CSI Compression Beyond Latents: End-to-End Hybrid Attention-CNN Networks with Entropy Regularization

Maryam Ansarifard*, Mostafa Rahmani†, Mohit K. Sharma‡, Kishor C. Joshi*, George Exarchakos*, Alister Burr†

*Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
Emails: {m.ansarifard, k.c.joshi, G.Exarchakos}@tue.nl
†School of Physics, Engineering, and Technology, University of York, York, YO10 5DD, UK
Emails: {rahmani.mostafa, alister.burr}@york.ac.uk
‡Directed Energy Research Center, Technology Innovation Institute, Abu Dhabi, UAE
Email: mohit.sharma@tii.ae

Abstract—**Massive MIMO systems rely on accurate Channel State Information (CSI) feedback to enable high-gain beamforming. However, the feedback overhead scales linearly with the number of antennas, presenting a major bottleneck. While recent deep learning methods have improved CSI compression, most overlook the impact of quantization and entropy coding, limiting their practical deployability. In this work, we propose an end-to-end CSI compression framework that integrates a Spatial Correlation-Guided Attention Mechanism with quantization and entropy-aware training. Our model effectively exploits the spatial correlation among the antennas, thereby learning compact, entropy-optimized latent representations for efficient coding. This reduces the required feedback bitrates without sacrificing reconstruction accuracy, thereby yielding a superior rate-distortion trade-off. Experiments show that our method surpasses existing end-to-end CSI compression schemes, exceeding benchmark performance by an average of 21.5% on indoor datasets and 18.9% on outdoor datasets. The proposed framework results in a practical and efficient CSI feedback scheme.**

Index Terms—**STQENet, CSI feedback, attention mechanism, entropy encoding, quantization, massive MIMO.**

I. INTRODUCTION

Massive Multiple-Input Multiple-Output (MIMO) systems have emerged as a cornerstone of next-generation wireless communication networks, offering significant gains in spectral and energy efficiency. A critical enabler of these gains is the availability of accurate Channel State Information (CSI) at the base station (gNB). However, in Frequency Division Duplex (FDD) massive MIMO systems, obtaining CSI feedback from the receiver to the transmitter introduces a prohibitive overhead which scales linearly with the number of antennas, thus posing a major bottleneck for practical deployments.

To address the CSI feedback problem, various approaches have been explored. Classical Compressive Sensing (CS) methods leverage the sparsity of the wireless channel to reduce feedback requirements [1] but often struggle under practical channel conditions and require complex recovery algorithms. Deep Learning (DL) techniques, particularly autoencoders,

have shown significant promise in learning compact CSI representations. CsiNet [2] pioneered Convolutional Neural Network (CNN)-based CSI compression using an Auto-Encoder (AE), with an encoder at the UE and a decoder at the gNB. However, CNN performance heavily depends on the size of the receptive field (or convolutional kernel). To address this, CsiNet+ [3] enhances the original CsiNet by enlarging the receptive field, better exploiting CSI sparsity in the angular-delay domain, while small kernels capture fine details more effectively. Based on this, CRNet [4] introduces a multi-resolution design, using different kernel sizes in both the encoder and decoder, enabling more adaptive and effective CSI feedback. Nevertheless, they still perform poorly in outdoor scenarios with high compression ratios (CRs), a common issue for existing algorithms. To address this, CsiNet+DNN [5] introduces additional layers and incorporates a different activation function to acquire more knowledge to enhance the performance of the RefineNet.

While many deep learning-based CSI compression methods enhance performance through architectural modifications, the CSI matrix is often treated as an unstructured data, thus neglecting the spatial correlations between antennas induced by the propagation environment. The attention mechanisms offer a more adaptive solution by explicitly modeling these correlations, enabling the network to focus on the most informative features. Notably, Attention-CSINet [6] pioneered the integration of attention mechanism into CNN-based CSI compression by introducing a module that generates a vector to describe the importance of each feature map, leading to superior performance. A two-layer transformer model, TransNet [7], demonstrated significant performance improvements but was ultimately deemed impractical due to its high computational complexity. To address this, CSIFormer [8] proposed a more efficient architecture employing locally grouped (windowed) self-attention, which lowered the computational burden, albeit with some reduction in performance. Furthermore, [9]
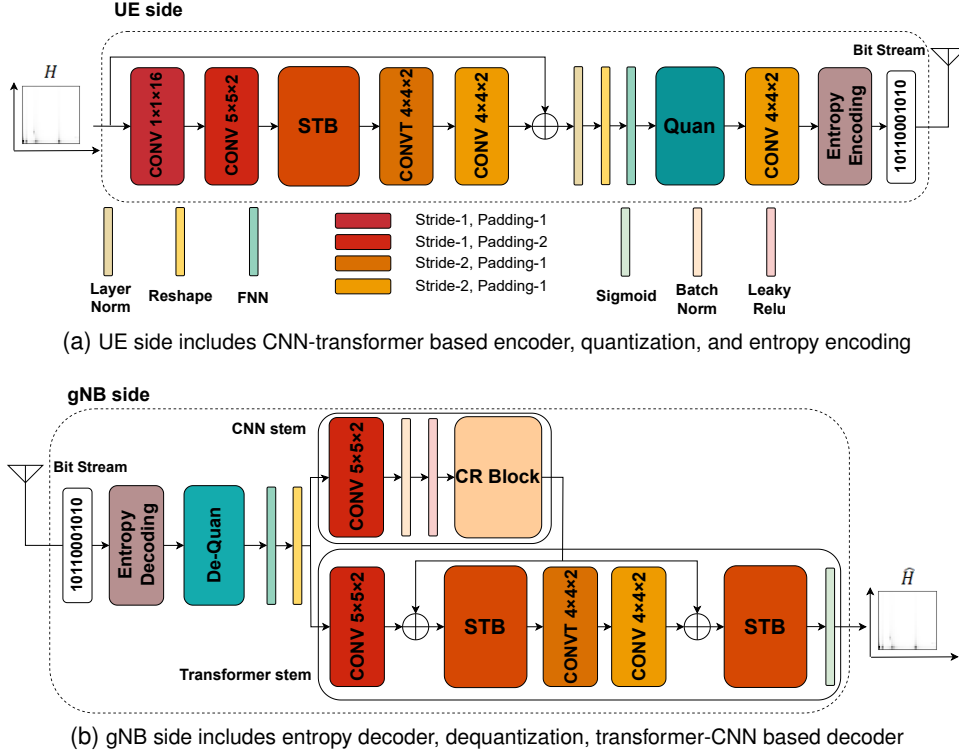
(a) UE side includes CNN-transformer based encoder, quantization, and entropy encoding

(b) gNB side includes entropy decoder, dequantization, transformer-CNN based decoder

Fig. 1: STQENet architecture for CSI feedback. "CONV" and "CONVT" denote convolutional and transposed convolutional layers, respectively. "STB" refers to a spatially separable attention transformer block. "CR Block" represents the multi-resolution CNN module from CRNet (see Fig. 2c.)

addresses the generalizability challenge of CSI compression by introducing AiANet, an attention-infused autoencoder that captures both global and local spatial features through self-attention mechanisms.

In practical systems, CSI is fed back as bitstreams. Directly transmitting 32-bit floating-point codewords from the encoder would incur excessive overhead. To reduce this, codewords are quantized before feedback [10], using uniform or non-uniform schemes to balance compression and accuracy. In [11], the authors propose a bit-level CSI feedback framework that compresses downlink CSI using an encoder, followed by non-uniform quantization to produce a discrete bitstream. To mitigate quantization noise, a quantizer-dequantizer pair is introduced, along with an offset neural network that reconstructs the compressed CSI before decoding. This approach improves feedback accuracy but involves a complex three-stage training process, including pre-training the autoencoder, training the offset network while freezing other parameters, and a final fine-tuning phase. [12] introduces a quantization module with bit allocation and propose a joint training method with an adaptive loss that balances quantization and reconstruction errors.

Beyond quantization, entropy coding is an essential step for further reducing the bitrate by exploiting the statistical redundancy in quantized outputs. By assigning shorter codes to more probable symbols, entropy coding—such as Huffman or arithmetic coding—enables bit-level compression without information loss. While such frameworks mark a step toward

practical CSI feedback, they typically treat quantization and entropy coding as separate post-processing steps, disconnected from the encoder-decoder training. This suboptimal separation can lead to performance degradation, as the encoder is not optimized to produce quantized representations that are both compact and entropy-efficient. [13] introduces a simple AE-based CSI compression framework incorporating an entropy bottleneck to optimize quantization, entropy coding, and reconstruction quality jointly. Although effective, our method surpasses theirs in terms of reconstruction accuracy at lower bit rates.

To address these limitations, we propose an end-to-end CSI compression framework, Spatially separable Transformer with Quantization and Entropy encoding Network (STQENet), that combines attention-based encoding, quantization, and entropy modeling to reduce feedback overhead in massive MIMO systems. The attention module captures inter-antenna correlations via spatially informative features, while quantization produces low-bit codewords for efficient transmission. A learned entropy model guides the compression by estimating the bit cost and introducing an entropy loss during training, enabling compact and accurate CSI representations.

The main difference between our work and the state-of-the-art lies in the joint integration of attention-based encoding, quantization, and entropy modeling within a single end-to-end trainable framework. Unlike AiANet [9], which focuses solely on attention mechanisms for spatial feature extraction without addressing quantization or entropy efficiency, our

method incorporates both, enabling practical low-bitrate feedback. Compared to [13], which applies an entropy bottleneck to a standard autoencoder without attention modeling, our approach employs a spatially separable transformer to better capture antenna-domain correlations. Experiments show that our method achieves superior rate-distortion performance compared to state-of-the-art transformer-based approaches, with lower bitrate and complexity.

The remainder of the paper is structured as follows. Section II introduces the system model. Section III describes the proposed architecture in detail. Section IV presents the experimental results along with a comprehensive discussion. Finally, Section V concludes the paper.

II. System Model

We consider an FDD system where the gNB is equipped with $N_t \gg 1$ transmit antennas and a single receiver antenna at a User Equipment (UE). The system operates in Orthogonal Frequency Division Multiplexing (OFDM) with $\tilde{N}_c$ subcarriers. The received signal at the n-th subcarrier is provided as follows:

$$y_n = \tilde{h}_n^H \mathbf{v}_n x_n + z_n, \tag{1}$$

where $\tilde{h}_n \in \mathbb{C}^{N_t \times 1}$, $\mathbf{v}_n \in \mathbb{C}^{N_t \times 1}$, $x_n \in \mathbb{C}$, $z_n \in \mathbb{C}$ denote the channel vector, precoding vector, data-bearing symbol, and additive noise of the n-th subcarrier, respectively. The CSI matrix in the spatial-frequency domain is given by $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \ldots, \tilde{\mathbf{h}}_{\tilde{N}_c}]^H \in \mathbb{C}^{\tilde{N}_c \times N_t}$. The number of feedback parameters is $2\tilde{N}_c N_t$. To reduce feedback overhead, we adopt the compression approach from [2], which sparsifies $\tilde{\mathbf{H}}$ in the angular-delay domain using a 2D Discrete Fourier Transform (DFT):

$$\bar{\mathbf{H}} = \mathbf{F_d} \tilde{\mathbf{H}} \mathbf{F_a^H}, \tag{2}$$

where $\mathbf{F_d}$ and $\mathbf{F_a}$ are 2D DFT matrices of dimensions $\tilde{N}_c \times \tilde{N}_c$ and $N_t \times N_t$, respectively. In the delay domain, the time delay between multipath arrivals is typically confined to a limited range, resulting in most of the significant energy being concentrated in the first $N_c \leq \tilde{N}_c$ rows of the channel matrix $\bar{\mathbf{H}}$. Assuming $\bar{\mathbf{H}}$ is complex-valued and exhibits similar sparsity, we retain its first N_c rows and then concatenate the real and imaginary parts to construct a real-valued matrix $\mathbf{H} \in \mathbb{R}^{2N_c \times N_t}$. With the sparsified channel matrix $\mathbf{H}$ obtained, it is processed through the encoder-decoder architecture depicted in Fig. 1 In this stage, $\mathbf{H}$ is compressed into a one-dimensional vector of size $M \times 1$. The compression ratio is defined as $\gamma = \frac{M}{2N_c N_t}$. To reduce the bitrate and prepare for practical transmission, the continuous-valued latent vectors are passed through a quantization module. This step maps the continuous values to discrete symbols. After quantization, the discrete symbols are entropy encoded using Huffman coding. Huffman encoding assigns shorter codewords to more frequent symbols, minimizing the average code length. This lossless compression step ensures that the quantized data is stored or transmitted with minimal redundancy.

This compressed representation is then transmitted from the UE to the gNB over the uplink. Upon reception, Huffman decoding and de-quantization are first applied to recover the compressed latent features, which are then passed through the decoder to reconstruct the channel matrix, denoted as $\hat{\mathbf{H}}$. The encoding and decoding processes are defined as follows:

$$\mathbf{s} = f_e(\mathbf{H}), \quad \hat{\mathbf{H}} = f_d(\mathbf{s}),$$

where f_e and f_d represent the functions of the encoder and decoder, respectively. Here, $\mathbf{s}$ denotes the compressed codeword, and $\hat{\mathbf{H}}$ is the reconstructed channel matrix estimated by the model.

III. STQENet architecture

We implement the CSI encoder and decoder within STQENet, as shown in Fig. 1, drawing inspiration from the STNet architecture [14]. Leveraging the attention mechanism, STNet achieves high performance in CSI feedback tasks. It effectively reduces the amount of feedback data while aggregating spatial-frequency domain CSI features. Moreover, its architecture captures long-range dependencies and exploits inter-antenna correlations within the channel matrix, enhancing the accuracy of the reconstructed CSI. [14] presents a detailed description of the STNet architecture. In this Section we explain each block in the architecture.

A. Spatially Separable Attention Transformer Block (STB)

The STB incorporates a hybrid attention mechanism consisting of three main stages: Locally-grouped Self-Attention (LSA), Global Sub-sampled Attention (GSA), and Multi-Layer Perceptron (MLP).

1) LSA: In the LSA stage, the input channel matrix (of size $L \times L$) is divided into non-overlapping windows of size $W \times W$, where $W = L/m$ and m is the number of partitions along each spatial dimension. Attention is computed independently within each window, reducing the complexity from $\mathcal{O}(L^4 d)$ to $\mathcal{O}(\frac{L^4}{m^4} d)$, where d is the feature dimension. However, because attention is local and restricted to each window, global dependencies across windows are lost.

2) GSA: To address this limitation without reverting to full global attention, the GSA mechanism is introduced. The output from LSA is passed through a CNN layer with stride W, effectively summarizing each window into a single spatial token and producing a feature map of size $m \times m$. This map serves as the keys and values for another layer of attention, while the LSA output continues to serve as the queries. The GSA thus reintroduces global context in a computationally efficient manner, with a complexity of $\mathcal{O}(m^2 L^2 d)$. The total complexity of the full attention mechanism becomes:

$$\mathcal{O}\left(\frac{L^4}{m^4} d + m^2 L^2 d\right) \tag{3}$$

3) MLP block: MLPs are essential for feature transformation and nonlinear mixing across channels (not across spatial locations, which is done by attention layers). This block has a linear layer followed by a Gaussian Error Linear Unit (GELU) non-linearity and another linear layer.
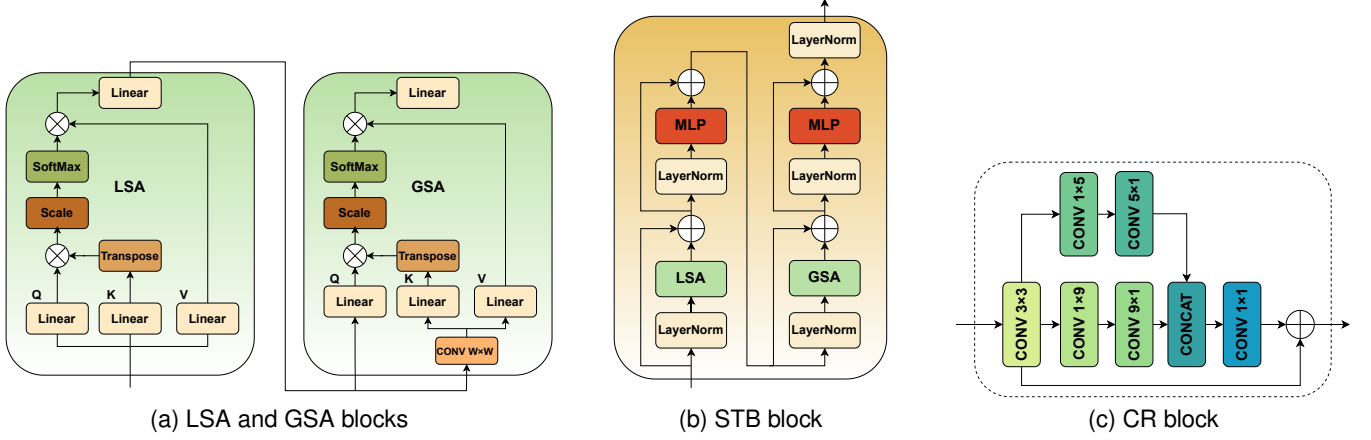
Fig. 2: Detailed blocks of the STQENet architecture, adopted from the STNet architecture [14].

B. CR Block

Beyond the attention block, the overall decoder architecture consists of two parallel stems: one transformer-based and the other convolution-based. This dual-branch design is intended to leverage the strengths of both components—while the transformer branch effectively captures global dependencies, the convolutional branch focuses on modeling local spatial features and enhances generalization through weight sharing and spatial invariance. As illustrated in Fig. 2c, the convolutional branch integrates multi-kernel convolutions to extract spatial features at multiple scales, further enriching the feature representation.

C. Quantization and Entropy

To enable end-to-end differentiable training while preserving the benefits of entropy coding, we adopt a learned probabilistic quantization model based on μ-law companding [15]. To address the non-differentiability of the quantization function, which poses a challenge for training neural networks via backpropagation, we employ the Straight-Through Estimator (STE) technique. In this approach, the true gradient is replaced with a constant during backpropagation, effectively enabling gradients to pass through the quantization step [16]. For a normalized output of the encoder, the μ-law compression function is defined as:

$$f(s) = \text{sgn}(s)\frac{\ln(1 + \mu|s|)}{\ln(1 + \mu)}. \tag{4}$$

Entropy is computed as the expected negative log-likelihood of the quantized latent variables given $\hat{s}$:

$$R = \mathbb{E}_{H \sim p_H}\left[-\log_2 \hat{p}_s(\hat{s})\right], \tag{5}$$

where p_H is the true distribution of CSI tensors and $\hat{p}_s$ models the probability of each quantized value. In our implementation, we adopt a factorized prior for the quantized latent variables. Specifically, $\hat{p}_s$ is modeled using an empirical histogram of symbol frequencies, and entropy coding is performed with Huffman coding.

IV. EXPERIMENTS AND NUMERICAL RESULTS

We examine a system configuration featuring 32 antennas at the gNB and a single antenna at the UE. For performance evaluation, we utilize the COST2100 dataset [17], focusing on two specific scenarios: an indoor picocellular environment operating at $5.3\,\text{GHz}$, and an outdoor rural setting at $300\,\text{MHz}$. The number of subcarriers is set to $N_c = 32$, with a window size $W = 8$, and the multi-head attention mechanism employs $P = 4$ heads. The dataset is divided into 100,000 samples for training, $30,000$ for validation, and $20,000$ for testing. We use a batch size of 200 and train the model for 1000 epochs. The optimization process is carried out using the Adam optimizer with a learning rate of 0.001, and the Mean Squared Error (MSE) and entropy loss are used as the loss function, as follows;
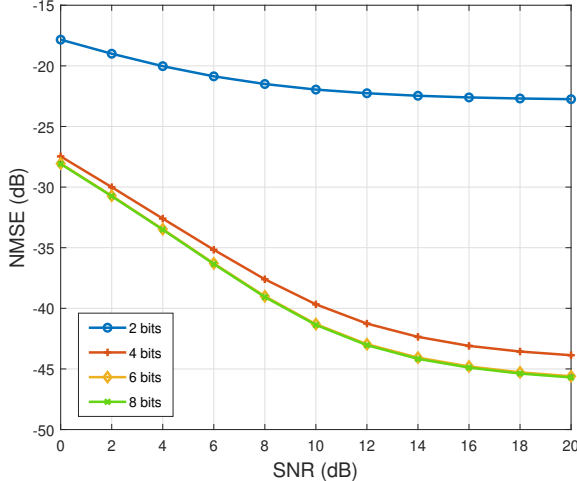
$$\mathcal{L}(\theta_e, \theta_d, \phi) = \frac{1}{B}\sum_{i=1}^{B}\|\mathbf{H}_i - \hat{\mathbf{H}}_i\|^2 + \lambda R, \tag{6}$$

where $\mathbf{H}$ is the input channel matrix, $\hat{\mathbf{H}}$ is the reconstructed channel matrix, and B is the batch size. θ_e, θ_d, and ϕ are the parameters of the encoder, decoder, and the entropy bottleneck, respectively. λ is the regularization parameter that decides the rate–distortion tradeoff, which is set to 10^{-3}. We adopt the Normalized Mean Squared Error (NMSE) as the evaluation metric defined as follows:
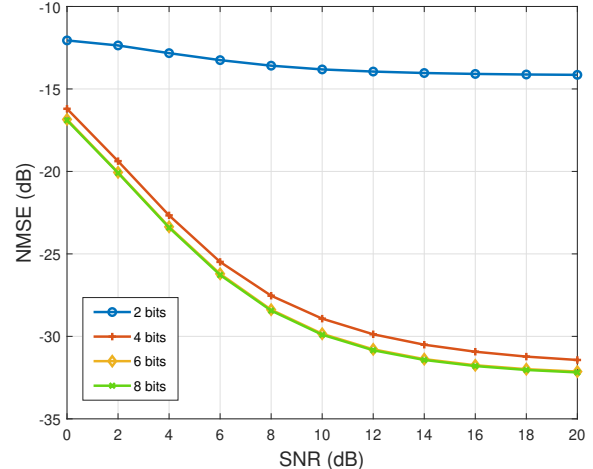
$$\text{NMSE} = \mathbb{E}\left\{\frac{\|\mathbf{H} - \hat{\mathbf{H}}\|^2}{\|\mathbf{H}\|^2}\right\}. \tag{7}$$

We apply $10\log_{10}(\cdot)$ to the expectation in Eq. 7 and report the NMSE in decibels (dB). Entropy is expressed in Bits Per Pixel (BPP), calculated by dividing the entropy value from Eq. 5 by 2×32^2 (2 refers to the real and imaginary part of the CSI data), which represents the spatial dimension of the CSI tensors. For the results in Fig. 4, the BPP is computed as:

$$\text{BPP} = \frac{k \times N}{2 \times 32^2}, \tag{8}$$

Fig. 3: NMSE vs. SNR at different quantization bit-rates.

where N and k are the number of quantization bits, and encoder output dimension. The source code for the simulation of our approach is available in [18].
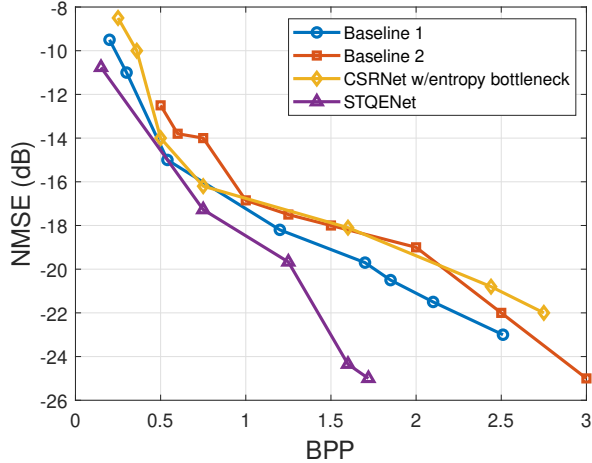
A. Result Analysis

We evaluate the reconstruction quality using the NMSE metric across a range of Signal-to-Noise Ratio (SNR) values from 0 dB to 20 dB. The evaluation is performed on both indoor and outdoor datasets under different quantization bit-rates (2, 4, 6, and 8 bits), as shown in Fig. 3.
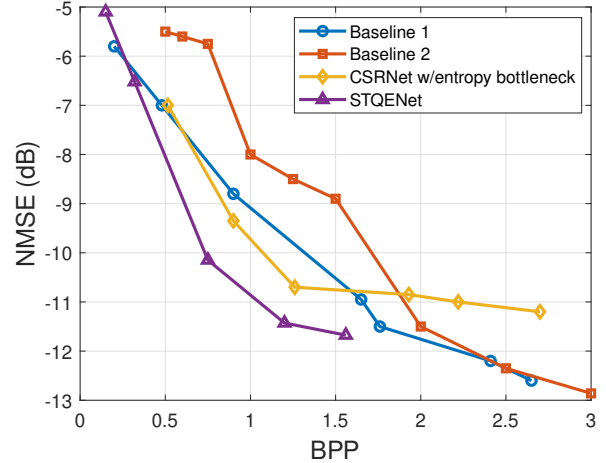
We observe in Fig. 3a that as the quantization bit-rate increases, the NMSE improves consistently across all SNR values. For example, the 8-bit quantization significantly outperforms the 2-bit version, especially at higher SNRs, confirming that finer quantization granularity leads to better channel reconstruction. For a fixed bit-rate, the NMSE decreases as SNR increases. This indicates that the model benefits from better signal quality, leading to more accurate CSI recovery. The curve corresponding to 2-bit quantization exhibits a performance floor, especially at higher SNRs. This suggests that coarse quantization becomes the dominant source of distortion, limiting the benefits from improved SNR. The trends observed in Fig. 3b for the outdoor dataset align closely with the indoor scenario, with some noticeable differences: The NMSE values in the outdoor dataset are generally higher than those of the indoor dataset at the same bit-rate and SNR. This is due to the more complex and variable propagation conditions in outdoor environments, which increases the difficulty of accurate CSI recovery. The performance gap between different quantization levels is more pronounced in the outdoor dataset. For example, the improvement from 2 bits to 8 bits is larger in the outdoor case compared to the indoor case, highlighting the importance of sufficient quantization resolution in challenging environments.

Figure 4 shows the NMSE performance versus BPP for various methods evaluated on both indoor and outdoor datasets.

The compared models include Baseline 1 [13], Baseline 2 [11], CSRNet [4], and the proposed STQENet architecture. To ensure a fair comparison, we integrated an entropy bottleneck into Baseline 2 and CSRNet, enabling them to function as end-to-end frameworks similar to our approach. For indoor dataset in Fig. 4a, STQENet consistently outperforms all baseline methods across most BPP values, especially in the mid-to-low rate region (0.5–1.5 BPP). This indicates that our method, achieves superior compression efficiency without sacrificing reconstruction quality. STQENet outperforms Baseline 1, which uses a simple autoencoder with different quantization but the same entropy model, by an average of 17.72%, demonstrating the benefits of using attention mechanisms in the autoencoder for better feature representation. Baseline 2, focusing solely on quantization without an entropy model, performs significantly worse, which STQENet outperforms it by 27.48%. This confirms the critical role of entropy modeling in achieving efficient compression. CSRNet with an entropy bottleneck shows competitive performance but still lags behind STQENet, which beats it by an average of 19.32%. This suggests that while CSRNet benefits from added entropy modeling, its underlying autoencoder may not be as expressive or adaptive as the attention-based design in STQENet. In the outdoor dataset, shown in Fig. 4b, STQENet once again delivers top-tier performance at low BPP levels (around 0.75), demonstrating strong robustness in handling more challenging compression scenarios. It exhibits a significant NMSE reduction as BPP rises, highlighting its effectiveness in environments with limited bandwidth. STQENet delivers moderate to substantial average improvements across benchmarks: 8.79% over Baseline 1, 40.11% over Baseline 2, and 7.68% over CSRNet with entropy bottleneck. The largest gain is observed against Baseline 2, highlighting the significant advantage of integrating entropy modeling with learned attention-based compression.

Fig. 4: NMSE vs BPP for different methods.

V. Conclusion

In this paper, we introduced STQENet, an end-to-end CSI compression framework designed to address the feedback bottleneck in massive MIMO systems. By leveraging a Spatial Correlation-Guided Attention Mechanism alongside quantization-aware and entropy-aware training, our approach effectively captures antenna dependencies and produces compact, efficiently coded representations. Experimental results across diverse datasets demonstrate that STQENet consistently outperforms existing compression methods. Our analysis reveals that performance improves steadily with finer quantization and higher SNR, while coarse quantization introduces distortion floors, especially in high-SNR or complex propagation scenarios. Compared to baselines with similar entropy modeling, STQENet achieves superior rate-distortion trade-offs due to its attention-based architecture and learned coding pipeline. These findings underscore the importance of jointly optimizing neural compression models with practical coding constraints in mind. Our work offers a promising step toward scalable and deployable CSI feedback solutions for next-generation wireless networks.

Acknowledgment

References

[1] P.-H. Kuo, H. Kung, and P.-A. Ting, "Compressive sensing based channel feedback protocols for spatially-correlated massive antenna arrays," in *2012 IEEE WCNC*. IEEE, 2012, pp. 492–497.

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

[3] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2827–2840, 2020.

[4] Z. Lu, J. Wang, and J. Song, "Multi-resolution CSI feedback with deep learning in massive MIMO system," in *IEEE ICC 2020*. IEEE, 2020, pp. 1–6.

[5] Y. Zhang, X. Zhang, and Y. Liu, "Deep learning based CSI compression and quantization with high compression ratios in FDD massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2101–2105, 2021.

[6] Q. Cai, C. Dong, and K. Niu, "Attention model for massive MIMO CSI compression feedback and recovery," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–5.

[7] Y. Cui, A. Guo, and C. Song, "Transnet: Full attention network for CSI feedback in FDD massive MIMO system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 903–907, 2022.

[8] X. Bi, S. Li, C. Yu, and Y. Zhang, "A novel approach using convolutional transformer for massive mimo csi feedback," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 1017–1021, 2022.

[9] K. Lou, H. Ji, and X. Wu, "Attention-infused autoencoder for massive MIMO CSI compression," in *2025 IEEE WCNC*. IEEE, 2025, pp. 1–6.

[10] C. Lu, W. Xu, S. Jin, and K. Wang, "Bit-level optimized neural network for multi-antenna channel quantization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 87–90, 2019.

[11] T. Chen, J. Guo, S. Jin, C.-K. Wen, and G. Y. Li, "A novel quantization method for deep learning-based massive MIMO CSI feedback," in *2019 IEEE GlobalSIP*. IEEE, 2019, pp. 1–5.

[12] M. Yin, S. Han, and C. Yang, "Quantization design for deep learning-based CSI feedback," *IEEE Wireless Commun. Lett.*, 2025.

[13] S. Ravula and S. Jain, "Deep autoencoder-based massive MIMO CSI feedback with quantization and entropy coding," in *2021 IEEE GLOBECOM*. IEEE, 2021, pp. 1–6.

[14] S. Mourya, S. Amuru, and K. K. Kuchi, "A spatially separable attention mechanism for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 40–44, 2022.

[15] Y. Jang, G. Kong, M. Jung, S. Choi, and I.-M. Kim, "Deep autoencoder based CSI feedback with feedback errors and feedback delay in FDD massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 833–836, 2019.

[16] J. Cai and L. Zhang, "Deep image compression with iterative non-uniform quantization," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 451–455.

[17] "COST2100 dataset," https://www.kaggle.com/datasets/forment/cost2100, MIT License.

[18] M. Ansarifard, "Entropy-aware CSI compression with hybrid CNN-attention autoencoder," https://github.com/Maryam-Ansarifard96, 2025, version 1.0.