CSIEVAL: A 6G STANDARD-ALIGNED BENCHMARK FOR AI-BASED CSI MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As 6G evolves toward an AI-native paradigm, channel state information (CSI) has become a key target for AI-enabled wireless communication, with growing research on CSI estimation, prediction, and compression feedback. However, existing evaluations lack systematic alignment with emerging 6G standards and use inconsistent protocols across studies, hindering fair comparison. Thus, we propose **CSIEval**, a standards-aligned benchmark for CSI processing that encompasses three representative tasks. CSIEval constructs and processes datasets in accordance with relevant communication protocols and evaluates models under a unified framework across four dimensions: performance, storage, efficiency, and generalization. The benchmark includes 8,000 CSI samples and more than 15 evaluation metrics, providing broad coverage of model capabilities. Experimental results demonstrate that CSIEval provides a comprehensive characterization of model behavior, enables fair comparisons across different models, and reveals the trade-offs among predictive accuracy, resource consumption, and generalization capability. Code and project resources are available at the anonymous project page: `https://anonymous.4open.science/r/csieval-8018`.

1 INTRODUCTION

Sixth-generation (6G) wireless communication systems are transitioning from traditional signal-processing algorithms to an AI-native paradigm supported by both domain knowledge and data-driven methods (O'Shea & Hoydis, 2017; He et al., 2019; Hoydis et al., 2021). This evolution drives the development of AI models for channel and resource control, as well as new 6G specifications for AI model training, testing, and deployment (International Telecommunication Union Radiocommunication Sector, 2023; 3rd Generation Partnership Project, 2025a). In massive multiple-input multiple-output (MIMO) systems for sixth-generation (6G) wireless communications, reliable Channel State Information (CSI) remains fundamental to link adaptation and efficient transmission. Consequently, CSI tasks constitute a critical foundational problem in intelligent 6G wireless communication scenarios.

AI-driven CSI tasks mainly consist of three categories: estimation, prediction, and compression feedback. CSI estimation recovers the true CSI from noisy observations (Ye et al., 2018; Neumann et al., 2018; Soltani et al., 2019; Honkala et al., 2021); CSI prediction leverages observed spatial or frequency information to infer CSI at unobserved antenna locations or frequency bands (Jiang & Schotten, 2020; Alrabeiah & Alkhateeb, 2019; Lin et al., 2021; Liu et al., 2025); and CSI compression feedback learns a compact representation of CSI and reduces transmission overhead through encoding and quantization (Wen et al., 2018; Wang et al., 2019; Guo et al., 2020; Mashhadi et al., 2021; Ravula & Jain, 2021; Shin et al., 2024). These tasks respectively examine the model's capabilities in observation recovery, structured completion, and efficient representation, and they differ in input conditions, error sources, and resource constraints. Meanwhile, corresponding AI models have been developed for all three tasks, making the evaluation of their accuracy and other capabilities critically important.

For the evaluation of CSI models, the data have expanded from standardized statistical channel models to large-scale, multi-scenario ray-tracing datasets and differentiable link-level simulation platforms (Liu et al., 2012; Jaeckel et al., 2014; Alkhateeb, 2019; Huangfu et al., 2022; Hoydis et al., 2022). In parallel, general ML benchmarking has emphasized reproducible workflows, standardized

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

system measurements, and explicit dataset and model documentation (Pineau et al., 2021; Mattson et al., 2020; Reddi et al., 2020; Gebru et al., 2021; Mitchell et al., 2019; Koh et al., 2021). However, existing works primarily offer high-level guidelines for data and evaluation requirements, and they have yet to establish directly actionable 6G evaluation standards. Specifically, at the data level, different studies lack unified conventions for data splitting, preprocessing pipelines, metric implementation, and result aggregation. At the metric level, evaluations tend to emphasize in-distribution accuracy while paying insufficient attention to computational and storage costs, and cross-scenario generalization. Therefore, there is an urgent need for an evaluation framework that covers CSI tasks—estimation, prediction, and compression feedback—and standardizes data processing as well as a multi-dimensional metric system, so as to promote the reproducibility, fair comparison, and practical value of CSI AI model results.

To address these issues, this paper proposes **CSIEval**, a standard-aligned evaluation benchmark for CSI models in 6G. CSIEval unifies the data, evaluation protocols, and experimental workflows across the three tasks, covering 8,000 CSI samples and more than 15 evaluation metrics. It assesses model performance, storage, efficiency, and generalization under a common evaluation standard, providing a reproducible and extensible basis for future model development. Furthermore, we conduct experiments using CSIEval to validate that this common standard can be instantiated across heterogeneous CSI tasks and model architectures. The main contributions of this paper are as follows:

- This paper constructs a CSI evaluation dataset based on 6G standard protocols, covering diverse scenarios and distribution shifts, and unifies the procedures for data generation, preprocessing, and splitting. This dataset supports both in-distribution and cross-scenario testing, providing consistent and reproducible evaluation inputs for different models.
- This paper establishes a multidimensional evaluation protocol covering performance, storage, efficiency, and generalization, with unified metric definitions, measurement methods, and result aggregation procedures.It also implements a complete evaluation pipeline spanning data loading, model integration, and result reporting.
- This paper conducts experiments on CSI estimation, prediction, and compression feedback models and systematically analyzes their performance and trade-offs across multiple capability dimensions. Furthermore, it develops an automated evaluation toolkit for reproducible evaluation.

2 Related Work

2.1 Evaluation Standards

Existing standards for AI-based CSI processing include competitions, benchmarks, and formal standardization. Competitions such as the ITU AI/ML in 5G Challenge and WAIC define fixed datasets, testing conditions, and ranking rules for specific tasks (International Telecommunication Union & North Carolina State University, 2020; Xiao et al., 2021; 2024). Benchmarks such as CeBed and CSI-4CAST improve reproducibility through public datasets, baselines, and evaluation pipelines (Feriani et al., 2023; Cheng et al., 2025). Broader ML studies also emphasize standardized measurement and transparent reporting (Pineau et al., 2021; Mattson et al., 2020; Reddi et al., 2020; Gebru et al., 2021; Mitchell et al., 2019). In addition, 3GPP TR 38.843 provides recommendations for AI/ML evaluation on the NR air interface (3rd Generation Partnership Project, 2025a). However, existing efforts remain task-specific or principle-oriented and do not define a common evaluation standard spanning multiple CSI tasks.

2.2 Evaluation Data

Existing CSI data can be broadly divided into model-driven and scenario-driven sources. Standardized channel models, including 3GPP TR 38.901, COST 2100, and QuaDRiGa, generate controllable and repeatable samples, while Sionna supports reproducible link-level simulation and neural receiver development (3rd Generation Partnership Project, 2024a; Liu et al., 2012; Jaeckel et al., 2014; Hoydis et al., 2022). In contrast, DeepMIMO and WAIR-D provide environment-specific ray-tracing data with greater scenario realism and distribution diversity (Alkhateeb, 2019; Huangfu

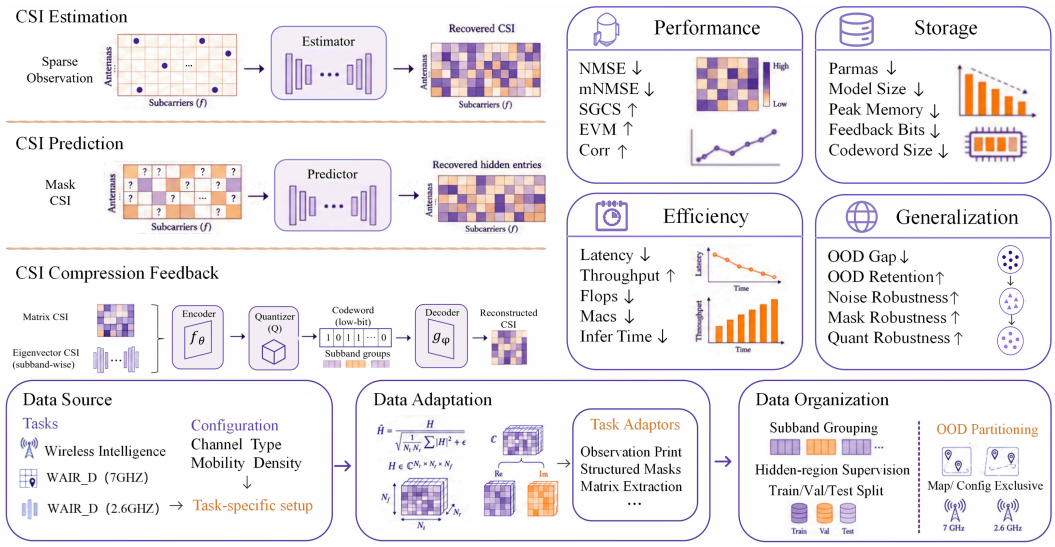


Figure 1: Overview of CSIEval, including task-specific data construction and four-dimensional model evaluation.

et al., 2022). However, these resources mainly provide raw channels, leaving task formulation, preprocessing, and data splitting to individual studies.

2.3 Evaluation Metrics

CSI evaluation has expanded from task accuracy to efficiency and distributional reliability. Estimation, prediction, and feedback studies commonly report MSE/NMSE and correlation-based metrics (Ye et al., 2018; Jiang & Schotten, 2020; Wen et al., 2018; Jin et al., 2022; Feriani et al., 2023), while lightweight-model studies additionally measure parameters, computation, and latency (Lu et al., 2020; Cao et al., 2021; Mattson et al., 2020; Reddi et al., 2020; Schwartz et al., 2020). Robustness benchmarks further encourage controlled corruption tests and explicit in- and out-of-distribution evaluation (Feriani et al., 2023; Cheng et al., 2025; Hendrycks & Dietterich, 2019; Koh et al., 2021). Nevertheless, metric definitions, execution settings, and aggregation procedures remain inconsistent across tasks, hindering fair comparison between models and studies.

3 CSIEval Benchmark

As illustrated in Fig. 1, CSIEval standardizes task-specific data construction and evaluates models across performance, storage, efficiency, and generalization.

3.1 CSI Data

CSIEval uses task-specific CSI sources under a common processing pipeline that standardizes complex-valued representation, normalization, sample organization, and split control. The radio concepts and evaluation axes follow relevant 6G/NR specifications where applicable (International Telecommunication Union Radiocommunication Sector, 2023; 3rd Generation Partnership Project, 2024a;b; 2025b); exact carrier profiles, tensor dimensions, split rules, masks, budgets, and metric selections remain disclosed benchmark choices. Detailed traceability is provided in Supplementary Secs. B and B.3.

3.1.1 CSI Estimation

For estimation, CSIEval adopts the Wireless Intelligence link-level dataset with varying channel, mobility, reference-signal, and observation conditions (Wireless Intelligence, 2021). To avoid target-

derived leakage, normalization is estimated only from *observable training inputs*. Let $\mathbf{O}_i$ denote the observable complex-valued input and Ω_i its valid observed entries:

$$s_{\text{train}} = \sqrt{\frac{\sum_{i\in\mathcal{D}_{\text{train}}} \|\mathcal{P}_{\Omega_i}(\mathbf{O}_i)\|_F^2}{\sum_{i\in\mathcal{D}_{\text{train}}} |\Omega_i|} + \epsilon}, \qquad \widetilde{\mathbf{O}}_i = \frac{\mathbf{O}_i}{s_{\text{train}}}, \qquad \widetilde{\mathbf{H}}_i = \frac{\mathbf{H}_i}{s_{\text{train}}}. \tag{1}$$

Here, $\mathcal{P}_{\Omega_i}$ retains only observable entries. For estimation, $\mathbf{O}_i$ contains pilot/reference-signal observations; prediction uses only masked observable CSI; and feedback uses the UE-side CSI available to the encoder. The training-derived s_{train} is frozen for validation and all test conditions, and the supervised target is scaled by the same constant. Real and imaginary components are represented as separate feature channels, while channel type, mobility, and observation configuration are retained as condition labels. Additional task-specific organization is given in Supplementary Sec. B.5.

3.1.2 CSI PREDICTION

For spatial–frequency prediction, CSIEval uses WAIR-D data at approximately 7 GHz and formulates prediction as structured missing-value recovery (Huangfu et al., 2022; Alrabeiah & Alkhateeb, 2019; Lin et al., 2021; Liu et al., 2025). With observation mask $\mathbf{M}$,

$$\mathbf{X}_{\text{obs}} = \mathbf{M} \odot \mathbf{X}, \qquad r_{\text{mask}} = 1 - \frac{\|\mathbf{M}\|_0}{|\mathbf{M}|}. \tag{2}$$

For spatial prediction, masks act along the antenna dimension, whereas frequency prediction masks subcarriers or subbands. CSIEval evaluates regular comb masks and non-uniform block and random patterns at declared observation ratios. Losses and metrics are computed only on the hidden region $\mathbf{1} - \mathbf{M}$, excluding already observed entries. Further mask-construction details are provided in Supplementary Sec. B.5.

3.1.3 CSI COMPRESSION FEEDBACK

For compression feedback, CSIEval uses WAIR-D at 2.6 GHz and supports full-matrix and subband-eigenvector representations (Huangfu et al., 2022; Wen et al., 2018; Guo et al., 2020; Mashhadi et al., 2021; Xiao et al., 2024). For the k-th subband,

$$\mathbf{H}_k^H \mathbf{H}_k \mathbf{w}_k = \lambda_k \mathbf{w}_k, \qquad \|\mathbf{w}_k\|_2 = 1. \tag{3}$$

With $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{C}^{N_t \times K}$, the real-valued eigenvector input is

$$\mathbf{x}_{\text{EV}} = \text{vec}(\text{concat}(\text{Re}(\mathbf{W}), \text{Im}(\mathbf{W}); \text{ axis} = 0)) \in \mathbb{R}^{2N_t K}. \tag{4}$$

Matrix feedback uses the standardized channel tensor directly, while the explicit concatenation above yields a $2N_t \times K$ real-valued eigenvector representation before vectorization. Adjacent subbands may be grouped when frequency correlation is modeled explicitly. To prevent leakage, all derivatives of the same source sample remain in one split; estimation separates channel, mobility, and observation conditions, whereas WAIR-D prediction and feedback use map-exclusive splits with unseen maps reserved for cross-scenario evaluation. Complete grouping and split rules are detailed in Supplementary Sec. B.5.

3.2 CSI METRICS

CSIEval evaluates models across four common dimensions—performance, storage, efficiency, and generalization—capturing output quality, resource requirements, execution cost, and robustness to controlled or distributional shifts. Task-specific metrics are therefore reported within the same four-dimensional schema, with more than 15 essential metrics standardized across the benchmark and additional implementation conventions given in Supplementary Sec. D.1.

3.2.1 PERFORMANCE

Performance metrics are task-specific. CSI estimation reports channel reconstruction error and, when the fixed link-level configuration is instantiated, BER; prediction reports reconstruction over hidden positions; and compression feedback reports reconstruction quality, SGCS, and correlation under a fixed feedback budget. NMSE measures relative numerical error,is calculated in the linear domain and reported in dB units while SGCS captures directional consistency in the complex-valued representation space.The definitions of other indicators can be found in the supplementary section D.2.

$$\mathrm{NMSE}_{\mathrm{dB}} = 10\log_{10}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\mathbf{H}_i - \hat{\mathbf{H}}_i\right\|_F^2}{\|\mathbf{H}_i\|_F^2 + \epsilon}\right], \tag{5}$$

where $\mathbf{H}_i$ and $\hat{\mathbf{H}}_i$ are the reference and model output. For dominant-eigenvector feedback, the arbitrary global complex phase is aligned before NMSE evaluation:

$$\phi_{i,k}^{\star} = \arg\left(\mathbf{w}_{i,k}^H \hat{\mathbf{w}}_{i,k}\right), \qquad \hat{\mathbf{w}}_{i,k}^{\mathrm{align}} = e^{-j\phi_{i,k}^{\star}}\hat{\mathbf{w}}_{i,k}. \tag{6}$$

Thus, phase-equivalent eigenvectors receive the same NMSE. For CSI prediction, errors are restricted to hidden entries:

$$e_i^{\mathrm{mask}} = \frac{\left\|\mathbf{M}_i^{\mathrm{hid}} \odot (\mathbf{H}_i - \hat{\mathbf{H}}_i)\right\|_F^2}{\left\|\mathbf{M}_i^{\mathrm{hid}} \odot \mathbf{H}_i\right\|_F^2 + \epsilon}. \tag{7}$$

with corresponding masked NMSE

$$\mathrm{mNMSE}_{\mathrm{dB}} = 10\log_{10}\left(\frac{1}{N}\sum_{i=1}^{N} e_i^{\mathrm{mask}}\right). \tag{8}$$

Directional consistency is measured by

$$\mathrm{SGCS} = \frac{1}{\sum_{i=1}^{N} J_i}\sum_{i=1}^{N}\sum_{j=1}^{J_i}\frac{\left|\mathbf{y}_{i,j}^H \hat{\mathbf{y}}_{i,j}\right|^2}{\|\mathbf{y}_{i,j}\|_2^2 \,\|\hat{\mathbf{y}}_{i,j}\|_2^2 + \epsilon}. \tag{9}$$

where J_i is the number of evaluation units in sample i. Estimation and matrix feedback vectorize the full channel, prediction uses hidden spatial or frequency slices, and eigenvector feedback treats each subband eigenvector as one unit. SGCS is invariant to global phase and therefore complements the phase-aligned NMSE definition. Phase alignment does not remove instability when the dominant eigenvalue is repeated or nearly repeated; accordingly, eigenvector NMSE is interpreted jointly with SGCS rather than as a stand-alone ranking criterion. Additional implementation conventions are given in Supplementary Sec. D.1.

3.2.2 STORAGE

Storage is characterized jointly by trainable parameter count, model file size, and peak inference memory (Reddi et al., 2020; Mattson et al., 2020; Schwartz et al., 2020). Parameter count describes learned model scale, file size additionally reflects parameter precision, and peak memory includes parameters, activations, and runtime buffers. Compression feedback additionally evaluates the quantized feedback payload:

$$\mathbf{z} = f_\theta(\mathbf{x}), \qquad \mathbf{q} = Q_b(\mathbf{z}), \qquad \widehat{\mathbf{x}} = g_\phi\left(Q_b^{-1}(\mathbf{q})\right), \tag{10}$$

$$B_{\mathrm{fb}} = Mb + B_{\mathrm{side}}, \tag{11}$$

where B_{side} denotes any transmitted side information required to decode the latent representation. A fixed-budget comparison is valid only when the quantizer rule, latent dimension M, precision b, and side-information accounting are declared and matched across rows; model-specific metadata cannot be omitted from B_{side}. The exact quantizer/calibration used to produce Table 1 must therefore be reported with the experimental configuration (Supplementary Sec. C).

3.2.3 Efficiency

CSIEval reports MACs/FLOPs together with measured latency and throughput under a fixed runtime protocol (Mattson et al., 2020; Reddi et al., 2020):

$$\mathrm{Throughput} = \frac{B}{T_B}, \tag{12}$$

where B is the batch size and T_B is the corresponding execution time. For compression feedback, encoder and decoder computation, latency, and memory are reported separately when the architecture exposes both components.

3.2.4 Generalization

Each task declares a baseline condition z_0 and one or more controlled or distribution-shifted conditions z: estimation varies noise, observation density, and mobility; prediction varies mask ratio and pattern; and feedback varies quantization precision, feedback budget, input noise, and deployment domain (Hendrycks & Dietterich, 2019; Koh et al., 2021).

$$D_E(z; z_0) = E_{\mathrm{dB}}(z) - E_{\mathrm{dB}}(z_0) = 10 \log_{10} \frac{E_{\mathrm{lin}}(z) + \epsilon}{E_{\mathrm{lin}}(z_0) + \epsilon}, \tag{13}$$

where $D_E > 0$ denotes degradation for error metrics; $|D_E|$ may additionally be reported when only shift magnitude is needed. For score-type metrics such as SGCS,

$$R_S(z; z_0) = \frac{S(z)}{S(z_0) + \epsilon} \times 100\%. \tag{14}$$

These definitions provide a common reporting convention while preserving task-specific physical quantities. Accordingly, feedback reports signed noise-induced dB change and OOD score retention, prediction reports signed cross-mask and cross-ratio changes, and estimation additionally reports raw NMSE at a designated low-SNR stress point. Further evaluator conventions are provided in Supplementary Sec. D.1.

4 Experiments

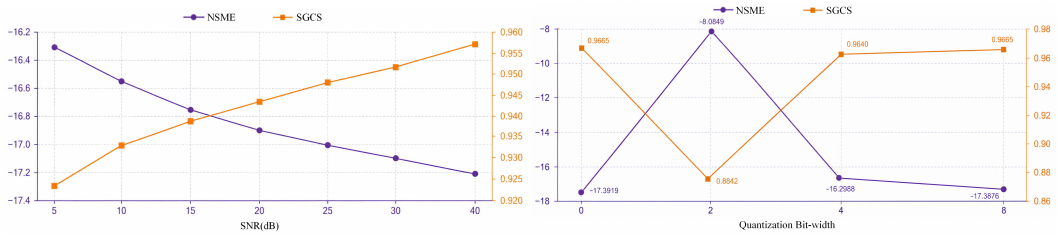
4.1 Experimental Setup

All experiments follow the common data-processing, split, model-standard, and runtime procedures in Section 3. Compression feedback uses WAIR-D at 2.6 GHz with $N_t = 32$, $K = 13$, $M = 104$, compression ratio $1/8$, and a common 4-bit operating point for the fixed-budget comparison. Robustness varies SNR from 5 to 40 dB and quantization from 2 to 8 bits. Prediction uses the approximately 7 GHz profile with 12.5%, 25%, and 50% observation ratios under comb, random, and block masks. Estimation varies SNR, mobility, reference-signal, and time-slot conditions. Full configuration and diagnostic conventions are given in Supplementary Secs. C and E.

Published paper values are generally not directly comparable because their data distributions, preprocessing, splits, observation patterns, feedback budgets, and metric implementations differ. When a published model can satisfy the standard directly, its architecture is interfaced with the benchmark; otherwise, a CSIEval implementation is adapted at the model/task interface and marked by *.

Table 1: Multidimensional evaluation of CSI feedback models at the declared common 4-bit operating point ($M = 104$). *, where present, denotes a CSIEval adaptation whose source task/data interface is not directly compatible with the benchmark.

	Performance			Storage			Efficiency			Generalization	
Model	**NMSE (dB) ↓**	**SGCS (%)↑**	**ρ (%)↑**	**Params (M) ↓**	**Size (MB) ↓**	**Peak Mem. (MB) ↓**	**Latency (ms) ↓**	**MACs (M) ↓**	**FLOPs (M) ↓**	**D_{noise} (dB) ↓**	**R_{OOD} (%) ↑**
CsiNet (Wen et al., 2018)	−6.420	38.6	54.2	**0.14**	**0.53**	**82.6**	**0.50**	**0.14**	**0.28**	3.842	69.82
CsiNet+ (Guo et al., 2020)	−8.156	54.8	62.3	1.05	4.01	96.4	0.87	2.13	4.26	3.276	75.64
CRNet (Lu et al., 2020)	−8.827	55.9	61.6	0.53	2.02	94.8	0.87	1.72	3.44	2.914	79.91
CLNet (Ji & Li, 2021)	−8.550	57.1	63.0	0.53	2.02	91.7	1.11	1.41	2.82	2.701	73.40
Attn-CSI (Cai et al., 2019)	−9.350	60.4	68.6	0.52	1.98	103.5	1.19	1.31	2.62	2.348	84.17
TransNet (Cui et al., 2022)	−10.996	61.9	72.0	1.45	5.53	136.2	1.78	15.60	31.20	1.786	87.63
EVCsiNet-CNN (Xiao et al., 2024)	−6.280	70.9	82.1	0.53	2.02	188.7	1.29	123.90	247.80	2.126	88.92
EVCsiNet-Transformer (Xiao et al., 2024)	**−16.30**	**96.1**	**98.1**	2.16	8.25	154.9	5.15	6.93	13.86	**0.943**	**91.75**



(a) NMSE and SGCS versus SNR

(b) NMSE and SGCS versus quantization bit-width

Figure 3: SNR and quantization robustness of EVCsiNet-Transformer.

4.2 CSI Compression Feedback Results

Table 1 compares eight CSI compression feedback models spanning autoencoder, multiresolution, lightweight complex-valued, attention, and Transformer architectures (Wen et al., 2018; Guo et al., 2020; Lu et al., 2020; Ji & Li, 2021; Cai et al., 2019; Cui et al., 2022; Xiao et al., 2024). It jointly reports performance, storage, efficiency, and generalization under a fixed feedback budget, with model size estimated from FP32 parameters.

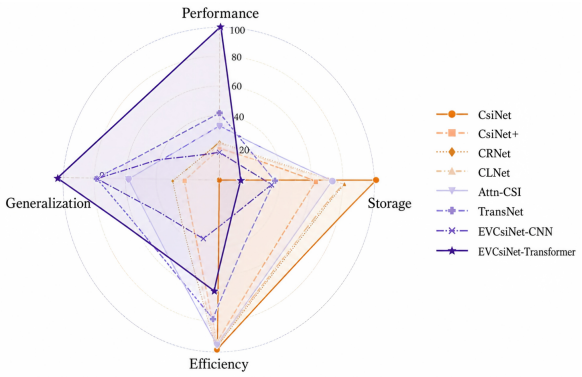


Figure 2: Radar chart of multidimensional performance for representative CSI compression feedback models.

Table 1 exposes distinct accuracy–resource operating points under the same declared feedback budget. In this run, EVCsiNet-Transformer reports lower NMSE and noise degradation together with higher SGCS, correlation, and OOD retention, but at higher latency than lightweight alternatives. CsiNet occupies a lower-resource operating point, while CRNet, CLNet, and Attn-CSI provide intermediate trade-offs. Figure 2 provides a compact qualitative view of these multidimensional trade-offs. These measurements illustrate multidimensional evaluation rather than statistical architecture ranking.

Figure 3 illustrates operating-point sensitivity for EVCsiNet-Transformer. As SNR increases from 5 to 40 dB, NMSE improves from -16.296 to -17.239 dB and SGCS from 92.09% to 95.87%. Four-bit quantization largely preserves directional similarity, whereas 2-bit quantization introduces clear degradation and 8-bit performance approaches the floating-point reference. Extended cross-domain and link-level diagnostics are provided in Supplementary Sec. E.1.1.

Table 2: Multidimensional evaluation of CSI prediction models.* denotes a model adapted because its original task or data interface is not directly compatible with the CSIEval protocol.

	Performance		Storage		Efficiency		Generalization	
Model	**NMSE (dB) ↓**	**SGCS (%) ↑**	**Params (M) ↓**	**Size (MB) ↓**	**Latency (ms) ↓**	**FLOPs (M) ↓**	**Cross-mask (dB) ↓**	**Cross-ratio (dB) ↓**
CsiNet* (Wen et al., 2018)	−4.15	74.2	0.38	1.45	**2.62**	1.87	0.05	**0.32**
WiFo-Little (Liu et al., 2025)	−5.49	81.7	1.60	6.10	4.23	3.70	0.04	0.69
STCMixerNet (He et al., 2025)	**−23.91**	**99.7**	**0.16**	**0.60**	2.99	**1.45**	**0.02**	0.41

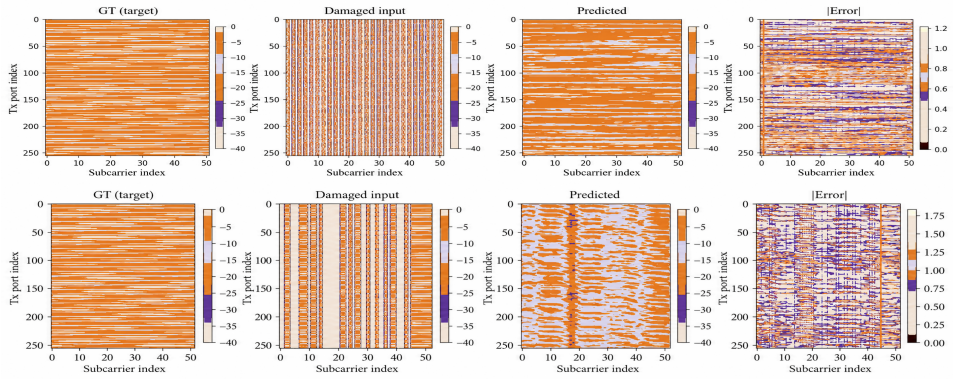


Figure 4: CSI reconstruction results under comb and random masking patterns. Each row shows, from left to right, the ground truth, damaged input, prediction, and absolute error. The top and bottom rows correspond to comb and random masking, respectively.

4.3 CSI Prediction Results

Table 2 summarizes the unified evaluation of CsiNet, WiFo, and STCMixer on synthetic CSI with $N_T = 256$, $N_R = 8$, and $N_F = 52$ (Wen et al., 2018; Liu et al., 2025; He et al., 2025). STCMixer achieves the best reconstruction performance, reaching -23.91 dB NMSE and 99.7% SGCS, while using only 0.16 M parameters and 0.60 MB storage. CsiNet has the lowest latency at 2.62 ms, whereas STCMixer shows the smallest cross-mask degradation (0.02 dB), indicating stronger robustness across mask patterns.

Figure 4 compares reconstruction results under comb and random masks. The comb mask uniformly samples subcarriers at fixed intervals, producing a regular periodic observation pattern that preserves the channel's slowly varying envelope. In contrast, the random mask creates a non-uniform distribution with both densely sampled and consecutively missing regions. These local gaps provide fewer constraints for reconstruction, making the fine channel structure more difficult to infer from neighboring observations.

4.4 CSI Estimation Results

Table 3 shows that ChannelNet provides a concrete published CNN anchor, while Transformer-PCE* provides a Transformer-based precedent adapted to the same output standard. In the current point estimates, Transformer-PCE* reports lower reconstruction error and BER at higher model and compute cost, whereas ChannelNet occupies a lower-resource operating point and CNN+LSTM* is intermediate.

Figure 5 examines two operating axes not captured by a single nominal operating point. In linear-domain NMSE, higher mobility yields larger estimation error, while additional time slots reduce error; CDLA-half reaches approximately 0.003 with four time slots. The result motivates reporting generalization across mobility and observation settings rather than at only one fixed condition. Complete SNR sweeps and deployment diagnostics remain in Supplementary Sec. E.3.

Table 3: Multidimensional evaluation of estimation models. * denotes a model adapted because its original task or data interface is not directly compatible with the CSIEval protocol. † BER is evaluated using the same fixed link-level chain for every model.

	Performance		Storage		Efficiency		Generalization
Model	**NMSE (dB) ↓**	**BER† (%) ↓**	**Params (M) ↓**	**Size (MB) ↓**	**Latency (ms) ↓**	**FLOPs (M) ↓**	**−10 dB NMSE (dB) ↓**
ChannelNet (Soltani et al., 2019)	−19.86	15.88	**1.05**	**4.01**	**3.70**	**2.85**	−1.28
RNN* (Gizzini & Chafii, 2024)	−11.50	28.10	1.80	6.84	13.80	4.12	−0.84
LSTM* (Pan et al., 2021)	−20.50	14.20	3.20	12.16	4.20	5.67	−8.50
GNN* (Ye et al., 2025)	−16.80	19.40	4.50	17.10	8.10	7.23	−4.65
GAN* (Balevi & Andrews, 2021)	−22.10	12.80	7.30	27.74	6.70	9.84	−9.80
CNN+LSTM* (Rapudu & Oyerinde, 2025)	−24.30	11.40	5.90	22.42	5.80	8.52	−10.72
CNN+GRU* (Hou et al., 2023)	−21.70	13.90	4.80	18.24	4.90	6.91	−8.94
Transformer-PCE* (Ju et al., 2024)	**−26.80**	**9.10**	12.40	47.10	7.20	15.60	**−12.10**

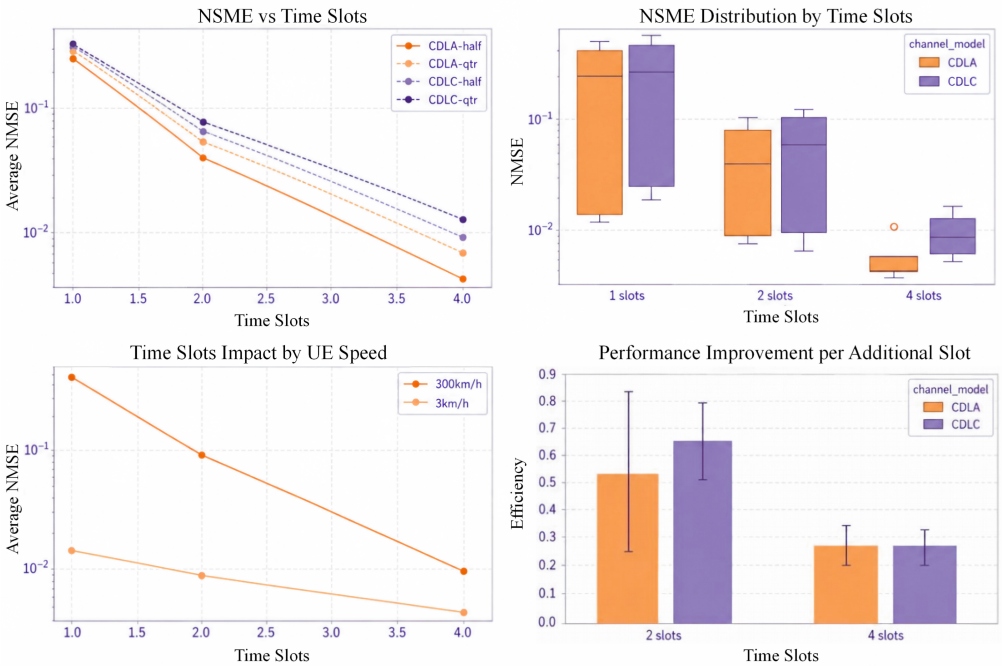


Figure 5: Estimation performance under different mobile speeds and time-slot configurations; plotted NMSE values are in the linear domain.

5 Conclusion

We presented CSIEval, a standard-aligned evaluation benchmark for CSI estimation, prediction, and compression feedback. Its central contribution is a common standard for leakage-aware data construction, splitting, model I/O, runtime measurement, and multidimensional reporting. Models are re-evaluated under this standard rather than compared through numerical results produced by incompatible source-paper protocols. Experiments show that the standard can characterize heterogeneous model architectures and expose accuracy–resource–robustness trade-offs while retaining task-specific physical metrics. Standard-derived concepts and CSIEval-specific numerical choices are explicitly separated through the traceability analysis.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and polishing the manuscript. This assistance was limited to improving grammar and word choice, refining sentence structure, and enhancing paragraph transitions. Generative AI tools were not used to generate the core research ideas, methodology, experimental conclusions, or scientific claims of this work. We reviewed all AI-assisted content to ensure that it accurately reflects the intended technical meaning and is consistent with the supporting evidence. We take full responsibility for the final content of this work, including any text, claims, or other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

CSIEval is specified through explicit data-processing, split, normalization, model-interface, runtime-measurement, and metric standards. The appendix provides standards-to-design traceability, task construction, model provenance, configuration requirements, measured supplementary metrics, and complementary diagnostics. For adapted rows, the source architectural precedent is separated from the exact CSIEval instantiation so that reported values are not confused with source-paper reproductions. To support reproducibility during double-blind review, an anonymized repository containing the executable CSIEval implementation and accompanying artifacts is available at `https://anonymous.4open.science/r/csieval-8018`. Identifying maintainer information and non-anonymous release endpoints remain withheld until the camera-ready version.

REFERENCES

3rd Generation Partnership Project. Study on channel model for frequencies from 0.5 to 100 GHz. Technical Report 3GPP TR 38.901 V18.0.0, 3rd Generation Partnership Project, 2024a.

3rd Generation Partnership Project. NR; physical channels and modulation. Technical Report 3GPP TS 38.211, 3rd Generation Partnership Project, 2024b.

3rd Generation Partnership Project. Study on artificial intelligence (AI)/machine learning (ML) for NR air interface. Technical Report 3GPP TR 38.843 V19.0.0, 3rd Generation Partnership Project, 2025a.

3rd Generation Partnership Project. NR; physical layer procedures for data. Technical Report 3GPP TS 38.214, 3rd Generation Partnership Project, 2025b.

Ahmed Alkhateeb. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications, 2019.

Muhammad Alrabeiah and Ahmed Alkhateeb. Deep learning for TDD and FDD massive MIMO: Mapping channels in space and frequency. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 1465–1470, 2019. doi: 10.1109/IEEECONF44664.2019.9048929.

Eren Balevi and Jeffrey G. Andrews. Wideband channel estimation with a generative adversarial network. *IEEE Transactions on Wireless Communications*, 20(5):3049–3060, 2021. doi: 10.1109/TWC.2020.3047100.

Qiuyu Cai, Chao Dong, and Kai Niu. Attention model for massive MIMO CSI compression feedback and recovery. In *Proceedings of the IEEE Wireless Communications and Networking Conference*, pp. 1–5, 2019. doi: 10.1109/WCNC.2019.8885897.

Zheng Cao, Wan-Ting Shih, Jiajia Guo, Chao-Kai Wen, and Shi Jin. Lightweight convolutional neural networks for CSI feedback in massive MIMO. *IEEE Communications Letters*, 25(8): 2624–2628, 2021. doi: 10.1109/LCOMM.2021.3076504.

Sikai Cheng, Reza Zandehshahvar, Haoruo Zhao, Daniel A. Garcia-Ulloa, Alejandro Villena-Rodriguez, Carles Navarro Manchón, and Pascal Van Hentenryck. CSI-4CAST: A hybrid deep learning model for CSI prediction with comprehensive robustness and generalization testing, 2025.

Yaodong Cui, Aihuang Guo, and Chunlin Song. TransNet: Full attention network for CSI feedback in FDD massive MIMO system. *IEEE Wireless Communications Letters*, 11(5):903–907, 2022. doi: 10.1109/LWC.2022.3149416.

Amal Feriani, Di Wu, Steve Liu, and Greg Dudek. CeBed: A benchmark for deep data-driven OFDM channel estimation, 2023.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723.

Abdul Karim Gizzini and Marwa Chafii. RNN-based channel estimation in doubly selective environments. *IEEE Transactions on Machine Learning in Communications and Networking*, 2:1–18, 2024. doi: 10.1109/TMLCN.2023.3332021.

Jiajia Guo, Chao-Kai Wen, Shi Jin, and Geoffrey Ye Li. Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis. *IEEE Transactions on Wireless Communications*, 19(4):2827–2840, 2020. doi: 10.1109/TWC.2020.2968430.

Hengtao He, Shi Jin, Chao-Kai Wen, Feifei Gao, Geoffrey Ye Li, and Zongben Xu. Model-driven deep learning for physical layer communications. *IEEE Wireless Communications*, 26(5):77–83, 2019. doi: 10.1109/MWC.2019.1800447.

Jilong He, Long Zhao, Chunhuan Wang, Hongrui Shen, and Zhan Peng. STCMixerNet for spatio-temporal CSI prediction in massive MIMO-OFDM systems. In *2025 11th International Conference on Computer and Communications (ICCC)*, pp. 1551–1556, 2025. doi: 10.1109/ICCC68654.2025.11437571.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Mikko Honkala, Dani Korpi, and Janne M. J. Huttunen. DeepRx: Fully convolutional deep learning receiver. *IEEE Transactions on Wireless Communications*, 20(6):3925–3940, 2021. doi: 10.1109/TWC.2021.3054520.

Jun Hou, Huaijie Liu, Yang Zhang, Wei Wang, and Jiaqian Wang. GRU-based deep learning channel estimation scheme for the IEEE 802.11p standard. *IEEE Wireless Communications Letters*, 12 (5):764–768, 2023. doi: 10.1109/LWC.2022.3187110.

Jakob Hoydis, Fayçal Ait Aoudia, Alvaro Valcarce, and Harish Viswanathan. Toward a 6G AI-native air interface. *IEEE Communications Magazine*, 59(5):76–81, 2021. doi: 10.1109/MCOM.001.2001187.

Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna: An open-source library for next-generation physical layer research, 2022.

Yourui Huangfu, Jian Wang, Shengchen Dai, Rong Li, Jun Wang, Chongwen Huang, and Zhaoyang Zhang. WAIR-D: Wireless AI research dataset, 2022.

International Telecommunication Union and North Carolina State University. ML5G-PHY [channel estimation]: Site-specific channel estimation with hybrid MIMO architectures. ITU Artificial Intelligence/Machine Learning in 5G Challenge, 2020. URL `https://research.ece.ncsu.edu/ai5gchallenge/`. Accessed: 2026-07-26.

International Telecommunication Union Radiocommunication Sector. Framework and overall objectives of the future development of IMT for 2030 and beyond. Technical Report Recommendation ITU-R M.2160-0, International Telecommunication Union, November 2023. URL `https://www.itu.int/rec/R-REC-M.2160-0-202311-I/en`.

Stephan Jaeckel, Leszek Raschkowski, Kai Börner, and Lars Thiele. QuaDRiGa: A 3-d multi-cell channel model with time evolution for enabling virtual field trials. *IEEE Transactions on Antennas and Propagation*, 62(6):3242–3256, 2014. doi: 10.1109/TAP.2014.2310220.

Sijie Ji and Mo Li. CLNet: Complex input lightweight neural network designed for massive MIMO CSI feedback. *IEEE Wireless Communications Letters*, 10(10):2318–2322, 2021. doi: 10.1109/LWC.2021.3100493.

Wei Jiang and Hans Dieter Schotten. Deep learning for fading channel prediction. *IEEE Open Journal of the Communications Society*, 1:320–332, 2020. doi: 10.1109/OJCOMS.2020.2982513.

Liqiang Jin, Qiuping Huang, Qiubin Gao, Yongqiang Fei, and Shaohui Sun. Scalable framework for deep learning-based CSI feedback, 2022.

Hyungyu Ju, Seokhyun Jeong, Byungju Lee, and Byonghyo Shim. Transformer-based predictive channel estimation for mmwave massive mimo systems. In *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*. IEEE, 2024. doi: 10.1109/VTC2024-Fall63153.2024.10758002.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021.

Bo Lin, Feifei Gao, Shun Zhang, Ting Zhou, and Ahmed Alkhateeb. Deep learning-based antenna selection and CSI extrapolation in massive MIMO systems. *IEEE Transactions on Wireless Communications*, 20(11):7669–7681, 2021. doi: 10.1109/TWC.2021.3087318.

Boxun Liu, Shijian Gao, Xuanyu Liu, Xiang Cheng, and Liuqing Yang. WiFo: Wireless foundation model for channel prediction. *Science China Information Sciences*, 68(6):162302, 2025. doi: 10.1007/s11432-025-4349-0.

Lingfeng Liu, Juho Poutanen, François Quitin, Katsuyuki Haneda, Fredrik Tufvesson, Philippe De Doncker, Pertti Vainikainen, and Claude Oestges. The COST 2100 MIMO channel model. *IEEE Wireless Communications*, 19(6):92–99, 2012. doi: 10.1109/MWC.2012.6393523.

Zhilin Lu, Jintao Wang, and Jian Song. Multi-resolution CSI feedback with deep learning in massive MIMO system. In *Proceedings of the IEEE International Conference on Communications*, pp. 1–6, 2020. doi: 10.1109/ICC40277.2020.9149229.

Mahdi Boloursaz Mashhadi, Qianqian Yang, and Deniz Gündüz. Distributed deep convolutional compression for massive MIMO CSI feedback. *IEEE Transactions on Wireless Communications*, 20(4):2621–2633, 2021. doi: 10.1109/TWC.2020.3043502.

Peter Mattson, Christine Cheng, Gregory Diamos, Cody Coleman, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debo Dutta, Udit Gupta, Kim Hazelwood, Andy Hock, Xinyuan Huang, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. MLPerf training benchmark. In *Proceedings of Machine Learning and Systems*, volume 2, 2020.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. doi: 10.1145/3287560.3287596.

David Neumann, Thomas Wiese, and Wolfgang Utschick. Learning the MMSE channel estimator. *IEEE Transactions on Signal Processing*, 66(11):2905–2917, 2018. doi: 10.1109/TSP.2018.2799164.

Timothy J. O'Shea and Jakob Hoydis. An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, 3(4):563–575, 2017. doi: 10.1109/TCCN.2017.2758370.

Jing Pan, Hangguan Shan, Rongpeng Li, Yingxiao Wu, Weihua Wu, and Tony Q. S. Quek. Channel estimation based on deep learning in vehicle-to-everything environments. *IEEE Communications Letters*, 25(6):1891–1895, 2021. doi: 10.1109/LCOMM.2021.3059922.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Thabang C. Rapudu and Olutayo O. Oyerinde. Channel estimation for wideband multi-RIS-assisted mmwave massive-MIMO OFDM system with beam squint effect. *IEEE Open Journal of the Communications Society*, 6:4804–4817, 2025. doi: 10.1109/OJCOMS.2025.3575947.

Sriram Ravula and Swayambhoo Jain. Deep autoencoder-based massive MIMO CSI feedback with quantization and entropy coding. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2021. doi: 10.1109/GLOBECOM46510.2021.9685912.

Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, et al. MLPerf inference benchmark. In *Proceedings of the 47th Annual International Symposium on Computer Architecture*, pp. 446–459, 2020. doi: 10.1109/ISCA45697.2020.00045.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020. doi: 10.1145/3381831.

Junyong Shin, Yujin Kang, and Yo-Seb Jeon. Vector quantization for deep-learning-based CSI feedback in massive MIMO systems. *IEEE Wireless Communications Letters*, 13(9):2382–2386, 2024. doi: 10.1109/LWC.2024.3415633.

Mehran Soltani, Vahid Pourahmadi, Ali Mirzaei, and Hamid Sheikhzadeh. Deep learning-based channel estimation. *IEEE Communications Letters*, 23(4):652–655, 2019. doi: 10.1109/LCOMM.2019.2898944.

Tianqi Wang, Chao-Kai Wen, Shi Jin, and Geoffrey Ye Li. Deep learning-based CSI feedback approach for time-varying massive MIMO channels. *IEEE Wireless Communications Letters*, 8 (2):416–419, 2019. doi: 10.1109/LWC.2018.2874264.

Chao-Kai Wen, Wan-Ting Shih, and Shi Jin. Deep learning for massive MIMO CSI feedback. *IEEE Wireless Communications Letters*, 7(5):748–751, 2018. doi: 10.1109/LWC.2018.2818160.

Wireless Intelligence. Channel estimation 1: AI-based link-level channel estimation dataset. Wireless Intelligence Dataset Platform, 2021. URL `https://wireless-intelligence.com/#/dataSet?id=2c92185c7e3f1aa4017e3f2c93e00001`. Accessed: 2026-07-26.

Han Xiao, Zhiqin Wang, Wenqiang Tian, Xiaofeng Liu, Wendong Liu, Shi Jin, Jia Shen, Zhi Zhang, and Ning Yang. AI enlightens wireless communication: Analyses, solutions and opportunities on CSI feedback. *China Communications*, 18(11):104–116, 2021. doi: 10.23919/JCC.2021.11.008.

Han Xiao, Zhiqin Wang, Dexin Li, Wenqiang Tian, Xiaofeng Liu, Wendong Liu, Shi Jin, Jia Shen, Zhi Zhang, and Ning Yang. AI enlightens wireless communication: A transformer backbone for CSI feedback. *China Communications*, 21(12):243–256, 2024. doi: 10.23919/JCC.ea.2022-0186.202401.

Hao Ye, Geoffrey Ye Li, and Biing-Hwang Fred Juang. Power of deep learning for channel estimation and signal detection in OFDM systems. *IEEE Wireless Communications Letters*, 7(1): 114–117, 2018. doi: 10.1109/LWC.2017.2757490.

Ming Ye, Xiao Liang, Cunhua Pan, Yinfei Xu, Ming Jiang, and Chunguo Li. Graph neural networks based channel estimation for mmwave massive MIMO systems. *IEEE Transactions on Vehicular Technology*, 74(12):19420–19435, 2025. doi: 10.1109/TVT.2025.3589434.

APPENDIX

APPENDIX CONTENTS

A Supplement Overview **15**

B Data Construction **15**

B.1 Source Data Organization 15

B.2 Standard-Aligned Radio Profiles 15

B.3 Standards-to-Benchmark Traceability 15

B.4 Channel Generation and Normalization 17

B.5 Task-Specific Construction Details 17

C Evaluation Configuration **18**

C.1 Model Integration Modes 18

C.2 Configuration Fields 18

C.3 Model standard and Validation 18

C.4 Reference-Model Provenance and Scope 19

C.5 Runtime Measurement Protocol 19

C.6 Execution and Reporting 20

C.7 Anonymous Artifact Availability and Publishing Plan 20

D Supplementary Metrics Used in Reported Experiments **21**

D.1 Core Metric Interpretation and Task-Specific Reporting 21

D.2 Link-Level and Reconstruction Metrics 21

D.3 Latency and System-Level Metrics 22

D.4 Measured Storage and Efficiency Diagnostics 22

D.5 Measured Adaptation and Robustness Diagnostics 23

E Additional Experimental Results **23**

E.1 CSI Compression Feedback 23

E.2 CSI Prediction 25

E.3 CSI Channel Estimation 25

A SUPPLEMENT OVERVIEW

This supplement provides standards-to-benchmark traceability, leakage-aware task construction, model provenance, evaluation configuration, metric conventions, and complementary results used to interpret the main paper, organized as detailed below. An anonymized implementation and accompanying artifacts are available at `https://anonymous.4open.science/r/csieval-8018`; identifying maintainer information and non-anonymous release endpoints remain withheld during double-blind review. Additional artifact and release details are specified below.

B DATA CONSTRUCTION

B.1 SOURCE DATA ORGANIZATION

The source data are generated from ray-tracing parameters associated with realistic urban maps. The collection contains two complementary deployment regimes. Part 1 consists of 10,000 diverse maps with sparse deployments, where each map contains five candidate base-station positions and 30 user-equipment positions. Part 2 contains 100 maps with dense deployments, where each map contains one base-station position and 10,000 user-equipment positions. The first regime provides broad environmental diversity, whereas the second provides dense spatial sampling within a fixed environment. Because the released representation contains ray parameters rather than only task-specific tensors, communication configurations can be regenerated while preserving the underlying geometry.

B.2 STANDARD-ALIGNED RADIO PROFILES

Two radio profiles are used to construct task-specific channel matrices. The approximately 7 GHz TDD profile supports spatial–frequency CSI prediction, while the 2.6 GHz FDD profile supports CSI compression feedback. Table 4 summarizes the principal settings.

Table 4: Radio profiles used for data generation. The table contains both standard-compatible settings and CSIEval-specific task-construction choices; their provenance is separated in Table 5.

Parameter	**Prediction**	**Compression feedback**
Duplex mode	TDD	FDD
Subcarrier spacing	30 kHz	15 kHz
Bandwidth	20 MHz	20 MHz
Resource blocks	52	104
Subcarriers	624	1248
PRG size	4	8
Subbands	13	13
BS/UE ports	256/8	32/4

B.3 STANDARDS-TO-BENCHMARK TRACEABILITY

In CSIEval, *standard-aligned* does not mean that every numerical benchmark parameter is mandated by ITU or 3GPP. It means that the benchmark preserves standard-defined radio concepts, channel-model assumptions, and AI/ML evaluation axes where applicable, while explicitly identifying dataset- and benchmark-specific choices. The external bases used for this mapping are ITU-R M.2160 (International Telecommunication Union Radiocommunication Sector, 2023), 3GPP TR 38.843 (3rd Generation Partnership Project, 2025a), 3GPP TR 38.901 (3rd Generation Partnership Project, 2024a), 3GPP TS 38.211 (3rd Generation Partnership Project, 2024b), 3GPP TS 38.214 (3rd Generation Partnership Project, 2025b), and the WAIR-D dataset description (Huangfu et al., 2022). Table 5 provides this one-to-one traceability.

Table 5: Traceability from standards and source-data assumptions to CSIEval benchmark decisions.

Benchmark component	CSIEval decision	Standards/document basis	Classification
6G evaluation scope	AI-based CSI estimation, prediction, and feedback are evaluated as physical-layer functions relevant to IMT-2030/6G development.	ITU-R M.2160	Contextual alignment
AI/ML evaluation standard	Separate training/evaluation conditions, explicit model/data configuration, reproducible testing, and deployment-oriented measurements.	3GPP TR 38.843	Standard-derived evaluation principle
Channel and mobility conditions	Estimation domains vary propagation/channel conditions and mobility using NR channel-model concepts.	3GPP TR 38.901	Standard-derived operating axes
OFDM numerology representation	The task data retain NR-style complex resource-grid structure and use 15/30 kHz subcarrier-spacing configurations.	3GPP TS 38.211	Standard-compatible representation
CSI/physical-layer procedure context	CSI-related evaluation is interpreted within NR physical-layer procedure and link-adaptation terminology.	3GPP TS 38.214	Standard-compatible context
Carrier/profile instantiation	Approximately 7 GHz for prediction and 2.6 GHz for feedback; 20 MHz bandwidth; the exact RB, port, PRG, and subband counts listed in Table 4.	WAIR-D plus CSIEval task construction	Benchmark choice; not claimed as standard-mandated values
Environment split	Derivatives of one source sample remain in one split; WAIR-D training and test domains are mutually exclusive at map level.	CSIEval	Benchmark-specific leakage-control rule
Prediction masks	Comb, block, and random observation masks at declared ratios.	CSIEval	Benchmark-specific task construction
Feedback quantization	4-bit fixed-budget main comparison; 2–8-bit sensitivity sweep.	CSIEval, informed by finite-rate CSI feedback literature	Benchmark-specific evaluation setting
Metric/report schema	Four common dimensions (performance, storage, efficiency, generalization) with task-specific metrics and a common output schema.	CSIEval, informed by reproducible ML/system benchmarking	Benchmark-specific protocol

This separation is important for interpreting the title claim: standards determine or motivate the radio concepts and evaluation axes, whereas frequencies, exact tensor dimensions, split rules, masks, quantization budgets, and metric selection are disclosed CSIEval design choices rather than being presented as requirements imposed by ITU or 3GPP.

For the prediction profile, the base station uses 256 transmit radio units and 1024 antenna elements, while the user equipment uses eight receive radio units. The horizontal and vertical element spacings are $(0.5, 0.8)\lambda$ at the base station and $(0.5, 0.5)\lambda$ at the user equipment. For compression feedback, the base station uses 32 transmit radio units and 128 antenna elements, and the user equipment uses

four receive radio units; all element spacings are 0.5λ. These configurations produce 13 frequency subbands for both tasks and retain the array scales required by their respective evaluation protocols.

B.4 Channel Generation and Normalization

Channel samples are stored as complex-valued arrays rather than images to avoid precision loss in CSI-sensitive tasks. The generator first forms the frequency-domain channel response from the ray parameters and array responses. Four vertically adjacent antenna elements are then aggregated into one logical port. For a port containing $N_e = 4$ elements, the aggregated response is

$$\mathbf{H}_{\text{port}} = \frac{1}{\sqrt{N_e}} \sum_{e=1}^{N_e} \mathbf{H}_e. \tag{15}$$

The factor $1/\sqrt{N_e}$ preserves the expected channel power after aggregation. Normalization is then derived only from quantities observable at model input. For each task, let $\mathbf{O}_i$ be the observable complex-valued input of training sample i and Ω_i its valid observed entries. CSIEval computes one fixed training statistic

$$s_{\text{train}} = \sqrt{\frac{\sum_{i \in \mathcal{D}_{\text{train}}} \|\mathcal{P}_{\Omega_i}(\mathbf{O}_i)\|_F^2}{\sum_{i \in \mathcal{D}_{\text{train}}} |\Omega_i|}} + \epsilon. \tag{16}$$

The statistic is frozen after training-data preprocessing. Estimation uses pilot/reference-signal observations as $\mathbf{O}_i$; prediction uses only entries selected by the observation mask; compression feedback uses the UE-side CSI available to the encoder. Validation/test inputs and supervised targets are divided by the same s_{train}, and no normalization statistic is recomputed from hidden positions, complete held-out targets, or test-domain samples. This prevents the preprocessing stage from disclosing target-derived power information.

The baseline generator uses the first 1,000 maps of the selected source scenario. Environment identifiers remain available during preprocessing, which allows training, validation, in-distribution testing, and out-of-distribution testing to be separated at the map level. For eigenvector feedback, dominant subband eigenvectors are computed during evaluation and cached in a task-local directory. The cache only reduces repeated feature-extraction cost and does not alter the samples or data split.

B.5 Task-Specific Construction Details

The main paper retains the equations required to define each task. The implementation details omitted there are collected here. For estimation, $\mathbf{O}_i$ consists of pilot/reference-signal observations; samples at different reference-signal densities are paired with the corresponding full-channel targets, and channel type, mobility, reference-signal configuration, and observation density are retained as condition labels. Real and imaginary components are represented as separate feature channels after the frozen observable-only normalization.

For prediction, $\mathbf{O}_i = \mathbf{M}_i \odot \mathbf{H}_i$ and Ω_i contains only observed positions. Spatial masks act along the antenna dimension and frequency masks act on subcarriers or subbands. Uniform observation uses a *comb* pattern with fixed spacing; non-uniform observation includes *block* and *random* patterns. At a matched operating point, the same masking ratio is used and observed entries are excluded from loss and metric computation.

For compression feedback, $\mathbf{O}_i$ is the UE-side CSI available to the encoder. Full-matrix feedback uses the complex channel representation directly, while eigenvector feedback uses the subband representation defined in the main paper. Models that explicitly exploit frequency correlation may group adjacent n subbands into one joint sample.

All derivatives of a source sample are assigned to the same split. Estimation partitions channel, mobility, and observation conditions, whereas WAIR-D prediction and feedback use mutually exclusive maps for train and test domains; unseen maps or deployment conditions form the cross-scenario evaluation domains.

Table 6: Supported model integration modes.

Mode	Operational requirement
Preloaded instance	Evaluate an instantiated model after strict checkpoint loading; preferred for custom architectures.
Checkpoint only	Construct the fixed reference task model and load a compatible state dictionary.
Class and arguments	Instantiate the submitted class from recorded constructor arguments and then load the checkpoint.

Table 7: Core evaluator configuration.

Field	Default	Function
`data`	Primary subset	Task-ready samples and feature cache.
`checkpoint`	None	Serialized model state and metadata.
`device`	CUDA	CPU fallback is enabled.
`seed`	42	Deterministic ordering and adaptation.
`latency_runs`	100	Repeated passes for latency statistics.
`modulation`	Explicit	Constellation and bit mapping used for BER evaluation.
`coding`	Explicit	Channel-code configuration; `none` denotes uncoded BER.
`detector`	Explicit	Equalizer/detector used after channel estimation.
`ber_stage`	Explicit	BER observation point, e.g., pre-decoding or post-decoding.
`report_formats`	JSON/HTML	Markdown can also be emitted.

C Evaluation Configuration

The evaluation configuration fixes the task, data subset, model/checkpoint, runtime controls, and output metrics. The primary in-distribution subset is the reference test domain, while unseen-map and dense-deployment subsets provide out-of-distribution targets. Task-specific features such as subband eigenvectors are computed consistently within the declared split.

C.1 Model Integration Modes

The evaluator supports three integration modes, summarized in Table 6. A preloaded model instance is preferred because it preserves the original constructor and permits strict checkpoint validation. Checkpoint-only loading is intended for a known reference architecture. Explicit class loading is used when the model class and constructor arguments are available but the model has not yet been instantiated.

C.2 Configuration Fields

Every run specifies the task, model source, dataset directory, and output directory. Runtime, robustness, adaptation, and reporting settings are controlled through a unified configuration object. Tables 7 and 8 list the complete set of fields used by the current evaluator.

C.3 Model standard and Validation

All models must satisfy the same complex-tensor convention and normalization used during training. For compression feedback, the expected input and output tensors have shape $(B, 2, N_{\mathrm{sub}}, N_t)$, where the second dimension stores real and imaginary components. Architectures using another layout are wrapped by a deterministic permutation layer before evaluation. The optional `encode` method exposes the latent representation, while optional compression-ratio and quantization-bit methods enable feedback-overhead and quantization tests.

Before metric computation, the evaluator performs the checks in Table 9. These checks prevent apparently valid execution from producing incomparable results.

Table 8: Robustness and adaptation configuration.

Field	Default	Function
snr_levels_db	5–40 dB	Controlled input-noise sweep.
quant_bits_sweep	0, 2, 4, 8	Precision-sensitivity sweep.
fewshot_samples	0–300	Target-domain support-set sizes.
fewshot_epochs	30	Maximum adaptation epochs.
fewshot_lr	10^{-4}	Adaptation learning rate.
ood_targets	Empty list	Optional user-defined OOD domains.
default_ood	Enabled	Register unseen-map and dense domains.

Table 9: Pre-evaluation validation protocol.

Check	Required condition
Dataset alignment	Antenna, subcarrier or subband dimensions, and complex-value layout match training.
Checkpoint compatibility	Parameter keys and tensor shapes load strictly; class metadata are recoverable.
Tensor standard	Output has the same layout and scale as the reference tensor.
Normalization consistency	Frozen training-derived observable-only normalization matches training; no held-out target or test-domain statistic is recomputed.
Inference consistency	Evaluation mode, disabled gradients, device, data type, and batch size are fixed.
Metadata completeness	Task, model identity, model size, and quantization state are recorded.

C.4 Reference-Model Provenance and Scope

All evaluated models are trained and tested under the common CSIEval data and evaluation protocol. A superscript * is used only when the cited source model additionally requires model/task-interface adaptation to operate under the CSIEval task definition; routine retraining, common preprocessing, split control, and metric computation alone do not trigger the mark. For prediction, the source models are CsiNet (Wen et al., 2018), WiFo (Liu et al., 2025), and STCMixerNet (He et al., 2025). For estimation, the cited sources are ChannelNet (Soltani et al., 2019), RNN (Gizzini & Chafii, 2024), LSTM (Pan et al., 2021), GNN (Ye et al., 2025), GAN (Balevi & Andrews, 2021), CNN–LSTM (Rapudu & Oyerinde, 2025), CNN–GRU (Hou et al., 2023), and Transformer-PCE (Ju et al., 2024). For starred rows, the defining modeling mechanism of the cited source is retained, while benchmark-dependent input/output representation, data organization, normalization, and evaluation settings are standardized as required by CSIEval. Accordingly, all values reported for starred rows are CSIEval results and should not be interpreted as reproduced source-paper numbers.

For every experimental row, the reproducibility record should include layer/block count, hidden dimensions, kernel or attention settings, activation/normalization choices, optimizer and schedule, batch size, epochs, training loss, random seed, checkpoint-selection rule, parameter count, and input/output tensor shapes. These fields distinguish the source architectural precedent from the exact CSIEval instantiation.

C.5 Runtime Measurement Protocol

Latency and peak memory are measured after warm-up under a fixed device, software stack, data type, batch size, and inference mode. Accelerator timing synchronizes the device before and after each timed region, and the timing boundary is held constant across models. The current configuration uses 100 timed repetitions. Table 11 records the environment needed to interpret the reported latency values.

Table 10: Source provenance and CSIEval standardization for evaluated models. Original sources are cited for every row. All values are obtained under CSIEval. * denotes a CSIEval adaptation whose source task/data interface is not directly compatible with the benchmark; the cited modeling mechanism is retained, while benchmark-dependent I/O and evaluation settings are standardized.

CSIEval row	Source model	Retained mechanism	Standardized by CSIEval
CsiNet* (pred.)	CsiNet (Wen et al., 2018)	Convolutional encoder–decoder mechanism	Prediction I/O, map split, masks, normalization, metrics
WiFo-Little (pred.)	WiFo (Liu et al., 2025)	Little-scale masked-autoencoder prediction mechanism	WAIR-D tensor layout, map split, observation ratios, masks, reporting
STCMixerNet (pred.)	STCMixerNet (He et al., 2025)	Spatio-temporal mixer prediction mechanism	WAIR-D observation ratios, split, normalization, metrics
ChannelNet	ChannelNet (Soltani et al., 2019)	CNN super-resolution/restoration estimation mechanism	Pilot tensor, normalization, split, training/evaluation interface
RNN*	RNN (Gizzini & Chafii, 2024)	Recurrent sequence-modeling mechanism	Observation/output tensor, normalization, split, reporting
LSTM*	LSTM (Pan et al., 2021)	LSTM-based temporal estimation mechanism	Observation/output tensor, normalization, split, reporting
GNN*	GNN (Ye et al., 2025)	Graph message-passing estimation mechanism	Graph input construction, condition labels, split, metrics
GAN*	GAN (Balevi & Andrews, 2021)	Generative channel-estimation mechanism	Common data split, tensor representation, normalization, reporting
CNN+LSTM*	CNN–LSTM (Rapudu & Oyerinde, 2025)	Convolutional front end with LSTM sequence modeling	Tensor, split, normalization, evaluation conditions
CNN+GRU*	CNN–GRU (Hou et al., 2023)	Convolutional front end with GRU sequence modeling	Tensor, split, normalization, evaluation conditions
Transformer-PCE*	Transformer-PCE (Ju et al., 2024)	Transformer-based predictive channel-estimation mechanism	Pilot-observation tensor, reconstruction head, split, normalization, reporting

Table 11: Runtime protocol for efficiency measurements.

Field	Setting
Accelerator / CPU	NVIDIA RTX A6000 (48 GB) / AMD EPYC 7Y83 64-Core Processor
CUDA / framework version	CUDA 12.1 / PyTorch 2.2.2
Batch size	256
Precision	FP32
Warm-up repetitions	10
Timed repetitions	100
Reported latency statistic	Median per-sample latency over synchronized runs
Included timing boundary	Model forward pass only; excludes I/O, preprocessing, and host–device transfer

C.6 Execution and Reporting

A run returns a structured report in which metrics are grouped by performance, storage, efficiency, and generalization. Individual values and OOD sub-results can be queried by key, while category-level filtering supports dimension-specific analysis. The evaluator writes a machine-readable metric file and a formatted report and records the resolved configuration with each output. Latency and memory are measured after warm-up under a fixed device, data type, batch size, and inference mode. Few-shot adaptation is repeated independently for each support-set size to avoid state leakage across adaptation points.

C.7 Anonymous Artifact Availability and Publishing Plan

The current submission is under double-blind review. To support reproducibility without disclosing author identity, an anonymized repository is available at `https://anonymous.4open.science/r/csieval-8018`. The repository provides the executable CSIEval implementation and accompanying artifacts needed to inspect and run the evaluation workflow, while identifying maintainer information and non-anonymous release endpoints remain withheld during review. The executable protocol is additionally specified in the manuscript and this supplement through

the model standard, configuration fields, input/output shapes, split rules, metric equations, runtime controls, and output schema. Each evaluation report records the resolved configuration, evaluator-version identifier, random seed, dependency/runtime information, and model/checkpoint metadata needed to associate a result with a reproducible run. If the paper is accepted, the camera-ready release will replace the anonymous review link with the permanent source repository and disclose the package release, tagged evaluator version, per-model configuration files, executable examples, and public leaderboard endpoint, where applicable.

D Supplementary Metrics Used in Reported Experiments

D.1 Core Metric Interpretation and Task-Specific Reporting

The main paper specifies the four shared evaluation dimensions, task-specific performance metrics, SGCS evaluation units, fixed-budget storage convention, and the mapping from generalization definitions to the reported tables. This supplement therefore retains only implementation details and additional metrics that are actually instantiated in the complementary experiments below. In particular, it specifies the link-level BER chain and the system-level diagnostics that cannot be inferred from CSI reconstruction quality alone.

Fixed-budget quantization accounting. A compression-feedback operating point is fully specified by $(M, b, Q_b, B_{\text{side}})$, including the quantizer type, calibration rule, clipping/range convention, and every transmitted side-information field. These quantities must be identical across rows in the fixed-budget table unless a difference is explicitly charged to B_{side}. For the course-project instantiation used in the completed table, Q_b is a symmetric uniform scalar quantizer applied element-wise to normalized latent codes. The clipping bound is fixed to $\pm 3\sigma_{\text{train}}$, where σ_{train} is estimated from encoder outputs on the training split only, and $\Delta = 6\sigma_{\text{train}}/(2^b - 1)$. Because the calibrated range is frozen and shared by encoder and decoder, no per-sample side information is transmitted and $B_{\text{side}} = 0$.

D.2 Link-Level and Reconstruction Metrics

For metrics not normalized by the reference power, mean squared error is

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{H}_i - \widehat{\mathbf{H}}_i\right\|_F^2. \tag{17}$$

For each sample, flatten the complex CSI tensors to vectors $\mathbf{h}_i$ and $\hat{\mathbf{h}}_i$, subtract their respective complex sample means, and compute the magnitude of the complex Pearson coefficient. The reported correlation is

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\left(\mathbf{h}_i - \bar{h}_i\mathbf{1}\right)^H\left(\hat{\mathbf{h}}_i - \bar{\hat{h}}_i\mathbf{1}\right)\right|}{\left\|\mathbf{h}_i - \bar{h}_i\mathbf{1}\right\|_2\left\|\hat{\mathbf{h}}_i - \bar{\hat{h}}_i\mathbf{1}\right\|_2 + \epsilon}, \tag{18}$$

and Table 1 reports $100\rho\%$. For the compression-feedback link diagnostic, error vector magnitude is

$$\text{EVM} = \frac{\sqrt{\mathbb{E}\{|s-\hat{s}|^2\}}}{\sqrt{\mathbb{E}\{|s|^2\}}} \times 100\%, \tag{19}$$

where s and $\hat{s}$ are the transmitted and recovered modulation symbols under the fixed link configuration.

BER protocol. BER is evaluated through an explicit link-level chain rather than inferred from CSI reconstruction error. A common transmitted bit sequence is mapped by the run-specified modulation, passed through the same channel/noise realization, equalized or detected using the evaluated channel estimate, and demapped using the same detector and coding configuration for every compared model. The configuration explicitly records `modulation`, `coding`, `detector`, and `ber_stage`; therefore, every reported BER can be identified as uncoded/pre-decoding or coded/post-decoding without ambiguity. Deterministic constellation phase or polarity ambiguities

are resolved from pilots or known reference symbols before bit demapping; the transmitted target bits are never used to select a favorable inversion. BER is then

$$\mathrm{BER} = \frac{\sum_{n=1}^{N_{\mathrm{bit}}} \mathbb{I}[\hat{b}_n \neq b_n]}{N_{\mathrm{bit}}} \times 100\%. \tag{20}$$

The same transmitted bits, mapper, channel/noise realization, detector, coding state, and BER observation point are used for all models at a given operating point. This protocol prevents a CSI-reconstruction metric, symbol-error rate, or block-error rate from being mislabeled as BER and prevents systematic bit-label or phase ambiguities from producing physically misleading low-SNR values.

D.3 Latency and System-Level Metrics

The stage-level end-to-end feedback latency reported in Table 12 is

$$T_{\mathrm{stage}} = T_{\mathrm{enc}} + T_{\mathrm{trans}} + T_{\mathrm{dec}}, \tag{21}$$

where the terms are measured encoding, transmission, and decoding latency under the same runtime configuration.

The fifth-percentile user throughput is

$$R_{5\%} = Q_{0.05}(\{R_u\}), \tag{22}$$

and characterizes the lower tail of the evaluated user-throughput distribution. Spectral efficiency is

$$\eta = \frac{R}{B} \quad [\mathrm{bit/s/Hz}], \tag{23}$$

where R is the achieved data rate and B is occupied bandwidth. The reported channel-capacity value uses the fixed transmit-covariance and SNR convention of the corresponding experiment. Estimation accuracy is the fraction of samples satisfying the predeclared amplitude-and-phase tolerance used by that experiment, while pilot-overhead ratio is the fraction of evaluated resource elements allocated to reference/pilot observations.

D.4 Measured Storage and Efficiency Diagnostics

The active-parameter ratio reported for compression feedback is

$$r_{\mathrm{act}} = \frac{N_{\mathrm{active}}}{N_{\mathrm{param}}}. \tag{24}$$

Model loading time is measured from initiation of checkpoint access until device-resident model initialization is complete. Runtime memory is the peak allocated model/runtime memory under the declared batch and precision configuration, and storage compression ratio is the reported stored-model-size ratio relative to the corresponding full-precision reference.

Parallel acceleration is

$$S_{\mathrm{parallel}} = \frac{T_{\mathrm{serial}}}{T_{\mathrm{parallel}}}, \tag{25}$$

under matched batch size and output accuracy. The preprocessing share is

$$R_{\mathrm{pre}} = \frac{\mathrm{FLOPs}_{\mathrm{pre}}}{\mathrm{FLOPs}_{\mathrm{total}}} \times 100\%. \tag{26}$$

Memory-access latency reduction is measured relative to the unoptimized execution path used in the same implementation. Training time is measured from initialization to the declared stopping criterion. CPU/GPU utilization values are time-averaged over the same measurement window, and the reported power reduction is measured at matched throughput relative to the implementation's reference runtime.

Table 12: Link-level diagnostics not reported in the main CSI compression-feedback results.

Diagnostic	**Value**
Final test loss	3.899×10^{-5}
Mean absolute error	0.0036
PSNR	44.0952 dB
Signal SNR	38.0886 dB
Cosine similarity	0.9999
Encoding latency	0.2790 ms
Transmission latency	0.3277 ms
Decoding latency	0.3402 ms
Stage-level end-to-end latency	0.9468 ms
Mean EVM	-38.09 dB
Mean subcarrier EVM	-39.29 dB

D.5 Measured Adaptation and Robustness Diagnostics

Adaptation latency is the interval between applying a declared domain transition and reaching the stable adapted state defined by the experiment's fixed acceptance criterion. Adaptation success rate is the fraction of repeated transitions that satisfy that criterion within the allowed adaptation budget. Channel-model adaptability and mobility robustness are reported as retained task performance over the declared channel-model and mobility sets, respectively. Temporal and spatial consistency summarize retained agreement across adjacent temporal indices and neighboring spatial units under the same task metric. Doppler tolerance and frequency-selectivity tolerance are reported at the declared acceptance threshold. These quantities appear in the supplementary experiments below; no additional unmeasured adaptation, codebook, update-bandwidth, block-error, energy-per-inference, occlusion-recovery, or measurement-overhead metrics are defined in this supplement.

E Additional Experimental Results

The following experiments report task-specific diagnostics that complement the main four-dimensional benchmark tables. They use the same data/model standard and reporting framework, but they are not mandatory core metrics for every task and are therefore interpreted within their corresponding task rather than merged into a cross-task scalar score.

E.1 CSI Compression Feedback

E.1.1 Additional Robustness Beyond the Main Sweep

The main paper reports the fixed-budget interpretation and SNR/quantization sweep. The additional cross-domain stress test below isolates the interaction between aggressive quantization and deployment shift.

Under 2-bit quantization, neither target domain exhibits significant additional degradation relative to the in-distribution reference; after adaptation with 300 target-domain samples, NMSE improves to -7.657 dB and -7.640 dB. In this setting, quantization is the dominant bottleneck, while the added effect of map and deployment shifts is comparatively small.

E.1.2 Link-Level Fidelity Diagnostics

A separate run at a representation ratio of $1/4$ is used only to characterize signal fidelity and end-to-end link behavior. Magnitude error is concentrated below 0.005, with more than 80% of the samples below 0.003. Phase error is centered near zero and is similarly concentrated below 0.005 rad. These distributions indicate that the reconstructed CSI preserves both amplitude and phase structure beyond what is conveyed by the aggregate reconstruction metrics in the main paper.

Table 13: Operational diagnostics for CSI compression feedback.

Operational diagnostic	**Value**
Active-parameter ratio	0.5382
Cold-loading time	156.8 ms
Preprocessing share	12.3%
Batch-32 parallel speedup	8.7×
Memory-access latency reduction	42%
CPU utilization	35–45%
Peak GPU utilization	78%
Power reduction at matched throughput	28%
Continuous stability test	72 h

Table 14: Cross-domain dynamic adaptation for CSI compression feedback.

Transition	**NMSE (dB)**	**SNR (dB)**	**Latency (ms)**	**Success (%)**
In→Out	−4.67	30.97	126.11	92.3
In→Low	−6.45	33.00	79.74	95.8
In→High	−5.36	32.67	81.81	94.2
Out→In	1.97	30.47	129.53	89.1
Out→Low	1.55	30.58	126.18	91.7
Out→High	1.28	30.51	127.63	90.4
Low→In	−5.17	31.13	124.41	93.6
Low→Out	−5.53	31.48	111.75	94.8
Low→High	−4.60	31.17	123.11	92.9
High→In	−5.07	31.05	126.46	91.2
High→Out	−6.71	31.60	117.10	93.4
High→Low	−4.16	31.12	124.26	92.1

The corresponding system-level evaluation yields an average user throughput of 31.46 Mbps and a fifth-percentile throughput of 30.96 Mbps. Ninety-five percent of users exceed 30.8 Mbps, and the fairness index is 0.92. Within the evaluated 800 m coverage region, throughput varies by approximately $\pm 5\%$, showing that the compressed representation maintains comparable service quality across user locations.

E.1.3 Operational Diagnostics

Table 13 reports implementation measurements that are not part of the unified model table in the main paper. The active-parameter ratio indicates that slightly more than half of the parameters contribute during a typical forward pass. Preprocessing remains a minor component of the execution cost, whereas batching, prefetching, and caching provide substantial acceleration. The 72-hour stability test exhibits no measurable degradation, supporting sustained deployment under the evaluated runtime conditions.

E.1.4 Mobility-Aware Transition Behavior

The main paper evaluates static cross-map and cross-deployment generalization. The following experiment instead measures online transition behavior among static indoor, static outdoor, low-mobility, and high-mobility domains. Table 14 reports only adaptation latency and success rate, thereby avoiding duplication of the reconstruction and retention metrics already used in the main benchmark.

The mean adaptation latency is 116.51 ms and the mean success rate is 92.63%. Indoor-to-low-mobility transfer is the fastest and most reliable transition, whereas outdoor-to-indoor transfer is

Table 15: Classical estimation references under the same CSIEval estimation test conditions.

Method	NMSE (dB)↓	BER (%)↓	Latency (ms)↓	Peak Mem. (MB)↓
LS	−12.84	24.60	0.18	74.2
LMMSE	−18.27	17.35	0.73	82.6

Table 16: Additional system and deployment measurements for CSI channel estimation.

Metric	Value
Spectral efficiency	4.73 bit/s/Hz
Channel capacity	8.94 bit/s/Hz
Estimation accuracy	89.42%
Pilot-overhead ratio	0.125
Runtime memory	12.3 MB
Quantization precision	16 bits
Storage compression ratio	0.68
Model loading time	245.6 ms
Initialization time	89.3 ms
Inference throughput	2,847 samples/s
MACs per sample	1,423,781
Training time	2.8 h

the slowest and least successful. The asymmetry indicates that adaptation difficulty depends on the direction of the propagation-domain shift rather than only on mobility level.

E.2 CSI Prediction

The three prediction rows in Table 2 use CsiNet, WiFo-Little, and STCMixerNet as their architectural bases (Wen et al., 2018; Liu et al., 2025; He et al., 2025). WiFo-Little and STCMixerNet are evaluated as prediction models, whereas CsiNet carries * because its published feedback/recovery architecture is cross-task adapted to the structured missing-value prediction standard. The set validates protocol coverage rather than reproducing source-paper performance claims. Cross-mask gaps are 0.05, 0.04, and 0.02 dB, respectively; without repeated-training confidence intervals, these differences are descriptive only.

E.3 CSI Channel Estimation

E.3.1 Reference-Model Scope and Additional Interpretation

The estimation experiment evaluates representative neural models after re-instantiation under the same tensor, split, normalization, observation, and reporting standard. Source-paper numerical results are intentionally not mixed with these values because they were produced under different evaluation conditions.

Classical estimators are nevertheless useful engineering anchors. We therefore report LS and LMMSE separately using the same reporting fields as the CSIEval estimation test condition. They are not included in the learned-model multidimensional table because trainable parameter count and model-file size are not comparable quantities for these estimators.

The following subsections provide complementary system measurements and SNR diagnostics.

E.3.2 System and Deployment Measurements

The channel-estimation experiment supplements the main comparison with link-level efficiency, deployment behavior, and detailed robustness measurements. Table 16 summarizes system and implementation quantities that are not included in the main model table.

Table 17: CSI channel-estimation reconstruction diagnostics across SNR. SNR and NMSE are in dB.

SNR	NMSE	MSE	Corr.
−10	−1.28	0.743	0.412
−5	−7.13	0.194	0.628
0	−12.05	0.062	0.794
5	−15.76	0.027	0.885
10	−18.13	0.015	0.936
15	−19.19	0.012	0.958
20	−19.65	0.011	0.969
25	−19.79	0.010	0.975
30	−19.84	0.010	0.978

Table 18: Representative model NMSE values across SNR, in dB. Original sources are cited in the column headers. All values are obtained under CSIEval. * denotes a CSIEval adaptation whose source task/data interface is not directly compatible with the benchmark; the cited modeling mechanism is retained, while benchmark-dependent I/O and evaluation settings are standardized.

SNR	**ChannelNet** (Soltani et al., 2019)	**LSTM*** (Pan et al., 2021)	**GAN*** (Balevi & Andrews, 2021)	**Trans.-PCE*** (Ju et al., 2024)
−10	−1.28	−8.50	−9.80	−12.10
−5	−7.13	−12.30	−14.20	−16.80
0	−12.05	−16.70	−18.60	−21.40
5	−15.76	−20.50	−22.10	−26.80
10	−18.13	−24.10	−25.30	−30.20
15	−19.19	−26.80	−27.80	−32.50
20	−19.65	−28.20	−29.40	−34.10
25	−19.79	−29.10	−30.20	−35.20
30	−19.84	−29.60	−30.80	−35.80

The estimator attains high spectral efficiency with a pilot-overhead ratio of 0.125, indicating that estimation quality is maintained while reserving most resource elements for data transmission. The 4.01 MB model requires 12.3 MB at runtime and loads in 245.6 ms, which supports rapid initialization on edge devices. The measured throughput and MAC count further indicate that the model can sustain real-time batch processing without a high-capacity accelerator.

E.3.3 Complete SNR Sweep

Table 17 reports reconstruction diagnostics across SNR. The ChannelNet estimator improves rapidly between -10 and 15 dB and then approaches saturation: NMSE decreases from -1.28 dB at -10 dB SNR to -19.19 dB at 15 dB, while the correlation coefficient rises from 0.412 to 0.958. Above 20 dB, NMSE remains near -19.8 dB and the correlation coefficient approaches saturation. BER is not included in this reconstruction-only SNR sweep. BER is reported only when the modulation, coding state, detector, transmitted bits, and BER observation point are fixed and recorded as required by the evaluator configuration.

A broader robustness comparison is given in Table 18. ChannelNet (Soltani et al., 2019) denotes the source CNN estimator, while Trans.-PCE* (Ju et al., 2024) denotes its CSIEval-adapted Transformer-based comparison counterpart. Unlike the main paper, which reports one low-SNR operating point, this table exposes the complete response curve of representative architectures. Transformer-PCE* and GAN* provide lower low-SNR NMSE in these CSIEval point estimates, while ChannelNet saturates earlier. The slope differences between -10 and 10 dB are substantially larger than those above 20 dB, confirming that the low-SNR region is the most discriminative robustness regime.

E.3.4 Additional Robustness Diagnostics

Additional measured diagnostics give 89.2% channel-model adaptability, 85.6% mobility robustness, 120.5 Hz Doppler tolerance, 92.3% frequency-selectivity tolerance, 93.4%/88.9% temporal/spatial consistency, 74.3% out-of-distribution retention, and 68.7% transfer efficiency using 128 labels. CDLA is more robust than CDLC at low SNR, and half precision consistently exceeds quarter precision. The dominant failure regime combines high mobility, low SNR, and coarse precision; BER-based mobility comparisons are reported only after applying the fixed link-level protocol in Eq. (20).