CHAST: Attention Aided SISO OFDM Channel Estimation

Mehmetcan Gok*
Department of Electrical Engineering
Northwestern University
Evanston, IL 60208
mehmetcan.gok@u.northwestern.edu

John Zhou
Apple
San Diego, CA 92121
johnzhou@apple.com

Supratik Bhattacharjee
Apple
San Diego, CA 92121
sbhattacharjee22@apple.com

Huaning Niu
Apple
Sunnyvale, CA 94085
huaning_niu@apple.com

Sharad Sambhwani
Apple
San Diego, CA 92121
ssambhwani@apple.com

Abstract

Next-generation wireless systems demand channel estimators that adapt to diverse propagation environments without requiring explicit channel statistics. We propose CHAST (Channel Attention Estimation), a lightweight deep learning architecture for OFDM channel estimation that operates directly on sparse pilot observations at DM-RS positions. CHAST combines a CNN feature extractor with a single multi-head self-attention block to capture both local and long-range time-frequency dependencies. Unlike existing transformer-based approaches such as CE-ViT, CHAST requires no knowledge of channel governing parameters (SNR, Doppler shift, delay spread) while achieving comparable performance with significantly reduced complexity. Extensive evaluation on 3GPP channel models demonstrates that CHAST outperforms traditional methods, particularly in high-mobility scenarios where Kronecker covariance assumptions break down. Attention visualization reveals that the model learns physically meaningful estimation strategies, with attention neighborhoods dynamically expanding as SNR increases and different heads specializing in specific spatial patterns.

1 Introduction

The transition towards 6G wireless networks envisions a diverse ecosystem of applications operating in increasingly complex and dynamic environments, from high-speed mobility to terahertz communications [1, 2]. This evolution renders traditional, model-based signal processing insufficient, as the underlying physical assumptions often fail to capture the richness of real-world channels [3]. This necessitates a paradigm shift towards an "AI-native" physical layer, where core receiver functions are data-driven, adaptive, and robust by design [4].

A fundamental component of any receiver is the channel estimator. In modern OFDM systems, this involves reconstructing the full time-frequency channel response from a sparse grid of Demodulation Reference Signals (DM-RS). Classical estimators, such as the Linear Minimum Mean Square Error (LMMSE) detector, rely on *a priori* knowledge of channel statistics, typically assuming Kronecker decomposition of the covariance matrix in time and frequency domains. However, these assumptions

*Work done during an internship at Apple

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

frequently break down in high-mobility or high-delay spread scenarios, creating significant model mismatch and performance degradation [5].

Recent deep learning approaches have shown promise by learning channel priors implicitly from data [6, 7]. However, many existing methods suffer from critical limitations: they either require explicit knowledge of channel governing parameters like SNR, Doppler shift, and delay spread, or they operate on pre-interpolated dense channel grids rather than raw sparse observations. For instance, the recent CE-ViT architecture [8], while achieving strong performance, requires many transformer blocks and assumes availability of channel parameters, limiting its practical deployment in parameter-agnostic scenarios.

To address these limitations, we propose CHAST (Channel Attention eSTimator), a lightweight attention-based architecture that operates directly on sparse DM-RS observations without requiring any channel governing parameters. Our contributions are: (1) A parameter-free channel estimation framework that achieves comparable performance to oracle-guided methods while using significantly reduced complexity, (2) Comprehensive evaluation demonstrating substantial improvements over traditional estimators across diverse 3GPP channel models, particularly in challenging high-mobility scenarios, and (3) Analysis of learned attention mechanisms revealing estimation strategies that dynamically adapt to channel conditions.

2 System Model and Estimation Baselines

2.1 OFDM System Model

We consider downlink SISO OFDM system in 5G-NR over a channel that is modeled as a resource grid spanning K subcarriers and L OFDM symbols. After cyclic prefix (CP) removal and FFT, the received signal $\mathbf{Y} \in \mathbb{C}^{K \times L}$ is related to the transmitted symbols $\mathbf{X} \in \mathbb{C}^{K \times L}$ by:

$$\mathbf{Y}[k, l] = \mathbf{H}[k, l]\mathbf{X}[k, l] + \mathbf{N}[k, l], \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{K \times L}$ is the channel's time-frequency response, and $\mathbf{N} \in \mathbb{C}^{K \times L}$ is additive white Gaussian noise (AWGN) with i.i.d. elements $\mathcal{CN}(0, \sigma_n^2)$. A known subset of resource elements at indices $\mathcal{P}$, contain DM-RS pilots for channel estimation. The objective is to leverage these pilots to estimate the full channel matrix $\widehat{\mathbf{H}}$, minimizing the normalized mean squared error (NMSE) with respect to the true channel $\mathbf{H}$:

$$\mathcal{L}_{\text{NMSE}}\left(\widehat{\mathbf{H}}\right) = \mathbb{E}\left[\frac{\|\widehat{\mathbf{H}} - \mathbf{H}\|_F^2}{\|\mathbf{H}\|_F^2}\right]. \tag{2}$$

Typically, this task involves denoising the noisy observations at the pilot locations and interpolating these values across the entire resource grid. In the 5G NR standard, the standard resource block (RB)—the basic unit of scheduling—spans 12 subcarriers in frequency and 14 OFDM symbols in time for the normal CP configuration.

2.2 Traditional Channel Estimation Baselines

Least Squares (LS) Estimation. In a vectorized form, the LS estimate for the channel at pilot locations, $\widehat{\mathbf{h}}_p^{\text{LS}} \in \mathbb{C}^{|\mathcal{P}|}$, is derived by solving the unconstrained minimization problem:

$$\widehat{\mathbf{h}}_p^{\text{LS}} = \arg\min_{\mathbf{h}_p} |\mathbf{y}_p - \mathbf{X}_p\mathbf{h}_p|_2^2 = \mathbf{X}_p^{-1}\mathbf{y}_p, \tag{3}$$

where $\mathbf{y}_p \in \mathbb{C}^{|\mathcal{P}|}$ is the vector of received pilot signals and $\mathbf{X}_p \in \mathbb{C}^{|\mathcal{P}| \times |\mathcal{P}|}$ is a diagonal matrix of the corresponding known DM-RS pilots. The full channel grid $\widehat{\mathbf{H}}_{\text{LS}}$ is then obtained by interpolating these sparse-noisy estimates $\widehat{\mathbf{h}}_p^{\text{LS}}$. While simple, LS estimation suffers from significant noise enhancement, especially in low-SNR regimes.

LMMSE Estimation. The LMMSE estimator is the optimal linear estimator that minimizes the MSE, provided the channel's second-order statistics and noise variance are known. It refines the LS

estimate by applying a Wiener filter that incorporates this prior knowledge. The LMMSE estimate of the full vectorized channel $\mathbf{h} \in \mathbb{C}^{KL}$ is:

$$\widehat{\mathbf{h}}_{\text{LMMSE}} = \mathbf{R}_{hh_p} \left(\mathbf{R}_{h_p h_p} + \sigma_n^2 (\mathbf{X}_p \mathbf{X}_p^H)^{-1} \right)^{-1} \widehat{\mathbf{h}}_p^{LS}, \tag{4}$$

where $\mathbf{R}_{h_p h_p}$ is the autocovariance matrix of the channel at the pilot locations, $\mathbf{R}_{hh_p}$ is the crosscovariance matrix between the full channel vector and the channel vector at pilot locations.The term $(\mathbf{X}_p \mathbf{X}_p^H)^{-1}$ is a diagonal matrix of inverse pilot powers. While theoretically optimal, the performance of the LMMSE estimator is critically contingent on the accuracy of the covariance matrices. In practical systems, these statistics are unknown and time-varying, making the acquisition of accurate covariance information a fundamental challenge for its implementation. To create meaningful benchmarks, we implement two distinct variants.

- **Model based LMMSE:** This estimator serves as a theoretical performance benchmark by assuming perfect, instantaneous knowledge of the channel's governing parameters (Doppler shift, delay spread) and noise variance. It relies on a separable covariance model, where the full channel covariance matrix $\mathbf{R}_h \in \mathbb{C}^{KL \times KL}$ is the Kronecker product of the time and frequency correlation matrices: $\mathbf{R}_h = \mathbf{R}_t \otimes \mathbf{R}_f$. We construct these using ground-truth parameters from our simulation environment. The required sub-matrices $\mathbf{R}_{h_p h_p}$ and $\mathbf{R}_{hh_p}$ are extracted from $\mathbf{R}_h$.
  - **Temporal Correlation ($\mathbf{R}_t$):** Following Jakes' model, the correlation between symbols l_1 and l_2 is given by the zeroth-order Bessel function:
    $$[\mathbf{R}_t]_{l_1, l_2} = J_0(2\pi f_D T_{\text{sym}} |l_1 - l_2|), \tag{5}$$
    where f_D is the maximum Doppler shift and T_{sym} is the OFDM symbol duration.
  - **Frequency Correlation ($\mathbf{R}_f$):** For an exponential power delay profile, the correlation between subcarriers k_1 and k_2 is:
    $$[\mathbf{R}_f]_{k_1, k_2} = \frac{1}{1 + (2\pi \tau_{\text{rms}} \Delta f |k_1 - k_2|)^2}, \tag{6}$$
    where τ_{rms} is the RMS delay spread and Δf is the subcarrier spacing.
- **Empirical LMMSE:** To create a practical but stronger baseline, we implement an LMMSE estimator using empirically estimated covariance matrices. For this, we compute sample covariance matrices from a large set of ground-truth channel realizations. During evaluation, the appropriate pre-computed matrices are used in (4). While avoiding explicit model assumptions, this approach is fundamentally limited. It assumes wide-sense stationarity within a coarse scenario definition, failing to adapt to the specific statistics of an individual channel realization. Its accuracy is limited by the finite number of samples used for estimation.

2.2.1 Practical Heuristic Estimator (MATLAB)

As a practical baseline, we include the `nrChannelEstimate` function from the MATLAB 5G Toolbox[2]. This estimator implements a multi-stage heuristic pipeline without requiring explicit channel statistics. The core process involves: *(i)* Initial LS estimation at pilot locations, followed by spline interpolation across the frequency axis. *(ii)* Denoising via a transform to the time domain, where a filter window is applied to the channel impulse response (CIR). *(iii)* A second denoising stage using an adaptive 2D averaging filter, with a window size determined by an internal SNR estimate. *(iv)* Final linear interpolation across the time axis to form the dense channel grid.

2.2.2 Comparative Deep Learning Model: CE-ViT

CE-ViT [8] tests the efficacy of providing explicit physical side-information to the network. Unlike our proposed model, which learns implicitly from the channel grid alone, CE-ViT defines a `TokenModule` that processes scalar values for SNR, Doppler, and delay spread. These scalars are converted into token embeddings and concatenated to each channel patch embedding before being processed by cascaded standard Transformer encoder blocks (see [8] for details). This design allows us to directly compare our implicit learning approach against a more complex model explicitly guided by known channel-governing parameters.

[2]https://www.mathworks.com/help/5g/ref/nrchannelestimate.html

3 Proposed Deep Learning Architecture

3.1 CHAST Model

Our proposed light-weight model, illustrated in Figure 1, learns the mapping from sparse, noisy pilot observations to a dense, clean channel grid in an end-to-end fashion. The input to the model is the LS estimate grid $\widehat{\mathbf{H}}_{\mathrm{p}}^{\mathrm{LS}}$, represented as a two-channel tensor (real and imaginary parts), which is non-zero only at the DM-RS pilot locations. The architecture proceeds in four stages:

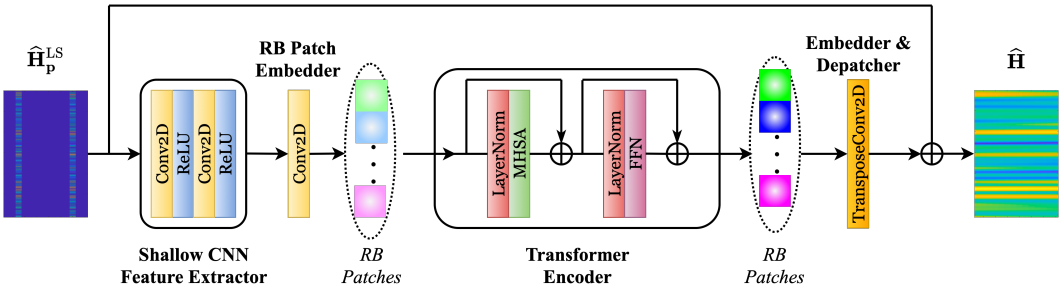


Figure 1: An overview of our proposed channel estimation architecture CHAST.

(i) Local Feature Extraction: Two convolutional layers process the sparse input grid, acting as learnable filters that extract local time-frequency features and produce a dense feature map by interpolating from sparse pilot signals in latent space. This CNN front-end enables the model to capture short-range correlations essential for initial channel reconstruction.

(ii) Patch Tokenization: Following ViT principles, we partition the feature map into non-overlapping patches, where each patch corresponds to a single Resource Block (RB). This is implemented efficiently using a `Conv2d` layer with kernel size and stride matching patch dimensions, projecting each patch into a d_{embed}-dimensional embedding.

(iii) Global Dependency Modeling: A single Transformer encoder block processes the patch embedding sequence. The multi-head self-attention mechanism computes relationships between all RB tokens, enabling capture of long-range time-frequency dependencies across the entire channel grid. This lightweight design maintains competitive performance while significantly reducing computational complexity.

(iv) Channel Reconstruction: A `TransposedConv2d` layer upsamples the processed patch sequence to reconstruct the full-resolution channel grid. A global residual connection adds the sparse input to the network output, allowing the model to focus on learning refinements for non-pilot elements, which stabilizes training and improves convergence.

4 Experimental Setup

4.1 Dataset Generation

We generated our dataset using MATLAB's 5G Toolbox to ensure realistic channel conditions (see Appendix Table 1 and Table 2 for details). To this end, all models were trained exclusively on a dataset generated the 3GPP-based scenarios (e.g. InO, UMi, UMa, RMa). For these scenarios, we randomized channel parameters such as carrier frequency, delay spread, and user velocity within the scenario-specific bounds. Our test set, however, comprised both held-out samples from these standard scenarios and data from three entirely unseen challenging scenarios designed to stress the model with extreme Doppler (High-Speed Train), high delay spread (Dense Urban Canyon), and doubly selective channels (Aerial Mobility). For all generated data, the SNR was swept from -6 to 21 dB in 3 dB increments.

4.2 Training and Evaluation

All models were trained to minimize the MSE loss using the AdamW optimizer [9]. We employed a CosineAnnealingLR schedule and an early stopping criterion to regularize training and prevent

overfitting; a comprehensive list of hyperparameters is provided in Appendix Table 3. Model performance is evaluated using the NMSE in dB.

5 Simulation Results

5.1 Overall Performance vs SNR

Figure 2 presents our primary performance evaluation, comparing CHAST against all baselines. The plots illustrate the NMSE (dB) as a function of SNR channel scenarios. To ensure statistical validity, each SNR point represents the average performance over independent channel realizations per scenario.

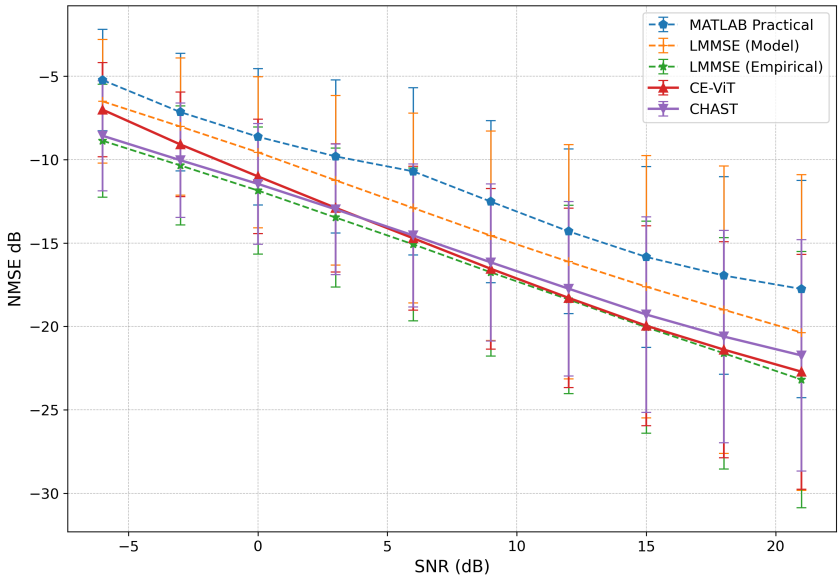


Figure 2: Overall channel estimation performances on the entire test

As illustrated in Figure 2, the proposed deep learning models demonstrate a significant performance advantage over traditional baselines. A key finding is that our lightweight CHAST model on par with the more complex CE-ViT baseline at high SNRs and outperforms in lower SNRs which highlights better denoising capability of CHAST. This is noteworthy because CE-ViT is a much deeper architecture (eight Transformer blocks vs. one) that requires oracle side-information (SNR, Doppler, etc.), whereas CHAST operates "blindly," implicitly learning the necessary channel characteristics from the data. This highlights our model's parameter efficiency and practicality. Moreover, the results also underscore the gains of data-driven approaches as it outperforms the model-based channel estimation techniques.

5.2 Analysis of the Learned Attention Mechanism

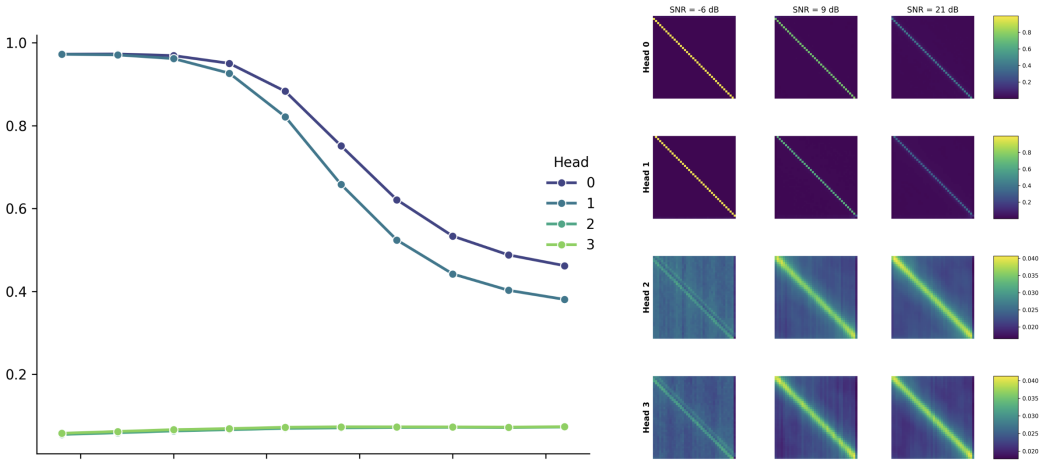
To understand *how* CHAST achieves its strong performance, we analyze its internal self-attention mechanism. We quantify the model's adaptive behavior by introducing the *local attention ratio*, a metric measuring the degree to which the model focuses on nearby versus distant channel patches. For a given attention matrix $\mathbf{A} \in \mathbb{R}^{N_p \times N_p}$, where N_p is the number of patches and A_{ij} is the attention from query RB patch i to key RB patch j, we define the local attention ratio $\mathcal{L}_W$ for a neighborhood window of radius W as:

$$\mathcal{L}_W(\mathbf{A}) = \frac{1}{N_p} \sum_{i=1}^{N_p} \sum_{j:|i-j|\leq W} A_{ij} \tag{7}$$

Fig. 3a plots this metric (using $W = 2$), for each attention head for UMa scenario, as a function of SNR. The results reveal a clear and consistent trend: the local attention ratio decreases as SNR

improves for main two heads. This quantitatively confirms that CHAST learns a sophisticated, adaptive strategy: it relies on a tight, local neighborhood in low-SNR conditions to mitigate noise, but expands its focus to capture long-range correlations (i.e., adopts a more global view) when the pilot measurements are more reliable.

To provide visual evidence for the model's adaptive capabilities, we also analyze the learned attention patterns as heatmaps from each of the four attention heads. Fig. 3b The model exhibits clear functional specialization. which illustrate a clear functional specialization among the individual attention heads. Rather than learning redundant patterns, the heads appear to decompose the estimation task. For instance, first two heads consistently learn to act as asymmetric local filters: one head may primarily attend to patches with lower frequency indices, while another attends to those with higher indices, effectively performing a directional interpolation from adjacent Resource Blocks (RBs). Conversely, other heads exhibit more dynamic, scenario-dependent behavior to capture complex channel characteristics. This learned decomposition—where some heads handle local refinement while others adapt to global channel statistics—is a key reason for the model's robust performance across diverse channel environments.



(a) Local attention ratio vs SNR

(b) Attention heatmaps per head vs SNR

Figure 3: Attention analysis of the heads vs SNR

6 Conclusion

In this work, we introduced CHAST, a lightweight attention-based architecture for OFDM channel estimation that operates directly on sparse pilot grids without requiring channel governing parameters. We demonstrated that CHAST achieves comparable performance to complex transformer-based approaches like CE-ViT while using significantly fewer resources—a single transformer block versus eight, and no dependency on SNR, Doppler shift, or delay spread knowledge. Our evaluation across diverse 3GPP channel models shows that CHAST substantially outperforms traditional LMMSE estimators in high-mobility scenarios where Kronecker covariance assumptions break down. The attention visualization reveals that the model learns physically meaningful estimation strategies, with attention patterns dynamically adapting to channel conditions. This parameter-free, adaptive approach addresses a critical need for next-generation wireless systems that must operate across diverse propagation environments without explicit channel statistics. The combination of competitive accuracy with reduced complexity makes CHAST particularly suitable for practical deployment in resource-constrained scenarios. Future work will extend this framework to MIMO systems, where learning joint spatio-temporal correlations will be essential for maximizing the potential of massive antenna arrays in next-generation networks.

References

[1] Walid Saad, Mérouane Bennis, and Mian Chen. A vision of 6g wireless systems: Applications, trends, technologies, and challenges. *IEEE network*, 34(3):134–142, 2020.

[2] Harsh Tataria, Mansoor Shafi, Andreas F Molisch, Mischa Dohler, Henrik Sjöland, and Fredrik Tufvesson. 6g wireless systems: Vision, requirements, challenges, enabling technologies, and new radio solutions. *IEEE Communications Magazine*, 59(7):26–32, 2021.

[3] Erik Björnson, Liesbet Van der Perre, Stefano Buzzi, and Erik G Larsson. Massive mimo in sub-6 ghz and mmwave: Physical, practical, and use-case differences. *IEEE wireless communications*, 26(2):100–108, 2019.

[4] Stamatis Karnouskos. Artificial intelligence in future-generation wireless networks. *IEEE Communications Magazine*, 58(3):10–11, 2020.

[5] Ove Edfors, Magnus Sandell, Jan-Jaap Van de Beek, Sarah Kate Wilson, and Per Ola Börjesson. Ofdm channel estimation by singular value decomposition. In *46th IEEE Vehicular Technology Conference*, volume 2, pages 923–927. IEEE, 1996.

[6] Mojtaba Soltani, Vahid Pouromoun, and Wolfgang Ertel. Deep learning-based channel estimation. In *2019 IEEE 2nd 5G World Forum (5GWF)*, pages 539–542. IEEE, 2019.

[7] Xiang Gao, Shi Jin, Chao-Kai Wen, and Geoffrey Ye Li. Deep learning for impaired wireless communications: A case study on channel estimation. *arXiv preprint arXiv:1807.08983*, 2018.

[8] Fangyu Liu, Jing Zhang, Peiwen Jiang, Chao-Kai Wen, and Shi Jin. Ce-vit: A robust channel estimator based on vision transformer for ofdm systems. In *GLOBECOM 2023-2023 IEEE Global Communications Conference*, pages 4798–4803. IEEE, 2023.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A Appendix

A.1 Dataset Parameters

Table 1: Summary of channel simulation scenarios used for training and evaluation, encompassing both standard 3GPP-based models and challenging, custom-defined environments.

Scenario	Channel Profiles	Frequency (GHz)	Delay Spread (ns)	Velocity (km/h)
Standard Scenarios				
Indoor Office (InO)	TDL-A, B, C	2.4, 5.0	10 – 39	0 – 20
Urban Micro (UMi)	TDL-A, B, C	2.4, 3.5, 5.0	65 – 180	10 – 60
Urban Macro (UMa)	TDL-A, B, C	2.1, 3.5, 5.0	65 – 300	30 – 100
Rural Macro (RMa)	TDL-A, B, C	0.85, 3.5, 5.0	32 – 37	100 – 200
UMa LOS	TDL-D, E	2.1, 3.5, 5.0	150 – 500	10 – 150
UMi LOS	TDL-D, E	2.4, 3.5, 5.0	100 – 350	10 – 150
Challenging Scenarios				
High-Speed Train	TDL-C, D	2.8, 3.6	100 – 300	250 – 400
Dense Urban Canyon	TDL-D, E	0.7, 5.0	500 – 1200	3 – 60
Aerial Mobility	TDL-D, E	3.5, 6.0	400 – 800	80 – 200

Table 2: Data generation parameters based of the 5G Toolbox simulation environment.

Parameter	Value / Description
Antennas (Tx, Rx)	(1, 1)
SNR	[-6, 21] dB
Carrier Configuration	40 RB, 30 kHz SCS, Normal CP
PDSCH Configuration	PRB:0-$N_{\text{RB}} - 1$, All symbols, Type A
DMRS Configuration	Type A Pos: 2, Length: 1, Configuration: 2, Additional Pos: 1
Number of samples per scenario	1024 per SNR point

A.2 Training and Model Parameters

Table 3: Training hyperparameters used for all deep learning models.

Parameter Group	Value / Configuration
Optimizer & Learning Rate Scheduler	
Optimizer	AdamW
Initial Learning Rate	0.001
Weight Decay	1e-5
LR Scheduler	CosineAnnealingLR
Early Stopping	
Patience	50 epochs
Min. Delta	1e-5
Metric	Validation Loss
Training Setup	
Loss Function	MSE
Epochs	300
Batch Size	64

Table 4: Architectural parameters for the evaluated deep learning models.

(a) CHAST

Parameter	Value
Input Channels	2
CNN Filters 1	32
CNN Filters 2	2
Patch Size	12×14
Embedding Dimension	64
Attention Heads	4
Dropout Rate	0.1
Number of Parameters	0.1 M

(b) CE-ViT

Parameter	Value
Input Channels	2
Patch Size	12×14
Embedding Dimension	64
Transformer Blocks (Depth)	8
Attention Heads	4
Auxiliary Token Dimension	128
Dropout Rate	0.1
Number of Parameters	0.46 M

A.3 Scenario Specific Results

Figure 4 presents our primary performance evaluation, comparing CHAST against all baselines. The plots illustrate the NMSE (dB) as a function of SNR across nine distinct channel scenarios, ranging from standard to challenging.

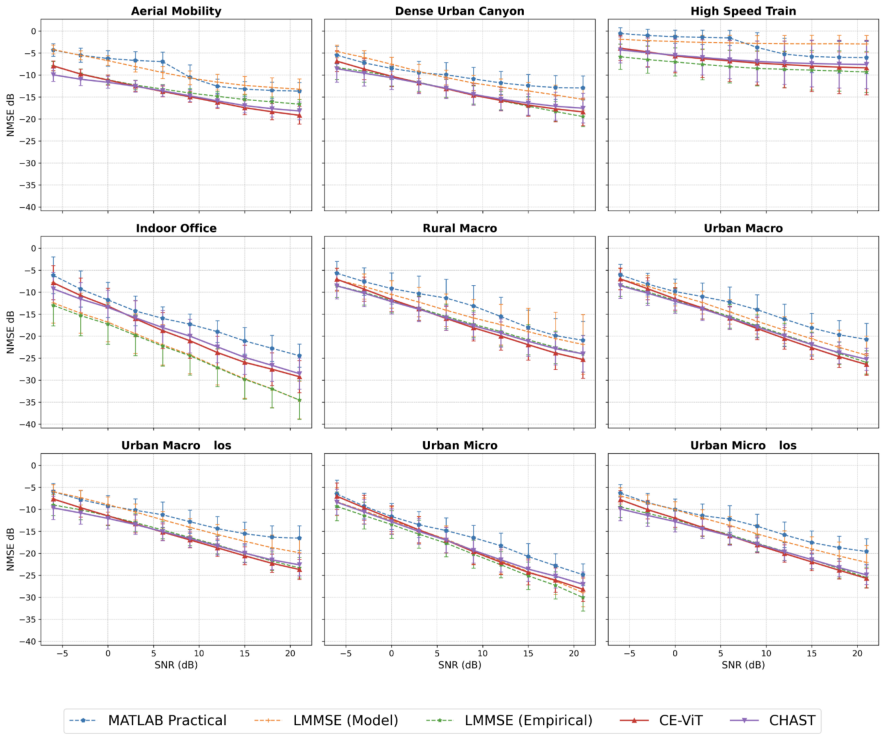


Figure 4: Per-scenario NMSE performance of CHAST against baseline methods. The top row shows challenging conditions with high mobility and/or large delay spreads while the middle and bottom rows present standard 3GPP scenarios.

As illustrated in Figure 4, the results underscore the limitations of model-based estimation. The Model based LMMSE, while strong in benign indoor conditions, fails in challenging scenarios with high mobility or delay spread (e.g., Rural Macro, Dense Urban Canyon) due to a fundamental model mismatch between its assumptions and the complex channel physics. In contrast, DL-based models exhibit robust adaptability across all tested environments. This advantage is particularly pronounced in the low-SNR regime, where the model's ability to jointly denoise and interpolate provides a substantial performance gain over traditional estimators that suffer from noise enhancement.