Transformer-based Pilot-to-Prediction for MIMO-OFDM Systems

1st Louis Lagona, 2nd Carsten Bockelmann, 3rd Armin Dekorsy *Department of Communications Engineering*
University of Bremen, Germany
Email: {lagona, bockelmann, dekorsy}@ant.uni-bremen.de

***Abstract*—Accurate channel state information (CSI) is essential for spectral efficiency (SE) in 6G multiple-input multiple-output orthogonal frequency-division multiplexing (MIMO-OFDM) systems, yet conventional methods incur substantial pilot overhead. This letter introduces KronFormer, a pilot-to-prediction (P2P) Transformer with factorized spatial-temporal attention aligned to the Kronecker structure of MIMO channel correlations. Unlike unfactorized Transformers that flatten spatial dimensions, KronFormer preserves the 4-D tensor structure throughout processing and decouples spatial and temporal attention, avoiding the quadratic complexity of joint spatial-temporal processing. Simulations on COST 259 channels demonstrate that KronFormer reduces attention complexity by $20\times$ and mean squared error (MSE) by up to $87\times$ over unfactorized Transformers in 8×8 MIMO. KronFormer limits SE degradation to 10–13% from 50 to 200 km/h, a $3\times$ reduction compared to the 34–39% loss of conventional Wiener filters. The architecture enables direct multi-step inference for scalable 6G channel prediction.**

***Index Terms*—Channel Prediction, Transformer, Frequency-Selective Fading, MIMO-OFDM, Pilot Overhead**

I. Introduction

Emerging beyond-5G and 6G demand substantial gains in spectral efficiency (SE), placing strict demands on reducing pilot overhead in orthogonal frequency-division multiplexing (OFDM) systems employing multiple-input multiple-output (MIMO) technology [1]. Traditional pilot-based channel estimation consumes time-frequency resources, limiting effective throughput and SE. Channel prediction is an alternative, enabling timely channel state information (CSI) acquisition from past pilots without requiring continuous pilot transmission [2].

Classical predictors such as the Wiener filter [3] and Kalman filter [4] rely on second-order statistics. While efficient implementations exist [3], these methods often exhibit lower accuracy than neural network (NN) solutions in complex MIMO-OFDM scenarios [5], [6]. The long short-term memory (LSTM)-based predictor [5] performs well but suffers from error propagation. Transformers are adopted in wireless sequence prediction because self-attention captures long-range dependencies and enables direct low-latency inference [7].

Recent work explored time sequences without capturing joint spatial-temporal-frequency dynamics. For instance, temporal prediction is formulated in [8], while LinFormer [9] addresses time-aware MIMO prediction but assumes statistically independent coefficients and overlooks frequency selectivity. Channelformer [10] emphasizes estimation rather than prediction. Generic mechanisms like crossover attention [11] ignore the separable Kronecker structure of MIMO covariance. Factorized attention has succeeded in video understanding [12], [13] but has not been applied to MIMO channel prediction.

However, existing methods [3], [5], [9] typically follow a two-stage estimate-then-predict pipeline that accumulates inter-stage errors, requires pilots throughout the observation window, and incurs high complexity or sequential processing delay, while failing to leverage Kronecker-separable spatial-temporal-frequency structure for pilot reduction, thus hindering 6G reliability and low-latency operation.

To address these limitations, extending the single-input single-output OFDM framework in [6] to MIMO systems, we propose KronFormer, a Transformer-based pilot-to-prediction (P2P) architecture with factorized spatial-temporal attention tailored for MIMO-OFDM. Leveraging the Transformer's self-attention, the P2P model simultaneously captures spatial-temporal-frequency channel dynamics and correlations across antennas. Unlike interpolation-based schemes, our predictor reconstructs the complete frequency-selective MIMO channel matrix for all transmit-receive antenna pairs and forthcoming OFDM symbols using only past pilot observations. This joint spatial-temporal-frequency prediction is enabled by keeping antenna dimensions explicit (rather than flattening), consistent with the Kronecker decomposition of MIMO covariance [14].

II. System Model

We consider a MIMO-OFDM system with N_t transmit and N_r receive antennas. The received signal at antenna r for symbol $m \in \{1,\dots,T\}$ and subcarrier $n \in \{1,\dots,K\}$ is given by

$$Y[m,n,r] = \sum_{t=1}^{N_t} H[m,n,r,t]X[m,n,t] + W[m,n,r], \quad (1)$$

where $H[m,n,r,t]$ is the channel coefficient from transmit antenna t; $X[m,n,t]$ denotes the transmitted symbol; and $W[m,n,r] \sim \mathcal{CN}(0,\sigma^2)$ is i.i.d. additive white Gaussian noise. The pilot subcarriers use quadrature phase shift keying (QPSK) modulation. Each OFDM symbol m comprises $|\Gamma|$ pilot subcarriers and $|\Lambda|$ data subcarriers (i.e., $|\Gamma|+|\Lambda| = K$). All OFDM symbols share the same comb-type pilot grid $\Gamma \subset \{1,\dots,K\}$ with $|\Gamma|$ uniformly spaced subcarriers, including both endpoints $n = 1$ and $n = K$. A time-orthogonal pilot pattern is employed: let $t^\dagger \triangleq 1 + ((m-1) \bmod N_t)$ denote

This research was supported in part by the German Federal Ministry of Research, Technology and Space (BMFTR) within the project Open6GHub under grant number 16KISK016

the active transmit antenna index for symbol m. In symbol m, only antenna $t^\dagger$ transmits pilots; others transmit zeros. Time-orthogonal pilots provide full frequency resolution per antenna. For $N_t = 2$ at 200 km/h, the pilot pattern occupies $\approx 13\%$ of the Nyquist bandwidth, ensuring stable tracking. Under these conditions (with 66.7 μs symbol duration), the inter-pilot correlation is high ($\rho_{\text{time}} \approx 0.96$); however, extrapolating $L = 5$ symbols ahead remains challenging as correlation drops to $\rho_{\text{time}} \approx 0.65$. The received pilot signal from the active antenna $t^\dagger$ is

$$Y_p[m,n,r] = H_p[m,n,r,t^\dagger]X_p[m,n,t^\dagger] + W[m,n,r]. \quad (2)$$

On pilot subcarriers $n \in \Gamma$, we employ the scalar least-squares (LS) estimator, which is computed separately per $(m,n,r,t^\dagger)$:

$$\begin{aligned}\hat{H}_p^{\text{LS}}[m,n,r,t^\dagger] &= \arg\min_h \left|Y_p[m,n,r] - hX_p[m,n,t^\dagger]\right|^2 \\ &= \frac{Y_p[m,n,r]}{X_p[m,n,t^\dagger]}, \quad n \in \Gamma. \end{aligned} \quad (3)$$

III. KronFormer: Factorized Attention for MIMO-OFDM Prediction

We introduce KronFormer, a Transformer predicting MIMO-OFDM channels from sparse observations via factorized spatial-temporal attention (Fig. 1). Motivated by the Kronecker-structured MIMO covariance, this factorization permits independent spatial-temporal processing. For each symbol m, a feature vector $\hat{\mathbf{H}}_p^{\text{LS}}[m,:]$ is formed. Due to the time-orthogonal pilot scheme, only the active transmit antenna $t^\dagger$ provides $|\Gamma|N_r$ LS estimates, which are embedded within a larger, zero-padded vector of pilot-domain dimension $D_p \triangleq |\Gamma|N_rN_t$. This sparse vectorization preserves a consistent spatial-frequency structure and forms a row in the estimate matrix $\hat{\mathbf{H}}_p^{\text{LS}} \in \mathbb{C}^{P\times D_p}$. Processing a batch of B realizations, the encoder input $\hat{\mathbf{H}}_{\text{enc}} \in \mathbb{R}^{B\times P\times 2D_p}$ contains pilot estimates from the last $P \le T$ OFDM symbols:

$$\hat{\mathbf{H}}_{\text{enc}}[b,m,:] = \left[\Re\{\hat{\mathbf{H}}_p^{\text{LS}}[b,m,:]\},\ \Im\{\hat{\mathbf{H}}_p^{\text{LS}}[b,m,:]\}\right]. \quad (4)$$

The decoder input $\hat{\mathbf{H}}_{\text{dec}} \in \mathbb{R}^{B\times(G+L)\times 2D_p}$ consists of the most recent $G \le P$ pilot-based estimates, followed by L zero-padded rows that reserve positions for future symbol prediction. Finally, the output $\hat{\mathbf{H}}_{\text{output}} \in \mathbb{R}^{B\times L\times 2D_c}$ is the predicted channel frequency response (CFR) for the next L OFDM symbols across all transmit-receive antenna pairs, where $D_c \triangleq KN_rN_t$ denotes the full CFR dimension.

A. Factorized Spatial-Temporal Attention

Under the Kronecker model [14] and assuming a common Doppler spectrum across all spatial links, the MIMO channel covariance shows the separable structure $\mathbf{R}_H \approx \mathbf{R}_{\text{Rx}} \otimes \mathbf{R}_{\text{Tx}} \otimes \mathbf{R}_{\text{time}}$ (extending the spatial model to the temporal domain), where $\mathbf{R}_{\text{Rx}} \in \mathbb{C}^{N_r\times N_r}$, $\mathbf{R}_{\text{Tx}} \in \mathbb{C}^{N_t\times N_t}$, $\mathbf{R}_{\text{time}} \in \mathbb{C}^{P\times P}$ are the spatial and temporal correlation matrices and $\otimes$ the Kronecker product.

Let d_{ant} denote the per-antenna embedding dimension and Z the number of attention heads with dimension $d_k = d_{\text{ant}}/Z$.

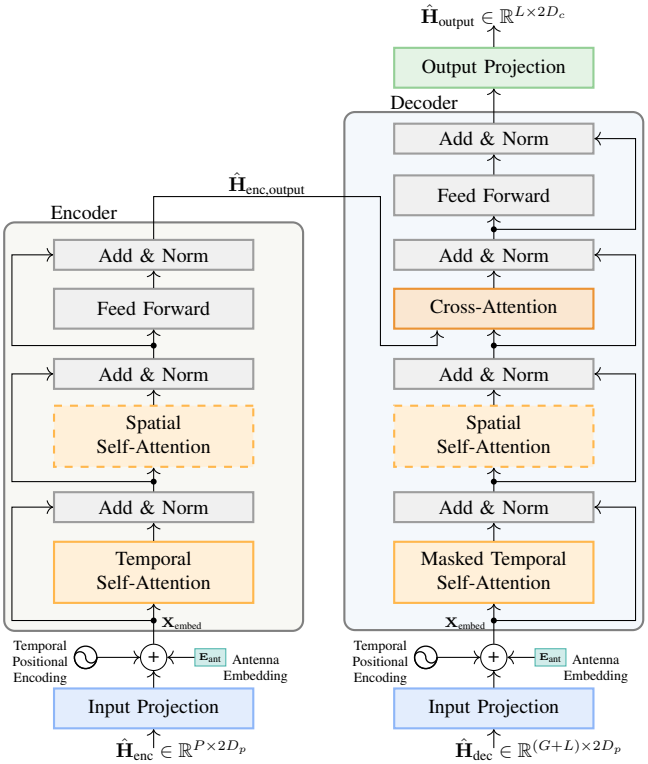

Fig. 1. KronFormer architecture. Temporal (solid) and spatial (dashed) attention in orange; $\mathbf{E}_{\text{ant}}$ (teal) encodes antenna identity.

The input to these attention layers incorporates a learnable antenna embedding $\mathbf{E}_{\text{ant}}$ (Fig. 1, teal) that encodes antenna-pair identity; its construction is detailed in Section III-B.

First, temporal attention (Fig. 1, solid orange) operates independently per antenna pair $a \in \{1,\dots,A\}$, where $A = N_rN_t$. Following the scaled dot-product attention [7], the output captures Doppler-induced temporal correlations:

$$\mathbf{Z}_a^{(\text{temp})} = \text{softmax}\left(\frac{\mathbf{Q}_a\mathbf{K}_a^\top}{\sqrt{d_k}}\right)\mathbf{V}_a, \quad (5)$$

where $\mathbf{Q}_a, \mathbf{K}_a, \mathbf{V}_a \in \mathbb{R}^{P\times d_k}$ and the softmax function operates row-wise (i.e., normalization along the key dimension); concatenating Z parallel heads yields $\mathbf{Z}_a^{(\text{temp})} \in \mathbb{R}^{P\times d_{\text{ant}}}$. Weights are shared across antenna pairs.

Second, spatial attention (Fig. 1, dashed orange) operates independently per OFDM symbol $m \in \{1,\dots,P\}$, learning inter-antenna correlations aligned with $\mathbf{R}_{\text{Rx}} \otimes \mathbf{R}_{\text{Tx}}$:

$$\mathbf{Z}_m^{(\text{spat})} = \text{softmax}\left(\frac{\mathbf{Q}_m\mathbf{K}_m^\top}{\sqrt{d_k}}\right)\mathbf{V}_m, \quad (6)$$

where $\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m \in \mathbb{R}^{A\times d_k}$; similarly, $\mathbf{Z}_m^{(\text{spat})} \in \mathbb{R}^{A\times d_{\text{ant}}}$.

This factorization reduces attention complexity from $O(P^2A^2)$ to $O(P^2A + A^2P)$, yielding an approximately 3.5× reduction for 2×2 MIMO ($A = 4$, $P = 30$), scaling to 10× and 20× for 4×4 and 8×8 arrays, respectively. KronFormer keeps antenna dimensions explicit throughout processing, enabling the architecture to align naturally with the Kronecker decomposition. Each attention block is followed by residual connections with layer normalization (Add&Norm) and a position-wise feed-forward network with Gaussian error

linear unit activation. A final dense linear projection flattens the antenna dimension to map the decoder output to $2D_c$ real dimensions, fusing information across all spatial streams.

B. Input Organization and Antenna Embeddings

The Input Projection (Fig. 1, blue) reshapes the 3-D encoder input $\hat{\mathbf{H}}_{\text{enc}} \in \mathbb{R}^{B\times P\times 2D_p}$ into A antenna-pair slices of $F \triangleq 2D_p/A = 2|\Gamma|$ features each and applies a shared linear map $f_{\text{proj}} : \mathbb{R}^F \to \mathbb{R}^{d_{\text{ant}}}$, where $d_{\text{model}} = A \cdot d_{\text{ant}}$. Adding sinusoidal temporal encoding $\mathbf{PE}_{\text{temp}} \in \mathbb{R}^{1\times P\times 1\times d_{\text{ant}}}$ and the learnable antenna embedding $\mathbf{E}_{\text{ant}} \in \mathbb{R}^{1\times 1\times A\times d_{\text{ant}}}$ (Fig. 1, teal) yields the input to the attention layers:

$$\mathbf{X}_{\text{embed}} = f_{\text{proj}}\big(\text{reshape}(\hat{\mathbf{H}}_{\text{enc}})\big) + \mathbf{PE}_{\text{temp}} + \mathbf{E}_{\text{ant}}. \quad (7)$$

This embedding enables the spatial attention in (6) to learn $\mathbf{R}_{\text{Rx}} \otimes \mathbf{R}_{\text{Tx}}$ correlations without explicit covariance estimation.

Error Propagation Mitigation. Unlike sequential predictors where $\hat{\mathbf{H}}[P+l]$ depends recursively on $\hat{\mathbf{H}}[P+l-1]$, KronFormer employs direct multi-step prediction. The decoder input $\hat{\mathbf{H}}_{\text{dec}}$ undergoes the same Input Projection and embedding pipeline as the encoder, with shared f_{proj} and $\mathbf{E}_{\text{ant}}$ weights, but uses zero-padding for prediction positions; thus, while self-attention remains causal, the input is fixed, preventing recursive error feedback.

C. Loss Function

Training minimizes the mean squared error (MSE) between predicted and true CFRs:

$$\mathcal{L}_{\text{MSE}}(\Theta) = \frac{1}{B\cdot L\cdot 2D_c}\left\|\hat{\mathbf{H}}_{\text{output}} - \mathbf{H}_{\text{true}}\right\|_F^2, \quad (8)$$

where $\mathbf{H}_{\text{true}} \in \mathbb{R}^{B\times L\times 2D_c}$ denotes the ground truth and $\|\cdot\|_F^2$ the squared Frobenius norm. The 2×2 configuration yields $|\Theta| = 360{,}608$ parameters (see Table III for scaling).

D. Spectral Efficiency

We compute the SE (bps/Hz) as $\text{SE}[l] = C[l]/K$ for each predicted symbol l, where the achievable sum-rate $C[l]$ (bits/symbol), assuming Gaussian signaling, is

$$C[l] = \mathbb{E}\left[\sum_{n=1}^{K}\log_2\det\left(\mathbf{I}_{N_r} + \tfrac{1}{N_t\sigma^2_{\text{eff}}[l,n]}\hat{\mathbf{H}}[l,n]\hat{\mathbf{H}}^H[l,n]\right)\right], \quad (9)$$

where $\mathbf{I}_{N_r} \in \mathbb{R}^{N_r\times N_r}$ denotes the identity matrix, $\hat{\mathbf{H}}[l,n], \mathbf{H}[l,n] \in \mathbb{C}^{N_r\times N_t}$ denote the predicted and true MIMO channel matrices, respectively, and the effective noise variance (assuming equal power allocation with $\mathbb{E}[\mathbf{x}\mathbf{x}^H] = (1/N_t)\mathbf{I}_{N_t}$) is

$$\sigma^2_{\text{eff}}[l,n] = \sigma^2 + \frac{\|\mathbf{H}[l,n] - \hat{\mathbf{H}}[l,n]\|_F^2}{N_tN_r}. \quad (10)$$

TABLE I
SIMULATION PARAMETERS

Parameter		Value			
MIMO-OFDM System					
Transmit Antennas	N_t	2, 4, 8			
Receive Antennas	N_r	2, 4, 8			
Subcarriers	K	128			
Pilot Subcarriers	$\lvert\Gamma\rvert$	32			
Subcarrier Spacing	Δf	15 kHz			
Center Frequency	f_c	2.5 GHz			
KronFormer Architecture			**MIMO Config.**		
			2×2	4×4	8×8
Encoder Symbols	$P{=}T$	30			
Decoder History	G	10			
Predicted Symbols	L	5			
Attention Heads	Z	4			
Model Dimension	d_{model}	—	256	1024	4096
Per-Antenna Dim.	d_{ant}	64		(fixed)	
Feed-Forward Dim.	d_{ff}	—	64	256	1024

Spatial Corr.		
	ρ_{Tx}	ρ_{Rx}
LOW:	0.04	0.06
MED:	0.22	0.90
HIGH:	0.88	0.90

IV. SIMULATION RESULTS

A single-layer KronFormer (Table I) is trained with AdamW ($\eta = 0.001$, $\lambda = 0.004$), learning rate decay $\eta_{\text{epoch}} = \eta\cdot\exp(-0.1\cdot\max(\text{epoch}-9,0))$, and early stopping (patience 20). Training uses 5×10^4 channels (80/20 split) for 200 epochs ($B = 64$, $E_b/N_0 = 13$ dB) with distinct train/test seeds. Ablation over 90 configurations shows deeper layers yield $< 2.5\%$ MSE gain at $\approx 40\%$ more parameters, validating single-layer sufficiency.

Frequency-selective channels are generated using the COST 259 typical urban (CTU) [15] at various user velocities v. To predict $L = 5$ future OFDM symbols (≈ 333 μs ahead), we compute the achievable SE via (9) and a fixed-$E_b/N_0 = 10$ dB MSE using 16-ary quadrature amplitude modulation, averaged over 10^4 channel realizations. For the bit error rate (BER)/MSE vs. E_b/N_0 sweeps, we employ QPSK over 5×10^4 realizations to ensure tractable maximum likelihood detection after equalizing the channel with the predicted CSI. SE for predictors is computed as described in Section III-D over all K subcarriers to isolate prediction quality. We compare against two perfect CSI bounds (i.e., $\hat{\mathbf{H}}[l,n] = \mathbf{H}[l,n]$): the pilot-overhead bound (summing over $|\Lambda| = 96$ data subcarriers) and the ideal upper bound (summing over all $K = 128$ subcarriers), both normalized by K.

Using the time-orthogonal pattern from Section II (effective pilot density $25\%/N_t$), KronFormer's 4-D architecture maintains antenna-pair identity via (B, P, A, F) tensors, independently capturing spatial and Doppler dynamics aligned to the Kronecker structure. Baselines include: i) unfactorized P2P Transformer [6] ($|\Theta| = 232$k) and ii) P2P LSTM (following [5], $|\Theta| = 378$k), both using flattened features; iii) Wiener filter [3] ($G = 29$); and iv) Kalman filter (state-space model from [4]), both after LS-interpolated estimation.

Fig. 2 shows SE versus user speed for all correlation levels (averaged over $L = 5$ symbols). KronFormer maintains the smallest gap to theoretical bounds, exhibiting 10–13% SE degradation from 50 to 200 km/h, compared to 15–19% for LSTM and 34–39% for the Wiener filter. High spatial correlation reduces multiplexing gain, degrading absolute SE for all methods.

This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2026.3676796

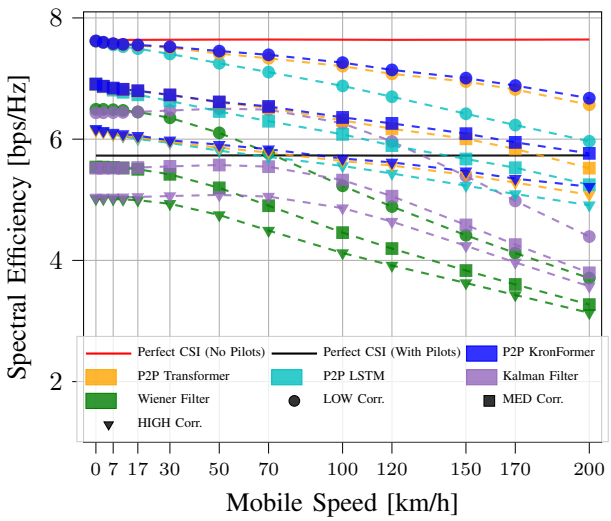

Fig. 2. SE vs. Mobile Speed over a CTU channel at 2.5 GHz for a 2×2 MIMO system ($N_t = N_r = 2$).

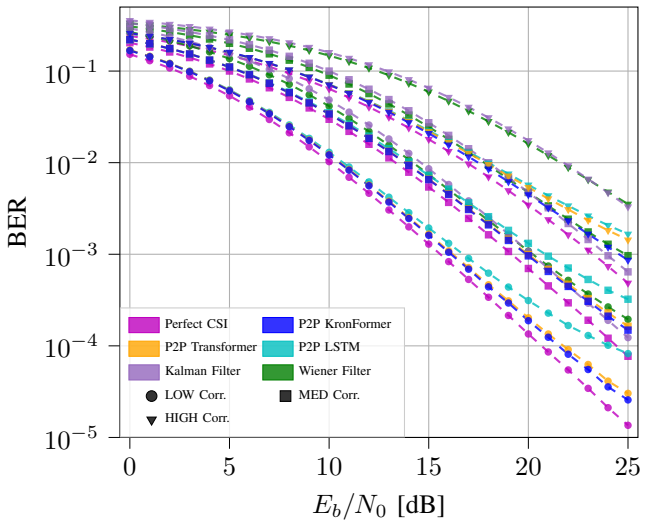

Fig. 4. BER vs. E_b/N_0 at 17 km/h over a CTU channel at 2.5 GHz for a 2×2 MIMO system ($N_t = N_r = 2$).

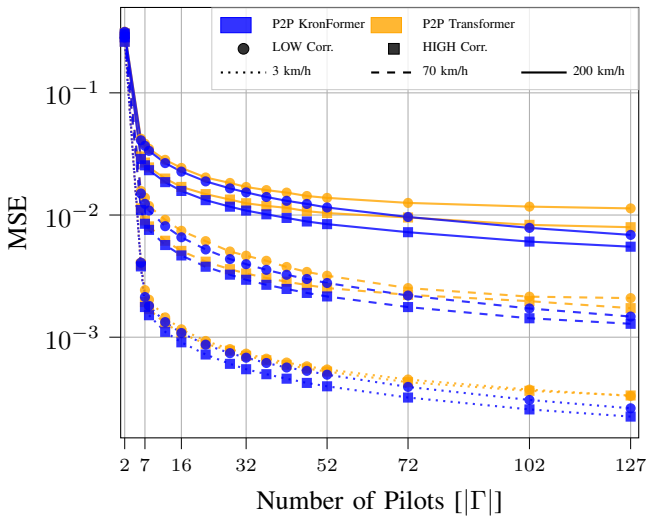

Fig. 3. MSE vs. Number of Pilots over a CTU channel at 2.5 GHz for a 2×2 MIMO system ($N_t = N_r = 2$).

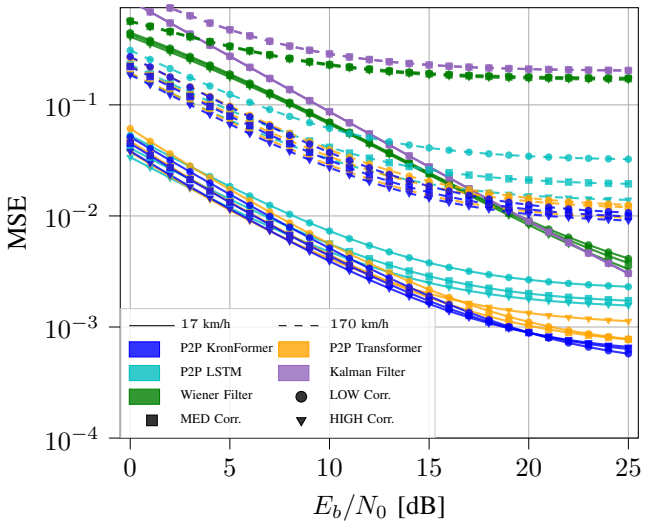

Fig. 5. MSE vs. E_b/N_0 over a CTU channel at 2.5 GHz for a 2×2 MIMO system ($N_t = N_r = 2$).

Diminishing MSE returns appear beyond $|\Gamma| = 32$ pilots (49–63% reduction, Fig. 3). KronFormer outperforms the unfactorized Transformer across all pilot counts; at 3 km/h with 32 pilots, increasing spatial correlation from LOW to HIGH reduces KronFormer's MSE by 20%, whereas the unfactorized Transformer gains only 4%, confirming that factorized attention more effectively exploits Kronecker-structured covariance.

Fig. 4 shows BER against E_b/N_0 at a user speed of 17 km/h. KronFormer's performance closely approaches the perfect CSI baseline, demonstrating a clear advantage over the Transformer, LSTM, and classical predictors. Consistent with the SE results, higher spatial correlation increases the BER for all models. Fig. 5 compares the MSE versus E_b/N_0 at speeds of 17 km/h and 170 km/h. KronFormer consistently achieves the lowest prediction error, with optimal performance under medium correlation at 17 km/h and under high correlation at 170 km/h. However, under high correlation and at low E_b/N_0 (< 5 dB), its advantage diminishes as all three NN models saturate at the pilot-noise floor, negating architectural differences. Crucially, at high mobility, both classical filters exhibit a significantly higher error floor than the NN models.

Mobility robustness is evaluated at $E_b/N_0 = 13$ dB (Fig. 6). KronFormer exhibits the smallest BER degradation (2.1–4.2×) across correlation levels. In contrast, the Transformer degrades by 2.2–4.7×, while classical and LSTM baselines suffer significantly higher degradation of up to 8.1×. KronFormer also achieves the lowest absolute BER at each correlation level. Higher correlation reduces relative degradation but worsens absolute BER, revealing correlation as the primary performance limiter. This reflects the MIMO trade-off: correlation improves predictability but reduces spatial diversity and multiplexing gain.

Table II compares three reshaping strategies over 39 configurations (13 speeds × 3 correlations). Batch-dimension reshaping ($B \times A$) isolates each antenna pair; its flat MSE versus correlation (-0.4%) confirms spatial correlation is

This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2026.3676796

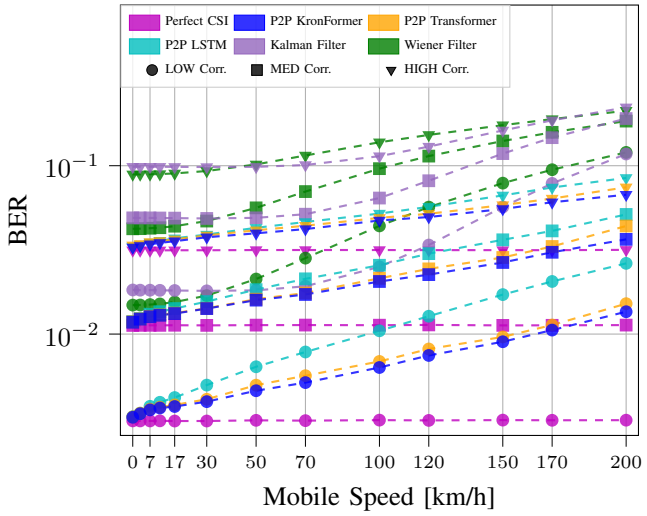

Fig. 6. BER vs. Mobile Speed at $E_b/N_0 = 13$ dB over a CTU channel at 2.5 GHz for a 2×2 MIMO system ($N_t = N_r = 2$).

TABLE II
ABLATION: RESHAPING STRATEGIES

Reshaping Strategy	**LOW**	**MED**	**HIGH**	**ΔMSE**
4-D Factorized (KronFormer)	9.76	**8.38**	**7.25**	—
Batch-dim ($B \times A$)	**9.69**	9.70	9.73	+14.7%
Feature-dim (5-D → 3-D)	10.93	9.34	8.06	+11.6%

2×2 MIMO, $E_b/N_0 = 10$ dB. MSE $\times 10^{-3}$, averaged over 13 speeds. ΔMSE: mean increase vs. KronFormer (all 39 conditions).

architecturally inaccessible. Feature flattening (5-D→3-D) enables only implicit coupling and exhibits $+11.6\%$ higher MSE than KronFormer (p-value $< 10^{-5}$, paired t-test), confirming the benefit of explicit spatial-temporal factorization.

A. MIMO Scalability

Table III compares KronFormer against the unfactorized Transformer across 2×2, 4×4, and 8×8 configurations at $E_b/N_0 = 10$ dB and 17 km/h, revealing three insights: (i) The unfactorized Transformer suffers severe degradation at 8×8 under low spatial correlation (MSE ≈ 0.5), while KronFormer maintains stable performance (MSE $= 5.71 \times 10^{-3}$), achieving an $87\times$ improvement. This validates the Kronecker-aligned factorization: by maintaining the explicit $(B, P, A, d_{\text{ant}})$ representation, KronFormer exploits the separable spatial-temporal correlations that flattening destroys.

(ii) Both architectures scale as $O(A^2)$ in total parameters. Since each configuration step quadruples A, the expected growth per step is $4^2 = 16\times$; observed ratios are $11.9\times$ ($2\times2 \to 4\times4$) and $15.6\times$ ($4\times4 \to 8\times8$), with the gap due to lower-order terms. Crucially, only the output projection drives this growth (up to 99.8% of parameters at 8×8); attention layers remain $O(1)$ with fixed $d_{\text{ant}} = 64$.

(iii) Inference efficiency scales favorably for KronFormer. At 2×2, both methods exhibit comparable inference times (1.07 versus 1.10 ms/link), confirming negligible factorization overhead. For 8×8, KronFormer achieves 0.84 ms/link despite 67 M parameters, outperforming the Transformer by 41% (1.42 ms/link). This stems from reduced attention FLOPs under factorization, enabling efficient parallelization across antenna pairs even when the output projection dominates.

TABLE III
FIXED-E_b/N_0 MSE PERFORMANCE AND MODEL COMPLEXITY

Config.	**Method**	**Params**	**Average Fixed-E_b/N_0 MSE**			**Inf.**[a,b]
			LOW	**MED**	**HIGH**	**[ms/link]**
2×2	Transformer	232,448	3.08×10^{-3}	2.54×10^{-3}	2.55×10^{-3}	1.10
	KronFormer	360,608	$\mathbf{2.62 \times 10^{-3}}$	$\mathbf{2.44 \times 10^{-3}}$	$\mathbf{2.25 \times 10^{-3}}$	**1.07**
4×4	Transformer	3,682,304	1.25×10^{-2}	3.96×10^{-3}	2.82×10^{-3}	**0.73**
	KronFormer	4,297,376	$\mathbf{2.47 \times 10^{-3}}$	$\mathbf{1.10 \times 10^{-3}}$	$\mathbf{7.34 \times 10^{-4}}$	0.74
8×8	Transformer	58,769,408	4.98×10^{-1}	1.76×10^{-2}	1.37×10^{-2}	1.42
	KronFormer	67,230,368	$\mathbf{5.71 \times 10^{-3}}$	$\mathbf{8.48 \times 10^{-4}}$	$\mathbf{6.62 \times 10^{-4}}$	**0.84**

[a]Inference time normalized per spatial link (antenna pair) for fair comparison.
[b]Measured on RTX 4090, Ryzen 9 7950X3D, 64 GB RAM, $B = 64$.

V. CONCLUSION

This letter presented KronFormer, a P2P architecture aligning Transformer attention with Kronecker-structured MIMO covariance. By preserving explicit tensor dimensions rather than flattening, it breaks the computational bottleneck of standard Transformers for scalable large-array inference. Simulations confirm superior mobility robustness and accuracy with reduced pilot overhead, offering a path for low-latency 6G prediction. Future work targets massive MIMO deployment.

REFERENCES

[1] P. Tarafder, C. Chun, A. Ullah, Y. Kim, and W. Choi, "Channel estimation in 5g-and-beyond wireless communication: A comprehensive survey," *Electronics*, vol. 14, 2025.
[2] B. D. Filippo, C. Amatetti, and A. Vanelli-Coralli, "Uplink ofdm channel prediction with hybrid cnn-lstm for 6g non-terrestrial networks," 2025.
[3] D. Schafhuber and G. Matz, "Mmse and adaptive prediction of time-varying channels for ofdm systems," *IEEE Trans. Wireless Commun.*, vol. 4, 2005.
[4] H. Shu, L. Ros, and E. P. Simon, "Simplified random-walk-model-based kalman filter for slow to moderate fading channel estimation in ofdm systems," *IEEE Trans. Signal Process.*, vol. 62, 2014.
[5] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open J. Commun. Soc.*, vol. 1, 2020.
[6] L. Lagona, M. Vakilifard, C. Bockelmann, and A. Dekorsy, "Transformer-based pilot-to-prediction for frequency-selective channels in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 15, 2026.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017.
[8] H. Jiang, M. Cui, D. W. K. Ng, and L. Dai, "Accurate channel prediction based on transformer: Making mobility negligible," *IEEE J. Sel. Areas Commun.*, vol. 40, 2022.
[9] Y. Jin, Y. Wu, Y. Gao, S. Zhang, S. Xu, and C.-X. Wang, "LinFormer: A linear-based lightweight transformer architecture for time-aware MIMO channel prediction," *IEEE Trans. Wireless Commun.*, vol. 24, 2025.
[10] D. Luan and J. S. Thompson, "Channelformer: Attention based neural solution for wireless channel estimation and effective online training," *IEEE Trans. Wireless Commun.*, vol. 22, 2023.
[11] K. He, T. X. Vu, L. Fan, S. Chatzinotas, and B. Ottersten, "Spatio-temporal predictive learning using crossover attention for communications and networking applications," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, 2025.
[12] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, 2021.
[13] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *Proc. ICCV*, 2021.
[14] J. Kermoal, L. Schumacher, K. Pedersen, P. Mogensen, and F. Frederiksen, "A stochastic mimo radio channel model with experimental validation," *IEEE J. Sel. Areas Commun.*, vol. 20, 2002.
[15] "Universal mobile telecommunications system (umts); deployment aspects (3gpp tr 25.943 version 17.0.0 release 17)," ETSI, Technical Report TR 125 943 V17.0.0, April 2022.