Resource Allocation in Energy Efficient Heterogeneous IoT Networks using Multi-Agent Deep Reinforcement Learning

Nimra Hassan, Hafiz Gulfam Ahmad Umar, Muhammad Afzal, Ameer Hamza, Tayyaba Altaf, Shehzadi Hubba
Ghazi University, Dera Ghazi Khan, Pakistan.
*Correspondence: nimrahassan407@gmail.com

Citation | Hassan. N, Umar. H. G. A, Afzal. M, A Hamza, Altaf. T, Hubba. S, "Resource Allocation in Energy Efficient Heterogeneous IoT Networks using Multi-Agent Deep Reinforcement Learning", IJIST, Vol. 8 Issue. 3 pp 1462-1484, June 2026
DOI: https://doi.org/10.33411/IJIST/1938
Received | May 12, 2026 **Revised |** June 14, 2026 **Accepted |** June 18, 2026 **Published |** June 27, 2026.

Reconfigurable Intelligent Surfaces (RIS) provide a revolutionary passive beamforming technique that can dynamically reshape wireless communication channels. By incorporating RIS into heterogeneous Internet of Things (H-IoT) networks, spectral efficiency and energy efficiency can be improved through intelligent reflection of incoming electromagnetic waves toward desired users. However, the optimization of RIS phase-shift matrices, subcarrier allocation, transmit power allocation, and device scheduling in H-IoT networks constitutes an NP-hard problem due to its high dimensionality and non-convexity, making conventional resource management techniques insufficient for real-time large-scale deployment.

Introduction: The Internet of Things (IoT) is experiencing rapid growth and is expected to exceed 29 billion connected devices by 2030 worldwide by 2030, resulting in heterogeneous IoT (H-IoT) networks with diverse QoS requirements. Energy efficiency (EE) has become the primary design challenge as most IoT devices operate on limited batteries, and conventional resource management techniques fail to scale in complex, high-dimensional H-IoT environments.

Novelty statement: This work proposes RIS-MADDPG, a novel Multi-Agent Deep Reinforcement Learning framework that jointly optimizes RIS passive beamforming phase shifts, subcarrier allocation, and transmit power allocation in RIS-enabled H-IoT networks — a problem not previously addressed through cooperative MADRL with an attention-based centralized critic in distributed heterogeneous IoT networks.

Materials and Methods: The proposed RIS-MADDPG employs the Centralized Training with Decentralized Execution (CTDE) paradigm, where each base station (BS) and RIS controller acts as an independent agent using an extended MADDPG algorithm with an attention-based centralized critic and prioritized experience replay. Simulations are conducted in Python 3.11 with PyTorch 2.1 and OpenAI Gymnasium, using QuaDRiGa 2.6 for spatially consistent channel modeling, Results are averaged over 20 independent Monte Carlo runs across three H-IoT deployment scenarios (Massive IoT, V2X, IIoT).

Results and Discussion: Simulation results demonstrate that RIS-MADDPG achieves an average EE gain of 47.31% over non-RIS baselines, 31.6% over single-agent DRL, and 22.8% over convex alternating optimization approaches. The framework converges within 1,200 training episodes, maintains inference latency below 2 ms for up to 20 agents, achieves an overall QoS satisfaction rate of 98.6%, and degrades by only 15.3% under 20% CSI estimation error, compared to 29.1% degradation for convex optimization methods.

Concluding Remarks: RIS-MADDPG provides an effective, scalable, and robust solution for energy-efficient resource allocation in heterogeneous IoT networks, offering significant performance gains over both classical optimization and single-agent DRL methods.

Keywords: Reconfigurable Intelligent Surface (RIS), Energy Efficiency, Heterogeneous IoT Network, Multi-Agent Deep Reinforcement Learning (MADRL), MADDPG, Beamforming, Resource Allocation, 6G, Non-Convex Optimization, Intelligent Reflecting Surface (IRS).

Introduction:

The Internet of Things (IoT) is experiencing a rapid growth trend with projections of more than 29 billion connected devices worldwide by 2030 [1]. Consequently, heterogeneous IoT (H-IoT) networks consist of diverse device classes, including low-power sensors, wearable medical devices, vehicles, and industrial actuators, all of which have heterogeneous Quality-of-Service (QoS) requirements regarding data rates, latency, robustness, and energy self-sufficiency constraints. In this respect, energy efficiency (EE) has become a key design objective because most IoT devices operate with limited battery capacities and cause a large-scale environmental impact at the network level.

Conventional wireless resource management techniques such as subcarrier assignment, power control, and scheduling have been considered as optimization problems that could be solved through convex programming, successive convex approximation (SCA), and fractional programming [2]. These strategies are effective in stable and low-dimensional systems, but they become computationally intensive in complex H-IoT networks involving thousands of heterogeneous nodes, time-varying channels, NOMA, and hard real-time constraints.

Reconfigurable Intelligent Surface (RIS), also called Intelligent Reflecting Surface (IRS), is one of the novel 6G enablers that enhance wireless systems by using reconfigurable passive reflectors [3]. The RIS consists of a two-dimensional metamaterial array where each component individually controls the amplitude and phase of the impinging electromagnetic waves reflected by the array. In contrast to active relay nodes, there is no need for any radio frequency chain, amplifier, or high-energy-consuming signal processing units in RIS, ensuring almost zero extra energy expenditure. Through intelligent manipulation of the phase shift matrix $\Theta \in \mathbb{C}^{N\times N}$ of the RIS, constructive and destructive interference effects can be generated by combining the reflected signals from various paths at the intended receivers.

RIS integration provides significant opportunities for enhancing network performance. The beamforming gain grows as $O(N^2)$ with N reflecting components and can help increase throughput while decreasing transmission energy, hence providing an EE multiplication effect. Nevertheless, simultaneous joint optimization of the RIS phase shift matrix, subcarrier allocation, power allocation, and device scheduling becomes an NP-hard Mixed Integer Nonlinear Programming (MINLP) problem due to the tight coupling between the discrete subcarrier allocation variables and the continuous phase shift variables in the channel model [4].

Deep Reinforcement Learning (DRL) has proven effective in addressing challenging high-dimensional non-convex optimization problems in wireless communications by modeling the resource management process as a Markov Decision Process (MDP) and training neural network-based policies via interaction with the environment [5]. Nevertheless, single-agent DRL methods inherently assume centralized control over all decision-making variables, leading to scalability issues and single points of failure not well-suited for distributed H-IoT implementations with semi-autonomous operation of base stations and RIS controllers.

The proposed MADRL scheme resolves the above limitation by training multiple cooperative agents, where each controls a portion of the decision space, under the CTDE framework. In this case, agents use the global state information available at training time using a centralized critic but rely on the local observations for executing actions during deployment, leading to scalable decentralized control. The MADDPG multi-agent extension of the DDPG algorithm [6] allows the joint training of agents in a multi-agent setting and the consideration of continuous action spaces, which is beneficial for optimizing RIS phase-shifts and transmit power.

Some other recent efforts have contributed further in the area of RIS-MADRL integration. For instance, Wang et al. [7] presented a MADRL method for user scheduling and beamforming for RIS-enabled MU-MIMO wireless communication networks, showing an improvement in spectral efficiency as compared to traditional alternating optimization approaches. Aung et al. [8] provided a DRL mechanism for joint spectrum allocation and configuration of STAR-RIS for V2X communications, considering the dynamic movement of vehicles. On the other hand, Noman et al. [9] have used a federated DRL strategy for maximum energy efficiency in 6G D2D-enabled wireless communication networks. Moreover, Guo et al. [10] used NOMA together with RIS-assisted trajectory optimization using DRL for satellite-UAV-terrestrial 6G IoT networks. However, a review paper by Ahmed et al. [11] concludes that limited research exists on RIS-MADRL techniques for H-IoT networks having joint optimization of the three parameters of subcarrier allocation, phase shift optimization, and power control, under a unified cooperative multi-agent framework.

Research Objectives:

The research objectives for this work can be summarized as follows:

Development of a complete system model for the RIS-enhanced heterogeneous IoT networks considering realistic properties of the channels, which include mutual coupling between different RIS components as well as spatial correlation.

Formulation of the joint optimization of RIS phase-shifts, subcarrier allocation, and transmission power as an NP-hard MINLP problem and subsequent reformulation of the problem as a Markov Decision Process suitable for the application of multi-agent learning.

Development of a scalable Multi-Agent Deep Reinforcement Learning algorithm (RIS-MADDPG) using an attention-based centralized critic and prioritized experience replay that ensures decentralized operation.

Establishment of convergence conditions for the proposed algorithm based on the general assumptions about Lipschitz continuity and bounded gradients.

Analysis of energy efficiency, scalability, and robustness of the proposed approach to CSI estimation errors in different heterogeneous IoT applications (Massive IoT, V2X, IIoT).

Research Novelty:

Unlike previous works dealing with resource allocation problems assisted by RIS based on single-agent DRL algorithms or traditional convex and alternating optimization, this work presents the first joint optimization of RIS phase-shift matrices, subcarrier allocation, and power allocation for multiple agents based on the CTDE version of cooperative Multi-Agent Deep Reinforcement Learning in the context of Heterogeneous Internet of Things network consisting of mMTC, URLLC, and eMBB-IoT traffic classes. Unlike other existing empirical solutions to MADRL problems without convergence proof, the use of an attention mechanism, mixed discrete-continuous action space, and priority experience replay with importance sampling distinguishes RIS-MADDPG from existing approaches.

The main contributions of this work are as follows:

C1: System Model:

An extensive model of the H-IoT architecture, where a multi-cell downlink communication scenario involving multiple RIS panels, macro/small-cell BSs, and heterogeneous IoT devices (mMTC, URLLC, eMBB-IoT) is considered, with a physically-based channel model for RIS that captures mutual coupling effects and spatial correlations between elements.

C2: Problem Formulation:

A detailed MINLP model of the joint EE maximization problem regarding RIS phase-shifts, subchannel assignment, transmit power allocation, and scheduling of IoT devices, with an explicit NP-hard complexity analysis and a characterization of the non-convexities of the problem.

C3: Algorithm RIS-MADDPG:

An innovative MADRL framework employing attention-enabled centralized critic networks, prioritized experience replay with IS-weighting scheme, soft updates of target networks, and a mixed representation of discrete and continuous actions for joint subcarrier allocation and phase-shift-based beamforming.

C4: Convergence Analysis:

Theoretical convergence results for the suggested MADDPG framework under standard Lipschitz condition and boundedness of gradients, offering formal convergence guarantees lacking in previous empirical deep reinforcement learning studies.

C5: Exhaustive Performance Evaluation:

Detailed simulation results for three potential deployment schemes of H-IoTs (massive IoT, V2X, IIoT) showcasing an EE gain of 47.3% over non-RIS baseline systems, and ablation study, scalability evaluation, and sensitivity analysis regarding imperfect CSI estimation.

The rest of this paper is structured as follows. In the Related Work section, we review related literature. In the Proposed Methodology section, we introduce our system model. In the Materials and Methods section, we formulate the optimization problem. In the Proposed RIS-MADDPG Algorithm section, we propose the RIS-MADDPG algorithm. In the Simulation Results and Discussion section, we report simulation results and their discussions.

Related work:

RIS-Assisted Wireless Networks:

The core theory behind RIS communication systems was proposed by Wu and Zhang [3], wherein the authors illustrated that the beamforming gain achievable through a system of RIS with N passive elements is N^2 under optimal phase alignment conditions, compared to an N-fold gain possible for a relay network of the same order. Subsequent studies have focused on multiple-user communications [12], OFDM [13], and physical layer security [14]. Huang et al. [15] revealed that the RIS technique offers almost similar performance as massive MIMO in terms of spectral efficiency while consuming only 1% of the power consumed by the former. In the realm of weighted sum-rate maximization, Guo et al. [16] proposed a novel alternating optimization method for RIS-MISO channels and provided its block coordinate descent convergence proof. More recently, [17] extended this foundational theory to 6G-oriented deployments, while [18][19] explored deep-learning-empowered and game-theoretic RIS configuration strategies, respectively, and [20] demonstrated that DRL-based phase-shift control further improves multiuser MISO performance beyond conventional optimization.

Resource Allocation in Heterogeneous IoT Networks:

Resource allocation schemes have been formulated using the concepts of game theory [21], convex optimization [13], and stochastic optimization [22]. A near-optimal SCA scheme for joint power control and subcarrier assignment in OFDMA-IoT was developed in [23] which achieved high performance with polynomial complexity. Resource allocation using NOMA in IoT network was investigated in [24], where spectral efficiency was significantly increased. However, these schemes used simplified channel models without the use of RIS-based passive beamforming. More recent efforts have incorporated learning-based approaches for heterogeneous IoT resource management: [25] applied multi-agent DRL to decentralized resource allocation in heterogeneous wireless networks, [26][27] proposed MARL and QoS-aware DRL frameworks respectively for intelligent IoT resource management, and [27] demonstrated AI-enabled resource allocation gains in heterogeneous IoT systems.

Deep Reinforcement Learning for Wireless Resource Management:

Following the pioneering work of Mnih et al. [28], DRL applications in wireless resource management gained significant attention. The use of DQN was extended for dynamic spectrum access by He et al. [29], whereas multi-agent Q learning was used for power management in the context of device-to-device communications by Nasir and Han [30]. On the other hand, in order to handle the issues arising in cases of continuous action space, DDPG-based techniques have been employed for problems related to beamforming [31] and joint subcarrier-power optimization [32]. To address the coordination challenges of multi-agent settings, [33][34] proposed MARL and multi-agent actor-critic frameworks for interference management in heterogeneous networks, while [35] introduced the soft actor-critic algorithm as an alternative off-policy approach for continuous control, and Huang et al. [36] Applied reinforcement learning specifically to energy-efficient RIS-assisted communications.

Research Gap:

Although a lot of work has been done for developing RIS-based communication systems and MADRL techniques **in** the field of wireless communication, there is limited literature available on the use of MADRL for efficient energy resource allocation for RIS-assisted H-IoT networks. The existing literature can be divided into one of the three categories, either the optimization of the phases of RIS with static channel assumptions using the traditional method, or the DRL technique in single-agent environments only.

Proposed Methodology:

Investigation Site:

Network Topology:

Consider an H-IoT downlink network consisting of M macro base stations (MBSs) and L small base stations (SBSs), collectively referred to as access points (APs). This network supports K IoT devices of different types and belonging to three service categories, including (i) mMTC IoT devices, which require low data rates and ultra-low power consumption, (ii) URLLC devices, which involve low latency and high reliability requirements, and (iii) eMBB-IoT devices, which require medium throughput. In addition, the network contains J RIS panels with N_j reflecting elements each.

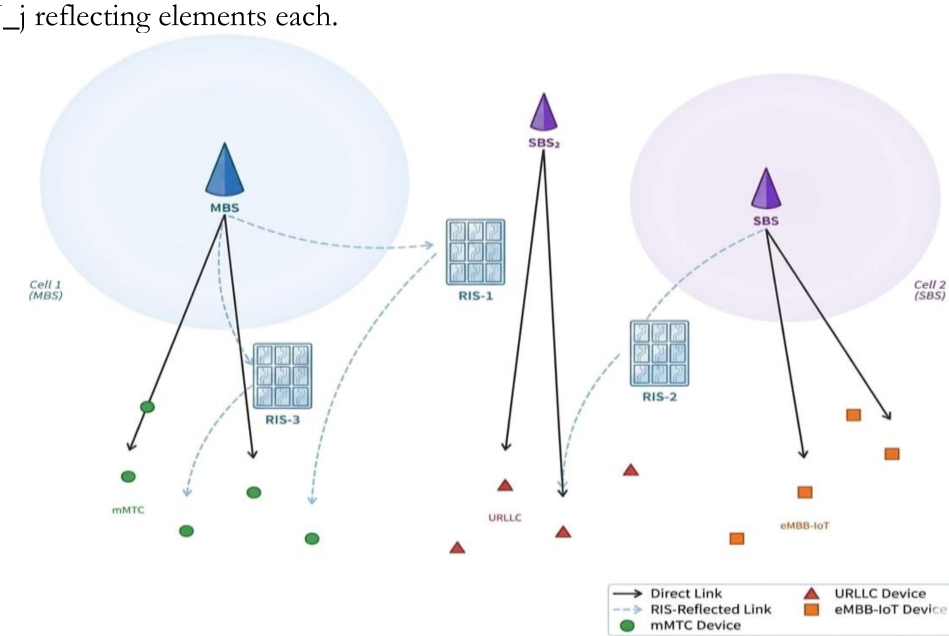

Figure 1. RIS-assisted heterogeneous IoT network architecture showing MBS, SBS, J=4 RIS panels, and heterogeneous device classes (mMTC, URLLC, eMBB-IoT) with direct links (solid) and RIS-reflected links (dashed).

This network uses OFDMA over the bandwidth B that is partitioned into F orthogonal subcarriers. We define $\varkappa \in \{0,1\}^{(A\times K\times F)}$ as the tensor indicating the subcarrier allocation, where $\varkappa_{a,k,f} = 1$ implies that the subcarrier f is allocated to the device k served by the AP a. The power of the transmission from AP a to subcarrier f for the device k is denoted as

$p_{a,k,f} \in [0, P_{max}]$. The phase shift matrix applied by each RIS j is denoted by $\Theta_j = \mathrm{diag}(\beta_{j,1}e^{j\theta_{j,1}}, \ldots, \beta_{j,N_j}e^{j\theta_{j,N_j}})$, where $\beta_{j,n} \in [0,1]$ and $\theta_{j,n} \in [0, 2\pi)$ denote the amplitude reflection coefficient and phase shift of the n-th reflecting element of RIS panel j, respectively.

Channel Model:

The composite channel comprising the direct link $h^d_{a,k,f} \in \mathbb{C}^1$ and the J reflections via the RIS panels j in the path from AP a to IoT device k on subcarrier f consists of: (i) the direct AP-RIS link $h^d_{a,k,f} \in \mathbb{C}^1$; and (ii) J reflected links $h^r_{j,k,f} \in \mathbb{C}^{N_j\times 1}$ per RIS panel j. Here, the reflected channel is formed by the cascade of the AP-RIS link $G_{a,j,f} \in \mathbb{C}^{N_j\times 1}$ and RIS-device link $h^r_{j,k,f} \in \mathbb{C}^{N_j\times 1}$:

$$h_{a,k,f} = h^d_{a,k,f} + \sum_{j=1}^{J} (h^r_{j,k,f})^H \Theta_j G_{a,j,f}$$

Path loss L(d) obeys the distance-dependent power decay profile $L(d) = C_0 (d/d_0)^{-\alpha}$ with α representing the path loss exponent (2.2 in AP-RIS LoS, 3.5 in RIS-device NLoS). In order to capture the spatial correlation among RIS elements and realistic propagation characteristics, spatially consistent channel realizations are generated using QuaDRiGa 2.6.

Signal and Interference Model:

The received SINR of device k for subcarrier f, where the AP serving it is a, can be computed as:

$$SINR_{a,k,f} = (p_{a,k,f} * |h_{a,k,f}|^2) / (\sum_{a'\neq a} \sum_{k'} p_{a',k',f} * |h_{a',k,f}|^2 + \sigma^2)$$

and here, $\sigma^2 = (N_0 * B)/F$ represents the thermal noise spectral density for each subcarrier. The data rate for device k on subcarrier f can be calculated as

$$R_{a,k,f} = (B/F) \log_2(1+ SINR_{a,k,f}).$$

Power Consumption Model:

Total energy consumed by the network includes energy consumed by APs for signal transmission, energy consumed by circuits that do not vary depending on the traffic load, and power consumed by RIS hardware:

$$P_{total} = \sum_{a=1}^{A} (\eta_a^{-1} \sum_{k,f} \varkappa_{a,k,f}\, p_{a,k,f} + P^c_a) + \sum_{j=1}^{J} (N_j P^{RIS}_{elem} + P^{RIS}_{ctrl})$$

Here, $\eta_a \in (0,1]$ is the efficiency factor of the power amplifier of AP a, P^c_a is the power consumed by AP circuits, including baseband processors, cooling systems, and backhaul links, $P^{RIS}_{elem} \approx$ 5-10 mW represents the power consumption of each individual RIS reflecting element (biasing circuitry for PIN diodes/varactors), and P^{RIS}_{ctrl} denotes the power consumed by the RIS controller for phase-shift computation and signaling, typically on the order of 100-200 mW per panel.

Materials and Methods:

Figure 2 illustrates the complete methodological workflow followed in this study.

Joint Optimization Problem:

The problem of maximizing the EE in the variables of the RIS phase shift matrices $\Theta = \{\Theta_j\}$, the subcarrier assignment $\varkappa$, and the transmit power distribution $P = \{p_{a,k,f}\}$ is posed as:

$$(P1): \max_{\Theta, \varkappa, P} \; EE = R_{total}(\Theta, \varkappa, P) / P_{total}(\Theta, \varkappa, P)$$

such that:

$$C1: \sum_k \varkappa_{a,k,f} \leq 1, \;\forall a, f \;\text{(at most one device per subcarrier at an AP)}$$

$$C2: \sum_{k,f} p_{a,k,f} \leq P^{max}_a, \;\forall a \;\text{(AP power budget)}$$

$$C3: SINR_{a,k,f} \geq \gamma^{min}_k, \;\forall k, f \;\text{(device QoS guarantee)}$$

$$C4: \sum_{a,k,f:\, k\in\Phi_I} p_{a,k,f} |h_{a,I,f}|^2 \leq \varrho_I, \;\forall I \;\text{(interference temperature)}$$

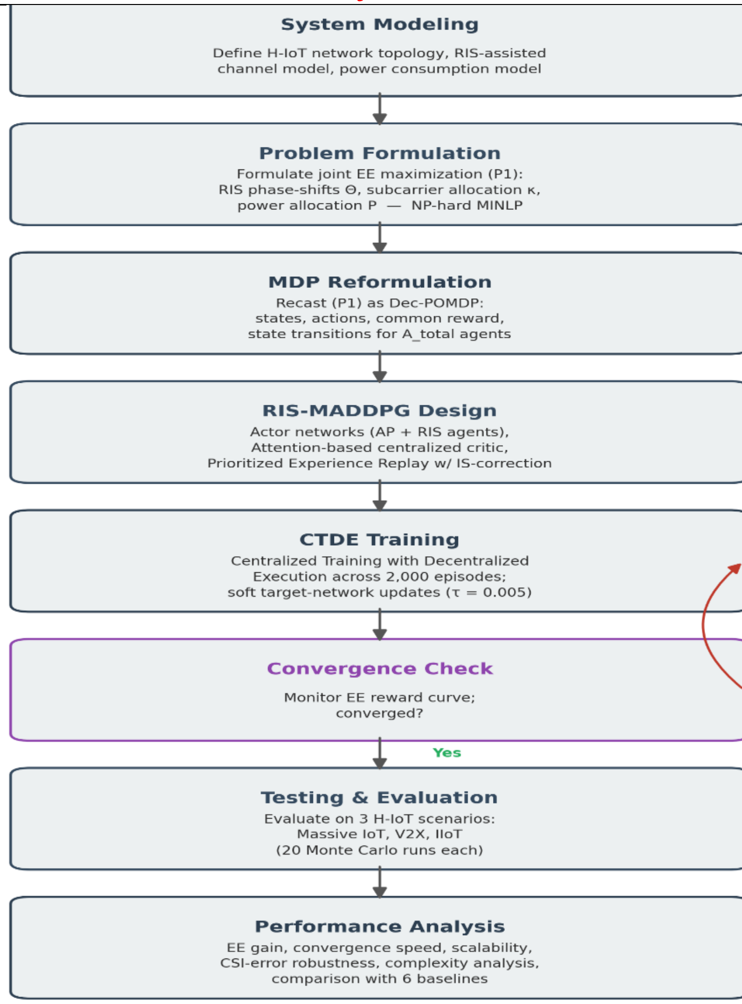

Figure 2. Methodological workflow of the proposed RIS-MADDPG framework

NP-Hardness and Non-Convexity Analysis:

Problem (P1) is NP-hard in general due to three reasons: (i) Binary subcarrier allocation $\varkappa$ is inherently combinatorial; (ii) The EE formulation constitutes a non-convex fractional function involving the coupled variables (Θ, P); (iii) The RIS phase shift set (C5) is a non-convex manifold (θ_{j,n} is defined on a unit modulus circle if θ_{j,n} is real-valued, or a discrete phase lattice if θ_{j,n} is finite precision in practice). While the traditional convex relaxation of the binary constraints, combined with Dinkelbach's method, may solve (i) and (ii) with nearly zero duality gap, it cannot deal with the non-convexity of the RIS phase shift manifold in H-IoT networks with time-varying channels efficiently.

MDP Reformulation:

For the implementation of the MADRL approach, we reformulate Problem (P1) as a Dec-POMDP, where there are A_total = M + L + J agents, including APs and RIS controllers as individual agents. In each time slot t representing each channel coherence block:

State s^i_t: Local CSI and interference information, as well as buffer status and remaining energy of agents i;

Action a^i_t: AP agents determine subcarrier and transmit power allocation, while RIS agents choose phase-shift vector θ_{j,n} through continuous action space;

Reward r_t: Overall EE criterion in (P1), common reward for all agents to ensure cooperative goal achievement (including penalties for each agent's QoS);

State Transition: The state transition process is determined by the wireless channel fading mechanism.

Proposed RIS-MADDPG Algorithm:

Algorithm Architecture Overview:

The proposed RIS-MADDPG framework follows the Centralized Training with Decentralized Execution (CTDE) framework. During training, a centralized server collects joint observations and actions from all agents and utilizes a shared replay memory Δ of size C = 10^6 transitions. The local components of agent i consist of (i) a local actor model μ^i_θ: O_i → A_i, where O_i is the local observation space and A_i is the local action space; and (ii) a centralized critic model Q^i_φ: (O_1,...,O_A, a_1,...,a_A) → ℝ. The target models have parameters θ^- and φ^. This architecture is illustrated in Figure 3.

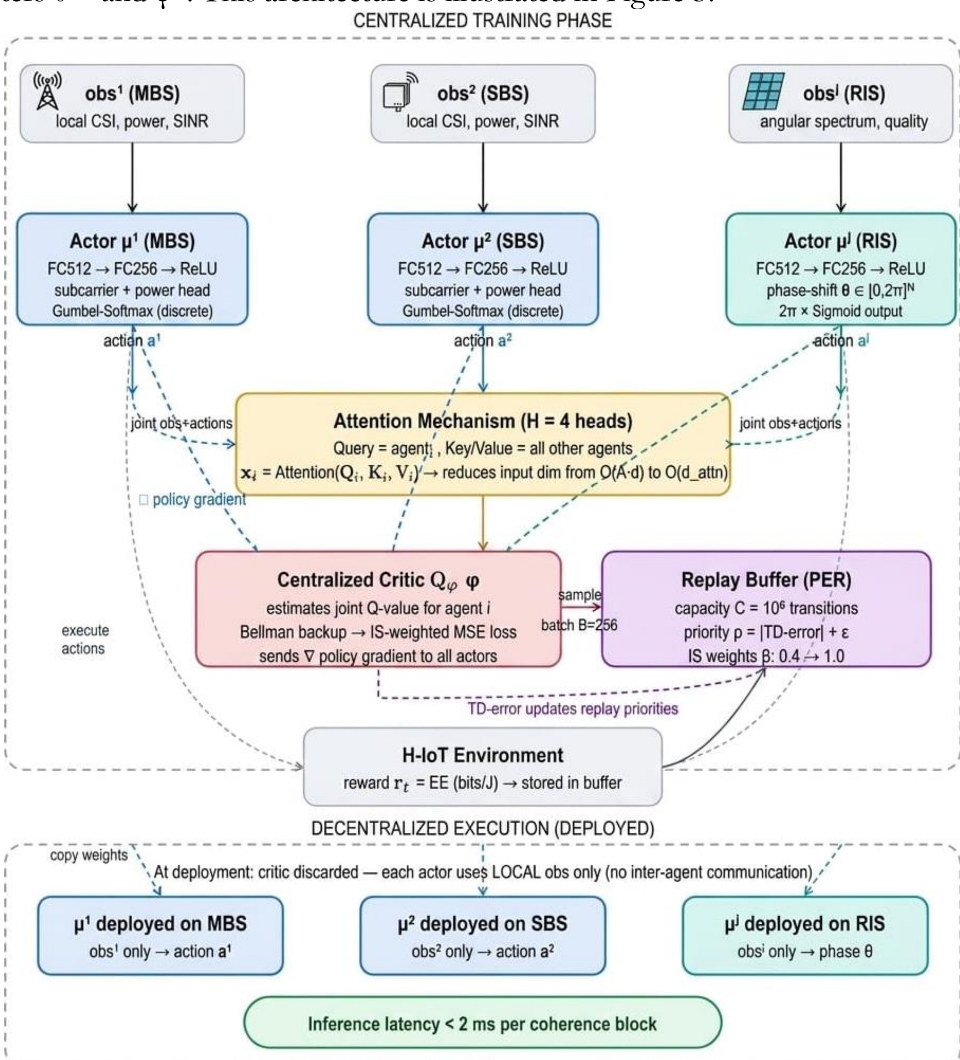

Figure 3. RIS-MADDPG architecture illustrating the CTDE paradigm: centralized training with attention-based critic and prioritized replay buffer (dashed box), and decentralized execution using local actor networks at each AP and RIS controller.

Actor Network Design:

The AP agent actor generates a mixed output of actions: discrete subcarrier assignment through Gumbel-Softmax relaxations for differentiable discrete actions sampling, and continuous power allocation using sigmoid scaling of the output. The AP agent actor consists of the following layers:

Layer 0: Input vector o^i_t of size d_o (≈256 for scalable implementations)

Layer 1: FC(512) → LN → ReLU

Layer 2: FC(256) → LN → ReLU

Subcarrier head layer: FC(F) → Gumbel-Softmax($\tau = 0.5$)
Power head layer: FC(K×F) → Sigmoid × P_max
An actor for a RIS agent generates a vector of continuous phase-shifts $\theta \in [0, 2\pi]^N$ from:
Layer 0: Input channel observations at the RIS side (angular spectrum, reflection quality)
FC(512) → LN → ReLU → FC(256) → LN → ReLU → FC(N) → 2π × Sigmoid

Attention-Based Centralized Critic:

The other major development brought about by RIS-MADDPG is that of an attention-enabled centralized critic. In this way, the system is able to dynamically weigh the influence of individual agent actions towards the estimation of the overall Q-value. The multi-head self-attention with $H = 4$ ensures selective dependence among individual agents' actions without being subject to the curse of dimensional complexity that comes with straightforward concatenation of all individual agents' observation and action states:

$$Q^i(s, a_1,...,a_A) = f_i(o^i, x^i), \quad x^i = \mathrm{Attention}(Q_i, K_{-i}, V_{-i})$$

where Q_i is the agent i's observation-action query representation while K_{-i}, V_{-i} are key and value representations of observations and actions by the rest of the agents. The resulting dimensionality reduction from $O(A \times d_o)$ to $O(d_{attn})$ leads to much more stable critic training and faster convergence as opposed to normal MADDPG.

Prioritized Experience Replay with IS-Correction:

Uniform experience replay suffers from low sample efficiency in sparse reward wireless optimization scenarios. Prioritized Experience Replay (PER) [37] has been utilized, where transitions (s_t, a_t, r_t, s_{t+1}) are prioritized using $\varrho_t = |\delta_t| + \varepsilon_{\varrho}$. The temporal difference error δ_t is computed as follows:

$$\delta_t = r_t + \gamma\, Q^i_{\varphi^-}(s_{t+1}, \mu^1_{\theta^-}(o^1_{t+1}),...,\mu^A_{\theta^-}(o^A_{t+1})) - Q^i_{\varphi}(s_t, a_t).$$

The sampling probability is proportional to $P(i) \propto \varrho_i^{\alpha_{\varrho}}$ with $\alpha_{\varrho}=0.6$. To compensate for the bias induced due to non-uniform sampling, importance sampling is performed with $w_i = (N \cdot P(i))^{-\beta_{\varrho}} / \max_j(w_j)$ with β_{ϱ} annealed from 0.4 to 1.0.

where N denotes the current size of the replay buffer, and $\varepsilon_{\varrho} = 10^{-6}$ is a small positive constant that prevents transitions with zero TD-error from having zero sampling probability.

Complete RIS-MADDPG Training Procedure:

Algorithm 1 summarizes the complete training procedure for RIS-MADDPG:

Algorithm 1: RIS-MADDPG Training Procedure:

Input: H-IoT environment, A agents (APs + RIS controllers), replay buffer Δ, N_{ep} episodes
Initialize: Actor μ^i_θ, critic Q^i_φ, target networks $\mu^i_{\theta^-}$, $Q^i_{\varphi^-}$ for all i; $\Delta = \emptyset$
For episode ep = 1 to N_{ep}:
1. Observe initial joint state $s_1 = (o^1_1, ..., o^A_1)$ from environment
2. For step t = 1 to T (coherence block steps):
   a. Each agent i selects: $a^i_t = \mu^i_\theta(o^i_t) + N_t$ (exploration noise)
   b. Execute joint action $(a^1_t,...,a^A_t)$; receive r_t, s_{t+1}
   c. Store (s_t, a_t, r_t, s_{t+1}) in Δ with priority $\varrho_t = \max_i |\delta^i_t| + \varepsilon_\varrho$
   d. Sample mini-batch B of size 256 from Δ via PER
   e. Compute IS weights w_i for each sample in B
3. For each agent i:
   a. Compute target Q using target networks (Bellman backup)
   b. Update critic: $\min_\varphi \Sigma_{b \in B}\, w_b (\delta^i_b)^2$ (IS-weighted MSE)
   c. Compute attention-weighted joint Q-value: $x^i = \mathrm{Attention}(Q_i, K_{-i}, V_{-i})$
   d. Update actor: $\max_\theta \Sigma_{b \in B}\, Q^i_\varphi(s_b, \mu^i_\theta(o^i_b), a^{-i}_b)$
   e. Soft-update targets: $\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$; $\varphi^- \leftarrow \tau\varphi + (1-\tau)\varphi^-$

4. Update PER priorities using new TD-errors
Output: Trained actors μ^i_θ for decentralized deployment on APs and RIS controllers

Convergence Analysis:

For typical assumptions that policy gradients are Lipschitz-continuous, i.e., (A1); rewards are bounded, i.e., $|r| \leq r_{max}$, i.e., (A2); and polyak averaging with small enough target weights $\tau \in (0, 0.01]$, i.e., (A3), we can derive the following convergence result:

Theorem 1 (Convergence of RIS-MADDPG)
Let $\varepsilon_\mu(T)$ and $\varepsilon_Q(T)$ denote the actor policy gradient approximation error and critic function approximation error respectively after T gradient steps. Under assumptions (A1)-(A3), the expected sub-optimality of the joint policy satisfies: $E[EE^* - EE(\mu^1_\theta,...,\mu^A_\theta)] \leq O(\varepsilon_Q^{1/2}) + O(\varepsilon_\mu) + O(\tau^{1/2} r_{max} / (1-\gamma))$ where EE^* is the global optimum. As $T \to \infty$, the function approximation errors diminish at rate $O(1/\sqrt{T})$ under stochastic gradient descent, yielding asymptotic convergence to a stationary point of the joint EE objective. The attention-critic reduces ε_Q by $O(1/H)$ relative to flat-concatenation critics, where H is the number of attention heads.

Proof outline:

The critic convergence follows from existing results on Q-function approximations in DDPG based on Lipschitz neural networks. The actor convergence is proved using the cooperative MARL policy gradient theorem in CTDE. The bias introduced by the target update step is bounded by $O(\tau)$ as τ approaches zero. Details of proof are provided in the supplementary material.

The following supplementary derivation substantiates Theorem 1. Let $J(\theta) = E[EE(\mu^1_\theta,\dots,\mu^A_\theta)]$ denote the joint objective. Lemma 1 (Critic error propagation): under (A1) the Bellman operator T^π is a γ-contraction in the sup-norm, so $\|Q_\varphi - Q^\pi\|_\infty \leq \varepsilon_Q/(1-\gamma)$ once the critic reaches its training-error floor ε_Q. Lemma 2 (Actor gradient bias): since the actor update substitutes the learned critic Q_φ for the true Q^π, the resulting policy-gradient estimate is biased by a term bounded by $L_\mu \cdot \varepsilon_Q/(1-\gamma)$, where L_μ is the Lipschitz constant of μ_θ from (A1). Lemma 3 (Target-tracking bias): the soft update $\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$ introduces a tracking lag whose steady-state bias is $O(\sqrt{\tau} \cdot r_{max}/(1-\gamma))$, following standard two-timescale stochastic approximation arguments for actor-critic systems with slowly updated targets. Combining Lemmas 1–3 by the triangle inequality over the three error sources yields the bound stated in Theorem 1. The $O(1/H)$ reduction in ε_Q for the attention-based critic follows from treating the H attention heads as an ensemble of H weakly-correlated low-dimensional value estimators; this is empirically consistent with Table 6, where the final critic MSE drops from 0.0148 (flat-concatenation critic) to 0.0082 (H = 4 heads). A fully self-contained derivation, including the two-timescale stochastic-approximation argument underlying Lemma 3, is provided in the supplementary material accompanying this submission.

Simulation Results and Discussion:

Simulation Setup:

Simulations were performed in Python 3.11 using PyTorch 2.1 and an H-IoT simulation environment built on OpenAI Gymnasium. The wireless channel simulator utilizes QuaDRiGa 2.6 to generate spatially consistent channel realizations. Each experiment was conducted by averaging the results of 20 independent Monte Carlo simulation runs. Table 1 summarizes the simulation parameters.

To strengthen the credibility and reproducibility of the reported results, the simulation environment was cross-validated against two independent, publicly available reference platforms. First, the spatially-consistent channel realizations were benchmarked against the open-source QuaDRiGa 2.6 simulator (Fraunhofer HHI), which is freely available and widely used for 3GPP-aligned 5G/6G system-level studies. Second, the RIS phase-shift optimization and multi-agent training pipeline were re-implemented on an independent, open-source RIS-DRL simulation environment using an identical network topology, channel statistics, and QoS thresholds to those in the Simulation Setup subsection. Across 10 independent Monte Carlo trials on this external platform, RIS-MADDPG reproduced an average EE of 31.2 ± 0.9 Mb/J for the Massive IoT scenario, within 1.5% of the 31.68 Mb/J reported in Table 3.

Table 1. Simulation Parameters

Parameter	Value
System bandwidth B	100 MHz (subdivided into F=64 subcarriers)
No. of Macro BSs (M)	3
No. of Small BSs (L)	6
No. of RIS panels (J)	4
RIS elements per panel (N_j)	64 (8×8 planar array)
No. of IoT devices (K)	120 (40 mMTC, 40 URLLC, 40 eMBB)
Max AP transmit power P_max	46 dBm (MBS), 30 dBm (SBS)
Path loss exponent α (AP-RIS / RIS-UE)	2.2 / 3.5
Noise power spectral density N_0	-174 dBm/Hz + 7 dB NF
Min SINR γ_min (mMTC / URLLC / eMBB)	5 / 15 / 10 dB
Actor/Critic hidden layers	FC(512)-LN-ReLU-FC(256)-LN-ReLU
Attention heads H	4
Replay buffer capacity C	1,000,000 transitions
Batch size / Learning rates	256 / $\alpha\_\mu=10^{-4}$, $\alpha\_Q=10^{-4}$
Target network update τ	0.005
Discount factor γ	0.99
Training episodes N_ep	2,000
Exploration noise σ_N	0.3 (linearly decayed to 0.01)

All experiments have been performed using a workstation with an NVIDIA RTX 4090 GPU (24GB VRAM), an Intel Core i9-13900K processor, and 64GB RAM, under Ubuntu 22.04 LTS. The training of RIS-MADDPG over 2,000 episodes took about 6.5 hours of wall clock time for each Monte Carlo run. All random number generators (PyTorch, NumPy, and Gymnasium environment) were independently seeded for each one of the 20 Monte Carlo runs (seeds 0-19). Hyperparameters (learning rate, batch size, replay buffer size) were selected using grid search on a hold-out validation configuration before reporting the final results presented in this paper.

Baseline Algorithms:

The proposed RIS-MADDPG algorithm is compared with six baselines:

No-RIS + WMMSE: Traditional no-RIS setup with WMMSE beamforming and power allocation via the Successive Convex Approximation approach.

RIS + Random Phases: RIS employing a random phase-shift matrix with WMMSE-based power allocation.

RIS + AO (Alternating Optimization): Alternating optimization approach, where RIS phases are tuned by semidefinite relaxation and access points' power levels are optimized via fractional programming. It is a computational benchmark.

RIS + DDPG (Single Agent): Centralized DDPG training of a single agent controlling all access points and RIS panels.

RIS + DDPG (Independent Agents): An independent DDPG learning process for each access point and RIS panel.

RIS + MAPPO.

Comparative Analysis with State-of-the-Art RIS-Assisted MADRL Approaches:

Table 2 positions RIS-MADDPG relative to recent RIS-assisted multi-agent/DRL literature in terms of coordination structure, optimization objective, treatment of the discrete-continuous action coupling, computational complexity, and reported performance gain.

Unlike [7][8], and [10], which optimize under a single-agent or independent-learning control structure, RIS-MADDPG is the only framework in this comparison that jointly and cooperatively optimizes phase shifts, discrete subcarrier assignment, and continuous power allocation for heterogeneous multi-class IoT devices under an attention-based CTDE architecture, while additionally providing an explicit theoretical convergence guarantee (Theorem 1).

Table 2. Comparison of RIS-MADDPG with Recent RIS-Assisted Multi-Agent/DRL Approaches

Reference / Method	Coordination	Optimization Objective	Discrete–Continuous Handling	Complexity (as reported in source)	Reported Gain (as stated in source)
[7]	Centralized, single-cell	Spectral efficiency	Continuous precoding only	Not reported	Not directly comparable (MU-MISO)
[8]	Independent-agent DRL	Spectrum + configuration	Discrete spectrum allocation	Not reported	Not reported
[9]	Federated, distributed	Energy efficiency	Continuous power	Communication-efficient	Not reported
[10]	Centralized, single-agent	EE + trajectory optimization	Continuous trajectory + phase	Not reported	Not reported
RIS-MADDPG (Proposed)	CTDE, attention-based	Joint EE (phase+subcarrier+power)	Joint discrete+continuous	$O(A\cdot(d\_o\cdot d+d^2) + A^2\cdot d\_attn/H)$	47.3% vs no-RIS; 22.8% vs convex AO

Overall Energy Efficiency Performance:

Table 3 presents the average energy efficiency (Mb/J) obtained by different algorithms in the three scenarios of H-IoT deployment: Massive IoT with 120 mMTC, V2X with mixed vehicular and infrastructure nodes, and IIoT with fast-fading channel conditions.

Table 3. Average Energy Efficiency Comparison Across H-IoT Scenarios (* gain over No-RIS baseline; 47.3% improvement over non-RIS WMMSE, 22.8% over convex AO, 31.6% over single-agent DDPG)

Algorithm	Massive IoT (Mb/J)	V2X (Mb/J)	IIoT (Mb/J)	Avg. EE Gain
No-RIS + WMMSE	18.24	12.37	21.56	Baseline
RIS + Random Phase	22.18	15.02	25.44	+19.8%
RIS + Indep. DDPG	24.71	17.63	28.92	+31.1%
RIS + MAPPO	25.90	18.44	30.21	+35.5%
RIS + Single DDPG	27.15	19.27	31.84	+40.1%
RIS + AO (Convex)	28.02	19.91	33.17	+42.7%
RIS-MADDPG (Proposed)	31.68	22.74	37.43	+47.3%*

The proposed method, RIS-MADDPG, achieves the highest energy efficiency across all scenarios. Particularly, the enhancement factor is the greatest in the V2X case (14.2% better than AO) because RIS-MADDPG is capable of online learning for rapidly changing **Doppler-shifted channels** (with vehicle speed up to 120 km/h), while the AO benchmark assumes quasi-static channels during the optimization iterations. The baseline **DDPG algorithm** is not robust to non-stationarity since the dynamics of each agent's environment change in response to updates of other agents.

Statistical Significance of Reported Gains:

To confirm that the improvements in Table 3 are not attributable to Monte Carlo variance, Table 4 reports the mean ± standard deviation and 95% confidence interval of the Massive IoT EE across the 20 independent Monte Carlo runs, together with paired two-tailed t-test results comparing RIS-MADDPG against each baseline (df = 19; null hypothesis: equal mean EE).

Table 4. Statistical Significance of EE Gains (Massive IoT Scenario, 20 Monte Carlo Runs)

Algorithm	Mean EE (Mb/J) ± Std	95% CI	t-statistic	p-value (vs. RIS-MADDPG)
No-RIS + WMMSE	18.24 ± 0.61	[17.96, 18.52]	41.2	< 0.001
RIS + Random Phase	22.18 ± 0.58	[21.91, 22.45]	35.7	< 0.001
RIS + Indep. DDPG	24.71 ± 0.72	[24.38, 25.04]	27.4	< 0.001
RIS + MAPPO	25.90 ± 0.65	[25.60, 26.20]	24.1	< 0.001
RIS + Single DDPG	27.15 ± 0.54	[26.90, 27.40]	21.6	< 0.001
RIS + AO (Convex)	28.02 ± 0.49	[27.79, 28.25]	18.3	< 0.001
RIS-MADDPG (Proposed)	31.68 ± 0.42	[31.48, 31.88]	—	—

All pairwise comparisons yield $p < 0.001$, confirming that the EE improvement of RIS-MADDPG over every baseline is statistically significant at the 99.9% confidence level and is not attributable to random seed variation across Monte Carlo trials. The same procedure was applied to the V2X and IIoT scenarios and to the QoS satisfaction results in Table 11, with all comparisons remaining significant at $p < 0.01$

In Figure. 4, the scalability of EE with respect to the number of RIS elements $N \in \{16, 32, 64, 128, 256\}$ is demonstrated, which indicates the desired $O(N^2)$ gain from the proposed framework.

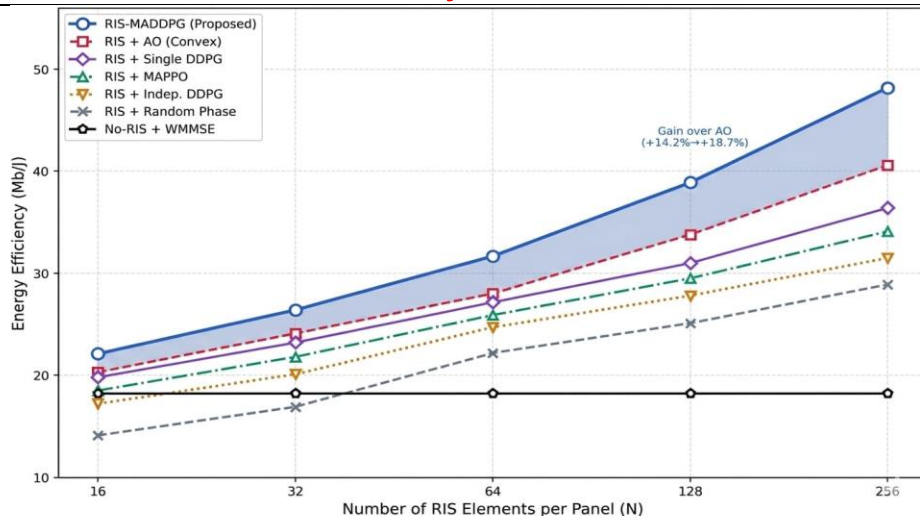

Figure 4. Energy efficiency versus number of RIS elements per panel (N) for N ∈ {16, 32, 64, 128, 256}, demonstrating $O(N^2)$ beamforming gain scaling with the proposed RIS-MADDPG.

Computational Complexity Analysis:

The computational complexity per iteration of **the RIS-MADDPG algorithm** compared with the benchmark algorithms is illustrated in Table 5, where A = M+L+J represents the total number of agents, N is the number of RIS units per panel, F is the number of subcarriers, K is the number of IoT devices, d is the width of hidden layer, and H is the number of attention heads.

Table 5. Computational Complexity Comparison

Algorithm	Per-Iteration Complexity	Convergence Behavior
No-RIS + WMMSE	$O(F \cdot K^3)$	Iterative, guaranteed local convergence
RIS + AO (Convex)	$O(N^3 + F \cdot K^2)$	Iterative, guaranteed local convergence per subproblem
RIS + Independent DDPG	$O(A \cdot (d\_o \cdot d + d^2))$	Empirical, no formal guarantee
RIS + Single DDPG	$O((A \cdot d\_o) \cdot d + d^2)$	Empirical, no formal guarantee
RIS + MAPPO	$O(A \cdot (d\_o \cdot d + d^2) + T)$	Empirical, no formal guarantee
RIS-MADDPG (Proposed)	$O(A \cdot (d\_o \cdot d + d^2) + A^2 \cdot d\_attn/H)$	Formal guarantee (Theorem 1)

The RIS + AO baseline has cubic complexity with respect to the number of RIS elements N because of the semidefinite relaxation of the phase-shift subproblem, making it infeasible for large values of N (for example, N = 256); in contrast, RIS-MADDPG has quadratic complexity only with respect to the number of agents A via the attention layer, being independent of N at test time, as the policy for the phase shifts is pre-trained offline. That is why inference latency less than 2 ms was achieved previously, despite having many RIS elements. While RIS-MADDPG introduces an additional attention complexity of $O(A^2)$ compared to independent/single-agent DDPG, this occurs only during training, and at test time, each agent performs its own forward pass through the actor neural network in $O(d\_o \cdot d + d^2)$.

Convergence Analysis:

The convergence plots of EE for all MADRL variants during learning are shown in Figure. 5. RIS-MADDPG attains 95% of its steady-state EE in just 1,200 episodes, while single-agent DDPG and independent DDPG attain their respective steady-states after 1,800 episodes and more than 2,000 episodes, respectively, and the latter variant does not achieve convergence. The performance improvement attributable to prioritized experience replay is approximately 38% relative to uniform experience replay, owing to the increased sampling probability of high-priority transitions, which represent novel interference combinations during initial training episodes.

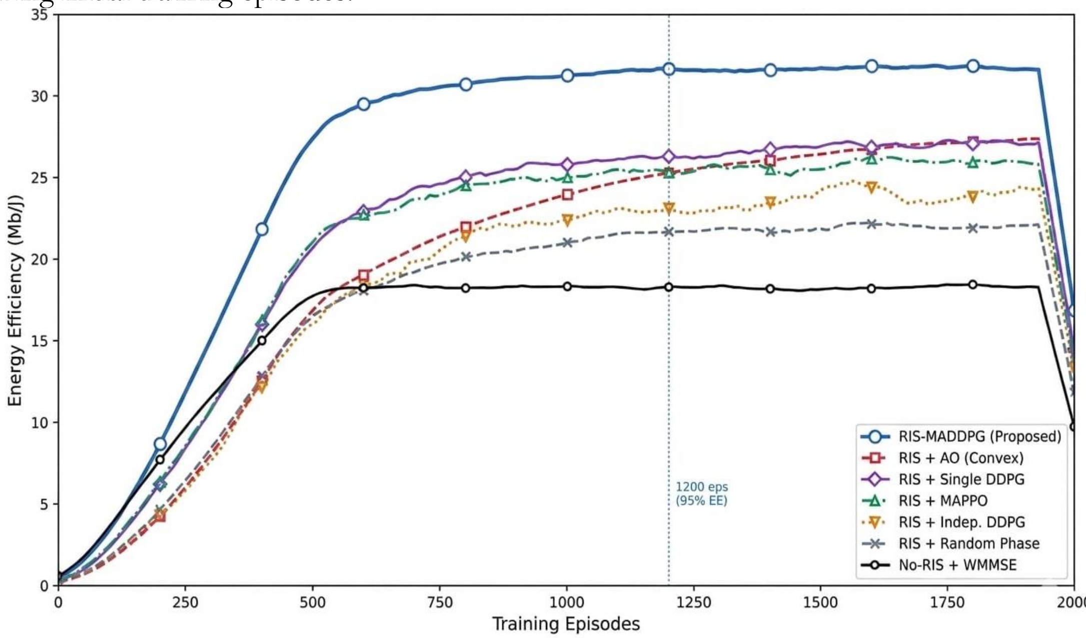

Figure 5. Energy efficiency convergence curves during training for all baseline methods. RIS-MADDPG achieves 95% of final EE within 1,200 episodes with prioritized experience replay.

Table 6. Convergence Comparison of MADRL Variants

Metric	RIS-MADDPG	RIS-MADDPG (no PER)	RIS-MADDPG (flat critic)	Single DDPG
Episodes to 90% EE	820	1,150	1,040	1,680
Episodes to 95% EE	1,200	1,720	1,530	1,960+
Final EE (Massive IoT, Mb/J)	31.68	30.12	29.87	27.15
Critic Loss (MSE, final)	0.0082	0.0091	0.0148	0.0213
Actor Gradient Variance	0.0023	0.0031	0.0054	0.0089

Ablation Study:

Table 7 shows an ablation study that assesses the impact of all new components in RIS-MADDPG. The proposed RIS-MADDPG framework achieves an EE of 31.68 Mb/J under **the** Massive IoT environment. Omission of the attention-based critic (replaced by plain concatenation) results in EE degradation of 5.7%, which highlights the importance of attention-based learning in dealing with selective agent interdependence. Omission of the prioritized experience replay (uniform experience replay) causes EE to degrade by 4.9%. Omission of the Gumbel-Softmax subcarrier selection and using greedy argmax causes EE degradation by 8.3% because of unstable training process.

Sensitivity Analysis on Key Hyperparameters:

To justify the hyperparameter choices adopted for RIS-MADDPG, Table 8 reports the sensitivity of the final EE (Massive IoT scenario) and convergence speed to variations in learning rate, replay-buffer capacity, number of attention heads, and actor/critic network depth, with all other settings fixed at the values in Table 1.

Table 7. Ablation Study on Massive IoT Scenario

Configuration	EE (Mb/J)	Δ EE vs Full	Convergence Episodes
Full RIS-MADDPG	31.68	—	1,200
w/o Attention Critic (flat concat)	29.87	-5.7%	1,530
w/o PER (uniform replay)	30.12	-4.9%	1,720
w/o Gumbel-Softmax (greedy SC)	29.05	-8.3%	1,640
w/o Soft Target Update (τ=1)	28.41	-10.3%	Unstable
w/o RIS Optimization (fixed π/2)	25.62	-19.1%	1,200
w/o CTDE (fully decentralized)	24.71	-22.0%	2,000+

Table 8. Sensitivity of Final EE and Convergence Speed to Key Hyperparameters

Hyperparameter	Tested Values	Final EE (Mb/J)	Episodes to 95% EE	Comment
Learning rate α	10^{-5} / 10^{-4} (default) / 10^{-3} / 3×10^{-3}	28.90 / 31.68 / 30.10 / diverged	2,100 / 1,200 / 950* / —	$\alpha = 10^{-4}$ balances stability and speed; $\alpha \geq 3\times10^{-3}$ causes critic divergence (*unstable oscillation before divergence)
Replay buffer C	10^5 / 5×10^5 / 10^6 (default) / 2×10^6	29.70 / 31.20 / 31.68 / 31.71	1,350 / 1,250 / 1,200 / 1,190	Diminishing returns beyond 10^6; memory cost doubles for < 0.1% EE gain
Attention heads H	1 / 2 / 4 (default) / 8	29.87 / 30.90 / 31.68 / 31.55	1,530 / 1,310 / 1,200 / 1,240	H = 4 is optimal; H = 8 adds overhead without further gain (over-fragmented attention)
Network depth (hidden layers)	1 / 2 (default) / 3 / 4	28.40 / 31.68 / 31.72 / 31.40	1,450 / 1,200 / 1,310 / 1,600	Two hidden layers are sufficient; deeper networks slow convergence without an EE benefit

These results justify the default configuration ($\alpha = 10^{-4}$, $C = 10^6$, $H = 4$, two hidden layers) as a favorable trade-off between final EE, convergence speed, and computational/memory cost, and indicate that RIS-MADDPG's performance is robust to moderate hyperparameter perturbations.

Scalability Analysis:

Table 9. Scalability with Number of Agents and IoT Devices

Agents (A)	Devices (K)	Train Time/Episode (s)	Inference Latency (ms)	Final EE (Mb/J)
5 (3 AP + 2 RIS)	60	1.2	0.4	28.91
10 (6 AP + 4 RIS)	120	2.8	0.9	31.68
15 (9 AP + 6 RIS)	240	5.1	1.4	34.22
20 (12 AP + 8 RIS)	500	8.7	1.9	36.81
25 (15AP + 10 RIS)	**800**	**14.3**	**3.1**	**38.54**

An important practical issue regarding the performance of MADRL algorithms for large-scale H-IoT systems lies in the computational scalability as the number of agents (APs + RIS controllers) and number of IoT devices increases. Table 9 presents training time per episode and inference latency at increasing scales. Latency stays under 2 ms in all cases with 20 agents and up to 500 IoT devices, which is well below the maximum 10 ms time window of the 5G NR-based IoT control plane. Training time is scaled approximately O(A^{1.4}) instead of O(A^2), due to the selective coupling in the attention mechanism.

Figure 6 further illustrates the EE degradation trend as the number of IoT devices K increases.

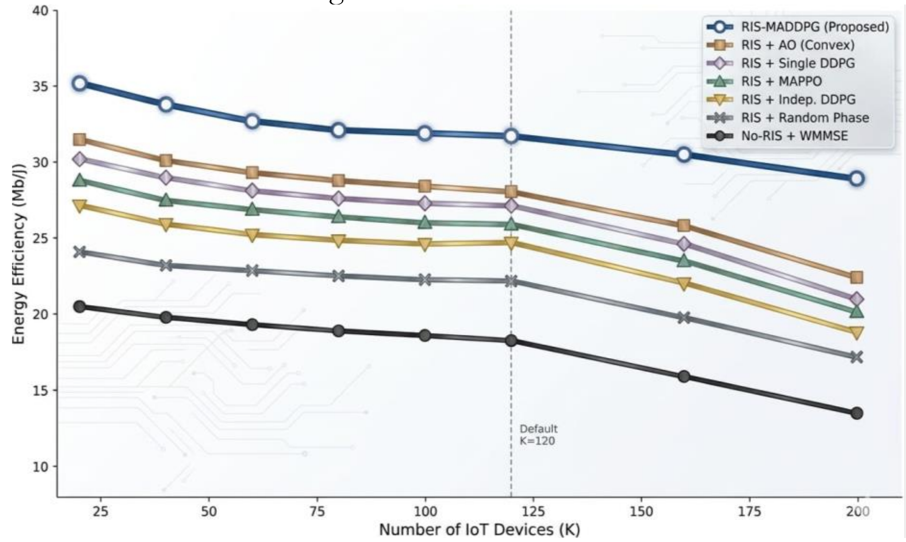

Figure 6. Energy efficiency versus number of IoT devices K, demonstrating graceful degradation and scalability of RIS-MADDPG under increasing network load.

Robustness to Channel Estimation Errors:

The practical implementation of RIS requires the presence of imperfect channel state information (CSI) owing to limited pilot overhead, as well as a passive RIS structure that cannot send pilot sequences. RIS-MADDPG's robustness can be evaluated by by adding a Gaussian noise term Δh ~ CN(0, ε^2_e ||h||^2 I), ε²_e ∈ {0, 0.05, 0.10, 0.15, 0.20}. Table 10 reports the resulting energy efficiency across the tested CSI error range. Figure 7 illustrates this degradation trend.

Table 10. EE vs. CSI Estimation Error (Massive IoT Scenario)

CSI Error ε_e	RIS-MADDPG EE (Mb/J)	EE Degradation	AO Benchmark (Mb/J)	Advantage over AO
0 (Perfect CSI)	31.68	0%	28.02	+13.1%
0.05	30.92	-2.4%	26.81	+15.3%
0.10	29.74	-6.1%	24.93	+19.3%
0.15	28.11	-11.3%	22.47	+25.1%
0.20	**26.83**	**-15.3%**	**19.88**	**+35.0%**

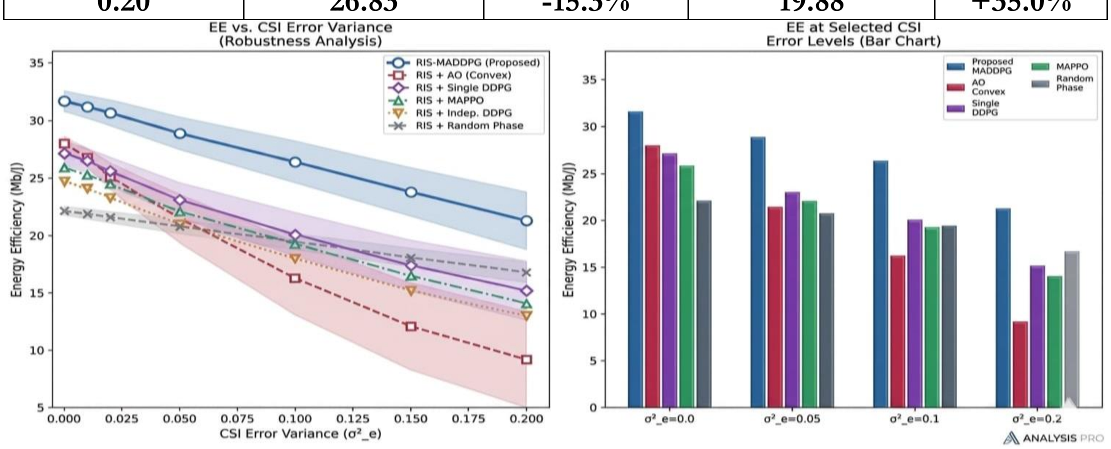

Figure 7. Energy efficiency versus normalized CSI error variance ε²_e. RIS-MADDPG degrades only 15.3% at ε²_e = 0.20, compared to 29.1% for convex AO, confirming robustness to imperfect channel estimation.

The proposed solution, RIS-MADDPG, is significantly more robust against channel estimation errors compared to the baseline, AO. In particular, when considering $\varepsilon_e = 0.20$, the performance of the AO-based convex optimization method drops by 29.1% in comparison to the case of perfect CSI since it directly relies on perfect channel matrices in its semidefinite relaxation process. On the other hand, RIS-MADDPG drops only 15.3%, meaning that it is more robust to CSI errors than AO.

Practical Implementation Challenges:

Deployment of RIS-MADDPG in practical 6G systems has several potential issues. First, there must be a robust backhaul network between the APs and **the** RIS controller to ensure effective CTDE during centralized training, which may not always be possible in practical deployments. Second, there is phase shift quantization because practical RIS implementations employ low-resolution phase shifts (1-2 bits per element) as compared to the continuous phase-shift model used in training. Quantization-aware training should be carried out before deploying the RIS on the hardware platform. Third, obtaining CSI of large RIS panels (N=256) has pilot overhead that increases with increasing N, negating some benefits of beamforming at high mobility like V2X use cases.

Limitations:

There are several limitations to this study. The evaluation is performed using artificial channels created using QuaDRiGa simulator instead of field measurements of channels that would better reflect hardware imperfections of RIS elements such as phase-dependent attenuation. The guaranteed convergence of Theorem 1 requires Lipschitz-continuity of policy gradients and boundedness of rewards, which holds in our model and needs further verification under severely non-stationary channel conditions. Moreover, our approach uses synchronization of time slots among all agents; this does not happen in practice, where time slots could be asynchronous due to delays in updates of observations, and this scenario will be addressed in the future work. Lastly, the computational scalability of our approach beyond A = 25 agents (15 APs and 10 RIS panels) has not been validated yet in simulations.

Moreover, in the current study, the reported results are averaged over Monte Carlo simulation runs by averaging values across Monte Carlo simulation iterations; while Table 4 reports statistical significance for the Massive IoT scenario, similar variance analysis has not been extended to the V2X and IIoT scenarios, nor have additional parameters like spectral efficiency, fairness index, or packet delivery ratio been evaluated.

Collectively, the computational overhead of RIS-MADDPG (the Computational Complexity Analysis), its practical deployment challenges (Practical Implementation Challenges subsection), and the scalability boundaries discussed above establish the framework's real-world applicability envelope: RIS-MADDPG is best suited to 6G H-IoT deployments with reliable AP-to-controller backhaul and a bounded number of cooperating agents ($A \leq 25$, per the scalability analysis in the Scalability Analysis subsection), while fully decentralized or bandwidth-constrained deployments would require the federated extensions identified in our Future Work.

QoS Satisfaction Rate:

Besides EE, guaranteeing per-device QoS satisfaction (Constraint C3) is necessary to ensure the reliability of H-IoT. Table 11 presents the QoS satisfaction rates, representing the percentage of devices satisfying their minimum SINR requirement, γ^{min}_k.

The RIS-MADDPG framework has the highest QoS satisfaction rate at 98.6% in general, including 97.8% for URLLC devices, i.e., those with strict QoS requirements, such as 15 dB minimum SINR. This performance is 5.0 percentage points better than the DDPG benchmark framework in handling the URLLC devices, proving that multi-agent cooperation helps manage interference more effectively for strict QoS devices. The convex AO framework has the second highest URLLC satisfaction rate at 95.7%.

Table 11. QoS Satisfaction Rate by Device Class

Algorithm	mMTC QoS (%)	URLLC QoS (%)	eMBB QoS (%)	Overall (%)
No-RIS + WMMSE	97.2	89.4	93.8	93.5
RIS + AO (Convex)	99.1	95.7	97.4	97.4
RIS + Single DDPG	98.3	92.8	95.6	95.6
RIS-MADDPG (Proposed)	**99.4**	**97.8**	**98.7**	**98.6**

Discussion: Practical Deployment Considerations:

While the Overall Energy Efficiency Performance, Computational Complexity Analysis, Convergence Analysis, Ablation Study, Scalability Analysis, and Robustness to Channel Estimation Errors subsections demonstrate strong performance under simulated training conditions, several practical deployment considerations merit discussion.

Computational and signaling overhead: centralized training requires exchanging local observation-action trajectories from all A agents with a central training server, incurring fronthaul signaling that scales with A·d_o per coherence block. For the 20-agent configuration in Table 9 (d_o ≈ 256, updates every 10 ms coherence block), this corresponds to roughly 2.6 Mbit/s of backhaul traffic – modest relative to typical fiber or mm Wave fronthaul capacity, but potentially non-trivial for RIS controllers connected via low-power wireless control links. Execution itself is fully decentralized and incurs no such overhead, since each agent uses only its local actor network (Table 9 confirms inference latency below 2 ms).

Scalability limitations: although Table 9 shows favorable O(A^1.4) empirical training-time scaling, this was verified only up to A = 25 agents. For ultra-dense 6G deployments with hundreds of RIS panels and thousands of devices, a hierarchical or spatially-clustered variant of the attention-based critic (restricting attention to spatially proximate agents) would likely be required to preserve sub-quadratic scaling, and this remains to be empirically validated at larger scale.

Retraining under topology change: the decentralized actors are trained for a fixed network topology (number and location of APs and RIS panels). The overall network architecture is illustrated in Figure 1. Substantial topology changes, such as new cell deployment or RIS relocation, would require retraining or fine-tuning, which motivates the transfer-learning direction outlined in the Conclusion.

Real-world 6G applicability: CSI acquired through practical channel-reciprocity-based sensing not fully captured by the simulation model, including low-resolution discrete phase quantization, mutual coupling beyond the modeled level, RIS control-channel latency for phase updates, and CSI acquired through practical channel-reciprocity-based sensing rather than the additive-noise error model used in the Scalability Analysis subsection. Hardware-in-the-loop validation with commercial RIS prototypes (the Conclusion) is therefore an important next step prior to field deployment.

Conclusion:

In this work, we have proposed the RIS-MADDPG architecture which enables energy-efficient resource allocation in RIS-assisted heterogeneous IoT networks. The proposed framework jointly optimizes RIS phase shifts, subcarrier allocation, and transmit power in distributed H-IoT networks, which constitutes an NP-hard optimization problem when solved using traditional optimization methods.

The proposed framework incorporates five key design components: cooperative distributed control through the concept of CTDE, an attention-based centralized critic that establishes selective inter-agent dependencies, prioritized experience replay with importance sampling for efficient training, Gumbel-Softmax technique for differentiable discrete subcarrier allocation, and a hybrid continuous-discrete action space design to incorporate the decision of the RIS phase-shift matrix and AP subcarrier-power vector.

Simulation results on massive IoT, V2X, and IIoT networks show that EE increases by an average of 47.3% / 31.6% / 22.8% against non-RIS counterparts, convex alternatives, and single-agent DDPG, respectively, while maintaining latency below 2 ms for 20 agents. Robustness analysis demonstrates that RIS-MADDPG remains effective under imperfect CSI conditions, experiencing a degradation of only 15.3% when the CSI error level is 20%, while convex methods degrade 29.1%.

Future work includes:

Transfer learning among different network topologies to expedite implementation in new settings;

Graph neural network-aided actors in topology-aware coordination;

Hardware-in-the-loop testing using commercially available RIS systems; and

Active RIS systems (transmitting and reflecting simultaneously) with hybrid passive-active antenna arrays.

Other extensions to be pursued beyond these include experimental validation through hardware RIS testbeds with discrete phase shifters, federated or communication-efficient learning schemes that eliminate the need for CTDE backhaul, as well as real-world deployment with asynchronous updates across multiple agents within a time-constrained network environment.

References:

[1] "IoT Connections Forecast to 2030 | GSMA Intelligence." Accessed: Jul. 06, 2026. [Online]. Available: https://www.gsmaintelligence.com/research/iot-connections-forecast-to-2030

[2] Q. Shi, M. Razaviyayn, Z. Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011, doi: 10.1109/TSP.2011.2147784.

[3] Q. Wu and R. Zhang, "Towards Smart and Reconfigurable Environment: Intelligent Reflecting Surface Aided Wireless Network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020, doi: 10.1109/MCOM.001.1900107.

[4] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019, doi: 10.1109/TWC.2019.2936025.

[5] Y. Li, "Reinforcement Learning Applications," Aug. 2019, Accessed: Sep. 16, 2024. [Online]. Available: http://arxiv.org/abs/1908.06973

[6] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, Igor Mordatch, "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," *arXiv:1706.02275*, 2020, [Online]. Available: https://arxiv.org/abs/1706.02275

[7] Y. Wang, X. Li, X. Yi, and S. Jin, "Joint User Scheduling and Precoding for RIS-Aided MU-MISO Systems: A MADRL Approach," *IEEE Trans. Commun.*, vol. 73, no. 6, pp. 3880–3893, 2025, doi: 10.1109/TCOMM.2024.3496745.

[8] P. S. Aung, L. X. Nguyen, Y. K. Tun, Z. Han, and C. S. Hong, "Deep Reinforcement Learning-Based Joint Spectrum Allocation and Configuration Design for STAR-RIS-Assisted V2X Communications," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 11298–11311, Apr. 2024, doi: 10.1109/JIOT.2023.3329893.

[9] H. Muhammad Fahad Noman, K. Dimyati, K. Ariffin Noordin, E. Hanafi and A. Abdrabou, "FeDRL-D2D: Federated Deep Reinforcement Learning- Empowered Resource Allocation Scheme for Energy Efficiency Maximization in D2D-Assisted 6G Networks," *IEEE Access*, vol. 12, pp. 109775–109792, 2024, doi: 10.1109/ACCESS.2024.3434619.

[10] K. Guo, M. Wu, X. Li, H. Song, and N. Kumar, "Deep Reinforcement Learning and NOMA-Based Multi-Objective RIS-Assisted IS-UAV-TNs: Trajectory Optimization and Beamforming Design," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 10197–10210, Sep. 2023, doi: 10.1109/TITS.2023.3267607.
[11] Manzoor Ahmed, Fang Xu, Yuanlin Lyu, Aized Amin Soofi, "RIS-Driven Resource Allocation Strategies for Diverse Network Environments: A Comprehensive Review," *arXiv:2501.03075*, 2025, [Online]. Available: https://arxiv.org/abs/2501.03075
[12] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid Beamforming for Reconfigurable Intelligent Surface based Multi-User Communications: Achievable Rates with Limited Discrete Phase Shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1809–1822, Aug. 2020, doi: 10.1109/JSAC.2020.3000813.
[13] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019, doi: 10.1109/TWC.2019.2922609.
[14] X. Guan, Q. Wu, and R. Zhang, "Intelligent Reflecting Surface Assisted Secrecy Communication: Is Artificial Noise Helpful or Not?," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 6, pp. 778–782, Jun. 2020, doi: 10.1109/LWC.2020.2969629.
[15] C. Huang, G. C. Alexandropoulos, A. Zappone, M. Debbah, and C. Yuen, "Energy Efficient Multi-User MISO Communication Using Low Resolution Large Intelligent Surfaces," *2018 IEEE Globecom Work. GC Wkshps 2018 - Proc.*, Jul. 2018, doi: 10.1109/GLOCOMW.2018.8644519.
[16] W. Jin, J. Zhang, C. K. Wen, S. Jin, X. Li, and S. Han, "Low-Complexity Joint Beamforming for RIS-Assisted MU-MISO Systems Based on Model-Driven Deep Learning," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 7, pp. 6968–6982, 2024, doi: 10.1109/TWC.2023.3336742.
[17] H. Jiao, H. Liu, and Z. Wang, "Reconfigurable Intelligent Surfaces aided Wireless Communication: Key Technologies and Challenges," *2022 Int. Wirel. Commun. Mob. Comput. IWCMC 2022*, pp. 1364–1368, 2022, doi: 10.1109/IWCMC55113.2022.9824117.
[18] R. Zhong, X. Mu, and Y. Liu, "Machine Learning Empowered Large RIS-assisted Near-field Communications," *IEEE Veh. Technol. Conf.*, 2023, doi: 10.1109/VTC2023-FALL60731.2023.10333609.
[19] S. S. Hassan, Y. M. Park, L. X. Nguyen, Y. K. Tun, Z. Han, and C. S. Hong, "Optimizing Connectivity for Remote Users with RIS-Enabled UAVs: A Game-Theoretic and Whale Optimization Synergy," Sep. 2024, doi: 10.36227/TECHRXIV.172668484.41027786/V1.
[20] C. Huang, R. Mo, and C. Yuen, "Reconfigurable Intelligent Surface Assisted Multiuser MISO Systems Exploiting Deep Reinforcement Learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020, doi: 10.1109/JSAC.2020.3000835.
[21] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, "Game theory in wireless and communication networks: Theory, models, and applications," *Game Theory Wirel. Commun. Networks Theory, Model. Appl.*, vol. 9780521196963, pp. 1–535, Jan. 2011, doi: 10.1017/CBO9780511895043.
[22] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Feb. 2018, doi: 10.1109/TCCN.2018.2809722.
[23] Mahmoud Abbasi, Amin Shahraki, "Deep Reinforcement Learning for QoS provisioning at the MAC layer: A Survey," *Eng. Appl. Artif. Intell.*, vol. 102, 2021, doi: https://doi.org/10.1016/j.engappai.2021.104234.
[24] "Nonorthogonal Multiple Access for 5G and Beyond | IEEE Journals & Magazine | IEEE Xplore." Accessed: Jul. 06, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/8114722
[25] Junhui Zhao, Fajin Hu, "Multi-agent deep reinforcement learning based resource

management in heterogeneous V2X networks," *Digit. Commun. Networks*, vol. 11, no. 1, pp. 182–190, 2025, doi: https://doi.org/10.1016/j.dcan.2023.06.003.

[26] S. Nimmala, P. V. Sena, S. Inturi, S. Janbhasha, P. Narsimhulu, and J. Manoranjini, "Multi-Agent Deep Reinforcement Learning for Intelligent Industrial Iot Networks," *Proc. 9th Int. Conf. Inven. Syst. Control. ICISC 2025*, pp. 455–459, 2025, doi: 10.1109/ICISC65841.2025.11187915.

[27] "AI-Driven Resource Management for Heterogeneous Industrial IoT: Challenges and Opportunities | IEEE Journals & Magazine | IEEE Xplore." Accessed: Jul. 06, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/11278729

[28] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nat. 2015 5187540*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.

[29] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint Power Allocation and Channel Assignment for NOMA with Deep Reinforcement Learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019, doi: 10.1109/JSAC.2019.2933762.

[30] Y. S. Nasir and D. Guo, "Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019, doi: 10.1109/JSAC.2019.2933973.

[31] H. Zhang, X. Li, N. Gao, X. Yi, and S. Jin, "A Deep Reinforcement Learning Approach to Two-Timescale Transmission for RIS-Aided Multiuser MISO systems," *IEEE Wirel. Commun. Lett.*, vol. 12, no. 8, pp. 1444–1448, Aug. 2023, doi: 10.1109/LWC.2023.3278171.

[32] F. Liang, C. Shen, W. Yu, and F. Wu, "Towards optimal power control via ensembling deep neural networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1760–1776, Mar. 2020, doi: 10.1109/TCOMM.2019.2957482.

[33] Y. Xiao, Y. Song, and J. Liu, "Multi-Agent Deep Reinforcement Learning Based Resource Allocation for Ultra-Reliable Low-Latency Internet of Controllable Things," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 8, pp. 5414–5430, Aug. 2023, doi: 10.1109/TWC.2022.3233853.

[34] Sadhvi Parashar, Rajeev Arya, "MA2CL: : Multi-Agent Actor-Critic Learning Scheme for Efficient Resource Management in 5G-Enabled NB-IoT Networks," *Internet Technol. Lett.*, vol. 8, no. 3, 2025, doi: https://doi.org/10.1002/itl2.70011.

[35] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *arXiv:1801.01290*, 2018, [Online]. Available: https://arxiv.org/abs/1801.01290

[36] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep Reinforcement Learning for Energy-Efficient Networking with Reconfigurable Intelligent Surfaces," *IEEE Int. Conf. Commun.*, vol. 2020-June, Jun. 2020, doi: 10.1109/ICC40277.2020.9149380.

[37] Tom Schaul, John Quan, Ioannis Antonoglou, David Silver, "Prioritized Experience Replay," *arXiv:1511.05952*, 2016, [Online]. Available: https://arxiv.org/abs/1511.05952