MUSE-FM: Multi-Task Environment-Aware Foundation Model for Wireless Communications

Tianyue Zheng, *Graduate Student Member, IEEE*, Jiajia Guo, *Member, IEEE*, Linglong Dai, *Fellow, IEEE*, Shi Jin, *Fellow, IEEE*, and Jun Zhang, *Fellow, IEEE*

***Abstract*—Recent advancements in foundation models (FMs) have attracted increasing attention in the wireless communication domain. Leveraging the powerful multi-task learning capability, FMs hold the promise of unifying multiple tasks of wireless communication with a single framework. Nevertheless, existing wireless FMs face limitations in the uniformity to address multiple tasks with diverse inputs/outputs across different communication scenarios. In this paper, we propose a MUlti-taSk Environment-aware FM (MUSE-FM) with a unified architecture to handle multiple tasks in wireless communications, while effectively incorporating scenario information. Specifically, to achieve task uniformity, we propose a unified prompt-guided data encoder-decoder pair to handle data with heterogeneous formats and distributions across different tasks. Besides, we integrate the environmental context as a multi-modal input, which serves as prior knowledge of environment and channel distributions and facilitates cross-scenario feature extraction. Simulation results illustrate that the proposed MUSE-FM outperforms existing methods for various tasks, and its prompt-guided encoder-decoder pair facilitates few-shot adaptation to new task configurations. Moreover, the incorporation of environment information improves the ability to adapt to different scenarios.**

***Index Terms*—Foundation models (FMs), multi-task learning, scene graph, environment awareness.**

Received 13 January 2026; revised 20 May 2026; accepted 21 June 2026. Date of publication 8 July 2026; date of current version 10 July 2026. This work was supported in part by the National Science and Technology Major Projects of China under Grant 2025ZD1301800, in part by the National Natural Science Foundation of China under Grant 62325106, and in part by the National Key Research and Development Program of China under Grant 2023YFB3811503. The work of Jiajia Guo was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62401640 and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515110732. The work of Shi Jin was supported in part by the Key Technologies Research and Development Program of Jiangsu (Prospective and Key Technologies for Industry) under Grant BE2023022-1 and Grant BE2023022, in part by the National Natural Science Foundation of China (NSFC) under Grant 62261160576, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242022k60004. The work of Jun Zhang was supported in part by Hong Kong Research Grants Council under the Areas of Excellence scheme under Grant AoE/E-601/22-R and in part by NSFC/RGC Collaborative Research Scheme under Grant CRS_HKUST603/22. The associate editor coordinating the review of this article and approving it for publication was H. T. Dinh. *(Corresponding author: Linglong Dai.)*

Tianyue Zheng and Linglong Dai are with the Department of Electronic Engineering and the State Key Laboratory of Space Network and Communications, Tsinghua University, Beijing 100084, China (e-mail: zhengty22@mails.tsinghua.edu.cn; daill@tsinghua.edu.cn).

Jiajia Guo and Jun Zhang are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: eejiajiaguo@ust.hk; eejzhang@ust.hk).

Shi Jin is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jinshi@seu.edu.cn).

Digital Object Identifier 10.1109/TWC.2026.3708709

I. Introduction

THE integration of artificial intelligence (AI) with wireless communication has become increasingly vital for the design and optimization of sixth-generation (6G) wireless communication networks [1], [2]. Specifically, the integration of AI demonstrates performance improvement or complexity reduction for various communication tasks, including channel estimation and prediction, channel state information (CSI) feedback, channel decoding, and resource allocation under dynamic conditions [3]. The recent paradigm shift towards Foundation Models (FMs), characterized by a large amount of parameters and pretraining on vast datasets, has dramatically reshaped the AI landscape [4]. These models exhibit remarkable reasoning capabilities and generalizability across multiple domains and tasks. Given their profound impact across diverse domains, exploring the applicability and potential of FMs in 6G wireless communications emerges as a crucial research imperative to unlock next-generation wireless intelligence.

A. Prior Works

There is an emerging trend of leveraging FMs to support massive device density, high-data-rate communications, and ubiquitous intelligence in 6G networks [5], [6], [7]. To be specific, existing works mainly have attempted to harness three key advantages of FMs, namely powerful feature extraction, few-shot learning, and generalization capability, to address challenges in wireless communication systems. Firstly, owing to the substantial parameters, FMs acquire superior feature extraction capabilities compared to traditional deep learning (DL) architectures. This inherent advantage has attracted a growing research focus on leveraging these models to address complex problems and scenarios for various tasks in wireless communications, such as channel prediction and scatter generation. Specifically, Liu et al. [8] introduce FM to physical layer communications, and propose LLM4CP to deal with the challenge of channel prediction in high-dynamic scenarios. LLM4CP unleashes the power of FMs to achieve accurate channel prediction. It is proposed in [9] that an FM-enabled decomposition-based multi-objective evolutionary algorithm can jointly optimize the communication and sensing objectives in integrated sensing and communication (ISAC) systems. Reference [10] focuses on beam prediction task for near-field communications, which demonstrates the advantages of FM in handling complex temporal dynamics due to the joint dependence on user angle and distance for near field channels.

1536-1276 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

Besides, FMs are also employed in fluid antenna systems [11] to tackle the challenges of user mobility.

Secondly, pretrained on extensive datasets, FMs possess inherent few-shot/zero-shot learning capabilities, compared to conventional DL-based models. The artificial general intelligence acquired by the FMs through pretraining enables direct task inference without the need to design and train dedicated networks from scratch. For example, in [12], the authors adopt a pretrained FM to perform power allocation using a few-shot learning approach. In this method, the channel gains and the corresponding transmit power strategies are contacted as prompts and fed to the FM. Then, the FM can provide the power allocation strategy without any fine-tuning.

Thirdly, FMs exhibit better generalization capability to different wireless environments, compared to traditional DL models. For instance, leveraging this generalization capability, Guo et al. [13] propose an FM-based CSI feedback method that is adaptable to various scenarios. The authors incorporate the channel distribution in different environments as input and enable the model to handle multiple scenarios simultaneously. Besides, Zhou et al. [14] introduce a predictive-foundation-model-based channel estimation framework trained on large-scale cross-domain data to extract universal channel representations and provide predictive priors with strong cross-scenario transferability, achieving significantly superior generalization performance across diverse configurations.

It has been observed that FMs have successfully improved the performance and robustness in various wireless communication tasks [8], [9], [10], [11], [12], [13]. Nevertheless, existing works typically design a dedicated FM for each individual task. Consequently, developing separate FMs for all the aforementioned tasks would incur substantial storage and deployment overhead. Recently, a few initial attempts have been made to propose a multi-task FM to unify multiple tasks with a single model. In [15], the authors propose an FM to extract features of wireless channels, which are used as common inputs of downstream models in downstream tasks. There are also studies [16], [17] focusing on channel-associated tasks such as channel estimation, beam selection and path loss estimation. They employ an FM to jointly learn a unified representation for multiple tasks, with task-specific adapters or fine-tuning for different downstream tasks. The authors in [18] attempt to integrate tasks with different data formats and distributions to fully demonstrate the multitasking capability of FMs. For this purpose, a multi-task FM is proposed to unify three tasks, namely channel prediction, signal detection and multi-user precoding, with the aid of task-specific adapters and elaborately designed prompts. Moreover, recent advances in multi-modal learning and aggregation techniques for FMs have also shown great potential in supporting heterogeneous data sources and collaborative knowledge extraction in 6G networks [19].

B. Motivations

Despite preliminary advances, existing wireless FMs face limitations in uniformity in addressing multiple tasks with diverse inputs/outputs in different communication scenarios.

1) Limitation in Task Uniformity: Existing works suffer from the narrow task scope, as they predominantly focus on channel-related tasks. These works utilize pilot signals or CSI as input, employing FMs to extract features for multiple downstream tasks. However, they fail to unify other critical components across the entire physical layer functionalities. Besides, existing works employ task-specific data encoders and decoders to handle diverse input/output formats and feature spaces across different tasks and parameters. However, given the diverse tasks and parameter variations, maintaining separate data encoder-decoder pairs for each task and parameter incurs high computational and deployment overhead. It also degrades the scalability of the model, contradicting the objectives of unified multi-task FM paradigms.

2) Limitation in Scenario Uniformity: Scenario uniformity refers to the environmental adaptability of the model to generalize across multiple scenarios, which is critical for real-world deployment. However, existing multi-task FMs are typically optimized for a specific scene. When encountered a new scenario with a different spatial structure, the shift in distribution of channel characteristics leads to severe performance distortion. In this case, data samples in the new environment are required to be collected and online fine-tuning is then performed, which imposes substantial computational overheads especially for FMs with large-scale parameters. Therefore, existing works significantly compromise the environmental adaptability of FMs, and increase deployment costs.

Furthermore, under **data scarcity**, task-specific wireless FMs trained on limited single-task data are prone to overfitting, resulting in performance drop [17]. This also motivates the design of multi-task wireless FMs that leverage intertask relationships to extract knowledge from diverse datasets, enabling superior joint optimization.

C. Our Contributions

To address the aforementioned limitations, in this work, we propose a MUlti-taSk Environment-aware FM (MUSE-FM) for wireless communications with a unified architecture to handle multiple tasks and incorporate scenario information. Specifically, the contributions are summarized as follows.

- A multi-task wireless FM is proposed to enhance the uniformity to address different tasks with heterogeneous data formats and distributions. It is achieved by a unified prompt-guided data encoder-decoder pair. To be specific, the parameters of the prompt-guided unified data encoder-decoder pair are dynamically generated by task-specific instructions through hypernetworks, allowing instruction to "guide" the feature extraction of data encoders (or output generation of data decoders).
- In addition, to improve the environmental uniformity, we integrate the environmental context as a multi-modal input. Specifically, prior knowledge of environment and channel distributions is obtained from scene graphs and facilitates cross-scenario feature extraction. Thus, we establish a multi-task environment-aware FM that dynamically adapts to varying propagation environments.
- A multi-task multi-scenario multi-modal dataset for physical layer communications is constructed to verify the

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

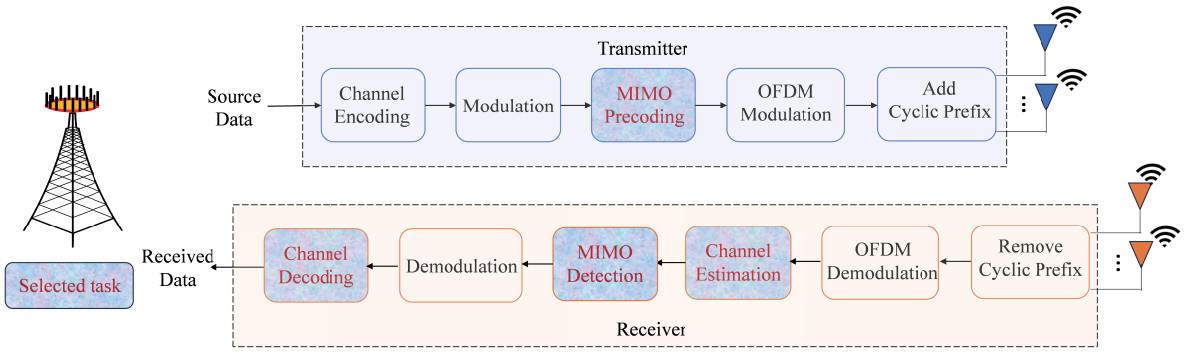


Fig. 1. Typical structure of MIMO-OFDM transmitter and receiver.

effectiveness of the proposed method. The dataset contains 2600 indoor scenarios with different layouts, in each of which 50 data samples are generated. The samples involve environmental information and data for various tasks for physical layer communications.

- Extensive experiments have been conducted to evaluate the performance of the proposed method. It outperforms baselines for various tasks. Besides, incorporating environmental information facilitates cross-scenario learning ability. The proposed prompt-guided unified encoder-decoder pair improves the scalability of the model while achieving comparable performance with task-specific encoder-decoder pairs. Furthermore, with limited data samples, MUSE-FM harnesses cross-task dependencies for knowledge transfer and enhances joint optimization compared with task-specific FMs.

The rest of the paper is organized as follows. Section II introduces the system model. We illustrate the typical structure of MIMO-OFDM transceivers and identify the essential yet challenging tasks in the transceivers. Then, the selected tasks are formulated, respectively. In Section III, we elaborate on the overall framework and specific design of the proposed MUSE-FM. The loss function and the training schedule are then specified. Simulation results are provided in Section IV. Finally, Section V concludes this paper.

Notation: $\mathbf{a}^T$, $\mathbf{a}^H$ denote the transpose, conjugate transpose of $\mathbf{a}$, respectively; $|\mathbf{a}|$ denotes the absolute value of $\mathbf{a}$ while $\|\mathbf{a}\|_2$ denotes the l_2 norm of $\mathbf{a}$; $\mathbb{R}$,$\mathbb{C}$ denote the set of real numbers and complex numbers, respectively; $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the probability density function of complex multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. "ln(BER)" denotes the natural logarithm of BER.

II. Systems Model

A multi-user (MU) multiple-input-single-output (MISO)-orthogonal-frequency-division-multiplexing (OFDM) system operating at mmWave frequency with M subcarriers is considered, where a base station (BS) simultaneously serves K users (UEs). The BS adopts $N_t = N_h \times N_v$ antennas arranged in a uniform planar array (UPA) configuration, where N_h and N_v denote the quantity of antennas along the horizontal and vertical dimensions, respectively. UEs are equipped with an omnidirectional antenna and can accommodate multiple antennas through parallel processing. In this work, we aim to deploy an FM at the BS to simultaneously handle most of the basic tasks for the transceiver. In the following, we first present a brief description of the typical structure of MIMO-OFDM transceivers and identify the basic and challenging tasks that require the employment of FM. Secondly, we present the problem formulations of the selected tasks sequentially.

A. The Structure of MIMO-OFDM Transceivers

The typical structure of a MIMO-OFDM transceiver, deployed at the BS, is depicted in Fig. 1. The MIMO-OFDM transmitter begins with channel encoding (e.g., LDPC or polar codes), which adds redundancy to the input bitstream $\mathbf{b}$ for robustness against channel errors. The encoded bits $\mathbf{s}$ are then mapped to complex symbols by modulation (e.g., QAM or PSK). For multi-user systems, multi-user precoding is applied to minimize inter-user interference and simultaneously transmit $\mathbf{x}_k, k = 1, 2, \cdots, K$ for K users, based on the multi-user CSI. The modulated symbols are subsequently allocated to different OFDM subcarriers, added by a cyclic prefix (CP) to mitigate inter-symbol interference (ISI) caused by multi-path fading. The data transmitted to the k-th user of the m-th subcarrier, $m = 1, 2, \cdots, M$, is referred to as $x_{m,k}$.

Due to the sparse scattering environment of the mmWave frequency [20], we adopt the Saleh-Valenzuela channel Model for the multi-path channel $\mathbf{h}_{m,k} \in \mathbb{C}^{N_t \times 1}$ between the BS and the user k at the m-th subcarrier as

$$\mathbf{h}_{m,k} = \sum_{l=1}^{L_k} \beta_{m,k,l} e^{-j2\pi f_m \tau_{k,l}} \mathbf{a}_m(\theta_{k,l}, \phi_{k,l}), \tag{1}$$

where L_k is the number of paths, and f_m is the frequency at the m-th subcarrier. $\beta_{m,k,l}$, $\tau_{k,l}$, $\theta_{k,l}$ and $\phi_{k,l}$ represent the complex path gain, delay, elevation angle, and azimuth angle of the l-th path of user k. $\mathbf{a}_m(\theta, \phi)$ represents the steering vector of the corresponding path, which is derived for UPA as

$$\mathbf{a}_m(\theta, \phi) = \frac{1}{\sqrt{N_t}} [1, \cdots, e^{jk_m d(n_v \sin\phi \sin\theta + n_h \cos\theta)}, \cdots, e^{jk_m d((N_v-1)\sin\phi\sin\theta + (N_h-1)\cos\theta)}]^T, \tag{2}$$

where $k_m = \frac{2\pi}{\lambda_m}$, λ_m, and d denote the wavenumber, wavelength of the m-th subcarrier, and spacing of adjacent antenna elements, respectively.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

Next, we illustrate the receiver of the BS. The BS antenna array simultaneously receives the transmitted symbols from K users. Then CP is removed to eliminate ISI and OFDM demodulation is performed. During the pilot transmission, the BS estimates the channel $\hat{\mathbf{H}}$. Based on the estimated channel, MIMO detection is performed to recover the transmitted symbol $\hat{\mathbf{x}}$ during data transmission. The detected symbols undergo demodulation to recover the encoded bits $\hat{\mathbf{s}}$ according to the modulation scheme. Finally, channel decoding corrects transmission errors by exploiting redundancy from the encoder, restoring the original bitstream $\hat{\mathbf{b}}$.

Building upon the aforementioned description of the MIMO-OFDM transceiver, we select critical yet challenging tasks including *channel estimation*, *MIMO detection*, *channel decoding*, and *multi-user precoding*, while ignoring the relatively trivial components (e.g., channel encoding, which requires only a matrix multiplication operation). The task selection aims to establish a unified FM capable of addressing the most demanding aspects of a transceiver. Furthermore, we incorporate *user localization* as an essential task, given its fundamental role in 6G ISAC systems, a pivotal direction for next-generation wireless networks. The designed multi-task FM aims to handle tasks across various physical layers as comprehensively as possible. In the next subsection, we will formulate the selected tasks individually.

B. Problem Descriptions of Selected Tasks

1) Channel Estimation: In MIMO-OFDM systems, the accuracy of channel estimation significantly affects the subsequent signal detection and demodulation process [21]. To obtain CSI, pilots are transmitted and the corresponding received signal $\mathbf{Y}_k^p$ of the k-th user is written as

$$\mathbf{Y}_k^p = \mathbf{W}_s \mathbf{H}_k + \mathbf{N}_k, \tag{3}$$

where $\mathbf{H}_k = [\mathbf{h}_{1,k}, \mathbf{h}_{2,k}, \cdots, \mathbf{h}_{M,k}] \in \mathbb{C}^{N_t \times M}$ is the channel of the M subcarriers for the user k, $\mathbf{W}_s \in \mathbb{C}^{L_p \times N_t}$ is the selection matrix at the BS with pilot length L_p, $\mathbf{N}_k$ is the additive Gaussian white noise (AWGN). A channel estimator is designed to construct the CSI $\mathbf{H}_k, \forall k$ from the received signal $\mathbf{Y}_k^p \in \mathbb{C}^{L_p \times M}$, that is.

$$\textbf{P1:} \min_{\Omega_{ce}} \quad ||\hat{\mathbf{H}}_k - \mathbf{H}_k||_2, \forall k \tag{4}$$

$$\text{s.t.} \quad \hat{\mathbf{H}}_k = f_{\Omega_{ce}}(\mathbf{Y}_k^p), \tag{5}$$

where the estimated channel is obtained from the neural network with a mapping function $f_{\Omega_{ce}}$. The neural network takes $\mathbf{Y}_k^p$ as input, with learnable parameters Ω_{ce}.

2) MIMO Detection: During uplink data transmission, the BS antenna array simultaneously receives the symbols transmitted by the K users. Specifically, denote the transmitted symbol vector by all users at the m-th subcarrier as $\mathbf{x}_m = [x_{m,1}, x_{m,2}, \cdots, x_{m,K}] \in \mathbb{C}^{K \times 1}$. The received signal $\mathbf{y}_m \in \mathbb{C}^{N_t \times 1}$ is given by

$$\mathbf{y}_m = \mathbf{H}_m \mathbf{x}_m + \mathbf{n}_m, \tag{6}$$

where $\mathbf{H}_m = [\mathbf{h}_{m,1}, \mathbf{h}_{m,2} \cdots, \mathbf{h}_{m,K}]$ is the uplink channel of the K users and $\mathbf{n}_m \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_{N_\mathrm{T}})$ is the AWGN at the m-th subcarrier.

The BS will recover the signals $\mathbf{x}_m$ from the received signals $\mathbf{y}_m$ and $\hat{\mathbf{H}}_m$. For the MIMO detection task, we can independently recover the signal of different subcarriers. We adopt the minimum mean squared error (MMSE) estimator to formulate the associated MIMO detection problem as

$$\textbf{P2:} \min_{\Omega_{det}} \quad ||\hat{\mathbf{x}}_m - \mathbf{x}_m||_2, \forall m \tag{7}$$

$$\text{s.t.} \quad \hat{\mathbf{x}}_m = f_{\Omega_{det}}(\hat{\mathbf{H}}_m, \mathbf{y}_m), \tag{8}$$

where $f_{\Omega_{det}}$ is the mapping function with parameters Ω_{det}.

3) Multi-User Precoding: Multi-user precoding aims to maximize the sum rate of K users through the optimization of the transmit precoders, based on the multi-user channel. The total power of all precoding vectors is limited due to the BS power budget. For simplicity, we parallelize the precoding for different subcarriers. That is, we design the precoder of each subcarrier separately. Therefore, the problem is mathematically formulated as

$$\textbf{P3:} \max_{\Omega_{precoding}} \quad \sum_{k=1}^{K} \log_2(1 + \gamma_{m,k}), \forall m \tag{9}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \|\mathbf{w}_{m,k}\|^2 \le P_{\max}, \tag{10}$$

$$\mathbf{W}_m = f_{\Omega_{precoding}}(\hat{\mathbf{H}}_m), \tag{11}$$

where $\mathbf{W}_m = [\mathbf{w}_{m,1}, \mathbf{w}_{m,2}, \cdots, \mathbf{w}_{m,K}]$ is the designed precoders, $\hat{\mathbf{H}}_m = [\hat{\mathbf{h}}_{m,1}, \hat{\mathbf{h}}_{m,2}, \cdots, \hat{\mathbf{h}}_{m,K}]$ is the estimated channel of K users, and $P_{\max}$ is the power budget. The precoder $\mathbf{W}_m$ is learned from the neural network with the mapping function $f_{\Omega_{precoding}}$, where $\Omega_{precoding}$ denotes the variable parameters. Besides, $\gamma_{m,k}$ represents the received signal-to-interference-plus-noise ratio (SINR) at user k:

$$\gamma_{m,k} = \frac{\left|\mathbf{h}_{m,k}^H \mathbf{w}_{m,k}\right|^2}{\sum_{k'=1,k' \neq k}^{K} \left|\mathbf{h}_{m,k}^H \mathbf{w}_{m,k'}\right|^2 + \sigma^2}, \tag{12}$$

where σ^2 represents the variance of AWGN.

4) Channel Decoding: Channel decoding aims to recover the information bitstream $\mathbf{b} \in \{0,1\}^m$ from the estimated encoded bitstream $\hat{\mathbf{s}}$, through a neural network with a parameterized mapping function $f_{\Omega_{decoding}}$. The recovered $\hat{\mathbf{s}}$ can be modeled as $\hat{\mathbf{s}} = \mathbf{s} + n_s$, with $\mathbf{s} \in \{\pm 1\}^n$ and n_s being noise. To achieve efficient channel decoding and facilitate fair comparison, we follow the pre-processing, post-processing, and problem descriptions of ECCT [22]. To be specific, the pre-processing process replaces $\hat{\mathbf{s}}$ with a vector of dimensionality $2n - m$ defined as

$$\tilde{\mathbf{s}} = [|\hat{\mathbf{s}}|, h(\hat{\mathbf{s}})], \tag{13}$$

where $|\hat{\mathbf{s}}|$ signifies the absolute value of $\hat{\mathbf{s}}$. $h(\hat{\mathbf{s}}) = \mathbf{P} \cdot \mathrm{bin}(\mathrm{sign}(\hat{\mathbf{s}})) \in \{0,1\}^{n-m}$ is the syndrome from hard-decoded $\hat{\mathbf{s}}$, where $\mathbf{P} \in \mathbb{R}^{(n-m) \times n}$ is the parity check matrix. Then $\tilde{\mathbf{s}}$ is fed to the neural network and the predicted noise $\hat{\mathbf{z}}$ is obtained (we suppose $\hat{\mathbf{s}} = \mathbf{s} \cdot \mathbf{z}$). In the post-processing step, the predicted noise $\hat{\mathbf{z}}$ is multiplied by $\hat{\mathbf{s}}$ to recover $\mathbf{b}$. That is, the predicted $\hat{\mathbf{b}}$ takes the form:

$$\hat{\mathbf{b}} = \mathrm{bin}(\mathrm{sign}(\hat{\mathbf{s}} \cdot \hat{\mathbf{z}})). \tag{14}$$

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

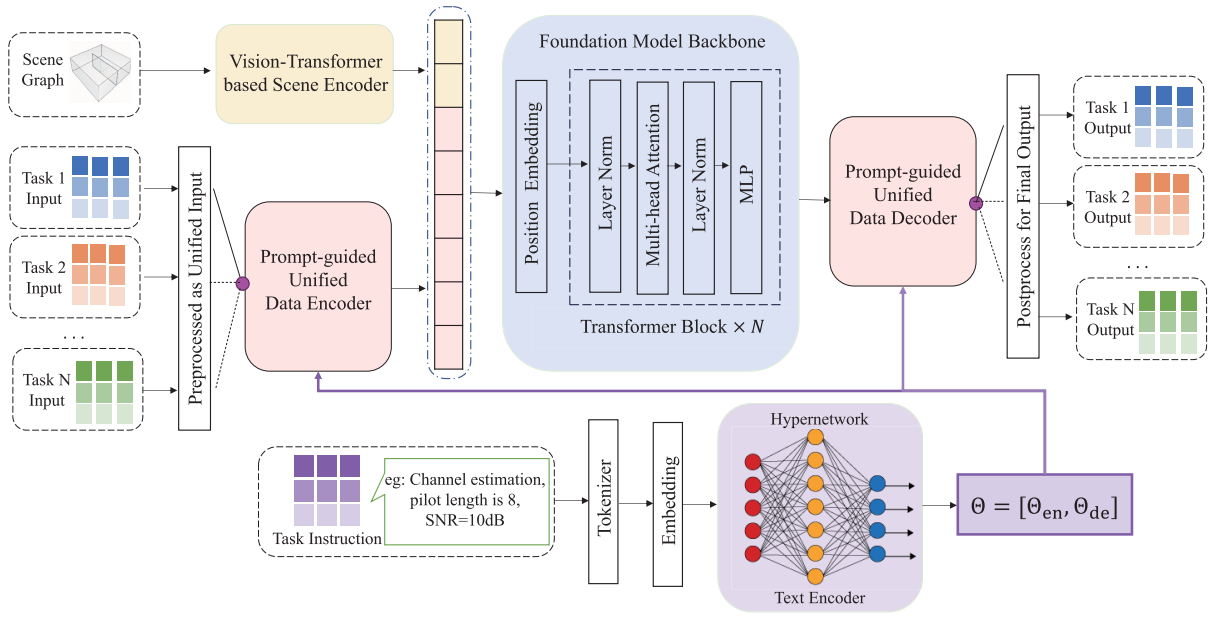


Fig. 2. The model framework of our method.

We employ binary cross-entropy as the loss function with the objective of learning to predict multiplicative noise $\mathbf{z}$. The corresponding target binary multiplicative noise is denoted as $\tilde{\mathbf{z}} = \text{bin}(\text{sign}(\mathbf{z})) = \text{bin}(\text{sign}(\hat{\mathbf{s}} \cdot \mathbf{s}))$. Therefore, the problem can be modeled as

$$\textbf{P4:} \min_{\Omega_{decodig}} \quad -\sum_{i=1}^{n} \tilde{z}_i \log(\hat{z}_i) + (1-\tilde{z}_i)\log(1-\hat{z}_i), \tag{15}$$

$$\text{s.t.} \quad \hat{\mathbf{z}} = f_{\Omega_{decoding}}(\tilde{\mathbf{s}}), \tag{16}$$

where $\tilde{z}_i$ and $\hat{z}_i$ are the i-th element of $\tilde{\mathbf{z}}$ and $\hat{\mathbf{z}}$, respectively.

5) User Localization: In 6G networks, ISAC emerges as a transformative paradigm, merging sensing and communication, where precise user localization is essential to enable dynamic resource allocation and intelligent network control. Therefore, in this paper, we not only attempt to estimate the channel from $\mathbf{Y}_k^p$, but also learn the user positions to facilitate subsequent ISAC applications. Similar to channel estimation, the user localization problem can be formulated as

$$\textbf{P5:} \min_{\Omega_{loc}} \quad ||\hat{\mathbf{pos}}_k - \mathbf{pos}_k||_2, \forall k \tag{17}$$

$$\text{s.t.} \quad \hat{\mathbf{pos}}_k = f_{\Omega_{loc}}(\mathbf{Y}_k^p), \tag{18}$$

where $\mathbf{pos}_k$ and $\hat{\mathbf{pos}}_k$ represent the actual and estimated position, respectively. Besides, Ω_{loc} is the trainable parameters of the mapping function $f_{\Omega_{loc}}$.

The problem formulations above highlight that the selected tasks exhibit substantial heterogeneity in input data formats, feature distributions, as well as distinct functional requirements and technical challenges. Consequently, an effective multi-task handling strategy is essential to jointly address these diverse tasks within a unified framework.

III. Multi-Task Environment-Aware Foundation Model (MUSE-FM)

In this section, we first introduce the overall framework of the proposed multi-task environment-aware FM (MUSE-FM). Then, the specific designs of the components in the framework are elaborated, respectively. Finally, the loss function and training schedule are illustrated.

A. Overall Framework

The framework of the proposed model is presented in Fig. 2, which includes the scene encoder, prompted-guided unified encoder-decoder pair, and FM backbone. To address scenario limitation, leveraging the strong correlation between physical environments and channel characteristics [23], we attempt to incorporate environmental information to allow cross-scenario generalization in multi-task FM. Specifically, the spatial structure of the physical environment, including the area building distribution as well as the BS position, determines the propagation path information of the wireless channel, and thus directly impacts channel characteristics. Therefore, involving the environmental knowledge as a multi-modal input provides a promising way to inform FMs of the channel distribution and improves the generalization ability in different scenarios. We implement it through a Vision Transformer-based encoder that extracts environmental and channel prior knowledge from scene graphs. The task-agnostic prior knowledge is subsequently fused with wireless data to enhance the model's environmental adaptability. The specific design of the scene encoder is described in Section III-B.

To handle data with heterogeneous formats and distributions across multiple tasks, we propose a prompt-guided unified data encoder whose parameters are dynamically generated by task-specific instructions through hypernetworks. In particular, the task instruction is fed to a hypernetwork to dynamically generate parameters, which is utilized in the data encoder. This architecture allows the instruction to "guide" the encoder to adaptively extract task-specific features according to each task's objectives and data characteristics. To achieve this, two key components are included: **a parameter generator**,

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

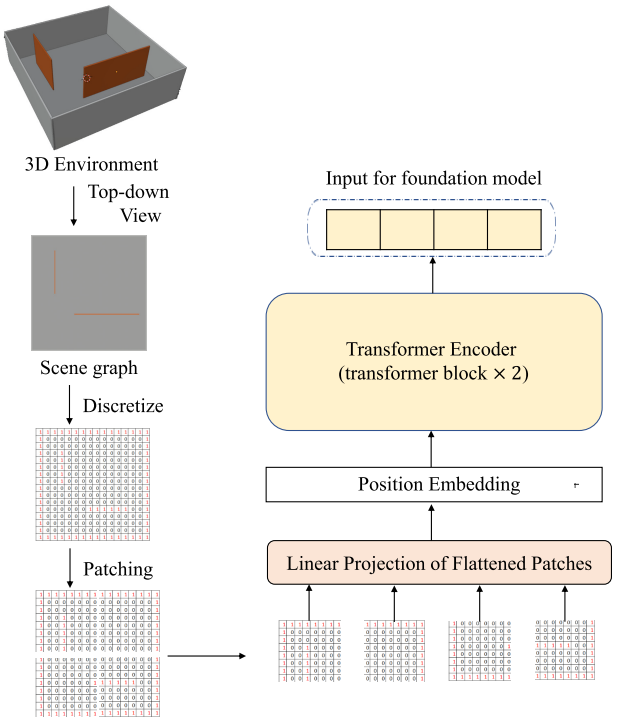


Fig. 3. Vision-transformer based scene encoder.

i.e., which is used to generate task-specific parameters from task instruction via a text encoder; **a prompted-guided unified encoder-decoder pair**, to utilize the generated parameters to project the data of multiple tasks into the shared FM space (or generate the final outputs in multiple tasks). The two components are elaborated in Section III-C and III-D.

The token embeddings from scene graphs and wireless data (e.g., CSI and received symbols) are concatenated to form the input embeddings for the FM backbone. Future extensions could incorporate advanced multi-modal learning and aggregation techniques [19] to further enhance feature fusion and knowledge transfer across diverse wireless tasks and scenarios. The FM backbone consists of position embedding and stacked classical transformer blocks, which is illustrated in Section III-E. Equipped with large-scale parameters and a layered structure, the FM backbone exhibits expressive learning capacity for multiple tasks. Besides, atthention mechanism enables effective fusion of multi-modal inputs. It assimilates related channel features from task-agnostic prior knowledge for different tasks and integrates it with wireless data, ensuring cross-scenario feature extraction. Finally, similar to the data encoder, equipped with task-aware parameters, the prompt-guided unified decoder generates the output for different tasks.

B. Scene Encoder

The scene encoder aims to extract task-agnostic priors of channel and environment from scene graphs. The structure of the scene encoder to effectively exploit environmental information is illustrated in Fig. 3. Direct utilization of raw 3D environmental data is challenging due to its high dimensionality. We therefore implement **pre-processing**, transforming environmental information into compact representations suitable for encoder input. Given that top-down views holistically encapsulate the spatial structure of the environment, that is, scene layouts and obstacle distributions, we extract it to construct the scene graph [24]. This scene graph is then discretized into a matrix form, where the value 1 represents the obstacles, and the value 0 represents free space.

Obtaining the discretized scene graph, denoted by $\mathcal{G} \in \mathbb{R}^{W\times W}$, a **vision transformer-based scene encoder** is performed to extract prior channel knowledge from the physical environment. Specifically, to handle the 2D scene graph, we reshape $\mathcal{G}$ into a sequence of flattened patches $\mathcal{G}_p \in \mathbb{R}^{L_s\times P^2}$, where (P, P) is the resolution of each image patch, and $L_s = W^2/P^2$ is the resulting number of patches. Then a trainable linear projection is utilized to map the flattened patches to the feature dimension of transformer encoder D. We refer to the output of this projection as the patch embeddings $\mathrm{X}_s^{\mathrm{patch}} \in \mathbb{R}^{L_s\times D}$:

$$\mathrm{X}_s^{\mathrm{patch}} = \mathrm{Linear}(\mathcal{G}_p). \tag{19}$$

Position embeddings are added to patch embeddings to retain positional information, and the resulting sequence of embedding $\mathrm{X}_s^{\mathrm{input}}$ serves as input to the transformer encoder. The Transformer encoder [25] consists of alternating layers of multi-head self-attention and multi-layer perceptron (MLP) blocks. Layer normalization (LN) is applied before every block, and residual connections are added after every block. The transformer encoder is summarized as

$$\mathrm{X}_s^{\mathrm{emb}} = \mathrm{Transformer}(\mathrm{X}_s^{\mathrm{input}}), \tag{20}$$

where $\mathrm{X}_s^{\mathrm{emb}}$ is then fed to the FM as part of the input.

C. Parameter Generator With Task Instruction

1) Design of Task Instruction: Concise and informative task instruction is essential to generate task-related parameters and thus "guide" the feature extraction of the data encoder and decoder. The designed task instruction includes two parts: **a task identifier**, and **key parameters**. The **task identifier** provides a distinct identifier for each task. In this work, the task identifier includes "Channel estimation", "MIMO detection", "Channel decoding", "Multi-user precoding", and "User localization". **Key parameters** aim to involve important domain knowledge as priors to facilitate better understanding of the task. Taking channel estimation as an example, the pilot length and SNR of the received signal are incorporated, i.e., "pilot length is x, SNR = x dB".[1]

2) Hypernetwork: A hypernetwork is designed to generate parameters for another neural network. In this work, we employ a hypernetwork-based architecture [26], which learns a mapping function that dynamically produces task-related parameters for the data encoder and decoder. By incorporating

[1]The designed task instructions for other tasks are "MIMO detection, transmitting antenna number is x, data length is x, SNR is x dB" for MIMO detection; "Channel decoding for polar codes with encoded bit length n = x and information bit length m = x" for channel decoding; "Multi-user precoding, user number is x, SNR is x dB" for multi-user precoding; "User localization, signal length for localization is x, SNR is x dB" for user localization.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

task instructions, the approach allows the data encoder-decoder pair to automatically adapt to diverse communication tasks and parameters. The specific designs are illustrated as follows.

Firstly, the task instruction is tokenized into vocabulary indices by the FM tokenizer (for example, we employ the pretrained tokenizer of GPT2). Then the vocabulary indices are mapped to high-dimensional token embeddings. The token embeddings, denoted as $\mathrm{X}_t^{\mathrm{emb}}$, serve as the input of the hypernetwork. The hypernetwork is an MLP module, consisting of alternating linear layers and ReLU activation functions. The generated task-aware parameters for the data encoder-decoder pair are written as:

$$\Theta = [\Theta_{en}, \Theta_{de}] = \mathrm{MLP}(\mathrm{X}_t^{\mathrm{emb}}), \tag{21}$$

where Θ_{en} and Θ_{de} are respectively the parameters for data encoder and data decoder.

D. Prompted-Guided Unified Data Encoder and Decoder

After the parameter generator produces the task-specific parameters, we focus on how the parameters can be used in the unified encoder and decoder.

1) Pre-Processor and Unified Data Encoder: To enable a unified data encoder to process multi-task data with different formats and distributions, pre-processing is required for each task to preprocess the input data. Since the transformer architecture inherently acquire the capability to process variable-length sequences, pre-processing primarily aims to standardize input feature dimensions via methods like zero-padding and unify distribution through batch normalization. Here, we process the input data to the same feature dimension $2N_t$ and the pre-processing process for the task n can be expressed as:

$$\mathrm{X}_n^{\mathrm{pre}} = \mathrm{Preprocessor}(\mathrm{X}_n^{\mathrm{input}}), \tag{22}$$

where n denotes the task identifier, and $\mathrm{X}_n^{\mathrm{pre}}$ and $\mathrm{X}_n^{\mathrm{input}}$ denote the preprocessed input and the original input, respectively.

For MIMO detection and multi-user precoding, the feature dimension of input real tensors is exactly $2N_t$ (the input of multi-user precoding is multi-user channel $\mathrm{X}_{precoding}^{\mathrm{input}} = \hat{\mathbf{H}}_m \in \mathbb{R}^{K\times 2N_t}, \forall m$; the input of MIMO detection is the concatenation of the multi-user channel and the received data $\mathrm{X}_{det}^{\mathrm{input}} = [\hat{\mathbf{H}}_m, \mathbf{Y}_m] \in \mathbb{R}^{(K+L_d)\times 2N_t}$, where L_d is the data length, $\mathbf{Y}_m = [\mathbf{y}_m^1, \mathbf{y}_m^2, \cdots, \mathbf{y}_m^{L_d}]$ is the concatenated received data of L_d time slots. Thus, batch normalization can be directly performed for the input data to facilitate convergence of network training. For channel estimation and user localization, the input data is received OFDM signal of pilot $\mathrm{X}_{ce}^{\mathrm{input}} = \mathrm{X}_{loc}^{\mathrm{input}} = \mathbf{Y}_k^p \in \mathbb{R}^{M\times 2L_p}, \forall k$, where L_p denotes pilot length. Thus, zero-padding is required to align feature dimension $2N_t$ and then batch normalization is performed. Finally, for channel decoding the one-dimensional input signal $\mathrm{X}_{decoding}^{\mathrm{input}} = \tilde{\mathbf{s}} \in \mathbb{R}^{2n-m}$ is converted to $\mathrm{X}_{decoding}^{\mathrm{pre}} = diag(\mathrm{X}_{decoding}^{\mathrm{input}})$, where $diag(\mathrm{s})$ is a diagonal matrix with s on the diagonal.

After preprocessing, $\mathrm{X}_n^{\mathrm{pre}}$ is fed into the unified encoder. Specifically, the generated parameters Θ_{en} consist of the weight matrix $\mathrm{W}_{en} \in \mathbb{R}^{D\times 2N_t}$ and the bias vector $\mathrm{b}_{en} \in \mathbb{R}^{D\times 1}$. The parameters are applied to project $\mathrm{X}_n^{\mathrm{pre}}$ of different tasks to a shared foundation model feature space:

$$\mathrm{X}_n^{\mathrm{emb}} = \mathrm{Linear}(\mathrm{X}_n^{\mathrm{pre}}; \mathrm{W}_{en}, \mathrm{b}_{en}). \tag{23}$$

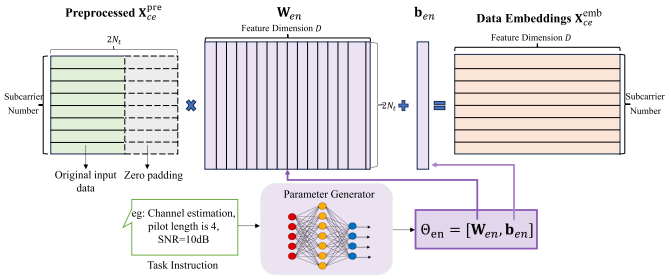


Fig. 4. Illustration of the unified data encoder.

In Fig. 4, we take channel estimation as an example to illustrate how the original data input $\mathrm{X}_n^{\mathrm{pre}}$ is preprocessed and handled by the unified data encoder, obtaining $\mathrm{X}_n^{\mathrm{emb}}$. As illustrated, the received pilot signals first undergo zero-padding and batch normalization to achieve a uniform feature dimension of $2N_t$, after which the prompt-guided unified encoder linearly projects the preprocessed data into the shared foundation model feature space using task-specific weights and bias generated from the task instruction.

Finnaly, the data embeddings $\mathrm{X}_n^{\mathrm{emb}}$ are concatenated with the scene embeddings $\mathrm{X}_s^{\mathrm{emb}}$. Besides, we use the standard approach of adding an extra learnable "classification token" (cls token) to the sequence, to form the final input of the FM backbone. We refer to the input as $\mathrm{X}_{\mathrm{FM}}^{\mathrm{emb}}$.

2) Post-Processor and Unified Data Decoder: Obtaining the output of the FM backbone, denoted as $\mathrm{X}_{\mathrm{FM}}^{O}$, the unified data decoder and the post-processor are employed to generate the final output. The parameters Θ_{de} capture the requirements of the corresponding task, enabling the data decoder to adaptively output the results for different tasks. Similarly, Θ_{de} includes the weight matrix $\mathrm{W}_{de} \in \mathbb{R}^{2N_t\times D}$ and the bias vector $\mathrm{b}_{de} \in \mathbb{R}^{2N_t\times 1}$. We first discard the prefix portion that is associated with the scene, while we only retain the output presentations of the data, $\mathrm{X}_n^{\mathrm{de}}$, for the decoder. The output of the data decoder is written as

$$\mathrm{X}_n^{\mathrm{post}} = \mathrm{Linear}(\mathrm{X}_n^{\mathrm{de}}; \mathrm{W}_{de}, \mathrm{b}_{de}). \tag{24}$$

Finally, the post-processor extracts the corresponding output for the target task, which can be expressed as

$$\mathrm{X}_n^{\mathrm{output}} = \mathrm{Postprocessor}(\mathrm{X}_n^{\mathrm{post}}). \tag{25}$$

For channel estimation and multi-user precoding, the resulted $\mathbf{X}_{ce}^{\mathrm{post}}$ and $\mathbf{X}_{precoding}^{\mathrm{post}}$ have the same feature dimension with required final outputs. Thus, $\mathbf{X}_{ce}^{\mathrm{output}} = \mathbf{X}_{ce}^{\mathrm{post}}[1:,:] = \hat{\mathbf{H}}_k$ is the estimated channel, while $\mathbf{X}_{precoding}^{\mathrm{output}} = \mathbf{X}_{precoding}^{\mathrm{post}}[1:,:] = \mathbf{W}_m$ is the designed beamforming vector. For user localization, we extract the estimated user localization from the cls token, i.e. $\mathbf{X}_{loc}^{\mathrm{output}} = \mathbf{X}_{loc}^{\mathrm{post}}[0,:2] = \hat{\mathbf{pos}}$. For MIMO detection, the recovered data is written as $\mathbf{X}_{det}^{\mathrm{output}} = \mathbf{X}_{det}^{\mathrm{post}}[-L_d:,:2K] = \hat{\mathbf{x}}_m$. Lastly, for channel decoding, the recovered bit stream is presented as $\mathbf{X}_{decoding}^{\mathrm{output}} = \mathbf{X}_{decoding}^{\mathrm{post}}[:,0] = \hat{\mathbf{b}}$.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

E. Foundation Model Backbone

The FM backbone effectively fuses the multi-modal inputs (environmental scene embeddings $\mathrm{X}_s^{\mathrm{emb}}$ and task-specific wireless data) for multi-task learning. The scene graph prior $\mathrm{X}_s^{\mathrm{emb}}$ primarily encodes physical layout features which strongly influence channel distributions. Therefore, it provides significant benefit to channel-related tasks such as channel estimation, and user localization.

Leveraging the Transformer's multi-head self-attention mechanism, it dynamically assigns varying importance weights to different tokens in the shared environmental prior $\mathrm{X}_s^{\mathrm{emb}}$ according to the task instruction and input characteristics. Consequently, the model selectively distills task-beneficial representations from the unified prior knowledge for each individual task. This attention-driven adaptive feature extraction ensures the robustness of the shared prior across heterogeneous tasks while maintaining strong cross-task generalization.

Architecturally, it is composed of stacked Transformer modules, which is elaborated in detail. Position embeddings are added to the combined tokens $\mathrm{X}_{\mathrm{FM}}^{\mathrm{emb}}$, providing sequential information. Thus, the input to the transformer blocks is denoted as $\mathrm{X}_{\mathrm{FM}}^{I(1)}$. As described in the scene encoder, each transformer block has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple MLP network. The l-th transformer block takes the output of the $l-1$-th transformer block as input, that is, $\mathrm{X}_{\mathrm{FM}}^{I(l)} = \mathrm{X}_{\mathrm{FM}}^{O(l-1)}$. LN operates independently across the feature dimension for each token in the sequence and each sample in the batch:

$$\mathrm{X}_{\mathrm{FM}}^{ln1(l)} = \mathrm{LayerNorm}(\mathrm{X}_{\mathrm{FM}}^{I(l)}), \tag{26}$$

where $\mathrm{LayerNorm}(\cdot)$ denotes the LN. Then $\mathrm{X}_{\mathrm{FM}}^{ln1(l)}$ is processed by the multi-head self-attention module with residual connection as

$$\mathbf{X}_{\mathrm{FM}}^{att(l)} = \mathrm{ATT}(\mathbf{X}_{\mathrm{FM}}^{ln1(l)}) + \mathbf{X}_{\mathrm{FM}}^{ln1(l)}, \tag{27}$$

where $\mathrm{ATT}(\cdot)$ denotes multi-head self-attention. Then another LN is applied, obtaining $\mathrm{X}_{\mathrm{FM}}^{ln2(l)}$, which is processed by the MLP module, that is

$$\mathbf{X}_{\mathrm{FM}}^{O(l)} = \mathrm{MLP}(\mathbf{X}_{\mathrm{FM}}^{ln2(l)}) + \mathbf{X}_{\mathrm{FM}}^{ln2(l)}, \tag{28}$$

where $\mathrm{MLP}(\cdot)$ is the MLP network and $\mathbf{X}_{\mathrm{FM}}^{O(l)}$ the output of the l-th transformer block. After progressively extracting and integrating features with L transformer blocks, the output of the FM is denoted as $\mathbf{X}_{\mathrm{FM}}^{O}$.

F. Loss Function and Training Schedule

The proposed MUSE-FM is trained on a mixed dataset with multiple tasks in different scenarios. The multi-task loss function for the training stage is written as

$$\mathrm{Loss}_{train} = \sum_n \alpha_n \mathrm{loss}_{train,n}, \tag{29}$$

where loss_n is the loss function of task n. Besides, due to the disparities in the convergence rates and loss scales across tasks, we implement a weighted linear combination scheme with α_n. Specifically, α_n are determined with Dynamic Weight Average (DWA) algorithm [27], to dynamically adjust each task's weight based on its loss every epoch.

TABLE I
SELECTED TASKS AND THE CORRESPONDING TRAINING/EVALUATION METRICS

Task	Training metric	Evaluation metric
CE	MSE	NMSE
Precoding	Negative sum rate	Sum rate
Det	MSE	NMSE
Decoding	cross entropy	BER
Loc	MSE	Error distance

During training, samples from multiple tasks are jointly presented within each batch to enable cross-task knowledge sharing. For each task, its own input data, together with the corresponding task-specific instruction, is fed into the model and processed independently by the shared FM backbone and the hypernetwork-generated prompt-guided encoder-decoder pair; while the joint exposure across tasks allows the model to learn rich inter-task relationships through parameter sharing. During inference, MUSE-FM operates in an independent way according to the demand of the base station: given any specific task (e.g., multi-user precoding at the transmitter, channel estimation at the receiver, or user localization for ISAC), only the corresponding task instruction and its input data need to be provided, allowing the model to be invoked flexibly using the same unified architecture.

Next, we briefly illustrate the loss function for each task. For multi-user precoding, we directly optimize the negative sum rate in an unsupervised manner, which is presented in (9). For regression problems including channel estimation, user localization, and MIMO detection, the MSE loss is selected as the loss function. For channel decoding, following [22], we utilize the cross entropy loss to optimize the estimated binary multiplicative noise. The selected tasks, the corresponding training metrics and evaluation metrics are summarized in Table I. During the training process, the model with the smallest validation loss is saved for the testing phase. Similar to the training loss, the validation loss is the weighted sum of the losses of individual tasks, i.e.

$$\mathrm{Loss}_{val} = \sum_n \beta_n \mathrm{loss}_{val,n}. \tag{30}$$

where β_n and $\mathrm{loss}_{val,n}$ denote the weight and loss of the task n, respectively. Specifically, NMSE is utilized to evaluate the performance of channel estimation and MIMO detection; while the normalized average positioning error is chosen to provide intuitive evaluation for user localization. For multi-user precoding, the loss function is the negative sum rate; for channel decoding, the loss function is the bit error ratio (BER).

MUSE-FM is designed to handle a wide range of physical-layer tasks. When encountering new parameter configurations (e.g., different pilot lengths, or user numbers), its uniform architecture allows for extension through minimal task-specific adaptation by updating the task instruction and pre-processor

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

according to the new parameters. For a novel task, the current architecture can be used directly in a zero-shot manner. However, the zero-shot performance may be limited. In such cases, we need to finetune the model with a modest amount of data from the new task.

IV. Simulation Results

In this section, extensive numerical simulations are presented to verify the effectiveness of the proposed method. First, we introduce the constructed multi-task multi-scenario multi-modal dataset. The detailed simulation setup and training configurations are then illustrated. Finally, we provide thorough performance evaluation and analysis of MUSE-FM.

A. Dataset Construction

To validate the proposed method, it is necessary to generate a multi-task dataset with corresponding environmental information across multiple scenarios. The constructed dataset consists of data collected in 2600 scenarios; for each scenario, 50 data samples are simulated. The dataset is partitioned into three subsets: 2000 scenarios for training, 300 scenarios for validation, and 300 scenarios for testing.

1) Environmental Information: To enable an environment-aware FM, environmental information should be included in the dataset. In the constructed dataset, we provide three data formats for a specific scenario, namely the **original 3D environment, scene graph, and discretized scene graph**. Specifically, indoor 3D environments are modeled using Blender,[2] a popular tool for creating 3D indoor scenes. All scenes share the same overall dimensions, i.e. $10m \times 10m \times 3m$, where the materials of the floor and the wall are assumed to be concrete and brick, respectively. In each room, $0 \sim 3$ internal marble walls and $0 \sim 2$ marble cylinders are placed. To be specific, the walls are rectangular partitions with randomized orientations, thicknesses uniformly sampled from [0.05, 0.1] m, and lengths from [2, 6] m. The cylinders have radius uniformly drawn from [0.5, 1.1] m. They are randomly placed such that the entire walls and cylinders remain fully within the room boundaries. This randomization of the number, size, orientation, and spatial distribution of obstacles generates rich variability in indoor layouts and propagation environments across all scenarios.

In Fig. 5, we illustrate a typical 3D scene generated by Blender, where the indoor scenario includes two internal walls and a cylinder. Given the computational complexity and high dimensionality of direct 3D environment processing, we extract top-view representations from 3D scenes that preserve both structural layouts and channel characteristics of the scene, denoted as the scene graph. The scene graph representation is further discretized in matrix format 100×100, which is denoted as the discretized scene graph.[3] We assign explicit physical meanings to the matrix elements: 1 indicates obstacle locations while 0 represents open space areas.

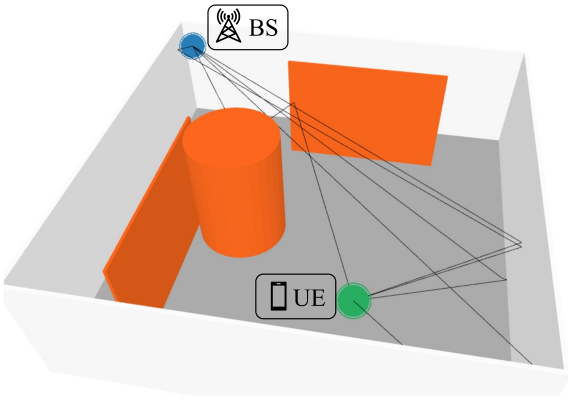


Fig. 5. Illustration of generated scenes and paths simulated by Sionna.

TABLE II
System Parameters During Dataset Generation

Parameter	Value
BS antenna number N_t	$64 = 16 \times 4$
UE antenna number	1
Subcarriers M	48
Center frequency	28 GHz
Subcarrier spacing	18 kHz
UE height	1 m
User number K	4

2) Multi-Task Dataset: Sionna [29], an open-source library for link-level simulations based on TensorFlow, is employed to obtain datasets for multiple tasks in constructed environments. In particular, the BS is randomly placed at $(\pm 4.75m, \pm 4.75m, 2.5m)$, and the element in the scene graph matrix is set as 2 to label the BS position. Nevertheless, the proposed method is generally applicable to base stations located at other positions as well. Besides, the users are randomly distributed in the scenario. Then the Sionna ray tracing simulator [30] is implemented to generate multipath channels in 3D environments, accurately representing propagation conditions by accounting for reflection, diffraction, and scattering. The simulated paths between the BS and the UE are also illustrated in Fig. 5. The parameters of the system are listed in Table II. Based on the generated channels, we can calculate the data samples for different tasks in the MIMO-OFDM transceiver, including the received signal, user positions, channels, etc.

B. Simulation Setup and Training Details

For hyperparameters in network training, we set the batch size as 100, and the initialized learning rate as 0.0001 with a cosine decay scheduler. We utilize Adam optimizer for model training with $\text{betas} = (0.9, 0.999)$. The model undergoes training for 500 epochs over the multi-task dataset. The smallest version of GPT-2 [4] with feature dimension $D = 768$ is adopted as the backbone, where all parameters are trained

[2] https://www.blender.org/

[3] Since DL models typically require data of uniform dimensions, we employ uniformly sized scenes and scene graphs in this study. However, it is important to note that our proposed method is applicable to scenes of varying sizes. In such cases, top-view maps are converted into square pixel images of the same size without altering their spatial characteristics [28]. Specifically, maps of different physical extents are first padded to form square shapes, then rescaled into uniform-size pixel images.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

TABLE III
PERFORMANCE OF TASK-SPECIFIC AND MULTI-TASK FM WITH/WITHOUT ENVIRONMENT INFORMATION

Method	CE (NMSE ↓)	Precoding (Rate ↑)	Det (NMSE ↓)	Decoding (ln(BER) ↓)	Loc (Error Dis ↓)
FM(s)	0.1387	19.44	0.001829	-10.34	0.1516
Env-aware FM(s)	**0.0982**	**19.71**	**0.001632**	**-10.54**	**0.1322**
Improvements	1.50 dB	1.38%	0.49 dB	0.20 dB	12.80%

to obtain a FM dedicated for wireless communications. All training and inference of the proposed model is conducted on four NVIDIA GeForce RTX 4090 24GB GPUs. To validate the superiority of the proposed method, several baselines are implemented which can be categorized into the following groups. To enhance the model's ability to process scene graphs of diverse sizes and resolutions, we apply data augmentation by resizing scene graphs to multiple scales during training.

1) Single-Task Small Models: This class of methods employ conventional DL architectures, such as MLPs, CNNs, and transformers, to address different tasks individually, which often have a relatively small number of parameters.

- **MLP**: MLP is a typical architecture applied to various wireless communication problems [31], [32].
- **CNN**: Convolutional networks (CNNs) are suitable for tasks with high-dimensional wireless data [33], such as the MIMO-OFDM channel matrix.
- **Transformer**: Transformer [22], [34], which is also the basic block for FMs, exhibits excellent performance for wireless tasks, leveraging the attention mechanism. We implement the transformer-based model as [22].

2) Dedicated Methods for Specific Tasks: The aforementioned typical DL architectures may not completely cover SOTA methods for all tasks. Thus, we have supplemented the baselines with additional effective approaches specific to each task, including both model-based methods and DL-based methods.

- *For multi-user precoding*, the eigen-based zero-forcing (ZF) algorithm [35] is a computationally efficient approach; while the WMMSE algorithm [36] is one of the most effective iterative algorithms.
- *For channel decoding*, the combination of the Belief Propagation (BP) model with hyper-graph network, namely hyper BP, is proposed in [37] to achieve effective channel decoding. We employ this method for 50 iterations, corresponding to a 100-layer neural network.
- *For MIMO detection*, ZF and linear minimum mean-squared error (LMMSE) detectors are classical methods with low complexity. In addition, DetNet in [38] and OAMP-Net [39] are popular model-driven deep learning networks for this task.
- *For channel estimation*, least squares (LS) constitutes a basic method.

3) Single-Task FMs: These methods typically finetune FMs for a single task, which leverage the powerful modeling capabilities of FMs to improve the performance.

- **Task-specific FM**: Following the main idea in existing work [8], we train the FM for a single task. The task-specific FM directly employs the FM backbone in our proposed model with elaborately designed data encoders and decoders for different tasks.
- **Environment-aware task-specific FM**: To better evaluate the impact of involving environmental information, we also train an environment-aware task-specific FM that takes the embeddings of scene graphs as parts of inputs.

4) Multi-Task FMs: We compare aforementioned initial attempts of multi-task FMs with the proposed method.

- **Multi-task FM in** [18]: The authors in [18] exploit the multitasking ability of FM and propose a multi-task FM to unify multiple tasks with task-specific encoder-decoder pairs and a shared FM backbone.

Furthermore, we elaborate on the evaluation metrics for the tasks. The sum rate of users, which is the objective of multi-user precoding, is utilized as performance metric to evaluate the multi-user precoding task. NMSE serves as a crucial indicator for assessing channel estimation and MIMO detection precision. For channel decoding, BER is utilized; while for user localization, normalized error distance is employed.

C. Performance Evaluation

1) Environment Awareness of the Proposed Model: First, we evaluate the efficacy of incorporating environmental information in various tasks. To be specific, we compare the proposed environment-aware task-specific FM (namely "Env-aware FM(s)") with task-specific but environment-agnostic FM (namely "FM(s)"). The simulation results are summarized in Table III, where the SNRs for the received signal and the estimated channel are set as 10 dB; the pilot lengths for channel estimation and user localization are set as 4 and 8, respectively. Besides, we employ (64,32) polar codes in the channel decoding task.

Given the strong correlation between the environment and the channel as discussed in Section III-B, the involvement of environmental information can provide prior knowledge of channel distributions. It is quite beneficial for channel estimation and user localization, the objectives of which are directly acquiring the CSI or positioning the users based on CSI. According to the results, the proposed environment-aware task-specific FM achieves better performance than task-specific FM for channel estimation and user localization. Specifically, the incorporation of environmental information reduces the NMSE of channel estimation by 1.50 dB and

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

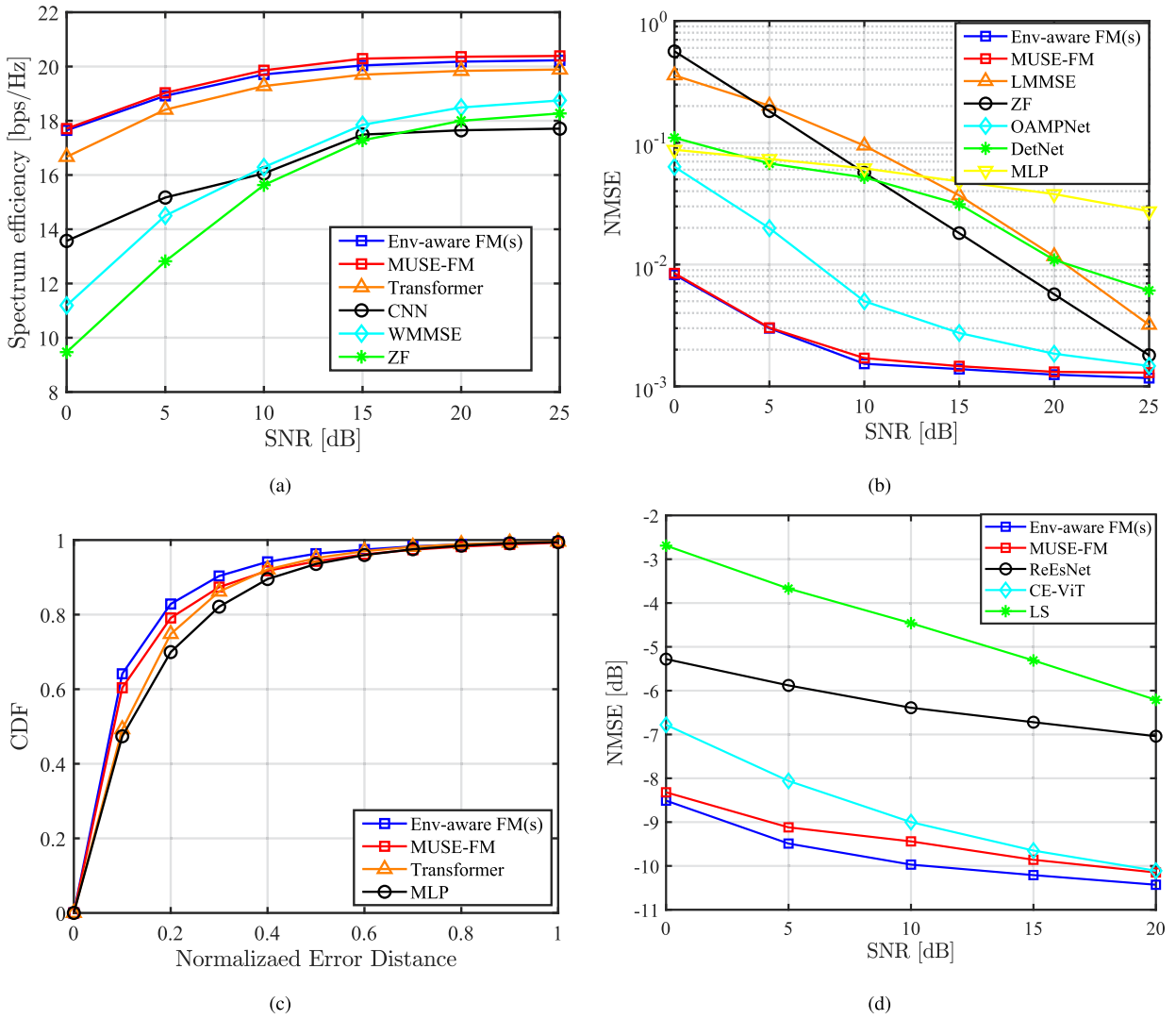


Fig. 6. Task-specific performance for: (a) v; (b) NMSE performance of MIMO detection under different SNRs; (c) CDF of normalized error distance for user localization; (d) NMSE performance of channel estimation under different SNRs.

improves the user localization accuracy by 18.67%. For multi-user precoding and MIMO detection which take CSI as input, the environment-aware FMs also slightly outperform the environment-agnostic ones. It may be attributed to the scene graph to provide side information for noisy CSI, thus obtaining more accurate CSI for feature extraction. Besides, since channel decoding is relatively independent of channel distributions, the two methods achieve comparable performance. As shown in Table III, the natural logarithm of BER achieved by environment-aware FM and environment-agnostic FM are -10.54 and -10.34, respectively.

Based on the analysis above, we summarize the main findings as follows.

- Leveraging the multi-modal processing ability of FMs, the involvement of the scene graph as a part of input is vital to improve the performance and facilitate constructing an environment-aware wireless FM.
- The tasks that are highly related to the channel may benefit more from being informed of the prior knowledge extracted from environmental information.

2) Task-Specific Performance Analysis: For multi-user precoding, we compare the proposed method with both model-driven methods (ZF [35] and WMMSE [36]) and data-driven methods (CNN [33] and transformer-based method). The sum rate performance is illustrated in Fig. 6 (a), where the multi-user channel is corrupted by noise with SNR values increasing from 0 dB to 25 dB. It is observed that the ZF, WMMSE, and CNN-based methods fail to achieve satisfactory performance. The proposed MUSE-FM significantly improves the sum rate, and outperforms shallow transformer and environment-aware task-specific FM-based methods with its strong feature extraction and joint learning capability. The advantage is more evident in the low SNR region, indicating the robustness of FM-based methods.

In Fig. 6 (b), the NMSE performance of the proposed MUSE-FM on MIMO detection under different SNRs is presented, where the SNR increases from 0 dB to 25 dB. We also compare the performance with several baselines. As illustrated in Fig. 6 (b), despite the low computational complexity, the ZF and LMMSE-based methods show relatively high NMSE

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
COMPARISON OF THE NEGATIVE NATURAL LOGARITHM OF BER FOR DIFFERENT E_b/N_0 IN CHANNEL DECODING TASK

$\mathbf{E_b/N_0}$	**4**	**5**	**6**
Hyper BP [37]	4.59	6.10	7.69
ECCT [22]	6.99	9.44	12.32
Env-aware FM(s)	**7.68**	**10.54**	**13.82**
MUSE-FM	7.56	10.38	13.72

performance in low SNR regime. To solve the problem, DL-driven methods, including OAMPNet [39], DetNet [38] and MLP-based [31] method, are proposed to efficiently improve the recover accuracy. In addition, the proposed FMs with environmental awareness further improve NMSE performance and outperform baselines, especially in the low SNR regime. Moreover, we observe that the exploitation of statistical information of noise (i.e. SNR) as prior knowledge helps improve the performance. It is validated by the superior performance of OAMPNet and proposed MUSE-FM over others, which utilize the noise variance in the model.

Next, we discuss the performance of user localization. The cumulative distribution function (CDF) curve of the normalized positioning error distance is presented in Fig. 6 (c). The SNR of the received data is set as 10 dB. The proposed environment-aware task-specific FM achieves the most accurate user localization, while the proposed MUSE-FM obtains comparable performance with marginal gap. They outperform the traditional transformer-based method and the simple low-cost MLP-based method [32]. Specifically, the ratio of samples with normalized error distance less than 0.1 is 64.13% for the proposed environment-aware task-specific FM and 60.38% for the proposed MUSE-FM. The ratio drops to 49.27% for the traditional transformer-based method and 45.37% for the MLP-based method, validating the advantage of the proposed method.

Fig. 6 (d) shows the NMSE performance versus SNR for channel estimation, where the SNR increases from 0 dB to 20 dB. Besides the traditional LS method, we also compare our method with the CNN-based method, ReEsNet [40], and transformer-based method, CE-ViT [41]. LS performs the worst as it ignores the effect of the noise. The performance of ReEsNet and CE-ViT exceeds that of LS, but still requires further improvements. The proposed FMs, both task-specific and multi-task, outperform other methods and maintain accurate estimation in low SNR regime.

Finally, we focus on the performance of channel decoding. The results are reported in Tab. IV where we present the negative natural logarithm of the BER under different normalized SNR (i.e. E_b/N_0). For transformer-based channel decoding, namely ECCT, we select the largest version presented in [22]. We observe that our approach significantly outperforms the current methods for all different SNR values. Despite the similar structure to ECCT, equipped with larger model size, the proposed model further improves the performance.

From the simulation of all selected tasks, we conclude that:

- The proposed environment-aware FMs, trained either on a single task or multiple tasks, outperform existing baselines for all tasks, indicating superior feature extraction capabilities of FMs and their potential in wireless domain.
- The advantages of FM-based models are more evident in challenging scenarios, such as the low SNR regime, which may be attributed to the robustness and generalization ability of FMs.
- With the multi-task learning ability of FMs, the proposed MUSE-FM achieves comparable performance with environment-aware task-specific FM.

Moreover, MUSE-FM demonstrates advantages in training and inference overhead compared to task-specific FMs. Regarding memory overhead, the proposed MUSE-FM contains a total of **134.44M parameters**. The detailed parameter distribution is summarized in Table VI, including FM backbone: 85.84M; scene encoder: 7.16M; Hypernetwork (for generating task-specific weights): 2.84M; text embeddings: 38.60M. Importantly, the prompt-guided unified data encoder and decoder introduce no additional parameters. Their weights and biases are dynamically generated by the hypernetwork according to the task instruction. This design enables a unified encoder-decoder pair to handle heterogeneous tasks without task-specific parameter overhead. For comparison, each environment-aware task-specific FM requires approximately 93.20M parameters. Deploying five separate task-specific FMs would thus incur around 466M parameters in total. By sharing the large backbone and scene encoder while using a lightweight hypernetwork for dynamic adaptation, MUSE-FM achieves approximately 71% fewer parameters than five task-specific FMs, significantly reducing storage and deployment overhead. Regarding memory overhead, with 134.44M parameters, MUSE-FM incurs significantly lower storage costs compared to deploying five task-specific FMs (each 93.20M parameters, totaling 466M parameters). Compared to the sequential training and inference of multiple task-specific FMs, MUSE-FM reduces training latency by 44.3% and inference latency by 48.9%.

3) Scaling Behavior: We next investigate how the model performs with different sizes of dataset. It is crucial for efficient deployment of DL-based models, which determines the trade-off between performance and the cost of data collection and network training. In this part, we evaluate the scaling behavior of the proposed MUSE-FM, compared with environment-aware task-specific FMs, as shown in Table V. Note that both task-specific FMs and multi-task FMs incorporate environmental information, while the notations "FM(s)" and "FM(m)" are shorthand for "Env-aware FM(s)" and "MUSE-FM", respectively, used for table conciseness. Besides training the model on the full dataset, we train the model with 2%, 10%, 20%, 50% and 80% of the dataset, respectively (i.e., we utilize 1, 5, 10, 25, and 40 samples for each training scenario). It is observed that for each task, as the data volume increases, the performance of both task-specific FM and multi-task FM gradually improves and asymptotically approaches their performance ceiling.

When trained on the full dataset, the environment-aware task-specific FM slightly outperforms MUSE-FM, benefiting from avoiding inter-task compromises and the higher

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

TABLE V
EVALUATION OF PERFORMANCE OF DIFFERENT TASKS TRAINED WITH DIFFERENT SIZES OF DATASET

Data size	**CE** (NMSE ↓)		**Precoding** (Rate ↑)		**Det** (NMSE ↓)		**Decoding** (ln(BER) ↓)		**loc** (Error Dis ↓)	
Ratio	FM(s)	FM(m)	FM(s)	FM(m)	FM(s)	FM(m)	FM(s)	FM(m)	FM(s)	FM(m)
2% dataset	0.2353	**0.2056**	13.73	**14.24**	0.029454	**0.023825**	-7.69	**-7.91**	0.3150	**0.2817**
10% dataset	0.1526	**0.1448**	17.28	**17.80**	**0.007424**	0.007726	-8.08	**-8.18**	0.2052	**0.1938**
20% dataset	0.1299	**0.1272**	18.36	**18.68**	**0.003105**	0.003218	-8.23	**-8.63**	0.1740	**0.1723**
50% dataset	**0.1126**	0.1180	19.20	**19.49**	**0.001888**	0.002097	-9.37	**-9.56**	**0.1422**	0.1505
80% dataset	**0.1044**	0.1164	19.59	**19.75**	**0.001611**	0.001702	-9.92	**-10.09**	**0.1347**	0.1481
full dataset	**0.0982**	0.1125	19.71	**19.76**	**0.001632**	0.001736	**-10.54**	-10.38	**0.1322**	0.1415

TABLE VI
PARAMETER COUNT COMPARISON BETWEEN MUSE-FM AND TASK-SPECIFIC FMS

Parameter(M)	MUSE-FM	Task-specific FM
FM Backbone	85.84	85.84
Scene Encoder	7.16	7.16
Hypernetwork	2.84	-
Text Embeddings	38.60	-
Data Encoder	-	0.99
Data Decoder	-	0.98
Total	**134.44**	**93.20**

complexity of multi-task training. As the dataset size decreases, the MUSE-FMs trained on 20% and 50% data samples achieve comparable performance with corresponding task-specific FMs; while MUSE-FMs even outperform task-specific ones in some tasks (multi-user precoding for instance). Furthermore, as shown in Table V, when data samples are extremely limited (only 2% and 10% of the dataset is accessible), the proposed MUSE-FM outperforms the task-specific FM on most tasks. For example, MUSE-FM achieves 17.80 bps/Hz for multi-user precoding, versus 17.24 bps/Hz for the task-specific FM, when trained on 10% of the dataset. When trained with only 2% dataset, MUSE-FM improves the performance for all selected tasks, compared with task-specific FMs. The main reason may be analyzed as follows: with limited data samples, the task-specific FM, accessing only one task's dataset, is more prone to overfitting, leading to performance degradation. In contrast, the proposed MUSE-FM leverages its multi-task learning capability to jointly learn from multiple datasets, enhancing joint feature representation.

It should be noted that, due to the large size of the constructed dataset, our proposed method trains all FM parameters to build a wireless communication-specific FM based on pre-trained parameters. For fair and direct comparison, we also train all FM parameters when training with only part of the dataset. Consequently, it results in overfitting and performance degradation even for multi-task FM. In practice, when data samples are scarce, we can train only a small portion of the parameters in the FMs to avoid overfitting and improve the performance.

We summarize the main insights for the experiments in this subsection as follows.

- Since FMs possess the capability to learn complex patterns from vast data, gradually increasing data volume enhances the performance, both for multi-task FM and task-specific FM.
- When data volume is sufficient, task-specific FMs achieve superior performance; however, with limited data, multi-task models benefit from joint learning across multiple tasks and datasets. This illustrates the capability of multi-task FMs to exploit different types of datasets of wireless communications.

4) Impact of Prompt-Guided Unified Data Encoder-Decoder Pair: In this subsection, the effectiveness of the proposed prompt-guided unified data encoder-decoder pair is evaluated. As discussed in Section III-C, designing separate data encoders and decoders as [18] for each task and parameter helps task alignment and facilitates multi-task learning, but it incurs high computational and deployment overhead, given the diverse tasks and parameter variations. Besides, when encountering a new task or parameter configuration with different input/output formats, FMs with task-specific data encoder-decoder pairs may not work. Instead, a new task-specific encoder-decoder pair is required to be designed and trained, which limits the scalability of the models. Therefore, we propose a unified prompt-guided data encoder-decoder pair to improve the scalability of multi-task FM, while retaining superior performance for all tasks. To validate the effectiveness, we compare the proposed MUSE-FM with multi-task FM with task-specific encoder-decoder pairs in [18].

The simulation results for the proposed MUSE-FM and that with task-specific encoder-decoder pairs are provided in Table VII. The SNRs for the received signal and estimated channel are set as 10 dB; the pilot lengths for channel estimation and user localization are set as 4 and 8, respectively. Although only a unified encoder-decoder pair is employed, the proposed method achieves comparable performance with multi-task FM with task-specific encoder-decoder pairs in [18]. It even slightly outperforms multi-task FM in [18] for channel estimation, multi-user precoding and MIMO detection. Its comparable performance mainly stems from two reasons, involving domain knowledge and learning joint feature

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
PERFORMANCE OF FM WITH PROPOSED FM WITH UNIFIED ENCODER-DECODER PAIR AND FM WITH TASK-SPECIFIC ENCODER-DECODER PAIRS

Method	CE (NMSE ↓)	Precoding (Rate ↑)	Det (NMSE ↓)	Decoding (ln(BER) ↓)	Loc (Error Dis ↓)
FM with task-specific encoder-decoder pairs [18]	0.1183	19.50	0.001778	-10.30	**0.1389**
MUSE-FM	**0.1125**	**19.76**	**0.001736**	**-10.38**	0.1415

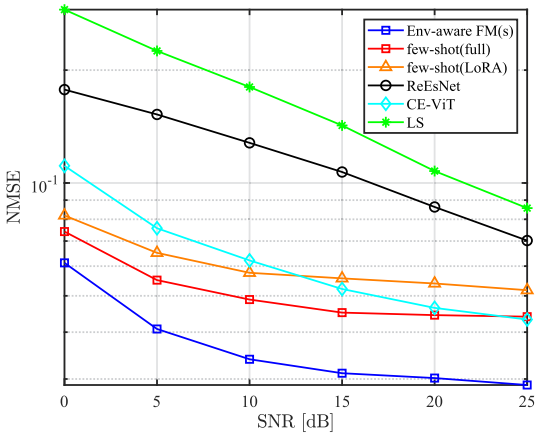


Fig. 7. NMSE performance for channel estimation with pilot length = 6.

representation. Specifically, the parameters of the prompt-guided unified data encoder-decoder pair are dynamically generated by task-specific instructions, which contain critical domain knowledge and key parameters. With extra information included, performance is improved. Besides, the proposed unified encoder-decoder pair extracts joint representation from various tasks, leveraging the relations among the tasks to achieve better performance.

In conclusion, we observe that the proposed prompt-guided unified data encoder-decoder pair achieves comparable performance with task-specific encoder-decoder pairs, while improves the adaptability of the multi-task FM.

5) Generalization Ability: Then we evaluate the generalization ability of MUSE-FM for a new parameter configuration (e.g. antenna number, pilot length, user number). Fig. 7 utilizes channel estimation for instance to illustrate the ability of the proposed MUSE-FM in handling new task configurations. During the training process, the pilot length is set as 8, while the pilot length is changed to 6 during evaluation. Varying pilot quantities alter input dimensions, rendering FM with task-specific encoder-decoder pairs inapplicable. In contrast, our proposed method remains directly applicable to new parameter configurations. To comprehensively demonstrate its capability, except the existing baselines, we compare: (1) FM trained specifically for pilot length = 6, (2) Few-shot performance where FM trained with pilot length = 8 is fully fine-tuned using 10% of the data for 5 epochs, (3) Few-shot performance where FM trained with pilot length = 8 is fine-tuned using 10% of the data for 5 epochs with low-rank adaptation (LoRA) [42]. The rank of LoRA is set as 8, which significant reduces the finetuning parameters of FM backbone from 85.84M to 0.44M. As shown in Fig. 7, with only minimal full-parameter fine-tuning, our method outperforms all baselines, indicating the generalization ability of the proposed MUSE-FM. Furthermore, the performance of LoRA finetuning also surpasses ReEsNet and LS-based methods and achieves comparable performance with CE-ViT, while significantly reducing the computational complexity compared to full finetuning. It validated that the fine-tuning approach can be readily extended to other strategies such as LoRA and partial parameter tuning.

V. CONCLUSION

In this work, we propose a multi-task environment-aware FM, namely MUSE-FM for wireless communications with a unified architecture to handle multiple tasks and incorporate scenario information. Through multi-modal input integration of environmental context and wireless data, the proposed model manages to facilitate cross-scenario feature extraction and enhances the environmental adaptability. The proposed prompt-guided data encoder-decoder pair employs instructions to dynamically generate task-specific parameters, and thus handles data with heterogeneous formats and distributions across multiple tasks with a unified encoder-decoder pair. Simulation results have demonstrated the multi-task learning and generalization capabilities of the proposed method. Besides, incorporating environmental information facilitates cross-scenario learning ability, while the proposed prompt-guided unified encoder-decoder pair improves the scalability of the model. Furthermore, with limited data samples, MUSE-FM harnesses cross-task dependencies for knowledge transfer and enhances joint optimization compared with task-specific FMs. Future works can focus on more effective use of environmental information. In addition, model quantization and pruning can be investigated to further improve inference efficiency, enabling deployment in real-world applications.

REFERENCES

[1] P. P. Ray, N. Kumar, and M. Guizani, "A vision on 6G-enabled NIB: Requirements, technologies, deployments, and prospects," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 120–127, Aug. 2021.
[2] J. Hoydis, F. A. Aoudia, A. Valcarce, and H. Viswanathan, "Toward a 6G AI-native air interface," *IEEE Commun. Mag.*, vol. 59, no. 5, pp. 76–81, May 2021.
[3] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.
[4] R. K. Alec, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[5] J. Guo, Y. Cui, S. Jin, and J. Zhang, "Large AI models for wireless physical layer," 2025, *arXiv:2508.02314*.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

[6] F. Jiang et al., "A comprehensive survey of large AI models for future communications: Foundations, applications and challenges," 2025, *arXiv:2505.03556*.
[7] F. Zhu et al., "Wireless large AI model: Shaping the AI-empowered future of 6G and beyond," 2025, *arXiv:2504.14653*.
[8] B. Liu, X. Liu, S. Gao, X. Cheng, and L. Yang, "LLM4CP: Adapting large language models for channel prediction," *J. Commun. Inf. Netw.*, vol. 9, no. 2, pp. 113–125, Jun. 2024.
[9] H. Li, M. Xiao, K. Wang, D. I. Kim, and M. Debbah, "Large language model based multi-objective optimization for integrated sensing and communications in UAV networks," *IEEE Wireless Commun. Lett.*, vol. 14, no. 4, pp. 979–983, Apr. 2025.
[10] W. Liu, C. Pan, H. Ren, W. Zhang, C.-X. Wang, and J. Wang, "Large-model AI for near field beam prediction: A CNN-GPT2 framework for 6G XL-MIMO," 2025, *arXiv:2510.22557*.
[11] Y. Zhang, H. Yin, W. Li, E. Björnson, and M. Debbah, "Port-LLM: A port prediction method for fluid antenna based on large language models," *IEEE Trans. Commun.*, vol. 73, no. 12, pp. 14534–14547, Dec. 2025.
[12] W. Lee and J. Park, "LLM-empowered resource allocation in wireless communications systems," 2024, *arXiv:2408.02944*.
[13] J. Guo, Y. Cui, C.-K. Wen, and S. Jin, "Prompt-enabled large AI models for CSI feedback," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 2654–2668, Dec. 2025.
[14] X. Zhou, L. Liang, H. Ye, J. Zhang, C.-K. Wen, and S. Jin, "Reducing pilots in channel estimation with predictive foundation models," *IEEE Trans. Commun.*, vol. 74, pp. 9360–9375, 2026.
[15] S. Alikhani, G. Charan, and A. Alkhateeb, "LWM: A pre-trained wireless foundation model for universal feature extraction," in *Proc. IEEE Int. Conf. Mach. Learn. Commun. Netw. (ICMLCN)*, Barcelona, Spain, May 2025, pp. 1–6.
[16] T. Yang et al., "WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication," *IEEE Netw.*, vol. 39, no. 5, pp. 58–65, Sep. 2025.
[17] X. Liu, S. Gao, B. Liu, X. Cheng, and L. Yang, "LLM4WM: Adapting LLM for wireless multi-tasking," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 835–847, Jul. 2025.
[18] T. Zheng and L. Dai, "Large language model enabled multi-task physical layer network," *IEEE Trans. Commun.*, vol. 74, pp. 307–321, 2026.
[19] J. Du, T. Lin, C. Jiang, Q. Yang, C. F. Bader, and Z. Han, "Distributed foundation models for multi-modal learning in 6G wireless networks," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 20–30, Jun. 2024.
[20] X. Ma, Z. Gao, F. Gao, and M. Di Renzo, "Model-driven deep learning based channel estimation and feedback for millimeter-wave massive hybrid MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2388–2406, Aug. 2021.
[21] X. Zhou, L. Liang, J. Zhang, P. Jiang, Y. Li, and S. Jin, "Generative diffusion models for high dimensional channel estimation," *IEEE Trans. Wireless Commun.*, vol. 24, no. 7, pp. 5840–5854, Jul. 2025.
[22] Y. Choukroun and L. Wolf, "Error correction code transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 38695–38705.
[23] L. Yu, L. Shi, J. Zhang, Z. Zhang, Y. Zhang, and G. Liu, "ChannelGPT: A large model toward real-world channel foundation model for 6G environment intelligence communication," *IEEE Commun. Mag.*, vol. 63, no. 10, pp. 68–74, Sep. 2025.
[24] J. Liu, J. Guo, Y. Cui, C.-K. Wen, and S. Jin, "AdapCsiNet: Environment-adaptive CSI feedback via scene graph-aided deep learning," 2025, *arXiv:2504.10798*.
[25] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6000–6010.
[26] D. Ha, A. M. Dai, and Q. V. Le, "Hypernetworks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.
[27] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1871–1880.
[28] T. Jiao et al., "Addressing the curse of scenario and task generalization in AI-6G: A multi-modal paradigm," *IEEE Trans. Wireless Commun.*, vol. 24, no. 9, pp. 7377–7391, Sep. 2025.
[29] J. Hoydis et al., "Sionna: An open-source library for next-generation physical layer research," 2022, *arXiv:2203.11854*.
[30] J. Hoydis et al., "Sionna RT: Differentiable ray tracing for radio propagation modeling," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2023, pp. 317–321.
[31] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.
[32] P. Ferrand, A. Decurninge, and M. Guillaud, "DNN-based localization from channel estimates: Feature design and experimental results," in *Proc. IEEE Global Commun. Conf.*, Taipei, Taiwan, Dec. 2020, pp. 1–6.
[33] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1866–1880, Mar. 2020.
[34] H. Jiang, M. Cui, D. W. K. Ng, and L. Dai, "Accurate channel prediction based on transformer: Making mobility negligible," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2717–2732, Sep. 2022.
[35] C. Zhang, Y. Jing, Y. Huang, and L. Yang, "Performance analysis for massive MIMO downlink with low complexity approximate zero-forcing precoding," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3848–3864, Sep. 2018.
[36] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
[37] E. Nachmani and L. Wolf, "Hyper-graph-network decoders for block codes," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 2329–2339.
[38] N. Samuel, T. Diskin, and A. Wiesel, "Deep MIMO detection," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Hokkaido, Japan, Jul. 2017, pp. 1–5.
[39] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "A model-driven deep learning network for MIMO detection," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 584–588.
[40] L. Li, H. Chen, H.-H. Chang, and L. Liu, "Deep residual learning meets OFDM channel estimation," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 615–618, May 2020.
[41] F. Liu, J. Zhang, P. Jiang, C.-K. Wen, and S. Jin, "CE-ViT: A robust channel estimator based on vision transformer for OFDM systems," in *Proc. IEEE Global Commun. Conf. (IEEE GLOBECOM)*, Dec. 2023, pp. 4798–4803.
[42] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

Tianyue Zheng (Graduate Student Member, IEEE) received the B.E. degree in information engineering from Southeast University, Nanjing, China, in 2022. She is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her research interests include extremely large scale MIMO (XL-MIMO), CSI acquisition, and AI for communications. She received the National Scholarship in 2019 and the Excellent Student of Jiangsu Province in 2021.

Jiajia Guo (Member, IEEE) received the B.S. degree from Nanjing University of Science and Technology, Nanjing, China, in 2016, the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2019, and the Ph.D. degree in information and communications engineering from Southeast University, Nanjing, in 2023. He has been a Research Assistant Professor with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology. His research interests focus on AI-native air interfaces, massive MIMO, ISAC, and large AI models. His achievements were selected as one of the Top ten science and technology progress in the information and communication field for 2022 in China. He is an Associate/Guest Editor of IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, *IET Signal Processing*, and *Digital Communications and Networks*.

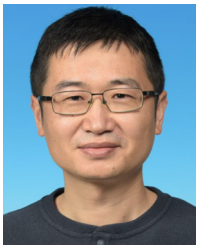
Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.

Linglong Dai (Fellow, IEEE) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2003, the M.S. degree (Hons.) from China Academy of Telecommunications Technology, Beijing, China, in 2006, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, in 2011.

From 2011 to 2013, he was a Post-Doctoral Research Fellow with the Department of Electronic Engineering, Tsinghua University, where he was an Assistant Professor from 2013 to 2016 and an Associate Professor from 2016 to 2022. Since 2022, he has been a Professor with Tsinghua University. He has co-authored the book *mmWave Massive MIMO: A Paradigm for 5G* (Academic Press, 2016). He has authored or co-authored more than 90 IEEE journal articles and more than 50 IEEE conference papers. He also holds more than 20 granted patents. His current research interests include massive MIMO, reconfigurable intelligent surface (RIS), millimeter-wave and terahertz communications, machine learning for wireless communications, near-field communications, and quantum information technologies.

Dr. Dai has received five IEEE Best Paper Awards from IEEE ICC 2013, IEEE ICC 2014, IEEE ICC 2017, IEEE VTC 2017-Fall, and IEEE ICC 2018. He has also received Tsinghua University Outstanding Ph.D. Graduate Award in 2011, Beijing Excellent Doctoral Dissertation Award in 2012, China National Excellent Doctoral Dissertation Nomination Award in 2013, the URSI Young Scientist Award in 2014, the IEEE Transactions on Broadcasting Best Paper Award in 2015, the Electronics Letters Best Paper Award in 2016, the National Natural Science Foundation of China for Outstanding Young Scholars in 2017, the IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award in 2017, the IEEE ComSoc Asia–Pacific Outstanding Paper Award in 2018, China Communications Best Paper Award in 2019, the IEEE Access Best Multimedia Award in 2020, the IEEE Communications Society Leonard G. Abraham Prize in 2020, the IEEE ICC Outstanding Demo Award in 2022, the National Science Foundation for Distinguished Young Scholars in 2023, and the IEEE ComSoc Stephen O. Rice Prize in 2025. He was listed as a Highly Cited Researcher by Clarivate Analytics from 2020 to 2025.

Shi Jin (Fellow, IEEE) received the B.S. degree in communications engineering from Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from Southeast University, Nanjing, in 2007. From June 2007 to October 2009, he was a Research Fellow with University College London, Adastral Park Research Campus, London, U.K. He is currently with the National Mobile Communications Research Laboratory, Southeast University. His research interests include wireless communications, random matrix theory, and information theory. He and his co-authors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory, the 2022 Best Paper Award and the 2010Young Author Best Paper Award by the IEEE Signal Processing Society, the IEEE Vehicular Technology Society 2023 Jack Neubauer Memorial Award, and the 2024 Marconi Prize Paper Award from the IEEE Communications Society and the IEEE Signal Processing Society. He was an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and *IET Communications*. He is serving as an Area Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and *Electronics Letters* (IET).



Jun Zhang (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin in 2009.

He is a Professor with the Department of Electronic and Computer Engineering (ECE) and the Associate Director of the Computer Engineering (CPEG) Program, The Hong Kong University of Science and Technology. He has co-authored/co-edited five books including *Fundamentals of LTE* (Prentice-Hall, 2010). His research interests include integrated communications and AI, generative AI, and edge AI systems. He is a Clarivate Highly Cited Researcher. He was an IEEE ComSoc Distinguished Lecturer from 2023 to 2024. He was a co-recipient of several best paper awards, including the 2025 IEEE Communications Society Katherine Johnson Young Author Best Paper Award, the 2021 Best Survey Paper Award of the IEEE Communications Society, the 2019 IEEE Communications Society and Information Theory Society Joint Paper Award, and the 2016 Marconi Prize Paper Award in Wireless Communications. He also received the 2016 IEEE ComSoc Asia–Pacific Best Young Researcher Award. He served as the Symposium Co-Chair for the IEEE Wireless Communications and Networking Conference (WCNC) in 2011 and 2026, respectively, and the IEEE International Conference on Communications (ICC) in 2021; and the TPC Co-Chair for The IEEE Hong Kong 6G Wireless Summit in 2023 and 2024, respectively. He is currently an Area Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (leading the area of machine learning and artificial intelligence) and IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING (leading the area of distributed learning and AI at the network edge).

Authorized licensed use limited to: Tsinghua University. Downloaded on July 17,2026 at 05:40:10 UTC from IEEE Xplore. Restrictions apply.