Reliability-Preserving Mixed-Precision Quantization Scheduling for Asymmetric CSI Feedback Autoencoders

Hyunjae Park and Wan Choi
Department of Electrical and Computer Engineering, Seoul National University
Seoul 08826, South Korea
Email: {hyunjae.park, wanchoi}@snu.ac.kr

***Abstract*—In frequency-division duplex (FDD) massive multiple-input multiple-output (MIMO) systems, downlink CSI must be estimated at the user equipment (UE) and fed back to the base station (BS). In practice, UE-side inference is constrained by latency, energy, and compute, making lightweight encoders and low-precision execution unavoidable. Beyond reconstruction metrics such as normalized mean-squared error (NMSE), CSI distortion under UE constraints directly translates into transmission-rate degradation and outage events when the reconstructed CSI is used for precoding and link adaptation. This paper adopts a reliability-oriented view by distinguishing between (i) an *encoder-induced information floor* due to structural distortion and (ii) an *operational degradation* that depends on the CSI realization and arises from mixed-precision execution. To lower the information floor under strict UE budgets, we propose an asymmetric CSI feedback autoencoder that employs a lightweight state-space-model (SSM) encoder at the UE and a high-capacity decoder at the BS. Building on this architecture, we develop Reliability-Preserving Mixed-Precision Quantization Scheduling (RP-MPQ): an offline stage constructs a compact set of candidate mixed-precision policies via sensitivity-aware pruning followed by distributional (KL-divergence) refinement, and an online stage selects a policy per CSI realization to minimize a weighted reliability-violation cost under a long-term average UE compute budget. Simulation results demonstrate that the proposed framework improves the accuracy–complexity trade-off and substantially reduces rate-based outage events under the same average UE-side budget.**

***Index Terms*—CSI feedback, mixed-precision quantization, reliability, massive MIMO, UE-side inference**

I. Introduction

A. Background and Motivation

In frequency-division duplex (FDD) massive multiple-input multiple-output (MIMO) systems, channel state information (CSI) feedback is essential for downlink precoding and rate adaptation. Unlike time-division duplex systems, FDD operation does not permit channel reciprocity; therefore, the user equipment (UE) must explicitly estimate downlink CSI and convey a compressed representation to the base station (BS) over a limited feedback link. Recent deep-learning-based CSI feedback methods employ autoencoders to learn compact latent representations that can be quantized and transmitted efficiently [1]–[4].

A practical bottleneck lies in UE-side feasibility during inference. CSI compression must be executed under stringent latency and energy constraints, which often forces lightweight encoder architectures and low-precision execution. Under these constraints, CSI distortion is unavoidable, and its impact is not limited to reconstruction error: distorted CSI directly affects downlink transmission rates and increases outage events when used for precoding and link adaptation [5]. This motivates CSI feedback design that explicitly accounts for transmission-rate-based reliability, rather than optimizing reconstruction metrics alone.

B. Problem Reframing: Structural Floor vs. Operational Degradation

We adopt a hierarchical view of performance degradation in UE-constrained CSI feedback.

First, a lightweight UE-side encoder inevitably discards part of the channel information when compressing high-dimensional CSI into a low-dimensional latent vector. This induces an *encoder-induced information floor*: even with an ideal BS-side decoder, the distortion cannot be reduced below what is implied by the information preserved in the latent representation. Second, mixed-precision execution introduces an additional *operational degradation* on top of a fixed encoder–decoder structure. Unlike the information floor, this degradation is runtime-dependent and varies across CSI realizations and compute budgets.

This separation is a design abstraction: the two effects need not sum additively for every realization, but the distinction enables architectural design (to lower the information floor) and runtime precision control (to mitigate operational degradation) to be optimized in a coordinated manner under practical UE constraints.

C. Structural Perspective: Asymmetric CSI Feedback Architecture

In the delay–angular domain, CSI exhibits energy concentration around dominant propagation paths, while also containing non-negligible long-range components due to off-grid effects and finite array resolution. Encoders with hard locality (e.g., limited receptive fields) may truncate these long-range components and incur irreversible structural loss under strict UE budgets. In contrast, state-space-model (SSM) encoders aggregate inputs through soft memory with exponentially

decaying influence, allowing long-range dependencies to be attenuated rather than abruptly truncated. Recent work also highlights opportunities for SSM-style architectures (including Mamba) in wireless communications and networking [8], [9].

Motivated by this structural match and the inherent UE–BS computational asymmetry, we propose an asymmetric CSI feedback architecture that places a lightweight SSM-based encoder at the UE and a higher-capacity decoder at the BS. In our implementation, the UE employs a Mamba-style SSM encoder and the BS employs a Transformer-based decoder.

D. Operational Perspective: Reliability-Preserving Mixed-Precision Quantization

On top of the fixed asymmetric architecture, we develop a Reliability-Preserving Mixed-Precision Quantization Scheduling (RP-MPQ) framework that enables runtime, reliability-aware precision adaptation with low online complexity.

In the offline stage, we reduce the exponentially large mixed-precision space to a compact set of candidate policies by combining a sensitivity-based pruning with a distributional refinement that captures inter-block quantization effects. In the online stage, the UE selects a policy per CSI realization by minimizing a reliability-violation cost under a Lagrangian relaxation, while satisfying a long-term average UE-side compute budget. This design supports lightweight runtime adaptation without increasing the encoder model size or requiring expensive online optimization.

E. Contributions

The main contributions of this paper are:

- A reliability-oriented formulation for UE-constrained CSI feedback that distinguishes an encoder-induced information floor from runtime mixed-precision degradation.
- An asymmetric CSI feedback autoencoder architecture using a lightweight SSM-based UE encoder and a high-capacity BS decoder, structurally aligned with delay–angular CSI under strict UE constraints.
- RP-MPQ, a two-stage mixed-precision framework that constructs a compact candidate policy set offline and performs lightweight, per-sample reliability-aware policy selection online under a long-term UE compute budget.

F. Organization

Section II describes the system model and problem formulation. Section III presents the proposed asymmetric architecture and its structural motivation. Sections IV and V detail the offline policy construction and online reliability-aware selection in RP-MPQ, respectively. Section VI reports experimental results, followed by concluding remarks in Section VII.

II. System Model and Problem Formulation

A. System Model

We consider an FDD massive MIMO downlink system where a BS with N_t transmit antennas serves a UE with N_r receive antennas. Let $\mathbf{H}[n] \in \mathbb{C}^{N_r \times N_t}$ denote the downlink MIMO channel on the n-th OFDM subcarrier, $n = 0, \ldots, N_f - 1$.

Applying an N_f-point *unitary inverse DFT (IDFT)* across frequency yields the delay-domain channel taps

$$\mathbf{H}_d[\ell] = \frac{1}{\sqrt{N_f}} \sum_{n=0}^{N_f-1} \mathbf{H}[n] e^{j2\pi n\ell/N_f}, \quad \ell = 0, \ldots, N_f - 1. \tag{1}$$

Each tap is further mapped to the angular domain using unitary DFT matrices

$$\mathbf{X}[\ell] = \mathbf{F}_r \mathbf{H}_d[\ell] \mathbf{F}_t^{\mathsf{H}}, \tag{2}$$

where $\mathbf{F}_r \in \mathbb{C}^{N_r \times N_r}$ and $\mathbf{F}_t \in \mathbb{C}^{N_t \times N_t}$ are unitary DFT matrices.

Due to limited delay spread, most channel energy is concentrated in the first $N_a \le N_f$ delay taps. We retain

$$\mathbf{X}_a \triangleq \big[\mathbf{X}[0], \mathbf{X}[1], \ldots, \mathbf{X}[N_a - 1]\big]. \tag{3}$$

At the UE, an encoder $f_\theta(\cdot)$ maps $\mathbf{X}_a$ to a latent vector

$$\mathbf{z} = f_\theta(\mathbf{X}_a), \quad \mathbf{z} \in \mathbb{R}^D, \quad D \ll 2N_a N_r N_t. \tag{4}$$

(Complex CSI is represented in a real-valued format in the neural network implementation, e.g., by stacking real and imaginary parts; the notation above omits this bookkeeping for clarity.)

The compression ratio is defined as

$$\mathrm{CR} \triangleq \frac{D}{2N_a N_r N_t}. \tag{5}$$

The latent vector is quantized and transmitted over the feedback link,

$$\tilde{\mathbf{z}} = Q_{\mathrm{fb}}(\mathbf{z}), \tag{6}$$

and the BS reconstructs the truncated CSI as

$$\hat{\mathbf{X}}_a = g_\phi(\tilde{\mathbf{z}}). \tag{7}$$

B. UE Constraints and Distortion Sources

In practical deployments, CSI feedback is constrained by UE-side latency, energy, and compute limits, while the BS typically has substantially greater computational resources. This UE–BS asymmetry motivates architectures with a lightweight UE-side encoder and a higher-capacity BS-side decoder.

Under such operation, distortion can be usefully viewed as arising from two conceptually distinct sources:

- **Structural distortion (encoder-induced floor):** distortion due to information discarded when $\mathbf{X}_a$ is compressed into the latent representation $\mathbf{z}$. This creates an encoder-determined performance floor that cannot be removed by increasing the decoder capacity alone [1], [2].
- **Operational degradation (precision-induced):** additional distortion due to low-precision/mixed-precision execution of the UE-side encoder at inference time. This degradation is runtime-dependent and varies across channel realizations and compute budgets.

This distinction is a design abstraction: the two effects do not necessarily add linearly for every realization. However, separating them clarifies a hierarchical design pathway under UE constraints: (i) lower the encoder-induced floor via an appropriate asymmetric architecture and (ii) mitigate runtime degradation via reliability-aware mixed-precision control.

C. Design Perspective and Problem Statement

We study the joint design of:

1) a UE-feasible asymmetric encoder–decoder architecture that improves the accuracy–complexity trade-off by reducing the encoder-induced structural distortion; and
2) a runtime-operable mixed-precision scheduling mechanism for the UE-side encoder that suppresses transmission-rate-based reliability violations under a long-term average UE compute budget.

In the following section, we focus on the *structural* aspect and motivate an SSM-based lightweight encoder for delay–angular CSI under UE constraints. Subsequent sections develop the mixed-precision scheduling framework on top of the resulting asymmetric architecture.

III. Asymmetric CSI Feedback Architecture under UE Constraints

A. UE–BS Asymmetry and an Encoder-Induced Information Floor

Let $\mathbf{X}_a$ denote the truncated delay–angular CSI defined in (3) and define the latent random variable

$$\mathbf{Z} \triangleq f_\theta(\mathbf{X}_a). \tag{8}$$

Given a decoder $g_\phi(\cdot)$, consider the reconstruction MSE

$$\mathcal{L}(f_\theta, g_\phi) \triangleq \mathbb{E}\Big[\|\mathbf{X}_a - g_\phi(\mathbf{Z})\|_F^2\Big]. \tag{9}$$

By the orthogonal decomposition property of conditional expectation, (9) admits the exact decomposition

$$\begin{aligned}\mathcal{L}(f_\theta, g_\phi) = &\underbrace{\mathbb{E}\Big[\|\mathbf{X}_a - \mathbb{E}[\mathbf{X}_a \mid \mathbf{Z}]\|_F^2\Big]}_{\text{encoder-induced information floor}} \\ &+ \underbrace{\mathbb{E}\Big[\|\mathbb{E}[\mathbf{X}_a \mid \mathbf{Z}] - g_\phi(\mathbf{Z})\|_F^2\Big]}_{\text{decoder refinement error}}.\end{aligned} \tag{10}$$

The first term is the minimum achievable distortion under an ideal decoder and depends only on the information preserved in $\mathbf{Z}$. It therefore characterizes an encoder-induced information floor. The second term captures approximation error due to a non-ideal decoder and can be reduced by increasing BS-side decoder capacity. This decomposition highlights that, under UE constraints, architectural choices for the UE encoder fundamentally limit the attainable reconstruction performance.

B. Long-Tailed Locality of Delay–Angular Domain CSI

In the delay–angular domain, CSI energy is typically concentrated around a small number of dominant angular components associated with physical propagation paths. However, locality is often not strictly compact: off-grid angles represented by a finite DFT basis produce spectral leakage with slowly decaying sidelobes.

For a fixed delay tap, let $\{x_i\}$ denote angular-domain coefficients around a dominant index i_0. Their magnitudes often exhibit qualitative long-tail decay of the form

$$|x_i| \lesssim \frac{c_0}{1+|i-i_0|}, \tag{11}$$

where $c_0 > 0$ absorbs path strength and normalization factors.

Although most energy is locally concentrated, distant angular coefficients can remain non-negligible. Consequently, encoders that impose hard locality may incur irreversible structural loss when operated under strict UE constraints.

C. Hard-Locality Encoding and Structural Truncation Loss

Encoders based on convolutional neural networks (CNNs) impose finite receptive fields and thus a hard-locality constraint. Let L denote an effective receptive-field radius. Components with $|i-i_0| > L$ are weakly represented or structurally excluded, depending on the architecture.

Under the long-tail behavior in (11), the residual energy beyond radius L satisfies the order-wise bound

$$E_{\text{hard}}(L) \triangleq \sum_{|i-i_0|>L} |x_i|^2 \lesssim \sum_{d>L} \frac{1}{(1+d)^2} = \mathcal{O}\big(L^{-1}\big), \quad d \triangleq |i-i_0|. \tag{12}$$

This indicates that the residual energy decreases only polynomially with the receptive-field size. Reducing this truncation loss thus requires substantially increasing L (e.g., deeper/wider CNNs), which increases UE-side compute and latency.

D. Soft-Memory Encoding via State-Space Models

State-space-model (SSM) encoders provide soft memory by recursively aggregating inputs with exponentially decaying influence, rather than hard truncation.

A linear time-invariant SSM can be written as

$$\mathbf{s}_{k+1} = \mathbf{A}\mathbf{s}_k + \mathbf{B}\mathbf{x}_k, \quad \mathbf{y}_k = \mathbf{C}\mathbf{s}_k, \tag{13}$$

which yields

$$\mathbf{y}_k = \sum_{\tau \geq 0} \mathbf{C}\mathbf{A}^\tau \mathbf{B}\mathbf{x}_{k-\tau}. \tag{14}$$

If $\mathbf{A}$ is stable, then there exist constants $c_1 > 0$ and $\alpha > 0$ such that $\|\mathbf{A}^\tau\| \leq c_1 e^{-\alpha\tau}$ for all $\tau \geq 0$. Hence, contributions from inputs beyond horizon L are exponentially attenuated, leading to a residual bound of the form

$$E_{\text{soft}}(L) \lesssim \mathcal{O}\big(e^{-\alpha L}\big). \tag{15}$$

Importantly, this soft-memory behavior is achieved with a fixed state dimension, so the compute scales with the state size rather than the effective memory length. This makes SSM-based encoders structurally suitable for delay–angular CSI with long-tailed locality under strict UE budgets.

E. Structural Implications and Asymmetric Architecture

For delay–angular CSI exhibiting long-tailed locality, hard-locality encoders may incur structural truncation loss unless receptive fields are substantially expanded, whereas SSM-based encoders attenuate long-range components smoothly under bounded complexity. Combined with the encoder-induced information floor in (10), this implies that increasing BS-side decoder capacity alone cannot compensate for information discarded by a UE-constrained encoder.

Motivated by this observation, we adopt an asymmetric CSI feedback architecture that deploys a lightweight SSM-based encoder at the UE and a higher-capacity decoder at the BS. In our implementation, the UE employs a Mamba-style selective SSM encoder [8] and the BS employs a Transformer-based decoder (in the spirit of TransNet [3]). The next sections build on this asymmetric structural framework and develop a reliability-preserving mixed-precision scheduling mechanism for UE-side inference.

IV. RP-MPQ: Offline Policy Set Construction

This section presents the offline stage of RP-MPQ. A *policy* specifies a mixed-precision configuration applied to the *UE-side encoder* during inference, while the feedback-link quantization $Q_{\mathrm{fb}}(\cdot)$ in (6) is kept fixed. Building on the asymmetric architecture in Section III, the objective of the offline stage is to compress an exponentially large mixed-precision space into a compact candidate set that supports lightweight, reliability-aware online selection.

A. Mixed-Precision Policy Space

Let the UE-side encoder consist of M quantizable blocks (e.g., attention/SSM/MLP blocks or their sub-block partitions). A mixed-precision policy is defined as a block-wise assignment of bit-widths

$$\pi \triangleq (b_1, b_2, \ldots, b_M), \tag{16}$$

where $b_m \in \mathcal{B}$ and $\mathcal{B}$ denotes the discrete set of supported bit-widths. Let Π denote the induced policy space. Its cardinality grows exponentially as

$$|\Pi| = |\mathcal{B}|^M, \tag{17}$$

which makes exhaustive evaluation infeasible for moderate M.

B. Intra-Block Sensitivity and a Hessian-Based Surrogate

To enable efficient coarse pruning, we adopt a block-wise second-order surrogate commonly used in sensitivity-aware mixed-precision quantization (e.g., [6]). For block m quantized at bit-width b, we approximate the induced reconstruction-loss increase by

$$\Omega_m(b) \;\triangleq\; \mathrm{Tr}(\mathbf{H}_m)\, \left\| \Delta\boldsymbol{\theta}_m^{(b)} \right\|_2^2, \tag{18}$$

where $\mathbf{H}_m$ denotes the block-wise Hessian (or a curvature proxy) of the reconstruction loss with respect to the encoder parameters of block m, and $\Delta\boldsymbol{\theta}_m^{(b)}$ denotes the effective perturbation induced by quantization at bit-width b. In RP-MPQ, (18) is used only for *offline coarse policy pruning*.

C. ILP-Based Coarse Candidate Generation

Let $\kappa_m(b)$ denote the normalized compute cost of executing block m at bit-width b (e.g., normalized BOPs). For a given UE-side budget level $c \in \mathcal{C}$, we construct surrogate-optimal candidates via an integer linear program (ILP).

Using binary assignment variables $x_{m,b} \in \{0,1\}$, we formulate

$$\min_{\{x_{m,b}\}} \quad \sum_{m=1}^{M} \sum_{b\in\mathcal{B}} x_{m,b}\, \Omega_m(b) \tag{19}$$

$$\text{s.t.} \quad \sum_{b\in\mathcal{B}} x_{m,b} = 1, \quad \forall m, \tag{20}$$

$$\sum_{m=1}^{M} \sum_{b\in\mathcal{B}} x_{m,b}\, \kappa_m(b) \le c. \tag{21}$$

A feasible ILP solution induces a policy π by selecting b_m such that $x_{m,b_m} = 1$.

Rather than retaining only a single solution per budget, we preserve multiple near-optimal policies under each budget level. Denote the ILP-pruned candidate set under budget c as

$$\Pi_{\mathrm{ILP}}^{(c)} \triangleq \left\{ \pi^{(c,1)}, \ldots, \pi^{(c,K)} \right\}. \tag{22}$$

Retaining multiple candidates is important because policies with similar surrogate scores in (19) can exhibit heterogeneous aggregate distortion after quantization noise propagates through subsequent blocks.

D. Inter-Block Effects and a Distributional Refinement Criterion

Quantization noise introduced at one block propagates through subsequent blocks, producing inter-block interaction effects that are not fully captured by a block-separable surrogate. To refine candidates in a tractable manner, we adopt a distributional view of the *encoder output*.

Let $p_{\mathrm{enc}}^{(\pi)}(\mathbf{z})$ denote the empirical distribution of encoder outputs $\mathbf{z} = f_\theta^{(\pi)}(\mathbf{X}_a)$ induced by policy π when $\mathbf{X}_a$ is drawn from a calibration set (i.e., the data distribution used for offline evaluation). Let $\pi^{(0)}$ denote the full-precision reference policy. We measure the policy-induced output shift by

$$J(\pi) \triangleq D_{\mathrm{KL}}\Big(p_{\mathrm{enc}}^{(\pi)}(\mathbf{z}) \;\big\|\; p_{\mathrm{enc}}^{(0)}(\mathbf{z}) \Big), \tag{23}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. In practice, $p_{\mathrm{enc}}^{(\pi)}$ is approximated using a lightweight density approximation on encoder outputs (e.g., histogramming in a low-dimensional projection or a simple parametric fit), and $J(\pi)$ is used only for ranking policies within the ILP-pruned candidate set.

The metric $J(\pi)$ reflects *aggregate* quantization-induced shift in the latent representation, including inter-block propagation effects, and thus serves as a refinement criterion beyond the block-separable surrogate in (19).

E. Representative Policy Set Construction

For each budget level $c \in \mathcal{C}$, the ILP step yields a reduced candidate set $\Pi_{\mathrm{ILP}}^{(c)}$. We then select a representative policy by minimizing the distributional criterion:

$$\pi^{(c)} \triangleq \arg\min_{\pi \in \Pi_{\mathrm{ILP}}^{(c)}} J(\pi). \tag{24}$$

Collecting representatives across budgets produces the final candidate policy set

$$\Pi_{\mathcal{C}} \triangleq \left\{ \pi^{(c)} \mid c \in \mathcal{C} \right\}. \tag{25}$$

The resulting set $\Pi_{\mathcal{C}}$ is compact and spans heterogeneous encoder-output distortion profiles, enabling effective reliability-aware online selection in the next section.

F. Computational Advantage of the Two-Stage Design

Direct KL-based evaluation over the full policy space Π would incur exponential complexity:

$$\mathrm{Cost}_{\mathrm{full}} = |\mathcal{B}|^M \cdot \mathcal{O}(|\mathcal{D}| \cdot \mathrm{FLOPs}_{\mathrm{enc}}), \tag{26}$$

where $|\mathcal{D}|$ denotes the calibration dataset size and $\mathrm{FLOPs}_{\mathrm{enc}}$ denotes encoder inference complexity. In contrast, RP-MPQ evaluates (23) only on the ILP-pruned sets of size $K \ll |\mathcal{B}|^M$, yielding

$$\mathrm{Cost}_{\mathrm{two-stage}} = \mathcal{O}(\mathrm{poly}(M, |\mathcal{B}|)) + |\mathcal{C}| K \cdot \mathcal{O}(|\mathcal{D}| \cdot \mathrm{FLOPs}_{\mathrm{enc}}). \tag{27}$$

This makes offline policy set construction computationally practical while retaining distributional diversity across candidate operating points.

V. RP-MPQ: Online Reliability-Aware Policy Selection

This section describes the online policy selection mechanism of RP-MPQ. Given the offline-constructed candidate set $\Pi_{\mathcal{C}}$ in (25), the UE selects, for each CSI realization, a mixed-precision policy that preserves transmission-rate-based reliability while satisfying a long-term average UE compute budget.

A. Candidate Policies and UE-Side Cost Model

Each policy $\pi \in \Pi_{\mathcal{C}}$ is associated with a normalized UE-side cost

$$\kappa_\pi \in [0, 1], \tag{28}$$

computed offline (e.g., via normalized BOPs). For the t-th CSI realization, the UE selects a policy $\pi_t \in \Pi_{\mathcal{C}}$. The resulting sequence must satisfy a long-term average budget constraint $\bar{c}$.

For a policy $\pi = (b_1, \ldots, b_M)$, we define its normalized UE-side cost as

$$\kappa_\pi \triangleq \frac{\sum_{m=1}^{M} \kappa_m(b_m)}{\sum_{m=1}^{M} \kappa_m(16)}, \qquad \kappa_\pi \in [0, 1]. \tag{29}$$

optionally normalized by the INT16 baseline so that $\kappa_{\pi^{(0)}} = 1$.

B. Transmission-Rate-Based Reliability Metric

Let $r_t(\pi)$ denote the achievable downlink transmission rate when precoding and link adaptation are performed using the reconstructed CSI obtained under policy π. Let r_t^{ref} denote the corresponding rate under ideal CSI. We define an outage event as

$$r_t(\pi) < \gamma\, r_t^{\mathrm{ref}}, \tag{30}$$

where $\gamma \in (0, 1)$ specifies the target reliability ratio.

C. Reliability Violation Cost and Sparsity-Aware Weighting

To capture both outage occurrence and severity, we define the reliability violation cost

$$V_{t,\pi} \triangleq \mathbb{I}\big(r_t(\pi) < \gamma\, r_t^{\mathrm{ref}}\big) + \beta \frac{\max\big(0,\, \gamma r_t^{\mathrm{ref}} - r_t(\pi)\big)}{\gamma r_t^{\mathrm{ref}} + \epsilon}, \tag{31}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, $\beta \geq 0$ balances outage frequency and magnitude and $\epsilon > 0$ is a small constant for numerical stability.

CSI distortion sensitivity can vary across channel conditions. To emphasize reliability for realizations that are more sensitive to distortion, we introduce a sparsity-aware importance weight

$$w_t \triangleq 1 + \alpha s_t, \tag{32}$$

where $\alpha \geq 0$ controls the influence of sparsity and s_t denotes the delay–angular sparsity level of the t-th realization.

In this work, we quantify sparsity via Hoyer's measure. Let $\mathbf{h}_t \in \mathbb{R}^N$ denote a real-valued vectorization of the truncated delay–angular CSI (e.g., stacking real and imaginary parts), where $N \triangleq 2N_a N_r N_t$. Then

$$s_t \triangleq \frac{\sqrt{N} - \|\mathbf{h}_t\|_1 \,/\, \|\mathbf{h}_t\|_2}{\sqrt{N} - 1}, \tag{33}$$

where larger s_t indicates stronger sparsity.

D. Online Optimization and Decision Rule

The online policy selection objective is to minimize the expected weighted reliability violation under a long-term average compute constraint:

$$\min_{\{\pi_t \in \Pi_{\mathcal{C}}\}} \mathbb{E}[w_t\, V_{t,\pi_t}] \quad \text{s.t.} \quad \mathbb{E}[\kappa_{\pi_t}] \leq \bar{c}. \tag{34}$$

Introducing a Lagrange multiplier $\lambda \geq 0$ yields the relaxed objective

$$\min_{\{\pi_t\}} \mathbb{E}[w_t\, V_{t,\pi_t} + \lambda\, \kappa_{\pi_t}], \tag{35}$$

Since $\Pi_{\mathcal{C}}$ is finite, the online decision decomposes into a per-sample rule:

$$\pi_t^\star = \arg\min_{\pi \in \Pi_{\mathcal{C}}} (w_t\, V_{t,\pi} + \lambda\, \kappa_\pi). \tag{36}$$

E. Practical Surrogate and Budget Calibration

The exact $V_{t,\pi}$ in (31) depends on downlink rate computations and is not directly available at the UE at decision time. Therefore, we evaluate a lightweight surrogate of the form

$$V_{t,\pi} \approx \tilde{V}(\pi; \boldsymbol{\xi}_t), \tag{37}$$

where $\boldsymbol{\xi}_t$ denotes observable UE-side features (e.g., SNR estimate and sparsity s_t). The surrogate $\tilde{V}(\pi; \boldsymbol{\xi})$ is constructed offline by averaging the empirical violation cost over discretized feature bins, resulting in a lookup table used online.

The multiplier λ is calibrated offline (e.g., via bisection) such that the induced policy sequence satisfies the long-term average constraint in (34). Once λ is fixed, online selection requires only evaluating the scalar objective in (36) over the finite set $\Pi_{\mathcal{C}}$.

F. Online Complexity

Since $|\Pi_{\mathcal{C}}| = |\mathcal{C}| \ll |\mathcal{B}|^M$, the online complexity scales as $\mathcal{O}(|\mathcal{C}|)$ per CSI realization. Therefore, RP-MPQ enables reliability-aware mixed-precision adaptation with negligible overhead relative to encoder inference.

VI. Experimental Results

This section validates the proposed asymmetric CSI feedback framework with RP-MPQ along three dimensions: (i) structural efficiency of the asymmetric UE encoder, (ii) effectiveness of the offline mixed-precision policy set construction, and (iii) reliability-aware online adaptation under identical long-term UE-side compute budgets.

A. Experimental Setup

We consider an FDD massive MIMO downlink system with $N_t = 32$ transmit antennas at the BS and a single-antenna UE ($N_r = 1$). CSI is represented in the delay–angular domain with $N_f = 1024$ OFDM subcarriers, and only the first $N_a = 32$ delay taps are retained for feedback. The BS employs maximum ratio transmission (MRT) based on reconstructed CSI, and performance is evaluated at SNR levels of 10, 20, and 30 dB.

We use COST 2100 outdoor channel realizations [7] (carrier frequency $f_c = 5.3$ GHz) and additionally evaluate generalization on FR1 TDL-A/B/C profiles from 3GPP TR 38.901 [10]. The UE encoder is a Mamba-style selective SSM [8], and the BS decoder is Transformer-based. Training uses AdamW with learning rate 10^{-3} for 200 epochs. Mixed-precision quantization is applied to UE-side encoder weights with candidate bit-widths $\{16, 8, 4, 2\}$, while internal activations use 16-bit and the latent uses 8-bit quantization for feedback. Unless stated otherwise, complexity is reported as encoder FLOPs (full precision) or encoder BOPs (quantized inference), normalized consistently across methods.

B. Structural Efficiency of the Asymmetric Encoder

We first evaluate structural reconstruction performance without encoder weight quantization. Tables II and III report NMSE and computational complexity under CR $\in \{1/4, 1/8\}$.

TABLE I
Simulation Setup

Item	Setting
System	FDD downlink, $N_t = 32$, $N_r = 1$
CSI representation	Delay–angular, $N_f = 1024$, $N_a = 32$
SNR (dB)	10 / 20 / 30
Dataset	COST 2100 (outdoor, $f_c = 5.3$ GHz) [7]; FR1 TDL-A/B/C [10]
Model	Mamba (UE encoder) [8] + Transformer (BS decoder)
Compression ratio	CR $= 1/4, 1/8$
Training	200 epochs, batch size 1000, AdamW (lr 10^{-3})
Weights (UE encoder)	16 / 8 / 4 / 2 bit (mixed precision)
Activations	16-bit internal, 8-bit latent
Budget set $\mathcal{C}$	75–95% (step 0.5%), normalized to INT16 BOPs
Reliability target γ	0.99 / 0.98 / 0.95

TABLE II
Structural Performance (CR $= 1/4$)

Model	NMSE (dB)	Enc. FLOPs (M)	Total FLOPs (M)
CsiNet	-8.95	2.71	5.42
CsiNet+	-12.40	12.29	24.57
TransNet	-14.86	17.86	35.72
Mamba-Transformer AE	**-15.34**	**4.87**	22.74

C. Quantization Robustness

We evaluate uniform weight quantization under INT16/INT8/INT4 at CR $= 1/4$. Table IV reports NMSE and encoder BOPs.

D. Ablation: Offline RP-MPQ Refinement

We evaluate the effectiveness of KL-based refinement in the offline stage. Fig. 1 compares ILP-predicted ranking with KL-based measurement under identical BOP budgets. The KL-based refinement improves the agreement between surrogate ranking and the observed encoder-output distortion, reducing inconsistencies that arise from the block-independence assumption in the surrogate.

E. Online RP-MPQ: Reliability under Identical Budgets

We compare uniform precision, static mixed-precision (offline-only), and the proposed online RP-MPQ under identical *average* UE-side encoder BOP budgets. Reliability is evaluated via outage probability under targets $\gamma \in \{0.99, 0.98, 0.95\}$ at SNR levels of 10, 20, and 30 dB.

1) Transmission Rate and Outage Definition: For the t-th CSI realization, we construct an MRT precoder per subcarrier using a channel estimate $\mathbf{x}[n] \in \mathbb{C}^{N_t}$. Since we consider $N_r = 1$ in this section, we denote the downlink channel on subcarrier n by a vector $\mathbf{h}_t[n] \in \mathbb{C}^{N_t}$ (equivalently, $\mathbf{H}_t[n] \in \mathbb{C}^{1\times N_t}$). The achievable rate is defined as

$$r_t(\mathbf{x}) \triangleq \frac{1}{N_f} \sum_{n=0}^{N_f-1} \log_2\Big(1 + \rho \left|\mathbf{h}_t[n]^{\mathsf{H}} \mathbf{w}_{\mathbf{x}}[n]\right|^2\Big), \tag{38}$$

where $\mathbf{h}_t[n]$ is the true channel, ρ is the received SNR, and $\mathbf{w}_{\mathbf{x}}[n] \triangleq \mathbf{x}[n] / \|\mathbf{x}[n]\|_2$ is the MRT beamforming vector. Equal power allocation across subcarriers is assumed.

TABLE III
STRUCTURAL PERFORMANCE (CR $= 1/8$)

Model	NMSE (dB)	Enc. FLOPs (M)	Total FLOPs (M)
CsiNet	—	—	—
CsiNet+	—	—	—
TransNet	—	—	—
Mamba-Transformer AE	—	—	—

TABLE IV
UNIFORM QUANTIZATION PERFORMANCE (CR $= 1/4$)

Model	Prec.	NMSE (dB)	Enc. BOPs (M)
CsiNet	INT16	-8.95	277.87
	INT8	0.68	138.94
	INT4	17.35	69.47
Mamba-Transformer AE	INT16	-15.34	622.85
	INT8	-15.12	311.43
	INT4	0.04	155.71

Under policy π, let $\hat{\mathbf{h}}_{t,\pi}[n]$ denote the reconstructed channel used to form the MRT precoder. We define $r_t(\pi) \triangleq r_t(\hat{\mathbf{h}}_{t,\pi})$ and the reference rate under perfect CSI as $r_t^{\mathrm{ref}} \triangleq r_t(\mathbf{h}_t)$. An outage event is declared when

$$r_t(\pi) < \gamma\, r_t^{\mathrm{ref}}. \tag{39}$$

In the online implementation, the surrogate $\tilde{V}(\pi;\boldsymbol{\xi})$ in Section V is realized as an offline-constructed lookup table by averaging violation costs over discretized feature bins (e.g., SNR and sparsity).

2) Outage Comparison: Fig. 2 presents outage probabilities under identical average encoder BOP budgets. The proposed online RP-MPQ suppresses outage events more effectively than static policies without increasing average UE-side encoder complexity, highlighting the benefit of runtime reliability-aware policy selection.

F. Budget Consistency Validation

To verify that the calibrated multiplier λ enforces the long-term compute constraint, we report the target versus achieved average encoder budgets in Table V. The achieved budget closely matches the target across operating points.

G. Validation under FR1 TDL Profiles and UE Runtime

To evaluate structural generalization, we train on a mixed set of FR1 TDL-A/B/C realizations and evaluate separately on each profile without profile-specific tuning. Table VI reports profile-wise NMSE and encoder complexity.

We also report measured runtime on a commercial UE platform in Table VII. The online policy selection overhead is designed to be negligible compared to total encoder execution time, confirming practical deployability.

VII. CONCLUSION

This paper studied CSI feedback for FDD massive MIMO under UE-side inference constraints from a two-level perspective. We distinguished (i) an encoder-induced information floor, which is fundamentally limited by the UE-side representation, from (ii) additional operational degradation introduced by low-precision execution.

To reduce the structural floor under UE constraints, we proposed an asymmetric CSI feedback autoencoder that deploys a lightweight SSM-based encoder at the UE and a higher-capacity decoder at the BS, aligning architectural roles with UE–BS computational asymmetry and delay–angular CSI characteristics. Building on this fixed structural framework, we developed RP-MPQ, which separates offline policy set construction (sensitivity-based pruning and KL-based refinement) from lightweight online reliability-aware policy selection under a long-term average UE compute budget. Experimental results demonstrated an improved accuracy–complexity trade-off and reduced rate-based outage events under identical average UE-side budgets. Future work includes dynamic budget adaptation and broader validation across heterogeneous channel and device conditions.

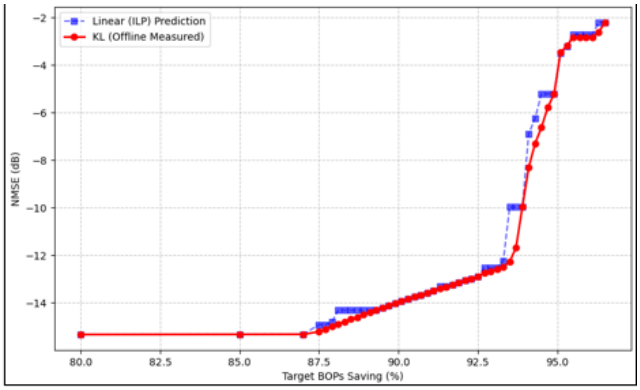

Fig. 1. ILP-based ranking versus KL-refined policy evaluation under identical computational budgets.

TABLE V
BUDGET CONSISTENCY UNDER ONLINE RP-MPQ

Target Avg. Budget	Achieved Avg. Budget	Deviation (%)
—	—	—
—	—	—
—	—	—

REFERENCES

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

[2] C.-K. Wen, S. Jin, and K.-K. Wong, "Deep learning-based CSI feedback for time-varying massive MIMO channels," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 881–884, Jun. 2019.

[3] Z. Liu, Y. Chen, and C.-K. Wen, "TransNet: Full-stack deep learning-based CSI feedback for massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 10, pp. 8261–8275, Oct. 2022.

[4] B. Cao, Y. Yang, P. Ran, D. He, and G. He, "ACCsiNet: Asymmetric convolution-based autoencoder framework for massive MIMO CSI feedback," *IEEE Communications Letters*, vol. 25, no. 12, pp. 3873–3877, Dec. 2021.

[5] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.

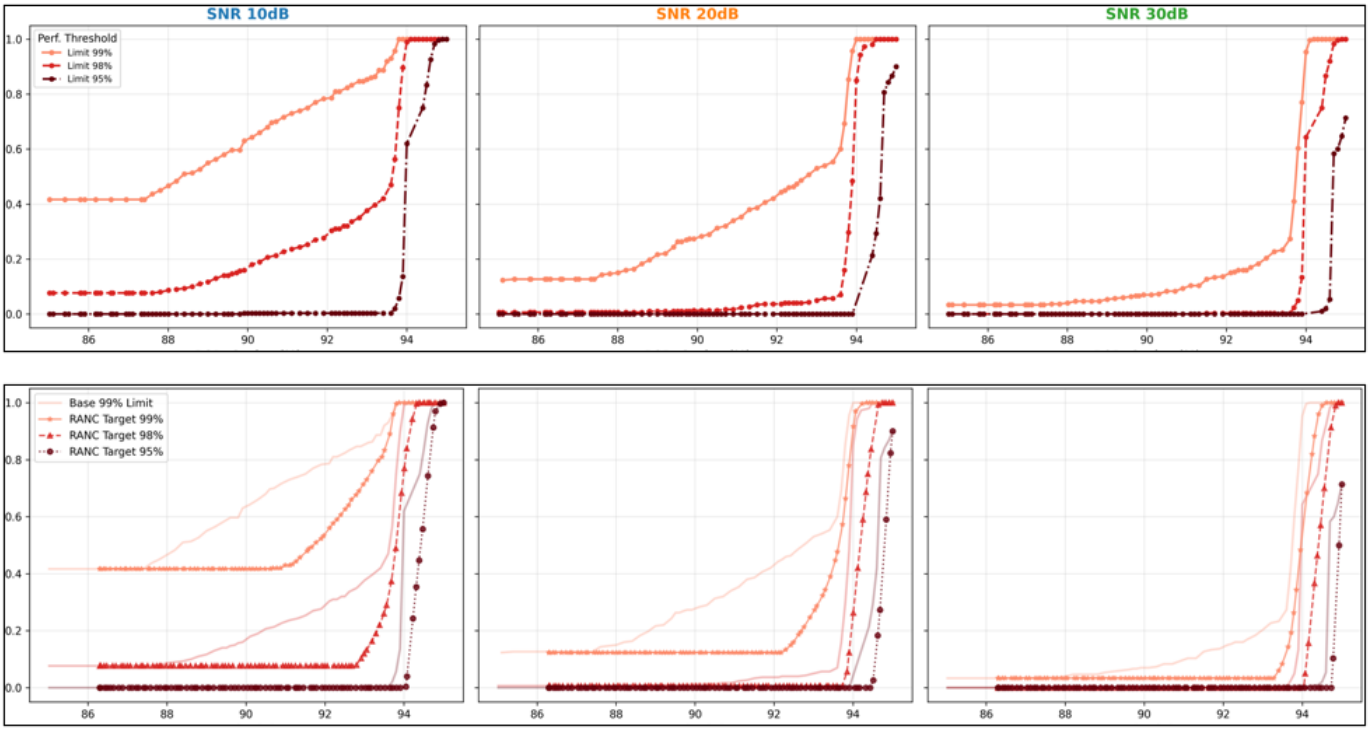

Fig. 2. Outage probability comparison under identical average UE-side computational budgets. Top: static policies. Bottom: online RP-MPQ.

TABLE VI
PROFILE-WISE GENERALIZATION UNDER FR1 TDL (CR=1/4)

Metric	TDL-A	TDL-B	TDL-C
NMSE (dB)	—	—	—
Encoder FLOPs (M)	—	—	—
Model size (MB)	—	—	—

TABLE VII
MEASURED RUNTIME OF ONLINE RP-MPQ ON UE

Component	Latency (ms)
Base encoder execution	—
Online precision selection	—
Total encoder latency	—

[6] H. Yao, Z. Dong, and K. Keutzer, "HAWQ-V3: Dyadic neural network quantization," in *Proc. ICML*, 2021.

[7] L. Liu *et al.*, "The COST 2100 MIMO channel model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, Dec. 2012.

[8] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint arXiv:2312.00752, 2023.

[9] R. Zhang, R. Zhang, Y. Lu, W. Chen, B. Ai, and D. Niyato, "Mamba for wireless communications and networking: Principles and opportunities," arXiv preprint arXiv:2508.00403, Aug. 2025, doi: 10.48550/arXiv.2508.00403.

[10] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP TR 38.901, V16.1.0, Dec. 2020.