

# SCA-LLM: Spectral-Attentive Channel Prediction with Large Language Models in MIMO-OFDM

Ke He *Member, IEEE*, Le He *Member, IEEE*, Lisheng Fan *Member, IEEE*, Xianfu Lei *Member, IEEE*, Thang X. Vu *Member, IEEE*, George K. Karagiannidis, *Fellow, IEEE* and Symeon Chatzinotas, *Fellow, IEEE*

**Abstract**—In recent years, the success of large language models (LLMs) has inspired growing interest in exploring their potential applications in wireless communications, especially for channel prediction tasks. However, directly applying LLMs to channel prediction faces a domain mismatch issue stemming from their text-based pre-training. To mitigate this, the “adapter + LLM” paradigm has emerged, where an adapter is designed to bridge the domain gap between the channel state information (CSI) data and LLMs. While showing initial success, existing adapters may not fully exploit the potential of this paradigm. To address this limitation, this work provides a key insight that learning representations from the spectral components of CSI features can more effectively help bridge the domain gap. Accordingly, we propose a spectral-attentive framework, named SCA-LLM, for channel prediction in multiple-input multiple-output orthogonal frequency division multiplexing (MIMO-OFDM) systems. Specifically, its novel adapter can capture finer spectral details and better adapt the LLM for channel prediction than previous methods. Extensive simulations show that SCA-LLM achieves state-of-the-art prediction performance and strong generalization, yielding up to  $-2.4$  dB normalized mean squared error (NMSE) advantage over the previous LLM based method. Ablation studies further confirm the superiority of SCA-LLM in mitigating domain mismatch.

**Index Terms**—channel prediction, large language models (LLMs), domain mismatch, MIMO-OFDM.

## I. INTRODUCTION

**M**ultiple-input multiple-output orthogonal frequency division multiplexing (MIMO-OFDM) [1] is a key technology offering reliable, high-speed wireless communication capabilities in 5G and future communication systems [2], [3]. This approach combines the spatial benefits of MIMO [4], [5] with the ability of OFDM to handle frequency-selective fading [6]. As a result, significant gains in spectral efficiency and link stability are achieved in MIMO-OFDM systems [1]. To further improve performance, adaptive techniques such as dynamic resource allocation [7], precoding [8], and beamforming [9], [10] are often employed in modern deployments.

K. He, T. X. Vu and S. Chatzinotas are with the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, L-1855 Luxembourg, Luxembourg (e-mail: ke.he, thang.vu, symeon.chatzinotas@uni.lu).

L. He, and L. Fan are both with School of Computer Science of Guangzhou University, Guangzhou 510006, China (e-mail: hele@gzhu.edu.cn, ls-fan@gzhu.edu.cn).

X. Lei is with the School of Information Science and Technology, Institute of Mobile Communications, Southwest Jiaotong University, Chengdu 610031, China (e-mail: xlei@home.swjtu.edu.cn).

G. K. Karagiannidis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece and also with Cyber Security Systems and Applied AI Research Center, Lebanese American University (LAU), Lebanon (e-mail: geokarag@auth.gr).

The corresponding author of this paper is Lisheng Fan.

The success of these techniques, depends heavily on the availability of accurate and timely channel state information (CSI) at the transceiver. However, the acquisition of CSI is often challenging in practice due to the rapid variations of wireless channels over time. As a result, the acquired CSI quickly becomes outdated—a phenomenon known as *channel aging* [11], particularly in high mobility scenarios or high round-trip scenarios such as non-terrestrial networks (NTN). In addition, frequent channel estimation procedures demand substantial pilot overhead consuming valuable communication resources [12]. To address these issues, channel prediction has emerged as a promising technique. By forecasting future channel conditions, proactive optimization of transmission strategies becomes possible. Consequently, the negative effects of channel aging can be mitigated and the need for frequent pilot transmissions can be significantly reduced. This allows adaptive MIMO-OFDM systems to better realize their full potential.

## A. Prior Works

Channel prediction methods can be broadly categorized into traditional methods and deep learning (DL) based methods. Traditional methods are typically based on explicit mathematical models derived from statistical assumptions or simplified representations of the underlying channel physics. However, it is often difficult for these methods to perform well in complex and dynamic wireless environments. For instance, although advanced model-based techniques, such as the Prony-based angle-delay domain (PAD) method [13], are designed to capture specific channel structures and potentially offer improved accuracy under certain conditions, fundamental limitations are often encountered. While analytical guarantees are provided by PAD, frequent updates to its model parameters are required for adaptation to environmental changes [14]. Substantial computational overhead is incurred by this process as well. There is still a need for more robust and generalizable channel prediction approaches capable of effective adaptation to diverse environments.

In response to this need, DL based methods [14]–[20] have emerged as a promising alternative. Rather than relying on predefined mathematical models, DL methods are data-driven approaches that learn complex dependencies and predictive patterns directly from historical channel observations. Architectures such as recurrent neural networks (RNNs) [16], long short-term memory (LSTM) networks [18], gated recurrent units (GRUs) [19], and transformers [20] have been

employed, attempting to model the complex spatio-temporal dynamics of wireless channels. For example, focusing on the angle-delay domain for high-mobility massive MIMO systems, a complex-valued neural network (CVNN) based prediction method has been introduced in [14]. This data-driven approach performs element-wise prediction of the angle-delay channel response matrix (ADCRM). By learning directly from historical channel data represented in the angle-delay domain, this method achieved improved prediction accuracy compared to the traditional model-based PAD method [13]. Furthermore, a transformer-based method for channel prediction was introduced in [20]. This approach uses the attention mechanism within the transformer model to establish the relationship between past CSI and future channels. It gives more importance to the most relevant past information, which improves the prediction accuracy and overall communication performance. Additionally, a unified pilot-to-precoder scheme has been proposed either in [20], using a single transformer model to integrate channel estimation, prediction, and precoding, thereby reducing computational complexity. While these approaches have often demonstrated improved accuracy over traditional methods, they still have limitations in practice. Specifically, these models are typically trained for specific tasks and environments. When encountering unseen communication environments without significant retraining, they often exhibit a lack of robustness due to limited sequence modeling capabilities. Therefore, there is a need to develop novel approaches that are more robust for unseen environments.

In recent years, large language models (LLMs) [21]–[24] have shown impressive abilities in pattern recognition and sequence modeling thanks to training on massive amounts of sequential data. Their success has led to a growing interest in using LLMs to solve problems in the physical layer of wireless communications [25]–[28]. One area of particular interest is channel prediction, which naturally fits the strength of LLMs as a sequential forecasting task. However, directly applying LLMs, which are pre-trained primarily on textual data to specialized physical layer signals such as wireless channels, often leads to a significant *domain mismatch*. LLMs, pre-trained mainly on text data, lack a good understanding of the complex spatio-temporal dynamics governed by wireless channel characteristics. Therefore, directly applying LLMs to raw or minimally processed CSI data significantly restricts their predictive capabilities for channel prediction tasks, which will be shown by the ablation study presented later in this work.

Recognizing this, the recent works such as LLM4CP [29] have proposed the “adapter + LLM” paradigm for channel prediction, where a specialized adapter preprocesses the CSI data before feeding them to a pre-trained LLM. The adapter of LLM4CP employs channel attention layers inspired by the SE-Net [30], termed CSI-attention modules, for feature analysis and modeling. These modules attempt to capture the spatio-temporal correlations within the CSI data and transform the data into a feature representation more easily interpreted by the subsequent LLM, aiming to mitigate the domain mismatch issue. As an effective initial effort, LLM4CP

validated the feasibility and potential of the “adapter + LLM” paradigm. The effectiveness of the LLM4CP’s adapter relies on the global average pooling (GAP) operation within its CSI-attention modules to generate the channel attention weights. By using GAP, only the lowest frequency component of the spatio-temporal feature spectrum is preserved, while all other frequency components are discarded [31]. Although preserving only the lowest frequency component, experiments in [29] suggested this representation learning method still offers some benefit for modeling multipath effect and extracting underlying physical propagation features.

However, the discarded spectral components also encode valuable information about the underlying wireless channel characteristics. This information reflects phenomena such as multipath effect within specific scattering environments, spatio-temporal correlations across the antenna array, mobility-induced Doppler effects, and environmental noise during propagation. Since accurate modeling of MIMO channel dynamics relies on such details, discarding them consequently restricts the performance potential of LLM4CP. Therefore, despite these early advances, there remains a significant gap for existing methods to fully exploiting LLMs for channel prediction.

## B. Contributions

To fill the gap, this work propose SCA-LLM, a spectral-attentive framework with a LLM well pre-trained on a very large corpus of general knowledge data, for accurate channel prediction in MIMO-OFDM systems. The proposed SCA-LLM comprises three sequential components: the spectral-attentive adapter, a pre-trained LLM, and the output head. Specifically, the main contributions of this work are summarized as follows.

- 1) We propose the SCA-LLM, a novel spectral-attentive framework for MIMO-OFDM channel prediction, which integrates a pre-trained LLM with a specialized spectral-attentive adapter to more effectively mitigate the domain-mismatch issue.
- 2) We provide a key insight that learning representations from the spectral components of CSI features can more effectively help bridge the gap between the CSI data and the LLM. Accordingly, we introduce the spectral channel adaptation (SCA) module within the adapter. This module adopts multi-spectral channel attention layers, which utilize the 2-dimensional (2-D) discrete cosine transform (DCT) [31], [32] for feature analysis and modeling. Unlike LLM4CP’s CSI-attention modules, the SCA module leverages multiple DCT bases to selectively preserve multiple spectral components within features, ranging from low frequency components to high frequency components. This helps the SCA-LLM to better model the wireless channel characteristics and bridge the domain gap between the CSI data and the LLM more effectively, thereby improving the prediction performance significantly.
- 3) Extensive simulations demonstrate that the proposed SCA-LLM achieves state-of-the-art (SOTA) prediction performance and exhibits strong robustness and generalization capabilities, across diverse communication

scenarios, mobility conditions, and signal-to-noise ratios (SNRs). Ablation studies further validate the effectiveness and superiority of the proposed SCA-LLM, confirming its significant contribution to adapting the LLM for channel prediction task.

### C. Organization and Notation

1) *Organization*: The remainder of this paper is organized as follows. Section II introduces the system model. Section III formulates the channel prediction problem. Section IV details the proposed SCA-LLM framework. Section V presents the numerical results and discusses the performance evaluation. Finally, Section VI concludes the paper.

2) *Notation*: Throughout this paper, we use bold uppercase letters, e.g.,  $\mathbf{H}$ , to denote matrices and bold lowercase letters, e.g.,  $\mathbf{a}$ , for vectors. The operators  $(\cdot)^T$ ,  $(\cdot)^*$ , and  $(\cdot)^H$  represent the transpose, complex conjugate, and hermitian transpose (conjugate transpose), respectively. The operators  $\text{Re}(\cdot)$  and  $\text{Im}(\cdot)$  extract the real and imaginary parts of a complex number or matrix, respectively. The  $\text{vec}(\cdot)$  operator vectorizes a matrix by stacking its columns into a single column vector. The squared Frobenius norm operator is denoted by  $\|\cdot\|_F^2$ . ZOA, ZOD, AOA, and AOD stand for zenith angle of arrival, zenith angle of departure, azimuth angle of arrival, and azimuth angle of departure, respectively.

## II. SYSTEM MODEL

### A. MIMO-OFDM System

In this work, we consider a MIMO-OFDM communication system where a base station (BS) equipped with  $N_{BS}$  antennas communicates with user equipments (UEs) with  $N_{UE}$  antennas. In order to optimize the data transmission to the users, the BS must conduct CSI estimation for all users. Assuming orthogonal pilot sequences across the users, the CSI estimation and prediction procedure can be applied to every user independently. Without loss of generality, we focus on a particular active user in the system, thus omit the user index [20]. The total bandwidth  $B$  is divided into  $K$  orthogonal subcarriers, indexed by  $k = 0, 1, \dots, K-1$ , with subcarrier spacing  $\Delta f$ . A cyclic prefix (CP), configured to be longer than the maximum channel delay spread, is added to each OFDM symbol. Moreover, the system operates in time-division duplexing (TDD) mode, where the uplink (UL) and downlink (DL) transmissions use the same frequency band but occur at different time slots within a communication frame. Each frame allocates specific time slots for transmitting uplink pilot signals, such as sounding reference signals (SRS) from the UE. These pilots enable the BS to periodically estimate the UL channel state information (CSI). Following the pilot transmissions, the frame structure provides configurable time slots designated for UL and DL data transmission.

### B. Channel Model

To practically reflect the wireless propagation environment, this work adopts the channel model based on the 3GPP TR

38.901 standard [33]. This channel model is the implementation of the geometry-based stochastic channel model (GSCM) framework [34]. Specifically, the UL channel between the BS and UE can be modeled as:

$$\mathbf{H}_{UL}(t, f) = \sum_{n=1}^N \sum_{m=1}^M \alpha_{n,m} e^{-j2\pi f \tau_{n,m}} e^{j2\pi \nu_{n,m} t} \mathbf{a}_{BS}(\theta_{n,m,ZOA}, \phi_{n,m,AOA}) \mathbf{a}_{UE}^H(\theta_{n,m,ZOD}, \phi_{n,m,AOD}), \quad (1)$$

where  $N$  is the number of propagation clusters, and  $M$  is the number of rays within each cluster. For the  $m$ -th ray of the  $n$ -th cluster,  $\alpha_{n,m}$  is its complex gain,  $\tau_{n,m}$  is its delay, and  $\nu_{n,m}$  is its Doppler shift.  $f$  denotes the frequency of the given carrier, and  $t$  represents the given time. The vectors  $\mathbf{a}_{BS}(\theta_{n,m,ZOA}, \phi_{n,m,AOA})$  and  $\mathbf{a}_{UE}(\theta_{n,m,ZOD}, \phi_{n,m,AOD})$  are the array response vectors of the BS and UE antennas, respectively. These two vectors depend on the ray's ZOA ( $\theta_{n,m,ZOA}$ ), AOA ( $\phi_{n,m,AOA}$ ), ZOD ( $\theta_{n,m,ZOD}$ ), and AOD ( $\phi_{n,m,AOD}$ ). The Doppler frequency component  $\nu_{n,m}$  depends on the arrival angles ( $\theta_{n,m,ZOA}, \phi_{n,m,AOA}$ ) and the UE velocity vector  $\mathbf{v}$ .

The above channel model can account for factors like multipath propagation through scattering clusters, path delays, and the angular spread of signals. This indicates that the channel model can capture dynamic spatio-temporal correlations of the MIMO-OFDM channel, which is influenced by factors such as user mobility and variations in the scattering environment. In addition, the specific parameters within the channel model, which dictate channel characteristics, are determined by the chosen deployment scenario, such as urban macro (UMa) or urban micro (UMi).

## III. PROBLEM FORMULATION OF CHANNEL PREDICTION

In this work, our goal is to predict the future DL CSI sequence based on the historical UL CSI sequence. Predicting future DL CSI allows the base station to prepare efficient transmission strategies, such as beamforming, ahead of time. This avoids the delays and overhead associated with measuring the current channel state, leading to faster and more reliable communication.

### A. TDD Operation and Channel Reciprocity

Since the system adopts the TDD mode, we can leverage the channel reciprocity. This property indicates that the DL CSI at a certain time  $t$  and a subcarrier  $k$  with frequency  $f$ , denoted as  $\mathbf{H}_{DL}(t, f) \in \mathbb{C}^{N_{UE} \times N_{BS}}$ , can be directly obtained from the corresponding UL CSI  $\mathbf{H}_{UL}(t, f) \in \mathbb{C}^{N_{BS} \times N_{UE}}$  by:

$$\mathbf{H}_{DL} = \mathbf{H}_{UL}^T. \quad (2)$$

Leveraging this reciprocity, the channel prediction problem is simplified from DL channel acquisition complexities, and predicting future UL CSI becomes sufficient to determine future DL CSI. This simplification therefore makes it much easier to directly evaluate our method's prediction performance and clearly demonstrate its core contribution.

### B. Channel Prediction Problem Formulation

Based on the channel reciprocity discussed above, our task simplifies to predicting future UL CSI using historical observations. This historical CSI sequence is typically acquired through channel estimation at previous time steps, or directly incorporating previously predicted CSI samples. For the estimation process, the UE transmits known pilot sequences, denoted by  $\mathbf{X}_p \in \mathbb{C}^{N_{UE} \times T_p}$ , over  $T_p$  symbol durations within designated time slots. The corresponding signal received at the BS,  $\mathbf{Y}_p \in \mathbb{C}^{N_{BS} \times T_p}$ , can be modeled as:

$$\mathbf{Y}_p(t) = \mathbf{H}_{UL}(t, f)\mathbf{X}_p + \mathbf{N}(t), \quad (3)$$

where  $\mathbf{H}_{UL}(t, f) \in \mathbb{C}^{N_{BS} \times N_{UE}}$  is the UL channel matrix at time  $t$  and frequency  $f$ , and  $\mathbf{N}(t) \in \mathbb{C}^{N_{BS} \times T_p}$  represents the additive noise. For notational clarity, we often omit the frequency index  $f$  from  $\mathbf{H}_{UL}(t, f)$  and use  $\mathbf{H}_{UL}(t)$ , assuming the operations are performed per frequency, or that the model implicitly handles the frequency dimension. Channel estimation techniques, such as least squares (LS) or minimum mean square error (MMSE), can then be applied to obtain an estimate  $\hat{\mathbf{H}}_{UL}(t)$  from  $\mathbf{Y}_p(t)$  and the known  $\mathbf{X}_p$ .

These channel estimates, possibly augmented with prior predictions, are from the historical data required for the current channel prediction task. Let the sequence of available UL CSI samples over a past window of  $L$  time steps ending at the current time  $t_0$  be denoted by:

$$\mathcal{H}_{\text{past}} = \{\mathbf{H}_{UL}(t_0 - L + 1), \dots, \mathbf{H}_{UL}(t_0 - 1), \mathbf{H}_{UL}(t_0)\}, \quad (4)$$

where each  $\mathbf{H}_{UL}(t)$  in the sequence represents the most recently available CSI for that time step, obtained via estimation or prediction.

The objective of this work is to predict the sequence of UL CSI samples for the next  $P$  future time steps, starting from  $t_0 + 1$ :

$$\mathcal{H}_{\text{future}} = \{\mathbf{H}_{UL}(t_0 + 1), \mathbf{H}_{UL}(t_0 + 2), \dots, \mathbf{H}_{UL}(t_0 + P)\}. \quad (5)$$

We aim to learn a predictive model, represented by a function  $\mathcal{F}$ , that maps the historical sequence  $\mathcal{H}_{\text{past}}$  to the future sequence  $\mathcal{H}_{\text{future}}$ :

$$\hat{\mathcal{H}}_{\text{future}} = \mathcal{F}(\mathcal{H}_{\text{past}}; \Theta), \quad (6)$$

where  $\hat{\mathcal{H}}_{\text{future}} = \{\hat{\mathbf{H}}_{UL}(t_0 + 1), \hat{\mathbf{H}}_{UL}(t_0 + 2), \dots, \hat{\mathbf{H}}_{UL}(t_0 + P)\}$  is the sequence of predicted future UL CSI samples, and  $\Theta$  represents the learnable parameters of the prediction model  $\mathcal{F}$ . The model  $\mathcal{F}$  is typically trained by minimizing a loss function that quantifies the discrepancy between the predicted sequence  $\hat{\mathcal{H}}_{\text{future}}$  and the ground-truth sequence  $\mathcal{H}_{\text{future}}$ , such as the mean squared error (MSE) or the normalized mean squared error (NMSE).

Once the future UL CSI sequence  $\hat{\mathcal{H}}_{\text{future}}$  is predicted, the corresponding future DL CSI sequence  $\hat{\mathcal{H}}_{\text{DL}, \text{future}} = \{\hat{\mathbf{H}}_{DL}(t_0 + 1), \dots, \hat{\mathbf{H}}_{DL}(t_0 + P)\}$  can be readily obtained via channel reciprocity using (2):

$$\hat{\mathbf{H}}_{DL}(t_0 + p) = \left(\hat{\mathbf{H}}_{UL}(t_0 + p)\right)^T, \quad \text{for } p = 1, \dots, P. \quad (7)$$

Therefore, the channel prediction problem is defined as a sequence forecasting task where the input is a multivariate

time series of past MIMO CSI samples, and the output is the predicted sequence of future MIMO CSI samples. These time series essentially reflect the complex spatio-temporal correlations within the MIMO channel, which arises from factors such as user mobility and the scattering environment. Accurately leveraging these correlations for prediction is the core challenge that needs to be addressed by the prediction model  $\mathcal{F}$ .

### IV. THE PROPOSED SCA-LLM FRAMEWORK

In this section, we introduce the proposed SCA-LLM framework, which is designed to effectively adapt pre-trained LLMs for channel prediction. The overall architecture of the SCA-LLM is shown in Fig. 1. This framework processes the historical CSI sequence sequentially through the following components: 1) the normalization layer; 2) the spectral-attentive adapter consisting of the CSI embedding layer and the spectral channel adaptation (SCA) module; 3) the pre-trained LLM consisting of the LLM embedding layer with positional embedding and the LLM backbone; 4) the output head; 5) and finally the de-normalization layer.

Since the SCA-LLM framework operates on real-valued inputs, the complex-valued CSI data should be converted into a suitable real-valued representation. To this end, the complex-valued CSI matrix  $\mathbf{H}_{UL}(t) \in \mathbb{C}^{N_{BS} \times N_{UE}}$  for each subcarrier  $k$  at time step  $t$  is first vectorized and converted into a real-valued representation  $\mathbf{h}(t) \in \mathbb{R}^N$ , which is given by:

$$\mathbf{h}(t) = \begin{bmatrix} \text{Re}(\text{vec}(\mathbf{H}_{UL}(t))) \\ \text{Im}(\text{vec}(\mathbf{H}_{UL}(t))) \end{bmatrix}, \quad (8)$$

where  $N = 2N_{BS}N_{UE}$ . After that, the historical CSI sequence is formed by collecting these vectors  $\mathbf{h}(t)$  over  $L$  time steps, denoted as  $\mathbf{H} \in \mathbb{R}^{B \times L \times N}$ , where  $B$  is the batch size and  $L$  is the number of past time steps. Additionally, since the subcarriers within OFDM are orthogonal, we can process each subcarrier in parallel [1] for computational efficiency. That is, we rearrange the CSI data such that the individual subcarriers within the historical CSI sequence are merged into the batch dimension  $B$ . After the above operations, the historical CSI sequence  $\mathbf{H} \in \mathbb{R}^{B \times L \times N}$  is prepared to serve as the input to the SCA-LLM and is processed sequentially by the following components.

#### A. Normalization

This step aims to standardize the input data distribution, which often helps stabilize model training and improve convergence speed. The normalization uses the mean and standard deviation calculated across all elements within the input tensor  $\mathbf{H} \in \mathbb{R}^{B \times L \times N}$ , which is expressed as:

$$\mathbf{H}_{\text{norm}} = \frac{\mathbf{H} - \mu_{\mathbf{H}}}{\sigma_{\mathbf{H}}}, \quad (9)$$

where  $\mu_{\mathbf{H}}$  and  $\sigma_{\mathbf{H}}$  denote the mean and standard deviation computed over all dimensions (batch, time, and feature) of the input tensor  $\mathbf{H}$ . The resulting normalized tensor  $\mathbf{H}_{\text{norm}}$  serves as the input to the subsequent spectral-attentive adapter. In addition, these calculated statistics,  $\mu_{\mathbf{H}}$  and  $\sigma_{\mathbf{H}}$ , are stored and

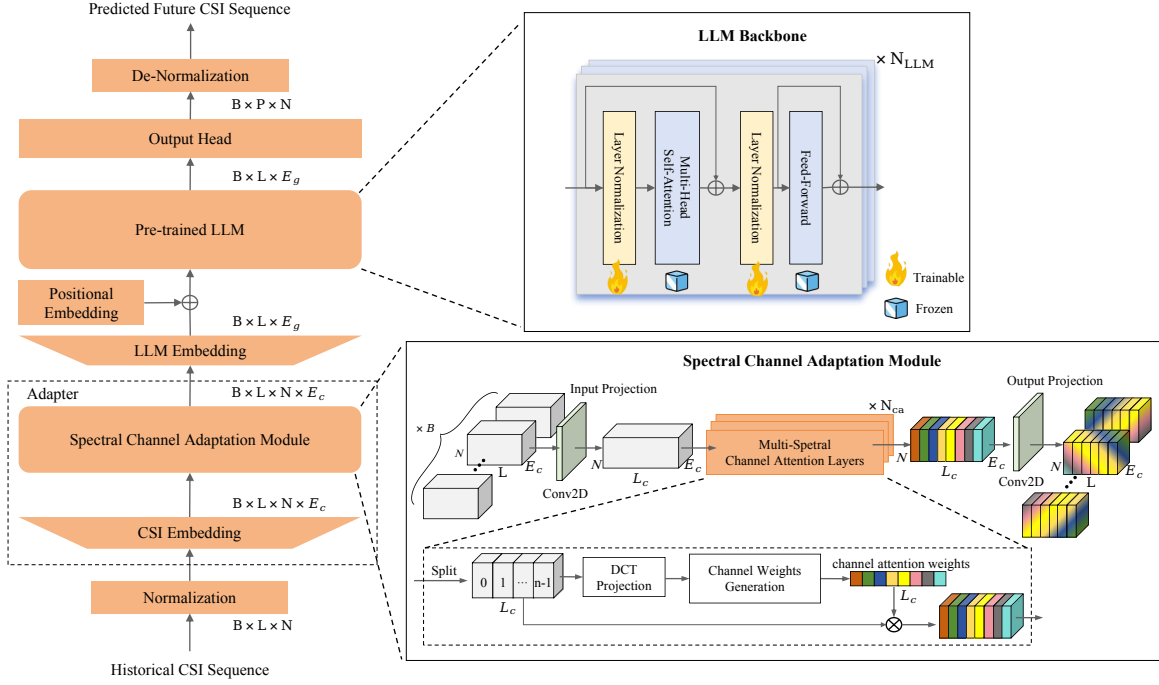


Fig. 1: The architecture of the proposed SCA-LLM framework. The adapter component encompasses the CSI embedding layer and the spectral channel adaptation (SCA) module.

subsequently used by the de-normalization layer to transform the model's final predictions back into the original scale of the input CSI data.

### B. Adapter

1) *CSI Embedding*: The CSI embedding layer receives the normalized historical CSI sequence,  $\mathbf{H}_{\text{norm}} \in \mathbb{R}^{B \times L \times N}$ , as the input. This layer transforms the  $N$ -dimensional ( $N$ -D) feature vector at each time step (corresponding to the last dimension of  $\mathbf{H}_{\text{norm}}$ ) into an embedding representation within a higher-dimensional latent space, resulting in a 4-D tensor  $\mathbf{H}_{\text{emb}} \in \mathbb{R}^{B \times L \times N \times E_c}$  as the output. Here,  $E_c$  is the target embedding dimension for each of the original  $N$  features.

Specifically, this transformation is achieved through two sequential point-wise linear projections, which operate independently on the feature vector at each time step. Mathematically, this process can be formulated as:

$$\mathbf{H}_{\text{emb}} = \text{Reshape}_{B,L,N,E_c}(\text{Linear}_2(\text{Linear}_1(\mathbf{H}_{\text{norm}}))), \quad (10)$$

where  $\text{Linear}_1$  represents the first point-wise linear projection (mapping  $N \rightarrow N_{hs}$  features) and  $\text{Linear}_2$  represents the second point-wise linear projection (mapping  $N_{hs} \rightarrow N \cdot E_c$  features). The  $\text{Reshape}_{B,L,N,E_c}(\cdot)$  function rearranges the last dimension of the output tensor from  $(N \cdot E_c)$  into  $(N, E_c)$ . In practice, these projections are implemented using 1-D convolutional layers with a kernel size of 1 and a stride of 1 for computational efficiency. This transformation is designed to prepare the data for the subsequent SCA module, and the embedded representation  $\mathbf{H}_{\text{emb}}$  serves as the input to the subsequent SCA module.

2) *Spectral Channel Adaptation Module*: This module is the core component designed to bridge the domain gap between the CSI data and the pre-trained LLM. As highlighted in the Introduction, directly applying text-pretrained LLMs to CSI data faces challenges due to domain mismatch. The SCA module addresses this by adaptively refining the features based on their spectral components. Specifically, it employs multi-spectral channel attention layers that utilize the 2-D discrete cosine transform (DCT) to capture richer information than methods relying solely on the lowest frequency component (like GAP used in [29]). This allows the module to better model wireless channel characteristics and adapt the features into a format more effectively processed by the subsequent LLM. The architecture of this module, detailed in the lower panel of Fig. 1, consists of an input projection layer, a stack of multi-spectral channel attention layers, and an output projection layer.

- **Input Projection**: The first step within the SCA module is an input projection layer. It receives the output from the CSI embedding layer,  $\mathbf{H}_{\text{emb}} \in \mathbb{R}^{B \times L \times N \times E_c}$ . Its purpose is to transform the input features while mapping the temporal dimension  $L$  to an intermediate dimension  $L_c$ , suitable for the subsequent multi-spectral channel attention layers. This transformation is achieved using a 2-D convolutional layer. Specifically, the convolution operates on the  $(N, E_c)$  spatial dimensions with a kernel size of  $(3, 3)$  and appropriate padding to maintain these dimensions. This operation takes the  $L$  feature maps corresponding to the time dimension as input and produces  $L_c$  output feature maps. Mathematically, this can

be expressed as:

$$\mathbf{H}_c^{(0)} = \text{Conv2D}_{k=(3,3)}^{(L \rightarrow L_c)}(\mathbf{H}_{\text{emb}}), \quad (11)$$

where  $\text{Conv2D}_{k=(3,3)}^{(L \rightarrow L_c)}(\cdot)$  denotes the 2-D convolution operation described above. This operation mixes local spatial information within the  $(N, E_c)$  plane while projecting the time dimension  $L$  to  $L_c$ , resulting in  $\mathbf{H}_c^{(0)} \in \mathbb{R}^{B \times L_c \times N \times E_c}$ . This tensor serves as the input to the first multi-spectral channel attention layer.

- **Multi-Spectral Channel Attention Layers:** After that, a stack of  $N_{ca}$  multi-spectral channel attention layers [31], are applied sequentially to these feature maps for feature analysis and modeling. Let  $\mathbf{H}_c^{(i)}$  be the input of the  $i$ -th multi-spectral channel attention layer ( $\mathbf{H}_c^{(0)} = \mathbf{H}_c$ ) and  $MSCA(\cdot)$  denote the operator of each layer, then we have:

$$\mathbf{H}_c^{(i+1)} = MSCA(\mathbf{h}_c^{(i)}) \in \mathbb{R}^{B \times L_c \times N \times E_c}. \quad (12)$$

For improved readability and to avoid excessive detail in this section, a full description of the multi-spectral channel attention layer is available in Appendix A.

- **Output Projection:** Following the multi-spectral channel attention layers, the resulting tensor  $\mathbf{H}_c^{(N_{ca})} \in \mathbb{R}^{B \times L_c \times N \times E_c}$  is processed by the output projection layer. This layer maps the intermediate dimension  $L_c$  back to the original time dimension  $L$ . Similar to the input projection, this transformation is implemented using a 2-D convolutional layer, mapping  $L_c$  feature maps as the input to  $L$  feature maps as the output. Mathematically, this process is given by:

$$\mathbf{H}_{\text{sca}} = \text{Conv2D}_{k=(3,3)}^{(L_c \rightarrow L)}(\mathbf{H}_c^{(N_{ca})}). \quad (13)$$

The resulting tensor  $\mathbf{H}_{\text{sca}} \in \mathbb{R}^{B \times L \times N \times E_c}$  is the final output of the SCA module, which is then passed to the LLM embedding layer.

### C. Pre-trained LLM

1) *LLM Embedding:* Following the SCA module, the resulting tensor, let's denote it as  $\mathbf{H}_{\text{sca}} \in \mathbb{R}^{B \times L \times N \times E_c}$ , is fed into the LLM embedding layer. The purpose of this layer is to project the adapted CSI features into the embedding space expected by the pre-trained LLM, which has a dimension of  $E_g$ .

Specifically, the transformation first involves flattening the last two dimensions ( $N$  and  $E_c$ ) of  $\mathbf{H}_{\text{sca}}$ . Then, a linear projection is applied to map the resulting  $N \cdot E_c$ -dimensional features to the target  $E_g$  dimension. Mathematically, this projection step can be represented as:

$$\mathbf{H}_{\text{LLM\_emb}} = \text{Linear}_{\text{Emb}}(\text{Flatten}_{N, E_c}(\mathbf{H}_{\text{sca}})), \quad (14)$$

where  $\text{Flatten}_{N, E_c}(\cdot)$  denotes the operation that combines the last two dimensions, and  $\text{Linear}_{\text{Emb}}$  represents the linear projection layer mapping  $N \cdot E_c \rightarrow E_g$  features. The output tensor  $\mathbf{H}_{\text{LLM\_emb}} \in \mathbb{R}^{B \times L \times E_g}$ . In practice,  $\text{Linear}_{\text{Emb}}$  is efficiently implemented using a 1-D convolutional layer too.

Furthermore, to provide the LLM with information about the order of time steps within the sequence, positional embeddings  $\mathbf{P} \in \mathbb{R}^{L \times E_g}$  are added to the projected features. The specific method for generating  $\mathbf{P}$  (e.g., learned embeddings, sinusoidal functions) depends on the chosen LLM architecture. The addition operation is broadcast across the batch dimension  $B$ :

$$\mathbf{X}_{\text{in}} = \mathbf{H}_{\text{LLM\_emb}} + \mathbf{P}. \quad (15)$$

The resulting tensor  $\mathbf{X}_{\text{in}} \in \mathbb{R}^{B \times L \times E_g}$  serves as the final input sequence to the LLM backbone.

2) *LLM Backbone:* The core sequence processing is performed by the pre-trained LLM backbone. In this work, we utilize GPT-2 [24] as the foundational LLM. The LLM takes the prepared input sequence  $\mathbf{X}_{\text{in}} \in \mathbb{R}^{B \times L \times E_g}$  and processes it through its stack of transformer layers, denoted collectively as  $LLM(\cdot)$ . This transformation yields the final hidden state sequence:

$$\mathbf{H}_{\text{LLM}} = LLM(\mathbf{X}_{\text{in}}), \quad (16)$$

where  $\mathbf{H}_{\text{LLM}} \in \mathbb{R}^{B \times L \times E_g}$  represents the sequence of output hidden states from the LLM.

As depicted in the upper panel of Fig. 1, the GPT-2 architecture consists of  $N_{\text{LLM}}$  stacked transformer decoder blocks built upon standard components: multi-head self-attention (MHSA), layer normalization (LN), and feed-forward networks (FFN). To effectively adapt the pre-trained model for the channel prediction task while preserving its powerful general sequence modeling capabilities acquired during large-scale pre-training, we employ a selective fine-tuning strategy. As indicated by the symbols in Fig. 1, the parameters of the MHSA and FFN layers are kept frozen (ice cube symbol), retaining the LLM's core knowledge. Conversely, the parameters within the LN layers are fine-tuned (flame symbol), allowing the model to adapt specifically to the statistical properties and nuances of the embedded CSI data. This strategy aims to bridge the domain gap effectively while leveraging the LLM's pre-trained strengths. Notably, the proposed SCA-LLM framework is flexible regarding the choice of the LLM backbone. Alternatives such as Qwen2 [21] or Llama3 [22] could potentially be integrated. The selection depends on the specific requirements concerning computational resources and desired prediction performance.

### D. Output Head

The output head module receives the final hidden state sequence from the pre-trained LLM, denoted as  $\mathbf{H}_{\text{LLM}} \in \mathbb{R}^{B \times L \times E_g}$ . Its purpose is to transform this sequence into the final normalized prediction sequence  $\hat{\mathbf{H}}_{\text{norm}} \in \mathbb{R}^{B \times P \times N}$ , where  $P$  is the desired future prediction length. This transformation involves two main steps: mapping the feature dimension from  $E_g$  back to  $N$ , and adjusting the sequence length from  $L$  to  $P$ .

First, the feature dimension is mapped using two sequential point-wise linear projections, similar to the CSI embedding layer but in reverse. Mathematically, this mapping step can be represented as:

$$\mathbf{H}' = \text{Linear}_4(\text{Linear}_3(\mathbf{H}_{\text{LLM}})), \quad (17)$$

where  $\mathbf{H}' \in \mathbb{R}^{B \times L \times N}$ .  $\text{Linear}_3$  maps features  $E_g \rightarrow N'_{hs}$  and  $\text{Linear}_4$  maps  $N'_{hs} \rightarrow N$ , with  $N'_{hs}$  being an intermediate feature dimension.

Second, the sequence length is transformed from  $L$  to  $P$ . This is achieved by applying linear transformations across the time dimension independently for each of the  $N$  features. To facilitate this, the tensor  $\mathbf{H}'$  is first permuted to  $(B, N, L)$ . Then, two sequential linear layers project the time dimension:

$$\hat{\mathbf{H}}_{\text{permuted}} = \text{Linear}_6 (\text{Linear}_5 (\mathbf{H}'_{\text{permuted}})), \quad (18)$$

where  $\mathbf{H}'_{\text{permuted}} \in \mathbb{R}^{B \times N \times L}$  and  $\hat{\mathbf{H}}_{\text{permuted}} \in \mathbb{R}^{B \times N \times P}$ .  $\text{Linear}_5$  maps the length  $L \rightarrow N''_{hs}$  and  $\text{Linear}_6$  maps  $N''_{hs} \rightarrow P$ , where  $N''_{hs}$  is an intermediate dimension for the sequence length transformation. Finally, the resulting tensor is permuted back to the desired shape  $\hat{\mathbf{H}}_{\text{norm}} \in \mathbb{R}^{B \times P \times N}$ .

Practically, all of these linear transformations are also efficiently implemented using 1-D convolutional layers.

### E. De-Normalization

The final step involves transforming the normalized predicted sequence  $\hat{\mathbf{H}}_{\text{norm}} \in \mathbb{R}^{B \times P \times N}$  back to the original scale of the input CSI data. This de-normalization process reverses the initial operation performed by the normalization layer. It utilizes the mean  $\mu_{\mathbf{H}}$  and standard deviation  $\sigma_{\mathbf{H}}$  computed and stored during the normalization step, as described in (9). The de-normalization is performed element-wise as follows:

$$\hat{\mathbf{H}}_{\text{pred}} = \hat{\mathbf{H}}_{\text{norm}} \cdot \sigma_{\mathbf{H}} + \mu_{\mathbf{H}}. \quad (19)$$

The resulting tensor  $\hat{\mathbf{H}}_{\text{pred}} \in \mathbb{R}^{B \times P \times N}$  represents the final predicted future CSI sequence in its original scale.

## V. NUMERICAL RESULTS

This section describes the environment setup, baseline methods, training details, evaluation results, and ablation studies.

### A. Environment Setup

To generate realistic time-varying channel data for training and evaluation, we use the widely adopted QuaDRiGa channel simulator [35], configured to generate datasets according to the 3GPP standards [33]. In particular, a MIMO-OFDM system is implemented in the TDD mode, and the BS and UE are equipped with  $N_{BS} = 4$  and  $N_{UE} = 4$  antennas, respectively. The aforementioned system uses an OFDM channel with the center frequency of 2.4 GHz and the bandwidth of 1.44 MHz, containing 12 subcarriers spaced by  $\Delta f = 120$  KHz. The time interval between consecutive CSI samples, defined by the SRS periodicity, is 0.625 ms. For the channel prediction task, we set the historical sequence length to  $L = 24$  and the future prediction length to  $P = 6$ .

In addition, the datasets are generated with a range of user mobility scenarios, specifically considering 13 different user velocities uniformly distributed between 0 km/h and 60 km/h. The models are trained and validated exclusively using data from the 3GPP UMa channel model under non-line-of-sight (NLOS). Specifically, the training dataset comprises 10400 samples (800 per velocity), while the validation dataset

contains 2600 samples (200 per velocity). During the training, the SNR for each sample is randomly selected from the range [0 dB, 20 dB] to enhance model robustness against varying noise levels. To evaluate performance and generalization capabilities, we prepare distinct test datasets for two different scenarios:

- **UMa NLOS Scenario:** This dataset serves to evaluate the model performance under conditions matching the training environment.
- **UMi NLOS Scenario:** This dataset assesses the model's ability to generalize to the 3GPP UMi channel model, which represents a different propagation environment not seen during the training.

The above two test datasets UMa and UMi consist of 13000 samples each, containing 1000 samples for every one of the 13 velocities (0 km/h to 60 km/h). This comprehensive testing setup allows us to assess both in-distribution accuracy and cross-scenario robustness.

### B. Baseline Methods

To comprehensively evaluate the performance of our proposed framework, we compare it against several relevant baseline methods. These include established sequence modeling architectures and the pioneering LLM4CP work [29], the first study to leverage LLMs for the channel prediction task. This method serves as a key LLM-based baseline in our evaluations. The specific baselines considered are:

- **RNN** [16]: A RNN based method. RNN based methods are frequently used in time-series analysis and serve as a fundamental baseline for sequence processing tasks like channel prediction.
- **LSTM** [18]: A LSTM based method. This method is specifically designed to capture long-range dependencies often present in time-varying channel data.
- **GRU** [19]: A GRU based method. It offers a variation on LSTMs, often providing comparable performance with greater computational efficiency.
- **Transformer** [20]: A transformer based architecture adapted for the sequence-to-sequence channel prediction task. This represents a non-recurrent approach to sequence modeling.
- **LLM4CP** [29]: The pioneering work that introduced the use of a fine-tuned pre-trained GPT-2 model for channel prediction in MISO-OFDM systems. We include LLM4CP as the key baseline representing the LLM-based approach in this area.

As to the network design of the proposed SCA-LLM framework, specific configurations are used in the experiments unless stated otherwise. The embedding size  $E_c = 64$  of the CSI embedding layer is used. The SCA module include  $N_{ca} = 4$  multi-spectral channel attention layers with an intermediate channel dimension  $L_c = 128$ . The DCT analysis dimensions  $H_{dct} = W_{dct} = 7$ , and  $n = 32$  DCT bases are adopted. Moreover, the smallest GPT-2 variant with the first  $N_{LLM} = 6$  decoder layers and embedding dimension  $E_g = 768$  is employed, and all convolutional layers share a hidden size of 128 (i.e.,  $N_{hs} = N'_{hs} = N''_{hs} = 128$ ).

The selection and configuration of RNN, LSTM, GRU, and Transformer mirrors the comparative analysis performed in the LLM4CP [29], facilitating a consistent evaluation context relative to prior work. For a fair comparison, the convolutional layers within the LLM4CP baseline is employed the same number of layers and hidden dimensions as our proposed SCA-LLM. The settings of GPT-2 model for LLM4CP is also the same.

### C. Training Details

The proposed framework and the baselines are first trained using the same training dataset before their performance are assessed on the test datasets. During the training phase, each method utilizes the ground truth future CSI sequences, denoted as  $\mathbf{H}_{\text{true}} \in \mathbb{R}^{B \times P \times N}$ . The objective of the training is to reduce the difference between the model's output predictions,  $\mathbf{H}_{\text{pred}}$ , and these ground truth targets. This difference is measured by the NMSE, which serves as the loss function of:

$$Loss = \frac{\|\mathbf{H}_{\text{pred}} - \mathbf{H}_{\text{true}}\|_F^2}{\|\mathbf{H}_{\text{true}}\|_F^2}. \quad (20)$$

The NMSE is also employed for calculating the validation loss during training. Ultimately, the version of each model exhibiting the lowest validation loss throughout the training process is preserved and utilized for the subsequent testing phase.

For the proposed SCA-LLM framework, it is worth noting that our adaptation strategy involves a selective fine-tuning process for the pre-trained backbone, as illustrated in Fig. 1. Specifically, within the pre-trained GPT-2, only the parameters associated with the learnable positional embedding and LN layers are fine-tuned (made trainable), while the parameters of the MHSA and FFN layers remain fixed (frozen) to preserve the general sequence modeling capabilities learned during pre-training. The other parts within the proposed framework, namely the adapter and the output head, are fully trainable. Consequently, despite the other parts being fully trainable, the total number of trainable parameters in our framework remains significantly smaller than its overall parameter count, as the frozen core components of the pre-trained LLM backbone account for the vast majority of the total parameters.

### D. Evaluation Results

In this part, we compare the performance of the proposed framework and the baselines using the NMSE metric in (20), a widely adopted metric that directly quantifies the accuracy of the channel prediction by measuring the normalized mean squared error relative to the true channel power. A smaller NMSE indicates a more accurate channel prediction, i.e., lower is better. In the comparisons, our proposed SCA-LLM framework is denoted as **Ours**.

#### 1) NMSE Performance Comparison under Varying SNR:

We evaluate the NMSE performance of our proposed framework against several baselines as the SNR varies. The results for UE velocities of 0 km/h and 50 km/h are presented in Fig. 2 and Fig. 3, respectively, covering both UMa and UMi NLOS scenarios. As depicted in these figures, a general trend

is observed across all methods and scenarios: the NMSE decreases significantly (indicating improved prediction accuracy) as the SNR increases from 0 dB to 20 dB. This aligns with the expectation that higher SNR levels facilitate more accurate channel prediction.

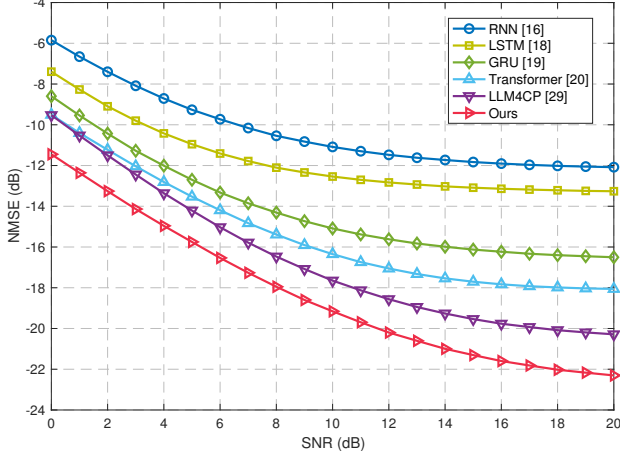
Focusing first on the static scenario (0 km/h), which is crucial for applications like fixed wireless access or low-mobility IoT devices, Fig. 2 details the performance comparison. In the UMa NLOS scenario (Fig. 2a), corresponding to the training distribution, our proposed method significantly outperforms all baselines. For instance, at an SNR of 20 dB, our method demonstrates a performance advantage of approximately  $-2.0$  dB over the next best baseline, LLM4CP (achieving NMSEs of roughly  $-22.3$  dB versus  $-20.3$  dB, respectively), and is substantially better than the traditional methods (RNN, LSTM, GRU, Transformer). When evaluating the generalization capability in the unseen UMi NLOS scenario (Fig. 2b), our method maintains its clear performance leadership. Interestingly, while all methods tend to exhibit a slight improvement in NMSE performance when generalizing from the trained UMa environment to the unseen UMi environment, our method still consistently achieves the lowest NMSE across the tested SNR range. Furthermore, the performance gap between our method and LLM4CP widens in the zero-shot UMi setting. In particular, at 20 dB SNR, this gap increases from approximately  $-2.0$  dB in UMa to  $-2.4$  dB in UMi, where the respective NMSEs are  $-22.7$  dB (our method) and  $-20.3$  dB (LLM4CP). This further highlights the superior robustness and generalization of our proposed framework.

Turning to the high-mobility scenario represented by a UE velocity of 50 km/h, Fig. 3 presents the corresponding NMSE versus SNR results. As expected, the faster channel variations at this higher speed lead to increased prediction difficulty, generally resulting in a noticeable degradation of NMSE performance across all methods compared to the static case. At the UE velocity of 50 km/h, transitioning from the trained UMa scenario to the unseen UMi scenario leads to a discernible performance degradation (increased NMSE) across all methods. This contrasts with the static case, where generalizing to the UMi scenario actually resulted in slight performance improvements for all methods. However, the fundamental performance patterns observed at 0 km/h are still persist. Our method continues to deliver the best performance in both the UMa NLOS (Fig. 3a) and the UMi NLOS (Fig. 3b) scenarios. For instance, in the UMa scenario at 10 dB SNR, our method demonstrates a clear advantage. It outperforms LLM4CP by approximately  $-0.5$  dB and also surpasses other baselines. Furthermore, this performance advantage widens in the zero-shot UMi setting. Here, at 10 dB SNR, our method leads LLM4CP by about  $-0.8$  dB. This larger performance lead over LLM4CP in the UMi scenario shows the robust generalization of our method, even under challenging high-mobility conditions.

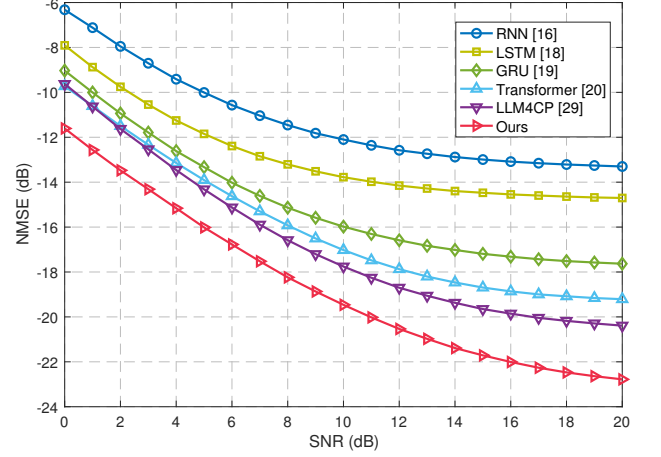
#### 2) NMSE Performance Comparison under Varying UE Velocity:

To further investigate the impact of user mobility on channel prediction accuracy, we analyze the NMSE performance across the full range of simulated UE velocities (0 km/h to 60 km/h) at a fixed SNR = 10 dB. Fig. 4 presents



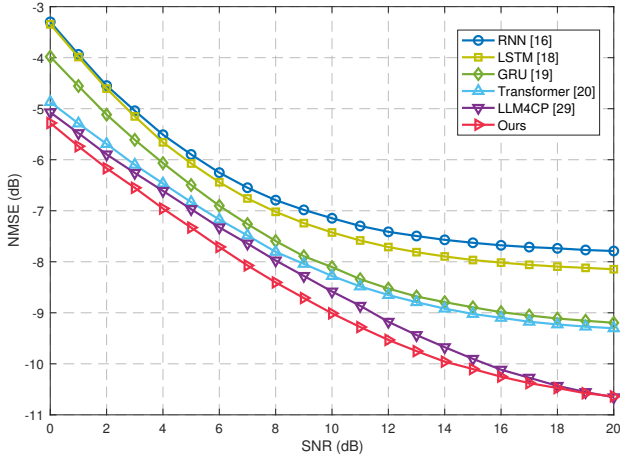


(a) UMa NLOS Scenario

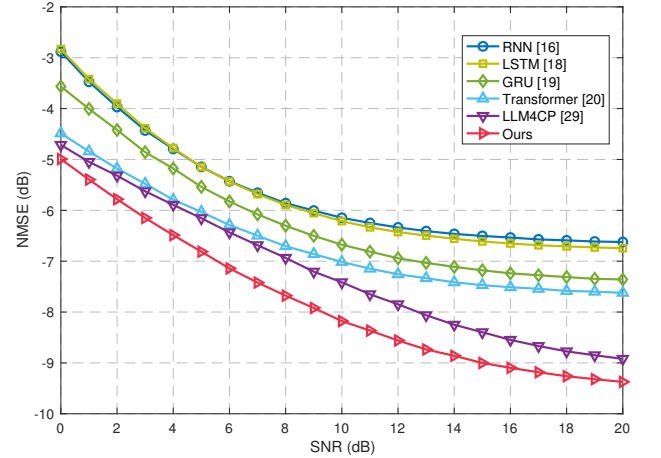


(b) UMi NLOS Scenario

Fig. 2: NMSE performance versus SNR between UMa and UMi NLOS scenarios with the UE velocity = 0 km/h.

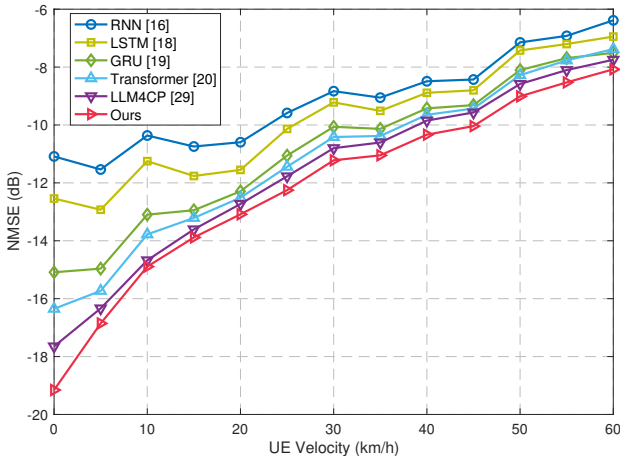


(a) UMa NLOS Scenario

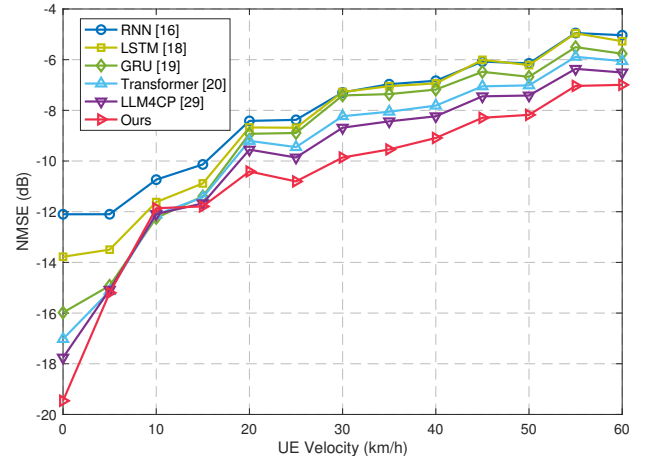


(b) UMi NLOS Scenario

Fig. 3: NMSE performance versus SNR between UMa and UMi NLOS scenarios with the UE velocity = 50 km/h.



(a) UMa NLOS Scenario



(b) UMi NLOS Scenario

Fig. 4: NMSE performance versus UE velocity between UMa and UMi NLOS scenarios with SNR = 10 dB.

these results for both UMa and UMi NLOS scenarios. A clear trend observed in both scenarios is that the performance generally degrades (NMSE increases) for all methods as the UE velocity rises. This is expected, as higher velocities induce faster channel variations, making accurate prediction more challenging. The rate of performance degradation tends to be steeper at lower velocities and somewhat flattens as the velocity approaches 60 km/h.

In the UMa NLOS scenario (Fig. 4a), which aligns with the training conditions, our proposed method consistently demonstrates superior performance, achieving the lowest NMSE across the entire velocity spectrum. Specifically, at 0 km/h, our method achieves an NMSE of approximately  $-19.2$  dB, outperforming the LLM4CP baseline (around  $-17.6$  dB) by about  $-1.6$  dB. As velocity increases, while the absolute performance of all methods degrades, our framework still maintains its advantage. For instance, at the UE velocities of 30 km/h and 60 km/h, our method leads LLM4CP by approximately  $-0.4$  dB and  $-0.3$  dB, respectively.

Turning to the generalization performance in the unseen UMi NLOS scenario (Fig. 4b), our proposed framework again maintains its clear lead over all baselines across all tested velocities. Similar to the observations in the SNR analysis, the performance gap between our method and LLM4CP is even more pronounced in this zero-shot scenario compared to the UMa case. At 0 km/h, the advantage over LLM4CP increases to about  $-1.7$  dB. This enhanced lead persists at higher velocities, with gaps of approximately  $-1.2$  dB at 30 km/h and  $-0.5$  dB at 60 km/h. These results again underscore the robustness and superior generalization capability of our proposed method across varying mobility conditions.

### 3) NMSE Performance per Prediction Step Comparison:

In this part, we delve deeper into the prediction performance by examining the NMSE specifically at each prediction step within the total prediction horizon  $P = 6$ . Fig. 5 presents this step-wise NMSE evaluation under challenging conditions (UE velocity = 50 km/h, SNR = 10 dB) for both UMa and UMi scenarios. In this figure, the prediction step ranges from the first step ( $p = 1$ ) to the last step ( $p = 6$ ). As expected, NMSE performance degrades with the prediction step for all methods in both scenarios, as predicting further into the future inherently involves a larger uncertainty and error accumulation.

In the UMa NLOS scenario (Fig. 5a), our proposed method consistently achieves the lowest NMSE at every prediction step. For the first prediction step with  $p = 1$ , our method yields an NMSE of approximately  $-12.7$  dB, surpassing LLM4CP ( $\sim -12.1$  dB) by about  $-0.6$  dB and significantly lower than the other baselines. This performance advantage is sustained across the prediction horizon, with the gap relative to LLM4CP ranging between  $-0.6$  dB and  $-0.3$  dB for steps  $p = 1$  to  $p = 6$ .

Analyzing the generalization performance in the unseen UMi NLOS scenario (Fig. 5b), our method maintains its superiority at each prediction step. While all methods exhibit slightly higher NMSE compared to the UMa scenario due to the zero-shot evaluation context, our method's lead over the LLM4CP baseline is consistently maintained and generally

more pronounced than that in the UMa case. In particular, for the first prediction step ( $p = 1$ ), the NMSE advantage over LLM4CP is about  $-0.6$  dB. This performance gap widens for intermediate prediction steps, reaching a maximum advantage of approximately  $-0.9$  dB at steps  $p = 3$  and  $p = 4$ , before settling back to a  $-0.6$  dB advantage when  $p = 6$ . This sustained lead across the prediction horizon in both scenarios indicates the robustness and effectiveness of our proposed method for multi-step channel prediction.

### E. Ablation Studies

To validate the effectiveness and necessity of the key components within our proposed framework, we conduct several ablation studies comparing our full proposed method against two variants:

- **w/o GPT-2:** This variant removes the entire LLM component (pre-trained LLM backbone and its embedding), relying solely on the proposed adapter (i.e., CSI embedding and SCA module) and the output head. Its primary purpose is to establish the performance baseline achieved without the LLM component, thereby directly demonstrating the indispensable role of the LLM backbone.
- **GPT-2:** This variant removes the *adapter* and instead processes the normalized CSI data directly by GPT-2 with full fine-tuning (FFT). The motivation is to create a strong baseline by evaluating a fully fine-tuned GPT-2 adapting all parameters directly to the task, without our specialized adapter. Comparing our method against this challenging baseline underscores the critical necessity and contribution of our specialized adapter.

The performance comparisons under the UMa NLOS scenario are presented in Fig. 6 (vs SNR), Fig. 7 (vs UE velocity), and Fig. 8 (vs prediction step). In these figures, the full version of our proposed method is denoted as **Ours**.

First, the results clearly establish the importance of the LLM backbone. Comparing our proposed method with the “w/o GPT-2” variant reveals a significant performance degradation across all conditions (Figs. 6-8) when the LLM is removed. This gap is substantial and notably widens for predictions further into the future (Fig. 8). This demonstrates the crucial role of the powerful sequence modeling capabilities of the pre-trained LLM in achieving high channel prediction performance. However, while LLM is indispensable, it is not sufficient for optimal performance in the channel prediction task. This is evident when comparing our proposed method against the “GPT-2” variant. Our method consistently and significantly outperforms “GPT-2” across all scenarios, despite the latter employing FFT across all its parameters. This substantial performance advantage powerfully underscores the critical importance and effectiveness of our proposed adapter. It indicates the adapter can effectively preprocess and transform the CSI data into a representation that the LLM can readily comprehend and model, thereby unlocking its predictive potential within the proposed framework. In addition, it also indicates that simply applying an LLM directly to the channel prediction task, even with full fine-tuning, is insufficient for achieving the promising performance.

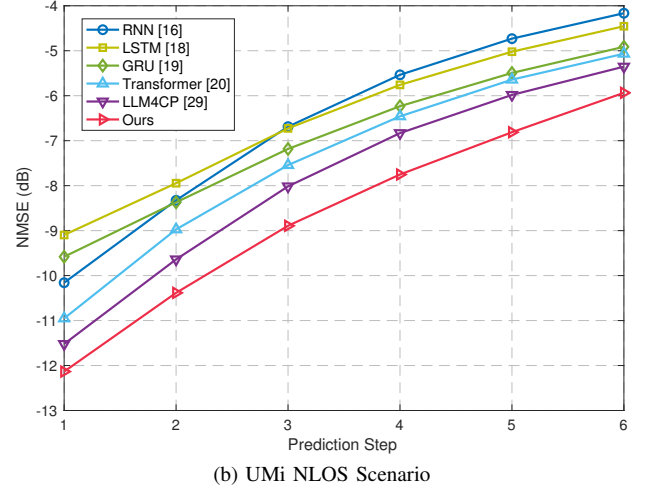
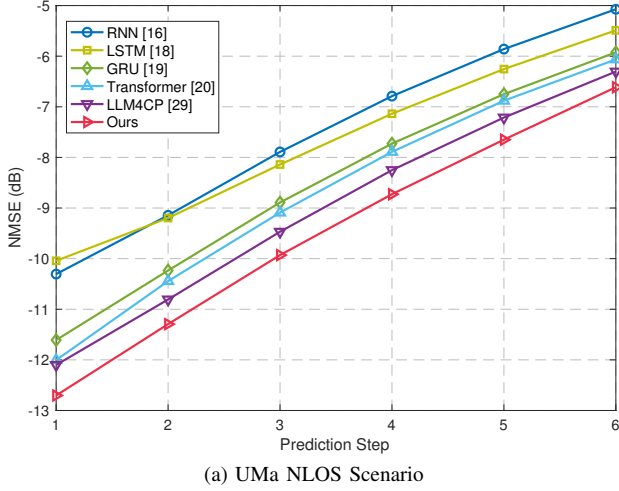


Fig. 5: NMSE performance versus prediction step  $p$  between UMa and UMi NLOS scenarios with UE velocity = 50 km/h and SNR = 10 dB. The total prediction horizon is  $P = 6$ .

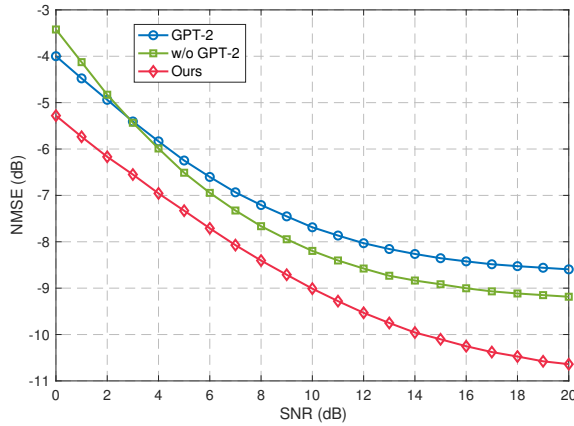


Fig. 6: NMSE performance versus SNR in the UMa NLOS scenario with UE velocity = 50 km/h.

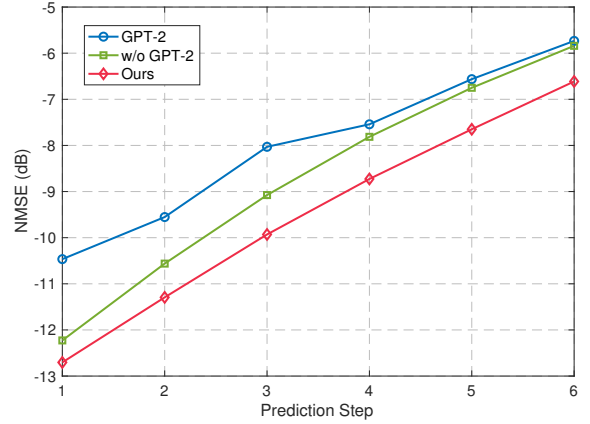


Fig. 8: NMSE performance versus prediction step  $p$  in the UMa NLOS scenario with UE velocity = 50 km/h and SNR = 10 dB. The total prediction horizon is  $P = 6$ .

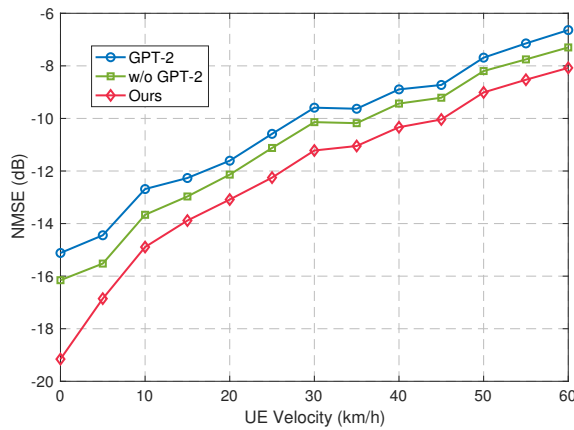


Fig. 7: NMSE performance versus UE velocity in the UMa NLOS scenario with SNR = 10 dB.

This insight is further emphasized by the comparison between the two ablated versions. Strikingly, the “w/o GPT-2” variant consistently outperforms the “GPT-2” variant across nearly all conditions (Figs. 6-8, except potentially at extremely low SNR). This compelling result strongly suggests that naive application of even a powerful, fully fine-tuned LLM yields mediocre performance, as the model struggles without domain-specific guidance. In contrast, it highlights the remarkable effectiveness of our proposed adapter in extracting useful information from CSI data, proving its value even in isolation.

Most importantly, our full method significantly outperforms the separated components across all tested SNRs, UE velocities, and prediction steps. This superiority highlights the effectiveness and superiority of the overall SCA-LLM framework. Within this framework, the LLM is adapted to achieve the SOTA performance compared to the baselines. This results compellingly validates that the proposed adapter can bridge the gap between the CSI data and the LLM more

effectively.

## VI. CONCLUSIONS

This paper proposed the SCA-LLM framework for accurate and robust channel prediction in MIMO-OFDM systems. The framework effectively adapted pre-trained LLMs to CSI data by mitigating domain mismatch through a novel spectral-attentive adapter. In this adapter, the SCA module employed multi-spectral channel attention with the 2-D DCT to capture a broader range of spectral components from CSI features compared to methods relying on global pooling operations. This, in turn, provided a more suitable input for the LLM and significantly improved its adaptation and prediction capabilities. Extensive simulations demonstrated that SCA-LLM achieved superior prediction performance, consistently outperformed existing methods, and yielded up to a notable NMSE advantage of  $-2.4$  dB over the prior LLM-based approach. Ablation studies further validated the critical contributions of both the spectral-attentive adapter and the LLM backbone. These findings underscored that the SCA-LLM framework strengthened the “adapter + LLM” paradigm, enabled better exploitation of LLMs for channel prediction, and suggested potential for broader LLM applications in wireless communications.

## APPENDIX A

### MULTI-SPECTRAL CHANNEL ATTENTION LAYER

The multi-spectral channel attention layer implements a variant of the channel attention mechanism, a technique widely used in deep learning to adaptively recalibrate feature responses [30], [31], [36]. It is crucial to note the distinction in terminology: in the context of this attention mechanism, the term *channel* refers to the *channel* dimension within the data tensor (e.g., the dimension  $L_c$  in  $\mathbf{H}_c^{(i)} \in \mathbb{R}^{B \times L_c \times N \times E_c}$ ), and should be distinguished from the physical wireless communication channel. Conversely, the dimensions  $(N, E_c)$  correspond to the spatial dimensions or the feature map upon which the subsequent analysis operates.

To incorporate richer information into the channel attention mechanism, this layer leverages the 2-D DCT [32] for feature analysis and modeling. For the purpose of DCT analysis, each feature map is considered to have dimensions  $H' \times W'$ . These dimensions might correspond directly to the input spatial dimensions  $(N, E_c)$ , or they might be target dimensions  $(H_{dct}, W_{dct})$  achieved through an optional pooling step described later. The 2-D DCT decomposes such a feature map into a sum of cosine functions oscillating at different frequencies. These basis functions form a complete orthogonal basis set, denoted as  $\mathcal{B}_{\text{DCT}} = \{\mathbf{F}_{u,v} | 0 \leq u < H', 0 \leq v < W'\}$ . The standard basis function  $\mathbf{F}_{u,v} \in \mathcal{B}_{\text{DCT}}$  corresponding to the frequency indices  $(u, v)$  for this  $H' \times W'$  map is defined as:

$$\mathbf{F}_{u,v}[h, w] = \cos\left(\frac{\pi u}{H'}\left(h + \frac{1}{2}\right)\right) \cos\left(\frac{\pi v}{W'}\left(w + \frac{1}{2}\right)\right), \quad (\text{A.1})$$

where  $(h, w)$  are the spatial indices ( $0 \leq h < H', 0 \leq w < W'$ ) and  $(u, v)$  are the frequency indices identifying each basis function within the set  $\mathcal{B}_{\text{DCT}}$ .

The projection of a generic  $H' \times W'$  feature map, let's denote it as  $\mathbf{Q} \in \mathbb{R}^{H' \times W'}$  (representing one such map from the input tensor, potentially after pooling), onto a specific frequency component  $(u, v)$  is achieved by calculating the corresponding DCT coefficient using the basis function  $\mathbf{F}_{u,v}$ . This projection is calculated as:

$$2\text{DDCT}^{u,v}(\mathbf{Q}) = \sum_{h=0}^{H'-1} \sum_{w=0}^{W'-1} \mathbf{Q}[h, w] \cdot \mathbf{F}_{u,v}[h, w], \quad (\text{A.2})$$

where  $\mathbf{Q}[h, w]$  denotes the element of  $\mathbf{Q}$  at spatial index  $(h, w)$ . Given the input tensor for the  $i$ -th layer,  $\mathbf{H}_c^{(i)} \in \mathbb{R}^{B \times L_c \times N \times E_c}$ , where  $L_c$  is the previously defined channel dimension, the layer performs the following steps:

#### A. Adaptive Pooling (Optional)

If the input spatial dimensions  $(N, E_c)$  differ from the target DCT analysis dimensions  $(H_{dct}, W_{dct})$ , adaptive average pooling is applied independently to each  $L_c$  channel map to obtain  $\mathbf{H}_{c,\text{pooled}}^{(i)} \in \mathbb{R}^{B \times L_c \times H_{dct} \times W_{dct}}$ . After that, the DCT analysis dimensions would be  $H' = H_{dct}, W' = W_{dct}$ .

#### B. Channel Splitting and DCT Projection

The  $L_c$  input channels are divided into  $n$  groups, with  $L'_c = L_c/n$  channels per group. A specific 2-D DCT frequency index  $(u_k, v_k)$ , corresponding to a basis function  $\mathbf{F}_{u_k, v_k} \in \mathcal{B}_{\text{DCT}}$ , is assigned to each group  $k \in \{0, \dots, n-1\}$ . These indices identify a pre-selected subset of  $n$  DCT basis functions, often chosen based on empirical findings for optimal performance as suggested in [31]. It is worth mentioning that using these multiple pre-defined DCT basis functions adds negligible computational cost and no extra trainable parameters compared to using only the lowest frequency component via GAP [31].

Let  $\mathbf{H}_{c,k}^{(i)} \in \mathbb{R}^{B \times L'_c \times H' \times W'}$  represent the feature maps for the  $k$ -th group (after optional pooling). The spectral features for this group are computed by applying the assigned 2-D DCT projection (Eq. (A.2)) using the basis function  $\mathbf{F}_{u_k, v_k}$  to the  $L'_c$  feature maps within the group  $k$  independently:

$$\mathbf{Freq}^k = 2\text{DDCT}^{u_k, v_k}(\mathbf{H}_{c,k}^{(i)}) \quad (\text{A.3})$$

This calculation results in a tensor  $\mathbf{Freq}^k \in \mathbb{R}^{B \times L'_c}$ , where each element represents the magnitude of the  $(u_k, v_k)$  frequency component for a specific channel map within group  $k$ .

#### C. Multi-Spectral Vector Concatenation

The frequency components calculated for all  $n$  distinct groups are concatenated along the channel dimension, denoted as  $\text{Concat}(\cdot)$ . This aggregation forms a comprehensive multi-spectral descriptor  $\mathbf{Y}_{\text{dct}}^{(i)}$  for each sample in the batch:

$$\mathbf{Y}_{\text{dct}}^{(i)} = \text{Concat}([\mathbf{Freq}^0, \mathbf{Freq}^1, \dots, \mathbf{Freq}^{n-1}]) \in \mathbb{R}^{B \times L_c}. \quad (\text{A.4})$$

These frequency components can help capture certain wireless channel characteristics.

### D. Channel Attention Weight Generation

This richer multi-spectral descriptor  $\mathbf{Y}_{\text{det}}^{(i)}$  is then fed into an attention weight generation block. This block typically consists of two fully connected (FC) layers forming a bottleneck structure (with a reduction factor  $R$  and a *rectified linear unit* (ReLU) activation after the first FC layer), followed by a final *sigmoid* activation  $\sigma(\cdot)$ . This process yields the channel attention weights:

$$\mathbf{W}_{\text{att}}^{(i)} = \sigma(\text{FC}_2(\text{ReLU}(\text{FC}_1(\mathbf{Y}_{\text{det}}^{(i)})))) \in \mathbb{R}^{B \times L_c}. \quad (\text{A.5})$$

### E. Feature Rescaling

Finally, the original input tensor  $\mathbf{H}_c^{(i)}$  (before any optional pooling) is recalibrated by element-wise multiplication ( $\odot$ ) with the generated attention weights  $\mathbf{W}_{\text{att}}^{(i)}$ , which can be expressed as:

$$\mathbf{H}_c^{(i+1)} = \mathbf{H}_c^{(i)} \odot \mathbf{W}_{\text{att}}^{(i)}. \quad (\text{A.6})$$

The result  $\mathbf{H}_c^{(i+1)} \in \mathbb{R}^{B \times L_c \times N \times E_c}$  is the output of this attention layer, ready for the next layer or the output projection step.

### REFERENCES

- [1] G. L. Stüber, J. R. Barry, S. W. McLaughlin, Y. G. Li, M. A. Ingram, and T. G. Pratt, "Broadband MIMO-OFDM wireless communications," *Proc. IEEE*, vol. 92, no. 2, pp. 271–294, 2004.
- [2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. E. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5g be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.
- [3] C. Wang, X. You, X. Gao, X. Zhu, Z. Li *et al.*, "On the Road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surv. Tutorials*, vol. 25, no. 2, pp. 905–974, 2023.
- [4] E. Biglieri, A. R. Calderbank, A. G. Constantinides, A. Goldsmith, and A. Paulraj, *MIMO Wireless Communications*. Cambridge University Press, 2010.
- [5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, 2014.
- [6] H. Bölcskei, D. Gesbert, and A. Paulraj, "On the capacity of ofdm-based spatial multiplexing systems," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 225–234, 2002.
- [7] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," *IEEE Commun. Surv. Tutorials*, vol. 23, no. 2, pp. 668–695, 2021.
- [8] M. A. M. Albreem, A. H. A. Habbash, A. M. Abu-Hudrouss, and S. S. Ikki, "Overview of precoding techniques for massive MIMO," *IEEE Access*, vol. 9, pp. 60 764–60 801, 2021.
- [9] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim *et al.*, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 4, pp. 3060–3097, 2018.
- [10] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen *et al.*, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, 2017.
- [11] J. Zheng, J. Zhang, E. Björnson, and B. Ai, "Impact of channel aging on cell-free massive MIMO over spatially correlated channels," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 10, pp. 6451–6466, 2021.
- [12] A. K. Kocharlakota, K. Upadhyay, and S. A. Vorobyov, "Impact of pilot overhead and channel estimation on the performance of massive MIMO," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8242–8255, 2021.
- [13] H. Yin, H. Wang, Y. Liu, and D. Gesbert, "Addressing the curse of mobility in massive MIMO with prony-based angular-delay domain channel predictions," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2903–2917, 2020.
- [14] C. Wu, X. Yi, Y. Zhu, W. Wang, L. You, and X. Gao, "Channel prediction in high-mobility massive MIMO: from spatio-temporal autoregression to deep learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1915–1930, 2021.
- [15] H. Kim, S. Kim, H. Lee, C. Jang, Y. Choi *et al.*, "Massive MIMO channel prediction: Kalman filtering vs. machine learning," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 518–528, 2021.
- [16] W. Jiang and H. D. Schotten, "Neural network-based fading channel prediction: A comprehensive overview," *IEEE Access*, vol. 7, pp. 118 112–118 124, 2019.
- [17] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proc. IEEE*, vol. 95, no. 12, pp. 2299–2313, 2007.
- [18] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 320–332, 2020.
- [19] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 2014, pp. 1724–1734.
- [20] 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, no. 9, pp. 2717–2732, 2022.
- [21] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu *et al.*, "Qwen2 technical report," *CoRR*, vol. abs/2407.10671, 2024.
- [22] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle *et al.*, "The llama 3 herd of models," *CoRR*, vol. abs/2407.21783, 2024.
- [23] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song *et al.*, "DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning," *CoRR*, vol. abs/2501.12948, 2025.
- [24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [25] J. Shao, J. Tong, Q. Wu, W. Guo, Z. Li *et al.*, "Wirelessllm: Empowering large language models towards wireless intelligence," *J. Commun. Inf. Networks*, vol. 9, no. 2, pp. 99–112, 2024.
- [26] H. Zhou, C. Hu, Y. Yuan, Y. Cui, Y. Jin *et al.*, "Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities," *CoRR*, vol. abs/2405.10825, 2024.
- [27] T. Yang, P. Zhang, M. Zheng, Y. Shi, L. Jing *et al.*, "WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication," *CoRR*, vol. abs/2502.06877, 2025.
- [28] Y. Chen, R. Li, Z. Zhao, C. Peng, J. Wu *et al.*, "NetGPT: An AI-native network architecture for provisioning beyond personalized generative services," *IEEE Netw.*, vol. 38, no. 6, pp. 404–413, 2024.
- [29] B. Liu, X. Liu, S. Gao, X. Cheng, and L. Yang, "LLM4CP: adapting large language models for channel prediction," *J. Commun. Inf. Networks*, vol. 9, no. 2, pp. 113–125, 2024.
- [30] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020.
- [31] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 763–772.
- [32] A. Shawahna, M. E. Haque, and A. Amin, "JPEG image compression using the discrete cosine transform: An overview, applications, and hardware implementation," *CoRR*, vol. abs/1912.10789, 2019.
- [33] 3GPP Radio Access Network Working Group, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 15)," 3GPP, Technical Report TR 38.901, 2018.
- [34] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surv. Tutorials*, vol. 15, no. 3, pp. 996–1019, 2013.
- [35] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, 2014.
- [36] M. Guo, T. Xu, J. Liu, Z. Liu, P. Jiang *et al.*, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.