Beyond Type-II: Site-Specific Subspace Inference for Efficient CSI Feedback

Anonymous Authors[1]

Abstract

We propose a site-specific Type-II feedback design for downlink limited-feedback beamforming. The paper first formulates an effective-rate objective that couples channel-state information (CSI) probing, user equipment (UE)-side compression, base station (BS)-side reconstruction, and online overhead. Guided by this objective, the BS uses a low-overhead synchronization signal block (SSB) reference signal received power (RSRP) fingerprint and a learned site-specific propagation prior to infer a UE-dependent dominant beam subspace before explicit CSI acquisition. The UE then estimates and reports only the effective channel coefficients in this subspace, avoiding full-dimensional online subspace search while retaining a Type-II-style multi-beam representation. We analyze the resulting subspace-capture behavior and jointly optimize the probing codebook and subspace inference network for channel-energy preservation. Experiments show Type-II-comparable capture quality with lower online overhead and UE-side complexity, leading to higher effective spectral efficiency.

1. Introduction

Beamforming is a key enabler of modern cellular systems, but its performance critically depends on how accurately the base station (BS) acquires user equipment (UE)-specific channel-state information (CSI) (Heath et al., 2016). Current 3rd Generation Partnership Project (3GPP) New Radio (NR) systems address this problem through standardized codebook-based feedback schemes including Type-I, Type-II, and port-selection codebook (PSC) (3GPP, 2018b; Fu et al., 2023). These schemes define different overhead-complexity-accuracy tradeoffs: Type-I is lightweight but limited in rich multipath, Type-II is more expressive but requires full-dimensional CSI acquisition, dominant-subspace search at the UE, and heavier feedback, and PSC is structured but constrained by the selected effective-port domain (Love et al., 2008; Giordani et al., 2019; Fu et al., 2023).

To improve the performance of existing codebooks with artificial intelligence, prior studies mainly follow two lines. One stays within standardized or near-standardized CSI feedback and improves beam management or codebook design while preserving signaling compatibility (Dreifuerst & Heath, 2024; 2025). The other exploits site-specific or environment-aware structure for probing, beam alignment, or beam synthesis (Heng et al., 2022; Heng & Andrews, 2024; Ning et al., 2023; Wu et al., 2024; Zhao et al., 2026a; Wang et al., 2026), but usually departs from the standardized limited-feedback architecture and therefore does not explain how site-specific priors should be incorporated into a practical CSI feedback pipeline.

This paper targets this missing middle ground. We refer to the learned propagation prior of a deployment site as *site-specific information (SSI)* and use it to assist, rather than replace, standardized feedback. The key idea is to move the dominant-subspace discovery task from the UE to the BS. The BS first combines SSI with a low-overhead synchronization signal block (SSB) reference signal received power (RSRP) fingerprint to infer a UE-dependent low-dimensional subspace. It then transmits CSI-reference signals (CSI-RS) only over this inferred subspace, so that the UE only needs to report the effective coefficients inside it. This preserves the Type-II-style multi-beam representation while reducing online CSI acquisition and UE-side processing.

The main contributions are threefold. First, we formulate an effective-rate-oriented limited-feedback objective that makes the tradeoff among CSI-RS probing, UE-side compression, BS-side reconstruction, and online overhead explicit. Second, we develop a site-specific Type-II feedback scheme in which the BS infers a UE-dependent dominant subspace before explicit CSI acquisition and the UE feeds back only low-dimensional coefficients. Third, we analyze the resulting subspace-capture behavior and co-optimize the SSB probing codebook with the BS-side inference network. Multi-scenario experiments show Type-II-like capture quality with a more favorable overhead-efficiency tradeoff.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. System Model

We consider a single-cell downlink (DL) system in which an N_t-antenna BS serves a single-antenna UE. The channel is assumed block fading and approximately constant within one coherence block.

2.1. Channel Model

We adopt a sparse geometric channel model in which the channel is a superposition of L dominant propagation paths (Ayach et al., 2014). The DL channel vector $\mathbf{h} \in \mathbb{C}^{N_t \times 1}$ is then represented as

$$\mathbf{h} = \sum_{l=1}^{L} \alpha_l \mathbf{a}(\varphi_l) = \mathbf{A}\boldsymbol{\alpha}, \tag{1}$$

where $\alpha_l \in \mathbb{C}$ denotes the complex gain of the l-th path and $\mathbf{a}(\varphi_l)$ is the transmit-array steering vector associated with the angle of departure φ_l. We define $\mathbf{A} = [\mathbf{a}(\varphi_1), \ldots, \mathbf{a}(\varphi_L)] \in \mathbb{C}^{N_t \times L}$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]^T \in \mathbb{C}^{L \times 1}$.

Assumption 2.1 (Near-orthogonal angular model)**.** The BS employs a uniform linear array (ULA) with normalized steering vector

$$\mathbf{a}(u_l) = \frac{1}{\sqrt{N_t}} \left[1, e^{j2\pi u_l}, \cdots, e^{j2\pi(N_t-1)u_l}\right]^T, \tag{2}$$

where $u_l = \frac{d}{\lambda}\sin(\varphi_l)$. Dominant spatial frequencies are sufficiently separated and the array aperture is large enough to resolve them, so $\mathbf{A}^H\mathbf{A} \approx \mathbf{I}_L$.

Assumption 2.1 allows the channel energy to be approximately decomposed across resolvable angular components. With unit-power beamformer $\mathbf{w}$ and data symbol s, the received signal is then given by

$$y = \sqrt{P_t}\mathbf{h}^H\mathbf{w}s + n, \tag{3}$$

where P_t is the transmit power and $n \sim \mathcal{CN}(0, \sigma_n^2)$.

2.2. SSB-Based Initial Access and RSRP Fingerprint

Before CSI feedback, the UE performs initial access through SSB beam sweeping. The BS transmits a predefined SSB probing codebook $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_K] \in \mathbb{C}^{N_t \times K}$ (3GPP, 2018a;c). For the i-th probing beam, the received SSB signal is

$$\mathbf{y}_{\mathrm{SSB},i} = \sqrt{P_{\mathrm{SSB}}}\mathbf{h}^H\mathbf{b}_i\mathbf{s}_{\mathrm{SSB}} + \mathbf{n}_{\mathrm{SSB},i}, \tag{4}$$

where P_{SSB} is the SSB transmit power, $\mathbf{s}_{\mathrm{SSB}} \in \mathbb{C}^{L_s \times 1}$ collects the SSB symbols used for this measurement, and $\mathbf{n}_{\mathrm{SSB},i} \sim \mathcal{CN}(\mathbf{0}_{L_s}, \sigma_n^2\mathbf{I}_{L_s})$ is the noise vector. The UE averages the received power and reports a decibel (dB)-domain RSRP vector. Following (Zhao et al., 2026a), the dB-domain RSRP fingerprint is modeled as

$$\mathbf{r}_{\mathbf{B}} = \mathbf{r}_{\mathbf{B}}^0 + \mathbf{n}_{\mathbf{B}}, \tag{5}$$

where $\mathbf{r}_{\mathbf{B}}^0 = [r_{\mathbf{b}_1}^0, \ldots, r_{\mathbf{b}_K}^0]^T$ is the noise-free RSRP fingerprint and $\mathbf{n}_{\mathbf{B}} = [n_{\mathbf{b}_1}, \ldots, n_{\mathbf{b}_K}]^T$ collects Gaussian perturbations with mean μ_b and variance σ_b^2. The detailed expressions of $r_{\mathbf{b}_i}^0$, μ_b, and σ_b^2 are given in Eq. (7), Eq. (8a), and Eq. (8b) of (Zhao et al., 2026a). During initial access, the UE uses $\mathbf{r}_{\mathbf{B}}$ to determine the access beam and associated random-access resources before CSI acquisition.

This stage provides only coarse beam-level information and is insufficient for high-quality downlink beamforming. Data transmission still requires a refined UE-specific CSI representation, which in practical limited-feedback systems entails additional CSI-RS transmission, UE-side channel estimation, CSI compression, and feedback. With large transmit arrays, the resulting online overhead can be substantial.

2.3. Problem Formulation

If the BS knew $\mathbf{h}$, the optimal single-user single-stream beamformer would be the maximum-ratio transmission (MRT) beamformer $\mathbf{w}^\star = \mathbf{h}/\|\mathbf{h}\|$. In practice, however, the BS does not directly access the UE-specific DL channel. In frequency-division duplex (FDD) systems, the DL CSI must therefore be estimated at the UE and fed back to the BS.

We abstract this DL CSI acquisition process as three coupled stages: BS-side CSI-RS transmission, UE-side CSI acquisition and compression, and BS-side CSI reconstruction. Let $\mathbf{C} \in \mathbb{C}^{N_t \times N_c}$ denote the CSI-RS precoder and $\mathbf{S}_{\mathrm{CSI}} \in \mathbb{C}^{N_c \times L_c}$ the training matrix. The UE receives

$$\mathbf{y}_{\mathrm{CSI}}^T(\mathbf{C}) = \sqrt{P_{\mathrm{CSI}}}\mathbf{h}^H\mathbf{C}\mathbf{S}_{\mathrm{CSI}} + \mathbf{n}_{\mathrm{CSI}}^T, \tag{6}$$

where P_{CSI} is the CSI-RS transmit power and $\mathbf{n}_{\mathrm{CSI}} \sim \mathcal{CN}(\mathbf{0}_{L_c}, \sigma_n^2\mathbf{I}_{L_c})$. Based on $\mathbf{y}_{\mathrm{CSI}}(\mathbf{C})$, the UE obtains CSI through an acquisition mapping $\mathcal{R}$, e.g., least-squares (LS) or minimum mean square error (MMSE) estimation (Hassibi & Hochwald, 2003), denoted by $\mathbf{h} = \mathcal{R}(\mathbf{y}_{\mathrm{CSI}})$. To focus on the feedback structure rather than estimation error, we assume perfect acquisition under sufficient training. Since reliable recovery typically requires L_c to scale at least with N_c, DL channel estimation incurs at least quadratic training overhead.

Because feeding back full CSI is generally prohibitive, the UE compresses the acquired CSI into $\mathbf{z} = \mathcal{Q}(\mathbf{h})$, where $\mathcal{Q} : \mathbb{C}^{N_t \times 1} \to \mathbb{C}^{N_q \times 1}$ is the UE-side compression mapping and N_q is an abstract proxy for feedback overhead. The BS then reconstructs $\hat{\mathbf{h}} = \mathcal{F}(\mathbf{z})$, where $\mathcal{F} : \mathbb{C}^{N_q \times 1} \to \mathbb{C}^{N_t \times 1}$ is the BS-side reconstruction mapping. Since $\mathcal{Q}$ is generally

lossy, $\mathcal{F}$ need not invert $\mathcal{Q}$. The resulting DL beamformer is

$$\hat{\mathbf{w}} = \frac{\hat{\mathbf{h}}}{\|\hat{\mathbf{h}}\|} = \frac{\mathcal{F}\big(\mathcal{Q}(\mathcal{R}(\mathbf{y}_{\mathrm{CSI}}(\mathbf{C})))\big)}{\|\mathcal{F}\big(\mathcal{Q}(\mathcal{R}(\mathbf{y}_{\mathrm{CSI}}(\mathbf{C})))\big)\|}. \tag{7}$$

Let $T_o = T_{\mathrm{SSB}} + T_{\mathrm{CSI}}$ denote the total online overhead in a coherence block of T_c channel uses, where T_{SSB} accounts for SSB probing and RSRP reporting when needed, and T_{CSI} accounts for CSI-RS transmission and CSI feedback. Denoting $\rho = P_t/\sigma_n^2$, the limited-feedback beamforming design problem for maximizing the effective rate can be abstracted as

$$\max_{\mathbf{C},\,\mathcal{Q},\,\mathcal{F}} \left(1 - \frac{T_o}{T_c}\right) \log_2\!\left(1 + \rho \left|\mathbf{h}^H \hat{\mathbf{w}}\right|^2\right). \tag{8}$$

This objective captures the central tradeoff: richer probing and feedback can improve beamforming quality, but also increase online overhead and UE-side processing. The proposed design addresses this tradeoff by using SSI and the SSB RSRP fingerprint to select a low-dimensional CSI-RS subspace at the BS, while the UE only estimates and feeds back the instantaneous coefficients within that subspace.

3. Site-Specific Type-II Feedback

Since the CSI acquisition, compression, and reconstruction mappings are non-parametric and coupled, the problem formulated above is highly non-convex and intractable to solve directly. Current 3GPP NR standards do not solve this problem in a direct optimization sense. Instead, they approximate the design objective through standardized limited-feedback schemes. A more detailed analysis can be found in (Zhao et al., 2026b). In particular, conventional Type-II feedback provides a flexible multi-beam representation, but it requires online dominant-subspace discovery from high-dimensional CSI at the UE. The proposed scheme addresses this burden by shifting dominant-subspace inference to the BS while retaining the useful Type-II-style coefficient representation.

3.1. Site-Specific Subspace Inference

Compared with conventional Type-II feedback, the key idea of the proposed feedback scheme is to infer the dominant transmit subspace at the BS before explicit UE-side CSI estimation and feedback. To do so, it uses the RSRP fingerprint $\mathbf{r_B}$ obtained during SSB probing and maps it to a UE-dependent low-dimensional basis

$$\mathbf{C}_p = \Psi(\mathbf{r_B}) = [\mathbf{c}_{p,1}, \cdots, \mathbf{c}_{p,Q}] \in \mathbb{C}^{N_t \times Q}, \tag{9}$$

where $Q \ll N_t$ and the columns of $\mathbf{C}_p$ are expected to align with the dominant propagation directions of the current UE. We assume that $\mathbf{C}_p$ is an orthonormal basis of the inferred subspace, since any linearly independent basis can be transformed into an equivalent orthonormal basis via QR decomposition without changing its span.

In this way, the dominant transmit directions are no longer searched exhaustively at the UE as in conventional Type-II, but are instead inferred at the BS by conditioning the offline learned SSI on the RSRP fingerprint. The remaining online feedback task is therefore reduced to refining the instantaneous low-dimensional coefficients over the inferred subspace. Concretely, the BS transmits beamformed CSI-RS only over this inferred subspace:

$$\mathbf{y}_{\mathrm{CSI}}^T(\mathbf{C}_p) = \sqrt{P_{\mathrm{CSI}}}\mathbf{h}^H \mathbf{C}_p \mathbf{S}_{\mathrm{CSI}} + \mathbf{n}_{\mathrm{CSI}}^T, \tag{10}$$

where $\mathbf{S}_{\mathrm{CSI}} \in \mathbb{C}^{Q \times L_c}$ is the training matrix. The UE then observes the effective channel $\mathbf{h}_p = \mathbf{C}_p^H \mathbf{h}$. Again assuming perfect recovery of $\mathbf{h}_p$ from the received CSI-RS, the UE directly feeds back this low-dimensional effective channel $\mathbf{z}_p = \mathcal{Q}_p(\mathbf{h}_p) = \mathbf{h}_p$. After receiving $\mathbf{z}_p$, the BS reconstructs

$$\hat{\mathbf{h}}_p = \mathcal{F}_p(\mathbf{z}_p; \mathbf{C}_p) = \mathbf{C}_p \mathbf{z}_p = \mathbf{C}_p \mathbf{C}_p^H \mathbf{h}. \tag{11}$$

The resulting beamforming vector is therefore $\hat{\mathbf{w}}_p = \hat{\mathbf{h}}_p / \|\hat{\mathbf{h}}_p\|$. Compared with conventional Type-II feedback, the CSI-RS dimension is reduced from N_t to Q, and the UE no longer performs full-dimensional dominant-subspace search or coefficient fitting. The proposed scheme therefore preserves the practical feedback pipeline while shifting the dominant online CSI-processing burden from the UE to the BS.

3.2. CSI-Capture Efficiency Analysis

The proposed framework essentially projects the DL CSI $\mathbf{h}$ onto the inferred subspace $\mathcal{U} = \mathrm{span}(\mathbf{C}_p)$ through the projector $\mathbf{P}_p = \mathbf{C}_p \mathbf{C}_p^H$. The resulting beamformer gives the instantaneous rate

$$R_p = \log_2\left(1 + \rho \left|\mathbf{h}^H \hat{\mathbf{w}}_p\right|^2\right) = \log_2\left(1 + \rho \|\mathbf{P}_p \mathbf{h}\|^2\right). \tag{12}$$

This rate is upper-bounded by the ideal MRT benchmark,

$$R_{\mathrm{MRT}} = \log_2\left(1 + \rho \|\mathbf{h}\|^2\right). \tag{13}$$

We therefore define the CSI-capture efficiency of the proposed scheme as

$$\eta_p \triangleq \frac{\|\mathbf{P}_p \mathbf{h}\|^2}{\|\mathbf{h}\|^2}, \tag{14}$$

which measures the fraction of channel energy preserved by the inferred subspace. Since a Q-dimensional CSI representation is generally lossy when $Q < N_t$, the ideal MRT benchmark is unattainable in general, and thus $\eta_p < 1$.

We then suppose that there exists an oracle Q-dimensional subspace $\mathcal{U}_Q^\star$ that captures the most channel energy among all Q-dimensional subspaces, i.e.,

$$\mathcal{U}_Q^\star = \arg \max_{\mathcal{U}:\dim(\mathcal{U})=Q} \|\mathbf{P}_{\mathcal{U}} \mathbf{h}\|^2. \tag{15}$$

The corresponding orthogonal projector is denoted by $\mathbf{P}_Q^\star$. Here, $\mathcal{U}_Q^\star$ is used only as an analytical benchmark, since such a subspace is practically inaccessible at the BS. For example, when $Q = 1$, $\mathcal{U}_Q^\star = \mathrm{span}(\mathbf{h})$ is an oracle subspace, but the BS cannot infer $\mathbf{h}$ from the RSRP vector. Alternatively, this oracle subspace may be interpreted as the subspace that best aligns with the dominant propagation directions. In particular, the oracle subspace can be spanned by the array response vectors of the Q strongest paths, i.e., $\mathcal{U}_Q^\star = \mathrm{span}([\mathbf{a}(u_l)]_{l\in\mathcal{L}_Q^\star})$, where $\mathcal{L}_Q^\star$ denotes the set of indices of the Q strongest paths. The captured channel energy is therefore given by

$$\|\mathbf{P}_Q^\star \mathbf{h}\|^2 = \boldsymbol{\alpha}^H \mathbf{A}^H \mathbf{P}_Q^\star \mathbf{A}\boldsymbol{\alpha} \approx \sum_{l\in\mathcal{L}_Q^\star} |\alpha_l|^2 g_{l,Q^\star}, \quad (16)$$

where

$$g_{l,Q^\star} \triangleq \|\mathbf{P}_Q^\star \mathbf{a}(u_l)\|^2 \approx 1, \forall l \in \mathcal{L}_Q^\star \quad (17)$$

denotes the corresponding oracle path-capture factor. In this case, the top-Q paths are fully captured by the oracle subspace without the grid mismatch induced by the DFT basis in Type-I/II. Note that this oracle subspace is determined by the specific application and is typically unavailable to the BS. It is used only for theoretical analysis.

Theorem 3.1 (Subspace-mismatch bound)**.** *Let $\mathbf{P}_p$ be the projector onto the subspace inferred by the proposed SSI-enhanced feedback scheme, and let $\mathbf{P}_Q^\star$ denote the projector onto an oracle Q-dimensional benchmark subspace. Define $\delta_p \triangleq \|\mathbf{P}_p - \mathbf{P}_Q^\star\|_2$ and $[x]_+ \triangleq \max\{x, 0\}$. Then, the proposed scheme satisfies*

$$\|\mathbf{P}_p\mathbf{h}\|^2 \geq \left[\|\mathbf{P}_Q^\star\mathbf{h}\|^2 - \delta_p\|\mathbf{h}\|^2\right]_+, \quad (18)$$

and its achievable rate is lower bounded by

$$R_p \geq \log_2\left(1 + \rho\left[\|\mathbf{P}_Q^\star\mathbf{h}\|^2 - \delta_p\|\mathbf{h}\|^2\right]_+\right). \quad (19)$$

Moreover, the CSI-capture efficiency satisfies

$$\eta_p = \frac{\|\mathbf{P}_p\mathbf{h}\|^2}{\|\mathbf{h}\|^2} \geq \left[\frac{\|\mathbf{P}_Q^\star\mathbf{h}\|^2}{\|\mathbf{h}\|^2} - \delta_p\right]_+. \quad (20)$$

Theorem 3.1 separates the performance of the proposed feedback scheme into two interpretable terms. The first term, $\|\mathbf{P}_Q^\star\mathbf{h}\|^2$, represents the best channel energy that can be preserved by an ideal Q-dimensional benchmark subspace. It is therefore a dimension-limited benchmark: it is upper bounded by $\|\mathbf{h}\|^2$ and becomes equal to the full-channel energy only when the chosen Q-dimensional subspace contains the channel direction. The second term, $\delta_p\|\mathbf{h}\|^2$, quantifies the mismatch between the SSI-inferred subspace and the benchmark oracle subspace. Hence, the theorem shows that the proposed design is limited by both the intrinsic dimension constraint of the reduced CSI-RS subspace and the accuracy with which SSI-conditioned inference identifies that subspace.

3.3. Joint Design and Realization

The remaining design question is how to choose the SSB probing codebook $\mathbf{B}$ and the BS-side inference mapping $\Psi(\cdot)$. The subspace-mismatch bound above motivates a task-oriented design criterion: the inferred subspace should preserve as large a fraction of the channel energy as possible for the channel distribution of the target site under a Q-dimensional subspace bottleneck. However, directly minimizing the oracle mismatch δ_p is not practical because the oracle projector is unavailable and can be non-unique. We therefore formulate the design in terms of CSI-capture efficiency, which directly measures the relative energy preserved by the inferred subspace.

The probing codebook $\mathbf{B}$ is coupled with this objective because the inferred subspace is computed only from the RSRP fingerprint $\mathbf{r_B}$ generated during SSB probing. A generic oversampled DFT probing codebook provides uniform directional sounding, but it does not necessarily produce measurements that are most informative for dominant-subspace inference as analyzed and illustrated in (Zhao et al., 2026a). Conversely, the information-maximizing probing design in (Zhao et al., 2026a) is not tailored to the subspace-capture objective considered here. Therefore, the probing stage and the BS-side inference mapping should be optimized as a single task-oriented module: the probing codebook should shape the RSRP fingerprint so that the inference mapping can recover a low-dimensional subspace with high CSI-capture efficiency.

Accordingly, for a given channel realization, the joint design of the probing codebook $\mathbf{B}$ and the inference mapping $\Psi(\cdot)$ can be formulated as the following optimization problem

$$\max_{\mathbf{B},\,\Psi} \ \eta_p(\mathbf{h};\mathbf{B},\Psi) = \frac{\left\|\Psi(\mathbf{r_B})\Psi^H(\mathbf{r_B})\mathbf{h}\right\|^2}{\|\mathbf{h}\|^2}. \quad (21)$$

For fixed K and Q in the proposed framework, maximizing this quantity is equivalent to problem (8). However, this problem remains difficult to solve by conventional optimization techniques. In particular, the probing codebook $\mathbf{B}$ and the inference mapping $\Psi(\cdot)$ are strongly coupled through the nonlinear RSRP measurement process and the resulting subspace projector. Moreover, optimizing over the function space of $\Psi(\cdot)$ is itself intractable. We therefore parameterize the BS-side inference mapping as $\Phi_\theta(\cdot)$ with trainable parameters θ and adopt a task-driven end-to-end learning framework that directly optimizes the probing codebook and the inference mapping from site-specific channel data. In this way, the offline design problem over the site-specific channel distribution p_s is formulated as

$$\max_{\mathbf{B},\,\theta} \ \mathbb{E}_{\mathbf{h}\sim p_s}[\eta_p(\mathbf{h};\mathbf{B},\Phi_\theta)]. \quad (22)$$

Since p_s is not available in closed form and can only be accessed through site-specific channel samples, we approximate (22) using a dataset $\mathcal{H}_s = \{\mathbf{h}^{(n)}\}_{n=1}^{N_h}$ with enough

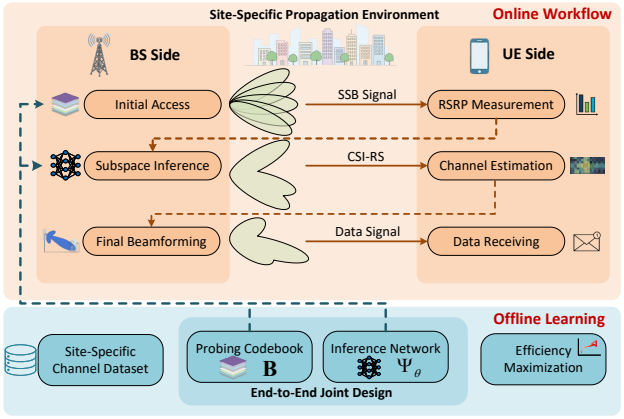


Figure 1. Illustration of the proposed site-specific Type-II feedback.

samples to represent p_s. The training objective is then formulated as

$$\min_{\mathbf{B},\theta} \mathcal{L}_{\mathcal{H}_s} = -\frac{1}{N_h}\sum_{n=1}^{N_h} \eta_p\Big(\mathbf{h}^{(n)}; \mathbf{B}, \Phi_\theta\Big). \tag{23}$$

However, when N_h is large, full-batch optimization becomes inefficient. We therefore adopt mini-batch SGD (Ghadimi & Lan, 2013). For each mini-batch $\mathcal{B} \subset \mathcal{H}_s$, the training objective is

$$\mathcal{L}_{\mathcal{B}}(\mathbf{B}, \theta) = -\frac{1}{|\mathcal{B}|}\sum_{\mathbf{h}^{(n)}\in\mathcal{B}} \eta_p\Big(\mathbf{h}^{(n)}; \mathbf{B}, \Phi_\theta\Big). \tag{24}$$

The overall workflow of the proposed limited-feedback scheme is shown in Fig. 1. Implementation details and convergence/complexity analysis are provided in Appendices B and C, respectively.

The overall implementation follows two stages: offline training and online deployment. During offline training, the probing codebook $\mathbf{B}$ and the inference network $\Phi_\theta(\cdot)$ are optimized jointly over site-specific channel samples, as summarized in **Algorithm 1**. During online deployment, the learned probing codebook $\mathbf{B}$ and subspace inference mapping $\Phi_\theta(\cdot)$ are embedded into the framework introduced in Section 3, whose procedure is summarized in **Algorithm 2**. A comparison between the conventional and proposed limited-feedback schemes in terms of performance and overhead is provided in Appendix D.

4. Numerical Results

We evaluate the effectiveness of the proposed end-to-end joint probing-and-inference design and the proposed site-specific Type-II feedback framework on DeepMIMO scenarios "asu_campus_3p5", "boston5g_3p5", "O1_28", and "O1B_28" (Alkhateeb, 2019). Unless otherwise stated, the experiments are conducted on "asu_campus_3p5" under the simulation settings summarized in Table 3 in Appendix E.

Algorithm 1 Offline Joint Training of the Probing Codebook and the Subspace Inference Mapping

Input Site-specific CSI dataset $\mathcal{H}_s$, batch size B, step β, iterations I
Output Optimized probing codebook and subspace inference mapping $\mathbf{B}^\star$ and $\Phi_{\theta^\star}$
1: Initialize probing codebook $\mathbf{B}^{(0)}$ and network parameters $\boldsymbol{\theta}^{(0)}$.
2: **for** $i = 1, 2, \cdots, I$ **do**
3: Sample a mini-batch of data $\mathcal{B}_i \subset \mathcal{H}_s$ with $|\mathcal{B}_i| = B$.
4: Generate probing RSRP measurements by (5).
5: BS infers the dominant subspace $\mathbf{C}_p$.
6: Compute the loss $\mathcal{L}_{\mathcal{B}_i}(\mathbf{B}^{(i-1)}, \boldsymbol{\theta}^{(i-1)})$ by (24).
7: Update $\mathbf{B}$ and $\boldsymbol{\theta}$ by $\mathbf{B}^{(i)} \leftarrow \mathbf{B}^{(i-1)} - \beta\nabla_{\mathbf{B}}\mathcal{L}_{\mathcal{B}_i}$ and $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{(i-1)} - \beta\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{B}_i}$.
8: **end for**
9: The trained probing codebook and inference module are obtained as $\mathbf{B}^\star = \mathbf{B}^{(I)}$ and $\Phi_{\theta^\star} = \Phi_\theta^{(I)}$.

Algorithm 2 Online Deployment of the Proposed Site-Specific Type-II Feedback

Input Trained probing codebook and inference mapping $\mathbf{B}^\star$ and $\Phi_{\theta^\star}$
Output Beamforming vector $\hat{\mathbf{w}}_p$
1: BS sweeps the learned probing codebook $\mathbf{B}^\star$ with SSB.
2: UE feeds back the measured RSRP vector $\mathbf{r}_{\mathbf{B}}$.
3: BS infers the dominant subspace $\mathbf{C}_p$.
4: BS transmits low-dimensional beamformed CSI-RS over the inferred subspace by (10).
5: UE feeds back the estimated effective channel.
6: BS reconstructs the DL CSI and applies the MRT beamformer.

4.1. Evaluation of the End-to-End Joint Design

Fig. 2 shows stable convergence of the proposed mini-batch SGD solver for all three representative (K, Q) settings, with most gains achieved within about 100 epochs before saturation. The small gap between training and validation curves indicates good generalization. The converged capture efficiency increases with both probing dimension and inferred subspace dimension, with $(16, 8)$ performing best, followed by $(16, 4)$ and $(8, 4)$.

Fig. 3 illustrates the overhead-performance tradeoff in the "asu_campus_3p5" scenario. Capture efficiency increases rapidly in the low-overhead regime but soon saturates, indicating clear diminishing returns. The Pareto frontier shows that high performance requires balancing K and Q: a small K limits fingerprint informativeness, while a small Q limits subspace expressiveness. As a result, increasing only one of them is inefficient once the other becomes the bottleneck. The effective spectral efficiency, shown by the color map, is therefore not maximized at the largest-overhead points, because the marginal capture gain is eventually outweighed by the overhead penalty.

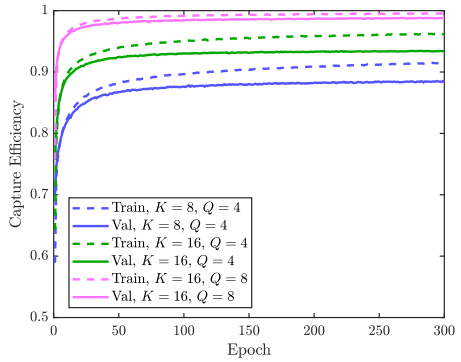


Figure 2. Convergence of the proposed end-to-end design

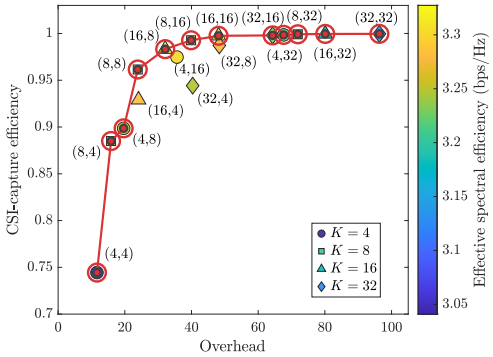


Figure 3. Pareto-optimal overhead-performance tradeoff

4.2. Evaluation of the Proposed Feedback Scheme

Table 1 and Fig. 4 compare the proposed scheme configured with $(K, Q) = (16, 8)$ against the Type-I, Type-II, and PSC baselines in the "boston5g_3p5" and "O1_28" scenarios under the same simulation settings. As expected, the proposed scheme attains a capture efficiency that is very close to, and in some cases even slightly higher than, that of Type-II. More importantly, despite this comparable raw efficiency, the proposed scheme consistently achieves the highest effective spectral efficiency over the entire SNR range. This result directly reflects the main advantage of the proposed SSI-enhanced framework: instead of pursuing marginal gains in instantaneous subspace optimality at the cost of full-dimensional CSI acquisition, it leverages site-specific inference to obtain Type-II-comparable subspace quality with much lower online overhead. Consequently, the proposed scheme converts Type-II-comparable capture efficiency into a strictly better system-level overhead-efficiency tradeoff. In other words, the key gain of the proposed framework comes not from universally outperforming Type-II in raw efficiency, but from matching Type-II in the relevant subspace-capture regime more efficiently.

Additional experimental results for the proposed site-specific Type-II feedback scheme are provided in Appendices G and H. These results further confirm the more favorable performance tradeoff achieved by the proposed design.

Table 1. CSI-capture efficiency of competing feedback schemes

Scheme	Boston	O1_28
Type-I	0.7017	0.7405
Type-II	**0.9951**	0.9992
PSC	0.2606	0.2510
Proposed	0.9860	**0.9999**

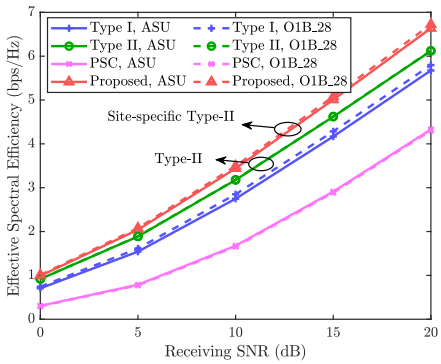


Figure 4. Effective spectral efficiency versus SNR

5. Conclusion

This paper develops a unified subspace-projection framework for limited-feedback beamforming and proposes a site-specific Type-II feedback scheme. By using offline learned SSI together with low-overhead RSRP fingerprints, the BS infers a UE-dependent dominant subspace in advance, so that the UE only needs to estimate and feed back low-dimensional effective CSI coefficients within that subspace. Simulation results show that the proposed scheme achieves a more favorable overhead-performance tradeoff than conventional feedback mechanisms, retaining Type-II-comparable subspace quality with substantially lower online CSI acquisition overhead and UE-side complexity. Future work may extend the framework to multi-user and multi-stream settings.

References

3GPP. NR; physical channels and modulation. Technical Specification TS 38.211, 3rd Generation Partnership Project (3GPP), 2018a.

3GPP. NR; physical layer procedures for data. Technical Specification TS 38.214, 3rd Generation Partnership Project (3GPP), 2018b.

3GPP. NR; physical layer measurements. Technical Specification TS 38.215, 3rd Generation Partnership Project (3GPP), 2018c.

Alkhateeb, A. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications. In *Proc. Inf. Theory Appl. Workshop (ITA)*, pp. 1–8, San Diego, CA, Feb. 2019.

Ayach, O. E., Rajagopal, S., Abu-Surra, S., Pi, Z., and Heath, R. W. Spatially sparse precoding in millimeter wave MIMO systems. *IEEE Trans. Wireless Commun.*, 13(3):1499–1513, Mar. 2014. doi: 10.1109/TWC.2014.011714.130846.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Dreifuerst, R. M. and Heath, R. W. Machine learning codebook design for initial access and CSI type-II feedback in sub-6-GHz 5G NR. *IEEE Trans. Wireless Commun.*, 23(6):6411–6424, Jun. 2024. doi: 10.1109/TWC.2023.3331313.

Dreifuerst, R. M. and Heath, R. W. Neural codebook design for MIMO network beam management. *IEEE Trans. Wireless Commun.*, 24(5):3909–3922, May 2025. doi: 10.1109/TWC.2025.3536290.

Fu, X., Le Ruyet, D., Visoz, R., Ramireddy, V., Grossmann, M., Landmann, M., and Quiroga, W. A tutorial on downlink precoder selection strategies for 3GPP MIMO codebooks. *IEEE Access*, 11:138897–138922, Dec. 2023. doi: 10.1109/ACCESS.2023.3338866.

Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Giordani, M., Polese, M., Roy, A., Castor, D., and Zorzi, M. A tutorial on beam management for 3GPP NR at mmwave frequencies. *IEEE Commun. Surv. Tutorials*, 21(1):173–196, Sep. 2019. doi: 10.1109/COMST.2018.2869411.

Hassibi, B. and Hochwald, B. M. How much training is needed in multiple-antenna wireless links? *IEEE Trans. Inf. Theory*, 49(4):951–963, Apr. 2003. doi: 10.1109/TIT.2003.809594.

Heath, R. W., González-Prelcic, N., Rangan, S., Roh, W., and Sayeed, A. M. An overview of signal processing techniques for millimeter wave mimo systems. *IEEE Journal of Selected Topics in Signal Processing*, 10(3):436–453, Apr. 2016. doi: 10.1109/JSTSP.2016.2523924.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Heng, Y. and Andrews, J. G. Grid-free MIMO beam alignment through site-specific deep learning. *IEEE Trans. Wireless Commun.*, 23(2):908–921, Feb. 2024. doi: 10.1109/TWC.2023.3283475.

Heng, Y., Mo, J., and Andrews, J. G. Learning site-specific probing beams for fast mmwave beam alignment. *IEEE Trans. Wireless Commun.*, 21(8):5785–5800, Jan. 2022. doi: 10.1109/TWC.2022.3143121.

Larsson, E. G., Edfors, O., Tufvesson, F., and Marzetta, T. L. Massive MIMO for next generation wireless systems. *IEEE Commun. Mag.*, 52(2):186–195, Feb. 2014. doi: 10.1109/MCOM.2014.6736761.

Love, D. J., Heath, R. W., N. Lau, V. K., Gesbert, D., Rao, B. D., and Andrews, M. An overview of limited feedback in wireless communication systems. *IEEE J. Sel. Areas Commun.*, 26(8):1341–1365, Oct. 2008. doi: 10.1109/JSAC.2008.081002.

Ning, X., Zhang, S., Xue, Y., Zheng, X., Shi, Q., and Chang, T.-H. Learning beams adaptive to the environment: An RSRP-based codebook design. In *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pp. 521–525, 2023. doi: 10.1109/SPAWC53906.2023.10304486.

Wang, Z., Zhou, Z., Zhao, C.-J., and Liu, Y. Generative site-specific beamforming for next-generation spatial intelligence. *arXiv preprint arXiv:2601.02301*, 2026.

Wu, D., Zeng, Y., Jin, S., and Zhang, R. Environment-aware hybrid beamforming by leveraging channel knowledge map. *IEEE Trans. Wireless Commun.*, 23(5):4990–5005, Oct. 2024. doi: 10.1109/TWC.2023.3323941.

Zhao, C.-J., Wang, Z., and Liu, Y. Generative site-specific beamforming via information-maximizing codebook. *arXiv preprint arXiv:2602.12552*, 2026a.

Zhao, C.-J., Wang, Z., Zhao, Z., and Liu, Y. Bridging standardized codebook and site-specific beamforming: A unified limited-feedback framework. *arXiv preprint arXiv:2604.14524*, 2026b.

A. Proof of Theorem 3.1

If $\mathbf{h} = \mathbf{0}$, then all three inequalities hold trivially. We therefore consider $\mathbf{h} \neq \mathbf{0}$.

Since $\mathbf{P}_p$ and $\mathbf{P}_Q^\star$ are orthogonal projectors, they are Hermitian and idempotent. Hence,

$$\|\mathbf{P}_p\mathbf{h}\|^2 = \mathbf{h}^H\mathbf{P}_p^H\mathbf{P}_p\mathbf{h} = \mathbf{h}^H\mathbf{P}_p\mathbf{h}, \tag{25}$$

and similarly

$$\|\mathbf{P}_Q^\star\mathbf{h}\|^2 = \mathbf{h}^H\mathbf{P}_Q^{\star H}\mathbf{P}_Q^\star\mathbf{h} = \mathbf{h}^H\mathbf{P}_Q^\star\mathbf{h}. \tag{26}$$

Subtracting the two terms gives

$$\|\mathbf{P}_p\mathbf{h}\|^2 - \|\mathbf{P}_Q^\star\mathbf{h}\|^2 = \mathbf{h}^H(\mathbf{P}_p - \mathbf{P}_Q^\star)\mathbf{h}. \tag{27}$$

Rearranging,

$$\|\mathbf{P}_p\mathbf{h}\|^2 = \|\mathbf{P}_Q^\star\mathbf{h}\|^2 + \mathbf{h}^H(\mathbf{P}_p - \mathbf{P}_Q^\star)\mathbf{h}. \tag{28}$$

Now define

$$\mathbf{A} \triangleq \mathbf{P}_p - \mathbf{P}_Q^\star. \tag{29}$$

Because both $\mathbf{P}_p$ and $\mathbf{P}_Q^\star$ are Hermitian, $\mathbf{A}$ is also Hermitian. For any Hermitian matrix $\mathbf{A}$, the Rayleigh quotient satisfies

$$-\|\mathbf{A}\|_2 \le \frac{\mathbf{h}^H\mathbf{A}\mathbf{h}}{\|\mathbf{h}\|^2} \le \|\mathbf{A}\|_2, \tag{30}$$

which implies

$$\mathbf{h}^H(\mathbf{P}_p - \mathbf{P}_Q^\star)\mathbf{h} \ge -\|\mathbf{P}_p - \mathbf{P}_Q^\star\|_2\,\|\mathbf{h}\|^2 = -\delta_p\|\mathbf{h}\|^2. \tag{31}$$

Substituting this bound into the previous equality yields

$$\|\mathbf{P}_p\mathbf{h}\|^2 \ge \|\mathbf{P}_Q^\star\mathbf{h}\|^2 - \delta_p\|\mathbf{h}\|^2. \tag{32}$$

Since $\|\mathbf{P}_p\mathbf{h}\|^2 \ge 0$, we further obtain

$$\|\mathbf{P}_p\mathbf{h}\|^2 \ge \left[\|\mathbf{P}_Q^\star\mathbf{h}\|^2 - \delta_p\|\mathbf{h}\|^2\right]_+, \tag{33}$$

which proves the first claim.

Next, the achievable rate of the proposed scheme is

$$R_p = \log_2\left(1 + \rho\|\mathbf{P}_p\mathbf{h}\|^2\right). \tag{34}$$

Because $\log_2(1+\rho x)$ is monotonically increasing for $x \ge 0$, applying the above lower bound gives

$$R_p \ge \log_2\left(1 + \rho\left[\|\mathbf{P}_Q^\star\mathbf{h}\|^2 - \delta_p\|\mathbf{h}\|^2\right]_+\right), \tag{35}$$

which proves the rate bound.

Finally, dividing the energy bound by $\|\mathbf{h}\|^2$ gives

$$\eta_p = \frac{\|\mathbf{P}_p\mathbf{h}\|^2}{\|\mathbf{h}\|^2} \ge \left[\frac{\|\mathbf{P}_Q^\star\mathbf{h}\|^2}{\|\mathbf{h}\|^2} - \delta_p\right]_+, \tag{36}$$

which completes the proof.

B. Detailed realization of the proposed end-to-end design

Since the probing encoder $\mathbf{B}$ is inherently trainable, it can be directly optimized in the learning process, subject only to the unit-power constraint $\|\mathbf{b}_k\|^2 = 1$ for each probing beam. The subspace decoder Φ_θ can be implemented by a multi-layer perceptron (MLP), which has a depth-D fully connected architecture with hidden width W, layer normalization (LN) (Ba et al., 2016), and Gaussian error linear unit (GELU) activation (Hendrycks & Gimpel, 2016). Let $\mathbf{f}^{(0)} = \bar{\mathbf{r}}_{\mathbf{B}}^{(n)}$. Then, for $d = 1, \ldots, D$, the hidden features are updated as

$$\mathbf{f}^{(d)} = \mathrm{GELU}\Big(\mathrm{LN}\Big(\mathbf{W}_d\mathbf{f}^{(d-1)} + \mathbf{b}_d\Big)\Big) \in \mathbb{R}^{W\times 1}, \tag{37}$$

where $\mathrm{GELU}(x) = x\Phi(x)$ and $\Phi(x)$ is the cumulative distribution function of the standard Gaussian distribution. The output layer yields

$$\mathbf{f}^{(\mathrm{out})} = \mathbf{W}_{\mathrm{out}}\mathbf{f}^{(D)} + \mathbf{b}_{\mathrm{out}} \in \mathbb{R}^{2N_tQ\times 1}. \tag{38}$$

The output vector is then reshaped into the real and imaginary parts of a raw complex Q-dimensional subspace representation $\widetilde{\mathbf{C}}_p$. To remove the scale ambiguity of the predicted basis, each column of $\widetilde{\mathbf{C}}_p$ is normalized to unit norm and orthogonalized before constructing the corresponding subspace projector.

C. Convergence and Complexity Analysis

Let $\boldsymbol{\xi} \triangleq (\mathbf{B}, \boldsymbol{\theta})$ collect all trainable variables. Suppose each mini-batch $\mathcal{B}_i$ is uniformly sampled from $\mathcal{H}_s$, and define the stochastic gradient

$$\mathbf{g}_i \triangleq \nabla_{\boldsymbol{\xi}}\mathcal{L}_{\mathcal{B}_i}(\boldsymbol{\xi}_i). \tag{39}$$

Assume that $\mathcal{L}_{\mathcal{H}_s}(\boldsymbol{\xi})$ is lower bounded and L_f-smooth, and that the stochastic gradient is an unbiased estimator of the full gradient with bounded variance that decreases with the mini-batch size, i.e.,

$$\mathbb{E}[\mathbf{g}_i] = \nabla_{\boldsymbol{\xi}}\mathcal{L}_{\mathcal{H}_s}(\boldsymbol{\xi}_i),\ \mathbb{E}\left[\|\mathbf{g}_i - \nabla_{\boldsymbol{\xi}}\mathcal{L}_{\mathcal{H}_s}(\boldsymbol{\xi}_i)\|^2\right] \le \frac{\sigma^2}{|\mathcal{B}_i|}. \tag{40}$$

Then, the mini-batch SGD update $\boldsymbol{\xi}_{i+1} = \boldsymbol{\xi}_i - \gamma_i\mathbf{g}_i$ with a standard diminishing or sufficiently small constant stepsize converges in expectation to a first-order stationary point. In particular, for the usual $\mathcal{O}(1/\sqrt{I})$ stepsize scaling, the average gradient norm obeys the standard nonconvex SGD rate

$$\frac{1}{I}\sum_{i=0}^{I-1}\mathbb{E}\left[\|\nabla_{\boldsymbol{\xi}}\mathcal{L}_{\mathcal{H}_s}(\boldsymbol{\xi}_i)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{I}} + \frac{\sigma^2}{|\mathcal{B}|\sqrt{I}}\right). \tag{41}$$

This result follows from the standard descent analysis for smooth nonconvex stochastic optimization and shows that

the proposed mini-batch joint solver approaches a stationary solution in expectation, with the gradient-variance term reduced by the mini-batch size.

We focus on the online deployment complexity, since the learning-based design is trained offline and its cost is amortized over the deployment site. During online deployment, the main complexity of the proposed scheme comes from BS-side subspace inference and UE-side effective channel estimation. For a depth-D MLP with hidden width W, the BS-side inference complexity is $\mathcal{O}\big(KW + (D-1)W^2 + 2N_tQW\big)$, while the UE-side LS estimation complexity is on the order of $\mathcal{O}(Q^2)$. By contrast, conventional feedback schemes place the dominant CSI-processing burden at the UE. Specifically, the UE-side complexity scales as $\mathcal{O}(N_t^2 O_D)$ for Type-I, $\mathcal{O}(N_t^2 Q O_D + Q^3)$ for Type-II with greedy beam selection, and $\mathcal{O}(N_t^2 + N_t \log N_t)$ for PSC, whereas the corresponding BS-side reconstruction complexity is only $\mathcal{O}(N_t)$, $\mathcal{O}(N_t Q)$, and $\mathcal{O}(N_t N_p)$, respectively.

D. Summary Comparison of Conventional and Proposed Limited-Feedback Schemes

Table 2 summarizes the conventional Type-I, Type-II, PSC, and proposed schemes using the subspace-capture notation above. Detailed derivations and discussions can be found in (Zhao et al., 2026b). It is shown that the proposed design retains Type-II-like subspace representation capability with only low-dimensional online refinement, thereby achieving a more favorable overhead-complexity-performance trade-off.

E. Simulation Settings

Key simulation parameters are summarized in Table 3, unless otherwise stated. By default, we use the "asu_campus_3p5" scenario, a 3.5 GHz outdoor deployment with rich multipath, moderate angular spread, and abundant UE samples. We also use the "O1_28", "O1B_28", and "boston5g_3p5" scenarios for additional validation, which correspond to a 28 GHz simple line-of-sight (LoS) outdoor scenario, a 28 GHz blocked non-LoS (NLoS) outdoor scenario, and a 3.5 GHz complex city scenario, respectively.

F. Ablation Study of the End-to-End Design

To validate whether the gain of the proposed end-to-end solver indeed comes from the jointly learned probing encoder, we compare the CSI-capture efficiency of the proposed end-to-end design with two baseline probing modes under the same MLP decoder and training pipeline: i) a fixed random probing codebook, and ii) a fixed DFT probing codebook. Table 4 reports the comparison across four representative DeepMIMO scenarios. The table shows that the proposed scheme consistently achieves the highest capture efficiency in all scenarios. Concretely, in the relatively easier "O1_28" and "O1B_28" scenarios, all methods already achieve high capture efficiencies, while the proposed design still preserves a clear advantage. In more challenging environments, especially "asu_campus_3p5", the gain becomes substantially larger, indicating that the learned probing codebook is more effective at generating informative fingerprints for downstream subspace inference when the propagation structure is more complex. Moreover, the relative ordering between the random and DFT baselines changes across scenarios, whereas the proposed design remains consistently superior, which further demonstrates its robustness and site-adaptive nature.

G. UE Performance Statistics

Fig. 5 shows the cumulative distribution function (CDF) of the effective spectral efficiency of four feedback schemes at an SNR of 10 dB in the "asu_campus_3p5" and "O1B_28" scenarios. The proposed scheme exhibits the rightmost CDF in both scenarios, indicating that it delivers higher effective spectral efficiency for the majority of users rather than only improving the average performance. This result is consistent with the previous average-rate comparison and further confirms that the gain of the proposed framework is population-wide. In particular, although the proposed scheme and Type-II often achieve comparable raw CSI-capture efficiency, the lower online overhead of the proposed framework shifts its effective-spectral-efficiency distribution consistently to the right of Type-II. Therefore, the benefit of the proposed method does not come from universally dominating Type-II in instantaneous subspace quality, but from realizing Type-II-comparable subspace quality with much lower overhead and converting this advantage into better UE-level effective spectral efficiency. By contrast, Type-I remains limited by its rank-one representation, which leads to a visibly broader and left-shifted distribution, while PSC performs the worst in both scenarios due to its insufficient ability to capture the dominant channel energy in the considered port-domain setting.

H. Physical Interpretation of the Inferred Subspace

Beyond rate and capture metrics, we further visualize the physical meaning of the learned subspace. Specifically, we select three representative test users in the "asu_campus_3p5" scenario and compare the angular responses of the proposed, Type-I, and Type-II subspace projectors. For each scheme $s \in \{p, \mathrm{I}, \mathrm{II}\}$, let $\mathbf{P}_s$ denote the corresponding subspace projector. Its angular response is evaluated over the azimuth

Table 2. Summary of Conventional and Proposed Limited-Feedback Schemes

Scheme	Subspace $\mathcal{U}$	Overhead T_o	UE Complexity	Efficiency η
Type-I	$\mathrm{span}(\mathbf{d}_{\mathbf{z}_\mathrm{I}})$	N_t+1	$\mathcal{O}(N_t^2 O_D)$	$\eta_\mathrm{I} \approx \frac{\lvert\alpha_{l_\mathrm{I}}\rvert^2 g_{l_\mathrm{I}}}{\sum_{l=1}^{L}\lvert\alpha_l\rvert^2}$
Type-II	$\mathrm{span}(\mathbf{D}_{\mathcal{S}_\mathrm{II}})$	N_t+2Q	$\mathcal{O}(N_t^2 Q O_D + Q^3)$	$\eta_\mathrm{II} \approx \frac{\sum_{l\in\mathcal{L}_{Q,\mathrm{II}}}\lvert\alpha_l\rvert^2 g_{l,\mathrm{II}}}{\sum_{l=1}^{L}\lvert\alpha_l\rvert^2}$
PSC	$\mathrm{span}(\mathbf{E}_{\mathcal{S}_\mathrm{PSC}})$	N_t+2N_p	$\mathcal{O}(N_t^2)$	$\eta_\mathrm{PSC} = \frac{\lVert\mathbf{P}_\mathrm{PSC}\mathbf{h}\rVert^2}{\lVert\mathbf{h}\rVert^2} = \frac{\sum_{i=1}^{N_p}\lvert\tilde{h}\rvert^2_{(i)}}{\lVert\mathbf{h}\rVert^2}$
Proposed	$\mathrm{span}(\mathbf{C}_p)$	$K+2Q$	$\mathcal{O}(Q^2)$	$\eta_\mathrm{P} = \frac{\lVert\mathbf{P}_\mathrm{P}\mathbf{h}\rVert^2}{\lVert\mathbf{h}\rVert^2}$

Table 3. Simulation settings

Parameter	**Description**	**Value**
f_c	Carrier frequency	3.5 GHz
BW	Bandwidth	10 MHz
P_t	Transmit power	40 dBm
K	Size of SSB codebook	8
Q	Dimension of the inferred subspace	4
N_t	Number of BS antennas	64
L_s	Number of SSB symbols	20
T_c	Coherence block length	1000
d	Antenna spacing	$\lambda/2$
S_n	Noise power spectrum density	-170 dBm/Hz
σ^2_{sh}	Log-variance of shadowing	1 dB
D	MLP depth	3
W	Hidden width	256

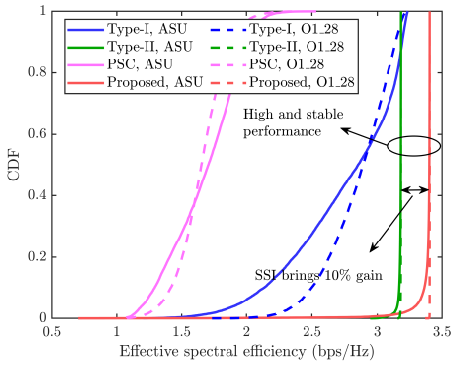


Figure 5. CDF of effective spectral efficiency

angle φ as

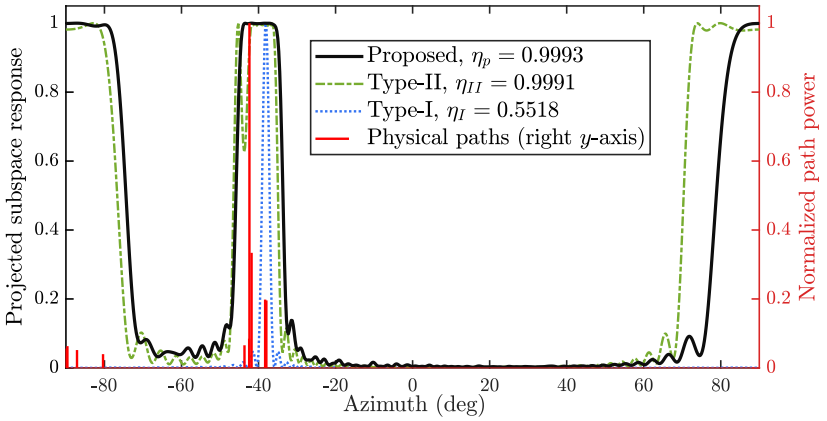
$$G_s(\varphi) = \mathbf{a}^H(\varphi)\mathbf{P}_s\mathbf{a}(\varphi). \tag{42}$$

For each user, the three responses and path powers are all normalized by their maximum value for visualization.

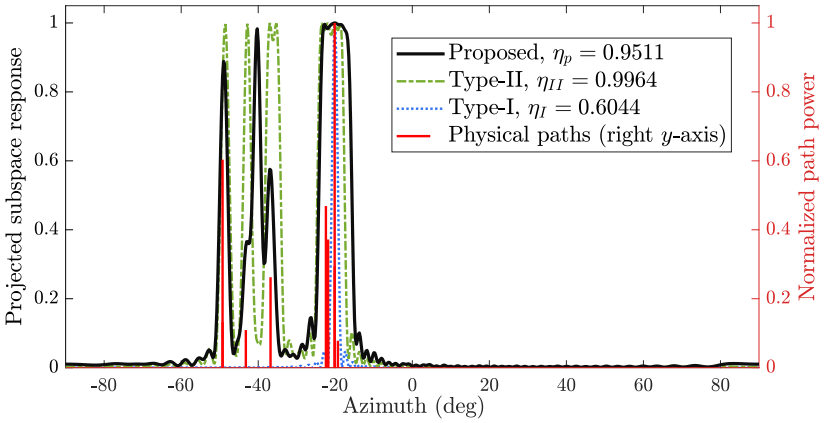
Fig. 6 visualizes these angular responses. First, the response of the proposed scheme is consistently concentrated around the dominant physical path clusters, which confirms that the inferred subspace is not an arbitrary latent representation but has a clear geometric interpretation in the angular domain. Second, the proposed response closely resembles that of Type-II in the dominant angular regions, which explains why the proposed scheme achieves CSI-capture efficiency that is close to Type-II in the quantitative results. Third, the Type-I response is much narrower and typically covers only one local direction, which explains its substantially lower capture efficiency. Overall, these figures provide physical evidence that the proposed SSI-enhanced framework succeeds in inferring a low-dimensional subspace aligned with the dominant propagation geometry, thereby achieving Type-II-comparable subspace quality with significantly reduced overhead.

Table 4. CSI-capture efficiency of different probing designs with the same inference network.

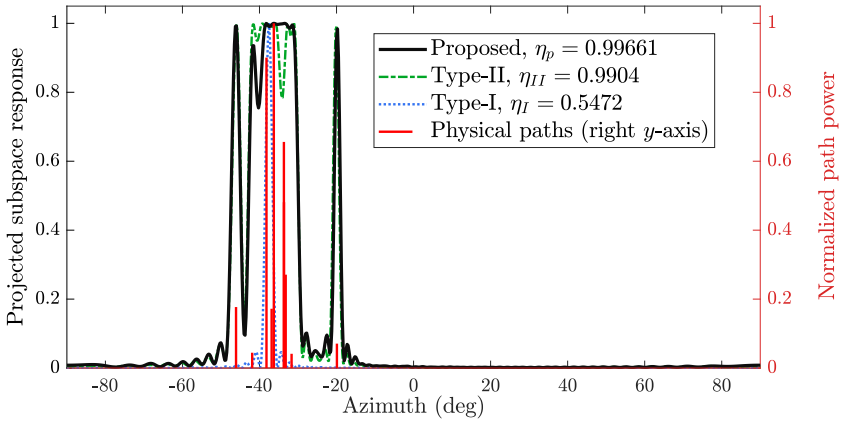
Scenario	Proposed	Random	DFT
O1_28	**0.99**	0.90	0.91
O1B_28	**0.99**	0.93	0.90
asu_campus_3p5	**0.89**	0.62	0.73
boston5g_3p5	**0.97**	0.81	0.86
Average	**0.96**	0.82	0.85



(a) User 1, max path power = -134.58 dB



(b) User 2, max path power = -105.52 dB



(c) User 3, max path power = -132.60 dB

Figure 6. Physical interpretation of the inferred subspace