IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 23 November 2025, accepted 4 January 2026, date of publication 12 January 2026, date of current version 15 January 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3651969

RESEARCH ARTICLE

Integrated SDR-Based Passive Beamforming and DRL-Driven Analog Beam Selection in IRS-Assisted Massive MIMO System

IRFAN AHMED[1], (Senior Member, IEEE), KHALIL SHAHID[1], MOATH AWAWDEH[1], (Senior Member, IEEE), ADNAN SHAHID[2], (Senior Member, IEEE), MAMOUN GUENACH[3,4], (Senior Member, IEEE), HEDI KHAMMARI[5], AND TARIG FAISAL[1], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Higher Colleges of Technology, Al Dhannah City, Abu Dhabi, United Arab Emirates
[2]IMEC, Department of Information Technology, Ghent University, 9052 Ghent, Belgium
[3]IMEC, 3001 Leuven, Belgium
[4]IMEC, Department of Telecommunications Information Processing, Ghent University, 9000 Ghent, Belgium
[5]Institut Supérieur des Sciences Appliquées et de Technologie de Kairouan, Kairouan 3100, Tunisia

Corresponding author: Irfan Ahmed (iahmed3@hct.ac.ae)

This work was supported in part by the Higher Colleges of Technology (HCT) Research Grant from HCT, Abu Dhabi, United Arab Emirates, under Grant 21060.

ABSTRACT This paper presents a deep reinforcement learning (DRL)-based hybrid beamforming (DHB) and semidefinite relaxation (SDR)-driven passive beamforming (SPB) for the intelligent reflective surface (IRS)-assisted multiuser massive multiple-input multiple-output (MIMO) downlink system. The conventional hybrid beamforming in massive MIMO provides multiple directional beams, but due to high absorption loss and low penetration power at millimeter wave (mmWave) frequencies, the distant or shadowed users suffer from low signal-to-noise ratio (SNR). The integration of IRS with massive MIMO enables extended communication within a beam, hence increasing the system capacity and coverage. The optimization problem tackled in this paper is inherently complex and nonconvex, primarily due to the coupling of multiple variables and strict unit-modulus constraints. The discrete nature of beam selection further complicates the problem, rendering it NP-hard, as it involves combinatorial optimization over binary variables representing beam-user associations. To address this, we reformulate the problem into a quadratically constrained quadratic program (QCQP) followed by semidefinite relaxation (SDR) for IRS phase shift optimization. Then, based on the knowledge of the optimized IRS phase shifters, DRL is used to obtain the analog beamformer at the base-station. The DQN agent selects the beam-user pair with the help of the optimized cascaded channel state information to form the analog beamformer. Simulation results show that the proposed knowledge-based DRL beamforming design outperforms the state-of-the-art beamforming schemes in terms of transmit power, energy efficiency, and spectral efficiency. Specifically, the proposed design reduces the transmit power 20.37% for the target SNR of 10 dB.

INDEX TERMS Hybrid beamforming, intelligent reflective surfaces, massive MIMO, reinforcement learning.

I. INTRODUCTION

The emerging design and development of the Internet of Things (IoTs) and their deployment demand for wireless and mobile data can only be met by the next generation ultra-high-speed system. The sixth generation (6G) communication

The associate editor coordinating the review of this manuscript and approving it for publication was Tarcisio Ferreira Maciel.

extends the frequency spectrum from 0.1 to 10 THz [2]. The 5G technology uses a high-frequency mmWave band ranging from 30 GHz to 300 GHz along with the massive multiple-input-multiple-output (MIMO) systems and the hybrid beamforming overcomes the high power consumption in the radio frequency (RF) chain. The 6G introduces the concept of intelligent reflective surfaces (IRS) in 3GPP release-18, which provides a low-cost low-power relay of electromagnetic waves with the capability of change of phase [3]. IRSs are considered a low-cost and energy-efficient substitute for traditional relays and repeaters. The IRS element consists of a low profile and lightweight programmable passive device such as positive-intrinsic-negative (PIN) diodes and phase shifters [4]. The IRS system consists of a sub-wavelength meta-material particle surface which is controlled by a PIN diode controller. The IRS does not require an extra power source and does not perform any signal processing, such as decoding.

The deployment of IRS in massive MIMO systems can effectively mitigate issues related to service dead zones, improving overall network performance and user experience in 6G environments [5]. Furthermore, IRS intelligent design can lead to substantial improvements in spectral efficiency and energy savings, making it a critical component in the future of wireless communications [6]. This integration not only addresses the challenges of high mobility and obstruction, but also enhances energy efficiency, making IRS a key enabler for 6G networks [7]. The potential of IRS to improve spectral efficiency and energy savings underscores its importance in the development of sustainable wireless communication systems for 6G networks.

IRS can play a significant role in enhancing the performance of low Earth orbit (LEO) satellite systems like Starlink. LEO satellites often suffer from line-of-sight (LoS) obstructions due to buildings, terrain, or foliage. IRS panels, strategically placed on rooftops or building facades, can intelligently reflect satellite signals to users in shadowed or hard-to-reach areas, ensuring continuous connectivity. IRS can support real-time data exchange essential for smart traffic management, environmental monitoring, and emergency response systems. This capability aligns with the anticipated demands of 6G networks, which are expected to support massive machine-type communications and ultra-reliable low-latency communications. As urban environments become increasingly complex, the ability of IRS to dynamically adjust signal pathways will be crucial in maintaining robust connectivity amidst physical obstructions and high user density. Consequently, this transformative technology not only promises to improve user experiences, but also contributes significantly to the operational efficiency and sustainability of future urban ecosystems.

Due to the advantages mentioned above, IRS has been extensively investigated as an integral part of the future 6G mobile communication systems. For example, in the case of a single user system, IRS-assisted communications provide received signal power gain of the order of $\mathcal{O}(N^2)$ compared to the conventional MIMO system which gives received signal power gain of the order of $\mathcal{O}(M)$, where N is the number of IRS reflector elements and M is the number of antennas at the transmitter [8]. This higher gain is due to the coherent combination of signals reflected by IRS.

II. LITERATURE REVIEW

The synergy between active and passive beamforming in IRS-assisted massive MIMO systems has been a subject of growing interest. By combining both techniques, researchers aim to leverage the strengths of each to achieve optimal performance. The literature on active and passive beamforming design in IRS-assisted massive MIMO communication highlights the significant advancements and ongoing challenges in the field.

In [9] the authors develop a joint design for IRS phase shift and hybrid beamforming for narrowband MIMO systems. The beamforming design utilizes the angular sparsity of frequency-selective mmWave channels. But this paper considers an MISO system with one user.

The paper [10] discusses the design of analog beamforming within hybrid beamforming (HBF) for multi-user millimeter-wave MIMO systems without IRS assistance. It transforms the non-convex analog precoder optimization problem into a quadratically constrained quadratic program (QCQP). An iterative coordinate ascent (ICA) algorithm is proposed to efficiently solve this problem, ensuring high-performance solutions. The study highlights the advantages of HBF in reducing the number of radio frequency chains while maximizing the sum rate, demonstrating its effectiveness for 5G networks.

It has been shown [11] that the analog precoder design, typically a high-dimensional integer programming problem, can be effectively divided into multiple low-dimensional subproblems. Using an analytical lower bound for each subproblem and employing the branch-and-bound method, the proposed approach achieves optimal solutions, enabling the hybrid precoder to achieve 98% of the performance of fully digital solutions in typical antenna settings.

At mmWave and higher frequencies, the NLOS components decay very fast, thereby creating a coverage problem for users behind the obstacles. The IRS solves this problem by intelligently reflecting the incident signal towards the shadowed users. In the same context, the paper [12] discusses the design of analog beamforming within a hybrid beamforming framework for IRS-assisted mmWave communications, focusing on minimizing the sum-mean-square error (sum-MSE) in multi-user MIMO systems. It proposes an accelerated Riemannian gradient algorithm using majorization minimization to optimize the unit-modulus constrained analog precoder and the IRS reflection matrix.

Ribeiro et al. [13] propose two low-complexity algorithms: Kronecker Factorization (KF) and Third-Order Tensor (TOT) for joint active and passive beamforming in IRS-assisted MIMO systems by exploiting the Kronecker structure of the cascaded channel matrices. The KF method uses singular value decomposition (SVD) on factorized horizontal and vertical channel components, while the TOT method applies high-order singular value decomposition (HOSVD) to third-order tensors representing the channel. Both approaches decompose the global optimization problem into smaller subproblems, This paper uses Kronecker structure-based over-simplified channel model with rank-one approximation. It assumes a negligible correlation between IRS and BS antenna elements and does not account for interuser interference.

The discrete PS-based IRS-aided mmWave MIMO communication is studied in [14]. The authors propose a cross-entropy-based probability-learning (CEPL) algorithm for designing reflecting beamforming with 1-bit resolution PSs and then extended to high-resolution PSs. The complexity of the optimization problem increases with the resolution of phase shifters, which can affect the overall performance and efficiency of the system. Their method focuses solely on discrete PS scenarios and relies on a CEPL algorithm, which is effective for discrete optimization but lacks the adaptability and learning capability of DRL in dynamic environments.

Xiong et al. [15] show the application of active and passive beamforming in IRS-assisted physical layer security. It formulates the problem of maximization of the secrecy rate. This is a non-convex problem due to the fractional objective function and the strict constant modulus constraints applied on both the transmit and IRS phase shift variables. To address this, the authors propose two algorithmic solutions: the Dinkelbach-BSUM algorithm and the product manifold conjugate gradient descent (PMCGD) algorithm. The PMCGD algorithm requires considerably more iterations to reach convergence, approximately 350 iterations compared to the SDR IRS algorithm, which converges in roughly 20 iterations [15].

In [16], the authors use an irregular IRS structure instead of regular uniform linear array (ULA) or planar linear array (PLA) to address the limitations of regular architectures to enhance spatial degrees of freedom and system capacity without increasing the number of IRS elements. The optimization problem, which is non-convex due to discrete phase shifts and topology constraints, is addressed using a DL framework that predicts optimal precoding vectors based on omni-received uplink pilot signals. The DL model is trained by implicit mapping between the received pilot sequence and the beamforming vectors without the conventional explicit CSI acquisition. While the DL framework significantly alleviates training overhead by predicting the best beamforming vectors using omni-received uplink pilots, it simultaneously introduces challenges related to model training convergence and the need for extensive datasets.

The paper [17] introduces a reduced WMMSE (RWMMSE) algorithm that optimizes the active beamforming process with lower complexity. The authors propose a combination of RWMMSE for active beamforming and gradient projection (GP) enhanced with Barzilai-Borwein step size selection for passive beamforming. RWMMSE reduces computational complexity by reformulating the optimization problem to eliminate the Lagrangian power constraint and reduce matrix dimensions, while GP efficiently updates IRS phase shifts without iterative eigenvalue computations. It might still face scalability challenges as the system sizes grow further. The complexity analysis provided indicates a dependency on factors such as the number of iterations for convergence and the outer AO loop, which might become prohibitive in dense scenarios.

In their study, Yue Ju et al. [18] aim to improve beamforming techniques to optimize the weighted sum rate (WSR) in a multi-user setting involving both active and passive components in the IRS architecture. They employ the WMMSE method to reformulate the WSR issue into a more manageable format. This allows them to implement alternating optimization (AO) for refining transmit precoders, receive equalizers, weighting matrices, and IRS phase shifts. Despite utilizing the WMMSE transformation and creating an AO algorithm grounded in majorization-minimization (MM) methods, the derived solution remains suboptimal, representing a stationary point rather than the global optimum. Additionally, the study considers a narrowband quasi-static block-fading channel.

In [19], the authors derive the weighted sum rate (WSR) and energy efficiency (EE) based on imperfect cascaded CSI, formulating a nonconvex optimization problem aimed at maximizing WSR while jointly optimizing power allocation, HBF, and phase and amplitude of IRS elements. The methodology employs an AO approach, integrating weighted minimum mean-square error (WMMSE) and block coordinate descent (BCD) techniques to iteratively derive closed-form solutions for each subproblem. It shows that increasing the number of IRS elements does not always yield performance improvements due to the adverse effects of estimation errors. The iterative update of multiple variables, such as power allocation, phase, amplitude, digital beamforming, and analog beamforming, requires frequent matrix inversions and convergence checks, which might not be practical in real-time communication systems or in scenarios with rapid CSI variations. The WSR and EE are interdependent, and the trade-off between WSR and EE is not evaluated in this paper. In addition, this paper does not employ any artificial intelligence-based technique.

Another paper [20] investigates the joint optimization problem of the base station beamforming and IRS phase control in a wireless communication system with multiple users, aiming to maximize the system's sum rate. The authors propose a deep deterministic policy gradient (DDPG) algorithm based on deep reinforcement learning (DRL) to solve this nonconvex optimization problem. The DDPG algorithm is a model-free off-policy method for learning continuous actions. This approach is utilized to optimize both the base station beamforming and the IRS phase control simultaneously, addressing the nonconvex optimization problem caused by IRS limitations. This work is close to our work, but its system model only considers the reflected path, also it does not consider the hybrid beamforming at BS which cannot be energy efficient due to the separate RF chain for each transmit antenna at BS. Moreover, the optimization problem maximizes the sum rate without considering QoS requirements.

In [21], the authors focus on hybrid beamforming design in IRS-assisted mmWave massive MU-MISO systems, utilizing a zero-force (ZF) beamforming technique for the digital component. It employs a probability learning algorithm based on a cross-entropy optimization (CEO) framework to optimize the analog part of hybrid beamforming and IRS phase shifts. Although it addresses energy efficiency and spectral reuse, this paper lacks learning-based adaptability to dynamic environments and does not address transmit power minimization directly.

A comprehensive overview of machine learning techniques in IRS-assisted communication is presented in [8]. In [22], the authors use supervised deep learning to train and predict reflection phases. They assume active reflective beamforming with RF chains deployed at the IRS, which is not feasible in most cases. In another work [23], they use DRL by assuming IRS as an agent. The IRS contains passive as well as active reflectors for channel estimation purposes. This design requires a large amount of computational power in the IRS. The paper [24] uses the supervised deep neural network (DNN) to reconfigure the IRS reflection phases and finds the precoder in BS with the help of pilot signals from a single user. Huang et al. [25] propose a DRL-based hybrid beamforming design for IRS-assisted THz multi-hop communications. The DRL consists of a deep deterministic policy gradient (DDPG) actor-critic based agent that approximates the digital beamforming vectors and the reflection phases. This paper assumes full digital precoding at the BS which is difficult to implement in the massive MIMO system because of the very large number of RF chains required. In [26], a DDPG-based actor-critic model-free DRL agent is used to optimize the reflective phases, this work assumes infinite resolution reflective phases which is impractical.

To overcome the limitations identified in prior work, this paper leverages DRL to perform analog beam selection at the BS using the cascaded channel that includes the BS–IRS and IRS–user links. Specifically, the DRL agent selects optimal beams from a predefined DFT-based codebook, which significantly reduces the dimensionality of the channel and facilitates the construction of a low-complexity digital precoder at the BS. For passive beamforming at the IRS, we assume access to perfect CSI for both the BS–IRS and IRS–user channels. This CSI is used to compute phase shift configurations that maximize the effective cascaded channel gain, thereby enhancing the overall system performance. It is important to note that accurate CSI acquisition in IRS-assisted systems is a nontrivial challenge due to the passive nature of IRS elements and the indirect channel structure. However, the development of robust CSI estimation techniques is beyond the scope of this work. We assume that perfect CSI is available, as enabled by existing estimation methods proposed in the literature for IRS-assisted wireless systems [27], [28].

Therefore, the main contributions of this paper can be summarized as follows.

- A novel framework is proposed that integrates semidefinite relaxation (SDR)-based passive beamforming at the IRS with deep reinforcement learning (DRL)-based analog beam selection at the base station (BS) for IRS-assisted massive MIMO systems.
- The paper formulates the IRS phase shift design as a quadratically constrained quadratic program (QCQP) and solves it using semidefinite relaxation to maximize the effective channel gain. A deep Q-network (DQN) agent is trained to select optimal beam-user pairs using cascaded channel state information, enabling efficient analog beamforming with reduced complexity.
- Simulation results demonstrate that the proposed RL-SDR-based beamforming scheme outperforms conventional greedy, stable matching, and Hungarian-based methods in terms of transmit power, energy efficiency, and spectral efficiency. The paper provides a detailed Simulink-based implementation of the DRL agent and validates its performance through extensive simulations, showing convergence and robustness.

The rest of the paper is organized as follows. The system, signal, and channel models are described in Section III. The problem is formulated in Section IV. Section V presents the proposed DRL-based beam-user selection and hybrid precoder design for IRS-assisted multiuser massive MIMO system. Simulation results are given in Section VI, followed by the conclusions in Section VII.

Notations: Bold upper/lower case letters denote vectors and matrices, respectively. The notations $\mathbf{X}^{\dagger}$, $\mathbf{X}^{T}$, $\mathbf{X}^{*}$, and $\mathbf{X}^{H}$ denote the pseudoinverse, transpose, conjugate, and conjugate transpose of a matrix $\mathbf{X}$. The element in the ith row and jth column of $\mathbf{X}$ is denoted by $[\mathbf{X}]_{i,j}$. For a set A, $Card(\mathcal{A})$ is the cardinality of set $\mathcal{A}$. For matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \odot \mathbf{B}$ and $\mathbf{A} \otimes \mathbf{B}$ are the Hadamard product and Kronecker product, respectively. For quick reference, the list of main symbols used in this paper is given in Table 1.

III. SYSTEM MODEL

We consider an IRS-assisted multiuser mmWave massive MIMO downlink system as shown in Fig. 1. The base station (BS) is located in the center of the cell and is equipped with M antennas uniform linear array (ULA) and N_{RF} RF chains to serve K, single antenna users. Due to the blockage, the communication between the BS and the users takes place either through the intelligent reflective surface (IRS) or direct or through both IRS and direct links. The IRS has N reflecting

TABLE 1. List of main symbols.

Symbol	Description
M	Number of BS antennas
N	Number of IRS antenna elements
N_{RF}	Number of RF chains at BS
K	Number of users
$\mathcal{K}$	Set of users
$\mathcal{B}$	Set of selected beams
$\mathcal{K}_n$	Set of users in beam n
χ	beam-user selection binary variable
$\mathcal{C}$	Correlation matrix
ϕ_k^l	Array response angle of users k in path l
θ_k^l	Physical angle of users k in path l
$\mathcal{I}$	Beams index set
λ	Carrier wavelength
ρ	Transmit SNR
Γ	SNR gap
$\boldsymbol{\Theta}$	Weights matrix of DNN
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
∇	Gradient operator
γ	Discount factor
α	Learning rate

elements in the form of a uniform planar array (UPA). The BS contains N_{RF} RF chains such that $K \leq N_{RF} < N$. At BS, each antenna is connected with all the RF chains through an independent phase shifter to form a fully-connected structure. The IRS consists only of phase shifters, and the phase shift is controlled by the IRS controller, which communicates with the BS via a backhaul link [4]. The BS is responsible for all the signal processing of IRS beamforming and sending it to the IRS controller. The user set $\mathcal{K}$ is defined as $\mathcal{K} = \{k|k = \{1, 2, \ldots, K\}\}$.

We assume a single stream per user, then, the symbol vector is

$$\mathbf{s} = [s_1, s_2, \ldots, s_K]^T. \tag{1}$$

It is assumed that $s_k, \forall k = \{1, \ldots, K\}$ are random variables with zero mean and unit variance. The hybrid precoder can be written as

$$\mathbf{F}_B = \mathbf{F}_{AB}\mathbf{F}_{DB} \in \mathbb{C}^{M\times K}, \tag{2}$$

such that $P = \|\mathbf{F}_B\|^2$ is the total transmit power of the BS. Upon receiving the data block, the BS first uses a low-dimensional digital precoder $\mathbf{F}_{DB} \in \mathbb{C}^{N_{RF}\times K}$ and then an analog precoder $\mathbf{F}_{AB} \in \mathbb{C}^{M\times N_{RF}}$ is used to create the $M \times 1$ transmit symbols vector. The digital precoder is given as

$$\mathbf{F}_{DB} = [\mathbf{f}_{DB,1}, \mathbf{f}_{DB,2}, \ldots, \mathbf{f}_{DB,K}] \quad \in \mathbb{C}^{N_{RF}\times K}, \tag{3}$$

and the analog precoder is

$$\mathbf{F}_{AB} = [\mathbf{f}_{AB,1}, \mathbf{f}_{AB,2}, \ldots, \mathbf{f}_{AB,N_{RF}}] \quad \in \mathbb{C}^{M\times N_{RF}}. \tag{4}$$

After passing through the digital and analog precoders, the transmitted symbol vector is $\mathbf{x} \in \mathbb{C}^{M\times 1}$. It can be written as

$$\mathbf{x} = \mathbf{F}_{AB}\mathbf{F}_{DB}\mathbf{s} \quad \in \mathbb{C}^{M\times 1} \tag{5}$$

$$= \sum_{k=1}^{K} \mathbf{F}_{AB}\mathbf{f}_{DB,k}s_k. \tag{6}$$

The $N \times 1$ received signal $\hat{\mathbf{x}}$ at IRS is given by

$$\hat{\mathbf{x}} = \mathbf{H}_{BI}^H\mathbf{x}, \tag{7}$$

where $\mathbf{H}_{BI} \in \mathbb{C}^{M\times N}$ is the wireless channel between the BS and IRS. The reflection beamforming of IRS is represented by a diagonal matrix $\boldsymbol{\Theta}$ and is defined as $\boldsymbol{\Theta} \triangleq \kappa\mathrm{diag}(\boldsymbol{\xi}) \in \mathbb{C}^{N\times N}$, where $0 \leq \kappa \leq 1$ is a reflection coefficient and $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^H$ are the diagonal elements of $\boldsymbol{\Theta}$, such that each ξ_n satisfies $\xi_n \in \{\xi_n||\xi_n| = \frac{1}{\sqrt{N}}, \forall n\}$, e.g., $\xi_n = e^{j\theta_n}$ is a phase shift of n^{th} element of the IRS. The beamforming vector of IRS can be obtained as $\boldsymbol{\xi} = \frac{1}{\kappa}\mathrm{diag}(\boldsymbol{\Theta}) \in \mathbb{C}^{N\times 1}$.

We assume that there is a direct path between BS and users, and IRS assists in the communication between BS and users.

The signal received y_k at the user k via IRS is given as

$$\begin{aligned} y_k &= \mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\hat{\mathbf{x}} + \mathbf{h}_{BU,k}^H\mathbf{x} + z_k, \\ &= (\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{BU,k}^H)\mathbf{x} + z_k, \\ &= (\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{BU,k}^H)\mathbf{F}_{AB}\mathbf{F}_{DB}\mathbf{s} \\ &\quad + z_k, \qquad \forall k \in \mathcal{K}, \end{aligned} \tag{8}$$

where $z_k \sim \mathcal{CN}(\mathbf{0}, 1)$ is the AWGN noise at the user k, $\mathbf{h}_{IU,k}$ is the $\mathbb{C}^{N\times 1}$ channel vector between the IRS and the user k, and $\mathbf{h}_{BU,k}$ is the $\mathbb{C}^{M\times 1}$ channel vector between the BS and the user k. The received signal of each user can be combined to form the composite received signal as $\mathbf{y} = [y_1, \ldots, y_K]^T$. The SINR of user k is given by

$$SINR_k = \frac{|(\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{BU,k}^H)\mathbf{f}_{B,k}|^2}{\sum_{j\neq k}^{K}|(\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{BU,k}^H)\mathbf{f}_{B,j}|^2 + \sigma_{z_k}^2}. \tag{9}$$

A. CHANNEL MODEL

In order to describe the wireless channel, we use the widely adopted extended Saleh-Valenzuela model [29]. Due to the directional nature of the mmWave propagation, we assume that the channel contributes limited propagation paths L_{BI} and L_{IU} for the BS-IRS and IRS-users links, respectively. The $M \times N$ MIMO channel between BS and IRS can be written as

$$\mathbf{H}_{BI} = \sqrt{\frac{MN}{L_{BI}}}\sum_{l=0}^{L_{BI}}\alpha_l\mathbf{a}_P(\phi_l, \varrho_l)\mathbf{a}_L(\varphi_l)^H, \tag{10}$$

where α_0 and α_l $(l = 1, \ldots, L_{BI})$ represent the complex gains of the LOS, the $l = 0$ path and non-line-of-sight (NLOS), $l = 1, \ldots, L_{BI}$ paths with i.i.d. $\mathcal{CN}(0, 1)$, respectively. Moreover, $\mathbf{a}_P$ and $\mathbf{a}_L$ are the respective array response vectors of UPA and ULA. The variables ϕ_l (ϱ_l) and φ_l are the l^{th} path's azimuth (elevation) angle-of-arrival (AoA) and angle-of-departure (AoD), respectively. We assume these angles are

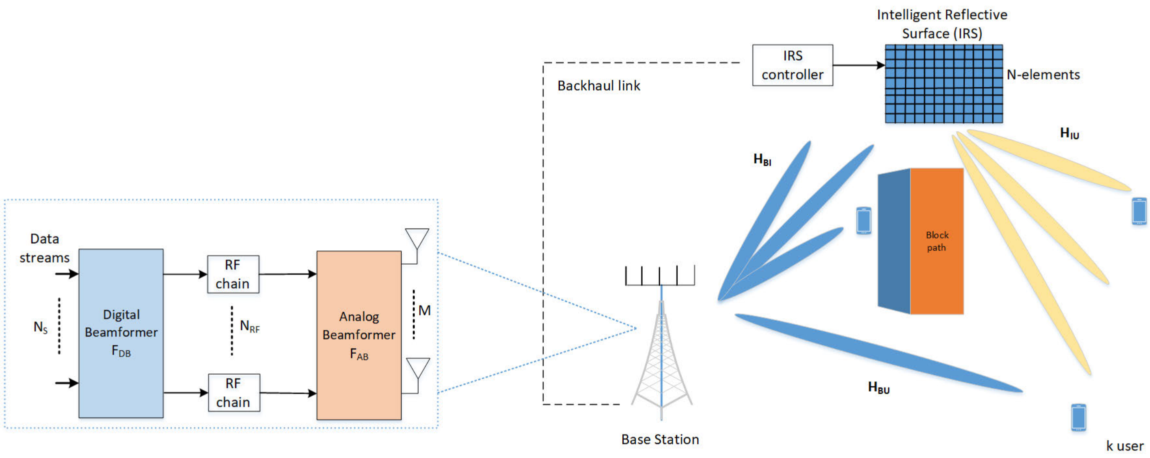

FIGURE 1. System model of IRS-assisted massive MIMO downlink.

uniformly distributed with $\sim \mathcal{U}(-\pi/2, \pi/2)$. The transmit antenna response vector of ULA is given by [30]

$$\mathbf{a}_L(\varphi_l) \triangleq \frac{1}{\sqrt{M}}\left[1, e^{j\frac{2\pi d_B}{\lambda}\sin(\varphi_l)}, \dots, e^{j\frac{2\pi d_B}{\lambda}\sin(\varphi_l)(M-1)}\right]^T, \tag{11}$$

where $j = \sqrt{-1}$, λ is the wavelength and d_B is the antenna spacing at the base-station. The receive antenna array response vector for an $N = P \times Q$-element UPA is defined as [30]

$$\mathbf{a}_P(\phi_l, \varrho_l) \triangleq \frac{1}{\sqrt{N}}\begin{bmatrix} 1 \\ e^{j\frac{2\pi d_I}{\lambda}(p\sin(\phi_l)\sin(\varrho_l)+q\cos(\varrho_l))} \\ \vdots \\ e^{j\frac{2\pi d_I}{\lambda}((P-1)\sin(\phi_l)\sin(\varrho_l)+(Q-1)\cos(\varrho_l))} \end{bmatrix} \tag{12}$$

where d_I is the antenna elements spacing at the IRS and as we assume $P = Q$, therefore, $P = Q = \sqrt{N}$ and $0 \le p \le \sqrt{N} - 1, 0 \le q \le \sqrt{N} - 1$.

Similarly, the channel between IRS and user k is

$$\mathbf{h}_{IU,k} = \sqrt{\frac{N}{L_{IU}}}\sum_{l=0}^{L_{IU}} \beta_{k,l}\mathbf{a}_P(\psi_{k,l}, \vartheta_l), \forall k \tag{13}$$

where $\beta_{k,0}$ and $\alpha_{k,l}$, $l = 1, \dots, L_{IU}$ represent the complex gains of the LOS, the $l = 0$ path and NLOS, $l = 1, \dots, L_p$ paths with i.i.d. $\mathcal{CN}(0, 1)$, respectively. The $\mathbf{a}_P$ is the array response vector of UPA, and the variables $\psi_{k,l}$ and ϑ_l are the l^{th} path's respective azimuth and elevation AoD for user k and are uniformly distributed with $\sim \mathcal{U}(-\pi/2, \pi/2)$. The response vector of the UPA transmit antenna $\mathbf{a}(\psi, \vartheta)$ can be obtained similarly to (12).

$$\mathbf{a}(\psi, \vartheta) = \frac{1}{\sqrt{N}}\left[exp(-j2\pi\theta_k^l i)\right]_{i\in\mathcal{I}(N)}^T, \tag{14}$$

where $\mathcal{I}(N) = \{i - (N-1)/2, i = 0, 1, \dots, N-1\}$ is a symmetric index set centered at zero [31], $\phi_k^l = \frac{d}{\lambda}\sin\theta_k^l$ is the array response angle, θ is the physical angle whose value ranges between $-\frac{\pi}{2} \le \theta \le \frac{\pi}{2}$, λ is the wavelength, and d is the antenna spacing at the base-station. The overall channel matrix $\mathbf{H}_{IU}$ between IRS and users is given as

$$\mathbf{H}_{IU} = [\mathbf{h}_{IU,1}, \dots, \mathbf{h}_{IU,K}] \in \mathbb{C}^{N\times K}. \tag{15}$$

Similarly, we can obtain the channel matrix between BS and the users, $\mathbf{H}_{BU} = [\mathbf{h}_{BU,1}, \dots, \mathbf{h}_{BU,K}] \in \mathbb{C}^{M\times K}$.

IV. POWER-EFFICIENT HYBRID BEAMFORMING VIA IRS PHASE CONTROL AND SDR TECHNIQUES

In this section, we formulate a joint IRS phase shifts and hybrid beamforming design problem for the IRS-assisted mmWave MIMO downlink system to minimize the sum power performance subjected to the SINR based quality of service (QoS) requirement per user SINR.

The optimization problem is formulated as $\mathcal{P}\mathbf{1}$:

$$\mathcal{P}\mathbf{1}: \quad \underset{\boldsymbol{\Theta},\mathbf{F}_{DB},\mathbf{F}_{AB}}{\text{minimize}} \quad P_{sum}(\boldsymbol{\Theta}, \mathbf{F}_{DB}, \mathbf{F}_{AB}) \tag{16}$$

$$\text{s.t.} \quad \frac{|(\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{BU,k}^H)\mathbf{f}_{B,k}|^2}{\sum_{j\neq k}^K |(\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}BU, k^H)\mathbf{f}_{B,j}|^2 + \sigma_{z_k}^2} \ge \tilde{\gamma}_k, \ \forall k, \tag{16a}$$

$$\boldsymbol{\Theta} = \text{diag}([e^{j\theta_1}, \dots, e^{j\theta_N}]), \tag{16b}$$

$$\left|[\mathbf{F}_{AB}]_{i,j}\right| = \frac{1}{\sqrt{M}}, \forall 1 \le i \le M, \forall 1 \le j \le Card(\mathcal{K}), \tag{16c}$$

where $\tilde{\gamma}_k$ is the required SINR of user k. The second and third constraints in (16b) and (16c) are uni-modulus constraints for phase-shifter implementation of the passive reflective beamformer and the active analog beamformer, respectively. Although the objective function of ($\mathcal{P}1$) is convex, but the coupling of optimization variables in constraint (16a) and unit-modulus constraints in (16b) and (16c) make the optimization problem non-convex. There is no standard method for solving this type of optimization problem. We transform the above problem into the homogeneous quadratically constrained quadratic program (QCQP) problem [32] and then apply semidefinite relaxation on the rank-one constraint. The reformulated convex relaxed SDP problem can be solved using standard CVX and AO tools [33] as shown in the following subsection IV-A2.

A. PASSIVE REFLECTIVE BEAMFORMING

In IRS, there is no RF chain, and beamforming is only controlled by the microcontroller to assist with the reflection angle change of each IRS element. This type of beamforming is called passive beamforming. The passive beamforming diagonal matrix $\boldsymbol{\Theta} \in \mathbb{C}^{N\times N}$ and passive beamforming vector $\boldsymbol{\xi} \in \mathbb{C}^{N\times 1}$ are related as $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\xi}) \in \mathbb{C}^{N\times N}$ and $\boldsymbol{\xi} = \mathrm{diag}(\boldsymbol{\Theta}) \in \mathbb{C}^{N\times 1}$, assuming the reflection coefficient $\kappa = 1$.

The passive beamforming of the UPA can be decoupled into horizontal and vertical ULA beamforming. This results in two N-point DFT codebooks, $\mathcal{C}_I^{(1)}$ and $\mathcal{C}_I^{(2)}$ for horizontal and vertical beamforming of the IRS, respectively. Once we have the horizontal (vertical) IRS azimuth (elevation) angles of departures, The IRS transmit array response can be expressed as the Kronecker product of array responses of horizontal and vertical beamforming.

$$\begin{aligned}\mathbf{a}_P(\phi_l,\varrho_l) = &\Big[1, e^{j\frac{2\pi d_I}{\lambda}\sin(\vartheta_l)\cos(\psi_{k,l})},\dots\\ &, e^{j\frac{2\pi d_I}{\lambda}\sin(\vartheta_l)\cos(\psi_{k,l})(N-1)}\Big]^T\\ &\otimes\Big[1, e^{j\frac{2\pi d_I}{\lambda}\cos(\vartheta_l)},\dots, e^{j\frac{2\pi d_I}{\lambda}\cos(\vartheta_l)(N-1)}\Big]^T\end{aligned} \tag{17}$$

The above AoDs can be estimated by knowing the geometrical locations of IRSs and users or by exhaustive search of the optimal beams for each user.

1) CLOSED-FORM HEURISTIC-BASED IRS PASSIVE BEAMFORMING

One of our objectives is to find the phase shifts $\theta_{l,n}$ between AoA and AoD to maximize the SINR of each user. To understand the impact of θ, consider the end-to-end channel of user k

$$\mathbf{h}_k = \mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{d,k}^H \tag{18}$$

To obtain the maximum channel gain for user k, we use the triangular inequality as

$$|(\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{d,k}^H)\mathbf{f}_{B,k}| \le |\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H\mathbf{f}_{B,k}| + |\mathbf{h}_{d,k}^H\mathbf{f}_{B,k}| \tag{19}$$

The channel gain of the user k is maximum when $\arg(\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H\mathbf{f}_{B,k}) = \arg(\mathbf{h}_{d,k}^H\mathbf{f}_{B,k})$. Thus, the optimal phase shift of n^{th} IRS element is given by

$$\theta_n^* = \arg(\mathbf{h}_{d,k}^H\mathbf{f}_{B,k}) - \arg(h_{IU,n,k}^H\mathbf{f}_{B,k}) - \arg(\mathbf{h}_{BI}^H\mathbf{f}_{B,k}) \tag{20}$$

where $h_{IU,n,k}^H$ is the n^{th} element of $\mathbf{h}_{IU,k}^H$ and $\mathbf{h}_{BI}^H$ is the n^{th} row of the channel matrix $\mathbf{H}_{BI}^H$ between BS and IRS.

Unlike optimization-based (SDR, Majorization-Minimization (MM), Manifold optimization, etc.) IRS phase shift designs, this method has a closed-form update rule. The simple closed-form update fits well in alternating optimization frameworks. In AO-based joint optimization of IRS phase shift and beamforming at BS with ZF, the ZF removes inter-user interference and this heuristic method ensures that the IRS constructively enhances each user's received signal. However, if users have correlated channels, IRS phase shifts may not fully decouple users' signals.

2) WEIGHTED SUM OF THE CHANNEL GAIN BASED IRS PASSIVE BEAMFORMING

It can be seen from the constraint (16a) of $\mathcal{P}\mathbf{1}$ that the optimal value of per user transmit power can be given as

$$p_k^* = \frac{(\sum_{j\neq k}^K |(\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{BU,k}^H)\mathbf{f}_{B,j}|^2 + \sigma_{zk}^2)\tilde{\gamma}_k}{|(\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{BU,k}^H)\mathbf{f}_{B,k}|^2} \tag{21}$$

i.e., maximization of channel gain of the user k minimizes the transmit power to achieve the required QoS. The precoder for user k can be written as $\mathbf{f}_{B,k} = p_k\hat{\mathbf{f}}_{B,k}$ where $\hat{\mathbf{f}}_{B,k}$ is the direction or unit vector of the user k precoder $\mathbf{f}_{B,k}$. The optimization problem is formulated as $\mathcal{P}\mathbf{2}$:

$$\mathcal{P}\mathbf{2}: \quad \underset{\boldsymbol{\Theta}}{\text{maximize}} \quad \sum_{k=1}^K w_k\|\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H + \mathbf{h}_{BU,k}^H\|^2 \tag{22}$$

$$\text{s.t.} \quad 0 \le \theta_n \le 2\pi, \quad n = 1,\dots,N, \tag{22a}$$

where w_k is the weight of user k channel gain. Using the change of variable $\mathbf{h}_{IU,k}^H\boldsymbol{\Theta}\mathbf{H}_{BI}^H = \boldsymbol{\xi}^H\boldsymbol{\Phi}_k$ where $\boldsymbol{\Phi}_k = \mathrm{diag}(\mathbf{h}_{IU,k}^H)\mathbf{H}_{BI}^H \in \mathbb{C}^{N\times M}$, the objective function term can be written as

$$\begin{aligned}\|\boldsymbol{\xi}^H\boldsymbol{\Phi}_k + \mathbf{h}_{BU,k}^H\|^2 &= (\boldsymbol{\xi}^H\boldsymbol{\Phi}_k + \mathbf{h}_{BU,k}^H)(\boldsymbol{\xi}^H\boldsymbol{\Phi}_k + \mathbf{h}_{BU,k}^H)^H\\ &= (\boldsymbol{\xi}^H\boldsymbol{\Phi}_k + \mathbf{h}_{BU,k}^H)(\boldsymbol{\Phi}_k^H\boldsymbol{\xi} + \mathbf{h}_{BU,k})\\ &= \boldsymbol{\xi}^H\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^H\boldsymbol{\xi} + \boldsymbol{\xi}^H\boldsymbol{\Phi}_k\mathbf{h}_{BU,k}\\ &\quad + \mathbf{h}_{BU,k}^H\boldsymbol{\Phi}_k^H\boldsymbol{\xi} + \|\mathbf{h}_{BU,k}^H\|^2\end{aligned} \tag{23}$$

Thus, the problem (23) is equivalent to the following QCQP problem

$$\begin{aligned}\mathcal{P}\mathbf{3}:&\\ \max_{\boldsymbol{\xi}} &\sum_{k=1}^K w_k(\boldsymbol{\xi}^H\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^H\boldsymbol{\xi}\\ &+ \boldsymbol{\xi}^H\boldsymbol{\Phi}_k\mathbf{h}_{BU,k} + \mathbf{h}_{BU,k}^H\boldsymbol{\Phi}_k^H\boldsymbol{\xi} + \|\mathbf{h}_{BU,k}^H\|^2)\end{aligned} \tag{24}$$

$$\text{s.t.} \quad |\xi_n| = 1, \quad n = 1,\dots,N, \tag{24a}$$

This is a non-convex QCQP problem because of the unit-modulus constraint. It can be transformed to homogeneous QCQP [34]. Define an auxiliary variable $\mathbf{q} \in \mathbb{C}^{N+1}$

$$\mathbf{q} = \begin{bmatrix}\boldsymbol{\xi}\\ t\end{bmatrix} \tag{25}$$

where $t \in \mathbb{C}$ and $|t| = 1$. With this variable, the above problem can be written as

$$\mathcal{P}4: \quad \max_{\mathbf{q}} \sum_{k=1}^{K} w_k(\mathbf{q}^H \bar{\mathbf{\Phi}}_k \mathbf{q} + \|\mathbf{h}_{d,k}^H\|^2) \tag{26}$$

$$\text{s.t. } |q_n| = 1, \quad n = 1, \ldots, N, \tag{26a}$$

where

$$\bar{\mathbf{\Phi}}_k = \begin{bmatrix} \mathbf{\Phi}_k \mathbf{\Phi}_k^H & \mathbf{\Phi}_k \mathbf{h}_{d,k} \\ \mathbf{h}_{d,k}^H \mathbf{\Phi}_k^H & \mathbf{0} \end{bmatrix} \tag{27}$$

In the above problem, the constraint is still non-convex. It can be shown that $\mathbf{q}^H \bar{\mathbf{\Phi}}_k \mathbf{q} = trace(\bar{\mathbf{\Phi}}_k \mathbf{q}\mathbf{q}^H)$. Now $\mathbf{Q} \triangleq \mathbf{q}\mathbf{q}^H$ is a rank one matrix with $\mathbf{Q} \succcurlyeq \mathbf{0}$. The rank-one can be relaxed using semidefinite relaxation to $\mathbf{Q} \succcurlyeq \mathbf{0}$ (positive semidefinite) which makes the resulting problem a convex problem as

$$\mathcal{P}5: \quad \max_{\mathbf{Q}} \sum_{k=1}^{K} w_k(\mathrm{Tr}(\bar{\mathbf{\Phi}}_k \mathbf{Q}) + \|\mathbf{h}_{d,k}^H\|^2) \tag{28}$$

$$\text{s.t.} \quad \mathbf{Q}_{n,n} = 1, \quad n = 1, \ldots, N+1, \tag{28a}$$

$$\mathbf{Q} \succcurlyeq \mathbf{0}. \tag{28b}$$

This is a convex relaxed SDP optimization problem which can be solved with any standard convex optimization tool like CVX. In general, the relaxed SDP does not provide a guaranteed rank-one solution. In case $rank(\mathbf{Q}) \neq 1$, the optimal objective value implies that the upper bound of the problem. We need post-processing of the obtained higher-rank solution to get the rank-one solution. We use the randomization method [34] to obtain the rank-one solution. It has been shown that a sufficiently large number of randomizations guarantees at least a $\pi/4$ -approximation of the optimal objective value [34].

Having knowledge of the CSI of BS-IRS, IRS-k, and the reflection phases, we next design the hybrid beamforming ($\mathbf{F}_{AB}$ and $\mathbf{F}_{DB}$) in BS. Due to the involvement of RF chains and power amplifier, this is also called active beamforming.

B. ACTIVE HYBRID BEAMFORMING

In hybrid beamforming architectures, the beamforming is divided into low-dimensional digital beamforming and analog beamforming. The analog precoder plays a critical role in shaping transmission beams using a limited number of RF chains. Analog beamforming is typically implemented using phase shifters, necessitating unit-modulus elements in the analog beamforming matrix. The efficient design of the analog beamformer is crucial in IRS-assisted systems, where the effective channel is a combination of direct and IRS-reflected links.

The equivalent channel between BS and users can be expressed as

$$\mathbf{H}_{eq} = \mathbf{H}_{BI} \mathbf{\Theta} \mathbf{H}_{IU} + \mathbf{H}_d \in \mathbb{C}^{M \times K}, \tag{29}$$

where $\mathbf{H}_{IU}$ is given by (15).

1) ANALOG BEAMFORMING

The analog beamforming uses DFT codebook of predefined analog beams. The best DFT beam is selected on the basis of the effective channel between BS and user. The BS has M transmit antennas; therefore, we use M-point DFT codebook $\mathcal{C}_B$ at the BS. The codebook divides the spatial domain into the beam domain with M orthogonal beams set $\mathcal{B}$. The beam domain channel between the BS and the users is given as

$$\mathbf{H}_{eq,b} = \mathbf{U}_M^H \mathbf{H}_{eq} \quad \in \mathbb{C}^{M \times K} \tag{30}$$

where $\mathbf{U}_M$ is a unitary M-point DFT matrix. The beam domain channel can be written in terms of users' channel vectors as $\mathbf{H}_{eq,b} = [\mathbf{h}_{eq,b,1}, \mathbf{h}_{eq,b,2}, \ldots, \mathbf{h}_{eq,b,K}]$

After optimization of IRS phase shifters, the received signal can be obtained using the equivalent channel. This signal can be decomposed in the desired signal, inter-beam interference, and the AWGN noise as follows:

$$y_k = \underbrace{\mathbf{h}_{eq,b,k}^H \mathbf{B} \mathbf{f}_{DB} s_k}_{\text{desired signal}} + \underbrace{\mathbf{h}_{eq,b,k}^H \mathbf{B} \sum_{j \neq k}^{Card(\mathcal{K})} \mathbf{f}_{DB,j} s_j}_{\text{interference}} + \underbrace{\tilde{z}_k}_{\text{noise}}. \tag{31}$$

where $\mathbf{B}$ is the beam selection matrix whose elements $b_{i,j} \in \{0, 1\}$. The SINR of the user k can be expressed as

$$\gamma_k = \frac{|\mathbf{h}_{eq,b,k}^H \mathbf{B} \mathbf{f}_{DB,k}|^2}{\sum_{j \in \mathcal{K} \setminus k} |\mathbf{h}_{eq,b,k}^H \mathbf{B} \mathbf{f}_{DB,j}|^2 + \sigma_{\tilde{z}_k}^2}. \tag{32}$$

We aim to design an analog precoder to minimize the transmit power while achieving the required signal-to-noise and interference ratio at each user.

$$P_{sum} = \sum_{k=1}^{Card(\mathcal{K})} p_k, \tag{33}$$

where $p_k = \|\mathbf{f}_{B,k}\|^2$ is the power of the precoder for user k. The optimization problem is given as $\mathcal{P}6$:

$$\mathcal{P}6: \quad \underset{\mathbf{F}_{AB}}{\text{minimize}} \quad P_{sum}(\mathbf{\Theta}, \mathbf{F}_{AB}) \tag{34}$$

$$\text{s.t.} \quad \frac{|\mathbf{h}_{eq,b,k}^H \mathbf{B} \mathbf{f}_{DB,k}|^2}{\sum_{j \in \mathcal{K} \setminus k} |\mathbf{h}_{eq,b,k}^H \mathbf{B} \mathbf{f}_{DB,j}|^2 + \sigma_{\tilde{z}_k}^2} \geq \tilde{\gamma}_k, \quad \forall k, \tag{34a}$$

$$\sum_{i=1}^{M} b_{i,j} = 1 \quad \forall j, \tag{34b}$$

$$\sum_{j=1}^{K} b_{i,j} \leq 1 \quad \forall i, \tag{34c}$$

$$b_{i,j} \in \{0, 1\} \quad \forall i, j, \tag{34d}$$

where the constraint (34a) ensures that one RF chain is allocated to one beam and the constraint (34c) guarantees that each beam has been chosen for at most one user. The above problem (34) is a non-convex optimization problem due to the discrete beam selection matrix **B**. We propose a model-based DRL solution for this challenging optimization problem.

After the beam-users selection (described in Section V), we get $Card(\mathcal{K})$ number of beams such that $\tilde{\mathbf{H}}_{eq} = \mathbf{H}_{eq,b}(b,:)|_{b\in\mathcal{B}} \in \mathbb{C}^{K\times K}$. This also gives us the DFT-based analog precoder implementation as $\mathbf{F}_{AB} = \mathbf{U}(:,b)|_{b\in\mathcal{B}}$ of size $M \times K$.

2) DIGITAL BEAMFORMING DESIGN

After optimization of IRS phase shifters and analog beamforming, similar to Equation (8), the received signal can be obtained using the equivalent channel. This signal can be decomposed into the desired signal, inter-beam interference, and the AWGN noise as follows:

$$y_k = \underbrace{\tilde{\mathbf{h}}^H_{eq,k}\mathbf{f}_{DB}s_k}_{\text{desired signal}} + \underbrace{\tilde{\mathbf{h}}^H_{eq,k}\sum_{j\neq k}^{Card(\mathcal{K})}\mathbf{f}_{DB,j}s_j}_{\text{interference}} + \underbrace{\tilde{z}_k}_{\text{noise}}. \tag{35}$$

The SINR of user k can be expressed as

$$\gamma_k = \frac{|\tilde{\mathbf{h}}^H_{eq,k}\mathbf{f}_{DB,k}|^2}{\sum_{j\in\mathcal{K}\setminus k}|\tilde{\mathbf{h}}^H_{eq,k}\mathbf{f}_{DB,j}|^2 + \sigma^2_{\tilde{z}_k}}. \tag{36}$$

We aim to design a digital precoder to minimize transmit power while achieving the required signal-to-noise and interference ratio at each user. The optimization problem is given as $\mathcal{P}7$:

$$\mathcal{P}7: \quad \underset{\mathbf{F}_{DB}}{\text{minimize}} \quad P_{sum}(\boldsymbol{\Theta}, \mathbf{F}_{DB}, \mathbf{F}_{AB}) \tag{37}$$

$$\text{s.t.} \quad \frac{|\tilde{\mathbf{h}}^H_{eq,k}\mathbf{f}_{DB,k}|^2}{\sum_{j\in\mathcal{K}\setminus k}|\tilde{\mathbf{h}}^H_{eq,k}\mathbf{f}_{DB,j}|^2 + \sigma^2_{\tilde{z}_k}} \geq \tilde{\gamma}_k, \quad \forall k, \tag{37a}$$

3) REGULARIZED ZERO-FORCING DIGITAL BEAMFORMING

The regularized-ZF (RZF) or minimum mean square error (MMSE) is a linear precoding technique used to manage multiuser interference without noise amplification. The RZF digital precoder is obtained by the pseudo-inverse of $\tilde{\mathbf{H}}_{eq}$ as $\mathbf{F}_{DB} = \tilde{\mathbf{H}}_{eq}(\tilde{\mathbf{H}}^H_{eq}\tilde{\mathbf{H}}_{eq} + \alpha\mathbf{I})^{-1}$, where α is a regularization parameter depends on the noise power and total transmit power. The precoder can be represented as a stack of $Card(\mathcal{K})$ digital precoders for users, $\mathbf{F}_{DB} = [\mathbf{f}_{DB,1}, \ldots, \mathbf{f}_{DB,K}]$. To meet the total power constraint, the normalized digital precoder is given as

$$\tilde{\mathbf{f}}_{DB,k} = \frac{\mathbf{f}_{DB,k}}{\|\mathbf{f}_{DB,k}\|}, \forall k \in \mathcal{K}. \tag{38}$$

4) SDR-BASED DIGITAL BEAMFORMING

After fixing IRS phase shifts and analog beamforming, the optimization problem $\mathcal{P}8$ can be formulated for digital beamforming as

$$\mathcal{P}8: \quad \min_{\mathbf{F}_{DB,k}} \sum_{k=1}^{K} \mathrm{Tr}(\mathbf{F}_{DB,k}) \tag{39}$$

$$\text{s.t.} \quad \mathrm{Tr}(\tilde{\mathbf{H}}_{eq,k}\mathbf{F}_{DB,k}) \geq \tilde{\gamma}_k\left(\sum_{j\neq k}\mathrm{Tr}(\tilde{\mathbf{H}}_{eq,k}\mathbf{F}_{DB,j} + \sigma^2_{\tilde{z}_k})\right), \ \forall k \tag{39a}$$

$$\mathbf{F}_{DB,k} \succcurlyeq \mathbf{0}, \mathrm{rank}(\mathbf{F}_{DB,k}) = 1 \tag{39b}$$

where $\mathbf{F}_{DB,k} = \mathbf{f}_{DB,k}\mathbf{f}^H_{DB,k}$ is a rank-one matrix, similarly, $\tilde{\mathbf{H}}_{eq,k} = \tilde{\mathbf{h}}^H_{eq,k}\tilde{\mathbf{h}}_{eq,k}$. The above problem is now a SDP convex optimization problem and can be solved using standard convex optimization tools like CVX.

V. KNOWLEDGE-BASED DEEP REINFORCEMENT LEARNING FOR BEAM SELECTION

This section explores the role of symbolic and algorithmic knowledge in complementing data-driven learning within the proposed DRL-based hybrid beamforming framework. Symbolic knowledge refers to explicit, human-understandable rules and constraints, such as unit-modulus conditions for phase shifters, beam-user assignment policies, and SINR thresholds that define the operational boundaries of the system. Algorithmic knowledge, on the other hand, encompasses mathematical models and optimization techniques, such as SDR, ZF precoding, and DFT-based codebooks, that provide efficient solutions to subproblems within the beamforming design.

By embedding this prior knowledge into the DRL agent's environment and training process, the system benefits from improved learning efficiency, reduced exploration overhead, and enhanced performance in terms of transmit power, energy efficiency, and spectral efficiency. This hybrid approach exemplifies the principles of knowledge-based DRL, where structured insights guide the agent toward more informed and effective decision-making. Reinforcement learning is a branch of machine learning that does not require a large labeled dataset. It iteratively sets the best output depending upon its interaction with the environment and the reward from the environment.

Formally, reinforcement learning has two main components, an *agent* and an *environment*. When a state s_t is given at the input of an agent, it takes action a_t on the environment, and, in return, the environment produces a reward r_t and the next state s_{t+1} as shown in Fig. 2. The objective of reinforcement learning is to iteratively devise a policy to maximize the expected cumulative reward in the long-term period. The policy could be an action function ($a_t = f(s_t)$) implemented with a neural network (NN) and it is called an *Actor*. The RL algorithm that updated this NN is called the policy-based RL algorithm. The policy could be a value function (Q-function, $Q = f(s_t, a_t)$) implemented with NN and it is called a *Critic*. The RL algorithm that updated this NN is called the value-based RL algorithm. The third type

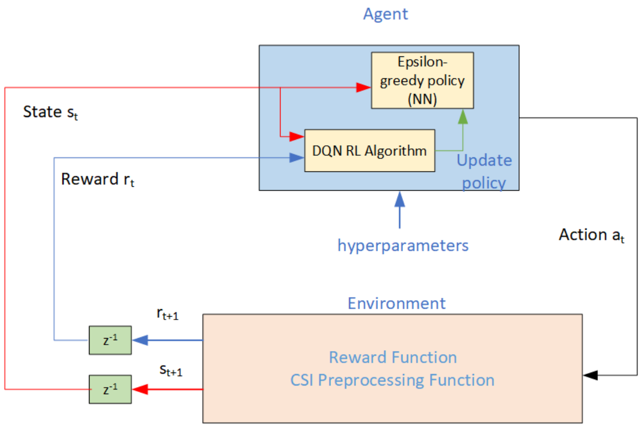

FIGURE 2. **Reinforcement learning block diagram.**

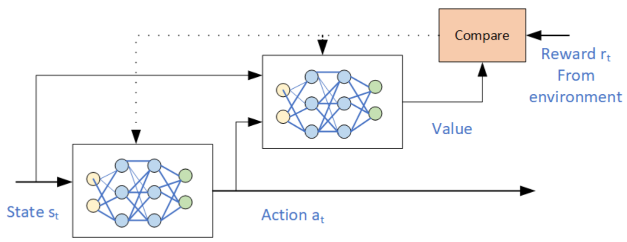

FIGURE 3. **Actor-critic neural networks based agent.**

of policy consists of actor-critic NNs. The actor takes state as input and produces an action, whereas critic takes state and action as input and produces a Q-value. This Q-value is compared with the reward produced by the environment to update the weights of NN as shown in Fig. 3.

In this paper, we use the deep Q-network (DQN) which uses an agent that consists only of the critic DNN. A critic takes state and action at the input and returns the predicted discounted value of the long-term reward. The DQN agent actions are based on Bellman's equation [35].

$$Q^{new}(s_t, a_t) = Q^{old}(s_t, a_t) + \alpha \overbrace{(\underbrace{r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})}_{\text{target value}} - Q^{old}(s_t, a_t))}^{\text{error}}, \tag{40}$$

where α is the learning rate ($0 < \alpha \leq 1$) and $\gamma \in [0, 1]$ is the discount factor. When s_{t+1} reaches the last or terminal state, an episode of the training algorithm ends.

The DQN agent takes the states and actions coarse information (type and range) and then repeatedly generates the actions to maximize the reward. The information tuples (s_t, a_t, r_t, s_{t+1}) are kept in the memory of the experience buffer for further training of the agent with the future expected rewards. The DQN networks implement the Q-value function, and they work better for discrete action space. Fig. 4 shows a typical implementation of a DQN with a Q-network and a target Q-network. The target Q-network is just a clone of the Q-network and is used to stabilize the training process [35]. Instead of training the DQN with one experience, the agent samples a group of batches of a random size from the replay memory for batch training of the Q-network. The weights Θ_Q^- of the target network are updated periodically with the weights Θ_Q of the Q-network according to a predefined hyperparameter 'update frequency'. Mathematically, the loss function is given by:

$$L(\Theta) = [y - Q(s_t, a_t; \Theta)]^2 \tag{41}$$

where y represents the output (target value) of the target DQN, which given by:

$$y = r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \Theta^-) \tag{42}$$

During training, the DQN explores the action space with a given exploration probability epsilon (ϵ). At each time step, either it selects a random action with probability (ϵ) or follows the value function to determine the action with probability ($1 - \epsilon$). The weights of DQN are updated after every mini-batch sample. These mini-batches are taken randomly from the experience buffer.

A. STATE SPACE AND OBSERVATION

The state space is derived from the channel state information $\mathbf{H} \in \mathbb{C}^{N\times K}$. The state space $\mathcal{S} \in \mathbb{R}^{2\times M\times K} \cup \mathbb{R}^{M\times 1} \cup \mathbb{R}^{M\times 1}$ consists of i) $M \times K$ absolute values of the channel vector, ii) $M \times K$ matrix containing angles of channel elements, iii) $M \times 1$ binary selection vector $\boldsymbol{\chi}$, and iv) $M \times 1$ vector containing all actions taken by the agent in terms of the selected beam numbers, including the repeated beams. The binary selection vector is initialized to all ones. The binary element of the n^{th} row is zero if the beam n is selected for the user k. At any time t, the state is represented by a $M \times 4$ tensor.

$$\mathbf{s}_t = [absH(:, selectedUser), angleH(:, selectedUser), binarySelectionVector, selectedBeam] \in \mathbb{R}^{M\times 4} \tag{43}$$

B. ACTION

The action space $\mathcal{A}$ is discrete and consists of $N \times 1$ beams tensor, $\mathcal{A} = \{1, 2, \ldots, N\}$. At any time t, only one beam is selected by the agent. In an episode, there are a total of $card(\mathcal{K})$ time steps, so a total of $card(\mathcal{K})$ beams are selected out of N beams. Once we have trained the policy $\pi^*(s)$, action can be found as

$$a_t = \max_a Q_{\pi^*}(s_t, a; \boldsymbol{\Theta}) \tag{44}$$

C. REWARD

The reward function consists of four components, i) the penalty $\vartheta(t)$ to avoid the same beam selection within an episode, ii) selected beam power, iii) SINR with the selected beam and user at each time step, and iv) the overall sum-rate at last time step of an episode. The reward is equal to the weighted sum of (ii) to (iv) if there is no duplicate beam is selected otherwise the reward is equal to a negative penalty. Mathematically, these functions are written as follows.

$$I(t) = \log_2(1 + |h_{n,k}|^2), \tag{45}$$

and

$$\vartheta(t) = w[(a_{t-1} == a_{t-2}) \vee (a_{t-1} == a_{t-3}) \vee \dots \vee (a_{t-1} == a_{t-K})]. \tag{46}$$

where w is the weight parameter for the penalty. Thus, the total reward is given by

$$r_t = I(t) - \vartheta(t). \tag{47}$$

DRL Agent Hyperparameters for Beam Selection:

1) EXPERIENCE REPLAY BUFFER

During DQN training, we calculate the target value as in (40) and the loss function as in (41). These expressions require information tuple (s_t, a_t, s_{t+1}, r_t). This information is stored in the experience buffer. The value of the buffer size is given in table 2. The agent takes a random mini-batch (table 2) from the experience buffer for an episode of training.

2) EPSILON GREEDY EXPLORATION ($0 \leq \epsilon \leq 1$)

In order to incorporate a suitable trade-off between exploration and exploitation of the action space, we use the epsilon greedy hyperparameter ($0 \leq \epsilon \leq 1$). The agent opts exploration (i.e., random selection of the action) with probability ϵ (table 2) or exploitation (i.e., select a greedy action by using (44)) with probability $1-\epsilon$. The greedy action is the action for which the value-function has a maximum value.

3) DISCOUNT FACTOR ($0 \leq \gamma \leq 1$)

The agent's goal is to maximize the expected cumulative reward within an episode. The expected reward at time t is the reward at time t and all future rewards up to the terminal time step. Since the future reward has less weight compared to the present reward, therefore, the agent tries to select actions so that the sum of discounted rewards in the future is maximized [35],

$$Q_t = r_t + \sum_{i=1}^{N_m} \gamma^i Q_{t+i}, \tag{48}$$

where γ is the discount factor and N_m is the number of time steps from the present time to the terminal time step. If $\gamma = 0$, the agent is concerned with maximizing the immediate reward irrespective of future rewards. We use $\gamma = 0.8$ to put more emphasis on the final accumulated reward.

4) LEARNING RATE ($0 < \alpha < 1$)

The learning rate (α) is used to control the step size during the learning of the neural network weights. A too small value of α increases the training time, and a too large value results in a suboptimal trained network. In our case, we use $\alpha = 0.01$.

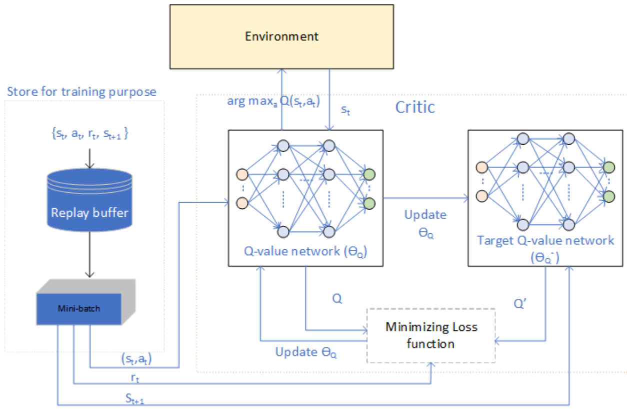

FIGURE 4. **Deep Q-network with target Q-network.**

Adam optimizer uses adaptive moments, and the effective learning rate for each parameter is scaled by the history of gradients.

The algorithm 1 outlines the hybrid beamforming strategy that combines passive beamforming in the IRS and DRL for analog beam selection at the BS. The algorithm begins with the input of channel matrices between BS–IRS and IRS–users (line 1), followed by initialization of the replay memory and neural network parameters (line 2). The effective cascaded channel is computed (line 3), and the DRL agent is trained over multiple episodes (lines 4 to 19). Within each episode, the agent selects beams for users based on the Q-values (line 8), calculates rewards (lines 9–19), and updates its policy using experience tuples (line 15). After training, the selected beams are used to compute the reduced-dimension equivalent channel (line 20), and the digital precoder is derived using regularized zero-forcing (RZF) (line 21). The final output includes both analog and digital precoders (line 22).

The computational complexity of the algorithm is primarily influenced by two components: the SDR-based passive beamforming and the DRL-based beam selection. The SDR step involves solving a semidefinite program (SDP), which has a worst-case complexity of approximately $\mathcal{O}(n^6)$ for an SDP with matrix size n, though practical solvers like CVX often perform better. The DRL component, specifically the DQN, has a complexity that scales with the number of episodes, batch size, and neural network parameters, typically $\mathcal{O}(E.B.P)$ where $E$is the number of episodes, B is the batch size, and P is the number of parameters in the network. Overall, the algorithm balances optimization accuracy with learning-based adaptability, making it suitable for real-time IRS-assisted MIMO systems.

VI. SIMULATION RESULTS

This section evaluates the performance of the proposed integrated SPB-DHB scheme using MATLAB. The proposed scheme is compared with hybrid beamforming designs using Greedy-based [36], Stable Matching-based [37] and

Algorithm 1 DQN-Based Hybrid Precoding in IRS-Assisted Massive MIMO System

Input: $\mathbf{H}_{BI} \in \mathbb{C}^{M \times N}$, $\mathbf{H}_{IU} \in \mathbb{C}^{N \times K}$,
Initialization: Replay memory, batch size, neural network parameters.
$\mathbf{H}_{eq} = |\mathbf{H}_{IU}||\mathbf{H}_{BI}|$
while do
  for *episode* $= 1$ to *maxEpisodes* **do**
    Generate $\mathbf{H}_{BI}$, $\mathbf{H}_{IU}$, $\boldsymbol{\chi} = \mathbf{1}^{M \times K}$
    for $t = 1$ to K **do**
      Generate action $a_t = \max_{a} Q_{\pi^*}(s_t, a; \boldsymbol{\Theta})$
      Calculate the reward
      if (*duplicate_penality* $== 0$) **then**
        $reward = selectedBeamPower + sumRate + (20 * SINR_k)$
      else
        $reward = -duplicate\_penality$
      end if
      Input reward along with the next state s_{t+1} to the RL agent
    end for
  end for
  Until: Reaches to *maxEpisodes* or convergent
end while
{**Active Precoder Design**}
$\tilde{\mathbf{H}}_{eq} = \mathbf{H}_{eq}(selectedBeams, :)$
$\mathbf{F_{DB}} = \tilde{\mathbf{H}}_{eq}\left(\tilde{\mathbf{H}}_{eq}^{H}\tilde{\mathbf{H}}_{eq} + \mathbf{I}_{Card(\mathcal{K})}\right)^{-1}$
Output: $\mathbf{F}_{AB}$, $\mathbf{F}_{DB}$

TABLE 2. List of hyperparameters.

Hyperparameter	Value
Optimizer	Adam
Learning rate α	0.01
Discount factor γ	0.8
Exploration decay	0.001
Experience replay buffer	20,000
Mini batch size	64

Hungarian-based [38] beam selection. We also compare with the DRL-based active analog precoder and passive IRS beamformer scheme [39].

The simulation scenario consists of a single cell with $K = 8$ users. The BS is equipped with $M = 32$ antennas and $N_{RF} = 8$ RF chains. The IRS consists of $N = 64$ reflective elements. We use Adam optimizer which is the standard optimizer for DQN implementations because of its stability, adaptiveness, and fast learning characteristics [40]. In the simulation, the results are obtained by averaging over 50 channel realizations.

Fig. 5 shows the Simulink diagram for a Deep Reinforcement Learning-based beam selection system for hybrid beamforming in a IRS-assisted massive MIMO system. The agent learns to select the best beam per user based on the CSI to maximize the reward. In each episode, CSI block generates CSI using (30) and (29). The MATLAB function block perUserCSI (preprocessing block) takes CSI, selected beam, and the selected user as input and generates the output $H_{beam} = [absH(:, selectedUser), angleH(:, selectedUser), selectionBeamVector, selectedBeam] \in \mathbb{R}^{M \times 4}$. The Reward function block takes the reshaped action and CSI as input and generates a scalar reward *reward_out* that contains both an immediate reward after each time step and a total reward at the end of the episode. The output *user* ensures that the unique users are eliminated by eliminating the user corresponding to the duplicate beam selected by the agent. *BeamVector* concatenates all selected beams till time step t. When the user variable reaches K, then *flag_isDone* $=$ *True* to terminate the episode.

Fig. 6 shows DQN training progress. The DQN consists of an input layer, a convolution layer, and then three alternating ReLU (Rectifier Linear Unit) and fully connected layers. It can be seen that the training process converges after 300 episodes. The light blue graph line is for the episode reward and the dark blue line is showing the average reward over 50 time steps. The orange graph is the estimate of long-term reward at the start of each episode, given the initial observation of the environment. One episode contains maximum of 15 time steps to process $K = 8$ users. We validate the trained DQN agent with the same training environment for 10 episodes. The DQN testing result is shown in Fig. 7. It shows a consistent performance of trained agent with the majority of the episodes' rewards within the standard deviation range. This consistent performance shows that the trained DQN is successfully avoiding the duplicate beam selection which is the hardest part of the RL agent training.

Exploration versus exploitation is adaptively controlled by the epsilon decay as shown in Fig. 8. In the start of the simulation, epsilon takes its initial value and then, at each interaction with the environment if Epsilon is greater than EpsilonMin, then it is updated using this formula.

$$\epsilon = \epsilon(1 - \epsilon\text{-decay}) \tag{49}$$

Epsilon is preserved throughout the transition from the conclusion of one episode to the commencement of the subsequent episode. Consequently, epsilon decreases uniformly across multiple episodes until it attains the value of epsilon minimum. As shown in Fig. 8, we set the initial epsilon to 1, epsilon decay is 0.001, and epsilon minimum is 0.01.

Beam-user selection can be visualized in Fig. 9. This figure shows the temporal behavior of the beam selection per user (BeamUser) and the corresponding reward received by the DQN agent during an episode. In Simulink, one time step is set to 0.025 second. The upper graph shows the beam selection with time steps (also the user index), e.g., beam number 5 is selected for user 1, beam number 1 is selected for user 2, and so on. The lower graph depicts the immediate reward after each time step within the episode. Because we use the sorted channel (based on the beam power), we get the lower beam index selection during the initial time steps.

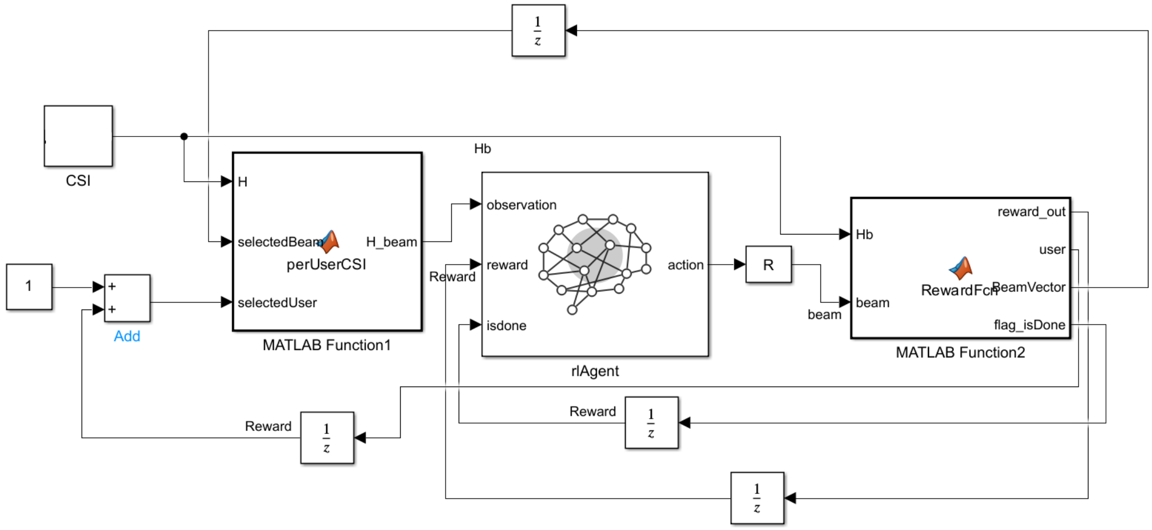

FIGURE 5. DRL agent and environment design.

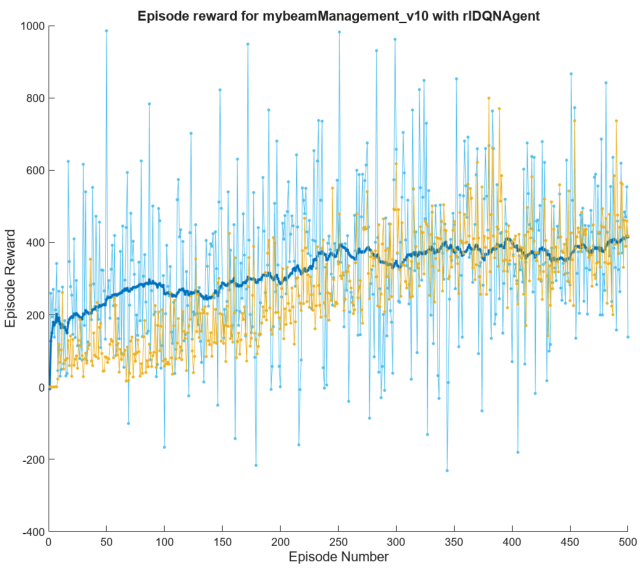

FIGURE 6. DQN agent training.

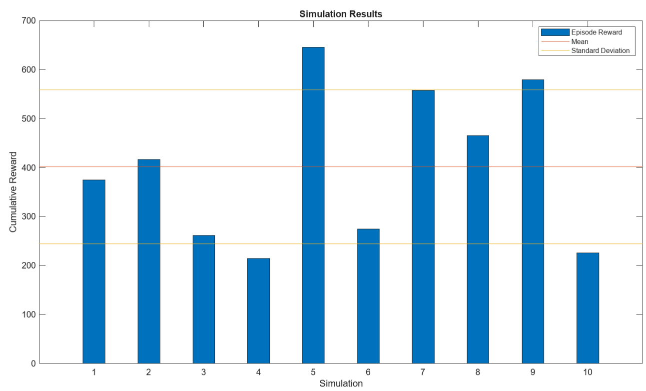

FIGURE 7. Testing of the trained RL agent.

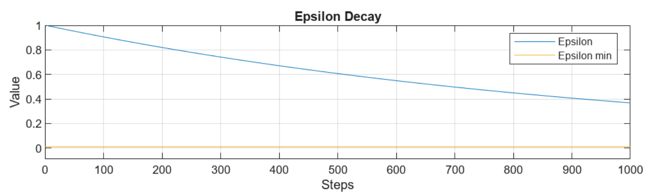

FIGURE 8. Epsilon decay.

The Fig. 10 illustrates the relationship between transmit power (in dBm) and target SNR (in dB) for various beam selection and user association algorithms in an IRS-assisted or massive MIMO communication system. As expected, all algorithms show an increasing trend in transmit power with rising target SNR values, since higher SNR requirements naturally demand more power. Among the evaluated methods, the RL-SDP-based approach demonstrates the most efficient performance, consistently requiring the least transmit power across the entire SNR range. This suggests that combining reinforcement learning with convex optimization techniques like semi-definite programming results in superior power minimization. The standalone RL-based algorithm also performs effectively, outperforming heuristic methods such as the Greedy and Hungarian-based algorithms, and coming close to the Stable Matching-based method. While the Stable Matching-based approach yields better results than Greedy and Hungarian methods, it is still less optimal than RL-based solutions. The Hungarian algorithm, despite being mathematically optimal in terms of assignment, performs the worst in power efficiency because it does not account for transmit power optimization. Overall, the RL-SDP-based method is the most power-efficient, followed closely

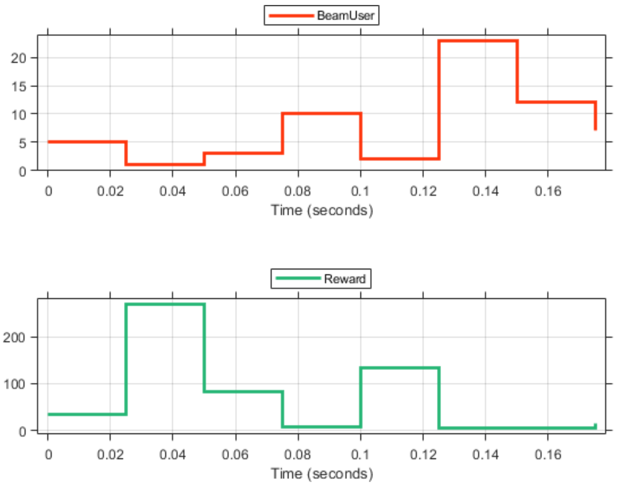

FIGURE 9. Beam-user selection within an episode during simulation.

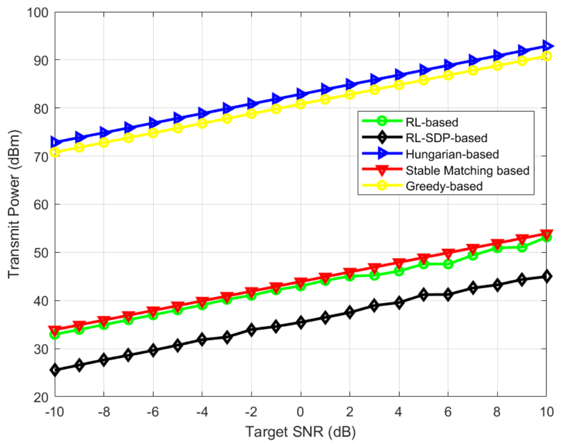

FIGURE 10. Transmit power versus target SNR with Greedy-based [36], Stable Matching-based [37], Hungarian-based [38], and DRL-based [39] beamforming for $K = 8$, $M = 32$ and $N = 64$.

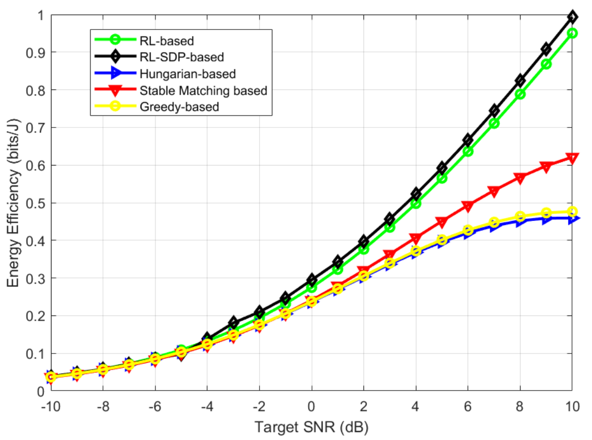

FIGURE 11. Energy efficiency versus target SNR with Greedy-based [36], Stable Matching-based [37], Hungarian-based [38], and DRL-based [39] beamforming for $K = 8$, $M = 32$ and $N = 64$.

by the RL-based agent, making them suitable candidates for energy-constrained or performance-critical wireless

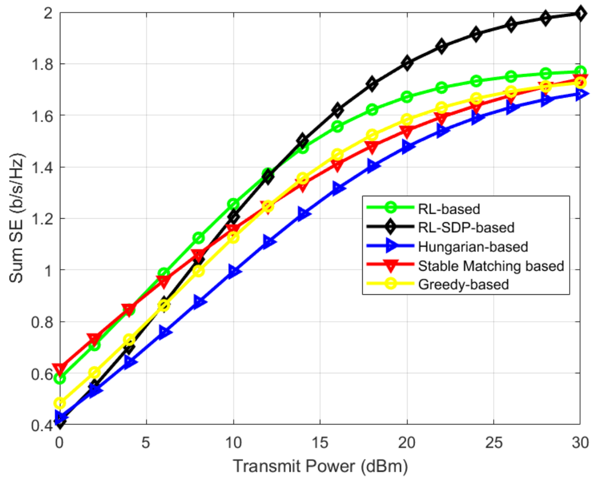

FIGURE 12. Sum-SE versus transmit power with Greedy-based [36], Stable Matching-based [37], Hungarian-based [38], and DRL-based [39] beamforming for $K = 8$, $M = 32$ and $N = 64$.

communication systems. The proposed RL-SDP-based scheme reduces the required transmit power to 20.37% at 10 dB SINR compared to Model-free DRL scheme.

Fig. 11 illustrates the energy efficiency in bits per Joule. Generally, the energy efficiency of all schemes tends to increase with the target SNR. This indicates that higher signal strength relative to noise allows for more efficient energy utilization in transmitting information. However, the Reinforcement Learning-based algorithms (RL-based and RL-SDP-based) consistently achieve the highest energy efficiency across most of the SNR range. It demonstrates superior performance compared to the Hungarian-based, Stable Matching based, and Greedy-based approaches. Notably, the Hungarian-based and Stable Matching based algorithms show a plateauing effect at higher SNR values, suggesting a limitation in their ability to further improve energy efficiency as the signal quality increases, while the Greedy-based algorithm exhibits the lowest overall energy efficiency. In particular, RL-SDP-based scheme outperforms Stable matching and greedy-based schemes by 1.67 and 2.1 times, respectively, at target SNR of 10 dB.

Fig. 12 depicts the sum SE with the transmit power. As the transmit power increases, the Sum SE also improves across all schemes, though a saturation effect is observed beyond 25 dBm, where the performance gains begin to plateau. Among the strategies compared, the RL-SDP-based approach achieves the highest Sum SE across the entire transmit power range, demonstrating its superior ability to optimize system performance by leveraging reinforcement learning in combination with semi-definite programming. The RL-based method follows closely, showing strong performance and outperforming traditional heuristic methods. The Greedy-based and Stable Matching-based approaches offer moderate performance, with the Greedy approach slightly edging out the Stable Matching method, particularly in the mid transmit power range. The Hungarian-based method yields the lowest Sum SE, indicating limited effectiveness in this scenario. Specifically, at 20 dBm transmit power, the RL-SDP-based scheme provides 13.92%, 16.88%, and 22.44% better sum-SE performance compared to the Greedy, Stable Matching, and Hungarian-based active and passive beamforming design methods, respectively.

VII. CONCLUSION

This paper studied a joint SDR-based passive beamforming and DRL-based beam selection and active hybrid beamforming for an IRS-assisted massive MIMO system. In the RL-SDR-based BF, we use the cascade channel BS-IRS-users for beam-users selections. The optimal IRS phase shifts are determined using semidefinite relaxation that cancel out the phase shifts of the BS-IRS and IRS-users channels. After the DRL-SDR-based beam-user selection, at BS a DFT-based analog precoder is obtained to mitigate the inter-beam interference. Finally, a ZF-based reduced dimension digital precoder is designed from the selected beam-based cascaded channel. It has been shown that the proposed RL-SDR-based beamforming outperforms the only RL-based, Greedy-based, Stable matching-based, and Hungarian-based beamforming schemes.

ACKNOWLEDGMENT

An earlier version of this paper was presented in part at the IEEE 97th Vehicular Technology Conference (VTC-Spring), in 2023 [DOI: 10.1109/VTC2023-Spring57618.2023.10200585].

REFERENCES

[1] I. Ahmed, K. Shahid, and H. Khammari, "DRL based beam selection and hybrid beamforming for intelligent reflective surface assisted massive MIMO system," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2023, pp. 1–6.

[2] Z. Chen, B. Ning, C. Han, Z. Tian, and S. Li, "Intelligent reflecting surface assisted terahertz communications toward 6G," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 110–117, Dec. 2021.

[3] Y. Sun, B. Duan, X. Su, H. Wang, Q. Gu, J. Jin, and Y. Yuan, "Performance analysis on reconfigurable intelligent surface and network-controlled repeater in 3GPP release-18," *Frontiers Inf. Technol. Electron. Eng.*, vol. 24, no. 12, pp. 1815–1828. [Online]. Available: https://link.springer.com/article/10.1631/FITEE.2300321

[4] C. Pradhan, A. Li, L. Song, B. Vucetic, and Y. Li, "Hybrid precoding design for reconfigurable intelligent surface aided mmWave communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 7, pp. 1041–1045, Jul. 2020.

[5] Z. Wang, L. Liu, S. Zhang, and S. Cui, "Massive MIMO communication with intelligent reflecting surface," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2566–2582, Apr. 2023.

[6] X. Zhang, X. Yu, S. H. Song, and K. B. Letaief, "IRS-aided MIMO systems over double-scattering channels: Impact of channel rank deficiency," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 2076–2081.

[7] T. Otsuki and G. Gui, "Deep learning aided intelligent sensing and identification for 6G," *ACM Comput. Surv.*, vol. 57, no. 3, pp. 1–37, 2024.

[8] B. Zheng, C. You, W. Mei, and R. Zhang, "A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications," *IEEE Commun. Surv. Tut.*, vol. 24, no. 2, pp. 1035–1071, 2nd Quart., 2022.

[9] S. H. Hong, J. Park, S.-J. Kim, and J. Choi, "Hybrid beamforming for intelligent reflecting surface aided millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7343–7357, Sep. 2022.

[10] X. Song, T. Kuhne, and G. Caire, "Fully-/partially-Connected hybrid beamforming architectures for mmWave MU-MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1754–1769, Mar. 2020.

[11] J. Lv, T. Wang, and S. Wang, "Optimal analog precoder design for hybrid beamforming is possible," *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 9573–9578, Jul. 2023.

[12] S. Gong, C. Xing, P. Yue, L. Zhao, and T. Q. S. Quek, "Hybrid analog and digital beamforming for RIS-assisted mmWave communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1537–1554, Mar. 2023.

[13] Y. S. Ribeiro, A. L. F. de Almeida, F.-E-Asim, B. Makki, and G. Fodor, "Low-complexity joint active and passive beamforming design for IRS-assisted MIMO," *IEEE Wireless Commun. Lett.*, vol. 13, no. 3, pp. 607–611, Mar. 2024.

[14] R. Zhang and Y. Wang, "Joint transceiver and reflecting beamforming design for IRS-aided mmWave MIMO systems with discrete phase shifters," *IEEE Trans. Veh. Technol.*, vol. 74, no. 5, pp. 8401–8406, May 2025.

[15] W. Xiong, J. Lin, Z. Xiao, and Q. Li, "Constant-modulus secure analog beamforming for an IRS-assisted communication system with large-scale antenna array," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 2957–2969, 2025.

[16] A. Ahmadinejad and S. Talebi, "Beamforming design via machine learning in intelligent reflecting surface-aided wireless communication," *Phys. Commun.*, vol. 68, Feb. 2025, Art. no. 102586, doi: 10.1016/j.phycom.2024.102586.

[17] S. F. Kimaryo, E. E. Bahingayi, and K. Lee, "Joint active/passive beamforming optimization via RWMMSE and gradient projection for downlink IRS-aided MU-MISO systems," *IEEE Access*, vol. 13, pp. 8779–8789, 2025.

[18] Y. Ju, S. Gong, H. Liu, C. Xing, J. An, and Y. Li, "Beamforming optimization for hybrid active-passive RIS assisted wireless communications: A rate-maximization perspective," *IEEE Trans. Commun.*, vol. 72, no. 9, pp. 5428–5442, Sep. 2024.

[19] X. Yu, C. Yang, J. Bai, K. Wang, Y. Rui, and X. Dang, "Joint design of power allocation and beamforming for IRS-assisted millimeter-wave communication system with imperfect CSI," *IEEE Trans. Mobile Comput.*, vol. 24, no. 7, pp. 6566–6582, Jul. 2025.

[20] L. Ma, X. Zhang, J. Sun, W. Zhang, and C.-X. Wang, "Joint optimization of reconfigurable intelligent surfaces and base station beamforming in MISO system based on deep reinforcement learning," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2023, pp. 1–5.

[21] J.-C. Chen, "Energy-efficient hybrid beamforming design for intelligent reflecting surface-assisted mmWave massive MU-MISO systems," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 1, pp. 330–344, Mar. 2024.

[22] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Deep learning for large intelligent surfaces in millimeter wave and massive MIMO systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[23] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, "Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.

[24] Ö. Özdogan and E. Björnson, "Deep learning-based phase reconfiguration for intelligent reflecting surfaces," in *Proc. 54th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2020, pp. 707–711.

[25] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, and M. Debbah, "Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663–1677, Jun. 2021.

[26] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, May 2020.

[27] Z. Wan, Z. Gao, and M.-S. Alouini, "Broadband channel estimation for intelligent reflecting surface aided mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[28] B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface assisted multi-user OFDMA: Channel estimation and training design," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8315–8329, Dec. 2020.

[29] B. Guo, R. Li, and M. Tao, "Joint design of hybrid beamforming and phase shifts in RIS-aided mmWave communication systems," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2404–2416, Apr. 2021.

[30] M. Soleimani, R. C. Elliott, W. A. Krzymien, J. Melzer, and P. Mousavi, "Hybrid beamforming for mmWave massive MIMO systems employing DFT-assisted user clustering," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11646–11658, Oct. 2020.

[31] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013. [Online]. Available: https://ieeexplore.ieee.org

[32] V. S. Mai, D. Maity, B. Ramasubramanian, and M. C. Rotkowitz, "Convex methods for rank-constrained optimization problems," in *Proc. SIAM Conf. Control Appl.*, 2015, pp. 123–130.

[33] M. Chaudhary, D. Kumari, and S. Kumar Bandari, "Joint active and passive beamforming in IRS-assisted wireless networks over generalized fading model," in *Proc. 2nd Int. Conf. Vis. Towards Emerg. Trends Commun. Netw. Technol. (ViTECoN)*, May 2023, pp. 1–6.

[34] A. M.-C. So, J. Zhang, and Y. Ye, "On approximating complex quadratic optimization problems via semidefinite programming relaxations," *Math. Program.*, vol. 110, no. 1, pp. 93–110, Mar. 2007.

[35] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., Cambridge, MA, USA: MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[36] R. Pal, A. K. Chaitanya, and K. V. Srinivas, "Low-complexity beam selection algorithms for millimeter wave beamspace MIMO systems," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 768–771, Apr. 2019.

[37] Z. Cheng, Z. Wei, and H. Yang, "Low-complexity joint user and beam selection for beamspace mmWave MIMO systems," *IEEE Commun. Lett.*, vol. 24, no. 9, pp. 2065–2069, Sep. 2020.

[38] I. Ahmed, M. K. Shahid, H. Khammari, and M. Masud, "Machine learning based beam selection with low complexity hybrid beamforming design for 5G massive MIMO systems," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 4, pp. 2160–2173, Dec. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9467318/

[39] Y. Zhu, Z. Bo, M. Li, Y. Liu, Q. Liu, Z. Chang, and Y. Hu, "Deep reinforcement learning based joint active and passive beamforming design for RIS-assisted MISO systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 477–482.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

IRFAN AHMED (Senior Member, IEEE) received the B.E. degree in electrical engineering and the M.S. degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 1999 and 2003, respectively, and the Ph.D. degree in telecommunication engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2008. He is currently an Associate Professor with the Higher Colleges of Technology, United Arab Emirates. Previously, he was an Assistant/Associate Professor with Taif University, from 2011 to 2017. He was a Postdoctoral Fellow with Qatar University, from April 2010 to March 2011, where he worked on two research projects, such as wireless mesh networks with Purdue University, USA; and radio resource allocation for LTE with Qtel. He was involved in the National ICT Pakistan funded Research Project "Design and Development of MIMO and Cooperative MIMO Test-Bed" with Iqra University, Islamabad, Pakistan, from 2008 to 2010. He is an Active Reviewer of IEEE, Springer, and Elsevier journals and conferences. His research interests include wireless LAN (WLAN) medium access control (MAC) protocol design and analysis, 5G communications, mmWave and massive MIMO communications, performance analysis of wireless channels, energy-constrained wireless networks, and radio resource allocation. He served as the Session Chair for the IEEE Wireless Communications, Networking and Mobile Computing conferences. He is an Associate Editor of IEEE ACCESS journal.

KHALIL SHAHID received the B.E. degree in electrical engineering from the University of Engineering and Technology, Lahore, in 1998, the M.S. degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2003, and the Ph.D. degree in engineering management from Beijing University of Posts and Telecommunications, Beijing, China, in 2008. He worked on LTE MiFi clouds, hotspots, Wingles, USB dongles, and drive testing for CDMA/EVDO network for checking QoS parameters using NEMO Analyzer and Genex Probe. He is currently an Assistant Professor with the Higher Colleges of Technology. He has over 15 years of experience in telecommunication industry. He is an author of several journals and conference papers in the field of communications and information technology. His current research interests include wireless communication, 5G communications, optical wireless communications, fiber optic systems and networks, optical transmission, optical fiber access networks, technology management, operational management, project management, and industrial organization.

MOATH AWAWDEH (Senior Member, IEEE) received the Ph.D. degree in control engineering from the University of Genoa, Italy. He is currently an Assistant Professor of electrical engineering with the Higher Colleges of Technology (HCT), United Arab Emirates. He has more than 15 years of combined academic and industry experience, with expertise in applied research, control systems, digital teaching and learning, curriculum development, and academic leadership. His career includes serving as the Division Chair and the Applied Research Coordinator, contributing to over $1M in funded projects across United Arab Emirates, Italy, Bahrain, Saudi Arabia, and Jordan. He has played a central role in aligning engineering programs with ABET accreditation standards and has implemented successful strategies to support at-risk students through targeted interventions and mentoring. He has been co-chaired the international ASET conferences, since 2022 and co-founded Ai-Enabled Bahrain Consultancy, where he serves on the Board of Advisors. His research and professional interests include strengthening academia–industry collaboration, innovation-driven engineering education, and real-world problem-solving. He is a Senior Fellow of the Higher Education Academy (HEA).

ADNAN SHAHID (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Sejong University, South Korea, in 2015. He is currently an Assistant Professor with the Internet Technology and Data Science Laboratory (IDLab), Ghent University, and IMEC, where he leads the "AI/ML for Wireless" Research within IDLab-Intelligent Wireless Networking (iWINe). He is an active contributor to several working groups, including the IEEE WG-P1900.8 Standard for Training, Testing, and Evaluating Machine-Learned Spectrum Awareness Models and the ATIS WG on Generative AI in Telecom. He has participated in numerous challenging projects, such as the DARPA Spectrum Collaboration Challenge (SC2), European H2020, 6G SNS, and ESA Projects. He leads several European and national projects (imec ICON and FWO). His research interests include wireless physical layer foundation models, decentralized learning, radio resource management, the Internet of Things, 5G/6G networks, and indoor localization.

MAMOUN GUENACH (Senior Member, IEEE) received the engineering degree from École Mohammadia d'Ingénieurs, Morocco, in 1995, and the Ph.D. degree from UCLouvain, Belgium, in 2002. He was a Postdoctoral Researcher with Ghent University, from 2002 to 2006. From 2006 to 2019, he was with Nokia Bell Laboratories. Since 2015, he has been a Visiting Research Professor with Ghent University. Since 2019, he has been a Research Scientist with IMEC, Leuven, Belgium. His research interests include coding, modulation, synchronization, and equalization for high-speed wired and wireless communications, including distributed massive MIMO.

HEDI KHAMMARI received the B.Eng. and M.Eng. degrees in electrical engineering (with a specialization in automatic control) and the Ph.D. degree in electrical engineering from Tunis University, Tunis, in 1988, 1990, and 1999, respectively. He is currently a Professor with the Institut Supérieur des Sciences Appliquées et de Technologie de Kairouan, Tunisia.

TARIG FAISAL (Senior Member, IEEE) received the bachelor's degree in electronics engineering technology (medical electronics) from the University of Gezira, the master's degree in mechatronics engineering from International Islamic University Malaysia, and the Ph.D. degree in biomedical signal processing from the University of Malaya, Malaysia. He is currently an Associate Professor of electrical engineering with the Higher Colleges of Technology (HCT), United Arab Emirates. He has more than 20 years of academic and professional experience spanning teaching, research, and academic leadership. He has served in several senior roles with HCT, including the Dean of academic operations, the Acting Director of Madinat Zayed and Ruwais campuses, the Engineering Division Chair, and an Academic Coordinator. His leadership portfolio includes overseeing academic and administrative operations, chairing quality and accreditation committees, leading ABET and CAA program accreditation activities, and establishing engineering laboratories and student chapters. He has secured multiple internal and external grants and contributed as the principal investigator and co-investigator to national and international funded projects. He has authored numerous journal and conference publications in reputable venues, including IEEE, Springer, and Elsevier, and has served as a reviewer for many international journals and conferences. His research interests include biomedical signal processing, intelligent diagnostic systems, mechatronics, control systems, the IoT-based sensing, and machine-learning-enabled engineering applications.