Spatial-Temporal Channel Prediction via a Dual-Domain Augmented VLM-based Framework

Anonymous ICME submission

***Abstract*—Channel state information (CSI) prediction is a crucial technology in future sixth-generation (6G) wireless communication systems. Although existing deep learning-based methods have achieved notable improvements in CSI prediction, they primarily focus on temporal or spectral feature extraction while neglecting spatial domain characteristics. This inherent limitation results in suboptimal performance when deployed in massive multi-input multi-output environments, where spatial correlation plays a critical role. Meanwhile, recent advances in vision-language models (VLMs) have enabled the extracting of complex spatial-temporal dependencies through multimodal representation learning. However, this paradigm remains unexplored in CSI prediction tasks. In this paper, we propose a novel dual-domain augmented VLM-based framework that explicitly integrates spatial-temporal coherence and spectral correlation for enhanced CSI estimation. Specifically, the architecture comprises two modules: a spatial-structural guidance block to extract geometric features from CSI data and a temporal-coherence guidance block to capture temporal cohenrence and physical consistency. Experimental results demonstrate that our proposed model outperforms the compared baseline methods in term of normalized mean squared error (NMSE) and squared generalized cosine similarity (SGCS), verifying its capability in capturing multi-dimensional channel structures.**

***Index Terms*—Channel prediction, vision-language models, channel state information, multimodal learning, spatial-temporal modeling**

I. Introduction

The accurate acquisition of channel state information (CSI) is crucial in communication systems, providing additional spatial gain for diversified wireless applications [1]. As shown in Figure 1, orthogonal pilot symbols embedded in the signal transmission frame are used as known reference signals for CSI estimation to obtain the unknown data. Multiple-input multiple-output (MIMO) technology leverages the spatial dimension of large antenna arrays to achieve significant multiplexing and diversity gains [2]. However, with the the evolution towards massive MIMO arrays, the excessive pilot overhead poses a serve challenge to accurate CSI acquisition. This is because the required pilot resources scale with the explosive growth in the number of antennas. To resolve this, channel prediction has become a key technology to reduce the overhead and enhance system performance. By inferring future CSI from historical data, it enables the system to obtain channel information without dedicated pilot symbols.

Fueled by the rapid advancements in deep learning, data-driven methods have achieved substantial performance gains in wireless communication systems [3], [4], particularly in CSI prediction tasks. For instance, recurrent neural networks (RNNs) and long short-term memory (LSTM) have demonstrated powerful performance in channel prediction by effectively capturing dynamic temporal dependencies [5], [6]. Studies on temporal feature modeling of CSI have been further advanced by Transformers [7] and large language model (LLM)-based approaches [8], which excel at modeling long-range sequential correlations.

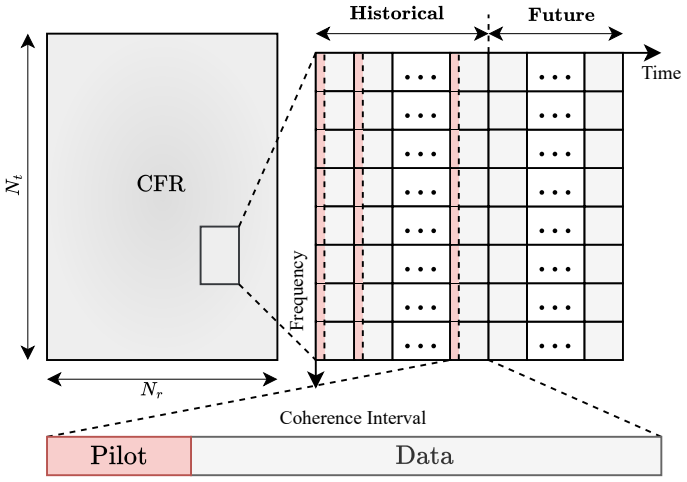

Fig. 1. The data structure in MIMO-OFDM system.

Beyond temporal dependence, parallel efforts have focused on the spatial dependencies. Complex-valued convolution neural networks have been specifically designed to preserve the phase-amplitude relationships of complex-valued CSI [9]. Researchers in [10] demonstrated that the superior spatial feature extraction capabilities of STEM GNN can be transferred to wireless channel modeling. WiFo [11] applies masked autoencoders (MAE) [12] to reconstruction tasks, treating CSI matrices as visual images to extract deep spatial structural features.

Despite these achievements, prior approaches fundamentally treats the spatial domain and temporal-spectral domains in isolation, ignoring the intrinsic spatiotemporal correlation of wireless channels. Specifically, above temporal-focused models simply flatten multi-dimensional CSI tensors into vectors, a process that inevitably obliterates the underlying spatial structure [5]–[8]. Conversely, spatial-focused models, while preserving structural features, often struggle to maintain precision across long prediction windows due to limited temporal modeling capabilities [9]–[11]. To bridge this gap, we introduce a dual-domain augmented framework that leverages

the pre-trained vision encoder and decoder-only LLMs to enhance spatial-temporal channel prediction. Our contributions are summarized as follows:

- We propose VLM-based framework to enhance spatial-temporal channel prediction by leveraging the robust representational capabilities of pre-trained vision encoder and large language model.
- We employ a spatial-structural guidance block and a temporal-coherence guidance block to extract multi-dimensional features, which are then integrated as multimodal inputs to leverage the VLM's powerful cross-modal reasoning capabilities
- The proposed method is validated on several open-source datasets featuring multi-dimensional CSI. Extensive experiments demonstrate that it achieves superior performance in spatio-temporal channel prediction tasks, outperforming baselines in terms of normalized mean squared error (NMSE) and squared generalized cosine similarity (SGCS).

II. Related work

A. Channel Prediction

In the field of channel prediction, traditional algorithms like the autoregressive (AR) model [13] and the kalman filtering [14] are computationally efficient and straightforward to implement. Nevertheless, their accuracy relies heavily on linear assumptions and requires stationary environments.

Related efforts in deep learning for channel prediction have followed two major research axes: temporal sequence modeling and spatial feature extraction. RNN and LSTM excel in channel prediction by capturing dynamic temporal features [5], [6]. Additionally, Transformer-based models significantly advance channel prediction in scenarios with long prediction windows, by facilitating parallel computation [7]. More recently, inspired by the advanced sequence-to-sequence modeling expertise of large language models in fields of natural language processing (NLP), several studies reflect their potential in CSI tasks. For instance, LLM4CP [8] fine-tuned a pre-trained GPT-2 [15] for CSI data and deployed a set of modules to boost model effectiveness. Similarly, the authors in [16] leveraged the powerful noise removal capability of LLM to improve CSI reconstruction performance.

In parallel, CNN-based approaches employ convolutional operations to extract the spatial structure of CSI data [17], [18]. Furthermore, CVCNN [9] utilizes complex-valued layers to effectively preserve the phase-amplitude relationships of CSI, thus capturing richer spatial features than conventional real-valued networks. Similarly, STEM GNN [10] utilizes its specialized graph convolutional architecture to jointly capture latent spatial correlations. More recently, WiFo [11] utilizes a self-supervised learning paradigm by applying the masked autoencoder (MAE) [12] to reconstruct CSI images, thereby effectively capturing intricate spatial structural features for channel tasks. However, current architectures typically treat temporal dependence and spatial structures as independent entities, thereby overlooking the complex spatio-temporal correlations and the multi-dimensional synergy of the CSI.

B. Vision-Language Models

Vision-language models (VLMs) advance multimodal learning by bridging the gap between visual and textual modalities CLIP [19] and ALIGN [20] demonstrate that contrastive learning effectively aligns image and text embeddings in a shared latent space. BLIP-2 [21] incorporate Query Transformer (Q-Former) with cross-attention mechanisms to improve cross-modal interaction. In contrast, LLaVA [22] employs a lightweight interface by feeding visual features into the LLM via a single learnable linear projection layer.

Beyond the field of computer vision (CV), authors in [23] have extended the application of VLMs to non-visual domains, transforming structured data into visual representations. Therefore VLMs are not limited to native images and can serve as universal spatial-temporal learners through cross-modal representation learning.

III. Method

A. Problem Formulation

As shown in Fig. 1, we consider a MIMO-OFDM system with N_t transmit and N_r receive antennas operating over N_c subcarriers. At time step t and subcarrier frequency $k \in \{1, \ldots, N_c\}$, the received signal $\mathbf{y}_{t,k} \in \mathbb{C}^{N_r}$ is modeled as:

$$\mathbf{y}_{t,k} = \mathbf{H}_{t,k}\mathbf{x}_{t,k} + \mathbf{n}_{t,k}, \tag{1}$$

where $\mathbf{x}_{t,k} \in \mathbb{C}^{N_t}$ denotes the transmitted vector, and $\mathbf{n}_{t,k} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ represents the additive complex Gaussian noise. The term $\mathbf{H}_{t,k} \in \mathbb{C}^{N_r \times N_t}$ is the channel frequency response (CFR) matrix, the frequency-domain expression of CSI. Aggregating the channel matrices over all subcarriers, we represent the CSI snapshot at time t as $\mathcal{H}_t \triangleq \{\mathbf{H}_{t,1}, \ldots, \mathbf{H}_{t,N_c}\}$. Standard schemes insert dense pilots in each coherence interval to estimate $\mathcal{H}_t$, which incurs substantial overhead and reduces spectral efficiency in large-scale or fast-varying channels. To mitigate this burden, we formulate channel prediction as a time-series forecasting task, where a model with parameters $\mathbf{\Theta}$ maps P historical CSI snapshots to L future ones:

$$\big(\widetilde{\mathcal{H}}_{t+1}, \ldots, \widetilde{\mathcal{H}}_{t+L}\big) = f_{\mathbf{\Theta}}\big(\mathcal{H}_{t-P+1}, \ldots, \mathcal{H}_t\big), \tag{2}$$

where $\widetilde{\mathcal{H}}_{t+\ell}$ denotes the predicted CSI at time index $t+\ell$, $\ell = 1, \ldots, L$.

B. Overall Architecture

The overall architecture of our proposed model has been illustrated in Fig. 2, employing a VLM-based framework with frozen vision encoder and frozen decoder-only LLM. In order to enhance the extraction of spatial-temporal CSI features by the VLM, we introduce a dual-domain guidance mechanism: the spatial-structural guidance block and the temporal-coherence guidance block. The following sections provide a detailed description of each block.

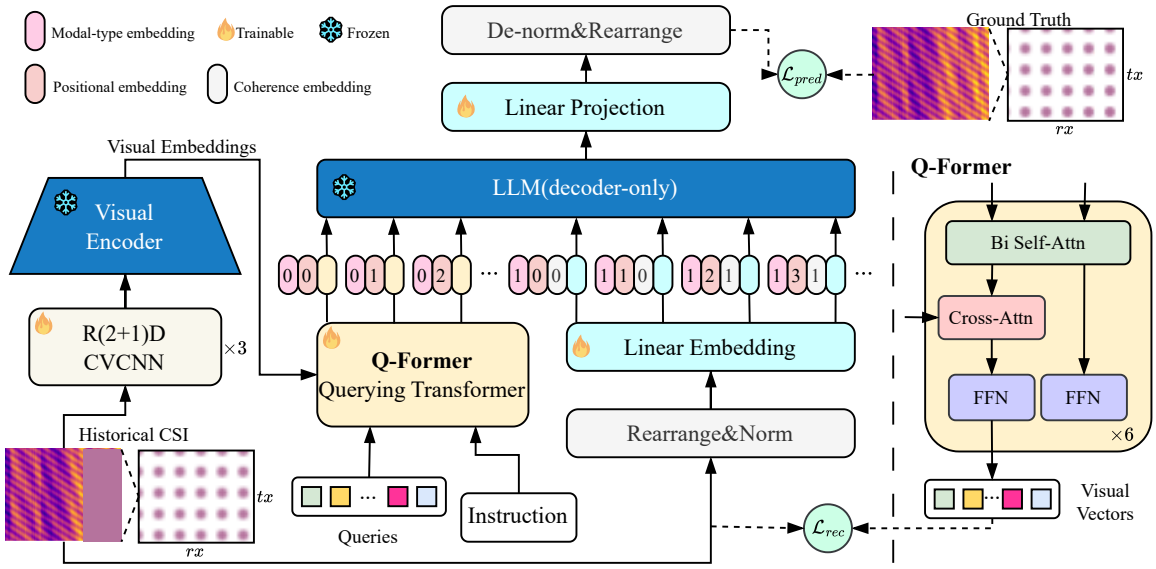

Fig. 2. Overview of our proposed method.

1) Spatial-Structural Guidance: The objective of this block is to extract geometric features of CSI data and extract visual features from frozen vision encoder.

Following research [24] proving early convolutions improve visual transformer (ViT) [25] optimization, we first process the historical input $\mathcal{X}_{\text{his}} = \{\mathcal{H}_{t-P+1}, \ldots, \mathcal{H}_t\} \in \mathbb{C}^{P \times N_c \times N_r \times N_t}$ through R(2+1)DCVCNN moudle. In this moudle, we utilize the residual complex-valued convolutional operations instead of standard CNN to explicitly preserve phase information, and then we factorize the standard 3D convolutional kernel size $3 \times 3 \times 3$ into $3 \times 3 \times 1$ and $1 \times 1 \times 3$ like study [26] to effectively capture the spatial correlations of $\mathcal{X}_{\text{his}}$ between the transmit antennas and receive antennas and the temporal dynamics in the time dimension. The operations abave are executed as follows:

$$\hat{\mathcal{X}} = \text{CVCNN}_{1D}(\text{CVCNN}_{2D}(\mathcal{X}_{\text{his}}, \mathbf{W}_S), \mathbf{W}_T) + \mathcal{X}_{\text{his}}, \quad (3)$$

where $\text{CVCNN}_{2D}(\cdot, \mathbf{W}_S)$ and $\text{CVCNN}_{1D}(\cdot, \mathbf{W}_T)$ denotes the complex-valued convolution with 2D kernel weights $\mathbf{W}_S$ and 1D kernel weights $\mathbf{W}_T$, respectively. Following this, the phase-amplitude representation $\hat{\mathcal{X}}$ is passed to the pre-trained ViT-based vision encoder, from where output visual embeddings $\mathbf{F}_{ve} \in \mathbb{R}^{N_v \times D_v}$.

Unlike BLIP-2 [21], we train querying transformer (Q-Former) from scratch. A set of N_q learnable query embeddings $\mathbf{Q} \in \mathbb{R}^{N_q \times D_q}$ is initialized randomly and optimized during training. The instruction prompt $\mathbf{I}$ is designed to guide the Q-Former to extract visual features relevant to CSI prediction tasks. The visual embeddings $\mathbf{F}_{ve}$ is then align to the dimension of Q-Former $\hat{\mathbf{F}}_{ve} \in \mathbb{R}^{N_v \times D_q}$ after a linear layer. Through 6 layers of Q-Former, the earnable query embeddings interact with the instruction prompt $\mathbf{I}$ through bidirectional self-attention layers, and interact with visual features $\hat{\mathbf{F}}_{ve}$ to compress spatial information into instruction-aware visual representations through cross-attention layers. and then the fully connected feed-forward network (FFN) [27], consists of two linear transformations with a ReLU activation in between, is employed. To this end, the operations within each l-th layers of Q-Former are executed as follows:

$$\tilde{\mathbf{Q}}^{(l)} = \mathbf{Q}^{(l-1)} + \text{Attention}\left(\text{LN}([\mathbf{Q}^{(l-1)}, \mathbf{I}])\right), \quad (4)$$

$$\hat{\mathbf{Q}}^{(l)} = \tilde{\mathbf{Q}}^{(l)} + \text{Attention}\left(\text{LN}(\tilde{\mathbf{Q}}^{(l)}), \hat{\mathbf{F}}_{ve}, \hat{\mathbf{F}}_{ve}\right), \quad (5)$$

$$\mathbf{Q}^{(l)} = \hat{\mathbf{Q}}^{(l)} + \max\left(0, \text{LN}(\hat{\mathbf{Q}}^{(l)})W_1 + b_1\right)W_2 + b_2, \quad (6)$$

where $\text{LN}(\cdot)$ denotes layer normalization [27], $[\cdot, \cdot]$ denotes the concatenation operator along the sequence dimension, $\mathbf{Q}^{(l)}$ represents the output of the l-th layer. The fundamental operation of this layers is the Scaled Dot-Product Attention [27]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \quad (7)$$

Thus we yield the Q-Former output $\mathbf{Z}_q \in \mathbb{R}^{N_q \times D_q} = \mathbf{Q}^{(6)}$.

2) Temporal-Coherence Guidance: The objective of this block is to calculate coherence embedding of CSI data and feed multi-input into frozen LLM.

We introduce the coherence embedding as the primary mechanism of this block, motivated by the theoretical channel correlation properties [28] and study [28]. The calculation method for coherence segmentation $\mathbf{S} \in \mathbb{Z}^P$ can be summarized by algorithm 1. We calculate the coherence embedding $\mathbf{E}_{coh} \in \mathbb{R}^{P \times D_l}$ by mapping the coherence segmentation $\mathbf{S}$ through a learnable embedding layer $\boldsymbol{Embedding}(\cdot)$, where D_l denotes the dimension of LLM.

We first reshape original historical input $\mathcal{X}_{his} \in \mathbb{C}^{P \times N_c \times N_r \times N_t}$ into a dense real-valued representation $\mathcal{X}_{rearrange} \in \mathbb{R}^{P \times (2N_c N_r N_t)}$, where the real and imaginary parts of each complex entry are decomposed and concatenated,

Algorithm 1 Coherence Segmentation

Input: Historical CSI sequence $\mathbf{X} \in \mathbb{C}^{P \times D}$, threshold η
Output: Coherence Segment Array $\mathbf{S} \in \mathbb{Z}^{P}$
$current\_id \leftarrow 0$
$\mathbf{S}[1] \leftarrow current\_id$
for $t = 2$ **to** T **do**
  $\boldsymbol{\delta} \leftarrow \|\mathbf{X}_t - \mathbf{X}_{t-1}\|_2$
  $ratio \leftarrow \dfrac{\boldsymbol{\delta}}{\|\mathbf{X}_{t-1}\|_2 + \epsilon}$
  if $ratio < \eta$ **then**
    $\mathbf{S}[t] \leftarrow current\_id$
  else
    $\mathbf{S}[t] \leftarrow current\_id + 1$
    $current\_id \leftarrow current\_id + 1$
  end if
end for
return $\mathbf{S}$

and then we apply instance normalization [29] to standardize each sample to zero mean and unit variance.

Thus, the composite embedding input $\mathbf{H}_{in}$ for frozen LLM can be represented as:

$$\mathbf{H}_{in} = [\mathbf{Z}_q, \mathcal{X}_{rearrange} + \mathbf{E}_{coh}] + \mathbf{E}_{pos} + \mathbf{E}_{mod}, \quad (8)$$

where $\mathbf{E}_{coh}$ represents our proposed coherence embedding, $[\cdot, \cdot]$ denotes the concatenation operator along the sequence dimension, and $\mathbf{E}_{mod}$ and $\mathbf{E}_{pos}$ denote standard modal-type and positional embeddings in research [30]. The LLM processes the composite embedding input $\mathbf{H}_{in}$ to generate output hidden states that encode future channel dynamics. To map these semantic representations back to the physical CSI space, we pass the output through a linear projection layer, followed by the denormalization and reshaping to yield the predicted output $\hat{\mathcal{Y}}_{pred} = \{\hat{\mathcal{H}}_{t+1}, \ldots, \hat{\mathcal{H}}_{t+L}\} \in \mathbb{C}^{L \times N_c \times N_r \times N_t}$.

C. Training

We form training samples by sliding a window [31] of length $P + L$ over the time dimension of training dataset $\mathcal{D}_{train}$, splitting each window $\mathcal{X} = \mathcal{H}_{t-P+1:t+L}$ into a historical segment $\mathcal{X}_{his} = \mathcal{H}_{t-P+1:t}$ and a target segment $\mathcal{X}_{gt} = \mathcal{H}_{t:t+L}$. With $\mathcal{X}_{\text{gt}}$ as supervision, we train the model to map $\mathcal{X}_{his}$ to the prediction $\hat{\mathcal{Y}}_{pred} = \mathcal{H}_{t:t+L}$.

The total loss of our model is formed by:

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda \mathcal{L}_{rec}. \quad (9)$$

The $\mathcal{L}_{pred}$ minimizes the normalized mean squared error (MSE) between the predicted and ground-truth CSI:

$$\mathcal{L}_{pred} = \frac{\|\hat{\mathcal{Y}}_{pred} - \mathcal{X}_{gt}\|_F^2}{\|\mathcal{X}_{gt}\|_F^2}, \quad (10)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Inspired by the concept [32], which employ discriminative classifiers for intermediate layers, we propose an auxiliary reconstruction objective $\mathcal{L}_{rec}$:

$$\mathcal{L}rec = \|\mathcal{R}(\mathbf{Z}_q \mathbf{W}_{rec} + \mathbf{b}_{rec}) - \mathcal{X}_{his}\|_F^2, \quad (11)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{Z}_q \in \mathbb{R}^{N_q \times D_q uantity}$ denotes the output of the Q-Former, and $\mathcal{R}(\cdot)$ denotes the reshaping operation that restores the spatial-temporal dimensions subsequent to the affine transformation of $\mathbf{Z}_q$ using the weight matrix $\mathbf{W}_{rec}$ and bias vector $\mathbf{b}_{rec}$. The hyperparameter λ balances the auxiliary loss $\mathcal{L}_{rec}$ with the primary loss $\mathcal{L}_{pred}$.

IV. Experiments

A. Experimental Setup

1) Datasets: We evaluated our model on the open mobile communication dataset[1], categorized into four subsets according to user velocity: 30km/h, 60km/h, 120km/h, and a mixture of samples from the aforementioned three speed levels. For each subset, we collect 21,000 samples structured as time-frequency grids across 32 transmit and 4 receive antennas. Each sample encompasses 20 time steps with a transmission time interval (TTI) of 5 ms and spans 8 physical resource blocks (PRBs) in the frequency domain. Specific simulation parameters are listed in Table I.

TABLE I
Parameters for Dataset

Parameters	Value
Scenario	Dense Urban (Macro only)
Channel Model	According to TR 38.901
Inter-BS Distance	200m
Frequency Range	FR1 only; 2GHz
Subcarrier Spacing	15kHz for 2GHz
Bandwidth	10M (52RB)
Speed	30/60/120/Mix. km/h
Data Size	(21000, 20, 2, 32, 4, 8)

2) Baseline: To evaluate the performance of our proposed model, we compare it with existing methods, such as CVCNN [9], LSTM [6], STEM GNN [10], and LLM4CP [8]. To ensure a fair comparison across all baselines, we adopt a unified experimental framework.

3) Evaluation Metrics: We employ normalized mean squared error(NMSE) and squared generalized cosine similarity(SGCS), which are standard metrics for channel prediction to quantifies the numerical and spatial discrepancy between the predicted channel states and the ground truth, respectively [33]. NMSE serves as the optimization objective in our framework, whose formulation is detailed in Eq. 10. SGCS is defined as:

$$\text{SGCS} = \frac{1}{N_{sp}} \sum_{i=1}^{N_{sp}} \frac{1}{N_{rb}} \sum_{j=1}^{N_{rb}} \frac{\|\mathbf{H}_{i,j} \hat{\mathbf{H}}_{i,j}\|^2}{\|\mathbf{H}_{i,j}\|^2 \|\hat{\mathbf{H}}_{i,j}\|^2}, \quad (12)$$

where N_{sp} represents the number of samples, N_{rb} is the number of resource blocks per sample, $\mathbf{H}_{i,j} \in \mathcal{C}^{N_r \times N_t}$ and $\hat{\mathbf{H}}_{i,j} \in \mathcal{C}^{N_r \times N_t}$ are the predicted channel matrices and ground truth, respectively.

[1] www.mobileai-dataset.com

TABLE II

THE SIZE OF OBSERVATION WINDOW IS SET AS 12 AND PREDICTION WINDOW SIZE $l_o \in \{2, 4, 8\}$. FOR NMSE, LOWER VALUES INDICATE BETTER PERFORMANCE, FOR SGCS, HIGHER VALUES INDICATE BETTER PERFORMANCE. BOLD: BEST, UNDERLINE: SECOND BEST

Models		**Ours**		**LSTM**		**CVCNN**		**STEM GNN**		**LLM4CP**	
Metric		**NMSE ↓**	**SGCS ↑**	**NMSE ↓**	**SGCS ↑**	**NMSE ↓**	**SGCS ↑**	**NMSE ↓**	**SGCS ↑**	**NMSE ↓**	**SGCS ↑**
30km/h	2	**0.017**	**0.944**	0.179	0.678	0.186	0.894	0.072	0.927	0.036	0.908
	4	**0.045**	**0.896**	0.216	0.646	0.490	0.628	0.228	0.784	0.072	0.849
	8	**0.113**	**0.799**	0.586	0.376	0.860	0.214	0.601	0.502	0.148	0.723
Average	-	**0.058**	**0.880**	0.327	0.566	0.512	0.579	0.300	0.738	0.086	0.827
60km/h	2	**0.136**	**0.740**	0.201	0.655	0.846	0.176	0.518	0.435	0.150	0.716
	4	**0.151**	**0.718**	0.212	0.650	0.883	0.177	0.471	0.491	0.161	0.693
	8	**0.174**	**0.680**	0.641	0.340	0.959	0.118	0.672	0.335	0.183	0.666
Average	-	**0.154**	**0.713**	0.351	0.548	0.896	0.157	0.554	0.420	0.165	0.692
120km/h	2	0.163	0.687	0.211	0.661	0.919	0.129	0.520	0.444	**0.159**	**0.702**
	4	0.167	0.678	0.472	0.474	0.952	0.139	0.470	0.495	**0.164**	**0.689**
	8	0.190	0.658	0.808	0.246	0.982	0.073	0.870	0.180	**0.178**	**0.669**
Average	-	0.173	0.674	0.497	0.460	0.951	0.114	0.620	0.373	**0.167**	**0.686**
x km/h	2	**0.116**	**0.785**	0.202	0.659	0.948	0.224	0.350	0.619	0.134	0.749
	4	**0.138**	**0.748**	0.483	0.447	0.960	0.200	0.300	0.635	0.153	0.708
	8	**0.164**	**0.694**	0.793	0.243	0.984	0.078	0.798	0.255	0.178	0.680
Average	-	**0.139**	**0.743**	0.492	0.450	0.964	0.167	0.483	0.503	0.155	0.712

B. Implementation Details

This experiment was implemented on 4 NVIDIA RTX 4090 GPUs with 24 GB memory under Ubuntu 22.04.3 LTS environment. The AdamW [34] optimizer is used with an initial learning rate of 0.001, combined with a cosine annealing scheduler and a warm-up phase of 10% epochs to adjust the learning rate. We adopted the pre-trained CLIP ViT-B/16 model [19] as vision encoder and the pre-trained GPT-2 [15] as decoder-only large language model. The hyperparameter λ was set to 0.1, and the sensitivity threshold η in coherence segmentation was set to 0.05 based on validation performance.

C. Results

The prediction window P is set within {2, 4, 8}, while the look-back window size L is set at 12 for all datasets. We calculate the NMSE and SGCS of baseline and our proposed model and show the results in TABLE II. It can be observed that our proposed consistently outperforms most baseline methods across different speed scenarios and prediction lengths. Specifically, in low-mobility scenarios with user speed in 30 km/h, we achieves significant improvements in prediction window 8, increasing SGCS by up to 10.5% compared to the second-best method. In dataset with user speed in 60 km/h and mixed-velocity scenarios, our model maintains robust performance, but in high-velocity scenario it is marginally outperformed by LLM4CP. This is because temporal models excel at capturing long-term evolutionary trends, while spatial models are adept at extracting instantaneous structural features. By leveraging a joint spatial-temporal modeling mechanism, our model ensures both reconstruction precision and robustness in high-dynamic settings in complex spatial environments, thereby achieving superior generalization.

TABLE III

ABLATION STUDY OF KEY COMPONENTS. "W/O" DENOTES WITHOUT.

Setting	**NMSE**	**SGCS**
Ours (Full Model)	**0.116**	**0.785**
w/o coherence embedding (CE)	0.123	0.774
w/o R(2+1)DCVCNN	0.324	0.613
w/o CE & R(2+1)DCVCNN	0.329	0.610

D. Ablation Study

To verify the effectiveness of the key components in our proposed framework, specifically the coherence embedding and the R(2+1)DCVCNN module, we conducted an ablation study on the mixed-velocity dataset, as summarized in Table III.

The most significant performance degradation is observed when the R(2+1)DCVCNN module is removed. Thus this module serves as the fundamental component of our system, playing a decisive role in extracting comprehensive spatial features from the CSI. Without this module, the model fails to capture the phase-amplitude features of CSI, leading to poor prediction capability. The exclusion of the coherence embedding leads to a slight performance degradation, with NMSE rising to 0.123. Although this impact is less drastic than removing the R(2+1)DCVCNN module, it underscores this embedding as critical role in enforcing temporal consistency. Ultimately, the superior performance of full model across all metrics, compared to the variant lacking both components—demonstrates, integrates that the synergy between

spatial feature extraction and temporal coherence modeling is vital for achieving channel prediction.

V. Conclusion

In this work, we addressed the limitation of decoupled spatial and temporal modeling in existing studies by proposing a dual-domain augmented VLM-based framework for spatial-temporal channel prediction. Through the integration of the spatial-structural and temporal-coherence guidance blocks, we ingeste multi-dimensional CSI snapshots in both original tensor and flattened formats. Our empirical results demonstrate that this cross-modal alignment leads to significant gains in prediction accuracy, particularly in complex-mobility scenarios. These outcomes highlight the promising potential of VLM-based model in wireless communication applications.

References

[1] David Tse and Pramod Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[2] Fredrik Rusek, Daniel Persson, Buon Kiong Lau, et al., "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, 2013.

[3] Jiajia Guo, Tong Chen, Shi Jin, et al., "Deep learning for joint channel estimation and feedback in massive MIMO systems," *Digit. Commun. Networks*, vol. 10, no. 1, pp. 83–93, 2024.

[4] Yunwu Zhang, Shibao Li, Dongyang Li, et al., "Transformer-based predictive beamforming for integrated sensing and communication in vehicular networks," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 20690–20705, 2024.

[5] Jie Wang, Ying Ding, Shujie Bian, et al., "UL-CSI data driven deep learning for predicting DL-CSI in cellular FDD systems," *IEEE Access*, vol. 7, pp. 96105–96112, 2019.

[6] Lemayian Joel Poncha and Jehad M. Hamamreh, "Recurrent neural network-based channel prediction in mmimo for enhanced performance in future wireless communication," in *Proc. Int. Conf. UK-China Emerging Technol. (UCET)*, 2020, pp. 1–4.

[7] Hao Jiang, Mingyao Cui, Derrick Wing Kwan Ng, and Linglong Dai, "Accurate channel prediction based on transformer: Making mobility negligible," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2717–2732, 2022.

[8] Boxun Liu, Xuanyu Liu, Shijian Gao, et al., "LLM4CP: adapting large language models for channel prediction," *J. Commun. Inf. Networks*, vol. 9, no. 2, pp. 113–125, 2024.

[9] Chi Wu, Xinping Yi, Yiming Zhu, Wenjin Wang, Li You, and Xiqi 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.

[10] Sharan Mourya, Pavan Reddy Manne, SaiDhiraj Amuru, and Kiran Kumar Kuchi, "Spectral temporal graph neural network for massive MIMO CSI prediction," *IEEE Wirel. Commun. Lett.*, vol. 13, no. 5, pp. 1399–1403, 2024.

[11] Boxun Liu, Shijian Gao, Xuanyu Liu, Xiang Cheng, and Liuqing Yang, "Wifo: Wireless foundation model for channel prediction," *Science China Information Sciences*, vol. 68, no. 6, pp. 162302, 2025.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2022, pp. 15979–15988.

[13] Kareem E. Baddour and Norman C. Beaulieu, "Autoregressive modeling for fading channel simulation," *IEEE Trans. Wirel. Commun.*, vol. 4, no. 4, pp. 1650–1662, 2005.

[14] Hwanjin Kim, Sucheol Kim, Hyeongtaek Lee, Chulhee Jang, Yongyun Choi, and Junil Choi, "Massive MIMO channel prediction: Kalman filtering vs. machine learning," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 518–528, 2021.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[16] Yiming Cui, Jiajia Guo, Chao-Kai Wen, et al., "Exploring the potential of large language models for massive MIMO CSI feedback," *CoRR*, vol. abs/2501.10630, 2025.

[17] Jie Wang, Ying Ding, Shujie Bian, et al., "UL-CSI data driven deep learning for predicting DL-CSI in cellular FDD systems," *IEEE Access*, vol. 7, pp. 96105–96112, 2019.

[18] Jingxiang Yang, Liyan Li, and Min-Jian Zhao, "A blind CSI prediction method based on deep learning for V2I millimeter-wave channel," in *IEEE International Conference on Network Protocols, ICNP*, 2020, pp. 1–6.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763.

[20] Chao Jia, Yinfei Yang, Ye Xia, et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916.

[21] Junnan Li, Dongxu Li, Silvio Savarese, et al., "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023, vol. 202, pp. 19730–19742.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, et al., "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[23] Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang, "Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[24] Tete Xiao, Mannat Singh, Eric Mintun, et al., "Early convolutions help transformers see better," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 30392–30400.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[26] Du Tran, Heng Wang, Lorenzo Torresani, et al., "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention is all you need," 2023.

[28] Kai Ying, Yuanxun Liu, Disheng Xiao, et al., "A data-driven channel prediction model with coherence time embedding," in *IEEE Conference on Standards for Communications and Networking, CSCN 2024, Belgrade, Serbia, November 25-27, 2024*, 2024, pp. 320–321.

[29] Taesung Kim, Jinhee Kim, Yunwon Tae, et al., "Reversible instance normalization for accurate time-series forecasting against distribution shift," in *International Conference on Learning Representations (ICLR)*, 2022.

[30] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, vol. 139, pp. 5583–5594.

[31] Jingzhe Shi, Qinwei Ma, Huan Ma, et al., "Scaling law for time series forecasting," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[32] Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, et al., "Deeply-supervised nets," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015, vol. 38 of *JMLR Workshop and Conference Proceedings*.

[33] 3rd Generation Partnership Project (3GPP), "Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface," Technical Report (TR) 38.843, 3rd Generation Partnership Project (3GPP), June 2024, Release 18.

[34] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.