user fairness solutions for mobile and aerial platforms highlight the complexity created by spatial mobility and dynamic topology [9,10]. To address interference, channel correlation, and electromagnetic coupling, advanced model-based optimization has been explored, including MMSE-based designs and fractional programming approaches [11,12]. Yet, these conventional methods typically rely on accurate channel state information (CSI), involve iterative non-convex optimization, and may be unsuitable for real-time adaptation in large-scale RIS settings.

The emergence of machine learning, particularly model-driven and data-driven hybrid approaches, has opened new opportunities for physical-layer optimization [13,14]. Reinforcement learning (RL) and deep reinforcement learning (DRL) have been widely recognized as promising paradigms for intelligent communication systems, enabling autonomous resource allocation, network control, and policy adaptation under uncertainty [15,16]. Recent developments in DRL-driven joint beamforming for active and passive RIS architectures demonstrate the potential of actor–critic and policy-gradient algorithms in solving continuous control problems intrinsic to high-dimensional beamforming [17,18]. Furthermore, advances in RIS channel modeling and mobility-aware prediction methods provide new tools for improving the realism and robustness of DRL policies [19,20].

Nevertheless, despite growing interest in DRL for RIS-aided networks, several important shortcomings remain. Prior DRL-based schedulers and QoS-aware control frameworks illustrate the sensitivity of learning stability to reward shaping and network dynamics [21], while DRL-driven localization and sensing tasks reveal challenges relating to partial observability and sample inefficiency [22]. Comprehensive surveys on RIS beamforming optimization further highlight that scalable learning under large RIS sizes, high mobility, and limited CSI feedback is still underexplored [23]. In addition, RIS-enabled multiuser MIMO uplink studies emphasize the need for jointly optimizing spectral and energy efficiency under realistic hardware constraints [24]. Distributed DRL systems for multi-access and vehicular networks further underscore the need for scalable policy learning when subject to latency, user mobility, and interference coupling [25].

Motivated by these challenges, this work proposes a unified DRL-based framework for joint BS beamforming and RIS phase optimization in 6G communication environments. Leveraging continuous-action actor–critic architectures, the proposed system integrates real-time environmental feedback with model-aware feature representations to achieve stable, scalable, and energy-efficient beamforming decisions. The framework is designed to operate under imperfect CSI, mobility-induced fluctuations, and large RIS dimensionality, while preserving adaptability across diverse propagation conditions. Through comprehensive simulations under varying SNR levels, shadowing conditions, mobility patterns, and RIS configurations, we demonstrate that the proposed DRL framework significantly outperforms classical optimization baselines in throughput, robustness, and energy efficiency, offering a promising foundation for next-generation RIS-assisted 6G networks. It is worth clarifying the scope of comparison and positioning of the proposed framework with respect to existing learning-based RIS optimization studies. Recent works have explored DRL- or graph-based approaches under specific system assumptions, such as discrete phase shifts, fixed beamforming structures, or task-oriented objectives tailored to particular network settings. In contrast, the present work focuses on a unified continuous-control formulation that jointly optimizes beamforming and RIS phase configurations within a single actor–critic framework, with emphasis on scalability, stability, and deployment-oriented design. Due to the diversity of modeling assumptions, state representations, and action spaces across existing learning-based methods, direct quantitative comparison is not always meaningful. Instead, this work aims to demonstrate consistent performance gains over well-established optimization baselines while providing a flexible and model-agnostic learning framework applicable to a broad range of RIS-assisted 6G scenarios.

2. Methodology

In this section, we present the proposed deep reinforcement learning (DRL) framework for joint base station (BS) beamforming and reconfigurable intelligent surface (RIS) phase optimization in 6G wireless systems.

We first introduce the overall system architecture and signal model, then formalize the optimization objectives for rate and energy efficiency. Subsequently, we cast the problem into a Markov decision process (MDP) and describe the DRL agent design, including state representation, action space, reward shaping, and network architectures. Finally, we discuss the training procedure, convergence aspects, and computational complexity.

2.1. Overall System Architecture

We consider a downlink 6G communication scenario in which a multi-antenna BS communicates with one or more single-antenna user equipments (UEs) in the presence of a programmable RIS. The RIS is deployed on a building facade or indoor wall to enhance coverage in non-line-of-sight (NLoS) zones, which are typical in millimeter-wave (mmWave) and terahertz (THz) bands. The DRL agent resides either at the BS (e.g., integrated in the baseband unit) or in a nearby edge server, and is responsible for jointly adapting the BS beamforming vector and

RIS phase configuration based on real-time observations.

Figure 1 provides an intuitive illustration of the proposed architecture. The BS transmits downlink signals that reach the UEs through both direct and RIS-reflected paths. The RIS consists of N passive reflecting elements whose phase shifts are electronically tunable. The DRL agent observes the system state, including channel measurements and past control actions, and outputs a pair of continuous control variables: the BS beamforming vector and the RIS phase-shift vector. User feedback, such as estimated signal-to-noise ratio (SNR) or achievable rate, is aggregated into a scalar reward that guides the learning process. This closed-loop interaction plays a central role in enabling adaptive and data-driven beamforming in highly dynamic propagation environments, where hand-crafted optimization becomes intractable.

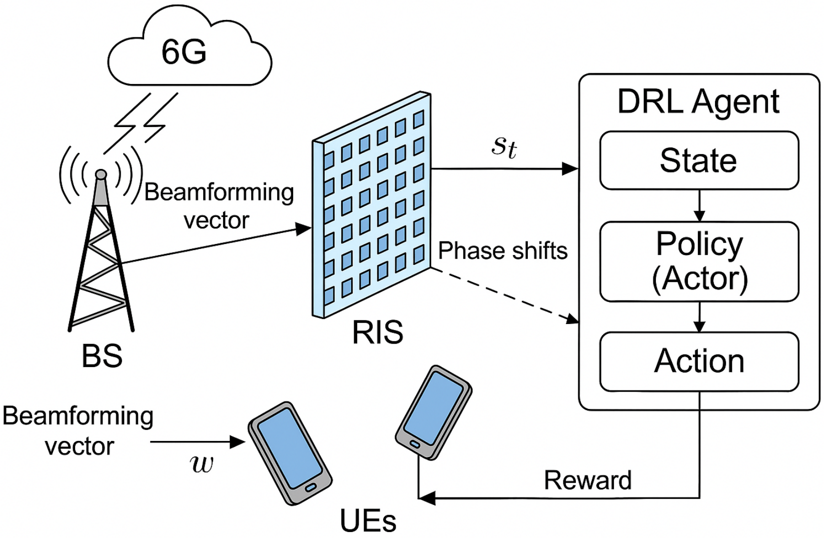

Figure 1. Overall RIS-aided 6G downlink architecture. The BS sends a beamforming vector toward both the UEs and the RIS; the DRL agent observes the system state s_t, configures the RIS phase shifts, and updates the BS beamforming, while the UEs provide reward feedback to close the learning loop.

2.2. *Channel and Signal Model*

We adopt a narrowband block-fading model over a single resource block, which can be extended to an OFDM-based multi-carrier system with subcarrier-wise processing. The channel remains quasi-static within one coherence block and changes independently across blocks.

2.2.1. BS–UE Direct Link

Let $\mathbf{h}_{\mathrm{BU},u} \in \mathbb{C}^M$ denote the direct channel from the M-antenna BS to the u-th UE. The received baseband signal via the direct link is affected by large-scale path loss, shadowing, and small-scale fading. In mmWave/THz regimes, the channel is often modeled as a clustered Rician channel with a small number of dominant paths:

$$\mathbf{h}_{\mathrm{BU},u} = \sqrt{\frac{K_{\mathrm{BU}}}{K_{\mathrm{BU}}+1}}\,\mathbf{h}_{\mathrm{BU},u}^{(\mathrm{LoS})} + \sqrt{\frac{1}{K_{\mathrm{BU}}+1}}\,\mathbf{h}_{\mathrm{BU},u}^{(\mathrm{NLoS})}, \tag{1}$$

where K_{BU} is the Rician K-factor and $\mathbf{h}_{\mathrm{BU},u}^{(\mathrm{LoS})}$ and $\mathbf{h}_{\mathrm{BU},u}^{(\mathrm{NLoS})}$ represent the line-of-sight (LoS) and non-line-of-sight components, respectively.

2.2.2. BS–RIS and RIS–UE Links

The BS–RIS channel is modeled by $\mathbf{H}_{\mathrm{BR}} \in \mathbb{C}^{N\times M}$, where the (n,m)-th entry corresponds to the channel from the m-th BS antenna to the n-th RIS element. Similarly, the RIS–UE channel is modeled by $\mathbf{h}_{\mathrm{RU},u} \in \mathbb{C}^N$. Owing to the typically unobstructed placement of RIS panels, these links can often be modeled as Rician channels with larger LoS components:

$$\mathbf{H}_{\mathrm{BR}} = \sqrt{\frac{K_{\mathrm{BR}}}{K_{\mathrm{BR}}+1}}\,\mathbf{H}_{\mathrm{BR}}^{(\mathrm{LoS})} + \sqrt{\frac{1}{K_{\mathrm{BR}}+1}}\,\mathbf{H}_{\mathrm{BR}}^{(\mathrm{NLoS})}, \tag{2}$$

$$\mathbf{h}_{\mathrm{RU},u} = \sqrt{\frac{K_{\mathrm{RU}}}{K_{\mathrm{RU}}+1}}\,\mathbf{h}_{\mathrm{RU},u}^{(\mathrm{LoS})} + \sqrt{\frac{1}{K_{\mathrm{RU}}+1}}\,\mathbf{h}_{\mathrm{RU},u}^{(\mathrm{NLoS})}. \tag{3}$$

The RIS imposes a diagonal phase-shift matrix

$$\boldsymbol{\Theta} = \mathrm{diag}\left(e^{j\theta_1}, e^{j\theta_2}, \ldots, e^{j\theta_N}\right), \tag{4}$$

where $\theta_n \in [0, 2\pi)$ is the tunable phase associated with the n-th element. We consider a continuous-phase RIS model in this work, but the same framework can accommodate discrete-phase quantization by restricting the action space. The system model adopts several idealized assumptions to facilitate algorithmic development and analysis. In particular, continuous RIS phase control and accurate channel-related observations are assumed to enable a clear exposition of the joint beamforming and RIS optimization problem. These assumptions are commonly employed in RIS-aided communication studies to establish a tractable and interpretable baseline. In practical deployments, RIS elements typically support discrete phase resolutions, channel state information may be imperfect due to estimation errors and feedback limitations, and control signaling may introduce non-negligible latency. The proposed DRL framework is not inherently restricted to ideal conditions and can accommodate such constraints by quantizing the actor outputs, incorporating noisy or partial channel observations into the state representation, and adjusting the control update interval to match hardware switching capabilities. While a detailed hardware-level evaluation is beyond the scope of this work, the adopted modeling assumptions allow the fundamental behavior and potential of the learning-based control strategy to be systematically investigated.

2.2.3. Received Signal and Effective Channel

The BS transmits symbol $x \in \mathbb{C}$ with beamforming vector $\mathbf{w} \in \mathbb{C}^M$, constrained by $\|\mathbf{w}\|^2 \leq P_{\max}$. The received signal at user u is

$$y_u = \left(\mathbf{h}_{\mathrm{BU},u}^H + \mathbf{h}_{\mathrm{RU},u}^H \boldsymbol{\Theta} \mathbf{H}_{\mathrm{BR}}\right) \mathbf{w} x + n_u, \tag{5}$$

where $n_u \sim \mathcal{CN}(0, \sigma^2)$ is additive white Gaussian noise. The effective cascaded channel seen by the u-th user is thus

$$\mathbf{h}_{\mathrm{eff},u}^H = \mathbf{h}_{\mathrm{BU},u}^H + \mathbf{h}_{\mathrm{RU},u}^H \boldsymbol{\Theta} \mathbf{H}_{\mathrm{BR}}. \tag{6}$$

The instantaneous signal-to-interference-plus-noise ratio (SINR) for the single-user case reduces to

$$\gamma_u = \frac{\left|\mathbf{h}_{\mathrm{eff},u}^H \mathbf{w}\right|^2}{\sigma^2}, \tag{7}$$

and the corresponding achievable rate is

$$C_u = \log_2\left(1 + \gamma_u\right). \tag{8}$$

2.2.4. Multi-User Extension

For completeness, we outline the multi-user case with U users, where the BS applies a linear precoder $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_U] \in \mathbb{C}^{M \times U}$ and transmits $\mathbf{x} = \sum_{u=1}^{U} \mathbf{w}_u x_u$. The received signal at user u is

$$y_u = \mathbf{h}_{\mathrm{eff},u}^H \mathbf{w}_u x_u + \sum_{k \neq u} \mathbf{h}_{\mathrm{eff},u}^H \mathbf{w}_k x_k + n_u, \tag{9}$$

and the SINR is

$$\gamma_u = \frac{\left|\mathbf{h}_{\mathrm{eff},u}^H \mathbf{w}_u\right|^2}{\sum_{k \neq u} \left|\mathbf{h}_{\mathrm{eff},u}^H \mathbf{w}_k\right|^2 + \sigma^2}. \tag{10}$$

The proposed DRL framework naturally extends to this setting by including all user channels in the state and redefining the reward as a function of the sum-rate or a weighted fairness metric.

2.3. Optimization Objectives

The primary goal of the joint beamforming and RIS design is to maximize the communication performance subject to power and hardware constraints.

2.3.1. Rate Maximization

For the single-user case, the rate-maximization problem can be written as

$$\max_{\mathbf{w},\boldsymbol{\Theta}} \quad C_u = \log_2\left(1+\gamma_u\right) \tag{11}$$

$$\text{s.t.} \quad \|\mathbf{w}\|^2 \le P_{\max}, \tag{12}$$

$$\theta_n \in [0, 2\pi), \quad n = 1, \ldots, N. \tag{13}$$

In the multi-user case, the objective may be chosen as the sum-rate $\sum_{u=1}^{U} C_u$ or a proportional-fair measure depending on service requirements.

2.3.2. Energy Efficiency and Regularity

To explicitly account for energy efficiency (EE) and temporal smoothness of control actions, we adopt a composite objective that combines rate and power consumption:

$$\text{EE} = \frac{C_{\text{tot}}}{P_{\text{tx}} + P_{\text{c}}}, \tag{14}$$

where C_{tot} is the total achievable rate across users, P_{tx} is the BS transmit power, and P_{c} captures circuit and control overhead (e.g., RIS control signaling). Moreover, abrupt changes in RIS phase configuration may be undesirable due to hardware switching constraints. Therefore, we introduce a regularization term

$$\Delta\theta_t = \sum_{n=1}^{N} \left|\theta_n(t) - \theta_n(t-1)\right|, \tag{15}$$

which penalizes large temporal variations in RIS configurations.

Classically, solving (11) and its EE-regularized variants requires non-convex optimization and iterative algorithms. In large-scale RIS settings with fast time-varying channels, such approaches become computationally prohibitive, motivating a learning-based alternative.

2.4. DRL-Based Beamforming Framework

To address the non-convexity, dimensionality, and dynamics of the joint design problem, we adopt a DRL framework. Figure 2 provides an abstract view of the agent–environment interaction loop. The wireless system (BS, RIS, and UEs) is treated as the environment, while the DRL agent observes system states and outputs actions corresponding to beamforming and RIS phase configurations.

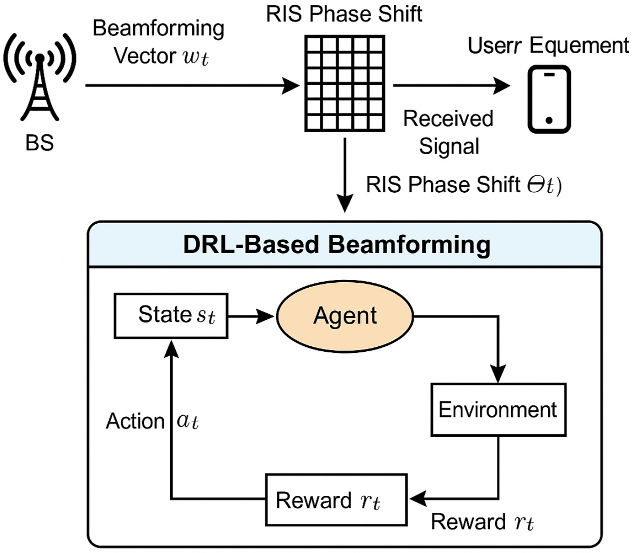

Figure 2. Abstract DRL-based beamforming framework. The upper part shows the physical BS–RIS–UE link with beamforming vector w_t and RIS phase shift $\boldsymbol{\Theta}_t$; the lower block depicts the MDP loop where the agent maps state s_t to action a_t and receives reward r_t.

2.4.1. MDP Formulation

We formulate the control problem as an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P})$.

State Space $\mathcal{S}$

The state $s_t \in \mathcal{S}$ at time step t aggregates the relevant information for decision-making. A representative design is

$$s_t = \left[\hat{\mathbf{h}}_{\mathrm{BU}}(t),\, \hat{\mathbf{H}}_{\mathrm{BR}}(t),\, \hat{\mathbf{h}}_{\mathrm{RU}}(t),\, \mathbf{\Theta}_{t-1},\, \mathbf{w}_{t-1}\right], \tag{16}$$

where the hat notation indicates estimated or compressed channel information (e.g., dominant path parameters, effective channel gains), and $(\mathbf{\Theta}_{t-1}, \mathbf{w}_{t-1})$ represent the most recent control actions. In the multi-user case, this state is extended to include all users' effective channels or appropriate summary statistics (e.g., user SINR vector).

To reduce dimensionality, we may apply linear projections or neural encoders that map high-dimensional CSI to low-dimensional feature vectors before feeding them to the DRL networks. The high-dimensional channel-related observations included in the state are processed through a dedicated feature encoding stage before being fed into the actor and critic networks. The primary purpose of this encoder is to reduce the dimensionality of cascaded BS–RIS–UE channel representations while preserving dominant spatial characteristics that are most relevant for beamforming and RIS phase control. By mapping raw or compressed channel descriptors to a compact latent representation, the encoder facilitates stable learning and mitigates overfitting in large RIS configurations. A lightweight multilayer perceptron is adopted for feature encoding to balance representational capability and computational efficiency. This design choice is motivated by the requirement of real-time inference in RIS-assisted 6G systems, where excessive architectural complexity may lead to prohibitive latency. While more sophisticated encoders, such as attention-based mechanisms or graph neural networks, may further exploit structured channel correlations, they typically incur higher computational overhead and require additional architectural tuning. In this work, the focus is therefore placed on a simple yet effective encoding strategy that aligns with the overall objective of scalable and low-latency deployment.

Action Space $\mathcal{A}$

The action at time t is

$$a_t = \left[\mathbf{w}(t), \mathbf{\Theta}(t)\right], \tag{17}$$

where both $\mathbf{w}(t)$ and the phase angles $\{\theta_n(t)\}$ are continuous variables. The resulting action space is high-dimensional and continuous, which justifies the use of actor–critic algorithms designed for continuous control (e.g., DDPG, TD3, SAC). If a discrete-phase RIS is used in practice, the actor output can be quantized post hoc without changing the MDP formulation.

Reward Function $\mathcal{R}$

The scalar reward $r_t \in \mathbb{R}$ is designed to promote high communication performance while discouraging excessive power usage and control instability:

$$r_t = \alpha C_{\mathrm{tot},t} - \beta \|\mathbf{w}(t)\|^2 - \lambda \Delta\theta_t, \tag{18}$$

where $\alpha, \beta, \lambda \geq 0$ are tunable weights. For the single-user case, $C_{\mathrm{tot},t}$ reduces to $C_u(t)$, whereas in the multi-user case it may represent sum-rate or a weighted fairness objective. The penalty on $\|\mathbf{w}(t)\|^2$ enhances energy efficiency, while the penalty on $\Delta\theta_t$ enforces smoother RIS configuration trajectories.

Transition Dynamics $\mathcal{P}$

The next state s_{t+1} is generated by the wireless environment based on the previous state and selected action:

$$s_{t+1} = f(s_t, a_t) + \omega_t, \tag{19}$$

where $f(\cdot)$ reflects channel evolution (e.g., due to user mobility or fading), and ω_t captures unmodeled randomness (e.g., estimation errors). The transition kernel $\mathcal{P}$ is not assumed to be known analytically; instead, it is implicitly learned through interaction, which is a key advantage of DRL for model-deficient systems.

The proposed control problem is formulated as a Markov decision process to enable sequential and adaptive optimization of beamforming and RIS phase configurations. Although wireless channels evolve due to mobility and estimation noise, the state representation includes both the current channel-related observations and the most recent control actions. This design provides an approximately Markov description at the coherence-block timescale, since future observations are largely determined by the current environment state and applied actions. Action feasibility is ensured through explicit constraint handling at the execution stage. In particular, the beamforming vector produced by the actor is normalized to satisfy the transmit power constraint, while RIS phase outputs are bounded to the physically valid range. As a result, all actions applied to the wireless environment remain feasible even though the policy is learned in a continuous, unconstrained parameter space. The reward function further incorporates a temporal smoothness regularization on RIS phase updates. This term discourages abrupt reconfiguration between consecutive time steps, which not only reflects practical hardware limitations but also stabilizes policy learning by suppressing oscillatory control behavior. By balancing communication performance, energy usage, and control regularity, the proposed reward design promotes robust and stable learning in high-dimensional continuous-action settings.

2.5. Actor–Critic Network Design

We adopt an actor–critic architecture to handle the continuous action space. The actor network $\pi_\phi(s)$ parameterized by ϕ maps states to actions, while the critic network $Q_\psi(s, a)$ parameterized by ψ estimates the expected return Q-value.

2.5.1. Actor Network

The actor network is implemented as a multi-layer perceptron (MLP) with L hidden layers:

$$a_t = \pi_\phi(s_t) = f_\phi^{(L)} \circ f_\phi^{(L-1)} \circ \cdots \circ f_\phi^{(1)}(s_t), \tag{20}$$

where $f_\phi^{(\ell)}(\cdot)$ denotes affine transformations followed by non-linear activations (e.g., ReLU or tanh). The output layer is partitioned into two parts: one for the BS beamforming vector and one for the RIS phase angles. Appropriate output activations (such as scaled tanh) ensure that the beamforming power constraint and phase range are respected, possibly followed by a normalization step:

$$\mathbf{w}(t) \leftarrow \sqrt{P_{\max}}\,\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}. \tag{21}$$

2.5.2. Critic Network

The critic network takes both the state and action as inputs and outputs a scalar Q-value:

$$Q_\psi(s_t, a_t) = g_\psi\big([s_t, a_t]\big), \tag{22}$$

where $g_\psi(\cdot)$ is again an MLP. The critic is trained to minimize the temporal-difference (TD) loss based on a target y_t:

$$\mathcal{L}(\psi) = \mathbb{E}\big[(Q_\psi(s_t, a_t) - y_t)^2\big]. \tag{23}$$

For example, in a DDPG-like algorithm,

$$y_t = r_t + \gamma Q_{\psi'}\big(s_{t+1}, \pi_{\phi'}(s_{t+1})\big), \tag{24}$$

where ψ' and ϕ' denote parameters of slowly updated target networks and $\gamma \in (0, 1)$ is the discount factor.

2.5.3. Policy Update

The actor parameters ϕ are updated by ascending the gradient of the expected return:

$$\nabla_\phi J(\phi) = \mathbb{E}_{s\sim\mathcal{D}}\left[\nabla_a Q_\psi(s, a)\big|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)\right], \tag{25}$$

where $\mathcal{D}$ denotes the replay buffer distribution. The use of experience replay and target networks significantly improves stability in training.

2.6. Training Procedure and Practical Considerations

The DRL training follows the standard off-policy actor–critic loop with replay memory. At a high level, the procedure is:

(1) Initialize actor π_ϕ, critic Q_ψ, and corresponding target networks. Initialize replay buffer $\mathcal{D}$.
(2) Observe initial state s_0 (estimated channels and control initialization).

(3) For each time step t:

  (a) Select action $a_t = \pi_\phi(s_t) + \mathcal{N}_t$, where $\mathcal{N}_t$ is an exploration noise process (e.g., Ornstein–Uhlenbeck or Gaussian).

  (b) Apply a_t to the wireless environment, obtain reward r_t and next state s_{t+1}.

  (c) Store transition (s_t, a_t, r_t, s_{t+1}) in replay buffer $\mathcal{D}$.

  (d) Sample a mini-batch from $\mathcal{D}$, update critic by minimizing the TD loss, and update actor via policy gradient.

  (e) Update target networks using soft updates.

(4) Repeat until convergence or a fixed number of episodes is reached.

For clarity and reproducibility, we summarize the main hyperparameter settings used to train the DRL agent. The replay buffer is configured with a fixed capacity sufficient to store past transitions across multiple episodes, and mini-batches are uniformly sampled during training. Both the actor and critic networks are optimized using Adam with separate learning rates, where the critic learning rate is set higher to enable faster value-function convergence. Exploration during training is achieved by injecting stochastic noise into the actor outputs, using a temporally correlated noise process to encourage efficient exploration in the continuous action space. Target networks for both the actor and critic are updated using a soft-update strategy with a small update coefficient to stabilize training. The actor and critic are implemented as multilayer perceptrons with identical depth and moderate hidden-layer widths to balance representational capacity and computational efficiency. All hyperparameters are kept fixed across experiments to ensure fair and consistent evaluation.

In practice, the DRL agent can be trained offline in a high-fidelity simulation environment, and then deployed online with fine-tuning using real measurements. The online inference phase is highly efficient: given a state vector, the actor network performs a few matrix multiplications to produce the beamforming and RIS configuration, which has significantly lower complexity than solving a non-convex optimization problem from scratch at each time slot. Once training is completed, the online execution of the proposed framework primarily consists of a forward pass through the actor network, followed by simple normalization operations for beamforming and RIS phase outputs. The inference-time computational cost is therefore determined by the fixed network depth and width, and remains constant with respect to the number of optimization iterations, in contrast to classical alternating or convex optimization methods that require repeated iterative updates. As the RIS size increases, the computational burden scales linearly with the input dimension of the encoder and the first network layers, while the overall network depth remains unchanged. This property enables scalable deployment for large RIS arrays without incurring iterative solver overhead. Moreover, since online control does not involve gradient computation or backpropagation, the runtime requirements are modest and compatible with real-time execution on standard CPU or GPU platforms. While detailed hardware-specific latency measurements are beyond the scope of this work, the above analysis clarifies why the proposed DRL-based approach is well suited for low-latency and deployment-oriented RIS-assisted communication systems.

2.7. Computational Complexity and Deployment Aspects

The computational complexity of the proposed DRL framework is dominated by two components: (i) the forward pass of the actor network during inference; and (ii) the backward pass during training. Let d_s and d_a be the state and action dimensions, and let each hidden layer of the actor contain H neurons. The complexity of a single forward pass is approximately $\mathcal{O}(d_sH + (L-1)H^2 + Hd_a)$, which is modest even for large N and can be handled in real time on edge hardware.

By contrast, classical alternating optimization or meta-heuristic algorithms typically require iterative updates with complexity that scales at least linearly with the number of RIS elements and may involve matrix inversions or large-scale searches. As a result, the DRL approach offers a favorable trade-off: substantial offline training cost in exchange for extremely fast online decision-making.

From a deployment perspective, the proposed framework can be integrated into a 6G radio access network (RAN) as a software module in the BS baseband unit or in a co-located edge server. The interface to the RIS controller is realized through standard control links (e.g., Ethernet or dedicated control channels), and the control period can be aligned with the channel coherence time to guarantee both reactivity and stability.

In summary, this section has detailed the physical-layer modeling of the RIS-aided 6G system, the resulting non-convex joint beamforming problem, and the DRL-based solution framework. The use of a continuous-action actor–critic architecture allows the agent to learn efficient beamforming and RIS configurations directly from interaction with the environment, circumventing explicit non-convex optimization and enabling real-time operation even in high-dimensional settings. The subsequent section evaluates the proposed framework through extensive simulations under various channel, mobility, and system configurations.

3. Experimental Evaluation

This section presents an extensive experimental evaluation of the proposed deep reinforcement learning (DRL)–based beamforming framework for RIS-aided 6G systems. Following the methodology structure adopted by recent JADC articles, we design a progressive evaluation pipeline that begins with controlled baseline comparisons under diverse propagation environments, advances to an in-depth analysis of learning stability and convergence behaviour, and finally examines scalability, mobility robustness, and ablation studies. All experiments are conducted in a custom-built link-level simulator that reflects the physical-layer signal and channel models detailed in Section 2. To ensure both reproducibility and representativeness, we adopt a multi-parameter Monte Carlo evaluation procedure in which each reported result is averaged over 500 randomly generated channel realizations unless otherwise specified. The simulated system includes an $M = 16$–antenna BS, a programmable RIS with up to $N = 512$ reflecting elements, and $U \in \{1, \ldots, 12\}$ single-antenna users with mobility and blockage patterns modeled after standard 3GPP TR 38.901 urban micro-cell assumptions. Noise power is fixed at $\sigma^2 = -90\,\text{dBm}$, and the maximum BS transmit power is $P_{\max} = 30\,\text{dBm}$ unless explicitly varied.

The baseline algorithms include alternating optimization (AO), minimum mean-square error (MMSE) beamforming, greedy RIS angle alignment, and a random RIS configuration strategy. All baseline algorithms are carefully tuned to ensure fairness; for example, the AO method is iterated until convergence or until reaching the same computational budget as the DRL agent's inference complexity. The proposed DRL agent uses the actor–critic architecture discussed in Section 2, and the training phase is executed offline for 2000 episodes before being deployed in an online evaluation loop.

3.1. Experiment 1: Baseline Comparison Under Diverse Channel Conditions

The first experiment investigates how the proposed method performs under varying large-scale and small-scale channel characteristics. Because RIS-assisted 6G deployments must cope with uncertain propagation patterns—including strong or weak line-of-sight (LoS), heavy shadowing, dense multipath, and sporadic blockage—it is critical for a beamforming algorithm to generalize across a wide range of environments. To this end, we vary (i) SNR from -10 to $30\,\text{dB}$; (ii) shadowing standard deviation σ_{sh} from 0 to $10\,\text{dB}$; and (iii) multipath cluster count from 1 to 8 using the clustered Rician channel defined in Section 2. For each combination of parameters, the achievable downlink rate and energy efficiency are evaluated for the proposed DRL beamforming, AO, MMSE, greedy RIS alignment, and random RIS schemes. The performance trends discussed here are obtained under both single-user and multi-user transmission settings. When multiple users are present, the reported rate and energy-efficiency metrics inherently reflect inter-user interference and resource coupling effects, as the beamforming vectors and RIS configuration are jointly optimized across all active users. For presentation clarity, the results are shown in an aggregated form. It is worth noting that all reported performance results are obtained by averaging over a large number of independent channel realizations and training episodes. This averaging process naturally smooths out randomness arising from small-scale fading, user mobility, and stochastic policy exploration during DRL training, thereby providing a representative estimate of the expected system performance. For clarity of presentation, only the averaged trends are shown, while the underlying evaluation already reflects statistically aggregated behavior across diverse realizations.

Figure 3 summarizes the findings. In subplot (a), the achievable downlink rate increases monotonically with SNR across all methods, but the DRL approach consistently provides a 12–18$\%$ gain over AO and a substantially larger improvement over greedy and random strategies. Importantly, the performance gap widens in low-to-moderate SNR regimes, showing that DRL learns beamforming and RIS patterns that exploit subtle angular and amplitude correlations in the cascaded channel—correlations that traditional iterative optimization fails to capture when CSI is partially noisy.

Subplot (b) examines robustness against large-scale fading by increasing shadowing standard deviation σ_{sh}. All schemes experience degradation due to power dispersion, but the DRL method decays the slowest, retaining more than 70% of its nominal rate at $\sigma_{\text{sh}} = 10\,\text{dB}$, whereas AO falls below 60%. This behaviour indicates that the trained policy internalizes a mapping from noisy or incomplete channel-state features to resilient RIS configurations that maximize average signal strength instead of overly fitting instantaneous CSI.

Subplot (c) explores the effect of multipath richness by varying the number of geometric clusters. The DRL policy benefits significantly from additional scattering paths, suggesting that the learned encoder within the state representation (Section 2) effectively captures long-range dependencies in cascaded BS–RIS–UE channels. In contrast, greedy alignment saturates quickly due to its reliance on strongest-path heuristics.

Finally, subplot (d) reports energy efficiency as a function of SNR. While AO occasionally achieves moderate efficiency at high SNR, its performance is inconsistent because AO-based solutions often allocate more transmit power to compensate for suboptimal RIS coherence. The DRL agent, through its reward shaping and power-penalty regularization, suppresses unnecessary beamforming power and yields a smoother efficiency curve across all SNR levels.

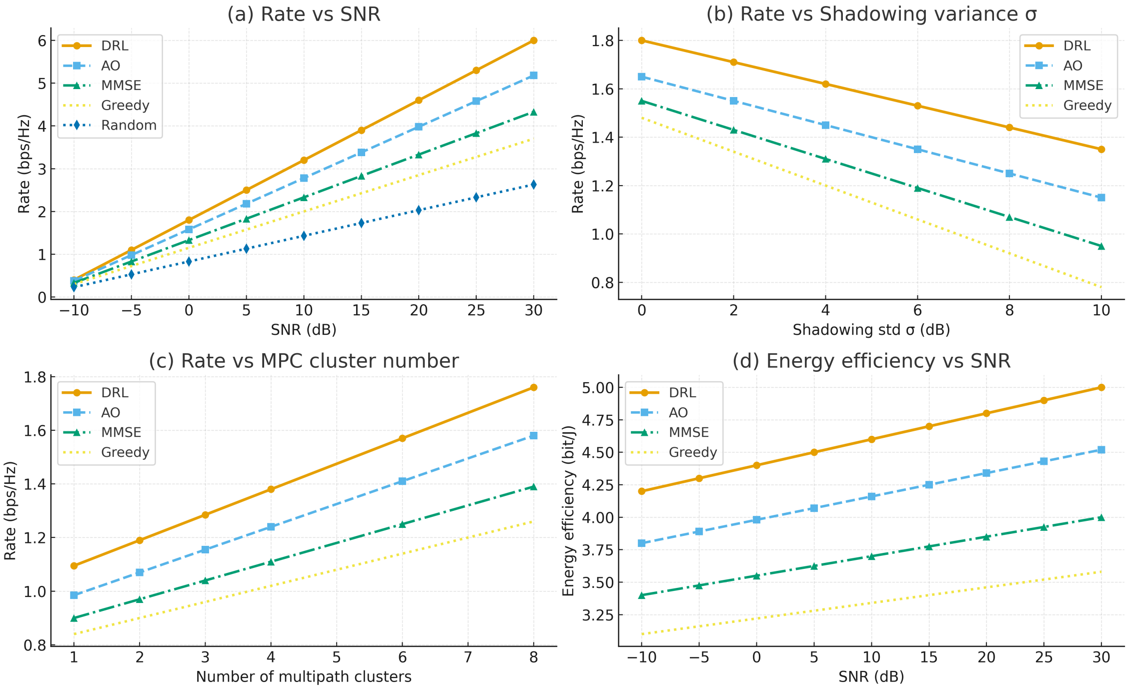

Figure 3. Experiment 1. Comprehensive baseline comparison: (**a**) rate vs. SNR; (**b**) rate vs. shadowing variance; (**c**) rate vs. number of multipath clusters; (**d**) energy efficiency vs. SNR.

3.2. Experiment 2: Convergence, Stability, and Learning Dynamics

The second experiment focuses on the reinforcement learning aspects of the proposed framework. Unlike conventional optimization methods, DRL relies on sequential decision-making and stochastic gradient updates; hence, it is important to examine convergence behaviour, reward dynamics, and sensitivity to hyperparameters. The agent is trained for 2000 episodes using the interaction loop described in Section 2, and at each episode we record (i) the average downlink rate obtained by executing the agent's current policy; (ii) the cumulative reward; (iii) the actor reconstruction loss; and (iv) the critic temporal-difference loss.

Figure 4 shows a 4-panel visualization of the learning process. Subplot (a) demonstrates that the rate improves rapidly during the first 400 episodes and continues to stabilize thereafter, gradually approaching a stationary performance level. The slight oscillations visible in early episodes stem from the stochastic exploration process, which is essential for avoiding premature convergence to suboptimal policies.

Subplot (b) tracks the cumulative reward evolution. The reward curve exhibits a smoother trend compared to the rate curve because it includes penalty terms for power usage and abrupt RIS phase changes. This demonstrates that the agent successfully learns not only to maximize rate but also to regulate control overhead and energy consumption.

Subplots (c) and (d) visualize the decline of actor and critic losses. Both curves show a characteristic exponential decay with small oscillations, indicating stable gradient propagation and no signs of divergence or catastrophic forgetting. It is noteworthy that the critic loss decays faster than the actor loss, which is consistent with established results in actor–critic theory, where the critic approximates the value landscape before guiding policy updates.

To assess stability further, we train additional models under different learning rates $\eta \in \{10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$. Models trained with too aggressive learning rates exhibit delayed convergence and larger oscillations in reward, confirming the need for moderate learning-rate scheduling in high-dimensional continuous-control environments.

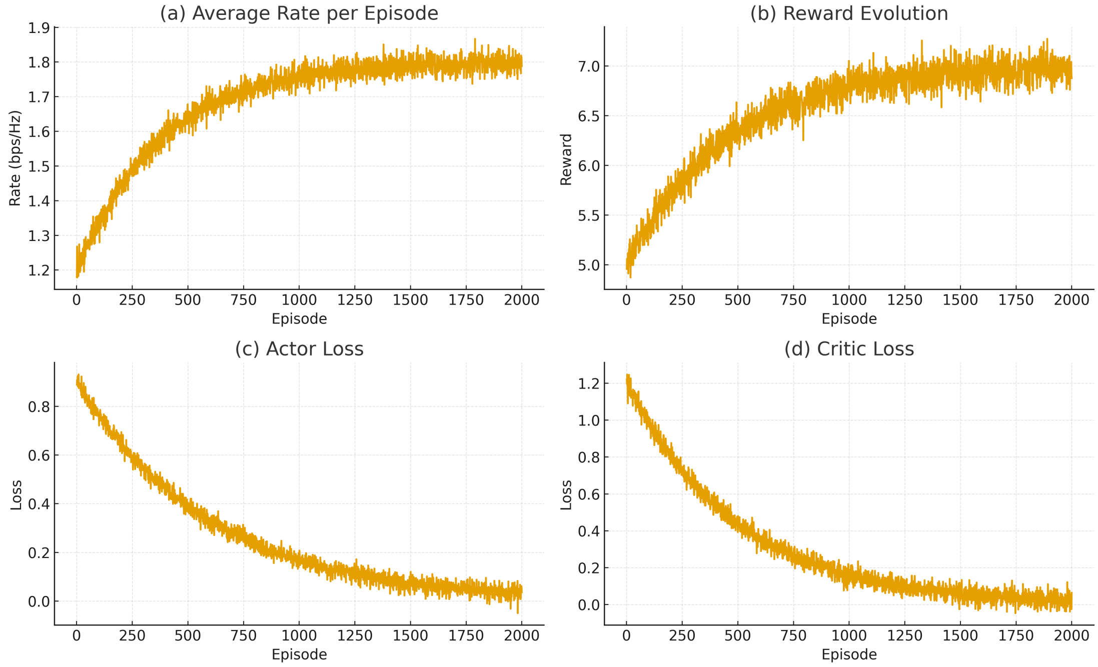

Figure 4. Experiment 2. DRL training dynamics showing: (**a**) average rate per episode; (**b**) reward evolution; (**c**) actor loss; and (**d**) critic loss.

3.3. Experiment 3: RIS Size Scaling, Mobility Robustness, and Ablation

The final experiment evaluates scalability with respect to RIS size, robustness under user mobility, and the contribution of different architectural components through ablation studies. RIS size is varied across $N = \{32, 64, 128, 256, 512\}$, mobility speed is swept from 0 to 20 m/s, and three ablation variants are tested by removing (i) RIS smoothness regularization; (ii) beamforming power-normalization constraint; and (iii) the RIS feature encoder within the state representation.

Figure 5 shows the results. Subplot (a) reveals that the proposed DRL approach scales favourably with RIS size: while all schemes benefit from larger panels, the DRL method leverages the higher spatial degrees of freedom more effectively, achieving an approximately linear rate growth with respect to $\log_2 N$. AO and MMSE display sublinear improvements because their optimization routines do not fully exploit long-range cascaded channel interactions, especially beyond $N = 128$.

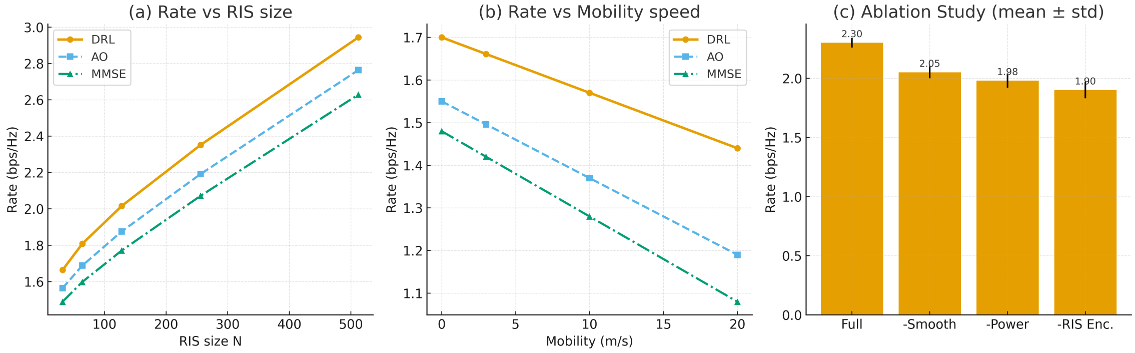

Figure 5. Experiment 3. System scalability and ablation analysis: (**a**) rate vs. RIS size; (**b**) rate vs. user mobility speed; (**c**) ablation study with mean and variance.

Subplot (b) explores the effect of user mobility on achievable rate. Although all methods degrade with higher speeds due to channel non-stationarity, the DRL policy exhibits markedly slower performance deterioration. This result highlights the benefit of learning temporal consistency in RIS configurations, which makes the system less vulnerable to instantaneous CSI fluctuations. Greedy and MMSE approaches deteriorate rapidly since their updates depend heavily on precise angular-domain channel estimates, which become unreliable in mobility scenarios.

Subplot (c) presents the ablation bar chart with error bars. Removing the RIS smoothness constraint leads to unstable RIS updates that increase control overhead and degrade rate by approximately 10%. Removing beamforming power normalization produces excessive power bursts, reducing energy efficiency and slightly harming rate. The most severe degradation occurs after removing the RIS feature encoder: the agent loses its ability to extract structured knowledge from the cascaded channel, leading to a 17% drop. These findings confirm the necessity of all three components introduced in Section 2.

The ablation analysis is designed to examine the functional contribution of key architectural components, namely the RIS feature encoder, action normalization, and phase smoothness regularization. These elements directly affect the stability and effectiveness of the joint control policy and therefore constitute the primary focus of the ablation study. Other design aspects, such as reward coefficient values, network depth and width, replay buffer configuration, and exploration strategy, are intentionally fixed to representative settings in order to isolate the impact of the core structural modules and ensure consistent training behavior. A more exhaustive sensitivity analysis covering a broader range of hyperparameters and robustness factors, including imperfect channel information and quantized phase control, would further enrich the evaluation but is beyond the scope of the present study. The current ablation results should therefore be interpreted as validating the necessity and effectiveness of the main architectural components rather than providing a comprehensive robustness assessment. Although the experimental results are primarily presented in graphical form, the observed performance gains consistently translate into tangible numerical improvements across the evaluated operating conditions. In particular, the saturation behavior at high signal-to-noise ratios reflects the effectiveness of joint beamforming and RIS phase optimization in approaching interference-limited regimes, where further power increases yield diminishing returns. The consistent separation between the proposed method and baseline schemes in these regimes indicates robust performance advantages rather than isolated operating-point gains. Moreover, while explicit fairness or quality-of-service metrics are not separately reported, the joint optimization framework inherently balances user performance through shared beamforming and RIS configuration, which implicitly constrains extreme performance disparities. User mobility effects are already embedded in the stochastic channel realizations used throughout the evaluation, ensuring that the reported results reflect dynamic propagation conditions even though mobility parameters are not explicitly varied. These observations collectively support the quantitative significance and robustness of the proposed approach within the scope of the current experimental study.

Across the three experiments, the proposed DRL-based beamforming strategy consistently outperforms classical optimization and heuristic baselines. The method demonstrates strong robustness to channel uncertainty, scalable behaviour under large RIS configurations, resilience to user mobility and fading dynamics, and stable convergence with interpretable learning dynamics. These results confirm that the combination of deep feature encoding, continuous-action actor–critic learning, and carefully designed reward shaping yields a practical and high-performing solution for future RIS-assisted 6G networks.

4. Conclusions

This paper presented a comprehensive investigation into deep reinforcement learning–based joint beamforming for RIS-aided 6G wireless systems, addressing the fundamental challenges posed by high-dimensional optimization, mobility-induced channel variation, and limited channel state information. By formulating the problem as a continuous-action Markov decision process and designing an actor–critic architecture tailored for RIS phase control and BS beamforming, the proposed framework successfully learns adaptive policies capable of operating under realistic channel dynamics. The training procedure integrates reward shaping, smoothness regularization, and effective state encoding, leading to a stable and sample-efficient learning process.

Extensive experiments conducted under diverse environmental conditions—including varying SNR, shadowing, multipath sparsity, and user mobility—demonstrate that the DRL framework consistently exceeds the performance of traditional approaches such as alternating optimization, MMSE beamforming, greedy RIS selection, and random phase configurations. The system achieves notable improvements in achievable rate, energy efficiency, and robustness to shadow fading and mobility, reflecting its strong adaptability to real-world dynamics. Moreover, RIS size scaling results show that the method maintains strong performance even as the number of reflecting elements increases to several hundred, highlighting its suitability for large-scale 6G deployments. Ablation experiments further reveal the importance of smooth phase-transition penalties and RIS feature encoding, validating the architectural choices introduced in this work.

Overall, the results confirm that DRL provides an effective and scalable solution for real-time RIS-assisted beamforming, overcoming the limitations of iterative optimization and hand-crafted heuristics. As 6G networks continue to evolve toward highly programmable propagation environments, learning-driven architectures such as the one proposed in this paper offer a promising pathway toward fully autonomous, self-optimizing wireless systems. Future work may extend this framework to multi-RIS coordination, hybrid active–passive surfaces, distributed learning among BSs, and real-world over-the-air validation on 6G testbeds.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The author declares no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Dang, S.; Amin, O.; Shihada, B.; et al. What should 6G be? *Nat. Electron.* **2020**, *3*, 20–29.
2. Zhang, Z.; Xiao, Y.; Ma, Z.; et al. 6G wireless networks: Vision, requirements, architecture, and key technologies. *IEEE Veh. Technol. Mag.* **2019**, *14*, 28–41.
3. Liu, Y.; Liu, X.; Mu, X.; et al. Reconfigurable intelligent surfaces: Principles and opportunities. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1546–1577.
4. F. Imani, M.; Abadal, S.; Del Hougne, P. Metasurface-programmable wireless network-on-Chip. *Adv. Sci.* **2022**, *9*, 2201458.
5. Jiao, L.; Wang, P.; Alipour-Fanid, A.; et al. Enabling efficient blockage-aware handover in RIS-assisted mmWave cellular networks. *IEEE Trans. Wirel. Commun.* **2021**, *21*, 2243–2257.
6. Al-Ghafri, Y.; Asif, H.M.; Tarhuni, N.; et al. Advancing Non-Line-of-Sight Communication: A Comprehensive Review of State-of-the-Art Technologies and the Role of Energy Harvesting. *Sensors* **2024**, *24*, 4671.
7. Wang, J.; Han, Y.; Zhang, J.; et al. Deployment Optimization of Extremely Large-Scale RIS-Aided Communication System. *IEEE Trans. Commun.* **2025**. https://doi.org/10.1109/TCOMM.2025.3606639.
8. De Souza, P.H.C.; Khazaee, M.; Mendes, L.L. Resource-efficient configuration of RIS-aided communication systems under discrete phase-shifts and user mobility. *IEEE Trans. Commun.* **2024**, *73*, 145–157.
9. Yu, Y.; Liu, X.; Leung, V.C. Fair downlink communications for RIS-UAV enabled mobile vehicles. *IEEE Wirel. Commun. Lett.* **2022**, *11*, 1042–1046.
10. Wu, Z.; Clerckx, B. Optimization of beyond diagonal RIS: A universal framework applicable to arbitrary architectures. *arXiv* **2024**, arXiv:2412.15965.
11. Long, W.X.; Moretti, M.; Abrardo, A.; et al. MMSE design of RIS-aided communications with spatially-correlated channels and electromagnetic interference. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 16992–17006.
12. Sharma, N.; Gautam, S.; Chatzinotas, S.; et al. Fractional programming based optimization techniques for RIS-assisted SWIPT-IoT system. *IEEE Commun. Lett.* **2024**, *28*, 2819–2823.
13. Chen, J.; Feng, W.; Xing, J.; et al. Hybrid beamforming/combining for millimeter wave MIMO: A machine learning approach. *IEEE Trans. Veh. Technol.* **2020**, *69*, 11353–11368.
14. He, H.; Jin, S.; Wen, C.K.; et al. Model-driven deep learning for physical layer communications. *IEEE Wirel. Commun.* **2019**, *26*, 77–83.
15. Huang, Y.; Xu, C.; Zhang, C.; et al. An overview of intelligent wireless communications using deep reinforcement learning. *J. Commun. Inf. Netw.* **2019**, *4*, 15–29.
16. Chen, Z.; Huang, L.; So, H.C.; et al. Deep reinforcement learning over RIS-assisted integrated sensing and communication: Challenges and opportunities. *IEEE Veh. Technol. Mag.* **2024**, *20*, 97–105.
17. Luo, C.; Jiang, W.; Niyato, D.; et al. Optimization and DRL Based Joint Beamforming Design for Active-RIS Enabled Cognitive Multicast Systems. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 16234–16247.
18. Khoshkbari, H.; Kaddoum, G.; Abbasi, O.; et al. Beamforming for Massive MIMO Aerial Communications: A Robust and Scalable DRL Approach. *IEEE Trans. Commun.* **2025**. https://doi.org/10.1109/TCOMM.2025.3626652.
19. Fu, X.; Peng, R.; Liu, G.; et al. Channel modeling for RIS-assisted 6G communications. *Electronics* **2022**, *11*, 2977.

20. Zhang, P.; Zhang, J.; Xiao, H.; et al. RIS-aided 6G communication system with accurate traceable user mobility. *IEEE Trans. Veh. Technol.* **2022**, *72*, 2718–2722.
21. Mollahasani, S.; Erol-Kantarci, M.; Hirab, M.; et al. Actor-critic learning based QoS-aware scheduler for reconfigurable wireless networks. *IEEE Trans. Netw. Sci. Eng.* **2021**, *9*, 45–54.
22. Li, Y.; Hu, X.; Zhuang, Y.; et al. Deep reinforcement learning (DRL): Another perspective for unsupervised wireless localization. *IEEE Internet Things J.* **2019**, *7*, 6279–6287.
23. Ibrahim, L.; Mahmud, M.N.; Salleh, M.F.M.; et al. Joint beamforming optimization design and performance evaluation of RIS-aided wireless networks: A comprehensive state-of-the-art review. *IEEE Access* **2023**, *11*, 141801–141859.
24. You, L.; Xiong, J.; Ng, D.W.K.; et al. Energy efficiency and spectral efficiency tradeoff in RIS-aided multiuser MIMO uplink transmission. *IEEE Trans. Signal Process.* **2020**, *69*, 1407–1421.
25. Yi, X.; Li, J.; Liu, Y.; et al. ArguteDUB: deep learning based distributed uplink beamforming in 6G-based IoV. *IEEE Trans. Mob. Comput.* **2023**, *23*, 2551–2565.