

Computationally Efficient Neural Receivers via Axial Self-Attention

SaiKrishna Saketh Yellapragada*, Atchutaram K. Kocharalakota*, Mário Costa†, Esa Ollila*, Sergiy A. Vorobyov*

*Aalto University †Nokia Technologies

Abstract—Deep learning-based neural receivers are redefining physical-layer signal processing for next-generation wireless systems. We propose an axial self-attention transformer neural receiver designed for applicability to 6G and beyond wireless systems, validated through 5G-compliant experimental configurations, that achieves state-of-the-art block error rate (BLER) performance with significantly improved computational efficiency. By factorizing attention operations along temporal and spectral axes, the proposed architecture reduces the quadratic complexity of conventional multi-head self-attention from $O((TF)^2)$ to $O(T^2F + TF^2)$, yielding substantially fewer total floating-point operations and attention matrix multiplications per transformer block compared to global self-attention. Relative to convolutional neural receiver baselines, the axial neural receiver achieves significantly lower computational cost with a fraction of the parameters. Experimental validation under 3GPP Clustered Delay Line (CDL) channels demonstrates consistent performance gains across varying mobility scenarios. Under non-line-of-sight CDL-C conditions, the axial neural receiver consistently outperforms all evaluated receiver architectures, including global self-attention, convolutional neural receivers, and traditional LS-LMMSE at 10% BLER with reduced computational complexity per inference. At stringent reliability targets of 1% BLER, the axial receiver maintains robust symbol detection at high user speeds, whereas the traditional LS-LMMSE receiver fails to converge, underscoring its suitability for ultra-reliable low-latency (URLLC) communication in dynamic 6G environments and beyond. These results establish the axial neural receiver as a structured, scalable, and efficient framework for AI-Native 6G RAN systems, enabling deployment in resource-constrained edge environments.

Index Terms—deep learning, transformers, axial attention, 6G, radio access networks, neural receivers, self attention

I. INTRODUCTION

As wireless communications advance towards Sixth Generation (6G) Radio Access Networks (RAN), Deep Learning (DL)-based neural receivers are emerging as transformative Physical Layer (PHY) solutions, achieving superior performance compared to traditional model-based approaches. 3GPP Release 20 establishes Artificial Intelligence (AI) as integral to future air interface enhancements and network intelligence frameworks [2], [3]. However, neural receiver deployment in hard real-time systems faces strict latency and computational constraints, especially when processing Two Dimensional (2D)

time-frequency grids in Orthogonal Frequency Division Multiplexing (OFDM) systems. Since 3GPP Release 20 adopts OFDM-based waveforms for 6G PHY design, neural receiver architectures under this framework are highly relevant.

Convolutional Neural Network (CNN)-based neural receivers formulate the task of jointly optimizing channel estimation, equalization, and demapping to produce Log-Likelihood Ratios (LLRs) as supervised learning within a single architecture [4], [5]. Extensions to Multiple-Input-Multiple-Output (MIMO) leverage convolutional layers to capture time-frequency correlations and Graph Neural Network (GNN)-based modules to mitigate multi-user interference [6], [7]. Recent studies have shown that CNN-based neural receivers when subjected to model efficiency techniques such as Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ) exhibit notable resilience to ultra-low bit quantization [8], [9]. This characteristic makes neural network architectures promising candidates for deployment in future 6G systems, particularly in hardware-constrained edge devices.

Advances in transformer architectures have achieved remarkable success in domains like natural language processing and computer vision, especially with Large Language Models (LLMs), motivating their exploration for wireless communication applications [10]–[14]. In transformers, the Multi-Head Self-Attention (MHSA) mechanism enables global context modeling by computing attention across all positions in the input sequence, providing crucial advantages for wireless applications where channel responses exhibit dependencies across both time and frequency domains due to multipath propagation and Doppler effects. The authors of [11] demonstrated that a transformer architecture can effectively process OFDM Resource Grid (RG) by operating on non-overlapping tiles of Resource Blocks (RBs), applying MHSA across flattened resource elements and leveraging 2D positional encodings to inject temporal and spectral context that captures dependencies across the time-frequency domain.

However, when processing 2D time-frequency grids, the standard MHSA first flattens the RG into a single sequence, leading to a computational complexity of $O((TF)^2)$, where T and F denote the temporal and spectral extents of the processed RG, respectively. While the authors of [11] consider $T = 14$ OFDM symbols and $F = 12$ subcarriers for individual resource blocks in the form of non-overlapping tiles, practical systems must process significantly larger dimensions. Modern 5G NR systems support bandwidth parts up to 400 MHz

Corresponding author: saikrishna.yellapragada@aalto.fi

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

with $\approx 3,200$ subcarriers and slots of 14 symbols each. This quadratic scaling leads to computational bottlenecks for large time-frequency dimensions in modern OFDM systems, particularly in carrier aggregation and massive MIMO deployments where RGs span multiple frequency bands and spatial dimensions.

To address these limitations, the contributions of this work introduce an axial-attention neural receiver that factorizes self-attention along the time and frequency axes to produce output LLRs that is subsequently fed to the decoder. This design critically reduces the computational complexity to $\mathcal{O}(T^2F + TF^2)$ while preserving the model's ability to capture long-range temporal and spectral dependencies across large resource grids, a capability essential for robust channel estimation and signal detection. By mitigating the quadratic cost of standard MHSA, the axial neural-receiver enables energy-efficient, low-latency inference suitable for AI-RAN in 6G.

II. SYSTEM MODEL AND NEURAL RECEIVER

A. System Model

We consider an uplink Single-Input-Multiple-Output (SIMO) OFDM system. At the transmitter, an input bitstream is Low-Density Parity-Check (LDPC) encoded, mapped to symbols, and are arranged into a RG spanning N_{sym} OFDM symbols and N_{sc} subcarriers. The resources within this grid are indexed by the symbol index n and the subcarrier index k . Demodulation Reference Signals (DMRSs) are embedded at known time-frequency locations to facilitate channel estimation. After applying the Inverse Fast Fourier Transform (IFFT) and inserting a cyclic prefix, the signal is transmitted over a 3GPP Clustered Delay Line (CDL) channel [15].

At the receiver, after synchronization and cyclic prefix removal, the Fast Fourier Transform (FFT) is applied to each OFDM symbol. The received signal at symbol n and subcarrier k is given by

$$\mathbf{y}_{n,k} = \mathbf{h}_{n,k} x_{n,k} + \mathbf{n}_{n,k}, \quad (1)$$

where $\mathbf{y}_{n,k} \in \mathbb{C}^{N_{\text{rx}} \times 1}$ is the received signal vector, $\mathbf{h}_{n,k} \in \mathbb{C}^{N_{\text{rx}} \times 1}$ is the true channel frequency response, and $x_{n,k}$ is the transmitted symbol, normalized such that $\mathbb{E}[|x_{n,k}|^2] = 1$, with $\mathbb{E}[\cdot]$ being expectation operator. The term $\mathbf{n}_{n,k} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_{\text{rx}}})$ represents the additive white Gaussian noise vector.

B. Neural receiver

Neural network-based receivers learn to replace the entire signal processing chain, including channel estimation, equalization, and demapping, operating directly on post-FFT RGs to produce LLRs for decoding [4]. This work introduces an axial attention transformer architecture that reduces computational complexity through time-frequency self-attention factorization.

To benchmark the proposed axial attention transformer receiver (see Section III), we compare against two baselines:

- A global MHSA receiver that applies attention to the fully flattened time-frequency grid.
- CNN ResNet receiver.

All neural receiver models are trained end-to-end under an identical training and regularization protocol to ensure a controlled comparison with the parameters in Table I. Let $\hat{L}$ denote the neural receiver that maps the post-FFT received signal to predicted LLR. Training maximizes bit-metric decoding performance using the binary cross-entropy loss [16]:

$$\mathcal{L}_{\text{BCE}} = -\mathbb{E}[B \log \sigma(\hat{L}) + (1 - B) \log(1 - \sigma(\hat{L}))],$$

where B denotes the ground-truth coded bits and $\sigma(\cdot)$ is the sigmoid activation. Note that both $\hat{L}$ and $\mathcal{L}_{\text{BCE}}$ are parameterized by the weights of the neural network, denoted by $\mathbf{W}$. Therefore, the training process minimizes the loss function with respect to the optimization parameter, $\mathbf{W}$.

TABLE I: Training Parameters and Randomization

Parameter	Training	Randomization
Carrier Frequency	3.5 GHz	None
Channel Model	CDL-[A,B,E]	Uniform
RMS Delay Spread	10–100 ns	Uniform
UE Velocity	0–50 m/s	Uniform
SNR	0–15 dB	Uniform
Subcarrier Spacing	30 kHz	None
Modulation Scheme	64-QAM	None
Receive Antennas (N_{rx})	2	None
Code Rate	0.5	None
DMRS Configuration	3 rd , 12 th symbol	None
Optimizer	Adam	None

TABLE II: Testing Parameters

Parameter	Tested on
Channel Model	CDL-[C,D]
UE Velocity	Low: 0–5.1 m/s Medium: 10–20 m/s High: 25–40 m/s
SNR	0–12 dB
Number of Sub-Carriers	128
Number of OFDM symbols	14

III. AXIAL ATTENTION ARCHITECTURE FOR NEURAL RECEIVER DESIGN

In this section, we introduce the axial attention transformer-based neural receiver, which processes a resource grid of T OFDM symbols and F subcarriers to predict log-likelihood ratios $\hat{L}$. This is shown in Fig. 1. The input RG, of size $T \times F$, is crucial for capturing long-range dependencies in both time and frequency dimensions, making scalability with respect to the input size an important factor in achieving excellent performance.

To assess computational efficiency and scalability, we compare the axial attention architecture with a global self-attention based neural receiver. Both designs share the same end-to-end structure, shown in Fig. 1, comprising a 2D convolutional input projection, 2D learned positional encoding, six transformer blocks, and a 2D convolutional output projection. *The difference lies solely in the attention mechanism within the transformer blocks.* In the following subsections, we outline each module and analyze the computational complexity of both approaches.

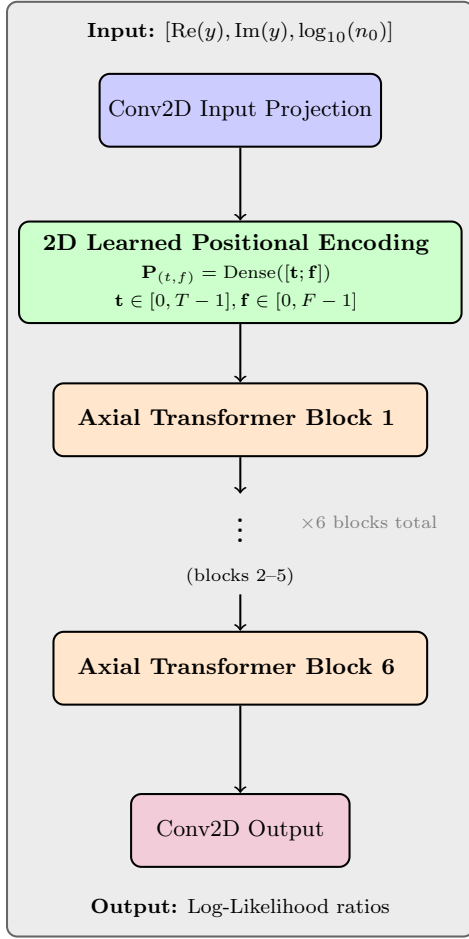


Fig. 1: Architecture of axial attention transformer-based neural receiver. It comprises a 2D convolutional input projection, 2D learned positional encoding, six transformer blocks, and a 2D convolutional output projection.

A. Convolutional 2D Input Projection

Following the standard neural receiver design, the complex-valued input tensor $\mathbf{Y} \in \mathbb{C}^{T \times F \times N_{\text{Rx}}}$ is decomposed into real ($\Re$) and imaginary parts ($\Im$), and concatenated with the logarithm of the noise power estimate N_0 . The resulting real-valued tensor is given by

$$\mathbf{Z} = [\Re(\mathbf{Y}), \Im(\mathbf{Y}), \log_{10}(N_0) \cdot \mathbf{1}_{T \times F \times 1}] \in \mathbb{R}^{T \times F \times (2N_{\text{Rx}}+1)}. \quad (2)$$

To project $\mathbf{Z}$ into the embedding space $\mathbb{R}^D$, we employ a 2D convolutional layer:

$$\text{Conv2D}(\mathbf{Z}) : \mathbb{R}^{T \times F \times (2N_{\text{Rx}}+1)} \rightarrow \mathbb{R}^{T \times F \times D}, \quad (3)$$

where D denotes the embedding dimension. The output of the 2D convolutional layer is denoted by $\mathbf{X}_{\text{conv}} \in \mathbb{R}^{T \times F \times D}$. Unlike linear embeddings commonly used in sequence tasks, the convolutional projection leverages the local spatial structure of OFDM resource grids shaped by channel coherence and spectral correlation. The convolution preserves spatial dimensions while mapping each time-frequency position (t, f)

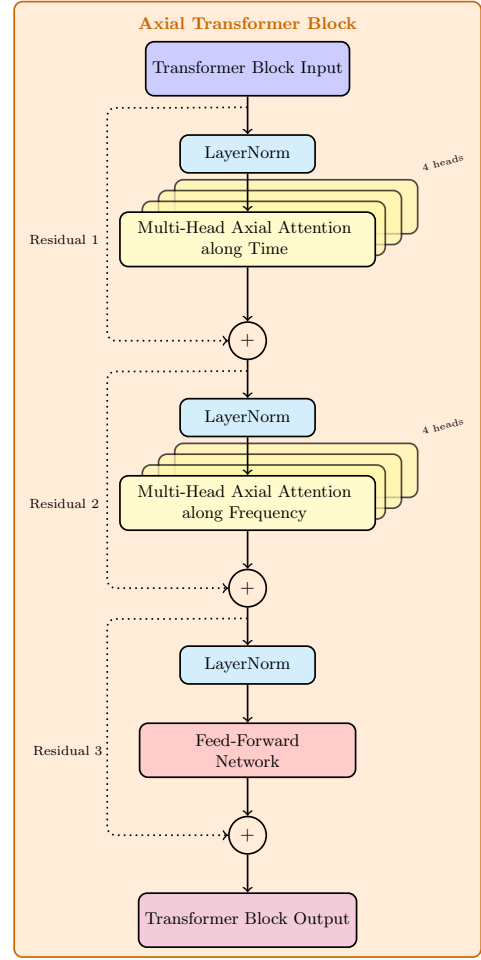


Fig. 2: Axial transformer block with sequential time-axis and frequency-axis multi-head attention preceded by layer normalization. Factorized attention operations reduce computational complexity while maintaining long-range dependency modeling through residual connections.

to a D -dimensional feature vector incorporating information from its local spatial neighborhood.

B. Learned Positional Encoding

Transformer architectures are inherently permutation-invariant, requiring explicit positional information to distinguish spatial locations in the time-frequency grid. We employ learned 2D positional encoding additively combined with the convolutional features:

$$\mathbf{X} = \mathbf{X}_{\text{conv}} + \mathbf{P} \in \mathbb{R}^{T \times F \times D}, \quad (4)$$

where $\mathbf{P} \in \mathbb{R}^{T \times F \times D}$ denotes the positional encoding tensor with learnable parameters. Unlike fixed sinusoidal encodings, learned embeddings adapt to the spatial correlation patterns of wireless channels during training. The resulting tensor $\mathbf{X}$ serves as input to the transformer blocks.

C. Preliminaries of Multi-Head Self-Attention Mechanisms

The attention mechanism operates on the positionally-encoded tensor $\mathbf{X} \in \mathbb{R}^{T \times F \times D}$, where D denotes the embedding dimension. Multi-head attention decomposes this representation into H parallel subspaces with head dimension $d_h = D/H$. For the h -th attention head, $h \in \{1, \dots, H\}$, three learnable linear transformations project the input features into query, key, and value representations:

$$\mathbf{Q}^{(h)} = \mathbf{X}\mathbf{W}_Q^{(h)}, \quad \mathbf{K}^{(h)} = \mathbf{X}\mathbf{W}_K^{(h)}, \quad \mathbf{V}^{(h)} = \mathbf{X}\mathbf{W}_V^{(h)}, \quad (5)$$

where $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{D \times d_h}$ are trainable weight matrices, and $\mathbf{Q}^{(h)}, \mathbf{K}^{(h)}, \mathbf{V}^{(h)} \in \mathbb{R}^{T \times F \times d_h}$. The query tensor $\mathbf{Q}^{(h)}$ encodes features to be attended to, the key tensor $\mathbf{K}^{(h)}$ provides similarity scores, and the value tensor $\mathbf{V}^{(h)}$ contains representations to be aggregated.

D. Axial Self-Attention

Exploiting the separable 2D structure of OFDM resource grids, axial attention factorizes global attention into sequential time-axis and frequency-axis operations, both operating on the projections defined in (5). We adopt the notation $\mathbf{Q}_{:,f}^{(h)}$ to denote the $T \times d_h$ matrix slice *along the time dimension* for fixed subcarrier f , and $\mathbf{Q}_{t,:}^{(h)}$ to denote the $F \times d_h$ matrix slice *along the frequency dimension* for fixed OFDM symbol t . The same notation applies to $\mathbf{K}^{(h)}$ and $\mathbf{V}^{(h)}$.

Time-Axis Attention. For each subcarrier $f \in \{1, \dots, F\}$, time-axis attention processes slices $\mathbf{Q}_{:,f}^{(h)}, \mathbf{K}_{:,f}^{(h)}, \mathbf{V}_{:,f}^{(h)} \in \mathbb{R}^{T \times d_h}$ as follows:

$$\mathbf{A}_{\text{time},f}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}_{:,f}^{(h)}(\mathbf{K}_{:,f}^{(h)})^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{T \times T}, \quad (6)$$

$$\mathbf{Y}_{\text{time},f}^{(h)} = \mathbf{A}_{\text{time},f}^{(h)} \mathbf{V}_{:,f}^{(h)} \in \mathbb{R}^{T \times d_h}. \quad (7)$$

Multi-head aggregation via concatenation and linear projection with learnable output matrix $\mathbf{W}_O \in \mathbb{R}^{D \times D}$ yields

$$\text{Att}_{\text{time}}(\mathbf{X})_{:,f} = \text{Concat}(\mathbf{Y}_{\text{time},f}^{(1)}, \dots, \mathbf{Y}_{\text{time},f}^{(H)})\mathbf{W}_O \in \mathbb{R}^{T \times D}. \quad (8)$$

Stacking across all F subcarriers produces $\text{Att}_{\text{time}}(\mathbf{X}) \in \mathbb{R}^{T \times F \times D}$.

Frequency-Axis Attention. Analogously, for each OFDM symbol $t \in \{1, \dots, T\}$, frequency-axis attention operates on slices $\mathbf{Q}_{t,:}^{(h)}, \mathbf{K}_{t,:}^{(h)}, \mathbf{V}_{t,:}^{(h)} \in \mathbb{R}^{F \times d_h}$ as follows:

$$\mathbf{A}_{\text{freq},t}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}_{t,:}^{(h)}(\mathbf{K}_{t,:}^{(h)})^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{F \times F}, \quad (9)$$

$$\mathbf{Y}_{\text{freq},t}^{(h)} = \mathbf{A}_{\text{freq},t}^{(h)} \mathbf{V}_{t,:}^{(h)} \in \mathbb{R}^{F \times d_h}. \quad (10)$$

Multi-head aggregation yields

$$\text{Att}_{\text{freq}}(\mathbf{X})_{t,:} = \text{Concat}(\mathbf{Y}_{\text{freq},t}^{(1)}, \dots, \mathbf{Y}_{\text{freq},t}^{(H)})\mathbf{W}_O \in \mathbb{R}^{F \times D}, \quad (11)$$

with stacking producing $\text{Att}_{\text{freq}}(\mathbf{X}) \in \mathbb{R}^{T \times F \times D}$.

Sequential Composition. The axial transformer block applies both operations sequentially with residual connections:

$$\mathbf{X} \leftarrow \mathbf{X} + \text{Att}_{\text{time}}(\mathbf{X}), \quad (12)$$

$$\mathbf{X} \leftarrow \mathbf{X} + \text{Att}_{\text{freq}}(\mathbf{X}). \quad (13)$$

Time-axis attention captures temporal dependencies among OFDM symbols, subsequently refined by frequency-axis attention modeling spectral correlations.

E. Axial Transformer Block Architecture

The axial transformer block employs a pre-normalization architecture with three sequential operations using $H = 4$ attention heads. This is illustrated in Fig. 2. First, the input undergoes Layer Normalization (LN) followed by multi-head axial attention along the time dimension. Second, another LN precedes multi-head axial attention along the frequency dimension. Third, LN is applied before a feed-forward network. Each operation includes a residual connection to facilitate gradient flow [17], while LN normalizes features across the embedding dimension to stabilize training [18].

The complete neural receiver architecture (Fig. 1) stacks six such transformer blocks between 2D convolutional input projection with learned positional encoding and 2D convolutional output projection. This end-to-end design outputs LLRs for subsequent error-correction decoding.¹

IV. GLOBAL SELF-ATTENTION

Conventional MHSA applies attention globally by flattening the time-frequency grid $\mathbf{X}$ into a one-dimensional sequence of length $N = TF$. The flattened projections $\bar{\mathbf{Q}}^{(h)}, \bar{\mathbf{K}}^{(h)}, \bar{\mathbf{V}}^{(h)} \in \mathbb{R}^{N \times d_h}$ undergo scaled dot-product attention as follows:

$$\mathbf{A}^{(h)} = \text{softmax}\left(\frac{\bar{\mathbf{Q}}^{(h)}(\bar{\mathbf{K}}^{(h)})^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{N \times N}, \quad (14)$$

$$\mathbf{Y}^{(h)} = \mathbf{A}^{(h)} \bar{\mathbf{V}}^{(h)} \in \mathbb{R}^{N \times d_h}. \quad (15)$$

The final output concatenates H head-specific representations with learned projection:

$$\text{Att}_{\text{global}}(\mathbf{X}) = \text{Concat}(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(H)})\mathbf{W}_O \in \mathbb{R}^{N \times D}. \quad (16)$$

V. COMPUTATIONAL COMPLEXITY AND MODEL EFFICIENCY

The dominant computational cost in attention mechanisms arises from the pairwise similarity computation. For global MHSA, the quadratic dependence $\mathcal{O}(N^2 D) = \mathcal{O}(T^2 F^2 D)$ scales poorly for large resource grids. Axial attention factorizes this computation along temporal and spectral dimensions, processing F independent temporal sequences of length T and T independent spectral sequences of length F , reducing complexity to

$$\mathcal{O}(FT^2 D + TF^2 D) = \mathcal{O}(TFD(T + F)). \quad (17)$$

¹Work in Progress : Code will be available at https://github.com/saiksaketh/efficient_neural_rx

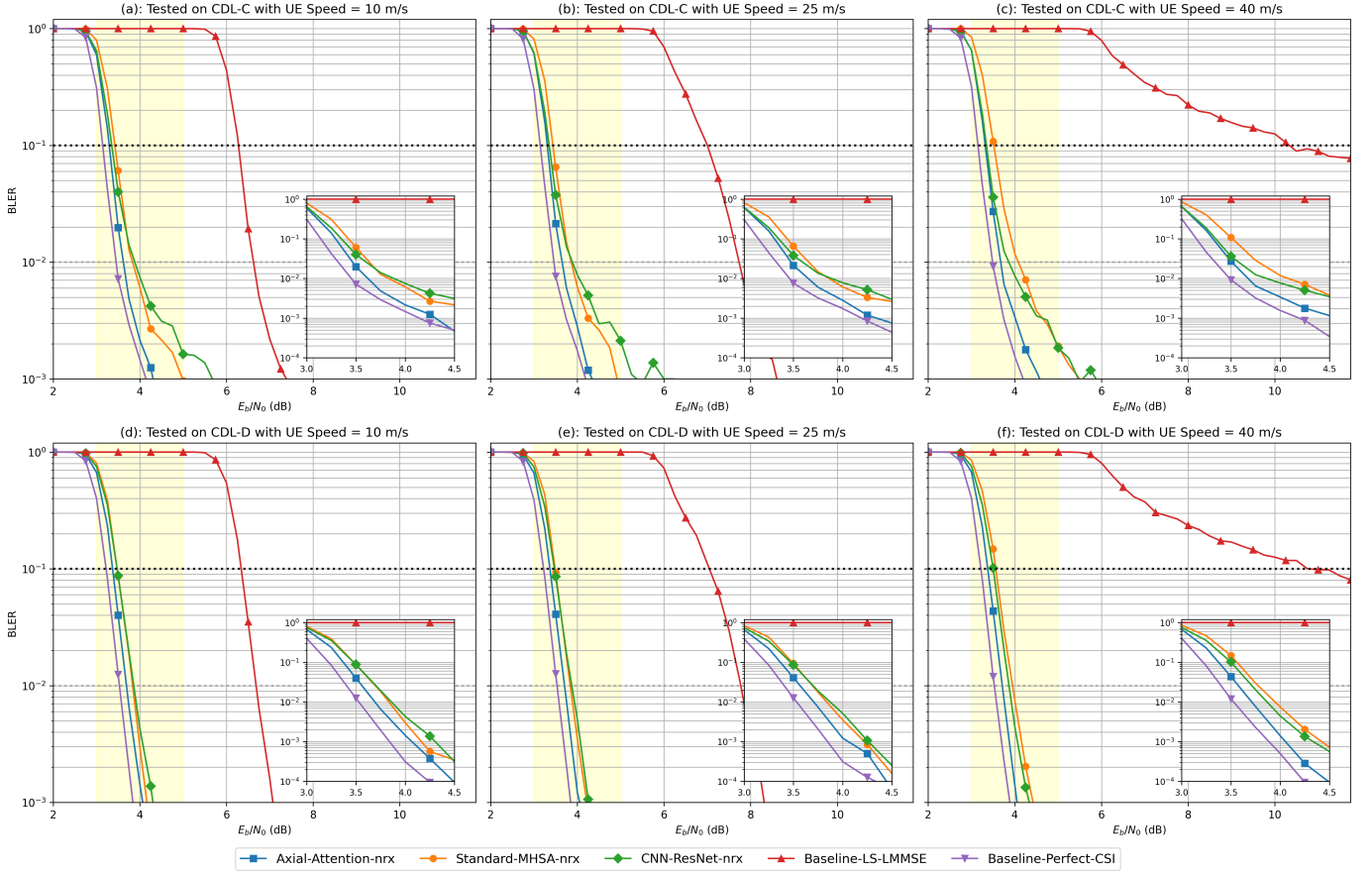


Fig. 3: BLER performance under CDL-C Non-LoS (NLoS) and CDL-D Line-of-Sight (LOS) channels at user velocities 10–40 m/s. Axial attention achieves 0.12–0.40 dB gain over standard MHSA and outperforms CNN-ResNet across all mobility scenarios at 10% and 1% BLER.

The reduction factor relative to global attention is $\frac{TF}{T+F}$. For 5G NR parameters ($T = 14$ symbols, $F = 128$ subcarriers, $D = 128$), axial attention achieves $12.6\times$ complexity reduction.

VI. NUMERICAL STUDY

In this section, we compare the proposed axial neural receiver against the standard MHSA receiver, CNN-ResNet, Least-Squares (LS)-Linear Minimum Mean Squared Error (LMMSE) receiver equalization with soft demapping and an ideal receiver with perfect Channel State Information (CSI). We report Block Error Rate (BLER) across varying User Equipment (UE) velocity for the aforementioned receiver architectures. The training has been done on NVIDIA A40 GPUs, and evaluations were done using Sionna, a hardware-accelerated differentiable open-source library for research on communication systems [19].

Performance in NLOS Scenario: The upper panel of Fig. 3 illustrates BLER performance under NLoS conditions using the CDL-C channel model. The axial neural receiver achieves 10% BLER at 3.25–3.30 dB across velocities from 10 to 40 m/s, outperforming LS-LMMSE by 2.96–6.80 dB

(increasing with mobility), standard MHSA by 0.15–0.25 dB, and CNN-ResNet by approximately 0.1 dB. The widening performance gap over LS-LMMSE at higher velocities indicates superior channel tracking under high-mobility urban NLoS propagation. At the stringent 1% BLER target, the axial receiver requires 3.60–3.70 dB, outperforming standard MHSA by 0.25–0.40 dB and CNN-ResNet by 0.20–0.30 dB. Notably, LS-LMMSE fails to reach 1% BLER at 40 m/s, whereas the axial receiver maintains reliable performance at 3.70 dB Signal-to-Noise-Ratio (SNR), demonstrating substantial advantages for ultra-reliable low-latency communication.

Performance in LOS Scenario: The lower panel of Fig. 3 shows performance under LOS conditions using the CDL-D channel model. The axial receiver achieves 10% BLER at 3.35–3.37 dB across mobility scenarios, outperforming LS-LMMSE by 2.90–7.63 dB (increasing with mobility), standard MHSA by 0.12–0.19 dB, and CNN-ResNet by 0.11–0.2 dB. At 1% BLER, the axial receiver requires 3.55–3.70 dB, achieving gains of 0.15–0.25 dB over both standard MHSA and CNN-ResNet. Similar to NLoS conditions, LS-LMMSE fails at 40 m/s while the axial receiver maintains robust performance. Across both LOS and NLoS scenarios, the axial

attention mechanism provides consistent improvements over standard MHSA and CNN-ResNet, with the primary advantage being substantially reduced computational complexity through factorized attention operations.

VII. MODEL EFFICIENCY

Table III summarizes model efficiency metrics for the evaluated architectures. Transformer-based receivers achieve substantial parameter reduction relative to the CNN-ResNet baseline, with axial and standard MHSA models requiring 16.5% and 12.3% of the CNN parameter count, respectively. Despite this parameter efficiency, standard MHSA incurs 9.40 GFLOPs per inference, nearly 80% of the CNN-ResNet cost (11.81 GFLOPs), due to the quadratic complexity of global attention over flattened time-frequency grids. In contrast, the axial architecture achieves 3.34 GFLOPs, representing a $2.81\times$ reduction compared to standard MHSA and $3.54\times$ reduction compared to CNN-ResNet, while maintaining superior BLER performance.

The axial architecture incurs a $1.3\times$ parameter increase over standard MHSA, arising from factorized attention operations requiring separate projection matrices for temporal and spectral dimensions. Specifically, standard MHSA employs shared query-key-value projections $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{D \times d_h}$ applied uniformly across the flattened time-frequency grid, while axial attention maintains separate projections $\{\mathbf{W}_{Q,\text{time}}^{(h)}, \mathbf{W}_{K,\text{time}}^{(h)}, \mathbf{W}_{V,\text{time}}^{(h)}\}$ and $\{\mathbf{W}_{Q,\text{freq}}^{(h)}, \mathbf{W}_{K,\text{freq}}^{(h)}, \mathbf{W}_{V,\text{freq}}^{(h)}\}$ for temporal and spectral operations, doubling the projection layer parameter count. Despite this modest parameter overhead, the substantial computational efficiency gain with 64% fewer FLOPs than standard MHSA with only 33% more parameters establishes axial neural receiver as a favorable trade-off to enable low power wireless communications.

TABLE III: Model complexity comparison of neural receiver architectures.

Receiver Model	Learnable Parameters	GFLOPs	Size (MB)
Axial-Transformer	1,600,902	3.34	6.11
Standard MHSA	1,196,898	9.40	4.57
CNN-ResNet	9,714,182	11.81	37.06

VIII. CONCLUSION AND FUTURE WORK

This work introduces axial attention as a computationally efficient foundation for neural receivers toward realizing the vision of an AI-native air interface for 6G. By factorizing self-attention along temporal and spectral dimensions, the proposed architecture overcomes the scalability limitations of conventional transformers while retaining the global context modeling essential for robust physical-layer processing. The demonstrated efficiency and consistent gains under high-mobility conditions highlight axial attention as a practical design for AI-based wireless systems, especially where traditional model-based receivers degrade under dynamic channels.

The proposed design achieves a balanced trade-off between complexity and detection accuracy, making it suitable for deployment in mobile edge environments with stringent latency and energy budgets. Its robustness across propagation conditions and ability to meet 1% BLER with the lowest computational cost demonstrate its potential for ultra-reliable low-latency and extended reality applications. These results collectively position axial neural receiver as a strong foundation for AI-native RAN systems, enabling intelligent and reliable signal processing at the network edge.

Future work will explore quantization and model compression. The reduced attention matrix dimensions inherent to axial attention are more amenable to low-precision arithmetic, potentially enabling low-bit quantization while preserving radio performance.

ACKNOWLEDGEMENTS

The work of first author has been supported in parts by Research Council of Finland (grant no:359848) and 6GARROW project: No. RS-2024-00435652. The authors also thank developers from NVIDIA for their contributions to the Sionna Open-Source project that enable this work. Thanks to DICE-ELEC-IT team at Aalto University for GPU support.

REFERENCES

- [1] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial Attention in Multidimensional Transformers," <https://arxiv.org/abs/1912.12180>, 2019.
- [2] X. Lin, "A Tale of Two Mobile Generations: 5G-Advanced and 6G in 3GPP Release 20," *IEEE Communications Standards Magazine*, pp. 1–9, 2025.
- [3] M. Shafi, E. Larsson, X. Lin, D. Panaitopol, S. Parkvall, F. Rosteix-Jacquet, and A. Toskala, "Industrial Viewpoints on RAN Technologies for 6G," <https://arxiv.org/pdf/2508.08225>, 2025.
- [4] M. Honkala, D. Korpi, and J. Huttunen, "DeepRx: Fully Convolutional Deep Learning Receiver," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3925–3940, 2021.
- [5] F. Ait Aoudia and J. Hoydis, "End-to-End Learning for OFDM: From Neural Receivers to Pilotless Communication," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 1049–1063, 2022.
- [6] S. Cammerer, F. A. Aoudia, J. Hoydis, A. Oeldemann, A. Roessler, T. Mayer, and A. Keller, "A Neural Receiver for 5G NR Multi-User MIMO," in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023.
- [7] D. Korpi, M. Honkala, J. Huttunen, and V. Starck, "DeepRx MIMO: Convolutional MIMO Detection with Learned Multiplicative Transformations," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–7.
- [8] S. S. Yellapragada, E. Ollila, and M. Costa, "Efficient Quantization-Aware Neural Receivers: Beyond Post-Training Quantization," *arXiv preprint arXiv:2509.13786*, 2025.
- [9] S. S. Yellapragada, E. Ollila, and M. Costa, "Efficient Deep Neural Receiver with Post-Training Quantization," *arXiv:2508.06275, Accepted for IEEE 59th Asilomar Conference on Signals, Systems, and Computers, October, 2025*.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, E. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.
- [11] Y. Kawai and R. Koodli, "A Unified Transformer Architecture for Low-Latency and Scalable Wireless Signal Processing," *arXiv preprint arXiv:2508.17960*, 2025.
- [12] A. K. Kocharalakota, S. A. Vorobyov, and R. W. Heath, "Attention neural network for downlink cell-free massive mimo power control," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, 2022.
- [13] A. K. Kocharalakota, S. A. Vorobyov, and R. W. Heath, "Pilot contamination aware transformer for downlink power control in cell-free massive mimo networks," <https://arxiv.org/abs/2411.19020>, 2024.

- [14] T Zhang, S. A. Vorobyov, D. J. Love, T. Kim, and K. Dong, "Pilot contamination-aware graph attention network for power control in cfm-mimo," <https://arxiv.org/abs/2506.00967>, 2025.
- [15] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," Tech. Rep. TR 38.901, 3rd Generation Partnership Project (3GPP), 2020, version 16.1.0.
- [16] K.P. Srinath and J. Hoydis, "Bit-Metric Decoding Rate in Multi-User MIMO Systems: Theory," *IEEE Transactions on Wireless Communications*, vol. 22, no. 11, pp. 7961–7974, 2023.
- [17] K. He and X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.
- [18] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [19] J. Hoydis, S. Cammerer, F. Ait Aoudia, M. Nimier-David, L. Maggi, G. Marcus, A. Vem, and A. Keller, "Sionna," 2022, <https://nvlabs.github.io/sionna/>.
- [20] Christopher M. Bishop, *Deep Learning Foundations and Concepts*, Springer, 2023.
- [21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.