Machine Learning-Enabled RIS-Assisted MmWave Multi-Hop Communications: Path Scheduling and Beamforming Design

Tongyi Wei, *Graduate Student Member, IEEE,* Beixiong Zheng, *Senior Member, IEEE,* Kun Tang, *Member, IEEE,* Wenjie Feng, *Senior Member, IEEE,* Wenquan Che, *Fellow, IEEE,* and Quan Xue, *Fellow, IEEE*

***Abstract*—In this paper, a reconfigurable intelligent surface (RIS)-assisted millimeter-wave (mmWave) multi-hop communication system is considered, where the base station (BS) transmits signals to remote users via multi-hop paths formed by multiple RISs. We aim to maximize the sum achievable rate by jointly optimizing multi-hop path scheduling, active beamforming, and passive beamforming, while satisfying the maximum transmit power of the BS. To tackle the formulated non-convex joint optimization problem, a machine learning (ML)-based two-stage algorithm is proposed. In the first stage, a multi-hop path scheduling algorithm based on graph neural networks (GNN) is investigated. Specifically, the considered system is first modeled as a graph topology, where the BS, RIS, and users are all served as nodes. Then, the weights between adjacent nodes associated with the path gain and the number of RIS reflection units are defined based on the expressions of RIS-assisted equivalent channels. Finally, a GNN-based multi-hop path scheduling algorithm is proposed. In the second stage, according to the obtained optimal multi-hop path, the deep deterministic policy gradient (DDPG) strategy is adopted to optimize the active beamforming of the BS and the passive beamforming of the selected RISs. Simulation results demonstrate that the proposed GNN-based multi-hop path scheduling algorithm maintains an error within 3% compared to the global optimal solution and outperforms the graph-based methods. Additionally, the proposed two-stage algorithm improves the sum achievable rate by approximately 27.6% compared to the alternating optimization (AO) algorithm.**

***Index Terms*—Reconfigurable intelligent surface (RIS), millimeter-wave (mmWave), multi-hop path scheduling, machine learning (ML), joint optimization.**

I. Introduction

With the rapid advancement of wireless communication technology, millimeter-wave (mmWave) communications with recognized for its abundant bandwidth resources, has emerged as a pivotal technology for satisfying future high-speed and high-capacity requirements. Nevertheless, mmWave communications encounter critical challenges in practical deployment, including substantial path-loss, signal blockage, and limited propagation distance [1]. Several strategies have been proposed to address these issues. For instance, increasing deployment density of base station (BS) can partially relieve path-loss and signal blockage, while the optimized beamforming enhances transmission performance by concentrating signal energy [2]. In addition, massive multiple-input multiple-output (MIMO) systems leverage multi-antenna arrays to expand coverage and improve signal stability, which can effectively counter path-loss [3]. However, these solutions involve high costs and intricate system designs [4].

Fortunately, the recently proposed reconfigurable intelligent surface (RIS) offers a promising solution to tackle the limited propagation distance of mmWave. RIS is a metasurface composed of numerous passive and low-cost reflective units that can dynamically adjust the propagation characteristics of electromagnetic waves [5]. Compared with traditional methods, RIS offers advantages such as low cost, low power consumption, and flexible deployment [6]. Therefore, in complex communication environments, such as urban areas, long-distance mmWave communication can be effectively implemented through a multi-hop manner with assistance of multiple RISs layed on the exterior walls of buildings. Several studies have shown that, compared to single-RIS systems, multi-RIS systems can significantly reduce multiplicative path loss and enhance system performance through the optimization of path scheduling and beamforming design [7]–[9]. However, multi-hop path scheduling based on multiple RISs presents a complex combinatorial optimization problem, as it involves the interconnection of multiple nodes and the selection of paths, resulting in a rapid increase in the dimensionality of the solution space. Traditional methods require exploring numerous potential path combinations to identify the optimal solution, which brings a higher computational overhead. Recently, machine learning (ML) has made significant advancements in wireless communications, and its powerful learning capability and adaptivity enable it to show an excellent performance in coping with complex optimization problems [10]. Consequently, we investigate the efficient use of ML algorithms to handle the problem of optimal path scheduling and beamforming design in RIS-assisted multi-hop mmWave communications.

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2024ZYGXZR057, Grant 2023ZYGXZR106, and Grant 2024ZYGXZR087, in part by the National Natural Science Foundation of China under Grant 62231014, Grant 62321002, Grant 62201214, and Grant 62331022, in part by the the Natural Science Foundation of Guangdong Province under Grant 2023A1515011753 and Grant 2024A1515011726. (*Corresponding authors: Kun Tang*.)

T. Wei, K. Tang, W. Feng, W. Che, and Q. Xue are with the Guangdong Provincial Key Laboratory of Millimeter-Wave and Terahertz, Guangdong-Hong Kong-Macao Joint Laboratory for Millimeter-Wave and Terahertz, School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China. (e-mail: eeweitongyi@mail.scut.edu.cn; tangkun@scut.edu.cn; fengwenjie1985@163.com; eewqche@scut.edu.cn; eeqxue@scut.edu.cn).

B. Zheng is with the School of Microelectronics, South China University of Technology, Guangzhou, 511442, China. (e-mail: bxzheng@scut.edu.cn).

A. Related Works

1) Single-RIS-Assisted Communication Systems: With the aid of RIS, both spectrum efficiency and energy efficiency of wireless communication systems have witnessed significant improvement. In [11], the authors designed an RIS-assisted wireless network by optimizing the passive beamforming of RIS to maximize the spectral efficiency. To improve the sum achievable rate of the RIS-assisted communication system, the successive convex approximation (SCA) and greedy algorithms was proposed in [12] to jointly optimize the active beamforming of the BS and the passive beamforming of RIS. Furthermore, the authors of [13] investigated the impact of RIS phase shift accuracy on the energy efficiency of wireless communication systems, which showed that optimal energy efficiency can be achieved with a phase shift accuracy of 1 or 2 bits. Recently, the effective use of RIS has emerged as promising solutions for enhancing mmWave communications. In [14], an alternating optimization (AO)-based algorithm was developed to maximize the sum achievable rate of RIS-assisted mmWave systems by jointly optimizing BS beamforming, discrete reflection coefficients of RIS, and user scheduling strategies. In [15], a distributed RIS-assisted massive MIMO system was studied, where the joint design of RIS phase shifts and beamforming was employed to improve the transmission rate of mmWave communications in areas with weak coverage. ML can utilize various algorithms and a large number of data samples to effectively solve non-convex optimization problems by continuously iterating towards the global optimal solution. The authors of [16] considered the beamforming design of RIS-assisted communication systems and proposed a gradient-based manifold meta-learning (GMML) method to enhance the overall spectral efficiency. To improve beamforming robustness in mmWave MIMO systems, a robust gradient-based liquid neural network (GLNN) framework was introduced in [17], which utilized liquid neurons based on ordinary differential equations. A beyond-diagonal RIS architecture was proposed in [18], where a deep reinforcement learning (DRL) algorithm was employed to jointly design the RIS phase shift matrix and BS beamforming to enhance the spectral efficiency of mmWave systems.

2) Multi-RIS-Assisted Communication Systems: Single RIS deployments have limited control over wireless channels, especially in complex urban environments, where a blockage-free reflection link between the BS and distant users may not exist with just one RIS. To address this issue, recent studies have focused on designing efficient systems by deploying more RISs within the networks. In [19], a multi RIS-assisted multi-cell communication system was considered to maximize the average capacity by jointly optimizing the BS-RIS-user association and the RIS deployment location. The authors of [20] investigated an uplink multi-cell non-orthogonal multiple access (NOMA) network with multi-RIS collaboration, where the transmit power was minimized by introducing an inter-group interference cancellation (IGIC)-based NOMA scheme and jointly designing RIS phase shift matrix and transmit power. The authors of [21] focused on the hybrid beamforming design for multi-RIS-assisted terahertz systems and proposed a DRL-based algorithm to jointly optimize digital and analog beamforming vectors. In [22], the authors considered a robust hybrid beamforming design for multi-RIS-assisted mmWave-MIMO systems with imperfect channel state information (CSI), where an AO-based method was proposed to jointly optimize BS beamforming and RIS reflection coefficients. Although the aforementioned multi-RIS transmission schemes significantly enhance the transmission performance of communication systems, their predefined signal transmission paths neglect the dynamic interactions among multiple RISs. Therefore, these schemes struggle to effectively cope with channel variations and link blockages in complex environments. A multi-hop path scheduling problem of multi-RIS-assisted single-user communication systems was studied in [23], where a graph-based algorithm was proposed to obtain an optimal multi-hop path. The authors extended this study to a multi-user scenario, where a recursive algorithm was adopted to maximize the minimum power of received signal among all users [24]. Furthermore, the authors in [8] considered scattering interference among RISs in multi-hop path scheduling and proposed a strategy that divides users into different activation groups to mitigate interference. To further mitigate path-loss, the authors in [9] studied the multi-hop path scheduling problem of multi-active and multi-passive RIS assisted communication systems, where a graph optimization method was proposed to jointly optimize the multi-hop path scheduling along with active and passive beamforming vectors.

B. Motivation and Contributions

Although previous studies [8], [9], [23], [24] have explored multi-hop path scheduling problem, they have primarily employed graph-based optimization methods. However, such methods are not suitable for RIS-assisted mmWave multi-hop communications with the following reasons. First, RIS-assisted mmWave multi-hop communications typically face complex channel environments and multipath propagation, which leads to a significant increase in the dimensionality of the solution space. Conventional graph-based methods are inefficient in addressing high-dimensional problems, making it challenging to obtain a global optimal solution. Second, the limited transmission distance of mmWave signals requires the deployment of a large number of RISs to enhance the coverage performance, which makes the multi-hop path scheduling of mmWave systems involve the connectivity and path selection of multiple nodes. Traditional graph-based methods require traversing numerous potential path combinations to identify the optimal solution, which may result in insufficient real-time performance in practical communications. Therefore, a more efficient approach is needed to handle the high-dimensional solution space and the complex multi-path propagation characteristics. In this context, graph neural networks (GNNs) have shown significant potential, as they can effectively capture the complex relationships between nodes through the graph structure, making them better suited to address the challenges of the mmWave environment. However, to the best of our knowledge, there has been limited research focused on this issue. Furthermore, for improving the transmission performance of the system, several strongly coupled variables need

to be optimized jointly. Although the AO methods [13]–[15] can obtain suboptimal solutions, the computational overhead will be increased with highly time-varying mmWave channels, which greatly limits their application in delay-sensitive communications. ML algorithms can adaptively respond to the dynamically changing mmWave environments due to their excellent learning and representation capabilities. They can solve non-convex optimization problems with minimal computational complexity through feature learning and model training. Therefore, we utilize ML algorithms to achieve better resource allocation results and reduce computational overhead. The contributions of this paper are shown as follows:

- We consider an RIS-assisted mmWave multi-hop communication system, where the BS transmits signals to remote users through a multi-RIS cascaded multi-hop path. Under the maximum transmission power constraint of the BS, a joint optimization problem for multi-hop path scheduling and corresponding design of beamforming is considered to maximize the sum achievable rate of the system. Furthermore, to more accurately reflect practical communication environments, both the imperfect CSI and a limited number of discrete RIS phase shifts are considered.
- To address the optimization problem, a two-stage ML-based algorithm is proposed. In the first stage, a novel GNN-based multi-hop path scheduling algorithm is investigated. To be specific, we first model the considered system as a graph topology, where the BS, RIS, and users are all serves as nodes. Subsequently, the expression of RIS-assisted equivalent channel is derived. Then, the weights between adjacent nodes are defined associated with path gain and the number of RIS reflecting uints. Finally, the multi-hop path scheduling problem can be transformed into a shortest path problem in graph theory, and an efficient GNN-based method is proposed to determine the optimal multi-hop path. In the second stage, based on the obtained optimal multi-hop path, the joint optimization of BS active beamforming and passive beamforming of the selected RIS can be modeled as a Markov decision process (MDP). Given that the optimization variables involve continuous variables, a DRL algorithm based on deep deterministic policy gradient (DDPG) is employed to obtain optimal beamforming designs.
- Simulation results have confirmed the performance advantages of the proposed two-stage ML-based algorithm. Regarding multi-hop path scheduling optimization, the proposed GNN-based algorithm maintains an error within 3% of the global optimum and outperforms the traditional graph-based methods. Moreover, the sum achievable rate obtained by using the DRL algorithm also surpasses that of other benchmark algorithms. Additionally, compared to the traditional AO algorithm, the proposed ML-based two-stage algorithm improves the sum achievable rate by approximately 27.6%.

The remainder of this paper is organized as follows. Section II describes the system model and formulates the optimization problem of maximizing the sum achievable rate. Section III presents the proposed two-stage ML-based algorithm in detail. Section IV validates the effectiveness of the proposed two-stage algorithm through numerical results. Finally, Section V concludes this paper.

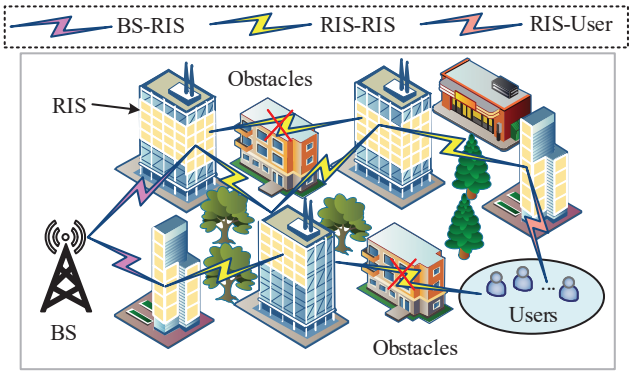

Fig. 1. RIS-assisted mmWave multi-hop communication system.

Notation: Lowercase letters, bold lowercase letters, and bold uppercase letters represent scalars, vectors, and matrices, respectively. $\mathcal{CN}\left(\mu,\sigma^2\right)$ denotes a circularly symmetric complex Gaussian distribution with mean μ and variance σ^2. $\|\cdot\|$ represents the Euclidean norm. $\text{diag}\left(\boldsymbol{a}\right)$ refers to a diagonal matrix whose main diagonal elements are those of the vector $\boldsymbol{a}$. $\text{real}\left\{\cdot\right\}$ and $\text{imag}\left\{\cdot\right\}$ denote the real and imaginary parts of the elements in matrix/vector, respectively. $\boldsymbol{x}^T$ and $\boldsymbol{x}^H$ are the transpose and conjugate transpose of vector $\boldsymbol{x}$, respectively.

II. System Model And Problem Formulation

A. System Model

As illustrated in Fig. 1, a multi-RIS-assisted downlink mmWave communication system is considered, where the direct links between the BS and remote users are blocked by building obstacles. Thus, the BS can only communicate with remote users through reflected multi-hop paths formed by the selected RISs. Without loss of generality, it is assumed that the BS is equipped with N_t antennas, each user is equipped with N_r antennas, and both employ uniform linear array (ULA) configurations. Each RIS, with M reflection elements, is configured in a uniform rectangular array (URA) layout. Let M_x and M_z denote the number of RIS reflection units in the horizontal and vertical directions, respectively, with $M = M_x \times M_z$. Given that long-distance transmission is implemented, it is assumed that each user can share the same multi-hop path to receive its own signal. Therefore, we focus on the design of optimal multi-hop path between the BS and a typical user[1].

Let $\mathcal{J} = \{1, 2, \cdots, J\}$ and $\mathcal{K} = \{1, 2, \cdots, K\}$ represent the sets of RISs and users, respectively. To facilitate graph representation, we define the nodes set as $\tilde{\mathcal{J}} =$

[1]The obtained results can be extended to scenarios with varying user requirements and spatial distributions by independently designing multi-hop paths for each user. However, it will inevitably introduce inter-user/path interference, where inter-path interference can be modeled by using conflict graphs [25] and users can be divided into different groups to alleviate interference [8], which will be explored in our future work.

$\{0, \mathcal{J}, \cdots, J+1\}$, where nodes 0 and $J+1$ stand for the BS and a typical user, respectively. We introduce a binary state indicator $l(i,j) \in \{0,1\}$ to represent the communication status between two nodes. Specifically, $l(i,j) = 1$ indicates that the node i and node j can communicate with each other; otherwise, $l(i,j) = 0$. In addition, we have $l(i,i) = 0, \forall i$, and $l(i,j) = l(j,i), \forall i, j$.

B. Channel Model and Data Transmission

Let $\boldsymbol{H}_{0,j}, j \in \mathcal{J}$, denote the channel matrix between the BS and j-th RIS, $\boldsymbol{S}_{i,j}, i, j \in \mathcal{J}, i \neq j$, represent the channel matrix between the i-th RIS and j-th RIS, and $\boldsymbol{g}_{J,k}^H, k \in \mathcal{K}$, denote the channel vector between the last RIS and k-th user. For the RIS j, let $\boldsymbol{\Phi}_j = \text{diag}\left(\beta_{j,1} e^{j\phi_{j,1}}, \cdots, \beta_{j,M} e^{j\phi_{j,M}}\right)$ represent the phase shift matrix, where $\beta_{j,m}$ and $\phi_{j,m}$ denote the amplitude and phase of the reflection units, respectively. Most existing studies on RIS assumed that each reflection units has infinite-bit resolution phase shifts [15], [18]. However, this idealized assumption is difficult to achieve due to hardware constraints. A common implementation is to restrict the amplitude and sample phase values from a finite set. Consequently, we set $\beta_{j,m} = 1$ to maximize the reflected power, and $\phi_{j,m} \in \mathcal{B}$, where $\mathcal{B} = \left\{\frac{2\pi n}{2^b}, n = 1, \cdots, 2^b\right\}$ with b being the number of quantization bits.

Given the limited scattering characteristics of the mmWave channels, we utilize the Saleh-Valenzuela geometric channel model to characterize the channel [26]. To be specific, if $l(0,j) = 1$, the channel matrix between the BS and RIS j can be given as

$$\boldsymbol{H}_{0,j} = \sqrt{\frac{N_t M}{L_H}} \sum\nolimits_{l=1}^{L_H} \alpha_{l,0,j} \left(\boldsymbol{h}_{l,0,j,r} \boldsymbol{h}_{l,0,j,t}^H\right), \tag{1}$$

where L_H is the number of scattering paths between the BS and RIS j, $\alpha_{l,0,j}$ represents the path gain of the l-th path, $\boldsymbol{h}_{l,0,j,r} = \boldsymbol{\alpha}_R\left(\varphi_{l,0,j}^a, \varphi_{l,0,j}^e, M\right)$ denotes the received array response of RIS j with $\varphi_{l,0,j}^a$ and $\varphi_{l,0,j}^e$ representing the azimuth and elevation angles-of-arrival (AoAs) from the BS to RIS j, respectively, and $\boldsymbol{h}_{l,0,j,t} = \boldsymbol{\alpha}_B(\vartheta_{l,0,j}, N_t)$ represents the transmit array response of the BS with $\vartheta_{l,0,j}$ being the angle-of-departure (AoD) from the BS to RIS j. Moreover, $\boldsymbol{\alpha}_B(\vartheta_{l,0,j}, N_t)$ and $\boldsymbol{\alpha}_R\left(\varphi_{l,0,j}^a, \varphi_{l,0,j}^e, M\right)$ can be calculated as

$$\boldsymbol{\alpha}_B(\vartheta_{l,0,j}, N_t) = \boldsymbol{u}\left(\frac{2d_B}{\lambda} \sin\vartheta_{l,0,j}, N_t\right), \tag{2}$$

and

$$\begin{aligned}\boldsymbol{\alpha}_R\left(\varphi_{l,0,j}^a, \varphi_{l,0,j}^e, M\right) = \boldsymbol{u}\left(\frac{2d_R}{\lambda} \sin\varphi_{l,0,j}^a \cos\varphi_{l,0,j}^e, M_x\right) \\ \otimes \boldsymbol{u}\left(\frac{2d_R}{\lambda} \cos\varphi_{l,0,j}^a, M_z\right),\end{aligned} \tag{3}$$

respectively, where d_B and d_R denote the antenna spacing and the spacing between two adjacent RIS reflecting units, respectively, $\boldsymbol{u}(\zeta, U) = \left[1, e^{-j\pi\zeta}, \cdots, e^{-j(U-1)\pi\zeta}\right]$ is the steering vector with ζ being the phase difference between two adjacent elements in observation, and U being the number of elements in the ULA. Accordingly, if $l(i,j) = 1$, the channel matrix between the RIS i and RIS j can be expressed as

$$\boldsymbol{S}_{i,j} = \sqrt{\frac{M^2}{L_s}} \sum\nolimits_{l=1}^{L_s} \alpha_{l,i,j} \left(\boldsymbol{s}_{l,i,j,r} \boldsymbol{s}_{l,i,j,t}^H\right), \tag{4}$$

where $\boldsymbol{s}_{l,i,j,r} = \boldsymbol{\alpha}_R\left(\varphi_{l,i,j}^a, \varphi_{l,i,j}^e, M\right)$ represents the receive array response of RIS j with $\varphi_{l,i,j}^a$ and $\varphi_{l,i,j}^e$ being the azimuth and elevation AoAs from RIS i to RIS j, respectively, and $\boldsymbol{s}_{l,i,j,t} = \boldsymbol{\alpha}_R\left(\vartheta_{l,i,j}^a, \vartheta_{l,i,j}^e, M\right)$ denotes the transmit array response of RIS i with $\vartheta_{l,i,j}^a$ and $\vartheta_{l,i,j}^e$ being the azimuth and elevation AoDs from RIS i to RIS j, respectively. Similarly, if $l(J, J+1) = 1$, the channel between RIS J and k-th user is modeled as

$$\boldsymbol{g}_{J,k}^H = \sqrt{\frac{MN_r}{L_g}} \sum\nolimits_{l=1}^{L_g} \alpha_{l,J,k} \left(\boldsymbol{g}_{l,J,k,r} \boldsymbol{g}_{l,J,k,t}^H\right), \tag{5}$$

where $\boldsymbol{g}_{l,J,k,r} = \boldsymbol{\alpha}_B(\varphi_{l,J,k}, N_r)$ denotes the receive array response of the k-th user with $\varphi_{l,J,k}$ being the AoA from the RIS J to k-th user, and $\boldsymbol{g}_{l,J,k,t} = \boldsymbol{\alpha}_R\left(\vartheta_{l,J,k}^a, \vartheta_{l,J,k}^e, M\right)$ represents the transmit array response of RIS J with $\vartheta_{l,J,k}^a$ and $\vartheta_{l,J,k}^e$ being the azimuth and elevation AoDs from the RIS J to k-th user, respectively.

Let $\Omega = \{a_1, a_2, \cdots, a_I\}$ signify the multi-hop path scheduling from the BS to users with $a_i, i \in \{1, \cdots, I\}$, and I denoting the index of the RIS and the number of selected RIS, respectively. Then, a multi-hop path scheduling Ω is feasible if and only if the following conditions are satisfied, i.e.,

$$a_i \in \mathcal{J}, a_i \neq a_{i'}, \forall i, i \neq i', \tag{6}$$

$$l(a_i, a_{i+1}) = 1, \forall i, i \neq I, \tag{7}$$

$$l(0, a_1) = l(a_I, J+1) = 1, \tag{8}$$

where (6) ensures that each RIS can reflect a signal at most once, (7) indicates the presence of an LoS path between two cascaded RISs in a multi-hop path scheduling Ω, and (8) guarantees LoS paths from the BS to the first RIS a_1 and from the last RIS a_I to users. According to (1)-(8), given a multi-hop path scheduling Ω, the equivalent channel model between the BS and k-th user is thus given by

$$\boldsymbol{h}_{0,k}(\Omega) = \boldsymbol{g}_{a_I,k}^H \boldsymbol{\Phi}_{a_I} \left(\prod\nolimits_{i=1}^{I-1} \boldsymbol{S}_{a_i,a_{i+1}} \boldsymbol{\Phi}_{a_i}\right) \boldsymbol{H}_{0,a_1}. \tag{9}$$

In considered systems, the BS obtains CSI through the transmission and feedback of pilot signals. Specifically, users first transmit pilot signals, which are reflected by the RIS and received by the BS. Then the BS estimates the BS-RIS-user equivalent channel [8]. However, due to the passive nature of the RIS, accurate CSI is difficult to be acquired at the BS [22]. Thus, we consider the CSI error existed and the equivalent imperfect channel is represented as $\boldsymbol{f}_{0,k}(\Omega) = \boldsymbol{h}_{0,k}(\Omega) + \boldsymbol{e}_{0,k}(\Omega)$, where $\boldsymbol{e}_{0,k}(\Omega)$ denotes the CSI error with independently and identically distributed zero-mean and unit-variance complex Gaussian entries that are independent of $\boldsymbol{h}_{0,k}(\Omega)$. Therefore, the received signal at the k-th user after experiencing multi-hop path transmission is given by

$$y_k = \boldsymbol{f}_{0,k}(\Omega) \boldsymbol{x} + n_k, \tag{10}$$

where $\boldsymbol{x} = \sum_{k\in\mathcal{K}} \boldsymbol{w}_k x_k$ represents the transmitted signal from the BS with $\boldsymbol{w}_k$ and x_k being the active beamforming vector and dedicated signal of the k-th user, respectively, $\mathbb{E}\left\{|x_k|^2\right\} = 1$, and $n_k \sim \mathcal{CN}\left(0, {\sigma_n}^2\right)$ represents the complex additive white Gaussian noise (AWGN) at the k-th user with zero mean and variance of σ_n. Then, the signal-to-interference-plus-noise ratio (SINR) of the k-th user can be expressed as

$$\gamma_k = \frac{\left|\boldsymbol{f}_{0,k}\left(\Omega\right)\boldsymbol{w}_k\right|^2}{\sum_{r\in\mathcal{K},r\neq k}\left|\boldsymbol{f}_{0,k}\left(\Omega\right)\boldsymbol{w}_r\right|^2 + {\sigma_n}^2}. \tag{11}$$

Correspondingly, the achievable rate of the k-th user is given by $R_k = \log\left(1+\gamma_k\right)$, while the sum achievable rate of the considered multi-RIS-assisted multi-hop mmWave communication system can be calculated as $R_{sum} = \sum_{k=1}^{K} R_k$.

C. Problem Formulation

We aim to maximize the sum achievable rate of the considered mmWave communication system by jointly optimizing the multi-hop path scheduling, active beamforming at the BS, and passive beamforming of selected RISs. Therefore, the optimization problem can be formulated as

$$\begin{aligned}
(\text{P1}):\ & \max_{\Omega,\{\boldsymbol{w}_k\},\{\boldsymbol{\Phi}_{a_i}\}} R_{\text{sum}} \\
\text{s.t.}\ & (6)-(8), && \text{(12a)}\\
& R_k \geq R_{\min}, \forall k, && \text{(12b)}\\
& \phi_{a_i,m} \in \mathcal{B}, && \text{(12c)}\\
& \sum\nolimits_{k\in\mathcal{K}} \|\boldsymbol{w}_k\|^2 \leq P_{\max}. && \text{(12d)}
\end{aligned}$$

In the optimization problem (P1), constraint (12a) ensures the feasibility of multi-hop path scheduling Ω, constraint (12b) is the minimum rate constraint for each user, while constraint (12d) limits the maximum transmit power $P_{\max}$ of the BS.

Solving the optimization problem (P1) involves the following challenges. First, there are many alternative multi-hop paths and it is necessary to choose an optimal path with minimum path-loss and maximum sum achievable rate of the system. This is a combinatorial optimization problem with a non-convex solution space, where path selection involves discrete variables and cannot be directly solved using convex optimization techniques. Although the traditional graph-based methods can solve such problem in small-scale networks, they are not suitable for multi-hop communication scenarios. This is because the increase in the number of alternative paths and nodes greatly increases the computational complexity. To address these challenges, a GNN-based approach is proposed, which can automatically learn and optimize the path selection process by capturing the complex dependencies between nodes, thus reduces the dependence on graph structure and path traversal. Different from the graph-based methods, GNN does not require explicit construction and updating of complex graph structures but instead improves the accuracy and efficiency of path selection through efficient feature representation learning. More importantly, GNN can perform fast inference on the graph after the training completed, and select the optimal multi-hop paths with an approximate constant time complexity. Second, the strong coupling of optimization variables and the fractional SINR expression in (P1) results in a non-convex problem. The conventional AO algorithms suffer from high computational overhead and tend to get trapped in local optima. To tackle this, a DRL-based approach is adopted, which models the problem as an MDP and refines the solution through policy iteration for approximating the global optimum. Therefore, we propose a method that integrates GNN and DRL to tackle the problem (P1).

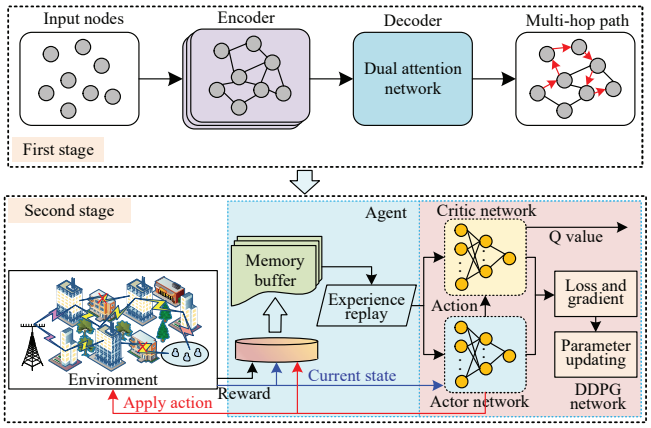

Fig. 2. The overall flowchart of the proposed two-stage algorithm.

III. Proposed Solutions

This section proposes a two-stage algorithm to jointly optimize multi-hop path scheduling, BS active beamforming, and RIS passive beamforming for maximizing the sum achievable rate of the considered system. The overall algorithmic flow is illustrated in Fig. 2.

A. Optimization of Multi-Hop Path Scheduling

In this subsection, we first derive the expression of the RIS-assisted cascaded equivalent channel and analyze its structural characteristics. Then, the weights between any two nodes are defined. Finally, a GNN-based multi-hop path scheduling algorithm is proposed to determine the optimal multi-hop path.

Given a multi-hop path scheduling $\Omega = \{a_1, a_2, \cdots, a_I\}$, the equivalent channel between the BS and k-th user can be obtained by substituting (1)-(5) into (9), which can be reformulated as[2]

$$\boldsymbol{f}_{0,k}\left(\Omega\right) = \sqrt{MN_r}\alpha_{a_I,k}\boldsymbol{g}_{a_I,k,r}\left(\prod_{i=1}^{I}\boldsymbol{G}_i\right)\boldsymbol{h}_{0,a_1,t}^H + \boldsymbol{e}_{0,k}\left(\Omega\right), \tag{13}$$

where

$$\boldsymbol{G}_i = \begin{cases} \sqrt{N_t M}\alpha_{0,a_1}\,\boldsymbol{s}_{a_1,a_2,t}^H\boldsymbol{\Phi}_{a_1}\boldsymbol{h}_{0,a_1,r}, & i=1,\\ M\alpha_{a_{i-1},a_i}\,\boldsymbol{s}_{a_i,a_{i+1},t}^H\boldsymbol{\Phi}_{a_i}\boldsymbol{s}_{a_{i-1},a_i,r}, & 2\leq i\leq I-1,\\ M\alpha_{a_{I-1},a_I}\,\boldsymbol{g}_{a_I,k,t}^H\boldsymbol{\Phi}_{a_I}\boldsymbol{s}_{a_{I-1},a_I,r}, & i=I. \end{cases} \tag{14}$$

[2]To simplify the derivation, we assume that the channel consists of only a single dominant path. However, the proposed method remains applicable to scenarios with multiple paths. Therefore, to enhance notational clarity, we omit the path index l in the subsequent derivations.

For obtaining the optimized multi-hop path, we aim to find a beam path scheduling that maximizes the user's effective channel gain, i.e., $\left|\boldsymbol{f}_{0,k}\left(\Omega\right)\right|^2, \forall k$. According to (13), maximizing $\left|\boldsymbol{f}_{0,k}\left(\Omega\right)\right|^2$ is equivalent to maximize the modulus of $\boldsymbol{G}_i$. In the following, we illustrate the processes of obtaining the maximum modulus of $\boldsymbol{G}_i$ with $2 \leq i \leq I-1$ as example. Specifically, $\left|\boldsymbol{G}_i\right|^2$ can be expressed as

$$\left|\boldsymbol{G}_i\right|^2 = \left| M\alpha_{a_{i-1},a_i} \boldsymbol{s}^H_{a_i,a_{i+1},t} \boldsymbol{\Phi}_{a_i} \boldsymbol{s}_{a_{i-1},a_i,r} \right|^2. \quad (15)$$

To facilitate explanation of (15), we further define $\boldsymbol{s}^H_{a_i,a_{i+1},t} = \text{vec}\left(\left[e^{-j\zeta_t}\right]_{1\times M}\right)$, $\boldsymbol{s}_{a_{i-1},a_i,r} = \text{vec}\left(\left[e^{-j\varsigma_t}\right]_{M\times 1}\right)$, and $\boldsymbol{\Phi}_{a_i} = \text{diag}\left(\left[e^{j\phi_{a_i,t}}\right]_{M\times M}\right)$, where vec$(\cdot)$ and diag$(\cdot)$ stand for vectorization and diagonalization, respectively, ζ_t and ς_t represent the transmission and reception phase shifts of the channel, respectively. Then we can obtain

$$\left|\boldsymbol{G}_i\right|^2 = \left| M\alpha_{a_{i-1},a_i} \sum\nolimits_{t=1}^{M} e^{j\Delta\delta_t} \right|^2, \quad (16)$$

where $\Delta\delta_t = \phi_{a_i,t} - \zeta_t - \varsigma_t$ represents the phase difference. Note that the $\left|\boldsymbol{G}_i\right|^2$ can reach its maximum value when the phase difference terms satisfy $\Delta\delta_1 = \Delta\delta_2 = \cdots = \Delta\delta_M = 0$. Thus, we further have $\left|\boldsymbol{G}_i\right|^2_{max} = M^4\alpha^2_{a_{i-1},a_i}$. Similarly, the same conclusion can be drawn for $i = 1$ and $i = I$.

According to the analyses in above, the individual channel gain between nodes a_i and a_{i+1} can be determined by $\left|\boldsymbol{G}_i\right|^2$. The optimal value of $\left|\boldsymbol{G}_i\right|^2$ is proportional to both the number of reflection units of the RIS and the path gain between two adjacent nodes. To formulate the maximization of the equivalent channel gain as a shortest path problem in graph theory, the channel weight between nodes a_i and a_{i+1} is thus defined as

$$\Gamma_{a_i,a_{i+1}} = \begin{cases} \ln\left(1 + \dfrac{l\left(a_i,a_{i+1}\right)}{M\alpha_{a_i,a_{i+1}}}\right), & \text{if } i \in \mathcal{I}^c, \\ \ln\left(1 + \dfrac{l\left(a_i,a_{i+1}\right)}{N_r\alpha_{a_i,a_{i+1}}}\right), & \text{if } i = I. \end{cases} \quad (17)$$

where $\mathcal{I}^c \triangleq \{0, 1, \ldots, I-1\}$ and $l\left(a_i,a_{i+1}\right)$ represents the blocking factor that dictates the availability of a given path. Specifically, if there is an obstruction between nodes a_i and a_{i+1}, then $l\left(a_i,a_{i+1}\right) = 0$; otherwise, $l\left(a_i,a_{i+1}\right) = 1$. Moreover, $\alpha_{a_i,a_{i+1}} = \tilde{\alpha}_{a_i,a_{i+1}}/\sqrt{PL_{a_i,a_{i+1}}}$ with $PL_{a_i,a_{i+1}}$ representing the path-loss and $\tilde{\alpha}_{a_i,a_{i+1}} \sim \mathcal{CN}\left(0,\sigma_f^2\right)$ denoting the small-scale fading, which follows a complex Gaussian distribution with a mean of zero and a variance of σ_f. It is worth noting that we utilize a logarithmic function involving the path gain and the number of RIS reflective elements to define the path weight. The primary reason for adopting the logarithmic transformation is its ability to linearize the multiplicative nature of path-loss and reflection effects, which is common in wireless communications, where path gain is typically expressed in logarithmic units (e.g., dB). Moreover, the logarithmic form not only helps to smooth out significant variations in path gain but also facilitates optimization scaling, which enhances the stability and efficiency of path selection.

Based on (17), we can model the considered system as a directed weighted graph. Then, the problem of obtaining optimal multi-hop path scheduling is transformed into finding the shortest path in the directed weighted graph. We define a directed weighted graph as $G = (V, E, W)$, where the set of vertex V consists of all nodes in the system, i.e., $V = \tilde{\mathcal{J}}$, the set of edges is denoted as

$$E = \left\{ (i,j) \,\middle|\, l\left(i,j\right) = 1, i,j \in \tilde{\mathcal{J}}, i \neq j \right\}, \quad (18)$$

and the weight function is defined as $W = \{\Gamma_{i,j} \,|\, l\left(i,j\right) = 1\}$. In addition, we denote the feature of node i as $\boldsymbol{n}_i, i \in V$, which represents the three-dimensional coordinates of node i. In this way, we establish a one-to-one correspondence between the multi-hop path from the BS to a typical user and the path from node 0 to node $J+1$ in the graph G.

1) GNN modeling: We propose an GNN-based algorithm to find the optimal path scheduling, which consists of two main components, i.e., an encoder and a decoder. The encoder transforms all nodes and edges into embeddings, which are high-dimensional vector representations encapsulating the features of the nodes and edges. The decoder selects one node at each time step and masks this node after generation to prevent the model from revisiting it in the subsequent steps.

Encoder: The encoder is designed based on the graph attention (GAT) network, which is a neural network architecture that transmits node information through an attention mechanism [27]. However, GAT updates the features of each node solely by assigning new weights to its neighbors without considering the edge information, which may result in the loss of some neighborhood information inherent in the graph. To address this issue, we propose a residual edge-graph attention (RE-GAT) network, which simultaneously considers the information of both the nodes and edges in the graph structure [28]. Additionally, a residual connection is incorporated into each RE-GAT sublayer to prevent gradient vanishing.

The inputs of the encoder consist of the node feature $\boldsymbol{n}_i$ and edge feature $\Gamma_{i,j},\ i,j \in V$. These features first pass through fully connected layers and then embed them into feature spaces of dimensions d_n and d_Γ, denoted as $\hat{\boldsymbol{n}}_i^{(0)}$ and $\hat{\Gamma}_{i,j}$, where the superscript 0 indicates that the features are before entering the RE-GAT. Then, the embedded features are input into the RE-GAT for the fusion of nodes and edges. We use the layer index $l \in \{1, 2, \cdots, L\}$ to represent the node embeddings $\hat{\boldsymbol{n}}^{(l)}$ obtained at the l-th layer in the RE-GAT. The inputs of the first layer of the RE-GAT include the node features $\hat{\boldsymbol{n}}^{(0)} = \left\{\hat{\boldsymbol{n}}_0^{(0)}, \hat{\boldsymbol{n}}_1^{(0)}, \cdots, \hat{\boldsymbol{n}}_{J+1}^{(0)}\right\}$ and edge features $\hat{\Gamma} = \left\{\hat{\Gamma}_{0,1}, \hat{\Gamma}_{1,2}, \cdots, \hat{\Gamma}_{J,J+1}\right\}$ with the output being the new node features $\hat{\boldsymbol{n}}^{(1)} = \left\{\hat{\boldsymbol{n}}_0^{(1)}, \hat{\boldsymbol{n}}_1^{(1)}, \cdots, \hat{\boldsymbol{n}}_{J+1}^{(1)}\right\}$ and the edge features $\hat{\Gamma}_{i,j}$ remaining unchanged. We take the first layer as an example to illustrate the calculation processes of RE-GAT:

- Firstly, the attention coefficient α_{ij} between node i and its neighboring node j can be calculated as

$$\alpha_{ij} = \frac{\exp(\sigma(\boldsymbol{g}^T[\boldsymbol{F}(\hat{\boldsymbol{n}}_i^{(0)} \left\| \hat{\boldsymbol{n}}_j^{(0)} \right\| \hat{\Gamma}_{i,j})]))}{\sum_{r\in\mathcal{N}_i} \exp(\sigma(\boldsymbol{g}^T[\boldsymbol{F}(\hat{\boldsymbol{n}}_i^{(0)} \left\| \hat{\boldsymbol{n}}_r^{(0)} \right\| \hat{\Gamma}_{i,r})]))}, \quad (19)$$

where $\mathcal{N}_i$ denotes the set of neighboring nodes of node i, $\boldsymbol{g}$ and $\boldsymbol{F}$ represent the learnable parameter vector and parameter matrix, respectively, $\sigma(\cdot)$ stands for the LeakyReLU activation function, and $(\cdot\|\cdot)$ indicates the concatenation operation.

- Secondly, the information of each node can be updated based on $\hat{\boldsymbol{n}}_{i,R}^{(1)} = \sum_{j\in\boldsymbol{N}_i} \alpha_{ij}\boldsymbol{F'}\hat{\boldsymbol{n}}_j^{(0)}$, where $\boldsymbol{F'}$ represents a learnable parameter matrix.
- Finally, with the use of residual connections, the input node and updated node features are added together to obtain the final node features, i.e., $\hat{\boldsymbol{n}}_i^{(1)} = \hat{\boldsymbol{n}}_{i,R}^{(1)} + \hat{\boldsymbol{n}}_i^{(0)}$.

The encoder consists of L RE-GAT layers, the execution process of each layer is the same as described above. These layers are connected in a cascade manner, where the output of the previous layer serves as the input to the next. The output node embedding of the L-th layer RE-GAT is denoted as $\hat{\boldsymbol{n}}_i^{(L)}$. Then, we utilize $\hat{\boldsymbol{n}}_i^{(L)}$ to compute the average embedding of the graph $\bar{\boldsymbol{n}} = \{\bar{\boldsymbol{n}}_1, \bar{\boldsymbol{n}}_2, \cdots, \bar{\boldsymbol{n}}_{d_n}\}$, where $\bar{\boldsymbol{n}}_j$ is given by

$$\bar{\boldsymbol{n}}_j = \frac{1}{J+2}\sum\nolimits_{i=0}^{J+1}\left(\hat{\boldsymbol{n}}_i^{(L)}\right)_j,\ j = 1, 2, \cdots, d_n. \tag{20}$$

Decoder: The decoder employs the attention mechanism, where the softmax probability distribution is used at each decoding step to select an output node from the input sequence. The decoder consists of two attention layers, i.e., a multi-head attention layer and a single-head attention layer. The multi-head attention layer is used to compute contextual information, while the single-head attention layer generates the probability distribution of the nodes that are selected. At each time step t, the decoder generates a node a_t based on the previous node $a_{t'}, t' < t$, and the embeddings of all nodes. To be specific, the context vector $\boldsymbol{c}_t^{(0)}$ can be obtained by utilizing the encoder's output $\bar{\boldsymbol{n}}$, features of the last selected node a_{t-1}, and features of the first selected node a_0, which is given as

$$\boldsymbol{c}_t^{(0)} = \bar{\boldsymbol{n}} + \boldsymbol{F}_n\left(\hat{\boldsymbol{n}}_{a_0}^{(L)}\left\|\hat{\boldsymbol{n}}_{a_{t-1}}^{(L)}\right.\right), \tag{21}$$

where $\boldsymbol{F}_n$ represents the learnable parameter matrix, and the superscript 0 indicates that the context vector is before entering the multi-head attention layer. Subsequently, $\boldsymbol{c}_t^{(0)}$ is fed into the first layer, i.e., multi-head attention layer, to generate a new contextual information $\boldsymbol{c}_t^{(1)}$. The detailed execution processes of the multi-head attention layer are presented as follows:

- Firstly, we define three vectors of dimension d_v, namely the query vector $\boldsymbol{q} = \boldsymbol{F}^Q\boldsymbol{c}_t^{(0)}$, the key vector $\boldsymbol{k}_i = \boldsymbol{F}^K\hat{\boldsymbol{n}}_i^{(L)}$, and the value vector $\boldsymbol{z}_i = \boldsymbol{F}^Z\hat{\boldsymbol{n}}_i^{(L)}$, where $\boldsymbol{F}^Q$, $\boldsymbol{F}^K$, and $\boldsymbol{F}^Z$ are learnable parameter matrices, and $d_v = d_n/P$ with P representing the number of heads.
- Secondly, the attention coefficients at the time step t are calculated by utilizing the query vector $\boldsymbol{q}$ and the key vector $\boldsymbol{k}_i$, which can be denoted as

$$\mathfrak{a}_{i,t}^{(1)} = \begin{cases}\frac{\boldsymbol{q}^T\boldsymbol{k}_i}{\sqrt{d_v}},\ if\ i \neq a_{t'}, \forall t' < t,\\ -\infty,\ otherwise.\end{cases} \tag{22}$$

- Thirdly, the attention coefficients are normalized by leveraging the softmax function as $\hat{\mathfrak{a}}_{i,t}^{(1)} = \text{softmax}(\mathfrak{a}_{i,t}^{(1)})$.
- Finally, independent attention computations are performed on each of the P heads, followed by sequential concatenation of the results and processing through a fully connected layer to produce the final context vector $\boldsymbol{c}_t^{(1)}$, which is given as $\boldsymbol{c}_t^{(1)} = \boldsymbol{F}_c\left(\left\|_{p=1}^{P}\sum_{r\in\mathcal{N}_i}\left(\hat{\mathfrak{a}}_{r,t}^{(1)}\right)^p\boldsymbol{z}_r^p\right.\right)$, where $\boldsymbol{F}_c$ represents the learnable parameter matrix with $\left(\hat{\mathfrak{a}}_{r,t}^{(1)}\right)^p$ and $\boldsymbol{z}_r^p$ denoting the attention coefficient and value vector of the p-th head, respectively.

The obtained context vector $\boldsymbol{c}_t^{(1)}$ is then put into the second layer, i.e., single-head attention layer, to calculate the selection probability of each node. In particular, we first compute the attention coefficient at the time step t, and then clip it within the range $[-C, C]$ by utilizing the tanh function, which can be expressed as

$$\mathfrak{a}_{i,t}^{(2)} = \begin{cases}C\tanh\left(\frac{\boldsymbol{c}_t^{(1)T}\boldsymbol{k}_i}{\sqrt{d_v}}\right),\ if\ i \neq a_{t'}, \forall t' < t,\\ -\infty,\ otherwise.\end{cases} \tag{23}$$

Subsequently, the softmax function is utilized to compute the selection probability of each node, which can be denoted as $p_{i,t} = \text{softmax}(\mathfrak{a}_{i,t}^{(2)})$. At last, a stochastic sampling strategy is employed to predict the next node that will be visit.

At this point, the construction of the GNN model has been completed. The next step involves training the model.

2) Training: An improved baseline REINFORCE algorithm is proposed to train the proposed GNN model, which is a type of actor-critic reinforcement learning algorithm [28]. We first define the length of multi-hop path Ω as $\mathcal{L}(\Omega) = \sum_{i=0}^{I}\Gamma_{a_i,a_{i+1}}$. Then, a random strategy $p(\Omega)$ that represents the selection probability of multi-hop path Ω is given as

$$p_\theta(\Omega) = \prod\nolimits_{i=0}^{I} p_\theta(a_i|a_{i'}), \forall i' < i, \tag{24}$$

where θ is the parameters of the GNN and $p_\theta(a_i|a_{i'})$ denotes the selection probability of node a_i, which can be obtained by the decoder. Consequently, the objective function of the improved baseline REINFORCE algorithm can be defined as $J_G(\theta) = \mathbb{E}_{\Omega\sim p_\theta(\Omega)}[\mathcal{L}(\Omega)]$. We utilize the policy gradient to update the parameters, which can be expressed as

$$g(\theta) = \mathbb{E}_{\Omega\sim p_\theta(\Omega)}\left[(\mathcal{L}(\Omega) - \mathfrak{b})\nabla_\theta\log(p_\theta(\Omega))\right], \tag{25}$$

where $\mathfrak{b}$ represents the baseline that can be estimated by the baseline policy network. The detailed processes of training the proposed GNN method by using the improved REINFORCE algorithm are shown in Algorithm 1.

B. Active and Passive Beamforming

In this subsection, based on the obtained optimal multi-hop path scheduling, we propose a DRL-based method to jointly optimize the active beamforming of the BS and the passive beamforming vectors of the selected RISs. The DRL algorithm can reformulate the joint optimization of active and passive beamforming as an MDP, which comprises four components, i.e., state $s^{(n)} \in \mathcal{S}$, action $a^{(n)} \in \mathcal{A}$, reward $r^{(n)}$, and next state $s^{(n+1)} \in \mathcal{S}$. The objective of the DRL algorithm is to find the optimal action $a^{(n)}$ for each state $s^{(n)}$ and uses the Q-value function to evaluate the long-term value of each action.

Algorithm 1 Improved Baseline REINFORCE Algorithm.

Input: Training parameter θ of actor, training parameter θ_c of baseline policy network, and significance level χ of t-test.
Output: Training parameter θ of actor.

```
Initialize θ and θ_c.
for each epoch do
  for each step do
    Execute the GNN model to obtain the node probability distributions p_θ(Ω) and p_{θ_c}(Ω).
    Sample a node a_i in p_θ(Ω) using a stochastic sampling.
    Sample a node a_i^{θ_c} in p_{θ_c}(Ω) using a greedy strategy.
    Calculate L(a_i) and L(a_i^{θ_c}) and normalize them.
    Compute advantage estimate Â_i = |L(a_i) − L(a_i^{θ_c})|.
    Update parameter θ using (25).
  end for
  if OneSidedPairedTTest(p_θ, p_{θ_c}) < χ then
    θ_c ← θ.
  end if
end for
return Training parameter θ of actor.
```

1) State, Action, and Reward Function: To solve the optimization problem (P1), we define the following state and action spaces and the reward function.

State Space: The state space is a set of observations that describe the characteristics of environment. We define the CSI and multi-hop paths Ω as the state space. Since the input data type of neural networks needs to be consistent, the state array can be represented as

$$s^{(n)} = \{\Omega, \text{real}\{\boldsymbol{H}, \boldsymbol{S}, \boldsymbol{g}\}, \text{imag}\{\boldsymbol{H}, \boldsymbol{S}, \boldsymbol{g}\}\}. \tag{26}$$

Action Space: The action space consists of two optimization variables, i.e., the active beamforming at the BS and the passive beamforming at the RIS involved in Ω, which is denoted as

$$a^{(n)} = \{\text{real}\{\boldsymbol{w}_k, \boldsymbol{\Phi}_{a_i}\}, \text{imag}\{\boldsymbol{w}_k, \boldsymbol{\Phi}_{a_i}\}\}. \tag{27}$$

Reward Function: Based on the objective function of problem (P1), the reward function of DRL is set to the sum achievable rate, which is denoted as $r^{(n)} = R_{sum}$.

2) Algorithm Flow: The optimization problem (P1) involves continuous variables and requires to effectively handle the large decision spaces. Therefore, the DDPG algorithm is employed for solving it. Compared to deep Q-network (DQN) and double deep Q-network (DDQN), the adopted DDPG algorithm performs better in handling continuous action spaces and demonstrates higher stability and faster convergence during training. The DDPG algorithm consists of two pairs of DNNs, i.e., the actor network and the critic network, along with their corresponding target networks [29]. The actor network and the critic network are responsible for selecting actions and evaluating the value of actions, respectively, while the corresponding target networks are used to stabilize the training process. The procedure of the active and passive beamforming based on DDPG is described as follows:

First, the agent observes the environment to obtain the CSI and the multi-hop path Ω to form the state $s^{(n)}$. Second, the state $s^{(n)}$ is input into the actor network to derive the action $a^{(n)}$. Third, the state $s^{(n)}$ and the action $a^{(n)}$ are fed into the critic network for action evaluation to output the Q-value. At last, after taking action $a^{(n)}$ in the communication environment, the state $s^{(n)}$ is transferred to $s^{(n+1)}$ and an immediate reward $r^{(n)}$ is obtained. Upon completing this process, the transitions $\left\{s^{(n)}, a^{(n)}, r^{(n)}, s^{(n+1)}\right\}$ are recorded in the memory buffer. During each training session, the agent utilizes experience replay to randomly sample a certain number of examples from the memory bank for training.

Notably, the DDPG outputs continuous phase shifts within the range of $[0, 2\pi]$, whereas the codebook only contains a finite set of discrete phase shifts, which may lead to quantization errors and thus affect the convergence and robustness of the algorithm. To mitigate the impact of quantization errors, we integrate quantization-aware training (QAT) into the training process, which explicitly models quantization effects during training, enabling the network to learn policies that adapt to quantization constraints and enhance the robustness and effectiveness of the final strategy.

In particular, we add a quantization layer behind the output layer of the actor network to quantize the continuous phase shift. Let $\boldsymbol{\phi}_{a_i} = \arg(\boldsymbol{\Phi}_{a_i})$ denote the continuous phase output of the actor network with $\arg(\cdot)$ being the phase extraction operation. After obtaining the continuous phase, we apply the quantization layer to discretize it, which can be expressed as

$$\tilde{\boldsymbol{\phi}}_{a_i} = \mathfrak{D}\left(\boldsymbol{\phi}_{a_i}\right) \triangleq \frac{2\pi}{2^b} \times \text{round}\left(\frac{2^b}{2\pi} \times \boldsymbol{\phi}_{a_i}\right), \tag{28}$$

where $\mathfrak{D}(\cdot)$ is the quantization function and $\text{round}(\cdot)$ represents the rounding operation. Consequently, the final action space can be reformulated as

$$\tilde{a}^{(n)} = \left\{\text{real}\left\{\boldsymbol{w}_k, \tilde{\boldsymbol{\Phi}}_{a_i}\right\}, \text{imag}\left\{\boldsymbol{w}_k, \tilde{\boldsymbol{\Phi}}_{a_i}\right\}\right\}, \tag{29}$$

where $\tilde{\boldsymbol{\Phi}}_{a_i} = e^{j\tilde{\boldsymbol{\phi}}_{a_i}}$.

During the training process, the critic network updates by minimizing the loss function, which can be expressed as

$$\varpi_c^{(n+1)} = \varpi_c^{(n)} - \delta_c \nabla_{\varpi_c^{(train)}} \mathcal{L}\left(\varpi_c^{(train)}\right), \tag{30}$$

where

$$\mathcal{L}\left(\varpi_c^{(train)}\right) = \left(y^{(n)} - Q\left(\varpi_c^{(train)} \middle| s^{(n)}, a^{(n)}\right)\right)^2 \tag{31}$$

with $y^{(n)} = r^{(n)} + \gamma Q\left(\varpi_c^{(target)} \middle| s^{(n+1)}, a'\right)$, δ_c represents the learning rate of the critic network, $\nabla_{\varpi_c^{(train)}} \mathcal{L}\left(\varpi_c^{(train)}\right)$ is the gradient of the loss function with respect to $\varpi_c^{(train)}$, $\varpi_c^{(target)}$ and $\varpi_c^{(train)}$ are the training parameters of the target critic network and the training critic network, respectively. The actor network is trained by using the policy gradient that is computed by the critic network, which can be given as

$$\varpi_a^{(n+1)} = \varpi_a^{(n)} - \delta_a \nabla_a Q^{(traget)} \nabla_{\varpi_a^{(train)}} \pi(\varpi_a^{(train)} \Big| s^{(n)}), \tag{32}$$

Algorithm 2 DRL-Based Active and Passive Beamforming Algorithm.

Input: $\boldsymbol{H}_{0,j}$, $\boldsymbol{S}_{i,j}$, $\boldsymbol{g}_{J,k}^H$ and Ω.
Output: $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{\Phi}_{a_i}\}$.

for each epoch **do**
  Reset the communication system and state space.
  for each step **do**
    Collecting CSI and Ω forms the set $s^{(n)}$.
    Obtain $a^{(n)}$ from the actor network.
    Obtain $\tilde{a}^{(n)}$ from the quantization layer.
    Apply the obtained $\tilde{a}^{(n)}$ to the environment, observe new state $s^{(n+1)}$, and get the instant reward $r^{(n)}$.
    Store $\left\{s^{(n)}, \tilde{a}^{(n)}, r^{(n)}, s^{(n+1)}\right\}$ in the memory buffer.
    Evaluate the action and obtain the Q-value using the critic network.
    Training the critic network using (30).
    Training the actor network using (32).
    Copy the parameters of the actor and critical networks to their respective target networks.
  end for
  Storing $\varpi_c^{(target)}$, $\varpi_c^{(train)}$, $\varpi_a^{(target)}$ and $\varpi_a^{(train)}$.
end for
return $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{\Phi}_{a_i}\}$.

where δ_a represents the learning rate of the actor network, $\nabla_{\varpi_a^{(train)}}\pi(\varpi_a^{(train)}\left|s^{(n)}\right.)$ denotes the gradient of the actor network with respect to parameter $\varpi_a^{(train)}$, and $\nabla_a Q^{(traget)}$ is the gradient of the target critic network with respect to the action. Since the target network and training network share an identical structure, the data of training network can be copied to the target network after a period of training. It should be noted that since $\mathfrak{D}\left(\phi_{a_i}\right)$ is a discontinuous function and thus it is non-differentiable. To mitigate the issue of vanishing gradients during backpropagation, we employ the straight-through estimator (STE) to approximate the gradient, which can be expressed as $\partial\mathfrak{D}(\boldsymbol{\phi}_{a_i})/\partial\boldsymbol{\phi}_{a_i} \approx 1$.

The detailed implementation processes of the active and passive beamforming algorithm based on DRL are shown in Algorithm 2. Note that achieving high-precision RIS control in large-scale communications presents certain challenges. As the network scale expands, the number of RISs and the complexity of control increase, which leads to higher computational resource demands. Additionally, mutual interference and cooperative control between RISs further complicate beamforming design. Therefore, employing a distributed multi-agent DRL algorithm, where each RIS is assigned an independent agent for phase shift learning, can alleviate computational burdens and enhance scalability [30].

Remark 1: Due to the multiplicative attenuation characteristic of multi-hop channels, the received signal may degrade after multiple reflections or refractions, which causes the achievable rate of certain users to fall below a predefined threshold. Hence, if the system detects that the rate of a user drops below the threshold, the algorithm will terminate execution and re-run after removing the user with the lowest rate to ensure the communication quality of the remaining users. This strategy effectively mitigates the impact of low-rate users on the overall system performance, thereby optimizing resource allocation and enhancing the operational efficiency and stability of the system.

C. Analyses of Complexity

The complexity of the proposed two-stage algorithm primarily arises from the GNN and DRL components. Specifically, the complexity of the GNN is attributed to the encoder and decoder sections. In the encoder, node and edge features are first processed through fully connected layers, with a complexity of $\mathcal{O}\left(nd_n^2 + ed_\Gamma^2\right)$, where n and e stand for the number of nodes and edges, respectively. Subsequently, feature embeddings are input to RE-GAT for node and edge fusion, which has a complexity of $\mathcal{O}\left(n^2 L\right)$. In the decoder, the complexity of multi-head attention is $\mathcal{O}\left(n^2 d_v P\right)$, while that of single-head attention is $\mathcal{O}\left(n^2 d_v\right)$. Therefore, the overall complexity of the GNN is $\mathcal{O}\left(nd_n^2 + ed_\Gamma^2 + n^2 L + (P+1)\, n^2 d_v\right)$. The complexity of the DRL component is mainly determined by the dimensions of the state and action spaces. Therefore, the complexity can be calculated as $\mathcal{O}\left(\sum_{q=1}^{Q} |\mathcal{S}|\,|\mathcal{A}|\right)$, where Q represents the number of network layers. Consequently, the sum complexity of the two-stage algorithm is $\mathcal{O}(nd_n^2 + ed_\Gamma^2 + n^2 L + (P+1)\, n^2 d_v + \sum_{q=1}^{Q} |\mathcal{S}|\,|\mathcal{A}|)$.

To facilitate a fair comparison with the proposed two-stage algorithm, we conduct a comprehensive complexity analysis of the conventional scheme. Specifically, during the path selection stage, the conventional method employs Dijkstra's algorithm to determine the shortest path between the BS and users, with a complexity denoted by $\mathcal{O}((n+e)\log n)$. In the joint beamforming stage, the conventional scheme adopts the AO method as described in [31], which proceeds in two steps. First, given the fixed passive beamforming, the active beamforming is optimized using the SCA method, resulting in a complexity of $\mathcal{O}(T_1(K^3 N_r(N_t+2)^3 + K^2 N_r(N_t+2)^2))$, where T_1 is the number of SCA iterations. Second, with active beamforming fixed, the passive beamforming problem is relaxed into a convex form and solved using standard convex optimization tools, yielding a complexity of $\mathcal{O}(T_2(JM+2)^2(JM+3+K))$, where T_2 represents the iteration count for the convex solver. Hence, the overall complexity of the conventional scheme is $\mathcal{O}((n+e)\log n + T[T_1(K^3 N_r(N_t+2)^3 + K^2 N_r(N_t+2)^2) + T_2(JM+2)^2(JM+3+K)])$, where T denotes the total number of AO iterations.

In summary, once the proposed two-stage algorithm has been trained, it enables efficient multi-hop path selection and joint beamforming without the need for path search or iterative optimization, thereby significantly reducing the online computational cost.

D. Analysis of Convergence

The proposed two-stage algorithm consists of GNN and DDPG, and its convergence analysis must be conducted separately for each component. First, we prove the convergence of GNN. The GNN is trained by using an improved REINFORCE

algorithm, and its convergence is ensured by satisfying three conditions, namely the unbiasedness of the gradient estimation, the learning rate meeting the Robbins-Monro conditions, and the convergence of the objective function. Specifically, we can expand (25) as

$$g(\theta) = \mathbb{E}_{\Omega}\left[L(\Omega)\nabla_{\theta}\log\left(p_{\theta}(\Omega)\right)\right] - \mathbb{E}_{\Omega}\left[\mathfrak{b}\nabla_{\theta}\log\left(p_{\theta}(\Omega)\right)\right], \tag{33}$$

where the first term is the gradient of the objective function, while the second term is the baseline. To demonstrate the unbiasedness of the gradient estimation, it suffices to show that the baseline term is equal to 0. Furthermore, we have

$$\mathbb{E}_{\Omega}\left[\mathfrak{b}\nabla_{\theta}\log\left(p_{\theta}(\Omega)\right)\right] = \mathbb{E}_{\Omega}\left[\mathfrak{b}\nabla_{\theta}\sum_{i=i'+1}^{I} p_{\theta}\left(a_i\left|a_{i'}\right.\right)\right]. \tag{34}$$

Since $\sum_{i=i'+1}^{I} p_{\theta}\left(a_i\left|a_{i'}\right.\right) = 1$, thus the baseline term is equal to 0. The Robbins-Monro condition is a sufficient condition for the convergence of stochastic gradient descent, which requires that the learning rate η_t satisfies $\sum_{t=1}^{\infty}\eta_t = \infty$ and $\sum_{t=1}^{\infty}\eta_t^2 < \infty$. Setting $\eta_t = \frac{1}{t^{\tau}}, 0.5 < \tau \leq 1$, can ensure compliance with this condition. In gradient-based optimization for DRL algorithms, the objective function is updated by

$$J_G(\theta_{t+1}) - J_G(\theta_t) \approx \eta_t \nabla_{\theta} J_G(\theta_t)\, g(\theta) + \mathcal{O}\left(\eta_t^2\right), \tag{35}$$

where the higher-order terms become negligible if $\eta_t \to 0$. Hence, the expression gradually approaches zero, which indicates that the objective function converges to its optimal value. Additionally, assuming that the objective function $J_G(\theta)$ is smooth and the variance of the gradient estimation is bounded, the GNN has a sublinear convergence rate according to stochastic gradient optimization theory, which can be expressed as

$$\mathbb{E}\left[J_G(\theta^*) - J_G(\theta_t)\right] \leq \frac{C_G}{\sqrt{t}}, \tag{36}$$

where C_G is a constant determined by the smoothness of the objective function and the variance of the gradient estimation.

Next, we provide a proof of the convergence of DDPG. The DDPG employs a deterministic policy gradient to maximize the long-term cumulative reward, which can be expressed as

$$J_R(\varpi) = \mathbb{E}_{s\sim p^{\pi}(s)}\left[Q^{\pi}\left(s, \pi_{\varpi}(s)\right)\right], \tag{37}$$

where $p^{\pi}(s)$ is the steady-state distribution of states and $\pi_{\varpi}(s)$ is a parameterized deterministic strategy. According to the deterministic strategy gradient theorem, we have

$$\nabla_{\varpi} J_R(\varpi) = \mathbb{E}_{s\sim p^{\pi}}\left[\nabla_a Q^{\pi}(s,a)\big|_{a=\pi_{\varpi}(s)} \nabla_{\varpi}\pi_{\varpi}(s)\right], \tag{38}$$

where $Q^{\pi}(s,a)$ is an action-valued function. If $Q^{\pi}(s,a)$ and $\pi_{\varpi}(s)$ are Lipschitz continuous with respect to ϖ, the parameter update can be expressed as $\varpi_{t+1} = \varpi_t + \delta\nabla_{\varpi} J(\varpi_t)$. Thus, if the learning rate δ satisfies the Robbins-Monro condition, the policy parameters will converge to the optimal solution. Moreover, the critic network is trained by minimizing the temporal difference error, which is shown as (31). The theoretical analysis of DQN shows that with experience replay and target networks, the update of the Q-function can be interpreted as a mean regression process, leading to the convergence of the Q-function to an approximate optimal solution [32]. Furthermore, based on mean regression theory, the convergence bound of the critic network can be expressed as

$$\|Q_t - Q^*\| \leq C_Q \chi_Q^t, \tag{39}$$

where $0 < \chi_Q < 1$, and C_Q is a constant.

In summary, the proposed two-stage algorithm is theoretically convergent with quantifiable error bounds and asymptotic convergence rates. In addition, we also experimentally demonstrate the convergence behavior and tightness of the GNN and DRL algorithms in Section IV.

IV. Numerical Results And Analysis

A. Simulation Settings

We consider a 3D scenario in which the BS is located on the xz-plane at coordinates (2, 0, 10) m. The users are randomly distributed within a semicircular region of the xy-plane, centered at (12, 105) m with a radius of 10 m, and the vertical height of the users is assumed to be fixed at 2 m. Multiple RISs are deployed between the BS and the users, oriented perpendicular to the xy-plane. In addition, we assume that the last RIS is located at (11, 100, 10) m. The path-loss model is given by

$$PL(d) = \varsigma_a + 10\varsigma_b \log_{10}(d) + X_{\sigma}, \tag{40}$$

where ς_a represents the reference path-loss at a distance of 1 m, ς_b denotes the path-loss exponent, d signifies the link distance, and $X_{\sigma} \sim \mathcal{CN}\left(0, \sigma_P^2\right)$ corresponds to the shadow fading. Based on the parameter setting in [33], we set $\varsigma_a = 61.4$, $\varsigma_b = 2$, and $\sigma_P = 5.8$ dB. Assuming a carrier frequency of 28 GHz and the noise power is ${\sigma_n}^2 = -90$ dBm. The antenna spacing of BS and spacing between two adjacent RIS reflection elements are set to $d_B = d_R = \lambda/2$. The numbers of quantization bits and paths are set to 3 and 7, respectively. The minimum rate for all users is $R_{\min} = 0.5$ bps/Hz.

We trained the proposed GNN model for multi-hop path scheduling with the number of RISs J = 5, 10, 15, and 20. The training, validation, and test datasets are generated from a three-dimensional region defined by $(0, 20)$ m $\times$ $(0, 150)$ m $\times$ $(0, 10)$ m, with the BS fixed at $(2, 0, 10)$ m. Since the focus of the study is long-range mmWave transmissions, the coordinates of RIS J are fixed at $(11, 100, 10)$ m, while the remaining $J-1$ RISs are randomly deployed between the BS and RIS J. In the vertical dimension, the heights of the RIS are uniformly distributed within the range of $(0, 10)$ m. Additionally, it was assumed that beams could only be transmitted from RIS i to a further RIS j, i.e., $d_{j,0} > d_{i,0}$, to reach the user as quickly as possible. We generate 5000, 10000, 15000, and 20000 instances for the cases where the number of RISs $J = 5, 10, 15,$ and 20, respectively[3]. To optimize the neural network training process, we referred to the hyperparameter

[3]The proposed framework relies on channel realizations that may not fully capture real-world distributions or dynamic changes. To address this, future work could explore online training, domain adaptation, or robust ML techniques, such as liquid neural networks and meta-learning, to improve performance under out-of-distribution conditions and enhance adaptability to real-world environments [16], [17].

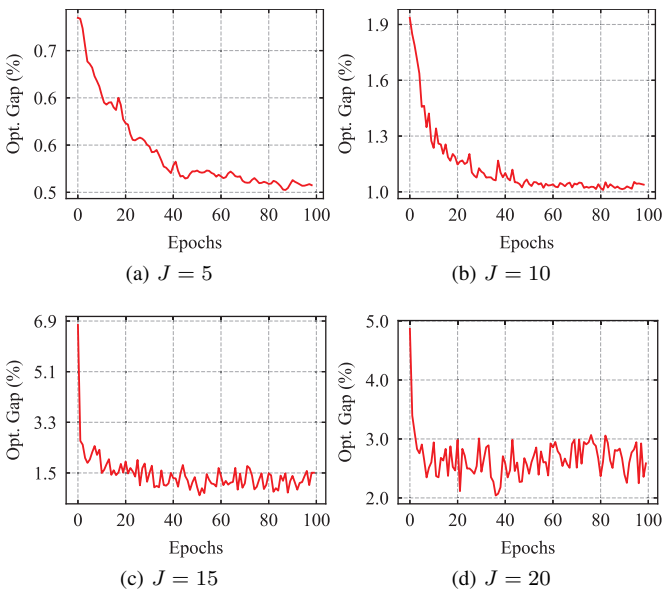

Fig. 3. The convergence properties of the proposed algorithm for obtaining multi-hop path scheduling under varying scales of RIS.

settings provided in [28] and [29] and systematically fine-tuned them. After extensive experimental validation, we finalize the optimized hyperparameter configuration. Specifically, the discount factor is set to 0.99, the learning rates of both the actor and critic networks are set to 10^{-4}, and the learning rate of the GNN is set to 10^{-3}. The simulations are conducted by using the PyTorch framework on a Linux-based computing server equipped with an Intel Xeon(R) Platinum 8474C @ 2.1 GHz CPU and a 24 GB NVIDIA RTX 4090D GPU.

B. Evaluation of Multi-Hop Path Scheduling

The convergence of the proposed multi-hop path scheduling algorithm is first demonstrated at different RIS scales, and the performance comparison of various path selection schemes is then discussed. For evaluating the convergence, we adopt the optimality gap (Opt. Gap) to measure the average performance difference between the proposed algorithm and the benchmark method, i.e., Gurobi solver, which utilizes mixed-integer programming to obtain the global optimal solution with high time complexity [28]. The Opt. Gap can be expressed as

$$\text{Opt. Gap} = \frac{1}{N_{test}} \sum\nolimits_{i=1}^{N_{test}} \frac{L_i\left(\tilde{\Omega}\right) - L_i\left(\Omega\right)}{L_i\left(\Omega\right)}, \tag{41}$$

where N_{test} denotes the number of experiments, $L_i(\tilde{\Omega})$ and $L_i\left(\Omega\right)$ represent the path lengths obtained by the Gurobi solver and the proposed algorithm, respectively.

Fig. 3 illustrates the convergence behavior of the proposed multi-hop path scheduling optimization algorithm under different number of RISs. Overall, the proposed algorithm converges with increase of training epochs across various RIS scales. Specifically, for a small-scale RIS deployment ($J = 5$), the Opt. Gap reaches to near 0.5%, which indicates that the proposed algorithm closely approximates the global optimal solution. The smooth convergence process further

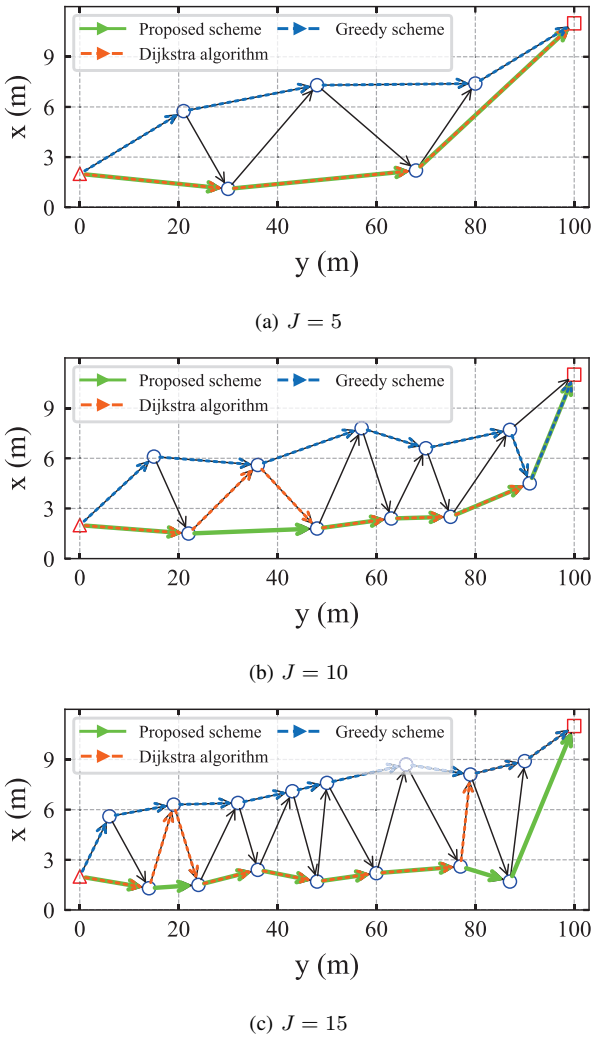

Fig. 4. Routing path selection results of different optimization schemes under various scales of RIS.

confirms the strong stability of the algorithm. As the number of RIS increases, the Opt. Gap value is gradually increased and oscillations appear in the convergence process. This is because that the increased number of RISs makes the multi-hop path optimization more complex, which leads to a larger gap between the proposed algorithm and the benchmark solution. Moreover, the complicated multi-hop path selection also causes slight oscillations during the convergence process. Even with a large-scale RIS deployment ($J = 20$), the Opt. Gap of the proposed algorithm remains below 3% to demonstrate the excellent availability of the proposed GNN-based multi-hop path scheduling algorithm.

In Fig. 4, we compare the proposed multi-hop path scheduling algorithm with other two benchmark schemes, i.e., Dijkstra algorithm [8] and Greedy-based beam routing scheme, where the red triangle represents the BS, the blue circle denotes the RIS, and the red square indicates the fixed position of the last RIS. The Dijkstra algorithm traverses all feasible routing paths and selects the path with the maximum equivalent channel gain as given by (14). The Greedy-based beam routing scheme sequentially chooses the RIS with the smallest edge weight for

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3579599

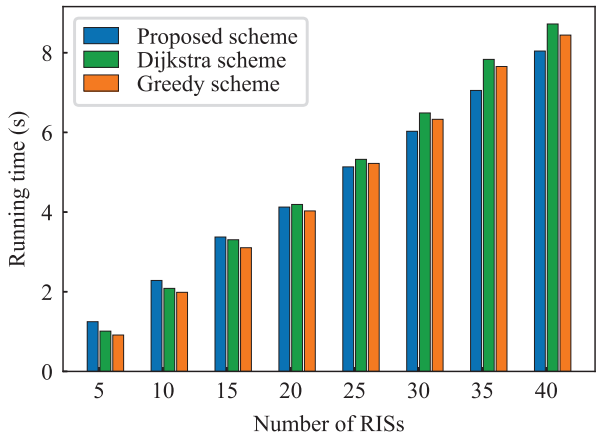

Fig. 5. Runtime of different schemes under various scales of RIS.

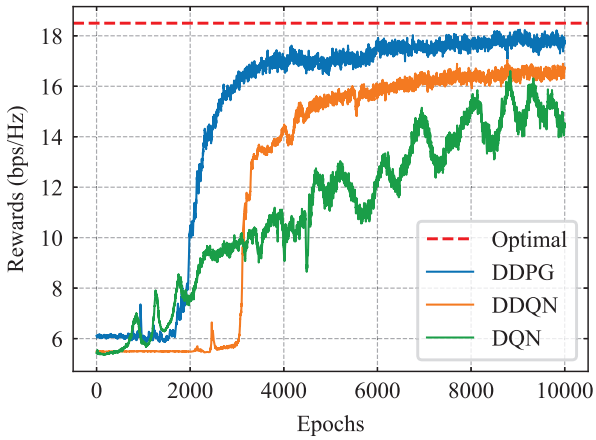

Fig. 6. The convergence of different DRL models. $J = 7$, $N_t = 16$, $N_r = 3$, $M = 9$, $K = 6$, and $P_{\max} = 30$ dBm.

routing until the signal reaches to users. Several important observations are highlighted as follows. In the case of small-scale RIS ($J = 5$), as illustrated in Fig. 4 (a), the proposed multi-hop path scheduling scheme generates the same beam routing path as the Dijkstra algorithm, while the greedy scheme generates a path containing more RISs. Such results indicate that in simple scenarios with small number of RISs, both the proposed algorithm and Dijkstra algorithm can achieve a better beam routing path than the greedy scheme. The reason lies in that both the proposed algorithm and Dijkstra algorithm account for global information when selecting paths, balancing the weights of nodes and paths to choose the overall optimal solution. In contrast, the greedy scheme focuses solely on the immediate optimal choice at each step without considering the impact of the entire path, which makes it prone to falling into local optima. Moreover, as the number of RIS increases shown in Figs. 4 (b) and 4 (c), the three schemes generate distinct beam routing paths, with the proposed algorithm demonstrating significantly superior performance compared to the other two schemes. This demonstrates that in complex RIS deployment environments, the proposed GNN-based multi-hop path scheduling algorithm significantly outperforms the Dijkstra algorithm and the greedy scheme. The advantage benefits from the GNN's ability that can better capture and utilize the complex relationships and global information in the network, thereby achieving superior multi-hop path selection.

In Fig. 5, we compare the running time of different optimization schemes under various scales of RIS. To guarantee the fairness, all calculations are performed on the CPU. From the figure, it can be seen that when the number of RISs is small, the running time of the proposed scheme is longer than that of the compared Dijkstra and greedy schemes. This is because the GNN needs to process more complex graph structure information during inference, which incurs higher computational overhead at smaller number of RISs. However, as the number of RISs increases, the running time of the proposed scheme gradually becomes shorter than that of the two baseline schemes. This is due to the proposed method can effectively model the dependency of the inter-RIS by using GNN, optimizing the path selection and reducing redundant computations. Specifically, the GNN learns the global network structure, which avoids the inefficient process of repetitive path searches inherent in the Dijkstra algorithm. Additionally, although the greedy scheme can select a multi-hop path quickly, its local optimization strategy will result in a lower transmission performance. Therefore, the proposed scheme demonstrates better performance in large-scale scenarios.

C. Sum Achievable Rate with Proposed Two-Stage Algorithm

Fig. 6 illustrates the convergence performance of different DRL models in the joint optimization of active and passive beamforming. As a performance upper bound, the optimal benchmark first exhaustively determines the best multi-hop path between the BS and users, and then applies the DDPG algorithm to jointly optimize the active beamforming at the BS and the passive beamforming at the RISs along with the selected path. To further evaluate the theoretical upper limit of system performance, the passive beamforming in the optimal benchmark adopts a continuous phase shift model. As the training epochs increase, the sum achievable rate of each model improves and eventually converges, demonstrating that DRL algorithms can effectively address non-convex optimization problems through interaction with the environments. In general, the DDPG significantly outperforms other comparison models and converges to near-optimal performance, which because that the DDPG utilizes the actor-critic structure and a deterministic policy to learn efficient optimization strategies and accurately estimate the value of actions, resulting in superior performance. Furthermore, the DDPG converges after approximately 4000 training steps, while DDQN takes about 6000 steps to converge. This is primarily because the DDPG can adjust the strategy more quickly with its deterministic policy in a continuous action space, accelerating the convergence process. In contrast, the DDQN leads to slower convergence due to its requiring more exploration steps to balance exploration and exploitation. Meanwhile, DQN shows oscillation and instability during the convergence, which is

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3579599

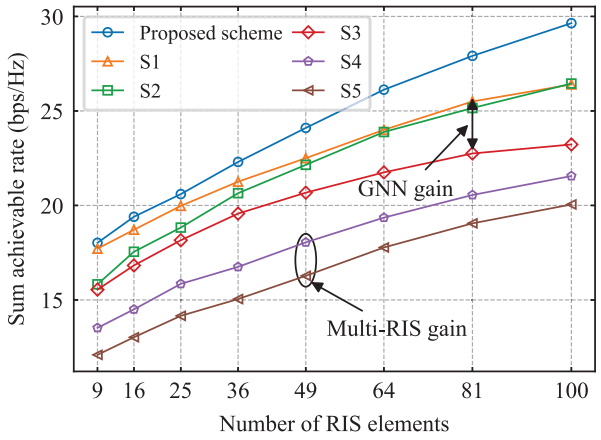

Fig. 7. The sum achievable rate versus the number of reflection units per RIS for different algorithms. $J = 7$, $N_t = 16$, $N_r = 3$, $K = 6$, and $P_{\max} = 30$ dBm.

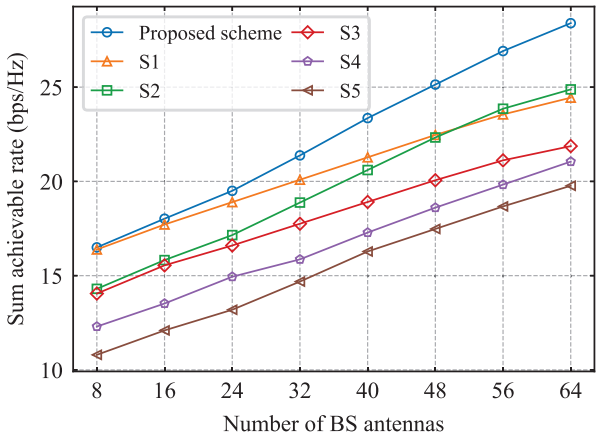

Fig. 8. The sum achievable rate versus the number of BS antennas for different algorithms. $J = 7$, $M = 9$, $N_r = 3$, $K = 6$, and $P_{\max} = 30$ dBm.

mainly due to the overestimation of value in the training process and leads to the instability of learning strategy.

In the following, we analyze the impacts of the system parameters on the sum achievable rate. To ensure a fair comparison of the proposed two-stage method, we select the AO method proposed in [31] as the baseline. The AO method alternately optimizes the active beamforming and passive beamforming by fixing one set of variables while optimizing the other. In particular, given a fixed passive beamforming configuration, the SCA method is employed to optimize active beamforming. Conversely, when active beamforming is determined, the passive beamforming problem is transformed into a convex problem through appropriate constraint relaxation and solved by using convex optimization toolboxs. This iterative process continues until convergence. The comparison schemes are shown as follows:

- Algorithm S1 utilizes the proposed GNN model for beam routing path selection and employs the AO algorithm to design active and passive beamforming.
- Algorithm S2 utilizes the Dijkstra algorithm for beam routing path selection and employs the proposed DRL model to obtain active and passive beamforming.
- Algorithm S3 utilizes the Dijkstra algorithm for beam routing path selection and employs the AO algorithm to design active and passive beamforming.
- Algorithm S4 utilizes a greedy strategy for beam routing path selection and employs the proposed DRL model to obtain active and passive beamforming.
- Algorithm S5 utilizes a single RIS to serve multiple users and employs the proposed DRL model to obtain active and passive beamforming.

Fig. 7 shows the comparison of the proposed two-stage algorithm with other algorithms in terms of sum achievable rate under different numbers of reflection units in each RIS. With the increase number of reflection units, the sum achievable rate of all algorithms are improved, while the proposed two-stage algorithm consistently outperforms other algorithms. The reason is that the proposed algorithm not only employs an GNN model to select better cascaded path for significantly enhancing passive beamforming gain, but also leverages an DRL algorithm to dynamically optimize both active and passive beamforming for further boosting the sum achievable rate of the system. Moreover, with a small number of reflection units in each RIS, the algorithm S1 significantly outperforms the algorithm S2. However, as the number of reflection units increases, the gap between algorithms S1 and S2 gradually narrows and their sum achievable rates become almost equal at $M = 100$. It is mainly because in a small number of RISs, the performance difference between the traditional AO algorithm and DRL algorithm is not significant, and the algorithm S1 utilizes the GNN model to select superior multi-hop path. As the number of RISs increases, the advantages of DRL algorithm gradually emerge, which results a narrow gap between them. In addition, the algorithm S4 consistently outperforms the algorithm S5 across the entire number range of reflection units, which indicates that even the greedy scheme is adopted for beam path selection, its performance surpasses that of single-RIS reflection. It is further demonstrated that multi-RIS reflections can provide greater path diversity between the BS and users, leading to more significant cooperative passive beamforming gains. Furthermore, the proposed algorithm achieves a sum achievable rate of 29.64 bps/Hz with $M = 100$, while algorithm S3 reaches only 23.22 bps/Hz, the proposed achievable demonstrates an improvement of approximately 27.6% in sum achievable rate compared to conventional algorithm.

Fig. 8 illustrates the sum achievable rate versus the number of BS antennas with different algorithms. Overall, as the number of BS antennas increases, the sum achievable rate under all algorithms is improved, while the proposed two-stage algorithm consistently performing the best. The superior performance is attributed to exceptional ability of the GNN model to obtain the optimal multi-hop path scheduling between the BS and users, and the precise beamforming design of the DRL algorithm, which further enhances the system's sum achievable rate. Another noteworthy observation is that

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3579599

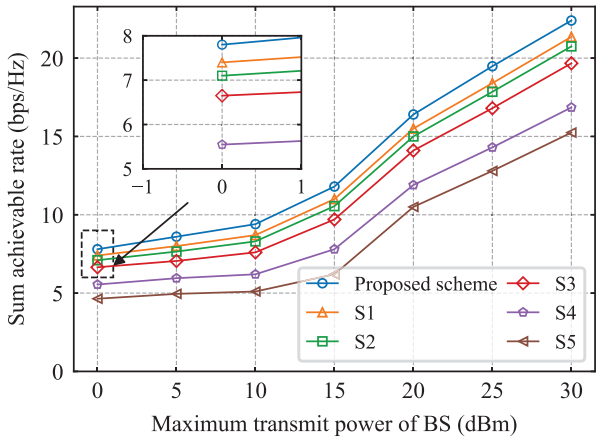

Fig. 9. The sum achievable rate versus the maximum transmit power for different algorithms. $J = 7$, $N_t = 16$, $N_r = 3$, $M = 36$, and $K = 6$.

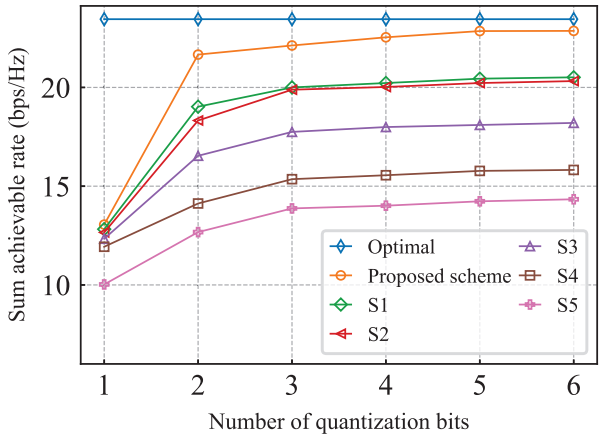

Fig. 10. The sum achievable rates versus number of quantization bits for different algorithms. $J = 7$, $N_t = 16$, $N_r = 3$, $M = 64$, $K = 6$, and $P_{\max} = 30$ dBm.

when the number of BS antennas is relatively small, the performance gap between the proposed two-stage algorithm and the algorithm S1, as well as between the algorithms S2 and S3, is not significant. However, as the number of BS antennas increases, these gaps are become widely, which is mainly due to the performance differences between the traditional AO algorithm and the DRL algorithm. Specifically, with a small number of BS antennas, the performance of the traditional AO algorithm and the DRL algorithm is similar. As the scale of the BS antennas increases, the advantage of the DRL algorithm becomes more evident, which results in a more significant difference in the sum achievable rate. In addition, we can find that the single-RIS reflection scheme performs the worst due to its inability to provide sufficient path diversity and passive beamforming gain to compensate the severe path-loss in complex mmWave communications.

Fig. 9 presents the sum achievable rate with respect to the maximum transmit power of BS for different algorithms. It can be observed that the sum achievable rate under all algorithms increases monotonically with the BS's maximum transmit power, and their performance curves exhibit similar trends. Moreover, the proposed two-stage algorithm significantly outperforms the other algorithms, which indicates that under high transmit power conditions, the proposed algorithm can utilize the BS's transmit power more efficiently and thus improve the sum achievable rate of the system. The algorithm S5 has the lowest sum achievable rate among all the schemes since the single-RIS scheme is unable to provide sufficient path diversity and beamforming gain to overcome signal fading and multi-path effects in mmWave communications.

Fig. 10 illustrates the sum achievable rate of various algorithms under different RIS phase quantization resolutions. It can be observed that, except for the proposed optimal scheme, the sum achievable rate of all other algorithms increases with the number of quantization bits and tends to saturate when the resolution reaches 4 bits. The optimal scheme consistently serves as the performance upper bound, while the proposed algorithm outperforms all compared schemes and approaches the performance of the optimal one. Notably, when the quantization resolution is 1 bit, the sum achievable rates of all multi-RIS schemes are nearly identical. This is because a 1-bit quantizer allows only two discrete phase levels (0° and 180°), which severely limits the beamforming capability, making it difficult to distinguish performance differences among the algorithms. In contrast, the single-RIS scheme consistently yields lower sum achievable rates than the multi-RIS schemes, primarily due to the lack of path diversity and beamforming gain in single-hop systems. As the quantization resolution increases, the controllability of the RIS improves significantly, enabling more effective adjustment of the reflection direction and thus enhancing the overall system performance. However, it should be noted that higher quantization resolutions can increase the implementation complexity. Therefore, adopting 3 to 4 bits of quantization precision provides a good trade-off between system performance and hardware complexity in most practical scenarios.

V. Conclusion

In this paper, we investigated multi-hop path scheduling and beamforming for RIS-assisted mmWave multi-hop communications, where multiple RISs collaborated to establish a virtual LoS multi-hop path from the BS to users. To guarantee the transmission performance of the considered system, we formulated a joint optimization problem to maximize the sum achievable rate of all users, which is a challenging task since it integrates multi-hop path scheduling, active beamforming of the BS, and passive beamforming of the selected RIS. An ML-based two-stage algorithm was designed to tackle the joint optimization problem. In particular, in the first stage, we constructed the weights between adjacent nodes by analyzing the compositional characteristics of the RIS-assisted cascaded equivalent channel. Then, a GNN-based multi-hop path scheduling algorithm was proposed to select an optimal multi-hop path that can minimize path-loss and maximize the equivalent channel gain. In the second stage, based on the identified optimal multi-hop path, we invoked a DDPG

algorithm based on the Markov decision process to optimize active beamforming at the BS and passive beamforming of the selected RISs. Simulation results demonstrated that the proposed GNN-based algorithm consistently selected superior multi-hop paths compared to traditional graph-based optimization methods. Moreover, the error of the proposed GNN-based multi-hop path scheduling optimization algorithm remained within 3% of the global optimum. Numerical results indicated that the proposed two-stage ML-based algorithm achieves an approximately 27.6% improvement in sum achievable rate compared to the conventional AO algorithm.

Due to space limitations, some practical issues are not addressed in this paper, which also provides motivation for the future research. This paper assumes a quasi-static communication environment, with the user location remaining constant over a short period. However, in real-world applications, the communication environment is often influenced by dynamic factors such as user movement, RIS position adjustments, and external environmental changes. Therefore, it is necessary to update the CSI in real-time, which, significantly increases the computational load. Therefore, future research could incorporate channel knowledge maps, utilizing historical user data to construct a knowledge base reflecting local wireless environment characteristics. This would allow to obtain the environmental prior information, reducing the need for repeated CSI acquisition and alleviating the computational burden associated with real-time CSI updates in high-dynamic environments. In addition, to mitigate the severe path-loss induced by multi-hop transmission, future research could also introduce a small number of active RISs, which would compensate for signal attenuation at lower energy and hardware costs.

REFERENCES

[1] F. Yu, C. Zhang, and T. Q. S. Quek, "Blockage correlation in IRS-assisted millimeter wave communication systems," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 6, pp. 5786-5800, Jun. 2023.

[2] W. Chen, C. Liu, W. Wang, M. Peng, and W. Zhang, "Adaptive hybrid beamforming for UAV mmWave communications against asymmetric jitter," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 8, pp. 9432–9445, Feb. 2024.

[3] V. B. Shukla, V. Bhatia and K. Choi, "Cascaded channel estimator for IRS-aided mmWave hybrid MIMO system," *IEEE Wirel. Commun. Lett.*, vol. 13, no. 3, pp. 622-626, Mar. 2024.

[4] B. Zheng, Q. Wu, and R. Zhang, "Rotatable antenna enabled wireless communication: modeling and optimization," *arXiv preprint arXiv:2501.02595*, Jan.2025.

[5] M. Zhou, Y. Li, Y. Sun and Z. Ding, "Outage performance of RIS-assisted V2I communications with inter-cell interference," *IEEE Wirel. Commun. Lett.*, vol. 12, no. 6, pp. 962-966, Jun. 2023.

[6] R. K. Fotock, A. Zappone and M. D. Renzo, "Energy efficiency optimization in RIS-aided wireless networks: Active versus nearly-passive RIS with global reflection constraints," *IEEE Transactions on Communications*, vol. 72, no. 1, pp. 257-272, Jan. 2024.

[7] D. Shen, Z. Zhang and L. Dai, "Joint beamforming design for RIS-assisted cell-free network with multi-hop transmissions," *Tsinghua Sci. Technol.*, vol. 28, no. 6, pp. 1115-1127, Dec. 2023.

[8] X. Ma, H. Zhang, X. Chen, Y. Fang and D. Yuan, "Multi-hop multi-RIS wireless communication systems: Multi-reflection path scheduling and beamforming," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 6778-6792, Jul. 2024.

[9] Y. Zhang, C. You, and B. Zheng, "Multi-active multi-passive (MAMP)-IRS aided wireless communication: A multi-hop beam routing design," *IEEE J. Sel. Areas Comm.*, vol. 41, no. 8, pp. 2497–2513, Dec. 2023.

[10] H. Zhou, M. Erol-Kantarci, Y. Liu and H. V. Poor, "A survey on model-based, heuristic, and machine learning optimization approaches in RIS-aided wireless networks," *IEEE Commun. Surv. Tutor.*, vol. 26, no. 2, pp. 781-823, Secondquarter 2024.

[11] Y. Liu, R. Wang and Z. Han, "Passive beamforming for practical RIS-assisted communication systems with non-ideal hardware," *IEEE Trans. Veh. Technol.*, vol. 73, no. 11, pp. 17743-17748, Nov. 2024.

[12] X. Wang et al., "Joint beamforming and reflecting elements optimization for segmented RIS assisted multi-user wireless networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 3820-3831, Mar. 2024.

[13] D. Jia, Y. Zhong, X. Zhou, A. Xu and M. Zhao, "Energy efficiency in RIS-assisted wireless networks: Impact of phase shift and deployment," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Dubai, United Arab Emirates, Apr. 2024, pp. 1-6.

[14] Z. Guo, Y. Niu, S. Mao, et al., "Sum rate maximization under AoI constraints for RIS-assisted mmWave communications," *IEEE Trans. Vehicular Technol.*, vol. 73, no. 4, pp. 5243-5258, Apr. 2024.

[15] Z. Wang, Y. Zhang, H. Zhou, J. Li, D. Wang, and X. You, "Performance analysis and optimization for distributed RIS-assisted mmWave massive MIMO with multi-antenna users and hardware impairments," *IEEE Trans. Commun.*, vol. 72, no. 8, pp. 4661–4676, Aug. 2024.

[16] F. Zhu et al., "Robust beamforming for RIS-aided communications: gradient-based manifold meta learning," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 11, pp. 15945-15956, Nov. 2024.

[17] X. Wang et al., "Robust beamforming with gradient-based liquid neural network," *IEEE Wirel. Commun. Lett.*, vol. 13, no. 11, pp. 3020-3024, Nov. 2024.

[18] S. Sobhi-Givi, M. Nouri, H. Behroozi, and Z. Ding, "Joint BS and beyond diagonal RIS beamforming design with DRL methods for mmWave 6G mobile communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Dubai, United Arab Emirates, Apr. 2024, pp. 1-6.

[19] Y. Wang, Y. Zhang, Y. Ren, L. Pang, Y. Chen and J. Li, "Joint BS-RIS-User Association and Deployment Design for Multi-RIS-Aided Wireless Networks," *IEEE Commun. Lett.*, vol. 28, no. 9, pp. 2181-2185, Sep. 2024.

[20] L. Liu, H. Wang and R. Song, "Optimization for multi-cell NOMA systems assisted by multi-RIS with inter-RIS reflection," *IEEE Commun. Lett.*, vol. 28, no. 1, pp. 123-127, Jan. 2024.

[21] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, et al., "Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design," *IEEE J. Sel. Areas Comm.*, vol. 39, no. 6, pp. 1663–1677, Mar. 2021.

[22] Z. Chen, J. Tang, X. Y. Zhang, Q. Wu, G. Chen and K. -K. Wong, "Robust hybrid beamforming design for multi-RIS assisted MIMO system with imperfect CSI," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 6, pp. 3913-3926, Jun. 2023.

[23] W. Mei and R. Zhang, "Cooperative beam routing for multi-IRS aided communication," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 2, pp. 426–430, Feb. 2020.

[24] W. Mei and R. Zhang, "Multi-beam multi-hop routing for intelligent reflecting surfaces aided massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1897–1912, Mar. 2022.

[25] A. Köse, H. Gökcesu, N. Evirgen, K. Gökcesu and M. Médard, "A novel method for scheduling of wireless ad hoc networks in polynomial time," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 1, pp. 468-480, Jan. 2021.

[26] Y. Song, S. Xu, R. Xu and B. Ai, "Weighted sum-rate maximization for multi-STAR-RIS-assisted mmwave cell-free networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 4, pp. 5304-5320, Apr. 2024.

[27] W. Lu, N. Jiang, D. Jin, H. Chen and X. Liu, "Learning distinct relationship in package recommendation with graph attention networks," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 6, pp. 3308-3320, Dec. 2023.

[28] K. Lei, P. Guo, Y. Wang, X. Wu, and W. Zhao, "Solve routing problems with a residual edge-graph attention neural network," *Neurocomputing*, vol. 508, no. 7, pp. 79–98, Oct. 2022.

[29] R. Zhong, Y. Liu, X. Mu, Y. Chen and L. Song, "AI empowered RIS-assisted NOMA networks: deep learning or reinforcement learning?" *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 182-196, Jan. 2022.

[30] A. Abdallah, A. Celik, M. M. Mansour and A. M. Eltawil, "Multi-agent deep reinforcement learning for beam codebook design in RIS-aided systems," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 7, pp. 7983-7999, Jul. 2024.

[31] J. Zuo, Y. Liu, E. Basar, and O. A. Dobre, "Intelligent reflecting surface enhanced millimeter-wave NOMA systems," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2632–2636, Nov. 2020.

[32] J. Chi, X. Zhou, F. Xiao, Y. Lim and T. Qiu, "Task offloading via prioritized experience-based double dueling DQN in edge-assisted IIoT," *IEEE. Trans. Mob. Comput.*, vol. 23, no. 12, pp. 14575-14591, Dec. 2024.

[33] M. R. Akdeniz et al., "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164-1179, Jun. 2014.

Tongyi Wei (Graduate Student Member, IEEE) received the B.S. degree from Bohai University, Jinzhou, China, and the M.S. from Lanzhou Jiaotong University, Lanzhou, China. He is currently pursuing the Ph.D. degree in Information and Communication Engineering with the Guangdong Provincial Key Laboratory of Millimeter-Wave and Terahertz, South China University of Technology, Guangzhou, China. His research interests include machine learning, reconfigurable intelligent surface, and millimeter-wave wireless communications.

Beixiong Zheng (Senior Member, IEEE) received his B.S. and Ph.D. degrees from the South China University of Technology, Guangzhou, China, in 2013 and 2018, respectively. He is currently an Associate Professor with the School of Microelectronics, South China University of Technology. He was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore from 2019 to 2022. From 2015 to 2016, he was a Visiting Student Research Collaborator with Columbia University, New York, NY, USA. His recent research interests include 5G/6G wireless communications, intelligent reflecting surface (IRS), satellite communication, and signal processing.

Dr. Zheng is now serving as an Editor for the IEEE COMMUNICATIONS LETTERS. He was the recipient of the IEEE COMMUNICATIONS SOCIETY Heinrich Hertz Award for Best Communications Letter in 2022, the IEEE COMMUNICATIONS SOCIETY Best Tutorial Paper Award in 2023, the Guangdong Provincial Electronic Information Science and Technology Award (First Prize of Natural Science Award) in 2022, the Best Ph.D. Thesis Award from the China Education Society of Electronics in 2018, the Best Paper Award from the IEEE International Conference on Computing, Networking and Communications (ICNC) in 2016, and the Best Paper Award from the International Conference on Ubiquitous Communication (Ucom) in 2023. In addition, he was listed as the World's Top 2% Scientist by Stanford University and Mendeley Data from 2021 to 2023. He was also recognized as an Exemplary Reviewer of the IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE COMMUNICATIONS LETTERS, and an Outstanding Reviewer of the PHYSICAL COMMUNICATION.

Kun Tang (Member, IEEE) received the B.S. degree in telecommunications from Wuhan University of Technology, Wuhan, China, in 2006, and M.S. degree in The University of New South Wales, Sydney, Australia, in 2011, and Ph.D. degree in Telecommunications from the Central South University, Changsha, China, in 2018. He is now an Associate Professor with the school of Electronic and Information Engineering at South China University of Technology. He was a Post-Doctor with the school of Electronic and Information Engineering, South China University of Technology from 2019 to 2022. His research interests are in the areas of wireless power transfer, mmWave communications, and network security.

Wenjie Feng (Senior Member, IEEE) was born in Shangqiu, Henan, China, in 1985. He received the B.Sc. degree from the First Aeronautic College of the Airforce, Xinyang, China, in 2008, and the M.Sc. and Ph.D. degrees from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2010 and 2013, respectively.

From July 2017 to September 2017, he was a Research Fellow with the City University of Hong Kong. From October 2010 to March 2011, he was an Exchange Student with the Institute of High Frequency Engineering, Technische Universität München, Munich, Germany. He is currently a Professor with NUST, and also a Professor with South China University of Technology, Guangzhou, China. He has authored or coauthored over 100 IEEE journal articles (including 70 IEEE TRANSACTIONS papers) and 80 conference papers. His research interests include wideband circuits and technologies, microwave and millimeter-wave circuits and components, circuits interconnection, and packaging. Dr. Feng was a recipient of the National Science Fund for Excellent Young Scholars in 2018, the Young Scientist Award of ACES-China 2018, and a reviewer for over 20 internationally refereed journals and conferences.

Wenquan Che (Fellow, IEEE) received the B.Sc. degree from the East China Institute of Science and Technology, Nanjing, China, in 1990, the M.Sc. degree from the Nanjing University of Science and Technology (NUST), Nanjing, in 1995, and the Ph.D. degree from the City University of Hong Kong (CITYU), Kowloon, Hong Kong, in 2003.

In 1999, she was a Research Assistant with CITYU. From March 2002 to September 2002, she was a Visiting Scholar with the Polytechnique de Montréal, Montréal, QC, Canada. She is currently a Professor with the South China University of Technology, Guangzhou, China. From 2007 to 2008, she conducted academic research with the Institute of High Frequency Technology, Technische Universität München. From summer 2005 to summer 2006 and from summer 2009 to summer 2012, she was with CITYU, as a Research Fellow and a Visiting Professor. She has authored or coauthored over 300 internationally refereed journal articles and international conference papers. Her research interests include electromagnetic computation, planar/coplanar circuits and subsystems in RF/microwave frequency, microwave monolithic integrated circuits (MMICs), and medical application of microwave technology.

Quan Xue (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1988, 1991, and 1993, respectively.

In 1993, he joined the UESTC, as a Lecturer. He became a Professor in 1997. From October 1997 to October 1998, he was a Research Associate and then a Research Fellow with The Chinese University of Hong Kong. In 1999, he joined the City University of Hong Kong and was a Chair Professor in microwave engineering. He also served the University as the Associate Vice President (Innovation Advancement and China Office) from June 2011 to January 2015, the Director of the Information and Communication Technology Center (ICTC Center), and the Deputy Director of the State Key Laboratory of Millimeter Waves (Hong Kong). In 2017, he joined the South China University of Technology, where he is currently a Professor and also serves as the Dean of the School of Electronic and Information Engineering, the Dean of the School of Microelectronics, and the Director the Guangdong Key Laboratory of Terahertz and Millimeter Waves. He has authored or coauthored over 300 internationally refereed journal articles and over 130 international conference papers. He is the co inventor of five granted Chinese patents and 15 granted U.S. patents, in addition with 26 filed patents. His research interests include microwave/millimeter-wave/THz passive components, active components, antenna, and microwave monolithic integrated circuits (MMIC, and radio frequency integrated circuits (RFIC).