Worst-case MSE Minimization for RIS-assisted mmWave MU-MISO Systems with Hardware Impairments and Imperfect CSI

Shao-Heng Chen[1], Hsin-Yuan Chang[1], Chih-Yu Wang[2], Ren-Hung Hwang[3], and Wei-Ho Chung[1, 2]
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[3]College of Artificial Intelligence, National Yang Ming Chiao Tung University, Taiwan
E-mail: paulchen.2713@gmail.com, hyuan.chang@outlook.com, cywang@citi.sinica.edu.tw, rhhwang@nycu.edu.tw, whchung@ee.nthu.edu.tw

2025 IEEE Wireless Communications and Networking Conference (WCNC) | 979-8-3503-6836-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/WCNC61545.2025.10978171

***Abstract*—Robustness of reconfigurable intelligent surface (RIS) has been a concern due to potential hardware impairments (HWI) and imperfect channel state information (CSI) measurements caused by the numerous passive elements on board. Recent studies observe that the impairments not only introduce misalignment in phase adjustments but also affect the amplitude of reflected signals, which further complicates the issue. To address this issue, we introduce a novel deep reinforcement learning (DRL)-based discrete optimization framework aimed at mitigating various HWI and CSI imperfections in RIS-assisted millimeter-wave (mmWave) multi-input-single-output (MU-MISO) systems. Employing proximal policy optimization (PPO), our method discretely addresses HWI and CSI challenges without continuous relaxation. Simulation results demonstrate the superiority of our approach over the traditional optimal beamforming baseline in minimizing the worst-case mean squared error (MSE) of the signal received by the users. The code has been made open-source on GitHub, serving as a valuable reference for further research and application in RIS-assisted communication systems.**

***Index Terms*—Deep reinforcement learning (DRL), hardware impairment (HWI), reconfigurable intelligent surface (RIS), imperfect channel state information (CSI), Von-Mises phase error.**

I. Introduction

Millimeter-wave (mmWave) communication has recently gained attention to meet the demands for extremely high data rates, ultra-reliability, and ultra-low latency. However, the unique characteristics of mmWave transmission, such as channel sparsity and substantial path and penetration losses, pose challenges for transceiver design [1]. Reconfigurable intelligent surface (RIS) is an emerging technology which composed of numerous low-cost, low-power, passive reflecting elements that can intelligently manipulate the radio propagation of the incident signal solely by adjusting its phase shifts. The channel manipulation capability offered by RIS can significantly improve the coverage and quality of mmWave communications in a low-cost and energy-efficient manner.

Existing studies have explored the joint design of RIS and transceivers to achieve various objectives, including maximizing the received signal-to-noise ratio (SNR), minimizing the bit error rate (BER) and optimizing the transmit power [2]–[4]. In [2], a closed-form solution, accompanied by performance analysis, was proposed for optimal active and passive beamforming to maximize the spectral efficiency (SE). In [3], a geometric mean decomposition (GMD)-based approach was presented for RIS and transceiver design aimed at minimizing the BER. In [4], semi-definite relaxation (SDR) and alternating optimization (AO) techniques were employed to optimize the beamforming vectors with the goal of minimizing transmit power. These conventional iterative algorithms struggle with high complexity and computation time, limiting their practicality. They also rely on impractical assumptions, neglecting phase shift errors [5], [6] introduced by unexpected channel interference [7] during transmission, and hardware impairment (HWI) [8] on the RIS or the transceiver, thereby reducing their applicability in real-world scenarios. Plus, considering the phase-dependent amplitude response, which is influenced by the phase shifts of the elements, and the presence of inevitable phase errors, the optimal RIS configuration should be determined by jointly addressing these factors. Moreover, the optimization problem is intricately linked with the mmWave channel, which is affected by the CSI uncertainties in channel estimation, necessitating a comprehensive consideration of HWI and CSI uncertainties in the design.

Incorporating complex and practical settings, deep reinforcement learning (DRL) has emerged as a popular and powerful alternative [9]–[11]. In [9], a deep deterministic policy gradient (DDPG)-based secure beamforming algorithm was proposed to maximize the sum secrecy rate, taking into account the impact of phase errors in both the transceiver and the RIS. In [10],

This work was supported by the National Science and Technology Council under Grants 111-2628-E-001 -002 -MY3, 113-2221-E-007 -136 -MY3, 113-2218-E-194-004-, and by the Academia Sinica under Thematic Research Grant AS-TP-110-M07-3 and Visible Project at the Research Center for Information Technology Innovation.

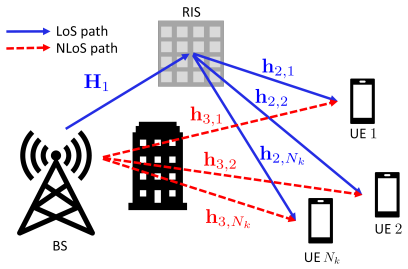

Fig. 1. A downlink RIS-assisted mmWave MU-MISO system.

a deep Q-network (DQN)-based RIS configuration algorithm, utilizing only 1-bit resolution, was introduced to maximize the SE of a downlink orthogonal frequency division multiplexing (OFDM) wireless system. In [11], a soft actor-critic (SAC)-based optimization algorithm was proposed to maximize the sum downlink rate while separately considering the phase-dependent amplitude model [8] and an imperfect CSI scenario.

In this study, we consider the collective influence of HWI, encompassing both phase-dependent amplitude response and phase shift errors, along with CSI imperfections in a downlink RIS-assisted multi-user, multi-input-single-output (MU-MISO) system. Subsequently, we introduce a DRL-based optimization framework aimed at minimizing the worst-case mean squared error (MSE) of the received signals. Our contributions are listed as follows:

- To the best of our knowledge, this is the first work to address the worst-case MSE of the received signals in multi-user RIS system considering the phase-dependent amplitude response and CSI imperfections. Our simulation results suggested that ignoring such effects will lead to robustness degradation in terms of worst-case MSE.
- There has been no study in the literature applying DRL to RIS systems with HWI and CSI imperfections. Our results suggest that DRL is capable of addressing the MSE minimization problem even in the presence of complex HWI and CSI effects.
- To further encourage the advance of related studies, we shared the developed model in open-source form on https://github.com/paulchen2713/RIS-MISO-HWI-DRL.

Notations: $\mathcal{CN}(\mu, \sigma^2)$ stands for a complex Gaussian with mean μ and variance σ^2. $\mathbb{E}[\cdot]$ is the mathematical expectation. $(\cdot)^T$, $(\cdot)^H$ denote transpose and conjugate transpose operators, respectively. $\mathrm{tr}\{\cdot\}$, $\|\cdot\|_2$ and $\|\cdot\|_\mathrm{F}$ stand for the the trace, l_2-norm and Frobenius norm of a matrix, respectively.

II. System Model and Problem Formulation

As depicted in Fig. 1, we consider a RIS-assisted mmWave MU-MISO system that comprises a base station (BS) with N_t antennas, N_k single-antenna user equipments (UE), and a RIS featuring N_s elements to improve the downlink data transmission. To entirely leverage channel diversity, a fully digital precoder $\mathbf{F} \in \mathbb{C}^{N_t \times N_k}$ is employed at the BS to map N_k data streams onto N_t antennas, enabling simultaneous transmission from the BS to the UE, where $N_k \leq N_t$ to satisfy sufficient degrees of freedom. Each user receives signals from two paths: *the direct path*, where signals propagate through obstacles between the BS and the UE, and *the reflected path*, wherein signals transmitted from the BS are combined and re-scattered via the RIS. Let $\mathbf{H}_1 \in \mathbb{C}^{N_s \times N_t}$, $\mathbf{\Phi} \triangleq \mathrm{Diag}\{\phi_1, ..., \phi_{N_s}\} \in \mathbb{C}^{N_s \times N_s}$, $\mathbf{h}_{2,k} \in \mathbb{C}^{N_s \times 1}$, and $\mathbf{h}_{3,k} \in \mathbb{C}^{N_t \times 1}$ denote the BS-RIS channel, the diagonal phase shift matrix of the RIS, the RIS-kth user channel and the BS-kth user channel, respectively. $\mathbf{h}_{n,k}$ represents the nth channel for the kth user. $\mathrm{Diag}\{\mathbf{h}\}$ represents a diagonal matrix with elements of $\mathbf{h}$ on the main diagonal. The signal received at the kth user can thus be expressed as:

$$y_k = (\underbrace{\mathbf{h}_{2,k}^T \mathbf{\Phi} \mathbf{H}_1}_{\text{reflected path}} + \underbrace{\mathbf{h}_{3,k}^T}_{\text{direct path}})\mathbf{F}\mathbf{x} + n_k, \tag{1}$$

where $\mathbf{x} = [x_1, \ldots, x_{N_k}]^T \in \mathbb{C}^{N_k \times 1}$ and $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ are the transmitted data stream and the received additive white Gaussian noise (AWGN), respectively. We assume the data stream is mutually independent and possess autocorrelations equal to 1, satisfying $\mathbb{E}[x_m x_n^H] = 0, \forall m \neq n$, and $\mathbb{E}[x_k x_k^H] = 1, \forall k = 1, \ldots, N_k$, respectively.

A. Channel Model with Imperfect CSI

Due to its short wavelength, mmWave faces limitations in diffracting around obstacles, leading to channels that exhibit a sparse multipath structure. These mmWave channels are typically characterized by the Saleh-Valenzuela (SV) channel model [3], [12]. Assuming the BS utilizes a uniform linear array (ULA) structure and the RIS adopts a uniform planar array (UPA) structure, the mmWave channels, which comprise a single line-of-sight (LoS) path and L non-line-of-sight (NLoS) paths, are modeled in the following manner:

$$\mathbf{H}_1 = \sqrt{\frac{N_t N_s}{L+1}} \sum_{l=0}^{L} \nu_{1,l}\ \mathbf{a}_{\mathrm{RIS}}(\gamma_l^r, \eta_l^r)\mathbf{a}_{\mathrm{BS}}(\theta_l^t)^H, \tag{2}$$

$$\mathbf{h}_{2,k} = \sqrt{\frac{N_s}{L+1}} \sum_{l=0}^{L} \nu_{2,l}\ \mathbf{a}_{\mathrm{RIS}}(\gamma_l^t, \eta_l^t)^H + \Delta\mathbf{h}_{2,k}, \tag{3}$$

$$\mathbf{h}_{3,k} = \sqrt{\frac{N_t}{L}} \sum_{l=1}^{L} \nu_{3,l}\ \mathbf{a}_{\mathrm{BS}}(\theta_l^t)^H + \Delta\mathbf{h}_{3,k}, \tag{4}$$

where $\nu_{1,l}$, $\nu_{2,l}$ and $\nu_{3,l}$ are the complex channel gains. For $l = 0$, this represents the LoS component modeled by $\mathcal{CN}(0, 0.1)$, while for $l = 1, \ldots, L$, these indicate the NLoS components modeled by $\mathcal{CN}(0, 10^{-0.1\mu})$ with Rician factor μ. The BS-kth user channel $\mathbf{h}_{3,k}$ consists of only NLoS paths, as the LoS is blocked by obstructions, as shown in Fig. 1. $\Delta\mathbf{h}_{2,k} \sim \mathcal{CN}(0, \psi)$ and $\Delta\mathbf{h}_{3,k} \sim \mathcal{CN}(0, \psi)$ represent the CSI uncertainties [7], [13] associated with the RIS-UE channel and BS-UE channel, respectively. $\mathbf{a}_{\mathrm{BS}}(\cdot)$ and $\mathbf{a}_{\mathrm{RIS}}(\cdot)$ denote the normalized steering vectors at the BS and RIS, respectively. θ_l^t denotes the angles of departure (AoD) at the BS for the lth path, and γ_l^r (η_l^t) and η_l^r (γ_l^t) represent the azimuth (elevation) angles of arrival (AoA) and AoD at the RIS in a similar

manner, respectively. The formulation of the steering vectors, accommodating a ULA configuration with N antennas at the BS, and a UPA setup at the RIS consisting of $M = M_y \times M_z$ elements, can be expressed as follows:

$$\mathbf{a}_{\mathrm{BS}}(\theta) = \frac{1}{\sqrt{N}}\Big[1, e^{j\frac{2\pi d_a}{\lambda}\sin(\theta)}, \ldots, e^{j\frac{2\pi d_a}{\lambda}(N-1)\sin(\theta)}\Big]^T, \tag{5}$$

$$\mathbf{a}_{\mathrm{RIS}}(\gamma,\eta) = \frac{1}{\sqrt{M}}\Big[1, \ldots, e^{j\frac{2\pi d_{\mathrm{RIS}}}{\lambda}(m_y\sin(\gamma)\cos(\eta)+m_z\sin(\eta))}, \ldots, e^{j\frac{2\pi d_{\mathrm{RIS}}}{\lambda}((M_y-1)\sin(\gamma)\cos(\eta)+(M_z-1)\sin(\eta))}\Big]^T, \tag{6}$$

where d_a and d_{RIS} specify the spacing between antennas at the BS and between elements at the RIS, respectively, each set to half the wavelength, i.e., $\frac{\lambda}{2}$. The angle parameters θ, γ and η are then substituted by the previously defined variables to characterize the directional components of the channel response in Eqs. (2)–(4). The values of AoD and AoA are randomly generated from the range of 0 to 2π.

B. Hardware Impairments

Considering a practical RIS-assisted mmWave system that incorporates phase-dependent amplitude variations [8], phase shift errors [5], [6], and limited resolution, the entries of the passive beamforming matrix are described as follows:

$$\mathbf{\Phi}(i,i) = \phi_i = \beta(\varphi_i)\cdot e^{j\varphi_i},\ \forall i = 1, ..., N_s \tag{7}$$

$$\beta(\varphi_i) = (1-\beta_{\min})\cdot\Big(\frac{\sin(\varphi_i-\mu_{\mathrm{PDA}})+1}{2}\Big)^{\kappa_{\mathrm{PDA}}} + \beta_{\min}, \tag{8}$$

where $\beta_{\min} \in [0,1]$, $\mu_{\mathrm{PDA}} \geq 0$, and $\kappa_{\mathrm{PDA}} \geq 0$ denote the constants that characterize the HWI effects of the RIS. The actual phase shift of the ith element is modeled as $\varphi_i = \hat{\varphi}_i + \Delta\varphi_i$, where $\Delta\varphi_i$ is the phase error, which follows a Von-Mises distribution with mean μ_{PE} and concentration factor κ_{PE} [5], [6]. The probability density function (PDF) is given by $f(\Delta\varphi_i \mid \mu_{\mathrm{PE}}, \kappa_{\mathrm{PE}}) = \frac{\exp(\kappa_{\mathrm{PE}}\cos(\Delta\varphi_i-\mu_{\mathrm{PE}}))}{2\pi I_0(\kappa_{\mathrm{PE}})}$, where $I_0(\kappa)$ is the modified Bessel function of the first kind of order 0. Our desired phase shifts are constrained to a discrete set of 2^B values, i.e., $\hat{\varphi}_i \in \mathcal{F} \triangleq \{-\frac{2\pi(2^{B-1}-1)}{2^B}, \ldots, 0, \ldots, \frac{2\pi(2^{B-1})}{2^B}\}$, for $i = 1, \ldots, N_s$. B is the bit resolution in the RIS setting.

C. Problem Formulation

This study focuses on jointly optimizing the downlink precoder and the discrete configuration of the RIS. Our goal is to minimize the worst-case MSE of the received signal under HWI and CSI uncertainties. This objective gives rise to a Min-Max MSE optimization problem, which is formulated as follows:

$$\min_{\mathbf{\Phi},\mathbf{F}} \max_{k=1,\ldots,N_k} \quad \mathrm{MSE}_k \tag{9a}$$

$$\text{s.t.} \quad \mathrm{tr}\{\mathbf{F}\mathbf{F}^H\} \leq P_t, \tag{9b}$$

$$\mathbf{\Phi}(i,i) \in \mathbb{U}, \forall i \tag{9c}$$

where constraints (9b) and (9c) are the transmit power constraint and finite discrete phase set constraint, respectively. We denote $\mathbb{U} = \{x = \beta(\frac{2\pi b}{2^B} + \Delta\varphi_i)e^{j(\frac{2\pi b}{2^B}+\Delta\varphi_i)} | b = -(2^{B-1}-1), \ldots, 0, \ldots, 2^{B-1}\}$ as a set of finite discrete phases. The MSE at the kth user is defined as $\mathrm{MSE}_k = \mathbb{E}[(x_k - y_k)(x_k - y_k)^H],\ \forall k$, where x_k and y_k represent the transmitted data and the received signal at the kth user, respectively. P_t in Eq. (9b) restricts the total transmit power. The diverse mathematical characteristics give rise to mixed discrete and continuous programming in the optimization problem (9), rendering it a non-deterministic polynomial-time hard (NP-hard) problem. As a scalar function of matrices $\mathbf{\Phi}$ and $\mathbf{F}$, the discrete function $\mathrm{MSE}_k(\mathbf{\Phi}, \mathbf{F})$ is not differentiable in the classical sense. To address this problem without resorting to any form of relaxation or quantization, we develop a DRL-based framework capable of directly searching for solutions within the discrete action space. The problem (9) can be restated equivalently as:

$$\min_{\mathbf{\Phi},\mathbf{F}} \max_{k=1,\ldots,N_k} \quad \mathbf{E}_{k,k} \tag{10a}$$

$$\text{s.t.} \quad \text{(9b), (9c)}, \tag{10b}$$

where $\mathbf{E} \in \mathbb{C}^{N_k \times N_k}$ is the MSE matrix encompassing all N_k users, with its diagonal elements representing the MSE of the kth user, i.e., $\mathrm{diag}(\mathbf{E}) \triangleq [\mathrm{MSE}_1, ..., \mathrm{MSE}_{N_k}]^T \in \mathbb{R}^{N_k \times 1}$. The MSE matrix is given by:

$$\mathbb{E}[\|\mathbf{x}-\mathbf{y}\|_2^2] = (\mathbf{I}_{N_k} - \widetilde{\mathbf{H}}\mathbf{F})(\mathbf{I}_{N_k} - \widetilde{\mathbf{H}}\mathbf{F})^H + \sigma_n^2\mathbf{I}_{N_k}, \tag{11}$$

where $\mathbf{y} = [y_1, \ldots, y_{N_k}]^T \in \mathbb{C}^{N_k \times 1}$ is the stacked received signal of all users, further expressed as $\mathbf{y} = \widetilde{\mathbf{H}}\mathbf{F}\mathbf{x} + \mathbf{n}$, where $\widetilde{\mathbf{H}} = \mathbf{H}_2\mathbf{\Phi}\mathbf{H}_1 + \mathbf{H}_3 \in \mathbb{C}^{N_k \times N_t}$ and $\mathbf{n} = [n_1, \ldots, n_{N_k}]^T \in \mathbb{C}^{N_k \times 1}$ are the effective channel and the AWGN noise vector, respectively. $\mathbf{H}_2 = \big[\mathbf{h}_{2,1}^T, \ldots, \mathbf{h}_{2,N_k}^T\big] \in \mathbb{C}^{N_k \times N_s}$ and $\mathbf{H}_3 = \big[\mathbf{h}_{3,1}^T, \ldots, \mathbf{h}_{3,N_k}^T\big] \in \mathbb{C}^{N_k \times N_t}$ are the stacked channel matrices for the RIS-UE channel and BS-UE channel, respectively.

III. DRL-BASED RIS CONFIGURATION ALGORITHM

In DRL, an agent iteratively learn by interacting with an environment across multiple episodes and time steps, each indicated by an additional time index t. At each step, the agent observes a state $s^{(t)}$, selects an action $a^{(t)}$ based on its policy $\pi(a^{(t)}, s^{(t)})$, evaluates its strategy using its value function $V(s^{(t)})$ and receives a reward $r^{(t)}$ from the environment for that action. Following this trial-and-error process, the agent aims to optimize its decision-making to enhance the total expected rewards over time. The policy and the value function are modeled using deep neural networks (DNN), denoted as $\pi_\theta = \pi(a^{(t)}, s^{(t)}|\theta)$ and $V_\theta = V(s^{(t)}|\theta)$, respectively, where θ denotes DNN's parameters.

A. Construction of RL Environment

1) State: Assuming the presence of a DRL agent functioning as a central controller with the capability to gather instantaneous CSI to construct the state. At time step t, the

state vector $\mathbf{s}^{(t)}$ comprises the channel matrix, the RIS matrix, and the previous action, as defined below

$$\begin{aligned}\mathbf{s}^{(t)} = \Big\{&\mathrm{Re}\{\mathbf{H}_1^{(t)}\}, \mathrm{Im}\{\mathbf{H}_1^{(t)}\}, \mathrm{Re}\{\mathbf{H}_2^{(t)}\}, \mathrm{Im}\{\mathbf{H}_2^{(t)}\},\\ &\mathrm{Re}\{\mathbf{H}_3^{(t)}\}, \mathrm{Im}\{\mathbf{H}_3^{(t)}\}, \mathrm{Re}\{\mathrm{diag}(\mathbf{\Phi}^{(t)})\},\\ &\mathrm{Im}\{\mathrm{diag}(\mathbf{\Phi}^{(t)})\}, \mathbf{a}^{(t-1)}\Big\}. \end{aligned} \tag{12}$$

The diagonal entries of the RIS phase shift matrix is denoted as $\mathrm{diag}(\mathbf{\Phi}^{(t)}) \triangleq [\phi_1^{(t)}, ..., \phi_{N_s}^{(t)}]^T \in \mathbb{C}^{N_s \times 1}$. Due to the default incapability of DNN to handle complex numbers, we flattened these matrices and separated them into their real and imaginary parts, except for the discrete action. As a result, we obtain $2N_sN_t$, $2N_kN_s$, and $2N_kN_t$ entries from the BS-RIS channel, RIS-UE channel, and BS-UE channel, respectively. Additionally, $2N_s$ entries are obtained from the RIS matrix, and N_s entries from the action vector. This results in a state space with dimensions $2N_sN_t + 2N_kN_s + 2N_kN_t + 2N_s + N_s$.

2) Action: The action $\mathbf{a}^{(t)}$ is defined as a vector of individual discrete actions with the dimension of N_s, expressed as $[a_1^{(t)}, \ldots, a_{N_s}^{(t)}]$, where each action $a_i^{(t)} \in \{0, \ldots, 2^B - 1\}, \forall i$ represents the index corresponding to the value in the phase set $\mathcal{F}$ for each RIS element. After obtaining the discrete phase $\hat{\varphi}_i$ from action $\mathbf{a}^{(t)}$, we add the Von-Mises phase error $\Delta\varphi_i$ to each phase, resulting in a continuous phase shift for the ith element. Subsequently, we use this actual phase φ_i, which includes phase error, to calculate the phase-dependent amplitude $\beta(\varphi_i)$ in Eq. (8), following Euler's formula, described as:

$$\phi_i = \beta(\varphi_i) \cdot (\cos\varphi_i + j \cdot \sin\varphi_i),\ \forall i = 1, ..., N_s. \tag{13}$$

Afterwards, we adopt maximum-ratio transmission (MRT) with equal power allocation in our downlink precoder design. Given $\mathbf{\Phi}^{(t)}$, the transmit beamformer at time t is expressed as:

$$\mathbf{F}^{(t)} = \sqrt{\frac{P_t}{N_k}} \cdot \frac{(\mathbf{H}_2^{(t)}\mathbf{\Phi}^{(t)}\mathbf{H}_1^{(t)} + \mathbf{H}_3^{(t)})^H}{\|\mathbf{H}_2^{(t)}\mathbf{\Phi}^{(t)}\mathbf{H}_1^{(t)} + \mathbf{H}_3^{(t)}\|_F}. \tag{14}$$

By following the process outlined above, we can ensure the constantly fulfillment of the constraints described in Sec. II-B. In the actual code, we employ action space shaping [14] to implement the multi-discrete action space.

3) Reward: The reward at the tth time step is determined by the negative of the maximum MSE across all users. Therefore, the reward function is defined as:

$$r^{(t)} = -\mathrm{MSE}_{\max}^{(t)}, \tag{15}$$

where $\mathrm{MSE}_{\max}^{(t)}$ is the maximum diagonal entry of the MSE matrix $\mathbf{E}$ and can be computed using Eq. (10) and Eq. (11).

B. The Proximal Policy Optimization Algorithm

We utilize the proximal policy optimization (PPO) algorithm [15] to tackle the optimization problem (10) within our custom RL environment, with the following training objective: $L^{\mathrm{PPO}}(\theta) = \hat{\mathbb{E}}^{(t)}\Big[L^{\mathrm{CLIP}}(\theta) - c_1 L^{\mathrm{VF}}(\theta) + c_2 S[\pi_\theta](s^{(t)})\Big]$, where c_1 and c_2 are hyper-parameters controlling the micro-adjustments of the optimization direction in the loss function. $S[\pi_\theta](s^{(t)})$ is the entropy of policy π_θ at state $s^{(t)}$ designed to ensure sufficient exploration during training. $L^{VF}(\theta) = (V(s^{(t)}|\theta) - \hat{r}^{(t)})^2$ denotes the squared-error loss between predicted and actual rewards. In each iteration, PPO seeks to optimize the policy network by maximizing the clipped objective:

$$\begin{aligned}L^{\mathrm{CLIP}}(\theta) = \hat{\mathbb{E}}^{(t)}\Big[&\min(\rho^{(t)}(\theta)\hat{A}^{(t)},\\ &\mathrm{clip}(\rho^{(t)}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}^{(t)})\Big], \end{aligned} \tag{16}$$

where $\hat{A}^{(t)} = \hat{r}^{(t)} - V(s^{(t)}|\theta)$ is the advantage function, which aims to estimate the relative value of the agent's action in the current state. The importance sampling ratio $\rho^{(t)}(\theta)$ reflects the probability change of an action under the updated policy versus the old policy, formulated as: $\rho^{(t)}(\theta) = \frac{\pi_\theta(a^{(t)}, s^{(t)})}{\pi_{\theta_{\mathrm{old}}}(a^{(t)}, s^{(t)})}$. The clipping function $\mathrm{clip}(\cdot)$ then bounds this ratio within $1-\epsilon$ and $1+\epsilon$ to mitigate excessive policy updates, where ϵ is the clipping parameter.

IV. Simulation Results

A. Experimental Settings and Baselines

We simulate a RIS-assisted mmWave MU-MISO system featuring $L = 4$ sparse paths, with a Rician factor of $\mu = 20$. Phase-dependent amplitude constants $(\beta_{\min}, \mu_{\mathrm{PDA}}, \kappa_{\mathrm{PDA}})$ are set to (0.9, 0.21, 3.4), and phase error constants $(\mu_{\mathrm{PE}}, \kappa_{\mathrm{PE}})$ are $(0.6\pi, 1.2)$. An uncertainty factor of $\psi = 0.001$ is used, with noise variance and total transmit power set at -30 dBm and 30 dBm, respectively. The discrete phase set $\mathcal{F}$, with a 1-bit resolution, is $\{0, \pi\}$. The PPO training includes a total of $N = 100$ episodes, each with $T = 20480$ time steps, using a learning rate of 0.0003 and a mini-batch size of 64. Each episode involves $K = 10$ epochs with a discount factor of $\gamma = 0.99$ and a clip range of $\epsilon = 0.2$ for policy updates. Additionally, the coefficients (c_1, c_2) are set to $(0.5, 0.01)$. We employ identical DNN architectures for policy and value function networks, which comprises 5 hidden layers, utilize $\tanh(\cdot)$ activation, and have the number of neurons in each layer being proportional to N_s. Specifically, the units are set to $32N_s$, $16N_s$, $8N_s$, $4N_s$, and $2N_s$ for each successive layer. This common design principle, wherein the neuron count is scaled with the number of RIS elements, aims to enhance the model's adaptability to increasingly complex systems [16]. All training and evaluations are performed on a personal computer equipped with an Intel Core i7-12700 processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3060 Ti graphics card.

The performance of the proposed PPO method is compared with the following benchmarks in terms of the average worst-case MSE, with error bars denoting 95% confidence intervals over 200 trials:

- *Conventional beamforming baseline* [2]: Employing the dominant eigenvector matching (DEM)-based method with continuous relaxation for designing the RIS matrix $\mathbf{\Phi}$, plus the MRT baseband precoder;

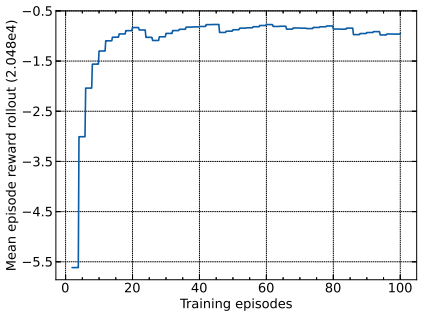

Fig. 2. The convergence behavior of the PPO algorithm, with $N_k = 2$, $N_t = 16$, and $N_s = 36$.

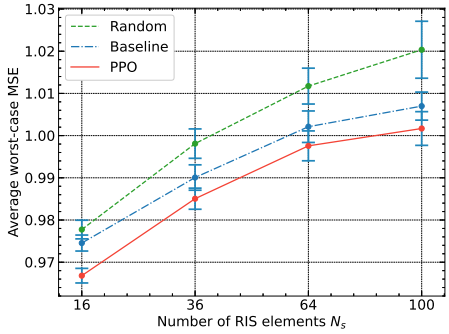

Fig. 3. Average worst-case MSE vs. number of RIS elements N_s with $N_k = 2$ and $N_t = 16$.

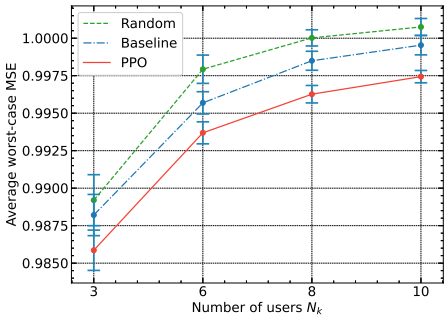

Fig. 4. Average worst-case MSE vs. number of users N_k with $N_t = 16$ and $N_s = 16$.

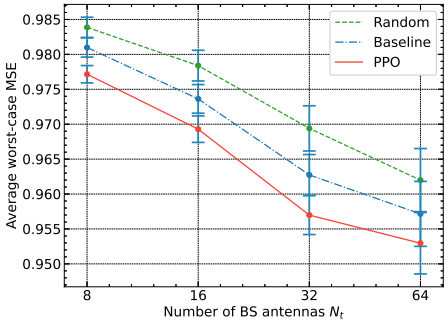

Fig. 5. Average worst-case MSE vs. number of BS antennas N_t with $N_k = 2$ and $N_s = 16$.

- *Random*: The phase shift values of $\mathbf{\Phi}$ are randomly chosen from the phase set $\mathcal{F}$, followed by a MRT-based precoder.

B. Performance Evaluation

Fig. 2 shows the convergence of the proposed PPO algorithm during the training process. The learning curve is obtained by rolling out the mean episode reward, which represents the average total cumulative reward acquired by the agent over several episodes during interactions with the environment. As depicted, PPO reaches a saturation state after only 20 episodes, demonstrating its efficient learning capability. Note that after the training process, the trained weights were recorded and therefore no re-training was required for dynamic channel states. Additionally, the computational complexity of the proposed method is low due to the lightweight framework consisting of five fully connected layers. In Fig. 3, the variation of the average worst-case MSE versus the number of RIS elements is shown. As observed, the MSE of all methods gradually increases as N_s increases. This can be attributed to the increasing difficulty of the optimization problem as the dimension of the state space grows, leading to the accumulation of more errors. The PPO algorithm consistently outperforms the Random and Baseline algorithms across all settings with varying numbers of elements, thus confirming the robustness of our proposed method.

Fig. 4 plots the average worst-case MSE vs. the number of users. As observed, an increase in N_k, corresponds to an elevation in the MSE. This uptrend suggests a direct relationship between the number of users and the complexity of the optimization challenge faced by the system. As can be seen, PPO consistently attains a lower worst-case MSE, showcasing its effectiveness in complex multi-user scenarios. The advantage of the PPO approach is even more apparent when observing the tightness of the confidence intervals. Notably, as N_k grows, the PPO method not only maintains its superiority but does so with increased certainty, as reflected by the smaller error bars.

In Fig. 5, the worst-case MSE vs. the number of BS antennas performance is presented. An examination of the graph reveals a consistent decrease in worst-case MSE as N_t increases for all three compared methods. The decrement in MSE is more pronounced as the antenna count progresses from 8 to 16 and then from 16 to 32, signaling considerable performance improvements with each augmentation in N_t. Comparing Fig. 3 and Fig. 5, it is seen that increasing the number of RIS elements and BS antennas has opposite effects on performance. This is because the design of RIS matrix is dominated by the accumulated CSI uncertainties in $\mathbf{h}_{2,k}$, and remains largely unaffected by the BS channels of $\mathbf{H}_1$ and $\mathbf{h}_{3,k}$.

Fig. 6 shows the impact of the phase-dependent amplitude constant $\beta_{\min}$ on the average worst-case MSE. Here, $\beta_{\min}$ varies from 0 to 1, reflecting the range of practical phase-dependent amplitude responses as specified in Eq. (8). The nonlinear relationship reveals that as $\beta_{\min}$ increases, the MSE tends to rise. This trend is anticipated since a higher $\beta_{\min}$ value indicates a stronger amplitude response, leading to an

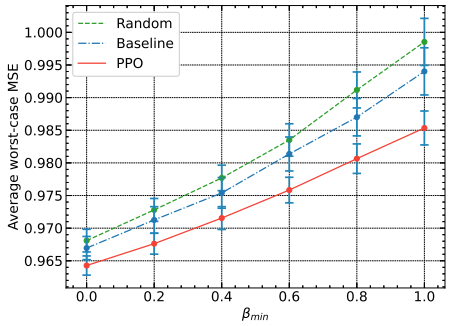

Fig. 6. Average worst-case MSE vs. phase-dependent amplitude constant $\beta_{\min}$ with $N_k = 2$, $N_t = 16$, and $N_s = 36$.

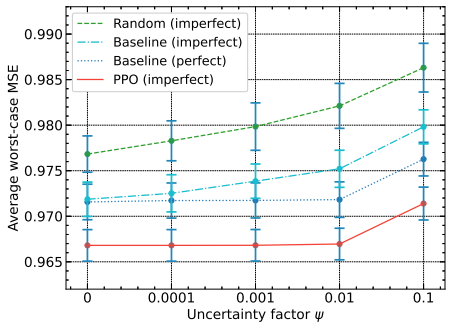

Fig. 7. Average worst-case MSE vs. uncertainty factor ψ with $N_k = 2$, $N_t = 16$, and $N_s = 16$.

increase in the worst-case error across the system in general. Furthermore, the performance gap widens as $\beta_{\min}$ increases, indicating that the continuous relaxation of the conventional Baseline degrades performance under HWI effects due to error propagation. Alternatively, the proposed method directly designs finite phase shift solutions for achieving stable performance.

In Fig. 7, we present the connection between the average worst-case MSE and the uncertainty factor ψ. As illustrated, both the Baseline and Random methods demonstrate sensitivity to the uncertainty level, while the proposed PPO method remains unaffected as ψ increases from 0 to 0.01. Moreover, the PPO outperforms the Baseline method by incorporating a feedback learning mechanism for both RIS and MRT beamformer design. Specifically, the reward for designing RIS, as defined in Eq. (15), is calculated using the MSE computed in Eq. (11), i.e., by utilizing the results of MRT along with RIS design to improve future iterations.

V. Conclusion

Our study introduces a DRL-based discrete optimization framework that can leverage traditional algorithms for the joint design of active and passive beamforming in RIS-assisted mmWave MU-MISO systems. Notably, we pioneers the use of PPO to optimize the discrete RIS configuration while addressing HWI and CSI inaccuracies. Subsequently, the MRT technique is applied to derive the corresponding active beamformer, aiming to minimize the worst-case MSE for all users. Simulation results demonstrate the robustness of our method compared to the conventional beamforming baseline across various settings. Our findings underscore the potential of developing DRL-based approaches to address scenarios involving HWI and CSI uncertainties. Furthermore, we have made our code publicly available on GitHub, providing a valuable resource for future studies and practical applications of advanced DRL techniques in RIS-assisted communication systems.

References

[1] S. He, Y. Zhang, J. Wang, J. Zhang, J. Ren, Y. Zhang, W. Zhuang, and X. Shen, "A survey of millimeter-wave communication: Physical-layer technology specifications and enabling transmission technologies," *Proc. IEEE*, vol. 109, no. 10, pp. 1666–1705, Oct. 2021.

[2] N. K. Kundu and M. R. McKay, "RIS-assisted MISO communication: Optimal beamformers and performance analysis," in *IEEE Global Commun. Conf. Workshops (GC Wkshps)*, Dec. 2020, pp. 1–6.

[3] K. Ying, Z. Gao, S. Lyu, Y. Wu, H. Wang, and M.-S. Alouini, "GMD-based hybrid beamforming for large reconfigurable intelligent surface assisted millimeter-wave massive MIMO," *IEEE Access*, vol. 8, pp. 19 530–19 539, Jan. 2020.

[4] Z. Peng, Z. Chen, C. Pan, G. Zhou, and H. Ren, "Robust transmission design for RIS-aided communications with both transceiver hardware impairments and imperfect CSI," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 528–532, Mar. 2022.

[5] M.-A. Badiu and J. P. Coon, "Communication through a large reflecting surface with phase errors," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 184–188, Feb. 2020.

[6] T. Wang, M.-A. Badiu, G. Chen, and J. P. Coon, "Outage probability analysis of RIS-assisted wireless networks with von mises phase errors," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2737–2741, Dec. 2021.

[7] Z. Peng, W. Xu, L.-C. Wang, and C. Zhao, "Achievable rate analysis and feedback design for multiuser MIMO relay with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 780–793, Feb. 2014.

[8] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Jun. 2020.

[9] Z. Peng, Z. Zhang, L. Kong, C. Pan, L. Li, and J. Wang, "Deep reinforcement learning for RIS-aided multiuser full-duplex secure communications with hardware impairments," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21 121–21 135, Nov. 2022.

[10] P. Chen, X. Li, M. Matthaiou, and S. Jin, "DRL-based RIS phase shift design for OFDM communication systems," *IEEE Wireless Commun. Lett.*, vol. 12, no. 4, pp. 733–737, Apr. 2023.

[11] B. Saglam, D. Gurgunoglu, and S. S. Kozat, "Deep reinforcement learning based joint downlink beamforming and RIS configuration in RIS-aided MU-MISO systems under hardware impairments and imperfect CSI," in *IEEE Int. Conf. on Commun. Workshops (ICC Workshops)*, May 2023, pp. 66–72.

[12] P. Wang, J. Fang, L. Dai, and H. Li, "Joint transceiver and large intelligent surface design for massive MIMO mmwave systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1052–1064, Oct. 2021.

[13] W.-Y. Chen, C.-Y. Wang, R.-H. Hwang, W.-T. Chen, and S.-Y. Huang, "Impact of hardware impairment on the joint reconfigurable intelligent surface and robust transceiver design in MU-MIMO system," *IEEE Trans. Mobile Comput.*, pp. 1–16, Jun. 2023.

[14] A. Kanervisto, C. Scheller, and V. Hautamäki, "Action space shaping in deep reinforcement learning," in *IEEE Conf. on Games (CoG)*, Aug. 2020, pp. 479–486.

[15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, Jul. 2017.

[16] J. Gao, C. Zhong, X. Chen, H. Lin, and Z. Zhang, "Unsupervised learning for passive beamforming," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1052–1056, Jan. 2020.