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[1,2,3*], Sheng Chen[3,4], Chun Hu[1], and Dezhi Zheng[1]

[1]MIIT Key Laboratory of Complex-Field Intelligent Sensing, Beijing Institute of Technology, Beijing, China

[2]Yangtze Delta Region Academy of Beijing Institute of Technology, Beijing Institute of Technology, Jiaxing, China

[3]Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Jinan, China

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

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

[*]Address correspondence to: gaozhen16@bit.edu.cn

Abstract

The reconfigurable intelligent surface (RIS) is a promising novel technology for Terahertz (THz) massive multiple-input multiple-output (MIMO) communication systems. However, the acquirement of the high-dimensional channel state information (CSI) and the efficient active/passive beamforming for RIS are challenging due to its cascaded channel structure and its lack of signal processing units. To address this, we propose a deep learning (DL)-based physical signal processing 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 DL-based channel estimation framework, which consists of pilot design, CSI feedback, sub-channel estimation, and channel extrapolation. Specifically, we first only activate partial RIS elements and estimate a sub-sampling RIS channel, and then utilize a DL-based extrapolation network to reconstruct the full-dimensional CSI. Moreover, to maximize the sum rate under imperfect CSI, a DL-based scheme is developed 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 greatly reducing 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, deep learning (DL).

Introduction

1 Introduction

The surge in demand for wireless data traffic in recent years, owing to the exponential growth of massive internet-of-things (IoT) devices and broadband multimedia applications, has necessitated the exploration of Terahertz (THz) communications as a viable solution [1]. However, extremely high free-space losses and strong atmospheric attenuation in the THz band pose a challenge to the long-range coverage of THz communication systems. To overcome this problem, the massive or ultra-massive multiple-input multiple-output (MIMO) technique has been considered to achieve high array gain and mitigate the high propagation loss [2]. Conventional massive MIMO systems require a dedicated radio frequency (RF) chain for each antenna (i.e., fully-digital architecture), which suffers from extremely high power consumption and hardware costs. To circumvent this technical hurdle, the hybrid analog-digital massive MIMO architecture has been widely adopted to reduce the number of RF chains while ensuring high array gains [3].

Besides, the advent of reconfigurable intelligent surfaces (RIS) has garnered attention as a potentially transformative technology for improving communication performance [4–10]. By manipulating the phase and amplitude of RIS phase shifters, RIS passively reflects incident electromagnetic (EM) signals towards desired directions and provides significant beamforming gain. More importantly, RIS does not require power-intensive RF chains, which contributes to a more environmentally friendly and cost-effective communication solution. Therefore, the integration of RIS and massive MIMO techniques holds promise for overcoming the limitations of THz communications and realizing 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 the THz band will increase the Rayleigh distance [11], both far and near-field need to be considered for THz massive MIMO systems. Therefore, the distance from each antenna of the BS to the UE needs to be considered under near-field conditions by the spherical-wave channel model [12]. On the other hand, the number of spherical-wave channel parameters is proportional to the number of massive antennas, which indicates that directly adopting the spherical-wave channel model in THz massive MIMO systems 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 [13]. For such a channel model, the EM signal is modeled as a spherical wave for the inter-subarray and a planer wave for the intra-subarray, based on different subarray architectures. Although the application of RISs has been widely researched recently [14–18], 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 [19–22]. However, accurately estimating high-dimensional CSI with limited pilot signals remains a formidable challenge [14]. To address this challenge, compressive sensing (CS)-based solutions have been proposed to reduce the pilot overhead by leveraging the channel sparsity [15, 16]. However, these solutions present challenges with regard to computational complexity and storage requirements, as the corresponding matrix inversion and iterative operations. Recently, the integration of deep learning (DL) in communication systems has garnered extensive attention. For instance, in [23], the authors proposed an effective pilot reduction technique by gradually pruning less significant neurons from the dense layers during training. In [17], the authors designed a DL-based channel estimation network to acquire the RIS-aided channel and the non-RIS-aided channel. In [18], a semi-passive RIS architecture was proposed, where the orthogonal match pursuit (OMP) algorithm and a denoising convolutional neural network (CNN) are applied to reconstruct the CSI. However, the deployment of RF chains negates the key benefits of the RIS, i.e., reducing hardware costs and power consumption.

In fact, due to the highly-dense arrangement of RIS elements [24], 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 [25]. Recently, there are some initial attempts to utilize the channel extrapolation for further reducing the pilot overhead. In [26], the authors proposed a DL-based extrapolation network to extrapolate the complete CSI by exploiting the correlation of the antenna domain, where partial antennas are activated by a selection network. In [27], the authors utilized a neural network structure modified by ordinary differential equations to improve the performance of extrapolation. Besides, the authors of [28] adopted a grouping strategy to reduce the dimension of the estimated channel and designed a CNN-based network to extrapolate the full-dimensional cascaded channel as well as eliminate the grouping interference. 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 EM wave 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 [29–31]. In [29], simultaneous orthogonal matching pursuit (SOMP)-based hybrid beamforming was proposed for RIS-aided mmWave MIMO systems. In [30], an iteration-based jointly active/passive beamforming algorithm was designed to maximize the sum rate of systems. Furthermore, the DL-based beamforming methods have also been studied in RIS-aided wireless communication systems. In [31], a deep neural network (DNN)-based beamforming approach was developed to jointly optimize the transmit/reflect beamforming vectors for achieving data rate maximization. However, further analysis of the aforementioned schemes with regard to adaptability is necessary, as the current analysis only considers the idealized CSI assumption.

1.2 Motivations

The current research on RIS has primarily centered on the development of two modes of operation, namely, reflective mode [29, 30, 32] and transmissive mode [33–35]. A number of studies have been conducted on RIS-aided communication in reflective mode, which is primarily utilized to address the blind coverage problem. By contrast, the main purpose of transmissive RIS is to improve the spectral efficiency of the networks, as the transmissive mode does not alter the direction of EM waves. Therefore, 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. In view of this, the transmissive RIS has the potential to provide indoor signal enhancement service.

Considering the hybrid-field channel model, the authors of [36] presented a two-stage channel estimation mechanism, where a CNN-based network is designed to estimate channel parameters and the complete channel is reconstructed by channel extrapolation based on geometric relationships of channel parameters. However, this parametric-based extrapolation method requires a large number of training labels containing accurate channel parameters. In [37], the authors proposed a sensor-assisted channel estimation and beamforming technique, where a LoS MIMO architecture is considered in the hybrid field. However, the channel estimation in [37] relies heavily on the awareness of sensors, which can prove challenging in obtaining accurate CSI. Therefore, similar to [26–28], 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 in the field of RIS-aided communication systems have made the assumption that the BS-RIS and RIS-UEs CSIs are perfect [29–32]. However, this assumption is impractical. Therefore, the channel estimation error should be considered when designing these systems. Recently, imperfect CSI conditions have been considered in some works [38, 39]. For instance, the authors of [38] utilized a penalty-based alternating algorithm to jointly design active beamforming and RIS phase under the presence of imperfect CSI. Similarly, the authors of [39] exploited a gradient projection-based alternating optimization algorithm to jointly design active beamforming, RIS placement, and RIS phase under imperfect CSI. While there are numerous DL-based methods available for RIS-aided communication systems with the perfect CSI, there are only a few DL-based methods that consider imperfect CSI [40]. Therefore, this work aims to provide a DL-based hybrid beamforming and RIS phase design solution that incorporates imperfect CSI in RIS-aided communication systems.

1.3 Contributions

This paper presents 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 a transmissive RIS on the window 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 a 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 as well as only one LoS path exists, the BS-RIS channel can be considered to be quasi-static and known. In contrast, due to the mobility of the UE, the RIS-UE channel is time-varying. Therefore, we only focus on estimating 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 [26–28] 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 maintain high reconstruction performance with a few pilot overhead. Specifically, by using the CSI feedback network, the UE-side feeds the quantized pilot information back to the BS, and the BS estimates the sub-sampling RIS-UE 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 maximizing the sum rate with E2E training, the proposed scheme can realize higher performance and better robustness than the existing state-of-the-art methods.

Notations: In this paper, scalars are denoted as lower-case letters, vectors are denoted as lower-case boldface letters, and matrices are denoted as upper-case boldface letters. The conjugate, transpose, conjugate transpose, inversion, and Moore-Penrose inversion operators are denoted as the superscripts $(\cdot)^{*}$, $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, $(\cdot)^{-1}$, and $(\cdot)^{\dagger}$, respectively. The diagonalization, block diagonalization, Kronecker product, and Hadamard product are represented by the operators $\mathrm{diag}(\cdot)$, $\mathrm{blkdiag}(\cdot)$, $\otimes$, and $\odot$, respectively. The Frobenius norm of $\mathbf{A}$ is denoted as $|\mathbf{A}|_F$. The identity matrix with size $n \times n$ is represented by $\mathbf{I}_n$, while the column vector of size n with all elements equal to 1 (0) is represented by $\mathbf{1}_n$ ($\mathbf{0}_n$). The real and imaginary parts of the corresponding argument are denoted as $\Re\{\cdot\}$ and $\Im\{\cdot\}$, respectively. The m-th row and n-th column element of $\mathbf{A}$ is represented by $\{\mathbf{A}\}_{m,n}$, and the m-th entry of $\mathbf{a}$ is represented by $\{\mathbf{a}\}_m$. The sub-matrix containing the m-th to n-th columns of $\mathbf{A}$ is represented by $\mathbf{A}_{[:,m:n]}$. The expectation operator is represented by $\mathbb{E}(\cdot)$, and the real (complex) Gaussian distribution with mean μ and variance σ^2 is denoted as $\mathcal{N}(\mu, \sigma^2)$ ($\mathcal{CN}(\mu, \sigma^2)$), where the matrix trace operator is represented by $\mathrm{Tr}\{\cdot\}$.

Materials and Methods

2 System Model

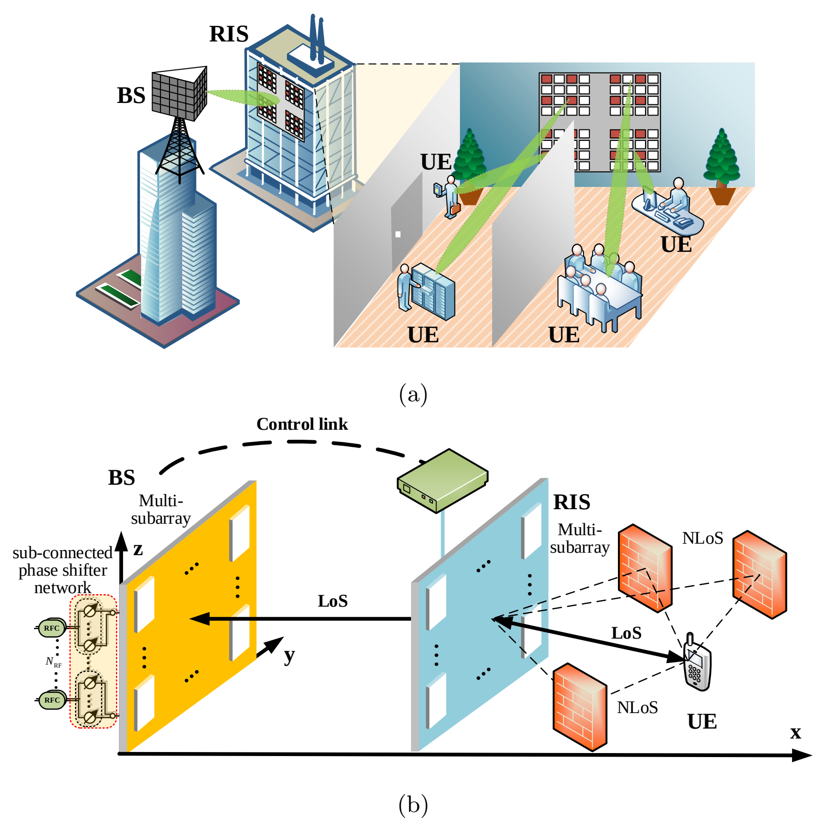


Figure 1: Schematic diagram of a RIS-aided THz massive MIMO system: (a) multiple indoor UEs are served by the BS with the help of a transmissive RIS deployed on the window, and (b) hardware architectures at the BS, RIS, and UEs.

2.1 System Description

As shown in Figure 1, we consider a downlink RIS-aided MIMO orthogonal frequency division multiplexing (OFDM) transmission system 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 BS (RIS)-side subarrays along the horizontal and vertical 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 BS (RIS)-side subarray antennas along the horizontal and vertical 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 $N^{\mathrm{R}} = M^{\mathrm{R}} N_{\mathrm{sub}}^{\mathrm{R}}$. To simplify the analysis, we assume that the normals of 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).

In this paper, we consider a BS-side sub-connected hybrid analog-digital array architecture. This architecture consists of M^{B} RF chains, which are capable of supporting $U \leq M^{\mathrm{B}}$ data streams. Each of these RF chains is connected to a subarray through $N_{\mathrm{sub}}^{\mathrm{B}}$ phase shifters. Furthermore, We set the number of subcarriers to K and sampling frequency (i.e., bandwidth) to f_s. The carrier frequency is f_c, which corresponds to central wavelength λ.

2.2 Channel Model

2.2.1 BS-RIS Channel Model

Due to the negligible non-LoS (NLoS) component energy in the THz band, we only consider the LoS path in the analysis of the BS-RIS channel. By utilizing the spherical wave propagation characteristic, we construct the LoS MIMO link between the BS and RIS with only one single LoS path, but it can support intra-path multiplexing for multi-stream transmission [41]. 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 [41, 42]. 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 [42]. We have the following proposition from [43].

Proposition 1 *Let the transceiver arrays be placed with a separation distance of D and be working at a carrier wavelength λ ($\lambda \ll D$). If the inter-antenna spacing and carrier wavelength λ are in* *the same order of magnitude, the planner wave model can be applied. Otherwise, the spherical wave model should be exploited.*

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:

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

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 horizontal and vertical antennas, respectively, while θ and ϕ are the horizontal and vertical 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 same path's direction difference in different subarrays is negligible. Therefore, all subarrays on either the BS or RIS-side can be assumed to share the identical array response vectors. However, as subarrays are widely spaced, the relative phase differences among subarrays are non-negligible [43]. Motivated by the above analysis, the downlink spatial-frequency BS-RIS channel $\mathbf{G}[k] \in \mathbb{C}^{N^\mathrm{R} \times N^\mathrm{B}}$ 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 AoD and AoA of the LoS path, respectively. Without loss of generality, we assume the LoS angles are fixed and known in advance since the BS and RIS are fixed. In (3), the entries of $\tilde{\mathbf{G}}[k] \in \mathbb{C}^{M^\mathrm{R} \times M^\mathrm{B}}$ are defined according to 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 $D^{(m_r, m_b)}$ represents 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 different from the array gain generated by beamforming [44]. 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 quasi-static BS-RIS channel is 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 [45]. The indoor RIS-UE channel model consists of one LoS path and L_p NLoS paths, where their three-dimensional (3D) distances are represented as d_0 and d_l, for $1 \le l \le L_p$, respectively [46]. The total EM wave propagation loss mainly consists of two parts: the free space path loss $\beta_\mathrm{spr}(f_k, d_l) = \frac{c}{4\pi f_k d_l}$ and the molecular absorption loss $\beta_\mathrm{abs}(f_k, d_l) = e^{-\frac{1}{2}\kappa(f_k) d_l}$, where $\kappa(f_k)$

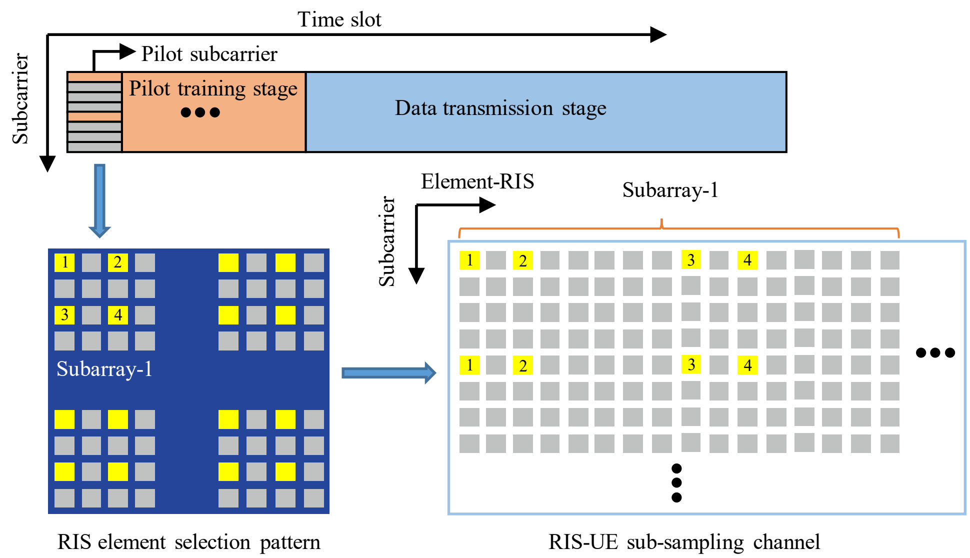


Figure 2: Block diagram of the frame structure, RIS element selection pattern, and RIS-UE sub-sampling channel, where the selected parts are marked in yellow blocks and the number in the yellow block indicates the index of the selected element.

denotes the frequency-dependent absorption coefficient [47]. Hence, the spatial-frequency channel $\mathbf{h}[k] \in \mathbb{C}^{1\times N^{\mathrm{R}}}$ for the RIS-UE link is

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

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

3 Problem Formulation and Proposed Channel Estimation Solution

3.1 Problem Formulation of Channel Estimation

In this subsection, the downlink channel estimation problem is formulated based on the considered RIS-aided 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 slots) dedicated to channel estimation. The m-th received signal at the UE-side[1] on the k-th

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

subcarrier is represented by

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

where $1 \leq k \leq K$, $1 \leq m \leq M$, 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 $\mathbf{v}_{m_r,m} \in \mathbb{C}^{1\times N_{\mathrm{sub}}^{\mathrm{R}}}$ for the m_r-th subarray elements of the RIS in the m-th time slot as

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

where $\mathbf{o}_{m_r,m} \in \mathbb{C}^{1\times N_{\mathrm{sub}}^{\mathrm{R}}}$ represents the amplitude control vector, $\tilde{\mathbf{v}}_{m_r,m} \in \mathbb{C}^{1\times N_{\mathrm{sub}}^{\mathrm{R}}}$ represents the phase control vector, and $1 \leq n_{\mathrm{sub}}^r \leq N_{\mathrm{sub}}^{\mathrm{R}}$, while $\eta_{n_{\mathrm{sub}}^r,m_r,m} \in [0,\ 1]$ and $\phi_{n_{\mathrm{sub}}^r,m_r,m} \in [0,\ 2\pi]$ are the amplitude/phase control coefficient, respectively. $\eta_{n_{\mathrm{sub}}^r,m_r,m}$ can control the switch of the refraction function for 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 provide array gain and eliminate the multi-stream interference. Since the sub-connected architecture, the analog beamformer implemented by phase shifters is written 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 $\mathbf{f}_{m_b} = \left[\mathrm{f}_{m_b,1}, \cdots, \mathrm{f}_{m_b,n_{\mathrm{sub}}^b}, \cdots, \mathrm{f}_{m_b,N_{\mathrm{sub}}^{\mathrm{B}}}\right]^{\mathrm{T}} \in \mathbb{C}^{N_{\mathrm{sub}}^{\mathrm{B}}\times 1}$ with $\left|\mathrm{f}_{m_b,n_{\mathrm{sub}}^b}\right|^2 = 1/N_{\mathrm{sub}}^{\mathrm{B}}$. Since the BS-RIS channel with the LoS path only is quasi-static and known, each analog beamforming vector can be designed as

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

where k can be set to $K/2$ for alleviating the beam squint problem induced by the large bandwidth [48]. 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 RIS subarrays, i.e.,

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

where $\tilde{\mathbf{G}}_{\mathrm{eq}}[k] = \left[\alpha[k]G_{\mathrm{T}}\tilde{\mathbf{G}}[k] \otimes \mathbf{a}_{\mathrm{B}}^{\mathrm{H}}(\theta_{\mathrm{B}}, \phi_{\mathrm{B}}, f_k)\right]\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{M^{\mathrm{R}}\times M^{\mathrm{B}}}$ is the equivalent BS-RIS channel obtained from the perspective of the first element of different subarrays at the RIS, and $\zeta = \sqrt{M^{\mathrm{B}}/\mathrm{Tr}\big\{\tilde{\mathbf{G}}_{\mathrm{eq}}^{\dagger}[k]\big(\tilde{\mathbf{G}}_{\mathrm{eq}}^{\dagger}[k]\big)^H\big\}}$ is a constant to meet the total transmit power constraint after beamforming. 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 written 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 in 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 $\sqrt{N^{\mathrm{B}}}\alpha[k]G_{\mathrm{T}}\mathrm{blkdiag}\big(\mathbf{1}^1_{N^{\mathrm{R}}_{\mathrm{sub}}}, \cdots, \mathbf{1}^u_{N^{\mathrm{R}}_{\mathrm{sub}}}, \cdots, \mathbf{1}^U_{N^{\mathrm{R}}_{\mathrm{sub}}}\big)$. 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 continuous measurements of M time slots, the aggregate received signal vector $\mathbf{y}[k] = [y_1[k], \cdots, y_M[k]] \in \mathbb{C}^{1\times M}$ is written as

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

where $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_M] = A_{\mathrm{T}}\mathbf{V} = A_{\mathrm{T}}\left[\mathbf{v}_1, \cdots, \mathbf{v}_M\right] \in \mathbb{C}^{N^{\mathrm{R}}\times M}$, 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 written as

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

where $\mathbf{H} = \left[\mathbf{h}^{\mathrm{T}}[1], \cdots, \mathbf{h}^{\mathrm{T}}[K]\right]^{\mathrm{T}} \in \mathbb{C}^{K\times N^{\mathrm{R}}}$ represents 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

As shown in Figure 2, we choose to activate only $N^{\mathrm{R}}_s \leq N^{\mathrm{R}}$ RIS elements at the pilot training stage, and define $\rho \triangleq N^{\mathrm{R}}/N^{\mathrm{R}}_s \geq 1$ as the element compression ratio. Furthermore, 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. Then, we estimate the sub-channels associated with the activated RIS elements and the selected subcarriers. 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 complete channel $\hat{\mathbf{H}} \in \mathbb{C}^{K\times N^{\mathrm{R}}}$ based on limited received pilot signals $\mathbf{Y}_s$, i.e., extrapolating the rest unknown channels from the acquired partial channels. Based on the non-linear function fitting capability of DL, We can

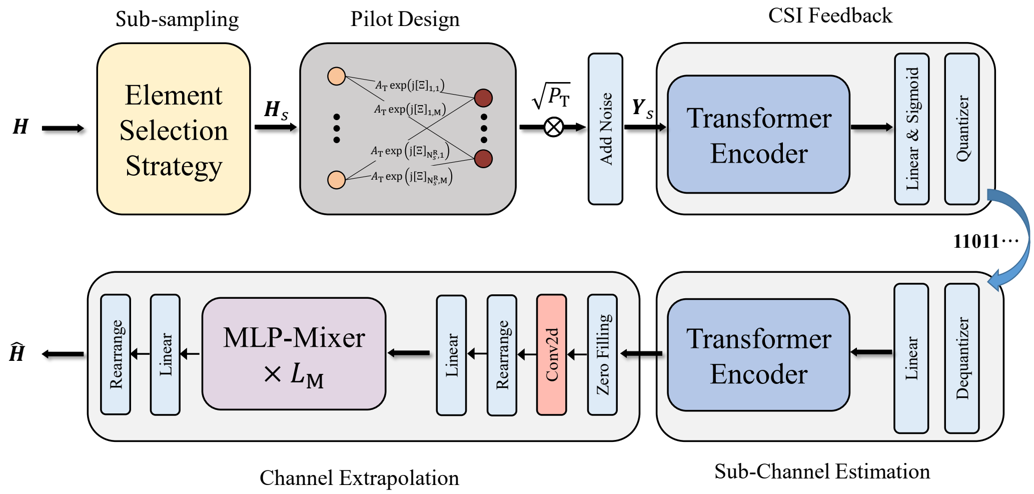


Figure 3: The overall block diagram of the proposed DL-based spatial-frequency domain channel extrapolation scheme.

learn a mapping to represent proximity correlations between different spatial/frequency locations of channels. 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 complete process of the proposed scheme can be represented as

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

where the mapping $f_{\mathrm{ESS}}(\cdot)$ represents the element selection strategy for deciding the sub-sampling channel $\mathbf{H}_s$, 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 Equation (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

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

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 different RIS element selection vectors 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 $\rho = \rho_y \times \rho_z$, where ρ_y and ρ_z are the compression ratios along the azimuth and elevation directions, respectively. To fairly sound the channel and ensure a balanced estimation performance along two directions, ρ_y and ρ_z are expected to be as close as possible. However, $\rho_y = \rho_z$ cannot always be guaranteed under all system parameter configurations. In cases where $\rho_y \neq \rho_z$, it is desirable to allocate more activated elements along the azimuth (y-xis) direction rather than the z-axis direction (i.e., $\rho_y \leq \rho_z$). This strategic choice aligns with the consideration of indoor UEs, which are more likely to be distributed across a wide azimuth range, as opposed to the elevation range, as indoor UEs are typically stationary in the vertical dimension. In light of this, the y-z compression ratio allocation can be solved from the following optimization problem

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

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 $\boldsymbol{\xi}_{m_r} \in \mathbb{C}^{1\times N_{\text{sub}}^{\text{R}}/\rho}$ of the m_r-th subarray can be expressed as

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

where $1 \leq m_r \leq M^{\text{R}}$, $0 \leq n_i^y \leq \frac{N_y^{\text{R}}}{\rho_y} - 1$, and $0 \leq n_i^z \leq \frac{N_z^{\text{R}}}{\rho_z} - 1$. The entire active element index vector or set of the RIS is defined as $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{m_r}, \cdots, \boldsymbol{\xi}_{M^{\text{R}}}]^{\text{T}} \in \mathbb{C}^{N_s^{\text{R}}\times 1}$. 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 and generates the active element index vector $\boldsymbol{\xi}$. 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 [26], 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 $\mathbf{g} = [g_1, g_2, \cdots, g_{N^{\text{R}}}]^{\text{T}} \in \mathbb{C}^{N^{\text{R}}\times 1}$ is generated, where g_i denotes the probability that the i-th element is selected. Thus, the active element index vector $\boldsymbol{\xi}$ can be defined as

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

where $\arg\,\text{top}_{N_s^{\text{R}}}\{\cdot\}$ is a function that finds the element index set of the first N_s^{R} largest selection probabilities. The details of the selection network can be found in [26].

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 $\mathbf{V}_s \in \mathbb{C}^{N_s^{\mathrm{R}} \times M}$ denotes the RIS phase matrix of selected 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. Since most DL frameworks, such as Tensorflow and Pytorch, have limited support for complex-valued operations, it becomes challenging to train the complex-valued pilot matrix $\mathbf{P}_s$ directly. To circumvent this issue, we adopt the real-valued RIS phase control matrix $\boldsymbol{\Xi}$ whose entries take values in $[0,\ 2\pi)$ as 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 trainable parameters of the PDN, i.e., $\boldsymbol{\Xi}$, are learned at the DL training stage.

3.2.3 CSI Feedback

Recently, DL-based solutions, such as CsiNet [49], have achieved good performance for CSI feedback. Furthermore, an emerging CSI feedback architecture based on the transformer [50] has been demonstrated to further reduce the feedback overhead and obtain more efficient compression performance than the CsiNet framework [51]. Therefore, we utilize the transformer as the backbone of CSI feedback network $f_{\mathrm{CsiFd}}(\cdot)$. The original transformer is divided into an encoder and a decoder. However, since we are dealing with the CSI without time-sequential information, there is no causality

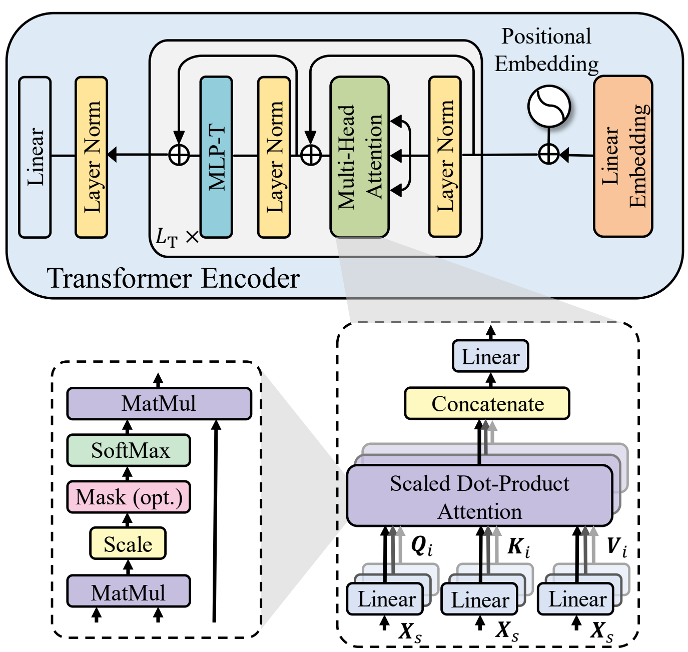


Figure 4: The structure of the transformer encoder.

constraint. Thus, we only exploit the encoder module of the transformer which obtains output in parallel. Since real-valued operations are more effective and the transformer can only extract the correlation between sequences, we convert 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

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

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

The schematic diagram of the transformer encoder is shown in Figure 4. Through the fully-connected linear embedding layer, input sequence $\bar{\mathbf{Y}}_s$ can be converted into $\mathbf{X}_s \in \mathbb{R}^{K_s \times d_{\mathrm{T}}}$, which merges the relative position information of the sub-carriers using the positional embedding layer. Then, multiple encoder layers are utilized to extract correlations between 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 plays a key role in 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 $d_m = d_{\mathrm{T}}/h$ and h is the number of heads. 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 simultaneously, where the weights on values can be obtained from a softmax function, 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 linear layer followed by a sigmoid function is used to generate a compressed codeword, which is then transformed into B bits as the feedback information through a uniform scalar quantization layer. 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, received CSI feedback bits are initially processed by a dequantization layer, which conducts the inverse operation of the quantizer and outputs a real-valued vector. Then, the initial coarse channel estimate is obtained by a linear layer. Finally, the transformer encoder extracts the spatial-frequency correlation of the channel and further improves the channel estimation performance. The sub-channel estimation process can be represented by

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

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 is represented by

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

The non-zero rows/columns in $\tilde{\mathbf{H}}$ 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, i.e., adding additional zeros around the input feature map.

Subsequently, we consider a competitive yet conceptually and technically simple architecture, called MLP-Mixer [52], 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 series of flattened 2D patches $\mathbf{X}_p \in \mathbb{R}^{N_p \times (2L^2)}$, where (K, N^{R}) represents the size of the original input, (L, L) represents each patch's length and width, as well as $N_p = KN^{\mathrm{R}}/L^2$ represents the number of patches. Then, all the patches are linearly projected

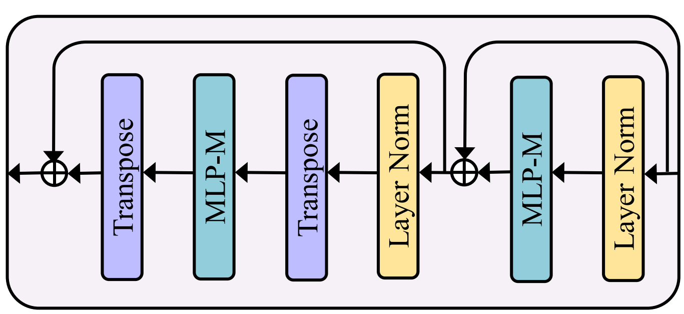


Figure 5: The structure of the mixer layer.

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 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. 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 $\tilde{\mathbf{X}}_t$ denotes the input matrix to the t-th mixer layer, $\mathbf{W}_{t,i}$, $1 \le i \le 4$, 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 denotes an 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 is represented by

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

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 off-line training dataset is represented as $\boldsymbol{\mathcal{H}}$, which comprises of $|\boldsymbol{\mathcal{H}}| = N_{\mathrm{set}}$ samples. Each sample in $\boldsymbol{\mathcal{H}}$ is an input-label pair represented by $(\mathbf{H}, \mathbf{H})$, where $\mathbf{H}$ serves as both the extrapolation target and the input for the 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 an E2E training for the pilot design network, CSI feedback network, sub-channel estimation network, and channel extrapolation network. Thus, the loss function can be represented by minimizing the normalized mean square error (NMSE) between the output $\hat{\mathbf{H}}$ and the target $\mathbf{H}$, i.e,

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

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, i.e., the loss function can be represented by

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

where $0 < \gamma \leq 1$ represents the weight used to balance channel extrapolation and ESS with $\gamma = 1$ denoting that the non-learning based $f_{\text{ESS}}(\cdot)$ is selected, and $\mathcal{L}_{\text{ESS}}$ is the loss function of the learning-based ESS. The details of $\mathcal{L}_{\text{ESS}}$ are available in [26].

4 Proposed Beamforming Solution

4.1 Problem Formulation of RIS-aided Multi-User Beamforming

The BS can simultaneously support U UEs with the aid of RIS at the data transmission stage, since the LoS MIMO architecture 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 represented by

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

where $\mathbf{h}[u,k] \in \mathbb{C}^{1\times N^{\text{R}}}$, $1 \leq u \leq U, 1 \leq k \leq K$, denotes the downlink RIS-UE channel of the u-th UE on the k-th subcarrier, $\mathbf{f}_{\text{BB}}[u,k] \in \mathbb{C}^{M^{\text{B}}\times 1}$ 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

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

Therefore, the sum rate R of total UEs is represented by

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

By utilizing the estimated RIS-UE channel at the pilot training stage, the BS can design the hybrid beamformer $\{\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}[k], \forall k\}$ 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 illustrated as

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

where $\hat{\mathbf{H}}[u]$ is the estimated spatial-frequency RIS-UE channel of the u-th UE, and $\mathcal{F}(\cdot)$ represents a function that maps the estimated RIS-UE channels onto the hybrid beamformer $\{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}\left[k\right], \forall k\}$ and the RIS refraction phase matrix $\mathbf{\Phi}$.

4.2 Deep Learning Based Hybrid Beamforming and RIS Phase Design

To solve this optimization Equation (34), some alternating iterative algorithms [29, 30, 32] 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. By using the non-linear function fitting capability of DL, we can 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 diagram depicting the design of the proposed scheme is presented in Figure 6.

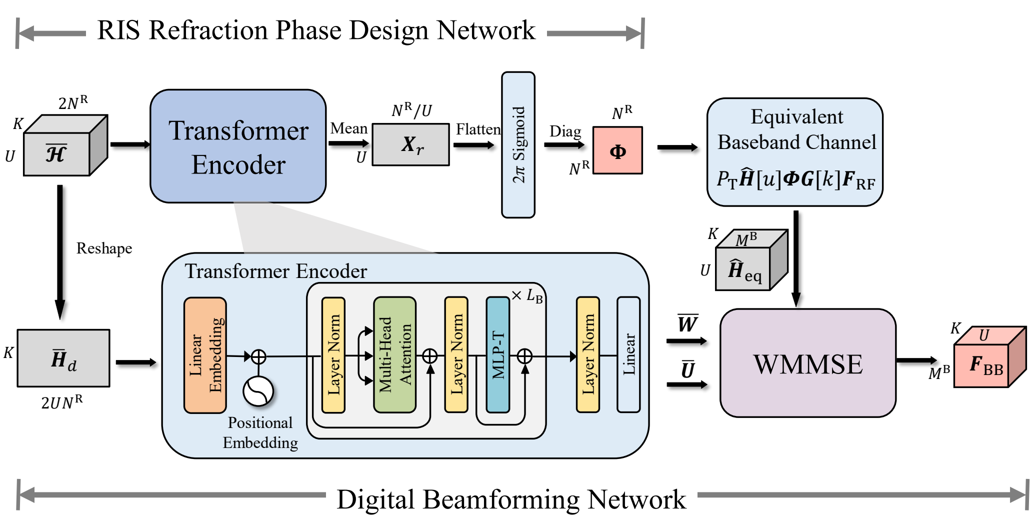


Figure 6: The overall structure of the proposed DL-based hybrid beamforming and RIS refraction phase design scheme.

4.2.1 Analog Beamformer Design

The integration of the active/passive beamforming at the BS and RIS is a non-convex optimization problem, which poses significant difficulties in finding a global optimum solution. Hence, we separately design the analog beamforming and the passive beamforming. Specifically, both BS analog beamforming and RIS passive beamforming are designed to focus energy for improving the received SINR of UEs. However, since the sub-connected structure 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 is focused on the RIS.

Since the BS-RIS channel with only the LoS path is quasi-static and known, we can utilize the angle information of the BS-RIS link to design the analog beamformer. Specifically, the transmit beam of the m_b-th subarray designed for the u-th UE should be aligned to the m_r-th subarray of the RIS, where the u-th UE is assisted by the m_r-th subarray of the RIS. Therefore, analog beamformer $\mathbf{F}_{\mathrm{RF}} = \mathrm{blkdiag}\big(\mathbf{f}_1, \cdots, \mathbf{f}_{m_b}, \cdots, \mathbf{f}_{M^{\mathrm{B}}}\big)$ can be simply designed for alignment between BS and RIS subarrays according to Equation (9).

4.2.2 DL-Based RIS Refraction Phase Design

Optimizing a common RIS phase shared by all the subcarriers is a crucial challenge in a RIS-aided OFDM system. In the THz broadband case, there exists a non-negligible beam squint effect for different subcarriers [48]. 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, a transformer-based RIS phase design network (RPDN) is proposed 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 $\bar{\boldsymbol{\mathcal{H}}} \in \mathbb{R}^{U\times K\times 2N^{\mathrm{R}}}$, i.e.,

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

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 constant modulus constraint, i.e.,

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

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

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

where $f_{\text{RIS}}(\cdot)$ 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 $\mathbf{\Phi}$ and the analog beamforming matrix $\mathbf{F}_{\text{RF}}$ as well as the estimated RIS-UE channel $\hat{\mathbf{h}}[u,k]$, the estimated equivalent baseband channel $\hat{\mathbf{h}}_{\text{eq}}[u,k] \in \mathbb{C}^{1\times M^{\text{B}}}$ can be represented by

$$\hat{\mathbf{h}}_{\text{eq}}[u,k] = P_{\text{T}}\hat{\mathbf{h}}[u,k]\mathbf{\Phi}\mathbf{G}[k]\mathbf{F}_{\text{RF}}. \tag{38}$$

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

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

Note that since the analog beamformer and the RIS refraction phase have been specifically designed in Subsections 4.2.1 and 4.2.2 respectively, the above problem (39) 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. Taking the latter as an example, the iterative WMMSE algorithm is designed to solve the optimization (39) by addressing the equivalent MMSE problem specified in (40) below, which has the identical optimal solution $\mathbf{F}_{\text{BB}}[k], \forall k$ to the problem (39).

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

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 $\hat{s}[u,k] = \bar{u}_{u,k}y[u,k]$ is 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 [53], the above problem is convex in each individual optimization variable. This property enables each subproblem to have a closed-form solution, given the other optimization variables. Then, the optimization (40) can be solved by a block coordinate descent (BCD) iterative algorithm. Here, we depict the iterative WMMSE beamforming design algorithm in Algorithm 1.

However, the iterative WMMSE algorithm typically imposes a large number of iterations with 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 $\bar{\boldsymbol{\mathcal{H}}} \in \mathbb{R}^{U\times K\times 2N^{\mathrm{R}}}$ is reshaped into a 2D matrix $\bar{\mathbf{H}}_d \in \mathbb{R}^{K\times 2UN^{\mathrm{R}}}$, which is inputted into the transformer encoder. The transformer encoder will output $\mathbf{X} \in \mathbb{R}^{K\times 4U}$, which 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 the running time can be reduced significantly. To satisfy transmit power constraint, the normalization operation can be represented by

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

The proposed knowledge-data dual-driven digital beamformer design can be represented by

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

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

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** $\{\bar{\mathbf{U}}\}_{u,k}$: $\bar{u}_{u,k} = \left(\sum_{i=1}^{U} |\mathbf{h}_{\mathrm{eq}}[u,k]\mathbf{f}_{\mathrm{BB}}[i,k]|^2 + \sigma_n^2\right)^{-1} \mathbf{h}_{\mathrm{eq}}[u,k]\mathbf{f}_{\mathrm{BB}}[u,k]$, $\forall u,k$;
4:     **Update** $\{\bar{\mathbf{W}}\}_{u,k}$: $\bar{w}_{u,k} = \left(1 - \bar{u}^*_{u,k}\mathbf{h}_{\mathrm{eq}}[u,k]\mathbf{f}_{\mathrm{BB}}[u,k]\right)^{-1}$, $\forall u,k$;
5:     **Update** $\mathbf{f}_{\mathrm{BB}}[u,k]$: $\mathbf{f}_{\mathrm{BB}}[u,k] = \bar{u}_{u,k}\bar{w}_{u,k}\left(\sum_{i=1}^{U} \bar{w}_{i,k}|\bar{u}_{i,k}|^2\mathbf{h}^{\mathrm{H}}_{\mathrm{eq}}[i,k]\mathbf{h}_{\mathrm{eq}}[i,k] + \mu_k\mathbf{I}\right)^{-1}\mathbf{h}^{\mathrm{H}}_{\mathrm{eq}}[u,k]$, where $\mu_k = \sum_{j=1}^{U} \frac{\sigma^2}{M^{\mathrm{B}}}\bar{w}_{j,k}|\bar{u}_{j,k}|^2$, $\forall u,k$;
6:     $t = t + 1$;
7: **until** $t \geq I_{\max}$
8: Scale $\mathbf{F}_{\mathrm{BB}}[k]$ to meet the transmit power constraint.

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 $|\boldsymbol{\mathcal{H}}_U| = N_{\text{set}}/U$. A sample in $\boldsymbol{\mathcal{H}}_U$ is an UE set $\{\mathbf{H}[u], 1 \leq u \leq U\}$, where $\mathbf{H}[u]$ is the spatial-frequency RIS-UE channel of the u-th UE.

$\{\mathbf{H}[u], 1 \leq u \leq U\}$ are inputted to the trained SFDCEtra network to obtain the estimated channels $\{\hat{\mathbf{H}}[u], 1 \leq u \leq U\}$, 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 [54] 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 hybrid beamforming and RIS phase, 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.

Results and Discussion

5 Numerical Results

In this section, we evaluate the effectiveness of our proposed spatial-frequency domain channel extrapolation scheme as well as hybrid beamforming and RIS phase design for a 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 $N_{\text{sub}}^{\mathrm{B}} = N_y^{\mathrm{B}} N_z^{\mathrm{B}} = 64$ ($N_{\text{sub}}^{\mathrm{R}} = N_y^{\mathrm{R}} N_z^{\mathrm{R}} = 64$) 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 $N^{\mathrm{B}} = M^{\mathrm{B}} N_{\text{sub}}^{\mathrm{B}} = 256$ ($N^{\mathrm{R}} = M^{\mathrm{R}} N_{\text{sub}}^{\mathrm{R}} = 256$). For simplicity, the BS and RIS are assumed to 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 $K = 128$ and the BS's antenna gain is $G_T = 10$ dBi. Given the above parameter settings, the subarray intervals of both the BS and 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 RIS-UE channel model for the indoor environment, where the positions of the RIS, UEs, and scatterers are indicated by blue, red, and green circles, respectively. The RIS-UE LoS path is depicted by a red solid line, and the NLoS link via a scatterer is represented by a black dotted line. We assume that $U = 4$ UEs are randomly distributed over the xy-plane of the rectangular room ($W_x = 5\,\text{m}$, $W_y = 10\,\text{m}$), and the height of UEs is 1 m lower than the RIS. The number of available NLoS paths (scatterers) is set to $L_p = 5$, implying that only a single-bounce scattering mode is considered. The reflection coefficient parameters β_{RC} are set to $\mu_{\text{R}} = -5$, $\sigma_{\text{R}} = 2$. The noise power spectrum density at the UEs is $\sigma^2_{\text{NSD}} = -174$ dBm/Hz. Thus, the power of the AWGN is $\sigma_n^2 = \sigma^2_{\text{NSD}} f_s / K = -105$ 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 $d_{\text{T}} = 256$ neurons. In the transformer encoder, the number of the encoder layers is $L_{\text{T}} = 3$, where the number of heads is $h = 8$ and the position-wise MLP sub-layer has 2 fully-connected layers with $4d_{\text{T}}$ 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 $2N_s^{\text{R}}$-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, a convolutional layer has 7×7 kernel and 2 filters. The parameters of the rearranged operation are $L = 16$ and $N_p = 128$, as well as the number of neurons in the linear layer is $d_{\text{M}} = 512$. The number of mixer layers is $L_{\text{M}} = 6$, 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, $2d_{\text{M}}$,

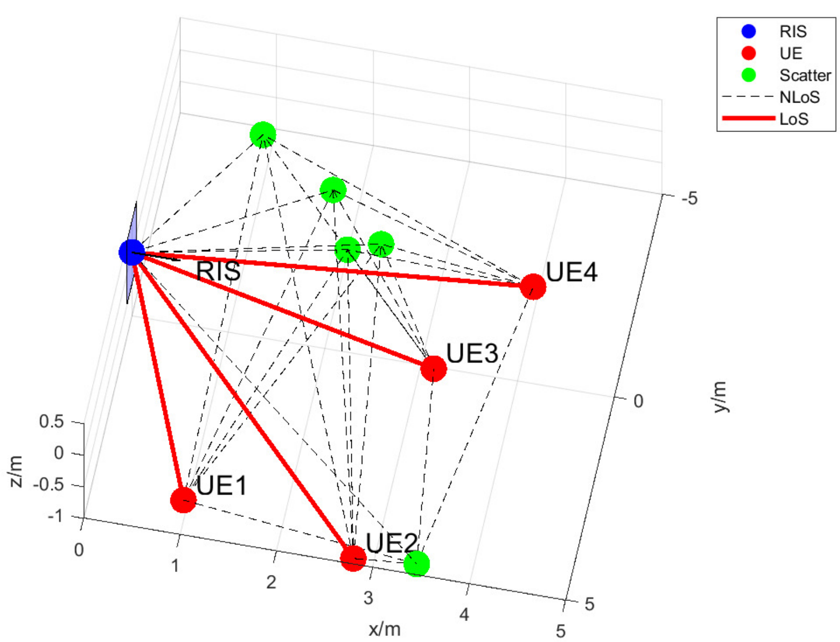


Figure 7: The schematic diagram of RIS-UE channel model for the indoor environment.

and d_{M}, respectively. The above structural parameters of the SFDCEtra network are empirically found to be appropriate.

The dataset is divided into three distinct subsets, namely the training set, validation set, and testing set, which consist of 102400, 10240, and 10240 samples respectively. Unless otherwise specified, simulations adopt the uniform element selection strategy. When considering the learning-based element selection strategy, the weight factor γ is 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 [50]. The batch size is set to 512 with 200 epochs.

5.1.3 HBFRPD Network Parameter Configuration

Again we determine the 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 dataset, which is composed of three parts, with a sample size of 25600, 2560, and 2560 respectively. The batch size is set to 32 with 180 epochs.

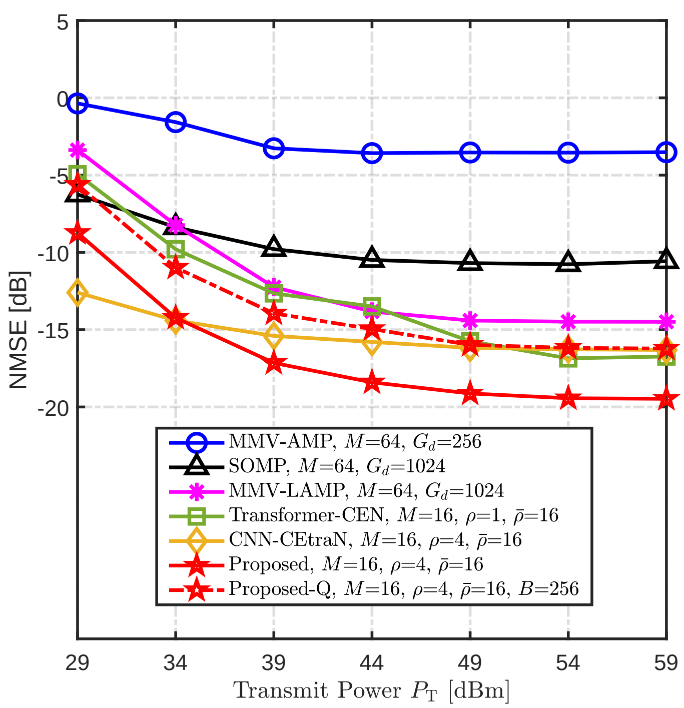


Figure 8: NMSE performance comparison of different channel estimation schemes versus transmit power P_{T}.

5.2 DL-Based Spatial-Frequency Domain Channel Extrapolation

Since the RIS element can only passively receive EM waves, selecting partial elements of the RIS 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 illustrated in Figure 8, we show the NMSE performance of the different schemes with different transmit power P_{T}. The number of NLoS paths is $L_p = 5$. We consider three model-based channel estimation benchmark algorithms, namely, the SOMP algorithm [55], the multiple-measurement-vector approximate message passing (MMV-AMP) algorithm [56], and the model-driven MMV learned AMP (MMV-LAMP) network [57], which utilize $M = 64$ OFDM symbols on all subcarriers and then directly estimate the complete channel. For the SOMP and MMV-LAMP schemes, the redundant dictionary with an oversampling ratio of 4 is utilized to further improve the performance, i.e., the number of codewords is $G_d = 1024$. However, for the MMV-AMP approach, the requirement of independent and identically distributed elements in the measurement matrix precludes the use of a 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) [51] and the CNN-based channel extrapolation network (CNN-CEtraN) [28]. For these methods, we set $M = 16$ OFDM symbols and $\bar{\rho} = 16$ subcarrier compression ratio. The transformer-based scheme turn on 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 our proposed channel extrapolation scheme outperforms the other schemes considerably in terms of NMSE performance while imposing a smaller pilot overhead. This is because exploiting spatial-frequency correlations 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 perception 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 a 2-bit quantizer, 'Proposed-Q', can still achieve very good performance. These results demonstrate that the proposed channel extrapolation scheme can achieve high reconstruction performance while ensuring low pilot and feedback overheads.

We further investigate the robustness of the proposed DL-based channel extrapolation scheme with respect to the number of multipath L_p in Figure 9. The proposed DL-based channel extrapolation scheme is trained offline using channel samples that contain $L_p = 5$ multipath components. As depicted in Figure 9, at the online estimation stage, the proposed scheme demonstrates its ability to estimate channels with different L_p without the need for retraining the entire network. Thus, our proposed scheme exhibits superior robustness and generalization capabilities in various channel conditions.

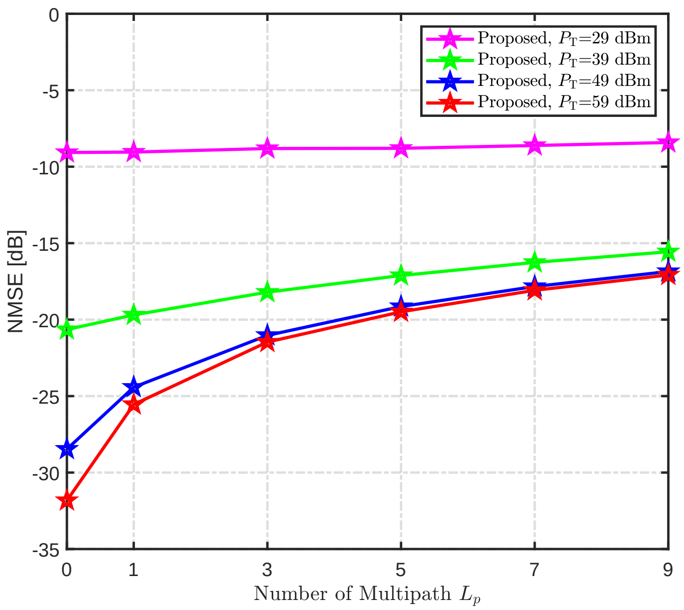


Figure 9: NMSE performance comparison of the proposed scheme versus the number of multipath L_p, given $\rho = 4$, $\bar{\rho} = 16$ and $M = 16$. Offline training is based on the channel samples with $L_p = 5$ multipath components.

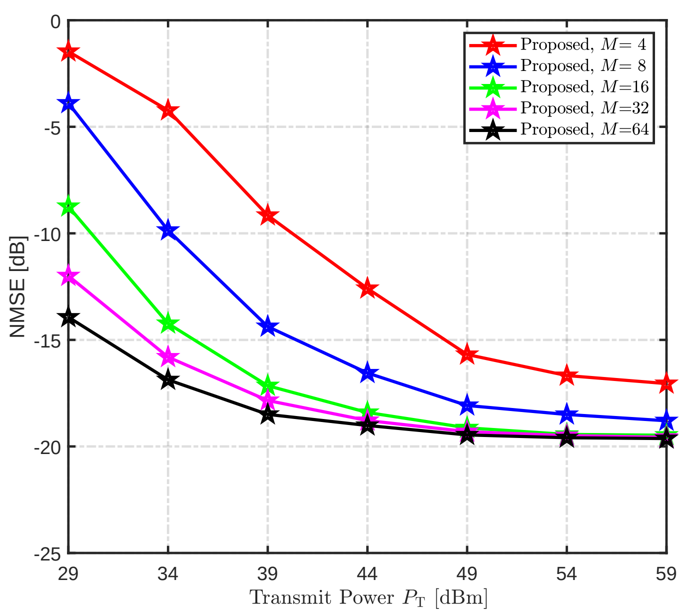


Figure 10: NMSE performance comparison of the proposed scheme with different pilot numbers versus transmit power P_{T}, given $\rho = 4$, $\bar{\rho} = 16$, $L_p = 5$.

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' corresponds to the uniform selection strategy, the curve labeled by 'Random' represents the random selection strategy, while the other three labeled by 'DL-based with 200 epochs', 'DL-based with 300 epochs', and 'DL-based with 400 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 amount of wireless communication transmission resources, the accuracy of the sub-channel estimation, and the amount 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, we can observe that the performance gap between different element selection strategies is not obvious at low compression ratios. However, at a high compression ratio (e.g., $\rho > 8$), the performance difference can be seen clearly as 'Uniform' < 'Random' < 'DL-based', which demonstrates the effectiveness of the proposed approach. Since the aperture of the random pattern is statistically larger than that of the fixed uniform pattern, the random selection strategy is better than that of the uniform selection strategy especially at a high compression ratio. The performance of the DL-based approach is relatively better than that of the first two approaches after reaching a sufficient number of training epochs—specifically 300 epochs in this scenario, as the learning of the selection network requires more epochs to converge.

To fully illustrate the effectiveness of our proposed scheme, its channel extrapolation module is verified 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 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

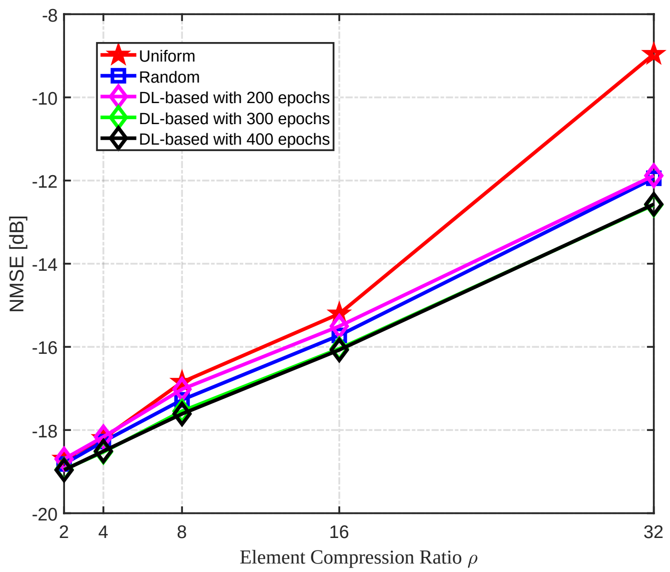
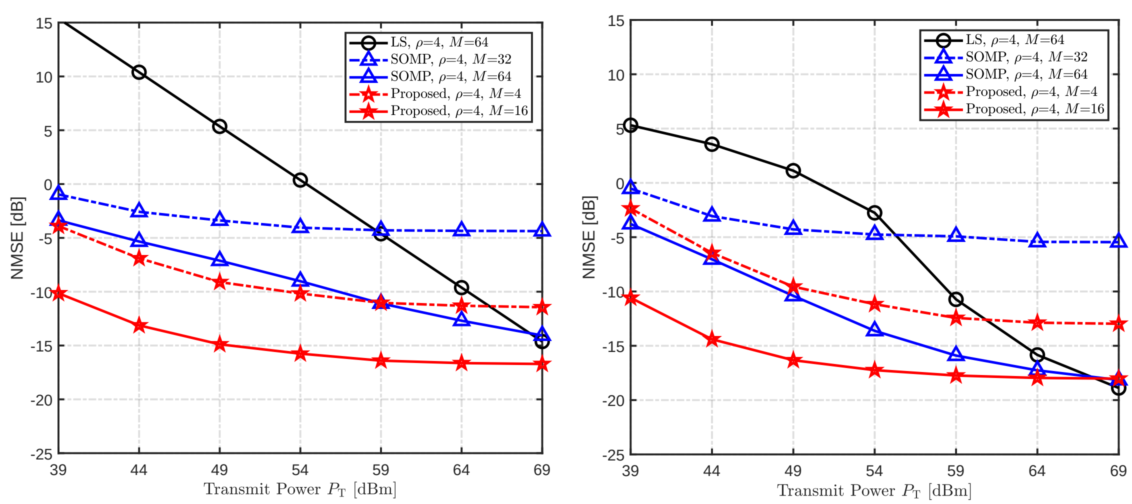


Figure 11: NMSE performance comparison of the proposed scheme with different element selection strategies versus element compression ratio ρ, given $\bar{\rho} = 16$, $L_p = 5$, $M = 16$, $P_{\mathrm{T}} = 44\,\mathrm{dBm}$.



(a) (b)

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.

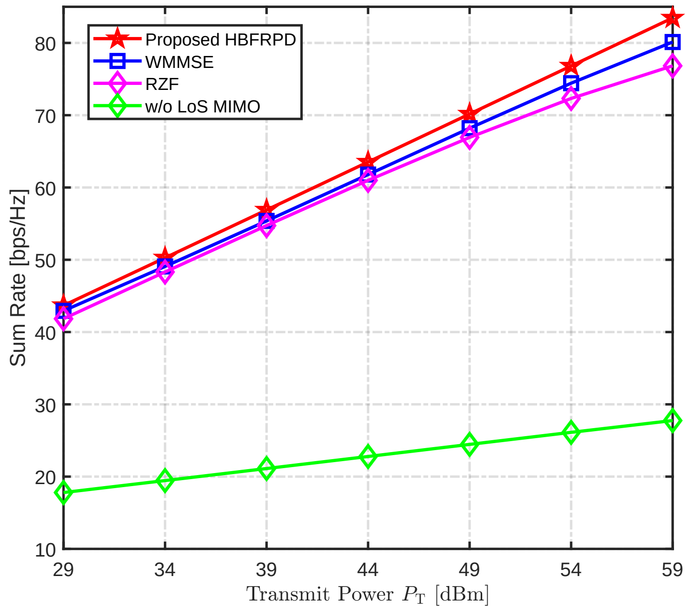
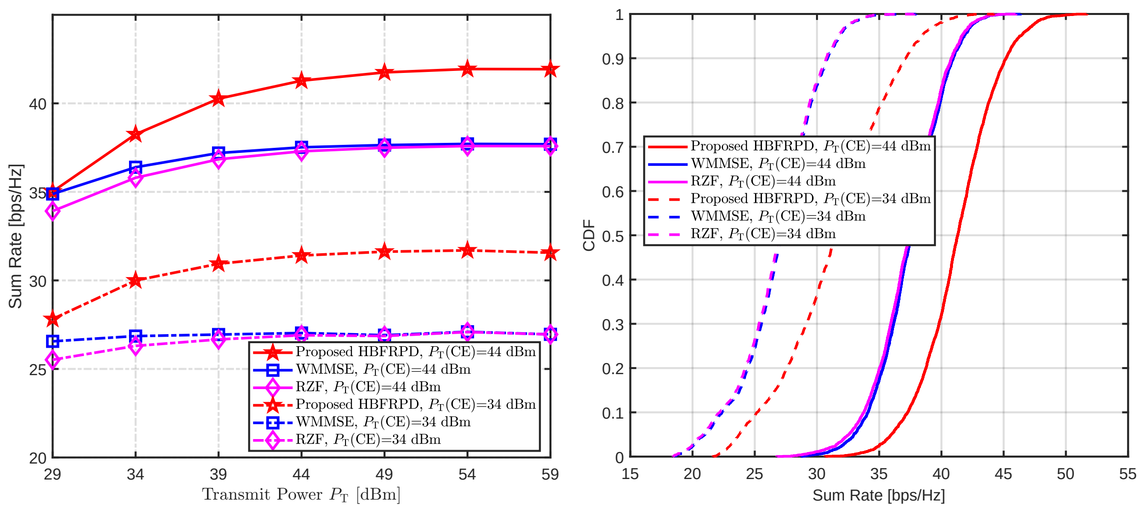


Figure 13: Sum rates achieved by different schemes versus transmit power P_T given the perfect CSI. The actual transmit power of the 'w/o LoS MIMO' case is $UP_\mathrm{T} = 4P_\mathrm{T}$.

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 learn a latent mapping among channel elements to significantly reduce the pilot overhead while achieving the same or better channel estimation performance.

5.3 DL-Based Hybrid Beamforming and RIS Phase Design

Figure 13 shows the sum rates of total UEs 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, the beam of each subarray is aligned to the corresponding associated UE. For digital beamforming design, these two comparison schemes adopt the RZF and iterative WMMSE algorithms, respectively, thus they are abbreviated as 'RZF' and 'WMMSE', respectively. We can observe that our proposed HBFRPD scheme has better performance than other schemes and the superiority is more evident as the transmit power increases. Besides, another advantage of our proposed HBFRPD scheme is that it does not require $\mathbf{F}_{\mathrm{BB}}[k], \forall k$, 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 without LoS MIMO



(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 $M = 16$.

array structure (i.e., both the BS and RIS use conventional 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 in the absence of LoS MIMO is equal to that with LoS MIMO, i.e., the transmit power in the absence of LoS MIMO 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 with LoS MIMO is much higher than that without LoS MIMO. This is because the LoS MIMO architecture can increase the sum rate linearly benefited from extra spatial multiplexing gain, while the conventional array architecture can only provide log-level growth as the SINR increases.

Although most schemes can achieve good sum rate performance under perfect CSI, the sum rate of multi-users is actually degraded due to 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 increase 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 result indicates that our proposed scheme can mitigate the interference caused by CSI errors and hence has better robustness to inaccurate CSI than the other schemes.

The cumulative distribution functions (CDFs) characterizing the sum rate performance achieved by different schemes are shown in Figure 14(b). Here, we consider the transmit power $P_{\mathrm{T}} = 44\,\mathrm{dBm}$ at the data transmission stage. Figure 14(b) shows 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 $30\,\mathrm{bps/Hz}$, 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 higher than the other two schemes. This result again confirms the superior performance of our proposed HBFRPD network over existing conventional schemes.

5.4 Computational Complexity Analysis

The computational complexity analysis of different schemes at the inference stage 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 [55], correlation operation imposes significant computational complexity, where I is the number of iterations. The MMV-AMP algorithm [56] mainly requires matrix multiplication operations, but a large number of iterations I increases its computational complexity. The MMV-LAMP algorithm [57] has a low computational complexity because DL reduces the required number of iterations. The Transformer-CEN [51] 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 [28], 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 [53], 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.

Conclusions

In this paper, we have proposed a DL-based transmission architecture 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

Table 1: Computational Complexity of Different Schemes.

Channel estimation scheme	**Complexity**	**FLOPs**	**Run time/s**
SOMP	$\mathcal{O}\left(G_dKMI+G_d^2KI\right)$	4.707 G	0.1130
MMV-AMP	$\mathcal{O}\left(MKN^{\mathrm{R}}I\right)$	5.337 G	0.7482
MMV-LAMP	$\mathcal{O}\left(MG_dKI\right)$	0.341 G	0.0689
Transformer-CEN	$\mathcal{O}\left(L_{\mathrm{T}}(Kd_{\mathrm{T}}^2+K^2d_{\mathrm{T}})\right)$	0.362 G	0.0103
CNN-CEtraN	$\mathcal{O}\left(Z_5^2N^{\mathrm{R}}KC_{32}^2\right)$	10.16 G	0.6039
Proposed	$\mathcal{O}\left(L_{\mathrm{M}}(N_p^2d_{\mathrm{M}}+N_pd_{\mathrm{M}}^2)\right)$	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	$\mathcal{O}\left(L_{\mathrm{B}}U(Kd_{\mathrm{B}}^2+K^2d_{\mathrm{B}})\right)$	0.512 G	0.0616

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 to 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 discrete phase shifter, the analysis of the complex near-field channel, and sensing-aided communications.

Acknowledgments

General: Thank Yifei Zhang, Ziwei Wan, and others for any contributions.

Funding: This work was supported by the Natural Science Foundation of China (NSFC) [Grant 62071044]; in part by the Shandong Province Natural Science Foundation [Grant ZR2022YQ62]; and in part by the Beijing Nova Program.

Author Contributions: Y. Wang, Z. Gao, and S. Chen contributed to the algorithms and analysis. Y. Wang and Z. Gao prepared the manuscript; all the authors contributed to the manuscript editing. C. Hu and D. Zheng initiated the presented concept and supervised the research.

Competing Interests: The authors declare that there is no conflict of interest regarding the publication of this paper.

Data Availability

Data are available from the corresponding author on reasonable request.

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 Journal of the Communications Society*, vol. 1, pp. 1–32, 2020.

[2] C. Lin and G. Y. Li, "Indoor Terahertz communications: How many antenna arrays are needed?" *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3097–3107, 2015.

[3] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, 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 Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450–2525, 2020.

[5] C. Huang *et al.*, "Holographic mimo surfaces for 6g wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Communications*, vol. 27, no. 5, pp. 118–125, 2020.

[6] M. D. Renzo *et al.*, "Smart radio environments empowered by reconfigurable ai meta-surfaces: An idea whose time has come," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, pp. 1–20, 2019.

[7] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157–4170, 2019.

[8] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450–2525, 2020.

[9] G. C. Alexandropoulos, K. Stylianopoulos, C. Huang, C. Yuen, M. Bennis, and M. Debbah, "Pervasive machine learning for smart radio environments enabled by reconfigurable intelligent surfaces," *Proceedings of the IEEE*, vol. 110, no. 9, pp. 1494–1525, 2022.

[10] M. Wu, Z. Gao, Y. Huang, Z. Xiao, D. W. K. Ng, and Z. Zhang, "Deep learning-based rate-splitting multiple access for reconfigurable intelligent surface-aided tera-hertz massive mimo," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1431–1451, 2023.

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

[12] M. Cui and L. Dai, "Channel estimation for extremely large-scale MIMO: Far-field or near-field?" *IEEE Transactions on Communications*, vol. 70, no. 4, pp. 2663–2677, 2022.

[13] 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 Transactions on Communications*, vol. 70, no. 2, pp. 1391–1406, 2022.

[14] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted miso wireless energy transfer," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4659–4663.

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

[16] 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 Transactions on Communications*, vol. 69, no. 6, pp. 4144–4157, 2021.

[17] A. M. Elbir, A. Papazafeiropoulos, P. Kourtessis, and S. Chatzinotas, "Deep channel learning for large intelligent surfaces aided mm-Wave massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 9, no. 9, pp. 1447–1451, 2020.

[18] 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 Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 9223–9228, 2020.

[19] J. An *et al.*, "Codebook-based solutions for reconfigurable intelligent surfaces and their open challenges," *IEEE Wireless Communications*, 2022.

[20] L. Wei *et al.*, "Joint channel estimation and signal recovery for ris-empowered multiuser communications," *IEEE Transactions on Communications*, vol. 70, no. 7, pp. 4640–4655, 2022.

[21] M. Liu, X. Li, B. Ning, C. Huang, S. Sun, and C. Yuen, "Deep learning-based channel estimation for double-ris aided massive mimo system," *IEEE Wireless Communications Letters*, vol. 12, no. 1, pp. 70–74, 2022.

[22] Z. Wan, Z. Gao, and M.-S. Alouini, "Broadband channel estimation for intelligent reflecting surface aided mmwave massive mimo systems," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, IEEE, 2020, pp. 1–6.

[23] M. B. Mashhadi and D. Gündüz, "Pruning the pilots: Deep learning-based pilot design and channel estimation for mimo-ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6315–6328, 2021.

[24] Z. Wan, Z. Gao, F. Gao, M. Di Renzo, and M.-S. Alouini, "Terahertz massive mimo with holographic reconfigurable intelligent surfaces," *IEEE Transactions on Communications*, vol. 69, no. 7, pp. 4732–4750, 2021.

[25] S. Zhang, Y. Liu, F. Gao, C. Xing, J. An, and O. A. Dobre, "Deep learning based channel extrapolation for large-scale antenna systems: Opportunities, challenges and solutions," *IEEE Wireless Communications*, vol. 28, no. 6, pp. 160–167, 2021.

[26] B. Lin, F. Gao, S. Zhang, T. Zhou, and A. Alkhateeb, "Deep learning-based antenna selection and CSI extrapolation in massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7669–7681, 2021.

[27] 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 Communications Letters*, vol. 25, no. 6, pp. 1921–1925, 2021.

[28] S. Zhang, S. Zhang, F. Gao, J. Ma, and O. A. Dobre, "Deep learning-based RIS channel extrapolation with element-grouping," *IEEE Wireless Communications Letters*, vol. 10, no. 12, pp. 2644–2648, 2021.

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

[30] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1809–1822, 2020.

[31] Y. Ahn and B. Shim, "Deep learning-based beamforming for intelligent reflecting surface-assisted mmWave systems," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 2021, pp. 1731–1734.

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

[33] S. Zhang, H. Zhang, B. Di, Y. Tan, Z. Han, and L. Song, "Beyond intelligent reflecting surfaces: Reflective-transmissive metasurface aided communications for full-dimensional coverage extension," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 905–13 909, 2020.

[34] Y. Youn *et al.*, "Demo: Transparent intelligent surfaces for sub-6 GHz and mmWave B5G/6G systems," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2022, pp. 1–2.

[35] D. Kitayama, Y. Hama, K. Goto, K. Miyachi, T. Motegi, and O. Kagaya, "Transparent dynamic metasurface for a visually unaffected reconfigurable intelligent surface: Controlling transmission/reflection and making a window into an RF lens," *Optics Express*, vol. 29, no. 18, pp. 29 292–29 307, 2021.

[36] Y. Chen, L. Yan, and C. Han, "Hybrid spherical- and planar-Wave modeling and DCNN-powered estimation of Terahertz ultra-massive MIMO channels," *IEEE Transactions on Communications*, vol. 69, no. 10, pp. 7063–7076, 2021.

[37] 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 Sensors Journal*, vol. 22, no. 14, pp. 14 540–14 552, 2022.

[38] S. Hong, C. Pan, H. Ren, K. Wang, K. K. Chai, and A. Nallanathan, "Robust transmission design for intelligent reflecting surface-aided secure communication systems with imperfect cascaded CSI," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2487–2501, 2021.

[39] Z. Chen, J. Tang, X. Y. Zhang, Q. Wu, G. Chen, and K.-K. Wong, "Robust hybrid beamforming design for multi-RIS assisted MIMO system with imperfect CSI," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.

[40] W. Xu, L. Gan, and C. Huang, "A robust deep learning-based beamforming design for RIS-assisted multiuser MISO communications with practical constraints," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 694–706, 2022.

[41] P. Larsson, "Lattice array receiver and sender for spatially orthonormal MIMO communication," in *2005 IEEE 61st Vehicular Technology Conference*, vol. 1, 2005, 192–196 Vol. 1.

[42] F. Bohagen, P. Orten, and G. E. Oien, "Optimal design of uniform planar antenna arrays for strong line-of-sight MIMO channels," in *2006 IEEE 7th Workshop on Signal Processing Advances in Wireless Communications*, 2006, pp. 1–5.

[43] X. Song, W. Rave, N. Babu, S. Majhi, and G. Fettweis, "Two-level spatial multiplexing using hybrid beamforming for millimeter-Wave backhaul," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4830–4844, 2018.

[44] L. Yan, C. Han, and J. Yuan, "Energy-efficient dynamic-subarray with fixed true-time-delay design for Terahertz wideband hybrid beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 10, pp. 2840–2854, 2022.

[45] C. Han, A. O. Bicen, and I. F. Akyildiz, "Multi-ray channel modeling and wideband characterization for wireless communications in the Terahertz band," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2402–2412, 2015.

[46] 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 Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1472–1486, 2021.

[47] J. M. Jornet and I. F. Akyildiz, "Channel modeling and capacity analysis for electromagnetic wireless nanonetworks in the Terahertz band," *IEEE Transactions on Wireless Communications*, vol. 10, no. 10, pp. 3211–3221, 2011.

[48] B. Wang, F. Gao, S. Jin, H. Lin, and G. Y. Li, "Spatial- and frequency-wideband effects in millimeter-Wave massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3393–3406, 2018.

[49] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 748–751, 2018.

[50] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

[51] Y. Wang, Z. Gao, D. Zheng, S. Chen, D. Gunduz, and H. V. Poor, "Transformer-empowered 6g intelligent networks: From massive MIMO processing to semantic communication," *IEEE Wireless Communications*, pp. 1–9, 2022.

[52] I. O. Tolstikhin *et al.*, "MLP-mixer: An all-MLP architecture for vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 261–24 272, 2021.

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

[54] 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, 1989.

[55] C.-R. Tsai, Y.-H. Liu, and A.-Y. Wu, "Efficient compressive channel estimation for millimeter-Wave large-scale antenna systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2414–2428, 2018.

[56] 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 Transactions on Signal Processing*, vol. 68, pp. 764–779, 2020.

[57] 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 Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2388–2406, 2021.