Task-Based Quantization for Channel Estimation in RIS Empowered MmWave Systems

Gyoseung Lee, In-soo Kim, Yonina C. Eldar, A. Lee Swindlehurst, Hyeongtaek Lee, Minje Kim, and Junil Choi

***Abstract*—In this paper, we investigate channel estimation for reconfigurable intelligent surface (RIS) empowered millimeter-wave (mmWave) multi-user single-input multiple-output communication systems using low-resolution quantization. Due to the high cost and power consumption of analog-to-digital converters (ADCs) in large antenna arrays and for wide signal bandwidths, designing mmWave systems with low-resolution ADCs is beneficial. To tackle this issue, we propose a channel estimation design using task-based quantization that considers the underlying hybrid analog and digital architecture in order to improve the system performance under finite bit-resolution constraints. Our goal is to accomplish a channel estimation task that minimizes the mean squared error distortion between the true and estimated channel. We develop two types of channel estimators: a cascaded channel estimator for an RIS with purely passive elements, and an estimator for the separate RIS-related channels that leverages additional information from a few semi-passive elements at the RIS capable of processing the received signals with radio frequency chains. Numerical results demonstrate that the proposed channel estimation designs exploiting task-based quantization outperform purely digital methods and can effectively approach the performance of a system with unlimited resolution ADCs. Furthermore, the proposed channel estimators are shown to be superior to baselines with small training overhead.**

***Index Terms*—Reconfigurable intelligent surface (RIS), channel estimation, task-based quantization, multi-user single-input multiple-output (MU-SIMO).**

I. Introduction

In millimeter-wave (mmWave) communication systems operating at high carrier frequencies ranging from 30-300 GHz, the availability of large bandwidths allows for high data rates, thereby supporting various requirements anticipated for future wireless communications [1]. However, high frequency propagation leads to significant path-loss and strong directivity, making the signals vulnerable to blockages. To tackle these issues,metasurface-based technologies, such as reconfigurable intelligent surfaces (RISs), which are planar metamaterial structures consisting of low-cost passive scattering elements [2]–[5], and reconfigurable holographic surfaces (RHSs), which operate as reconfigurable antenna arrays at a base station (BS) with a reduced number of radio frequency (RF) chains to enhance cost and energy efficiency [6], [7], have emerged as promising solutions for beyond-5G communications. In the case of RIS, the propagation environment can be adaptively adjusted in beneficial ways by intelligently modifying its reflection properties [2]–[5]. For instance, an RIS can effectively construct a virtual line-of-sight link between a BS and user equipment (UE) that would otherwise be blocked, thus enhancing mmWave system coverage despite the presence of obstacles [2], [4].

To fully leverage the benefits of an RIS, it is necessary to acquire accurate channel state information (CSI) at the BS or UE. However, an RIS typically lacks baseband processing capability as it comprises purely passive elements without RF chains, making CSI acquisition for the RIS-related channels inherently difficult. To tackle this challenge, most existing works propose estimating the cascaded UE-RIS-BS channel, which is generally sufficient when designing the RIS phase shifts for data transmission [8]–[11]. However, in some applications such as those involving UE localization, obtaining CSI for the individual BS-RIS and RIS-UE channels is necessary [12]–[14]. To address this issue, some existing works have suggested integrating the RIS with a few semi-passive elements equipped with their own RF chain and analog-to-digital converters (ADCs), enabling the reception and processing of training signals [15]–[19].

Prior work on channel estimation for both types of RIS has mostly assumed infinite resolution ADCs at the BS or RIS. In practical applications, developing systems that operate with low-resolution ADCs helps mitigate the overall cost and power consumption, particularly in mmWave massive multiple-input multiple-output (MIMO) systems that employ large antenna arrays and high bandwidths. Prior work that has considered low-resolution ADCs for this problem includes [20], which developed a bilinear generalized approximate message passing (BiG-AMP) approach for cascaded channel estimation in single-user multiple-input single-output systems. In [21], a leakage structure orthogonal matching pursuit (LS-OMP) algorithm was proposed to estimate the cascaded channel

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) under 6G·Cloud Research and Education Open Hub (IITP-2025-RS-2024-00428780) grant funded by the Korea government (MSIT), by IITP grant funded by the Korea government (MSIT) (No.2021-000269, Development of sub-THz band wireless transmission and access core technology for 6G Tbps data rate), and by the U.S. National Science Foundation under grants CNS-2107182 and CCF-2225575. *(Junil Choi and A. Lee Swindlehurst are the corresponding authors of this article.)*

Gyoseung Lee, Minje Kim, and Junil Choi are with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea (e-mail: {iee4432; mjkim97; junil}@kaist.ac.kr).

In-soo Kim is with Wireless Research & Development (WRD), Qualcomm Technologies, Inc., San Diego, CA, USA (e-mail: insookim@qti.qualcomm.com).

Yonina C. Eldar is with the Faculty of Math and Computer Science, Weizmann Institute of Science, Rehovot, Israel (e-mail: yonina.eldar@weizmann.ac.il)

A. Lee Swindlehurst is with the Center for Pervasive Communications and Computing, Henry Samueli School of Engineering, University of California, Irvine, CA 92697, USA (e-mail: swindle@uci.edu).

Hyeongtaek Lee is with the Department of Electronic and Electrical Engineering, Ewha Womans University, Seoul 03760, South Korea (e-mail: htlee@ewha.ac.kr).

taking into account leakage caused by grid mismatch. Also, [22] proposed a Bayesian estimator to estimate the separate RIS channels for multi-user single-input multiple-output (MU-SIMO) systems, exploiting additional information from semi-passive elements equipped with low-resolution ADCs. In [23], a generalized multiple measurement vector (GMMV) problem and matrix completion were used to estimate the individual RIS channels, and a hierarchical message passing-based algorithm was proposed based on AMP-like approximations. However, in mmWave systems, deploying large antenna arrays at the BS still results in significant overhead in terms of the total number of quantized bits that must be processed during the channel estimation phase. Our goal in this work is to design a channel estimator using low-resolution ADCs that achieves high estimation accuracy with a minimal number of quantization bits at the receiver.

Among the various quantization strategies that have been proposed, [24]–[28] have shown that co-designing the quantizer together with a hybrid analog and digital architecture can dramatically improve the performance. In conventional "task-ignorant" systems where the quantizer is designed solely to discretize the measurements, the desired task (recovering the transmitted signal vector) is performed separately in the digital domain. On the other hand, *task-based quantization* captures lower-dimensional information from signals in the analog domain based on the specific system task, allowing for a reduction in the overall number of bits required to accomplish the task, thereby minimizing the power consumption and memory requirements. Task-based quantization has been applied in various applications such as channel estimation [24], [25], [28], quadratic function recovery [27], MIMO radar [29], and dual functional radar and communication systems [30]. However, prior work has not addressed task-based quantization in RIS-aided communication systems, except for our preliminary work in [31] which only addressed cascaded channel estimation.

In this paper, motivated by the success of task-based quantization over conventional all-digital approaches independent of the task [24]–[28], we apply task-based quantization to the problem of channel estimation for RIS-empowered MU-SIMO systems, where identical scalar ADCs are used to mitigate hardware cost and energy limitations. Specifically, we consider two approaches: 1) cascaded channel estimation for cases involving an RIS with purely passive elements, and 2) separate estimation of the RIS channels by exploiting an RIS with a few semi-passive elements connected with low-resolution ADCs that provide additional local information. Note that deploying low-resolution ADCs for the semi-passive elements of an RIS is consistent with the goal that such surfaces operate with minimal additional cost and power consumption. For both channel estimation approaches, the goal is to minimize the mean squared error (MSE) in estimating the channel for a hybrid analog and digital architecture by proper choice of the analog and digital combining matrices and the ADC quantization range or support. Our main contributions are as follows:

- For an RIS equipped with purely passive elements, we first propose a cascaded channel estimation framework using task-based quantization to minimize the total number of quantization bits at the BS while employing low-resolution ADCs. Combining the mmWave channel structure with task-based quantization, we demonstrate that, under typical mmWave system conditions, the dimensionality of the BS observations in the analog domain can be significantly reduced, depending on the number of propagation paths that comprise the RIS-related channels. Due to the limited RF scattering in mmWave bands, the number of propagation paths is limited, which enables the proposed channel estimator to achieve a significant reduction in the total number of quantization bits at the BS compared to task-ignorant approaches.
- For an RIS with semi-passive elements, we propose a two-stage estimator of the RIS channel components using task-based quantization that first estimates the UE-RIS channels based on observations from the semi-passive elements of the RIS, and subsequently estimates the BS-RIS channel using the UE-RIS channel estimate and observations at the BS. This estimator is the first to consider low-resolution quantization at both the BS and the semi-passive elements of the RIS. We also demonstrate that the total number of quantization bits required to estimate the individual channel components at the BS can be further reduced compared to the proposed cascaded channel estimator.
- Our numerical results verify that, in terms of channel estimation accuracy, the proposed cascaded channel estimator closely approaches the performance of the minimum MSE (MMSE) estimator without quantization [32], and outperforms a purely digital approach assuming identical bit-resolution ADCs in which a task-ignorant MMSE estimator based on [24] is applied to the quantized measurements. Subsequently, we demonstrate that the proposed channel estimator for the separate RIS channels shows performance comparable to an MMSE estimator without quantization even with a small number of semi-passive elements. Finally, we show that the proposed channel estimators outperform baseline approaches including the methods of [20], [22] that use low-resolution ADCs without task-based quantization, achieving a significant reduction in the total number of quantization bits and a small training overhead.

The rest of the paper is organized as follows. Section II presents the system model for the assumed RIS-aided MU-SIMO system. In Section III, the problem formulations for obtaining the cascaded and separate channel estimates are investigated. The basic concept of task-based quantization is explained in Section IV, and the proposed channel estimators based on task-based quantization are developed in Sections V and VI. Numerical results for the proposed approaches are provided in Section VII, and we conclude the paper in Section VIII.

$Notation$: Lower- and upper-case boldface letters represent column vectors and matrices, respectively. The conjugate, transpose, and conjugate transpose of a matrix $\mathbf{A}$ are denoted by $\mathbf{A}^*$, $\mathbf{A}^{\mathrm{T}}$, and $\mathbf{A}^{\mathrm{H}}$, respectively. For a square matrix $\mathbf{A}$,

TABLE I: Summary of notation used in the paper.

Notation	Description
N	Number of BS antennas
$L = L_\mathrm{h} L_\mathrm{v}$	Number of total RIS elements
$L_\mathrm{h}, L_\mathrm{v}$	Number of horizontal and vertical elements of RIS
$L_\mathrm{p}, L_\mathrm{a}$	Number of passive and semi-passive RIS elements
K	Number of UEs
$T_\mathrm{p} = T\tau$	Number of total time slots for pilot training
T	Number of subblocks in T_p
τ	Number of time slots in each subblock
P_k	Transmit power at the k-th UE
M_RB	Number of propagation paths in the RIS-BS channel
$M_{\mathrm{UR},k}$	Number of propagation paths in the k-th UE-RIS channel
$\mathbf{G}$	Uplink channel from the RIS to the BS
$\mathbf{f}_k$	Uplink channel from the k-th UE to the RIS
$\mathbf{F}$	Uplink channel from the UEs to the RIS
$\mathbf{C}$	Cascaded channel between the BS and the UEs
$\boldsymbol{\theta}[t]$	Vector of RIS reflection coefficients
$\boldsymbol{\Omega}$, $\boldsymbol{\Omega}^\mathrm{c}$	Index vectors of semi-passive and passive RIS elements
$\mathbf{s}[t]$	Vector of passive RIS reflection coefficients
$\mathbf{y}[t], \mathbf{z}[t]$	Received signals at the BS and semi-passive elements
$\boldsymbol{\pi}_\mathbf{z}$	Quantized signals at the semi-passive RIS elements
$x_k[t]$	Transmit signal from the k-th UE
$\mathbf{n}_\mathrm{B}[t], \mathbf{n}_\mathrm{R}[t]$	AWGN noise at the BS and RIS
$Q_{\tilde{\nu}}(\cdot)$	Scalar ADC with resolution $\log_2 \tilde{\nu}$ bits
$\gamma_\mathbf{c}, \gamma_\mathbf{g}$	ADC thresholds for $\mathbf{C}$ and $\mathbf{G}$
$\mathbf{B}_\mathbf{c}, \mathbf{B}_\mathbf{g}$	Analog combining matrices for $\mathbf{C}$ and $\mathbf{G}$
$\mathbf{D}_\mathbf{c}, \mathbf{D}_\mathbf{f}, \mathbf{D}_\mathbf{g}$	Digital processing matrices for $\mathbf{C}$, $\mathbf{F}$, and $\mathbf{G}$
$G_\mathbf{c}, G_\mathbf{g}$	Number of scalar ADCs for $\mathbf{C}$ and $\mathbf{G}$
$\boldsymbol{\pi}_\mathbf{c}, \boldsymbol{\pi}_\mathbf{g}$	Outputs of scalar ADCs for $\mathbf{C}$ and $\mathbf{G}$

$\mathrm{tr}(\mathbf{A})$ and $\mathbf{A}^{-1}$ are respectively the trace and inverse of $\mathbf{A}$. The quantities $[\mathbf{A}]_{i,:}$ and $[\mathbf{A}]_{i,j}$ denote the i-th row and the (i,j)-th entry of matrix $\mathbf{A}$, respectively. The operator $\mathrm{vec}(\mathbf{A})$ denotes the vectorization of matrix $\mathbf{A}$, and $\mathrm{unvec}(\mathbf{a})$ denotes the inverse operation. The notation $\mathrm{diag}(\mathbf{a})$ represents a diagonal matrix whose diagonal elements correspond to the entries of the vector $\mathbf{a}$, and $\mathrm{blkdiag}(\cdot)$ denotes a block-diagonal matrix with blocks defined by the argument. The ℓ_p-norm of a vector $\mathbf{a}$ and the Frobenius-norm of a matrix $\mathbf{A}$ are respectively denoted by $\|\mathbf{a}\|_p$ and $\|\mathbf{A}\|_\mathrm{F}$. A circularly symmetric complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$ is represented using $\mathcal{CN}(\boldsymbol{\mu}, \mathbf{K})$. The quantities $\mathbf{0}_m$, $\mathbf{1}_m$, and $\mathbf{I}_m$ represent the $m \times 1$ all-zeros vector, the $m \times 1$ all-ones vector, and the $m \times m$ identity matrix, respectively. The expressions $|a|$, $\angle a$, $\mathrm{Re}(a)$, and $\mathrm{Im}(a)$ represent the magnitude, angle, real part, and imaginary part of a complex number a, respectively. The expression a^+ denotes $\max(a, 0)$ for a real number a. The Hadamard product, Kronecker product, and Khatri-Rao product of matrices $\mathbf{A}$ and $\mathbf{B}$ are denoted as $\mathbf{A} \odot \mathbf{B}$, $\mathbf{A} \otimes \mathbf{B}$, and $\mathbf{A} \diamond \mathbf{B}$, respectively. For convenience, we summarize the notation that will be used throughout the paper in Table I.

II. System Model

A. Signal model

We investigate an uplink RIS-aided MU-SIMO scenario as illustrated in Fig. 1, where a BS with N antennas in a uniform linear array (ULA) communicates with K single-antenna UEs. We assume a mmWave setting with large N, so the BS adopts both hybrid analog and digital beamforming together with low-resolution ADCs to reduce the required cost and power consumption. A single RIS composed of a uniform planar array (UPA) with $L = L_\mathrm{h} L_\mathrm{v}$ elements is present, where L_h and L_v denote the number of horizontal rows and vertical columns, respectively. The configuration of the RIS is under the control of the BS via a low-rate network connection. Among the L RIS elements, L_p passive elements reflect the incoming signals, while $L_\mathrm{a} = L - L_\mathrm{p} \ll L$ semi-passive elements operate in sensing mode during the channel estimation phase and in reflecting mode during the data transmission phase[1]. Since the RIS is intended to be a low-power device, the baseband outputs of the semi-passive elements are sampled with low-resolution ADCs [18], [22]. In general, the goal is to use as few semi-passive elements as possible to minimize the additional cost and power consumption at the RIS.

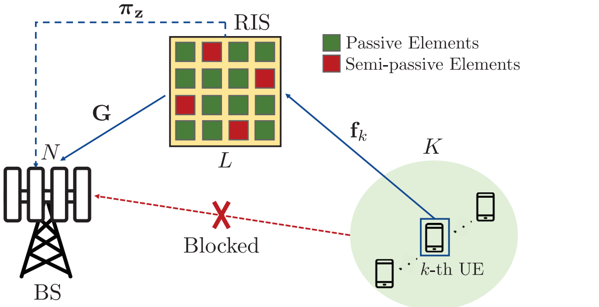

Fig. 1: An example of an uplink RIS-aided mmWave MU-SIMO system with semi-passive elements. There are N BS antennas, K UEs with a single antenna, and L total RIS elements.

As in [18], [34], [35], the channels between the BS and UEs are assumed to be entirely blocked. Based on a block fading channel model, we assume that the BS-RIS and RIS-UE channels are constant during T_p time slots. To indicate how the RIS is partitioned into semi-passive and passive elements, let $\boldsymbol{\Omega} \in \{0,1\}^L$ and $\boldsymbol{\Omega}^\mathrm{c} = \mathbf{1}_L - \boldsymbol{\Omega}$ denote the index vectors of the semi-passive and the passive elements, respectively, such that $\|\boldsymbol{\Omega}\|_0 = L_\mathrm{a}$. Focusing on uplink transmissions, the signal received at the BS prior to the hybrid analog/digital combining is given by

$$\mathbf{y}[t] = \sum_{k=1}^{K} \mathbf{G}\,\mathrm{diag}(\underbrace{\boldsymbol{\Omega}^\mathrm{c} \odot \boldsymbol{\theta}[t]}_{=\mathbf{s}[t]})\mathbf{f}_k x_k[t] + \mathbf{n}_\mathrm{B}[t], \tag{1}$$

where $x_k[t] \in \mathbb{C}$ is the transmit signal from the k-th UE satisfying $\mathbb{E}[|x_k[t]|^2] \leq P_k$, and $\mathbf{n}_\mathrm{B}[t] \sim \mathcal{CN}(\mathbf{0}_N, \sigma_\mathrm{B}^2 \mathbf{I}_N)$ is additive white Gaussian noise (AWGN) at the BS with variance σ_B^2. The uplink channel from the RIS to the BS is

[1] A semi-passive RIS element is equipped with a sensor implemented with a low cost receive RF chain, allowing it to process the received signal [18], [33]. In the sensing mode, the reflecting element is turned off, and the sensor is activated to receive pilot signals. In the reflection mode, the sensor is deactivated, and the reflecting element is turned on, operating like a conventional passive RIS element.

denoted by $\mathbf{G} \in \mathbb{C}^{N\times L}$, and the uplink channel from the k-th UE to the RIS is represented by $\mathbf{f}_k \in \mathbb{C}^{L\times 1}$. The vector of RIS reflection coefficients is $\boldsymbol{\theta}[t] = [\theta_1[t], \cdots, \theta_L[t]]^{\mathrm{T}} \in \mathbb{C}^{L\times 1}$, with reflection amplitudes $|\theta_\ell[t]| = 1$ and phase shifts $\angle\theta_\ell[t] \in [0, 2\pi)$. The model for the hybrid beamforming and low-resolution quantization of $\mathbf{y}[t]$ will be discussed in Sections IV and V. Note that, although we assume a narrowband channel model throughout this paper, the proposed channel estimators are also applicable without loss of generality to wideband systems employing orthogonal frequency division multiplexing (OFDM), since each subcarrier typically experiences a narrowband channel.

If the RIS is equipped with a few semi-passive elements, it is possible to estimate the individual RIS channels. In this case, the RIS sends the locally sampled information from these elements to the BS. The signal received at time t by the semi-passive RIS elements prior to quantization is denoted by $\mathbf{z}[t] = \mathbf{\Omega} \odot (\mathbf{F}\mathbf{x}[t] + \mathbf{n}_{\mathrm{R}}[t])$, with $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_K] \in \mathbb{C}^{L\times K}$, $\mathbf{x}[t] = [x_1[t], \cdots, x_K[t]]^{\mathrm{T}} \in \mathbb{C}^{K\times 1}$, and where $\mathbf{n}_{\mathrm{R}}[t] \sim \mathcal{CN}(\mathbf{0}_L, \sigma_{\mathrm{R}}^2\mathbf{I}_L)$ represents AWGN at the RIS with variance σ_{R}^2. Stacking these signals over the T_{p} time slots leads to

$$\begin{aligned}\mathbf{Z} &= [\mathbf{z}[1], \cdots, \mathbf{z}[T_{\mathrm{p}}]] \\ &= \bar{\mathbf{\Omega}} \odot (\mathbf{F}\mathbf{X}_{\mathrm{I}} + \mathbf{N}_{\mathrm{R}}),\end{aligned} \tag{2}$$

where $\bar{\mathbf{\Omega}} = \mathbf{\Omega} \otimes \mathbf{1}_{T_{\mathrm{p}}}^{\mathrm{T}} \in \mathbb{C}^{L\times T_{\mathrm{p}}}$, $\mathbf{X}_{\mathrm{I}} = [\mathbf{x}[1], \cdots, \mathbf{x}[T_{\mathrm{p}}]] \in \mathbb{C}^{K\times T_{\mathrm{p}}}$, and $\mathbf{N}_{\mathrm{R}} = [\mathbf{n}_{\mathrm{R}}[1], \cdots, \mathbf{n}_{\mathrm{R}}[T_{\mathrm{p}}]] \in \mathbb{C}^{L\times T_{\mathrm{p}}}$. Both the real and imaginary parts of each element in $\mathbf{Z}$ are quantized using identical low-resolution ADCs. Denoting the vectorized representation of (2) as $\mathbf{z} = \mathrm{vec}(\mathbf{Z})$, the quantized version of $\mathbf{z}$ is denoted by $\boldsymbol{\pi}_{\mathbf{z}}$, which is forwarded to the BS. The i-th element of $\boldsymbol{\pi}_{\mathbf{z}}$ is given by

$$\pi_{\mathbf{z},i} = \begin{cases} Q_{\tilde{\nu}}(z_i) & \text{if } [\mathrm{vec}(\bar{\mathbf{\Omega}})]_i = 1 \\ 0 & \text{if } [\mathrm{vec}(\bar{\mathbf{\Omega}})]_i = 0, \end{cases} \tag{3}$$

where $Q_{\tilde{\nu}}(\cdot)$ denotes the ADC quantization operation with $\log_2 \tilde{\nu}$ bits of resolution, applied separately to the real and imaginary parts. The specific ADC model will be discussed in Section IV-A.

Remark: As described in (1) and (2), we consider a model in which the location of the semi-passive elements is fixed. This is different from the model in [22], [23], where a switching network between the RF chains and RIS elements is introduced to allow the selection of semi-passive elements to change at each time instant and improve channel estimation performance. However, such a network requires analog connections from the RF chains to all RIS elements via switches, which would be difficult to implement in practice, especially for a large RIS. As will be shown in Section VII-C, the proposed individual channel estimator based on task-based quantization without the switching network achieves superior channel estimation performance compared to [22], [23].

B. Channel model

To properly describe the mmWave channels, we adopt a geometric model [18], [36], [37] due to the limited scattering environment in mmWave bands. The channel $\mathbf{G}$ from the RIS to the BS is then represented by

$$\mathbf{G} = \sqrt{\frac{NL}{M_{\mathrm{RB}}}} \sum_{m=1}^{M_{\mathrm{RB}}} \alpha_{\mathrm{RB},m}\mathbf{a}_{\mathrm{B}}(\phi_m)\mathbf{a}_{\mathrm{R}}^{\mathrm{H}}(\theta_{\mathrm{RB},m}^{\mathrm{Azi}}, \theta_{\mathrm{RB},m}^{\mathrm{Ele}}), \tag{4}$$

where M_{RB} denotes the number of propagation paths between the BS and RIS, and $\alpha_{\mathrm{RB},m} \sim \mathcal{CN}(0, \sigma_{\mathrm{RB}}^2)$ is the complex gain of the m-th path, which is independent and identically distributed (i.i.d.) with variance σ_{RB}^2 that depends on the path-loss. The array steering vectors at the BS and RIS are expressed as $\mathbf{a}_{\mathrm{B}}(\cdot) \in \mathbb{C}^{N\times 1}$ and $\mathbf{a}_{\mathrm{R}}(\cdot) \in \mathbb{C}^{L\times 1}$, respectively, and $\mathbf{a}_{\mathrm{B}}(\phi_m)$ is represented by

$$\mathbf{a}_{\mathrm{B}}(\phi_m) = \frac{1}{\sqrt{N}}[1, e^{j\omega_m}, \cdots, e^{j(N-1)\omega_m}]^{\mathrm{T}}, \tag{5}$$

where ϕ_m is the angle of arrival (AoA) of the m-th path and $\omega_m = 2\pi d_{\mathrm{B}} \sin(\phi_m)/\lambda_{\mathrm{c}}$ represents the spatial frequency with carrier wavelength λ_{c} and BS antenna spacing d_{B}. The steering vector at the RIS is

$$\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{RB},m}^{\mathrm{Azi}}, \theta_{\mathrm{RB},m}^{\mathrm{Ele}}) = \mathbf{a}_{\mathrm{R,v}}(\psi_m) \otimes \mathbf{a}_{\mathrm{R,h}}(\varphi_m), \tag{6}$$

where $\theta_{\mathrm{RB},m}^{\mathrm{Azi}}$ and $\theta_{\mathrm{RB},m}^{\mathrm{Ele}}$ are the azimuth and elevation angles of departure (AoD) of the m-th path, respectively. The RIS steering vectors along the vertical and horizontal directions are respectively given by

$$\begin{aligned}\mathbf{a}_{\mathrm{R,v}}(\psi_m) &= \frac{1}{\sqrt{L_{\mathrm{v}}}}\left[1, e^{j\psi_m}, \cdots, e^{j(L_{\mathrm{v}}-1)\psi_m}\right]^{\mathrm{T}} \\ \mathbf{a}_{\mathrm{R,h}}(\varphi_m) &= \frac{1}{\sqrt{L_{\mathrm{h}}}}\left[1, e^{j\varphi_m}, \cdots, e^{j(L_{\mathrm{h}}-1)\varphi_m}\right]^{\mathrm{T}},\end{aligned} \tag{7}$$

where $\psi_m = 2\pi d_{\mathrm{R,v}} \sin(\theta_{\mathrm{RB},m}^{\mathrm{Ele}})/\lambda_{\mathrm{c}}$ is the spatial frequency along the vertical direction with vertical spacing $d_{\mathrm{R,v}}$, and $\varphi_m = 2\pi d_{\mathrm{R,h}} \cos(\theta_{\mathrm{RB},m}^{\mathrm{Ele}}) \sin(\theta_{\mathrm{RB},m}^{\mathrm{Azi}})/\lambda_{\mathrm{c}}$ is the horizontal spatial frequency with horizontal spacing $d_{\mathrm{R,h}}$. For simplicity, we reformulate $\mathbf{G}$ in (4) as

$$\mathbf{G} = \mathbf{A}_{\mathrm{B,RB}}\, \mathrm{diag}(\boldsymbol{\alpha}_{\mathrm{RB}})\mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}}, \tag{8}$$

where $\mathbf{A}_{\mathrm{B,RB}} = [\mathbf{a}_{\mathrm{B}}(\phi_1), \cdots, \mathbf{a}_{\mathrm{B}}(\phi_{M_{\mathrm{RB}}})] \in \mathbb{C}^{N\times M_{\mathrm{RB}}}$, $\mathbf{A}_{\mathrm{R,RB}} = [\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{RB},1}^{\mathrm{Azi}}, \theta_{\mathrm{RB},1}^{\mathrm{Ele}}), \cdots, \mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{RB},M_{\mathrm{RB}}}^{\mathrm{Azi}}, \theta_{\mathrm{RB},M_{\mathrm{RB}}}^{\mathrm{Ele}})] \in \mathbb{C}^{L\times M_{\mathrm{RB}}}$, and $\boldsymbol{\alpha}_{\mathrm{RB}} = \sqrt{\frac{NL}{M_{\mathrm{RB}}}}[\alpha_{\mathrm{RB},1}, \cdots, \alpha_{\mathrm{RB},M_{\mathrm{RB}}}]^{\mathrm{T}} \in \mathbb{C}^{M_{\mathrm{RB}}\times 1}$.

Similarly, the channel $\mathbf{f}_k$ from the k-th UE to the RIS is expressed as

$$\mathbf{f}_k = \sqrt{\frac{L}{M_{\mathrm{UR},k}}} \sum_{m=1}^{M_{\mathrm{UR},k}} \alpha_{\mathrm{UR},k,m}\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{UR},k,m}^{\mathrm{Azi}}, \theta_{\mathrm{UR},k,m}^{\mathrm{Ele}}), \tag{9}$$

where $M_{\mathrm{UR},k}$ is the number of propagation paths for this channel, $\alpha_{\mathrm{UR},k,m} \sim \mathcal{CN}(0, \sigma_{\mathrm{UR},k}^2)$ is the i.i.d. complex gain of the m-th path with variance $\sigma_{\mathrm{UR},k}^2$, and the RIS steering vector $\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{UR},k,m}^{\mathrm{Azi}}, \theta_{\mathrm{UR},k,m}^{\mathrm{Ele}})$ is defined similarly to (6) with azimuth and elevation AoAs $\theta_{\mathrm{UR},k,m}^{\mathrm{Azi}}$ and $\theta_{\mathrm{UR},k,m}^{\mathrm{Ele}}$, respectively. Based on (9), $\mathbf{f}_k$ can be reformulated as

$$\mathbf{f}_k = \mathbf{A}_{\mathrm{R,UR},k}\boldsymbol{\alpha}_{\mathrm{UR},k}, \tag{10}$$

where $\mathbf{A}_{\mathrm{R,UR},k} = [\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{UR},k,1}^{\mathrm{Azi}}, \theta_{\mathrm{UR},k,1}^{\mathrm{Ele}}), \cdots, \mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{UR},k,M_{\mathrm{UR},k}}^{\mathrm{Azi}}, \theta_{\mathrm{UR},k,M_{\mathrm{UR},k}}^{\mathrm{Ele}})] \in \mathbb{C}^{L\times M_{\mathrm{UR},k}}$, and $\boldsymbol{\alpha}_{\mathrm{UR},k} = \sqrt{\frac{L}{M_{\mathrm{UR},k}}}[\alpha_{\mathrm{UR},k,1}, \cdots, \alpha_{\mathrm{UR},k,M_{\mathrm{UR},k}}]^{\mathrm{T}} \in \mathbb{C}^{M_{\mathrm{UR},k}\times 1}$.

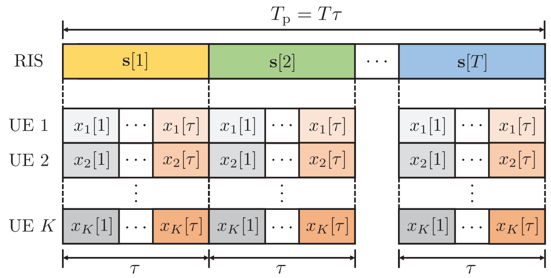

Fig. 2: The uplink time-domain transmission design for estimating the cascaded channel.

III. Problem Formulation

In the following, we will consider two types of channel estimation problems for the above RIS-aided system: *1) Cascaded channel estimation* and *2) Individual RIS channel estimation.* The first problem is encountered in scenarios where all RIS elements are purely passive, and no information is available at the RIS from the training signals. Cascaded channel CSI is generally sufficient in many situations such as BS transceiver design, but there are applications where knowledge of the individual RIS channels is necessary [12]–[14]. One approach for separately estimating the BS-RIS and RIS-UE channels involves the use of semi-passive elements at the RIS that can receive and process the training data to eliminate ambiguities normally present in the cascaded channel.

A. Cascaded channel estimation

In this case, we assume the RIS consists of purely passive elements, i.e., $L_\mathrm{p} = L$ and $\boldsymbol{\Omega} = \mathbf{0}_L$. The uplink signal design used to estimate the cascaded channel is depicted in Fig. 2, where a block of length T_p is composed of T subblocks of τ time slots each. Within each subblock, the RIS configuration remains unchanged, and all UEs transmit the same pilot sequences in each of the T subblocks. The RIS passive reflection vector and the transmit signal from the k-th UE in the u-th time slot of the t-th subblock are then respectively given by

$$\begin{aligned} \mathbf{s}[t,u] &= \mathbf{s}[t], \;\; 1 \leq u \leq \tau, \\ x_k[t,u] &= x_k[u], \;\; 1 \leq t \leq T, \;\; \forall k = 1, \cdots, K. \end{aligned} \tag{11}$$

Based on (1) and (11), the signal received at the BS in the u-th time slot of the t-th subblock, $\mathbf{y}[t,u] = \mathbf{y}[(t-1)\tau + u]$, is expressed as

$$\begin{aligned} \mathbf{y}[t,u] &= \sum_{k=1}^{K} \mathbf{G}\,\mathrm{diag}(\mathbf{s}[t])\mathbf{f}_k x_k[u] + \mathbf{n}_\mathrm{B}[t,u] \\ &= \mathbf{G}\,\mathrm{diag}(\mathbf{s}[t])\mathbf{F}\mathbf{x}[u] + \mathbf{n}_\mathrm{B}[t,u]. \end{aligned} \tag{12}$$

The observation matrix at the BS is obtained by stacking (12) over τ time slots for the t-th subblock:

$$\begin{aligned} \mathbf{Y}[t] &= [\mathbf{y}[t,1], \cdots, \mathbf{y}[t,\tau]] \\ &= \mathbf{G}\,\mathrm{diag}(\mathbf{s}[t])\mathbf{F}\mathbf{X}_\mathrm{C} + \mathbf{N}_\mathrm{B}[t], \end{aligned} \tag{13}$$

where $\mathbf{X}_\mathrm{C} = [\mathbf{x}[1], \cdots, \mathbf{x}[\tau]] \in \mathbb{C}^{K \times \tau}$, and $\mathbf{N}_\mathrm{B}[t] = [\mathbf{n}_\mathrm{B}[t,1], \cdots, \mathbf{n}_\mathrm{B}[t,\tau]] \in \mathbb{C}^{N \times \tau}$. Using the identities $\mathrm{vec}(\mathbf{M}_1\mathbf{M}_2\mathbf{M}_3) = (\mathbf{M}_3^\mathrm{T} \otimes \mathbf{M}_1)\mathrm{vec}(\mathbf{M}_2)$ and $\mathrm{vec}(\mathbf{M}_1\,\mathrm{diag}(\mathbf{m})\mathbf{M}_2) = (\mathbf{M}_2^\mathrm{T} \diamond \mathbf{M}_1)\mathbf{m}$, $\mathbf{Y}[t]$ in (13) can be vectorized as

$$\mathrm{vec}(\mathbf{Y}[t]) = (\mathbf{X}_\mathrm{C}^\mathrm{T} \otimes \mathbf{I}_N)(\mathbf{F}^\mathrm{T} \diamond \mathbf{G})\mathbf{s}[t] + \mathrm{vec}(\mathbf{N}_\mathrm{B}[t]). \tag{14}$$

Finally, the complete set of BS observations over all T subblocks in (14) is collected in the matrix $\mathbf{Y} \in \mathbb{C}^{N\tau \times T}$:

$$\begin{aligned} \mathbf{Y} &= [\mathrm{vec}(\mathbf{Y}[1]), \cdots, \mathrm{vec}(\mathbf{Y}[T])] \\ &= (\mathbf{X}_\mathrm{C}^\mathrm{T} \otimes \mathbf{I}_N)(\mathbf{F}^\mathrm{T} \diamond \mathbf{G})\mathbf{S} + \mathbf{N}_\mathrm{B}, \end{aligned} \tag{15}$$

where $\mathbf{S} = [\mathbf{s}[1], \cdots, \mathbf{s}[T]] \in \mathbb{C}^{L \times T}$, and $\mathbf{N}_\mathrm{B} = [\mathrm{vec}(\mathbf{N}_\mathrm{B}[1]), \cdots, \mathrm{vec}(\mathbf{N}_\mathrm{B}[T])] \in \mathbb{C}^{N\tau \times T}$.

From (15), let $\mathbf{C} = \mathbf{F}^\mathrm{T} \diamond \mathbf{G}$ be the cascaded channel between the BS and the UEs. Using the identity $\mathrm{vec}(\mathbf{M}_1\mathbf{M}_2\mathbf{M}_3) = (\mathbf{M}_3^\mathrm{T} \otimes \mathbf{M}_1)\mathrm{vec}(\mathbf{M}_2)$, the received signals at the BS in (15) can be vectorized as

$$\begin{aligned} \mathrm{vec}(\mathbf{Y}) &= (\mathbf{S}^\mathrm{T} \otimes \mathbf{X}_\mathrm{C}^\mathrm{T} \otimes \mathbf{I}_N)\mathrm{vec}(\mathbf{C}) + \mathrm{vec}(\mathbf{N}_\mathrm{B}) \\ &= \bar{\mathbf{S}}\mathbf{c} + \mathbf{n}_\mathrm{B} \\ &\triangleq \mathbf{y}, \end{aligned} \tag{16}$$

where $\bar{\mathbf{S}} = (\mathbf{S}^\mathrm{T} \otimes \mathbf{X}_\mathrm{C}^\mathrm{T} \otimes \mathbf{I}_N)$, $\mathbf{c} = \mathrm{vec}(\mathbf{C})$, and $\mathbf{n}_\mathrm{B} = \mathrm{vec}(\mathbf{N}_\mathrm{B})$. Our objective is to design a channel estimator for $\mathbf{c}$ from low-resolution quantized observations of $\mathbf{y}$ using a hybrid analog and digital architecture at the BS. Specifically, let $\mathbf{B}_\mathbf{c} \in \mathbb{C}^{G_\mathbf{c} \times NT\tau}$ be an analog combining matrix that projects $\mathbf{y}$ onto a lower-dimensional space $\mathbb{C}^{G_\mathbf{c} \times 1}$ for which $G_\mathbf{c} \leq NT\tau$, and let $\mathbf{D}_\mathbf{c} \in \mathbb{C}^{NKL \times G_\mathbf{c}}$ be a digital processing matrix that reconstructs $\mathbf{c}$. Our approach presented in Section IV will employ task-based quantization with the goal of minimizing the following MSE distortion between the true and estimated channel:

$$\min_{\mathbf{B}_\mathbf{c}, \mathbf{D}_\mathbf{c}, \gamma_\mathbf{c}} \mathbb{E}\left[\|\mathbf{c} - \hat{\mathbf{c}}\|_2^2\right], \tag{17}$$

where $\hat{\mathbf{c}}$ is the estimate of $\mathbf{c}$, and $\gamma_\mathbf{c}$ is the support for the ADC.

B. Estimation of individual RIS channels

When an RIS consists entirely of passive elements, it is not possible to extract the individual CSI components of the BS-RIS and RIS-UE channels from the cascaded channels without additional information. This is due to the inherent ambiguity in the cascaded channel, since $\mathbf{G}\,\mathrm{diag}(\mathbf{f}_k) = (\mathbf{G}\boldsymbol{\Lambda})(\boldsymbol{\Lambda}^{-1}\,\mathrm{diag}(\mathbf{f}_k))$ for any invertible diagonal matrix $\boldsymbol{\Lambda}$ [9]. To address this ambiguity, we exploit the availability of a few semi-passive elements in the RIS. Here, our goal is to estimate $\{\mathbf{G},\ \mathbf{F}\}$ from low-resolution quantized observations of $\{\mathbf{Y},\ \mathbf{Z}\}$, employing a task-based approach that minimizes the MSE channel estimation error for both $\mathbf{G}$ and $\mathbf{F}$.

Since the signals forwarded from the RIS to the BS are already quantized and only bear information about the RIS-UE link, jointly estimating both the RIS-UE and BS-RIS channels is a challenging task. For this reasons, we formulate a two-stage channel estimation approach, where in Stage I the RIS-UE channel $\mathbf{F}$ is estimated based on the quantized observations from the semi-passive RIS elements, and in Stage II the BS-RIS channel $\mathbf{G}$ is determined using the quantized observations at the BS and the estimate of $\mathbf{F}$ obtained in Stage I. Since the signals received by the semi-passive elements are independent of the RIS phase shifts, we will assume $\tau = 1$ in this case such that $T_{\mathrm{p}} = T$, and each UE transmits a different pilot symbol in each subblock.

1) Stage I: In this stage, our goal is to design an estimator for $\mathbf{F}$ based on the quantized observations at the semi-passive elements $\boldsymbol{\pi}_{\mathbf{z}}$. Note that, consistent with the low cost and low power design of the RIS, we assume it has no analog combining capability. Consequently, the optimization problem involves designing only a digital processing matrix at the BS $\mathbf{D}_{\mathbf{f}} \in \mathbb{C}^{KL \times LT}$, as follows:

$$\min_{\mathbf{D}_{\mathbf{f}}} \mathbb{E}[\|\mathbf{f} - \hat{\mathbf{f}}\|_2^2], \tag{18}$$

where $\hat{\mathbf{f}}$ is the estimate of $\mathbf{f} = \mathrm{vec}(\mathbf{F})$.

2) Stage II: In this stage, the goal is to estimate $\mathbf{G}$ based on the quantized version of $\mathbf{Y}$ and the estimate $\hat{\mathbf{F}}$ obtained in Stage I. Stacking the received signal vectors at the BS over the T subblocks, the overall measurement matrix in (15) can be equivalently written as

$$\begin{aligned}\mathbf{Y} &= [\mathbf{y}[1], \cdots, \mathbf{y}[T]] \\ &= \mathbf{G}(\mathbf{S} \odot \mathbf{F}\mathbf{X}_{\mathrm{I}}) + \mathbf{N}_{\mathrm{B}},\end{aligned} \tag{19}$$

where $\mathbf{N}_{\mathrm{B}} = [\mathbf{n}_{\mathrm{B}}[1], \cdots, \mathbf{n}_{\mathrm{B}}[T]] \in \mathbb{C}^{N \times T}$. Denoting $\bar{\mathbf{X}}_{\mathrm{I}} = (\mathbf{S} \odot \mathbf{F}\mathbf{X}_{\mathrm{I}})$, the observations at the BS in (19) can be vectorized as

$$\begin{aligned}\mathrm{vec}(\mathbf{Y}) &= \left(\bar{\mathbf{X}}_{\mathrm{I}}^{\mathrm{T}} \otimes \mathbf{I}_N\right) \mathrm{vec}(\mathbf{G}) + \mathrm{vec}(\mathbf{N}_{\mathrm{B}}) \\ &= \mathbf{W}_{\mathbf{y}}\mathbf{g} + \mathbf{n}_{\mathrm{B}} \\ &\triangleq \mathbf{y},\end{aligned} \tag{20}$$

where $\mathbf{W}_{\mathbf{y}} = \left(\bar{\mathbf{X}}_{\mathrm{I}}^{\mathrm{T}} \otimes \mathbf{I}_N\right)$, $\mathbf{g} = \mathrm{vec}(\mathbf{G})$, and $\mathbf{n}_{\mathrm{B}} = \mathrm{vec}(\mathbf{N}_{\mathrm{B}})$. In this case, given $\mathbf{y}$ and $\hat{\mathbf{f}}$, we aim to jointly design an analog combining matrix $\mathbf{B}_{\mathbf{g}} \in \mathbb{C}^{G_{\mathbf{g}} \times NT}$ for which $G_{\mathbf{g}} \leq NT$, a digital processing matrix $\mathbf{D}_{\mathbf{g}} \in \mathbb{C}^{NL \times G_{\mathbf{g}}}$, and the support for the ADC $\gamma_{\mathbf{g}}$ to minimize the MSE distortion between $\mathbf{g}$ and the estimate $\hat{\mathbf{g}}$. The optimization problem is formulated assuming $\mathbf{f}$ is known and is given by

$$\min_{\mathbf{B}_{\mathbf{g}}, \mathbf{D}_{\mathbf{g}}, \gamma_{\mathbf{g}}} \mathbb{E}\left[\|\mathbf{g} - \hat{\mathbf{g}}\|_2^2 \middle| \mathbf{f}\right]. \tag{21}$$

IV. Task-Based Quantization

The goal of task-based quantization [24]–[28] is to jointly design the hybrid analog and digital combining matrices and the quantization to minimize the MSE distortion between the true and estimated channel. We focus on a system operating with identical scalar ADCs, referred to as a hardware-limited task-based quantizer [24], since a corresponding vector quantizer, despite its superior performance, would be computationally infeasible in practice for a large array. In the following, we will first explain the basic idea of the hardware-limited task-based quantizer and establish a theoretical basis for the proposed channel estimation design.

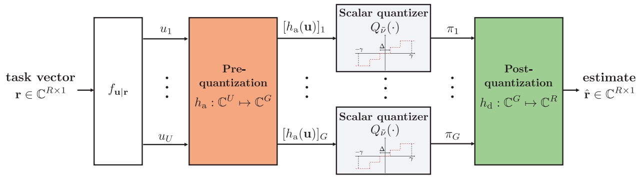
task vector
$\mathbf{r} \in \mathbb{C}^{R\times 1}$
$f_{\mathbf{u}|\mathbf{r}}$
u_1
u_U
Pre-quantization
$h_{\mathrm{a}} : \mathbb{C}^U \mapsto \mathbb{C}^G$
$[h_{\mathrm{a}}(\mathbf{u})]_1$
$[h_{\mathrm{a}}(\mathbf{u})]_G$
Scalar quantizer
$Q_{\tilde{\nu}}(\cdot)$
$-\gamma$
Δ
γ
Scalar quantizer
$Q_{\tilde{\nu}}(\cdot)$
π_1
π_G
Post-quantization
$h_{\mathrm{d}} : \mathbb{C}^G \mapsto \mathbb{C}^R$
estimate
$\hat{\mathbf{r}} \in \mathbb{C}^{R\times 1}$

Fig. 3: An example of a hardware-limited task-based quantization system.

A. Hardware-limited task-based quantizer

Fig. 3 depicts a general hardware-limited task-based quantization system, where $\mathbf{r} \in \mathbb{C}^{R\times 1}$ denotes the task vector we aim to recover, $\mathbf{u} \in \mathbb{C}^{U\times 1}$ is the measurement vector, and the statistical relationship between $\mathbf{r}$ and $\mathbf{u}$ is represented by the conditional probability $f_{\mathbf{u}|\mathbf{r}}$. First, a pre-quantization mapping $h_{\mathrm{a}}(\cdot)$ carried out in the analog domain projects $\mathbf{u}$ onto a lower-dimensional space $\mathbb{C}^{G\times 1}$ for which $G \leq U$. Subsequently, each component of $h_{\mathrm{a}}(\mathbf{u}) \in \mathbb{C}^{G\times 1}$ is quantized element-wise using identical scalar ADCs. Due to their useful statistical properties, the ADCs are modeled as non-subtractive uniform dithered quantizers [38], which facilitates a tractable analysis for the channel estimator design to be presented later.[2]

We define $\tilde{\nu} = \lfloor \nu^{\frac{1}{2G}} \rfloor$, where $\log_2 \nu$ is the total number of quantization bits sampled by the system at any given time instant, and the individual ADC resolution is thus $\log_2 \tilde{\nu}$ bits. The ADC with dithering is modeled by the operation $Q_{\tilde{\nu}}(I_g) = q(\mathrm{Re}(I_g + \beta_g)) + j \cdot q(\mathrm{Im}(I_g + \beta_g))$, $g = 1, \cdots, G$. Note that there are $2G$ ADCs to separately quantize the real and imaginary parts of I_g. Here, $\{\beta_1, \cdots, \beta_G\}$ represent i.i.d. dither signals whose real and imaginary parts are uniformly distributed over $[-\Delta/2, \Delta/2]$ for quantization step size $\Delta = \frac{2\gamma}{\tilde{\nu}}$ and mutually independent inputs, and $q(\cdot)$ is defined as

$$q(I) = \begin{cases} -\gamma + \Delta\left(\ell + \frac{1}{2}\right) & I - \ell \cdot \Delta + \gamma \in [0, \Delta], \\ & \ell \in \{0, \cdots, \tilde{\nu} - 1\}, \\ \mathrm{sign}(I)\left(\gamma - \frac{\Delta}{2}\right) & |I| > \gamma, \end{cases} \tag{22}$$

where γ is the support for the ADC. To guarantee that the inputs of the ADC in (22) fall within the range $[-\gamma, \gamma]$ with high probability, γ is designed as some multiple η of the maximum standard deviation of the inputs. Finally, a post-quantization mapping $h_{\mathrm{d}}(\cdot)$ carried out in the digital domain reconstructs $\mathbf{r}$ based on the G outputs of the identical scalar ADCs. Denoting the g-th output of the scalar ADC as $\pi_g = Q_{\tilde{\nu}}([h_{\mathrm{a}}(\mathbf{u})]_g)$, $\mathbf{r}$ is reconstructed as $\hat{\mathbf{r}} = h_{\mathrm{d}}(\boldsymbol{\pi}) \in \mathbb{C}^{R\times 1}$, where $\boldsymbol{\pi} = [\pi_1, \cdots, \pi_G]^{\mathrm{T}} \in \mathbb{C}^{G\times 1}$.

We focus on the problem of recovering a zero-mean random vector $\mathbf{r}$ from a zero-mean random measurement vector $\mathbf{u}$, where all entries in both $\mathbf{r}$ and $\mathbf{u}$ have finite variances. We use MSE distortion as the task recovery error, aiming to design h_{a},

[2] Assuming a non-overloaded quantizer, the output of a uniform scalar ADC modeled by dithering can be equivalently described as the sum of the input and uncorrelated quantization noise. However, this characteristic can be approximately achieved in systems where dithering is not used for various types of inputs, such as Gaussian signals [39]. The impact of dithering will be investigated in Section VII-A.

h_{d}, and γ to minimize the MSE between $\mathbf{r}$ and its estimate $\hat{\mathbf{r}}$. This leads to the following optimization problem:

$$\begin{aligned}&\min_{h_{\mathrm{a}},h_{\mathrm{d}},\gamma}\ \mathbb{E}\left[\|\mathbf{r}-\hat{\mathbf{r}}\|_2^2\right]\\&\overset{(a)}{=}\mathbb{E}\left[\|\mathbf{r}-\tilde{\mathbf{r}}\|_2^2\right]+\min_{h_{\mathrm{a}},h_{\mathrm{d}},\gamma}\ \mathbb{E}\left[\|\tilde{\mathbf{r}}-\hat{\mathbf{r}}\|_2^2\right], \qquad (23)\end{aligned}$$

where $\tilde{\mathbf{r}}=\mathbb{E}[\mathbf{r}|\mathbf{u}]$ is the MMSE estimate of $\mathbf{r}$ based on $\mathbf{u}$, and (a) follows from the orthogonality principle stating that the MMSE estimate is uncorrelated with the estimation error [40]. The first term in (23) is independent of the quantizer design, indicating the irreducible estimation error of $\mathbf{r}$ solely based on $\mathbf{u}$. This implies that the optimization problem boils down to minimizing the second term, which represents the minimal distortion in quantizing the MMSE estimate of $\mathbf{r}$ [24], [41], and this term would only be zero when the designed quantizer exactly recovers the MMSE estimate. We will assume linear mappings for both $h_{\mathrm{a}}(\cdot)$ and $h_{\mathrm{d}}(\cdot)$, i.e., $h_{\mathrm{a}}(\mathbf{u})=\mathbf{B}\mathbf{u}$ with the analog combining matrix $\mathbf{B}\in\mathbb{C}^{G\times U}$ and $h_{\mathrm{d}}(\boldsymbol{\pi})=\mathbf{D}\boldsymbol{\pi}$ with the digital processing matrix $\mathbf{D}\in\mathbb{C}^{R\times G}$.

B. Task-ignorant quantizer

While task-based quantization systems are designed to minimize the MSE between $\mathbf{r}$ and $\hat{\mathbf{r}}$, the ADCs in task-ignorant systems are designed to simply reconstruct the given observation $\mathbf{u}$ independently of the desired task, and the desired vector $\mathbf{r}$ is recovered from the quantized data by applying the MMSE estimator.

V. Cascaded Channel Estimation

In this section, we develop a cascaded channel estimator based on hardware-limited task-based quantization under finite bit-resolution constraints. Based on (23), the optimization problem formulated in (17) boils down to

$$\min_{\mathbf{B}_{\mathbf{c}},\mathbf{D}_{\mathbf{c}},\gamma_{\mathbf{c}}}\ \mathbb{E}\left[\|\tilde{\mathbf{c}}-\hat{\mathbf{c}}\|_2^2\right], \qquad (24)$$

where $\tilde{\mathbf{c}}=\mathbb{E}[\mathbf{c}|\mathbf{y}]$ is the MMSE estimate of $\mathbf{c}$ given $\mathbf{y}$.

To solve (24), analyzing the MMSE estimate $\tilde{\mathbf{c}}$ is necessary. To facilitate the analysis, we provide the following lemma to reformulate $\mathbf{c}$.

Lemma 1. *The vectorized cascaded channel* $\mathbf{c}$ *can be expressed as*

$$\begin{aligned}\mathbf{c}&=((\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}}\diamond\mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}}\diamond\tilde{\mathbf{A}}_{\mathrm{B,RB}})(\boldsymbol{\alpha}_{\mathrm{UR}}\otimes\boldsymbol{\alpha}_{\mathrm{RB}})\\&\triangleq\mathbf{W}_{\mathbf{c}}\boldsymbol{\alpha}_{\mathbf{c}}, \qquad (25)\end{aligned}$$

where $\mathbf{A}_{\mathrm{R,UR}}=[\mathbf{A}_{\mathrm{R,UR},1},\cdots,\mathbf{A}_{\mathrm{R,UR},K}]$, $\tilde{\mathbf{A}}_{\mathrm{B,RB}}=\mathrm{blkdiag}(\mathbf{1}_{M_{\mathrm{UR},1}}^{\mathrm{T}},\cdots,\mathbf{1}_{M_{\mathrm{UR},K}}^{\mathrm{T}})\otimes\mathbf{A}_{\mathrm{B,RB}}$, $\boldsymbol{\alpha}_{\mathrm{UR}}^{\mathrm{T}}=[\boldsymbol{\alpha}_{\mathrm{UR},1}^{\mathrm{T}},\cdots,\boldsymbol{\alpha}_{\mathrm{UR},K}^{\mathrm{T}}]$, $\mathbf{W}_{\mathbf{c}}=(\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}}\diamond\mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}}\diamond\tilde{\mathbf{A}}_{\mathrm{B,RB}}$, *and* $\boldsymbol{\alpha}_{\mathbf{c}}=\boldsymbol{\alpha}_{\mathrm{UR}}\otimes\boldsymbol{\alpha}_{\mathrm{RB}}$.

Proof. See Appendix A. □

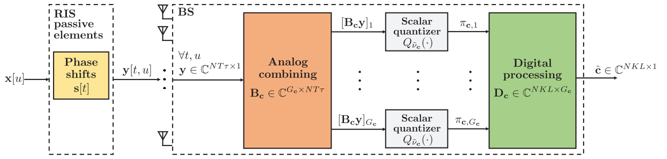
RIS passive elements
Phase shifts
$\mathbf{s}[t]$
$\mathbf{x}[u]$
$\mathbf{y}[t,u]$
BS
$\forall t,u$
$\mathbf{y}\in\mathbb{C}^{NT\tau\times 1}$
Analog combining
$\mathbf{B}_{\mathbf{c}}\in\mathbb{C}^{G_{\mathbf{c}}\times NT\tau}$
$[\mathbf{B}_{\mathbf{c}}\mathbf{y}]_1$
$[\mathbf{B}_{\mathbf{c}}\mathbf{y}]_{G_{\mathbf{c}}}$
Scalar quantizer
$Q_{\tilde{\nu}_{\mathbf{c}}}(\cdot)$
Scalar quantizer
$Q_{\tilde{\nu}_{\mathbf{c}}}(\cdot)$
$\pi_{\mathbf{c},1}$
$\pi_{\mathbf{c},G_{\mathbf{c}}}$
Digital processing
$\mathbf{D}_{\mathbf{c}}\in\mathbb{C}^{NKL\times G_{\mathbf{c}}}$
$\hat{\mathbf{c}}\in\mathbb{C}^{NKL\times 1}$

Fig. 4: Architecture for cascaded channel estimation using hardware-limited task-based quantization.

Note that in general, the AoAs/AoDs in $\mathbf{W}_{\mathbf{c}}$ change much more slowly than the complex channel gains in $\boldsymbol{\alpha}_{\mathbf{c}}$, and it can be assumed that $\mathbf{W}_{\mathbf{c}}$ remains fixed across multiple channel coherence blocks [36], [42]. Based on this, we can assume that $\mathbf{W}_{\mathbf{c}}$ is deterministic, and thus the distribution of $\mathbf{c}$ is based on a linear transformation of the random vector $\boldsymbol{\alpha}_{\mathbf{c}}$ by $\mathbf{W}_{\mathbf{c}}$. Note that each element in $\boldsymbol{\alpha}_{\mathbf{c}}$ is the product of independent complex channel gains that follow a zero-mean Gaussian distribution. In [43], it is shown that if $x_1\sim\mathcal{CN}(0,\sigma_1^2)$ and $x_2\sim\mathcal{CN}(0,\sigma_2^2)$, then the mean and variance of the product $y=x_1x_2$ are zero and $\sigma_1^2\sigma_2^2$, respectively. In our analysis below, we will approximate y as a complex Gaussian random variable with this mean and variance. Based on this result, $\boldsymbol{\alpha}_{\mathbf{c}}$ approximately follows the following complex Gaussian distribution: $\boldsymbol{\alpha}_{\mathbf{c}}\sim\mathcal{CN}(\mathbf{0}_{M_{\mathrm{RB}}M_{\mathrm{UR}}},\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{\mathbf{c}}})$, where $M_{\mathrm{UR}}=\sum_{k=1}^{K}M_{\mathrm{UR},k}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{\mathbf{c}}}=\mathrm{blkdiag}(\bar{\sigma}_{\mathrm{RB}}^2\bar{\sigma}_{\mathrm{UR},1}^2\mathbf{I}_{M_{\mathrm{RB}}M_{\mathrm{UR},1}},\cdots,\bar{\sigma}_{\mathrm{RB}}^2\bar{\sigma}_{\mathrm{UR},K}^2\mathbf{I}_{M_{\mathrm{RB}}M_{\mathrm{UR},K}})$ with $\bar{\sigma}_{\mathrm{RB}}^2=\frac{NL}{M_{\mathrm{RB}}}\sigma_{\mathrm{RB}}^2$ and $\bar{\sigma}_{\mathrm{UR},k}^2=\frac{L}{M_{\mathrm{UR},k}}\sigma_{\mathrm{UR},k}^2$ denoting the variances of the elements in $\boldsymbol{\alpha}_{\mathrm{RB}}$ and $\boldsymbol{\alpha}_{\mathrm{UR},k}$, respectively. This means that $\mathbf{c}\sim\mathcal{CN}(\mathbf{0}_{NKL},\boldsymbol{\Sigma}_{\mathbf{c}})$ approximately holds, where $\boldsymbol{\Sigma}_{\mathbf{c}}=\mathbb{E}[\mathbf{c}\mathbf{c}^{\mathrm{H}}]=\mathbf{W}_{\mathbf{c}}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_{\mathbf{c}}}\mathbf{W}_{\mathbf{c}}^{\mathrm{H}}$. Thus, the relationship between $\tilde{\mathbf{c}}$ and $\mathbf{y}$ is linear, i.e., $\tilde{\mathbf{c}}$ can be represented by $\tilde{\mathbf{c}}=\boldsymbol{\Gamma}_{\mathbf{c}}\mathbf{y}$, where $\boldsymbol{\Gamma}_{\mathbf{c}}$ is given by

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{c}}&=\mathbb{E}\left[\mathbf{c}\mathbf{y}^{\mathrm{H}}\right]\left(\mathbb{E}\left[\mathbf{y}\mathbf{y}^{\mathrm{H}}\right]\right)^{-1}\\&=\boldsymbol{\Sigma}_{\mathbf{c}}\bar{\mathbf{S}}^{\mathrm{H}}\left(\bar{\mathbf{S}}\boldsymbol{\Sigma}_{\mathbf{c}}\bar{\mathbf{S}}^{\mathrm{H}}+\sigma_{\mathrm{B}}^2\mathbf{I}_{NT\tau}\right)^{-1}. \qquad (26)\end{aligned}$$

Based on the above discussions, we develop here a hardware-limited task-based quantization approach for estimating $\mathbf{c}$ to minimize the MSE in (24). Denoting $\log_2\nu_{\mathbf{c}}$ as the total number of quantization bits used to estimate $\mathbf{c}$, the resolution of the ADCs that separately quantize the real and imaginary parts of each element in $\mathbf{B}_{\mathbf{c}}\mathbf{y}$ is $\tilde{\nu}_{\mathbf{c}}=\lfloor\nu_{\mathbf{c}}^{\frac{1}{2G_{\mathbf{c}}}}\rfloor$. First, the digital processing matrix that minimizes the MSE in (24) for a given analog combining matrix is found via the following lemma.

Lemma 2. *For any analog combining matrix* $\mathbf{B}_{\mathbf{c}}$*, the digital processing matrix that minimizes the MSE in (24) is*

$$\mathbf{D}_{\mathbf{c}}^{\mathrm{o}}(\mathbf{B}_{\mathbf{c}})=\boldsymbol{\Gamma}_{\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}_{\mathbf{c}}^{\mathrm{H}}\left(\mathbf{B}_{\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}_{\mathbf{c}}^{\mathrm{H}}+\frac{4\gamma_{\mathbf{c}}^2}{3\tilde{\nu}_{\mathbf{c}}^2}\mathbf{I}_{G_{\mathbf{c}}}\right)^{-1}, \qquad (27)$$

and the resulting minimum MSE distortion is given by

$$\begin{aligned}&\min_{\mathbf{D}_{\mathbf{c}}}\mathbb{E}\left[\|\tilde{\mathbf{c}}-\hat{\mathbf{c}}\|_2^2\right]=\mathrm{tr}\left(\boldsymbol{\Gamma}_{\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{y}}\boldsymbol{\Gamma}_{\mathbf{c}}^{\mathrm{H}}\right)\\&-\mathrm{tr}\left(\boldsymbol{\Gamma}_{\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}_{\mathbf{c}}^{\mathrm{H}}\left(\mathbf{B}_{\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}_{\mathbf{c}}^{\mathrm{H}}+\frac{4\gamma_{\mathbf{c}}^2}{3\tilde{\nu}_{\mathbf{c}}^2}\mathbf{I}_{G_{\mathbf{c}}}\right)^{-1}\mathbf{B}_{\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{y}}\boldsymbol{\Gamma}_{\mathbf{c}}^{\mathrm{H}}\right),\\&\qquad (28)\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{y}}=\mathbb{E}[\mathbf{y}\mathbf{y}^{\mathrm{H}}]$.

Proof. The proof is based on Lemma 1 in [24], and details are provided in Appendix B. □

Using the results of Lemma 2, the following proposition characterizes the analog combining matrix that minimizes the MSE distortion in (28).

Proposition 1. *The optimal analog combining matrix that minimizes the MSE in (28) is given by* $\mathbf{B}_{\mathbf{c}}^{\mathrm{o}} = \mathbf{U}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}\mathbf{V}_{\mathbf{c}}^{\mathrm{H}}\mathbf{\Sigma}_{\mathbf{y}}^{-\frac{1}{2}}$*, where:*

1) $\mathbf{V}_{\mathbf{c}} \in \mathbb{C}^{NT\tau \times NT\tau}$ *is the matrix of right singular vectors of* $\widetilde{\mathbf{\Gamma}}_{\mathbf{c}} \triangleq \mathbf{\Gamma}_{\mathbf{c}}\mathbf{\Sigma}_{\mathbf{y}}^{\frac{1}{2}}$.
2) $\mathbf{\Lambda}_{\mathbf{c}} \in \mathbb{C}^{G_{\mathbf{c}} \times NT\tau}$ *is a diagonal matrix with diagonal entries given by*
$$([\mathbf{\Lambda}_{\mathbf{c}}]_{g,g})^2 = \frac{4\kappa_{\mathbf{c}}}{3\tilde{\nu}_{\mathbf{c}}^2 \cdot G_{\mathbf{c}}}\left(\zeta_{\mathbf{c}} \cdot \lambda_{\widetilde{\mathbf{\Gamma}}_{\mathbf{c}},g} - 1\right)^{+}, \qquad (29)$$
where $\kappa_{\mathbf{c}} = \eta_{\mathbf{c}}^2\left(1 - \frac{2\eta_{\mathbf{c}}^2}{3\tilde{\nu}_{\mathbf{c}}^2}\right)^{-1}$, $\eta_{\mathbf{c}}$ *represents a design parameter to guarantee that the inputs to the ADCs fall within the desired dynamic range,* $\{\lambda_{\widetilde{\mathbf{\Gamma}}_{\mathbf{c}},g}\}$ *are the singular values of* $\widetilde{\mathbf{\Gamma}}_{\mathbf{c}}$ *arranged in descending order, and* $\zeta_{\mathbf{c}}$ *is chosen to satisfy*
$$\frac{4\kappa_{\mathbf{c}}}{3\tilde{\nu}_{\mathbf{c}}^2 \cdot G_{\mathbf{c}}}\sum_{g=1}^{G_{\mathbf{c}}}\left(\zeta_{\mathbf{c}} \cdot \lambda_{\widetilde{\mathbf{\Gamma}}_{\mathbf{c}},g} - 1\right)^{+} = 1. \qquad (30)$$
3) $\mathbf{U}_{\mathbf{c}} \in \mathbb{C}^{G_{\mathbf{c}} \times G_{\mathbf{c}}}$ *is a unitary matrix that forces* $\mathbf{U}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}^{\mathrm{H}}\mathbf{U}_{\mathbf{c}}^{\mathrm{H}}$ *to have identical diagonal entries.*
4) *The resulting support of the ADCs is* $\gamma_{\mathbf{c}} = \sqrt{\frac{\kappa_{\mathbf{c}}}{G_{\mathbf{c}}}}$.

Proof. The proof is based on Theorem 1 in [24], and details are provided in Appendix C. □

The unitary matrix $\mathbf{U}_{\mathbf{c}}$ in Proposition 1 can be computed using Algorithm 2.2 in [44], which guarantees that each ADC input has an identical variance and allows for the same quantization rule to be applied to each channel. By employing $\mathbf{\Lambda}_{\mathbf{c}}$, it is possible to balance the estimation and quantization errors by suppressing the less significant singular values $\{\lambda_{\widetilde{\mathbf{\Gamma}}_{\mathbf{c}},g}\}$, following the water-filling type formula specified in (29).

Note that, as shown in Lemma 2 and Proposition 1, the quantizer in the proposed channel estimator is jointly designed with both the analog combining matrix and the digital processing matrix to minimize the channel estimation error. In contrast, the performance metric for designing a standard task-ignorant quantizer is solely the accuracy of the digital representation with respect to its input. This implies that the proposed task-specific estimator design can achieve superior performance compared to task-ignorant designs, even under the same per-ADC bit resolution constraints.

The overall proposed cascaded channel estimation design is depicted in Fig. 4, where the g-th element of $\mathbf{B}_{\mathbf{c}}\mathbf{y}$ is quantized using the low-resolution ADC discussed in Section IV-A, and the output of the scalar ADC is denoted by $\pi_{\mathbf{c},g} \triangleq Q_{\tilde{\nu}_{\mathbf{c}}}([\mathbf{B}_{\mathbf{c}}\mathbf{y}]_g)$. Subsequently, $\mathbf{c}$ is reconstructed as $\hat{\mathbf{c}} = \mathbf{D}_{\mathbf{c}}\boldsymbol{\pi}_{\mathbf{c}}$, where $\boldsymbol{\pi}_{\mathbf{c}} = [\pi_{\mathbf{c},1}, \cdots, \pi_{\mathbf{c},G_{\mathbf{c}}}]^{\mathrm{T}} \in \mathbb{C}^{G_{\mathbf{c}} \times 1}$. Based on the estimate $\hat{\mathbf{c}}$, the cascaded channel for the k-th UE, $\mathbf{C}_k = \mathbf{G}\,\mathrm{diag}(\mathbf{f}_k)$, is reconstructed as
$$\hat{\mathbf{C}}_k = [\hat{\mathbf{C}}]_{(k-1)N+1:kN,:}\,, \qquad (31)$$
where $\hat{\mathbf{C}} = \mathrm{unvec}(\hat{\mathbf{c}})$. Note that in order to minimize the MSE, the output dimension of the analog combining matrix $G_{\mathbf{c}}$ should not be larger than the rank of the covariance matrix of $\tilde{\mathbf{c}}$, according to Corollary 1 in [24], and the optimal $G_{\mathbf{c}}$ is equivalent to the number of non-zero singular values $\{\lambda_{\widetilde{\mathbf{\Gamma}}_{\mathbf{c}},g}\}$. This implies that, based on (26), the choice of G_{c} heavily depends on the rank of $\mathbf{\Sigma}_{\mathbf{c}}$, which is determined by the rank of $\mathbf{W}_{\mathbf{c}}$ in (25) since $\mathbf{\Sigma}_{\boldsymbol{\alpha}_{\mathbf{c}}}$ is a full rank matrix. For further analysis, we provide the following lemma regarding the rank of $\mathbf{W}_{\mathbf{c}}$.

Lemma 3. *If the following conditions are satisfied,* $\mathbf{W}_{\mathbf{c}}$ *has full column rank* $M_{\mathrm{RB}}M_{\mathrm{UR}}$*:*

1) $L \geq M_{\mathrm{RB}}M_{\mathrm{UR}}$,
2) $\theta_{\mathrm{UR},k,i_k}^{\mathrm{Azi}} \neq \theta_{\mathrm{RB},j}^{\mathrm{Azi}}$, $\theta_{\mathrm{UR},k,i_k}^{\mathrm{Ele}} \neq \theta_{\mathrm{RB},j}^{\mathrm{Ele}}$, $\forall k \in \{1, \cdots K\}$, $\forall i_k \in \{1, \cdots, M_{\mathrm{UR},k}\}$, $\forall j = \{1, \cdots, M_{\mathrm{RB}}\}$.

Proof. See Appendix D. □

In mmWave systems that employ an RIS, the RIS should have a large number of elements to overcome the multiplicative path-loss, and thus since the number of propagation paths for the RIS-related channels is limited, the first condition will generally be satisfied. Furthermore, the AoAs/AoDs of different propagation paths can be treated as independent continuous random variables, from which the second condition for Lemma 3 is satisfied with probability one [35], [45], [46].

Based on the above discussion, in large-scale systems where L is sufficiently large, the rank of the covariance matrix of $\tilde{\mathbf{c}}$ depends on the rank of $\mathbf{W}_{\mathbf{c}}$ and is upper bounded by $M_{\mathrm{RB}}M_{\mathrm{UR}}$ for proper design of the UE pilot sequence $\mathbf{X}_{\mathrm{C}}$ and the passive RIS reflection coefficients $\mathbf{S}$, which defines $\mathbf{\Gamma}_{\mathbf{c}}$ in (26). To investigate the number of training time slots required to achieve $\mathrm{rank}(\mathbf{\Gamma}_{\mathbf{c}}) = M_{\mathrm{RB}}M_{\mathrm{UR}}$, we provide the following lemma.

Lemma 4. *If the conditions in Lemma 3 hold and both* $\mathbf{S}$ *and* $\mathbf{X}_{\mathrm{C}}$ *have their maximum ranks with* $\tau \geq K$*, the necessary conditions for* $\mathrm{rank}(\mathbf{\Gamma}_{\mathbf{c}}) = M_{\mathrm{RB}}M_{\mathrm{UR}}$ *are given by*

1) $NT\tau \geq M_{\mathrm{RB}}M_{\mathrm{UR}}$,
2) $KT \geq M_{\mathrm{UR}}$.

Proof. See Appendix E. □

We can see that the first condition in Lemma 4 is easily satisfied when the number of BS antennas is large. The second condition indicates it is sufficient that T be at least the average number of propagation paths in the UE-RIS links, which is usually small in mmWave systems. Consequently, unlike task-ignorant systems, the analog combining matrix derived from Proposition 1 can reduce the dimensionality of the observation vector $\mathbf{y}$ at the BS from $NT\tau$ to $M_{\mathrm{RB}}M_{\mathrm{UR}}$ with a small training overhead, which significantly reduces quantization error when the total number of quantization bits at the BS is highly limited.

In Proposition 1, the matrix $\mathbf{B}_{\mathbf{c}}$ linearly combines the elements of the vector $\mathbf{y}$ consisting of received signals at the BS over multiple time instances during the channel estimation phase. In [47] it was shown that it is feasible to combine signals received over multiple time instances in the analog domain using the concept of virtual channel extension, in which an analog combining matrix for multiple time instances is constructed by sequentially reusing a simple hardware architecture that generates an analog combining matrix. Furthermore, hardware prototypes for analog combining with dynamic weights have been demonstrated in [48], [49]. In [48], dynamically adjustable complex gains were implemented using highly reconfigurable noise-canceling constant-Gm vector modulators, while in [49] analog vector multipliers were used to control the gain and phase of the analog signals.

Finally, we analyze the computational complexity required to implement the proposed cascaded channel estimator discussed thus far. To simplify the analysis, we assume $T_\mathrm{p} \ll L$ and $G_\mathbf{c} = M_\mathrm{RB} M_\mathrm{UR} \ll NT_\mathrm{p}$. To compute $\mathbf{\Gamma_c}$, the dominant complexity comes from the matrix multiplications involved and is given by $\mathcal{O}(N^3K^2L^2T_\mathrm{p})$. The complexity required to compute the analog combining matrix $\mathbf{B_c}$ based on Proposition 1 is $\mathcal{O}(N^3KLT_\mathrm{p}^2)$. Computing the digital processing matrix $\mathbf{D_c}$ in Lemma 2 requires a complexity of $\mathcal{O}(N^3KLT_\mathrm{p}^2)$. Note that, the remaining operations including matrix-vector multiplications to obtain $\hat{\mathbf{c}}$ based on $\mathbf{B_c}$ and $\mathbf{D_c}$ have negligible computational complexity. Thus, the total complexity of the proposed cascaded channel estimator is $\mathcal{O}(N^3K^2L^2T_\mathrm{p})$.

VI. Individual RIS Channel Estimation

In this section, we design an estimator for the individual RIS-related channels using hardware-limited task-based quantization, where information from the semi-passive elements of the RIS is utilized, and low-resolution ADCs are present at both the BS and the RIS. As discussed in Section III-B, we apply task-based quantization to the two-stage estimation problem, where in Stage I the RIS-UE channel $\mathbf{F}$ is estimated based on the quantized observations from the semi-passive RIS elements, and in Stage II the BS-RIS channel $\mathbf{G}$ is estimated using the quantized observations at the BS and the estimate of $\mathbf{F}$ obtained in Stage I.

A. Stage I: Estimation of $\mathbf{F}$

We first develop a channel estimator for $\mathbf{F}$ based on quantized observations at the semi-passive elements equipped with low-resolution ADCs. To facilitate the analysis, we use a pseudo-measurement model as in [22], where a statistically equivalent expression for $\mathbf{Z}$ in (2) is given by

$$\begin{aligned}\hat{\mathbf{Z}} &= [\hat{\mathbf{z}}[1], \cdots, \hat{\mathbf{z}}[T]] \\ &= \bar{\mathbf{\Omega}} \odot \mathbf{F}\mathbf{X}_\mathrm{I} + \mathbf{N}_\mathrm{R}. \end{aligned} \tag{32}$$

In (32), $\hat{\mathbf{Z}}$ contains the same information about $\mathbf{F}$ as $\mathbf{Z}$ for the semi-passive elements, and no information about $\mathbf{F}$ at the passive element locations. This implies that the quantized values of $\hat{\mathbf{Z}}$ corresponding to the zeros of $\bar{\mathbf{\Omega}}$ can be chosen arbitrarily from among the possible ADC outputs. The vectorized representation of $\hat{\mathbf{Z}}$ is given by

$$\begin{aligned}\mathrm{vec}(\hat{\mathbf{Z}}) &= \mathrm{diag}(\mathrm{vec}(\bar{\mathbf{\Omega}}))(\mathbf{X}_\mathrm{I}^\mathrm{T} \otimes \mathbf{I}_L)\mathrm{vec}(\mathbf{F}) + \mathrm{vec}(\mathbf{N}_\mathrm{R}) \\ &= \mathbf{W}_{\hat{\mathbf{z}}}\mathbf{f} + \mathbf{n}_\mathrm{R} \\ &\triangleq \hat{\mathbf{z}}, \end{aligned} \tag{33}$$

where $\mathbf{W}_{\hat{\mathbf{z}}} = \mathrm{diag}(\mathrm{vec}(\bar{\mathbf{\Omega}}))(\mathbf{X}_\mathrm{I}^\mathrm{T} \otimes \mathbf{I}_L)$, $\mathbf{f} = \mathrm{vec}(\mathbf{F})$, and $\mathbf{n}_\mathrm{R} = \mathrm{vec}(\mathbf{N}_\mathrm{R})$.

Our goal is to design an estimator of $\mathbf{f}$ from $\hat{\mathbf{z}}$ by minimizing the MSE distortion under finite bit-resolution constraints using hardware-limited task-based quantization. While in principle an analog combining matrix such as that in Proposition 1 could be implemented at the RIS to obtain an accurate estimate of $\mathbf{f}$ as discussed in Section III-B, we assume that no analog processing is performed at the RIS to maintain the low cost and power consumption of the RIS. Mathematically this is equivalent to setting the RIS analog combiner as $\mathbf{I}_{LT}$. In this case, the optimization problem boils down to solely designing a digital processing matrix $\mathbf{D_f}$ under finite bit-resolution constraints, leading to the following optimization problem based on the result in (23):

$$\min_{\mathbf{D_f}} \mathbb{E}[\|\tilde{\mathbf{f}} - \hat{\mathbf{f}}\|_2^2], \tag{34}$$

where $\tilde{\mathbf{f}} = \mathbb{E}[\mathbf{f}|\hat{\mathbf{z}}]$ is the MMSE estimate of $\mathbf{f}$ given $\hat{\mathbf{z}}$. Note that the digital processing matrix $\mathbf{D_f}$ found in (34) is applied at the BS after the RIS forwards $\boldsymbol{\pi}_\mathbf{z}$ to the BS.

From (10), with fixed $\mathbf{A}_{\mathrm{R,UR},k}$, $\mathbf{f}_k$ is a linear transformation of the random vector $\boldsymbol{\alpha}_{\mathrm{UR},k}$ by $\mathbf{A}_{\mathrm{R,UR},k}$, and thus is distributed as $\mathbf{f} \sim \mathcal{CN}(\mathbf{0}_{KL}, \mathbf{\Sigma_f})$ with covariance matrix $\mathbf{\Sigma_f} = \mathrm{blkdiag}\ (\bar{\sigma}_{\mathrm{UR},1}^2 \mathbf{A}_{\mathrm{R,UR},1}\mathbf{A}_{\mathrm{R,UR},1}^\mathrm{H}, \cdots, \bar{\sigma}_{\mathrm{UR},K}^2 \mathbf{A}_{\mathrm{R,UR},K}\ \mathbf{A}_{\mathrm{R,UR},K}^\mathrm{H})$. This suggests that the MMSE estimate $\tilde{\mathbf{f}}$ is linear in $\hat{\mathbf{z}}$, i.e., $\tilde{\mathbf{f}} = \mathbf{\Gamma_f}\hat{\mathbf{z}}$, where

$$\begin{aligned}\mathbf{\Gamma_f} &= \mathbb{E}[\mathbf{f}\hat{\mathbf{z}}^\mathrm{H}](\mathbb{E}[\hat{\mathbf{z}}\hat{\mathbf{z}}^\mathrm{H}])^{-1} \\ &= \mathbf{\Sigma_f}\mathbf{W}_{\hat{\mathbf{z}}}^\mathrm{H}(\mathbf{W}_{\hat{\mathbf{z}}}\mathbf{\Sigma_f}\mathbf{W}_{\hat{\mathbf{z}}}^\mathrm{H} + \sigma_\mathrm{R}^2\mathbf{I}_{LT})^{-1}. \end{aligned} \tag{35}$$

Assume that each low-resolution ADC at the RIS has resolution $\tilde{\nu}_\mathbf{f} = \lfloor \nu_\mathbf{f}^{\frac{1}{2L_\mathrm{a}}} \rfloor$, where $\log_2 \nu_\mathbf{f}$ represents the total number of quantization bits available at the RIS, of which each RIS ADC produces $\log_2 \tilde{\nu}_\mathbf{f}$ bits. The digital processing matrix minimizing the MSE distortion in (34) for this case is characterized by the following corollary:

Corollary 1. *Assuming no analog combining at the RIS, the digital processing matrix which minimizes the MSE in (34) is given by*

$$\mathbf{D}_\mathbf{f}^\mathrm{o}(\mathbf{I}_{LT}) = \mathbf{\Gamma_f}\mathbf{\Sigma}_{\hat{\mathbf{z}}}\left(\mathbf{\Sigma}_{\hat{\mathbf{z}}} + \frac{4\kappa_\mathbf{f}\sigma_{\hat{\mathbf{z}},\max}^2}{3\tilde{\nu}_\mathbf{f}^2}\mathbf{I}_{LT}\right)^{-1}, \tag{36}$$

where $\kappa_\mathbf{f} = \eta_\mathbf{f}^2\left(1 - \frac{2\eta_\mathbf{f}^2}{3\tilde{\nu}_\mathbf{f}^2}\right)^{-1}$ *with* $\eta_\mathbf{f}$ *defined similarly to (29),* $\mathbf{\Sigma}_{\hat{\mathbf{z}}} = \mathbb{E}[\hat{\mathbf{z}}\hat{\mathbf{z}}^\mathrm{H}]$*, and* $\sigma_{\hat{\mathbf{z}},\max}^2 = \max_{i=1,\cdots,LT}[\mathbf{\Sigma}_{\hat{\mathbf{z}}}]_{i,i}$.

Proof. The proof follows directly from Lemma 2. □

The overall estimation process for $\mathbf{f}$ is depicted in Fig. 5, where $\hat{\mathbf{f}} = \mathbf{D_f}\boldsymbol{\pi}_\mathbf{z}$ and $\hat{\mathbf{F}} = \mathrm{unvec}(\hat{\mathbf{f}})$. Note that when the AoAs in the UE-RIS channels are distinct and $\mathbf{X}_\mathrm{I}$ has its maximum rank, it can be shown that the necessary condition for $\mathrm{rank}(\mathbf{\Gamma_f}) = M_\mathrm{UR}$ is $L_\mathrm{a}T \geq M_\mathrm{UR}$.

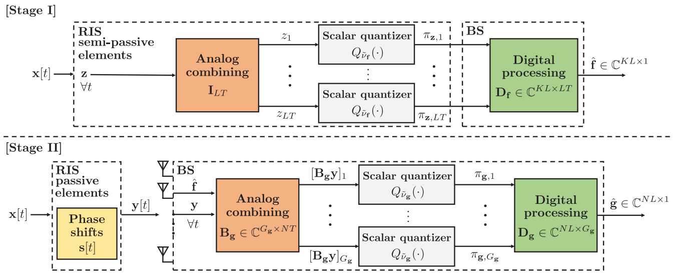
[Stage I]
RIS semi-passive elements
$\mathbf{x}[t]$
$\mathbf{z}$
$\forall t$
Analog combining $\mathbf{I}_{LT}$
z_1
z_{LT}
Scalar quantizer $Q_{\tilde{\nu}_\mathbf{f}}(\cdot)$
Scalar quantizer $Q_{\tilde{\nu}_\mathbf{f}}(\cdot)$
$\pi_{\mathbf{z},1}$
$\pi_{\mathbf{z},LT}$
BS
Digital processing $\mathbf{D}_\mathbf{f} \in \mathbb{C}^{KL\times LT}$
$\hat{\mathbf{f}} \in \mathbb{C}^{KL\times 1}$
[Stage II]
RIS passive elements
$\mathbf{x}[t]$
Phase shifts $\mathbf{s}[t]$
$\mathbf{y}[t]$
BS
$\hat{\mathbf{f}}$
$\mathbf{y}$
$\forall t$
Analog combining $\mathbf{B}_\mathbf{g} \in \mathbb{C}^{G_\mathbf{g}\times NT}$
$[\mathbf{B}_\mathbf{g}\mathbf{y}]_1$
$[\mathbf{B}_\mathbf{g}\mathbf{y}]_{G_\mathbf{g}}$
Scalar quantizer $Q_{\tilde{\nu}_\mathbf{g}}(\cdot)$
Scalar quantizer $Q_{\tilde{\nu}_\mathbf{g}}(\cdot)$
$\pi_{\mathbf{g},1}$
$\pi_{\mathbf{g},G_\mathbf{g}}$
Digital processing $\mathbf{D}_\mathbf{g} \in \mathbb{C}^{NL\times G_\mathbf{g}}$
$\hat{\mathbf{g}} \in \mathbb{C}^{NL\times 1}$

Fig. 5: Architecture for estimating the individual RIS channels using hardware-limited task-based quantization.

B. Stage II: Estimation of $\mathbf{G}$

In Stage II we estimate $\mathbf{G}$ based on the quantized version of $\mathbf{Y}$ and the estimate $\hat{\mathbf{F}}$ obtained in Stage I, where the optimization problem in (21) is reformulated based on the result in (23):

$$\min_{\mathbf{B_g},\mathbf{D_g},\gamma_\mathbf{g}} \mathbb{E}\left[\|\tilde{\mathbf{g}}-\hat{\mathbf{g}}\|_2^2 \middle| \mathbf{f}\right], \tag{37}$$

where $\tilde{\mathbf{g}} = \mathbb{E}[\mathbf{g}|\mathbf{y},\mathbf{f}]$ is the MMSE estimate of $\mathbf{g}$ given $\mathbf{y}$ and $\mathbf{f}$. From (8), $\mathbf{g}$ can be represented as $\mathbf{g} = (\mathbf{A}^*_{\mathrm{R,RB}} \diamond \mathbf{A}_{\mathrm{B,RB}})\boldsymbol{\alpha}_{\mathrm{RB}}$, implying that with fixed $\mathbf{A}_{\mathrm{R,RB}}$ and $\mathbf{A}_{\mathrm{B,RB}}$, $\mathbf{g}$ is distributed as $\mathbf{g} \sim \mathcal{CN}(\mathbf{0}_{NL}, \boldsymbol{\Sigma}_\mathbf{g})$, with covariance $\boldsymbol{\Sigma}_\mathbf{g} = \bar{\sigma}^2_{\mathrm{RB}}(\mathbf{A}^*_{\mathrm{R,RB}} \diamond \mathbf{A}_{\mathrm{B,RB}})(\mathbf{A}^*_{\mathrm{R,RB}} \diamond \mathbf{A}_{\mathrm{B,RB}})^{\mathrm{H}}$. This implies that, given $\mathbf{f}$, $\tilde{\mathbf{g}}$ is linear in $\mathbf{y}$ and can be represented as $\tilde{\mathbf{g}} = \boldsymbol{\Gamma}_{\mathbf{g}|\mathbf{f}}\mathbf{y}$, where

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{g}|\mathbf{f}} &= \mathbb{E}[\mathbf{g}\mathbf{y}^{\mathrm{H}}|\mathbf{f}](\mathbb{E}[\mathbf{y}\mathbf{y}^{\mathrm{H}}|\mathbf{f}])^{-1}\\ &= \boldsymbol{\Sigma}_\mathbf{g}\mathbf{W}^{\mathrm{H}}_\mathbf{y}(\mathbf{W}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{g}\mathbf{W}^{\mathrm{H}}_\mathbf{y} + \sigma^2_{\mathrm{B}}\mathbf{I}_{NT})^{-1}.\end{aligned} \tag{38}$$

Assuming a total of $\log_2 \nu_\mathbf{g}$ quantization bits available at the BS to estimate $\mathbf{g}$, the resolution of the individual ADCs is given by $\tilde{\nu}_\mathbf{g} = \lfloor \nu_\mathbf{g}^{\frac{1}{2G_\mathbf{g}}} \rfloor$ bits. Based on (38), the optimal $\mathbf{B_g}$, $\mathbf{D_g}$, and $\gamma_\mathbf{g}$ that minimize the MSE in (37) can be derived using the results from Proposition 1 and Lemma 2. Following Lemma 3, it is straightforward to verify that $\mathbf{A}^*_{\mathrm{R,RB}} \diamond \mathbf{A}_{\mathrm{B,RB}}$ in $\boldsymbol{\Sigma}_\mathbf{g}$ has full column rank M_{RB} when $N \geq M_{\mathrm{RB}}$, implying that the analog combining matrix $\mathbf{B_g}$ can reduce the dimensionality of the observations at the BS from NT to M_{RB}, given proper designs of $\mathbf{X}_{\mathrm{I}}$ and $\mathbf{S}$. Furthermore, using an approach similar to that in Lemma 4, it can be shown that a necessary condition for $\mathrm{rank}(\boldsymbol{\Gamma}_{\mathbf{g}|\mathbf{f}}) = M_{\mathrm{RB}}$ is $NT \geq M_{\mathrm{RB}}$, and this condition can be satisfied even with $T = 1$ when N is large. This implies that, in large-scale systems, the RIS-BS channel can be accurately estimated with extremely small training overhead, while simultaneously achieving a significant reduction in the total number of quantization bits at the BS compared to task-ignorant estimators.

In large-scale systems, where a large number of BS antennas and RIS elements are deployed, the number of scalar ADCs required at the BS for the proposed algorithm to estimate the individual channels is M_{UR} times less than the number required by the proposed cascaded channel estimator in Section V. This is a significant reduction, illustrating the clear advantage of the proposed individual channel estimator when a few semi-passive elements are available at the RIS.

The architecture for estimating $\mathbf{g}$ is depicted in Fig. 5, where $\boldsymbol{\Gamma}_{\mathbf{g}|\mathbf{f}}$ in (38) is constructed using $\hat{\mathbf{F}}$ obtained from Stage I. The estimate of $\mathbf{g}$ is given by $\hat{\mathbf{g}} = \mathbf{D_g}\boldsymbol{\pi}_\mathbf{g}$, where the g-th element in $\boldsymbol{\pi}_\mathbf{g} = [\pi_{\mathbf{g},1}, \cdots, \pi_{\mathbf{g},G_\mathbf{g}}]^{\mathrm{T}} \in \mathbb{C}^{G_\mathbf{g}\times 1}$ is $\pi_{\mathbf{g},g} \triangleq Q_{\tilde{\nu}_\mathbf{g}}([\mathbf{B_g y}]_g)$, and $\mathbf{G}$ can be reconstructed as $\hat{\mathbf{G}} = \mathrm{unvec}(\hat{\mathbf{g}})$.

The discussion above has assumed that the BS knows the AoAs/AoDs for the RIS-related channels since, as discussed in Section V, these angles change relatively slowly. If the angles are unknown, they can be estimated given data from the semi-passive elements at the RIS using various techniques such as compressed sensing (CS) [15] or the method in [18]. In Section VII-C, we will compare the performance for cases where the angles must be estimated.

To analyze the computational complexity of the proposed individual channel estimator, we make assumptions similar to those used in the analysis of the proposed cascaded channel estimator, namely that $T \ll L$ and $G_\mathbf{g} \ll NT$. In Stage I, the complexity required to compute $\boldsymbol{\Gamma}_\mathbf{f}$ is $\mathcal{O}(KL^3T^2 + K^2L^3T + L^3T^3)$, while computing the digital processing matrix $\mathbf{D_f}$ based on Corollary 1 requires a complexity of $\mathcal{O}(KL^3T^2 + L^3T^3)$, implying that the total complexity of Stage I is $\mathcal{O}(KL^3T^2 + K^2L^3T + L^3T^3)$. In Stage II, computing $\boldsymbol{\Gamma}_{\mathbf{g}|\mathbf{f}}$ requires a complexity of $\mathcal{O}(N^3L^2T)$. Computing the analog combining and digital processing matrices has complexity $\mathcal{O}(N^3LT^2)$ and $\mathcal{O}(N^3LT^2)$, respectively, which leads to a total complexity for Stage II of $O(N^3L^2T)$. Thus, the total computational complexity of the proposed individual channel estimator is $\mathcal{O}(KL^3T^2+K^2L^3T+L^3T^3+N^3L^2T)$.

VII. Numerical Results

In this section, we investigate the performance of the proposed channel estimators based on hardware-limited task-based quantization. We assume a system with uplink carrier frequency $f_c = 24$ GHz, $N = 16$ antennas at the BS, $L = 100$ elements at the RIS with $L_\mathrm{h} = L_\mathrm{v} = 10$, and $K = 3$ UEs transmitting uncorrelated Gaussian signals. The BS and RIS are located at (0 m, 0 m) and (20 m, 10 m), respectively, and the UEs are distributed around a circle centered at (40 m, 0 m) with radius 5 m. The following system parameters are set based on the Dense Urban-eMBB scenario specified in ITU-R M.2412-0 [50]. The UE transmit power is $P_k = 23$ dBm$, \forall k = 1, \cdots, K$. Assuming a noise spectral density $N_0 =$ -174 dBm/Hz, bandwidth $W = 80$ MHz, and noise figure $\mathrm{NF} = 7$ dB, the noise variances at the BS and RIS are $\sigma^2_\mathrm{B} = \sigma^2_\mathrm{R} = W \times N_0 \times \mathrm{NF}$. Considering the line-of-sight dominant environments in mmWave bands, we set the the path-loss model as $\mathrm{PL} = 31.4 + 20\log_{10}(r)$ dB, where r represents the distance of the link in meters [22]. The antenna spacing at the BS and RIS is $d_\mathrm{B} = d_{\mathrm{R,v}} = d_{\mathrm{R,h}} = \frac{\lambda_c}{2}$, where λ_c is the wavelength corresponding to f_c. For cascaded channel estimation, the number of time slots for each subblock is set to $\tau = K$. The output dimensions of the analog combining matrices for cascaded and individual channel estimation are set to $G_\mathbf{c} = M_{\mathrm{RB}}M_{\mathrm{UR}}$ and $G_\mathbf{g} = M_{\mathrm{RB}}$, respectively. Each RIS reflection coefficient in $\mathbf{S}$ is randomly selected from the binary set $\{-1, 1\}$ to account for the hardware limitations of the RIS. Unless otherwise specified, the number of propagation paths

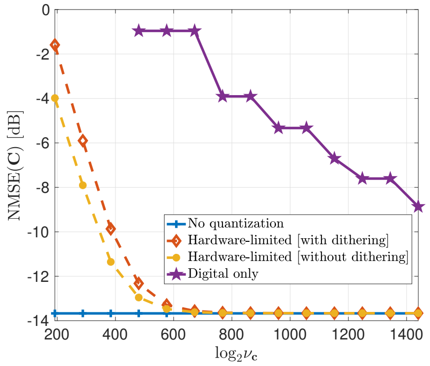

Fig. 6: NMSE performance comparison for cascaded channel estimation versus total number of quantization bits.

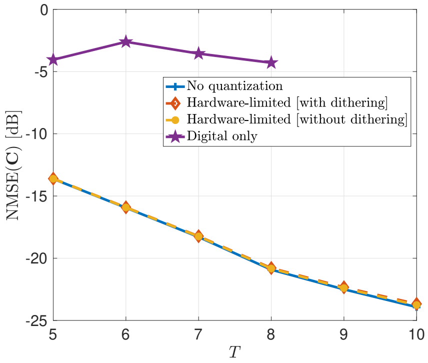

Fig. 7: NMSE comparison for cascaded channel estimation versus number of pilot training subblocks.

in the channel model is $M_{\mathrm{RB}} = M_{\mathrm{UR},1} = \cdots = M_{\mathrm{UR},K} = 4$ considering the limited number of clusters in mmWave bands [51], and the number of subblocks for pilot training is $T = 5$.

A. Cascaded channel estimation performance

In this section, we evaluate the performance of the proposed cascaded channel estimator compared to the following baseline schemes:

- No quantization [32]: The MMSE estimate $\tilde{\mathbf{c}} = \mathbf{\Gamma_c y}$ is applied without quantization. This estimate characterizes the minimum possible MSE distortion.
- Digital only: In this approach, the MMSE estimator $\mathbf{\Gamma_c}$ is applied to the quantized observations without analog combining, i.e., the ADCs operate independently of the channel estimation task. In particular, the uniform quantizer proposed in [52], [53] with the fixed resolution $\lfloor \nu_{\mathbf{c}}^{\frac{1}{2NT\tau}} \rfloor$ is applied separately to the real and imaginary parts of each entry in $\mathbf{y}$.

We adopt the normalized MSE (NMSE) to measure the cascaded channel estimation error, which is defined as

$$\mathrm{NMSE}(\mathbf{C}) = \mathbb{E}\left[\frac{\|\mathbf{C} - \hat{\mathbf{C}}\|_{\mathrm{F}}^2}{\|\mathbf{C}\|_{\mathrm{F}}^2}\right]. \tag{39}$$

Fig. 6 illustrates the NMSE performance versus the total number of ADC quantization bits at the BS, $\log_2 \nu_{\mathbf{c}}$. It is observed that the NMSE difference between the proposed technique and the digital-only approach is quite large, suggesting that incorporating knowledge of the system task in the analog domain can yield significant performance improvement. This gain is mainly due to the design of the analog combining matrix that reduces the dimensionality of the input to the ADCs, allowing for more accurate quantization even with a small number of total quantization bits. In particular, when each ADC uses at least six bits, the quantization error becomes negligible, and the NMSE of the proposed technique effectively approaches that of the MMSE estimate achievable with unlimited resolution ADCs. For small $\log_2 \nu_{\mathbf{c}}$, the proposed technique shows improved performance without dithering. As discussed in Section IV-A, Gaussian signals lead to approximately uncorrelated quantization noise without dithering [39]. Since ADCs with dithering generally produce higher quantization noise than those without dithering, additional distortion is introduced to the inputs, resulting in a slight performance degradation for small $\log_2 \nu_{\mathbf{c}}$, as discussed in [24]. As $\log_2 \nu_{\mathbf{c}}$ increases, this additional quantization noise becomes negligible, leading to nearly the same performance with or without dithering.

Fig. 7 depicts the NMSE performance versus the number of subblocks T assuming $\log_2 \nu_{\mathbf{c}} = 800$. The performance of the digital-only approach does not appreciably improve with T since the ADC resolution per time instant decreases due to the increased dimension of the observation vector $\mathbf{y}$. Conversely, in the proposed technique, the resolution of each ADC is identical regardless of T for a fixed $\log_2 \nu_{\mathbf{c}}$, resulting in comparable performance to the ideal system without quantization when $\log_2 \nu_{\mathbf{c}}$ is sufficiently large.

B. Individual RIS channel estimation performance

Next we evaluate the performance achieved by the proposed two-stage approach for estimating the individual RIS channels. In this case, we will use the following definitions of NMSE to quantify the estimation error for the individual channels:

$$\mathrm{NMSE}(\mathbf{F}) = \mathbb{E}\left[\frac{\|\mathbf{F} - \hat{\mathbf{F}}\|_{\mathrm{F}}^2}{\|\mathbf{F}\|_{\mathrm{F}}^2}\right], \tag{40}$$

$$\mathrm{NMSE}(\mathbf{G}) = \mathbb{E}\left[\frac{\|\mathbf{G} - \hat{\mathbf{G}}\|_{\mathrm{F}}^2}{\|\mathbf{G}\|_{\mathrm{F}}^2}\right]. \tag{41}$$

As in the previous section, approaches without quantization are used as the baselines. Specifically, in Stage I the no-quantization approach utilizes the MMSE estimate $\tilde{\mathbf{f}} = \mathbf{\Gamma_f z}$ without quantization, while in Stage II the MMSE estimate $\tilde{\mathbf{g}} = \mathbf{\Gamma_{g|f} y}$ is employed without quantization based on perfect knowledge of $\mathbf{F}$.

In Fig. 8, we evaluate the NMSE performance for estimation of $\mathbf{F}$ versus the ADC resolution $\log_2 \tilde{\nu}_{\mathbf{f}}$ of the semi-passive elements, whose locations are randomly chosen and fixed per iteration. The overall NMSE performance improves with the number of semi-passive elements L_{a} due to the increased number of observations available for estimation. Since no analog combining occurs at the RIS, the NMSE of the proposed

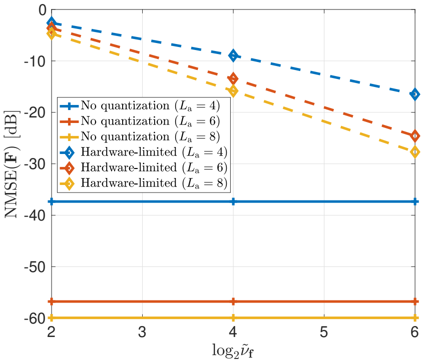

Fig. 8: NMSE comparison of the RIS-UE link channel estimates versus the number of quantization bits at each semi-passive element.

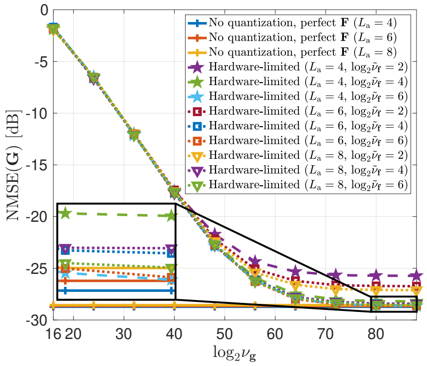

Fig. 9: NMSE comparison of the BS-RIS channel estimate versus number of quantization bits at the BS.

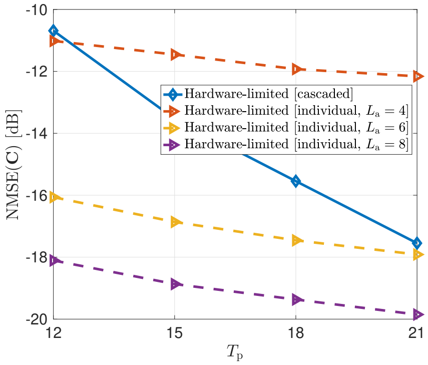

Fig. 10: NMSE comparison for cascaded channel estimation versus number of time slots.

technique is unable in these cases to achieve performance similar to the solution without quantization. Nevertheless, when $L_{\mathrm{a}} = 8$ and $\log_2 \tilde{\nu}_{\mathbf{f}} = 6$, the NMSE of the proposed technique is approximately -30 dB, indicating that the channel is estimated with high accuracy.

In Fig. 9, we investigate the NMSE performance for estimation of $\mathbf{G}$ versus the number of quantization bits $\log_2 \nu_{\mathbf{g}}$ at the BS. When applying the proposed technique, the estimates of $\mathbf{F}$ obtained from Stage I based on different L_{a} and $\log_2 \tilde{\nu}_{\mathbf{f}}$ are employed. We see that the NMSE without quantization increases slightly with L_{a} due to the reduced number of observations obtained through the passive elements. Unlike Stage I for estimating $\mathbf{F}$, the BS employs analog combining, resulting in a reduced NMSE difference between the performance of the proposed technique and that achievable without quantization when $\log_2 \nu_{\mathbf{g}}$ is sufficiently large. Despite its suboptimal performance in estimating $\mathbf{F}$ as seen in Fig. 8, the NMSE of the proposed technique for estimating $\mathbf{G}$ effectively approaches the no-quantization lower bound based on perfect knowledge of $\mathbf{F}$ when at least a 4-bit ADC is employed at each semi-passive RIS element, even for small L_{a}. This demonstrates that quality of the $\mathbf{F}$ estimate obtained by the proposed technique is sufficient to match the performance of systems using unlimited resolution ADCs, even with only a small number of semi-passive elements.

We investigate the NMSE performance of the proposed cascaded and individual channel estimators versus the number of time slots T_{p} in Fig. 10. In this case, the BS requires a total of $\log_2 \nu_{\mathbf{c}} = 576$ quantization bits for the cascaded channel estimator and $\log_2 \nu_{\mathbf{g}} = 48$ bits for the individual channel estimator, while each semi-passive element employs $\log_2 \tilde{\nu}_{\mathbf{f}} = 4$ bits. For individually estimated RIS channels, the cascaded channel estimate is constructed from the individual estimates. For small L_{a}, the proposed cascaded channel estimate outperforms the one formed from the individual estimates since the limited number of observations at the semi-passive elements degrades the estimation performance. However, when at least $L_{\mathrm{a}} = 6$ semi-passive elements are employed, the proposed individual channel estimator achieves performance superior to the cascaded channel estimate, with a significant reduction in the total number of quantization bits at the BS.

C. Performance with estimated angles

The examples thus far have assumed that the BS has perfect knowledge of the AoAs/AoDs for the RIS-related channels since these angles vary relatively slowly compared to the complex channel gains, implying that they can be accurately estimated. Here we consider a scenario where these angles need to be estimated to demonstrate the robustness of the systems based on cascaded and individual channel estimation. We will consider the following baseline approaches:

- LS: The well-known least squares (LS) estimator is applied to quantized observations at the BS to estimate the cascaded channel without the RIS semi-passive elements.
- DS-OMP [11]: This approach employs the double-structured OMP (DS-OMP) algorithm to estimate the cascaded channel leveraging the fact that the BS-RIS link channel is shared across the cascaded channels of all UEs. In this approach, the BS is equipped with infinite resolution ADCs without the RIS semi-passive elements.
- LS-OMP [21]: This approach uses the LS-OMP algorithm for cascaded channel estimation with low-resolution ADCs at the BS without the RIS semi-passive elements, taking into account the potential leakage caused by grid mismatch.

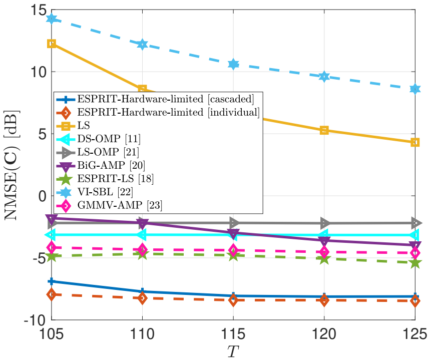

Fig. 11: NMSE comparison of cascaded channel estimates versus number of time slots with estimated angles.

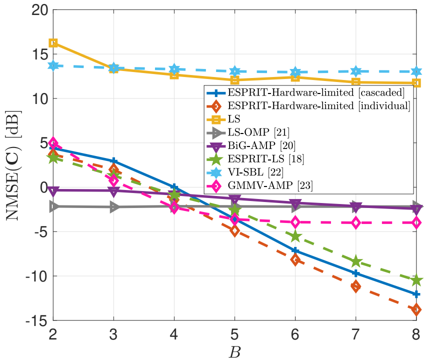

Fig. 12: NMSE comparison for cascaded channel estimates versus per-ADC bit resolution with estimated angles.

- BiG-AMP [20]: This approach uses bilinear matrix completion to estimate the cascaded channel with low-resolution ADCs at the BS using known prior distributions, implemented with a modified BiG-AMP algorithm without the RIS semi-passive elements.
- ESPRIT-LS [18]: This approach uses the signals received by the semi-passive elements to estimate the angles related to the BS-RIS and RIS-UE channels using the estimation of signal parameter via rotational invariance technique (ESPRIT). The complex channel gains are subsequently estimated using LS. We assume that the semi-passive elements are arranged as a $\sqrt{L_{\mathrm{a}}} \times \sqrt{L_{\mathrm{a}}}$ block at the bottom corner of the RIS.
- VI-SBL [22]: This approach employs variational inference (VI) in the sparse Bayesian learning (SBL) framework to approximate the posterior distribution of the channel based on received signals at both the BS and the RIS semi-passive elements.
- GMMV-AMP [23]: This approach solves a CS-based GMMV problem together with matrix completion based on a system with RIS semi-passive elements equipped with low-resolution ADCs. It is implemented via a hierarchical message passing algorithm based on AMP-like approximations.

In the proposed technique, denoted as ESPRIT-hardware-limited, the AoAs/AoDs for all links are estimated using the ESPRIT-based approach, and subsequently the proposed hardware-limited task-based quantization algorithm is applied to reconstruct the channels using these estimates. Note that in VI-SBL and GMMV-AMP, the BS is assumed to have unlimited resolution ADCs, and the locations of the semi-passive elements are randomly chosen at each time instance.

Fig. 11 plots the NMSE of the cascaded channel estimates versus the number of time slots $T_{\mathrm{p}} = T$ assuming a single UE, $L_{\mathrm{a}} = 9$, $M_{\mathrm{RB}} = 1$, $M_{\mathrm{UR}} = 2$, $\log_2 \nu_{\mathbf{c}} = 24$ bits, $\log_2 \nu_{\mathbf{g}} = 12$ bits, and $\log_2 \tilde{\nu}_{\mathbf{f}} = 6$ bits. For the LS, LS-OMP, and BiG-AMP approaches, 6-bit ADCs are employed at the BS. In the proposed channel estimators, the AoAs/AoDs are estimated during $T_{\mathrm{ESPRIT}} = 100$ time slots, and task-based quantization is applied at the BS during the remaining time slots. While the estimates of the number of propagation paths for the RIS-related channels may differ from the ground truth, if the estimates match the ground truth, the number of quantization bits per ADC at the BS is six bits for the proposed channel estimators, the same as for the baseline algorithms. From Fig. 11, we see that the proposed channel estimators achieve the lowest NMSE of all the considered approaches including the CS-based algorithms with a small training overhead and a significant reduction in the total number of quantization bits at the BS, demonstrating the advantages of task-based quantization over conventional approaches. Since the algorithm in [18] uses the results of the eigenvalue decomposition of a sample covariance matrix, quantization effects resulting from using low-resolution ADCs degrade the accuracy of the angle estimates, making the performance of ESPRIT-based approaches, including the proposed channel estimators, relatively stable compared to other algorithms as T increases. However, the performance gap between the proposed channel estimators and ESPRIT-LS clearly shows the advantage of task-based quantization. In particular, the proposed estimators provide more accurate estimates of the complex path gains compared to ESPRIT-LS even for small T. BiG-AMP, VI-SBL, and GMMV-AMP focus on approximating the posterior distributions of the channels rather than specific channel parameters such as the AoAs/AoDs. Thus these approaches will require a larger number of observations to accurately estimate the channels. The performance of DS-OMP and LS-OMP is inferior to some of the baselines since the cascaded channel gains are not Gaussian distributed, and grid mismatch becomes more severe for these algorithms in large-scale systems. Note that, as discussed in Sections V and VI, although the proposed channel estimators have relatively high computational complexity, they require a very small total number of quantization bits, which highlights their clear advantages over task-ignorant approaches.

In Fig. 12, we investigate algorithm performance versus the per-ADC bit resolution B used to estimate the cascaded channel, where $L_{\mathrm{a}} = 9$, $M_{\mathrm{RB}} = 1$, $M_{\mathrm{UR}} = 2$, $T = 105$, and for the proposed channel estimators, $\log_2 \nu_{\mathbf{c}} = 2M_{\mathrm{RB}}M_{\mathrm{UR}}B$ bits for the cascaded channel estimator and $\log_2 \nu_{\mathbf{g}} = 2M_{\mathrm{RB}}B$ bits for the individual channel case. This implies that when the number of propagation paths is accurately estimated, the per-ADC bit resolution is the same for all approaches, except for VI-SBL and GMMV-AMP which use infinite-resolution ADCs at the BS. We can see that when B is at least five bits, the proposed channel estimators again show the lowest NMSE, achieved with significantly fewer total quantization bits at the BS.

VIII. Conclusion

In this paper, we developed channel estimators for RIS-aided mmWave MU-SIMO systems using hardware-limited task-based quantization. In the proposed approaches, an analog combining matrix, a digital processing matrix, and the support for the ADCs are jointly designed to minimize the MSE distortion of the channel estimate. In one approach, the cascaded channel is estimated based only on quantized observations at the BS. In the other, the observations at the BS are augmented by quantized observations at a few semi-passive elements at the RIS, and estimates of the individual RIS-related channels are obtained. Numerical results verify that, with relatively low-resolution ADCs, the proposed channel estimators can closely approach the performance of the MMSE estimate without quantization and outperform a purely digital approach. Furthermore, the proposed estimators are shown to outperform baseline approaches with a low training overhead in scenarios requiring angle estimates at the RIS.

Appendix

A. Proof of Lemma 1

From (15), the cascaded channel is $\mathbf{C} = \mathbf{F}^{\mathrm{T}} \diamond \mathbf{G}$, and $\mathbf{F}$ can be reformulated as $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_K] = \mathbf{A}_{\mathrm{R,UR}}\mathbf{\Lambda}_{\mathrm{UR}}$, where $\mathbf{\Lambda}_{\mathrm{UR}} = \mathrm{blkdiag}(\boldsymbol{\alpha}_{\mathrm{UR},1}, \cdots, \boldsymbol{\alpha}_{\mathrm{UR},K})$. Based on the expressions (8) and (10), $\mathbf{C}$ can be rewritten as

$$\begin{aligned}\mathbf{C} &= (\mathbf{\Lambda}_{\mathrm{UR}}^{\mathrm{T}}\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}}) \diamond (\mathbf{A}_{\mathrm{B,RB}}\,\mathrm{diag}(\boldsymbol{\alpha}_{\mathrm{RB}})\mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})\\ &\overset{(a)}{=}(\mathbf{\Lambda}_{\mathrm{UR}}^{\mathrm{T}} \otimes \mathbf{A}_{\mathrm{B,RB}}\,\mathrm{diag}(\boldsymbol{\alpha}_{\mathrm{RB}}))(\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})\\ &\overset{(b)}{=}(\mathbf{I}_K \otimes \mathbf{A}_{\mathrm{B,RB}})(\mathbf{\Lambda}_{\mathrm{UR}}^{\mathrm{T}} \otimes \mathrm{diag}(\boldsymbol{\alpha}_{\mathrm{RB}}))(\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}}),\end{aligned} \tag{42}$$

where (a) and (b) follow from the identities $\mathbf{M}_1\mathbf{M}_2 \diamond \mathbf{M}_3\mathbf{M}_4 = (\mathbf{M}_1 \otimes \mathbf{M}_3)(\mathbf{M}_2 \diamond \mathbf{M}_4)$ and $\mathbf{M}_1\mathbf{M}_2 \otimes \mathbf{M}_3\mathbf{M}_4 = (\mathbf{M}_1 \otimes \mathbf{M}_3)(\mathbf{M}_2 \otimes \mathbf{M}_4)$, respectively. Using the identity $\mathrm{vec}(\mathbf{M}_1\mathbf{M}_2\mathbf{M}_3) = (\mathbf{M}_3^{\mathrm{T}} \otimes \mathbf{M}_1)\mathrm{vec}(\mathbf{M}_2)$, $\mathbf{C}$ in (42) can be vectorized as

$$\begin{aligned}\mathbf{c} &= ((\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}} \otimes (\mathbf{I}_K \otimes \mathbf{A}_{\mathrm{B,RB}}))\\ &\qquad \cdot \mathrm{vec}(\mathbf{\Lambda}_{\mathrm{UR}}^{\mathrm{T}} \otimes \mathrm{diag}(\boldsymbol{\alpha}_{\mathrm{RB}}))\\ &= ((\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}} \diamond \tilde{\mathbf{A}}_{\mathrm{B,RB}})(\boldsymbol{\alpha}_{\mathrm{UR}} \otimes \boldsymbol{\alpha}_{\mathrm{RB}}),\end{aligned} \tag{43}$$

which completes the proof.

B. Proof of Lemma 2

In the considered system for hardware-limited task-based quantization, the scalar ADCs are modeled as non-subtractive uniform dithered quantizers, where the dithered signals are uniformly distributed over $(-\Delta/2, \Delta/2]$. Assuming that the inputs to the ADCs lie within their dynamic range with probability one, the outputs of the ADCs can be expressed as $\mathbf{B_c y} + \mathbf{e}$, where $\mathbf{e}$ denotes the quantization noise. Based on the results in [38], it is clear that $\mathbf{e}$ is uncorrelated with $\mathbf{B_c y}$, and the real and imaginary parts of each entry of $\mathbf{e}$ are independent and have zero-mean and variance $\frac{\Delta^2}{6}$. Thus, given $\mathbf{B_c}$, the optimal approach for estimating $\hat{\mathbf{c}}$ is the linear MMSE estimator of $\tilde{\mathbf{c}} = \mathbf{\Gamma_c y}$ from $\mathbf{B_c y} + \mathbf{e}$, given by

$$\begin{aligned}\mathbf{D}_{\mathbf{c}}^{\mathrm{o}}(\mathbf{B_c}) &= \mathbb{E}[\tilde{\mathbf{c}}(\mathbf{B_c y} + \mathbf{e})^{\mathrm{H}}] \cdot (\mathbb{E}[(\mathbf{B_c y} + \mathbf{e})(\mathbf{B_c y} + \mathbf{e})^{\mathrm{H}}])^{-1}\\ &= \mathbb{E}[\mathbf{\Gamma_c y}(\mathbf{B_c y})^{\mathrm{H}}] \cdot (\mathbb{E}[(\mathbf{B_c y})(\mathbf{B_c y})^{\mathrm{H}}] + \mathbb{E}[\mathbf{e}\mathbf{e}^{\mathrm{H}}])^{-1}\\ &\overset{(a)}{=} \mathbf{\Gamma_c \Sigma_y B_c^{\mathrm{H}}} \left(\mathbf{B_c \Sigma_y B_c^{\mathrm{H}}} + \frac{4\gamma_{\mathbf{c}}^2}{3\tilde{\nu}_{\mathbf{c}}^2}\mathbf{I}_{G_{\mathbf{c}}}\right)^{-1},\end{aligned} \tag{44}$$

where (a) follows from $\Delta = \frac{2\gamma_{\mathbf{c}}}{\tilde{\nu}_{\mathbf{c}}}$. The resulting MSE in (28) can be computed based on the result in (44), which completes the proof.

C. Proof of Proposition 1

Based on the discussion in Section IV-A, we first define the ADC threshold $\gamma_{\mathbf{c}}$, which is taken to be a multiple $\eta_{\mathbf{c}}$ of the maximum standard deviation of its input, given by

$$\begin{aligned}\gamma_{\mathbf{c}}^2 &= \eta_{\mathbf{c}}^2 \max_{g=1,\cdots,G_{\mathbf{c}}} \mathbb{E}\left[|[\mathbf{B_c y}]_g + \beta_g|^2\right]\\ &= \eta_{\mathbf{c}}^2 \max_{g=1,\cdots,G_{\mathbf{c}}} \mathbb{E}\left[|[\mathbf{B_c y}]_g|^2\right] + \eta_{\mathbf{c}}^2 \frac{2\gamma_{\mathbf{c}}^2}{3\tilde{\nu}_{\mathbf{c}}}\\ &= \kappa_{\mathbf{c}} \max_{g=1,\cdots,G_{\mathbf{c}}} \mathbb{E}\left[|[\mathbf{B_c y}]_g|^2\right],\end{aligned} \tag{45}$$

where β_g is the dither signal, which is independent of $\mathbf{y}$ and has variance $\frac{\Delta^2}{6} = \frac{2\gamma_{\mathbf{c}}^2}{3\tilde{\nu}_{\mathbf{c}}}$ [38], and $\kappa_{\mathbf{c}} = \eta_{\mathbf{c}}^2\left(1 - \frac{2\eta_{\mathbf{c}}^2}{3\tilde{\nu}_{\mathbf{c}}}\right)^{-1}$. Using (45), given an analog combining matrix $\mathbf{B_c}$, the achievable MSE defined in (28) can be rewritten as

$$\begin{aligned}\mathrm{MSE}(\mathbf{B_c}) = \mathrm{tr}\Bigg(&\mathbf{\Gamma_c \Sigma_y \Gamma_c^{\mathrm{H}}} - \mathbf{\Gamma_c \Sigma_y B_c^{\mathrm{H}}}\Bigg(\mathbf{B_c \Sigma_y B_c^{\mathrm{H}}}\\ &+ \frac{4\kappa_{\mathbf{c}}^2}{3\tilde{\nu}_{\mathbf{c}}^2} \max_{g=1,\cdots,G_{\mathbf{c}}} \mathbb{E}\left[|[\mathbf{B_c y}]_g|^2\right]\mathbf{I}_{G_{\mathbf{c}}}\Bigg)^{-1}\mathbf{B_c \Sigma_y \Gamma_c^{\mathrm{H}}}\Bigg).\end{aligned} \tag{46}$$

To simplify (46), we use the result in Lemma C.1 from [24], which states that for any $\mathbf{B_c}$, there exists a unitary matrix $\mathbf{U_c} \in \mathbb{C}^{G_{\mathbf{c}} \times G_{\mathbf{c}}}$ such that $\mathrm{MSE}(\mathbf{B_c}) \geq \mathrm{MSE}(\mathbf{U_c B_c})$. Furthermore, using Corollary 2.4 from [44], we have

$$\min_{\mathbf{U_c}} \max_{g=1,\cdots,G_{\mathbf{c}}} \left[\mathbf{U_c B_c \Sigma_y B_c^{\mathrm{H}} U_c^{\mathrm{H}}}\right]_{g,g} = \frac{1}{G_{\mathbf{c}}}\mathrm{tr}\left(\mathbf{B_c \Sigma_y B_c^{\mathrm{H}}}\right), \tag{47}$$

where the $\mathbf{U_c}$ that achieves this minimum value can be derived using Algorithm 2.2 in [44]. Combining the results from (46) and (47), and defining $\widetilde{\mathbf{B}}_{\mathbf{c}} = \mathbf{B_c}\mathbf{\Sigma}_{\mathbf{y}}^{\frac{1}{2}}$ and $\widetilde{\mathbf{\Gamma}}_{\mathbf{c}} = \mathbf{\Gamma_c}\mathbf{\Sigma}_{\mathbf{y}}^{\frac{1}{2}}$, the MSE minimization problem in (46) simplifies to

$$\begin{aligned}\max_{\widetilde{\mathbf{B}}_{\mathbf{c}}}\ \mathrm{tr}\Bigg(&\widetilde{\mathbf{\Gamma}}_{\mathbf{c}}\widetilde{\mathbf{B}}_{\mathbf{c}}^{\mathrm{H}}\left(\widetilde{\mathbf{B}}_{\mathbf{c}}\widetilde{\mathbf{B}}_{\mathbf{c}}^{\mathrm{H}} + \frac{4\kappa_{\mathbf{c}}}{3\tilde{\nu}_{\mathbf{c}}G_{\mathbf{c}}}\mathrm{tr}(\widetilde{\mathbf{B}}_{\mathbf{c}}\widetilde{\mathbf{B}}_{\mathbf{c}}^{\mathrm{H}})\mathbf{I}_{G_{\mathbf{c}}}\right)^{-1}\\ &\times \widetilde{\mathbf{B}}_{\mathbf{c}}\widetilde{\mathbf{\Gamma}}_{\mathbf{c}}^{\mathrm{H}}\Bigg).\end{aligned} \tag{48}$$

To proceed further, we use the fact that the MSE in (48) remains unchanged when $\widetilde{\mathbf{B}}_{\mathbf{c}}$ is replaced with $\alpha\mathbf{U}\widetilde{\mathbf{B}}_{\mathbf{c}}$, where $\mathbf{U}$ is a unitary matrix and $\alpha > 0$, allowing us to set

$\mathrm{tr}(\widetilde{\mathbf{B}}_{\mathbf{c}}\widetilde{\mathbf{B}}_{\mathbf{c}}^{\mathrm{H}}) = 1$. Letting $\widetilde{\mathbf{B}}_{\mathbf{c}} = \mathbf{\Lambda}_{\mathbf{c}}\mathbf{V}_{\mathbf{c}}^{\mathrm{H}}$ with diagonal matrix $\mathbf{\Lambda}_{\mathbf{c}} \in \mathbb{C}^{G_{\mathbf{c}} \times NT\tau}$ and unitary matrix $\mathbf{V}_{\mathbf{c}} \in \mathbb{C}^{NT\tau \times NT\tau}$, the problem in (48) can be reformulated as

$$\begin{aligned} \max_{\mathbf{\Lambda}_{\mathbf{c}}, \mathbf{V}_{\mathbf{c}}} \quad & \mathrm{tr}\left(\widetilde{\mathbf{\Gamma}}_{\mathbf{c}}^{\mathrm{H}}\widetilde{\mathbf{\Gamma}}_{\mathbf{c}}\mathbf{V}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}^{\mathrm{H}}\left(\mathbf{\Lambda}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}^{\mathrm{H}} + \frac{4\kappa_{\mathbf{c}}}{3\tilde{\nu}_{\mathbf{c}}G_{\mathbf{c}}}\mathbf{I}_{G_{\mathbf{c}}}\right)^{-1}\mathbf{\Lambda}_{\mathbf{c}}\mathbf{V}_{\mathbf{c}}^{\mathrm{H}}\right) \\ \text{s.t.} \quad & \mathrm{tr}\left(\mathbf{\Lambda}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}^{\mathrm{H}}\right) = 1. \end{aligned} \tag{49}$$

Finally, let $\widetilde{\mathbf{\Lambda}}_{\mathbf{c}} = \mathbf{\Lambda}_{\mathbf{c}}^{\mathrm{H}}\left(\mathbf{\Lambda}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}^{\mathrm{H}} + \frac{4\kappa_{\mathbf{c}}}{3\tilde{\nu}_{\mathbf{c}}G_{\mathbf{c}}}\mathbf{I}_{G_{\mathbf{c}}}\right)^{-1}\mathbf{\Lambda}_{\mathbf{c}}$, which is a diagonal matrix. Based on Theorem II.1 in [54], it can be shown that the optimal $\mathbf{V}_{\mathbf{c}}$ for the problem in (49) is the matrix of right singular vectors corresponding to $\widetilde{\mathbf{\Lambda}}_{\mathbf{c}}$ when its entries are arranged in descending order. Based on this result, it is clear that the objective function in (49) is concave with respect to $\{([\mathbf{\Lambda}_{\mathbf{c}}]_{g,g})^2\}_{g=1}^{G_{\mathbf{c}}}$, and the resulting solution satisfying the Karush–Kuhn–Tucker (KKT) conditions leads to the expression given in (29), where we implicitly assume $NT\tau > G_{\mathbf{c}}$. Additional details can be found in [24]. Thus, combining the results discussed so far, the optimal analog combining matrix minimizing the MSE is given by $\mathbf{B}_{\mathbf{c}}^{\mathrm{o}} = \mathbf{U}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}\mathbf{V}_{\mathbf{c}}^{\mathrm{H}}\mathbf{\Sigma}_{\mathbf{y}}^{-\frac{1}{2}}$, and the corresponding ADC threshold is $\gamma_{\mathbf{c}}^2 = \frac{\kappa_{\mathbf{c}}}{G_{\mathbf{c}}}\mathrm{tr}\left(\mathbf{\Lambda}_{\mathbf{c}}\mathbf{\Lambda}_{\mathbf{c}}^{\mathrm{H}}\right) = \frac{\kappa_{\mathbf{c}}}{G_{\mathbf{c}}}$, which completes the proof.

D. Proof of Lemma 3

To analyze the rank of $\mathbf{W}_{\mathbf{c}} = (\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}} \diamond \tilde{\mathbf{A}}_{\mathrm{B,RB}}$, we first investigate the rank of $(\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}}$, which is the cascaded array response matrix at the RIS [9], [55], whose s-th column is given by

$$\begin{aligned} & \left[(\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})_{s,:}\right]^{\mathrm{T}} \\ & = [\mathbf{a}_{\mathrm{R}}^{\mathrm{T}}(\theta_{\mathrm{UR},k,i_k}^{\mathrm{Azi}}, \theta_{\mathrm{UR},k,i_k}^{\mathrm{Ele}}) \diamond \mathbf{a}_{\mathrm{R}}^{\mathrm{H}}(\theta_{\mathrm{RB},j}^{\mathrm{Azi}}, \theta_{\mathrm{RB},j}^{\mathrm{Ele}})]^{\mathrm{T}} \\ & = \mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{UR},k,i_k}^{\mathrm{Azi}} - \theta_{\mathrm{RB},j}^{\mathrm{Azi}}, \theta_{\mathrm{UR},k,i_k}^{\mathrm{Ele}} - \theta_{\mathrm{RB},j}^{\mathrm{Ele}}), \end{aligned} \tag{50}$$

where $i_k = i - \sum_{r=1}^{k-1} M_{\mathrm{UR},r}$ with $i = \lceil \frac{s}{M_{\mathrm{RB}}} \rceil$ when i is within the range $\sum_{r=1}^{k-1} M_{\mathrm{UR},r} < i \leq \sum_{r=1}^{k} M_{\mathrm{UR},r}$, and $j = ((s-1) \bmod M_{\mathrm{RB}})+1$. From (50), when the second condition in Lemma 3 is satisfied, all columns in $(\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}}$ are linearly independent, implying that $(\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}}$ has full column rank $M_{\mathrm{RB}}M_{\mathrm{UR}}$ when $L \geq M_{\mathrm{RB}}M_{\mathrm{UR}}$. To proceed further, we introduce the following lemma regarding the rank of a Khatri-Rao product.

Lemma 5. *If* $\mathrm{rank}((\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}}) + \mathrm{rank}(\tilde{\mathbf{A}}_{\mathrm{B,RB}}) \geq M_{\mathrm{RB}}M_{\mathrm{UR}} + 1$, $\mathbf{W}_{\mathbf{c}}$ *has full column rank* $M_{\mathrm{RB}}M_{\mathrm{UR}}$.

Proof. The proof is based on Lemma 1 in [56]. □

From Lemma 5, since $\mathrm{rank}(\tilde{\mathbf{A}}_{\mathrm{B,RB}}) \geq 1$ always holds as long as it is not an all-zero matrix, $\mathbf{W}_{\mathbf{c}}$ has full column rank $M_{\mathrm{RB}}M_{\mathrm{UR}}$ when $(\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}}$ is a full column rank matrix, which completes the proof.

E. Proof of Lemma 4

Under the considered conditions, the rank of $\mathbf{\Gamma}_{\mathbf{c}}$ in (26) depends on the rank of $\mathbb{E}[\mathbf{c}\mathbf{y}^{\mathrm{H}}]$, which is upper bounded by $\min(M_{\mathrm{RB}}M_{\mathrm{UR}}, \mathrm{rank}(\bar{\mathbf{S}}\mathbf{W}_{\mathbf{c}}))$. Defining $\mathbf{A}_{\mathbf{c}} = (\mathbf{A}_{\mathrm{R,UR}}^{\mathrm{T}} \diamond \mathbf{A}_{\mathrm{R,RB}}^{\mathrm{H}})^{\mathrm{T}}$, $\bar{\mathbf{S}}\mathbf{W}_{\mathbf{c}}$ can be reformulated as

$$\begin{aligned} \bar{\mathbf{S}}\mathbf{W}_{\mathbf{c}} &= (\mathbf{S}^{\mathrm{T}} \otimes \mathbf{X}_{\mathrm{C}}^{\mathrm{T}} \otimes \mathbf{I}_N)(\mathbf{A}_{\mathbf{c}} \diamond \tilde{\mathbf{A}}_{\mathrm{B,RB}}) \\ &= (\mathbf{S}^{\mathrm{T}}\mathbf{A}_{\mathbf{c}}) \diamond ((\mathbf{X}_{\mathrm{C}}^{\mathrm{T}} \otimes \mathbf{I}_N)\tilde{\mathbf{A}}_{\mathrm{B,RB}}). \end{aligned} \tag{51}$$

Let $\mathbf{R} = (\mathbf{X}_{\mathrm{C}}^{\mathrm{T}} \otimes \mathbf{I}_N)\tilde{\mathbf{A}}_{\mathrm{B,RB}}$. Now, to derive the tightest conditions for T, we consider the case where $T \leq M_{\mathrm{RB}}M_{\mathrm{UR}}$ and $KM_{\mathrm{RB}} \leq N\tau$, leading to $\mathrm{rank}(\mathbf{S}\mathbf{A}_{\mathbf{c}}) = T$ and $\mathrm{rank}(\mathbf{R}) = KM_{\mathrm{RB}}$. The rank of $\bar{\mathbf{S}}\mathbf{W}_{\mathbf{c}}$ is then upper bounded by

$$\begin{aligned} & \mathrm{rank}(\bar{\mathbf{S}}\mathbf{W}_{\mathbf{c}}) \\ & = \mathrm{rank}(((\mathbf{S}^{\mathrm{T}}\mathbf{A}_{\mathbf{c}}) \diamond \mathbf{R})^{\mathrm{H}}((\mathbf{S}^{\mathrm{T}}\mathbf{A}_{\mathbf{c}}) \diamond \mathbf{R})) \\ & \overset{(a)}{=} \mathrm{rank}((\mathbf{S}^{\mathrm{T}}\mathbf{A}_{\mathbf{c}})^{\mathrm{H}}(\mathbf{S}^{\mathrm{T}}\mathbf{A}_{\mathbf{c}})) \odot (\mathbf{R}^{\mathrm{H}}\mathbf{R})) \\ & \overset{(b)}{\leq} KTM_{\mathrm{RB}}, \end{aligned} \tag{52}$$

where (a) holds due to $(\mathbf{M}_1 \diamond \mathbf{M}_2)^{\mathrm{H}}(\mathbf{M}_1 \diamond \mathbf{M}_2) = (\mathbf{M}_1^{\mathrm{H}}\mathbf{M}_1) \odot (\mathbf{M}_2^{\mathrm{H}}\mathbf{M}_2)$, and (b) follows from $\mathrm{rank}(\mathbf{M}_1 \odot \mathbf{M}_2) \leq \mathrm{rank}(\mathbf{M}_1) \cdot \mathrm{rank}(\mathbf{M}_2)$. From (52), it is clear that $\mathrm{rank}(\bar{\mathbf{S}}\mathbf{W}_{\mathbf{c}}) \leq \min(KTM_{\mathrm{RB}},\ NT\tau, M_{\mathrm{RB}}M_{\mathrm{UR}})$. Thus, the necessary conditions for $\mathbf{\Gamma}_{\mathbf{c}}$ to have its maximum rank are: 1) $NT\tau \geq M_{\mathrm{RB}}M_{\mathrm{UR}}$, and 2) $KT \geq M_{\mathrm{UR}}$, which completes the proof.

References

[1] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, "Millimeter Wave Communication: A Comprehensive Survey," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 3, pp. 1616–1653, Aug. 2018.

[2] Q. Wu and R. Zhang, "Towards Smart and Reconfigurable Environment: Intelligent Reflecting Surface Aided Wireless Network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[3] M. Di Renzo *et al.*, "Smart Radio Environments Empowered by Reconfigurable AI Meta-Surfaces: An Idea Whose Time Has Come," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–20, May 2019.

[4] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless Communications Through Reconfigurable Intelligent Surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.

[5] M. Di Renzo, F. H. Danufane, and S. Tretyakov, "Communication Models for Reconfigurable Intelligent Surfaces: From Surface Electromagnetics to Wireless Networks Optimization," *Proc. IEEE*, vol. 110, no. 9, pp. 1164–1209, Sep. 2022.

[6] Q. Li, M. El-Hajjar, C. Xu, J. An, C. Yuen, and L. Hanzo, "Stacked Intelligent Metasurfaces for Holographic MIMO-Aided Cell-Free Networks," *IEEE Trans. Commun.*, vol. 72, no. 11, pp. 7139–7151, Nov. 2024.

[7] Q. Li, M. El-Hajjar, Y. Sun, I. Hemadeh, A. Shojaeifard, and L. Hanzo, "Energy-Efficient Reconfigurable Holographic Surfaces Operating in the Presence of Realistic Hardware Impairments," *IEEE Trans. Commun.*, vol. 72, no. 8, pp. 5226–5238, Aug. 2024.

[8] A. L. Swindlehurst, G. Zhou, R. Liu, C. Pan, and M. Li, "Channel Estimation With Reconfigurable Intelligent Surfaces-A General Framework," *Proc. IEEE*, vol. 110, no. 9, pp. 1312–1338, Sep. 2022.

[9] C. Pan, G. Zhou, K. Zhi, S. Hong, T. Wu, Y. Pan, H. Ren, M. D. Renzo, A. L. Swindlehurst, R. Zhang, and A. Y. Zhang, "An Overview of Signal Processing Techniques for RIS/IRS-Aided Wireless Systems," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 5, pp. 883–917, Aug. 2022.

[10] S. Kim, H. Lee, J. Cha, S.-J. Kim, J. Park, and J. Choi, "Practical Channel Estimation and Phase Shift Design for Intelligent Reflecting Surface Empowered MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6226–6241, Aug. 2022.

[11] X. Wei, D. Shen, and L. Dai, "Channel Estimation for RIS Assisted Wireless Communications—Part II: An Improved Solution Based on Double-Structured Sparsity," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1403–1407, May 2021.

[12] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted Sum-Rate Maximization for Reconfigurable Intelligent Surface Aided Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064–3076, May 2020.

[13] S. Zhang and R. Zhang, "Capacity Characterization for Intelligent Reflecting Surface Aided MIMO Communication," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1823–1838, Jun. 2020.

[14] M. Hua, Q. Wu, C. He, S. Ma, and W. Chen, "Joint Active and Passive Beamforming Design for IRS-Aided Radar-Communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2278–2294, Apr. 2023.

[15] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling Large Intelligent Surfaces With Compressive Sensing and Deep Learning," *IEEE Access*, vol. 9, pp. 44 304–44 321, Mar. 2021.

[16] S. Liu, Z. Gao, J. Zhang, M. D. Renzo, and M.-S. Alouini, "Deep Denoising Neural Network Assisted Compressive Channel Estimation for mmWave Intelligent Reflecting Surfaces," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 9223–9228, Aug. 2020.

[17] Y. Jin, J. Zhang, X. Zhang, H. Xiao, B. Ai, and D. W. K. Ng, "Channel Estimation for Semi-Passive Reconfigurable Intelligent Surfaces With Enhanced Deep Residual Networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11 083–11 088, Oct. 2021.

[18] X. Hu, R. Zhang, and C. Zhong, "Semi-Passive Elements Assisted Channel Estimation for Intelligent Reflecting Surface-Aided Communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1132–1142, Feb. 2022.

[19] G. Lee, H. Lee, J. Oh, J. Chung, and J. Choi, "Channel Estimation for Reconfigurable Intelligent Surface With a Few Active Elements," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 8170–8174, Jun. 2023.

[20] R. Wang, H. Ren, C. Pan, J. Fang, M. Dong, and O. A. Dobre, "Channel Estimation for RIS-Aided mmWave Massive MIMO System Using Few-Bit ADCs," *IEEE Commun. Lett.*, vol. 27, no. 3, pp. 961–965, Mar. 2023.

[21] S. Han, Y. Liao, S. Chen, and Y.-C. Liang, "Joint Channel Estimation for RIS-Aided mmWave MIMO Wireless Communication Systems With Mixed-Resolution Quantization Schemes," *IEEE Internet Things J.*, vol. 12, no. 16, pp. 33 756–33 768, Aug. 2025.

[22] I.-S. Kim, M. Bennis, J. Oh, J. Chung, and J. Choi, "Bayesian Channel Estimation for Intelligent Reflecting Surface-Aided mmWave Massive MIMO Systems With Semi-Passive Elements," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9732–9745, Dec. 2023.

[23] Y. Cao, C. Xing, Y. Wu, J. An, D. W. K. Ng, and X.-G. Xia, "RIS-Assisted Massive Access With Semi-Passive Elements," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 10 546–10 561, Sep. 2024.

[24] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues, "Hardware-Limited Task-Based Quantization," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5223–5238, Oct. 2019.

[25] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues, "Asymptotic Task-Based Quantization With Application to Massive MIMO," *IEEE Trans. Signal Process.*, vol. 67, no. 15, pp. 3995–4012, Aug. 2019.

[26] N. Shlezinger and Y. C. Eldar, "Task-Based Quantization With Application to MIMO Receivers," *Commun. Inf. Syst.*, vol. 20, pp. 131–162, 2020.

[27] S. Salamatian, N. Shlezinger, Y. C. Eldar, and M. Médard, "Task-Based Quantization for Recovering Quadratic Functions Using Principal Inertia Components," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 390–394.

[28] N. I. Bernardo, J. Zhu, Y. C. Eldar, and J. Evans, "Design and Analysis of Hardware-Limited Non-Uniform Task-Based Quantizers," *IEEE Trans. Signal Process.*, vol. 71, pp. 1551–1562, May 2023.

[29] F. Xi, N. Shlezinger, and Y. C. Eldar, "BiLiMO: Bit-Limited MIMO Radar via Task-Based Quantization," *IEEE Trans. Signal Process.*, vol. 69, pp. 6267–6282, Nov. 2021.

[30] D. Ma, N. Shlezinger, T. Huang, Y. Liu, and Y. C. Eldar, "Bit Constrained Communication Receivers In Joint Radar Communications Systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2021, pp. 8243–8247.

[31] G. Lee, I.-s. Kim, Y. C. Eldar, A. L. Swindlehurst, and J. Choi, "Channel Estimation for RIS-Aided Communication Systems: A Task-Based Quantization Approach," in *Proc. 19th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Jul. 2024.

[32] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[33] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent Reflecting Surface-Aided Wireless Communications: A Tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.

[34] B. Guo, C. Sun, and M. Tao, "Two-Way Passive Beamforming Design for RIS-Aided FDD Communication Systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2021.

[35] P. Wang, J. Fang, L. Dai, and H. Li, "Joint Transceiver and Large Intelligent Surface Design for Massive MIMO mmWave Systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1052–1064, Feb. 2021.

[36] G. Zhou, C. Pan, H. Ren, P. Popovski, and A. L. Swindlehurst, "Channel Estimation for RIS-Aided Multiuser Millimeter-Wave Systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 1478–1492, Mar. 2022.

[37] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[38] R. Gray and T. Stockham, "Dithered Quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.

[39] B. Widrow, I. Kollar, and M.-C. Liu, "Statistical Theory of Quantization," *IEEE Trans. Instrum. Meas.*, vol. 45, no. 2, pp. 353–361, Apr. 1996.

[40] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

[41] J. Wolf and J. Ziv, "Transmission of Noisy Information to a Noisy Receiver with Minimum Distortion," *IEEE Trans. Inf. Theory*, vol. 16, no. 4, pp. 406–411, Jul. 1970.

[42] S. Kim, J. Wu, and B. Shim, "Efficient Channel Probing and Phase Shift Control for mmWave Reconfigurable Intelligent Surface-Aided Communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 231–246, Jan. 2024.

[43] R. Ware and F. Lad, "Approximating the Distribution for Sums of Products of Normal Variables," *Univ. Canterbury, U.K., Tech. Rep., UCDMS*, 2003.

[44] D. P. Palomar and Y. Jiang, *MIMO Transceiver Design via Majorization Theory*. Delft, The Netherlands: Now Publ., 2007.

[45] S. H. Hong, J. Park, S.-J. Kim, and J. Choi, "Hybrid Beamforming for Intelligent Reflecting Surface Aided Millimeter Wave MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7343–7357, Sep. 2022.

[46] G. Casella and R. Berger, *Statistical Inference*. Boston, MA, USA: Cengage Learning, 2001.

[47] T. Gong, N. Shlezinger, S. S. Ioushua, M. Namer, Z. Yang, and Y. C. Eldar, "RF Chain Reduction for MIMO Systems: A Hardware Prototype," *IEEE Syst. J.*, vol. 14, no. 4, pp. 5296–5307, Dec. 2020.

[48] T. Zirtiloglu, P. Crary, E. Tasci, A. Riaz, Y. C. Eldar, N. Shlezinger, and R. T. Yazicigil, "Task-Specific Low-Power Beamforming MIMO Receiver Using 2-Bit Analog-to-Digital Converters," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2023.

[49] X. Zhang, H. Zhang, N. Glazer, O. Cohen, E. Reznitskiy, S. Savariego, M. Namer, and Y. C. Eldar, "Hardware Implementation of Task-Based Quantization in Multiuser Signal Recovery," *IEEE Trans. Ind. Electron.*, vol. 71, no. 7, pp. 7716–7724, Jul. 2024.

[50] *Guidelines for Evaluation of Radio Interface Technologies for IMT-2020*. Document ITU-R M.2412-0, 2017.

[51] Q. C. Li, G. Wu, and T. S. Rappaport, "Channel Model for Millimeter-Wave Communications Based on Geometry Statistics," in *Proc. IEEE Globecom Workshops*, Dec. 2014, pp. 427–432.

[52] J. Mo, P. Schniter, and R. W. Heath, "Channel Estimation in Broadband Millimeter Wave MIMO Systems With Few-Bit ADCs," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1141–1154, Mar. 2018.

[53] J. Max, "Quantizing for Minimum Distortion," *IRE Trans. Inf. Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.

[54] J. Lasserre, "A Trace Inequality for Matrix Product," *IEEE Trans. Autom. Control*, vol. 40, no. 8, pp. 1500–1501, Aug. 1995.

[55] J. Chen, Y.-C. Liang, H. V. Cheng, and W. Yu, "Channel Estimation for Reconfigurable Intelligent Surface Aided Multi-User mmWave MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 10, pp. 6853–6869, Oct. 2023.

[56] N. Sidiropoulos, R. Bro, and G. Giannakis, "Parallel Factor Analysis in Sensor Array Processing," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2377–2388, Aug. 2000.

Gyoseung Lee (Graduate Student Member, IEEE) received the B.S. degree in Electrical Engineering from Korea University in 2021 and the M.S. degree from the School of Electrical Engineering, KAIST, South Korea, in 2023, where he is currently pursuing the Ph.D. degree. He was a Visiting Scholar at the University of California, Irvine, in 2023, and a Visiting Student Research Collaborator at Princeton University in 2025. He was a recipient of the NRF Korea Research Subsidies of Ph.D. Candidates from 2023 to 2024, and the 2024 KAIST Graduate Student Outstanding Paper Award. His research interests include the design and analysis of massive multiple-input multiple-output (MIMO) communications, reconfigurable intelligent surface (RIS)-aided communication systems, and integrated sensing and communications (ISAC).

In-soo Kim (Member, IEEE) received the B.S. degree in electrical engineering from Hanyang University, Seoul, South Korea, in 2017, the M.S. degree in electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2019, and the Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2022. He is currently with Wireless Research and Development (WRD), Qualcomm Technologies Inc., San Diego, CA, USA, as a Senior Systems Engineer. His research interests include transceiver architectures for low-power wireless communication systems with low-resolution ADCs.

Yonina C. Eldar (Fellow, IEEE) is the Aoun Chair Professor of Electrical and Computer Engineering at Northeastern University and the Dorothy and Patrick Gorman Professorial Chair of Mathematics and Computer Science at the Weizmann Institute where she founded and heads the Signal Acquisition Modeling Processing and Learning Lab (SAMPL) and the Center for Biomedical Engineering. She is also a Visiting Professor at MIT and Princeton, a Visiting Scientist at the Broad Institute, and an Adjunct Professor at Duke University and was a Visiting Professor at Stanford. She is a member of the Israel Academy of Sciences and Humanities and of the Academia Europaea, an IEEE Fellow and a EURASIP Fellow. She received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering from Tel-Aviv University, and the Ph.D. degree in electrical engineering and computer science from MIT. She has received many awards for excellence in research and teaching, including the Israel Prize (2025), Landau Prize (2024), IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014) and the IEEE Kiyo Tomiyasu Award (2016). She received the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), and the Award for Women with Distinguished Contributions. She was selected as one of the 50 most influential women in Israel, and was a member of the Israel Committee for Higher Education. She is the Editor in Chief of Foundations and Trends in Signal Processing, a member of several IEEE Technical Committees and Award Committees, and heads the Committee for Promoting Gender Fairness in Higher Education Institutions in Israel.

A. Lee Swindlehurst (Fellow, IEEE) received the B.S. (1985) and M.S. (1986) degrees in Electrical Engineering from Brigham Young University (BYU), and the PhD (1991) degree in Electrical Engineering from Stanford University. He was with the Department of Electrical and Computer Engineering at BYU from 1990-2007, where he served as Department Chair from 2003-06. During 1996-97, he held a joint appointment as a visiting scholar at Uppsala University and the Royal Institute of Technology in Sweden. From 2006-07, he was on leave working as Vice President of Research for ArrayComm LLC in San Jose, California. Since 2007 he has been with the Electrical Engineering and Computer Science (EECS) Department at the University of California Irvine, where he is a Distinguished Professor and currently serving as Department Chair. Dr. Swindlehurst is a Fellow of the IEEE, during 2014-17 he was also a Hans Fischer Senior Fellow in the Institute for Advanced Studies at the Technical University of Munich, and in 2016, he was elected as a Foreign Member of the Royal Swedish Academy of Engineering Sciences (IVA). He received the 2000 IEEE W. R. G. Baker Prize Paper Award, the 2006 IEEE Communications Society Stephen O. Rice Prize in the Field of Communication Theory, the 2006, 2010 and 2021 IEEE Signal Processing Society's Best Paper Awards, the 2017 IEEE Signal Processing Society Donald G. Fink Overview Paper Award, a Best Paper award at the 2020 and 2024 IEEE International Conferences on Communications, the 2024 IEEE Communications Society SPCC Best Paper Award, the 2022 Claude Shannon-Harry Nyquist Technical Achievement Award from the IEEE Signal Processing Society, and the 2024 Fred W. Ellersick Prize from the IEEE Communications Society. His research focuses on array signal processing for radar, wireless communications, and biomedical applications.

Hyeongtaek Lee (Member, IEEE) received the B.S. (Hons.) degree in electrical engineering from POSTECH in 2018 and the Ph.D. degree in the School of Electrical Engineering from KAIST in 2023.

He is currently working as an assistant professor with the Department of Electronic and Electrical Engineering, Ewha Womans University. From 2023 to 2025, he served as a post-doctoral researcher at KAIST. His research interests include the practical design and analysis of massive/mmWave multiple-input multiple-output (MIMO) communications, reconfigurable intelligent surface (RIS)-aided communication systems, artificial intelligence (AI)-based communications, and integrated sensing and communication (ISAC) systems.

Minje Kim (Graduate Student Member, IEEE) received the B.S. (Hons.) degree in Electrical Engineering from Pohang University of Science and Technology (POSTECH) in 2020. He is currently pursuing the integrated M.S. and Ph.D degree in the School of Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST). His research interests include the design and analysis of massive MIMO communications and satellite communications. He was a co-recipient of the IITP (Information and Communication Technology Planning and Evaluation Institute) Director's Award at the ICT Challenge 2021, the Excellence Award at the Electronic Newspaper ICT Paper Contest Grand Exhibition in 2023. He was awarded the KEPCO (Korea Electric Power Corporation) Electrical Engineering Scholarship in 2021.

Junil Choi (Senior Member, IEEE) received the B.S. (Hons.) and M.S. degrees in electrical engineering from Seoul National University in 2005 and 2007, respectively, and the Ph.D. degree in electrical and computer engineering from Purdue University in 2015.

He is currently working as a KAIST Endowed Chair Associate Professor with the School of Electrical Engineering, KAIST. From 2007 to 2011, he was a member of technical staff at the Samsung Advanced Institute of Technology (SAIT) and Samsung Electronics Company Ltd., South Korea, where he contributed to advanced codebook and feedback framework designs for the 3GPP LTE/LTE-Advanced and IEEE 802.16m standards. Before joining KAIST, he was a post-doctoral fellow at The University of Texas at Austin from 2015 to 2016 and an assistant professor at POSTECH from 2016 to 2019. His research interests include the design and analysis of massive MIMO, mmWave communications, satellite communications, visible light communications, and communication systems using machine-learning techniques.

Dr. Choi was a co-recipient of the 2022 IEEE Vehicular Technology Society Best Vehicular Electronics Paper Award, the 2021 IEEE Vehicular Technology Society Neal Shepherd Memorial Best Propagation Award, the 2019 IEEE Communications Society Stephen O. Rice Prize, the 2015 IEEE Signal Processing Society Best Paper Award, and the 2013 Global Communications Conference (GLOBECOM) Signal Processing for Communications Symposium Best Paper Award. He was awarded the IEEE ComSoc AP Region Outstanding Young Researcher Award in 2017, the NSF Korea and Elsevier Young Researcher Award in 2018, the KICS Haedong Young Researcher Award in 2019, the IEEE Communications Society Communication Theory Technical Committee Early Achievement Award in 2021, and the 6th Next-Generation Scientist Award the S-OIL Science and Culture Foundation in 2024. He is an Area Editor of IEEE Open Journal of the Communications Society and an Associate Editors of IEEE Transactions on Wireless Communications and IEEE Transactions on Communications.