Normalization-Free Knowledge Distillation for Variable-Bit CSI Feedback in Massive MIMO

Anonymous Authors[1]

Abstract

Efficient Channel State Information (CSI) feedback is critical for beamforming in Frequency Division Duplexing (FDD) Massive MIMO, yet variable-bit deep learning schemes suffer from trailing-bit information loss under constrained training budgets (Ji & Chung, 2024). While Knowledge Distillation (KD) is a natural remedy, we identify a previously overlooked problem: standard Z-score normalization applied during KD destroys the physical scale information embedded in wireless channel tensors, degrading reconstruction quality across variable feedback lengths. We propose Normalization-Free Knowledge Distillation (NF-KD), a training framework that preserves physical scale by applying Kullback–Leibler Divergence (KLD) directly to raw, temperature-scaled tensors without any normalization step. Experiments on the DeepMIMO dataset show that NF-KD achieves -12.87 dB Normalized Mean Squared Error (NMSE) at 256 bits (vs. -11.98 dB baseline), consistent cosine similarity gains across all feedback lengths, and a 0.056 dB reduction in beamforming gain loss that translates to $+0.019$ bps/Hz spectral efficiency under Zero-Forcing (ZF) precoding at 20 dB Signal-to-Noise Ratio (SNR)—all within a 500-epoch training budget. Ablation confirms that normalization is the critical failure factor.

1. Introduction

Massive MIMO systems under FDD require precise CSI at the Base Station (BS) for effective beamforming, yet the feedback overhead scales prohibitively with the number of antennas. Deep learning-based autoencoders have emerged as a dominant approach (Guo et al., 2022), compressing the channel matrix $\mathbf{H}$ at the User Equipment (UE) and reconstructing it at the base station (Wen et al., 2018). More recently, variable-bit schemes have been proposed to adapt feedback granularity to instantaneous channel conditions (Ji & Chung, 2024); however, the Feedback Bit Masking Unit (FBMU) truncation mechanism causes structural information loss when training resources are limited, degrading reconstruction quality at short feedback lengths.

KD offers a principled path to recovering this lost structure. Prior KD work for CSI feedback follows the classical model-compression paradigm, where a high-capacity Teacher transfers knowledge to a parameter-limited Student via soft labels. Although our setting targets bit-width rather than parameter count, the two regimes share a common abstraction: in both cases the Student is constrained by a narrow information channel—whether defined by model capacity or by feedback bits—and the resulting information loss cannot be fully recovered by hard Mean Squared Error (MSE) supervision alone. Soft-label distillation addresses precisely this gap by conveying distributional structure (e.g., global phase relationships) that persists through the bottleneck. We therefore adopt a fixed-length, 256-bit Teacher to guide the variable-bit Student across 1–256 bit budgets, transferring global phase knowledge that would otherwise be lost to FBMU truncation.

In this paper, we propose NF-KD, a training framework that preserves physical scale by applying KLD directly to raw, temperature-scaled channel tensors without any normalization step. A Phase-Aware Teacher, pre-trained under the full 256-bit budget, guides the Student via a KLD auxiliary loss, allowing the Student to align its global energy distribution with the Teacher's without sacrificing absolute magnitude. The main contributions of this paper are as follows:

- We identify the scale distortion problem caused by Z-score normalization in KD for variable-bit CSI feedback—a previously overlooked problem that degrades reconstruction performance, especially in the low-bit regime.
- We propose NF-KD, which to the best of our knowledge is the first framework to simultaneously preserve physical scale and global phase structure in variable-bit

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

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

CSI feedback under a constrained training budget.

- We experimentally demonstrate consistent gains across all feedback lengths in NMSE, cosine similarity, beamforming gain loss, and spectral efficiency under a 500-epoch budget, and confirm via ablation that normalization is the critical failure factor.
- We bridge reconstruction-level metrics to system-level utility: NF-KD reduces beamforming gain loss by up to 0.056 dB and improves spectral efficiency by up to $+0.019$ bps/Hz at 20 dB SNR, validating that scale preservation translates into deployable 6G beamforming gains.

The remainder of this paper is organized as follows. Section 2 provides background on Massive MIMO CSI feedback and analyzes the scale distortion problem of normalization-based KD. Section 3 describes the proposed NF-KD framework. Section 4 presents experimental results and ablation studies, and Section 5 concludes the paper.

2. Background and Motivation

2.1. Theoretical Limitations of CSI Feedback in Massive MIMO

In Massive MIMO systems, the performance of ZF precoding for achieving multiplexing gain is directly determined by CSI estimation accuracy. Channel estimation errors increase inter-user interference and degrade system throughput (Jindal, 2006), making the acquisition of high-fidelity CSI essential. To minimize the substantial feedback overhead, deep learning-based autoencoder methods have been proposed as an alternative (Wen et al., 2018), in which the original channel matrix $\mathbf{H}$ is compressed into a low-dimensional vector $\mathbf{s}$ through an encoder and reconstructed as $\hat{\mathbf{H}}$ at the base station via a decoder:

$$\mathbf{s} = f_{en}(\mathbf{H}, \Theta_{en}), \qquad \hat{\mathbf{H}} = f_{de}(\mathbf{s}, \Theta_{de}) \tag{1}$$

2.2. Scale Distortion of Normalization in KD

Wireless channel tensors encode intrinsic physical magnitude: the absolute scale of $\mathbf{H}$ reflects path loss, antenna gain, and beamforming energy, all of which are central to massive MIMO system design (Lu et al., 2014). For a channel drawn from a distribution with variance σ_h^2, the magnitude σ_h governs the energy budget that the base station must allocate when computing precoding weights. Distorting this scale during training therefore prevents the Student from learning the correct energy allocation across variable feedback lengths.

Scale invariance under Z-score normalization. A standard KD pipeline applies Z-score normalization to the Teacher and Student outputs before computing the KLD:

$$\tilde{v} = \frac{v - \mu_v}{\sigma_v} \tag{2}$$

where μ_v and σ_v are the sample mean and standard deviation of the flattened tensor v. This mapping is *scale-invariant*: for any scalar $\alpha > 0$, replacing v with αv yields $\widetilde{\alpha v} = \frac{\alpha v - \alpha\mu_v}{\alpha\sigma_v} = \tilde{v}$. Consequently, the normalized tensor is identical regardless of the channel's absolute magnitude. The resulting distillation loss $\mathcal{L}_{\mathrm{KD}} = D_{\mathrm{KL}}(P_T \,\|\, P_S)$ where $P_T = \mathrm{Softmax}(\tilde{v}_T/\tau)$ and $P_S = \mathrm{LogSoftmax}(\tilde{v}_S/\tau)$ follow (Hinton et al., 2015), is therefore invariant to the global scale of the channel. Formally, letting σ_h denote this scale factor:

$$\frac{\partial \mathcal{L}_{\mathrm{KD}}}{\partial \sigma_h} = 0, \tag{3}$$

confirming that the KLD loss provides *no gradient signal* for scale alignment. The Student is thus free to converge to an arbitrary energy level, irrespective of the physical magnitude of the Teacher's output.

Scale sensitivity without normalization. When KLD is applied directly to temperature-scaled raw tensors, the Softmax distribution $P = \mathrm{Softmax}(v/\tau)$ retains sensitivity to the absolute scale of v through its entropy. By the standard properties of the Softmax distribution and Shannon entropy (Cover & Thomas, 2006), scaling the input by $\alpha > 1$ yields a more concentrated distribution and thus strictly lower entropy:

$$H(\mathrm{Softmax}(\alpha v/\tau)) < H(\mathrm{Softmax}(v/\tau)), \quad \alpha > 1. \tag{4}$$

meaning that channels with larger absolute magnitude produce more concentrated, lower-entropy distributions, while high path-loss channels (small $\|v\|$) yield flatter, higher-entropy distributions. Because both Teacher and Student distributions inherit this scale dependence, the resulting KLD becomes a function of the input magnitude—in direct contrast to Eq.(3), where the gradient with respect to σ_h vanishes. Intuitively, low-entropy (high-magnitude) distributions sharpen any mismatch between the Teacher's and Student's modes, providing stronger gradient signals on precisely those channels where accurate energy allocation matters most—a regime where high-fidelity CSI has long been recognized as essential for effective beamforming (Lu et al., 2014; Wen et al., 2018).

In the high-dimensional channel space ($D = 262{,}144$), Teacher and Student outputs are virtually never exactly parallel, and FBMU-induced residual mismatch persists throughout training, particularly at low bit budgets. As the Student converges toward the Teacher, the KLD itself diminishes, ensuring that this scale-aware supervision fades naturally only when no longer needed.

Severity in the low-bit regime. At small feedback budgets, FBMU truncation (Ji & Chung, 2024) weakens MSE supervision over the sparse code, leaving scale alignment to the distillation loss alone. Section 4.3 confirms this: under Normalized KD, the Student degrades below the no-KD baseline at 32 bits.

2.3. Related Work

Deep learning-based CSI compression has evolved from fixed-rate autoencoders (Wen et al., 2018) to multi-rate (Guo et al., 2020), variable-length (Ji & Chung, 2024), and Transformer-based schemes (Cui et al., 2022).

Prior research on KD for CSI feedback has primarily targeted network compression under fixed compression ratios (Tang et al., 2021; Cui et al., 2024), focusing on UE-side model lightweighting rather than the structural information loss arising in variable-bit environments. More critically, these methods overlook the channel scale distortion introduced by standard KD pipelines: applying Z-score normalization before the KLD computation destroys absolute scale, acting as a bottleneck that degrades reconstruction performance. To the best of our knowledge, NF-KD is the first work to jointly address the variable-bit regime and absolute-scale preservation in KD-based CSI feedback.

3. Proposed Method

In this section, we describe the proposed NF-KD framework. The core objective is to transfer the *global phase map* and structural robustness from a high-capacity Teacher model to a variable-bit Student model without compromising the physical scale of the CSI data.

3.1. Phase-Aware Static Knowledge Distillation

KD is employed as a methodology to overcome the structural information loss inherent in the FBMU's truncation process. Unlike conventional KD that targets model compression, our approach focuses on structural restoration of information lost through FBMU truncation.

The framework consists of a Phase-Aware Teacher (f_T), which is pre-trained for 1,000 epochs to understand the full-band global phase, and a Student model (f_S), which learns to reconstruct CSI across 1–256 bits. During training, the Teacher's parameters are frozen, and its output $\hat{H}_T$ serves as a *soft label* or structural guide for the Student's output $\hat{H}_S$.

3.2. Global Phase Map via High-Dimensional Flattening

To capture the intricate spatial-frequency correlations of Massive MIMO channels, we treat the entire channel tensor as a single Global Phase Map.

Both $\hat{H}_T$ and $\hat{H}_S$ (shape: $2 \times N_t \times N_r \times N_c$) are flattened into a high-dimensional vector $\mathbf{v} \in \mathbb{R}^D$, where $D = 262{,}144$.

By flattening the tensor before distillation, the model learns the global dependency across all antennas and subcarriers. This is crucial for FBMU environments where trailing bit truncation often leads to local phase collapses; the distillation process forces the Student to align its global energy distribution with the Teacher's.

3.3. KLD Formulation

While KD is the overall training methodology, KLD is the mathematical indicator used to measure the discrepancy between the two models' distributions. Following the formulation by Hinton et al. (2015), we apply temperature scaling ($\tau = 5$) to soften the high-dimensional flattened outputs into probability distributions:

$$P_T = \mathrm{Softmax}(v_T/\tau), \quad P_S = \mathrm{Softmax}(v_S/\tau) \tag{5}$$

The KLD loss is then defined as:

$$\mathcal{L}_{\mathrm{KD}} = D_{\mathrm{KL}}(P_T \,\|\, P_S) = \sum_{i=1}^{D} P_{T,i} \log \frac{P_{T,i}}{P_{S,i}} \tag{6}$$

The Student aligns its softened energy distribution with the Teacher's, capturing global phase and energy ratios rather than memorizing raw values.

The Student aligns its softened energy distribution with the Teacher's, capturing global phase and energy ratios rather than memorizing raw values.

3.4. Normalization-Free Architecture and Loss Ensemble

A key contribution of this work is the omission of Z-score normalization to preserve the critical physical information embedded in absolute magnitudes. In our NF-KD framework, we apply KLD directly to the raw, temperature-scaled tensors, ensuring that the channel's absolute magnitudes—which encode path-loss and beamforming-energy structure—are preserved in the gradient signal.

We combine the Ground Truth (GT) reconstruction loss with the KD loss through a weight ratio of $\alpha_{\mathrm{KD}} = 0.3$, keeping the GT signal as the primary objective and KD as auxiliary structural guidance:

$$\mathcal{L}_{\text{total}} = (1 - \alpha_{\mathrm{KD}}) \cdot \mathcal{L}_{\mathrm{MSE}}(\hat{H}_S, H) + \alpha_{\mathrm{KD}} \cdot (\mathcal{L}_{\mathrm{KD}} \cdot \tau^2) \tag{7}$$

where $\mathcal{L}_{\mathrm{MSE}} = \|\hat{H}_S - H\|_F^2$, and τ^2 restores the gradient scale reduced by temperature softening.

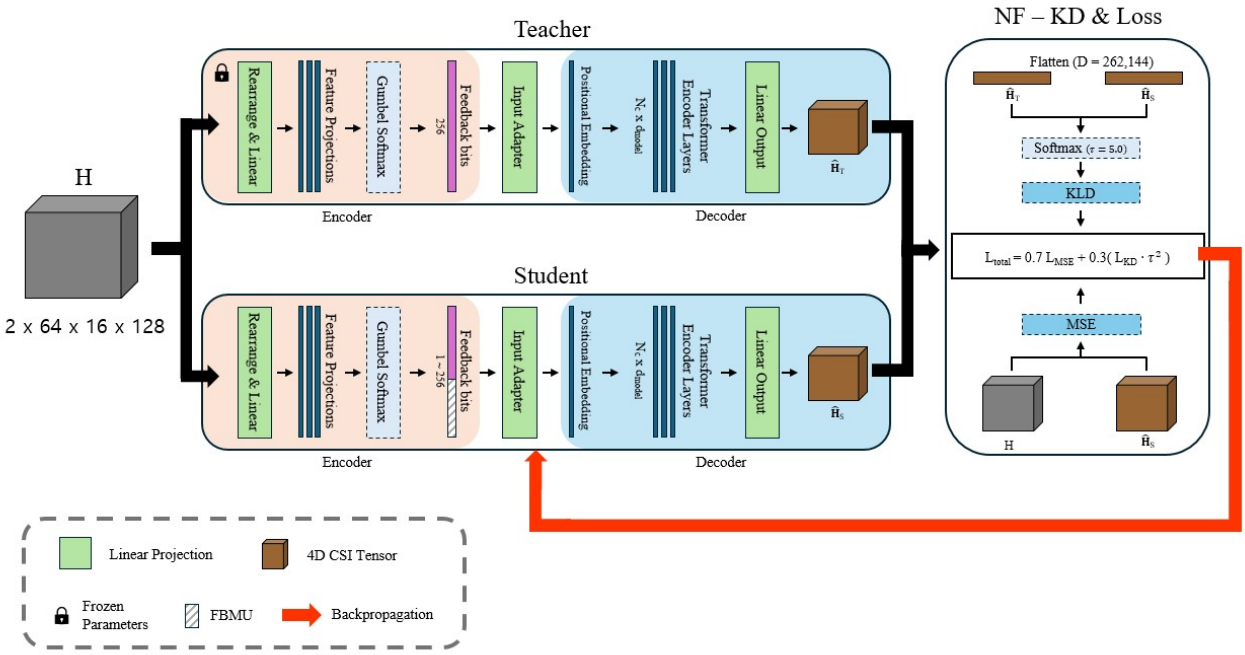

Figure 1. Architecture of NF-KD. The frozen Teacher provides a soft label; the Student with FBMU reconstructs CSI under variable-bit constraints (1–256 bits). KD loss is computed from flattened vectors without Z-score normalization.

Table 1. Simulation Parameters. UPA: Uniform Planar Array.

Parameter	Value
Scenario	O1
Frequency Band	140 GHz
Bandwidth	100 MHz
BS Antenna (N_t)	8×8 UPA
UE Antenna (N_r)	4×4 UPA
Subcarriers (N_c)	128
Number of Paths (P)	25

Table 2. Training Parameters

Parameter	Value
Optimizer	AdamW
Batch Size	256
Student Epochs	500
Teacher Epochs	1000
Scheduler	Cosine Annealing w/ Warm Restarts
Learning Rate ($\eta_{\max}/\eta_{\text{base}}$)	$10^{-3}/10^{-5}$
KLD Loss Weight(α_{KD})	0.3
KD Temperature (τ)	5
Concrete τ (initial $\rightarrow$ final)	$10 \rightarrow 0.01$

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Dataset and Channel Model

To objectively validate the proposed method, we use the DeepMIMO dataset simulation environment (Alkhateeb, 2019). The scenario and channel parameters used for data generation are summarized in Table 1. The complex channel matrix is separated into real and imaginary components and transformed into a tensor of shape $(2 \times N_t \times N_r \times N_c)$. The dataset is randomly split into training, validation, and test sets in a 6:2:2 ratio.

4.1.2. Training Environment and Hyperparameters

All Student and baseline models are evaluated under identical conditions. Specifically, these models are implemented in PyTorch with a batch size of 256 and trained for 500 epochs using the AdamW optimizer (Loshchilov & Hutter, 2017). Cosine Annealing with Warm Restarts (Loshchilov & Hutter, 2016) is used for learning rate scheduling. The temperature parameter τ controlling the discretization of the Concrete Feedback Layer is decayed exponentially from an initial value of 10 to a final value of 0.01. Detailed

hyperparameters are listed in Table 2.

4.1.3. Evaluation Metrics

We evaluate CSI reconstruction performance using four metrics: NMSE, cosine similarity (ρ), Beamforming Gain Loss (Δ_{BG}), and Spectral Efficiency (SE).

1. **NMSE**
   NMSE measures power and scale error on a decibel (dB) scale; lower values indicate better reconstruction:

$$\mathrm{NMSE}(\hat{\mathbf{H}}, \mathbf{H}) = \mathbb{E}\left\{ \frac{\|\hat{\mathbf{H}} - \mathbf{H}\|_F^2}{\|\mathbf{H}\|_F^2} \right\} \tag{8}$$

   where $\mathbb{E}\{\cdot\}$ denotes the expectation and $\|\cdot\|_F$ denotes the Frobenius norm.

2. **Cosine Similarity (ρ)**
   Cosine similarity evaluates the subspace alignment and phase agreement between two channel matrices. It is computed per subcarrier and defined as:

$$\rho(\hat{\mathbf{H}}, \mathbf{H}) = \frac{1}{N_c} \sum_{c=1}^{N_c} \frac{\langle \hat{\mathbf{H}}^c, \mathbf{H}^c \rangle_F}{\|\hat{\mathbf{H}}^c\|_F \, \|\mathbf{H}^c\|_F} \tag{9}$$

   where N_c is the total number of subcarriers and $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. A value of ρ close to 1 indicates that the beamforming direction and phase of the reconstructed channel are in perfect agreement with the ground truth.

3. **Beamforming Gain Loss (Δ_{BG})**
   Beamforming gain loss quantifies the reduction in effective beamforming energy at the base station due to imperfect CSI reconstruction. Under the standard ZF precoding model (Jindal, 2006), it is derived directly from the cosine similarity as:

$$\Delta_{\mathrm{BG}} = -20 \log_{10}(\rho) \quad [\mathrm{dB}] \tag{10}$$

   A value of Δ_{BG} close to 0 indicates near-perfect beam alignment, while larger values reflect greater energy misdirection caused by CSI reconstruction error.

4. **Spectral Efficiency (SE)**
   Spectral efficiency measures the achievable data rate per unit bandwidth under imperfect CSI feedback. Following the massive MIMO analysis in Lu et al. (2014), the effective SNR is approximated as $\rho^2 \cdot \mathrm{SNR}$, yielding:

$$\mathrm{SE} = \log_2\left(1 + \rho^2 \cdot \frac{P}{\sigma^2}\right) \quad [\mathrm{bps/Hz}] \tag{11}$$

   where P/σ^2 denotes the transmit SNR. We evaluate SE at $P/\sigma^2 = 20\,\mathrm{dB}$ as a representative operating point. Higher values indicate greater achievable throughput.

4.2. Performance Evaluation

For evaluation, we use as the baseline a model that employs the Concrete Feedback Layer and FBMU architecture from Ji & Chung (2024) but trained without KD. The proposed NF-KD distills knowledge from a Teacher pre-trained for 1,000 epochs under a fixed 256-bit budget, with a loss ratio of $\alpha_{\mathrm{KD}} = 0.3$ (KLD : GT $= 0.3 : 0.7$). All models are evaluated under the same 500-epoch budget.

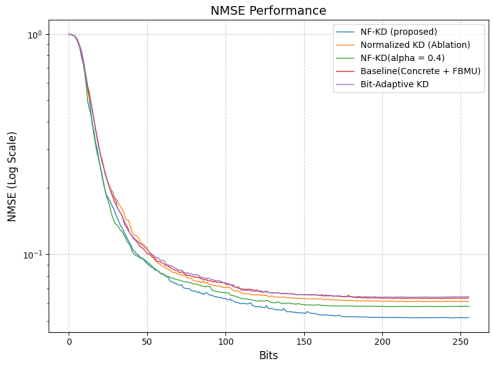

Figure 2. NMSE performance analysis.

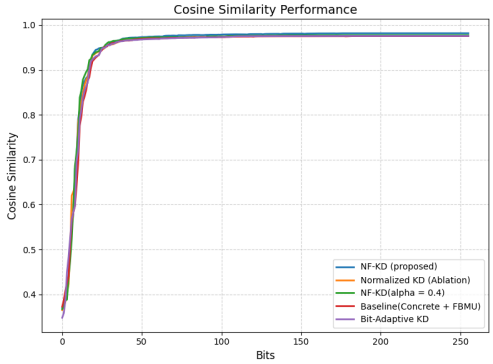

Figure 3. Cosine similarity performance analysis.

As shown in Table 3, Figs. 2 and 3, NF-KD achieves the strongest NMSE and cosine similarity at 64–256 bits, mitigating the trailing-bit saturation observed in the baseline. The improvement is most pronounced at 256 bits, where NF-KD reaches $-12.87\,\mathrm{dB}$ NMSE versus $-11.98\,\mathrm{dB}$ for the baseline.

KLD vs. MSE-based distillation. To isolate the contribution of the KLD formulation, we introduce *Bit-Adaptive KD*, which replaces the KLD term with an MSE loss against the Teacher output. Its bit-dependent weight $\alpha_i = 0.5(1 - b_i/b_{\max})$ is designed to increase Teacher reliance at smaller bit budgets. Despite this adaptive design, Bit-Adaptive KD performs slightly below the baseline at 64–256 bits, indicating that pointwise MSE alignment cannot capture the inter-element energy ratios encoded in the Teacher's

Table 3. Performance comparison across variable feedback bit lengths.

	NMSE (dB)				Cosine Similarity			
Model	32 bit	64 bit	128 bit	256 bit	32 bit	64 bit	128 bit	256 bit
Baseline (Concrete + FBMU)	−7.88	−10.53	−11.73	−11.98	0.9594	0.9704	0.9748	0.9754
Normalized KD (ablation)	−7.59	−10.66	−11.92	−12.13	0.9567	0.9714	0.9757	0.9762
Bit-Adaptive KD	−7.90	−10.46	−11.74	−11.92	0.9580	0.9701	0.9753	0.9756
NF-KD ($\alpha_{\text{KD}} = 0.4$)	**−8.69**	−10.99	−12.10	−12.36	**0.9631**	0.9727	0.9762	0.9773
NF-KD (proposed)	−8.38	**−11.02**	**−12.43**	**−12.87**	0.9607	**0.9745**	**0.9800**	**0.9817**

Table 4. Beamforming performance at SNR = 20 dB across varying bit lengths.

Model	**32-bit**			**64-bit**			**128-bit**			**256-bit**		
	ρ	BG loss	SE	ρ	BG loss	SE	ρ	BG loss	SE	ρ	BG loss	SE
Baseline	0.9594	0.361	6.540	0.9704	0.262	6.573	0.9748	0.223	6.585	0.9754	0.217	6.587
NF-KD (proposed)	**0.9607**	**0.350**	**6.543**	**0.9745**	**0.225**	**6.585**	**0.9800**	**0.175**	**6.601**	**0.9817**	**0.161**	**6.606**
Improvement (Δ)	+0.0013	−0.011	+0.003	+0.0041	−0.037	+0.012	+0.0052	−0.048	+0.016	+0.0063	−0.056	+0.019

soft-label distribution—the very structure that KLD-based NF-KD exploits.

Effect of the KD weight. NF-KD with $\alpha_{\text{KD}} = 0.4$ achieves the best NMSE at 32 bits (−8.69 dB vs. −8.38 dB for $\alpha_{\text{KD}} = 0.3$), as stronger Teacher guidance compensates for severe information loss at low bit budgets. From 64 bits onward, however, $\alpha_{\text{KD}} = 0.3$ consistently surpasses $\alpha_{\text{KD}} = 0.4$, reaching −12.87 dB versus −12.36 dB at 256 bits: a smaller KD weight preserves the GT reconstruction signal needed for fine-tuning when the bit budget is sufficient. The proposed $\alpha_{\text{KD}} = 0.3$ thus offers the best trade-off across the full 1–256 bit range.

To bridge reconstruction accuracy and system utility, we evaluate the downstream beamforming performance of the reconstructed channel. Table 4 summarizes results at a representative transmit SNR of 20 dB. NF-KD consistently reduces beamforming gain loss across all bit lengths; at 256 bits, the loss drops from 0.217 dB to 0.161 dB, translating to a spectral efficiency gain of +0.019 bps/Hz. This validates the core hypothesis of Section 2.2: preserving the physical scale of CSI tensors during distillation is vital for accurate energy allocation. While prior deep-learning-based CSI feedback methods established the importance of reconstruction fidelity for downstream system performance (Wen et al., 2018; Guo et al., 2020), our results further show that scale preservation in the distillation loss itself yields measurable beamforming gains.

4.3. Ablation Study on Scale Preservation

To substantiate the necessity of the normalization-free design, we evaluate "Normalized KD," which shares the NF-KD configuration but applies Z-score normalization (zero mean, unit variance) before the KLD computation.

As shown in Table 3 and Figs. 2–3, Normalized KD underperforms NF-KD across all feedback bit lengths, and at 32 bits even degrades below the baseline in both NMSE and cosine similarity. This is consistent with the scale distortion analyzed in Section 2.2: removing absolute magnitude eliminates the gradient signal needed for low-bit energy alignment.

5. Conclusion

We proposed NF-KD, a variable-bit CSI feedback framework that preserves the channel's physical scale by omitting Z-score normalization in KD. Within a 500-epoch budget, NF-KD achieves −12.87 dB NMSE at 256 bits and a +0.019 bps/Hz spectral efficiency gain over the baseline, with consistent improvements across all feedback bit lengths. Ablation studies confirm that (i) Z-score normalization is the critical failure factor in KD-based CSI feedback, and (ii) KLD-based distillation outperforms pointwise MSE alignment by capturing the inter-element energy ratios encoded in the Teacher's soft-label distribution. These results indicate that scale-preserving KD is a lightweight, deployable enhancement for variable-bit CSI feedback in 6G Massive MIMO systems, requiring no additional UE-side computation. Future work will extend NF-KD to multi-cell and time-varying channel scenarios, and explore adaptive temperature scheduling for further low-bit improvements.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be

specifically highlighted here.

References

Alkhateeb, A. Deepmimo: A generic deep learning dataset for millimeter wave and massive MIMO applications. *arXiv preprint arXiv:1902.06435*, 2019.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. ISBN 978-0-471-24195-9.

Cui, Y., Guo, A., and Song, C. Transnet: Full attention network for csi feedback in fdd massive mimo system. *IEEE Wireless Communications Letters*, 11(5):903–907, 2022.

Cui, Y., Guo, J., Cao, Z., Tang, H., Wen, C.-K., Jin, S., Wang, X., and Hou, X. Lightweight neural network with knowledge distillation for CSI feedback. *IEEE Transactions on Communications*, 72(8):4917–4929, 2024.

Guo, J., Wen, C.-K., Jin, S., and Li, G. Y. 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.

Guo, J., Wen, C.-K., Jin, S., and Li, G. Y. Overview of deep learning-based CSI feedback in massive MIMO systems. *IEEE Transactions on Communications*, 70(12): 8017–8045, 2022.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ji, D. J. and Chung, B. C. Concrete feedback layers: Variable-length, bit-level CSI feedback optimization for FDD wireless communication systems. *IEEE Transactions on Wireless Communications*, 23(10):15353–15366, 2024.

Jindal, N. MIMO broadcast channels with finite-rate feedback. *IEEE Transactions on information theory*, 52(11): 5045–5060, 2006.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, L., Li, G. Y., Swindlehurst, A. L., Ashikhmin, A., and Zhang, R. An overview of massive MIMO: Benefits and challenges. *IEEE journal of selected topics in signal processing*, 8(5):742–758, 2014.

Tang, H., Guo, J., Matthaiou, M., Wen, C.-K., and Jin, S. Knowledge-distillation-aided lightweight neural network for massive MIMO csi feedback. In *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pp. 1–5. IEEE, 2021.

Wen, C.-K., Shih, W.-T., and Jin, S. Deep learning for massive MIMO CSI feedback. *IEEE Wireless Communications Letters*, 7(5):748–751, 2018.

Algorithm 1 Phase-Aware Teacher Pre-Training

Input: Training set $\mathcal{D} = \{H^{(i)}\}$, feedback bit budget $B = 256$
Output: Pre-trained Teacher parameters Θ_T

Initialize Teacher encoder $f_{enc}(\cdot;\Theta_{en})$ and decoder $f_{dec}(\cdot;\Theta_{de})$
for $epoch = 1, 2, ..., 1000$ **do**
  for each mini-batch $\{H_b\} \subset \mathcal{D}$ **do**
    $s_T \leftarrow f_{enc}(H_b;\Theta_{en})$
    $\hat{H}_T \leftarrow f_{dec}(s_T;\Theta_{de})$
    $\mathcal{L}_{MSE} \leftarrow ||\hat{H}_T - H_b||_F^2$ ▷ Frobenius reconstruction loss
    Update $\Theta_T \leftarrow \Theta_T - \eta \cdot \nabla_{\Theta_T}\mathcal{L}_{MSE}$ ▷ AdamW + cosine annealing
  end for
end for
Freeze Teacher parameters: $\Theta_T \leftarrow \Theta_T$ (no gradient)
return Θ_T

Algorithm 2 Normalization-Free Knowledge Distillation (NF-KD)

Input: Training set $\mathcal{D}$, frozen Teacher Θ_T, temperature $\tau = 5$, KD weight $\alpha_{\text{KD}} = 0.3$
Output: Trained Student parameters Θ_S

Initialize Student encoder/decoder Θ_S (with Concrete Feedback Layer + FBMU)
for $epoch = 1, 2, ..., 500$ **do**
  for each mini-batch $\{H_b\} \subset \mathcal{D}$ **do**
    $\hat{H}_T \leftarrow f_T(H_b;\Theta_T)$ ▷ Teacher forward pass (frozen)
    $\hat{H}_S \leftarrow f_S(H_b;\Theta_S)$ ▷ Student forward pass
    $v_T \leftarrow \text{flatten}(\hat{H}_T) \in \mathbb{R}^D$ ▷ Global phase map, $D = 262{,}144$
    $v_S \leftarrow \text{flatten}(\hat{H}_S) \in \mathbb{R}^D$ ▷ No normalization applied
    $P_T \leftarrow \text{Softmax}(v_T/\tau)$, $P_S \leftarrow \text{LogSoftmax}(v_S/\tau)$ ▷ Temperature scaling only
    $\mathcal{L}_{KD} \leftarrow \sum_i P_{T,i} \cdot \log(P_{T,i}/P_{S,i})$ ▷ KLD on raw tensors
    $\mathcal{L}_{MSE} \leftarrow ||\hat{H}_S - H_b||_F^2$ ▷ Ground-truth reconstruction loss
    $\mathcal{L}_{\text{total}} \leftarrow (1 - \alpha_{\text{KD}}) \cdot \mathcal{L}_{MSE} + \alpha_{\text{KD}} \cdot (\mathcal{L}_{KD} \cdot \tau^2)$ ▷ Loss ensemble
    Update $\Theta_S \leftarrow \Theta_S - \eta \cdot \nabla_{\Theta_S}\mathcal{L}_{\text{total}}$ ▷ AdamW optimizer
  end for
end for
return Θ_S

A. Source code

The complete PyTorch implementation for the proposed NF-KD framework, including the variable-bit FBMU environment, is anonymously available at:
https://anonymous.4open.science/r/NF-KD-C1CA

B. Algorithms

This section provides the detailed pseudocodes for the Phase-Aware Teacher Pre-Training (Algorithm 1) and the proposed Normalization-Free Knowledge Distillation (Algorithm 2).