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

Anonymous ICME submission

Abstract—**Channel prediction is a crucial task in various communications applications. Although existing methods have employed large language models (LLMs) with robust modeling and generalization abilities to improve prediction performance, their practical use is limited by overlooking the underlying visual information of CSI. Recently, vision-language models (VLMs) have transformed multimodal learning by mapping images and text into a unified semantic space. In this paper, we investigate a VLM-based channel prediction framework (CPVLM), aiming to bridge CSI and image-text data. Specifically, we utilize inherent structural alignment between complex-valued CSI and visual data. Subsequently, we devise a coherence embedding method, enabling VLMs to interpret the entire CSI sequence as a coherent linguistic representation. Experimental results demonstrate that CPVLM outperforms the compared schemes and establishes a new direction for channel prediction. [1] [2] [3] [4] [5]**

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

I. Introduction

Accurately acquiring channel state information (CSI) is pivotal in a variety of wireless communication technologies and applications [6], such as precoding, beamforming and power allocation, aiming to improve communication quality and throughput. However, since wireless channels are highly dynamic and are affected by multiple factors such as multipath propagation, user mobility, and environmental changes, the accurate acquisition of CSI presents considerable challenges. Therefore, channel prediction, as a key technology for estimating future channel states based on historical or current observations, has become a promising research prospect.

In recent years, artificial intelligence (AI) has witnessed remarkable progress, particularly driven by advances in deep learning. These developments have significantly improved performance in wireless communications such as channel estimation [7], channel feedback [8], signal detection [9], and beamforming [10], which brings hope to solve the problems of the channel prediction. Recurrent neural networks (RNN) have demonstrated strong performance in channel prediction by effectively capturing dynamic temporal dependencies [11], [12]. In parallel, convolutional neural networks (CNNs) have been employed to model the spatial characteristics of CSI data [13], [14]. Moreover, Transformer-based architectures [15] have further advanced the field by enabling parallel processing and leveraging attention mechanisms to focus on salient patterns, thus achieving superior performance under complex and dynamic channel conditions. Fortunately, the emergence of pre-trained large language models (LLMs), distinguished by their powerful modeling and generalization capabilities, has facilitated their application to channel prediction. Unlike classical deep learning approaches, they design customized preprocessor, embedding, and output modules to bridge the gap between CSI data and textual information, aiming to fully exploit the transferable knowledge embedded in pre-trained language models.

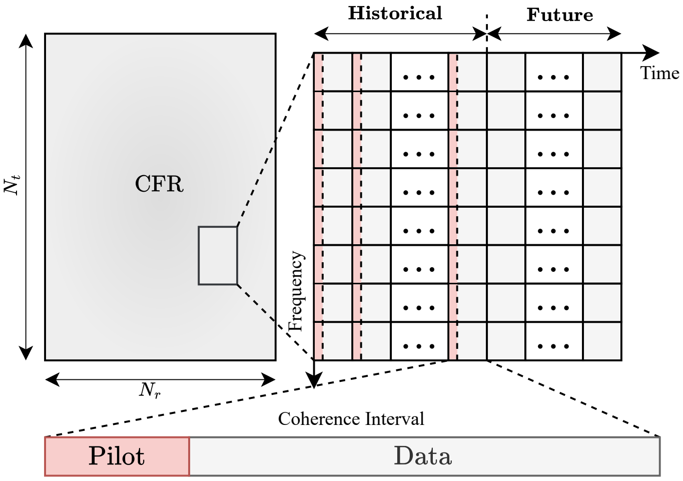

Fig. 1. The data structure in MIMO-OFDM system. Pilot symbols are predefined reference signals used for channel estimation, data symbols are unknown user information that must be detected based on the estimated channel.

Despite these achievements, existing approaches still face several critical limitations. Most notably, they often ignore the natural structural similarities between complex-valued CSI data and image data, and fail to exploit the combined advantages of multiple modalities. These limitations hinder their applicability in more complex or real-world scenarios.

The main challenge at present is how to combine image-text pattern with CSI data to improve prediction performance. To bridge this gap, we introduce CPVLM, a novel framework that use a pre-trained vision-language model to enhance channel prediction by jointly integrating visual, textual, and CSI modalities. Leveraging the strong cross-modal alignment capabilities of vision-language models, CPVLM effectively embeds CSI data into the shared visual-linguistic semantic space, thereby enabling seamless integration across the three modalities. This unified representation fosters cross-modal interactions and allows each modality to contribute complementary information. Specifically, complex-valued CSI data—comprising in-phase (I) and quadrature (Q) components—can be naturally organized into a two-dimensional matrix across multiple time steps and subcarriers, exhibiting a structural resemblance to visual data. Moreover, motivated by the observation that model like [16] maps time-series features into the embedding space of large language models for unified semantic modeling, coherence embedding, aligning CSI embeddings space with VLMs embeddings space, is adopted.

- We propose VLM-based framework (CPVLM) to enhance channel prediction by leveraging the complementary strengths of CSI data, visual, and textual modalities.
- We employ a **Visual Processor** module, a **Textual Processor** module, and a **Multimodal Alignment Block** module to integrate heterogeneous modalities, thereby enhancing the accuracy of channel state information prediction.
- Experimental results show that this method achieves optimal performance on channel prediction tasks and exhibits excellent generalization capabilities of **few-shot**.

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 [11], [12].Additionally, CNN-based approaches [13], [14], which model spatial characteristics, is introduced. By facilitating parallel computation and paying attention to important patterns, Transformer-based models have significantly advanced channel prediction in challenging environments [15]. 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) [17], some studies reflects their potential in CSI tasks. For instance, LLM4CP [18] fine-tunes a pre-trained GPT-2 for CSI data and deploy a set of modules to boost model effectiveness. Similarly, method [19] 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 [20] and ALIGN [21] demonstrate that contrastive learning effectively aligns image and text embeddings in a shared latent space. Studies such as Flamingo [22] and BLIP [23] further improve cross-modal interaction by incorporating cross-attention mechanism. Beyond conventional computer vision, recent research [24] 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 $k \in \{1, \ldots, N_c\}$, 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, i.e., the channel state information (CSI). Aggregating the channel matrices over all subcarriers, we represent the CSI snapshot at time t as $\mathcal{H}_t \triangleq \{\mathbf{H}_{t,1}, \ldots, \mathbf{H}_{t,N_c}\}$. 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 forzen vision encoder and frozen decoder-only LLM. In order to improve the VLM's extraction of CSI spatiotemporal features, we introduce a dual-guidance mechanism: spatial-structural guidance block and 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 [25] 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 pretrained 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.

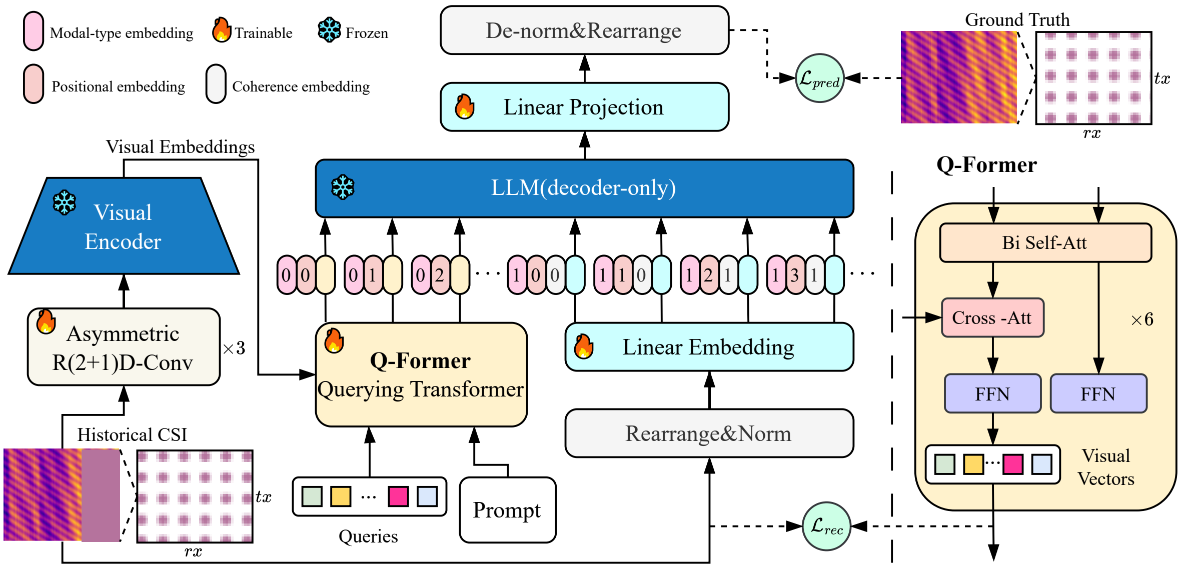

Fig. 2. Overview of proposed method.

We fine-tune a pre-trained Querying Transformer (Q-Former). 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 prompt-aware visual features from the output embeddings of the frozen visual encoder $\mathbf{F}_{ve}$. The learnable queries interact with the visual embedding $\mathbf{F}_{ve}$ through 6 alternating layers of bidirectional self-attention layers, cross-attention layers, and feed-forward layers, consisting of two linear transformations with a ReLU activation in between, to compress spatial information into K encoded visual vectors $\mathbf{F}_{vv}$:

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

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

$$\mathbf{Q}^{(l)} = \hat{\mathbf{Q}}^{(l)} + \max(0, \mathrm{LN}(\hat{\mathbf{Q}}^{(l)})W_1 + b_1)W_2 + b_2, \tag{5}$$

$$\mathbf{F}_{vv} = \mathbf{Q}^{(6)}, \tag{6}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization, $\mathrm{MSA}(Q, K, V)$ represents the Multi-Head Self-Attention mechanism, $\mathrm{MCA}(Q, K, V)$ represents the Multi-Head Cross-Attention mechanism, and $\mathbf{Q}^{(n)}$ represents the output of the n-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 form $\mathcal{X}_{rearrange} \in \mathbb{R}^{P \times (2 \cdot N_c \cdot N_r \cdot N_t)}$, where each complex entry is decomposed into its real and imaginary part, and then apply instance normalization [26] to the input $\mathcal{X}_{rearrange}$, standardizing each sample to zero mean and unit variance.

We introduce the Coherence Embedding as the primary mechanism of this block, inspired 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.

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

$$\mathbf{H}_{in} = [\mathbf{F}_{vv}, \mathbf{F}_a + \mathbf{E}_{coh}] + \mathbf{E}_{pos} + \mathbf{E}_{modal}, \tag{7}$$

where $\mathbf{E}_{modal}$ and $\mathbf{E}_{pos}$ denote standard modal-type and positional embeddings in [27], and $\mathbf{E}_{coh}$ represents the novel coherence embedding method, calculated from $\boldsymbol{Embedding}(\mathbf{S})$.

is a learnable coherence embedding vector designed to guide the frozen LLM in maintaining the temporal causality and physical consistency of the predicted CSI series. The LLM output hidden states, which encode future channel dynamics, are passed through a linear projection layer. We then apply denormalization and reshaping to map these semantic linear representations back to 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 [28] 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 $\mathcal{X}_{\mathrm{gt}}$ 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:

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda \mathcal{L}_{rec}, \tag{8}$$

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

$$\mathcal{L}_{pred} = \frac{\|\hat{\mathcal{Y}}_{pred} - \mathcal{X}_{gt}\|_F^2}{\|\mathcal{X}_{gt}\|_F^2}, \tag{9}$$

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
    $\boldsymbol{\delta} \leftarrow \|\mathbf{X}_t - \mathbf{X}_{t-1}\|_2$
    $ratio \leftarrow \frac{\boldsymbol{\delta}}{\|\mathbf{X}_{t-1}\|_2 + \epsilon}$
    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 $\|\cdot\|_F$ denotes the Frobenius norm. Inspired by the concept of Deep Supervision [29], which employ discriminative classifiers for intermediate layers, we propose an auxiliary reconstruction objective $\mathcal{L}_{rec}$:

$$\mathcal{L}rec = \|\mathcal{R}(\mathbf{Z}_q\mathbf{W}_{rec} + \mathbf{b}_{rec}) - \mathcal{X}_{his}\|_F^2, \tag{10}$$

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 spatiotemporal 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 II.

2) Baseline: To evaluate the effectiveness of LM-net, we compared it against several methods, such as convolutional neural networks (CNNs) [13], and long short-term memory (LSTM) networks [12]. To ensure a fair comparison across all baselines, we adopt a unified experimental framework.

3) Evaluation Metrics: In this study, we use **N**ormalized **M**ean **S**quared **E**rror (NMES) and **S**patial-**G**rouped **C**osine **S**imilarity (SGCS) as evaluation metrics to measure the prediction performance of the models.

NMSE provides a scale-invariant and interpretable metric that effectively reflects the accuracy of predicted magnitudes.

$$\text{NMSE} = \frac{\|\widehat{\mathbf{H}} - \mathbf{H}\|_2^2}{\|\mathbf{H}\|_2^2}, \tag{11}$$

where $\mathbf{H} \in \mathbb{C}^{N_s\times N}$ and $\hat{\mathbf{H}}$ represents the target channel matrix and the model output, respectively and $\|\cdot\|_2$ represents the Frobenius norm.

SGCS quantifies the angular difference between two vectors, with values ranging from -1 to 1.

$$SGCS = \frac{1}{N_s}\frac{1}{N}\sum_{i=0}^{N_s-1}\sum_{j=0}^{N-1}\frac{\mathbf{H_{i,j}}\widehat{\mathbf{H_{i,j}}}}{\|\mathbf{H_{i,j}}\|\|\widehat{\mathbf{H_{i,j}}}\|}, \tag{12}$$

where $\mathbf{H} \in \mathbb{C}^{N_s\times N}$. These two evaluation metrics assess channel prediction from two complementary aspects: the directional angle and the complex amplitude, respectively, demonstrating their feasibility and effectiveness.

B. Implement Details

This experiment was conducted on a machine running Ubuntu 22.04.3 LTS, equipped with four NVIDIA RTX 4090 GPUs (24 GB of video memory each). For the frozen backbone, we adopted the base CLIP ViT model [20] as visual encoder and GPT-2 as decoder-only large language model.

C. Result

D. Abtion Study

V. Conclusion

In this paper, we introduces CPVLM, a novel framework leveraging vision-language models (VLMs) for channel state information (CSI) prediction. By harnessing the intrinsic structural similarity between complex-valued CSI and visual data, CPVLM effectively bridges CSI representations with the visual-linguistic semantic space. To strengthen the alignment, we develop a coherence embedding technique that transforms the entire CSI sequence into a unified linguistic representation, enabling the model to capture both temporal dynamics and semantic relationships. Experimental results demonstrate that CPVLM consistently outperforms baseline methods. These outcomes highlight the promising potential of VLMs in wireless communication applications and open new directions for multimodal modeling of CSI data.

References

[1] Jiaming Cheng, Wei Chen, Jialong Xu, Yiran Guo, Lun Li, and Bo Ai, "Swin transformer-based CSI feedback for massive MIMO," in *23rd IEEE International Conference on Communication Technology, ICCT 2023, Wuxi, China, October 20-22, 2023*. 2023, pp. 809–814, IEEE.

[2] 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.

[1] www.mobileai-dataset.com

Models		CPVLM		CNN		RNN		LSTM		Transformer		LLM4CP	
Metric		NMSE	SGCS	NMSE	SGCS	NMSE	SGCS	NMSE	SGCS	NMSE	SGCS	NMSE	SGCS
30 km/h	5	**0.012**	**0.972**	0.096	0.837	0.070	0.877	0.019	0.959	0.024	0.960	0.022	0.959
	8	**0.006**	**0.987**	0.076	0.868	0.055	0.900	0.014	0.965	0.027	0.953	0.016	0.968
	10	**0.005**	**0.988**	0.075	0.870	0.054	0.902	0.014	0.967	0.032	0.955	0.023	0.961
	16	**0.006**	**0.988**	0.092	0.849	0.068	0.891	0.022	0.975	0.040	0.946	0.026	0.946
60 km/h	5	**0.120**	**0.776**	0.207	0.688	0.188	0.709	0.151	0.752	0.140	0.766	0.145	0.759
	8	**0.122**	0.771	0.193	0.698	0.179	0.713	0.150	0.744	0.134	0.771	0.127	**0.778**
	10	**0.126**	0.764	0.188	0.702	0.174	0.723	0.147	0.765	0.130	**0.787**	0.138	0.776
	16	0.137	0.745	0.203	0.695	0.192	0.712	0.169	0.745	0.145	0.768	**0.135**	**0.777**
120 km/h	5	**0.176**	0.707	0.241	0.650	0.225	0.666	0.193	0.699	0.182	**0.722**	0.188	0.710
	8	0.175	0.708	0.215	0.663	0.207	0.673	0.192	0.692	0.173	**0.728**	**0.166**	0.723
	10	0.174	0.702	0.209	0.672	0.202	0.681	0.188	0.698	**0.172**	**0.732**	0.180	0.715
	16	0.178	0.703	0.212	0.668	0.209	0.677	0.202	0.696	0.186	**0.724**	**0.169**	0.719
x km/h	5	**0.116**	**0.802**	0.198	0.708	0.180	0.731	0.144	0.775	0.140	0.788	0.142	0.782
	8	**0.114**	0.792	0.178	0.722	0.166	0.742	0.141	0.784	0.133	0.792	0.119	**0.803**
	10	**0.118**	0.787	0.174	0.719	0.164	0.743	0.143	0.791	0.132	**0.796**	0.138	0.793
	16	**0.131**	0.775	0.184	0.711	0.174	0.733	0.156	0.778	0.142	0.774	0.149	**0.798**
average	-	**0.107**	**0.810**	0.171	0.732	0.157	0.755	0.128	0.799	0.121	0.810	0.118	0.810

TABLE I

FULL-SHOT LEARNING ON ALL TRAINING DATA. THE SIZE OF OBSERVATION WINDOW IS SET AS 12 AND PREDICTION WINODW 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

TABLE II

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)

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "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, OpenReview.net.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong, "TEST: text prototype aligned embedding to activate llm's ability for time series," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. 2024, OpenReview.net.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.