

Article

A Deep Hybrid Intelligent Framework for Dynamic Downlink Power Allocation in Cell-Free Massive MIMO Systems

Hussein A. Jasim ^{1,2,*}, Mohd Fadlee A Rasid ^{1,3}, Fazirulhisyam Hashim ^{1,3}  and Syamsiah Mashohor ¹

¹ Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Selangor, Malaysia; fadlee@upm.edu.my (M.F.A.R.); fazirul@upm.edu.my (F.H.); syamsiah@upm.edu.my (S.M.)

² Computer Sciences Department, College of CSIT, University of Basrah, Basrah 61001, Iraq

³ Wireless and Photonics Networks Research Centre (WiPNET), Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Selangor, Malaysia

* Correspondence: gs61712@student.upm.edu.my or hussein.jasim@uobasrah.edu.iq

Abstract

Cell-free massive multiple-input multiple-output (CF-mMIMO) systems have emerged as a promising architecture for beyond-5G wireless networks because they can provide user-centric coverage, improved spectral efficiency, and reduced cell-boundary limitations. However, dynamic downlink power allocation remains challenging due to user mobility, time-varying channel conditions, interference coupling, and the need to maintain Quality of Service (QoS) under practical transmit-power constraints. This paper proposes a Deep Hybrid Intelligent (DHI) framework for dynamic downlink power allocation in CF-mMIMO systems. The proposed framework integrates Soft Actor–Critic (SAC) reinforcement learning with three power-control strategies: DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate. The SAC agent learns adaptive power-allocation policies from the network state, while L-BFGS-B refinement is applied to the Max-Product and Max-Sum-Rate strategies to improve the power-allocation decisions under bounded transmit power. The framework is evaluated using a CF-mMIMO scenario with 64 access points and 32 pieces of user equipment distributed over a $1000 \times 1000 \text{ m}^2$ area. The simulation results show that DHI-Max-Sum-Rate achieves the highest sum spectral efficiency, while DHI-Max-Min provides the strongest QoS-oriented performance with a QoS satisfaction rate of 93.75%. In addition, DHI-Max-Product and DHI-Max-Sum-Rate achieve mean computational times of 0.0690 s and 0.0696 s, respectively, compared with 0.63 s for the DDPG benchmark. These results demonstrate that the proposed DHI framework provides an adaptive and computationally efficient solution for QoS-aware downlink power allocation in dynamic CF-mMIMO networks.

Keywords: cell-free massive MIMO; downlink power allocation; deep hybrid intelligent framework; soft actor–critic; reinforcement learning; spectral efficiency; QoS satisfaction; SINR; L-BFGS-B; dynamic resource allocation



Academic Editor: Zuhaib Khan

Received: 5 May 2026

Revised: 24 May 2026

Accepted: 27 May 2026

Published: 2 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Multiple-input multiple-output (MIMO) technology has become one of the main foundations of modern wireless communication systems because it can improve spectral efficiency, link reliability, and network capacity by exploiting spatial diversity and multiplexing [1]. In conventional massive MIMO systems, a large number of antennas are usually deployed at a centralized base station to serve many users within a cell. Although

this architecture can provide high throughput, its performance may degrade for cell-edge users due to path loss, shadowing, inter-cell interference, and non-uniform service quality, as shown in Figure 1. These limitations become more critical in dense beyond-5G networks, where a large number of users require reliable connectivity, high spectral efficiency, and stable Quality of Service (QoS) [2].

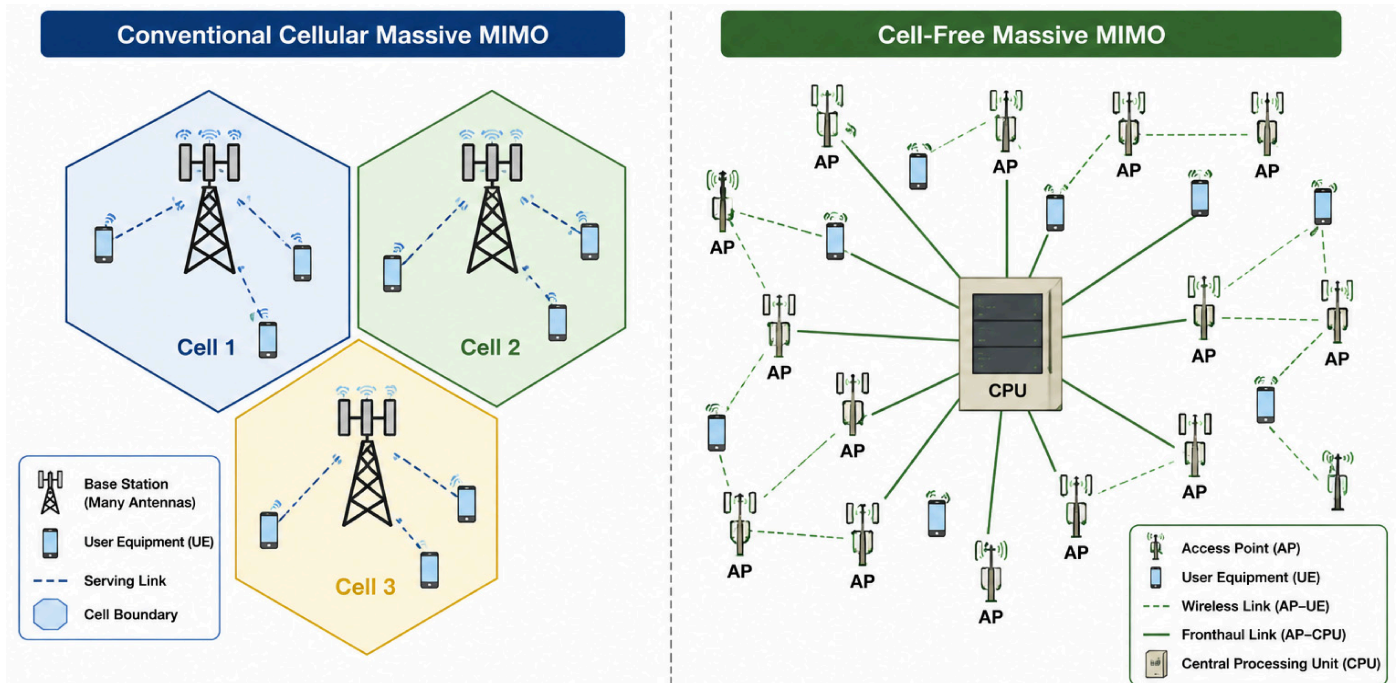


Figure 1. Comparison of the conventional massive MIMO and cell-free massive MIMO.

Cell-free massive MIMO (CF-mMIMO) has been introduced as a promising distributed architecture to overcome the limitations of conventional cell-based networks [3]. In CF-mMIMO, a large number of geographically distributed access points (APs) jointly serve all user equipment (UE) over the same time-frequency resources with the assistance of a central processing unit (CPU). By removing cell boundaries, CF-mMIMO can provide more uniform coverage, reduce the cell-edge effect, and improve the reliability of user-centric transmission [4]. However, these advantages depend strongly on efficient power allocation, especially in the downlink, where the transmit power must be distributed among users while controlling interference and maintaining QoS requirements [5].

Dynamic downlink power allocation in CF-mMIMO remains a challenging problem. The channel conditions vary continuously due to user mobility, shadow fading, changing AP-UE distances, and interference coupling among users [6]. Classical optimization-based power-control methods, such as max-min fairness, product-based SINR optimization, and sum-rate maximization, can improve specific performance objectives, but they often require repeated optimization and may not adapt efficiently to fast network changes [7]. Moreover, purely throughput-oriented strategies may improve the total spectral efficiency while reducing the service quality of weak users, whereas QoS-oriented strategies may protect weak users but sacrifice aggregate throughput. Therefore, a flexible power-allocation framework is needed to adapt to dynamic network conditions while balancing spectral efficiency, SINR behavior, QoS satisfaction, and computational cost.

Recently, artificial intelligence-based resource management has attracted significant attention for beyond-5G and 6G wireless networks. Deep reinforcement learning (DRL) is particularly suitable for dynamic wireless optimization because it can learn control policies from interaction with the environment instead of solving a full optimization problem

from the beginning at every time step [8]. Recent studies on AI-driven wireless edge networks and reinforcement learning-based CF-mMIMO access-point clustering have shown the potential of learning-based methods for scalable resource management. In parallel, reconfigurable intelligent surface (RIS)-assisted MIMO systems have also been studied as an emerging 5G/6G technology for improving wireless propagation and resource allocation [9]. However, RIS-based studies mainly focus on modifying the propagation environment, while the present work focuses on dynamic downlink power allocation in CF-mMIMO using a hybrid learning-and-optimization framework [10].

To address the above challenges, this paper proposes a Deep Hybrid Intelligent (DHI) framework for dynamic downlink power allocation in CF-mMIMO systems. The proposed framework integrates Soft Actor–Critic (SAC) reinforcement learning with three power-control strategies: DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate. The SAC agent learns adaptive power-allocation decisions from the network state, while the optimization objectives guide the allocation process toward different operating priorities. DHI-Max-Min emphasizes QoS-oriented service reliability, DHI-Max-Product provides a balanced trade-off between SINR improvement and spectral-efficiency performance, and DHI-Max-Sum-Rate prioritizes aggregate throughput. In addition, L-BFGS-B refinement is applied to the Max-Product and Max-Sum-Rate strategies to improve the SAC-generated power allocation under bounded transmit-power constraints. The main contributions of this paper are summarized as follows:

- A DHI-based dynamic downlink power-allocation framework is proposed for CF-mMIMO systems under user mobility, time-varying channels, interference, and transmit-power constraints.
- The proposed framework integrates SAC-based adaptive learning with three downlink power-control objectives, namely DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate, to support different QoS and spectral-efficiency requirements.
- An optimization refinement stage based on L-BFGS-B is incorporated into the DHI-Max-Product and DHI-Max-Sum-Rate strategies to improve the power-allocation decisions while respecting bounded transmit-power constraints.
- A QoS-aware reward and evaluation structure is adopted to assess the relationship between spectral efficiency, SINR, QoS satisfaction rate, outage probability, and computational time.

The remainder of this paper is organized as follows. Section 2 reviews related work on CF-mMIMO power allocation, optimization-based approaches, DRL-based resource management, and recent AI/RIS-assisted wireless studies. Section 3 presents the system model, mathematical formulation, SAC-based learning framework, and optimization objectives. Section 4 describes the simulation setup and parameter configuration. Section 5 discusses the simulation results, including spectral efficiency, SINR, QoS satisfaction, computational time, and convergence behavior. Section 6 discusses limitations and practical considerations, and Section 7 concludes the paper.

2. Related Work

Downlink power allocation in cell-free massive multiple-input multiple-output (CF-mMIMO) systems has received considerable attention because it directly affects spectral efficiency, interference control, Quality of Service (QoS), and user-centric coverage. Unlike conventional cellular massive MIMO, CF-mMIMO removes cell boundaries by allowing many distributed access points (APs) to jointly serve user equipment (UE) over the same time-frequency resources. This distributed structure improves coverage uniformity and reduces the cell-edge effect, but it also makes power allocation more challenging because

AP–UE channel conditions, interference levels, and QoS requirements vary dynamically across the network [11].

Early power-allocation methods mainly relied on classical optimization techniques. Water-filling and convex-optimization-based approaches have been widely used to distribute power efficiently under ideal or semi-ideal channel assumptions. These methods can provide strong mathematical solutions for static or slowly varying systems, but their performance depends heavily on accurate channel state information (CSI) and repeated optimization. In practical CF-mMIMO systems, where users may move and channel conditions change over time, repeatedly solving the optimization problem can become computationally expensive and less suitable for real-time operation [12].

Max-min-based power control has also been widely studied to improve user fairness and protect weak users. In this approach, the objective is to improve the minimum user performance so that users with poor channel conditions are not severely degraded. Although this method is useful for QoS-oriented operation, it may reduce the total spectral efficiency because more resources are allocated to weaker users. On the other hand, sum-rate maximization aims to improve the total system throughput by allocating power to users or links that contribute more strongly to the aggregate spectral efficiency. However, this may reduce QoS satisfaction for users with poor channel conditions. Product-based SINR optimization provides an intermediate strategy by improving the overall SINR distribution while avoiding extreme imbalance among users. Therefore, max-min, max-product, and max-sum-rate strategies represent different trade-offs among QoS reliability, SINR improvement, and spectral-efficiency maximization [13].

Recently, reinforcement learning (RL) and deep reinforcement learning (DRL) have become promising tools for wireless resource management. Unlike conventional optimization methods, DRL can learn adaptive control policies through interaction with the environment and can reuse learned experience when network conditions change. Q-learning and deep Q-network-based methods have been applied to power control and resource allocation in wireless networks. However, discrete-action RL methods may be less suitable for downlink power allocation because transmit power is naturally continuous. For this reason, actor–critic methods such as Soft Actor–Critic (SAC) are attractive because they can handle continuous action spaces and support exploration through entropy regularization [14].

Recent AI-based studies for beyond-5G and 6G networks have also highlighted the importance of pushing learning and decision-making capabilities closer to the wireless edge. Such works show that AI can support integrated sensing, communication, computation, and distributed resource management [15]. In CF-mMIMO, access-point clustering using conventional and federated multi-agent reinforcement learning has also been investigated, showing the relevance of distributed learning for scalable cell-free networks [16]. These studies support the general direction of AI-driven wireless resource allocation, but many of them focus on clustering, network-edge intelligence, or general resource management rather than dynamic downlink power allocation with explicit QoS-aware evaluation.

In parallel, reconfigurable intelligent surface (RIS)-assisted MIMO has become an important research direction for 5G and 6G systems. RIS can improve the wireless propagation environment by controlling signal reflection, phase shifts, or operating modes. Recent RIS-assisted studies have investigated BER minimization without CSI and joint operating-mode and resource-allocation optimization in wireless-powered multi-user systems [17,18]. These studies are relevant because they show how 5G/6G systems can use intelligent control to improve wireless performance. However, RIS-assisted works mainly focus on modifying the propagation channel, while the present paper focuses on adaptive downlink power allocation in CF-mMIMO systems.

Although the above studies have contributed significantly to wireless power control and AI-based resource management, several limitations remain. Classical optimization methods may be computationally demanding under dynamic network conditions. Fairness-oriented methods may reduce aggregate throughput, while throughput-oriented methods may reduce QoS satisfaction for weaker users. DRL-based approaches can improve adaptability, but many existing works do not clearly combine learning-based power allocation with multiple optimization objectives and post-learning refinement. Moreover, several studies do not provide sufficient discussion of QoS satisfaction, computational time, convergence behavior, or practical deployment feasibility [12,14,16].

To address these gaps, this paper proposes a Deep Hybrid Intelligent (DHI) framework for dynamic downlink power allocation in CF-mMIMO systems. The proposed framework integrates SAC-based learning with three power-control strategies: DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate. The DHI-Max-Min strategy emphasizes QoS-oriented service reliability, DHI-Max-Product provides a balanced SINR and spectral-efficiency behavior, and DHI-Max-Sum-Rate focuses on maximizing aggregate throughput. In addition, L-BFGS-B refinement is applied to the Max-Product and Max-Sum-Rate strategies to improve the power-allocation decisions under bounded transmit-power constraints. This hybrid structure allows the proposed framework to adapt to dynamic channel conditions while evaluating the trade-off among spectral efficiency, QoS satisfaction, SINR behavior, and computational cost. Table 1 presents a summary of the related work on power allocation and intelligent resource optimization in cell-free massive MIMO systems.

Table 1. Summary of related work.

Cit.	Year	Objective	Problem Addressed	Algorithm/Method	Limitation	DHI Response
[11]	2017	Introduce the cell-free massive MIMO concept.	Distributed AP cooperation and user-centric service.	Cell-free massive MIMO architecture.	Does not study dynamic downlink power allocation using SAC.	✓
[12]	2020	Improve scalable CF-mMIMO operation.	Scalability, cooperation, and power control.	Scalable CF-mMIMO modeling and optimization.	Mainly analytical and not focused on SAC-based dynamic control.	✓
[13]	2023	Provide CF-mMIMO modeling foundations.	Downlink/uplink modeling, interference, and spectral efficiency.	Mathematical CF-mMIMO framework.	Does not include DHI strategy selection or L-BFGS-B refinement.	✓
[14]	2019	Learn adaptive continuous control policies.	Continuous action-space decision-making.	Soft Actor–Critic.	Requires careful reward design and training stability.	✓
[15]	2023	Discuss AI-enabled 6G resource management.	Edge intelligence and wireless resource control.	AI-assisted network resource management.	Broad 6G focus, not specific to CF-mMIMO downlink power allocation.	✓
[16]	2023	Apply learning to CF-mMIMO organization.	AP clustering and scalable network management.	Multi-agent and federated reinforcement learning.	Focuses on AP clustering rather than downlink power allocation.	✓

Table 1. Cont.

Cit.	Year	Objective	Problem Addressed	Algorithm/Method	Limitation	DHI Response
[17]	2024	Improve wireless performance using RIS.	RIS-assisted communication and propagation control.	Dynamic RIS-assisted optimization.	Different system model from CF-mMIMO power allocation.	×
[18]	2025	Optimize RIS-assisted multi-user systems.	Wireless-powered RIS resource allocation.	RIS-assisted resource optimization.	Focuses on RIS resource allocation, not CF-mMIMO downlink power control.	×
[19]	2021	Maximize downlink sum spectral efficiency in dynamic CF-mMIMO.	Time-varying downlink power allocation and high WMMSE complexity.	DQN and DDPG with WMMSE comparison.	Does not include SAC-based DHI strategy selection, QoS-aware refinement, or L-BFGS-B optimization.	✓

3. DHI Model Description

This section presents the system model, mathematical formulation, SAC-based learning structure, DHI power-control strategies, and complexity analysis of the proposed downlink power-allocation framework for CF-mMIMO systems.

3.1. System Model and Mathematical Formulation

This study considers a downlink cell-free massive multiple-input multiple-output (CF-mMIMO) system consisting of M distributed access points (APs) and K single-antenna user equipment (UE). The APs are geographically distributed over the coverage area and connected to a central processing unit (CPU) through fronthaul links. The CPU coordinates the downlink transmission and power-allocation decisions, while all APs jointly serve the UE using the same time-frequency resources. In order to maintain consistency with the adopted simulation model and scalar SINR formulation, each AP–UE connection is represented by an effective scalar channel coefficient. This assumption corresponds to a single-antenna AP configuration or an equivalent scalar channel after local AP processing. Similar scalar-link representations are commonly used in CF-mMIMO power-control studies to evaluate distributed transmission, large-scale fading, and user-centric power allocation [19].

Let $h_{m,k}(t)$ denote the effective downlink channel coefficient between AP m and UE k at time step t . The channel coefficient is modeled as [20]

$$h_{m,k}(t) = \sqrt{\beta_{m,k}(t)}g_{m,k}(t), \quad (1)$$

where $\beta_{m,k}(t)$ represents the large-scale fading coefficient, including path loss and shadow fading, and $g_{m,k}(t)$ denotes the small-scale fading coefficient. Since the users are mobile, the AP–UE distance changes over time, and therefore the large-scale fading coefficient $\beta_{m,k}(t)$ is updated dynamically according to the user position and shadowing model.

The downlink transmit power allocated from AP m to UE k at time step t is denoted by $\rho_{m,k}(t)$, where

$$\rho_{m,k}(t) \geq 0. \quad (2)$$

The total transmit power of each AP is constrained by the maximum available AP power P_m^{max} , such that [21]

$$\sum_{k=1}^K \rho_{m,k}(t) \leq P_m^{max}, m = 1, 2, \dots, M. \quad (3)$$

If the implementation uses an effective UE-level power vector, the AP-level power coefficient can be written as

$$\rho_{m,k}(t) = \eta_{m,k}(t) p_k(t), \quad (4)$$

where $p_k(t)$ is the effective power assigned to UE k , and $\eta_{m,k}(t)$ is a non-negative AP-UE weighting coefficient that distributes the UE-level power across the serving APs.

The received downlink signal at UE k can be expressed as [19]

$$y_k(t) = \sum_{m=1}^M \sqrt{\rho_{m,k}(t)} h_{m,k}(t) S_k + \sum_{\substack{i=1 \\ i \neq k}}^K \sum_{m=1}^M \sqrt{\rho_{m,i}(t)} h_{m,i}(t) S_i + n_k(t), \quad (5)$$

where S_k is the intended data symbol for UE k , S_i is the interfering data symbol intended for UE i , and $n_k(t)$ is the additive noise at UE k , with variance σ_k^2 . The first term represents the useful signal received by UE k , while the second term represents multi-user interference caused by downlink transmission to other UE.

Accordingly, the downlink signal-to-interference-plus-noise ratio (SINR) of UE k is defined as [21]

$$\gamma_k(t) = \frac{|\sum_{m=1}^M \sqrt{\rho_{m,k}(t)} h_{m,k}(t)|^2}{\sum_{\substack{i=1 \\ i \neq k}}^K |\sum_{m=1}^M \sqrt{\rho_{m,i}(t)} h_{m,i}(t)|^2 + \sigma_k^2} \quad (6)$$

This expression defines the UE-level SINR $\gamma_k(t)$. Therefore, the notation SINR $R_{m,k}$ is not used as the final performance metric; instead, each UE is evaluated using the aggregated SINR $\gamma_k(t)$, which includes the useful contribution from distributed APs and the interference generated by transmissions to other users.

The spectral efficiency of UE k is calculated as [19]

$$SE_k(t) = (1 - \frac{\tau_p}{\tau_c}) \log_2(1 + \gamma_k(t)), \quad (7)$$

where τ_p is the pilot length and τ_c is the coherence block length. If pilot overhead is not explicitly considered in the simulation, the pre-log factor can be set to one, and the simplified expression becomes

$$SE_k(t) = \log_2(1 + \gamma_k(t)), \quad (8)$$

The total downlink spectral efficiency of the system is then given by

$$SE_{sum}(t) = \sum_{k=1}^K SE_k(t) \quad (9)$$

To evaluate the Quality of Service (QoS) performance, a minimum SINR threshold γ_{min} is defined. UE k is considered to satisfy the QoS requirement if

$$\gamma_k(t) \geq \gamma_{min} \quad (10)$$

The QoS satisfaction rate is computed as

$$Q_{sat}(t) = \frac{1}{K} \sum_{k=1}^K \Pi(\gamma_k(t) \geq \gamma_{min}) \times 100\% \quad (11)$$

where $\Pi(\cdot)$ is an indicator function equal to one when the condition is satisfied and zero otherwise. The outage probability is then given by

$$P_{out}(t) = 100\% - Q_{sat}(t) \quad (12)$$

In this work, QoS is treated as a QoS-aware performance requirement rather than an absolute hard guarantee for every UE at every time step. This is because dynamic channel variation, user mobility, and transmit-power limitations may prevent all users from simultaneously satisfying the SINR threshold. Therefore, QoS performance is evaluated using the QoS satisfaction rate and outage probability.

The general downlink power-allocation problem can be formulated as

$$\max_{\rho_{m,k}(t)} \sum_{k=1}^K SE_k(t) \quad (13)$$

subject to

$$\sum_{k=1}^K \rho_{m,k}(t) \leq P_m^{max}, \quad m = 1, 2, \dots, M, \quad (14)$$

$$\rho_{m,k}(t) \geq 0, \quad m = 1, 2, \dots, M, \quad k = 1, 2, \dots, K, \quad (15)$$

$$\gamma_k(t) \geq \gamma_{min}, \quad k = 1, 2, \dots, K. \quad (16)$$

However, because the QoS constraint may not always be feasible under dynamic channel and power-limited conditions, the proposed DHI framework handles QoS using a QoS-aware reward penalty and evaluates the final performance using QoS satisfaction rate and outage probability. This avoids claiming unrealistic full QoS guarantee while still guiding the learning model toward QoS-supportive power allocation. Figure 2 illustrates the architecture of the proposed DHI framework, which integrates deep reinforcement learning with hybrid optimization strategies for dynamic downlink power allocation in CF-mMIMO systems.

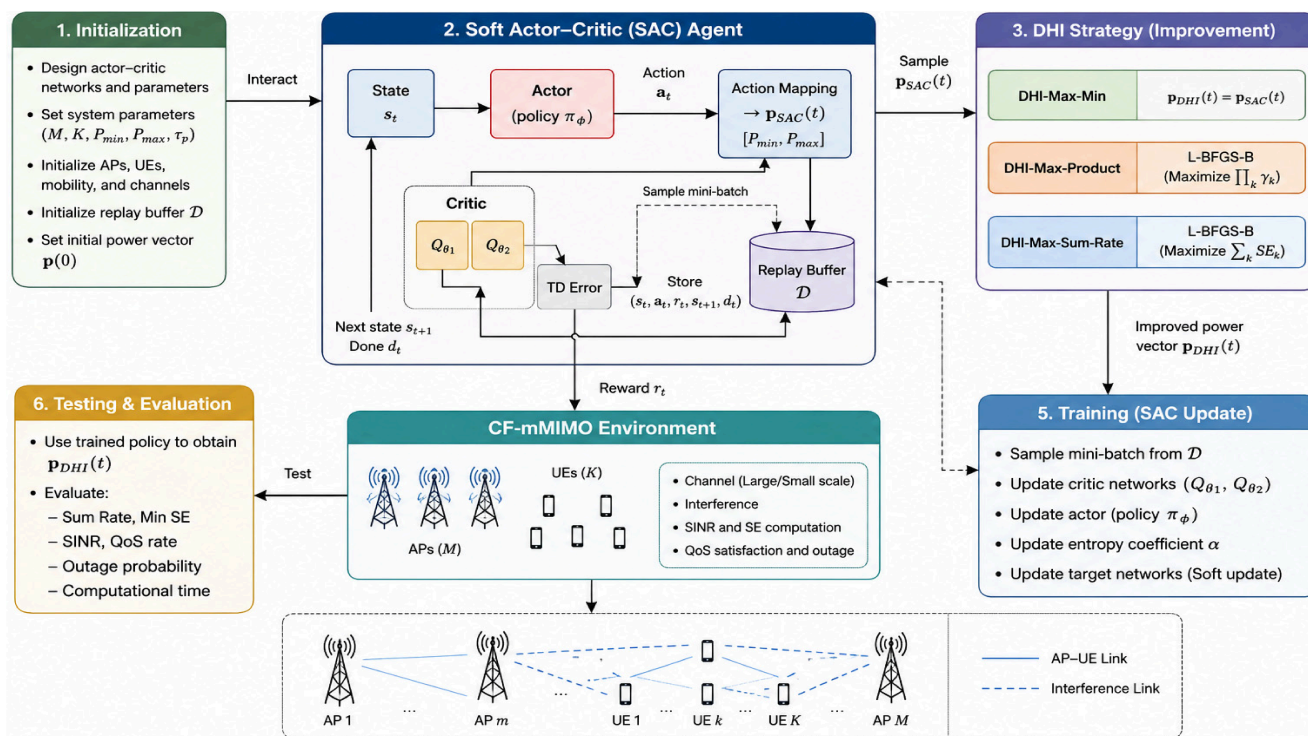


Figure 2. Architecture of the proposed DHI framework for dynamic downlink power allocation in CF-mMIMO systems.

3.2. SAC-Based Learning Algorithm Description

The proposed DHI framework uses the Soft Actor–Critic (SAC) algorithm to learn a dynamic downlink power-allocation policy for the considered CF-mMIMO system. SAC is an off-policy actor–critic reinforcement learning method designed for continuous action spaces, making it suitable for downlink power control because transmit power is a continuous variable rather than a discrete decision. In the proposed framework, the SAC agent observes the current network condition, selects the downlink power-allocation action, receives a reward based on spectral efficiency and QoS behavior, and updates its policy using experience replay.

At each time step t , the interaction between the SAC agent and the CF-mMIMO environment is formulated as a Markov decision process (MDP), defined by the tuple

$$m = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{r}, \gamma_{SAC}), \quad (17)$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the continuous action space, $\mathcal{P}$ represents the state-transition probability caused by channel evolution and UE mobility, $\mathcal{r}$ is the reward function, and γ_{SAC} is the discount factor.

3.2.1. State Space

The state vector contains the network information required by the SAC agent to make downlink power-allocation decisions. In this work, the state at time step t is defined as [22]

$$s_t = \left[\text{vec} \left(\tilde{\beta}_{m,k}(t), \tilde{u}(t), \tilde{\gamma}(t-1), \tilde{I}(t-1), \tilde{P}(t-1) \right) \right], \quad (18)$$

where $\tilde{\beta}_{m,k}(t)$ is the normalized large-scale fading coefficient between AP m and UE k , $\tilde{u}(t)$ contains the normalized UE location coordinates, $\tilde{\gamma}(t-1)$ is the normalized SINR vector from the previous time step, $\tilde{I}(t-1)$ represents the previous interference level of the UE, and $\tilde{P}(t-1)$ is the previous downlink power-allocation vector. The operator $\text{vec}(\cdot)$ converts the AP–UE fading matrix into a vector.

For a system with M APs and K UE, the state dimension is [23]

$$D_s = MK + 2K + K + K + K = MK + 5K \quad (19)$$

For the simulation setup used in this paper, where $M = 64$ and $K = 32$, the state dimension becomes

$$D_s = (64)(32) + 5(32) = 2208 \quad (20)$$

This explicit state representation improves reproducibility by clarifying which network parameters are used by the SAC agent during training and inference.

3.2.2. Action Space

The action generated by the SAC actor network is a continuous vector representing the downlink power-control decision for all UE. The action vector is defined as [24]

$$a_t = [a_1(t), a_2(t), \dots, a_K(t)]^T, \quad (21)$$

where each action component is bounded as

$$a_k(t) \in [-1, 1], k = 1, 2, \dots, K. \quad (22)$$

The normalized SAC action is then mapped to the physical transmit-power range as

$$p_k(t) = P_{min} + \frac{a_k(t) + 1}{2}(P_{max} - P_{min}), \quad (23)$$

where P_{min} and P_{max} denote the minimum and maximum allowable downlink transmit power, respectively. In this study, $P_{min} = 0$ mW and $P_{max} = 100$ mW. This mapping ensures that the SAC-generated power allocation always remains inside the permitted transmit-power range.

If AP-level power allocation is required, the effective UE-level power $p_k(t)$ is distributed across the serving APs using [25]

$$\rho_{m,k}(t) = \eta_{m,k}(t)p_k(t), \quad (24)$$

where $\eta_{m,k}(t)$ is a non-negative AP–UE weighting coefficient.

3.2.3. Reward Function

Since reward design is central to SAC-based power control, the proposed framework defines the reward as a QoS-aware objective that balances spectral efficiency, SINR performance, QoS satisfaction, and power-constraint violation. The general reward at time step t is formulated as

$$r_t^{(j)} = w_1 \hat{F}_j(t) + w_2 Q_{sat}(t) - w_3 V_{QoS}(t) - w_4 V_P(t), \quad (25)$$

where $j \in \{\text{Max} - \text{Min}, \text{Max} - \text{Product}, \text{Max} - \text{Sum} - \text{Rate}\}$ denotes the selected DHI strategy, $\hat{F}_j(t)$ is the normalized objective value of the selected strategy, $Q_{sat}(t)$ is the QoS satisfaction rate, $V_{QoS}(t)$ is the QoS violation penalty, $V_P(t)$ is the power-constraint violation penalty, and w_1, w_2, w_3, w_4 are non-negative weighting factors.

The QoS violation penalty is defined as [26]

$$V_{QoS}(t) = \frac{1}{K} \sum_{k=1}^K \max\left(0, \frac{\gamma_{min} - \gamma_k(t)}{\gamma_{min}}\right), \quad (26)$$

where γ_{min} is the minimum SINR threshold required for QoS satisfaction. This term penalizes the agent when the SINR of a UE falls below the QoS threshold.

The power violation penalty is expressed as

$$V_P(t) = \frac{1}{K} \sum_{k=1}^K \max\left(0, \frac{p_k(t) - P_{max}}{P_{max}}\right), \quad (27)$$

In the current implementation, the action mapping in Equation (23) already bounds the power values between P_{min} and P_{max} . Therefore, $V_P(t)$ is normally zero, but it is retained in the formulation to make the constraint-handling mechanism explicit.

The strategy-dependent objective $\hat{F}_j(t)$ is defined according to the selected DHI operating mode. For DHI-Max-Min, the objective emphasizes the weakest UE:

$$F_{Max-Min}(t) = \min_k SE_k(t) \quad (28)$$

For DHI-Max-Product, the objective improves the overall SINR distribution using a logarithmic product form:

$$F_{Max-Product}(t) = \frac{1}{K} \sum_{k=1}^K \log(\gamma_k(t) + \epsilon) \quad (29)$$

where ϵ a small positive is a constant used to avoid numerical instability.

For DHI-Max-Sum-Rate, the objective maximizes the total downlink spectral efficiency:

$$F_{Max-Sum-Rate}(t) = \sum_{k=1}^K SE_k(t) \quad (30)$$

The objective value is normalized before being used in the reward:

$$\hat{F}_j(t) = \frac{F_j(t) - F_j^{min}}{F_j^{max} - F_j^{min} + \epsilon} \quad (31)$$

This reward formulation allows the SAC agent to learn power-allocation policies that improve the selected objective while reducing QoS violations and maintaining bounded transmit power.

3.2.4. SAC Policy and Critic Updates

The SAC agent consists of an actor network, two critic networks, and their corresponding target networks. The actor network generates a stochastic policy $\pi_{\phi}(a_t|s_t)$, where ϕ denotes the actor parameters. The critic networks estimate the soft state-action value functions $Q_{\theta_1}(s_t|a_t)$ and $Q_{\theta_2}(s_t|a_t)$, where θ_1 and θ_2 denote the critic parameters.

For each transition (s_t, a_t, r_t, s_{t+1}) , the target soft value is computed as [22]

$$y_t = r_t + \gamma_{SAC} \left[\min_{i=1,2} Q_{\bar{\theta}_i}(s_{t+1}, a_{t+1}) - \alpha \log \pi_{\phi}(a_{t+1}|s_{t+1}) \right], \quad (32)$$

where $a_{t+1} \sim \pi_{\phi}(\cdot|s_{t+1})$, $\bar{\theta}_i$ are the target critic parameters, and α is the entropy-temperature coefficient.

The critic loss is defined as [27]

$$J_Q(\theta_i) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1})} \sim \mathcal{D} \left[(Q_{\theta_i}(s_t, a_t) - y_t)^2 \right], i = 1, 2, \quad (33)$$

where $\mathcal{D}$ is the replay buffer.

The actor loss is formulated as [24]

$$J_{\pi}(\phi) = \mathbb{E}_{s_t} \sim \mathcal{D}, a_t \sim \pi_{\phi} \left[\alpha \log \pi_{\phi}(a_t|s_t) - \min_{i=1,2} Q_{\theta_i}(s_t, a_t) \right] \quad (34)$$

The entropy coefficient α controls the balance between exploration and exploitation. A larger α encourages more exploration, while a smaller α encourages more deterministic power allocation. In this work, automatic entropy tuning is used to adjust α during training.

The target critic networks are updated using Polyak averaging:

$$\bar{\theta}_i \leftarrow \tau \theta_i + (1 - \tau) \bar{\theta}_i, \quad i = 1, 2, \quad (35)$$

where τ is the Polyak update coefficient.

3.2.5. Replay Buffer and Training Procedure

The SAC agent stores each interaction with the environment in a replay buffer as

$$(s_t, a_t, r_t, s_{t+1}, d_t), \quad (36)$$

where d_t is the terminal-state indicator. During training, mini-batches are sampled randomly from the replay buffer to update the actor and critic networks. This off-policy

learning mechanism improves sample efficiency and reduces the correlation between consecutive training samples.

The training process begins with environment initialization, including AP and UE deployment, channel generation, mobility initialization, and initial power setting. At each time step, the SAC actor selects a downlink power action, the environment computes the resulting SINR, spectral efficiency, QoS satisfaction rate, and reward, and the transition is stored in the replay buffer. Once the replay buffer contains enough samples, the critic networks, actor network, entropy coefficient, and target networks are updated.

The actor and critic networks are implemented as multilayer perceptrons. The network architecture is selected through hyperparameter optimization, where candidate architectures include [64,64], [256,256], [400,300], [128,256,128], and [256,256,256]. The search also includes the learning rate, batch size, discount factor, replay-buffer size, learning-start steps, training frequency, and Polyak coefficient.

3.2.6. Convergence Monitoring

To evaluate the convergence behavior of the proposed SAC-based DHI framework, the training performance is monitored using the moving average of the episode reward, sum spectral efficiency, QoS satisfaction rate, and critic loss. The moving average reward over a window of W episodes is calculated as

$$\bar{R}_e = \frac{1}{W} \sum_{i=e-W+1}^e R_i, \quad (37)$$

where R_i is the total reward of episode i . The policy is considered stable when the moving average reward and QoS satisfaction rate no longer show large fluctuations over consecutive training windows. Figure 3 illustrates the SAC action-generation mechanism used to transform the unbounded policy output into a bounded downlink power allocation vector within the predefined power constraints.

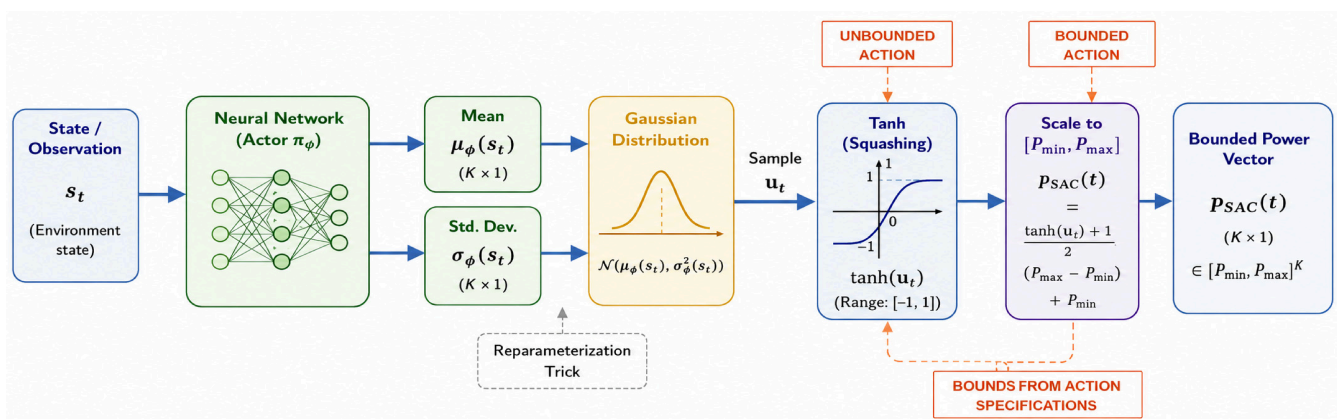


Figure 3. SAC action-generation mechanism for bounded downlink power allocation.

3.3. DHI Power-Control Strategies and L-BFGS-B Refinement

The proposed DHI framework combines SAC-based learning with three downlink power-control strategies: DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate. These strategies are designed to guide the learning model toward different operating objectives. DHI-Max-Min focuses on QoS-oriented service reliability, DHI-Max-Product improves the overall SINR balance among users, and DHI-Max-Sum-Rate maximizes the aggregate spectral efficiency of the CF-mMIMO system. Therefore, the proposed framework does not rely on a single fixed power-control objective; instead, it provides flexible operating modes according to the required network performance target.

3.3.1. DHI-Max-Min Strategy

The DHI-Max-Min strategy aims to improve the performance of the weakest UE by maximizing the minimum spectral efficiency among all users. This strategy is suitable for QoS-oriented operation because it prevents users with poor channel conditions from being severely degraded. The objective is formulated as [28]

$$\max_{P(t)} \min_{k \in \{1, \dots, K\}} SE_k(t), \quad (38)$$

subject to

$$P_{min} \leq p_k(t) \leq P_{max}, \quad k = 1, 2, \dots, K. \quad (39)$$

In this strategy, the SAC agent learns a power-allocation policy that improves the minimum user performance. However, because the objective prioritizes weak users, the total sum spectral efficiency may be lower than that achieved by throughput-oriented strategies. Therefore, DHI-Max-Min is mainly suitable for scenarios where QoS reliability and minimum user service are more important than maximum aggregate throughput.

3.3.2. DHI-Max-Product Strategy

The DHI-Max-Product strategy aims to improve the overall SINR distribution by maximizing the product of user SINR values. Direct multiplication of SINR values may cause numerical instability, especially when the number of users is large. Therefore, the logarithmic form is used [29]:

$$\max_{P(t)} \sum_{k=1}^K \log(\gamma_k(t) + \epsilon), \quad (40)$$

subject to

$$P_{min} \leq p_k(t) \leq P_{max}, \quad k = 1, 2, \dots, K, \quad (41)$$

where ϵ is a small positive constant used to avoid $\log(0)$. This objective encourages the model to improve the SINR of all users while avoiding excessive concentration of resources on only a few strong users. As a result, DHI-Max-Product provides a balanced operating mode between QoS-oriented allocation and spectral-efficiency maximization.

3.3.3. DHI-Max-Sum-Rate Strategy

The DHI-Max-Sum-Rate strategy aims to maximize the total downlink spectral efficiency of the CF-mMIMO system. The objective is written as [26]

$$\max_{P(t)} \sum_{k=1}^K SE_k(t), \quad (42)$$

subject to

$$P_{min} \leq p_k(t) \leq P_{max}, \quad k = 1, 2, \dots, K. \quad (43)$$

This strategy is suitable for throughput-priority scenarios because it allocates power in a way that increases the total system spectral efficiency. However, this may reduce the QoS satisfaction of users with weaker channel conditions. Therefore, in the proposed DHI framework, the reward function includes a QoS violation penalty to reduce the negative effect of throughput maximization on weak users.

3.3.4. L-BFGS-B Refinement

After the SAC agent generates the initial power-allocation vector, the DHI-Max-Product and DHI-Max-Sum-Rate strategies apply an additional L-BFGS-B refinement step. L-BFGS-B is selected because it is suitable for bound-constrained optimization problems, where each user power must remain within the allowable interval $[P_{min}, P_{max}]$.

Let the SAC-generated power vector be [30]

$$\mathbf{P}_{SAC}(t) = [p_1(t), p_2(t), \dots, p_k(t)]^T. \quad (44)$$

The refined power vector is obtained as

$$\mathbf{P}^*(t) = \arg \min_{\mathbf{p}(t)} \mathcal{L}_j(\mathbf{P}(t)), \quad (45)$$

subject to

$$P_{min} \leq p_k(t) \leq P_{max}, \quad k = 1, 2, \dots, K. \quad (46)$$

where $j \in \{\text{Max} - \text{Product}, \text{Max} - \text{Sum} - \text{Rate}\}$ and $\mathcal{L}_j(\cdot)$ is the negative form of the selected objective. For DHI-Max-Product, the loss function is

$$\mathcal{L}_{\text{Max-Product}}(\mathbf{P}(t)) = -\sum_{k=1}^K \log(\gamma_k(t) + \epsilon), \quad (47)$$

For DHI-Max-Sum-Rate, the loss function is

$$\mathcal{L}_{\text{Max-Sum-Rate}}(\mathbf{P}(t)) = -\sum_{k=1}^K SE_k(t). \quad (48)$$

The L-BFGS-B update can be generally expressed as

$$\mathbf{P}_{q+1} = \prod_{[P_{min}, P_{max}]} (\mathbf{P}_q - \alpha_q \mathbf{B}_q^{-1} \nabla \mathcal{L}_j(\mathbf{P}_q)), \quad (49)$$

where q is the internal L-BFGS-B iteration index, α_q is the step size, $\mathbf{B}_q^{-1}$ is the limited-memory approximation of the inverse Hessian matrix, and $\prod_{[P_{min}, P_{max}]}$ denotes projection onto the feasible power interval.

The DHI-Max-Min strategy does not require L-BFGS-B refinement in this framework because its main purpose is QoS-oriented minimum user protection rather than gradient-based throughput maximization. Therefore, the SAC-generated power vector is directly evaluated using the Max-Min objective and QoS satisfaction metrics.

The final output of the DHI framework is therefore written as

$$\mathbf{P}_{DHI}(t) = \begin{cases} \mathbf{P}_{SAC}(t), & \text{for DHI - Max - Min,} \\ \mathbf{P}^*(t), & \text{for DHI - Max - Product} \\ \mathbf{P}^*(t), & \text{for DHI - Max - Sum - Rate} \end{cases} \quad (50)$$

This hybrid design allows the SAC agent to provide adaptive power-allocation decisions under dynamic network conditions, while L-BFGS-B further refines the throughput-oriented strategies within the bounded transmit-power range. As a result, the proposed DHI framework can support different downlink operating requirements, including QoS-oriented service reliability, balanced SINR improvement, and sum spectral-efficiency maximization. The main power-control strategies integrated within the proposed DHI framework are summarized in Figure 4, showing how the framework applies different optimization objectives for downlink power allocation.

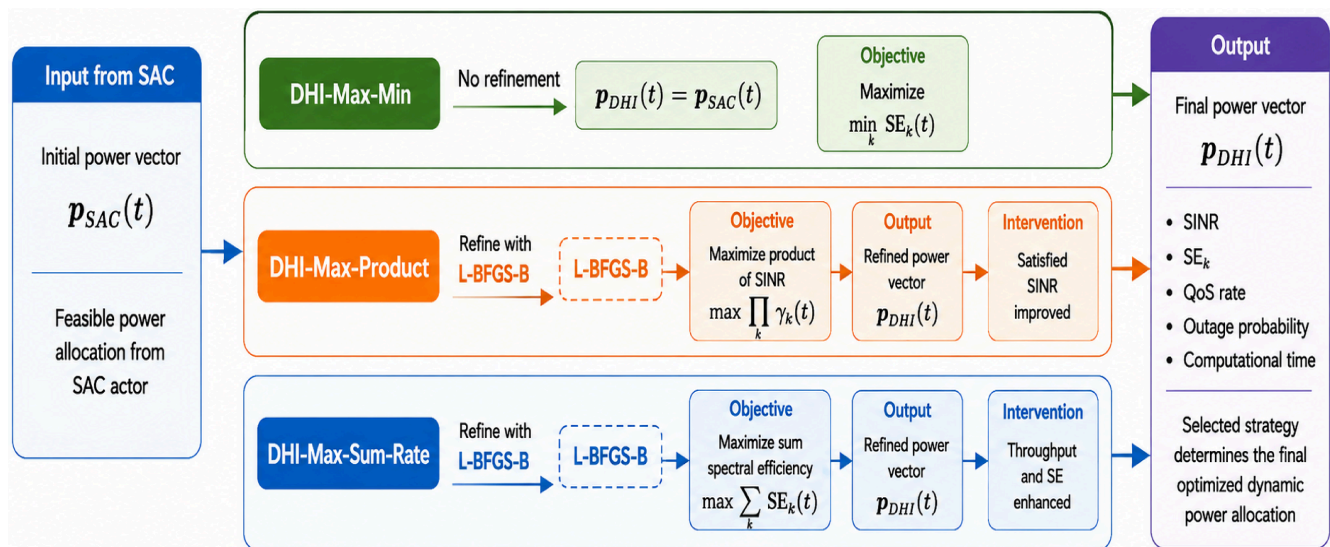


Figure 4. DHI power-control strategies.

3.4. Proposed DHI Model

The overall procedure of the proposed DHI framework is summarized in Algorithm 1. The algorithm begins by initializing the CF-mMIMO environment, including AP and UE locations, channel coefficients, mobility parameters, transmit-power limits, and SAC learning parameters. At each time step, the SAC agent observes the current network state and generates a continuous power-allocation action. The action is then mapped into the feasible transmit-power range and evaluated using the selected DHI strategy. For DHI-Max-Min, the SAC-generated power vector is directly evaluated because the objective focuses on minimum user QoS-oriented performance. For DHI-Max-Product and DHI-Max-Sum-Rate, the SAC-generated power vector is further refined using the L-BFGS-B optimizer to improve the objective value while maintaining the transmit-power bounds.

After applying the final power-allocation vector, the environment computes the downlink SINR, spectral efficiency, QoS satisfaction rate, outage probability, and reward. The transition is stored in the replay buffer, and the SAC actor and critic networks are updated using mini-batch samples from the buffer. This process continues until the maximum number of training steps or episodes is reached. The trained policy is then used for online downlink power allocation under dynamic channel and mobility conditions.

Algorithm 1 shows how the proposed DHI framework combines SAC-based adaptive learning with strategy-dependent power-control objectives. The SAC agent provides the initial power-allocation decision based on the current network state, while the selected DHI strategy determines how the decision is evaluated and refined. The DHI-Max-Min strategy directly evaluates the SAC-generated power vector for QoS-oriented minimum user performance, whereas DHI-Max-Product and DHI-Max-Sum-Rate use L-BFGS-B refinement to improve SINR-product and sum-rate objectives under bounded transmit-power constraints. This structure allows the proposed framework to adapt to user mobility and channel variation while supporting different downlink operating requirements.

Algorithm 1: Proposed DHI Framework for Dynamic Power Allocation**Require:** $M, K, P_{\min}, P_{\max}, \tau_p$, SAC parameters, and DHI strategy j **Ensure:** Optimized downlink power vector $p_{DHI}(l)$

1. Initialize AP/UE locations, mobility model, channel parameters, SAC networks, and replay buffer D .
2. Set the initial downlink power vector $p(0)$.
3. For each episode do:
4. Reset the CF-mMIMO environment and observe the initial state s_0 .
5. For each time step l do:
6. Update UE positions, channel coefficients, and interference conditions.
7. Observe the current state s_l .
8. Generate SAC action $a_l \sim \pi_\phi(\cdot | s_l)$.
9. Map a_l to the feasible power vector $p_{SAC}(l)$ within $[P_{\min}, P_{\max}]$.
10. If $j = \text{DHI-Max-Min}$, then:
11. Set $p_{DHI}(l) \leftarrow p_{SAC}(l)$.
12. Else if $j = \text{DHI-Max-Product}$, then:
13. Refine $p_{SAC}(l)$ using L-BFGS-B with the SINR-product objective.
14. Obtain $p_{DHI}(l)$.
15. Else if $j = \text{DHI-Max-Sum-Rate}$, then:
16. Refine $p_{SAC}(l)$ using L-BFGS-B with the sum-rate objective.
17. Obtain $p_{DHI}(l)$.
18. End if.
19. Apply $p_{DHI}(l)$ to the downlink CF-mMIMO environment.
20. Compute SINR, spectral efficiency, QoS satisfaction rate, outage probability, and reward.
21. Store $(s_l, a_l, r_l, s_l + 1, d_l)$ in D .
22. If D contains enough samples, then:
23. Update SAC critic networks, actor network, entropy coefficient, and target networks.
24. End if.
25. End for.
26. End for.
27. Return the trained DHI policy and $p_{DHI}(l)$.

3.5. Complexity and Practical Feasibility Analysis

The computational complexity of the proposed DHI framework depends on three main components: the CF-mMIMO environment evaluation, the SAC learning process, and the optional L-BFGS-B refinement used in the DHI-Max-Product and DHI-Max-Sum-Rate strategies. Since the system consists of M distributed APs and K UE, the computational cost increases with the number of AP–UE channel links, the number of users whose downlink powers are optimized, the neural network size of the SAC agent, and the number of refinement iterations required by L-BFGS-B.

At each time step, the environment computes the channel-dependent quantities, downlink SINR values, spectral efficiency, QoS satisfaction rate, outage probability, and reward. The dominant environment cost comes from evaluating the useful signal and multi-user in-

interference terms for all UE. Since each UE receives useful and interfering contributions from M APs and $K - 1$ interfering user streams, the SINR evaluation can be approximated as

$$C_{env}(M, K) = \mathcal{O}(MK^2). \quad (51)$$

This cost reflects the repeated computation of AP–UE contributions for each user and interfering stream. Therefore, increasing the number of users has a stronger effect on the environment evaluation cost than increasing the number of APs.

For the SAC component, the offline training cost depends on the number of training time steps, the environment evaluation cost, and the actor–critic neural network update cost. If T_{tr} denotes the total number of training time steps and C_{NN} denotes the cost of one SAC neural network update, the total SAC training complexity can be expressed as

$$C_{train} = \mathcal{O}(T_{tr}(C_{env}(M, K) + C_{NN})) \quad (52)$$

The neural network update cost depends on the selected multilayer perceptron architecture, batch size, and number of critic and actor updates. For a network with L_{NN} layers and layer widths $d_0, d_1, \dots, d_{L_{NN}}$ the forward-pass cost can be approximated as

$$C_{forward} = \mathcal{O}\left(\sum_{\ell=1}^{L_{NN}} d_{\ell-1}d_{\ell}\right). \quad (53)$$

Because SAC uses an actor network, two critic networks, and target critic networks, the practical training cost is higher than a single-network forward pass. Considering a mini-batch size B , the neural network update cost can be approximated as

$$C_{NN} = \mathcal{O}\left(B\sum_{\ell=1}^{L_{NN}} d_{\ell-1}d_{\ell}\right) \quad (54)$$

This cost is mainly incurred during offline training. Once training is completed, online deployment does not require full actor–critic updates. Instead, the trained actor network generates the downlink power-allocation action using only a forward pass. Therefore, the online SAC inference complexity is

$$C_{infer} = \mathcal{O}(C_{forward}) \quad (55)$$

This is significantly lower than the offline training cost and makes the trained policy suitable for real-time or near-real-time downlink power control.

For DHI-Max-Product and DHI-Max-Sum-Rate, L-BFGS-B refinement is applied after the SAC actor generates the initial power vector. Let K be the number of optimization variables, corresponding to the UE-level downlink power vector, I_{LBFGS} be the number of L-BFGS-B iterations, and m_{LBFGS} be the limited-memory correction parameter. The complexity of the refinement stage can be approximated as

$$C_{LBFGS} = \mathcal{O}\left(I_{LBFGS}\left(C_{obj}(M, K) + m_{LBFGS}K\right)\right), \quad (56)$$

where $C_{obj}(M, K)$ is the cost of evaluating the selected objective function. Since the objective evaluation depends on SINR and spectral-efficiency computation, it is linked to $C_{env}(M, K)$. Therefore,

$$C_{LBFGS} \approx \mathcal{O}\left(I_{LBFGS}\left(MK^2 + m_{LBFGS}K\right)\right), \quad (57)$$

The total online complexity of the proposed DHI framework depends on the selected strategy. For DHI-Max-Min, no L-BFGS-B refinement is used; therefore, the online complexity is mainly the SAC actor inference and environment evaluation:

$$C_{online}^{Max-Min} = \mathcal{O}(C_{forward} + C_{env}(M, K)) \quad (58)$$

For DHI-Max-Product and DHI-Max-Sum-Rate, the online complexity includes the SAC forward pass, environment evaluation, and L-BFGS-B refinement:

$$C_{online}^j = \mathcal{O}(C_{forward} + C_{env}(M, K) + C_{LBFGS}), \quad (59)$$

where $j \in \{Max - Product, Max - Sum - Rate\}$.

In terms of scalability, the number of AP-UE links grows as MK , while the interference-related SINR computation grows approximately as MK^2 . This means that user density has a stronger impact on computational cost than AP density. As K increases, the action dimension of the SAC policy also increases because the actor outputs one power-control value for each UE. Therefore, for larger CF-mMIMO deployments, practical scalability can be improved by using user clustering, AP selection, parameter sharing, or distributed/federated learning structures. These approaches reduce the effective number of AP-UE links and make the learning problem more manageable.

The proposed framework is practical because the computationally expensive SAC training process can be performed offline using a CPU, GPU, or cloud/edge server. The trained actor policy can then be deployed online at the CPU or an edge controller to generate power-allocation decisions through fast inference. The code implementation supports automatic device selection among CPU, CUDA-enabled GPU, and Apple MPS when available, which allows the training process to benefit from hardware acceleration when supported.

From a deployment perspective, the CPU collects large-scale fading, UE location, SINR, and interference-related information through the fronthaul network. Since the SAC state mainly relies on slowly varying large-scale fading, mobility information, previous SINR, previous interference, and previous power allocation, the signaling overhead can be reduced compared with schemes that require full instantaneous CSI exchange at every time step. However, the framework still requires reliable fronthaul signaling between APs and the CPU to update state information and apply the optimized downlink power-control decisions.

The measured computational times support the feasibility of the proposed approach. In the reported results, DHI-Max-Product and DHI-Max-Sum-Rate achieved mean computational times of 0.0690 s and 0.0696 s, respectively, while the DDPG benchmark required 0.63 s. DHI-Max-Min required a higher mean duration of 0.3757 s because it focuses on minimum user QoS-oriented behavior. These results indicate that the proposed DHI-Max-Product and DHI-Max-Sum-Rate strategies can provide faster online execution than the benchmark while maintaining strong spectral-efficiency performance.

Overall, the DHI framework separates offline learning from online inference. The offline phase handles policy training, hyperparameter tuning, and convergence monitoring, while the online phase uses the trained actor and optional L-BFGS-B refinement to generate bounded downlink power-allocation decisions. This structure improves deployment feasibility by avoiding the need to train the full SAC model during real-time operation. Nevertheless, the practical implementation of the proposed framework still depends on fronthaul capacity, state-update frequency, network size, and the ability of the CPU or edge controller to perform inference and optional refinement within the required scheduling interval.

4. Simulation Setup

The proposed DHI framework is evaluated using a dynamic downlink CF-mMIMO simulation environment. The simulated network consists of $M = 64$ distributed APs and $K = 32$ UE deployed within a $1000 \times 1000 \text{ m}^2$ square area. The APs and UE are distributed over the service region, and all APs jointly serve the UE through coordinated downlink transmission. The system uses $\tau_p = 20$ orthogonal pilots. The large-scale fading model includes path loss and spatially correlated shadow fading, while the small-scale fading component follows Rayleigh fading. The shadow fading standard deviation is set to 8 dB, the decorrelation distance is 100 m, and the shadow fading decorrelation parameter is $\delta = 0.5$.

The system bandwidth is set to 20 MHz and the receiver noise figure is 9 dB. Therefore, the noise variance is calculated as

$$\alpha_{\text{dBm}}^2 = -174 + 10\log_{10}(B) + NF \quad (60)$$

where B is the system bandwidth in Hz and NF is the receiver noise figure in dB. For $B = 20 \text{ MHz}$ and $NF = 9 \text{ dB}$, the resulting noise variance is approximately -92 dBm .

User mobility is enabled to represent time-varying AP–UE distances and dynamic channel conditions. The UE speed is randomly selected from the range 0.5–5 m/s, and the mobility model is updated at every time step of 1 s. A maximum pause time of 5 s is considered to model temporary UE stopping behavior. At each time step, the UE positions are updated, the AP–UE distances are recalculated, and the corresponding large-scale fading coefficients are updated. This allows the simulation to capture mobility-driven channel variation and interference changes over time.

The SAC-based DHI model uses continuous power-control actions. The SAC action is mapped to the physical transmit-power interval $[P_{\min}, P_{\max}]$, where $P_{\min} = 0 \text{ mW}$ and $P_{\max} = 13 \text{ mW}$. The initial transmit power is also set to 13 mW. The temporal reward window size is set to 10 steps, and the positive and negative temporal reward weights are set to $r_\alpha = 1$ and $r_\beta = 1$, respectively.

The SAC hyperparameters are tuned using Optuna. The implementation uses a TPE sampler, median pruning, and a maximum of 200 trials with 2000 time steps per trial. The candidate hyperparameters include the discount factor, learning rate, batch size, replay-buffer size, learning-start steps, training frequency, Polyak coefficient, and MLP network architecture. In the implementation, the SAC entropy coefficient and target entropy are set to automatic tuning, and SGD is used as the optimizer. The code samples the SAC hyperparameter search space using Optuna and fixes the temporal reward window at 10 with $r_\alpha = r_\beta = 1$. The optimizer implementation also supports TPE sampling, median pruning, device selection, and saving the trial results to CSV and PKL files. The main system simulation parameters adopted for evaluating the proposed downlink CF-mMIMO system are listed in Table 2. Table 3 presents the simulation hyperparameters used to configure the learning environment and optimization process. The implementation details and training configuration of the proposed DHI framework are summarized in Table 4.

Table 2. System simulation parameters.

Parameter	Symbol/Setting	Value
Number of APs	M	64
Number of UE	K	32
Number of orthogonal pilots	τ_p	20
Simulation area	—	$1000 \times 1000 \text{ m}^2$

Table 2. Cont.

Parameter	Symbol/Setting	Value
System bandwidth	B	20 MHz
Receiver noise figure	NF	9 dB
Noise variance	σ^2	approximately -92 dBm
Shadow fading standard deviation	σ_{sf}	8 dB
Shadow fading decorrelation factor	δ	0.5
Shadow fading decorrelation distance	ddecorr	100 m
Antenna spacing	—	0.5 wavelength
Angular standard deviation	—	15°
Minimum transmit power	Pmin	0 mW
Maximum transmit power	Pmax	100 mW
Initial transmit power	Pinit	100 mW
Pre-log factor	—	1
UE speed range	—	0.5–5 m/s
Mobility time step	Δt	1 s
Maximum pause time	—	5 s
Pause probability	Ppause	0.3
Temporal reward window	Q	10 steps
Positive temporal reward weight	$r\alpha$	1
Negative temporal reward weight	$r\beta$	1

Table 3. System simulation hyperparameters.

Hyperparameter	Search Space
Discount factor γ_{SAC}	{0.9, 0.95, 0.98, 0.99, 0.995, 0.999, 0.9999}
Learning rate	$[10^{-5}, 1]$, log-uniform
Batch size	{16, 32, 64, 128, 256, 512, 1024, 2048}
Replay-buffer size	$\{10^4, 10^5, 10^6\}$
Learning starts	{0, 10, 100, 1000}
Train frequency	{1, 4, 8, 16, 32, 64, 128, 256, 512}
Gradient steps	Equal to train frequency
Polyak coefficient τ	{0.001, 0.005, 0.01, 0.02, 0.05, 0.08}
Network architecture	[64,64], [256,256], [400,300], [128,256,128], [256,256,256]
Entropy coefficient	Auto
Target entropy	Auto
Optimizer	SGD

Table 4. DHI implementation and training configuration.

Component	Setting Used in the Implementation
Learning algorithm	Soft Actor–Critic
Policy type	MLP policy
Learning type	Off-policy actor–critic
Action type	Continuous downlink power-control action
Replay buffer	Stores (st, at, rt, st+1, dt)
Entropy coefficient	Auto
Target entropy	Auto
Optimizer	SGD
Temporal reward window size	10
Positive temporal reward weight	$r\alpha = 1$
Negative temporal reward weight	$r\beta = 1$
Hyperparameter optimizer	Optuna
Sampler	TPE sampler
Pruner	Median pruner
Number of trials	200
Time steps per trial	2000
Evaluation mode	Deterministic evaluation
Training device	GPU

5. Results and Discussion

This section evaluates the performance of the proposed DHI framework for dynamic downlink power allocation in CF-mMIMO systems. The evaluation focuses on convergence behavior, spectral efficiency, SINR distribution, QoS satisfaction, outage probability, spatial SINR behavior, and computational time. The three proposed DHI strategies, namely DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate, are compared to show how different power-control objectives affect downlink performance. DHI-Max-Min is evaluated as a QoS-oriented strategy, DHI-Max-Product as a balanced SINR-aware strategy, and DHI-Max-Sum-Rate as a throughput-oriented strategy. The DDPG benchmark is also included in the sum spectral-efficiency and computational-time comparisons to assess the advantage of the SAC-based DHI design.

The purpose of this section is not only to compare numerical values but also to explain how each result supports the motivation of the paper. Specifically, the results examine whether the proposed DHI framework can adapt to dynamic downlink conditions, improve spectral-efficiency behavior, support QoS satisfaction, and reduce computational time compared with the benchmark method.

Figure 5 presents the empirical cumulative distribution function (CDF) and probability density function (PDF) of the downlink spectral efficiency obtained using DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate. The CDF curves show the probability that the spectral efficiency is below a given value, while the PDF curves indicate the concentration of spectral-efficiency values across the simulated downlink conditions. Therefore, these distributions provide more complete information than a single average value because they show how each strategy affects both low-performance and high-performance users.

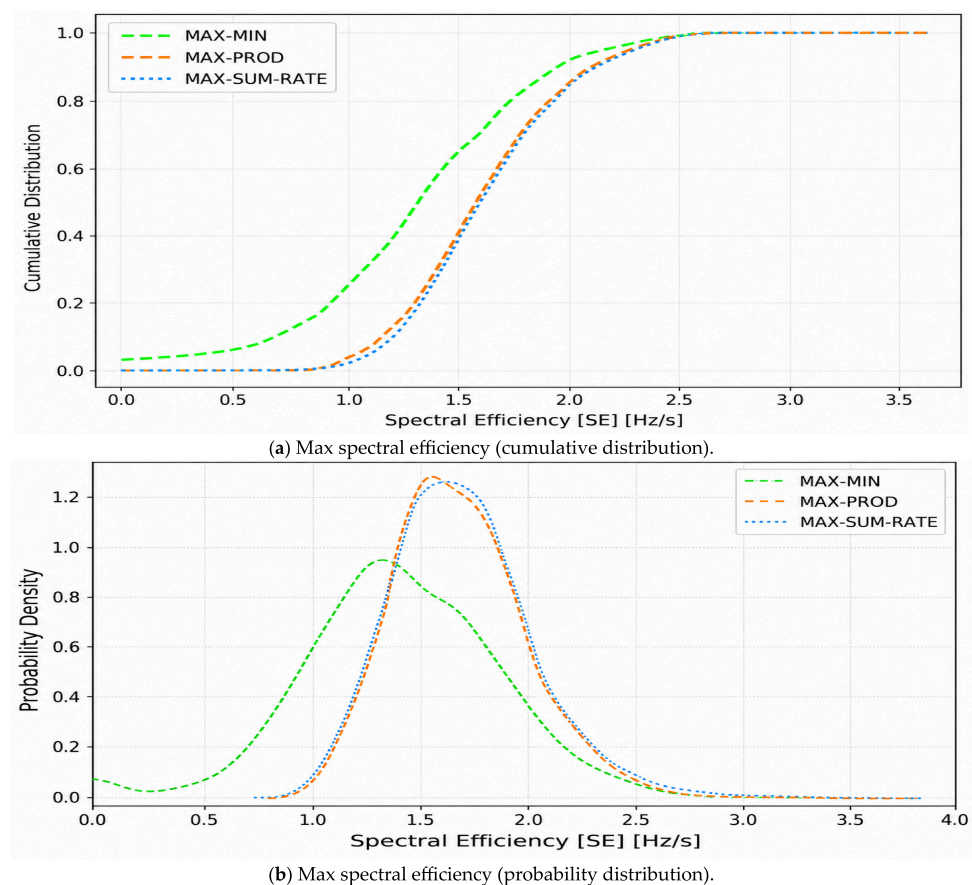
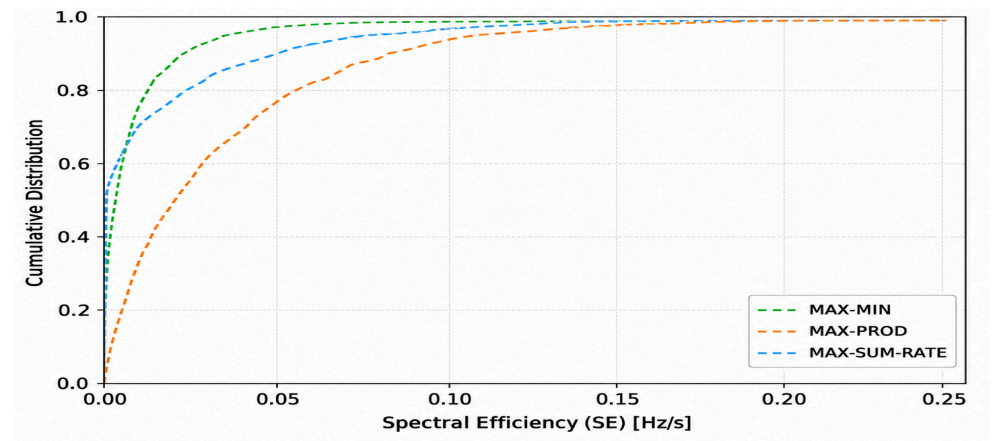
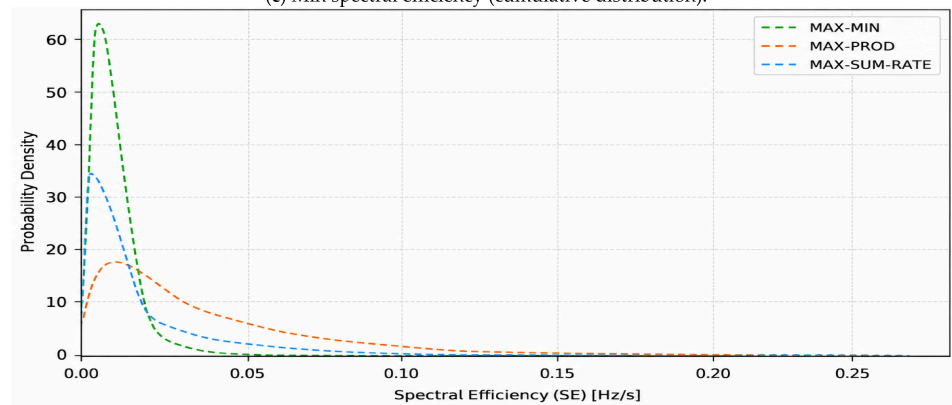


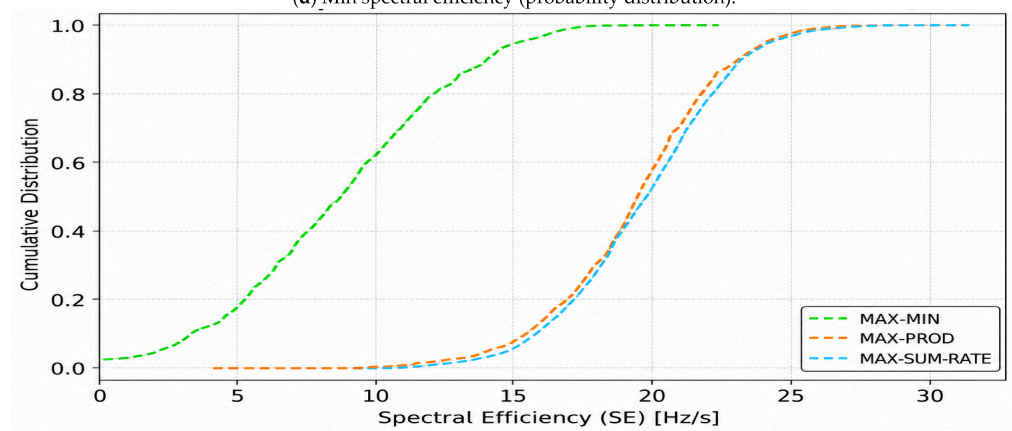
Figure 5. Cont.



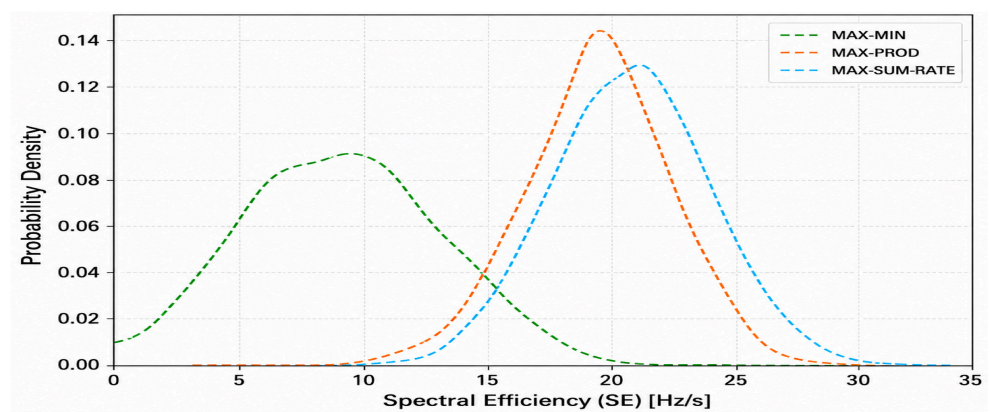
(c) Min spectral efficiency (cumulative distribution).



(d) Min spectral efficiency (probability distribution).



(e) Total spectral efficiency (cumulative distribution).



(f) Total spectral efficiency (probability distribution).

Figure 5. Empirical CDF and PDF of downlink spectral efficiency for DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate.

The DHI-Max-Min strategy produces a more conservative spectral-efficiency distribution because its objective prioritizes the weakest UE. This behavior improves minimum user service and supports QoS-oriented operation, but it also limits the aggregate spectral-efficiency gain because additional power is assigned to weaker users instead of only maximizing throughput. In contrast, DHI-Max-Sum-Rate shifts the spectral-efficiency distribution toward higher values because it directly maximizes the total system spectral efficiency. This confirms its suitability for throughput-priority operation. However, the improvement in aggregate spectral efficiency may come with reduced QoS satisfaction for users experiencing weaker channel conditions.

DHI-Max-Product provides an intermediate behavior between these two strategies. By improving the product-based SINR objective, it avoids the excessive throughput bias of DHI-Max-Sum-Rate while still achieving better spectral-efficiency behavior than the QoS-oriented DHI-Max-Min strategy. This makes DHI-Max-Product suitable for balanced downlink operation where both SINR improvement and spectral-efficiency performance are required.

Overall, Figure 5 confirms that the three DHI strategies provide different operating modes rather than a single fixed solution. DHI-Max-Min is more suitable when minimum user service and QoS reliability are prioritized, DHI-Max-Product provides a balanced operating point, and DHI-Max-Sum-Rate is preferable when the main objective is to maximize aggregate downlink throughput.

Figure 6 compares the empirical cumulative distribution function of the sum spectral efficiency achieved by the proposed DHI strategies and the DDPG benchmark. The sum spectral efficiency is an important metric because it measures the aggregate downlink throughput of the CF-mMIMO system and directly reflects how efficiently the available transmit power and distributed AP resources are used.

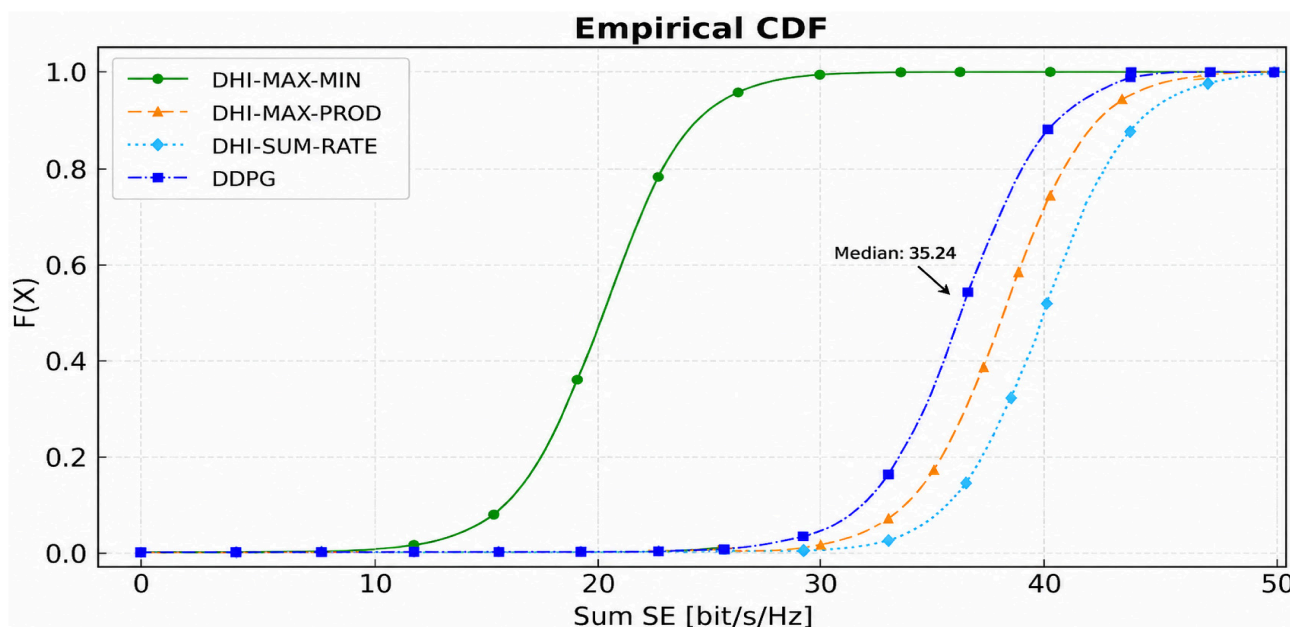


Figure 6. Empirical CDF comparison of sum spectral efficiency between the proposed DHI strategies and the DDPG benchmark [19].

The DHI-Max-Min strategy achieves the lowest sum spectral-efficiency distribution among the proposed strategies. This behavior is expected because DHI-Max-Min prioritizes the weakest users and attempts to improve minimum user service rather than maximizing the total throughput. As a result, part of the available power is allocated to users with

weaker channel conditions, which improves QoS-oriented behavior but reduces aggregate spectral-efficiency gain.

The DHI-Max-Product strategy shifts the distribution toward higher spectral-efficiency values compared with DHI-Max-Min. This indicates that maximizing the logarithmic SINR-product objective improves the overall SINR balance among users while still maintaining better aggregate throughput than the minimum user-focused strategy. Therefore, DHI-Max-Product provides a balanced operating point between QoS-oriented service reliability and sum spectral-efficiency improvement.

The DHI-Max-Sum-Rate strategy achieves the highest sum spectral-efficiency performance because its objective directly maximizes the total downlink spectral efficiency. Compared with the DDPG benchmark, the DHI-Max-Sum-Rate curve is shifted toward higher spectral-efficiency values, showing that the proposed hybrid learning-and-optimization structure provides stronger throughput-oriented performance. The DDPG benchmark achieves a median sum spectral efficiency of approximately 35.24 bit/s/Hz, while the DHI-Max-Sum-Rate strategy provides higher overall spectral-efficiency behavior under the same CF-mMIMO simulation setting.

The improvement of DHI-Max-Sum-Rate can be explained by two factors. First, SAC uses entropy-regularized stochastic exploration, which helps the agent explore different continuous power-allocation actions during training instead of converging too early to a narrow deterministic policy. Second, the L-BFGS-B refinement stage further improves the SAC-generated power vector under bounded transmit-power constraints. This combination allows the proposed DHI-Max-Sum-Rate strategy to exploit both learning-based adaptability and optimization-based refinement.

The DDPG benchmark is selected because it is also an actor-critic DRL method designed for continuous control problems. Therefore, it provides a reasonable comparison for continuous downlink power allocation. However, unlike SAC, DDPG uses a deterministic policy and may be more sensitive to exploration noise and local convergence behavior. This explains why the SAC-based DHI framework can achieve stronger and more stable sum spectral-efficiency performance in the dynamic CF-mMIMO environment.

Figure 7 presents the empirical CDF and PDF distributions of the downlink SINR under the three proposed DHI strategies: DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate. The SINR distribution is an important indicator because it directly affects both spectral efficiency and QoS satisfaction. While spectral efficiency measures the achievable data rate, SINR shows how effectively each strategy controls interference and supports reliable downlink reception.

The maximum SINR distribution shows the highest SINR values achieved under each strategy. DHI-Max-Sum-Rate generally produces higher maximum SINR values because its objective prioritizes aggregate throughput. This means that users with favorable channel conditions can receive stronger power allocation and achieve higher link quality. However, high maximum SINR alone does not necessarily indicate better QoS for all users, because it may occur when the strategy benefits strong users more than weak users.

The minimum SINR distribution provides a more direct view of QoS-oriented behavior. DHI-Max-Min performs better in this part because it is designed to improve the weakest user performance. By increasing the lower tail of the SINR distribution, DHI-Max-Min reduces the probability that users experience very poor link quality. This explains why DHI-Max-Min achieves stronger QoS satisfaction and lower outage behavior, even though its aggregate spectral efficiency is lower than the throughput-oriented strategy.

The mean SINR distribution shows the average SINR behavior across users and provides a balanced view of network-level link quality. DHI-Max-Product is expected to perform strongly in this metric because the logarithmic product-based objective encourages

improvement across multiple users rather than focusing only on the weakest user or only on the strongest throughput contributors. Therefore, DHI-Max-Product can be interpreted as a balanced strategy that improves SINR distribution while maintaining better spectral-efficiency behavior than DHI-Max-Min.

Overall, the SINR results confirm that the three DHI strategies produce different downlink operating behaviors. DHI-Max-Min is more suitable for QoS-oriented operation because it improves the minimum SINR. DHI-Max-Product provides a balanced SINR distribution by reducing extreme imbalance among users. DHI-Max-Sum-Rate is more suitable for throughput-oriented operation because it increases the higher-SINR region and improves aggregate spectral efficiency. Therefore, Figure 7 supports the main motivation of the paper by showing that the proposed DHI framework can adapt the power-allocation behavior according to the required downlink objective.

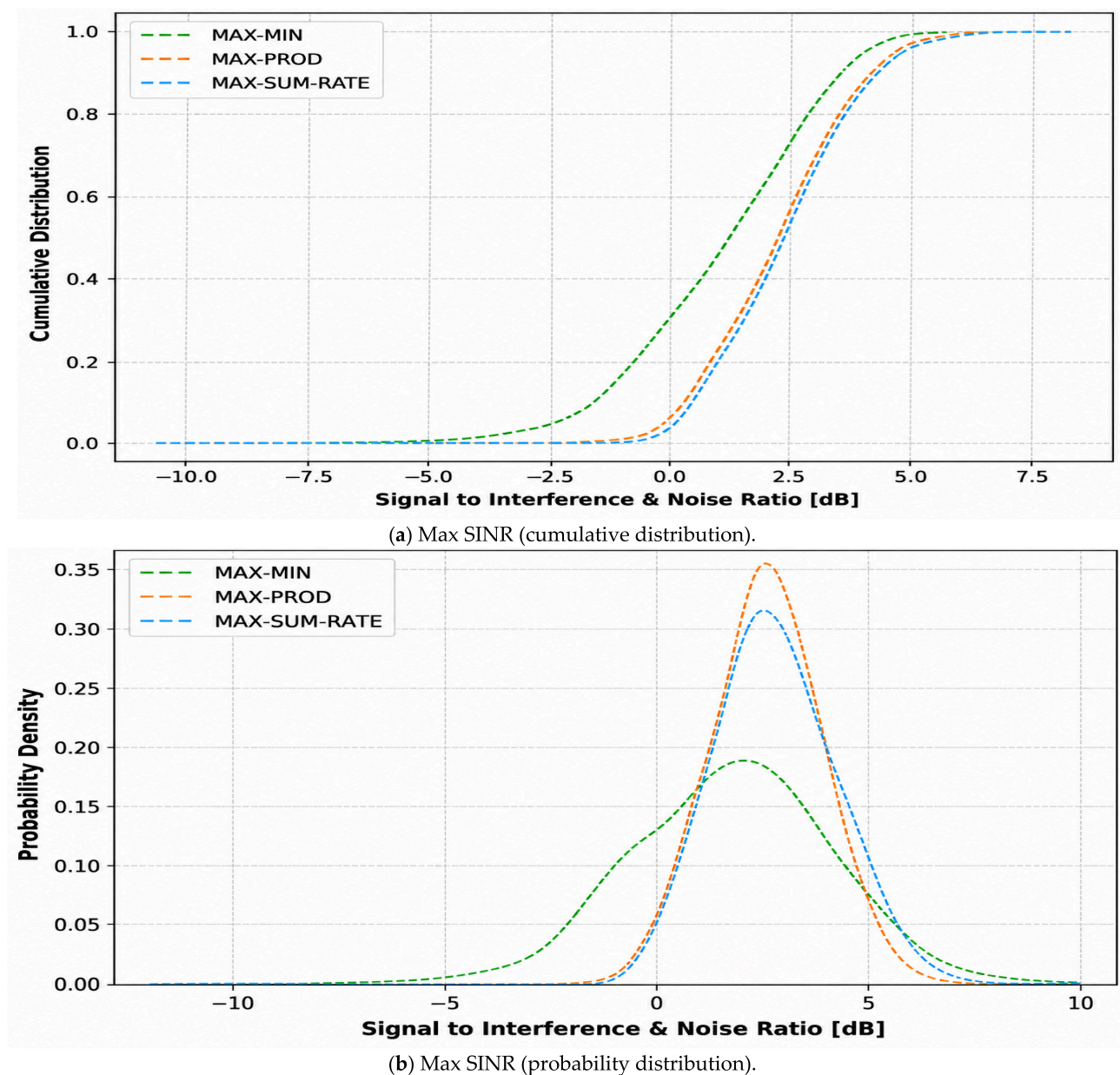
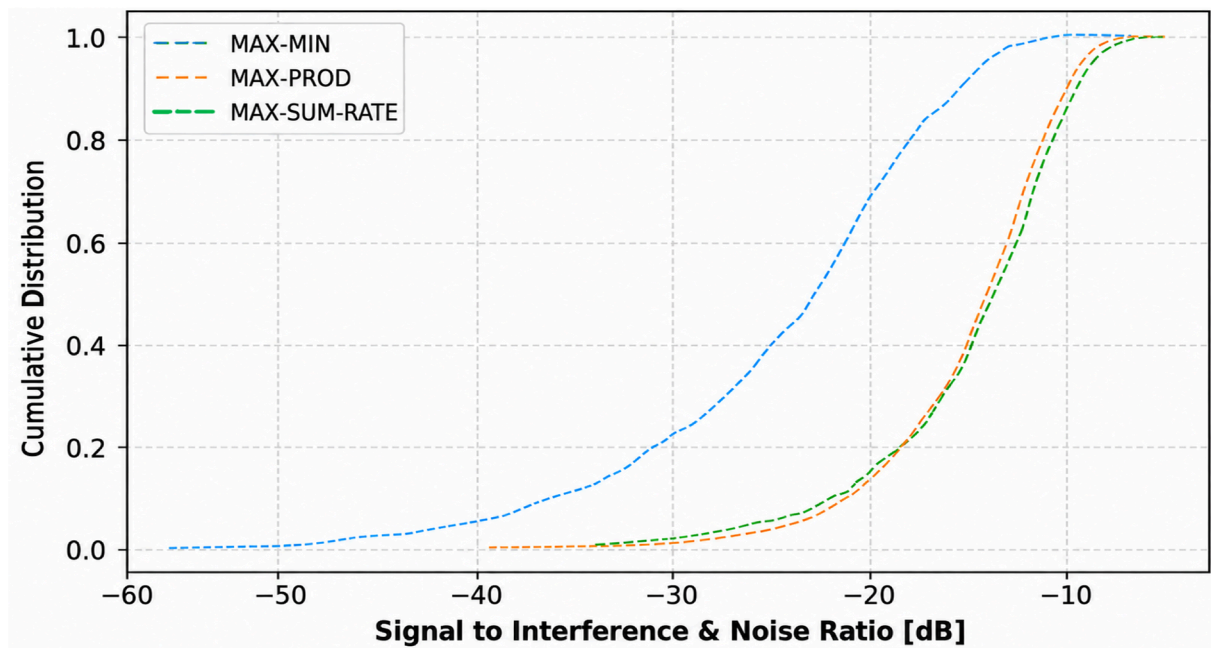
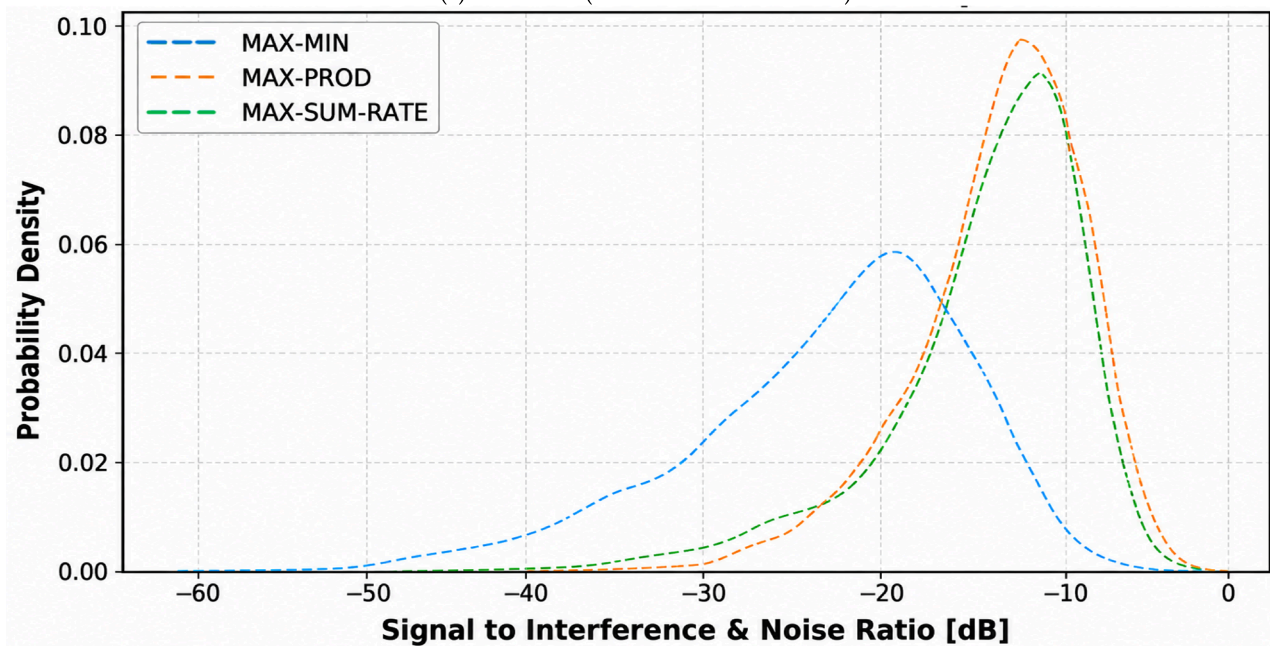


Figure 7. Cont.

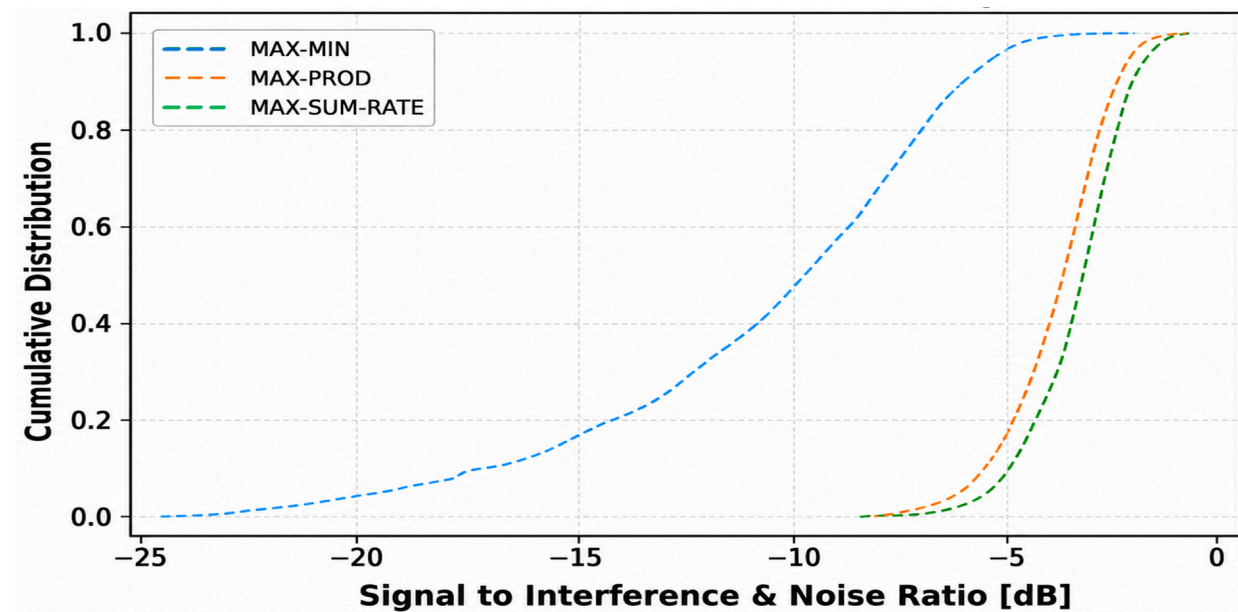


(c) Min SINR (cumulative distribution).

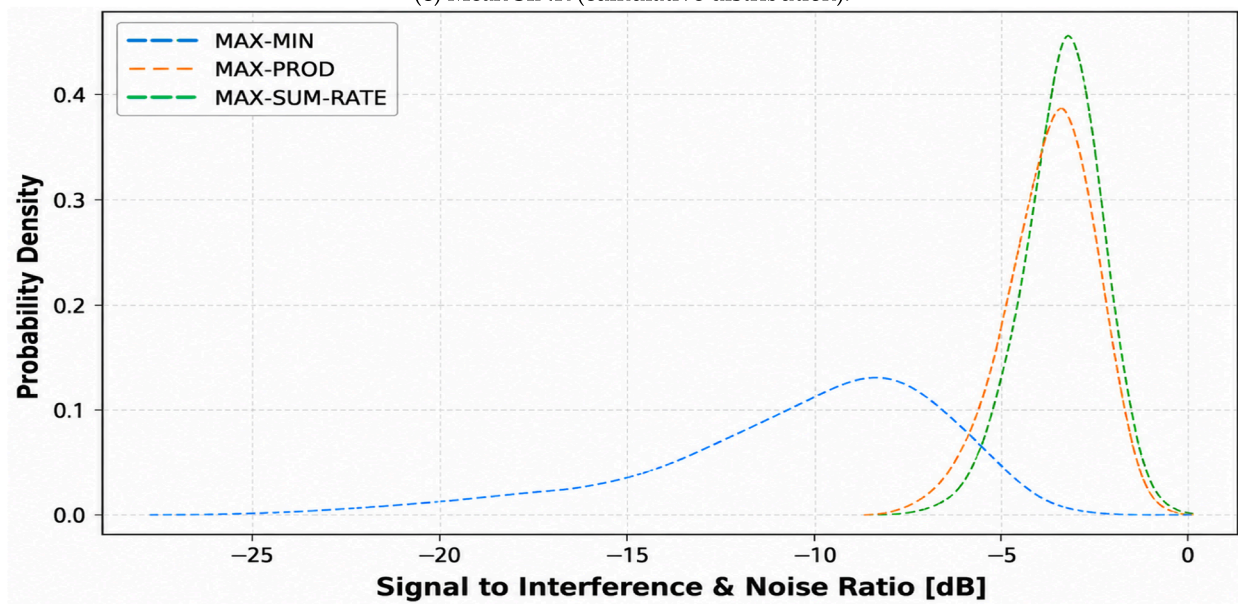


(d) Min SINR (probability distribution).

Figure 7. Cont.



(e) Mean SINR (cumulative distribution).



(f) Mean SINR (probability distribution).

Figure 7. Empirical CDF and PDF distributions of downlink SINR for DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate.

Table 5 summarizes the QoS-oriented performance of the proposed DHI downlink power-allocation strategies. The QoS satisfaction rate is calculated based on the percentage of UE whose SINR values satisfy the required SINR threshold. The outage probability represents the percentage of UE that do not satisfy this threshold. Therefore, these two metrics provide a direct evaluation of how well each strategy supports downlink service reliability under dynamic channel and mobility conditions.

The results show that DHI-Max-Min achieves the highest QoS satisfaction rate of 93.75%, corresponding to an average of 30 users satisfying the required SINR threshold out of 32 UE. It also achieves the lowest outage probability of 6.25%. This confirms that DHI-Max-Min is the most suitable strategy when the main objective is QoS-oriented service reliability and minimum user protection. However, this improvement is obtained at the cost of lower aggregate spectral efficiency, with a mean sum SE of 17.50 bit/s/Hz.

DHI-Max-Product achieves a QoS satisfaction rate of 91.04% and an outage probability of 8.96%. Its mean sum spectral efficiency reaches 19.89 bit/s/Hz, which is higher than DHI-Max-Min. This shows that DHI-Max-Product provides a balanced trade-off between QoS support and throughput improvement. It does not prioritize the weakest users as strongly as DHI-Max-Min, but it avoids the more aggressive throughput bias of DHI-Max-Sum-Rate.

Table 5. QoS-oriented performance comparison of the proposed DHI downlink strategies.

Strategy	Avg. Users Meeting QoS	QoS Satisfaction Rate (%)	Outage Probability (%)	Avg. Minimum SINR (dB)	Mean SINR (dB)	Mean Sum SE (bit/s/Hz)	Main Interpretation
DHI-Max-Min	30.00	93.75	6.25	−3.30	−13.50	17.50	Best QoS reliability and lowest outage; supports minimum user service
DHI-Max-Product	29.13	91.04	8.96	−3.38	−14.74	19.89	Balanced QoS and spectral-efficiency performance
DHI-Max-Sum-Rate	28.15	87.98	12.02	−3.26	−16.24	20.14	Highest throughput-oriented spectral-efficiency performance

DHI-Max-Sum-Rate achieves the highest mean sum spectral efficiency of 20.14 bit/s/Hz. However, its QoS satisfaction rate decreases to 87.98%, with an outage probability of 12.02%. This behavior is expected because the sum-rate objective prioritizes aggregate throughput, which may allocate more resources to users with better channel conditions. As a result, some weaker users may fail to satisfy the SINR threshold.

These results clarify that QoS in the proposed framework should be interpreted as a QoS-aware performance objective, not as an absolute hard guarantee for every UE at every time step. This is because mobility, interference, and transmit-power limitations may prevent all users from satisfying the SINR threshold simultaneously. Therefore, the proposed framework evaluates QoS using QoS satisfaction rate and outage probability rather than claiming complete QoS guarantee.

Figure 8 presents the spatial SINR heat maps obtained using the three proposed DHI downlink power-allocation strategies: DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate. The heat maps provide a visual interpretation of how each strategy distributes downlink link quality across the CF-mMIMO coverage area. Unlike the CDF and PDF plots, which summarize the statistical behavior of SINR, the heat maps show the geographical distribution of SINR and help identify regions with strong or weak service quality.

The DHI-Max-Min heat map shows a more uniform SINR distribution across the service area. This behavior is consistent with the objective of DHI-Max-Min, which focuses on improving the weakest user performance. As a result, extremely low-SINR regions are reduced, and the minimum user service is improved. However, because the strategy allocates more resources to weaker users, it does not produce the highest SINR peaks across the network. This confirms that DHI-Max-Min is more suitable for QoS-oriented downlink operation rather than throughput-maximizing operation.

The DHI-Max-Product heat map shows a more balanced spatial pattern. Compared with DHI-Max-Min, it provides higher SINR values in several regions while still avoiding severe imbalance among users. This is because the product-based SINR objective encourages the improvement of multiple users simultaneously instead of focusing only on the weakest user or only on the strongest channel conditions. Therefore, DHI-Max-Product can

be considered a balanced strategy for scenarios where both SINR improvement and QoS support are required.

The DHI-Max-Sum-Rate heat map shows stronger SINR values in selected areas, especially where AP–UE channel conditions are more favorable. This behavior is expected because the sum-rate objective prioritizes the maximization of aggregate spectral efficiency. However, this may create more spatial variation because users in weaker regions may receive less relative support compared with users in stronger channel conditions. Therefore, DHI-Max-Sum-Rate is more suitable for throughput-priority scenarios, while DHI-Max-Min remains more appropriate for QoS-oriented service coverage.

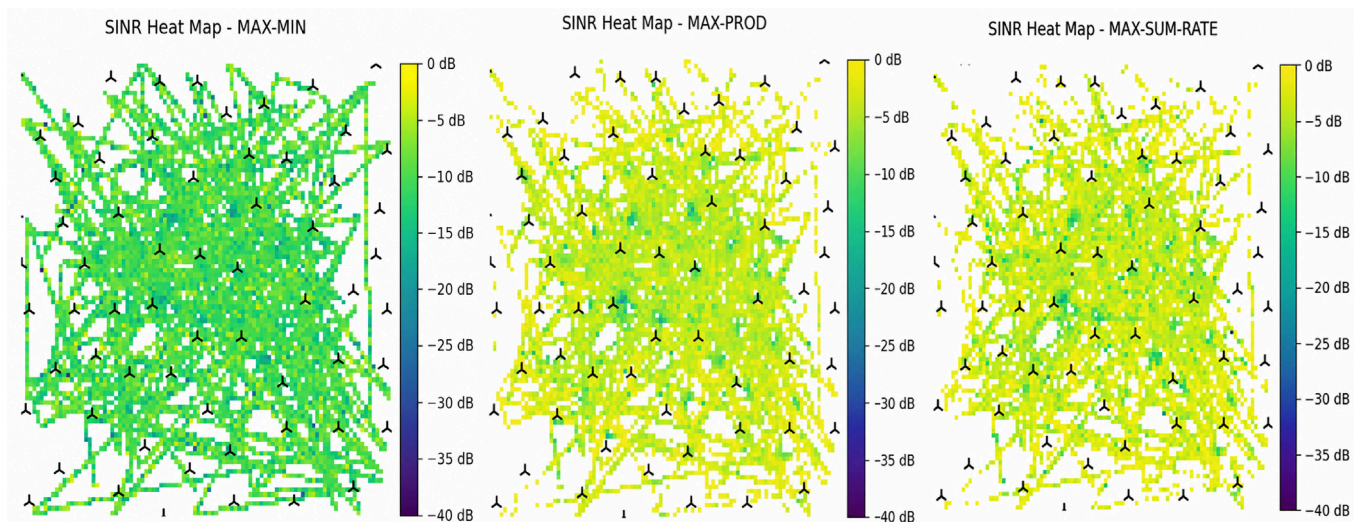


Figure 8. Spatial SINR heat map distribution for the proposed DHI downlink power-allocation strategies.

Overall, the heat map analysis confirms that the proposed DHI framework provides different spatial power-allocation behaviors according to the selected strategy. DHI-Max-Min improves spatial service uniformity, DHI-Max-Product provides a balanced SINR distribution, and DHI-Max-Sum-Rate increases high-SINR regions to improve total downlink throughput. These observations support the main objective of the proposed framework, which is to provide adaptive downlink power allocation under dynamic CF-mMIMO conditions.

$$R_{time}(\%) = \frac{T_{DDPG} - T_{DHI}}{T_{DDPG}} \times 100. \quad (61)$$

Using this equation:

$$R_{time}^{DHI-Max-Product} = \frac{0.63 - 0.0690}{0.63} \times 100. = 89.05\% \quad (62)$$

$$R_{time}^{DHI-Max-Sum-Rate} = \frac{0.63 - 0.0696}{0.63} \times 100. = 88.95\% \quad (63)$$

$$R_{time}^{DHI-Max-Min} = \frac{0.63 - 0.3757}{0.63} \times 100. = 40.37\% \quad (64)$$

Table 6 compares the computational time of the proposed DHI strategies with the DDPG benchmark. Computational time is an important metric for dynamic downlink power allocation because the power-control decision must be updated within a practical scheduling interval, especially under user mobility and time-varying channel conditions.

The results show that DHI-Max-Product and DHI-Max-Sum-Rate achieve the lowest mean computational times, with 0.0690 s and 0.0696 s, respectively. Compared with the DDPG benchmark, which requires 0.63 s, DHI-Max-Product reduces the mean computa-

tional time by approximately 89.05%, while DHI-Max-Sum-Rate reduces it by approximately 88.95%. This confirms that the proposed DHI framework can provide faster online execution for throughput-oriented downlink power allocation.

Table 6. Computational-time comparison of the proposed DHI strategies and DDPG [19].

Model	Mean Duration (s)	Max Duration (s)	Min Duration (s)	Main Interpretation
DHI-Max-Min	0.3757	0.7430	0.0001	Strongest QoS-oriented behavior, but higher computational cost
DHI-Max-Product	0.0690	0.3321	0.0191	Lowest mean computational time with balanced performance
DHI-Max-Sum-Rate	0.0696	0.3638	0.0175	Very low computational time with highest throughput-oriented SE
DDPG [19]	0.6300	0.9900	-----	Higher computational time than the proposed DHI strategies

DHI-Max-Min records a higher mean computational time of 0.3757 s compared with DHI-Max-Product and DHI-Max-Sum-Rate. This is expected because DHI-Max-Min focuses on minimum user service and QoS-oriented power balancing, which requires more careful allocation to weaker users. However, DHI-Max-Min is still faster than the DDPG benchmark, reducing the mean computational time by approximately 40.37%.

The maximum duration values further support this observation. DHI-Max-Min reaches a maximum duration of 0.7430 s, while DHI-Max-Product and DHI-Max-Sum-Rate require only 0.3321 s and 0.3638 s, respectively. In comparison, DDPG reaches a maximum duration of 0.99 s. These values show that the proposed DHI strategies reduce not only the average execution time but also the worst-case computational delay compared with the benchmark.

The lower computational time of DHI-Max-Product and DHI-Max-Sum-Rate can be explained by the hybrid design of the proposed framework. The SAC actor provides an initial power-allocation decision through fast neural network inference, while L-BFGS-B refinement improves the objective within bounded power constraints. This avoids solving the full power-allocation problem from the beginning at every time step. Therefore, the proposed framework separates the expensive offline training phase from the faster online decision-making phase.

Overall, the computational-time results confirm that the proposed DHI framework is suitable for dynamic CF-mMIMO downlink power allocation. DHI-Max-Sum-Rate is preferable when the main objective is high spectral efficiency with low execution time, DHI-Max-Product is suitable for balanced SINR and spectral-efficiency performance with very low computational cost, and DHI-Max-Min is appropriate when QoS-oriented service reliability is prioritized despite a higher computational time.

Figure 9 presents the downlink performance evolution of the proposed DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate strategies over the simulation steps. Figure 9a shows the moving average sum spectral efficiency, which smooths short-term fluctuations and highlights the general performance trend of each strategy. Figure 9b presents the instantaneous sum spectral efficiency, showing the step-by-step variation caused by mobility, channel changes, and dynamic power-allocation decisions. The results show that DHI-Max-Sum-Rate achieves the highest sum spectral-efficiency behavior, while DHI-Max-Product provides a close and stable performance. In contrast, DHI-Max-Min produces lower aggregate spectral efficiency because it prioritizes minimum user service and QoS-oriented allocation.

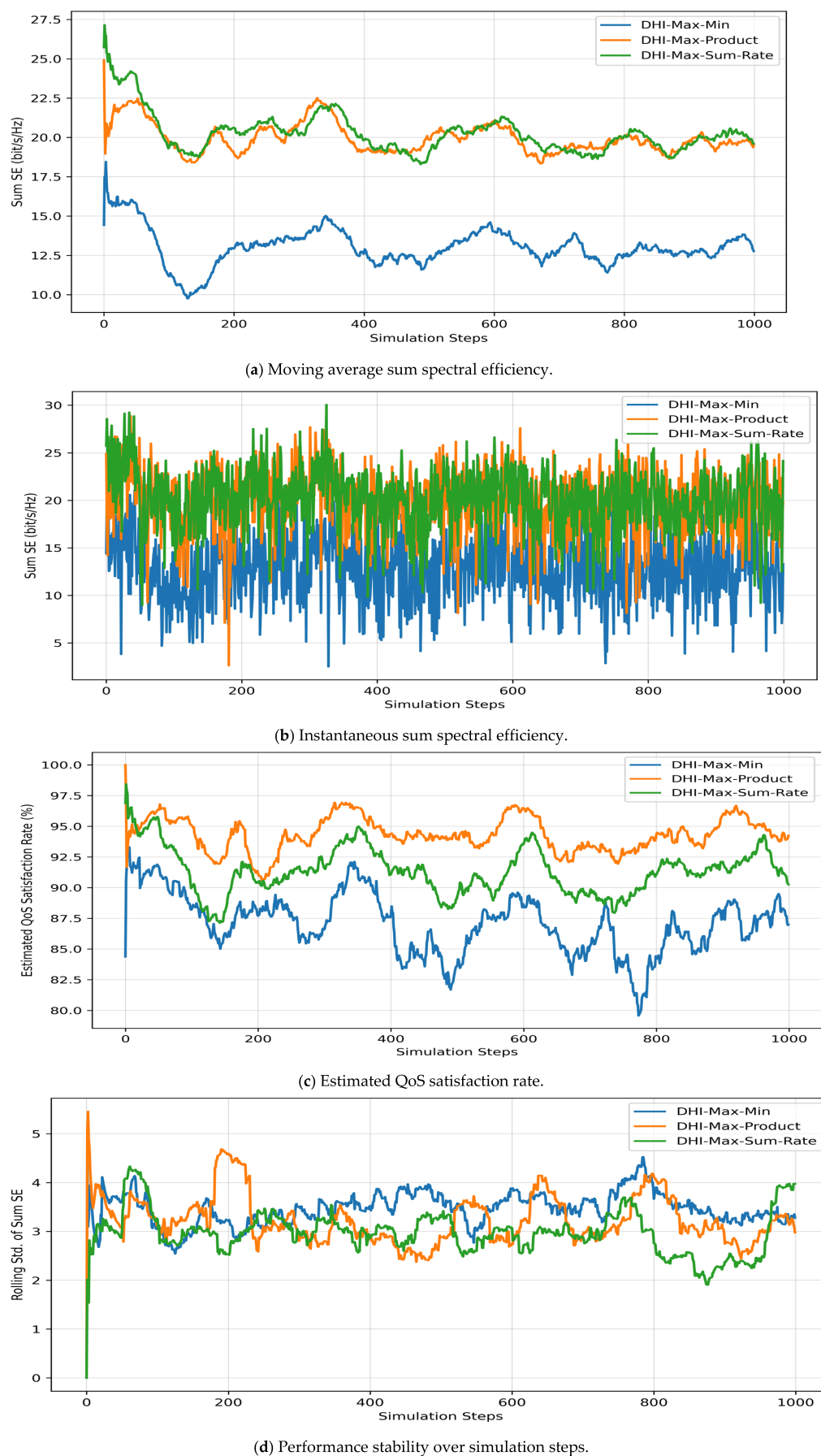


Figure 9. Evolution of downlink performance for the proposed DHI strategies.

Figure 9c shows the estimated QoS satisfaction behavior based on the spectral-efficiency threshold used in the evaluation, while Figure 9d presents the rolling standard deviation of the sum spectral efficiency as a stability indicator. A lower fluctuation level indicates more stable downlink performance over time. Overall, the figure confirms that the proposed DHI framework can maintain stable downlink performance under dynamic network conditions, with DHI-Max-Sum-Rate being more suitable for throughput-oriented operation and DHI-Max-Product providing a balanced performance trend.

The overall results confirm that the proposed DHI framework provides a flexible and adaptive solution for dynamic downlink power allocation in CF-mMIMO systems. Instead of using a single fixed objective, the proposed framework supports three operating strategies: DHI-Max-Min, DHI-Max-Product, and DHI-Max-Sum-Rate. Each strategy produces a different performance behavior according to the selected downlink objective as shown in Table 7.

Table 7. Summary of key findings.

Strategy	Main Strength	Main Limitation	Recommended Scenario
DHI-Max-Min	Highest QoS satisfaction and lowest outage	Lower sum spectral efficiency and higher computational time than the other DHI strategies	QoS-oriented downlink service and weak user protection
DHI-Max-Product	Balanced SINR, QoS support, spectral efficiency, and very low computational time	Does not achieve the highest throughput	Balanced dynamic downlink operation
DHI-Max-Sum-Rate	Highest sum spectral efficiency and very low computational time	Lower QoS satisfaction compared with DHI-Max-Min	Throughput-priority CF-mMIMO operation

DHI-Max-Min provides the strongest QoS-oriented performance. It achieves the highest QoS satisfaction rate and the lowest outage probability because it prioritizes the weakest users and improves minimum user service. This makes it suitable for scenarios where service reliability is more important than maximizing total throughput. However, this strategy provides the lowest mean sum spectral efficiency because more power is assigned to users with weaker channel conditions.

DHI-Max-Product provides the most balanced behavior among the three strategies. It improves the SINR distribution across users while maintaining better spectral-efficiency performance than DHI-Max-Min. Its low computational time also makes it suitable for practical dynamic downlink operation. Therefore, DHI-Max-Product can be considered an appropriate strategy when the network requires a balance between QoS support, SINR improvement, spectral-efficiency performance, and execution time.

DHI-Max-Sum-Rate achieves the highest sum spectral efficiency. This confirms that the sum-rate objective is effective when the main goal is to maximize aggregate downlink throughput. However, its QoS satisfaction rate is lower than DHI-Max-Min and DHI-Max-Product because throughput-oriented allocation may favor users with stronger channel conditions. Therefore, DHI-Max-Sum-Rate is most suitable for throughput-priority scenarios where maximizing total system spectral efficiency is the main requirement.

The comparison with the DDPG benchmark further confirms the advantage of the proposed SAC-based DHI framework. DHI-Max-Sum-Rate provides stronger sum spectral-efficiency behavior than DDPG, while DHI-Max-Product and DHI-Max-Sum-Rate achieve much lower computational time. This improvement is mainly due to the combination of SAC-based adaptive learning and L-BFGS-B refinement, which allows the proposed

framework to generate effective power-allocation decisions without solving the complete optimization problem from the beginning at every time step.

The results also clarify that QoS in this paper should be interpreted as a QoS-aware performance objective rather than a strict guarantee for all users at all time steps. Under mobility, interference, and bounded transmit power, it is not always feasible for every UE to satisfy the SINR threshold simultaneously. For this reason, QoS is evaluated using QoS satisfaction rate and outage probability. This interpretation is consistent with the reported results and avoids over-claiming absolute QoS guarantee.

Overall, the proposed DHI framework demonstrates three main advantages. First, it adapts to dynamic downlink conditions through SAC-based learning. Second, it provides flexible strategy selection according to QoS-oriented, balanced, or throughput-oriented requirements. Third, it reduces computational time compared with the DDPG benchmark, especially for DHI-Max-Product and DHI-Max-Sum-Rate. These findings support the suitability of the proposed framework for dynamic downlink power allocation in CF-mMIMO systems.

6. Limitations and Practical Considerations

Although the proposed DHI framework shows promising performance for dynamic downlink power allocation in CF-mMIMO systems, several limitations should be acknowledged. First, the evaluation is based on simulation results with 64 APs and 32 UE. Larger network sizes, heterogeneous user demands, and different mobility patterns should be investigated in future work to further validate scalability.

Second, the system model uses an effective scalar AP–UE channel representation. This assumption is suitable for evaluating power allocation, large-scale fading, and mobility effects, but it does not fully capture practical multi-antenna precoding, pilot contamination, synchronization errors, hardware impairments, or fronthaul capacity limitations.

Third, the proposed framework requires offline SAC training before online deployment. Although the trained actor can generate power-allocation decisions with low inference time, the training phase may require significant computational resources when the network size or state dimension increases.

Fourth, QoS is treated as a QoS-aware performance objective rather than a strict guarantee for every UE at every time step. Under mobility, interference, and bounded transmit power, some UE may not always satisfy the SINR threshold. Therefore, QoS is evaluated using QoS satisfaction rate and outage probability.

Finally, the benchmark comparison is mainly based on DDPG and the proposed DHI strategies. Future work should include additional comparisons with WMMSE, fractional programming, convex optimization, multi-agent reinforcement learning, and federated learning approaches to provide broader validation.

7. Conclusions

In this paper, we studied the dynamic downlink power allocation in cell-free massive multiple-input multiple-output (MIMO) networks and proposed a Deep Hybrid Intelligent (DHI) framework, which integrates deep reinforcement single-agent learning with max-min, max-product, and max-sum-rate strategies. The proposed framework demonstrated promising results, with improved long-term downlink performance in spectral efficiency, QoS support, and computational efficiency. Compared with the DDPG benchmark, the DHI-based methods showed better spectral-efficiency behavior, achieving a mean computational time of 0.0690 s for DHI-MAX-PROD and 0.0696 s for DHI-MAX-SUM-RATE compared to 0.63 s for DDPG. The numerical results strengthen this intuitive conclusion by highlighting how the coupling between learning-adaptive policies and the optimization-based control

can provide tangible gains in CF-mMIMO downlink power allocation. This work is limited to simulation-based evaluation, the fixed network that was considered in this study, and the limited set of benchmark comparisons. Moreover, as SAC and L-BFGS-B are integrated together, perhaps it makes it too complex to implement in real-time, especially over big scales. These limitations indicate helpful directions for future work: testing on larger scenarios, more general benchmark comparisons, and implementing our solutions under real deployment conditions.

Author Contributions: Conceptualization, H.A.J. and M.F.A.R.; methodology, H.A.J.; software, H.A.J.; validation, H.A.J., M.F.A.R., F.H. and S.M.; formal analysis, H.A.J.; investigation, H.A.J.; resources, M.F.A.R., F.H. and S.M.; data curation, H.A.J.; writing—original draft preparation, H.A.J.; writing—review and editing, M.F.A.R., F.H. and S.M.; visualization, H.A.J.; supervision, M.F.A.R., F.H. and S.M.; project administration, M.F.A.R.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. Any external funding body did not fund the APC.

Data Availability Statement: The data supporting the findings of this study were generated through simulation experiments conducted by the authors. The simulation parameters and methodological details are provided within the article. Additional data may be made available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to acknowledge the Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia, for providing academic support and research facilities during the preparation of this work.

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

References

1. Wiffen, F.D. Advanced MIMO Techniques for Future Wireless Communications. Doctoral Dissertation, University of Bristol, Bristol, UK, 2021.
2. Xu, Y.; Larsson, E.G.; Jorswieck, E.A.; Li, X.; Jin, S.; Chang, T.-H. Distributed signal processing for extremely large-scale antenna array systems: State-of-the-art and future directions. *IEEE J. Sel. Top. Signal Process.* **2025**, *19*, 304–330. [\[CrossRef\]](#)
3. Ahmed, S.K. Cell-Free Massive Multiple-Input Multiple-Output Under Open Radio Access Network Flexible Functional Splits Towards Efficient Cellular Network. Doctoral Dissertation, University of British Columbia, Vancouver, BC, Canada, 2025.
4. Sadrani, L.K.; Rajput, R.S. Cell-Free Massive MIMO: Concepts, Advances, and Future Research Challenges. *Int. J. Multidiscip. Res. (IJFMR)* **2026**, *8*, 1–18. [\[CrossRef\]](#)
5. Wang, Y.; Wu, S.; Lei, C.; Jiao, J.; Zhang, Q. A review on wireless networked control system: The communication perspective. *IEEE Internet Things J.* **2023**, *11*, 7499–7524. [\[CrossRef\]](#)
6. Mohammadi, M.; Mobini, Z.; Ngo, H.Q.; Matthaiou, M. Next-Generation Multiple Access With Cell-Free Massive MIMO. *Proc. IEEE* **2024**, *112*, 1372–1420. [\[CrossRef\]](#)
7. Conceição, F.; Antunes, C.H.; Gomes, M.; Silva, V.; Dinis, R. Max-Min Fairness Optimization in Uplink Cell-Free Massive MIMO Using Meta-Heuristics. *IEEE Trans. Commun.* **2022**, *70*, 1792–1807. [\[CrossRef\]](#)
8. Mendoza, C.F. Deep Reinforcement Learning for Cell-Free Massive MIMO Network Optimization. Doctoral Dissertation, Technische Universität Wien, Vienna, Austria, 2025.
9. Ramezani, P.; Khorsandmanesh, Y.; Björnson, E. Joint discrete precoding and RIS optimization for RIS-assisted MU-MIMO communication systems. *IEEE Trans. Commun.* **2024**, *73*, 1531–1546. [\[CrossRef\]](#)
10. Sui, Z.; Ngo, H.Q.; Van Chien, T.; Matthaiou, M.; Hanzo, L. RIS-assisted cell-free massive MIMO relying on reflection pattern modulation. *IEEE Trans. Commun.* **2024**, *73*, 968–982. [\[CrossRef\]](#)
11. Ngo, H.Q.; Ashikhmin, A.; Yang, H.; Larsson, E.G.; Marzetta, T.L. Cell-free massive MIMO versus small cells. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 1834–1850. [\[CrossRef\]](#)
12. Björnson, E.; Sanguinetti, L. Scalable Cell-Free Massive MIMO Systems. *IEEE Trans. Commun.* **2020**, *68*, 4247–4261. [\[CrossRef\]](#)
13. Demir, Ö.T.; Björnson, E.; Sanguinetti, L. Foundations of user-centric cell-free massive MIMO. *Found. Trends Signal Process.* **2021**, *14*, 162–472. [\[CrossRef\]](#)

14. Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. Soft Actor-Critic Algorithms and Applications. *arXiv* **2019**, arXiv:1812.05905. [[CrossRef](#)]
15. Zhu, G.; Lyu, Z.; Jiao, X.; Liu, P.; Chen, M.; Xu, J.; Cui, S.; Zhang, P. Pushing AI to wireless network edge: An overview on integrated sensing, communication, and computation towards 6G. *Sci. China Inf. Sci.* **2023**, *66*, 130301. [[CrossRef](#)]
16. Banerjee, B.; Elliott, R.; Krzymien, W.; Medra, M. Access Point Clustering in Cell-Free Massive MIMO Using Conventional and Federated Multi-Agent Reinforcement Learning. *IEEE Trans. Mach. Learn. Commun. Netw.* **2023**, *1*, 107–123. [[CrossRef](#)]
17. Gong, B.; Huang, G.; Tu, W. Minimize BER without CSI for dynamic RIS-assisted wireless broadcast communication systems. *Comput. Netw.* **2024**, *253*, 110729. [[CrossRef](#)]
18. Yuan, M.; Zhang, W.; Huang, G.; Tu, W. Joint operating mode and resource allocation optimization in wireless-powered RIS-assisted multiuser communication systems. *Comput. Netw.* **2025**, *272*, 111650. [[CrossRef](#)]
19. Zhao, Y.; Niemegeers, I.G.; De Groot, S.M.H. Dynamic power allocation for cell-free massive MIMO: Deep reinforcement learning methods. *IEEE Access* **2021**, *9*, 102953–102965. [[CrossRef](#)]
20. Sheela Rani, N.; Vishnu Vardhan, D.; Mamatha, G.; Gunjan, V.K.; Shaik, F. Cell-Free Massive MIMO Versus Small Cells. In *International Conference on Information and Management Engineering; Cognitive Science and Technology*; Springer: Singapore, 2023; pp. 23–32. [[CrossRef](#)]
21. Nayebi, E.; Ashikhmin, A.; Marzetta, T.L.; Yang, H.; Rao, B.D. Precoding and Power Optimization in Cell-Free Massive MIMO Systems. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 4445–4459. [[CrossRef](#)]
22. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [[CrossRef](#)]
23. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. In *Proceedings of the 4th International Conference on Learning Representations (ICLR 2016)*, San Juan, Puerto Rico, 2–4 May 2016. Available online: <https://arxiv.org/abs/1509.02971> (accessed on 26 May 2026).
24. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, Stockholm, Sweden, 10–15 July 2018; *Proceedings of Machine Learning Research*; PMLR: Cambridge, MA, USA, 2018; Volume 80, pp. 1861–1870.
25. Zaher, M.; Demir, O.T.; Björnson, E.; Petrova, M. Learning-Based Downlink Power Allocation in Cell-Free Massive MIMO Systems. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 174–188. [[CrossRef](#)]
26. Chakraborty, S.; Demir, Ö.T.; Björnson, E.; Giselsson, P. Efficient downlink power allocation algorithms for cell-free massive MIMO systems. *IEEE Open J. Commun. Soc.* **2020**, *2*, 168–186. [[CrossRef](#)]
27. Fujimoto, S.; Van Hoof, H.; Meger, D. Addressing Function Approximation Error in Actor-Critic Methods. In *International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2018.
28. Peng, Q.; Ren, H.; Pan, C.; Liu, N.; El Kashlan, M. Resource allocation for cell-free massive MIMO-enabled URLLC downlink systems. *IEEE Trans. Veh. Technol.* **2023**, *72*, 7669–7684. [[CrossRef](#)]
29. Ghazanfari, A.; Member, S.; Cheng, H.V.; Björnson, E.; Member, S.; Erik, G. Enhanced Fairness and Scalability of Power Control Schemes in Multi-Cell Massive MIMO. *IEEE Trans. Commun.* **2020**, *68*, 2878–2890. [[CrossRef](#)]
30. Byrd, R.H.; Lu, P.; Nocedal, J.; Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.