Spatial-Temporal Channel Prediction via a Dual-Guided VLM-based Framework

Anonymous ICME submission

***Abstract*—Channel state information (CSI) prediction is a crucial technology for massive multiple-input multiple-output (MIMO) systems in future sixth-generation (6G) wireless communication networks. Existing advancements predominantly focus on spatial or temporal features in isolation, thereby failing to capture the inherent spatial-temporal correlations of high-dimensional CSI. In this paper, leveraging the superior capability of vision-language models (VLMs) in capturing spatial-temporal dependencies via multimodal representation learning, we propose a novel dual-guided VLM-based framework for spatial-temporal CSI prediction. The framework integrates a spatial-structural guidance block to extract geometric features from CSI data and a temporal-coherence guidance block to capture temporal continuity 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 high-dimensional channel structures.**

***Index Terms*—Channel prediction, spatial-temporal modeling, vision-language models, multimodal learning, coherence embedding**

I. Introduction

The accurate acquisition of channel state information (CSI) is pivotal in Massive Multiple-Input Multiple-Output(MIMO) communication systems [1], facilitating advanced transmission strategies, such as precoding, beamforming, and power allocation, thereby significantly improving system performance. Typically, as illustrated in Fig. 1, CSI is acquired via pilot symbols embedded in the transmission frame. However, in high-mobility scenarios, channel aging renders the channel information outdated, and increasing pilot density to mitigate this incurs severe overhead. Therefore, channel prediction, defined as inferring future channel information based on historical data, has become a key technology for enhancing system performance.

Driven by recent breakthroughs in deep learning (DL), artificial intelligence (AI) has significantly improved performance in wireless communications, such as channel estimation [2], channel feedback [3], signal detection [4], and beamforming [5], while also encompassing channel prediction. Recurrent Neural Networks(RNNs) and Long Short-Term Memory(LSTM) have demonstrated strong performance in channel prediction by effectively capturing dynamic temporal dependencies [6], [7]. Studies on temporal feature modeling of CSI have been further advanced by Transformers [8] and Large Language Model (LLM)-based approaches [9], which excel at modeling long-range sequential correlations. In parallel, researchs including complex-valued convolution neural network (CVCNN) [10] and STEM GNN [11] have demonstrated superior capability in modeling the spatial structure of wireless channels. WiFo [12] innovatively applies masked autoencoders (MAE) [13] to reconstruction tasks, treating CSI matrices as visual images to extract deep spatial structural features.

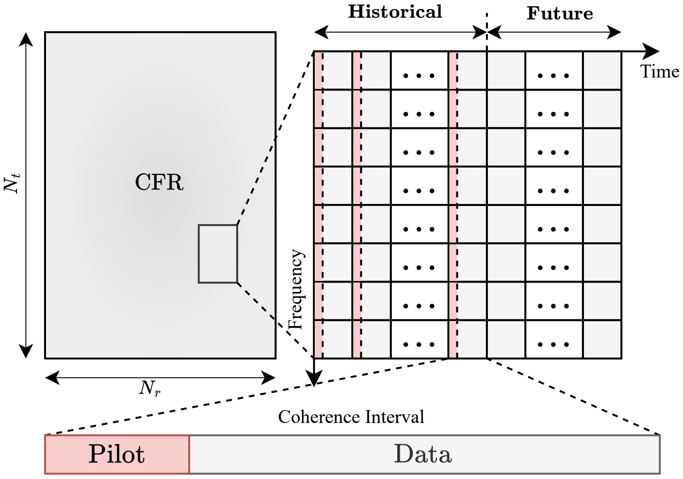

Fig. 1. The data structure in MIMO-OFDM system. Pilot symbols are predefined reference signals used for channel estimation, data symbols are unknown user information that must be detected based on the estimated channel.

Despite these achievements, prior research fundamentally treats the spatial and temporal domains in isolation, ignoring the intrinsic spatiotemporal coupling of wireless channels. Specifically, temporal-centric approaches typically flatten high-dimensional CSI tensors into vectors, a process that inevitably obliterates the underlying spatial structure. Conversely, spatial-focused models, while preserving structural features, often struggle to maintain precision across long prediction windows due to limited temporal modeling capabilities. To bridge this gap, we introduce a dual-guided 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 decoder-only LLMs.
- We employ a **Spatial-Structural Guidance** block and a **Temporal-coherence Guidance** block to capture the spatial-temporal feature of wireless channels in Massive MIMO system.

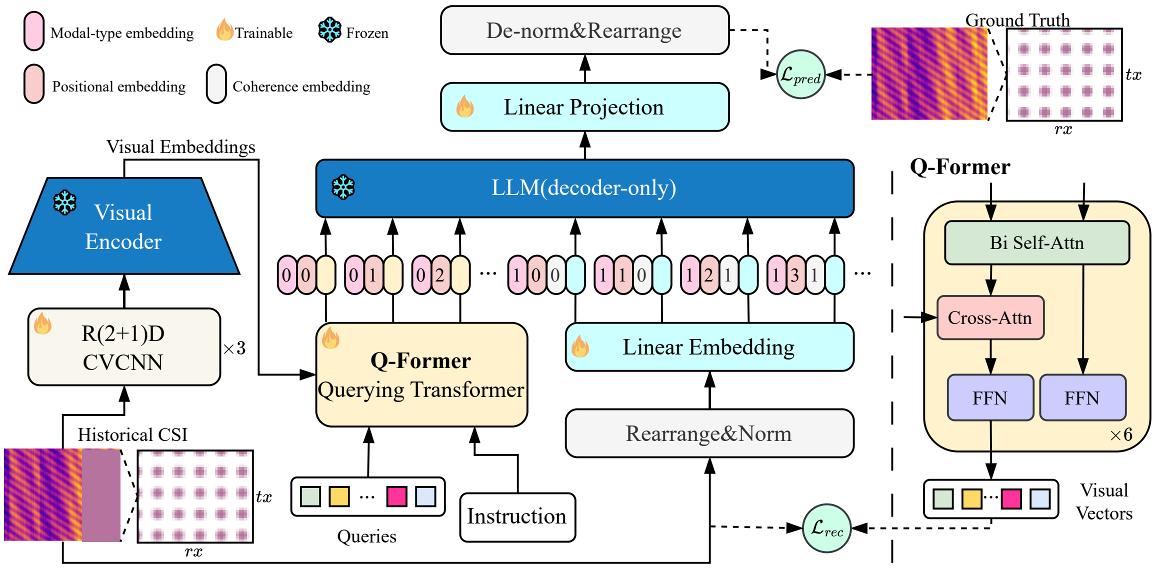

Fig. 2. Overview of proposed method.

- Experimental results show that this method achieves superior performance on spatial-temporal channel prediction tasks, measured by NMSE and SGCS.

II. Related Work

A. Channel prediction (CP)

Deep learning-based strategies have been increasingly applied to CP. In particular, recurrent neural network, such as LSTM, excel in channel prediction by capturing dynamic temporal features [6], [7]. Additionally, CNN-based approaches [14], [15] employ convolutional neural networks(CNNs) with complex-valued convolutional layers to extract the spatial structure of CSI data. By facilitating parallel computation and paying attention to important patterns, Transformer-based models have significantly advanced channel prediction in challenging environments [8]. Nevertheless, the lack of accurate modeling and robust generalization remains a key limitation of these models. More recently, inspired by the success of large language modals in fields of natural language processing (NLP) [16], some studies reflects their potential in CSI tasks. For instance, LLM4CP [9] fine-tunes a pre-trained GPT-2 for CSI data and deploy a set of modules to boost model effectiveness. Similarly, method [17] leverages the powerful noise removal capability of LLM to improve CSI reconstruction performance. However, these methods remain limited in their ability to align CSI data with the textual input required by LLMs and ignore the inherent structural similarities between CSI and computer vision (CV) data.

B. Vision-Language Models (VLMs)

VLMs are fundamental to multimodal learning, enabling joint understanding of visual and textual modalities. CLIP [18] and ALIGN [19] demonstrate that contrastive learning effectively aligns image and text embeddings in a shared latent space. Studies such as Flamingo [20] and BLIP [21] further improve cross-modal interaction by incorporating cross-attention mechanism. Beyond conventional computer vision, recent research [22] has begun to extended the application of VLMs to non-visual domains, transforming structured data into visual representations and enabling the reuse of pre-trained visual backbones. However, the exploration of VLMs in channel prediction is still in its infancy. These advances demonstrate that VLMs are not limited to native images and can serve as universal cross-modal learners across diverse and data-intensive tasks.

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 $k \in \{1, \dots, N_c\}$, 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, i.e., the channel state information (CSI). Aggregating the channel matrices over all subcarriers, we represent the CSI snapshot at time t as $\mathcal{H}_t \triangleq \{\mathbf{H}_{t,1}, \dots, \mathbf{H}_{t,N_c}\}$. 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:

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

Algorithm 1 Coherence Segmentation

```
**Input:** Historical CSI sequence $\mathbf{X} \in \mathbb{C}^{T \times D}$, sensitivity threshold $\eta$ (Hyperparameter)
**Output:** Coherence Segment Indices $\mathbf{S} \in \mathbb{Z}^{T}$
$current\_id \leftarrow 0$
$\mathbf{S}[1] \leftarrow current\_id$
**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** $\mathbf{S}$
```

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 forzen vision encoder and frozen decoder-only LLM. In order to enhance the extraction of spatiotemporal CSI features by the VLM, we introduce a dual-guidance mechanism: the Spatial-Structural Guidance Block and the Temporal-Coherence Guidance Block. The following sections provide a detailed description of each block.

C. 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 [23] that proving early convolutions improve Visual Transformer(ViT) optimization, we first process the historical CSI $\mathcal{X}_{\text{his}} = \{\mathcal{H}_{t-P+1}, \ldots, \mathcal{H}_t\} \in \mathbb{C}^{P \times N_c \times N_r \times N_t}$ through residual (2+1) dimensional complex-valued convolutional layers before feeding it into the pre-trained vision encoder. We utilize a complex-valued convolutional neural network(CVCNN) instead of whole CNN to explicitly preserve phase information and factorize the standard 3D convolutional kernel size $3 \times 3 \times 3$ into $3 \times 3 \times 1$ and $1 \times 1 \times 3$ to effectively capture the spatial correlations of $\mathcal{X}_{\text{hist}}$ between the transmit antennas and receive antennas and the local temporal dynamics in the P dimension.

We train Querying Transformer (Q-Former) from scratch. The input to the Q-Former contains a set of K learnable query embeddings $\mathbf{Q} \in \mathbb{R}^{K \times D_q}$ and the instruction-aware visual features from the output embeddings of the frozen vision encoder $\mathbf{F}_{ve}$. The query embeddings are 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. Through $L = 6$ stacked layers, the queries interact with the visual features to compress spatial information into instruction-aware visual representations. Each layer consists of a Multi-Head Self-Attention(MHSA) module, a Multi-Head Cross-Attention(MHCA) module, and a Feed-Forward Network(FFN). The update process for the l-th layer is formulated as:

$$\tilde{\mathbf{Q}}^{(l)} = \mathbf{Q}^{(l-1)} + \mathrm{MHSA}\left(\mathrm{LN}([\mathbf{Q}^{(l-1)}, \mathbf{I}])\right), \tag{3}$$

$$\hat{\mathbf{Q}}^{(l)} = \tilde{\mathbf{Q}}^{(l)} + \mathrm{MHCA}\left(\mathrm{LN}(\tilde{\mathbf{Q}}^{(l)}), \mathbf{F}_{ve}, \mathbf{F}_{ve}\right), \tag{4}$$

$$\mathbf{Q}^{(l)} = \hat{\mathbf{Q}}^{(l)} + \mathrm{FFN}\left(\mathrm{LN}(\hat{\mathbf{Q}}^{(l)})\right), \tag{5}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization, $\mathbf{I}$ denotes the instruction prompt, FFN consists of two linear layers with a ReLU activation: $\mathrm{FFN}(x) = \max(0, x\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$, and $\mathbf{Q}^{(l)}$ represents the output of the l-th Q-Former layer.

D. Temporal-coherence guidance block

To be fed into frozen LLM model, we first reshape historical CSI $\mathcal{X}_{his}$ into dense real-valued form $\mathcal{X}_{rearrange} \in \mathbb{R}^{P \times (2 \cdot N_c \cdot N_r \cdot N_t)}$, where the real and imaginary parts of each complex entry are decomposed and concatenated, and then apply instance normalization [24] to standardize each sample to zero mean and unit variance.

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

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

$$\mathbf{H}_{in} = [\mathbf{F}_{vv}, \mathbf{F}_a + \mathbf{E}_{coh}] + \mathbf{E}_{pos} + \mathbf{E}_{modal}, \tag{6}$$

where $\mathbf{E}_{modal}$ and $\mathbf{E}_{pos}$ denote standard modal-type and positional embeddings in [25], and $\mathbf{E}_{coh}$ represents the coherence embedding. The LLM processes these embeddings 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 CSI $\hat{\mathcal{Y}}_{pred} = \{\hat{\mathcal{H}}_{t+1}, \ldots, \hat{\mathcal{H}}_{t+L}\} \in \mathbb{C}^{L \times N_c \times N_r \times N_t}$.

E. Training

We form training samples by sliding a window [26] 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:

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda \mathcal{L}_{rec}, \tag{7}$$

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)

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

$$\mathcal{L}_{pred} = \frac{\|\hat{\mathcal{Y}}_{pred} - \mathcal{X}_{gt}\|_F^2}{\|\mathcal{X}_{gt}\|_F^2}, \tag{8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Inspired by the concept of Deep Supervision [27], which employ discriminative classifiers for intermediate layers, we propose an auxiliary reconstruction objective $\mathcal{L}_{rec}$:

$$\mathcal{L}rec = \|\mathcal{R}(\mathbf{Z}_q \mathbf{W}_{rec} + \mathbf{b}_{rec}) - \mathcal{X}_{his}\|_F^2, \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{Z}_q \in \mathbb{R}^{N_q \times D}$ 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.

2) Baseline: To evaluate the performance of our proposed model, we compared it with existing methods, such as complex-valued convolutional neural networks (CVCNN) [10], long short-term memory (LSTM) networks [7], STEM GNN [11], and LLM4CP [9]. 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 [28] to quantifies the numerical and spatial discrepancy between the predicted channel states and the ground truth, respectively. NMSE serves as the optimization objective in our framework, whose formulation is detailed in Eq. 8. SGCS is defined as:

$$\mathrm{SGCS} = \frac{1}{N_{sp}} \sum_{i=1}^{N_{sp}} \frac{1}{N_{rb}} \sum_{j=1}^{N_{rb}} \frac{\|\mathbf{H}_{i,j}\hat{\mathbf{H}}_{i,j}\|^2}{\|\mathbf{H}_{i,j}\|^2 \|\hat{\mathbf{H}}_{i,j}\|^2}, \tag{10}$$

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.

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 [29] 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 [18] as vision encoder and the pre-trained GPT-2 [30] 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. Result

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 compelx-mobility scenarios(x km/h), our model maintains robust performance, but in high-moblity scenario(120 km/h). 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 spatiotemporal modeling mechanism, our model ensures both robustness in high-dynamic settings and reconstruction precision in complex spatial environments, thereby achieving superior generalization.

D. Ablation Study

To verify the effectiveness of the key components in our proposed framework, specifically the Coherence Embedding and the R(2+1)D CVCNN 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)D CVCNN module is removed. indicating that

[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	-	**0.131**	**0.752**	0.417	0.506	0.831	0.254	0.489	0.508	0.143	0.730
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	-	**0.131**	**0.752**	0.417	0.506	0.831	0.254	0.489	0.508	0.143	0.730
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	-	**0.131**	**0.752**	0.417	0.506	0.831	0.254	0.489	0.508	0.143	0.730

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

TABLE IV
THE IMPACT OF THRESHOLD η ACROSS ALL DATASETS IN TERM OF NMSE.

Threshold (η)	30km/h	60km/h	120km/h	Mix.
0.01	0.120	0.176	0.192	0.169
0.05	**0.113**	0.174	**0.190**	**0.164**
0.10	0.115	**0.172**	0.194	0.170
0.20	0.117	0.175	0.193	0.166

the R(2+1)D CVCNN serves as the fundamental backbone of our system, playing a decisive role in extracting local spatiotemporal features from the CSI sequence. Without this module, the model fails to capture the intricate physical structures of the channel, leading to poor reconstruction capability. The omission of the Coherence Embedding leads to a noticeable decline in performance, with NMSE rising to 0.123. Although this impact is less drastic than removing the CNN backbone, it highlights the embedding's role in enforcing temporal alignment. By explicitly modeling the continuity of channel evolution, this component is essential for fine-tuning prediction accuracy. Furthermore, the full model's superior performance across all metrics—compared to the variant lacking both components—demonstrates that integrating spatiotemporal feature extraction (via R(2+1)D) with temporal coherence modeling is vital for achieving precise channel prediction

As shown in Table IV, we evaluate the NMSE performance under varying threshold values $\eta \in \{0.01, 0.05, 0.10, 0.20\}$. We observe that the performance peaks at $\eta = 0.05$, achieving the best NMSE of 0.113, 0.190, and 0.164 for the 30km/h, 120km/h, and Mixed datasets, respectively. Both lower ($\eta = 0.01$) and higher ($\eta = 0.20$) thresholds lead to increased prediction errors. Therefore, $\eta = 0.05$ is selected as the optimal hyperparameter to balance feature retention and noise suppression.

V. CONCLUSION

In this work, we addressed the limitation of decoupled spatial and temporal modeling in existing studies by proposing a unified VLM-based framework. Through the synergistic integration of the Spatial-Structural and Temporal-Coherence Guidance Blocks, we successfully aligned the high-dimensional CSI features with the semantic reasoning capabilities of LLMs. Our empirical results demonstrate that this cross-modal alignment leads to significant gains in prediction accuracy and structural fidelity, particularly in complex, high-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] Jiajia Guo, Tong Chen, Shi Jin, Geoffrey Ye Li, Xin Wang, and Xiaolin Hou, "Deep learning for joint channel estimation and feedback in massive MIMO systems," *Digit. Commun. Networks*, vol. 10, no. 1, pp. 83–93, 2024.

[3] Muhan Chen, Jiajia Guo, Chao-Kai Wen, Shi Jin, Geoffrey Ye Li, and Ang Yang, "Deep learning-based implicit CSI feedback in massive MIMO," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 935–950, 2022.

[4] Pengxuan Gao, Disheng Xiao, Ruiheng Zou, and Kai Ying, "A self-supervised UAV detection method based on channel state information," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*. 2025, pp. 1–5, IEEE.

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

[6] Jie Wang, Ying Ding, Shujie Bian, Yang Peng, Miao Liu, and Guan Gui, "UL-CSI data driven deep learning for predicting DL-CSI in cellular FDD systems," *IEEE Access*, vol. 7, pp. 96105–96112, 2019.

[7] Lemayian Joel Poncha and Jehad M. Hamamreh, "Recurrent neural network-based channel prediction in mmimo for enhanced performance in future wireless communication," in *2020 International Conference on UK-China Emerging Technologies, UCET 2020, Glasgow, United Kingdom, August 20-21, 2020*. 2020, pp. 1–4, IEEE.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. 2022, pp. 15979–15988, IEEE.

[14] Jie Wang, Ying Ding, Shujie Bian, Yang Peng, Miao Liu, and Guan Gui, "UL-CSI data driven deep learning for predicting DL-CSI in cellular FDD systems," *IEEE Access*, vol. 7, pp. 96105–96112, 2019.

[15] Jingxiang Yang, Liyan Li, and Min-Jian Zhao, "A blind CSI prediction method based on deep learning for V2I millimeter-wave channel," in *28th IEEE International Conference on Network Protocols, ICNP 2020, Madrid, Spain, October 13-16, 2020*. 2020, pp. 1–6, IEEE.

[16] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[17] Yiming Cui, Jiajia Guo, Chao-Kai Wen, Shi Jin, and En Tong, "Exploring the potential of large language models for massive MIMO CSI feedback," *CoRR*, vol. abs/2501.10630, 2025.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Marina Meila and Tong Zhang, Eds. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR.

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Marina Meila and Tong Zhang, Eds. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916, PMLR.

[20] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, Eds., 2022.

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, Eds. 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, PMLR.

[22] 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," *CoRR*, vol. abs/2502.04395, 2025.

[23] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross B. Girshick, "Early convolutions help transformers see better," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, Eds., 2021, pp. 30392–30400.

[24] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo, "Reversible instance normalization for accurate time-series forecasting against distribution shift," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. 2022, OpenReview.net.

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Marina Meila and Tong Zhang, Eds. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594, PMLR.

[26] Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li, "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*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, Eds., 2024.

[27] Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang, and Zhuowen Tu, "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*, Guy Lebanon and S. V. N. Vishwanathan, Eds. 2015, vol. 38 of *JMLR Workshop and Conference Proceedings*, JMLR.org.

[28] 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.

[29] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. 2019, OpenReview.net.

[30] 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.