Foundation Model for Intelligent Wireless Communications

Boxun Liu[1], Xuanyu Liu[1], Shijian Gao[2], Xiang Cheng[1*], Liuqing Yang[2,3]

[1*]State Key Laboratory of Photonics and Communications, School of Electronics, Peking University, Beijing, 100871, China.
[2]Internet of Things Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511400, China.
[3]Intelligent Transportation Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511400, China.

*Corresponding author(s). E-mail(s): xiangcheng@pku.edu.cn;
Contributing authors: boxunliu@stu.pku.edu.cn;
xyliu25@stu.pku.edu.cn; shijiangao@hkust-gz.edu.cn; lqyang@ust.hk;

Abstract

The evolution toward intelligent next-generation wireless systems promises unprecedented spectral efficiency and reliability but is hindered by a paradigm of narrow and data-hungry AI models. Breaking from this constraint, this work introduces WiFo-2, a revolutionary wireless foundation model that establishes a new state of the art for extensive channel state information (CSI)-based tasks. Uniquely architected as a sparse mixture of experts, WiFo-2 effectively manages heterogeneous data and tasks while enabling highly efficient inference. It is pre-trained on a massive and diverse dataset of 11.6 billion CSI points, which enables the acquisition of profound and generalizable channel knowledge. WiFo-2 demonstrates remarkable zero-shot capabilities, not only matching but surpassing the full-shot performance of task-specific baselines on unseen configurations, all while providing reliable confidence estimates. Furthermore, the model achieves exceptional performance on eight key downstream tasks with minimal fine-tuning. A functional hardware prototype demonstrates its real-world deployment feasibility and significant system gains, highlighting WiFo-2's superiority and paving the way for a paradigm shift in AI-based wireless systems.

1 Introduction

In the contemporary digitized age, wireless communication systems have become indispensable, forming the backbone of both the mobile Internet and the Internet of Things (IoT) [1]. Traditional mobile systems, including the fifth-generation (5G) and its predecessors, rely heavily on parametric algorithms crafted for each physical layer module, which are based on intricate wireless channel models. However, real-world wireless propagation environments are filled with numerous complex factors that are challenging to model precisely, thereby limiting the effectiveness of these parametric models. With its robust data-driven modeling capabilities, deep learning presents a promising toolkit for 6G [2, 3], designed to overcome these challenges and accommodate broader and more intricate communication demands. In recent decades, deep learning has been applied to various physical-layer modules [4], such as channel state information (CSI) estimation and prediction [5], and automatic modulation recognition [6], and has demonstrated improvements in transmission efficiency in both theory and practice. Nevertheless, existing AI-based physical-layer designs are predominantly task-specific models tailored for particular modules and system configurations. Since practical systems typically demand diverse configurations (e.g., antennas and subcarriers) and a wide range of physical-layer modules, they require numerous models and substantially increase storage, computational, and management overhead. Moreover, current AI-based physical-layer models are trained for specific scenarios with limited generalization. Adapting to new environments typically demands extensive labeled data and retraining, hindering rapid deployment in dynamic settings.

Foundation models have reshaped the landscape of AI [7–11], offering new avenues to address the above challenges. Specifically, pretrained with self-supervision on massive datasets, foundation models [12–15] adapt to diverse downstream tasks through fine-tuning or zero-shot inference, consistently surpassing task-specific models. Given the central role of CSI in the physical layer, a wireless foundation model is expected to serve as a versatile model for all CSI-related tasks, thereby greatly reducing the number of required models. Pretrained on massive CSI data, it acquires general wireless representations and generalizes across heterogeneous configurations and diverse CSI distributions, showing strong zero-shot and few-shot performance even beyond task-specific models. Moreover, the wireless foundation model should be sufficiently computationally efficient for deployment on computational resource-constrained wireless devices.

A few recent studies [13, 16–20] have preliminarily validated the effectiveness of the foundation model paradigm in wireless physical-layer design. Despite this, related studies remain nascent, and several key factors limit their practical utility. First, existing pretrained CSI datasets suffer from limited scale, narrow sources, insufficient system diversity, and a lack of open access, which constrain the performance ceiling of wireless foundation models. Second, prior studies have examined only the few-shot fine-tuning performance of wireless foundation models across a limited set of CSI-related tasks, leaving their zero-shot reasoning capabilities largely unexplored. While our earlier study, WiFo [13], first demonstrates strong zero-shot performance on channel prediction, it still does not provide unified and reliable zero-shot channel reconstruction,

including channel estimation. In addition, existing studies on wireless foundation models are confined to simulation-based validation of single physical-layer tasks, lacking evaluation of link-level performance gains in practical wireless hardware platforms.

This work presents WiFo-2, our next-generation wireless foundation model (WiFo), which is the first to enable unified zero-shot channel reconstruction while supporting the broadest set of downstream wireless tasks to date through few-shot fine-tuning. We have developed the largest open-source heterogeneous 3D space–time–frequency CSI dataset (LH-CSI), consisting of 11.6 billion CSI points from 8 different data sources, for pretraining. WiFo-2 utilizes a masked denoising autoencoder (MDAE) and a two-stage pretraining regimen comprising mixed masking and denoising pretraining and confidence-enhanced pretraining, resulting in generalizable CSI representations and reliable zero-shot channel reconstruction. Moreover, it incorporates a CSI-sparse mixture of experts (CSI-SMoE) design to effectively address task and data heterogeneity while substantially reducing computational overhead, thereby enabling deployment on resource-constrained communication devices. A wireless foundation model evaluation (WFME) benchmark was constructed to comprehensively evaluate its zero-shot performance on time-domain channel prediction, frequency-domain channel prediction, and channel estimation tasks, as well as few-shot fine-tuning performance on 8 CSI-related downstream tasks across 10 constructed datasets. Across diverse reconstruction ratios and signal-to-noise ratios, WiFo-2 exhibits superior zero-shot channel reconstruction across diverse system configurations and unseen CSI distributions, consistently surpassing the full-shot performance of task-specific models. It also provides accurate predictions of the reconstruction NMSE, with a mean absolute prediction error below 2.8 dB. For the downstream tasks, in terms of both task accuracy metrics and communication performance metrics, WiFo-2 demonstrates significantly superior few-shot fine-tuning performance compared to existing wireless foundation models, even surpassing the full-shot performance of task-specific baselines. In addition, we have developed the first wireless foundation model-empowered hardware validation prototype for the over-the-air experiment, verifying both the performance advantages and deployment feasibility of WiFo-2 in practical communication systems. Overall, WiFo-2 is vital in catalyzing a paradigm shift in wireless communication system design.

2 Results

2.1 WiFo-2 possesses general CSI knowledge via pretraining

We introduce WiFo-2, a wireless foundation model developed for both channel reconstruction and diverse CSI-related tasks. To fully exploit its generalizable CSI representation capability, we curate LH-CSI, an unprecedented heterogeneous space-time-frequency 3D CSI dataset (Fig. 1a). LH-CSI comprises 78 sub-datasets, either coarse-grained or fine-grained, supporting tasks at different time–frequency sampling granularities. LH-CSI contains 11.6 billion CSI points drawn from 8 distinct sources [21–27] , covering all three acquisition modalities, namely statistical channel modeling, real-world measurements, and ray-tracing. It offers rich system variability across carrier frequencies from sub-6 GHz to terahertz, antenna configurations from single-antenna to massive multiple-input multiple-output (MIMO), and subcarrier counts

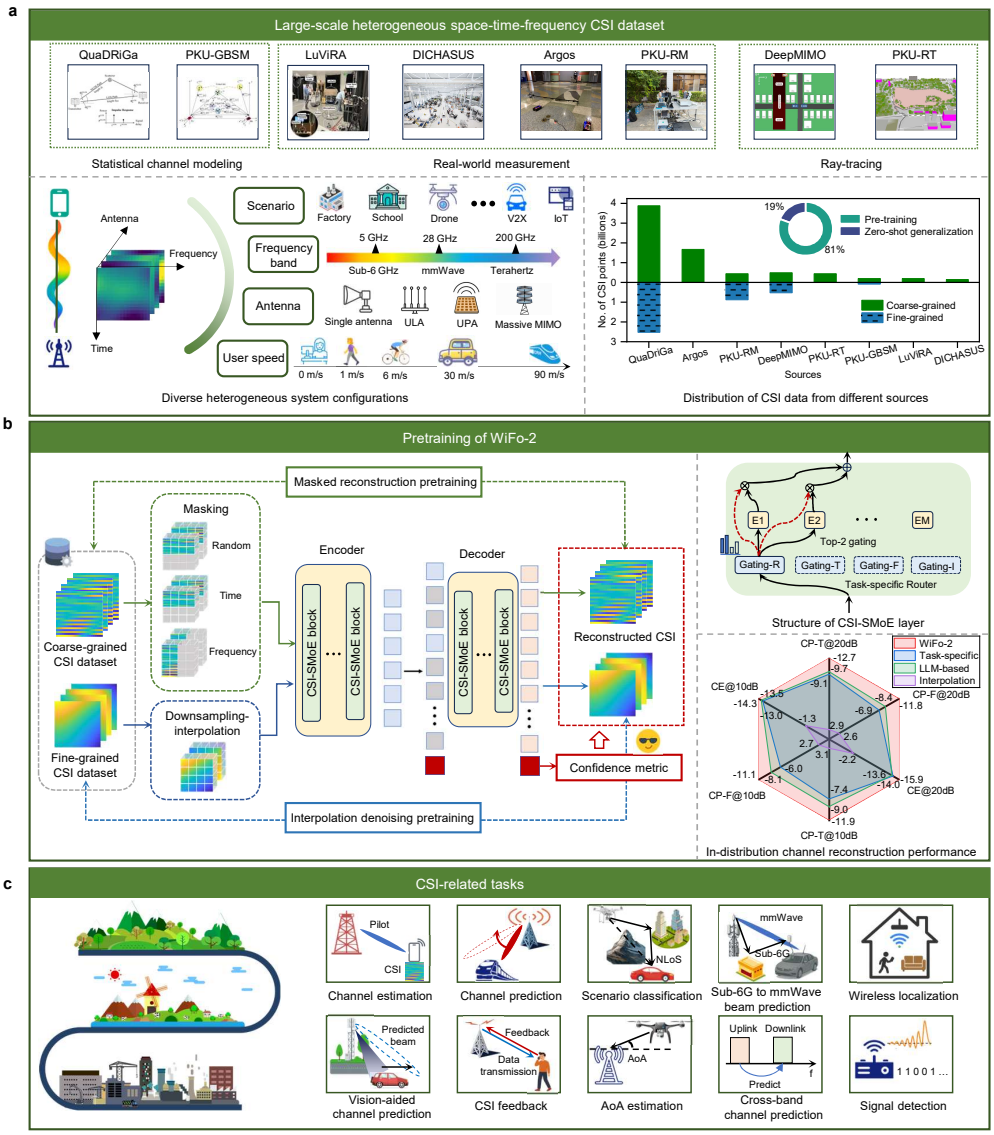

Fig. 1 Overview of WiFo-2. **a**, The constructed large-scale heterogeneous space-time-frequency CSI dataset for the WiFo-2's pretraining and zero-shot generalization. It encompasses 8 data sources and features a rich set of system configurations. **b**, WiFo-2 is based on the MDAE architecture and undergoes two-stage pretraining. The NMSE performance (dB) of different models on the time-domain channel prediction (CP-T), frequency-domain channel prediction (CP-F), and channel estimation (CE) across the LH-CSI pretraining dataset under SNR of 10dB and 20dB is given. For each SNR, we report the mean performance over both the low and high prediction/estimation ratios. For channel prediction tasks, the low and high prediction ratios are 25% and 50%. For the channel estimation task, the low and high ratios correspond to pilot-symbol placements with proportions across (time, antenna, subcarrier) of (1/4, 1, 1/12) and (1/8, 1, 1/24). **c**, WiFo-2 is applicable to multiple CSI-related tasks, including 3 channel reconstruction tasks and 8 additional CSI-related downstream tasks.

and temporal sampling densities that span a wide range. The dataset further covers diverse CSI distributions, encompassing representative factory and drone scenarios as well as a wide range of user mobility speeds. LH-CSI was partitioned into an 81% subset as pretraining sub-datasets and a 19% subset reserved as zero-shot generalization sub-datasets for CSI reconstruction tasks. Details of the LH-CSI are described in Supplementary Note 2.

WiFo-2 adopts the proposed transformer-based MDAE architecture, which comprises a preprocessor for task-specific processing, an encoder for general CSI representation, and a decoder for CSI reconstruction (Fig. 1b). To enable unified learning across diverse CSI distributions and heterogeneous pretraining tasks, we introduce a CSI-SMoE layer for both the encoder and decoder, replacing the feedforward networks (FFN) in standard transformer blocks. Specifically, each CSI-SMoE layer comprises multiple expert networks that are selectively activated, with the expert activation determined by a task-related gating mechanism conditioned on the input tokens. Moreover, this partial-activation mechanism preserves high model capacity while substantially reducing computational cost, enabling energy-efficient, low-latency communication. WiFo-2 comprises small, base, and large versions, where the large version has 100.86 million total parameters, with 50.86 million active (Supplementary Note 3). Unless otherwise specified, the large version is utilized.

WiFo-2 utilizes a two-phase pretraining strategy. The first phase involves four mixed masking and denoising pretraining tasks. Specifically, random, time, and frequency-masked reconstruction are conducted on the coarse-grained CSI dataset, while interpolation denoising is applied to the fine-grained CSI dataset. This enables WiFo-2 to learn general CSI masked reconstruction and denoising capabilities, effectively supporting CSI-related tasks. During pretraining, we randomize the prediction and downsampling ratio and add Gaussian noise with random power to improve WiFo-2's adaptability to varying reconstruction ratios and noise levels. In the second phase, confidence-enhanced pretraining produces confidence estimates tailored to zero-shot channel reconstruction, which is crucial for reliability in real deployments. Specifically, we freeze the backbone and add an auxiliary output head for further training, enabling NMSE metric prediction while preserving the original performance.

Fig. 1b presents the in-distribution performance of WiFo-2 across the three channel reconstruction tasks. We adopt interpolation, LLM-based approaches, and SOTA task-specific AI models as baselines, with the latter two trained separately for each dataset and reconstruction ratio setting, substantially increasing the number of models needed. Notably, WiFo-2 achieves SOTA performance across all three channel-reconstruction tasks at both SNR levels, validating its ability to handle multiple channel-reconstruction tasks and heterogeneous CSI distributions simultaneously (Extended Data Table 1).

To comprehensively assess its generalization capability, we establish the WFME benchmark covering 3 channel-reconstruction tasks and 8 CSI-related downstream tasks, providing the most comprehensive evaluation of wireless foundation models to date (Fig. 1c). For channel reconstruction tasks, we adopt the zero-shot generalization split of the LH-CSI dataset, which contains nine coarse-grained CSI sub-datasets for channel prediction and 6 fine-grained sub-datasets for channel estimation. These

tasks evaluate the WiFo-2's ability to reconstruct CSI in new scenarios, which is crucial for establishing communication links during disaster recovery and under other adverse conditions. For downstream CSI-related tasks, we consider 8 representative tasks: scenario classification, sub-6G to mmWave beam prediction, wireless localization, vision-aided frequency-domain channel prediction, CSI feedback, angle of arrival (AoA) estimation, cross-band channel prediction, and signal detection. For each task, we curated datasets either derived from publicly available CSI datasets or constructed via simulation, spanning diverse data sources. These tasks rigorously evaluate WiFo-2's rapid adaptability under label-scarce conditions, with important implications for accelerating model deployment and improving everyday convenience and industrial productivity. In a word, the WFME benchmark affords a comprehensive evaluation of wireless foundation models across diverse wireless tasks and is essential for advancing the field.

2.2 WiFo-2 offers high reliability in zero-shot CSI reconstruction

We evaluate the zero-shot generalization of WiFo-2 across three channel-reconstruction tasks. We compared four categories of methods, namely task-specific AI models (transformer [28], 3D ResNet [29], channelNet [30], channelformer [31], and LSTM-based schemes [32]), LLM-based approaches [5], parametric models (interpolation and PAD[33]), and Time-MoE [34] as a universal time-series foundation model (Supplementary Note 4). Both the WiFo-2 and Time-MoE are utilized without any fine-tuning, whereas all task-specific AI models and the LLM-based scheme are trained separately for each dataset and reconstruction ratio. Table 2.2 reports the NMSE performance for channel reconstruction at an SNR of 10 dB with the low reconstruction ratio. Notably, across nearly all tasks and datasets, WiFo-2 achieves SOTA zero-shot performance, surpassing even full-shot baselines. Parametric models and Time-MoE perform worse on most datasets. For the average NMSE metric, WiFo-2 exceeds the next-best full-shot baseline by 3.24 dB for frequency-domain channel prediction, 2.67 dB for time-domain channel prediction, and 2.53 dB for channel estimation. Fig. 2a further compares average NMSE across a range of reconstruction ratios and SNRs, where WiFo-2 consistently achieves the lowest NMSE (Extended Data Table 2-4). These results validate WiFo-2's superior zero-shot channel reconstruction capability, which is vital to reducing the number of models required in deployment and to lowering the costs of additional data collection and fine-tuning.

To assess the efficiency benefits of the proposed CSI-SMoE architecture, we replaced the CSI-SMoE layer with a dense layer containing an equivalent number of activated parameters. Using the same training protocol, we constructed three corresponding dense models, i.e., small-dense, base-dense, and large-dense. Fig. 2b reports the NMSE performance of zero-shot inference and inference FLOPs of the three versions of WiFo-2 and their dense counterparts. For each version, the CSI-SMoE architecture improves performance while keeping inference FLOPs and the number of activated parameters essentially unchanged. Compared with the large version with a dense structure, the base version with the CSI-SMoE lowers the NMSE by 4.5% while

Table 1 Zero-shot channel reconstruction performance of WiFo-2 compared to other baselines across various generalization sub-datasets. A low prediction and estimation ratio and a 20 dB SNR are used.

		Full-shot						Zero-shot			
Dataset	Task	Transformer	3D ResNet	LSTM	ChannelNet	Channelformer	LLM-based	Interpolation	PAD	Time-MoE	WiFo-2
QC17	CP-T	-4.046	-9.782	-3.971	-	-	-8.631	4.911	-2.929	-4.294	**-13.559**
	CP-F	-1.274	-3.611	-2.164	-	-	-5.265	2.556	-	-0.855	**-11.813**
QC18	CP-T	-6.854	-10.737	-5.692	-	-	-10.775	3.977	-5.632	-4.610	**-15.130**
	CP-F	-8.717	-9.135	-4.989	-	-	-9.112	0.369	-	-4.903	**-15.233**
QC19	CP-T	-2.358	-6.588	-2.024	-	-	-6.075	3.050	-2.093	-1.502	**-10.586**
	CP-F	-0.078	-4.550	-4.735	-	-	-5.464	-0.230	-	-4.963	**-13.710**
AC6	CP-F	-1.173	-12.669	-13.206	-	-	**-14.032**	4.900	-	-2.948	-12.282
AC7	CP-F	-3.432	-9.847	-7.287	-	-	**-10.242**	4.584	-	-4.479	-9.216
RMC8	CP-T	-17.419	-21.152	-18.835	-	-	**-21.173**	-18.504	-10.538	-17.815	-20.944
	CP-F	-11.929	-14.074	-12.540	-	-	**-16.099**	0.048	-	-11.136	-14.096
RMC9	CP-T	-13.931	-15.894	-14.957	-	-	-15.829	-13.788	-6.438	-6.479	**-15.919**
	CP-F	-11.410	-11.938	-9.515	-	-	**-12.605**	0.059	-	1.128	-11.917
O1drone	CP-T	-5.644	-9.439	-7.403	-	-	-9.060	2.608	-6.203	-1.768	**-9.951**
	CP-F	-5.929	**-17.052**	-8.745	-	-	-15.270	-5.743	-	-11.113	-15.170
RTC3	CP-T	-16.402	-22.736	-20.880	-	-	-18.242	-15.261	**-23.873**	-9.029	-21.660
	CP-F	-17.233	-21.160	-19.335	-	-	-18.738	-11.939	-	-7.259	**-22.974**
QF17CE	CE	-	-	-	-11.420	-6.644	-10.898	-8.037	-	-	**-14.525**
QF18CE	CE	-	-	-	-19.333	-6.075	-17.783	-15.315	-	-	**-24.468**
QF19CE	CE	-	-	-	-11.792	-5.213	-10.482	-8.918	-	-	**-14.010**
RMF8	CE	-	-	-	-17.640	-16.365	-18.048	1.647	-	-	**-20.588**
RMF9	CE	-	-	-	-14.456	-13.273	-15.368	1.702	-	-	**-15.994**
O1droneCE	CE	-	-	-	-19.520	-13.237	-18.775	-15.633	-	-	**-21.515**
Average	CP-F	-5.003	-8.764	-6.431	-	-	-9.715	1.306	-	-3.559	**-12.953**
	CP-T	-6.589	-10.958	-6.596	-	-	-10.386	1.337	-5.529	-4.546	**-13.626**
	CE	-	-	-	-14.473	-8.326	-13.882	-2.619	-	-	**-17.007**

The table reports NMSE values in dB. CP-T and CP-F refer to the time-domain and frequency-domain channel prediction tasks, respectively, and CE refers to the channel estimation task. Values in **bold** indicate the best performance in each column.

reducing FLOPs by 69.9%. These results chart a scalable path to deploy larger, more capable models on compute- and energy-constrained communication devices.

To validate the effectiveness of WiFo-2's confidence prediction, we assess the accuracy of the predicted NMSE on zero-shot generalization datasets over various SNRs and reconstruction ratios. Fig. 2c shows the error cumulative distribution function for the three tasks, with more than 80% of the samples having a prediction error below 5 dB. The average MAE for time-domain channel prediction, frequency-domain channel prediction, and channel estimation is 2.12 dB, 2.76 dB, and 2.48 dB, respectively. These results indicate that WiFo-2 provides accurate confidence estimates that inform model updates and network handovers, substantially enhancing reliability in real-world deployments.

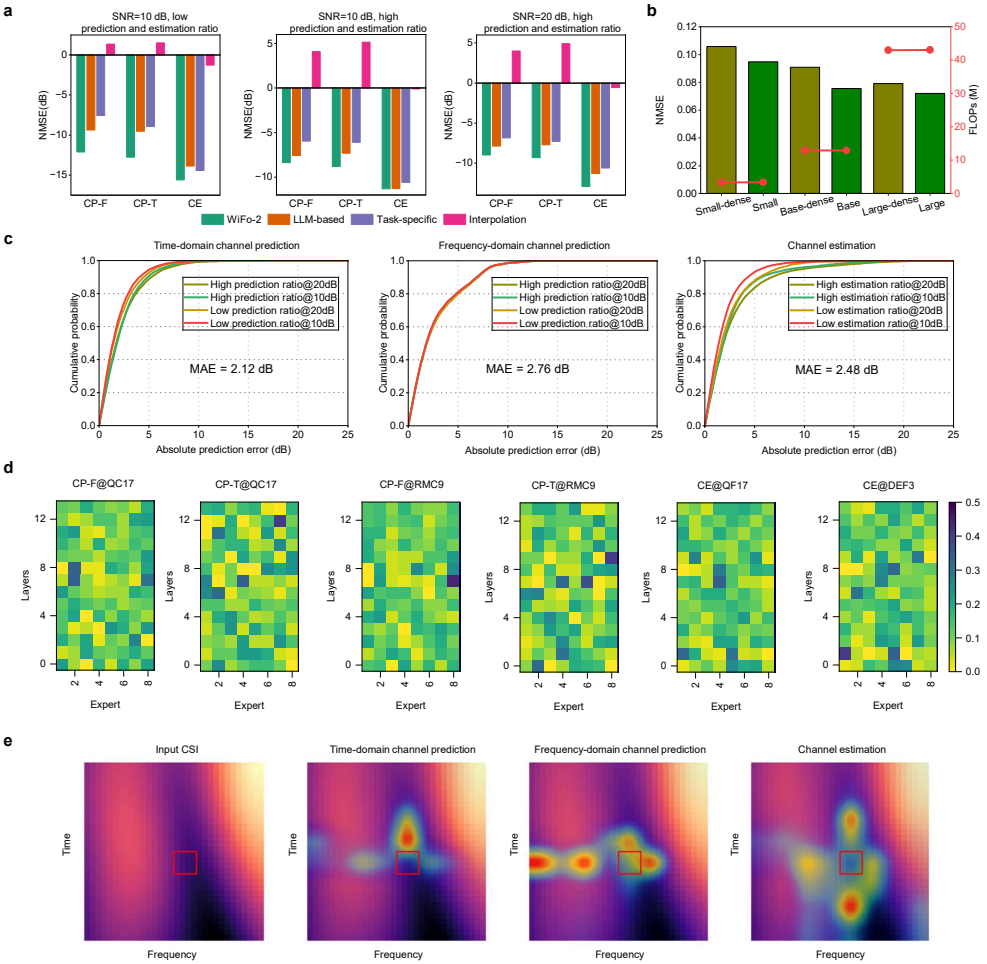

Fig. 2 Performance analysis of WiFo-2 on zero-shot channel reconstruction tasks. **a**, The zero-shot performance of WiFo-2 across three tasks under different SNR and reconstruction ratios. **b**, At an SNR of 20 dB, the average zero-shot performance of different versions of WiFo-2 on the three channel reconstruction tasks under the high- and low-ratio settings is reported. The model FLOPs for a single-token input are also provided to evaluate the inference cost. **c**, The cumulative distribution function of the confidence prediction error across three tasks. **d**, The expert activation visualization of diverse channel reconstruction tasks and datasets. The SNR is set as 20dB, and the low reconstruction ratio is considered. **e**, Visualization of self-attention heat maps of WiFo-2 for different channel reconstruction tasks.

Fig. 2d visualizes the distribution of expert activations across tasks and datasets, namely the fraction of tokens assigned to each expert of every CSI-SMoE layer. We observe that all experts are extensively utilized, and the activation patterns vary with both task and dataset. This demonstrates that the CSI-SMoE architecture effectively accommodates task and dataset heterogeneity, strengthening WiFo-2's generalization.

Fig. 2e presents self-attention heat maps that visualize the regions WiFo-2 focuses on across different tasks. We derive visualizations from the first decoder layer's query–key dot product, treating the reference token's patch as the query and displaying its attended patches. The left image shows the real part of the time–frequency domain of the processed CSI, with the red box marking the region of the reference token. For time-domain and frequency-domain channel prediction, the network primarily attends to tokens adjacent along the time or frequency axis, respectively. For channel estimation, the network focuses on surrounding tokens. These results indicate that WiFo-2 effectively captures task-adaptive time–frequency dependencies.

2.3 WiFo-2 facilitates CSI-related downstream tasks via fine-tuning

We compare WiFo-2 with pretrained wireless foundation models (including LWM [16] and SWIT [17]) and SOTA task-specific AI models on CSI-related downstream tasks. Each wireless foundation model is fine-tuned on the constructed datasets in a few-shot manner with its backbone kept frozen, whereas task-specific models are trained from scratch in either few-shot or full-shot settings. For WiFo-2, we leverage its encoder–decoder architecture and design task-oriented fine-tuning strategies for various tasks, as illustrated in Fig. 3a-b. In particular, we propose a task-specific gating fine-tuning scheme, where a new router dedicated to each task is added to the CSI-SMoE layer during fine-tuning (Fig. 3c). Only the router parameters are updated while the expert weights remain frozen, enabling effective transfer of general CSI knowledge across eight CSI-related downstream tasks. As shown in Table 2.3, WiFo-2 achieves SOTA performance across eight downstream tasks, even surpassing task-specific baselines trained under full-shot settings, clearly demonstrating its superior few-shot adaptation capability. For each task, both random initialization (scratch) and the removal of WiFo-2 lead to a significant degradation in model performance, which strongly verifies the effectiveness of transferring the pretrained general CSI knowledge. The analyses of experimental results for each downstream task are provided below, and the implementation details are illustrated in Supplementary Note 4.

Scenario classification. The task aims to infer from CSI samples whether a direct LoS path is present between the wireless link. We construct a dataset based on QuaDRiGa and use the F1 score to evaluate LoS/NLoS classification accuracy. We adopt ST-CNN [35] as the SOTA task-specific AI baseline. WiFo-2 surpasses the other baselines by at least 10.3%, evidencing its superior discrimination of multipath components.

Sub-6G to mmWave beam prediction. For a dual-band communication system, this task aims to predict the optimal mmWave beam index according to the CSI of the sub-6 GHz frequency band. We construct a QuaDRiGa-based dataset and evaluate performance using top-1 classification accuracy and spectral efficiency (SE), the latter obtained via link-level simulations. AI models referenced in [36] are used as task-specific baselines for comparison. WiFo-2 exceeds the other baselines by 18.7% in accuracy and 9.7% in spectral efficiency, demonstrating its superior spatial feature extraction capability.

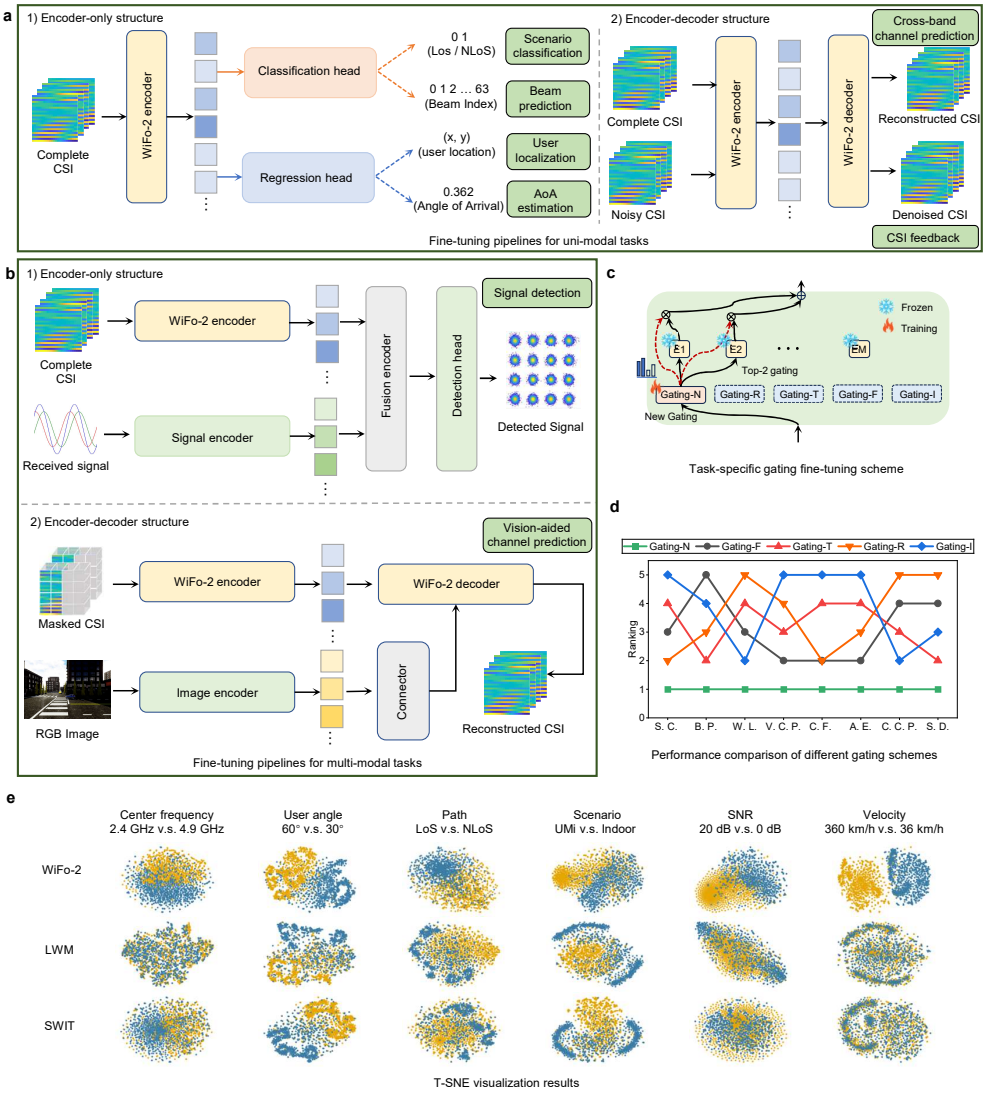

Fig. 3 Results of WiFo-2 on CSI-related downstream tasks. **a**, Fine-tuning pipelines of WiFo-2 for unimodal tasks. **b**, Fine-tuning pipelines of WiFo-2 for multi-modal tasks. **c**, An illustration of the proposed task-specific gating fine-tuning scheme. **d**, Performance ranking of different gating schemes across eight tasks. **e**, The t-SNE visualization of final-layer encoder tokens for WiFo-2, LWM, and SWIT across six datasets.

Wireless localization. The task aims to perform two-dimensional user localization from CSI. We construct a wireless localization dataset based on the open-source real-measurement dataset DICHASUS and use root-mean-square error (RMSE) to quantify localization performance. WiT [37] serves as the SOTA task-specific baseline.

Table 2 Few-shot of full-shot performance of different models on 8 downstream CSI-related tasks.

Downstream Tasks	Dataset source	Metric	WiFo-2	LWM	SWIT	Scratch	w/o WiFo-2	Task-specific (few/full shot)	Samples (few/full shot)
Scenario classification	QuaDRiGa	F1 score	**0.914**	0.639	0.762	0.550	0.549	0.800 / 0.829	1800 / 9000
Sub-6G to mmWave beam prediction	QuaDRiGa	Acc.@1	**0.871**	0.695	0.730	0.690	0.324	0.680 / 0.734	900 / 4500
		SE (bps/Hz)	**5.955**	5.124	5.370	5.157	3.333	5.077 / 5.430	
Wireless localization	DICHASUS	RMSE (m)	**0.969**	1.706	1.479	1.421	2.407	1.633 / 1.500	900 / 4500
Vision-aided frequency-domain channel prediction	SynthSoM	NMSE	**0.053**	0.940	0.877	0.980	1.020	1.226 / 1.219	180 / 900
		SE (bps/Hz)	**10.589**	8.202	9.775	7.947	6.944	6.256 / 6.478	
CSI feedback	QuaDRiGa	NMSE	**0.088**	0.193	0.175	0.646	1.380	1.253 / 1.027	900 / 4500
		SE (bps/Hz)	**11.596**	11.521	11.574	9.832	10.858	10.360 / 9.106	
	Argos	NMSE	**0.715**	0.841	0.905	1.018	1.718	2.353 / 1.011	900 / 4500
		SE (bps/Hz)	**10.043**	9.584	9.375	7.848	8.453	6.953 / 8.625	
AoA estimation	DeepMIMO	MAE (rad)	**0.019**	0.033	0.124	0.068	0.536	0.230 / 0.189	900 / 4500
Cross-band channel prediction	QuaDRiGa	NMSE	**0.328**	0.499	0.614	0.502	0.915	0.692 / 0.497	9000 / 45000
		SE (bps/Hz)	**9.367**	9.030	9.284	9.062	8.930	9.004 / 9.294	
	DICHASUS	NMSE	**0.485**	0.927	0.522	0.962	0.800	0.832 / 0.672	9000 / 45000
		SE (bps/Hz)	**10.989**	8.703	10.952	8.032	9.476	9.247 / 9.854	
Signal detection	QuaDRiGa	SER	**0.027**	0.459	0.273	0.222	0.129	0.554 / 0.438	900 / 4500

Values in **bold** indicate the best performance in each column.

WiFo-2 reduces RMSE by at least 31.8% relative to the other baselines, highlighting its ability to extract localization-relevant features directly from raw CSI.

Vision-aided frequency-domain channel prediction. The task aims to enhance frequency-domain channel prediction by exploiting environmental images captured by co-located cameras. We construct a dataset based on the open-source multi-modal sensing-communication dataset SynthSoM [38], and we evaluate models with the prediction NMSE and the system SE. We construct a task-specific baseline using a similar vision-CSI fusion network in [39]. With only 180 paired CSI and images, WiFo-2 achieves SOTA performance, with NMSE decreasing by 94.0% and SE increasing by 8.3% relative to the full-shot performance of the task-specific model.

CSI feedback. Typical CSI feedback includes compressing CSI at the transmitter and reconstructing the original CSI at the receiver. This task focuses on post-processing the CSI reconstructed by the classical CSI feedback algorithm [40] to further improve reconstruction fidelity. We construct two datasets: one based on QuaDRiGa simulations and another derived from the open-source Argos measurement dataset. We evaluate the reconstruction performance via NMSE of the reconstructed CSI and the system SE. We adopt TransNet [41] as a competitive task-specific baseline. WiFo-2 reduces NMSE by at least 15.0% and increases SE by at least 0.2% relative to the other baselines, which validates its strong denoising ability.

AoA estimation. The task aims to estimate the AoA of the dominant path between the transmitter and the receiver from CSI. We construct a ray-tracing CSI

dataset based on DeepMIMO and utilize the mean absolute error (MAE) of the estimated angle as the evaluation metric. We build a task-specific AI model based on ResNet similar to [42]. WiFo-2 achieves SOTA performance and reduces MAE by 42.4% relative to the next best baseline, which validates its ability to capture dominant path information from CSI.

Cross-band channel prediction. The task aims to infer the CSI at a nonadjacent frequency based on the CSI from another band. We construct two datasets for evaluation, one generated by QuaDRiGa simulations and another built from the DICHASUS dataset, and we evaluate model performance using both the NMSE of the prediction and the system SE. On both datasets, WiFo-2 reduces NMSE by at least 7.1% and increases SE by at least 0.3% relative to other baselines, which validates its strong capability for extrapolation across frequency.

Signal detection. The task aims to recover the transmitted symbols from the received symbols using the estimated CSI. We construct a simulation dataset with QuaDRiGa and evaluate performance using the symbol error rate (SER) of the demodulated recovered symbols. We adopt OAMPNet [43] as the SOTA task-specific baseline for the signal detection. WiFo-2 delivers the lowest SER, reducing it by at least 79.1% compared with the next best method, which we attribute to its stronger ability to handle noisy CSI.

To validate the effectiveness of the proposed task-specific gating fine-tuning scheme, we compare it with the direct application of existing gating networks without fine-tuning. Relative to using existing gating networks, the additional fine-tuning overhead of the proposed gating network is nearly negligible. As shown in Fig. 3d, the proposed gating fine-tuning scheme achieves SOTA performance across 8 tasks, confirming its ability to transfer general CSI knowledge to diverse CSI-related downstream tasks efficiently.

In Fig. 3e, we use t-SNE [44] to compare token embeddings from the final encoder layer of WiFo-2 with those from two wireless foundation model baselines. We construct 6 QuaDRiGa-based simulation datasets for controlled comparisons. In each set only one factor varies, namely center frequency, user angle, the presence or absence of a LoS path, scenario category, SNR, or user velocity. Across all sets, WiFo-2 yields more clearly separated clusters, whereas the baselines show mixed and poorly separated patterns. These results intuitively demonstrate the strong representational capacity of WiFo-2 for heterogeneous CSI.

2.4 WiFo-2's validation via over-the-air experiments

To validate the performance gains and deployment feasibility of WiFo-2 in a practical communication system, we built a wireless foundation model-empowered hardware validation platform for wireless video transmission, aiming to evaluate WiFo-2 on frequency-domain channel prediction and channel estimation tasks. As shown in Fig. 4a, the transmitter and the receiver are implemented using software-defined radios (Ettus USRP X410), each equipped with a single whip omnidirectional antenna. The radios are controlled by a Nuvo-9006E industrial PC (IPC) and synchronized with a rubidium clock using a 10-MHz reference signal. The receiver hosts an NVIDIA edge platform (Jetson AGX Orin) for AI model deployment, connected to the IPC through

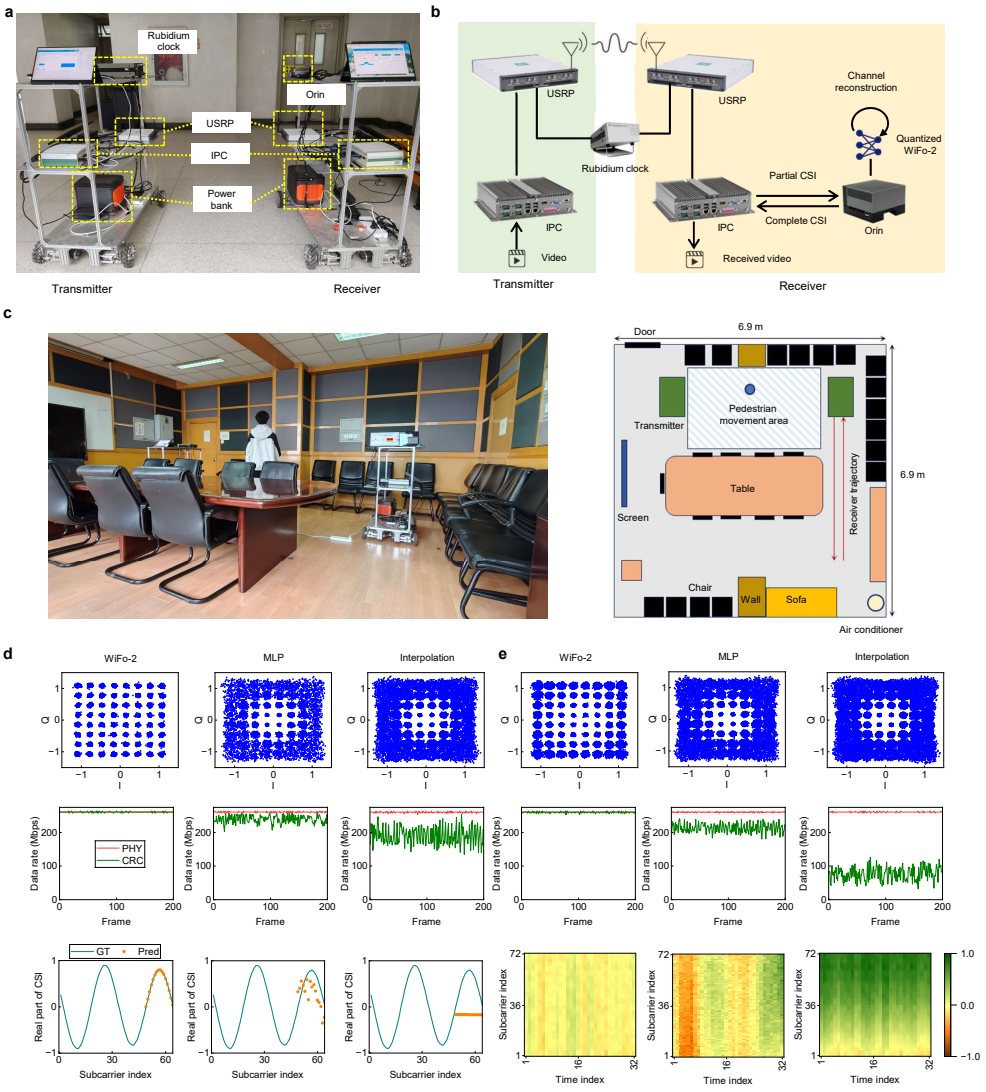

Fig. 4 Experimental results of hardware validation platform. a, Prototype of the constructed wireless hardware validation platform. **b**, System block diagram and workflow for WiFo-2 deployment. **c**, Indoor experimental setup: stationary transmitter, slow receiver motion, pedestrians randomly traversing the link. **d**, Results for the channel prediction task: receiver constellation diagrams, data rates, and CSI prediction visualizations. **e**, Results for the channel estimation task: receiver constellation diagrams, data rates, and CSI estimation error.

a wired TCP link. The overall deployment pipeline of WiFo-2 is shown in Fig. 4b. The system follows the 5G NR frame structure and inserts Demodulation Reference Signal (DM-RS) pilot symbols on selected time–frequency resources for channel estimation. The transmitter encodes and modulates the video with 64-QAM and sends it to the

receiver's USRP. To validate the effectiveness of channel prediction and estimation, the receiver estimates CSI only at a subset of pilot positions and invokes WiFo-2 to infer the full CSI. Finally, the IPC performs channel equalization with the inferred CSI and demodulates to recover the original video stream. The receiver displays the constellation of the received symbols and the system throughput in real time to indicate the accuracy of the model's channel prediction or estimation (Supplementary Note 7).

We conducted tests in an indoor environment measuring 6.9m $\times$ 6.9m, featuring a rich set of scatterers including walls, tables, chairs, and moving pedestrians (Fig. 4c). The transmitter remained stationary, and the receiver moved back and forth along a straight line at approximately 1 m/s. Pedestrians moved randomly between the transmitter and receiver at 1 m/s, frequently occluding the LoS path of the link. The system operates at a center frequency of 4.0 GHz with a bandwidth of 97.6 MHz and comprises 3,252 subcarriers. For each task, we process data in units of ten consecutive frames of the time-frequency resource grid, and each unit spans all subcarriers in frequency and 200 slots in time (100 ms). For the channel prediction task, we partition the time–frequency grid into basic prediction blocks that cover 64 subcarriers with a subcarrier spacing of 120 kHz and 32 time samples with an interval of 2.5 ms. Using the CSI of the first 48 subcarriers estimated from pilots, we predict the CSI of the remaining 16 subcarriers, reducing pilot overhead by 25%. Similarly, for the channel estimation task, we partition the grid into basic estimation blocks that cover 72 subcarriers with a spacing of 30 kHz and 32 time samples with an interval of 2.5 ms. We simulate sparse-pilot channel estimation with pilot densities of 1/8 in time and 1/24 in frequency, reducing the pilot overhead by 99.5%.

We adopt MLP and linear interpolation as baselines, which are suited for edge deployment. We collect 1,000 CSI samples in the testing scenario to fine-tune WiFo-2 and train the MLP model. In addition, to meet real-time requirements, we utilize NVIDIA TensorRT to perform INT8 quantization of WiFo-2, which accelerates inference while maintaining accuracy largely unchanged. For the frequency-domain channel prediction task, Fig. 4d shows receiver-side constellations after channel equalization, the physical layer (PHY) data rate together with the cyclic redundancy check (CRC)-verified data rate, and visual comparisons between predicted CSI and ground truth. The CSI visualizations are obtained by offline inference on the same collected samples. For a fair comparison, the constellation diagrams and data rates were obtained under essentially identical conditions, whereas the CSI visualizations are the results of offline inference on the same measured sample. We observe that WiFo-2 yields tightly clustered constellation points with error-free transmission and highly accurate CSI predictions. In contrast, the MLP- and interpolation-based predictors produce larger errors, causing constellation dispersion and throughput degradation. Similarly, Fig. 4e presents the constellation diagrams, data rates, and CSI-error visualizations for the channel-estimation task. Under an extremely sparse pilot configuration, WiFo-2 still delivers accurate estimates and enables error-free transmission. In contrast, the MLP- and interpolation-based estimators incur large prediction errors that severely reduce throughput. These experiments demonstrate WiFo-2's practical deployability and key

advantage that it achieves high-accuracy channel prediction and estimation with minimal samples. This capability drastically reduces pilot overhead while maintaining data rates, a crucial step toward building energy-efficient green 6G systems.

3 Discussion

In this study, we propose a wireless foundation model, WiFo-2, as a general-purpose model for a wide range of wireless tasks. By self-supervised pretraining across the constructed multi-source and large-scale dataset LH-CSI, WiFo-2 exhibits strong generalization ability across diverse scenarios, system configurations, and wireless tasks. Bridging the intrinsic discrepancy between channel prediction and channel estimation, for the first time, we demonstrate that a single model can achieve unified zero-shot channel reconstruction, reaching or even surpassing the full-shot performance of task-specific AI baselines. In addition, WiFo-2 can output highly accurate confidence estimates for reconstruction performance evaluation, substantially improving reliability in practical deployment. We further design task-specific fine-tuning strategies for WiFo-2 to empower 8 different types of downstream tasks, achieving performance that significantly surpasses existing wireless foundation models and task-specific AI models with fewer labeled samples. Moreover, we pioneeringly validate the feasibility of the paradigm and its system gain through hardware prototype. By maintaining transmission quality while reducing pilot overhead, WiFo-2 has the potential to improve the overall throughput rates. The above simulations and over-the-air experiments collectively highlight the transformative impact of WiFo-2 on AI-native air-interface design. On the one hand, it greatly reduces the number of deployed models via a single model serving multiple purposes, and on the other hand, it reduces or even eliminates the cost of data collection and fine-tuning in new scenarios.

Despite these advances, research on wireless foundation models still faces several open issues. On the data side, the performance of WiFo-2 is more constrained by the scale and diversity of the data than by the model size (Supplementary Note 6). Although we have built an unprecedentedly large and heterogeneous CSI dataset, it remains dominated by simulated data and is largely confined to sub-6 GHz bands. Therefore, it is essential for industrial and academic partners in wireless communications to jointly develop large-scale, open-source, measurement-based CSI datasets, thereby fundamentally lifting the performance ceiling of wireless foundation models. On the model-design side, the limited computational resources of communication devices must be fully considered. Although WiFo-2 adopts a sparse MoE architecture to scale capacity under a fixed FLOPs budget, further physics-informed model designs that exploit prior knowledge of electromagnetic propagation are needed to reduce reliance on purely data-driven learning.

In summary, WiFo-2 drives a paradigm shift in AI-native air-interface design from task-specific AI models to wireless foundation models, and holds substantial promise for building green, energy-efficient 6G communication networks.

4 Methods

4.1 Large-scale dataset construction

To pretrain WiFo-2 and evaluate its zero-shot generalization performance on channel-reconstruction tasks, we curated LH-CSI, a large-scale heterogeneous three-dimensional CSI dataset comprising 11.6 billion CSI points. The dataset integrates data from 8 distinct sources, covering all 3 CSI acquisition modalities: statistical channel modeling, real-world measurements, and ray-tracing. Each source dataset comprises several sub-datasets of either coarse-grained or fine-grained types, with 9,000, 1,000, and 2,000 CSI samples designated for training, validation, and testing, respectively. The dataset covers a broad range of wireless communication scenarios, including factories, schools, and unmanned aerial vehicle (UAV) communications. It features rich system configurations with carrier frequencies spanning from sub-6 GHz to the terahertz band and a diverse spectrum of user-mobility speeds. Detailed descriptions of each source are provided below and the system configurations are given in Supplementary Note 2.

QuaDRiGa. The dataset was generated by the 3GPP-compliant generator QuaDRiGa [21] through statistical channel modeling, which allows for the flexible adjustment of various system configurations and thereby increases the dataset's diversity. The dataset includes 19 coarse-grained sub-datasets QC1 to QC19, and 19 fine-grained CSI sub-datasets QF1 to QF19, covering typical scenarios such as UMa, UMi, indoor, and RMa. It spans multiple 5G NR frequency bands from 1.5 GHz to 28 GHz, supports antenna arrays ranging from 4 to 32 elements, and accommodates user speeds ranging from 0 to 300 km/h.

PKU-GBSM. The dataset is self-developed based on a spatiotemporal-consistent and space-time-frequency triple non-stationary channel model [22], serving as a supplement to QuaDRiGa. It is designed for dynamic vehicle-to-vehicle (V2V) scenarios in vehicular networks and can effectively model continuous arbitrary trajectories of connected vehicles. The dataset includes one coarse-grained sub-dataset VC and one fine-grained CSI sub-dataset VF, with a frequency of 5.9 GHz and 16 antennas.

LuViRA. The dataset was constructed by processing the radio modality of the open-source LuViRA [23] dataset, which is a synchronized dataset spanning vision, radio, and audio for indoor localization research. LuViRA contains 89 trajectories that are either straight lines or random. For each trajectory, it provides three-dimensional CSI with 100 samples along the time, subcarrier, and antenna dimensions. Data were collected at 100 Hz with a 20 MHz transmission bandwidth using an irregular planar antenna array, and the receiver moved at low speed. We continuously segmented the data into contiguous CSI samples of size $16\times16\times64$ along the time, antenna, subcarrier dimensions and randomly selected 12,000 samples to construct a coarse-grained CSI dataset LC.

DICHASUS. The dataset was constructed by processing the open-source massive MIMO CSI dataset DICHASUS [24], which provides channel measurements and position labels across a variety of indoor and outdoor scenarios. The coarse-grained CSI sub-datasets DC1, DC2, and DC3 are drawn from the indoor LoS dataset dichasus-015x, and from the two factory datasets dichasus-adxx and dichasus-cf0x, respectively.

From the original 1024 subcarriers, we retained the first 512 and applied smoothing to obtain 64 subcarriers. In addition, we selected the first 64 antenna elements from the original 100-element array. Since the original CSI samples contain only the subcarrier and antenna dimensions, we replicated each sample four times along the time axis to construct 3D CSI.

Argos. The dataset was constructed by processing the open-source Argos [25] dataset, which contains many-antenna MU-MIMO channel measurements collected in indoor and outdoor scenarios. We used subsets 9, 34, 38, 39, 41, 43, and 56 to construct the coarse-grained CSI sub-datasets AC1 to AC7, respectively. The measurements operated at 2.4 GHz with 96 antennas. Although the raw data include periodic temporal sampling, the sampling interval is too large to support time-domain channel prediction. We therefore used only the two-dimensional subcarrier-by-antenna CSI at each time sample and augment along the time dimension.

PKU-RM. Although existing publicly available CSI datasets are relatively large, their system configurations, such as frequency points, are limited. Additionally, the temporal sampling intervals are quite large, which makes it difficult to construct a fine-grained CSI dataset for channel estimation tasks. To address this, we developed a channel measurement platform based on a software-defined radio system (Ettus USRP X410) and conducted CSI measurements in dynamic indoor and outdoor environments on the Peking University campus to create the PKU-RM dataset. This dataset consists of 9 coarse-grained sub-datasets (RMC1 to RMC9) and 9 fine-grained sub-datasets (RMF1 to RMF9). For the fine-grained sub-datasets, the temporal sampling interval of the CSI is the inverse of the subcarrier spacing, in accordance with the time-frequency relationship of the resource block. The dataset covers 9 different frequency points and various subcarrier spacing configurations. Since the measurement platform only supports a single-input single-output (SISO) link, we replicated each sample four times along the antenna dimension to construct 3D CSI.

DeepMIMO. Our dataset was constructed by processing DeepMIMO [26], an open-source ray-tracing dataset generated with Remcom's Wireless InSite. DeepMIMO allows flexible adjustment of system configurations and follows 5G NR time and frequency settings. We selected the O1_3p4, O1_28, and O1_drone_200 scenarios to simulate the coarse-grained datasets DEC1, DEC2, DEC3 and the fine-grained CSI datasets DEF1, DEF2, DEF3, respectively. The corresponding carrier frequencies are 3.4 GHz, 28 GHz, and 200 GHz, with diverse user velocities. During generation, the channel model for all three scenarios was set to CDL-D.

PKU-RT. Although DeepMIMO offers a rich set of scenarios, its scene modeling is relatively coarse and the environments lack static scatterers, leading to scattering characteristics that differ markedly from reality. To address this, we constructed a simulated Peking University campus and performed ray tracing with SionnaRT [27], which is a lightning-fast stand-alone ray tracer for radio propagation modeling. The scene is modeled at higher fidelity and includes dynamic scatterers such as vehicles and pedestrians. We built three coarse-grained CSI sub-datasets, RTC1, RTC2, and RTC3. The first two represent communication between a base station and vehicles, and the third represents communication between a base station and pedestrians. Vehicles and pedestrians move at low, nonuniform speeds, with a temporal sampling interval of 2 ms.

4.2 Model architecture

Building on the strong performance of the MAE [45] design, we propose a MDAE as the network architecture of WiFo-2, in order to accommodate masked reconstruction and denoising. The architecture consists of three modules: a preprocessor that converts raw CSI into tokens through task-specific processing, an encoder that transforms these tokens to capture general CSI representations, and a decoder that reconstructs the CSI from the encoded features.

4.2.1 Preprocessor

To facilitate neural network processing, we convert the three-dimensional complex CSI tensor into $\boldsymbol{x} \in \mathbb{R}^{2\times T\times K\times N}$, where the two channels along the first dimension correspond to the real and imaginary parts of the original CSI. Here, T, K, and N denote the numbers of time samples, subcarriers, and antennas, respectively. To accommodate inputs with varying CSI sizes, we partition $\boldsymbol{x}$ into non-overlapping 3D patches of size $(p_{\mathrm{t}}, p_{\mathrm{f}}, p_{\mathrm{s}})$ along the time, frequency, and spatial axes, respectively. The patching operation facilitates modeling of local CSI structure and substantially reduces the number of tokens, thereby lowering computational cost. The generated $L = \left\lceil \frac{T}{p_t} \right\rceil \left\lceil \frac{K}{p_f} \right\rceil \left\lceil \frac{N}{p_a} \right\rceil$ patches are further mapped to input tokens $\boldsymbol{x}_{\mathrm{e}} \in \mathbb{R}^{L\times D}$ via a linear layer, where $\lceil\cdot\rceil$ is the ceiling operator and D is the hidden dimension. For masked reconstruction pretraining tasks, only the visible tokens are fed to the network. The masking process is defined as

$$[\boldsymbol{x}_{\mathrm{vis}}, \boldsymbol{x}_{\mathrm{mask}}] = \mathrm{MASK}\left(\boldsymbol{x}_{\mathrm{e}}\right), \tag{1}$$

where $\boldsymbol{x}_{\mathrm{mask}}$ denotes the tokens to be masked, and $\mathrm{MASK}(\cdot)$ is a task-dependent masking operator. For the denoising pretraining task, the input $\boldsymbol{x}$ is downsampled and interpolated before being fed into the network, with all tokens passed to the encoder without masking.

4.2.2 Encoder

The encoder first applies positional encoding to the preprocessed tokens and then processes them with a stack of transformer blocks. The positional encoding adopts the same space–time–frequency (STF) absolute positional encoding as WiFo, thereby embedding the three-dimensional coordinates of each token's associated patch in the original 3D CSI. This process can be formulated as $\boldsymbol{x}_{\mathrm{enc}} = f_{\mathrm{enc}}(\boldsymbol{x}_{\mathrm{vis}} + \boldsymbol{P}_{\mathrm{enc}})$, where $\boldsymbol{x}_{\mathrm{enc}} \in \mathbb{R}^{L\times D}$ is the output of the encoder, $\boldsymbol{P}_{\mathrm{enc}} \in \mathbb{R}^{L\times D}$ represents the positional encoding for the encoder, and f_{enc} denotes the encoder operation. To handle diverse CSI distributions and heterogeneous pretraining tasks, we propose a CSI-SMoE architecture that replaces the FFN in the vanilla transformer block. It comprises M expert networks $\{E_1, ..., E_M\}$ and four gating networks $\left\{G^1, G^2, G^3, G^4\right\}$, which respectively correspond to four pretraining tasks: random-masked reconstruction, time-masked reconstruction, frequency-masked reconstruction, and interpolation denoising. Each expert network is implemented with an identical FFN, while the gating network selectively activates K_{e} out of M experts and assigns weights according to the input token

$\tilde{\boldsymbol{x}}$. For j-th pretraining tasks, the output of its gating network is derived as

$$G^j(\tilde{\boldsymbol{x}}) = \text{Top-K}(\text{softmax}(g(\tilde{\boldsymbol{x}})), K_\text{e}). \tag{2}$$

Here, $g(\cdot)$ denotes the learnable network within the gating network, which in practice is implemented as a fully-connected layer. The combination of the softmax function and the Top-K$(\cdot, K)$ operator remains K largest elements of the vector while setting the remaining entries to zero. The resultant output $\hat{\boldsymbol{x}}$ of the MoE layer is the weighted sum of the outputs from the K_e selected activated experts, derived as

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{M} G^j(\tilde{\boldsymbol{x}})_i E_i(\tilde{\boldsymbol{x}}), \tag{3}$$

where $G^j(\tilde{\boldsymbol{x}})_i$ denotes the activation weight of the i-th expert for the j-th pretraining task, and $E_i(\tilde{\boldsymbol{x}})$ denotes the output of the i-th expert.

4.2.3 Decoder

In the masked reconstruction task, the encoder output is concatenated with learnable mask tokens. In contrast, the interpolation denoising task proceeds without them. Subsequently, the decoder applies STF positional encoding to all input tokens and processes them through multiple transformer blocks, which can be formulated as $\boldsymbol{x}_\text{dec} = f_\text{dec}(\boldsymbol{x}_\text{dec-in} + \boldsymbol{P}_\text{dec})$, where $\boldsymbol{x}_\text{dec-in} \in \mathbb{R}^{L\times D}$ and $\boldsymbol{x}_\text{dec} \in \mathbb{R}^{L\times D}$ denote the decoder input and output, respectively. $\boldsymbol{P}_\text{dec} \in \mathbb{R}^{L\times D}$ denotes the positional encoding, and f_dec denotes the decoder mapping. The decoder uses the same CSI-SMoE architecture as the encoder, but is lighter. Finally, the decoder maps $\boldsymbol{x}_\text{dec}$ to the complete reconstructed CSI via a linear projection and dimensional reshaping.

4.3 Pretraining scheme and applications

4.3.1 Phase 1: Mixed masking and denoising pretraining

We first pretrain WiFo-2 using a hybrid self-supervised framework that simultaneously learns expressive CSI representations and boosts performance on downstream channel prediction and estimation. The pretraining suite comprises four tasks: random-masked reconstruction, time-masked reconstruction, frequency-masked reconstruction, and interpolation denoising. On the coarse-grained CSI dataset, we pretrain WiFo-2 with three masked-reconstruction objectives to build generic masked-modeling capacity and improve performance on time and frequency domain prediction. On the fine-grained dataset, we apply the interpolation denoising pretraining scheme to enhance CSI denoising performance specifically for channel estimation.

Random-masked reconstruction. At a high masking ratio r_R, tokens are randomly omitted. This objective requires the network to reconstruct the full CSI from incompletely observed inputs, strengthening its modeling of the joint space–time–frequency structure.

Time-masked reconstruction. Along the time axis, a proportion r_T of tokens is masked, requiring the model to infer the final r_T portion of CSI in the temporal

dimension from the observed $1 - r_{\mathrm{T}}$ portion. This objective trains the network to infer future CSI dynamics from historical observations, improving time-domain channel prediction.

Frequency-masked reconstruction. Tokens with frequency indices above a specified threshold are masked at a ratio r_{F}. This objective compels the network to reconstruct CSI from partially observed spectra and strengthens learning of inter-band dependencies across neighboring subbands, thereby improving performance on frequency domain channel prediction.

Interpolation denoising. We preprocess the raw 3D CSI by retaining CSI points at a subset of space-time–frequency locations to emulate partial pilot placement. A coarse complete CSI is then produced by linear interpolation and passed to the network without masking. In this regime, WiFo-2 functions as a denoiser, yielding a refined CSI estimate and enhancing channel estimation.

During pretraining, the loss function combines a reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ and an auxiliary load-balancing loss $\mathcal{L}_{\mathrm{load}}$. Denote the reconstructed CSI on the target subset ω and its corresponding ground truth by $\boldsymbol{H}_{\omega}^{\mathrm{rec}}$ and bmH_{ω}, respectively. The MSE is adopted as the reconstruction loss, i.e.,

$$\mathcal{L}_{\mathrm{rec}} = \|\boldsymbol{H}_{\omega} - \boldsymbol{H}_{\omega}^{\mathrm{rec}}\|_F^2. \tag{4}$$

In masked-reconstruction tasks, ω comprises the masked positions. In the interpolation denoising task, ω is the entire set. In addition, to mitigate load imbalance induced by the sparse MoE, we add a load-balancing loss. For task j at an MoE layer, given a batch $\mathcal{B}$ with T_{B} tokens, the load-balancing loss is given by

$$\mathcal{L}_{\mathrm{load}} = M \sum_{i=1}^{M} \mathcal{D}_i P_i. \tag{5}$$

Here, $\mathcal{D}_i = \frac{1}{K_{\mathrm{e}} T_{\mathrm{B}}} \sum_{x \in \mathcal{B}} \mathbf{1}\left\{ \arg\max_k G^j(\tilde{\boldsymbol{x}})_k = i \right\}$ denotes the fraction of tokens routed to expert i across the batch, $P_i = \frac{1}{T_{\mathrm{B}}} \sum_{x \in \mathcal{B}} G^j(\tilde{\boldsymbol{x}})_i$ denotes the average activation weight assigned to expert i by the j-th gating network.

4.3.2 Phase 2: Confidence-enhanced pretraining

Since the zero-shot generalization performance of a model largely depends on the degree of distributional mismatch between the test scenario and the pretraining dataset. Despite extensive pretraining on a large-scale dataset in the first stage, certain corner cases can cause performance degradation, which is unacceptable for reliability-oriented communication systems. To address this, a confidence-enhanced pretraining scheme is designed to enable the pretrained WiFo-2 to actively assess and quantify its inference performance. This approach ensures more robust zero-shot generalization and offers precise channel quality priors for subsequent communication processes.

As shown in Fig. 1b, a learnable channel-quality indicator token is added at the beginning of the input token sequence for the WiFo-2 decoder. During the decoder's

forward inference, this token leverages the self-attention mechanism to perceive both the feature distribution of the input data and the proportion of masked tokens, thereby endowing the model with inference-awareness capability. To maintain the robust channel representations acquired during the first phase pretraining of WiFo-2, without modifying its original output for channel prediction and estimation, we introduce a prefix-isolated attention mask. In this design, the original tokens attend exclusively to one another, effectively isolating them from the influence of the confidence token. An additional MLP-based output head is introduced to predict the NMSE of channel reconstruction, with the MSE adopted as the loss function. Notably, this pretraining scheme allows an easy decoupling between the confidence output capability and different channel reconstruction tasks. Therefore, we train a separate learnable token and confidence output head for each task to avoid the common objective conflicts in multi-task training, thereby improving the accuracy and stability of the confidence outputs.

4.3.3 Applications

After two-stage pretraining, WiFo-2 can be directly applied to CSI reconstruction tasks. For channel prediction tasks, the target CSI segment is zero-padded along the temporal or frequency axis, and the corresponding masking type is applied. The task-aware gating network is selected for expert activation, and the predicted CSI is then obtained from the corresponding portion of the reconstructed CSI. For channel estimation, the coarsely estimated CSI is fed into the network, and the interpolation denoising gating network is activated. In addition to the reconstruction result, a confidence output is also produced.

Moreover, WiFo-2 can be efficiently applied to diverse downstream tasks via few-shot fine-tuning. As shown in Fig. 3a-b, leveraging the encoder's feature extraction ability and the decoder's CSI reconstruction ability, we design fine-tuning architectures for both uni-modal and multi-modal tasks.

Fine-tuning for uni-modal tasks. The encoder-only design is adopted for CSI feature extraction, wherein the WiFo-2 encoder is frozen and a trainable classification or regression output head is attached to enable scenario classification, beam prediction, user localization, and AoA estimation. The encoder–decoder architecture is employed for tasks whose outputs are CSI, leveraging WiFo-2's intrinsic CSI reconstruction capability to perform cross-band channel prediction and CSI feedback directly.

Fine-tuning for multi-modal tasks. The encoder-only architecture is employed for signal detection, where a frozen WiFo-2 encoder and a trainable signal encoder process the CSI and received signal, respectively, followed by a fusion encoder and detection head to output the detected symbol. The encoder-only architecture is applied for vision-aided channel prediction, with a frozen image encoder and a trainable connector converting images into tokens that are concatenated to the WiFo-2 decoder's input token sequence.

During fine-tuning, we apply a task-specific gating scheme, as shown in Fig. 3c. Since the added gating network imposes only minor adaptation overhead, this design enables efficient transfer of general CSI knowledge across various tasks.

Data and code availability

Once the paper is accepted, the data and code will be open-sourced via GitHub at https://github.com/liuboxun/WiFo-2.

Author contributions

B. L.: methodology, model pretraining, dataset construction, hardware prototype, writing original draft. X. L.: downstream tasks fine-tuning, dataset construction, hardware prototype, review, and editing. S. G.: methodology, supervision, review, and editing. X. C.: conceptualization, overall planning, supervision, review, and editing. L. Y.: supervision, review, and editing.

Competing interests

The authors declare no competing interests.

Extended data

Extended Data Table 1

Table 3 In-distribution channel reconstruction performance of WiFo-2 and other baselines.

Configurations	Task	Transformer	3D ResNet	LSTM	ChannelNet	Channelformer	LLM-based	Interpolation	PAD	Time-MoE	WiFo-2
SNR=20dB &low ratio	CP-F	-6.514	-7.792	-4.985	-	-	-9.429	1.232	-	-4.907	**-13.321**
	CP-T	-7.534	-11.253	-5.423	-	-	-11.445	0.793	-5.619	-5.560	**-15.384**
	CE	-	-	-	-16.229	-15.098	-15.016	-3.683	-	-	**-18.629**
SNR=20dB &high ratio	CP-F	-4.962	-6.239	-3.852	-	-	-7.582	3.636	-	-2.381	**-10.639**
	CP-T	-5.575	-7.662	-4.028	-	-	-8.404	4.313	5.790	-2.999	**-11.077**
	CE	-	-	-	-12.514	-10.550	-12.479	-1.162	-	-	**-14.190**
SNR=10dB &low ratio	CP-F	-6.177	-6.741	-4.954	-	-	-9.059	1.297	-	-3.384	**-12.583**
	CP-T	-7.227	-8.987	-5.415	-	-	-10.491	0.960	-3.138	-3.816	**-14.278**
	CE	-	-	-	-14.816	-15.097	-14.839	-2.185	-	-	**-16.873**
SNR=10dB &high ratio	CP-F	-4.697	-5.407	-3.840	-	-	-7.331	3.746	-	-1.865	**-10.016**
	CP-T	-5.404	-6.309	-4.026	-	-	-7.923	4.550	-1.625	-2.327	**-10.322**
	CE	-	-	-	-11.778	-10.552	-12.427	-0.635	-	-	**-12.747**

The table reports NMSE values across all pretraining datasets in dB. CP-T and CP-F refer to the time-domain and frequency-domain channel prediction tasks, respectively, and CE refers to the channel estimation task. Values in **bold** indicate the best performance in each column.

Table 4 Zero-shot channel reconstruction performance of WiFo-2 compared to other baselines across various generalization sub-datasets. A low prediction and estimation ratio and a 10 dB SNR are used.

		Full-shot						Zero-shot			
Dataset	Task	Transformer	3D ResNet	LSTM	ChannelNet	Channelformer	LLM-based	Interpolation	PAD	Time-MoE	WiFo-2
QC17	CP-T	-3.626	-6.819	-3.954	-	-	-7.024	4.986	-0.707	-1.315	**-11.825**
	CP-F	-1.183	-3.114	-2.136	-	-	-4.997	2.586	-	0.019	**-10.580**
QC18	CP-T	-6.410	-7.978	-5.705	-	-	-9.289	4.064	-3.915	-3.012	**-13.492**
	CP-F	-7.870	-7.545	-4.965	-	-	-8.428	0.420	-	-4.701	**-13.320**
QC19	CP-T	-2.225	-5.038	-2.024	-	-	-5.521	3.163	-0.319	-0.078	**-9.887**
	CP-F	-4.508	-4.016	-2.391	-	-	-5.387	-0.102	-	-6.666	**-12.924**
AC6	CP-F	-1.070	-7.201	-13.145	-	-	**-13.296**	4.929	-	-1.978	-11.281
AC7	CP-F	-3.139	-7.981	-7.278	-	-	**-9.856**	4.627	-	-3.973	-8.685
RMC8	CP-T	-16.775	-18.719	-18.636	-	-	**-20.503**	-11.896	-4.305	-11.582	-20.295
	CP-F	-11.521	-12.902	-12.326	-	-	**-15.400**	0.115	-	-8.486	-13.438
RMC9	CP-T	-13.680	-15.221	-14.838	-	-	-15.519	-10.257	-3.665	-5.068	**-15.636**
	CP-F	-11.137	-11.337	-9.354	-	-	**-12.204**	0.127	-	1.785	-11.480
O1drone	CP-T	-5.505	-7.974	-7.407	-	-	-8.728	2.730	-5.163	-1.536	**-9.464**
	CP-F	-5.674	-14.050	-8.742	-	-	-14.055	-5.544	-	-8.279	**-14.359**
RTC3	CP-T	-16.261	-19.897	-20.872	-	-	-17.835	-11.051	-20.461	-6.086	**-20.924**
	CP-F	-17.128	-18.922	-19.281	-	-	-18.528	-11.155	-	-5.315	**-21.864**
QF17CE	CE	-	-	-	-11.420	-6.644	-10.898	-5.928	-	-	**-13.426**
QF18CE	CE	-	-	-	**-19.333**	-6.075	-17.783	-8.285	-	-	-18.001
QF19CE	CE	-	-	-	-11.792	-5.213	-10.482	-6.251	-	-	**-12.778**
RMF8	CE	-	-	-	-17.640	-16.365	-18.048	2.671	-	-	**-20.096**
RMF9	CE	-	-	-	-14.456	-13.273	-15.368	2.719	-	-	**-15.788**
O1droneCE	CE	-	-	-	**-18.634**	-13.224	-18.628	-8.761	-	-	-18.014
Average	CP-F	-4.795	-7.535	-6.398	-	-	-9.364	1.363	-	-2.834	**-12.104**
	CP-T	-6.327	-8.938	-6.590	-	-	-9.507	1.499	-3.266	-2.949	**-12.758**
	CE	-	-	-	-14.422	-8.325	-13.874	-1.255	-	-	**-15.581**

The table reports NMSE values in dB. CP-T and CP-F refer to the time-domain and frequency-domain channel prediction tasks, respectively, and CE refers to the channel estimation task. Values in **bold** indicate the best performance in each column.

Extended Data Table 2

Extended Data Table 3

Extended Data Table 4

Supplementary information

Supplementary Note 1. System model and problem description

Space-time-frequency channel model

We study a canonical multiple-input multiple-output (MIMO) and orthogonal frequency division multiplexing (OFDM) system where the transmitter uses a planar

Table 5 Zero-shot channel reconstruction performance of WiFo-2 compared to other baselines across various generalization sub-datasets. A high prediction and estimation ratio and a 20 dB SNR are used.

		Full-shot						Zero-shot			
Dataset	Task	Transformer	3D ResNet	LSTM	ChannelNet	Channelformer	LLM-based	Interpolation	PAD	Time-MoE	WiFo-2
QC17	CP-T	-2.508	-5.627	-1.683	-	-	-5.264	8.791	-0.340	-0.044	**-8.749**
	CP-F	-0.571	-2.175	-1.149	-	-	-3.088	5.596	-	0.782	**-7.405**
QC18	CP-T	-4.386	-6.299	-2.862	-	-	-7.301	7.777	-1.714	-2.182	**-9.656**
	CP-F	-0.100	-5.974	-2.610	-	-	-6.809	3.478	-	-1.979	**-10.223**
QC19	CP-T	-0.985	-3.414	-1.415	-	-	-3.576	6.347	-0.217	0.903	**-6.835**
	CP-F	-0.071	-3.429	-1.198	-	-	-4.228	2.587	-	-6.134	**-10.574**
AC6	CP-F	-0.670	-12.328	-12.019	-	-	**-12.592**	8.071	-	0.556	-7.862
AC7	CP-F	-0.865	-7.069	-5.380	-	-	**-7.957**	7.613	-	-1.716	-5.589
RMC8	CP-T	-16.267	-18.493	-10.582	-	-	-19.444	-14.788	-9.105	-13.921	**-19.636**
	CP-F	-10.443	-9.726	-9.179	-	-	**-14.002**	0.071	-	-4.104	-9.919
RMC9	CP-T	-12.988	-14.176	-6.100	-	-	**-15.009**	-10.954	-5.224	-4.246	-14.949
	CP-F	-9.282	-9.058	-2.367	-	-	**-11.128**	0.105	-	0.590	-8.469
O1drone	CP-T	-3.636	-5.398	-5.388	-	-	**-6.965**	5.656	-0.002	-2.302	-5.038
	CP-F	-5.547	-13.409	-6.841	-	-	**-13.913**	-2.865	-	-5.044	-11.100
RTC3	CP-T	-16.603	**-20.916**	-19.252	-	-	-18.242	-12.066	-20.145	-4.814	-19.016
	CP-F	-13.916	-19.563	-17.508	-	-	-16.360	-8.615	-	-5.675	**-19.801**
QF17CE	CE	-	-	-	-6.266	-2.178	-7.091	-2.416	-	-	**-8.689**
QF18CE	CE	-	-	-	-18.625	-2.741	-16.861	-15.666	-	-	**-21.840**
QF19CE	CE	-	-	-	-8.240	-2.253	-7.990	-5.059	-	-	**-10.121**
RMF8	CE	-	-	-	-14.367	-9.424	-17.307	3.312	-	-	**-18.940**
RMF9	CE	-	-	-	-11.609	-8.651	-14.836	3.314	-	-	**-15.104**
O1droneCE	CE	-	-	-	**-17.855**	-10.575	-17.199	-12.936	-	-	-16.209
Average	CP-F	-2.517	-6.859	-4.288	-	-	-7.860	3.989	-	-1.759	**-8.987**
	CP-T	-4.927	-7.269	-4.382	-	-	-7.726	4.919	-2.450	-2.259	**-9.325**
	CE	-	-	-	-10.605	-4.620	-11.294	-0.578	-	-	**-12.915**

The table reports NMSE values in dB. CP-T and CP-F refer to the time-domain and frequency-domain channel prediction tasks, respectively, and CE refers to the channel estimation task. Values in **bold** indicate the best performance in each column.

array and the receiver has a single antenna. The planar array contains $N = N_h N_v$ elements, with N_h and N_v along the horizontal and vertical axes. Under a widely adopted geometric channel model with P paths, the CSI between the transmitter and the receiver at time t and frequency f is modeled as

$$\boldsymbol{h}(t,f) = \sum_{p=1}^{P} \beta_p \boldsymbol{a}(\phi_p, \theta_p) e^{j2\pi(\nu_p t - \tau_p f)}, \tag{6}$$

where β_p, ν_p, and τ_p represent the complex path gain, Doppler frequency shift, and the delay of the p-th path. The UPA steering vector $\boldsymbol{a}(\phi_p, \theta_p)$ is parameterized by the azimuth angle ϕ_p and elevation angle θ_p. Assume that T time samples are collected uniformly with spacing Δt, and K subcarriers are uniformly spaced by Δf. The

Table 6 Zero-shot channel reconstruction performance of WiFo-2 compared to other baselines across various generalization sub-datasets. A high prediction and estimation ratio and a 10 dB SNR are used.

		Full-shot						Zero-shot			
Dataset	Task	Transformer	3D ResNet	LSTM	ChannelNet	Channelformer	LLM-based	Interpolation	PAD	Time-MoE	WiFo-2
QC17	CP-T	-2.255	-4.027	-1.672	-	-	-4.543	8.887	0.326	0.584	**-7.664**
	CP-F	-0.518	-1.838	-1.136	-	-	-2.925	5.644	-	0.848	**-6.691**
QC18	CP-T	-4.110	-4.871	-2.857	-	-	-6.560	7.894	-1.286	-1.739	**-8.745**
	CP-F	-0.099	-4.924	-2.595	-	-	-6.238	3.555	-	-1.246	**-9.075**
QC19	CP-T	-0.956	-2.547	-1.422	-	-	-3.400	6.511	0.874	0.787	**-6.378**
	CP-F	-0.068	-3.094	-1.188	-	-	-4.155	2.683	-	-5.856	**-9.973**
AC6	CP-F	-0.621	-7.754	-12.009	-	-	**-12.104**	8.114	-	1.071	-7.119
AC7	CP-F	-0.816	-5.315	-5.370	-	-	**-7.830**	7.677	-	-1.444	-5.174
RMC8	CP-T	-15.594	-15.437	-10.550	-	-	-18.760	-7.058	-3.590	-8.965	**-18.778**
	CP-F	-9.744	-8.915	-9.143	-	-	**-12.576**	0.258	-	-3.099	-9.355
RMC9	CP-T	-12.767	-13.044	-6.035	-	-	**-14.666**	-6.084	-2.613	-3.478	-14.587
	CP-F	-9.226	-8.629	-2.329	-	-	**-10.351**	0.292	-	0.563	-8.069
O1drone	CP-T	-3.611	-4.594	-5.414	-	-	**-6.894**	5.846	-0.248	-1.784	-4.928
	CP-F	-5.359	-11.028	-6.790	-	-	**-13.070**	-2.531	-	-3.855	-10.500
RTC3	CP-T	-16.395	-18.654	**-19.185**	-	-	-17.915	-6.537	-17.409	-3.419	-18.300
	CP-F	-13.694	-17.284	-17.422	-	-	-16.122	-7.497	-	-4.612	**-18.539**
QF17CE	CE	-	-	-	-6.266	-2.178	-7.091	-1.644	-	-	**-8.034**
QF18CE	CE	-	-	-	**-18.625**	-2.741	-16.861	-9.476	-	-	-15.676
QF19CE	CE	-	-	-	-8.240	-2.253	-7.990	-3.736	-	-	**-9.168**
RMF8	CE	-	-	-	-14.367	-9.424	-17.307	3.519	-	-	**-17.780**
RMF9	CE	-	-	-	-11.609	-8.651	**-14.836**	3.519	-	-	-14.544
O1droneCE	CE	-	-	-	**-17.256**	-10.582	-17.075	-8.164	-	-	-10.649
Average	CP-F	-2.463	-5.940	-4.268	-	-	-7.557	4.073	-	-1.342	**-8.364**
	CP-T	-4.788	-6.104	-4.376	-	-	-7.321	5.144	-1.468	-1.724	**-8.816**
	CE	-	-	-	-10.585	-4.620	-11.289	-0.095	-	-	**-11.314**

The table reports NMSE values in dB. CP-T and CP-F refer to the time-domain and frequency-domain channel prediction tasks, respectively, and CE refers to the channel estimation task. Values in **bold** indicate the best performance in each column.

resulting space–time–frequency CSI forms a tensor $\boldsymbol{H} \in \mathbb{C}^{T\times K\times N}$ satisfying

$$\boldsymbol{H}_{i,k,1:N} = \boldsymbol{h}(t_i, f_k), \qquad i = 1,\dots,T,\ k = 1,\dots,K, \tag{7}$$

where $t_i = i\Delta t$ and $f_k = f_1 + (k-1)\Delta f$ denote the i-th time instant and the k-th subcarrier, respectively. Here f_1 denotes the frequency of the first subcarrier.

CSI reconstruction tasks

Channel estimation. It aims to reconstruct the complete 3D CSI from partial observations at pilot locations. In practice, communication systems adopt diverse pilot configurations, including block, comb and scattered pilots, which makes it difficult

to process them with a single network. To address this challenge, we employ a two-stage estimation scheme. First, a coarse estimate of the full CSI, $\hat{\boldsymbol{H}}^{\mathrm{co}}$, is obtained by three-dimensional interpolation of the pilot-location estimates. Second, a denoising refinement mapping function f_{CE} maps $\hat{\boldsymbol{H}}^{\mathrm{co}}$ to the final estimate $\hat{\boldsymbol{H}}^{\mathrm{ce}}$. Accordingly, we model channel estimation task as minimizing the mean squared error (MSE) between the estimate and the ground truth $\boldsymbol{H}$:

$$\min \|\boldsymbol{H} - \hat{\boldsymbol{H}}^{\mathrm{ce}}\|_F^2, \quad s.t. \quad \hat{\boldsymbol{H}}^{\mathrm{ce}} = f_{\mathrm{CE}}\left(\hat{\boldsymbol{H}}^{\mathrm{co}}\right). \tag{8}$$

Here $\|\cdot\|_F$ is the Frobenius norm.

Time-domain channel prediction. It predicts the CSI over the next T_h time samples from the CSI over the preceding $T - T_h$ samples. Therefore, we formulate the time-domain channel prediction task as minimizing the MSE between the predicted CSI $\hat{\boldsymbol{H}}^{\mathrm{TCP}}$ and the ground truth, expressed as

$$\min \|\boldsymbol{H}_{T-T_h+1:T,1:K,1:N} - \hat{\boldsymbol{H}}^{\mathrm{TCP}}_{T-T_h+1:T,1:K,1:N}\|_F^2, \tag{9}$$

$$s.t. \quad \hat{\boldsymbol{H}}^{\mathrm{TCP}}_{T-T_h+1:T,1:K,1:N} = f_{\mathrm{TCP}}(\boldsymbol{H}_{1:T-T_h,1:K,1:N}), \tag{10}$$

where f_{TCP} represents the mapping function for time-domain channel prediction.

Frequency-domain channel prediction. Similar to the prediction in the time domain, we aim to predict the CSI over the last $K - K_u$ sub-carriers based on the CSI of the first K_u sub-carriers. The frequency-domain channel prediction is also modeled as the MSE minimization problem as

$$\min \|\boldsymbol{H}_{1:T,K_u+1:K,1:N} - \hat{\boldsymbol{H}}^{\mathrm{FCP}}_{1:T,K_u+1:K,1:N}\|_F^2, \tag{11}$$

$$s.t. \quad \|\hat{\boldsymbol{H}}^{\mathrm{FCP}}_{1:T,K_u+1:K,1:N}\|_F^2 = f_{\mathrm{FCP}}(\boldsymbol{H}_{1:T,1:K_u,1:N}), \tag{12}$$

where f_{FCP} is the mapping function in the frequency-domain and $\hat{\boldsymbol{H}}^{\mathrm{FCP}}$ denotes the predicted CSI for the frequency domain.

It is observed that the above three tasks can be uniformly formulated as a CSI reconstruction task, which obtains complete and accurate CSI with partial or coarse estimated CSI. Denote the estimated partial CSI as $\hat{\boldsymbol{H}}_{\Omega}$, where Ω represents the subset of all elements. The unified channel reconstruction task can be modeled as

$$\min \|\boldsymbol{H} - \boldsymbol{H}^{\mathrm{rec}}\|_F^2, \; s.t. \, \boldsymbol{H}^{\mathrm{rec}} = f_{\mathrm{rec}}(\hat{\boldsymbol{H}}_{\Omega}), \tag{13}$$

where f_{rec} represents the mapping function of the unified channel reconstruction and $\boldsymbol{H}^{\mathrm{rec}}$ denotes the reconstructed CSI.

CSI-related downstream tasks

We evaluate 8 CSI-related downstream tasks to provide a comprehensive assessment of the enabling capability of the pretrained general-purpose representation across diverse applications.

Scenario Classification. This task aims to determine from CSI whether the link between the transmitter and the receiver contains a LoS path. This deepens understanding of the propagation environment and guides precoding and related strategy choices. The input is a three-dimensional CSI sample, and the output is a scene label, LoS or NLoS. The task evaluates the model's ability to extract multipath characteristics from CSI.

Sub-6G to mmWave Beam Prediction. MmWave bands are critical for 5G and future 6G systems. Communication devices are typically equipped with antennas that cover both low-frequency (sub-6 GHz) and mmWave frequencies. This task uses sub-6 GHz CSI to predict the optimal mmWave beam, thereby reducing beam-sweeping overhead. The input is an instantaneous two-dimensional subcarrier-by-antenna CSI sample. The output is the codebook index of the optimal mmWave beam. The task assesses the model's capability for cross-frequency, geometry-aware alignment.

Wireless Localization. Wireless localization leverages CSI to estimate user position while preserving privacy and delivering precise location services. It serves as an important complement in environments where global navigation satellite systems (GNSS) are degraded or unavailable, especially indoors. The input is a two-dimensional subcarrier-by-antenna CSI sample of the communication link, and the output is the user's two-dimensional position coordinates. The task comprehensively assesses the model's ability to extract delay and angular information from raw CSI.

Vision-aided Frequency-domain Channel Prediction. Communication devices such as smartphones and embodied robots typically integrate wireless transceivers and multi-modal sensing modules. Because both multi-modal sensory data and the wireless channel observe the same physical space, they exhibit inherent correlation [12]. This task leverages visual information from the environment to enhance frequency-domain channel prediction. Given a co-temporal pair comprising an instantaneous two-dimensional subcarrier-by-antenna CSI sample and an RGB image, the output is a frequency-domain CSI prediction. The task assesses the model's ability to leverage auxiliary side information to enhance channel prediction.

CSI Feedback. CSI feedback aims to compress and reconstruct channel state information efficiently so that accurate recovery is achieved with minimal overhead. Following [46], we cast the problem as channel denoising and apply post-processing to channels reconstructed by existing compression and feedback methods to further improve accuracy. The task requires general channel denoising capability and assesses the model's ability to extract generic channel characteristics.

AoA Estimation. AoA estimation aims to infer the angle of arrival (AoA) of the channel's dominant path from CSI, providing critical priors for subsequent beamforming and localization. The task takes as input a two-dimensional CSI sample at a given time instant and outputs the AoA of the current channel's dominant path. It assesses the model's spatial-spectrum super-resolution and dominant-path separability, and further evaluates its robust generalization in complex propagation environments.

Cross-band channel prediction. In frequency-division duplexing (FDD) systems, the uplink and downlink work on non-adjacent frequency bands, which necessitates additional CSI feedback. Predicting downlink CSI from uplink CSI can reduce or

even eliminate downlink CSI feedback and thereby substantially improve spectral efficiency. Unlike frequency-domain channel prediction, this task seeks to predict CSI in one band from CSI observed in another, non-adjacent band, posing a stricter challenge for modeling the frequency-domain correlations of CSI.

Signal detection. CSI is critical for accurately recovering the transmitted symbols at the receiver. Consider a link with an N_a-antenna transmitter and a single-antenna receiver. Let the transmit symbol be $x \in \mathbb{C}$, the spatial-domain channel be $\boldsymbol{h}_{\mathrm{s}} \in \mathbb{C}^{N_a}$, and the received signal be $\boldsymbol{y} = \boldsymbol{h}_{\mathrm{s}} x + \boldsymbol{n}$, where $\boldsymbol{n} \in \mathbb{C}^{N_a}$ denotes noise. The signal detection task aims to recover x from the observed $\boldsymbol{y}$ and the channel $\boldsymbol{h}_{\mathrm{s}}$. This task evaluates the model's spatial-domain detection and interference suppression under imperfect CSI.

Supplementary Note 2. Details of the LH-CSI dataset

In this section, the detailed system configurations of the constructed LH-CSI are illustrated, including the source, center frequency f_{C}, subcarrier number K and spacing Δf, time sample number T and interval Δt, antenna configuration, scenario, and user velocity of each sub-dataset. The datasets from the statistical channel modeling, real-world measurement, and ray-tracing parts are presented separately in Tables 7, 8, and 9, respectively. It can be observed that the constructed LH-CSI dataset exhibits pronounced heterogeneity in both system configurations and CSI distributions, providing strong support for WiFo's pre-training and generalization testing.

Supplementary Note 3. Details of network and pretraining settings

WiFo-2 comprises 3 versions, including small, base, and large, and detailed parameters for each are listed in Table 10. For each version of WiFo-2, 2 experts are selectively activated out of the 8. We use a patch size of 4 along the antenna, time, and subcarrier dimensions. During the two-phase pretraining, we add complex Gaussian noise to the input CSI, with SNR randomly set to 10–25dB. For the time- and frequency-masked reconstruction pretraining tasks, the mask ratio is randomly sampled from 10%–25%. For the random-masked reconstruction pretraining tasks, the mask ratio is fixed as 85%. For the interpolation–denoising task, pilot-symbol placement proportions over (time, antenna, subcarrier) are randomly chosen in the range (1/8, 1, 1/24) to (1/4, 1, 1/6). All models are trained on 8 NVIDIA GeForce RTX 5090-32G GPUs with BF16 precision.

For first-phase pretraining, we trained for 250 epochs with a batch size of 128. We used AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a weight decay of 0.05. The load-balancing loss is weighted by 0.03. We employed a cosine-decay learning-rate schedule with 5 warmup epochs. For the base and small versions, the base learning rate is 5×10^{-4} and the minimum learning rate is 3×10^{-4}. For the large version, the base learning rate is 3×10^{-4} and the minimum learning rate is 6×10^{-5}.

Table 7 An illustration of the system configurations for the part of LH-CSI generated using statistical channel modeling.

Source	Dataset	f_C(GHz)	K	Δf(kHz)	T	Δt(ms)	Antenna	Scenario	User speed(km/h)
QuaDRiGa	QC1	1.5	128	60	32	1	1×4	UMi NLoS	3-50
	QC2	1.5	128	90	32	0.5	2×4	RMa NLoS	100-200
	QC3	1.5	64	90	16	1	1×8	Indoor LoS	0-10
	QC4	1.5	32	180	16	0.5	4×8	UMa LoS	30-100
	QC5	2.5	64	90	16	0.5	2×2	RMa LoS	100-200
	QC6	2.5	128	90	32	1	2×4	UMi LoS	3-50
	QC7	2.5	32	360	16	0.5	4×8	UMa LoS	30-100
	QC8	2.5	64	90	16	1	4×4	Indoor NLoS	0-10
	QC9	4.9	128	60	32	1	1×4	UMi NLoS	3-50
	QC10	4.9	64	180	16	0.5	2×4	UMa LoS	100-200
	QC11	4.9	64	60	16	0.5	4×4	UMa NLoS	30-100
	QC12	4.9	32	180	16	1	4×8	Indoor LoS	0-10
	QC13	5.9	64	90	16	0.5	2×8	UMa NLoS	100-200
	QC14	5.9	128	60	32	1	2×4	UMi NLoS	3-50
	QC15	5.9	64	90	16	1	4×4	Indoor +LoS	0-10
	QC16	5.9	32	360	16	0.5	4×8	UMa NLoS	30-100
	QC17	3.5	64	60	16	0.5	2×4	UMa NLoS	30-100
	QC18	6.7	64	60	16	1	4×4	UMi LoS	3-50
	QC19	28	64	360	16	0.5	2×8	UMa LoS	30-100
	QF1	1.5	72	60	28	0.017	1×4	UMi NLoS	3-50
	QF2	1.5	72	15	28	0.067	2×4	RMa NLoS	100-200
	QF3	1.5	72	15	28	0.067	1×8	Indoor LoS	0-10
	QF4	1.5	36	30	12	0.033	4×8	UMa LoS	30-100
	QF5	2.5	72	30	28	0.033	2×2	RMa LoS	100-200
	QF6	2.5	72	15	28	0.067	2×4	UMi LoS	3-50
	QF7	2.5	36	60	12	0.017	4×8	UMa LoS	30-100
	QF8	2.5	36	15	12	0.067	4×4	Indoor NLoS	0-10
	QF9	4.9	72	60	28	0.017	1×4	UMi NLoS	3-50
	QF10	4.9	72	30	28	0.033	2×4	UMa LoS	100-200
	QF11	4.9	36	60	12	0.017	4×4	UMa NLoS	30-100
	QF12	4.9	36	30	12	0.033	4×8	Indoor LoS	0-10
	QF13	5.9	36	15	12	0.067	2×8	UMa NLoS	100-200
	QF14	5.9	72	60	28	0.017	2×4	UMi NLoS	3-50
	QF15	5.9	36	15	12	0.067	4×4	Indoor LoS	0-10
	QF16	5.9	36	15	12	0.067	4×8	UMa NLoS	30-100
	QF17	3.5	72	15	12	0.067	2×4	UMa NLoS	30-100
	QF18	6.7	36	15	12	0.067	4×4	UMi LoS	3-50
	QF19	28	72	30	16	0.033	2×8	UMa LoS	30-100
PKU-GBSM	VC	5.9	64	360	16	1	4×4	V2V	108 & 36
	VF	5.9	72	60	12	0.017	2×4	V2V	108 & 36

Table 8 An illustration of the system configurations for the part of LH-CSI generated via real-world measurement.

Source	Dataset	f_C(GHz)	K	Δf(kHz)	T	Δt(ms)	Antenna	Scenario	User speed(km/h)
LuViRA	LC	3.7	64	200	16	10	64	Indoor	~ 0.5
DICHASUS	DC1	1.3	64	391	4	-	64	Indoor LoS	-
	DC2	3.4	64	391	4	-	64	Factory	-
	DC3	5.9	64	391	4	-	64	Factory	-
Argos	AC1	2.4	52	313	4	-	96	Indoor	-
	AC2	2.4	52	313	4	-	96	Indoor	-
	AC3	2.4	52	313	4	-	96	Outdoor	-
	AC4	2.4	52	313	4	-	96	Outdoor	-
	AC5	2.4	52	313	4	-	96	Outdoor	-
	AC6	2.4	52	313	4	-	96	Outdoor	-
	AC7	5.0	52	313	4	-	96	Outdoor	-
PKU-RM	RMC1	3.8	64	180	16	0.5	4	Indoor LoS	~ 2
	RMC2	4.6	64	180	16	0.5	4	Indoor NLoS	~ 2
	RMC3	3.3	64	90	16	1	4	Indoor NLoS	~ 2
	RMC4	2.0	64	90	16	1	4	Indoor LoS	~ 2
	RMC5	5.5	64	360	16	0.25	4	Indoor LoS	~ 2
	RMC6	1.6	64	180	16	0.5	4	Outdoor LoS	~ 5
	RMC7	2.1	64	180	16	0.5	4	Outdoor LoS	~ 5
	RMC8	3.7	64	180	16	0.5	4	Outdoor LoS	~ 5
	RMC9	5.2	64	180	16	0.5	4	Outdoor LoS	~ 5
	RMF1	3.8	72	30	28	0.033	4	Indoor LoS	~ 2
	RMF2	4.6	72	30	28	0.033	4	Indoor NLoS	~ 2
	RMF3	3.3	72	15	28	0.067	4	Indoor NLoS	~ 2
	RMF4	2.0	72	15	28	0.067	4	Indoor LoS	~ 2
	RMF5	5.5	72	60	28	0.017	4	Indoor LoS	~ 2
	RMF6	1.6	72	30	28	0.033	4	Outdoor LoS	~ 5
	RMF7	2.1	72	30	28	0.033	4	Outdoor LoS	~ 5
	RMF8	3.7	72	30	28	0.033	4	Outdoor LoS	~ 5
	RMF9	5.2	72	30	28	0.033	4	Outdoor LoS	~ 5

For second-stage pretraining, we trained for 50 epochs with a batch size of 128, again using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a weight decay of 0.001 and a cosine-decay schedule with 5 warmup epochs. Across all model sizes, the base learning rate was 1×10^{-4} with a minimum of 1×10^{-5}.

Table 9 An illustration of the system configurations for the part of LH-CSI generated via ray-tracing.

Source	Dataset	f_C(GHz)	K	Δf(kHz)	T	Δt(ms)	Antenna	Scenario	User speed(km/h)
DeepMIMO	DEC1	3.4	64	120	16	0.233	8	Outdoor	$100-120$
	DEC2	28	32	480	16	0.029	4×8	Outdoor	$100-300$
	DEC3	200	64	960	16	0.029	2×8	Drone	$30-50$
	DEF1	3.4	72	30	28	0.022	8	Outdoor	$100-120$
	DEF2	28	36	120	12	0.008	4×8	Outdoor	$100-300$
	DEF3	200	72	120	12	0.008	2×8	Drone	$30-50$
PKU-RT	RTC1	5.9	64	180	16	2	8	V2I	~10
	RTC2	5.9	64	180	16	2	8	V2I	~10
	RTC3	5.9	64	180	16	2	8	Pedestrian	~5

Table 10 The network configuration of WiFo-2 variants. The "enc." and "dec." denote the encoder and decoder, respectively.

Version	Depth (enc. /dec.)	Heads	Experts (active/total)	Expert dimension	Params (active/total)
Small	6/4	8	2/8	256	4.2M/8.3M
Base	6/4	8	2/8	512	16.5M/32.4M
Large	8/6	12	2/8	768	50.9M/100.9M

Supplementary Note 4. Dataset, baselines, and implementation details of the WFME benchmark

In this section, we provide a detailed account of the dataset construction, the implementation details, and the baseline settings for each task in the proposed WFME benchmark.

Channel reconstruction. We evaluate the zero-shot generalization performance of WiFo-2 on the zero-shot split of the LH-CSI dataset. For each channel reconstruction task, we evaluate the zero-shot performance at both the 10 dB and 20 dB SNR, and under both the high and low prediction/estimation ratios. Details of the adopted baselines are presented below.

(1) **Transformer**[28]: The transformer-based predictor in [28] is proposed for parallel channel prediction. For a fair comparison, it uses the same CSI embedding pipeline as WiFo-2. Specifically, the raw 3D CSI is first partitioned into patches and embedded into tokens, which are then processed by the network. The token dimension is set to 128, and the encoder and decoder depths are 5 and 8, respectively. In our experiments, the model is trained and evaluated for both time-domain and frequency-domain CSI prediction.

(2) **3D ResNet**[29]: A 3D ResNet (SlowFast-style) originally devised for video recognition [29] leverages spatiotemporal convolutions to model volumetric dependencies. In our configuration, the 50-layer network ingests CSI tensors to capture

three-dimensional correlations across time, frequency, and antenna dimensions, and is trained to perform both time-domain and frequency-domain channel prediction.

(3) **LSTM**[32]: Long short-term memory (LSTM) is proposed for sequential processing to overcome the vanishing gradient problem. For time-domain channel prediction, the antenna and frequency dimensions are flattened and fed as features to the LSTM. For frequency-domain prediction, we analogously flatten the time and antenna dimensions.

(4) **LLM-based scheme**[5]: LLM4CP fine-tunes GPT-2 for cross-modal knowledge transfer in channel prediction [5]. We equip LLM4CP with the same 3D patching module as WiFo-2 so that it can handle 3D CSI. By analogy, we also evaluate a fine-tuned LLM for channel estimation.

(5) **Interpolation**: We include interpolation baselines for both prediction and estimation. For channel prediction, future CSI is obtained by first-order linear extrapolation from historical observations. For channel estimation, we use bilinear interpolation to reconstruct missing CSI entries.

(6) **PAD**[47]: PAD is a Prony-based angular-delay domain channel prediction method that is based on the angle-delay-doppler structure of the channel. We use PAD only for time-domain prediction and set the predictor order to $N = 4$.

(7) **Time-MoE**[34]: Time-MoE is a large-scale time-series foundation model with strong zero-shot forecasting ability [34]. We evaluate the Time-MoE-200M checkpoint strictly under a zero-shot protocol, using the publicly released pretrained model without fine-tuning, prompt tuning, or in-domain adaptation. By parallelizing the remaining two dimensions, we transform time-domain and frequency-domain channel prediction into a one-dimensional sequence prediction problem, which can be readily handled by Time-MoE.

(8) **ChannelNet**[30]: ChannelNet formulates channel estimation using image super-resolution and restoration techniques within a CNN framework. Following [30], we employ a 2-layer CNN for super-resolution and a 20-layer CNN for restoration. To handle 3D CSI, the antenna dimension is processed in parallel across channels.

(9) **Channelformer**[31]: Channelformer adopts an encoder–decoder architecture tailored to channel estimation. In line with the configuration in [31], we use 5 encoder layers and 12 decoder layers, and parallelize over the antenna dimension to accommodate 3D CSI.

For all task-specific AI models and LLM-based schemes, the SNR during training is randomly chosen in the range 10–25 dB, identical to WiFo-2 for a fair comparison. We use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 0.001. For channel prediction tasks, all task-specific and LLM-based schemes are trained for 50 epochs with a fixed learning rate of 4×10^{-4} and a batch size of 64. For channel estimation tasks, the learning rate is adjusted to 1×10^{-4}, with other settings unchanged.

Scenario classification. We used QuaDRiGa [21] to synthesize a scenario-classification dataset with 1,200 UMi-LoS and 1,200 UMi-NLoS CSI samples at a center frequency of 2.5 GHz and a bandwidth of 10 MHz. Each CSI sample is annotated with a scenario label and is organized along three axes, temporal, frequency, and spatial, with 24 consecutive time samples, 128 OFDM subcarriers, and measurements

Table 11 The fine-tuning setting of WiFo-2 across scenario classification, sub-6G to mmWave beam prediction, wireless localization, and vision-aided frequency-domain channel prediction.

Task	Scenario classification	Sub-6G to mmWave beam prediction	Wireless localization	Vision-aided frequency-domain channel prediction
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	2e-4	1e-4	2e-05	1e-4
Weight decay	0.0001	0.0001	0.0001	0.0001
Optimizer momentum	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Batch size	256	256	256	256
Max epoch	100	100	200	100
Loss function	CrossEntropy	CrossEntropy	MSE	MSE

Table 12 The fine-tuning setting of WiFo-2 across CSI feedback, AoA estimation, cross-band channel prediction, and signal detection.

Task	CSI feedback	AoA estimation	Cross-band channel prediction	Signal detection
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	6e-05	2e-05	4e-05	2e-05
Weight decay	0.0001	0.0001	0.0001	0.0001
Optimizer momentum	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Batch size	256	256	256	256
Max epoch	300	200	100	100
Loss function	MSE	MSE	MSE	CrossEntropy

over a 2×4 UPA. The temporal sampling interval is 1 ms, and the subcarrier spacing is 90 kHz. The dataset is partitioned into training, validation, and test sets containing 1,800, 200, and 400 samples, respectively. Each wireless foundation model serves as an encoder for feature extraction, upon which we attach the same linear classification head to produce scenario labels. As for task-specific AI models, we adopt ST-CNN [35], a 3D CNN tailored to capture the joint temporal–frequency–spatial structure of CSI. The F1 score was used to evaluate the classification accuracy. Further training details are provided in Table 11.

Sub-6G to mmWave beam prediction. This dataset was also generated using the QuaDRiGa [21] simulator under the UMi-LoS scenario. The dataset comprises 900 training samples, 100 validation samples, and 200 test samples, each containing paired sub-6 GHz CSI and mmWave CSI acquired at the same transmitter–receiver location. For the sub-6 GHz link, the carrier frequency is set to 2.5 GHz with a bandwidth of 10 MHz, employing a 1×16 ULA antenna configuration. For the mmWave link, the carrier frequency is 28 GHz with a bandwidth of 0.5 GHz, using a 1×64 ULA antenna configuration. Both the sub-6 GHz link and the mmWave link comprise 64 subcarriers. The mmWave system uses a 64-entry DFT codebook, and the optimal beam is the codeword that maximizes the average magnitude across all subcarriers. Each wireless foundation model serves as an encoder to extract features from the Sub-6 GHz channel, and the same linear layer is subsequently applied to predict the corresponding mmWave beam. For task-specific AI models, we adopt the same neural network architecture proposed in [36], which consists of five fully-connected layers with ReLU activation functions. Both the top-1 classification accuracy and SE are

adopted as evaluation metrics. Specifically, the SE is computed based on the true millimeter-wave channel $\boldsymbol{h} \in \mathbb{C}^{N_t \times K}$ and the beamforming vector $\boldsymbol{w}_b \in \mathbb{C}^{N_t}$ from the DFT codebook corresponding to the beam index b predicted by each model, where N_t and K denote the number of antennas and subcarriers, respectively. The SE can be expressed as

$$R = \sum_{k=1}^{K} \log_2 \left(1 + \frac{|\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{w}_b|^2}{\sigma_n^2} \right), \tag{14}$$

where $\boldsymbol{h}_k^{\mathrm{H}}$ is kth column of $\boldsymbol{h}$. More training details are provided in Table 11.

Wireless Localization. We evaluate wireless localization on the dichasus-005x scenario [24] using 1,200 sample pairs consisting of CSI and user 2D positions. The system operates at the center frequency of 1.272 GHz with a bandwidth of 50 MHz. Each CSI sample spans 64 OFDM subcarriers in frequency and a 4×8 UPA in space. The dataset is divided into 900, 100, and 200 samples for training, validation, and testing. Wireless foundation models serve as an encoder for feature extraction. For the task-specific AI models, we adopt WiT [37], which utilizes multi-head self-attention to capture long-range dependencies across subcarriers and antenna elements jointly, learns discriminative representations from raw CSI, and regresses positions via a lightweight head. Root-mean-square error (RMSE) is adopted as the metric of the localization accuracy, and training details are provided in Table 11.

Vision-aided Frequency-domain Channel Prediction. In the urban cross-road scenario of the SynthSoM [38] dataset, we perform vision-aided frequency-domain channel prediction. The dataset comprises 240 samples, each containing spatially aligned 2D CSI data and RGB images with a resolution of 720×1280. The channel data is measured at 28 GHz with a 10 MHz bandwidth, including 64 subcarriers and 16 antennas. The dataset is partitioned into training, validation, and test subsets containing 180, 20, and 40 samples, respectively. For WiFo-2, we employ a pretrained ResNet-18 [48] to extract visual features, which are linearly projected and concatenated to the decoder's input token sequence. During fine-tuning, only the output head, the RGB feature projection layer, and the first layer of the decoder are updated. For LWM and SWIT, which lack a channel reconstruction decoder, we concatenate the extracted channel and RGB features and use a linear layer to reconstruct the predicted channel. As a task-specific baseline, we adopt a vision and CSI fusion architecture similar to [39], and the CSI encoder extracts features with the Swin Transformer architecture. We adopt the NMSE, which measures the reconstruction accuracy, and the SE, which directly reflects the capacity of the communication system, as the evaluation metrics. Specifically, we compute the matched-filter precoder $\boldsymbol{P} \in \mathbb{C}^{N \times K}$ from the predicted channel $\hat{\boldsymbol{h}} \in \mathbb{C}^{N \times K}$. Assuming the true channel is $\bar{\boldsymbol{h}} \in \mathbb{C}^{N \times K}$, the resulting SE can be computed as

$$R = \sum_{k=1}^{K} \log_2 \left(1 + \frac{|\bar{\boldsymbol{h}}_k^{\mathrm{H}} \boldsymbol{P}_k|^2}{\sigma_n^2} \right), \tag{15}$$

where $\bar{\boldsymbol{h}}_k^{\mathrm{H}}$ and $\boldsymbol{P}_k$ denote the k-th column of $\bar{\boldsymbol{h}}$ and $\boldsymbol{P}$, respectively. More training details are provided in Table 11.

CSI Feedback. For the CSI feedback task, we constructed both the statistical channel modeling-based dataset and the real-world measurement-based dataset for validation. For the first dataset, we utilize QuaDRiGa [21] to generate the dataset under an Indoor LoS scenario. Each CSI sample corresponds to a carrier frequency of 3.5 GHz and a 1 MHz bandwidth, including 32 subcarriers and 32 antennas. For the second dataset, we derived it by processing the subset 40 of Argos [25], which corresponds to an outdoor scenario. Each CSI sample corresponds to a carrier frequency of 2.4 GHz and a 20 MHz bandwidth, including 32 subcarriers and 32 antennas. Both datasets contain 1,200 antenna-frequency 2D samples, which are split into 900 for training, 100 for validation, and 200 for testing. We adopt the same experimental setting as CsiNet [40]. The transmitter compresses the CSI to 25% and sends it, and the receiver reconstructs the original CSI. Because CSI compression and feedback typically use a two-network design, with an encoder deployed at the transmitter and a decoder deployed at the receiver, the wireless foundation model is employed on the receiver side to denoise the reconstructed CSI. Specifically, we first train CsiNet [40] on the constructed dataset, and then the wireless foundation model further refines the CSI reconstructed by CsiNet to improve reconstruction accuracy. During this process, only the input embedding, output head, and the first transformer block are updated, while most parameters remain frozen. As a task-specific baseline, we adopt the TransNet [41] scheme, which represents the SOTA in compact CSI feedback networks. Similarly, we employ both NMSE and SE as performance metrics, where SE is computed exactly as for the vision-aided frequency-domain channel prediction task. More training details are given in Table 12.

AoA Estimation. For the AoA estimation task, we select the I1_2p4 scenario of DeepMIMO [26] to generate 1,200 samples, where 900, 100, and 200 samples are utilized for training, validation, and testing, respectively. Each sample contains a CSI and a corresponding AoA of its dominant path. The center frequency is 2.5 GHz, and the BS is equipped with a ULA with 32 antennas. Each CSI sample spans a 40 MHz bandwidth with 64 subcarriers. Each wireless foundation model acts as a CSI feature-extraction encoder, followed by a linear layer that outputs the AoA estimate. For the task-specific AI baseline, we build a ResNet-based network similar to [42] for AoA estimation. We adopt the mean absolute error (MAE) between the predicted AoA and the ground truth as the evaluation metric. Details of the training configuration are provided in Table 12.

Cross-band channel prediction. We construct two datasets for evaluation. The first dataset for cross-band channel prediction is generated using QuaDRiGa [21] under the RMa-LoS scenario at 2.5 GHz center frequency. Each sample contains paired 3D CSI from two bands whose center frequencies are 3 MHz apart, where both CSI have 4 time samples and 8 antennas. The second dataset Another dataset is generated according to the dichasus-ca0x dataset with a center frequency of 1.27GHz. The measured CSI is collected in an industrial environment, specifically in a LoS area within the ARENA2036 research factory campus. Each sample contains paired 3D CSI from two bands whose centre frequencies are 12.5 MHz apart, where both CSI have 4 time

samples and 32 antennas. Both dataset consists of 1,200 samples, partitioned into 900 for training, 100 for validation, and 200 for testing. The task aims to infer the latter band's centre-subcarrier CSI from 32 subcarriers of the former band. In this task, each wireless foundation model primarily serves as a feature extractor. The extracted channel representations are linearly mapped to the corresponding target-band channels through a fully connected layer. As a task-specific baseline, we employ RF-Diffusion [37], a powerful diffusion-based generative model for RF data synthesis. It demonstrates strong performance in cross-band channel prediction by effectively capturing temporal–frequency domain dependencies through diffusion processes. NMSE and SE are adopted as evaluation metrics to assess the performance of these models, with SE computed in the same way as in the CSI feedback task. More training details are given in Table 12.

Signal detection. The QuaDRiGa [21] simulator is employed to generate the dataset for signal detection under the UMi-NLoS scenario at 2.5 GHz. A MISO system with single-carrier transmission and quadrature phase shift keying (QPSK)-modulated data symbols is considered. The receiver is equipped with 16 antennas, and the dataset comprises 1,200 samples, partitioned into 900 for training, 100 for validation, and 200 for testing. Each sample contains the received signal, the estimated channel matrix, and the transmitted symbols. To emulate practical system interference, complex Gaussian white noise with an SNR of 10 dB is added to the channel matrix. Wireless foundation models are utilized to extract channel features, while a parallel DNN branch extracts features from the received signal. The two feature vectors are concatenated to form a fused representation, which is then passed through a linear output head to produce the detection results. As a task-specific AI baseline, we adopt OAMPNet [43], a model-driven signal detection network that achieves efficient training and strong performance with minimal parameter overhead, particularly under limited data scenarios. We use the symbol error rate (SER) to quantify the detection accuracy, defined as the ratio of incorrectly detected symbols to the total number of transmitted symbols. More training details are shown in Table 12.

Supplementary Note 5. Ablation Study of WiFo-2

To assess the effectiveness of several key modules in the WiFo-2 design, we conduct an ablation study on the base version of WiFo-2 and evaluate performance both on zero-shot channel reconstruction and downstream tasks via fine-tuning. For zero-shot channel reconstruction, Table 13 reports the NMSE performance on three channel reconstruction tasks of high and low reconstruction ratios with the SNR fixed at 20 dB. For the downstream evaluation, we consider the wireless localization task with RMSE as the metric and the sub-6G to mmWave beam prediction task with top-1 classification accuracy and SE as metrics, and the ablation results are shown in Table 14. It is observed that the version without any ablation achieves the best average performance on both the zero-shot channel reconstruction tasks and downstream tasks, which validates the effectiveness of the proposed key modules.

Replacing the proposed task-specific gating with a unified gate results in a significant performance degradation, including a 5.70 dB increase in the NMSE of zero-shot channel reconstruction, a 32.27% rise in the RMSE of wireless localization, and a 5.69%

decrease in the accuracy of beam prediction. This clearly demonstrates the crucial role of the proposed task-specific gating in handling conflicts among tasks. Removing the random masking pretraining task leads to an overall degradation in performance, demonstrating that this task plays a key role in learning general CSI representations. In addition, removing the load balance loss likewise leads to a significant drop in performance, which can be attributed to an imbalance in expert assignment. Moreover, the dynamic ratios of masking and interpolation denoising used during pretraining are replaced with fixed values for ablation experiments. Specifically, the masking ratios of both the time- and frequency-masked reconstruction tasks are fixed at 25%, and the pilot-symbol placement proportions over (time, antenna, subcarrier) are fixed at (1/4, 1, 1/12) for the interpolation denoising tasks. It is observed that pretraining with fixed ratios severely impairs generalization to unseen ratios and also leads to degraded performance on downstream tasks. Therefore, the proposed dynamic-ratio pretraining is essential for improving the foundation model's generalization across different ratios.

Table 13 Ablation results of WiFo-2 across three zero-shot channel reconstruction tasks. The SNR is fixed at 20 dB, and the reconstruction NMSE results in dB are reported.

	High ratio			Low ratio			
	CP-T	CP-F	CE	CP-T	CP-F	CE	Average
WiFo-2	**-8.809**	-9.254	**-13.027**	-12.271	-12.591	-16.818	**-12.128**
w/ unified gating	-5.433	-4.114	-3.165	-8.847	-7.465	-9.544	-6.428
w/o random masking	-8.709	**-9.306**	-12.289	-12.248	-12.469	-16.602	-11.937
w/o load balance loss	-8.553	-8.581	-12.805	-11.959	-12.287	-16.824	-11.835
w/o dynamic ratio	-3.558	-4.498	-2.567	**-12.619**	**-13.016**	**-16.860**	-8.853

Table 14 Ablation results of WiFo-2 on CSI-related downstream tasks, including wireless localization and sub-6G to mmWave beam prediction.

	Wireless localization	Sub-6G to mmWave beam prediction	
	RMSE (m)	Acc. @1	SE (bps/Hz)
WiFo-2	**1.156**	**84.4%**	**5.910**
w/ unified gating	1.529	79.6%	5.751
w/o random masking	1.324	81.3%	5.789
w/o load balance loss	1.313	82.8%	5.839
w/o dynamic ratio	1.399	81.6%	5.732

Supplementary Note 6. Scaling analysis of WiFo-2

Scaling analysis [49] is crucial for guiding the design of wireless foundation models and the construction of corresponding datasets. Here, the zero-shot channel reconstruction and downstream task performance of WiFo-2 under different model sizes and dataset scales are illustrated in Table 15 and Table 16. For the 2nd, 4th, and 5th rows, the base-version WiFo-2 model is adopted while the number of sub-datasets in the pretraining corpus is varied, with their proportion increasing from 1/4 to 1. It can be observed that, as more sub-datasets are included, the performance on both the zero-shot channel reconstruction tasks and the downstream tasks improves significantly. This indicates that the constructed model can learn general CSI representations from large-scale heterogeneous datasets that are beneficial for generalization. For the 1st, 2nd, and 3rd rows, the full pretraining dataset is used, and the model size is scaled from small to large, where a notable performance improvement is observed. This can be attributed to the stronger modeling capacity of larger models. Consequently, WiFo-2 exhibits strong scalability with respect to both data volume and the number of model parameters. Even for relatively small models, further increasing the amount of pretraining data can still substantially enhance generalization performance, providing a meaningful path toward deploying wireless foundation models on computation-constrained communication devices.

Table 15 Scaling results of WiFo-2 across three zero-shot channel reconstruction tasks. The SNR is fixed at 20 dB, and the reconstruction NMSE results in dB are reported.

Model scale	Pretraining dataset ratio	High ratio			Low ratio			Average
		CP-T	CP-F	CE	CP-T	CP-F	CE	
Small	1	-8.259	-7.986	-12.297	-11.399	-11.101	-16.491	11.255
Base	1	-8.809	**-9.254**	**-13.027**	-12.271	-12.591	-16.818	-12.128
Large	1	**-9.308**	-9.022	-12.936	**-13.552**	**-12.682**	**-16.977**	**-12.413**
Base	1/2	-7.469	-5.490	-10.792	-10.887	-8.556	-14.684	-9.646
Base	1/4	-4.505	-3.393	-11.303	-6.315	-5.223	-14.027	-7.461

Supplementary Note 7. Implementation details of the over-the-air experiments

The constructed wireless video transmission system adopts the 5G NR frame structure for downlink transmission, as illustrated in Fig. 5. Each frame has a duration of 10 ms and consists of 20 slots. Three types of slots are employed: synchronization signal block (SSB) slots for synchronization, physical downlink shared channel (PDSCH) slots for data transmission, and empty slots serving as guard intervals. Each PDSCH slot spans 14 OFDM symbols and 4096 subcarriers, of which 3252 consecutive subcarriers are activated, with a subcarrier spacing of 30 kHz. Within each PDSCH slot, one OFDM symbol is allocated to the demodulation reference signal (DM-RS) for channel estimation, while the remaining 13 OFDM symbols are used for data-bit transmission.

Table 16 Scaling results of WiFo-2 on CSI-related downstream tasks, including wireless localization and sub-6G to mmWave beam prediction.

Model scale	Pretraining dataset ratio	Wireless localization	Sub-6G to mmWave beam prediction	
		RMSE (m)	Acc. @1	SE (bps/Hz)
Small	1	1.236	79.6%	5.738
Base	1	1.156	84.4%	5.910
Large	1	**0.969**	**85.1%**	**5.955**
Base	1/2	1.174	83.6%	5.897
Base	1/4	1.264	82.0%	5.877

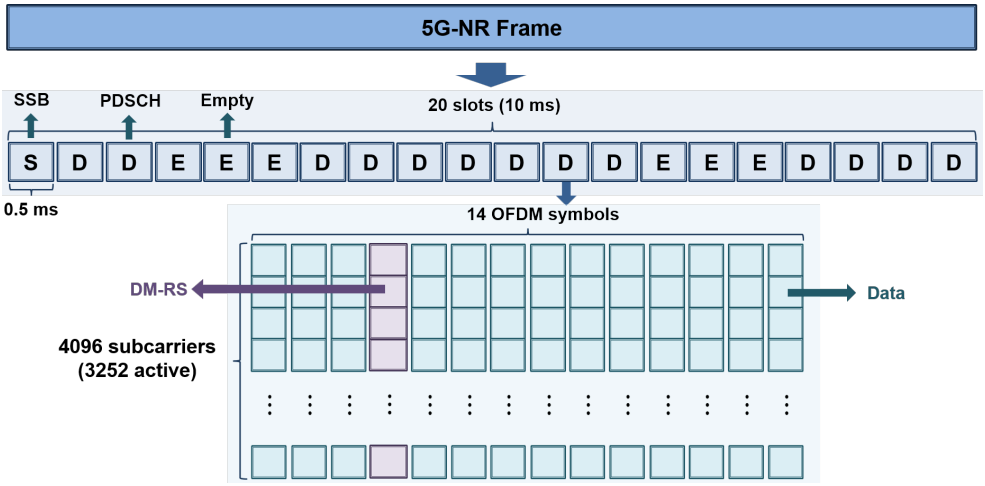

Fig. 5 An illustration of the adopted 5G NR frame structure.

We process every 10 frames as one time–frequency unit, which introduces a buffering delay of 100 ms into the overall system. For model inference, we partition each unit into basic prediction or estimation blocks for parallel processing. For the channel prediction task, we construct 96 basic blocks, and within each basic block, the prediction ratio in the frequency domain is 25%. Specifically, in the frequency domain, the unit is divided into 48 groups, each consisting of 64 subcarriers with a subcarrier spacing of 120 kHz. In the time domain, the unit is divided into 2 groups, each consisting of 32 time instants with a sampling interval of 2.5 ms. For the channel estimation task, we construct 48 basic blocks and emulate sparse channel estimation with pilot densities of 1/8 and 1/24 in the time and frequency domains, respectively. Concretely, in the frequency domain, the unit is divided into 21 groups, each containing 72 subcarriers with a subcarrier spacing of 30 kHz. In the time domain, the unit is divided into 2 groups of 32 time instants with a sampling interval of 2.5 ms. In addition, we repeat

the time-frequency 2D CSI four times along the antenna dimension before feeding them into WiFo-2 for processing.

We adopt the dense architecture of the base version of WiFo-2 for deployment, which contains 16.46M parameters. In the testing scenario, we collect 1,000 CSI samples for fine-tuning, including 500 basic channel prediction blocks and 500 basic channel estimation blocks. We perform 8-bit quantization and calibration using TensorRT. The pre-quantization model achieves a per-token inference latency of 6.250 ms on an NVIDIA GeForce RTX 4090-24G GPU, while the quantized model attains an inference latency of 3.242 ms on the Jetson AGX Orin-64G edge device. The channel prediction and estimation performance remains essentially unchanged before and after quantization. For the task-specific AI models, we employ a three-layer MLP with a hidden dimension of 64, comprising 6k parameters, which is likewise thoroughly trained on the 1,000 CSI samples. In addition, the first-order linear extrapolation is utilized for frequency-domain channel prediction, and the bilinear interpolation is adopted for channel estimation.

References

[1] Choi, H.W., *et al.*: Smart Textile Lighting/Display System With Multifunctional Fibre Devices for Large Scale Smart Home and IoT Applications. Nat. Commun. **13**, 814 (2022)

[2] Cheng, X., *et al.*: Intelligent Multi-Modal Sensing-Communication Integration: Synesthesia of Machines. IEEE Commun. Surv. Tutor. **26**, 258–301 (2024)

[3] Chen, S., *et al.*: Vision, Requirements, and Technology Trend of 6G: How to Tackle the Challenges of System Coverage, Capacity, User Data-Rate and Movement Speed. IEEE Wireless Commun. **27**(2), 218–228 (2020) https://doi.org/10.1109/mwc.001.1900333

[4] Liu, X., Gao, S., Liu, B., Cheng, X., Yang, L.: LLM4WM: Adapting LLM for Wireless Multi-Tasking. IEEE Trans. Mach. Learn. Commun. Netw. **3**, 835–847 (2025) https://doi.org/10.1109/TMLCN.2025.3585845

[5] Liu, B., Liu, X., Gao, S., Cheng, X., Yang, L.: LLM4CP: Adapting Large Language Models for Channel Prediction. J. Commun. Inf. Netw. **9**(2), 113–125 (2024)

[6] Li, Y., *et al.*: Multi-Representation Domain Attentive Contrastive Learning Based Unsupervised Automatic Modulation Recognition. Nat. Commun. **16**, 5951 (2025)

[7] Abramson, J., *et al.*: Accurate Structure Prediction of Biomolecular Interactions With AlphaFold 3. Nature **630**, 493–500 (2024)

[8] Moor, M., Banerjee, O., Shakeri Hossein Abad, Z., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., *et al.*: Foundation Models for Generalist Medical

Artificial Intelligence. Nature **616**, 259–265 (2023)

[9] Wu, K., *et al.*: A Semantic-Enhanced Multi-Modal Remote Sensing Foundation Model for Earth Observation. Nat. Mach. Intell. **7**, 1235–1249 (2025)

[10] Binz, M., *et al.*: A Foundation Model to Predict and Capture Human Cognition. Nature **644**, 1002–1009 (2025)

[11] Xue, B., *et al.*: Deep Spectral Component Filtering as a Foundation Model for Spectral Analysis Demonstrated in Metabolic Profiling. Nat. Mach. Intell. **7**, 743–757 (2025)

[12] Cheng, X., Liu, B., Liu, X., Liu, E., Huang, Z.: Foundation Model Empowered Synesthesia of Machines (SoM): AI-Native Intelligent Multi-Modal Sensing-Communication Integration. IEEE Trans. Netw. Sci. Eng. (2025) https://doi.org/10.1109/TNSE.2025.3587238 . Early Access

[13] Liu, B., Gao, S., Liu, X., Cheng, X., Yang, L.: WiFo: Wireless Foundation Model for Channel Prediction. Sci. China Inf. Sci. **68**, 162302 (2025) https://doi.org/10.1007/s11432-025-4349-0

[14] He, Y., *et al.*: Generalized Biological Foundation Model With Unified Nucleic Acid and Protein Language. Nat. Mach. Intell. (2025) https://doi.org/10.1038/s42256-025-01044-4

[15] Pai, S., *et al.*: Foundation Model for Cancer Imaging Biomarkers. Nat. Mach. Intell. **6**(3), 354–367 (2024) https://doi.org/10.1038/s42256-024-00807-9

[16] Alikhani, S., Charan, G., Alkhateeb, A.: Large Wireless Model (LWM): A Foundation Model for Wireless Channels. arXiv (2024) 2411.08872

[17] Salihu, A., Rupp, M., Schwarz, S.: Self-Supervised and Invariant Representations for Wireless Localization. IEEE Trans. Wireless Commun. **23**(8), 8281–8296 (2024) https://doi.org/10.1109/TWC.2023.3348203

[18] Catak, F.O., Kuzlu, M., Cali, U.: BERT4MIMO: A Foundation Model Using BERT Architecture for Massive MIMO Channel State Information Prediction. arXiv (2025) 2501.01802

[19] Zhao, Z., et al.: CSI-BERT2: A BERT-Inspired Framework for Efficient CSI Prediction and Classification in Wireless Communication and Sensing. arXiv (2024) 2412.06861

[20] Jiang, J., Yu, W., Li, Y., Gao, Y., Xu, S.: A MIMO Wireless Channel Foundation Model via CIR-CSI Consistency. arXiv (2025) 2502.11965

[21] Jaeckel, S., Raschkowski, L., Börner, K., Thiele, L.: QuaDRiGa: A 3-D Multi-Cell Channel Model With Time Evolution for Enabling Virtual Field Trials. IEEE

Trans. Antennas Propag. **62**(6), 3242–3256 (2014)

[22] Huang, Z., *et al.*: A Mixed-Bouncing Based Non-Stationarity and Consistency 6G V2V Channel Model With Continuously Arbitrary Trajectory. IEEE Trans. Wireless Commun. **23**(2), 1634–1650 (2023)

[23] Yaman, I., *et al.*: The LuViRA Dataset: Synchronized Vision, Radio, and Audio Sensors for Indoor Localization. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 11920–11926 (2024)

[24] Euchner, F., Gauger, M., Dörner, S., Brink, S.: A Distributed Massive MIMO Channel Sounder for "Big CSI Data"-Driven Machine Learning. In: WSA 2021; 25th International ITG Workshop on Smart Antennas (2021)

[25] Shepard, C., Ding, J., Guerra, R.E., Zhong, L.: Understanding Real Many-Antenna MU-MIMO Channels. In: 2016 50th Asilomar Conference on Signals, Systems and Computers, pp. 461–467 (2016)

[26] Alkhateeb, A.: DeepMIMO: A Generic Deep Learning Dataset for Millimeter Wave and Massive MIMO Applications. arXiv (2019) 1902.06435

[27] Hoydis, J., Cammerer, S., Ait Aoudia, F., Nimier-David, M., Maggi, L., Marcus, G., Vem, A., Keller, A.: Sionna. https://nvlabs.github.io/sionna/

[28] Jiang, H., Cui, M., Ng, D.W.K., Dai, L.: Accurate Channel Prediction Based on Transformer: Making Mobility Negligible. IEEE J. Sel. Areas Commun. **40**(9), 2717–2732 (2022)

[29] Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast Networks for Video Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)

[30] Soltani, M., Pourahmadi, V., Mirzaei, A., Sheikhzadeh, H.: Deep Learning-Based Channel Estimation. IEEE Commun. Lett. **23**(4), 652–655 (2019)

[31] Luan, D., Thompson, J.S.: Channelformer: Attention-Based Neural Solution for Wireless Channel Estimation and Effective Online Training. IEEE Trans. Wireless Commun. **22**(10), 6562–6577 (2023)

[32] Jiang, W., Schotten, H.D.: Deep Learning for Fading Channel Prediction. IEEE Open J. Commun. Soc. **1**, 320–332 (2020)

[33] Yin, H., Wang, H., Liu, Y., Gesbert, D.: Addressing the Curse of Mobility in Massive MIMO With Prony-Based Angular-Delay Domain Channel Predictions. IEEE J. Sel. Areas Commun. **38**(12), 2903–2917 (2020)

[34] Xiaoming, S., Shiyu, W., Yuqi, N., Dianqi, L., Zhou, Y., Qingsong, W., Jin, M.: Time-MoE: Billion-Scale Time Series Foundation Models With Mixture of

Experts. In: ICLR 2025: The Thirteenth International Conference on Learning Representations (2025). International Conference on Learning Representations

[35] Sun, Z., Wang, K., Sun, R., Chen, Z.: Channel State Identification in Complex Indoor Environments With ST-CNN and Transfer Learning. IEEE Commun. Lett. **27**(2), 546–550 (2023)

[36] Alrabeiah, M., Alkhateeb, A.: Deep Learning for mmWave Beam and Blockage Prediction Using Sub-6 GHz Channels. IEEE Trans. Commun. **68**(9), 5504–5518 (2020)

[37] Salihu, A., Schwarz, S., Rupp, M.: Attention Aided CSI Wireless Localization. In: 2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1–5 (2022)

[38] Cheng, X., *et al.*: SynthSoM: A Synthetic Intelligent Multi-Modal Sensing-Communication Dataset for Synesthesia of Machines (SoM). Sci. Data **12**, 819 (2025) https://doi.org/10.1038/s41597-025-05065-x

[39] Nam, Y., Choi, J.: Multi-Modal Variable-Rate CSI Reconstruction for FDD Massive MIMO Systems. arXiv (2025) 2501.11926

[40] Wen, C.-K., Shih, W.-T., Jin, S.: Deep Learning for Massive MIMO CSI Feedback. IEEE Wireless Commun. Lett. **7**(5), 748–751 (2018) https://doi.org/10.1109/LWC.2018.2818160

[41] Cui, Y., Guo, A., Song, C.: TransNet: Full Attention Network for CSI Feedback in FDD Massive MIMO System. IEEE Wireless Commun. Lett. **11**(5), 903–907 (2022)

[42] Pan, G., Huang, K., Chen, H., Zhang, S., Häger, C., Wymeersch, H.: Large Wireless Localization Model (LWLM): A Foundation Model for Positioning in 6G Networks. arXiv (2025) 2505.10134

[43] He, H., Wen, C.-K., Jin, S., Li, G.Y.: Model-Driven Deep Learning for MIMO Detection. IEEE Trans. Signal Process. **68**, 1702–1715 (2020)

[44] Maaten, L., Hinton, G.E.: Visualizing Data Using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)

[45] He, K., *et al.*: Masked Autoencoders Are Scalable Vision Learners. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 16000–16009 (2022)

[46] Cui, Y., Guo, J., Wen, C.-K., Jin, S., Tong, E.: Leveraging Pre-Trained Large Language Models for CSI Feedback in Massive MIMO Systems. Authorea Prepr.

[47] Yin, H., Wang, H., Liu, Y., Gesbert, D.: Addressing the Curse of Mobility in

Massive MIMO With Prony-Based Angular-Delay Domain Channel Predictions. IEEE J. Sel. Areas Commun. **38**(12), 2903–2917 (2020)

[48] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778 (2016)

[49] Kaplan, J., et al.: Scaling Laws for Neural Language Models. arXiv (2020) 2001.08361