• RESEARCH PAPER •

WiFo: Wireless Foundation Model for Channel Prediction

Boxun LIU, Shijian GAO, Xuanyu LIU, Xiang CHENG* & Liuqing YANG

Abstract Channel prediction permits to acquire channel state information (CSI) without signaling overhead. However, almost all existing channel prediction methods necessitate the deployment of a dedicated model to accommodate a specific configuration. Leveraging the powerful modeling and multi-task learning capabilities of foundation models, we propose the first space-time-frequency (STF) wireless foundation model (WiFo) to address time-frequency channel prediction tasks in a one-for-all manner. Specifically, WiFo is initially pre-trained over massive and extensive diverse CSI datasets. Then, the model will be instantly used for channel prediction under various CSI configurations without any fine-tuning. We propose a masked autoencoder (MAE)-based network structure for WiFo to handle heterogeneous STF CSI data, and design several mask reconstruction tasks for self-supervised pre-training to capture the inherent 3D variations of CSI. To fully unleash its predictive power, we build a large-scale heterogeneous simulated CSI dataset consisting of 160K CSI samples for pre-training. Simulations validate its superior unified learning performance across multiple datasets and demonstrate its state-of-the-art (SOTA) zero-shot generalization performance via comparisons with other full-shot baselines.

1 Introduction

Massive multi-input multi-output (MIMO) and orthogonal frequency division multiplexing (OFDM) have been two cornerstone technologies over the past decade. [1]. The performance of MIMO-OFDM systems hinges upon accurate channel state information (CSI), typically obtained through channel estimation [2]. However, in highly-dynamic scenarios, the channel coherence time is significantly limited, resulting in a sharp increase in channel estimation overhead. Therefore, channel prediction is proposed as a promising technology to reduce the overhead of CSI acquisition, where the estimated CSI is extrapolated in time or frequency. Extensive studies have been conducted on model-based channel prediction methods, including traditional autoregressive and polynomial extrapolation methods, as well as advanced vector Prony-based prediction methods [3] and high-resolution parameter estimation schemes [4]. Nevertheless, due to the difficulty of modeling complicated practical channels, the prediction accuracy of model-based approaches is limited.

Deep learning has emerged as a powerful tool to capture complex patterns in a data-driven manner without any prior assumption. It has been widely applied in various aspects of wireless communications, including time-domain and frequency-domain channel prediction addressed in this paper. For instance, several neural networks, including transformer [5], ConvLSTM [6], and variational autoencoders (VAE) [7], have been explored for channel prediction to learn temporal and frequency variations. In addition, several channel prediction methods based on physics-informed deep learning have been proposed to incorporate prior knowledge of the CSI or the propagation environment [8]. Recently, pre-trained large language models (LLMs) have been applied to MIMO-OFDM channel prediction [9] for enhanced accuracy and generalizability. However, existing deep learning-based channel prediction research is still at the starting stage and awaiting several improvements. First, due to the limited model scale, existing small model-based methods are not well poised to predict complex practical spatial-temporal-frequency channels accurately. Secondly, existing deep learning-based schemes are only trained under specific CSI distributions and specific system configurations. Correspondingly, these methods require retraining when scenarios and system parameters change, which leads to significant training overhead. Moreover, separate

---

networks are designed and deployed for time-domain and frequency-domain channel prediction, increasing the storage, computation, and management overhead at the base station (BS).

Recently, the emergence of foundational models [10] has led to a paradigm shift in deep learning. Specifically, large-scale neural networks pre-trained on vast and diverse data in a self-supervised manner can achieve remarkable generalization capabilities across various downstream tasks, significantly outperforming task-specific models. Its powerful modeling capabilities and multi-task learning potential offer promise for addressing the limitations of existing channel prediction methods [11]. Most recently, a few studies [12, 13] apply self-supervised pre-training schemes to wireless channel representation learning for CSI-related downstream tasks, such as fingerprint localization and cross-band feature extraction. Specifically, the pre-trained model is fine-tuned for specific wireless tasks through few-shot learning. Nevertheless, these cannot be directly adopted for channel prediction. First, these schemes can only handle space-frequency two-dimensional CSI and cannot address temporal variations. Secondly, they utilize the encoder-only network instead of an autoencoder, rendering these approaches suitable for classification or regression tasks but less effective for reconstruction tasks such as channel prediction. Last but not least, they can only reduce, but not eliminate, the retraining overhead for different tasks and scenarios.

Recognizing the limitations of existing studies, We make the pioneering attempt to apply foundation models to channel prediction tasks. Specifically, we aim to establish a large-scale wireless foundation model pre-trained on extensive CSI data, and directly apply it to various channel prediction tasks under different CSI configurations and scenarios without any fine-tuning. Evidently, building such a foundational model is quite challenging. First, it is challenging to address different types of channel prediction tasks with a single network. Secondly, it is also difficult to use a single network to simultaneously cope with diverse CSI data, where the number of antennas, sub-carriers, and time samples vary.

In order to overcome these challenges, we propose the first wireless foundation model for channel prediction tasks, termed as WiFo. Different from all existing works, we formulate time-domain and frequency-domain channel prediction tasks as a unified channel reconstruction problem, which captures the complete CSI based on partial CSI. Inspired by the masked autoencoders (MAE) for image and video self-supervised learning, we propose an MAE-based network structure suitable for CSI reconstruction. We apply 3D patching and embedding to convert diverse CSI into varying numbers of tokens, enabling efficient processing by the transformer blocks. For both the encoder and the decoder, we propose a novel positional encoding (STF-PE) structure to learn the 3D position information associated with the CSI. To capture the inherent 3D variations, we propose four self-supervised pre-training tasks, namely random, time, and frequency-masked reconstruction. The pre-trained model can be directly applied to zero-shot inference for both channel prediction tasks. To evaluate the performance of WiFo, we construct a heterogeneous CSI dataset with 16 different configurations of time sampling, sub-carrier, and antenna, containing 160K training samples. Preliminary results validate its superior multi-dataset unified learning performance and zero-shot generalization capability. It is also worth emphasizing that its zero-shot prediction performance in unseen scenarios surpasses the full-shot performance of all baselines trained on 10K samples, completely eliminating the retraining or fine-tuning costs.

The primary contributions of our work are summarized as follows.

• We propose the first wireless foundation model (WiFo) designed to uniformly facilitate time-domain and frequency-domain channel prediction tasks. To the best of our knowledge, it is the first one-for-all model capable of simultaneously tackling different types of channel prediction tasks and diverse CSI configurations.

• We develop an MAE-based network structure to cope with heterogeneous CSI data and introduce three mask reconstruction tasks for self-supervised pre-training, aimed at capturing the inherent space-time-frequency correlations of CSI. The pre-trained model can be directly utilized for inference without the need for fine-tuning.

• We construct 16 diverse datasets with various CSI configurations using the QuaDRiGa channel generator for pre-training, consisting of 160K training samples. Simulations confirm that WiFo can effectively learn across different channel prediction tasks and heterogeneous datasets and demonstrates strong zero-shot performance in new scenarios.

Notation: $\| \cdot \|_F$ denotes the Frobenius norm. $\boldsymbol{a}[i]$ is the i-th element of a vector $\boldsymbol{a}$ and $\boldsymbol{A}[i,j]$ denotes the element of matrix $\boldsymbol{A}$ at the i-th row and the j-th column. $\mathbb{R}$ and $\mathbb{C}$ denote the set of real numbers and complex numbers, respectively.

2 System Model and Problem Formulation

In this paper, we consider a MISO-OFDM system, where the BS and the user are equipped with a uniform planar array (UPA) and a single antenna, respectively. The UPA contains $N = N_h \times N_v$ elements, with N_h and N_v being the number of antennas along the horizontal and vertical directions, respectively.

2.1 3D Channel Model

We adopt the classical geometric channel model consisting of P paths. The CSI between the BS and the user sampled at frequency f and time t can be expressed as

$$\boldsymbol{h}(t,f) = \sum_{p=1}^{P} g_p \boldsymbol{a}(\phi_p, \theta_p) e^{-j2\pi f \tau_p} e^{j2\pi \nu_p t}, \tag{1}$$

where g_p, τ_p, and ν_p represent the complex amplitude, the delay, and the Doppler shift associated with the p-th path. $\boldsymbol{a}(\phi_p, \theta_p) \in \mathbb{C}^{N \times 1}$ is the steering vector of UPA [9], where ϕ_p and θ_p denote the azimuth and elevation angles, respectively.

Assume that the considered time-frequency region [9] spans T and K resource blocks (RBs) along the time and frequency dimensions, respectively. The pilot placement pattern is identical for all antennas and each RB contains a pilot. The pilot intervals along time and frequency are Δt and Δf, respectively. Without loss of generality, we consider the CSI only at the pilot positions. The considered space-time-frequency CSI is denoted as $\boldsymbol{H} \in \mathbb{C}^{T \times K \times N}$, which satisfies

$$\boldsymbol{H}[i,j,:] = \boldsymbol{h}(i\Delta t, f_1 + (j-1)\Delta f), \quad i = 1, \ldots, T, \quad j = 1, \ldots, K, \tag{2}$$

where f_1 represents the frequency at the pilot position of the first RB in the frequency domain.

2.2 Problem Description

We consider two types of channel prediction tasks: time-domain prediction and frequency-domain prediction.

1) **Time-domain channel prediction**: We aim to predict CSI for the future $T - T_h$ RBs based on the historical T_h RBs along the time dimension. Denote the mapping function of time-domain channel prediction as Φ_t, the prediction process is derived as

$$\boldsymbol{H}[T_h + 1 : T, :, :] = \Phi_t(\boldsymbol{H}[1 : T_h, :, :]). \tag{3}$$

2) **Frequency-domain channel prediction**: We focus on channel prediction for adjacent frequency bands. Without loss of generality, we aim to predict the last $K - K_u$ RBs via the first K_u RBs along the frequency dimension. Such an operation can be represented as

$$\boldsymbol{H}[:, K_u + 1 : K, :] = \Phi_f(\boldsymbol{H}[:, 1 : K_u, :]), \tag{4}$$

where Φ_f is the mapping function in the frequency-domain.

Previous studies have designed separate networks to manage individual prediction tasks and various CSI configurations. However, since a BS must simultaneously handle multiple channel prediction tasks and diverse user configurations, this approach leads to significant overhead in network deployment. To address this issue, we propose to unify the channel prediction tasks into a single channel reconstruction task using our wireless foundation model. Specifically, the formulated channel reconstruction task seeks to derive the complete CSI from partial CSI. We denote the partial CSI as $H[\Omega]$, where Ω represents the subset of all elements. Our goal is to develop a reconstruction function Φ_{rec} that facilitates a universal mapping as follows:

$$\boldsymbol{H} = \Phi_{\text{rec}}(\boldsymbol{H}[\Omega]), \tag{5}$$

where Ω represents the temporal or frequency subset for time-domain and frequency-domain channel prediction, respectively. It is worth noting that the function Φ_{rec} needs to handle diverse CSI data with arbitrary T, K, and N, while simultaneously performing reconstruction in the time or frequency domain. However, traditional deep learning-based channel prediction schemes and LLM-empowered schemes [9] are designed for fixed-size CSI and specific prediction tasks, which limits their ability to achieve this universal mapping. Consequently, we aim to establish a foundation model to address this challenge.
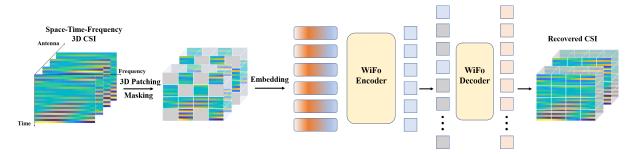
Figure 1 An illustration of the network structure of the proposed WiFo.

3 Wireless Foundation Model

In this section, a novel wireless foundation model, termed WiFo, is proposed to realize the universal channel prediction. First, the network structure of WiFo is introduced based on masked autoencoders (MAE). Then, a self-supervised pre-training scheme is proposed based on multiple reconstruction tasks. Finally, the pre-trained WiFo is directly applied for channel prediction without any fine-tuning.

3.1 Network Structure

MAE [14–16] has been demonstrated to be a data-efficient self-supervised pre-training framework for image and video representation learning. It adopts an asymmetric encoder-decoder architecture, where an encoder operates on unmasked tokens and a lightweight decoder recovers the original image or video. Inspired by MAE, WiFo adopts a transformer-based encoder-decoder network structure, as shown in Fig. 1. It consists of four major blocks, namely CSI embedding, masking, encoder, and decoder. Each of them will be described in detail below.

3.1.1 CSI Embedding

To facilitate neural network processing, the complex $\boldsymbol{H}$ is first converted into a real-valued tensor $\tilde{\boldsymbol{H}} \in \mathbb{R}^{2 \times T \times K \times N}$, which consists of two channels for the real and imaginary parts. To convert $\tilde{\boldsymbol{H}}$ into 1D sequential data suitable for transformer processing, we apply 3D patching [15] to the last three dimensions. Specifically, the size of each 3D patch is (t, k, n), where t, k, and n represent the patch size along the time, frequency, and space dimensions. Then the non-overlapping 3D patches, with a total number of $L = \frac{T}{t} \times \frac{K}{k} \times \frac{N}{n}$, are flattened and embedded into a series of tokens with dimension D_{enc}, where D_{enc} is the hidden size of the encoder. The above operation can be implemented using 3D convolution, i.e.,

$$\boldsymbol{H}_{\text{conv}} = Conv3d(\tilde{\boldsymbol{H}}), \tag{6}$$

where $Conv3d(\cdot)$ represents the 3D convolution operator with 2 input channels and D_{enc} output channels. Then the convolution result $\boldsymbol{H}_{\text{conv}} \in \mathbb{R}^{D_{\text{enc}} \times \frac{T}{t} \times \frac{K}{k} \times \frac{N}{n}}$ is flattened into $\boldsymbol{H}_{\text{emb}} \in \mathbb{R}^{D_{\text{enc}} \times L}$ as the CSI tokens.

3.1.2 Masking

As shown in Fig. 1, several 3D patches are masked. It is equivalent to mask partial tokens of $\boldsymbol{H}_{\text{emb}}$ and retains only a subset of tokens to be processed by the encoder module. Denote the visible tokens and the masked tokens as $\boldsymbol{H}_{\text{vis}} \in \mathbb{R}^{D_{\text{enc}} \times L_{\text{vis}}}$ and $\boldsymbol{H}_{\text{mask}} \in \mathbb{R}^{D_{\text{enc}} \times (L - L_{\text{vis}})}$, respectively, where L_{vis} is the number of visible tokens. Then the masking process is represented as

$$[\boldsymbol{H}_{\text{vis}}, \boldsymbol{H}_{\text{mask}}] = Mask(\boldsymbol{H}_{\text{emb}}), \tag{7}$$

where $Mask(\cdot)$ represents the masking operation with a certain masking strategy and masking ratio.

The role of the masking operation differs between the model pre-training and inference stages. During the pre-training stage, masking strategies are performed as self-supervised pre-training reconstruction tasks for better CSI representation learning. During the inference stage, masking strategies are performed to apply the proposed model to specific channel prediction tasks.
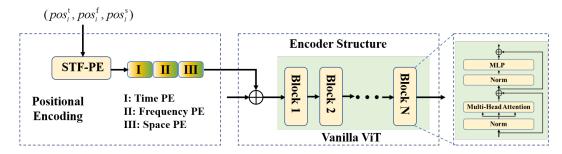
Figure 2 An illustration of the architecture of the encoder module.

3.1.3 *Encoder*

The architecture of the encoder is shown in Fig. 2. Consistent with MAE [14–16], the visible tokens are sequentially added with positional encoding and processed by a series of transformer blocks implemented by vanilla ViT backbone [17]. Denote the operation of the transformer blocks as f_{enc}, the encoder output $\boldsymbol{H}_{\mathrm{enc}} \in \mathbb{R}^{D_{\mathrm{enc}} \times L_{\mathrm{vis}}}$ is derived as

$$\boldsymbol{H}_{\mathrm{enc}} = f_{\mathrm{enc}}(\boldsymbol{H}_{\mathrm{vis}} + \boldsymbol{P}_{\mathrm{enc}}), \tag{8}$$

where $\boldsymbol{P}_{\mathrm{enc}} \in \mathbb{R}^{D_{\mathrm{enc}} \times L_{\mathrm{vis}}}$ denotes the positional encoding.

To enable the model to learn the three-dimensional position information of CSI, we propose an STF positional encoding (STF-PE). Specifically, separate positional encoding for time, frequency, and space, denoted as $\boldsymbol{P}_{\mathrm{enc}}^{\mathrm{t}} \in \mathbb{R}^{L_{\mathrm{vis}} \times D_{\mathrm{enc}}^{\mathrm{t}}}$, $\boldsymbol{P}_{\mathrm{enc}}^{\mathrm{f}} \in \mathbb{R}^{L_{\mathrm{vis}} \times D_{\mathrm{enc}}^{\mathrm{f}}}$, and $\boldsymbol{P}_{\mathrm{enc}}^{\mathrm{s}} \in \mathbb{R}^{L_{\mathrm{vis}} \times D_{\mathrm{enc}}^{\mathrm{s}}}$, are concatenated along the feature dimension, satisfying $D_{\mathrm{enc}} = D_{\mathrm{enc}}^{\mathrm{t}} + D_{\mathrm{enc}}^{\mathrm{f}} + D_{\mathrm{enc}}^{\mathrm{s}}$. Without loss of generality, let $D_{\mathrm{enc}}^{\mathrm{t}} = D_{\mathrm{enc}}^{\mathrm{f}} = \lfloor D_{\mathrm{enc}}/3 \rfloor$ and $D_{\mathrm{enc}}^{\mathrm{s}} = D_{\mathrm{enc}} - 2\lfloor D_{\mathrm{enc}}/3 \rfloor$, respectively. For better generalization across different sizes of CSI, each separate positional encoding of STF-PE adopts absolute SinCos positional encoding [14] instead of learnable encoding [16]. As an example, consider the time positional encoding. For i-th visible token, denote the temporal, spatial, and frequency coordinates of its corresponding 3D patch as $(pos_i^{\mathrm{t}}, pos_i^{\mathrm{f}}, pos_i^{\mathrm{s}})$, we have:

$$\boldsymbol{P}_{\mathrm{enc}}^{\mathrm{t}}[i, 2j] = \sin \frac{pos_i^{\mathrm{t}}}{10000^{\frac{2j}{D_{\mathrm{enc}}^{\mathrm{t}}}}}, \ \text{ and } \ \boldsymbol{P}_{\mathrm{enc}}^{\mathrm{t}}[i, 2j+1] = \cos \frac{pos_i^{\mathrm{t}}}{10000^{\frac{2j}{D_{\mathrm{enc}}^{\mathrm{t}}}}}. \tag{9}$$

3.1.4 *Decoder*

The decoder is designed to reconstruct the original $\boldsymbol{H}$ from the encoder output. $\boldsymbol{H}_{\mathrm{enc}}$ is first converted to $\bar{\boldsymbol{H}}_{\mathrm{enc}} \in \mathbb{R}^{D_{\mathrm{dec}} \times L_{\mathrm{vis}}}$ via a fully connected layer to align with the feature dimension D_{dec} of the decoder transformer blocks. $\bar{\boldsymbol{H}}_{\mathrm{enc}}$ is then concatenated with learnable mask tokens $\boldsymbol{M} \in \mathbb{R}^{D_{\mathrm{dec}} \times (L - L_{\mathrm{vis}})}$, and added with decoder positional encoding $\boldsymbol{P}_{\mathrm{dec}} \in \mathbb{R}^{D_{\mathrm{dec}} \times L}$ before processed by lightweight ViT transformer blocks, denoted as f_{dec}. $\boldsymbol{P}_{\mathrm{dec}}$ adopts the same STF-PE as the encoder. Denote the output of the decoder transformer blocks as $\boldsymbol{H}_{\mathrm{dec}} \in \mathbb{R}^{D_{\mathrm{dec}} \times L}$, and we have
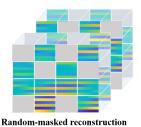
$$\boldsymbol{H}_{\mathrm{dec}} = f_{\mathrm{dec}}([\boldsymbol{H}_{\mathrm{vis}}, \boldsymbol{M}] + \boldsymbol{P}_{\mathrm{dec}}). \tag{10}$$
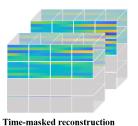
Then, $\boldsymbol{H}_{\mathrm{dec}}$ is transformed into $\tilde{\boldsymbol{H}}_{\mathrm{pred}} \in \mathbb{R}^{2 \times T \times K \times N}$ by a fully connected layer and reshape operation. Finally, we obtain the reconstructed complex CSI $\boldsymbol{H}_{\mathrm{pred}} \in \mathbb{C}^{T \times K \times N}$, i.e.,

$$\boldsymbol{H}_{\mathrm{pred}} = \tilde{\boldsymbol{H}}_{\mathrm{pred}}[1, :, :, :] + 1j \times \tilde{\boldsymbol{H}}_{\mathrm{pred}}[2, :, :, :]. \tag{11}$$

3.2 Self-Supervised Pre-Training

Unlike existing channel and prediction schemes [5–7, 9] which are trained and tested on a specific dataset, the proposed WiFo is first pre-trained across multiple heterogeneous datasets and then directly applied to unseen scenarios and new CSI configurations. Specifically, several self-supervised training tasks are designed to capture the intricate inherent space-time-frequency correlations of CSI. Masked reconstruction [14–16] has been proven to be an effective pre-training task for downstream visual tasks, which masks random patches of an image and then reconstructs the complete image. Notably, temporal-domain and
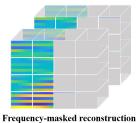
Random-masked reconstruction   **Time-masked reconstruction**   **Frequency-masked reconstruction**

Figure 3   An illustration of the proposed masked reconstruction tasks.

frequency-domain channel prediction are special types of masked reconstruction tasks, where patches are masked along the time or frequency dimensions. Therefore, we propose three masked reconstruction tasks to capture the intricate inherent space-time-frequency correlations of CSI, as shown in Fig. 3. Their detailed operations are introduced as follows. To simplify the description, we denote the temporal, spatial, and frequency coordinates of the 3D patch corresponding to the i-th token in $\boldsymbol{H}_{\text{emb}}$ as $(\overline{pos}_i^{\text{t}}, \overline{pos}_i^{\text{f}}, \overline{pos}_i^{\text{s}})$.

1) **Random-masked reconstruction**: It randomly masks tokens from all tokens with a R_{r} ratio. This masking strategy is isotropic across the space-time-frequency dimensions to capture the 3D structured features.

2) **Time-masked reconstruction**: It is designed to enhance the time-domain channel prediction task. In this masking strategy, future tokens are masked with a certain ratio R_{t}, i.e., all tokens with $\overline{pos}_i^{\text{t}} \geqslant (\frac{T}{t} - \lfloor R_{\text{t}} \frac{T}{t} \rfloor)$ are masked. It helps the model learn the causal relationships of CSI over time.

3) **Frequency-masked reconstruction**: This task is designed to improve Wifo's performance in frequency-domain channel prediction. Unlike the time-masked reconstruction, frequency masking strategy masks tokens along frequency dimension with a ratio R_{f}, so all tokens with $\overline{pos}_i^{\text{f}} \geqslant (\frac{K}{k} - \lfloor R_{\text{f}} \frac{K}{k} \rfloor)$ are masked. This helps the model learn the variations between adjacent frequency bands.

During self-supervised pre-training, the three pre-training tasks mentioned above are executed sequentially for each batch. Our objective is to minimize the reconstruction error, for which we adopt the mean squared error (MSE) as the loss function. The loss is computed using only the reconstructed CSI points and is defined as

$$\mathcal{L} = \frac{1}{|\omega|} \|\boldsymbol{H}[\omega] - \boldsymbol{H}_{\text{pred}}[\omega]\|_F^2, \tag{12}$$

where ω represents the temporal or frequency predicted CSI subset.

3.3   Model Inference

Once the pre-training stage is complete, the pre-trained WiFo can perform zero-shot inference on both the time-domain and frequency-domain tasks, as shown in Eq. 5. Specifically, the portion to be predicted is first zero-padded, and then the completed CSI is input into WiFo. Subsequently, the corresponding time-masked or frequency-masked strategy is applied with a certain masking ratio, allowing the network to output the reconstructed part as the predicted CSI.

4   Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of the proposed WiFo. First, we built a series of simulated datasets covering various system configurations for network pre-training and inference. Then, experimental setups are introduced, including network and training parameters, baselines, and the performance metric. Finally, the channel prediction performance of the proposed WiFo is comprehensively evaluated.

4.1   Datasets

To fully unleash WiFo's prediction capabilities across various CSI configurations, we have constructed a series of diverse 3D CSI datasets, generated through channel generator QuaDRiGa [18] compliant with the 3GPP standards. Consistent with our system model in Section 2.1, we consider MISO-OFDM systems, where the BS is equipped with a UPA and the user has a single antenna. The adjacent antenna spacing

Table 1   An illustration of the system configurations of the constructed 3D CSI datasets.

Dataset	f_C(GHz)	K	Δf(kHz)	T	Δt(ms)	UPA	Scenario	User speed(km/h)
D1	1.5	128	90	24	1	1×4	UMi+NLoS	3-50
D2	1.5	128	180	24	0.5	2×4	RMa+NLoS	120-300
D3	1.5	64	90	16	1	1×8	Indoor+LoS	0-10
D4	1.5	32	180	16	0.5	4×8	UMa+LoS	30-100
D5	2.5	64	180	24	0.5	2×2	RMa+NLoS	120-300
D6	2.5	128	90	24	1	2×4	UMi+LoS	3-50
D7	2.5	32	360	16	0.5	4×8	UMa+LoS	30-100
D8	2.5	64	90	16	1	4×4	Indoor+NLoS	0-10
D9	4.9	128	180	24	1	1×4	UMi+NLoS	3-50
D10	4.9	64	180	24	0.5	2×4	RMa+LoS	120-300
D11	4.9	64	90	16	0.5	4×4	UMa+NLoS	30-100
D12	4.9	32	180	16	1	4×8	Indoor+LoS	0-10
D13	5.9	64	90	24	0.5	2×8	RMa+LoS	120-300
D14	5.9	128	180	24	1	2×4	UMi+NLoS	3-50
D15	5.9	64	90	16	1	4×4	Indoor+LoS	0-10
D16	5.9	32	360	16	0.5	4×8	UMa+NLoS	30-100
D17	3.5	32	180	16	0.5	4×8	UMa+NLoS	30-100
D18	6.7	64	180	24	1	4×4	UMi+LoS	3-50

is half the wavelength at the central frequency. A total of 18 datasets are simulated, termed D1 to D18, covering various space-time-frequency CSI configurations, scenarios, and user speeds. Among them, the first 16 datasets are used for pre-training, and the last two datasets are used for generalization testing. The detailed simulation configurations of each dataset are shown in Table 1, where f_C represents the center frequency. There are six 5G New Radio (NR) frequency bands, eight 3GPP [19] scenarios, and seven user speed ranges considered. For each CSI sample, The user has a random initial position and a straight-line motion trajectory, where the speed is uniformly selected within the corresponding speed range. Each dataset contains 12000 samples, which are randomly split into 9000, 1000, and 2000 samples for training, validation, and inference, respectively. All CSI samples are pre-standardized using the mean and variance of the corresponding dataset. To simulate the imperfect factors of practical CSI acquisition, complex Gaussian noise with 20dB is added to the CSI samples during both the training and inference process.

4.2   Experimental Settings

We consider both the time-domain and the frequency-domain channel prediction tasks. For time-domain channel prediction, we predict the CSI of the future $\frac{T}{2}$ RBs according to historical $\frac{T}{2}$ RBs along the time dimension. For frequency-domain channel prediction, we predict the CSI of the last $\frac{K}{2}$ RBs using the first $\frac{K}{2}$ RBs as input along the frequency dimension.

4.2.1   *Network and Pre-training Settings*

To investigate the impact of model size on performance, we consider WiFo of 5 different sizes, with the specific parameters listed in Table 2. In the experiments, we set the size of each 3D patch as $(4, 4, 4)$. In addition, we set the masking ratio of random masking, time-domain masking, and frequency-domain masking as $R_r = 85\%$, $R_t = 50\%$, and $R_f = 50\%$, respectively.

We conduct all experiments on the same machine with 4 NVIDIA GeForce RTX4090 GPUs, AMD EPYC 7763 64-Core CPU, and 256 GB of RAM. WiFo and other baselines are trained with TF32 precision. The pre-training settings are illustrated in Table 3. During the pre-training process, the 16 datasets are split into batches with a certain batch size and shuffled. For each batch, the proposed three masked reconstruction tasks are applied sequentially. The final loss value for gradient descent is the mean loss of the three reconstruction tasks.

Table 2   Network parameters of WiFo with different sizes.

Model	Enc. depth	Enc. width	Enc. heads	Dec. depth	Dec. width	Dec. heads	Parameters
WiFo-Tiny	6	64	8	4	64	8	0.3M
WiFo-Little	6	128	8	4	128	8	1.4M
WiFo-Small	6	256	8	4	256	8	5.5M
WiFo-Base	6	512	8	4	512	8	21.6M
WiFo-Large	8	768	8	4	768	8	86.1M

Table 3   Pre-training parameters of WiFo.

Parameter	Value
Optimizer	AdamW ($\beta_1 = 0.9, \beta_1 = 0.999$, weight decay=0.05)
Batch size	128
Epochs	200
Learning rate schedule	Cosine decay [22] (warmup epochs = 5)
Base learning rate	5×10^{-4}

4.2.2   *Baselines*

For a comprehensive comparison, we provide the following five baselines, covering model-based, traditional deep learning-based, and advanced LLM-powered methods.

1) **PAD** [3]: PAD is a Prony-based angular-delay domain channel prediction method. It is only applied to time-domain prediction tasks. In our experiments, the predictor order is set as $N = 4$ and 6 for the case of $T = 16$ and 24, respectively.

2) **Transformer** [5]: The transformer-based prediction scheme is proposed in [5] for parallel channel prediction. For a fair comparison, it adopts the same CSI embedding method as WiFo. Specifically, 3D patching and embedding are first applied to the original CSI, and the embedded tokens are processed by the network. The feature dimension is set as 128, while the depth of the encoder and decoder are set as 5 and 8, respectively.

3) **LSTM** [21]: Long short-term memory (LSTM) is proposed for sequential processing to overcome the vanishing gradient problem. In our experiments, we consider a two-layer LSTM. For time-domain channel prediction, antenna and frequency dimensions are flattened and input into the network. For frequency-domain channel prediction, time and antenna dimensions are flattened similarly.

4) **3D ResNet** [20]: As a network specifically designed for handling 3D data, a ResNet-style model [20] proposed for video recognition tasks is considered for comparison. It consists of 50 weighted layers to capture the three-dimensional relationships within the CSI.

5) **LLM4CP** [9]: LLM4CP is a LLM-empowered channel prediction scheme, where GPT-2 is fine-tuned for cross-modality knowledge transfer. Since the original LLM4CP method cannot directly handle 3D CSI, we consider two implementation approaches. To ensure a fair comparison, the first approach considers the same 3D patching method as WiFo, termed LLM4CP. In addition, the second implementation approach adopts antenna parallel processing [9], termed LLM4CP*.

4.2.3   *Performance Metric*

In our experiments, normalized mean square error (NMSE) is adopted to measure the prediction accuracy directly. Let $\boldsymbol{H}_{\mathrm{P}}$ and $\boldsymbol{H}_{\mathrm{GT}}$ denote the predicted part of the CSI and its corresponding ground truth, respectively. The performance metric NMSE is derived as

$$\mathrm{NMSE}(\boldsymbol{H}_{\mathrm{P}}, \boldsymbol{H}_{\mathrm{GT}}) = \frac{\|\boldsymbol{H}_{\mathrm{P}} - \boldsymbol{H}_{\mathrm{GT}}\|_F^2}{\|\boldsymbol{H}_{\mathrm{GT}}\|_F^2}. \tag{13}$$
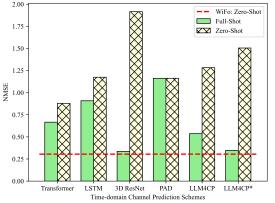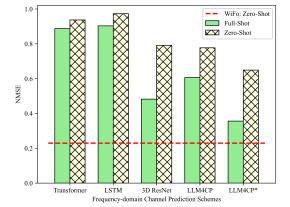
4.3   Performance Evaluation

4.3.1   *Multi-Dataset Unified Learning*

To validate the multi-dataset unified learning capability, we first evaluate the time-domain and frequency-domain prediction performance of WiFo across the 16 pre-training datasets. WiFo is pre-trained on all training and validation samples of Dataset D1 to D16, with a total of 160k diverse CSI samples. Then, the

Table 4 The NMSE performance of WiFo-Base and other baselines on the time-domain channel prediction task across the D1-D16 datasets. The best results are highlighted in **bold**, while the second-best results are <u>underlined</u>.

Dataset	WiFo-Base	Transformer	LSTM	3D ResNet	PAD	LLM4CP	LLM4CP*
D1	<u>0.082</u>	0.112	0.356	0.088	0.529	0.117	**0.074**
D2	**0.260**	0.416	0.797	0.351	1.074	0.451	<u>0.305</u>
D3	0.016	0.016	0.027	<u>0.014</u>	0.038	0.015	**0.013**
D4	**0.048**	0.107	0.418	0.055	0.317	0.106	<u>0.060</u>
D5	**0.494**	0.638	0.788	0.751	5.008	0.637	<u>0.510</u>
D6	**0.095**	0.174	0.542	0.157	0.568	0.206	<u>0.133</u>
D7	**0.081**	0.219	0.576	<u>0.103</u>	0.617	0.198	0.112
D8	<u>0.018</u>	0.024	0.092	**0.016**	0.073	0.025	**0.016**
D9	0.347	0.483	0.835	<u>0.349</u>	1.087	0.475	**0.312**
D10	**0.467**	0.649	0.689	0.869	3.863	0.709	<u>0.563</u>
D11	**0.227**	0.440	0.834	<u>0.274</u>	1.017	0.405	<u>0.273</u>
D12	**0.023**	0.035	0.166	<u>0.025</u>	0.132	0.035	0.026
D13	**0.482**	0.718	0.876	0.815	5.213	0.758	<u>0.648</u>
D14	<u>0.369</u>	0.546	0.884	0.388	1.021	0.562	**0.358**
D15	**0.029**	0.039	0.156	0.032	0.151	0.038	<u>0.030</u>
D16	**0.318**	0.591	0.944	0.329	1.034	0.545	<u>0.349</u>
Average	**0.210**	0.325	0.561	0.289	1.359	0.330	<u>0.236</u>

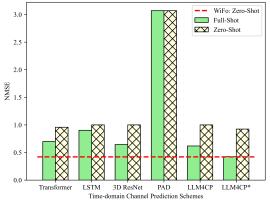(a) The time-domain channel prediction performance.

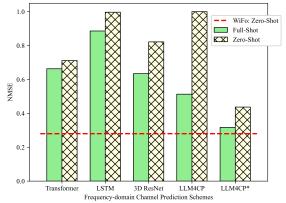(b) The frequency-domain channel prediction performance.

Figure 4 The zero-shot performance of WiFo and the full shot/zero-shot performance of other baselines on the D17 dataset.

pre-trained WiFo is evaluated on test samples of these 16 datasets. In contrast, for deep learning-based approaches, the network is individually trained and tested on each dataset for either time-domain or frequency-domain prediction tasks. The NMSE performance on the time-domain and frequency-domain task is illustrated in Table 4 and Table 5, respectively. Due to space limitations, only the performance of the base version of WiFo on each dataset is presented here, and the impact of the model size on performance is analyzed in Section 4.3.4. For both the time-domain and frequency-domain prediction tasks, WiFo achieves the SOTA average NMSE performance and shows the best or the second-best prediction performance on most datasets. The results show that WiFo can effectively perform unified learning across multiple datasets with different CSI configurations, while simultaneously mastering both time-domain and frequency-domain prediction capabilities. It is worth noting that WiFo is a one-for-all model, whereas other deep learning-based approaches require training 32 separate models for the above tasks, which have additional significant training and deployment overhead.

Table 5 The NMSE performance of WiFo-Base and other baselines on the frequency-domain channel prediction task across the D1-D16 datasets. The best results are highlighted in **bold**, while the second-best results are underlined.

Dataset	WiFo-Base	Transformer	LSTM	3D ResNet	LLM4CP	LLM4CP*
D1	**0.318**	0.532	0.705	0.839	0.392	0.375
D2	**0.181**	0.556	0.763	0.647	0.419	0.223
D3	0.027	**0.016**	0.037	0.071	0.023	0.025
D4	**0.073**	0.270	0.475	0.215	0.211	0.151
D5	**0.152**	0.315	0.577	0.386	0.267	0.165
D6	**0.081**	0.310	0.540	0.458	0.193	0.140
D7	**0.092**	0.392	0.578	0.354	0.318	0.189
D8	0.061	**0.024**	0.348	0.139	0.068	0.069
D9	0.436	0.481	0.895	0.918	0.574	**0.418**
D10	**0.087**	0.261	0.451	0.257	0.163	0.096
D11	**0.245**	0.723	0.859	0.823	0.621	0.349
D12	**0.023**	0.048	0.131	0.029	0.032	0.026
D13	0.068	0.238	0.531	0.177	0.165	**0.067**
D14	**0.395**	0.744	0.911	0.924	0.637	0.414
D15	**0.023**	0.053	0.083	0.045	0.024	0.024
D16	**0.270**	0.855	0.929	0.723	0.712	0.456
Average	**0.158**	0.364	0.551	0.438	0.301	0.199

(a) The time-domain channel prediction performance.

(b) The frequency-domain channel prediction performance.

Figure 5 The zero-shot performance of WiFo and the full shot/zero-shot performance of other baselines on the D18 dataset.

4.3.2 *Zero-shot Generalization*

To evaluate the generalization ability of WiFo, its zero-shot prediction capability is evaluated. Specifically, the pre-trained WiFo performs inference directly on unseen datasets without any fine-tuning. The D17 and D18 datasets are used for zero-shot testing because their operating carrier frequencies are not included in the training set. For other deep learning-based baselines, we consider both zero-shot and full-shot learning scenarios. For zero-shot learning, given that these methods struggle to generalize across CSI with varying shapes, they are trained on the D7 and D13 datasets and then perform inference on the D17 and D18 datasets, respectively. For full-shot learning, these methods are trained and tested on the same dataset. The performance of WiFo-Base and other baselines on the D17 and D18 datasets is illustrated in Fig. 4 and Fig. 5, respectively.

It is observed that most deep learning-based methods struggle to perform zero-shot inference. Additionally, the transformer and LSTM show poor full-shot performance and even fail to learn, indicating that prediction on D17 and D18 datasets is highly challenging. This is because the two datasets have large sub-carrier intervals and a large number of antennas, making it difficult for these models to learn the 3D variations of the CSI. Nevertheless, the zero-shot performance of WiFo outperforms the zero-shot

Table 6 Results of ablation experiments. The best results are highlighted in **bold**, while the second-best results are underlined.

	Time-domain prediction			Frequency-domain prediction			
	D1-D16	D17	D18	D1-D16	D17	D18	**Average**
WiFo-Base	0.210	**0.305**	0.420	0.158	0.229	0.280	**0.267**
w/ learnable PE [16]	0.224	0.354	0.442	0.154	0.220	0.267	0.277
w/o random-masked reconstruction	0.214	0.317	0.435	0.165	0.233	0.294	0.276
w/o time-masked reconstruction	0.497	0.754	0.723	**0.145**	**0.199**	**0.257**	0.429
w/o frequency-masked reconstruction	**0.205**	0.310	**0.412**	0.472	0.819	0.656	0.479

Table 7 The performance of WiFo is evaluated across different model sizes and various pre-training dataset scales. The best results are highlighted in **bold**, while the second-best results are underlined.

Pre-training Dataset	Model	Time-domain Prediction			Frequency-domain Prediction		
		D1,D5,D9,D15	D17	D18	D1,D5,D9,D15	D17	D18
	WiFo-Tiny	0.315	0.444	0.506	0.343	0.440	0.399
	WiFo-Little	0.271	0.371	0.446	0.260	0.284	0.301
D1-D16	WiFo-Small	0.245	0.326	0.421	0.239	0.245	0.299
	WiFo-Base	0.237	**0.305**	0.420	0.232	**0.229**	0.280
	WiFo-Large	**0.234**	0.314	**0.416**	**0.226**	0.232	**0.276**
D1,D2,D5,D6,D9,D12,D15,D16	WiFo-Base	0.261	0.356	0.477	0.258	0.269	0.345
D1,D5,D9,D15	WiFo-Base	0.306	0.960	0.636	0.320	1.036	0.523

and even full-shot performance of all other methods, demonstrating its remarkable cross-frequency generalization ability. Therefore, once WiFo is trained on large-scale datasets, it can potentially be deployed instantly at the BS, significantly reducing the costs associated with data collection and model fine-tuning.

4.3.3 *Ablation Study*

To verify the effectiveness of the specialized designs in WiFo, we perform ablation studies based on the WiFo-Base model. The ablation results are shown in Table 6, where the corresponding multi-dataset unified learning and zero-shot generalization performance are given. The average performance is measured by the mean NMSE of columns 2 to 7, representing overall performance. Replacing the proposed STF-PE with learnable space-time positional encoding designed for video would degrade overall performance. Additionally, removing the random-masked reconstruction task degrades the unified learning and generalization performance for both the time-domain and frequency-domain channel prediction, highlighting its effectiveness for self-supervised pre-training. It can be attributed that the additional random masking strategy facilitates WiFo's ability to capture the intrinsic 3D relationships of CSI. Moreover, removing time-masked or frequency-masked reconstruction enhances frequency-domain or time-domain channel prediction performance but significantly degrades the other. Therefore, both time-masked and frequency-masked reconstruction tasks are essential for the model pre-training to address both prediction tasks simultaneously.

4.3.4 *Scaling Analysis*

Scaling analysis is essential for foundation models as it highlights the effects of model size and pre-training dataset scale on performance. In our scaling experiments, we consider five model sizes and three dataset scales to evaluate their overall performance. We assess unified learning performance using the average prediction NMSE across the D1, D5, D9, and D15 datasets, and evaluate zero-shot performance on the D17 and D18 datasets, as presented in Table 7.

From the first five rows of Table 7, we observe that, with a fixed pre-training dataset scale, increasing the model size generally enhances both the unified learning performance and zero-shot generalization capabilities within the observed range. This improvement is attributed to larger models' ability to capture more complex patterns, resulting in better learning. However, once the model reaches a certain size, its zero-shot generalization ability plateaus, constrained by the scale of the pre-training dataset.

Conversely, as seen in rows 4, 6, and 7, expanding the pre-training dataset scale significantly boosts both unified learning performance and zero-shot performance simultaneously. In summary, the unified learning performance and generalization performance of WiFo are improved [23] with both the increased

Table 8 Network parameters and inference time per batch.

	WiFo-Base	Transformer	LSTM	3D ResNet	PAD	LLM4CP	LLM4CP*
Parameters(M)	21.60	0.91	1.13	31.73	0	82.35	83.32
Inference time(ms)	9.659	6.238	5.209	86.865	119.397	4.718	7.342

model size and the scale of the pre-training datasets. Therefore, WiFo is a scalable model that has the potential for further improvement as computational power and dataset scale increase.

4.3.5 *Network Storage and Inference Cost*

The parameters and inference time of models are closely tied to their storage and computational overhead, directly impacting the practical deployment of channel prediction models at the BS. The number of parameters and the inference time per batch of these models are shown in Table 8, where the batch size is set as 8 and inference samples are taken from the dataset D1. As a model-based method, PAD has a lower storage cost but a longer inference time due to the matrix inversion and autoregressive prediction process. For 3D ResNet, complex 3D convolution operations and the large number of parameters result in significant inference overhead. For LLM4CP, its antenna-parallelized version has a higher inference time because the number of CSI samples processed per batch increases proportionally with the number of antennas. It is observed that WiFo has an acceptable parameter count and comparable inference time among these channel prediction schemes. Considering that WiFo is a versatile model capable of replacing multiple specialized channel prediction models at the BS, it offers significant advantages in practical deployment.

5   Conclusions

In this paper, we have introduced a novel wireless foundation model, WiFo, designed to simultaneously facilitate time-domain and frequency-domain channel prediction tasks as well as diverse 3D CSI configurations. We developed an MAE-based network structure and implemented several mask reconstruction tasks for self-supervised pre-training to capture the intrinsic space-time-frequency features of CSI. WiFo, pre-trained on large-scale heterogeneous datasets, can be directly deployed for inference without any need for fine-tuning. Simulations demonstrate its exceptional performance in unified multi-dataset training and its superb zero-shot generalization capabilities with reasonable inference overhead.

References

1   CHENG X, ZHANG H, ZHANG J, et al. Intelligent multi-modal sensing-communication integration: synesthesia of machines. IEEE Commun Surveys Tuts, 2024, 26: 258-301

2   Gao S, Cheng X, Yang L. Estimating doubly-selective channels for hybrid mmWave massive MIMO systems: A doubly-sparse approach. IEEE Trans Wirel Commun, 2020, 19: 5703-5715.

3   YIN H, WANG H, LIU Y, et al. Addressing the curse of mobility in massive MIMO with prony-based angular-delay domain channel predictions. IEEE J Select Areas Commun, 2020, 38: 2903-2917

4   ROTTENBERG F, CHOI T, LUO P, et al. Performance analysis of channel extrapolation in FDD massive MIMO systems. IEEE Trans Wirel Commun, 2020, 19: 2728-2741

5   JIANG H, CUI M, NG D W K, et al. Accurate channel prediction based on transformer: making mobility negligible. IEEE J Select Areas Commun, 2022, 40: 2717-2732

6   LIU G, HU Z, WANG L, et al. Spatio-temporal neural network for channel prediction in massive MIMO-OFDM systems. IEEE Trans Commun, 2022, 70: 8003-8016

7   Liu Z, Singh G, Xu C, et al. FIRE: enabling reciprocity for FDD MIMO systems. In: Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, New York, 2021. 628-641

8   Zhang H, Gao S, Cheng X, et al. Integrated sensing and communications towards proactive beamforming in mmWave V2I via multi-modal feature fusion (MMFF). IEEE Trans Wirel Commun, 2024, 23: 15721-15735

9   Liu B, Liu X, Gao S, et al. LLM4CP: Adapting Large Language Models for Channel Prediction. J Commun Inf Netw, 2024, 9: 113-125

10   Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models. 2021. ArXiv:2108.07258

11   Fontaine J, Shahid A, De Poorter E. Towards a Wireless Physical-Layer Foundation Model: Challenges and Strategies. 2024. ArXiv: 2403.12065

12   Salihu A, Rupp M, Schwarz S. Self-Supervised and Invariant Representations for Wireless Localization. IEEE Trans Wirel Commun, 2024, 23: 8281-8296

13   Alikhani S, Charan G, Alkhateeb A. Large Wireless Model (LWM): A Foundation Model for Wireless Channels. 2024. ArXiv: 2411.08872

14   He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, 2022. 16000-16009

15   Tong Z, Song Y, Wang J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in neural information processing systems, New Orleans,2022. 10078-10093

16   Feichtenhofer C, Li Y, He K. Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems, 2022, 35: 35946-35958.

17   Alexey D. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. ArXiv: 2010.11929

18 Jaeckel S, Raschkowski L, Börner K, et al. QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials. IEEE Trans Antennas Propagat, 2014, 62: 3242-3256.
19 3GPP Radio Access Network Working Group. Study on channel model for frequencies from 0.5 to 100 GHz (Release 15). 3GPP TR 38.901, 2018.
20 Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, Seoul, 2019. 6202-6211.
21 Jiang W, Schotten H D. Deep learning for fading channel prediction. IEEE Open J Commun Soc, 2020, 1: 320-332.
22 Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. 2016. ArXiv: 1608.03983
23 Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. 2020. ArXiv: 2001.08361