Deep Learning-Based Channel Extrapolation and Multi-User Beamforming for RIS-aided Terahertz Massive MIMO Systems over Hybrid-Field Channels

Yang Wang[1], Zhen Gao[1], and Sheng Chen[2,3]

[1]School of Information and Electronics, Beijing Institute of Technology, Beijing, China.

[2]School of Electronics and Computer Science, University of Southampton, Southampton, U.K.

[3]Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China

Abstract

The reconfigurable intelligent surface (RIS) is a promising new technology for Terahertz (THz) massive multiple input multiple output (MIMO) communication systems. However, due to the cascaded channel structure and its lack of signal processing ability, it is difficult for RIS to obtain the high-dimensional channel state information (CSI) and optimize the active/passive beamforming. Therefore, in this paper, we propose a DL-based transmission scheme for RIS-aided THz massive MIMO systems over hybrid far-near field channels, where a channel estimation scheme with low pilot overhead and a robust beamforming scheme are conceived. Specifically, we first propose an end-to-end deep learning (DL)-based channel estimation framework, which consists of pilot design, CSI feedback, sub-channel estimation, and channel extrapolation. Specifically, we firstly only turn on a fraction of all the RIS elements and estimate a sub-sampling RIS channel, and then design a DL-based scheme to extrapolate the full-dimensional CSI from the partial one. Moreover, to maximize the sum rate of all users under imperfect CSI, we develop a DL-based scheme to simultaneously design the hybrid active beamforming at the BS and passive beamforming at the RIS. Simulation results show that our proposed channel extrapolation scheme has better CSI reconstruction performance than conventional schemes while imposing a much reduced pilot overhead and our proposed beamforming scheme has superior performance over conventional schemes in terms of robustness to imperfect CSI.

Keywords

Reconfigurable intelligent surface (RIS), Terahertz (THz), hybrid-field, channel extrapolation, hybrid beamforming, orthogonal frequency division multiplexing (OFDM), multiple-input multiple-output

(MIMO), deep learning (DL).

1 Introduction

Over the past few years, the demand for wireless data traffic has increased significantly due to the explosive growth of mobile devices and multimedia applications [1]. To accommodate these demands, Terahertz (THz) communications have attracted great interest from both industry and academia. However, there exists strong atmospheric attenuation and extremely high free-space losses in the THz band. These disadvantages may severely degrade the service coverage of THz communication systems. The deployment of massive multiple input multiple output (MIMO) or even ultra-massive MIMO has been recognized as an achievable solution to provide high array gain, so as to overcome the high propagation loss [2]. However, conventional massive MIMO systems with fully-digital architecture require a dedicated radio frequency (RF) chain for each antenna, which results in excessive power consumption and extremely high RF hardware costs. In order to circumvent this technical hurdle, the hybrid analog-digital massive MIMO architecture has been widely adopted to achieve large array gains with a much lower number of RF chains [3].

Besides, the emerging technology of reconfigurable intelligent surface (RIS) has also been considered as a promising technique to enhance communication performance [4]. The RIS can manipulate both the phase and amplitude of the incident electromagnetic (EM) signals so as to passively reflect them towards the desired directions and provide significant beamforming gain. More importantly, RIS does not need power-hungry RF chains, which is beneficial for developing green and cost-efficient communications. Therefore, the application of RIS and massive MIMO techniques is expected to overcome the limitations of THz communications and realize its full potential.

Generally, a simplified planar-wave channel model is appropriate in the case that the user equipment (UE) works in the far-field of the base station (BS). However, since severe path loss will reduce the effective coverage while the increasing array size in THz band will increase the Rayleigh distance [5], both far-field and near-field need to be considered for THz massive MIMO systems. Therefore, the spherical-wave is necessary to accurately analyze the propagation of THz waves under near-field conditions, where the distance from each antenna of the BS to the UE needs to be considered [6]. On the other hand, in THz massive MIMO systems, the number of channel parameters is proportional to the number of massive antennas, which indicates that directly applying the spherical-wave channel model is unrealistic. To this end, a hybrid-field (hybrid spherical- and planar-wave) channel model characterized by a smaller number of parameters while maintaining high accuracy has been proposed for THz massive MIMO systems [7]. For such a channel model, given the subarray structure, the inter-subarray is modeled as spherical-wave and the intra-subarray is modeled as planer-wave. Although the application of RISs has been widely investigated recently [8, 9, 10, 11, 12], the utilization of RIS for THz massive MIMO communications over hybrid-field channels is still at its early study stage.

1.1 Related Work

Acquiring accurate channel state information (CSI) is critical in establishing RIS-aided communication systems. However, it is challenging to accurately estimate high-dimensional channels with a few pilots [8]. By exploiting the sparsity of the channels in the angular and/or delay domains, compressive sensing (CS)-based solutions have been proposed to reduce the pilot overhead [9, 10]. Nevertheless, since the dimension of the CSI to be estimated is extremely large, the involved matrix inversion operations and the iterative nature of CS-based techniques result in prohibitively high computational complexity and storage requirements. With the development of artificial intelligence, the application of deep learning (DL) in RIS-aided communication systems has attracted extensive attention. In [11], the authors designed a twin convolutional neural network (CNN) to estimate the direct (BS-UE) and the cascaded (BS-RIS-UE) channels. In [12], the authors proposed a hybrid passive/active RIS architecture with a few RF chains, where the orthogonal match pursuit (OMP) algorithm and denoising CNN are applied to reconstruct the complete channel matrix. However, the deployment of RF chains at the RIS defeats the original purpose of reducing hardware cost and power consumption by deploying the RIS.

In fact, due to the highly-dense placement of RIS elements, there is a strong correlation between the different elements of the CSI matrix, which makes it possible to extrapolate the complete channel from a partial one, i.e., channel extrapolation [13]. Recently, there are some initial attempts to utilize the channel extrapolation for further reducing the pilot overhead. In [14], the authors proposed an antenna domain extrapolation network to perform channel extrapolation, where an antenna selection network is designed to choose the optimal antennas for the extrapolation. In [15], the neural network structure modified by ordinary differential equation was used to describe the latent relations between different data layers and improve the performance gains of antenna extrapolation. Besides, the authors of [16] adopted the element-grouping strategy to reduce the pilot overhead and the CNN-based channel extrapolation network to extrapolate the full-dimensional cascaded channels from the partial one. However, the above extrapolation schemes only consider the extrapolation process from the known sub-channels, while ignoring how to estimate the sub-channel. Moreover, the hybrid-field channel modeling of RISs has more complex propagation characteristics, which will hinder the sub-channel acquisition and the following extrapolation of complete channels.

How to properly and effectively design the hybrid beamforming and RIS phase according to the CSI is one of the major engineering challenges in the design of RIS-aided communication systems. Recently, some work has been conducted to investigate hybrid beamforming and RIS design problems [17, 18, 19, 20]. The authors in [17] proposed a geometric mean decomposition-based beamforming for RIS-assisted mmWave hybrid MIMO systems. The authors in [18] focused on the hybrid beamforming and RIS phase design for RIS-aided multi-user mmWave communication systems and aimed to minimize the mean squared error (MSE) using the gradient-projection method. In [19], the authors proposed an iterative algorithm to jointly optimize the RIS phase and the hybrid beamforming for maximizing the sum rate. Furthermore, the DL-based beamforming methods have also been studied in RIS-assisted wireless communication systems. In [20], a DL-based approach was proposed to jointly optimize the active beamforming at the BS and the passive beamforming at the

RIS for achieving rate maximization. However, a further adaptability analysis of the aforementioned schemes is desired because the current analysis only assumes the idealized perfect CSI condition.

1.2 Motivations

In the current research, RIS is mainly constructed to have two modes: reflective mode [17, 18, 19] and transmissive mode [21, 22, 23]. At present, more researches focus on RIS-assisted communication in reflective mode. The RIS in reflective mode is mainly used to solve the blind coverage problem and enhance the network coverage. By contrast, the transmissive RIS is mainly used to enhance the spectral and energy efficiency of the networks. Since the transmissive mode does not change the propagation direction of EM waves, it is suitable to deploy transmissive RIS in the case that a line-of-sight (LoS) path exists but the propagation attenuation is high, e.g., the case that the outdoor BS serves indoor UEs, to improve the energy of the received signals. Therefore, we consider the transmissive RIS for indoor signal enhancement service.

Considering the hybrid-field channel model, the authors of [24] proposed a two-stage channel estimation mechanism, where a CNN-based network is designed to estimate the parameters of the reference subarrays and the complete channel is reconstructed by channel extrapolation based on geometric relationships of channel parameters. However, this parametric-based extrapolation method needs to obtain a large number of accurate channel parameters as labels before training. In [25], the authors proposed a sensor-based channel estimation and beamforming for RIS-aided THz systems, where a LoS MIMO architecture is considered in hybrid-field. However, the channel estimation in [25] is mainly based on the awareness of sensors, and it is difficult to obtain accurate CSI. Therefore, similar to [14, 15, 16], we adopt a DL-based channel extrapolation method to address the performance limitations of conventional channel estimation methods for indoor hybrid-field propagation environments. Besides, in this paper, we consider the LoS MIMO architecture under the assumption of the hybrid-field channel model, where the LoS MIMO architecture can support multi-stream transmission in the pure LoS BS-RIS channel.

Most existing works assume that the CSIs between BS and RIS as well as between RIS and UEs are perfect [17, 18, 19, 20]. However, this assumption is impractical. Therefore, it is imperative to take the channel estimation error into consideration when designing RIS-aided communication systems. Recently, there have been some works on beamforming with imperfect CSI [26, 27]. Specifically, in [26], the authors exploited the penalty-based alternating optimization to design secure wireless communications assisted by RIS under the imperfect CSI. In [27], the authors exploited a gradient projection-based alternating optimization algorithm to jointly optimize transmit beamforming, RIS placement, and reflect beamforming of the RIS under the imperfect CSI. Currently, there are extensive DL-based methods for the optimization of RIS phase with the perfect CSI, but to our best knowledge, there exists only few DL-based methods considering imperfect CSI [28]. Therefore, this work aims to provide a DL-based hybrid beamforming and RIS phase design with imperfect CSI in RIS-aided communication systems.

1.3 Contributions

This paper proposes a DL-based spatial-frequency domain channel extrapolation (SFDCEtra) network as well as the DL-based hybrid beamforming and RIS phase design (HBFRPD) scheme for RIS-aided downlink multi-user THz massive MIMO systems over hybrid-field channels. The main contributions of this paper are summarized as follows.

- We deploy the transmissive RIS on the window of a building to reduce the penetration loss and thus achieve indoor enhanced communication. In addition, due to the negligible non-LoS (NLoS) component energy in the THz band, the BS-RIS channel is dominated by the LoS path. To achieve multi-stream transmission in the LoS case, we consider an LoS MIMO architecture under hybrid-field channel modeling, where the BS and RIS adopt the same subarray structures, and the subarray spacing is optimized to satisfy the LoS MIMO condition.

- Since the BS and the RIS are fixed and only the LoS path exists, the BS-RIS channel can be considered to be quasi-static and it is known. The RIS-UE channel by contrast is time-varying due to the mobility of the UEs. Therefore, we only need to estimate the RIS-UE channel, which significantly reduces the pilot overhead.

- To further reduce the pilot overhead for estimating the RIS-UE channel, we propose a DL-based channel extrapolation scheme, where the RIS only activates part of its elements at the channel estimation stage. Unlike the existing extrapolation schemes [14, 15, 16] that only focus on the CSI extrapolation process, we design a complete channel extrapolation framework, including the pilot design network, CSI feedback network, sub-channel estimation network, and channel extrapolation network. By adopting the end-to-end (E2E) training strategy, the proposed channel estimation scheme can reduce the pilot overhead while maintaining high reconstruction performance. Specifically, the UE-side feeds the quantized pilot information back to the BS using the CSI feedback network, and the BS estimates the sub-channel and then extrapolates the complete RIS-UE channel using the channel extrapolation network. In addition, for the RIS element selection, we discuss the impact of three different strategies, uniform selection, random selection and learning-based selection, on the final channel estimation performance.

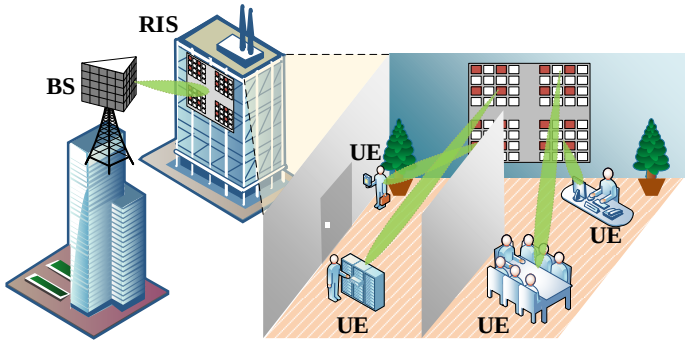
- To solve the multi-user interference problem under imperfect CSI, we propose a DL-based hybrid beamforming and RIS phase design scheme, which consists of the analog beamformer design, DL-based RIS phase design network, and knowledge-data dual-driven digital beamforming network. By adopting the sum rate as the loss function to perform E2E training, the proposed scheme can realize higher performance and better robustness than the existing state-of-the-art methods.

Notations: This paper uses lower-case letters for scalars, lower-case boldface letters for vectors, and upper-case boldface letters for matrices. Superscripts $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, $(\cdot)^{-1}$ and $(\cdot)^{\dagger}$ denote the conjugate, transpose, conjugate transpose, inversion, and Moore-Penrose inversion operators, respectively. The operators $\mathrm{diag}(\cdot)$, $\mathrm{blkdiag}(\cdot)$, $\otimes$ and $\odot$ represent the diagonalization, block diagonalization, Kronecker product and Hadamard product, respectively, while $\|\mathbf{A}\|_F$ denotes the

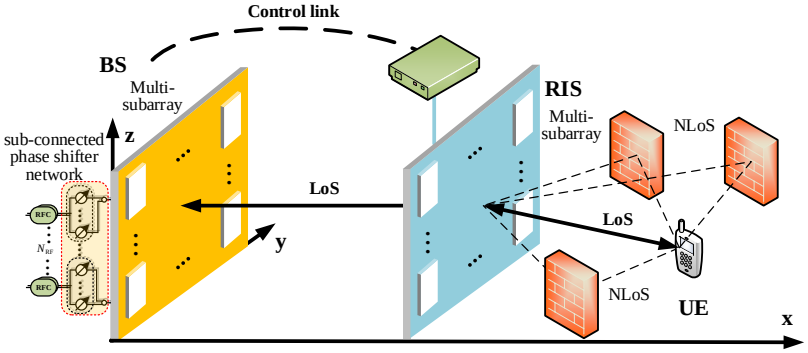
Frobenius norm of $\mathbf{A}$. $\mathbf{I}_n$ denotes the identity matrix with size $n \times n$, while $\mathbf{1}_n$ ($\mathbf{0}_n$) denotes the column vector of size n with all the elements being 1 (0). $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real part and imaginary part of the corresponding argument, respectively. $\{\mathbf{A}\}_{m,n}$ denotes the m-th row and n-th column element of $\mathbf{A}$, and $\{\mathbf{a}\}_m$ is the m-th entry of $\mathbf{a}$, while $\mathbf{A}_{[:,m:n]}$ is the sub-matrix containing the m-th to n-th columns of $\mathbf{A}$. The expectation is denoted by $\mathbb{E}(\cdot)$, and $\mathcal{N}(\mu, \sigma^2)$ ($\mathcal{CN}(\mu, \sigma^2)$) denotes the real (complex) Gaussian distribution with mean μ and variance σ^2, while $\mathrm{Tr}\{\cdot\}$ represents the matrix trace operator.

2 System Model

We first describe the RIS-assisted downlink THz multi-user MIMO system over frequency-selective fading channels, and then introduce the transmission channel model of this system.



(a)



(b)

Figure 1: Model of RIS-aided THz massive MIMO system: (a) multiple indoor UEs are served by the BS with the help of a transmissive RIS deployed on the window, and (b) hardware architectures at the BS, RIS, and UEs.

2.1 System Description

As shown in Figure 1, we consider a downlink transmission scenario in an indoor environment, where a transparent RIS is attached to the window surface to refract outdoor THz signals from the BS

into the room for serving U single-antenna UEs. Thus, the transparent transmissive RIS helps to enhance indoor coverage. Let the BS (RIS) have $M^{\mathrm{B}} = M_y^{\mathrm{B}} \times M_z^{\mathrm{B}}$ ($M^{\mathrm{R}} = M_y^{\mathrm{R}} \times M_z^{\mathrm{R}}$) uniformly spaced subarrays, where M_y^{B} (M_y^{R}) and M_z^{B} (M_z^{R}) are the numbers of subarrays along the azimuth and elevation directions, respectively. Each subarray of the BS (RIS) is a uniform planar array (UPA) with $N_{\mathrm{sub}}^{\mathrm{B}} = N_y^{\mathrm{B}} \times N_z^{\mathrm{B}}$ ($N_{\mathrm{sub}}^{\mathrm{R}} = N_y^{\mathrm{R}} \times N_z^{\mathrm{R}}$) isotropically radiating elements, where N_y^{B} (N_y^{R}) and N_z^{B} (N_z^{R}) are the numbers of antennas along the azimuth and elevation directions, respectively. Therefore, the complete antenna dimension of the BS is $N^{\mathrm{B}} = M^{\mathrm{B}} N_{\mathrm{sub}}^{\mathrm{B}}$, and the element dimension of the RIS is $N^{\mathrm{R}} = M^{\mathrm{R}} N_{\mathrm{sub}}^{\mathrm{R}}$. Without loss of generality, we assume that the normals of the central elements of both the BS and RIS are coaxial, i.e., meeting the parallel symmetric array arrangements, with a distance of D, as illustrated in Figure 1(b).

To reduce the power consumption and hardware cost, a sub-connected hybrid analog-digital array architecture is considered at the BS, i.e., there are only M^{B} RF chains to support $U \leq M^{\mathrm{B}}$ data streams, and each of RF chains is connected to a subarray through $N_{\mathrm{sub}}^{\mathrm{B}}$ phase shifters. An orthogonal frequency division multiplexing (OFDM) transmission scheme with K subcarriers and sampling period $T_s = 1/f_s$ is adopted. The cyclic prefix (CP) of length $N_{\mathrm{CP}} T_s$ is added before each OFDM symbol to avoid inter-symbol interference. The center-carrier frequency is f_c corresponding to a wavelength λ.

2.2 Channel Model

2.2.1 BS-RIS Channel Model

By considering the spherical wave propagation of THz signals, we investigate the LoS MIMO link between the BS and RIS with only the LoS path, but it can support intra-path multiplexing for multi-stream transmission [29]. In THz channels, the path loss of the NLoS paths is known to be much larger than that of the LoS path. Therefore, we can neglect the NLoS paths in the channel between the BS and the RIS. The inter-antenna spacing in each subarray is $d = \lambda/2$. In order to satisfy the LoS MIMO characteristic, the BS subarray spacing d_{sy}^{B} and d_{sz}^{B} are set to the following optimal LoS MIMO spacing

$$d_{sy}^{\mathrm{B}} = \sqrt{\frac{\lambda D}{M_y^{\mathrm{B}}}} - \frac{\lambda}{2}(N_y^{\mathrm{B}} - 1), d_{sz}^{\mathrm{B}} = \sqrt{\frac{\lambda D}{M_z^{\mathrm{B}}}} - \frac{\lambda}{2}(N_z^{\mathrm{B}} - 1), \tag{1}$$

i.e., d_{sy}^{B} and d_{sz}^{B} should satisfy the condition $\lambda \ll d_{sy}^{\mathrm{B}}, d_{sz}^{\mathrm{B}} \ll D$. The detailed explanation of Equation (1) can be found in [29, 30]. The RIS subarray spacing d_{sy}^{R} and d_{sz}^{R} can be obtained by using a similar definition. Note that self-orthogonal LoS MIMO not only is obtained from parallel symmetric antenna arrangements but also can be obtained with symmetrical/unsymmetrical arrangements on tilted non-parallel lines/planes [30]. We have the following proposition from [31].

Proposition 1 *Let the transceiver arrays be separated by D and be communicating at a carrier wavelength of λ that is much smaller than D, i.e., $\lambda \ll D$. If the inter-antenna spacing is in the order of $\mathcal{O}(\lambda)$, the first-order approximation of the spherical wave by the planner wave model can be applied. Otherwise, the spherical wave model should be used.*

According to Proposition 1, the subarray response vectors $\mathbf{a}(\theta,\phi,f_k)\in\mathbb{C}^{N_{\mathrm{H}}N_{\mathrm{V}}\times 1}$ can be approximated by a planner wave model:

$$\begin{aligned}\mathbf{a}(\theta,\phi,f_k)=&\mathbf{a}_h(\theta,\phi,f_k)\otimes\mathbf{a}_v(\phi,f_k)\\ =&\big[1,\ldots,e^{-\mathrm{j}2\pi\frac{f_k}{c}d(n_h\sin\theta\cos\phi+n_v\sin\phi)},\ldots,e^{-\mathrm{j}2\pi\frac{f_k}{c}d((N_{\mathrm{H}}-1)\sin\theta\cos\phi+(N_{\mathrm{V}}-1)\sin\phi)}\big]^{\mathrm{T}},\end{aligned}\quad(2)$$

where $f_k=f_c-\frac{f_s}{2}+\frac{kf_s}{K}$, $1\le k\le K$, is the k-th subcarrier frequency, c is the speed of light, $0\le n_h\le(N_{\mathrm{H}}-1)$, $0\le n_v\le(N_{\mathrm{V}}-1)$, N_{H} and N_{V} are the numbers of antennas along the azimuth and elevation directions in the subarray, respectively, while θ and ϕ are the azimuth and elevation angles of the departure or arrival (AoD or AoA) of the path, respectively.

Since $d^{\mathrm{B}}_{sy},d^{\mathrm{B}}_{sz},d^{\mathrm{R}}_{sy},d^{\mathrm{R}}_{sz}\ll D$, the direction difference of the same path in different subarrays is negligible, and all the subarrays on either the BS or RIS-side share the same set of array response vectors. However, the relative phase differences among subarrays are of non-negligible values, as subarrays are widely spaced. Motivated by the above analysis, the downlink spatial-frequency channel $\mathbf{G}[k]\in\mathbb{C}^{N^{\mathrm{R}}\times N^{\mathrm{B}}}$ from the BS to the RIS on the k-th subcarrier can be modeled as

$$\mathbf{G}[k]=\alpha[k]G_{\mathrm{T}}\tilde{\mathbf{G}}[k]\otimes\left[\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{R,A}},\phi_{\mathrm{R,A}},f_k)\mathbf{a}^{\mathrm{H}}_{\mathrm{B}}(\theta_{\mathrm{B}},\phi_{\mathrm{B}},f_k)\right],\quad(3)$$

where $\alpha[k]$ is the channel attenuation coefficient on the k-th subcarrier, $(\theta_{\mathrm{B}},\phi_{\mathrm{B}})$ and $(\theta_{\mathrm{R,A}},\phi_{\mathrm{R,A}})$ are the LoS AoD and LoS AoA of the BS-RIS channel, respectively. Without loss of generality, the LoS angles between the BS and the RIS are assumed to be fixed and known in advance based on their locations. In (3), the entries of $\tilde{\mathbf{G}}[k]\in\mathbb{C}^{M^{\mathrm{R}}\times M^{\mathrm{B}}}$ are given by the spherical wave model as

$$\{\tilde{\mathbf{G}}[k]\}_{m_r,m_b}=e^{-\mathrm{j}2\pi f_k\cdot\frac{D^{(m_r,m_b)}}{c}},\quad(4)$$

where $D^{(m_r,m_b)}$ denotes the distance between the m_r-th RIS-side subarray and the m_b-th BS-side subarray. Furthermore, the subarray response vectors $\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{R,A}},\phi_{\mathrm{R,A}},f_k)\in\mathbb{C}^{N^{\mathrm{R}}_{\mathrm{sub}}\times 1}$ and $\mathbf{a}_{\mathrm{B}}(\theta_{\mathrm{B}},\phi_{\mathrm{B}},f_k)\in\mathbb{C}^{N^{\mathrm{B}}_{\mathrm{sub}}\times 1}$ are defined in Equation (2). The constant coefficient G_{T} represents the antenna gain at the BS, which is the gain of the single antenna element, and is different from the array gain generated by beamforming [32]. The only unknown parameter in Equation (3) is the channel coefficient $\alpha[k]$, which can be obtained by placing a power detector at the RIS-side. Therefore, it is reasonable to assume that the BS-RIS channel is quasi-static and known.

2.2.2 RIS-UE Channel Model

As illustrated in Figure 1(b), we consider a multi-path THz channel model for indoor environments by using deterministic ray-tracing techniques [33]. For a RIS-UE link in a room, the multi-path propagation model consists of one LoS path and L_p NLoS paths reflected or scattered by the surrounding walls in the room, where the paths reflected by the ceiling and the floor can be neglected [34]. The three-dimensional (3D) distances for the LoS path and the NLoS paths between RIS and UE are denoted as d_0 and d_l, for $1\le l\le L_p$, respectively. The total attenuation of EM wave propagation in the THz band is composed of molecular absorption and free space path

loss. The free space path loss is $\beta_{\text{spr}}(f_k, d_l) = \frac{c}{4\pi f_k d_l}$, and the molecular absorption is given by $\beta_{\text{abs}}(f_k, d_l) = e^{-\frac{1}{2}\kappa(f_k)d_l}$, where $\kappa(f_k)$ stands for the frequency-dependent absorption coefficient [35]. Hence, the spatial-frequency channel $\mathbf{h}[k] \in \mathbb{C}^{1\times N^{\text{R}}}$ for the RIS-UE link can be written as

$$\mathbf{h}[k] = \beta[k]\tilde{\mathbf{h}}_{\text{LoS}}[k] \otimes \mathbf{a}_{\text{R}}^{\text{H}}(\theta_{\text{R,D}}^{\text{LoS}}, \phi_{\text{R,D}}^{\text{LoS}}, f_k) + \frac{1}{\sqrt{L_p}} \sum_{l=1}^{L_p} \beta_l[k]\tilde{\mathbf{h}}^l[k] \otimes \mathbf{a}_{\text{R}}^{\text{H}}(\theta_{\text{R,D}}^l, \phi_{\text{R,D}}^l, f_k), \tag{5}$$

where $\beta[k] = \beta_{\text{spr}}(f_k, d_0)\beta_{\text{abs}}(f_k, d_0)$ and $\beta_l[k] = \beta_{\text{spr}}(f_k, d_l)\beta_{\text{abs}}(f_k, d_l)\beta_{\text{RC}}$ are the channel attenuation coefficients of the LoS component and the l-th NLoS component, respectively, $(\theta_{\text{R,D}}^{\text{LoS}}, \phi_{\text{R,D}}^{\text{LoS}})$ and $(\theta_{\text{R,D}}^l, \phi_{\text{R,D}}^l)$ are the LoS AoD and the NLoS AoD of the l-th NLoS path, respectively. We model the approximate reflection coefficient β_{RC} in dB unit as an independent Gaussian random variable, i.e., $10\log\beta_{\text{RC}}[\text{dB}] \sim \min\{\mathcal{N}(\mu_{\text{R}}, \sigma_{\text{R}}^2), 0\}$. The entries of $\tilde{\mathbf{h}}_{\text{LoS}}[k] \in \mathbb{C}^{1\times M^{\text{R}}}$ are given as $\{\tilde{\mathbf{h}}_{\text{LoS}}[k]\}_{m_r} = e^{-\text{j}2\pi f_k \cdot \frac{d^{(m_r)}}{c}}$, where $d^{(m_r)}$ denotes the 3D distance between the UE and the m_r-th RIS-side subarray. $\tilde{\mathbf{h}}^l[k]$ can be written by using a similar notation and assumptions.

3 Problem Formulation and Proposed Channel Estimation Solution

3.1 Problem Formulation of Channel Estimation

In this subsection, we formulate the downlink channel estimation problem for the considered RIS-assisted THz massive MIMO communication system over hybrid-field channels. As shown in Figure 2, we consider the two-stage frame structure consisting of the pilot training and data transmission stages. At the pilot training stage, the BS transmits M pilot OFDM symbols (i.e., M time

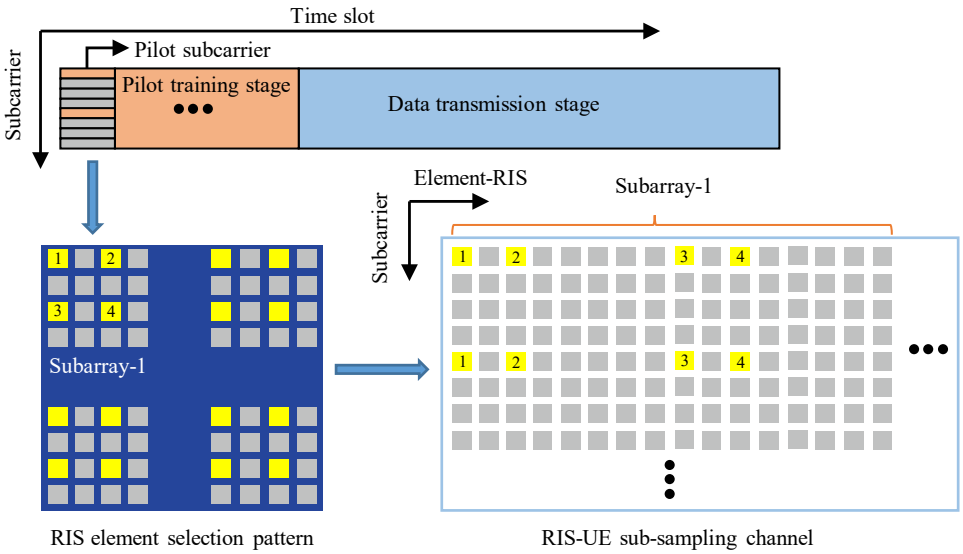


Figure 2: The schematic diagram of the frame structure, RIS element selection pattern, and RIS-UE sub-sampling channel, where the selected parts are marked in yellow blocks and the number in the yellow block indicates the index of the selected element.

slots) dedicated to channel estimation. The received signal at the UE-side[1] in the m-th time slot on the k-th subcarrier can be written as

$$y_m[k] = \sqrt{P_{\mathrm{T}}}\mathbf{h}[k]\mathbf{\Phi}_m\mathbf{G}[k]\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k]\mathbf{s}_m[k] + n_m[k], \tag{6}$$

where $1 \le k \le K$, $1 \le m \le M$, P_{T} is the transmit power of the BS, $\mathbf{s}_m[k] \in \mathbb{C}^{U\times 1}$ denotes the transmitted symbol vector with $\mathbb{E}\{\mathbf{s}_m[k]\mathbf{s}_m^{\mathrm{H}}[k]\} = \mathbf{I}_U$, and $n_m[k] \sim \mathcal{CN}(0, \sigma_n^2)$ is the effective complex additive white Gaussian noise (AWGN) at the UE, while $\mathbf{h}[k] \in \mathbb{C}^{1\times N^{\mathrm{R}}}$ and $\mathbf{G}[k] \in \mathbb{C}^{N^{\mathrm{R}}\times N^{\mathrm{B}}}$ are the downlink RIS-UE and BS-RIS channels on the k-th subcarrier, respectively. Denote the control vector $\mathbf{v}_{m_r,m} \in \mathbb{C}^{1\times N^{\mathrm{R}}_{\mathrm{sub}}}$ for the m_r-th subarray elements of the RIS in the m-th time slot as

$$\mathbf{v}_{m_r,m} = \mathbf{o}_{m_r,m} \odot \tilde{\mathbf{v}}_{m_r,m} = \left[\cdots, \eta_{n^r_{\mathrm{sub}},m_r,m}, \cdots\right] \odot \left[\cdots, e^{\mathrm{j}\phi_{n^r_{\mathrm{sub}},m_r,m}}, \cdots\right], \tag{7}$$

where $\mathbf{o}_{m_r,m} \in \mathbb{C}^{1\times N^{\mathrm{R}}_{\mathrm{sub}}}$ represents the amplitude control vector, $\tilde{\mathbf{v}}_{m_r,m} \in \mathbb{C}^{1\times N^{\mathrm{R}}_{\mathrm{sub}}}$ represents the phase control vector, and $1 \le n^r_{\mathrm{sub}} \le N^{\mathrm{R}}_{\mathrm{sub}}$, while $\eta_{n^r_{\mathrm{sub}},m_r,m} \in [0,\ 1]$ and $\phi_{n^r_{\mathrm{sub}},m_r,m} \in [0,\ 2\pi]$ are the amplitude control coefficient and phase control coefficient, respectively. $\eta_{n^r_{\mathrm{sub}},m_r,m}$ can be utilized to turn on/off the refraction functions of each RIS element. The entire RIS elements can be expressed as $\mathbf{v}_m = \mathbf{o}_m \odot \tilde{\mathbf{v}}_m = \left[\mathbf{v}_{1,m}, \cdots, \mathbf{v}_{m_r,m}, \cdots, \mathbf{v}_{M^{\mathrm{R}},m}\right]^{\mathrm{T}} \in \mathbb{C}^{N^{\mathrm{R}}\times 1}$, where $\mathbf{o}_m = \left[\mathbf{o}_{1,m}, \cdots, \mathbf{o}_{M^{\mathrm{R}},m}\right]^{\mathrm{T}} \in \mathbb{C}^{N^{\mathrm{R}}\times 1}$ and $\tilde{\mathbf{v}}_m = \left[\tilde{\mathbf{v}}_{1,m}, \cdots, \tilde{\mathbf{v}}_{M^{\mathrm{R}},m}\right]^{\mathrm{T}} \in \mathbb{C}^{N^{\mathrm{R}}\times 1}$. Then the RIS's refraction phase matrix is defined as $\mathbf{\Phi}_m = \mathrm{diag}(\mathbf{v}_m) = \mathbf{O}_m \odot \tilde{\mathbf{V}}_m \in \mathbb{C}^{N^{\mathrm{R}}\times N^{\mathrm{R}}}$, where $\mathbf{O}_m = \mathrm{diag}(\mathbf{o}_m) \in \mathbb{C}^{N^{\mathrm{R}}\times N^{\mathrm{R}}}$ is the RIS selection matrix and $\tilde{\mathbf{V}}_m = \mathrm{diag}(\tilde{\mathbf{v}}_m) \in \mathbb{C}^{N^{\mathrm{R}}\times N^{\mathrm{R}}}$ is the RIS phase matrix.

$\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N^{\mathrm{B}}\times M^{\mathrm{B}}}$ and $\mathbf{F}_{\mathrm{BB}}[k] \in \mathbb{C}^{M^{\mathrm{B}}\times U}$ are respectively analog and digital beamforming matrices that are used at the BS to perform hybrid beamforming. Due to the sub-connected hybrid MIMO architecture, the analog beamformer implemented by the phase shifters can be expressed as

$$\mathbf{F}_{\mathrm{RF}} = \mathrm{blkdiag}\big(\mathbf{f}_1, \cdots, \mathbf{f}_{m_b}, \cdots, \mathbf{f}_{M^{\mathrm{B}}}\big), \tag{8}$$

where $\mathbf{f}_{m_b} = \left[\mathrm{f}_{m_b,1}, \cdots, \mathrm{f}_{m_b,n^b_{\mathrm{sub}}}, \cdots, \mathrm{f}_{m_b,N^{\mathrm{B}}_{\mathrm{sub}}}\right]^{\mathrm{T}} \in \mathbb{C}^{N^{\mathrm{B}}_{\mathrm{sub}}\times 1}$ with $\left|\mathrm{f}_{m_b,n^b_{\mathrm{sub}}}\right|^2 = 1/N^{\mathrm{B}}_{\mathrm{sub}}$. Since the BS-RIS channel with the LoS path only is quasi-static and known, each analog beamforming vector can be designed as

$$\mathbf{f}_{m_b} = \mathbf{a}_{\mathrm{B}}(\theta_{\mathrm{B}}, \phi_{\mathrm{B}}, f_k),\ 1 \le m_b \le M^{\mathrm{B}}, \tag{9}$$

where k can be set to $K/2$ for alleviating the beam squint problem caused by the large bandwidth [36]. The digital beamformer $\mathbf{F}_{\mathrm{BB}}[k]$ is designed according to the zero forcing (ZF) precoding in order to eliminate the multi-stream interference between the BS and the RIS subarrays, i.e.,

$$\mathbf{F}_{\mathrm{BB}}[k] = \zeta\tilde{\mathbf{G}}^{\dagger}_{\mathrm{eq}}[k] = \zeta\tilde{\mathbf{G}}^{\mathrm{H}}_{\mathrm{eq}}[k]\left(\tilde{\mathbf{G}}_{\mathrm{eq}}[k]\tilde{\mathbf{G}}^{\mathrm{H}}_{\mathrm{eq}}[k]\right)^{-1}, \tag{10}$$

where $\tilde{\mathbf{G}}_{\mathrm{eq}}[k] = \big[\alpha[k]G_{\mathrm{T}}\tilde{\mathbf{G}}[k] \otimes \mathbf{a}^{\mathrm{H}}_{\mathrm{B}}(\theta_{\mathrm{B}}, \phi_{\mathrm{B}}, f_k)\big]\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{M^{\mathrm{R}}\times M^{\mathrm{B}}}$ is the equivalent BS-RIS channel obtained from the perspective of the first element of different subarrays at the RIS, and $\zeta = \sqrt{M^{\mathrm{B}}/\mathrm{Tr}\big\{\tilde{\mathbf{G}}^{\dagger}_{\mathrm{eq}}[k]\big(\tilde{\mathbf{G}}^{\dagger}_{\mathrm{eq}}[k]\big)^{H}\big\}}$ is a constant to meet the total transmit power constraint after beam-

[1]Note that since each UE can perform channel estimation independently, UE subscripts are omitted.

forming. In this way, the multi-stream interference between the BS and the RIS subarrays can be eliminated, i.e., $\mathbf{G}_{\mathrm{eq}}[k] = \mathbf{G}[k]\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k] \in \mathbb{C}^{N^{\mathrm{R}} \times U}$, $\forall k$, is a block diagonal constant matrix.

Therefore, the equivalent pilot signal $\mathbf{p}_m \in \mathbb{C}^{N^{\mathrm{R}} \times 1}$ can be expressed as

$$\mathbf{p}_m = \underbrace{\left[\mathbf{O}_m \odot \tilde{\mathbf{V}}_m\right]}_{\mathbf{\Phi_m}} \underbrace{\left[\mathbf{G}[k]\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k]\right]}_{\mathbf{G}_{\mathrm{eq}}[k]} \mathbf{s}_m[k], \tag{11}$$

where $\mathbf{p}_m$ is identical for different subcarriers since we set the transmit symbol $\mathbf{s}_m[k]$ to be $\mathbf{1}_U$, $\forall m, k$, and the ZF digital beamformer Equation (10) for $\mathbf{G}\,[k]$. Under the assumption that the BS and RIS meet the parallel symmetric array arrangements, $\mathbf{G}_{\mathrm{eq}}[k]$ is defined by $\sqrt{N^{\mathrm{B}}}\alpha[k]G_{\mathrm{T}}\mathrm{blkdiag}\big(\mathbf{1}^1_{N^{\mathrm{B}}_{\mathrm{sub}}}, \cdots, \mathbf{1}^u_{N^{\mathrm{B}}_{\mathrm{sub}}}, \cdots, \mathbf{1}^U_{N^{\mathrm{B}}_{\mathrm{sub}}}\big)$. Thus, the effective pilot signals can be further expressed as the RIS element vector given by $\mathbf{p}_m = \sqrt{N^{\mathrm{B}}}\alpha[k]G_{\mathrm{T}}\mathbf{v}_m \approx \sqrt{N^{\mathrm{B}}}\alpha G_{\mathrm{T}}\mathbf{v}_m = A_{\mathrm{T}}\mathbf{v}_m$, where the approximation $\alpha[k] \approx \alpha$, $\forall k$, is further applied and $A_{\mathrm{T}} = \sqrt{N^{\mathrm{B}}}\alpha G_{\mathrm{T}}$ represents the total attenuation from the BS to the RIS.

By collecting $y_m[k]$ for $1 \le m \le M$ together, the aggregate received signal vector $\mathbf{y}[k] = [y_1[k], \cdots, y_M[k]] \in \mathbb{C}^{1\times M}$ can be expressed as

$$\mathbf{y}[k] = \sqrt{P_{\mathrm{T}}}\mathbf{h}[k]\mathbf{P} + \mathbf{n}[k], \tag{12}$$

where $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_M] = A_{\mathrm{T}}\mathbf{V} = A_{\mathrm{T}}[\mathbf{v}_1, \cdots, \mathbf{v}_M] \in \mathbb{C}^{N^{\mathrm{R}}\times M}$, and $\mathbf{n}[k] = [n_1[k], \cdots, n_M[k]] \in \mathbb{C}^{1\times M}$. Thus, the received signal matrix $\mathbf{Y} = \left[\mathbf{y}^{\mathrm{T}}[1], \cdots, \mathbf{y}^{\mathrm{T}}[K]\right]^{\mathrm{T}} \in \mathbb{C}^{K\times M}$ can be expressed as

$$\mathbf{Y} = \sqrt{P_{\mathrm{T}}}\mathbf{H}\mathbf{P} + \mathbf{N}, \tag{13}$$

where $\mathbf{H} = \left[\mathbf{h}^{\mathrm{T}}[1], \cdots, \mathbf{h}^{\mathrm{T}}[K]\right]^{\mathrm{T}} \in \mathbb{C}^{K\times N^{\mathrm{R}}}$ denotes the downlink spatial-frequency domain RIS-UE channel matrix, and $\mathbf{N} = \left[\mathbf{n}^{\mathrm{T}}[1], \cdots, \mathbf{n}^{\mathrm{T}}[K]\right]^{\mathrm{T}} \in \mathbb{C}^{K\times M}$.

3.2 Deep Learning Based Spatial-Frequency Domain Channel Extrapolation

We choose to activate only $N^{\mathrm{R}}_s = \frac{N^{\mathrm{R}}}{\rho}$ RIS elements and estimate the sub-channels associated with the activated RIS elements, where $\rho > 1$ is the element compression ratio. Furthermore, as shown in Figure 2, only $K_s = \frac{K}{\bar{\rho}}$ uniformly selected subcarriers are used for pilot training, where $\bar{\rho}$ is the frequency compression ratio, and the remaining subcarriers can be used for transmitting control signals. We also give an example of the RIS element pattern selected uniformly and the corresponding RIS-UE side sub-sampling spatial-frequency channel in Figure 2, where the yellow blocks indicate the selected elements and the selected subcarriers. Thus, the practical received pilot signal $\mathbf{Y}_s \in \mathbb{C}^{K_s\times M}$ is defined as

$$\mathbf{Y}_s = \sqrt{P_{\mathrm{T}}}\mathbf{H}_s\mathbf{P}_s + \mathbf{N}_s, \tag{14}$$

where $\mathbf{H}_s \in \mathbb{C}^{K_s\times N^{\mathrm{R}}_s}$ is the sub-sampling of the spatial-frequency channel, $\mathbf{P}_s \in \mathbb{C}^{N^{\mathrm{R}}_s\times M}$ is the corresponding equivalent pilot signal, and $\mathbf{N}_s$ is the noise. Our goal is to recover the complete channel $\hat{\mathbf{H}} \in \mathbb{C}^{K\times N^{\mathrm{R}}}$ from the limited received pilot signals $\mathbf{Y}_s$, i.e., extrapolating the rest unknown

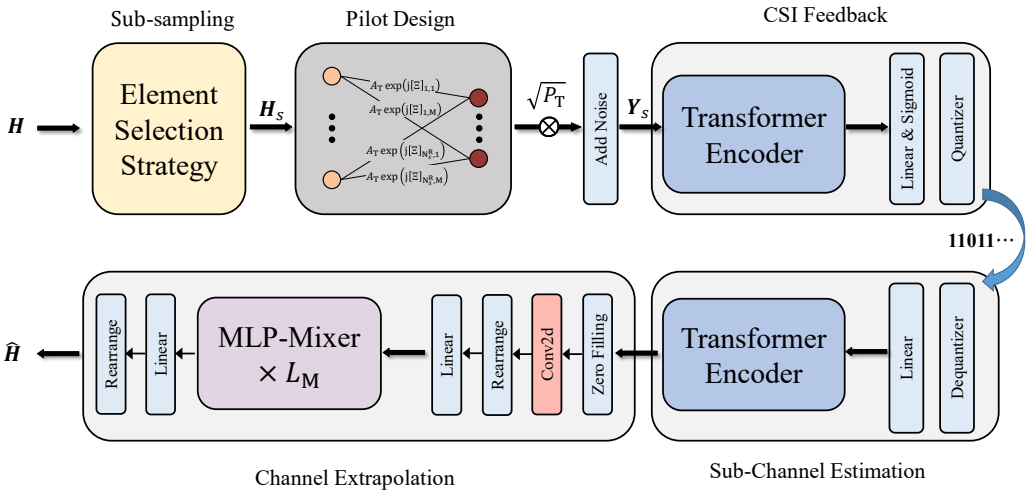


Figure 3: The overall structure of the proposed DL-based spatial-frequency domain channel extrapolation scheme.

channels from the acquired partial channels. Based on the universal approximation capability of DL, we can use DL to characterize the mapping among the channels at different space/frequency locations. Thus, we propose a DL-based spatial-frequency domain channel extrapolation network, which consists of the element selection strategy (ESS), pilot design, CSI feedback, sub-channel estimation, and spatial-frequency domain channel extrapolation modules, as illustrated in Figure 3. The overall flow of the proposed scheme is expressed as

$$\hat{\mathbf{H}} = f_{\mathrm{SFDE}}(f_{\mathrm{SCE}}(f_{\mathrm{CsiFd}}(\sqrt{P_{\mathrm{T}}}\mathbf{H}_s\mathbf{P}_s + \mathbf{N}_s))) = f_{\mathrm{SFDE}}(f_{\mathrm{SCE}}(f_{\mathrm{CsiFd}}(\sqrt{P_{\mathrm{T}}} f_{\mathrm{ESS}}(\mathbf{H})\,\mathbf{P}_s + \mathbf{N}_s))), \quad (15)$$

where the mapping $f_{\mathrm{ESS}}(\cdot)$ represents the element selection strategy for deciding the sub-sampling channel, and the equivalent pilot signal $\mathbf{P_s}$ can be learned as the trainable parameters, while $f_{\mathrm{CsiFd}}(\cdot)$, $f_{\mathrm{SCE}}(\cdot)$ and $f_{\mathrm{SFDE}}(\cdot)$ represent the CSI feedback network, the sub-channel estimation network and the spatial-frequency domain extrapolation network, respectively. We now detail each component.

3.2.1 Element Selection Strategy

With only N_s^{R} activated RIS elements, from (7), the RIS element selection vector $\mathbf{o} = \mathbf{o}_m = \left[\mathbf{o}_{1,m}, \cdots, \mathbf{o}_{m_r,m}, \cdots, \mathbf{o}_{M^{\mathrm{R}},m}\right]^{\mathrm{T}} \in \{0,1\}^{N^{\mathrm{R}}\times 1}$ is an N_s^{R}-hot vector with N_s^{R} elements being '1' and the other elements being '0', where the subscript 'm' can be dropped since $\mathbf{o}$ is fixed at the pilot training stage. Also since only K_s subcarriers are uniformly selected for pilot training, the frequency selection vector $\boldsymbol{\kappa} \in \{0,1\}^{K\times 1}$ is defined by $\{\boldsymbol{\kappa}\}_{\bar{\rho}k+1} = 1$, $0 \le k \le K_s - 1$, and the other elements being '0'. Thus, the selection operation of the sub-sampling function $f_{\mathrm{ESS}}(\cdot)$ can be expressed as

$$\mathbf{H}_s = f_{\mathrm{ESS}}(\mathbf{H}) = \mathbf{S} \odot \mathbf{H}, \quad (16)$$

where $\mathbf{S} = \boldsymbol{\kappa} \otimes \mathbf{o}^{\mathrm{T}} \in \{0,1\}^{K\times N^{\mathrm{R}}}$ is the spatial-frequency selection matrix, and the zero rows/columns in $\mathbf{S} \odot \mathbf{H}$ are deleted directly to yield $\mathbf{H}_s$. Note that the selection of the activated elements, i.e., the RIS element selection vector, can affect the extrapolation performance. Thus, we consider the

following three element selection strategies.

1) Uniform Selection Strategy: Since each subarray in the RIS is a UPA, its element compression ratio is expressed as $\rho = \rho_y \times \rho_z$. For fairness, ρ_y and ρ_z should be as close as possible. When $\rho_y \neq \rho_z$, the z-axis compression should be considered first to maintain the y-axis angular resolution of indoor UEs. Thus, the y-z compression ratio allocation can be solved from the following optimization

$$\begin{aligned} \min_{\{\rho_y,\rho_z\}} \quad & |\rho_z - \rho_y|, \\ \text{s.t.} \quad & \rho_y \times \rho_z = \rho, \\ & 1 \leq \rho_y \leq \rho_z. \end{aligned} \tag{17}$$

Some allocation examples as $\rho(2,4,8,16) = \rho_y(1,2,2,4) \times \rho_z(2,2,4,4)$. Given ρ_y and ρ_z, the active element index vector $\boldsymbol{\xi}_{m_r} \in \mathbb{C}^{1 \times N^{\mathrm{R}}_{\mathrm{sub}}/\rho}$ of the m_r-th subarray can be expressed as

$$\{\boldsymbol{\xi}_{m_r}\}_{n_i^y N_z^{\mathrm{R}}/\rho_z + n_i^z + 1} = N^{\mathrm{R}}_{\mathrm{sub}}(m_r - 1) + N_z^{\mathrm{R}} \rho_y n_i^y + \rho_z n_i^z + 1, \tag{18}$$

where $1 \leq m_r \leq M^{\mathrm{R}}$, $0 \leq n_i^y \leq \frac{N_y^{\mathrm{R}}}{\rho_y} - 1$, and $0 \leq n_i^z \leq \frac{N_z^{\mathrm{R}}}{\rho_z} - 1$. The entire active element index vector or set of the RIS is defined as $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{m_r}, \cdots, \boldsymbol{\xi}_{M^{\mathrm{R}}}]^{\mathrm{T}} \in \mathbb{C}^{N_s^{\mathrm{R}} \times 1}$. Thus, we set the entries of the RIS element selection vector $\mathbf{o}$ corresponding to the index set $\boldsymbol{\xi}$ to '1', i.e., $\{\mathbf{o}\}_\xi = 1$ for $\xi \in \boldsymbol{\xi}$, and the other elements of $\mathbf{o}$ to '0'.

2) Random Selection Strategy: It randomly selects N_s^{R} elements from the RIS as the random pattern. We can use the function 'random.choice(·)' of the NumPy library in Python to generate the active element index vector $\boldsymbol{\xi}$. Given that the elements of the pattern are randomly selected from the whole RIS, if the element compression ratio ρ is not large, then the aperture of a random pattern is usually comparable to that of the RIS.

3) Learning-Based Selection Strategy: In addition to the above two fixed selection strategies, the learning-based element selection strategy has also been widely studied. In [14], a differentiable selection network is proposed to learn the element selection vector $\mathbf{o}$. The input of this network is a random initialization vector. By utilizing several fully-connected layers and the softmax function, a probability vector $\mathbf{g} = \{g_1, g_2, \cdots, g_{N^{\mathrm{R}}}\}$ is generated, where g_i represents the probability of the i-th element being selected. Thus, the active element index vector $\boldsymbol{\xi}$ can be defined as

$$\boldsymbol{\xi} = \arg\ \mathrm{top}_{N_s^{\mathrm{R}}}\{\mathbf{g}\}, \tag{19}$$

where $\arg\ \mathrm{top}_{N_s^{\mathrm{R}}}\{\cdot\}$ is a function that finds the element index set of the first N_s^{R} largest selection probabilities. The details of the antenna selection network can be found in [14].

3.2.2 Pilot Design

As aforementioned in Equation (11), under the assumption that the BS and RIS meet the parallel symmetric array arrangements, the equivalent downlink pilots can be defined as $\mathbf{P}_s = A_{\mathrm{T}} \mathbf{V}_s$, where $\mathbf{V}_s \in \mathbb{C}^{N_s^{\mathrm{R}} \times M}$ is the phase matrix of selected RIS elements at the pilot training stage. Thus, the pilot matrix $\mathbf{P}_s$ can be obtained by adjusting the RIS phase at different time slots, which is

given by

$$\mathbf{P}_s = A_{\mathrm{T}} \exp^{(\mathrm{j}\boldsymbol{\Xi})} = A_{\mathrm{T}} \left(\cos(\boldsymbol{\Xi}) + \mathrm{j}\sin(\boldsymbol{\Xi})\right), \tag{20}$$

where $\boldsymbol{\Xi} \in \mathbb{R}^{N_s^{\mathrm{R}} \times M}$ is the phase control matrix of selected RIS elements. As it is well known that complex-valued outputs are not well supported by most deep learning frameworks (e.g., Tensorflow, Pytorch), it would be difficult to directly train the complex-valued pilot matrix $\mathbf{P}_s$. Hence, we adopt the real-valued RIS phase control matrix $\boldsymbol{\Xi}$ whose entries take values in $[0,\ 2\pi)$ as the real-valued trainable parameters of the pilot design network (PDN) and the pilot matrix $\mathbf{P}_s$ can be obtained from Equation (20). The structure of the PDN is shown in Figure 3, where the trainable parameters of the PDN, i.e., $\boldsymbol{\Xi}$, are learned at the DL training stage.

3.2.3 CSI Feedback

At the uplink CSI feedback stage, the UE extracts the CSI from the received pilot signals and feeds it back to the BS. However, the large number of array elements results in excessive feedback overhead. Recently, DL-based solutions, such as CsiNet [37], have achieved good performance for CSI feedback. Furthermore, an emerging DL architecture, known as the transformer [38], is utilized as the novel CSI feedback network to further reduce the feedback overhead and obtain more efficient compression performance than the CsiNet framework [39]. Therefore, in this paper, we utilize the transformer as the backbone of the CSI feedback network $f_{\mathrm{CsiFd}}(\cdot)$. The original transformer is composed of an encoder and a decoder. However, since we are dealing with the CSI without time-sequential information, there is no causality constraint. Thus, we only exploit the encoder module of the transformer which obtains output in parallel. Since neural networks are more effective for real-valued operations than complex-valued operations and the transformer can only extract the

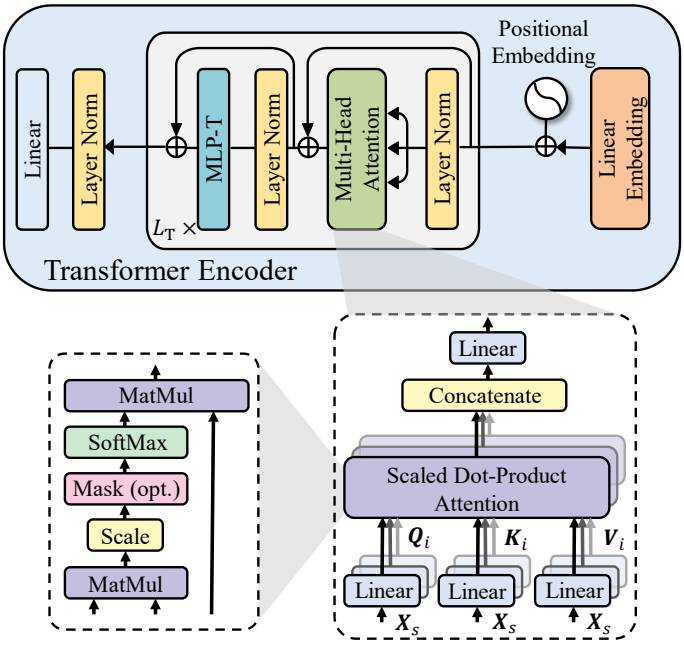


Figure 4: The schematic diagram of the transformer encoder.

correlation between sequences, we reshape the received pilot signal into a real-valued two-dimensional (2D) sequence $\bar{\mathbf{Y}}_s \in \mathbb{R}^{K_s \times 2M}$, which can be expressed as

$$\bar{\mathbf{Y}}_s = [\Re\{\mathbf{Y}_s\}, \Im\{\mathbf{Y}_s\}], \tag{21}$$

where the number of subcarriers K_s represents the input sequence length of the transformer.

The schematic diagram of the transformer encoder is shown in Figure 4. Through the fully-connected linear embedding layer, the input sequence $\bar{\mathbf{Y}}_s$ is converted to $\mathbf{X}_s \in \mathbb{R}^{K_s \times d_\mathrm{T}}$, which merges the relative position information of the sub-carriers using the positional embedding layer. Then, multiple encoder layers are utilized to extract features from the input sequences. Each encoder layer has the same structure which is composed of a multi-head self-attention sub-layer followed by a position-wise multi-layer perceptron (MLP) sub-layer. Layernorm is applied before every block and the residual connection is applied after every block. Among them, the multi-head attention mechanism is the key component for the performance improvement of the transformer. As shown in Figure 4, the input sequence $\mathbf{X}_s$ is first projected onto three different sequential vectors: the queries, keys, and values with different learned linear projections, respectively, namely, $\{\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\} \in \mathbb{R}^{K_s \times d_m}$, $1 \le i \le h$, where h is the number of heads and $d_m = d_\mathrm{T}/h$. Then, each value $\mathrm{head}_i \in \mathbb{R}^{K_s \times d_m}$, $1 \le i \le h$, is outputted by performing the scaled dot-product attention in parallel, where a softmax function is applied to obtain the weight on the value, which is given by

$$\mathrm{head}_i = \mathrm{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^\mathrm{T}}{\sqrt{d_m}}\right)\mathbf{V}_i,\ 1 \le i \le h. \tag{22}$$

These output values are concatenated and projected back to a d_T-dimensional representation using the linear projection matrix $\mathbf{W}^O \in \mathbb{R}^{K_s \times d_\mathrm{T}}$ as

$$\mathrm{MultiHead}(\mathbf{X}_\mathrm{s}) = \mathrm{Concat}(\mathrm{head}_i, \cdots, \mathrm{head}_h)\mathbf{W}^O. \tag{23}$$

After the transformer encoder, a fully-connected linear layer and sigmoid function are used to generate a real-valued compressed codeword. The codeword is then converted to B bits as the feedback information through a quantization layer, which is constructed by a B-bit uniform scalar quantizer. The above feedback process generates the binary vector $\mathbf{q} \in \{0,1\}^B$ as

$$\mathbf{q} = f_\mathrm{CsiFd}\left(\bar{\mathbf{Y}}_s; \mathcal{W}_F\right), \tag{24}$$

where $\mathcal{W}_F$ denotes the trained parameter set of the CSI feedback network.

3.2.4 Sub-Channel Estimation

When the BS receives the feedback bits, the sub-channel estimation network is used to reconstruct the sub-sampling of the complete spatial-frequency channel. Similar to Subsection 3.2.3, we also consider the transformer encoder as the backbone of this part. As shown in Figure 3, the received CSI feedback bit vector is first inputted into a dequantization layer, which conducts the

inverse operation of the quantizer and outputs a real-valued vector. Then, a fully-connected layer is utilized to obtain an initial coarse channel estimation. Finally, the transformer encoder extracts the spatial-frequency correlation of the channel and further improves the channel estimation performance. The output of the sub-channel estimation network is expressed as

$$\bar{\mathbf{H}}_s = \left[\Re\{\hat{\mathbf{H}}_s\}, \Im\{\hat{\mathbf{H}}_s\}\right] = f_{\mathrm{SCE}}\left(\mathbf{q}; \mathcal{W}_S\right), \tag{25}$$

where $\hat{\mathbf{H}}_s \in \mathbb{C}^{K_s \times N_s^{\mathrm{R}}}$ is the estimated sub-sampling channel, $\bar{\mathbf{H}}_s \in \mathbb{R}^{K_s \times N_s^{\mathrm{R}} \times 2}$ is a real-valued 3D matrix, and $\mathcal{W}_S$ is the trained parameter set of the sub-channel estimation network.

3.2.5 Spatial-Frequency Domain Channel Extrapolation

First, the initial input $\tilde{\mathbf{H}} \in \mathbb{R}^{K \times N^{\mathrm{R}} \times 2}$ to the channel extrapolation network is constructed from the estimated sub-sampling channel $\bar{\mathbf{H}}_s \in \mathbb{R}^{K_s \times N_s^{\mathrm{R}} \times 2}$ with the known RIS spatial-frequency selection pattern $\mathbf{S}$. Specifically, we copy the entries of $\bar{\mathbf{H}}_s$ to the corresponding positions in $\tilde{\mathbf{H}}$ and fill the other elements of $\tilde{\mathbf{H}}$ with zeros according to the known RIS spatial-frequency selection pattern $\mathbf{S}$. This initial operation can be expressed as

$$\tilde{\mathbf{H}} = f_{\mathrm{zfi}}\left(\bar{\mathbf{H}}_s; \mathbf{S}\right). \tag{26}$$

The non-zero rows/columns of $\tilde{\mathbf{H}}$ are consistent with $\bar{\mathbf{H}}_s$ and their locations are the same as the locations of elements '1' in $\mathbf{S}$. The neighborhood information in the receptive field is then extracted using a convolutional layer for initial interpolation. To guarantee that the output dimensions from the convolution layer remain unchanged, we employ zero padding in the convolution layer.

Subsequently, we consider a competitive yet conceptually and technically simple architecture, called MLP-Mixer [40], as the backbone of the channel extrapolation network. The architecture of this MLP-Mixer is based entirely on MLPs, which can extract and reconstruct 2D features by repeatedly applying them to either spatial locations or feature channels. Specifically, the input $\tilde{\mathbf{H}} \in \mathbb{R}^{K \times N^{\mathrm{R}} \times 2}$ is rearranged as a sequence of flattened 2D patches $\mathbf{X}_p \in \mathbb{R}^{N_p \times (2L^2)}$, where (K, N^{R}) is the resolution of the original input, (L, L) is the resolution of each patch, and $N_p = KN^{\mathrm{R}}/L^2$ is the resulting number of patches. Then, all the patches are linearly projected with the same projection matrix. This results in a 2D real-valued matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{N_p \times d_{\mathrm{M}}}$. Next the input matrix $\tilde{\mathbf{X}}$ is fed into

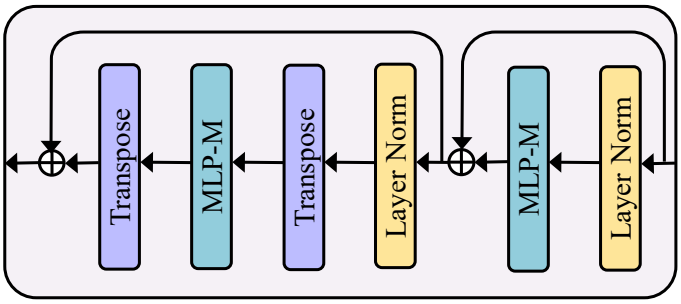


Figure 5: The structure of the mixer layer.

several mixer layers to extrapolate the complete channel. As illustrated in Figure 5, each mixer layer consists of two MLP blocks. The first one acts on the columns of $\tilde{\mathbf{X}}$, maps $\mathbb{R}^{N_p} \mapsto \mathbb{R}^{2N_p} \mapsto \mathbb{R}^{N_p}$, and is shared across all the columns. The second acts on the rows of $\tilde{\mathbf{X}}$, i.e., on the transposed input matrix $\tilde{\mathbf{X}}^{\mathrm{T}}$, maps $\mathbb{R}^{d_{\mathrm{M}}} \mapsto \mathbb{R}^{2d_{\mathrm{M}}} \mapsto \mathbb{R}^{d_{\mathrm{M}}}$, and is shared across all the rows. Each MLP block contains two fully-connected layers and a nonlinear activation function. Denote the input matrix to the t-th mixer layer as $\tilde{\mathbf{X}}_t$. The mapping of the t-th mixer layer can be expressed as

$$\begin{aligned} \mathbf{U} &= \tilde{\mathbf{X}}_t + \mathbf{W}_{t,2} f_\sigma(\mathbf{W}_{t,1}\mathrm{LayerNorm}(\tilde{\mathbf{X}}_{\mathrm{t}})), \\ \tilde{\mathbf{X}}_{t+1} &= \mathbf{U} + \left(\mathbf{W}_{t,4} f_\sigma(\mathbf{W}_{t,3}\mathrm{LayerNorm}(\mathbf{U})^{\mathrm{T}})\right)^{\mathrm{T}}, \end{aligned} \tag{27}$$

where $\mathbf{W}_{t,i}$, $1 \le i \le 4$, are the parameter matrices of the fully-connected layers in the t-th mixer layer for $1 \le t \le L_{\mathrm{M}}$, and L_{M} is the number of mixer layers, while f_σ is the activation function.

Finally, the output of the last mixer layer is linearly projected back to the original dimension $\mathbb{R}^{N_p \times d_{\mathrm{M}}} \mapsto \mathbb{R}^{N_p \times (2L^2)}$, and the 2D patches are rearranged back to $\mathbb{R}^{N_p \times (2L^2)} \mapsto \mathbb{R}^{K \times N^{\mathrm{R}} \times 2}$ for obtaining the final extrapolation result $\bar{\mathbf{H}} \in \mathbb{R}^{K \times N^{\mathrm{R}} \times 2}$, which is a real-valued 3D matrix. Thus, the extrapolation process can be expressed as

$$\hat{\mathbf{H}} = \bar{\mathbf{H}}_{[:,:,1]} + \mathrm{j}\bar{\mathbf{H}}_{[:,:,2]} = f_{\mathrm{SFDE}}\left(\bar{\mathbf{H}}_s; \mathcal{W}_E\right), \tag{28}$$

where $\hat{\mathbf{H}} \in \mathbb{C}^{K \times N^{\mathrm{R}}}$ is the estimated complete complex-valued channel, and $\mathcal{W}_E$ is the trained parameter set of the spatial-frequency domain extrapolation network.

3.2.6 Training Strategy

The data set for off-line training is denoted as $\boldsymbol{\mathcal{H}}$, where $|\boldsymbol{\mathcal{H}}| = N_{\mathrm{set}}$ is the number of off-line training samples. Furthermore, a sample in $\boldsymbol{\mathcal{H}}$ is an input-label pair written as ($\mathbf{H}$, $\mathbf{H}$), where $\mathbf{H}$ is the extrapolation target and is also the input of the entire SFDCEtra network. The input will go through the RIS array element and subcarrier sub-sampling strategy, since we need to extrapolate the original complete channel by receiving only the pilot signal of the sub-sampling channel.

With the uniform or random ESS $f_{\mathrm{ESS}}(\cdot)$, at the off-line training stage, we consider E2E training to jointly optimize the pilot design network, CSI feedback network, sub-channel estimation network, and channel extrapolation network, by minimizing the normalized mean square error (NMSE) between the output $\hat{\mathbf{H}}$ and the target $\mathbf{H}$. Thus, the loss function is written as

$$\mathcal{L}_c = \frac{1}{B_e} \sum_{i=1}^{B_e} \frac{\|\mathbf{H} - \hat{\mathbf{H}}\|_F^2}{\|\mathbf{H}\|_F^2}, \tag{29}$$

where B_e is the batch size for off-line training.

When the learning-based ESS is adopted, the parameters for the ESS and the above networks are optimized jointly. The loss function for the joint optimization problem is given by

$$\mathcal{L} = \gamma \mathcal{L}_c + (1 - \gamma)\mathcal{L}_{\mathrm{ESS}}, \tag{30}$$

where $0 < \gamma \leq 1$ is the weight to balance the penalties of channel extrapolation and ESS with $\gamma = 1$ denoting that the non-learning based $f_{\mathrm{ESS}}(\cdot)$ is selected, and $\mathcal{L}_{\mathrm{ESS}}$ is the loss function of the learning-based ESS. The details of $\mathcal{L}_{\mathrm{ESS}}$ are available in [14].

4 Proposed Beamforming Solution

4.1 Problem Formulation of RIS-aided Multi-User Beamforming

At the data transmission stage, the BS can simultaneously supports U UEs with the aid of RIS, since the LoS MIMO architecture between the BS and RIS can support multi-stream transmission via intra-path multiplexing. Similar to Equation (6), the received signal at the u-th UE on the k-th subcarrier can be expressed as

$$\begin{aligned} y[u,k] =& \sqrt{P_{\mathrm{T}}}\mathbf{h}[u,k]\mathbf{\Phi}\mathbf{G}[k]\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}}[u,k]s[u,k] \\ &+ \sum_{i=1,i\neq u}^{U} \sqrt{P_{\mathrm{T}}}\mathbf{h}[i,k]\mathbf{\Phi}\mathbf{G}[k]\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}}[i,k]s[i,k] + n[u,k], \end{aligned} \tag{31}$$

where $\mathbf{h}[u,k] \in \mathbb{C}^{1\times N^{\mathrm{R}}}$, $1 \leq u \leq U, 1 \leq k \leq K$, represents the downlink channel vector between the RIS and the u-th UE on the k-th subcarrier, $\mathbf{f}_{\mathrm{BB}}[u,k] \in \mathbb{C}^{M^{\mathrm{B}}\times 1}$ denotes the digital baseband beamforming vector associated with the u-th UE on the k-th subcarrier. Thus, the signal-to-interference plus-noise-ratio (SINR) of the u-th UE on the k-th subcarrier can be expressed as

$$\mathrm{SINR}[u,k] = \frac{P_{\mathrm{T}}|\mathbf{h}[u,k]\mathbf{\Phi}\mathbf{G}[k]\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}}[u,k]|^2}{P_{\mathrm{T}}\sum\limits_{i=1,i\neq u}^{U}|\mathbf{h}[i,k]\mathbf{\Phi}\mathbf{G}[k]\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}}[i,k]|^2 + \sigma_n^2}. \tag{32}$$

Therefore, the sum rate R in the downlink multi-user transmission can be expressed as

$$R = \frac{1}{K}\sum_{u=1}^{U}\sum_{k=1}^{K}\log_2\left(1+\mathrm{SINR}[u,k]\right). \tag{33}$$

Based on the estimated RIS-UE channel at the pilot training stage, the BS can design the hybrid beamformer $\{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}[k], \forall k\}$ and the RIS refraction phase matrix $\mathbf{\Phi}$ to maximize the sum rate R, where $\mathbf{F}_{\mathrm{BB}}[k] = \left[\mathbf{f}_{\mathrm{BB}}[1,k], \cdots, \mathbf{f}_{\mathrm{BB}}[U,k]\right]$. This design process is formulated as the following optimization problem

$$\begin{aligned} \max_{\mathcal{F}(\cdot)} \quad & R, \\ \text{s.t.} \quad & \{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}[k], \forall k, \mathbf{\Phi}\} = \mathcal{F}\left(\hat{\mathbf{H}}[u], \forall u\right), \\ & \mathbf{F}_{\mathrm{RF}} \in (8), \\ & \|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k]\|_F^2 = M^{\mathrm{B}}, \forall k, \\ & \{\mathbf{\Phi}\}_{i,i} = \{\mathbf{v}\}_i = e^{\mathrm{j}\phi_i}, \phi_i \in [0,\ 2\pi), \forall i, \end{aligned} \tag{34}$$

where $\hat{\mathbf{H}}[u]$ is the estimated spatial-frequency channel between the RIS and the u-th UE, and

$\mathcal{F}(\cdot)$ represents a function that maps the estimated RIS-UE channels onto the hybrid beamformer $\{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}[k], \forall k\}$ and the RIS refraction phase matrix $\mathbf{\Phi}$.

4.2 Deep Learning Based Hybrid Beamforming and RIS Phase Design

In order to solve the optimization Equation (34), some alternating iterative algorithms [17, 18, 19] have been proposed to obtain the analog beamformer, digital beamformer, and RIS phase, respectively. Unfortunately, all the aforementioned approaches are based on the idealized case that the CSI is known accurately. However, perfect CSI is usually unavailable, especially for indoor channel cases where the channel characteristics are complex due to rich scatterers. Inspired by the universal approximation capability of DL, it is possible to use DL to learn the complicated and unknown mapping from the estimated channels to the hybrid beamformers and RIS refraction phase. Thus, we propose a DL-based hybrid beamforming and RIS phase design scheme, which consists of analog beamformer design, DL-based RIS refraction phase design, and knowledge-data dual-driven digital beamformer design. The block diagram of the proposed scheme is shown in Figure 6.

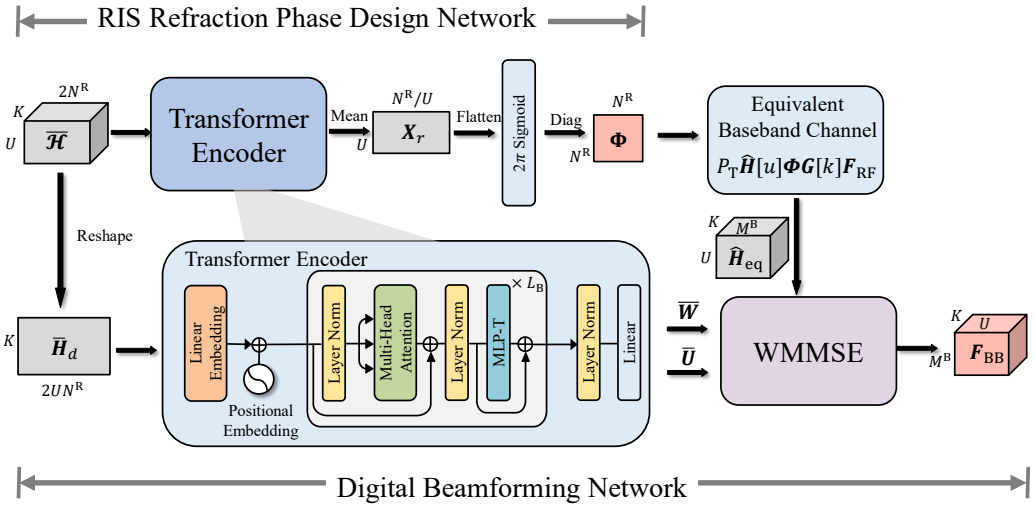


Figure 6: The overall structure of the proposed DL-based hybrid beamforming and RIS refraction phase design scheme.

4.2.1 Analog Beamformer Design

Since the BS's active beamforming and RIS's passive beamforming are coupled, the optimization problem is non-convex, and it is very challenging to find a global optimum. Hence, we separately design the analog beamforming of the BS and the passive beamforming of the RIS. Specifically, the BS analog beamforming and the RIS passive beamforming are designed to improve the received SINR of UEs. However, due to the sub-connected structure of phase shifters in the LoS MIMO architecture, the interference among beams from the BS subarrays to the RIS subarrays cannot be eliminated. Fortunately, this part of interference can be removed by appropriately designing the digital beamforming. Therefore, when designing the analog beamforming on the BS-side, it is sufficient to assume that the transmit energy of the BS is focused on the RIS.

Since the BS-RIS channel with only LoS path is quasi-static and known, we can utilize the angle information of the BS-RIS link to design the analog beamformer. Without loss of generality, we assume that the u-th UE is assisted by the m_r-th subarray of the RIS. Thus, the active beamforming designed for the u-th UE should be aligned to the m_r-th subarray of the RIS. Therefore, the transmit beam of the m_b-th subarray $\mathbf{f}_{m_b}$ is designed according to Equation (9).

4.2.2 DL-Based RIS Refraction Phase Design

The key challenge in the RIS-aided communication system is to optimize a common RIS phase shared by all the subcarriers. In the THz broadband case, there exists a non-negligible beam squint effect for different subcarriers [36]. Therefore, when designing the common RIS phase, it is necessary to consider this effect on all subcarriers, which makes the RIS phase design much more difficult than the narrowband case. To solve this challenging problem, we propose a transformer-based RIS phase design network (RPDN), as shown in Figure 6, to design the RIS refraction phase matrix.

We first convert all the estimated RIS-UE channels $\hat{\mathbf{H}}[u] \in \mathbb{C}^{K\times N^{\mathrm{R}}}$ for $1 \le u \le U$ into a real-valued 3D matrix $\bar{\boldsymbol{\mathcal{H}}} \in \mathbb{R}^{U\times K\times 2N^{\mathrm{R}}}$, which is expressed as

$$\bar{\boldsymbol{\mathcal{H}}} = \left[\bar{\mathbf{H}}[1], \cdots, \bar{\mathbf{H}}[u], \cdots, \bar{\mathbf{H}}[U]\right], \tag{35}$$

where $\bar{\mathbf{H}}[u] = [\Re\{\hat{\mathbf{H}}[u]\}, \Im\{\hat{\mathbf{H}}[u]\}] \in \mathbb{R}^{K\times 2N^{\mathrm{R}}}$ and $\hat{\mathbf{H}}[u]$ is the estimated RIS-UE channel of the u-th UE obtained from the DL-based SFDCEtra network. $\bar{\boldsymbol{\mathcal{H}}}$ is inputted into the transformer encoder, which globally extracts the inter-subcarrier correlation. To consider the beam squint effect for different subcarriers, the 2D matrix $\mathbf{X}_r \in \mathbb{R}^{U\times N^{\mathrm{R}}/U}$ is obtained by the mean operation over the subcarrier dimension of the transformer encoder's output. Then $\mathbf{X}_r$ is flattened as $\mathbf{x}_r \in \mathbb{R}^{N^{\mathrm{R}}\times 1}$, and passes through the activation function to generate the RIS phase vector $\mathbf{v} \in \mathbb{C}^{N^{\mathrm{R}}\times 1}$ that satisfies the the constant modulus constraint. The corresponding activation function is defined as

$$\mathbf{v} = e^{\mathrm{j}2\pi\cdot\mathrm{Sigmoid}(\mathbf{x}_r)}. \tag{36}$$

Finally, the RIS phase matrix $\boldsymbol{\Phi} \in \mathbb{C}^{N^{\mathrm{R}}\times N^{\mathrm{R}}}$ is obtained through diagonalization. The overall process of the RIS refraction phase design, namely, the transformer-based RPDN, can be expressed as

$$\boldsymbol{\Phi} = f_{\mathrm{RIS}}\left(\bar{\boldsymbol{\mathcal{H}}}; \mathcal{W}_R\right), \tag{37}$$

where $f_{\mathrm{RIS}}(\cdot)$ denotes the mapping of the RPDN, whose trainable parameter set is $\mathcal{W}_R$.

4.2.3 Knowledge-Data Dual-Driven Digital Beamformer Design

With the known BS-RIS channel $\mathbf{G}[k]$, the designed RIS refraction phase matrix $\boldsymbol{\Phi}$ and the analog beamforming matrix $\mathbf{F}_{\mathrm{RF}}$ as well as the estimated RIS-UE channel $\hat{\mathbf{h}}[u,k]$, the BS can obtain the estimated equivalent baseband channel $\hat{\mathbf{h}}_{\mathrm{eq}}[u,k] \in \mathbb{C}^{1\times M^{\mathrm{B}}}$ as

$$\hat{\mathbf{h}}_{\mathrm{eq}}[u,k] = P_T\hat{\mathbf{h}}[u,k]\boldsymbol{\Phi}\mathbf{G}[k]\mathbf{F}_{\mathrm{RF}}. \tag{38}$$

The true equivalent baseband channel $\mathbf{h}_{\mathrm{eq}}[u,k]$ has the similar form to Equation (38), given the designed $\mathbf{\Phi}$ and $\mathbf{F}_{\mathrm{RF}}$. Thus, the optimization problem Equation (34) can be simplified as

$$\begin{aligned} \max_{\mathbf{F}_{\mathrm{BB}}[k],\forall k} \quad & \frac{1}{K}\sum_{u=1}^{U}\sum_{k=1}^{K}\log_2\left(1+\mathrm{SINR}[u,k]\right), \\ \text{s.t.} \quad & \mathrm{SINR}[u,k]=\frac{|\mathbf{h}_{\mathrm{eq}}[u,k]\mathbf{f}_{\mathrm{BB}}[u,k]|^2}{\sum\limits_{i=1,i\neq u}^{U}|\mathbf{h}_{\mathrm{eq}}[i,k]\mathbf{f}_{\mathrm{BB}}[i,k]|^2+\sigma_n^2}, \\ & \|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k]\|_F^2=M^{\mathrm{B}},\ \forall k. \end{aligned} \tag{39}$$

The above problem is a classic baseband beamforming problem, which can be solved with standard liner beamforming schemes, such as the regularized ZF (RZF) or iterative weighted minimum mean-square error (WMMSE) algorithm.

The iterative WMMSE algorithm solves the optimization (39) by solving the MMSE problem given in (40) below, which has the identical optimal solution $\mathbf{F}_{\mathrm{BB}}[k],\forall k$ to the problem (39).

$$\begin{aligned} \max_{\bar{\mathbf{U}},\bar{\mathbf{W}},\mathbf{F}_{\mathrm{BB}}[k],\forall k} \quad & \sum_{u=1}^{U}\sum_{k=1}^{K}\left(\bar{w}_{u,k}e_{u,k}-\log_2\bar{w}_{u,k}\right), \\ \text{s.t.} \quad & \|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k]\|_F^2\leq M^{\mathrm{B}},\ \forall k, \end{aligned} \tag{40}$$

where $\bar{w}_{u,k}=\{\bar{\mathbf{W}}\}_{u,k}$ is the weight of the u-th user on the k-th subcarrier, $e_{u,k}=\mathbb{E}\{|\hat{s}[u,k]-s[u,k]|^2\}$ is the MSE between the transceiver symbols under the independence assumption of $s[u,k]$ and $n[u,k]$, while $\hat{s}[u,k]=\bar{u}_{u,k}y[u,k]$ denotes the estimated data symbol at the UE-side, and $\bar{u}_{u,k}=\{\bar{\mathbf{U}}\}_{u,k}$ is the receiver gain of the u-th UE on the k-th subcarrier. According to [41], the above problem is convex in individual optimization variable. Hence each of the optimization subproblems has a closed-form solution given the other optimization variables, and a block coordinate descent (BCD) iterative algorithm is adopted to solve the optimization (40). This algorithm is summarized in Algorithm 1, where we omit the iteration index t on the variables for clarity.

However, the iterative WMMSE algorithm typically imposes a large number of iterations with

Algorithm 1 Iterative WMMSE beamforming design algorithm

1: **Initialize** $\mathbf{F}_{\mathrm{BB}}[k]$ that meets $\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k]\|_F^2=M^{\mathrm{B}}$, set the maximum iteration number $I_{\max}$, and the current iteration index $t=0$;
2: **repeat**
3: **Update** $\{\bar{\mathbf{U}}\}_{u,k}$: $\bar{u}_{u,k}=\Big(\sum\limits_{i=1}^{U}|\mathbf{h}_{\mathrm{eq}}[u,k]\mathbf{f}_{\mathrm{BB}}[i,k]|^2+\sigma_n^2\Big)^{-1}\mathbf{h}_{\mathrm{eq}}[u,k]\mathbf{f}_{\mathrm{BB}}[u,k]$, $\forall u,k$;
4: **Update** $\{\bar{\mathbf{W}}\}_{u,k}$: $\bar{w}_{u,k}=\big(1-\bar{u}_{u,k}^{*}\mathbf{h}_{\mathrm{eq}}[u,k]\mathbf{f}_{\mathrm{BB}}[u,k]\big)^{-1}$, $\forall u,k$;
5: **Update** $\mathbf{f}_{\mathrm{BB}}[u,k]$: $\mathbf{f}_{\mathrm{BB}}[u,k]=\bar{u}_{u,k}\bar{w}_{u,k}\Big(\sum\limits_{i=1}^{U}\bar{w}_{i,k}|\bar{u}_{i,k}|^2\mathbf{h}_{\mathrm{eq}}^{\mathrm{H}}[i,k]\mathbf{h}_{\mathrm{eq}}[i,k]+\mu_k\mathbf{I}\Big)^{-1}\mathbf{h}_{\mathrm{eq}}^{\mathrm{H}}[u,k]$, where $\mu_k=\sum\limits_{j=1}^{U}\frac{\sigma^2}{M^{\mathrm{B}}}\bar{w}_{j,k}|\bar{u}_{j,k}|^2$, $\forall u,k$;
6: $t=t+1$;
7: **until** $t\geq I_{\max}$
8: Scale $\mathbf{F}_{\mathrm{BB}}[k]$ to meet the transmit power constraint.

long running time. Furthermore, the BS can only acquire the imperfect estimated CSI $\hat{\mathbf{h}}_{\mathrm{eq}}[u,k]$, and it is difficult for the traditional digital beamforming algorithms, such as Algorithm 1, to overcome the interference induced by the imperfect CSI. Thus, we propose the knowledge-data dual-driven digital beamforming network, as shown in Figure 6, which utilizes the transformer encoder to directly learn the parameters of the iterative WMMSE algorithm from the imperfect CSI for better interference elimination and shorter running time.

Specifically, the real-valued 3D matrix $\bar{\boldsymbol{\mathcal{H}}}$ is reshaped into a 2D matrix $\bar{\mathbf{H}}_d \in \mathbb{C}^{K\times 2UN^{\mathrm{R}}}$, which is inputted into the transformer encoder. The output of the transformer encoder $\mathbf{X} \in \mathbb{C}^{K\times 4U}$ is converted into the weight matrix $\bar{\mathbf{W}}$ and the receiver gain matrix $\bar{\mathbf{U}}$, i.e.,

$$\bar{\mathbf{W}} = \mathbf{X}^{\mathrm{T}}_{[:,:U]} + \mathrm{j}\mathbf{X}^{\mathrm{T}}_{[:,U:2U]}, \tag{41}$$

$$\bar{\mathbf{U}} = \mathbf{X}^{\mathrm{T}}_{[:,2U:3U]} + \mathrm{j}\mathbf{X}^{\mathrm{T}}_{[:,3U:]}. \tag{42}$$

Then, we can obtain $\mathbf{F}_{\mathrm{BB}}[k]$, $\forall k$, based on the learned $\bar{\mathbf{W}}$ and $\bar{\mathbf{U}}$ by the update function of $\mathbf{f}_{\mathrm{BB}}[u,k]$, i.e., line 5 of Algorithm 1. Compared with the iterative WMMSE beamforming design, our proposed scheme does not involve an iterative process so that running time can be reduced significantly. To satisfy the transmit power constraint, the normalization operation can be expressed as

$$\mathbf{F}_{\mathrm{BB}}[k] = \frac{\sqrt{M^{\mathrm{B}}}\mathbf{F}_{\mathrm{BB}}[k]}{\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k]\|_F}, \forall k. \tag{43}$$

The proposed knowledge-data dual-driven digital beamformer design can be expressed as

$$\{\mathbf{F}_{\mathrm{BB}}[k], \forall k\} = f_{\mathrm{DBF}}\left(\bar{\mathbf{H}}_d; \mathcal{W}_D\right), \tag{44}$$

where $f_{\mathrm{DBF}}(\cdot)$ is the map of the digital beamforming network with a trainable parameter set $\mathcal{W}_D$.

4.2.4 Training Strategy

We take every U channel samples (i.e., the channels of U UEs) in the training set of the channel estimation stage as a group to form a training set at the beamforming design stage, which is denoted as $\boldsymbol{\mathcal{H}}_U$. The number of off-line training samples is $|\boldsymbol{\mathcal{H}}_U| = N_{\mathrm{set}}/U$. A sample in $\boldsymbol{\mathcal{H}}_U$ is an UE set $\{\mathbf{H}[u], 1 \le u \le U\}$, where $\mathbf{H}[u]$ is the spatial-frequency channel between the RIS and the u-th UE.

$\{\mathbf{H}[u], 1 \le u \le U\}$ are inputted to the trained SFDCEtra network to obtain the estimated channels $\{\hat{\mathbf{H}}[u], 1 \le u \le U\}$, which form the input to the proposed network. Since imperfect CSI will reduce the sum rate upper bound, to ensure a faster learning process, we apply a teacher forcing technique [42] at the early stage of training by feeding the perfect CSI $\{\mathbf{H}[u], \forall u\}$ to the proposed network. At the off-line training stage, we consider E2E training to jointly optimize the proposed hybrid beamforming and RIS phase design network, i.e., the parameters of the entire network are trained by minimizing the negative sum rate. Thus, the loss function is written as

$$\mathcal{L}_b = -\frac{1}{B_b}\sum_{i=1}^{B_b} R, \tag{45}$$

where R is the sum rate defined in Equation (33) and B_b is the batch size for off-line training.

5 Results and Discussion

In this section, we evaluate the performance of the proposed spatial-frequency domain channel extrapolation scheme as well as hybrid beamforming and RIS phase design for the RIS-aided THz massive MIMO system through numerical simulations.

5.1 Simulation Settings

5.1.1 Communication Scenario Set up

In simulations, the BS is deployed on the top of a building of height 30 m, and the RIS is installed on a window surface on one floor of another building. As shown in Figure 1(b), the BS (RIS) is equipped with $M^{\mathrm{B}} = M_y^{\mathrm{B}} M_z^{\mathrm{B}} = 4$ ($M^{\mathrm{R}} = M_y^{\mathrm{R}} M_z^{\mathrm{R}} = 4$) subarrays on the yz-plane, where $M_y^{\mathrm{B}} = 2$ ($M_y^{\mathrm{R}} = 2$) and $M_z^{\mathrm{B}} = 2$ ($M_z^{\mathrm{R}} = 2$). Each subarray is a UPA with $N_{\mathrm{sub}}^{\mathrm{B}} = N_y^{\mathrm{B}} N_z^{\mathrm{B}} = 64$ ($N_{\mathrm{sub}}^{\mathrm{R}} = N_y^{\mathrm{R}} N_z^{\mathrm{R}} = 64$) isotropically radiating elements, where $N_y^{\mathrm{B}} = 8$ ($N_y^{\mathrm{R}} = 8$) and $N_z^{\mathrm{B}} = 8$ ($N_z^{\mathrm{R}} = 8$). Therefore, the number of elements of the complete array at the BS (RIS) is $N^{\mathrm{B}} = M^{\mathrm{B}} N_{\mathrm{sub}}^{\mathrm{B}} = 256$ ($N^{\mathrm{R}} = M^{\mathrm{R}} N_{\mathrm{sub}}^{\mathrm{R}} = 256$). For simplicity, we assume that the BS and RIS meet the parallel symmetric array arrangement with a distance of $D = 20$ m. The central frequency is $f_c = 0.3$ THz with bandwidth $f_s = 1$ GHz. The number of OFDM subcarriers is set to $K = 128$ and the antenna gain of the BS is $G_T = 10$ dBi. Given the above parameter settings, the subarray intervals of both the BS and

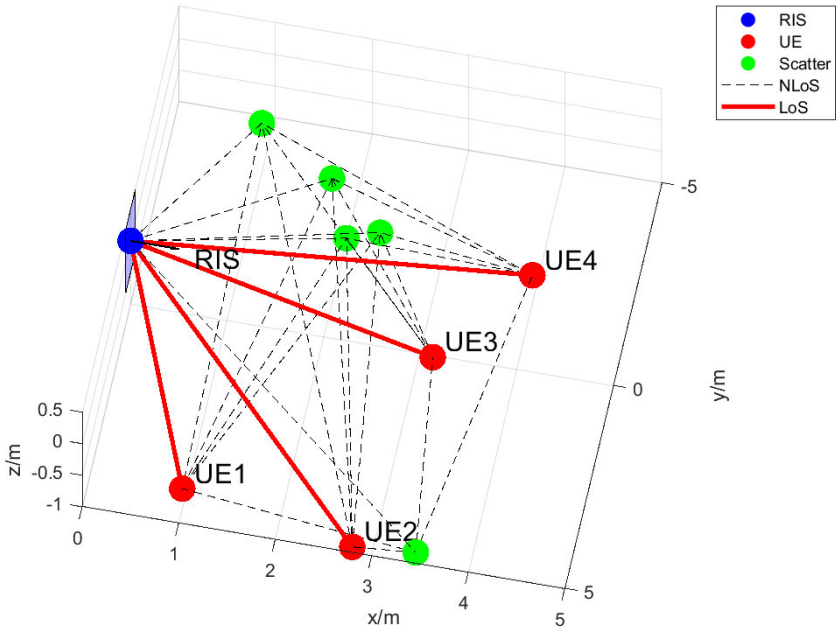


Figure 7: Multi-ray THz channel model for the indoor scenario: the NLoS rays are reflected by the scatters.

the RIS are calculated from Equation (1) as $d_{sy}^{\mathrm{B}}, d_{sz}^{\mathrm{B}}, d_{sy}^{\mathrm{R}}, d_{sz}^{\mathrm{R}} = 96.5\lambda$ for obtaining the multi-stream multiplexing gain over the LoS path.

Figure 7 depicts the schematic diagram of the fixed scattering environment, where the positions of the RIS, UEs, and scatterers are marked by blue, red, and green circles, respectively. The red solid line represents the LoS link between the RIS and an UE, and the black dotted line indicates the NLoS link via an scatterer. We assume that $U = 4$ UEs are randomly distributed over the xy-plane of the rectangular room ($W_x = 5\,\mathrm{m}$, $W_y = 10\,\mathrm{m}$), and the height of UEs is 1 m lower than the RIS. Due to rich scatterers for indoor environment as well as the high scattering and diffraction losses in the THz band, the number of available NLoS paths (scatterers) in the THz indoor channel is set to $L_p = 5$, implying that only a single-bounce scattering mode is considered. We set the parameters of the reflection coefficient β_{RC} as $\mu_{\mathrm{R}} = -5$, $\sigma_{\mathrm{R}} = 2$. The noise power spectrum density at the UEs is $\sigma_{\mathrm{NSD}}^2 = -174$ dBm/Hz. Thus, the power of the AWGN is $\sigma_n^2 = \sigma_{\mathrm{NSD}}^2 f_s/K = -105$ dBm. The RIS-UE channel samples are generated using Equation (5), where the UEs and scatterers are distributed randomly each time.

5.1.2 SFDCEtra Network Parameter Configuration

In the CSI feedback network, the linear embedding layer of the transformer encoder has $d_{\mathrm{T}} = 256$ neurons. In the transformer encoder, the number of the encoder layers is $L_{\mathrm{T}} = 3$, where the number of heads is $h = 8$ and the position-wise MLP sub-layer has 2 fully-connected layers with $4d_{\mathrm{T}}$ and d_{T} neurons, respectively, while the dimension of the output linear layer is $2M$. In the sub-channel estimation network, the linear layer is a $2N_s^{\mathrm{R}}$-dimensional fully-connected layer and the hyperparameters of the transformer encoder are the same as those of the CSI feedback network. As for the channel extrapolation network, the output of the zero filling is processed by the convolutional layer with the kernel size of 7×7 and the number of filters is 2. The patch size of the rearranged operation is $L = 16$, the number of patches is $N_p = 128$ and the number of neurons in the linear layer is $d_{\mathrm{M}} = 512$. We set the number of mixer layers as $L_{\mathrm{M}} = 6$, where each mixer layer consists of two MLP blocks, and the numbers of neurons in the MLP blocks are set to $2N_p$, N_p, $2d_{\mathrm{M}}$, and d_{M}, respectively. The above structural parameters of the SFDCEtra network are empirically found to be appropriate.

We divide the data set into the training set, validation set, and test set, which contain 102400, 10240, and 10240 samples, respectively. Unless otherwise specified, the uniform element selection strategy is adopted in the simulations. When considering the learning-based element selection strategy, the weight factor γ is set to 0.9. At the network training stage, the Adam optimizer is adopted to update the network weight parameters and the learning rate varies depending on the *warmup* mechanism [38]. We set the batch size of the training set to 512, and 200 epochs for training.

5.1.3 HBFRPD Network Parameter Configuration

Again we determine appropriate structural parameters of the HBFRPD network empirically. Specifically, in the RIS phase design network, the linear embedding layer of the transformer encoder has $d_{\mathrm{B}} = 128$ neurons. In the transformer encoder, the number of the encoder layers is $L_{\mathrm{B}} = 3$,

where the number of heads is $h = 8$ and the position-wise MLP sub-layer has 2 fully-connected layers with $4d_{\mathrm{B}}$ and d_{B} neurons, respectively, while the output linear layer of the transformer encoder has $N^{\mathrm{R}}/U = 64$ neurons. In the digital beamforming network, the hyperparameters of the transformer encoder are the same as those of the RIS phase design network, and the output linear layer of the transformer encoder has $4U$ neurons.

We take each U channel samples as a group to form a data set, which is divided into the training set, validation set, and testing set, which contain 25600, 2560, and 2560 samples, respectively. We set the batch size of the training set to 32, and 180 epochs for training.

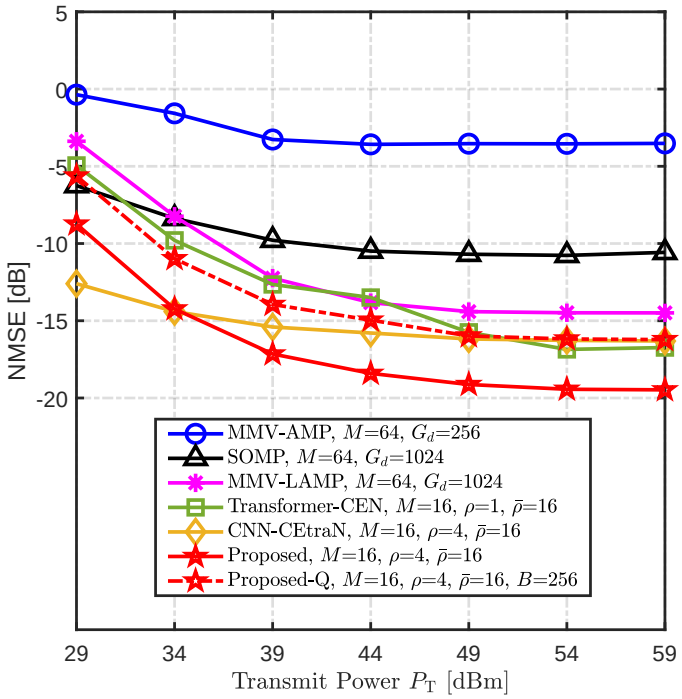


Figure 8: NMSE performance comparison of different channel estimation schemes versus transmit power P_{T}.

5.2 DL-Based Spatial-Frequency Domain Channel Extrapolation

Since the RIS element can only passively receive EM waves, selecting partial elements of array would reduce the signal energy radiated into the room. For the fair comparison between different schemes, we adopt the same transmit power instead of the same SNR as the comparison criterion to avoid ignoring the performance differences induced by the number of activated RIS elements. Specifically, as shown in Figure 8, we plot the NMSE performance of the different schemes as a function of transmit power P_{T}. The number of NLoS paths is $L_p = 5$. We consider three model-based channel estimation benchmark algorithms, namely, the simultaneous orthogonal match pursuit (SOMP) algorithm [43], the multiple-measurement-vector approximate message passing (MMV-AMP) algorithm [44], and the model-driven DL-based channel estimation scheme using the MMV learned AMP (MMV-LAMP) network [45], which utilize $M = 64$ OFDM symbols on all subcarriers and then di-

rectly estimate the complete channel. For the SOMP and MMV-LAMP schemes, the redundant dictionary with an oversampling ratio of 4 is considered to further improve the performance, i.e., the number of codewords is $G_d = 1024$. Note that the MMV-AMP scheme requires the measurement matrix's elements to be independent and identically distributed, and hence we cannot consider the redundant dictionary (i.e., $G_d = N^{\mathrm{R}} = 256$). Since data-driven DL algorithms have the potential to achieve better performance, we also compare our proposed DL-based SFDCEtra network with the transformer-based channel estimation network (Transformer-CEN) [39] and the CNN-based channel extrapolation network (CNN-CEtraN) [16]. For these methods, we set the number of OFDM symbols to $M = 16$ and the subcarrier compression ratio to $\bar{\rho} = 16$. The transformer-based scheme collects the signals from all RIS elements, i.e., $\rho = 1$, and directly estimates the complete channel. Both the CNN-based and our proposed channel extrapolation schemes consider the element compression ratio of $\rho = 4$ to perform partial channel extrapolation. Note that for fairness, the above model- and data-driven algorithms do not consider the quantization of CSI feedback information. Therefore, we additionally consider the proposed scheme with $B = 256$ feedback bits generated via a 2-bit quantizer, denoted as 'Proposed-Q'.

It can be observed from Figure 8 that the proposed channel extrapolation scheme outperforms the other channel estimation schemes considerably in terms of NMSE performance while imposing a smaller pilot overhead. This is because exploiting the spatial-frequency correlation allows our DL-based channel extrapolation scheme to recover the unobserved channel part from the estimated low-dimensional sub-channel, thus reducing the training overhead while improving the NMSE performance. In particular, our extrapolation scheme significantly improves the NMSE performance compared with the state-of-the-art CNN-based channel extrapolation scheme. Unlike local percep-

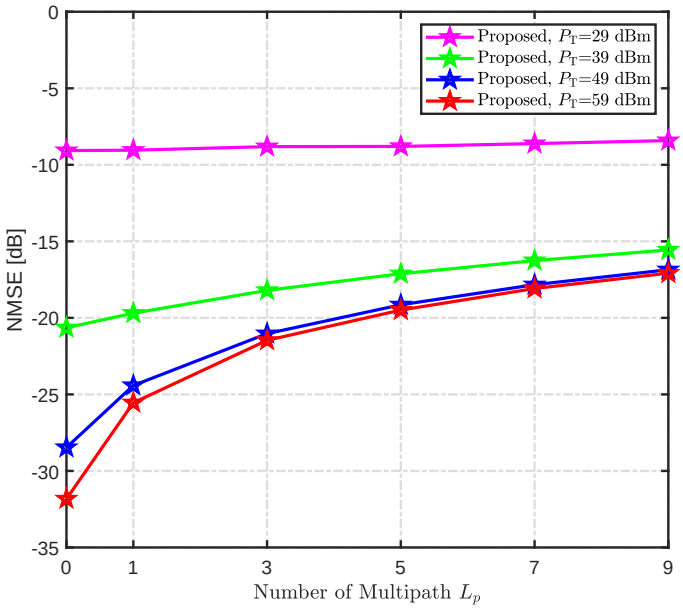


Figure 9: NMSE performance comparison of the proposed scheme versus the number of multipath L_p, given $\rho = 4$, $\bar{\rho} = 16$ and $M = 16$. Offline training is based on the channel samples with $L_p = 5$ multipath components.

tion in CNN, the MLP-mixer is utilized as the backbone of our channel extrapolation module and it can extract the global features of the channel for enhanced extrapolation accuracy. Considering the actual situation of finite quantized feedback, we can see that our proposed scheme with 2-bit quantizer, 'Proposed-Q', can still achieve very good performance. These results demonstrate that the proposed channel extrapolation scheme can learn latent features from the data more effectively to achieve better channel estimation accuracy with less pilot and feedback overhead.

We further investigate the robustness of the proposed channel extrapolation scheme as a function of the number of multipath L_p in Figure 9. Note that the proposed DL-based channel extrapolation scheme is trained at the offline training stage, based on the channel samples with $L_p = 5$ multipath components. It can be clearly seen from Figure 9 that at the online estimation stage, the proposed scheme can adapt to estimate multipath channels with $L_p \neq 5$, without having to retrain the entire network architecture. Therefore, the proposed DL-based channel extrapolation enjoys better robustness and generalization capability to various channel conditions.

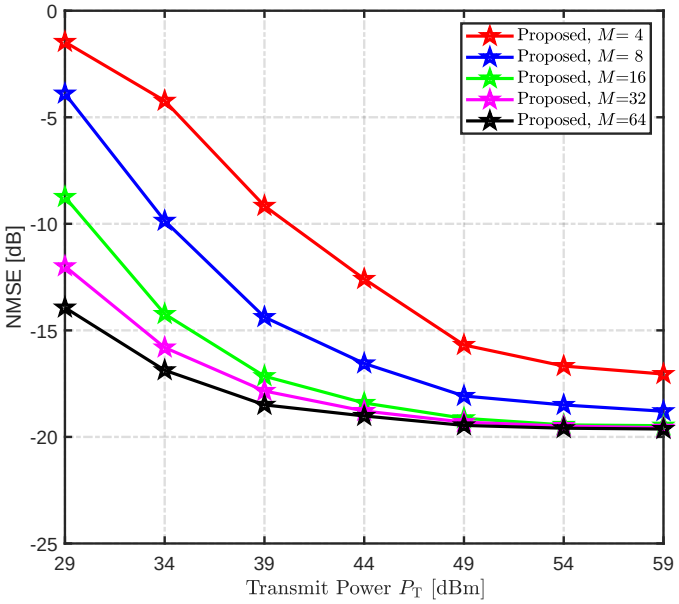


Figure 10: NMSE performance comparison of the proposed scheme with different pilot numbers versus transmit power P_{T}, given $\rho = 4$, $\bar{\rho} = 16$, $L_p = 5$.

In Figure 10, we investigate the channel extrapolation NMSE performance of our proposed scheme with different numbers of pilot OFDM symbols, $M = 4, 8, 16, 32$ and 64. As expected, the channel extrapolation performance improves with the increase of the number of pilot OFDM symbols. This is because more pilot OFDM symbols can improve the accuracy of sub-channel estimation, thus reducing the error propagation and improving the reconstruction of the extrapolation module. Furthermore, we can see that the proposed scheme can provide more significant performance gain by increasing the number of pilot OFDM symbols in the case of low transmit power. This is because the increase in the number of observations can improve the received SNR.

Figure 11 depicts the NMSE performance of the proposed DL-based channel extrapolation scheme versus the element compression ratio ρ, with three ESEs. Specifically, the curve labeled by 'Uniform'

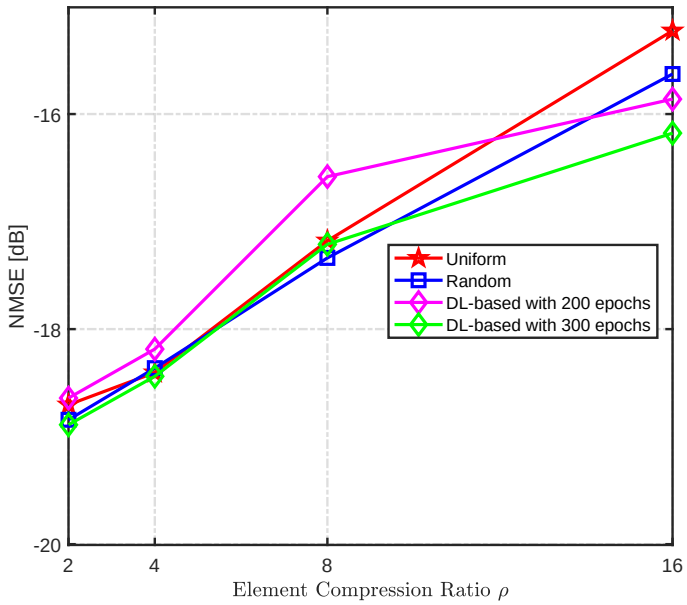


Figure 11: NMSE performance comparison of the proposed scheme with different element selection strategies versus element compression ratio ρ, given $\bar{\rho} = 16$, $L_p = 5$, $M = 16$, $P_{\mathrm{T}} = 44\,\mathrm{dBm}$.

corresponds to the uniform selection strategy, the curve labeled by 'Random' represents the random selection strategy, while the other two marked by 'DL-based with 200 epochs' and 'DL-based with 300 epochs' use the DL-based element selection strategy. As expected, the NMSE improves as the element compression ratio ρ decreases. This is largely due to two reasons: 1) As the number of selected RIS elements increases, or the element compression ratio ρ decreases, the received signal power will increase, thus improving the estimation accuracy of channel extrapolation input (i.e., sub-channel estimate), and 2) The received pilot signal can provide more channel information when more RIS elements are selected. However, this does not imply that we can obtain the best performance by choosing the lowest element compression ratio (or performing complete observations directly without extrapolation). Indeed, the channel extrapolation performance heavily depends on the number of transmission resources, the accuracy of the sub-channel estimate, and the number of selected RIS elements (i.e., the dimension of the sub-channel). Only when the transmission resources are sufficient, can the gain provided by more selected RIS elements be seen clearly. Moreover, it can be seen that the performance difference between different element selection strategies is not obvious at low compression ratios. Only at high compression ratio ($\rho > 8$), can the performance difference be seen clearly as 'Uniform' < 'Random' < 'DL-based'. Since the aperture of the random pattern is usually larger than that of the uniform pattern, the random selection strategy is slightly better than that of the uniform selection strategy. The performance of the DL-based approach is relatively better than that of the first two approaches only when its training reaches sufficient epochs, as the learning of the selection network needs sufficient number of epochs to converge.

To fully illustrate the effectiveness of the proposed DL-based channel extrapolation solution, we verify its channel extrapolation module separately. To do so, we fix the compression ratio of RIS elements to 4, i.e., $N_s^{\mathrm{R}} = 64$. First, the least squares (LS), the SOMP, and the proposed

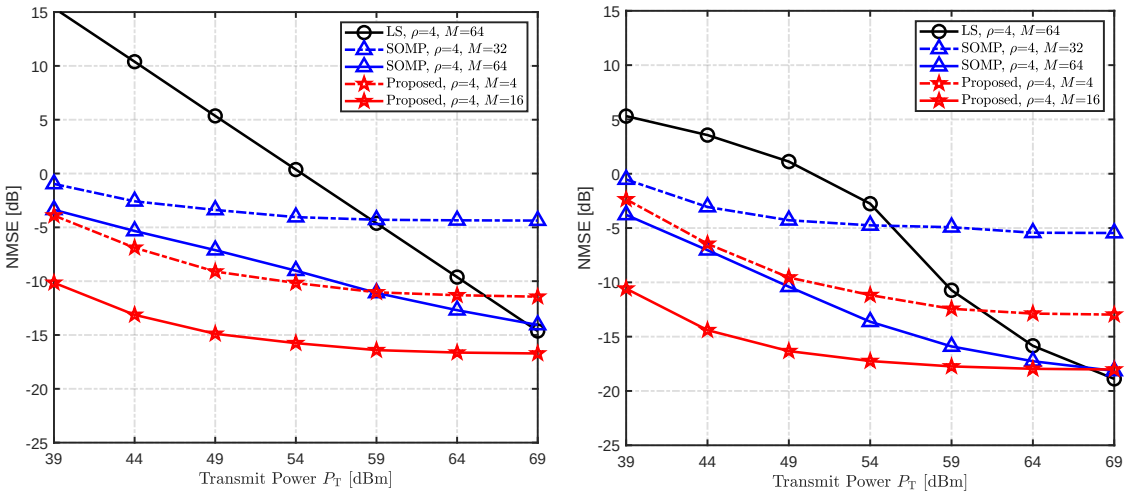


Figure 12: (a) NMSE performance comparison of different sub-channel estimation schemes versus transmit power P_{T}; and (b) NMSE performance of channel extrapolation versus transmit power P_{T} for different sub-channel estimation schemes.

transformer-based algorithm are utilized for sub-channel estimation, and the results are shown in Figure 12(a). Observe that the NMSE of the SOMP-based sub-channel estimation with $M = 64$ pilot symbols is significantly better than that of the LS-based sub-channel estimation with $M = 64$ pilot symbols, particularly at low transmit power P_{T}. Furthermore, the NMSE of our transformer-based sub-channel estimation algorithm with only $M = 16$ pilot symbols is considerably better than that of the SOMP-based sub-channel estimation with $M = 64$ pilot symbols. Then, we input the sub-channels estimated by different algorithms into the trained channel extrapolation network $f_{\mathrm{SFDE}}(\cdot)$, which outputs the estimation of the complete channel. The corresponding results are shown in Figure 12(b). Observe that the NMSE performance of the complete channel extrapolated from our channel extrapolation network is even better than the NMSE of the estimated low-dimensional sub-channel, without any additional pilot overhead. This shows that our proposed channel extrapolation network can not only be used for DL-based communication architecture, but also be combined with traditional algorithms to significantly reduce resource overhead. Therefore, we conclude that the proposed DL-based spatial-frequency domain channel extrapolation scheme can significantly reduce the pilot overhead while achieving the same or better channel estimation NMSE performance.

5.3 DL-Based Hybrid Beamforming and RIS Phase Design

Figure 13 shows the sum rates achieved by different schemes under the perfect CSI case. We considered two comparison schemes, both of which adopt the analog beamforming design discussed in Subsection 4.2.1 as well as the beam alignment-based RIS phase design. In the beam alignment-based RIS phase design, each subarray of the RIS selects one UE and performs beam alignment according to the RIS-UE CSI on the central subcarrier to concentrate and refract the signal energy to it, while ignoring interference to other UEs. For digital beamforming design, these two comparison

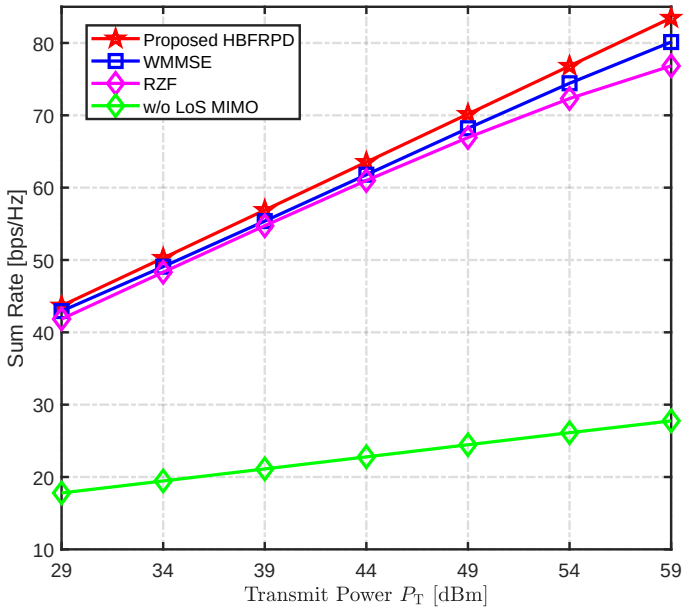
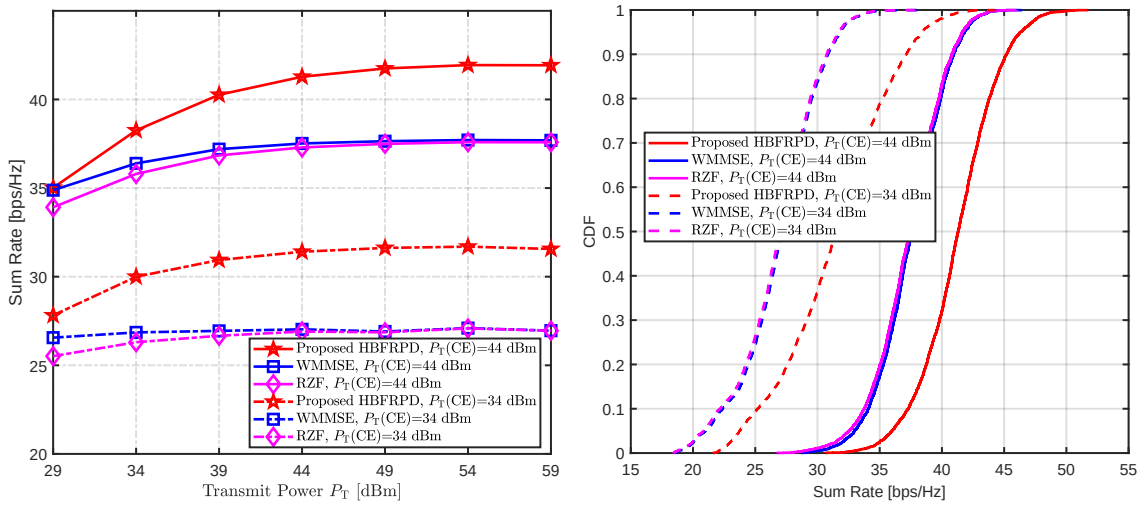


Figure 13: Sum rates achieved by different schemes versus transmit power P_{T} given the perfect CSI. The actual transmit power of the 'w/o LoS MIMO' case is $UP_{\mathrm{T}} = 4P_{\mathrm{T}}$.

schemes adopt the RZF and iterative WMMSE algorithms, respectively, thus they are abbreviated as 'RZF' and 'WMMSE', respectively. It can be observed that our proposed HBFRPD scheme has the better performance than other schemes and the superiority is more evident as the transmit power increases. In addition, the proposed HBFRPD scheme does not require to obtain $\mathbf{F}_{\mathrm{BB}}[k], \forall k$, in an iterative manner. Thus, it runs much faster than the iterative WMMSE algorithm. We also analyze the performance gain provided by LoS MIMO architecture. Considering the case of the BS and the RIS w/o LoS MIMO array structure (i.e., both use UPA arrays), the BS-RIS channel is a single LoS path with rank 1, which only provides single stream data transmission. To ensure fairness, the transmit power of the 'w/o LoS MIMO' case is equal to the total transmit power of the 'w/ LoS MIMO' cases, i.e., the transmit power of the 'w/o LoS MIMO' case is actually $UP_{\mathrm{T}} = 4P_{\mathrm{T}}$. By calculating the sum rate, we obtain the green curve in Figure 13. It can be seen that the sum rate in the 'w/ LoS MIMO' case is much higher than that of 'w/o LoS MIMO' case. This is because the LoS MIMO architecture can increase the sum rate linearly benefited from extra spatial multiplexing gain, while 'w/o LoS MIMO' case can only provide log-level growth as the SINR increases.

Although most schemes can achieve high sum rate performance under perfect CSI, the sum rate of multi-users is actually limited by the inter-user interference induced by CSI error. Figure 14(a) illustrates the sum rate performance of the different schemes with imperfect CSIs estimated at two different transmit powers $P_{\mathrm{T}}(\mathrm{CE})$. Compared with the case of perfect CSI, the sum rate degrades significantly with the decrease of CSI estimation accuracy, i.e., with the decrease of the transmit power at the channel estimation stage. It can be clearly seen that due to the inter-user interference induced by CSI errors, the sum rates of the RZF and iterative WMMSE schemes barely increases with transmit power. Moreover, our proposed HBFRPD scheme exhibits a significant performance gain over the RZF and iterative WMMSE algorithms in the presence of CSI estimation errors. This



(a) (b)

Figure 14: (a) Sum rates achieved by different schemes versus transmit power P_{T} under the imperfect CSI case, and (b) The CDFs of the sum rates achieved by different schemes under the imperfect CSI case, given $P_{\mathrm{T}} = 44\,\mathrm{dBm}$. We have $\rho = 4$, $\bar{\rho} = 16$, $L_p = 5$ and $M = 16$.

result indicates that our proposed scheme is capable of mitigating the interference caused by CSI errors and hence has better robustness to inaccurate CSI than the other schemes.

Figure 14(b) shows the cumulative distribution functions (CDFs) that characterize the sum rate performance achieved by the different schemes. Here, we consider the transmit power $P_{\mathrm{T}} = 44\,\mathrm{dBm}$ at the data transmission stage. It can be seen from Figure 14(b) that when the transmit power is $P_{\mathrm{T}}(\mathrm{CE}) = 34\,\mathrm{dBm}$ at the channel estimation stage, the proposed HBFRPD network has a probability of about 64.6% to achieve a sum rate exceeding 30 bps/Hz, while the other two schemes can only achieve 16.3%. When the transmit power is $P_{\mathrm{T}}(\mathrm{CE}) = 44\,\mathrm{dBm}$ at the channel estimation stage, our HBFRPD network has a probability of about 68.8% to achieve a sum rate exceeding 40 bps/Hz, which is significantly better than the other two schemes. This result again confirms the superior performance of the proposed DL-based HBFRPD network over the existing conventional schemes.

5.4 Computational Complexity Analysis

We now investigate the computational complexity. For the DL-based schemes, since there is no strict time limit at the offline training stage, we only consider the computational complexity at the inference stage. The computational complexity analysis of different schemes is presented in Table 1. All the numerical results are obtained on a PC with Intel(R) Core(TM) i9-10980XE CPU @ 3.00GHz and an Nvidia GeForce RTX 3090 GPU. The DL-based methods and the existing solutions are implemented on the PyCharm framework. The details are further elaborated as follows.

1) Channel estimation schemes: In the SOMP algorithm [43], correlation operation imposes significant computational complexity, where G_d is the dimension of the redundant dictionary and I is the number of iterations. The MMV-AMP algorithm [44] mainly requires matrix multiplication operations, but a large number of iterations I increases its computational complexity. The MMV-

Table 1: Computational Complexity of Different Schemes.

Channel estimation scheme	**Complexity**	**FLOPs**	**Run time/s**
SOMP	$\mathcal{O}\left(G_d K M I + G_d^2 K I\right)$	4.707 G	0.1130
MMV-AMP	$\mathcal{O}\left(M K N^{\mathrm{R}} I\right)$	5.337 G	0.7482
MMV-LAMP	$\mathcal{O}\left(M G_d K I\right)$	0.341 G	0.0689
Transformer-CEN	$\mathcal{O}\left(L_{\mathrm{T}}(K d_{\mathrm{T}}^2 + K^2 d_{\mathrm{T}})\right)$	0.362 G	0.0103
CNN-CEtraN	$\mathcal{O}\left(Z_5^2 N^{\mathrm{R}} K C_{32}^2\right)$	10.16 G	0.6039
Proposed	$\mathcal{O}\left(L_{\mathrm{M}}(N_p^2 d_{\mathrm{M}} + N_p d_{\mathrm{M}}^2)\right)$	1.066 G	0.0248
Beamforming scheme	**Complexity**	**FLOPs**	**Run time/s**
RZF	$\mathcal{O}\left((2U(M^{\mathrm{B}})^2 + (M^{\mathrm{B}})^3)K\right)$	24.58 K	0.1389
WMMSE	$\mathcal{O}\left(IK(U^2(M^{\mathrm{B}})^2 + U(M^{\mathrm{B}})^3)\right)$	8.192 M	0.9184
Proposed	$\mathcal{O}\left(L_{\mathrm{B}} U(K d_{\mathrm{B}}^2 + K^2 d_{\mathrm{B}})\right)$	0.512 G	0.0616

LAMP algorithm [45] has a low computational complexity because DL reduces the required number of iterations. The Transformer-CEN [39] also has a low computational complexity, and the main sources of its computational complexity come from self-attention and MLP sublayers. In the CNN-CEtraN [16], convolutional layers introduce significant computational complexity. By contrast,the MLP-mixer layers provide the majority of the computational complexity in our proposed SFDCEtra network, which is much lower than that of the CNN-CEtraN. We further meticulously count the numbers of floating-point operations per second (FLOPs) and run times per sample on CPU for different schemes in Table 1. Observe that at the inference stage, the FLOPs and run time per sample of the proposed scheme are lower than most benchmarks. Specifically, our SFDCEtra network imposes the second lowest run time per sample, and only the MMV-LAMP and Transformer-CEN have lower FLOPs than our proposed scheme.

2) Beamforming schemes: A matrix inversion is required in the RZF algorithm, which is its main source of computational complexity. In the iterative WMMSE algorithm [41], a large number of iterations increases the computational complexity and the run time per sample. In the proposed DL-based HBFRPD Network, self-attention and MLP sublayers impose higher computational complexity and FLOPs than the other two algorithms. However, the run time per sample of our proposed scheme is significantly lower than that of the two model-based schemes. This is due to the fact that the DL-based HBFRPD network just needs matrix multiplication operations and does not requires an iterative procedure. This is a superior advantage of our DL-based HBFRPD network.

6 Conclusions

In this paper, we have proposed a DL-based transmission scheme for RIS-aided THz massive MIMO systems over hybrid-field channels. Our novel twofold contribution has been to develop a channel estimation scheme with low pilot overhead and to design a robust beamforming scheme. More specifically, we have first proposed an E2E DL-based channel estimation framework, which consists of pilot design, CSI feedback, sub-channel estimation, and channel extrapolation. Then, to maximize the sum rate of all UEs under imperfect CSI, we have developed a DL-based scheme to simultaneously design the hybrid beamforming and RIS phase. Simulation results have shown that our proposed channel extrapolation scheme significantly outperforms the existing state-of-the-art

schemes, in terms of reconstruction performance, while imposing a much reduced pilot overhead. Moreover, the results have also demonstrated that our proposed beamforming scheme is superior over the existing designs in terms of achievable sum rate performance and robustness to imperfect CSI. Potential future research directions based on the outcomes of this paper include the practical phase shift model of reflecting elements, the analysis of hardware impairments, the analysis of the complex near-field channel, and sensing-aided communications.

References

[1] H. Elayan, O. Amin, B. Shihada, R. M. Shubair, and M.-S. Alouini, "Terahertz band: The last piece of RF spectrum puzzle for communication systems," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1-32, 2020.

[2] C. Lin and G. Y. Li, "Indoor terahertz communications: How many antenna arrays are needed?," *IEEE Trans. Wirel. Commun.*, vol. 14, no. 6, pp. 3097-3107, Jun. 2015.

[3] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 3, pp. 501-513, Apr. 2016.

[4] M. Di Renzo, *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450-2525, Nov. 2020.

[5] K. T. Selvan and R. Janaswamy, "Fraunhofer and fresnel distances: Unified derivation for aperture antennas," *IEEE Antennas Propag. Mag.*, vol. 59, no. 4, pp. 12–15, Aug. 2017.

[6] M. Cui and L. Dai, "Channel estimation for extremely large-scale MIMO: Far-field or near-field?," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2663-2677, Apr. 2022.

[7] L. Yan, Y. Chen, C. Han, and J. Yuan, "Joint inter-path and intra-path multiplexing for terahertz widely-spaced multi-subarray hybrid beamforming systems," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1391-1406, Feb. 2022.

[8] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," *Proc. ICASSP 2019* (Brighton, UK), May 12-17, 2019, pp. 4659-4663.

[9] P. Wang, J. Fang, H. Duan, and H. Li, "Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems," *IEEE Signal Process. Lett.*, vol. 27, pp. 905–909, May 2020.

[10] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications", *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144–4157, Jun. 2021.

[11] A. M. Elbir, A. Papazafeiropoulos, P. Kourtessis, and S. Chatzinotas, "Deep channel learning for large intelligent surfaces aided mm-Wave massive MIMO systems," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 9, pp. 1447-1451, Sep. 2020.

[12] S. Liu, Z. Gao, J. Zhang, M. D. Renzo, and M. -S. Alouini, "Deep denoising neural network assisted compressive channel estimation for mmWave intelligent reflecting surfaces," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 9223-9228, Aug. 2020.

[13] S. Zhang, *et al.*, "Deep learning based channel extrapolation for large-scale antenna systems: Opportunities, challenges and solutions," *IEEE Wirel. Commun.*, vol. 28, no. 6, pp. 160-167, Dec. 2021.

[14] B. Lin, F. Gao, S. Zhang, T. Zhou, and A. Alkhateeb, "Deep learning-based antenna selection and CSI extrapolation in massive MIMO systems," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 11, pp. 7669-7681, Nov. 2021.

[15] M. Xu, S. Zhang, C. Zhong, J. Ma, and O. A. Dobre, "Ordinary differential equation-based CNN for channel extrapolation over RIS-assisted communication," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 1921-1925, Jun. 2021.

[16] S. Zhang, S. Zhang, F. Gao, J. Ma, and O. A. Dobre, "Deep learning-based RIS channel extrapolation with element-grouping," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 12, pp. 2644-2648, Dec. 2021.

[17] K. Ying, *et al.*, "GMD-based hybrid beamforming for large reconfigurable intelligent surface assisted millimeter-wave massive MIMO," *IEEE Access*, vol. 8, pp. 19530-19539, Jan. 2020.

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

[19] B. Di, *et al.*, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1809-1822, Aug. 2020.

[20] Y. Ahn and B. Shim, "Deep learning-based beamforming for intelligent reflecting surface-assisted mmWave systems," *Proc. ICTC 2021* (Jeju Island, Korea), Oct. 20-22, 2021, pp. 1731-1734.

[21] S. Zhang, *et al.*, "Beyond intelligent reflecting surfaces: Reflective-transmissive metasurface aided communications for full-dimensional coverage extension," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13905–13909, Sep. 2020.

[22] Y. Youn, *et al.*, "Demo: Transparent intelligent surfaces for sub-6 GHz and mmWave B5G/6G systems," in *Proc. ICC Workshops 2022* (Seoul, Korea), May 16-20, 2022, pp. 1-2.

[23] D. Kitayama, *et al.*, "Transparent dynamic metasurface for a visually unaffected reconfigurable intelligent surface: Controlling transmission/reflection and making a window into an RF lens," *Opt. Express*, vol. 29, no. 18, pp. 29292-29307, Aug. 2021.

[24] Y. Chen, L. Yan, and C. Han, "Hybrid spherical- and planar-wave modeling and DCNN-powered estimation of terahertz ultra-massive MIMO channels," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 7063-7076, Oct. 2021.

[25] X. Wang, Z. Lin, F. Lin, and L. Hanzo, "Joint hybrid 3D beamforming relying on sensor-based training for reconfigurable intelligent surface aided terahertz-based multiuser massive MIMO systems," *IEEE Sens. J.*, vol. 22, no. 14, pp. 14540-14552, Jul. 2022.

[26] S. Hong, *et al.*, "Robust transmission design for intelligent reflecting surface-aided secure communication systems with imperfect cascaded CSI," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 4, pp. 2487-2501, Apr. 2021.

[27] Z. Chen, *et al.*, "Robust hybrid beamforming design for multi-RIS assisted MIMO system with imperfect CSI," *IEEE Trans. Wirel. Commun.*, Early Access, Nov. 2022. DOI:10.1109/TWC.2022.3222218

[28] W. Xu, L. Gan, and C. Huang, "A robust deep learning-based beamforming design for RIS-assisted multiuser MISO communications with practical constraints," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 694-706, Jun. 2022.

[29] P. Larsson,"Lattice array receiver and sender for spatially orthonormal MIMO communication," in *Proc. VTC 2005*, (Stockholm, Sweden), May. 30-Jun. 1, 2005, pp. 192-196.

[30] F. Bohagen, P. Orten, and G. E. Oien, "Optimal design of uniform planar antenna arrays for strong line-of-sight MIMO channels," *Proc. SPAWC 2006* (Cannes, France), Jul. 02-05, 2006, pp. 1-5.

[31] X. Song, W. Rave, N. Babu, S. Majhi, and G. Fettweis, "Two-level spatial multiplexing using hybrid beamforming for millimeter-wave backhaul," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 7, pp. 4830-4844, Jul. 2018.

[32] L. Yan, C. Han, and J. Yuan, "Energy-efficient dynamic-subarray with fixed true-time-delay design for terahertz wideband hybrid beamforming" *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 2840-2854, Oct. 2022.

[33] C. Han, A. O. Bicen, and I. F. Akyildiz, "Multi-ray channel modeling and wideband characterization for wireless communications in the terahertz band," *IEEE Trans. Wirel. Commun.*, vol. 14, no. 5, pp. 2402–2412, May 2015.

[34] Y. Wu, J. Kokkoniemi, C. Han, and M. Juntti, "Interference and coverage analysis for terahertz networks with indoor blockage effects and line-of-sight access point association," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 3, pp. 1472–1486, Mar. 2021.

[35] J. M. Jornet and I. F. Akyildiz, "Channel modeling and capacity analysis for electromagnetic wireless nanonetworks in the terahertz band," *IEEE Trans. Wirel. Commun.*, vol. 10, no. 10, pp. 3211–3221, Oct. 2011.

[36] B. Wang, F. Gao, S. Jin, H. Lin, and G. Y. Li, "Spatial- and frequency-wideband effects in millimeter-wave massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3393-3406, Jul. 2018.

[37] C. Wen, W. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wirel. Commun. Lett.*, vol. 7, no. 5, pp. 748-751, Oct. 2018.

[38] A. Vaswani, *et al.*, "Attention is all you need," in *Proc. NIPS 2017* (Long Beach, CA, USA), Dec. 4-9, 2017, pp. 6000–6010.

[39] Y. Wang, *et al.*, "Transformer-empowered 6G intelligent networks: From massive MIMO processing to semantic communication," *IEEE Wirel. Commun.*, Early Access, Nov. 2022. DOI:10.1109/MWC.008.2200157

[40] I. O. Tolstikhin, *et al.*, "Mlp-mixer: An all-mlp architecture for vision," in *Proc. NIPS 2021*, Dec. 6-14, 2021, pp. 24261-24272.

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

[42] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270-280, Jun. 1989.

[43] C. Tsai, Y. Liu, and A. Wu, "Efficient compressive channel estimation for millimeter-wave large-scale antenna systems," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2414-2428, May 2018.

[44] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, Jan. 2020.

[45] X. Ma, Z. Gao, F. Gao, and M. Di Renzo, "Model-driven deep learning based channel estimation and feedback for millimeter-wave massive hybrid MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2388-2406, Aug. 2021.