DOI: 10.1109/LWC.2024.3425150

Efficient Data Transmission With Compressed Channel Estimation in RIS-Assisted mmWave MIMO Systems

Wuqiong Zhao, Yifang Dai, You You, *Member, IEEE*, Li Zhang, *Senior Member, IEEE*, Xiaohu You, *Fellow, IEEE*, and Chuan Zhang, *Senior Member, IEEE*

***Abstract*—To improve wireless connectivity, reconfigurable intelligent surface (RIS) offers an energy-efficient solution in millimeter wave (mmWave) multi-input multi-output (MIMO) systems. However, the achievable spectrum efficiency (SE) has been limited by challenges associated with channel estimation (CE) and hybrid beamforming design. To address these issues, we propose an efficient data transmission (DT) scheme with two-stage compressed sensing (CS)-based CE by exploiting the sparse angular domain channel structure and pruning insignificant components. Simulation results demonstrate that the proposed method achieves reduction in pilot overhead and complexity of CE, and higher SE of DT via iterative optimization.**

***Index Terms*—Reconfigurable intelligent surface (RIS), channel estimation (CE), beamforming design, millimeter wave (mmWave).**

I. Introduction

In WIRELESS communications, a reconfigurable intelligent surface (RIS) serves as an efficient bridge reflecting signals [1], [2]. Thanks to the large number of reconfigurable elements, RIS can significantly enhance spectrum efficiency (SE) through reflection and beamforming designs in millimeter wave (mmWave) multiple-input multiple-output (MIMO) systems [3], [4]. As considered by this letter, passive RIS is both energy and cost-effective. However, due to the large number of RIS elements, the required pilot overhead during channel estimation (CE) as well as the computational complexity have multiplied. Therefore, it is crucial to develop an efficient CE method for channel state information (CSI) acquisition.

In RIS-assisted mmWave MIMO systems, the angular domain sparsity of the channel enables simplified CE by employing compressed sensing (CS), which significantly reduces the pilot overhead [5]. However, high computational complexity and limited performance issues still exist. To mitigate this, [2] proposes a formulation incorporating beam and reflection pattern designs. Moreover, the joint design of CS-based CE scheme and beam patterns shows promise for enabling demanding technologies including holographic communications [6]. Additionally, deep learning methods like deep denoising neural networks can further enhance the performance of CS-based CE [7]. However, few existing studies have jointly optimized CS-based CE and data transmission (DT), despite the importance of this joint optimization. Firstly, designing a beamforming scheme during DT does not require complete CSI involving RIS, so the complexity of CE can be cut down on. Secondly, existing hybrid beamforming and RIS reflection design methods are based on perfect knowledge of the two separate channels rather than the cascaded channel, which is difficult to obtain with fully passive RISs.

This letter presents a joint design of CE and DT for RIS-assisted mmWave MIMO systems to efficiently obtain a beamforming design to maximize SE with cascaded CE. The contributions are summarized as follows:

1) Exploiting angular domain sparsity, we propose an efficient two-stage CS-based CE scheme for DT by pruning insignificant CSI components.
2) Hybrid beamforming and RIS reflection patterns are iteratively optimized for DT with high SE based on partial knowledge of the cascaded channel.
3) Simulation results demonstrate the proposed method can achieve high SE in DT with reduced complexity and pilot overhead.

Notations: $\jmath \triangleq \sqrt{-1}$ is the imaginary unit, and $\mathrm{Arg}(\cdot)$ denotes the argument of a complex number. Boldface lower-case $\mathbf{a}$ and upper-case $\mathbf{A}$ letters stand for a vector and a matrix, respectively. $[\mathbf{a}]_i$ is the i-th element of vector $\mathbf{a}$. $(\cdot)^{\mathsf{T}}$, $(\cdot)^{\mathsf{H}}$, $(\cdot)^*$ denote the transpose, the conjugate transpose, the conjugate. $|\mathbf{A}|$ is the determinant of square matrix $\mathbf{A}$. $\mathrm{diag}(\mathbf{a})$ denotes the diagonal matrix with vector $\mathbf{a}$ on its diagonal. $\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}^{-1}_{M,N}(\mathbf{a})$ denote the vectorization of matrix $\mathbf{A}$ and the reshaping of vector $\mathbf{a}$ into an $M \times N$ matrix. $\mathbf{I}_M$ is the identity matrix of size $M \times M$ and $\mathbf{O}_{M\times N}$ is an $M \times N$ zero matrix.

II. System Model

In a RIS-assisted mmWave MIMO system, the uplink cascaded channel transmitting from the user to the base station (BS) reflected on RIS is defined as

$$\mathbf{H} \triangleq \mathbf{H}_r \mathrm{diag}(\mathbf{\Psi}) \mathbf{H}_t, \tag{1}$$

where $\mathbf{H}_t \in \mathbb{C}^{M\times N_t}$ and $\mathbf{H}_r \in \mathbb{C}^{N_r\times M}$ are the user–RIS and RIS–BS channel in a narrowband geometry channel model, respectively, which can be formulated according to [2], [5], [9]. $\mathbf{\Psi} \in \mathbb{C}^{M\times 1}$ is the reflection vector of RIS equipped with $M_x \times M_y = M$ uniform planar array (UPA) reflection elements. By considering the angular domain sparsity of mmWave channels due to the sparse scattering effect [10], the cascaded channel $\mathbf{H}$ in (1) can be written as [2]:

$$\mathbf{H} = \mathbf{V}_r \mathbf{\Gamma}_r \mathbf{U}^{\mathsf{H}} \mathrm{diag}(\mathbf{\Psi}) \mathbf{U} \mathbf{\Gamma}_t \mathbf{V}_t^{\mathsf{H}}, \tag{2}$$

This work was supported in part by National Key R&D Program of China under Grant 2020YFB2205503, in part by NSFC under Grants 62331009, 62122020, and 62001108, in part by the Jiangsu Provincial NSF under Grant BK20211512, in part by the Major Key Project of PCL under Grant PCL2021A01-2, and in part by the Fundamental Research Funds for the Central Universities. *(Wuqiong Zhao and Yifang Dai contributed equally to this letter.) (Corresponding author: Chuan Zhang.)*

Wuqiong Zhao, Yifang Dai, You You, Xiaohu You, and Chuan Zhang are with the LEADS, the National Mobile Communications Research Laboratory, and the Frontiers Science Center for Mobile Information Communication and Security, Southeast University, Nanjing 211189, China; and also with Purple Mountain Laboratories, Nanjing 211100, China. (email: chzhang@seu.edu.cn)

Li Zhang is with the School of Electronic and Electrical Engineering, University of Leeds, LS2 9JT Leeds, U.K. (email: l.x.zhang@leeds.ac.uk)

where the beamspace dictionaries for the user, BS, and RIS $\mathbf{V}_t \in \mathbb{C}^{N_t \times N_t^G}$, $\mathbf{V}_r \in \mathbb{C}^{N_r \times N_r^G}$, and $\mathbf{U} \in \mathbb{C}^{M \times M^G}$ are comprised of N_t^G, N_r^G, $M^G = M_x^G \times M_y^G$ steering vectors of uniformly distributed grids as in [2]. $\mathbf{\Gamma}_r$ and $\mathbf{\Gamma}_t$ are the sparse beamspace channels representing physical angles, with L_1 and L_2 non-zero elements corresponding to the number of paths which are typically small in mmWave [10]. To reduce the complexity and pilot overhead, CE is divided into estimations of K consecutive reflection patterns (i.e., K differently designed reflection vectors) [2]. The received signal with the k-th ($1 \le k \le K$) reflection pattern during CE by omitting the power coefficient can be expressed as

$$\mathbf{y}_k = \left((\widetilde{\mathbf{V}}\mathbf{\Psi}_k)^{\mathsf{T}} \otimes \left((\mathbf{F}_k^{\mathsf{T}} \otimes \mathbf{W}_k^{\mathsf{H}})(\mathbf{V}_t^* \otimes \mathbf{V}_r)\right)\right)\mathbf{j} + \mathbf{n}_k, \quad (3)$$

where $\otimes$ denotes the Kronecker product, $\mathbf{j} \in \mathbb{C}^{N_t^G N_r^G M^G \times 1}$ is a sparse vector that can be estimated by CS methods [5], [9], $\mathbf{n}_k$ is additive white Gaussian noise (AWGN) with variance σ^2, and $\mathbf{F}_k$ and $\mathbf{W}_k$ are the beam pattern matrices in [2]. $\widetilde{\mathbf{V}}$ is the first M^G rows of Khatri-Rao product $\mathbf{U}^{\mathsf{T}} \circledcirc \mathbf{U}^{\mathsf{H}}$, whose $((p-1)M_y^G + q, (r-1)M_y + s)$ element is

$$\widetilde{\mathbf{V}}(\{p,q\},\{r,s\}) = \frac{1}{M}\exp\left(-2\pi\jmath\left(\frac{(r-1)(p-1)}{M_x^G} + \frac{(s-1)(q-1)}{M_y^G}\right)\right). \quad (4)$$

Thus, according to [2], (3) is transformed to

$$\mathbf{y}_k = \mathbf{Q}_k\mathbf{x}_k + \mathbf{n}_k, \quad (5)$$

where $\mathbf{Q}_k \triangleq (\mathbf{F}_k^{\mathsf{T}} \otimes \mathbf{W}_k^{\mathsf{H}})(\mathbf{V}_t^* \otimes \mathbf{V}_r)$ is the sensing matrix and the sparse vector $\mathbf{x}_k$ satisfies the following expression:

$$[\mathbf{x}_k]_j = \sum\nolimits_{m=1}^{M^G} [(\widetilde{\mathbf{V}}\mathbf{\Psi}_k)^{\mathsf{T}}]_m [\mathbf{j}]_{N_t^G N_r^G \cdot (m-1)+j}, \quad (6)$$

for $j = 1, 2, \ldots, N_t^G N_r^G$. To further illustrate the transformation from (3) to (5), we define $\mathbf{J}$ as

$$\mathbf{J} \triangleq \mathrm{vec}^{-1}_{N_t N_r, M^G}(\mathbf{j}), \quad (7)$$

whose i-th column $\mathbf{j}_i$ is mutually orthogonal with each other. By controlling the distribution of zero and non-zero elements in $\widetilde{\mathbf{V}}\mathbf{\Psi}_k$, we can determine the way in which $\mathbf{j}_1, \mathbf{j}_2, \ldots, \mathbf{j}_M$ are superimposed into $\mathbf{x}_k$. When only one element of $\widetilde{\mathbf{V}}\mathbf{\Psi}_k$ is non-zero as in [2], (6) can be further formulated as

$$\mathbf{x}_k = \sum\nolimits_{i=1}^{L} [(\widetilde{\mathbf{V}}\mathbf{\Psi}_k)^{\mathsf{T}}]_{s_i} \mathbf{j}_{s_i}, \quad (8)$$

where s_i ($i = 1, 2, \ldots, L$) is the position (index) of a non-zero element in $\widetilde{\mathbf{V}}\mathbf{\Psi}_k$. Thus, $\mathbf{x}_k$ is a linear superposition of $\mathbf{j}_{s_i}$.

During DT, we utilize CSI obtained from CE to design a reflection pattern $\mathbf{\Psi}$ that can differ from ones in the CE phase to maximize SE. By utilizing $\widetilde{L} \le L \triangleq L_1 \times L_2$ estimated paths, the channel $\mathbf{H}$ for DT can be approximated as

$$\mathbf{H} = \mathrm{vec}^{-1}_{N_r, N_t}((\mathbf{V}_t^* \otimes \mathbf{V}_r)\mathbf{X}\boldsymbol{\beta}), \quad (9)$$

where[1] $\mathbf{X} \triangleq [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\widetilde{L}}]$ and coefficients $\boldsymbol{\beta} \triangleq [\beta_1, \beta_2, \ldots, \beta_{\widetilde{L}}]^{\mathsf{T}}$ can be determined with the specified reflection vector $\mathbf{\Psi}$:

$$\widetilde{\mathbf{V}}\mathbf{\Psi} = \sum\nolimits_{l=1}^{\widetilde{L}} \beta_l \widetilde{\mathbf{V}}\mathbf{\Psi}_l, \quad (10)$$

[1]Without loss of generality, we denote the first $\widetilde{L}$ estimations corresponding to $\widetilde{L}$ estimated paths for notational simplicity.

where $\mathbf{\Psi}_l$ is the reflection pattern corresponding to the l-th path, which will be elaborated in Section III-B. Then (9) can be further expressed as

$$\mathbf{H} = \sum\nolimits_{l=1}^{\widetilde{L}} \beta_l \mathbf{H}_l, \quad (11)$$

where $\mathbf{H}_l \triangleq \mathrm{vec}^{-1}_{N_r, N_t}((\mathbf{V}_t^* \otimes \mathbf{V}_r)\mathbf{x}_l)$. From (10) we can obtain $\beta_l = \widetilde{\mathbf{V}}([\tilde{\mathbf{i}}]_l, :)\mathbf{\Psi}$, where $\tilde{\mathbf{i}}$ is the position prior vector, which will be elaborated in Section III-A. With $\mathbf{\Psi} \triangleq [\psi_1, \psi_2, \ldots, \psi_M]^{\mathsf{T}}$, we can rewrite (11) as

$$\mathbf{H} = \sum\nolimits_{m=1}^{M} \psi_m \mathbf{N}_m, \quad (12)$$

where $\mathbf{N}_m \triangleq \sum_{l=1}^{\widetilde{L}} \widetilde{\mathbf{V}}([\tilde{\mathbf{i}}]_l, m)\mathbf{H}_l$. When $N_r^{\mathrm{RF}} = N_t^{\mathrm{RF}}$ and $N_r^B = N_t^B$, SE R can be approximated as

$$R \approx \log_2 \left| \frac{1}{N_t^B} \mathbf{R}^{-1} \mathbf{W}^{\mathsf{H}} \mathbf{H} \mathbf{F} \mathbf{F}^{\mathsf{H}} \mathbf{H}^{\mathsf{H}} \mathbf{W} \right|, \quad (13)$$

where $\mathbf{R} \triangleq \sigma^2 \mathbf{W}^{\mathsf{H}}\mathbf{W}$. To reduce the hardware implementation complexity, we adopt hybrid beamforming [3], [8] as $\mathbf{W} \triangleq \mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}$, $\mathbf{F} \triangleq \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$, where $\mathbf{W}_{\mathrm{RF}} \in \mathbb{C}^{N_r \times N_r^{\mathrm{RF}}}$ and $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_t \times N_t^{\mathrm{RF}}}$ represent the analog beamforming, $\mathbf{W}_{\mathrm{BB}} \in \mathbb{C}^{N_r^{\mathrm{RF}} \times N_r^B}$ and $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_t^{\mathrm{RF}} \times N_t^B}$ represent the digital beamforming, and N_t^{RF} and N_r^{RF} are the number of radio frequency (RF) chains at the user and BS, respectively.

III. Proposed Method

A. Channel Prior Acquisition

To reduce overall pilot overhead and computational complexity, the first stage of CE is utilized to acquire prior information for the second stage of CE. Though the formulation in [2] already simplifies the CS problem by designing M reflection patterns, only a few of them contribute to the real CSI due to the sparsity. Therefore, it is possible to filter out insignificant reflection patterns by obtaining the essential prior including rough path directions. Thus, around $L \ll M$ reflection patterns are required for accurate channel estimation.

The question is how to efficiently gain the big picture of the RIS-assisted transmission system with only a limited number of pilot overhead. Thus, the chosen reflection vectors should contribute to each element of $\mathbf{j}$ evenly and distinguish them as much as possible, posing constraint to $\widetilde{\mathbf{V}}\mathbf{\Phi}_n$ for $1 \le n \le N$, where N is the number of reflection patterns used in the first stage of CE. $\mathbf{\Phi}_n$ denotes the n-th reflection vector in the first-stage CE, which is different from the reflection vector in the second-stage CE ($\mathbf{\Psi}_l$) in notation. To meet this constraint, we employ the truncated discrete Fourier transform (DFT) pattern, characterized by a unit amplitude and evenly distributed phases, denoted as $\widetilde{\mathbf{V}}\mathbf{\Phi}_n = \mathbf{T}(:, n)$, where $\mathbf{T}$ is an $M \times M$ DFT matrix. Clearly, $\widetilde{\mathbf{V}}\mathbf{\Phi}_n$ satisfies

$$\mathrm{Arg}\left(\frac{[\widetilde{\mathbf{V}}\mathbf{\Phi}_2]_i}{[\widetilde{\mathbf{V}}\mathbf{\Phi}_1]_i}\right) = \frac{2\pi}{M^G}(i-1), \quad (14)$$

for $i = 1, 2, \ldots, M^G$. The maximum phase error (14) allows is $\frac{\pi}{M^G}$, which can result in significant pilot overhead when the number of RIS elements is large. To further optimize the design, dimension reduction can be applied by separately

considering the RIS's planar array in the x and y dimensions. With $\widetilde{\mathbf{V}}\mathbf{\Phi}_n$ for the RIS having $M = M_x \times M_y$ elements and considering (4), we can adopt the following dimension-reduced design

$$\begin{cases} \mathbf{\Phi}_{3n-2} = \mathbf{I}_M(:, nM_y - 1), \\ \mathbf{\Phi}_{3n-1} = \mathbf{I}_M(:, nM_y), \\ \mathbf{\Phi}_{3n} = \mathbf{I}_M(:, nM_y + 1), \end{cases} \tag{15}$$

which forms $(M_x - 1)$ L-shape groups across the RIS UPA. Importantly, this design uses the minimum of 3 patterns (one L-shape group), since it requires at least 3 to obtain an estimation on 2 dimensions. And the multiple-group scenario reduces errors by averaging the results of repeated single-group processes. By using this dimension-reduced design which achieves optimal resolution, the upper limit of allowable error is increased to $\min\{\frac{\pi}{M_x^G}, \frac{\pi}{M_y^G}\}$ enabled by each L-shaped design. This can be expressed as

$$\begin{cases} \operatorname{Arg}\left(\dfrac{[\widetilde{\mathbf{V}}\mathbf{\Phi}_2]_i}{[\widetilde{\mathbf{V}}\mathbf{\Phi}_1]_i}\right) - \operatorname{Arg}\left(\dfrac{[\widetilde{\mathbf{V}}\mathbf{\Phi}_2]_{i-1}}{[\widetilde{\mathbf{V}}\mathbf{\Phi}_1]_{i-1}}\right) = \dfrac{2\pi}{M_x^G}, \\ \operatorname{Arg}\left(\dfrac{[\widetilde{\mathbf{V}}\mathbf{\Phi}_3]_i}{[\widetilde{\mathbf{V}}\mathbf{\Phi}_2]_i}\right) - \operatorname{Arg}\left(\dfrac{[\widetilde{\mathbf{V}}\mathbf{\Phi}_3]_{i-M_x^G}}{[\widetilde{\mathbf{V}}\mathbf{\Phi}_2]_{i-M_x^G}}\right) = \dfrac{2\pi}{M_y^G}. \end{cases} \tag{16}$$

Therefore, x, y two-dimensional information $[\mathbf{i}_x]_l$, $[\mathbf{i}_y]_l$ can be calculated as

$$\begin{cases} [\mathbf{i}_x]_l = \underset{1\le i\le M_x^G}{\arg\min} \left| \operatorname{Arg}\left(\dfrac{[\mathbf{x}_2]_{s_l}}{[\mathbf{x}_1]_{s_l}}\right) - \dfrac{2\pi}{M_y^G} i \right|, \\ [\mathbf{i}_y]_l = \underset{1\le j\le M_y^G}{\arg\min} \left| \operatorname{Arg}\left(\dfrac{[\mathbf{x}_3]_{s_l}}{[\mathbf{x}_2]_{s_l}}\right) - \dfrac{2\pi}{M_x^G} j \right|, \end{cases} \tag{17}$$

where s_l is the position of non-zero elements in $\mathbf{x}_n$, and $\mathbf{x}_n$ is estimated using (5) and (6). Therefore, the i-th element of the position prior vector is given as

$$[\mathbf{i}]_l = [\mathbf{i}_x]_l M_y^G + [\mathbf{i}_y]_l, \tag{18}$$

where $[\mathbf{i}]_l$ directly corresponds to the required estimation of the reflection patterns in the second stage as

$$\mathbf{\Psi}_l = \mathbf{\Upsilon}(:, [\mathbf{i}]_l), \tag{19}$$

where $\mathbf{\Upsilon} = \frac{M^G}{M}\widetilde{\mathbf{V}}^\dagger$ and $\widetilde{\mathbf{V}}^\dagger$ is the pseudo-inverse of $\widetilde{\mathbf{V}}$.

The above-mentioned method can also be extended to a fully-on scheme for RIS, i.e., the reflection vector $\mathbf{\Phi}_n$ is a phase shift vector and each element has unit amplitude. Similar to (15), reflection vectors can be designed as

$$\begin{cases} [\mathbf{\Phi}_1]_{iM_y+j} = \exp\left(2\pi\jmath\left(\dfrac{3}{2M_x} i + \dfrac{3}{2M_y} j\right)\right), \\ [\mathbf{\Phi}_2]_{M-iM_y-j} = \exp\left(2\pi\jmath\left(\dfrac{3}{2M_x} i - \dfrac{3}{2M_y} j\right)\right), \\ [\mathbf{\Phi}_3]_{iM_y+j} = \exp\left(2\pi\jmath\left(\dfrac{3}{2M_x} i - \dfrac{3}{2M_y} j\right)\right), \end{cases} \tag{20}$$

where $1 \le i \le M_x^G$ and $1 \le j \le M_y^G$. Similar to (17),

$$\begin{cases} [\mathbf{i}_x]_l = \underset{1\le i\le M_x^G}{\arg\min} \left| \left|\dfrac{[\mathbf{x}_2]_{s_l}}{[\mathbf{x}_1]_{s_l}}\right| - \dfrac{[\widetilde{\mathbf{V}}\mathbf{\Phi}_2]_{iM_y^G}}{[\widetilde{\mathbf{V}}\mathbf{\Phi}_1]_{iM_y^G}} \right|, \\ [\mathbf{i}_y]_l = \underset{1\le j\le M_y^G}{\arg\min} \left| \left|\dfrac{[\mathbf{x}_3]_{s_l}}{[\mathbf{x}_1]_{s_l}}\right| - \dfrac{[\widetilde{\mathbf{V}}\mathbf{\Phi}_3]_j}{[\widetilde{\mathbf{V}}\mathbf{\Phi}_2]_j} \right|. \end{cases} \tag{21}$$

Thus $[\mathbf{i}]_l$ can be computed with (18).

In the first stage of CE, it is also possible to obtain the power prior of each path as

$$[\mathbf{p}]_l \propto |[\mathbf{x}]_{s_l}|^2, \tag{22}$$

where $\mathbf{x} = \sum_{n=1}^{N} \mathbf{x}_n$. Notably, designs in (15) and (20) can be readily extended to cases where $N > 3$. This is done by replicating their structure with element index shifts, exploiting the cyclic symmetry property.

B. Efficient Channel Estimation With Prior

Using the power prior of each path obtained in (22), it is feasible to estimate the L paths from strongest to weakest in the second stage of CE. In practice, estimating all L paths is unnecessary, as some paths are negligible for beamforming design, which will be elaborated in Section III-C. Let $\tilde{\mathbf{i}}$ be the sorted position prior vector with descending power prior. Therefore, with the position prior $\tilde{\mathbf{i}}$, estimation of the $\widetilde{L}$ paths with (5) can be simplified to

$$\mathbf{y}_l = \widetilde{\mathbf{Q}}_l \widetilde{\mathbf{x}}_l + \mathbf{n}_l, \tag{23}$$

where $\widetilde{\mathbf{Q}}_l \in \mathbb{C}^{N_t^B N_r^B \times \widetilde{L}}$, $\widetilde{\mathbf{x}} \in \mathbb{C}^{\widetilde{L}\times 1}$, $\widetilde{\mathbf{Q}}_l(:, i) = \mathbf{Q}_l(:, s_i)$ and $[\widetilde{\mathbf{x}}_l]_i = [\mathbf{x}_l]_{s_i}$. The reflection vector $\mathbf{\Psi}_l$ can be expressed as (19). Since $\tilde{\mathbf{i}}$ may not be accurate enough, i.e., with the prior grid potentially offset from the real non-zero index. Therefore, multiple searches of the path in the neighboring grids need to be conducted. To enable this, $\widehat{L} > \widetilde{L}$ patterns are employed in the second-stage CE, among which $\widetilde{L}$ patterns are effective for obtaining CSI. This is facilitated by $\widetilde{\mathbf{V}}\mathbf{\Psi}_l$ in the second stage of CE which has only one non-zero element, ensuring the sparse character of $\widetilde{\mathbf{x}}_k$ [2, Fig. 1]. The overall process of CE is summarized in Fig. 1.

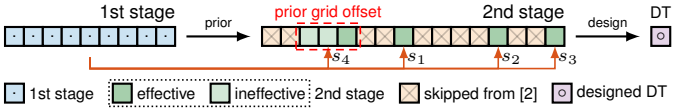

Fig. 1. Overall process of CE. Each square represents a reflection pattern.

The computational complexity comparison for a widely employed CS algorithm orthogonal matching pursuit (OMP) [11] is shown in Table I, where L' is the sparsity of $\mathbf{x}_l$, and Q, Q' and $Q_{1,2}$ are the average number of pilots for each reflection pattern in [5], [2] and the first/second stage of CE in this work. The overall complexity of the proposed method consists of two parts: $\mathcal{O}(LQ_1NN_tN_r)$ for the first stage, and $\mathcal{O}(L'Q_2\widehat{L}\widetilde{L})$ for the second stage. Here, N represents the number of reflection patterns estimated in the first stage ($N \ll M$), $\widehat{L}$ denotes the number of reflection patterns estimated in the second stage ($\widehat{L} \ll \min\{N_tN_r, M\}$), and $\widetilde{L}$ is the dimension of variable $\widetilde{\mathbf{x}}_l$ with $\widetilde{L} < L$. It clearly shows the reduced complexity compared with existing works [2], [5].

C. Joint Hybrid Beamforming Design

We employ an iterative approach to obtain an optimized reflection and beam pattern for DT, where it initially solves

TABLE I
COMPUTATIONAL COMPLEXITY OF OMP-BASED FORMULATIONS

Sparse Channel Formulation	Computational Complexity
Eq. (3) from [5]	$\mathcal{O}(LQMN_tN_r)$
Eq. (5) from [2]	$\mathcal{O}(L'Q'MN_tN_r)$
This Work (2 Stages)	$\mathcal{O}(LQ_1NN_tN_r + L'Q_2\widehat{L}\widetilde{L})$

$L' < \widetilde{L} < L \ll N_tN_r, \quad \widehat{L} \ll \min\{N_tN_r, M\}, \quad Q_2 \ll Q' < Q_1 \ll Q.$

the reflection pattern optimization problem using fixed hybrid beamforming. Subsequently, the optimized hybrid beamforming design using a fixed reflection pattern is obtained. This process is iterated until convergence. When $N_r^{\mathrm{RF}} \neq N_t^{\mathrm{RF}}$ and/or $N_r^B \neq N_t^B$, (13) is transformed into

$$R \approx N_r^B \log_2 \frac{1}{N_t^B \sigma^2} + \log_2 \big|\mathbf{W}_{\mathrm{BB}}^{\mathsf{H}}\mathbf{W}_{\mathrm{RF}}^{\mathsf{H}}\mathbf{H} \times \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^{\mathsf{H}}\mathbf{F}_{\mathrm{RF}}^{\mathsf{H}}\mathbf{H}^{\mathsf{H}}\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\big|, \tag{24}$$

and with the fixed analog beamforming, we can formulate the reflection pattern optimization as

$$\max_{\boldsymbol{\Psi}} \big\{ \log_2 \big|\mathbf{W}_{\mathrm{BB}}^{\mathsf{H}}\mathbf{W}_{\mathrm{RF}}^{\mathsf{H}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^{\mathsf{H}}\mathbf{F}_{\mathrm{RF}}^{\mathsf{H}}\mathbf{H}^{\mathsf{H}} \times \mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\big| \big\}, \quad \text{s.t. } \boldsymbol{\Psi} \triangleq [\psi_1, \psi_2, \ldots, \psi_M]^{\mathsf{T}}. \tag{25}$$

The reflecting elements are refined one by one, optimizing ψ_i while fixing the other $(M-1)$ elements for $1 \leq i \leq M$. (25) is further written as

$$\begin{aligned} &\log_2 \big|\mathbf{W}_{\mathrm{BB}}^{\mathsf{H}}\mathbf{W}_{\mathrm{RF}}^{\mathsf{H}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^{\mathsf{H}}\mathbf{F}_{\mathrm{RF}}^{\mathsf{H}}\mathbf{H}^{\mathsf{H}}\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\big| \\ &\overset{(a)}{=} \log_2 \Big|\Big(\psi_i\mathbf{M}_i + \sum_{1\leq m\leq M, m\neq i}\psi_m\mathbf{M}_m\Big)\Big(\psi_i^{\mathsf{H}}\mathbf{M}_i^{\mathsf{H}} + \sum_{1\leq m\leq M, m\neq i}\psi_m^*\mathbf{M}_m^{\mathsf{H}}\Big)\Big| \\ &\overset{(b)}{=} \log_2 \big|\mathbf{Z}^{-1}\big| + \log_2 \Big|\mathbf{I}_{N_r^B} + \mathbf{Z}^{-1}\psi_i\mathbf{M}_i\sum_{m=1,m\neq i}\psi_m^*\mathbf{M}_m^{\mathsf{H}} \\ &\quad + \mathbf{Z}^{-1}\Big(\sum_{1\leq m\leq M, m\neq i}\psi_m\mathbf{M}_m\Big)\psi_i^*\mathbf{M}_i^{\mathsf{H}}\Big| \\ &\overset{(c)}{=} \log_2 \big|\mathbf{Z}^{-1}\big| + \log_2 \big|\mathbf{I}_{N_r^B} + \psi_i\mathbf{Z}^{-1}\mathbf{m}_i\mathbf{n}_i^{\mathsf{H}} + \psi_i^*\mathbf{Z}^{-1}\mathbf{n}_i\mathbf{m}_i^{\mathsf{H}}\big|, \end{aligned} \tag{26}$$

where (12) is used in (a) and $\mathbf{M}_m \triangleq \mathbf{W}_{\mathrm{BB}}^{\mathsf{H}}\mathbf{W}_{\mathrm{RF}}^{\mathsf{H}}\mathbf{N}_m\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$. In (b) the property $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$ is employed for square matrices $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{Z} \triangleq \mathbf{M}_i\mathbf{M}_i^{\mathsf{H}} + (\sum_{1\leq m\leq M, m\neq i}\psi_m\mathbf{M}_m)(\sum_{1\leq m\leq M, m\neq i}\psi_m^*\mathbf{M}_m^{\mathsf{H}})$. In (c) $\mathbf{M}_i \triangleq \mathbf{U}_i\boldsymbol{\Sigma}_i\mathbf{V}_i$ by using singular value decomposition (SVD) where the first diagonal element of $\boldsymbol{\Sigma}$ is the dominant element, thus $\boldsymbol{\Sigma}_i \approx \widetilde{\mathbf{m}}\widetilde{\mathbf{n}}^{\mathsf{H}}$ and $\mathbf{m}_i \triangleq \mathbf{U}_i\widetilde{\mathbf{m}}$, $\mathbf{n}_i^{\mathsf{H}} \triangleq \widetilde{\mathbf{n}}^{\mathsf{H}}\mathbf{V}_i\sum_{1\leq m\leq M, m\neq i}\psi_m^*\mathbf{M}_m^{\mathsf{H}}$. Therefore, the optimization problem is transformed into

$$\max_{\psi_i} \big\{\log_2 \big|\mathbf{I}_{N_r^B} + \psi_i\mathbf{Z}^{-1}\mathbf{m}_i\mathbf{n}_i^{\mathsf{H}} + \psi_i^*\mathbf{Z}^{-1}\mathbf{n}_i\mathbf{m}_i^{\mathsf{H}}\big|\big\}, \quad \text{s.t. } \psi_i = \exp(\jmath\theta_i). \tag{27}$$

With $\mathbf{Z}^{-1} = (\mathbf{Z}^{-1})^{\mathsf{H}}$, $|\mathbf{I}_{N_r^B} + \psi_i\mathbf{Z}^{-1}\mathbf{m}_i\mathbf{n}_i^{\mathsf{H}}|$ and $|\mathbf{I}_{N_r^B} + \psi_i^*\mathbf{Z}^{-1}\mathbf{n}_i\mathbf{m}_i^{\mathsf{H}}|$ have the same maximum point. Therefore, (27) is simplified to

$$\max_{\psi_i} \big\{\log_2 \big|1 + 2\psi_i\mathbf{n}_i^{\mathsf{H}}\mathbf{Z}^{-1}\mathbf{m}_i\big|\big\}, \quad \text{s.t.} \quad \psi_i = \exp(\jmath\theta_i), \tag{28}$$

where the property $|\mathbf{I} + \mathbf{AB}| = |\mathbf{I} + \mathbf{BA}|$ is used. Therefore, $\theta_i = -\arg(\mathbf{n}_i^{\mathsf{H}}\mathbf{Z}^{-1}\mathbf{m}_i)$. With an optimized $\boldsymbol{\Psi}_o$, the corresponding channel $\mathbf{H}_o$ can be expressed using SVD as

$$\mathbf{H}_o = \mathbf{V}_o\boldsymbol{\Sigma}_o\mathbf{U}_o^{\mathsf{H}}, \tag{29}$$

where $\mathbf{V}_o \in \mathbb{C}^{N_r\times N_r}$ and $\mathbf{U}_o \in \mathbb{C}^{N_t\times N_t}$ are unitary matrices and $\boldsymbol{\Sigma}_o$ is a rectangular diagonal matrix consisting of descending singular values. With the sparse nature of mmWave channels, the cascaded channel can be approximated by

$$\mathbf{H}_o \approx \widetilde{\mathbf{V}}_o\widetilde{\boldsymbol{\Sigma}}_o\widetilde{\mathbf{U}}_o^{\mathsf{H}}, \tag{30}$$

where $\widetilde{\mathbf{V}}_o \triangleq \mathbf{V}_o(:, 1:N_t^B)$, $\widetilde{\mathbf{U}}_o \triangleq \mathbf{U}_o(:, 1:N_r^B)$, $\widetilde{\boldsymbol{\Sigma}}_o \triangleq \boldsymbol{\Sigma}_o(1:N_t^B, 1:N_r^B)$. Therefore, the optimized unconstrained beamformers for $\mathbf{H}$ can be given by $\mathbf{F}_o = \widetilde{\mathbf{V}}_o$, $\mathbf{W}_o = \widetilde{\mathbf{U}}_o$ and $\mathbf{F}_o$ and $\mathbf{W}_o$ can be easily formulated by hybrid beamforming [3], [8]. The final beamforming design can be obtained by iterating operations in (28) and (30).

D. Pilot Allocation Analysis for CE

As a two-stage method, it is significant to achieve a tradeoff between the allocated pilots in the first and second stages. In the CE sense, for a given $\widetilde{L}$, the optimal pilot allocation can be expressed as

$$\min_{N, Q_{1,2}} \{Q_1 \cdot N + Q_2 \cdot \mathbb{E}[\widehat{L}]\}, \tag{31}$$

where the average number of reflection patterns in the second stage can be expressed as

$$\mathbb{E}[\widehat{L}] = \widetilde{L} + \sum_{l=1}^{\widetilde{L}} \left\lfloor 2\frac{|\mathbf{er}(l\,; N, Q_1)|}{\alpha} \right\rfloor, \tag{32}$$

where $\mathbf{er}(l\,; N, Q_1)$ denotes the prior grid offset in the first stage for the l-th path, which depends on channel conditions (number of scatters, average signal-to-noise ratio (SNR), etc.) as well as N and Q_1. Larger N and Q_1 contribute to smaller $\mathbf{er}(l\,; N, Q_1)$. α denotes the first-stage allowable error explained in Section III-A. Therefore, (31) is transformed to

$$\min_{N, Q_{1,2}} \left\{ Q_1 \cdot N + Q_2 \left(\widetilde{L} + 2\sum_{l=1}^{\widetilde{L}} \left\lfloor 2\frac{|\mathbf{er}(l\,; N, Q_1)|}{\alpha} \right\rfloor \right) \right\}, \tag{33}$$

which is a multi-variable optimization problem that can be solved according to the specific applied scenarios. For a specific scenario with Q_2 fixed, the total pilot overhead is minimized by solving (33) to obtain Q_1:

$$\frac{\partial \sum_{l=1}^{\widetilde{L}} \left\lfloor 2\frac{|\mathbf{er}(l\,; N, Q_1)|}{\alpha} \right\rfloor}{\partial Q_1} = -\frac{N}{Q_2}, \tag{34}$$

where $|\mathbf{er}(l\,; N, Q_1)|$ can be readily obtained via simulation. Then, Q_1 is selected as the one corresponding to the smallest function value among all the stationary points obtained from (34). Moreover, $\widetilde{L}$ in (34) is constrained by the SE maximization sense as

$$\widetilde{L} \geq N_t^B N_r^B, \tag{35}$$

so that the approximate performance limit in (30) can be achieved. Generally, when the number of pilot overhead is small, focusing on estimating a few dominant paths with higher power yields better performance.

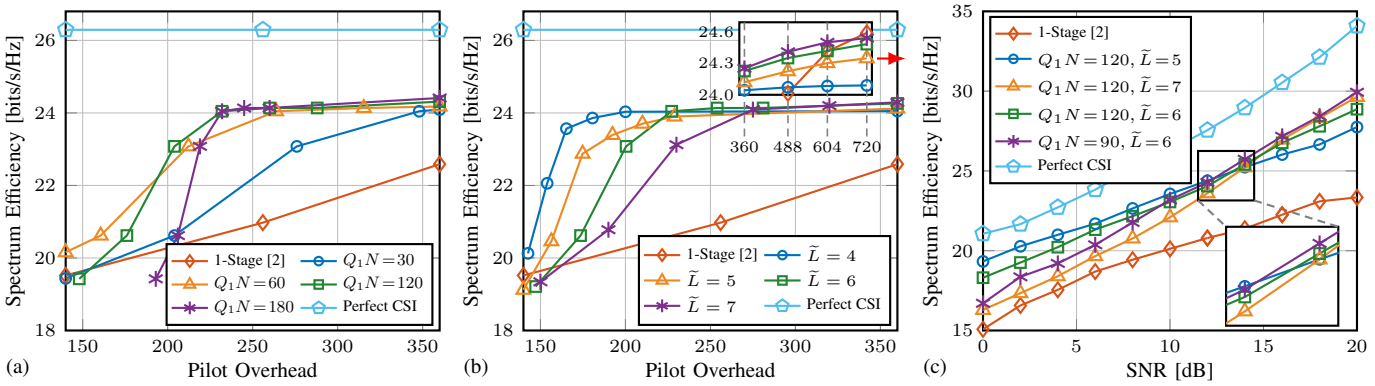

Fig. 2. Simulation results of SE in relation to different CE schemes with the OMP algorithm. (a) Different Q_1N with $\mathrm{SNR}=10\,\mathrm{dB}$. (b) Different $\widetilde{L}$ with $\mathrm{SNR}=10\,\mathrm{dB}$. (c) SE v.s. SNR with pilot overhead 200.

IV. Simulation Results

In this section, simulations are conducted to evaluate the performance of different pilot allocation schemes and different number of estimated paths for channel estimation. The OMP algorithm is employed, and the system parameters are $N_t = N_t^G = 8$, $N_r = N_r^G = 16$, $N_t^B = N_r^B = 2$, $M_x = M_y = M_x^G = M_y^G = 8$ and $L_1 = L_2 = 3$.

The proposed method is compared with existing work including [2], which designs the beam and reflection pattern in the CE stage. As shown in Fig. 2, compared with [2], our proposed method can achieve approximately $40\% \sim 50\%$ pilot reduction to achieve the same SE. Our method has more than 30% SE higher than [5] with low pilot/SNR, and the result is not shown in Fig. 2 due to a large margin. Notably, inadequate CE can significantly impact DT SE performance which results from poor divergence. Fig. 2(a) shows the relationship between varying Q_1N and the obtained SE with different pilot overheads, when SNR is $10\,\mathrm{dB}$. With a lower required SE, moderately reducing the pilot overhead in the first stage (for example $Q_1N = 60$) can effectively reduce the total pilot overhead. When the total number of pilot overhead is relatively large, the impact of different pilot allocation schemes on SE is negligible, which implies the stability of the proposed two-stage method. Fig. 2(b) depicts the relationship between different $\widetilde{L}$ and SE with different pilot overheads and $\mathrm{SNR} = 10\,\mathrm{dB}$, showing that reducing the number of estimated paths while satisfying (35) can cut down the total pilot overhead. The relationship between SNR and SE is shown in Fig. 2(c) with 200 total pilots, where the received power is assumed to be the same among all SNRs. It can be observed that larger $\widetilde{L}$ or smaller Q_2 provides better performance in high SNRs. In summary, simulation verifies the effectiveness of the proposed method and the flexibility in pilot allocation for CE. Adjustable parameters $\widetilde{L}$, N, and $Q_{1,2}$ further contribute to flexibility, making the proposed method applicable to multiple scenarios.

V. Conclusion

In this letter, we propose the efficient joint design of channel estimation and data transmission for RIS-assisted mmWave MIMO systems. Simulation results demonstrate that, with the same SE, the proposed method can reduce the pilot overhead by $40\% \sim 50\%$ compared to existing methods. Moreover, this method can adapt to various environments by adjusting the pilot allocation and the number of estimated paths.

References

[1] S. Basharat, S. A. Hassan, H. Pervaiz *et al.*, "Reconfigurable intelligent surfaces: Potentials, applications, and challenges for 6G wireless networks," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 184–191, Dec. 2021.

[2] Y. You, W. Zhao, L. Zhang, X. You, and C. Zhang, "Beam pattern and reflection pattern design for channel estimation in RIS-assisted mmwave MIMO systems," *IEEE Trans. Veh. Technol.*, 2023, to be published, doi: 10.1109/TVT.2023.3309950.

[3] Q. Zhu, H. Li, R. Liu, M. Li, and Q. Liu, "Hybrid beamforming and passive reflection design for RIS-assisted mmWave MIMO systems," in *Proc. IEEE ICC Workshops*, Jun. 2021, pp. 1–6.

[4] Q. Zhu, R. Liu, Y. Liu, M. Li, and Q. Liu, "Joint design of hybrid and reflection beamforming for RIS-aided mmWave MIMO communications," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2021, pp. 1–6.

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

[6] Z. Wan, Z. Gao, F. Gao, M. Di Renzo, and M.-S. Alouini, "Terahertz massive MIMO with holographic reconfigurable intelligent surfaces," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4732–4750, Jul. 2021.

[7] S. Liu, Z. Gao, J. Zhang, M. Di 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.

[8] Z. Wang, M. Li, Q. Liu, and A. L. Swindlehurst, "Hybrid precoder and combiner design with low-resolution phase shifters in mmWave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 256–269, May 2018.

[9] W. Zhao, Y. You, L. Zhang, X. You, and C. Zhang, "OMPL-SBL algorithm for intelligent reflecting surface-aided mmWave channel estimation," *IEEE Trans. Veh. Technol.*, vol. 72, no. 11, pp. 15121–15126, Nov. 2023.

[10] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2nd Quart. 2018.

[11] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.