Large Multimodal Model-Based Environment-Aware Channel Estimation

Seungnyun Kim, *Member, IEEE*, Seokhyun Jeong, *Student Member, IEEE*, Jiao Wu, *Member, IEEE*, Byonghyo Shim, *Fellow, IEEE*, and Moe Z. Win, *Fellow, IEEE*

***Abstract*—Recently, large multimodal models (LMMs) have been successfully adopted in various fields due to their outstanding adaptability and reasoning abilities. Despite their potential to automate diverse tasks in communications systems, application to the physical layer remains underexplored. The primary reason is that the traditional physical layer relies on analytic channel measurements (e.g., pilot measurements), which capture only quantitative changes in transmitted signals and fail to characterize qualitative physical interactions (e.g., reflections and blockages) with the environment. In this paper, we propose an LMM-based environment-aware channel estimation framework that captures the contextual channel information by leveraging both perceptual sensor data and numerical pilot measurements. The main idea of the proposed scheme is to utilize visual channel parameters (VCPs), i.e., positions of user equipment (UE), reflection points, and scatterers. Since VCPs provide a direct visualization of the propagation environment, we can identify how signals propagate and physically interacts with the surrounding objects. To extract the VCPs and learn their probability distributions, we develop a reflection learning technique based on LMM. By incorporating these parameters, we establish a fundamental channel knowledge map (CKM) between the UE position and the channel. Simulation results demonstrate that the proposed scheme can effectively predict the channel throughout the wireless environments.**

***Index Terms*—Large multimodal model, channel knowledge map, environment awareness, integrated sensing and communications.**

I. Introduction

GENERATIVE artificial intelligence (AI) is catalyzing disruptive innovation across a wide range of industries. In particular, large language models (LLMs), fueled by the remarkable success of ChatGPT, are increasingly permeating our daily lives [1], [2]. Owing to their vast network size and extensive pre-training, LLMs can effectively perform question and answering (Q&A) tasks either without examples (i.e., zero-shot prompting) or with a smaller number of examples (i.e., few-shot prompting) [3]. Recently, the capabilities of LLMs have been further extended to process not only language but also multimodal data including images, videos, and audio. This multimodal extension of LLMs, often called LMMs, can handle more complicated and diverse tasks than traditional LLMs, broadening their potential applications across various fields such as humanoids, autonomous driving, factory machinery, and healthcare [4].

Despite the vast amount of research integrating LMMs into different industries, research on combining LMMs with wireless communications is relatively scarce [5], [6], [7]. Some efforts have been made to utilize LMMs for medium access control (MAC) layer tasks (e.g., user scheduling and resource allocation), but the exploration of LMMs in physical layer tasks (e.g., channel estimation and beamforming) is still in its infancy [8], [9], [10]. The primary reason is that the traditional physical layer relies on analytic channel measurements (e.g., pilot measurements), which are fundamentally different from the input modalities (e.g., texts and images) typically processed by LMMs. For instance, in 5G New Radio (NR), channel state information (CSI) is obtained by applying linear estimation techniques, such as least squares (LS) estimation, to pilot measurements (e.g., channel state information reference signal (CSI-RS) and sounding reference signal (SRS)). Also, semi-blind channel estimation approaches that leverage both pilot and data symbols have been proposed in [11] and [12]. In millimeter wave (mmWave) and terahertz (THz) systems, based on the sparse scattering property that channels are characterized by a small number of geometric channel parameters (GCPs) (e.g., angles, delays, and path gains), approaches that extract these parameters from the pilot measurements have been proposed [13], [14], [15], [16], [17], [18]. In addition, deep learning (DL)-based methods that use deep neural networks (DNNs) to learn the relationship between pilot measurements and channel information have been developed in [19], [20], [21], and [22].

The analytic channel measurements can capture quantitative changes in the transmitted signal, however, they fail to capture contextual channel information that elucidates the qualitative physical interactions (e.g., reflections and blockage) between the transmitted signal and the surrounding physical environment (e.g., scatterers and obstacles). Since the fundamental relationship between the transmitted and received signals cannot be identified from the analytic channel measurements, one has to restart the entire channel estimation

Received 23 November 2024; revised 1 May 2025; accepted 9 June 2025. Date of publication 12 January 2026; date of current version 23 January 2026. The fundamental research described in this paper was supported, in part, by the National Research Foundation of Korea under Grant RS-2023-00252789 and Grant RS-2024-00409492, by the Institute of Information & Communications Technology Planning & Evaluation of Korea under Grant RS-2024-00418784, and by the Robert R. Taylor Professorship. *(Corresponding author: Moe Z. Win.)*

Seungnyun Kim is with the Wireless Information and Network Sciences Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: snkim94@mit.edu).

Seokhyun Jeong and Byonghyo Shim are with the Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, Seoul 08826, Republic of Korea (e-mail: shjeong@islab.snu.ac.kr; bshim@snu.ac.kr).

Jiao Wu is with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia (e-mail: jiao.wu@kaust.edu.sa).

Moe Z. Win is with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: moewin@mit.edu).

Digital Object Identifier 10.1109/JSAC.2025.3623163

process even with a small movement of the mobile. Due to this reason, traditional communications systems relying on analytic channel measurements fall short in promptly responding to changes in wireless environments. For example, in 5G NR, user equipment (UE) mobility is supported by the handover process [23]. However, since the handover process is initiated only after detecting gradual attenuation of the received signal power, it cannot predict sudden link deterioration, resulting in performance degradation in high-mobility environments such as non-terrestrial networks (NTN) [24], [25].

Recently, channel mapping approaches that eliminate the need for pilot measurements by learning the spatial distributions of the channel (i.e., channel knowledge map (CKM)) have gained much attention [26]. In [27] and [28], Kriging interpolation-based channel mapping techniques have been proposed. In [29], a channel mapping technique that uses autoencoders to generate CKM from partially observed channel measurements has been developed. Furthermore, recent studies have explored the use of digital twin (DT) of wireless systems for constructing CKM [30], [31]. In the DT-based channel mapping techniques, interactions between the transmitted signal and the surrounding environment can be directly identified as well as CKM through system emulators and ray tracing simulators. For example, in [32] and [33], DT-based channel prediction techniques using generative adversarial network (GAN) and diffusion models have been proposed. One major issue of conventional channel mapping approaches is that, due to the large output dimension which scales with the numbers of antennas and time-frequency resources, they require a substantial amount of training data to directly learn the end-to-end mapping between the UE position and the channel, as well as for continuous adaptation to dynamic changes. As a result, they might not work well in rapidly changing wireless environments such as vehicle-to-everything (V2X) networks where collecting sufficiently large datasets is challenging.

The fundamental questions related to the environment-aware channel estimation are: 1) how to acquire the contextual channel information; and 2) how to establish a fundamental relationship between the UE position and the channel information using the acquired contextual channel information. The answers to these questions will facilitate *environment-awareness*, the ability to recognize the surrounding environments and promptly adapt to changes in signal propagation caused by environmental factors such as mobilities and obstructions.

The aim of this paper is to propose an environment-aware multimodal channel estimation framework based on sensing and LMM. Recently, sensing technologies that observe the surrounding environments using various sensing modalities (e.g., camera and LiDAR) have been widely adopted in autonomous driving, robotics, and urban air mobility (UAM) [34], [35], [36], [37]. In accordance with this trend, computer vision (CV) technique for interpreting the sensor measurements has made a significant advancement in performing object detection and semantic segmentation. Motivated by these developments, the proposed scheme, henceforth referred to as LMM-based channel estimation (LMM-CE), leverages both perceptual sensor data and numerical pilot measurements to capture the contextual channel information. The key idea of LMM-CE is to utilize *visual channel parameters (VCPs)*, i.e., positions of UE, reflection points, and scatterers, that visually characterize the scattering geometry of the surrounding environment. In contrast to the conventional GCPs (i.e., angles, delays, and path gains), the VCPs provide a direct visualization of the propagation environment. Using the VCPs, we can identify how transmitted signals propagate and physically interact with the objects. Moreover, by leveraging VCPs as a bridge between the UE position and the channel, we can effectively construct the CKM with minimal training overhead.

The intriguing feature of LMM-CE is that we exploit LMMs to capture VCPs and learn their probability distributions from sensing and pilot measurements. As a generative AI model, LMMs excel at capturing the underlying probability distribution of sequential data and generating subsequent elements. In our context, this means that the LMM can effectively extract the VCPs from sensing and pilot measurements and learn their complex relationships with the UE position. The core components of LMMs for integrating and analyzing multimodal data are the Transformer architecture and its attention mechanism. The attention mechanism quantifies the correlations across the entire input data sequence, allowing LMMs to map relevant information from one modality to another. By measuring the correlations between sensing and pilot measurements, the LMM can learn how specific visual features (e.g., UE, scatterers, and obstacles) in the sensing measurements correspond to numerical features (e.g., GCPs) in the pilot measurements and then effectively integrate them as environmental information. Also, due to its significantly large network size and extensive pretraining on diverse datasets, LMMs can effectively handle a broad range of tasks, encompassing CV tasks (e.g., object detection and semantic segmentation) as well as analytic tasks (e.g., estimation, classification, and regression), even without updating the network parameter (i.e., in-context learning). This versatility is pivotal for the proposed environment-aware channel estimation framework in analyzing multiple modalities. One notable benefit of LMMs is that we can provide wireless system information and detailed task instructions in natural language (i.e., prompting) so that LMMs can solve complex wireless tasks. For example, when a prompt "*Perform a regression task to determine the reflection point for a given UE position. The reflection point must satisfy the law of reflection with respect to the base station (BS) and UE positions*" is given, the LMM can enhance regression performance by leveraging the additional task-specific information.

The main contributions of this paper are described in the follows. Specifically, we

- propose an environment-aware multimodal channel estimation framework that utilizes sensing and pilot measurements to capture contextual channel information;
- establish a fundamental mapping between UE position and channel (i.e., CKM) using VCPs (i.e., position of UE, reflection points, and scatterers); and

- develop a reflection learning technique that extracts the VCPs from multimodal data and learns their probability distributions using LMM.

The rest of this article is organized as follows: Section II presents sensing-aided mmWave/THz systems. Section III explains the CKM. Section IV proposes LMM-CE. Section V presents the numerical results. Section VI concludes the paper.

Notations: Random variables are displayed in sans serif, upright fonts; their realizations in serif, italic fonts. Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. For example, a random variable and its realization are denoted by x and x for scalars, $\mathbf{x}$ and $\boldsymbol{x}$ for vectors, and $\mathbf{X}$ and $\boldsymbol{X}$ for matrices. Sets and random sets are denoted by upright sans serif and calligraphic font, respectively. For example, a random set and its realization are denoted by X and $\mathcal{X}$, respectively. The m-by-n matrix of zeros is denoted by $\mathbf{0}_{m\times n}$; when $n=1$, the m-dimensional vector of zeros is simply denoted by $\mathbf{0}_m$. The m-by-m identity matrix is denoted by $\boldsymbol{I}_m$. The operator $\|\boldsymbol{x}\|_2$ denotes the Euclidean norm of $\boldsymbol{x}$. The transpose, conjugate, and conjugate transpose of $\boldsymbol{X}$ are denoted by $(\cdot)^{\mathrm{T}}$, $(\cdot)^*$, and $(\cdot)^{\mathrm{H}}$, respectively. The operations $\otimes$ and $*$ denote the Kronecker product and Khatri-Rao product, respectively. The notation $\mathrm{diag}(\cdot)$ represents a diagonal matrix with the arguments being its diagonal elements. The nth element of $\boldsymbol{x}$ is denoted by $[\,\boldsymbol{x}\,]_n$. The notation $\boldsymbol{X}_{\mathcal{S}}$ represents a matrix whose ith column is the $\mathcal{S}[i]$th column of $\boldsymbol{X}$.

II. Sensing-Aided MmWave/THz Systems

In this section, we present the sensing-aided mmWave/THz system model and discuss the geometric channel model. We also provide a brief overview of the conventional radio-frequency (RF)-based channel estimation schemes and channel mapping schemes.

A. Sensing-Aided MmWave/THz System Model

We consider a mmWave/THz multiple-input single-output (MISO) system where a BS equipped with $N = N_{\mathrm{h}} \times N_{\mathrm{v}}$ uniform rectangular array (URA) antenna serves a single-antenna UE. In this system, multiple RGB-depth (RGB-D) cameras are mounted on the BS to capture the surrounding wireless environments. The BS is equipped with a hybrid analog-digital beamforming architecture with N_{RF} RF chains, each connected with N analog phase shifters. Additionally, we consider an orthogonal frequency-division multiplexing (OFDM) system with S subcarriers operating at a carrier frequency of f_{c} and a system bandwidth of B. The subcarrier of the sth subcarrier is given by $f_s = f_{\mathrm{c}} + \left(s - \frac{S}{2}\right)\Delta f$ for $s = 0, 1, \ldots, S-1$ where $\Delta f \triangleq \frac{B}{S}$ is the subcarrier spacing. We consider a global coordinate system (GCS) where the BS and UE are located at $\boldsymbol{p}_{\mathrm{bs}} \in \mathbb{R}^3$ and $\boldsymbol{p}_{\mathrm{ue}} \in \mathcal{P}_{\mathrm{ue}}$, respectively, where $\mathcal{P}_{\mathrm{ue}}$ is the UE trajectory. Also, the position vector of the nth BS antenna is $\boldsymbol{p}_{\mathrm{bs},n} \in \mathbb{R}^3$ ($\boldsymbol{p}_{\mathrm{bs}} = \boldsymbol{p}_{\mathrm{bs},1}$).[1] The BS antenna orientation is specified by three angles: the bearing angle α, the downtilt angle β, and the slant angle γ, which correspond to rotations about the z-, y-, and x-axes, respectively [38].[2] We assume a time-division duplexing (TDD) system where the UE transmits uplink pilot signals (e.g., SRS), and the BS estimates the uplink channel information using the received pilot signals and the RGB and depth images obtained from the RGB-D camera. By exploiting the channel reciprocity of TDD system, the BS can recycle the acquired uplink channel information for downlink data transmission.

1) Pilot Measurement: The uplink pilot transmission is scheduled over M OFDM symbols and P pilot subcarriers. The sets of OFDM symbols and pilot subcarriers are denoted by $\mathcal{T} \triangleq \{t_1, t_2, \ldots, t_M\}$ and $\mathcal{S} \triangleq \{s_1, s_2, \ldots, s_P\}$, respectively. At each tth OFDM symbol, the UE transmits the pilot signal $x_t[s] = \sqrt{P_{\mathrm{tx}}}$ for all $s \in \mathcal{S}$ where P_{tx} is the transmit power per subcarrier. The received pilot signal $\mathbf{r}_t[s] \in \mathbb{C}^N$ at the tth symbol for the sth subcarrier is given by

$$\mathbf{r}_t[s] = \sqrt{P_{\mathrm{tx}}}\,\boldsymbol{h}[s] + \mathbf{n}_t[s] \tag{1}$$

where $\boldsymbol{h}[s] \in \mathbb{C}^N$ is the frequency-domain channel vector and $\mathbf{n}_t[s] \sim \mathcal{CN}(\mathbf{0}_N, \sigma_{\mathrm{n}}^2 \boldsymbol{I}_N)$ is the additive Gaussian noise. In turn, the BS combines the received pilot signal using a hybrid combiner $\boldsymbol{W}_t[s] = \boldsymbol{W}_{\mathrm{RF},t}\boldsymbol{W}_{\mathrm{BB},t}[s] \in \mathbb{C}^{N\times N_{\mathrm{RF}}}$ where $\boldsymbol{W}_{\mathrm{RF},t} \in \mathbb{C}^{N\times N_{\mathrm{RF}}}$ and $\boldsymbol{W}_{\mathrm{BB},t}[s] \in \mathbb{C}^{N_{\mathrm{RF}}\times N_{\mathrm{RF}}}$ are the RF and digital combiner matrices, respectively. Then the post-processed signal $\mathbf{y}_t[s] \in \mathbb{C}^{N_{\mathrm{RF}}}$ is given by

$$\mathbf{y}_t[s] \triangleq \boldsymbol{W}_t^{\mathrm{H}}[s]\,\mathbf{r}_t[s] \tag{2}$$

$$= \sqrt{P_{\mathrm{tx}}}\,\boldsymbol{W}_t^{\mathrm{H}}[s]\,\boldsymbol{h}[s] + \boldsymbol{W}_t^{\mathrm{H}}[s]\,\mathbf{n}_t[s]. \tag{3}$$

By concatenating $\mathbf{y}_t[s]$ for all $t \in \mathcal{T}$ to $\mathbf{y}[s] \triangleq \left[\mathbf{y}_{t_1}^{\mathrm{T}}[s]\ \mathbf{y}_{t_2}^{\mathrm{T}}[s]\ \cdots\ \mathbf{y}_{t_M}^{\mathrm{T}}[s]\right]^{\mathrm{T}} \in \mathbb{C}^{N_{\mathrm{RF}}M}$, we obtain

$$\mathbf{y}[s] = \sqrt{P_{\mathrm{tx}}}\,\boldsymbol{W}^{\mathrm{H}}[s]\,\boldsymbol{h}[s] + \mathbf{z}[s] \tag{4}$$

where $\boldsymbol{W}[s] \triangleq \left[\boldsymbol{W}_{t_1}[s]\ \boldsymbol{W}_{t_2}[s]\ \cdots\ \boldsymbol{W}_{t_M}[s]\right] \in \mathbb{C}^{N\times N_{\mathrm{RF}}M}$, $\mathbf{z}[s] \triangleq \left[\mathbf{z}_{t_1}^{\mathrm{T}}[s]\ \mathbf{z}_{t_2}^{\mathrm{T}}[s]\ \cdots\ \mathbf{z}_{t_M}^{\mathrm{T}}[s]\right]^{\mathrm{T}} \in \mathbb{C}^{N_{\mathrm{RF}}M}$, and $\mathbf{z}_t \triangleq \boldsymbol{W}_t^{\mathrm{H}}[s]\,\mathbf{n}_t[s] \in \mathbb{C}^{N_{\mathrm{RF}}}$. Finally, by combining $\mathbf{y}[s]$ for all $s \in \mathcal{S}$, we have the pilot measurement matrix $\mathbf{Y}_{\mathrm{pilot}} \triangleq \left[\mathbf{y}[s_1]\ \mathbf{y}[s_2]\ \cdots\ \mathbf{y}[s_P]\right] \in \mathbb{C}^{N_{\mathrm{RF}}M\times P}$ as

$$\mathbf{Y}_{\mathrm{pilot}} = \sqrt{P_{\mathrm{tx}}}\left[\boldsymbol{W}^{\mathrm{H}}[s_1]\,\boldsymbol{h}[s_1]\ \boldsymbol{W}^{\mathrm{H}}[s_2]\,\boldsymbol{h}[s_2]\right. \left.\cdots\ \boldsymbol{W}^{\mathrm{H}}[s_P]\,\boldsymbol{h}[s_P]\right] + \mathbf{Z} \tag{5}$$

$$= \sqrt{P_{\mathrm{tx}}}\left[\boldsymbol{W}^{\mathrm{H}}[s_1]\ \boldsymbol{W}^{\mathrm{H}}[s_2]\ \cdots\ \boldsymbol{W}^{\mathrm{H}}[s_P]\right] \times \mathrm{diag}\big(\boldsymbol{h}[s_1], \boldsymbol{h}[s_2], \ldots, \boldsymbol{h}[s_P]\big) + \mathbf{Z} \tag{6}$$

$$= \sqrt{P_{\mathrm{tx}}}\,\boldsymbol{W}^{\mathrm{H}}(\boldsymbol{I}_P * [\boldsymbol{H}]_{\mathcal{S}}) + \mathbf{Z} \tag{7}$$

where $\boldsymbol{H} = [\boldsymbol{h}[1]\ \boldsymbol{h}[2]\ \cdots\ \boldsymbol{h}[S]] \in \mathbb{C}^{N\times S}$ is the channel matrix, $\boldsymbol{W} \triangleq \left[\boldsymbol{W}^{\mathrm{T}}[s_1]\ \boldsymbol{W}^{\mathrm{T}}[s_2]\ \cdots\ \boldsymbol{W}^{\mathrm{T}}[s_P]\right]^{\mathrm{T}} \in \mathbb{C}^{NP\times N_{\mathrm{RF}}}$, and $\mathbf{Z} \triangleq \left[\mathbf{z}[s_1]\ \mathbf{z}[s_2]\ \cdots\ \mathbf{z}[s_P]\right] \in \mathbb{C}^{N_{\mathrm{RF}}M\times P}$.

2) Sensing Measurement: We use a circular array consisting of N_{cam} RGB-D cameras, each covering a specific angular sector. For instance, the Intel RealSense L515 RGB-D camera provides horizontal and vertical field of views (FoVs) of 70° and 45°, respectively, so the BS can cover the entire angular space with $\lceil\frac{360}{70}\rceil \times \lceil\frac{90}{45}\rceil = 12$ cameras. The RGB-D image capture is synchronized with the uplink pilot transmission of

[1]Note that the BS antenna located at the xth row and yth column of the antenna array is referred to as the nth BS antenna where $n = (x-1)N_{\mathrm{v}} + y$.

[2]For example, if the BS antenna array lies in the yz-plane, then the orientation angles are given by $(\alpha, \beta, \gamma) = (0°, 90°, 0°)$.

the UE. Specifically, the RGB-D image with $W \times H$ pixels is expressed as a three-dimensional tensor $\mathcal{D}_{\text{sens}} \in \mathbb{R}^{H\times W\times 4}$ containing the RGB data $\mathcal{D}_{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$ and the depth data $\mathcal{D}_{\text{depth}} \in \mathbb{R}^{H\times W\times 1}$:

$$\mathcal{D}_{\text{sens}} = [\, \mathcal{D}_{\text{rgb}};\ \mathcal{D}_{\text{depth}} \,] \tag{8}$$

where $[\cdot;\,\cdot]$ is concatenation along the third dimension.

B. Geometry-Based MmWave/THz Channel Model

In mmWave/THz systems, due to the high directivity and significant attenuation, the signal power is primarily concentrated along the line-of-sight (LOS) path and a few non-line-of-sight (NLOS) paths, which are determined by dominant scatterers near the BS. For example, in mmWave Frequency Range 2 (FR2) band, the number of dominant paths is lower than 6 and it further decreases to 4 in sub-THz band [39], [40], [41]. Based on the sparse scattering property, we adopt a geometric frequency-selective mmWave/THz channel model consisting of L discrete rays, where the first ray represents the LOS path and the remaining $L-1$ rays represent the NLOS paths [14], [38]. In this channel model, each ray is characterized by its azimuth angle of arrival (AOA) ϕ, zenith angle of arrival (ZOA) θ, delay τ, and path gains β.

Specifically, the frequency-domain channel response $h_n[s] \in \mathbb{C}$ from the UE to the nth BS antenna for the sth subcarrier is given by

$$\begin{aligned} h_n[s] = {} & \beta_{\text{los}}[s]\, e^{-j2\pi s\Delta f \tau_{\text{los}}} e^{j\frac{2\pi f_s}{c}\hat{\boldsymbol{r}}^{\mathrm{T}}(\theta_{\text{los}},\phi_{\text{los}})\,\boldsymbol{p}_{\text{bs},n}} \\ & + \sum_{l=1}^{L-1} \beta_l[s]\, e^{-j2\pi s\Delta f \tau_l} e^{j\frac{2\pi f_s}{c}\hat{\boldsymbol{r}}^{\mathrm{T}}(\theta_l,\phi_l)\,\boldsymbol{p}_{\text{bs},n}} \end{aligned} \tag{9}$$

where θ_l and ϕ_l are the ZOA and AOA, respectively, $\tau_l \in [0, \tau_{\max}]$ is the delay with $\tau_{\max}$ being the delay spread, and $\beta_l[s]$ is the path gain of the lth NLOS path for the sth subcarrier. Similarly, those of the LOS path are denoted as $(\theta_{\text{los}}, \phi_{\text{los}}, \tau_{\text{los}}, \beta_{\text{los}}[s])$. Also, $\hat{\boldsymbol{r}}(\theta_l,\phi_l) \in \mathbb{R}^3$ is the spherical unit vector associated with (θ_l,ϕ_l), defined as

$$\hat{\boldsymbol{r}}(\theta_l,\phi_l) \triangleq [\, \sin\theta_l\cos\phi_l \;\; \sin\theta_l\sin\phi_l \;\; \cos\theta_l \,]^{\mathrm{T}}. \tag{10}$$

The frequency-selective path gain $\beta_l[s]$, which accounts for the free-space path loss, antenna radiation pattern, reflection loss, and molecular absorption, is modeled as [42]

$$\beta_l[s] = \frac{F(\theta_l,\phi_l)\,|\Gamma_l(f_s)|e^{-\frac{1}{2}k_{\text{abs}}(f_s)c\tau_l}}{4\pi f_s \tau_l} \tag{11}$$

where $F(\theta_l,\phi_l)$ is the BS antenna radiation pattern and $k_{\text{abs}}(f_s)$ is the molecular absorption coefficient [43]. Also, $\Gamma_l(f_s)$ is the reflection coefficient defined as [44]

$$\Gamma_l(f_s) \triangleq \frac{\cos\psi_l - \sqrt{\eta_l^2 - \sin^2\psi_l}}{\cos\psi_l + \sqrt{\eta_l^2 - \sin^2\psi_l}} e^{-\frac{(4\pi f_s \sigma_l \cos\psi_l)^2}{2c^2}} \tag{12}$$

where c is the speed of light, ψ_l is the angle of incidence on the reflecting surface, $\eta_l \triangleq \frac{Z_0}{Z_l}$ is the refractive index where Z_0 and Z_l are the impedances of free-space and the reflecting material, respectively, and σ_l is the standard deviation of the rough surface height which characterizes the roughness of the reflecting surface of the scatterer associated with the lth NLOS path [17]. Due to the significant scattering and diffraction losses of mmWave/THz band signals, diffused and diffracted rays are heavily attenuated over distances greater than a few meters. Consequently, we only consider single-bounce reflected rays for the NLOS paths. Note that in the case of the LOS path, $\psi_{\text{los}} = 0$ and $|\Gamma_{\text{los}}(f_s)| = 1$.

By concatenating the channel responses of all BS antenna elements, we obtain the frequency-domain channel vector $\boldsymbol{h}[s] \triangleq [\, h_1[s]\; h_2[s] \cdots h_N[s] \,]^{\mathrm{T}} \in \mathbb{C}^N$ as

$$\begin{aligned} \boldsymbol{h}[s] = {} & \beta_{\text{los}}[s]\, e^{-j2\pi s\Delta f \tau_{\text{los}}} \boldsymbol{a}(\theta_{\text{los}},\phi_{\text{los}}) \\ & + \sum_{l=1}^{L-1} \beta_l[s]\, e^{-j2\pi s\Delta f \tau_l} \boldsymbol{a}(\theta_l,\phi_l) \end{aligned} \tag{13}$$

where $\boldsymbol{a}(\theta_l,\phi_l) \in \mathbb{C}^N$ is the array response vector defined as

$$\begin{aligned} \boldsymbol{a}(\theta_l,\phi_l) \triangleq \Big[& e^{j\frac{2\pi f_s}{c}\hat{\boldsymbol{r}}^{\mathrm{T}}(\theta_l,\phi_l)\,\boldsymbol{p}_{\text{bs},1}} \; e^{j\frac{2\pi f_s}{\lambda}\hat{\boldsymbol{r}}^{\mathrm{T}}(\theta_l,\phi_l)\,\boldsymbol{p}_{\text{bs},2}} \\ & \cdots \; e^{j\frac{2\pi f_s}{c}\hat{\boldsymbol{r}}^{\mathrm{T}}(\theta_l,\phi_l)\,\boldsymbol{p}_{\text{bs},N}} \Big]^{\mathrm{T}}. \end{aligned} \tag{14}$$

The combined channel matrix $\boldsymbol{H} \in \mathbb{C}^{N\times S}$ is given by

$$\boldsymbol{H} = [\, \boldsymbol{h}[1]\, \boldsymbol{h}[2] \cdots \boldsymbol{h}[S] \,]. \tag{15}$$

C. Conventional Channel Estimation Techniques

1) RF-Based Channel Estimation: In traditional communications systems, linear estimation techniques such as LS and linear minimum mean square error (LMMSE) are used to extract $\boldsymbol{H}$ from $\mathbf{Y}_{\text{pilot}}$. While the linear channel estimation techniques have been effective in Long-Term Evolution (LTE) and mmWave Frequency Range 1 (FR1) band systems with relatively small antenna arrays, they might not perform well in mmWave FR2 and sub-THz band systems with massive antenna arrays due to the significant pilot overhead which scales with the number of antennas. As a remedy, parametric channel estimation techniques that estimate the GCPs (i.e., angles, delays, and path gains) instead of full-dimensional channel matrix have been popularly used [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. Since the dominant paths are scarce in high-frequency bands, these approaches can effectively reduce the pilot overhead.

The major limitation of the parametric channel estimation techniques is that, while the GCPs can capture the specific channel for a given UE position, they fall short when describing the channel for arbitrary UE positions. Specifically, these parameters vary continuously within the same set of scatterers, but shift discontinuously when the scatterers change, exhibiting piecewise continuity. To learn the piecewise continuous distributions of GCPs and uncover the fundamental relationship between the UE position and the channel, it is essential to consider not only the analytic channel parameters but also the physical characteristics of the surrounding environments.

2) Channel Mapping: Recently, there has been growing interest in channel mapping approaches that learn the spatial distributions of channel (i.e., CKM) have gained much attention [26]. By directly inferring CSI from the UE position

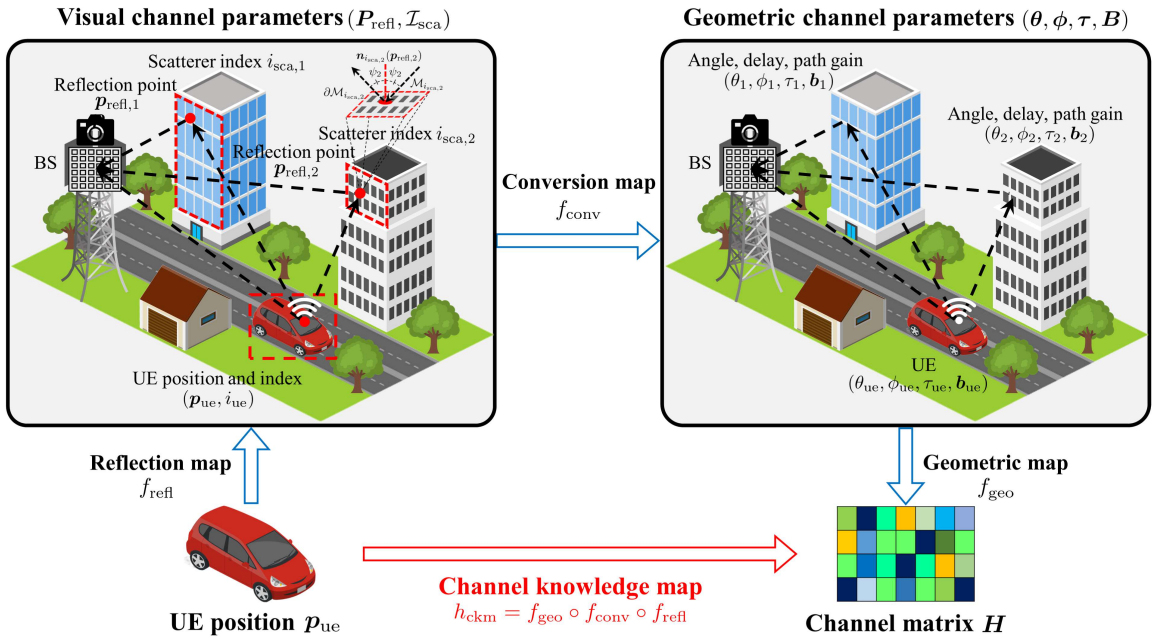


Fig. 1. Illustration of CKM, which represents the mapping between the UE position and the frequency-domain channel matrix. The CKM h_{ckm} is expressed as a composition of the reflection map f_{refl}, the conversion map f_{conv}, and the geometric map f_{geo}.

using the learned CKM, these methods eliminate the need for pilot measurements, thereby significantly reducing latency and resource overhead. To construct the CKM, various methods have been employed, including Kriging interpolation, matrix completion, and DL models (e.g., convolutional neural networks (CNNs) and autoencoder). However, the conventional channel mapping techniques learn the direct end-to-end mapping between the UE position and the channel information, which requires a prohibitive amount of training data due to the high dimensionality of the CSI. To address this issue, channel mapping techniques that exploit DT, a digital replica of the real world, have been proposed [30]. In DT-based channel mapping techniques, the physical properties of the surrounding environment are emulated through channel emulators, using which the ray-tracing simulator identifies the signal propagation paths [31]. This allows direct acquisition of channel information for any UE position.

However, to accurately emulate real-world conditions, the DT-based channel mapping techniques require a substantial volume of high-quality training data. Beyond initial training, these emulators must undergo continuous adjustments and retraining to remain aligned with system dynamics. This need for ongoing data collection and recalibration introduces significant computational overhead, posing challenges to real-time decision-making in rapidly changing wireless environments.

III. Visual Channel Parameter-Based Channel Knowledge Map

In essence, a wireless channel characterizes the propagation of radio waves as they interact (e.g., reflection, scattering, and diffraction) with the environment. These interactions are modeled using contextual information about the surrounding environment. The fundamental relationship between the UE position and the wireless channel is called the CKM [26]. The CKM is determined by the scattering geometry around the BS and UE. Traditionally, multipath propagation was viewed as a hindrance that causes signal degradation, and various diversity techniques have been developed to combat the multipath propagation [45], [46], [47], [48], [49], [50], [51]. However, when the scattering geometry is properly identified, it facilitates multipath diversity, thereby improving channel estimation accuracy and enhancing system capacity [52]. To formally define CKM, we introduce the notion of VCPs, i.e., the reflection points and scatterer indices, which visually describe the propagation dynamics. We then define the CKM as a composition of three key mappings: reflection, conversion, and geometric maps (see Fig. 1).

A. Geometric Map From Geometric Channel Parameters to Channel Matrix

Recall that the frequency-domain channel matrix $\boldsymbol{H}$ is a function of the GCPs, i.e., ZOA θ, AOA ϕ, delay τ, and path gain β (see (13)). That is, the relationship between the GCPs and the frequency-domain channel matrix $\boldsymbol{H}$ can be expressed as a geometric map $f_{\text{geo}} : \mathbb{R}^L \times \mathbb{R}^L \times \mathbb{R}^L \times \mathbb{R}^{L\times P} \to \mathbb{C}^{N\times P}$:

$$\boldsymbol{H} = f_{\text{geo}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{B}) \tag{16}$$

where $\boldsymbol{\theta} = [\theta_{\text{los}}\, \theta_1\, \theta_2 \cdots \theta_{L-1}]^{\text{T}}$, $\boldsymbol{\phi} = [\phi_{\text{los}}\, \phi_1\, \phi_2 \cdots \phi_{L-1}]^{\text{T}}$, $\boldsymbol{\tau} = [\tau_{\text{los}}\, \tau_1\, \tau_2 \cdots \tau_{L-1}]^{\text{T}}$, and $\boldsymbol{B} \in \mathbb{R}^{L\times S}$ is a matrix whose (l, s)th element is $\beta_{\text{los}}[s]$ for $l = 1$ and $\beta_{l-1}[s]$ for $l \geq 2$.

B. Conversion Map From Visual Channel Parameters to Geometric Channel Parameters

The primary factors determining the GCPs are the scatterers. For example, AOAs, ZOAs, and time delays are determined by the relative positions of the reflection points with respect

to the BS and UE, while the path gains are affected by the material properties of the scatterers as well. Let $\mathcal{I}_{\rm obj} \triangleq \{1, 2, \ldots, N_{\rm obj}\}$ be the set of objects (e.g., buildings and walls) within the wireless environment where $N_{\rm obj}$ is the number of objects. Each object i is characterized by its boundary $\partial\mathcal{M}_i \subseteq \mathbb{R}^3$ and class c_i. We collectively refer to these as the abstract environmental information $\mathcal{E}_{\rm abs}$:

$$\mathcal{E}_{\rm abs} \triangleq \left\{(\partial\mathcal{M}_i, c_i) \,:\, i \in \mathcal{I}_{\rm obj}\right\}. \tag{17}$$

Note that $\mathcal{E}_{\rm abs}$ can be readily extracted from RGB-D images using object detection or semantic segmentation techniques.

Among the objects, let $i_{\rm ue}$ and $i_{{\rm sca},l}$ be the indices of objects that correspond to the UE and the scatterer for the lth NLOS path, respectively.[3] Also, let $\boldsymbol{p}_{{\rm refl},l} \in \mathbb{R}^3$ be the position of the reflection point at the scatterer $i_{{\rm sca},l}$. Then, the VCPs are defined as the combined reflection point matrix $\boldsymbol{P}_{\rm refl} \in \mathbb{R}^{3\times L}$ and the scatterer index set $\mathcal{I}_{\rm sca} \in \mathcal{I}_{\rm obj}^L$:

$$\boldsymbol{P}_{\rm refl} \triangleq [\, \boldsymbol{p}_{\rm ue}\ \boldsymbol{p}_{{\rm refl},1}\ \boldsymbol{p}_{{\rm refl},2}\ \cdots\ \boldsymbol{p}_{{\rm refl},L-1}\,] \tag{18}$$

$$\mathcal{I}_{\rm sca} \triangleq \{\, i_{\rm ue},\, i_{{\rm sca},1},\, i_{{\rm sca},2},\, \ldots,\, i_{{\rm sca},L-1}\,\}. \tag{19}$$

In contrast to the GCPs $(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{B})$, which provide only analytic descriptions of the signal propagation, the VCPs $(\boldsymbol{P}_{\rm refl}, \mathcal{I}_{\rm sca})$ enable a direct visualization of the propagation environment. Hence, by exploiting $(\boldsymbol{P}_{\rm refl}, \mathcal{I}_{\rm sca})$ to model $\boldsymbol{H}$, one can effectively capture the interactions between the signals and the wireless environment. Specifically, θ_l, ϕ_l, and τ_l can be expressed as functions of $\boldsymbol{p}_{\rm ue}$ and $\boldsymbol{p}_{{\rm refl},l}$ as

$$\theta_l = \arccos\left(\frac{\left[\,\boldsymbol{R}^{-1}(\boldsymbol{p}_{\rm bs} - \boldsymbol{p}_{{\rm refl},l})\,\right]_3}{\left\|\boldsymbol{R}^{-1}(\boldsymbol{p}_{\rm bs} - \boldsymbol{p}_{{\rm refl},l})\right\|_2}\right) \tag{20}$$

$$\phi_l = {\rm atan2}\Big(\left[\,\boldsymbol{R}^{-1}(\boldsymbol{p}_{\rm bs} - \boldsymbol{p}_{{\rm refl},l})\,\right]_2, \left[\,\boldsymbol{R}^{-1}(\boldsymbol{p}_{\rm bs} - \boldsymbol{p}_{{\rm refl},l})\,\right]_1\Big) \tag{21}$$

$$\tau_l = \frac{1}{c}\Big(\left\|\boldsymbol{p}_{\rm bs} - \boldsymbol{p}_{{\rm refl},l}\right\|_2 + \left\|\boldsymbol{p}_{{\rm refl},l} - \boldsymbol{p}_{\rm ue}\right\|_2\Big) \tag{22}$$

where $\boldsymbol{R} \triangleq \boldsymbol{R}_x(\alpha)\boldsymbol{R}_y(\beta)\boldsymbol{R}_z(\gamma) \in \mathbb{R}^{3\times 3}$ is the rotation matrix for the BS antenna orientation [38]. Also, recall that $\beta_l[s]$ depends on τ_l, the angle of incidence on the reflecting surface ψ_l, the refractive index η_l, and the surface roughness σ_l (see (11) and (12)). Since $\psi_l = \frac{1}{2}\arccos\left(\frac{\|\boldsymbol{p}_{\rm bs}-\boldsymbol{p}_{{\rm refl},l}\|_2^2+\|\boldsymbol{p}_{{\rm refl},l}-\boldsymbol{p}_{\rm ue}\|_2^2-\|\boldsymbol{p}_{\rm bs}-\boldsymbol{p}_{\rm ue}\|_2^2}{2\|\boldsymbol{p}_{\rm bs}-\boldsymbol{p}_{{\rm refl},l}\|_2\|\boldsymbol{p}_{{\rm refl},l}-\boldsymbol{p}_{\rm ue}\|_2}\right)$ and (η_l, σ_l) can be obtained from the object class $c_{i_{{\rm sca},l}}$, $\beta_l[s]$ is expressed as a function of $(\boldsymbol{p}_{\rm ue}, \boldsymbol{p}_{{\rm refl},l}, i_{{\rm sca},l})$, parametrized by $\mathcal{E}_{\rm abs}$ as[4]

$$\beta_l[s] = \beta(\boldsymbol{p}_{\rm ue}, \boldsymbol{p}_{{\rm refl},l}, i_{{\rm sca},l}, f_s; \mathcal{E}_{\rm abs}). \tag{23}$$

Similarly, the GCPs of the LOS path can be expressed as functions of $\boldsymbol{p}_{\rm ue}$ and $i_{\rm ue}$.

Remark 1: The conversion map $f_{\rm conv} : \mathbb{R}^{3\times L} \times \mathcal{I}_{\rm obj}^L \to \mathbb{R}^L \times \mathbb{R}^L \times \mathbb{R}^L \times \mathbb{R}^{L\times P}$ from the VCPs $(\boldsymbol{P}_{\rm refl}, \mathcal{I}_{\rm sca})$ to the GCPs $(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{B})$ is defined as (20)-(23):

$$(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{B}) = f_{\rm conv}(\boldsymbol{P}_{\rm refl}, \mathcal{I}_{\rm sca}; \mathcal{E}_{\rm abs}). \tag{24}$$

Also, $f_{\rm conv}$ is parametrized by the abstract environmental information $\mathcal{E}_{\rm abs}$, encompassing the boundaries and classes of the objects. □

C. Reflection Map From Ue Position to Visual Channel Parameters

The conversion map $f_{\rm conv}$ in (24) is deterministic and can be characterized using only the abstract information $\mathcal{E}_{\rm abs}$ about the objects. In contrast, obtaining $\boldsymbol{P}_{\rm refl}$ and $\mathcal{I}_{\rm sca}$ requires comprehensive and detailed knowledge of the physical characteristics of the objects. Specifically, let $\mathcal{M}_i \subseteq \mathbb{R}^3$ be the set of points on the reflection surface and $\boldsymbol{n}_i : \mathcal{M}_i \to \mathbb{R}^3$ be the normal vector field of the ith object.[5] We collectively refer to these as the detailed environmental information $\mathcal{E}_{\rm env}$:

$$\mathcal{E}_{\rm det} \triangleq \left\{(\mathcal{M}_i, \boldsymbol{n}_i) \,:\, i \in \mathcal{I}_{\rm obj}\right\}. \tag{25}$$

Essentially, $\mathcal{E}_{\rm det}$ governs the physical interactions of radio waves with the surrounding environment, thereby determining the signal propagation paths.

Using $\mathcal{E}_{\rm det}$, the scatterer index $i_{{\rm sca},l} \in \mathcal{I}_{\rm obj}$ and the reflection point $\boldsymbol{p}_{{\rm refl},l} \in \mathcal{M}_{i_{{\rm sca},l}}$ of the lth NLOS path can be obtained by identifying the object index and the point on the reflection plane $\mathcal{M}_{i_{{\rm sca},l}}$ satisfying the law of reflection with respect to the position vectors of the BS and UE as[6]

$$\begin{aligned}&\angle\big(\boldsymbol{p}_{\rm ue} - \boldsymbol{p}_{{\rm refl},l}, \boldsymbol{n}_{i_{{\rm sca},l}}(\boldsymbol{p}_{{\rm refl},l})\big)\\ &\qquad = \angle\big(\boldsymbol{p}_{\rm bs} - \boldsymbol{p}_{{\rm refl},l}, \boldsymbol{n}_{i_{{\rm sca},l}}(\boldsymbol{p}_{{\rm refl},l})\big)\end{aligned} \tag{26}$$

where $\angle(\boldsymbol{a}, \boldsymbol{b}) \triangleq \arccos\left(\frac{\boldsymbol{a}^{\rm T}\boldsymbol{b}}{\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2}\right)$ denotes the angle between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Thus, the relationship between $\boldsymbol{p}_{\rm ue}$ and $(\boldsymbol{P}_{\rm refl}, \mathcal{I}_{\rm sca})$ can be expressed as the reflection map $f_{\rm refl} : \mathcal{P}_{\rm ue} \to \mathbb{R}^{3\times L} \times \mathcal{I}_{\rm obj}^L$, parametrized by $\mathcal{E}_{\rm det}$:

$$(\boldsymbol{P}_{\rm refl}, \mathcal{I}_{\rm sca}) = f_{\rm refl}(\boldsymbol{p}_{\rm ue}; \mathcal{E}_{\rm det}). \tag{27}$$

Finally, by substituting (24) and (27) into (17), we obtain the CKM $h_{\rm ckm} : \mathcal{P}_{\rm ue} \to \mathbb{C}^{N\times P}$ from $\boldsymbol{p}_{\rm ue}$ to $\boldsymbol{H} \in \mathbb{C}^{N\times P}$ as

$$\boldsymbol{H} = f_{\rm geo}(f_{\rm conv}\big(\boldsymbol{P}_{\rm refl}, \mathcal{I}_{\rm sca}; \mathcal{E}_{\rm abs})\big) \tag{28}$$

$$= f_{\rm geo}\big(f_{\rm conv}(f_{\rm refl}(\boldsymbol{p}_{\rm ue}; \mathcal{E}_{\rm det}); \mathcal{E}_{\rm abs})\big) \tag{29}$$

$$= h_{\rm ckm}(\boldsymbol{p}_{\rm ue}; \mathcal{E}_{\rm env}) \tag{30}$$

where $\mathcal{E}_{\rm env} = \left\{\mathcal{E}_{\rm abs}, \mathcal{E}_{\rm det}\right\}$ represents the comprehensive environmental information.

Remark 2: The CKM $h_{\rm ckm} : \mathcal{P}_{\rm ue} \to \mathbb{C}^{N\times P}$ between the UE position $\boldsymbol{p}_{\rm ue}$ and the channel matrix $\boldsymbol{H}$ is expressed as □

$$h_{\rm ckm} = f_{\rm geo} \circ f_{\rm conv} \circ f_{\rm refl}. \tag{31}$$

[3]Since we only consider single-bounce reflected rays, each NLOS path is associated with a single scatterer.

[4]3GPP specifies five material classes: concrete, glass, wood, metal, and brick [38]. Furthermore, to acquire more specific material properties, one can employ deep texture recognition models that classify materials based on texture patterns observed in images [53].

[5]The normal vector field is a map on $\mathcal{M}_i$ that assigns to each point $\boldsymbol{p} \in \mathcal{M}_i$ a vector $\boldsymbol{n}_i(\boldsymbol{p})$ that is perpendicular to the tangent space $\mathcal{T}_{\boldsymbol{p}}\mathcal{M}_i$ at $\boldsymbol{p}$.

[6]For planar reflection surfaces (e.g., building and wall), the image method is commonly used to find $\boldsymbol{p}_{{\rm refl},l}$. This method creates a virtual image of the UE and determines the intersection of the line connecting the virtual UE to the BS with the reflection plane.

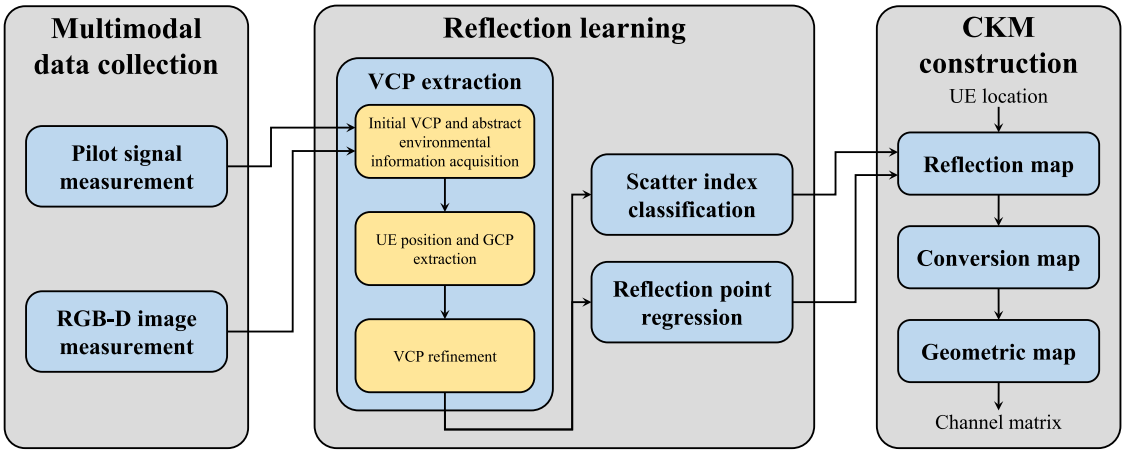


Fig. 2. Overall process of LMM-CE.

It is worth mentioning that h_{ckm} is parametrized by $\mathcal{E}_{\mathrm{env}}$, which includes both abstract and detailed information about the objects. Unfortunately, traditional analytic pilot measurements do not explicitly convey this contextual environmental information. Therefore, to determine the CKM and achieve environment-awareness, one should exploit both high-resolution sensing measurements containing semantic information and pilot measurements containing analytic information.

IV. Large Multimodal Model-Based Environment-Aware Channel Estimation

The primary objective of the proposed LMM-CE scheme is to identify the intrinsic relationship (i.e., CKM) between the UE position $\boldsymbol{p}_{\mathrm{ue}}$ and the frequency-domain channel matrix $\boldsymbol{H}$. As shown in Remark 2, the CKM h_{ckm} is a composition of three key maps: 1) the reflection map f_{refl} which locates the reflection points and scatterer indices $(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}})$ for a given UE position $\boldsymbol{p}_{\mathrm{ue}}$; 2) the conversion map f_{conv} which converts the VCPs $(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}})$ to the GCPs $(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{B})$; and 3) the geometric map f_{geo} which computes $\boldsymbol{H}$ using $(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{B})$. Since f_{geo} and f_{conv} are explicit mappings fully defined in (16) and (24), the main challenge is to determine f_{refl} in (27), which is an implicit mapping characterized by the detailed environmental information $\mathcal{E}_{\mathrm{det}}$. Unfortunately, acquiring $\mathcal{E}_{\mathrm{det}}$ is very difficult in practice since it requires extensive measurement campaigns to characterize surrounding environments along with continuous retraining to adapt to environmental changes.

To obtain the reflection map f_{refl} without acquiring $\mathcal{E}_{\mathrm{det}}$, we propose a *reflection learning* technique that extracts labeled samples $(\boldsymbol{p}_{\mathrm{ue}}, (\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}}))$ from the sensing and pilot measurements $(\mathcal{D}_{\mathrm{sens}}, \mathbf{Y}_{\mathrm{pilot}})$ and then learns their complex relationship $f(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}})$ using DL:

$$f_{\mathrm{refl}}(\boldsymbol{p}_{\mathrm{ue}}; \mathcal{E}_{\mathrm{det}}) = \underset{\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}}}{\operatorname{argmax}} \; f(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}}). \tag{32}$$

As a main DL engine for the reflection learning, we use an LMM, a generative AI model specialized in extracting correlated features across multimodal data and generating subsequent data. By extracting the visual features (e.g., boundaries and classes of the objects) from $\mathcal{D}_{\mathrm{sens}}$ and aligning them with the geometric features (e.g., angles, delays, and path gains) in $\mathbf{Y}_{\mathrm{pilot}}$, LMM-CE can identify the target VCPs $(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}})$ among the visual feature set. Furthermore, by capturing the underlying probability distribution of the VCPs and predicting the subsequent parameters, LMM-CE can learn f_{refl}, thereby obtaining CKM h_{ckm} and achieving environment-awareness. The overall process of LMM-CE is illustrated in Fig. 2.

A. Basics of Large Multimodal Model

The LMM architecture consists of three key components: 1) a multimodal encoder that converts multimodal inputs to unified latent representations; 2) a backbone that processes these representations to extract the features; and 3) a language modelling (LM) head that generates the subsequent words.

1) Multimodal Encoder: In the multimodal encoder, a sequence of inputs (e.g., texts and images) from I modalities are initially converted into discrete numeric representations, referred to as tokens, $\{v_{1,n}\}_{n=1}^{N_1}, \{v_{2,n}\}_{n=1}^{N_2}, \ldots, \{v_{I,n}\}_{n=1}^{N_I} \subseteq \mathcal{V}$ where $\mathcal{V} = \{1, 2, \ldots, V\}$ is the token vocabulary and N_i is the number of tokens of the ith modality. Then, each token $v_{i,n}$ is encoded into an embedding vector $\boldsymbol{e}_{i,n} \in \mathbb{R}^{e_i}$ using a modality-specific encoder $f_{\mathrm{encoder},i} : \mathcal{V} \to \mathbb{R}^{e_i}$ as

$$\boldsymbol{e}_{i,n} = f_{\mathrm{encoder},i}(v_{i,n}) \tag{33}$$

Next, $\boldsymbol{e}_{i,n}$ is projected into a latent vector $\boldsymbol{y}_{i,n}$ within the unified latent space $\mathbb{R}^d$ as

$$\boldsymbol{y}_{i,n} = \boldsymbol{W}_{\mathrm{proj},i} \boldsymbol{e}_{i,n} + \boldsymbol{b}_{\mathrm{proj},i} \tag{34}$$

where $\boldsymbol{W}_{\mathrm{proj},i} \in \mathbb{R}^{d \times e_i}$ is the projection matrix and $\boldsymbol{b}_{\mathrm{proj},i} \in \mathbb{R}^d$ is the bias vector. The latent vectors are combined into the latent matrix $\boldsymbol{Y}_i = [\, \boldsymbol{y}_{i,1} \; \boldsymbol{y}_{i,2} \cdots \boldsymbol{y}_{i,N_i} \,]^{\mathrm{T}} \in \mathbb{R}^{N_i \times d}$. Lastly, the latent matrices $\{\boldsymbol{Y}_i\}_{i=1}^{I}$ of all modalities are concatenated into the latent representation matrix $\boldsymbol{Y} \in \mathbb{R}^{N_{\mathrm{token}} \times d}$ as

$$\boldsymbol{Y} = \left[\, \boldsymbol{Y}_1^{\mathrm{T}} \; \boldsymbol{Y}_2^{\mathrm{T}} \cdots \boldsymbol{Y}_I^{\mathrm{T}} \, \right]^{\mathrm{T}} \tag{35}$$

where $N_{\mathrm{token}} = \sum_{i=1}^{I} N_i$ is the total number of tokens.

2) Backbone: To extract intra-modal and cross-modal features from the latent representation matrix $\boldsymbol{Y}$, the backbone employs L transformer blocks, each consisting multi-head attention layer, layer normalization layer, feed-forward layer, and another layer normalization layer [54]. In the multi-head attention layer, $\boldsymbol{Y}$ passes through H attention heads connected in parallel. The role of the attention head is to measure the correlations between the input elements and then assign weights (i.e., attention score) based on these correlations. Specifically, each hth attention head constructs three different embedding matrices from $\boldsymbol{Y}$, i.e., the query $\boldsymbol{Q}_h = \boldsymbol{Y}\boldsymbol{W}_{\mathrm{q},h} \in \mathbb{R}^{N_{\mathrm{token}} \times d_h}$, the key $\boldsymbol{K}_h = \boldsymbol{Y}\boldsymbol{W}_{\mathrm{k},h} \in \mathbb{R}^{N_{\mathrm{token}} \times d_h}$, and the value $\boldsymbol{V}_h = \boldsymbol{Y}\boldsymbol{W}_{\mathrm{v},h} \in \mathbb{R}^{N_{\mathrm{token}} \times d_h}$ where $\boldsymbol{W}_{\mathrm{q},h}, \boldsymbol{W}_{\mathrm{k},h}, \boldsymbol{W}_{\mathrm{v},h} \in \mathbb{R}^{d \times d_h}$ are the weight matrices and $d_h = \frac{d}{H}$. Then the attention score $\boldsymbol{S}_h \in \mathbb{R}^{N_{\mathrm{token}} \times N_{\mathrm{token}}}$ is obtained from the scaled dot product of $\boldsymbol{Q}_h$ and $\boldsymbol{K}_h$, followed by a row-wise softmax operation as

$$\boldsymbol{S}_h = f_{\mathrm{smax}}\left(\frac{\boldsymbol{Q}_h \boldsymbol{K}_h^{\mathrm{T}}}{\sqrt{d_h}} + \boldsymbol{M}\right). \tag{36}$$

Note that in (36), to ensure that only the present and past information is used for predicting future tokens, a mask matrix $\boldsymbol{M} \in \mathbb{R}^{N_{\mathrm{token}} \times N_{\mathrm{token}}}$ is employed whose (i,j)th element is 0 if $i \leq j$ and $-\infty$ otherwise. Then, the output $\boldsymbol{Y}_{\mathrm{att},h} \in \mathbb{R}^{N_{\mathrm{token}} \times d_h}$ of the hth attention head is computed as the product of $\boldsymbol{S}_h$ and $\boldsymbol{V}_h$ as

$$\boldsymbol{Y}_{\mathrm{att},h} = \boldsymbol{S}_h \boldsymbol{V}_h \,. \tag{37}$$

The outputs $\{\boldsymbol{Y}_{\mathrm{att},h}\}_{h=1}^{H}$ from all attention heads are concatenated and then multiplied by the output weight matrix $\boldsymbol{W}_{\mathrm{o}} \in \mathbb{R}^{d \times d}$ to generate the attention output $\boldsymbol{Y}_{\mathrm{att}} \in \mathbb{R}^{N_{\mathrm{token}} \times d}$:

$$\boldsymbol{Y}_{\mathrm{att}} = [\, \boldsymbol{Y}_{\mathrm{att},1}\ \boldsymbol{Y}_{\mathrm{att},2} \cdots \boldsymbol{Y}_{\mathrm{att},H} \,]\, \boldsymbol{W}_{\mathrm{o}} \,. \tag{38}$$

Once $\boldsymbol{Y}_{\mathrm{att}}$ is obtained, the residual connection is applied, followed by the layer normalization which scales and shifts the input to have zero mean and unit variance [54]:

$$\bar{\boldsymbol{Y}} = f_{\mathrm{ln}}(\boldsymbol{Y} + \boldsymbol{Y}_{\mathrm{att}}) \,. \tag{39}$$

Then, $\bar{\boldsymbol{Y}}$ is passed through the feed-forward network (FFN) as

$$\tilde{\boldsymbol{Y}} = f_{\mathrm{ffn}}(\bar{\boldsymbol{Y}}) \,. \tag{40}$$

Subsequently, the output matrix $\boldsymbol{Z}^{(1)} \in \mathbb{R}^{N_{\mathrm{token}} \times d}$ for the first transformer block is obtained by applying another residual connection and layer normalization as

$$\boldsymbol{Z}^{(1)} = f_{\mathrm{ln}}(\bar{\boldsymbol{Y}} + \tilde{\boldsymbol{Y}}) \,. \tag{41}$$

After passing through L transformer blocks, the final feature matrix $\boldsymbol{Z} \in \mathbb{R}^{N_{\mathrm{token}} \times d}$ is obtained as $\boldsymbol{Z} = \boldsymbol{Z}^{(L)}$.

3) Language Modelling Head: The LM head utilizes the extracted feature matrix $\boldsymbol{Z}$ to compute the probability distribution for the next token, conditioned on the previous tokens. To do so, each nth row vector $\boldsymbol{z}_n$ of $\boldsymbol{Z}$ is transformed into a logit vector $\boldsymbol{l}_n \in \mathbb{R}^V$, which represents a vector of unnormalized log-probabilities over the token vocabulary $\mathcal{V}$ for the nth token v_n, conditioned on the previous tokens $\{v_{n'}\}_{n'=1}^{n-1}$ as

$$\boldsymbol{l}_n = \boldsymbol{W}_{\mathrm{lm}} \boldsymbol{z}_n + \boldsymbol{b}_{\mathrm{lm}} \tag{42}$$

where $\boldsymbol{W}_{\mathrm{lm}} \in \mathbb{R}^{V \times d}$ is the weight matrix and $\boldsymbol{b}_{\mathrm{lm}} \in \mathbb{R}^V$ is the bias. Then, the token probability vector $\boldsymbol{p}_n \in \mathbb{R}^V$ is obtained by applying the softmax function to $\boldsymbol{l}_n$ as

$$\boldsymbol{p}_n = f_{\mathrm{smax}}(\boldsymbol{l}_n) \,. \tag{43}$$

Finally, each v_n is sequentially determined as the token with the highest probability for $n = 1, 2, \ldots, N_{\mathrm{token}}$:

$$v_n = \underset{v \in \mathcal{V}}{\mathrm{argmax}}\ [\, \boldsymbol{p}_n \,]_v \,. \tag{44}$$

This next-token prediction is a pivotal operation of LMMs that enables capturing complex patterns and relationships in multimodal data and generating subsequent words [2]. Furthermore, domain-specific knowledge of communications system such as mathematical channel models and standard specifications can be incorporated as natural language instructions so that LMMs can address complex problems in wireless communications.

B. Overview of Reflection Learning

As mentioned, the primary task in establishing CKM is to identify the reflection map f_{refl} between the UE position $\boldsymbol{p}_{\mathrm{ue}}$ and the VCPs $(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}})$ (see (27)). To handle the task at hand, we model the conditional probability distribution $f(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}})$ using an LMM. In the following remark, we introduce key properties of $(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}})$ used to develop our reflection learning technique.

Remark 3: The conditional probability distribution $f(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}})$ can be expressed as

$$f(\boldsymbol{P}_{\mathrm{refl}}, \mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}}) = f(\boldsymbol{P}_{\mathrm{refl}} \,|\, \mathcal{I}_{\mathrm{sca}}, \boldsymbol{p}_{\mathrm{ue}})\, f(\mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}}) \,. \tag{45}$$

This implies that the scatterer indices $\mathcal{I}_{\mathrm{sca}}$ and the reflection points $\boldsymbol{P}_{\mathrm{refl}}$ can be determined sequentially. To be specific, since $\mathcal{I}_{\mathrm{sca}}$ exhibit discrete change with UE mobility, identifying $f(\mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}})$ is a classification task. In contrast, as $\boldsymbol{P}_{\mathrm{refl}}$ change continuously when $\mathcal{I}_{\mathrm{sca}}$ remains fixed, identifying $f(\boldsymbol{P}_{\mathrm{refl}} \,|\, \mathcal{I}_{\mathrm{sca}}, \boldsymbol{p}_{\mathrm{ue}})$ is a regression task. □

Remark 4: The reflection point $\boldsymbol{p}_{\mathrm{refl},l}$ and its corresponding scatterer index $i_{\mathrm{sca},l}$ are determined independently for each NLOS path. Consequently, the conditional probabilities $f(\boldsymbol{P}_{\mathrm{refl}} \,|\, \mathcal{I}_{\mathrm{sca}}, \boldsymbol{p}_{\mathrm{ue}})$ and $f(\mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}})$ can be decomposed as

$$f(\boldsymbol{P}_{\mathrm{refl}} \,|\, \mathcal{I}_{\mathrm{sca}}, \boldsymbol{p}_{\mathrm{ue}}) = \prod_{i \in \mathcal{I}_{\mathrm{obj}}} f(\boldsymbol{p}_{\mathrm{refl},i} \,|\, S_i, \boldsymbol{p}_{\mathrm{ue}}) \tag{46}$$

$$f(\mathcal{I}_{\mathrm{sca}} \,|\, \boldsymbol{p}_{\mathrm{ue}}) = \prod_{i \in \mathcal{I}_{\mathrm{sca}}} f(S_i \,|\, \boldsymbol{p}_{\mathrm{ue}}) \prod_{j \in \mathcal{I}_{\mathrm{obj}} \setminus \mathcal{I}_{\mathrm{sca}}} \big(1 - f(S_j \,|\, \boldsymbol{p}_{\mathrm{ue}})\big) \tag{47}$$

where $S_i \in \{0, 1\}$ is a binary indicator defined as

$$S_i = \begin{cases} 1 & \text{if the object } i \text{ is used as a scatterer} \\ 0 & \text{otherwise} \,. \end{cases} \tag{48}$$

Also, $\boldsymbol{p}_{\mathrm{refl},i} \in \mathbb{R}^3$ is the reflection point on the object i. □

Based on these observations, we divide the reflection learning process into three main stages as follows (see Fig. 3):

- **VCP extraction**: The labeled samples $\mathcal{X}_{\mathrm{samp}} = \{(\boldsymbol{p}_{\mathrm{ue}}^{(n)}, (\boldsymbol{P}_{\mathrm{refl}}^{(n)}, \mathcal{I}_{\mathrm{sca}}^{(n)}))\}_{n=1}^{N_{\mathrm{samp}}}$ are extracted from the diverse measurements $\{(\mathcal{D}_{\mathrm{sens}}^{(n)}, \mathbf{Y}_{\mathrm{pilot}}^{(n)})\}_{n=1}^{N_{\mathrm{samp}}}$ using LMM-based semantic segmentation.

- **Scatterer index classification**: For each object $i \in \mathcal{I}_{\text{obj}}$, the conditional probability $f(S_i \,|\, \boldsymbol{p}_{\text{ue}})$ of the object i is identified using LMM-based binary classification.
- **Reflection point regression**: For each object $i \in \mathcal{I}_{\text{obj}}$, the conditional probability $f(\boldsymbol{p}_{\text{refl},i} \,|\, S_i, \boldsymbol{p}_{\text{ue}})$ of the reflection point $\boldsymbol{p}_{\text{refl}}$ is identified using LMM-based regression.

C. Visual Channel Parameter Extraction

To establish the intrinsic mapping f_{refl} between the UE position $\boldsymbol{p}_{\text{ue}}$ and the VCPs $(\boldsymbol{P}_{\text{refl}}, \mathcal{I}_{\text{sca}})$, we construct labeled samples $\mathcal{X}_{\text{samp}} = \{(\boldsymbol{p}_{\text{ue}}^{(n)}, (\boldsymbol{P}_{\text{refl}}^{(n)}, \mathcal{I}_{\text{sca}}^{(n)}))\}_{n=1}^{N_{\text{samp}}}$ from the sensing and pilot measurements $\{(\mathcal{D}_{\text{sens}}^{(n)}, \mathbf{Y}_{\text{pilot}}^{(n)})\}_{n=1}^{N_{\text{samp}}}$, which are collected for various UE positions. This process includes three steps: 1) UE position and GCP extraction; 2) initial VCP and abstract environmental information acquisition; and 3) VCP refinement.

1) UE Position and Geometric Channel Parameter Extraction: Initially, the UE position $\boldsymbol{p}_{\text{ue}}^{(n)}$ and rough estimates $(\hat{\boldsymbol{\theta}}^{(n)}, \hat{\boldsymbol{\phi}}^{(n)}, \hat{\boldsymbol{\tau}}^{(n)}, \hat{\boldsymbol{B}}^{(n)})$ of the GCPs are derived from the pilot measurements $\mathbf{Y}_{\text{pilot}}^{(n)}$. First, $\boldsymbol{p}_{\text{ue}}^{(n)}$ can be acquired by exploiting the global navigation satellite system (GNSS) data along with the advanced positioning techniques using time-based measurements and Doppler-based measurements [55], [56], [57], [58], [59], [60], [61], [62], [63], [64]. Also, $(\hat{\boldsymbol{\theta}}^{(n)}, \hat{\boldsymbol{\phi}}^{(n)}, \hat{\boldsymbol{\tau}}^{(n)}, \hat{\boldsymbol{B}}^{(n)})$ can be obtained via various GCP estimation techniques including subspace-based methods [65], compressed sensing (CS)-based methods [13], [14], [15], [16], [17], [18] and DL-based methods [19], [20], [21], [22]. However, due to practical issues such as noise and constraints on the numbers of antennas and RF chains, $(\hat{\boldsymbol{\theta}}^{(n)}, \hat{\boldsymbol{\phi}}^{(n)}, \hat{\boldsymbol{\tau}}^{(n)}, \hat{\boldsymbol{B}}^{(n)})$ may deviate from their true values and require further refinement.

2) Initial Visual Channel Parameter and Abstract Environmental Information Acquisition: After obtaining $(\hat{\boldsymbol{\theta}}^{(n)}, \hat{\boldsymbol{\phi}}^{(n)}, \hat{\boldsymbol{\tau}}^{(n)}, \hat{\boldsymbol{B}}^{(n)})$, they are converted to the initial VCPs $(\hat{\boldsymbol{P}}_{\text{refl}}^{(n)}, \hat{\mathcal{I}}_{\text{sca}}^{(n)})$ using the sensing measurements $\mathcal{D}_{\text{sens}}^{(n)}$. First, $\hat{\boldsymbol{P}}_{\text{refl}}^{(n)}$ is derived directly from $(\hat{\boldsymbol{\theta}}^{(n)}, \hat{\boldsymbol{\phi}}^{(n)})$ by leveraging the fact that $(\hat{\theta}_l^{(n)}, \hat{\phi}_l^{(n)})$ are the spherical coordinates of $\boldsymbol{p}_{\text{bs}} - \hat{\boldsymbol{p}}_{\text{refl},l}^{(n)}$. Second, the scatterer indices $\hat{\mathcal{I}}_{\text{sca}}^{(n)}$ are obtained by identifying the object index $\hat{i}_{\text{sca},l}^{(n)}$ associated with each $\hat{\boldsymbol{p}}_{\text{refl},l}^{(n)}$. This object identification is conducted through semantic segmentation, a CV task that assigns a specific class to each pixel in an image. In addition, this process facilitates the extraction of abstract environmental information $\mathcal{E}_{\text{abs}}^{(n)} = \{(\partial\mathcal{M}_i, c_i) : i \in \hat{\mathcal{I}}_{\text{sca}}^{(n)}\}$, which includes the boundary $\partial\mathcal{M}_i$ and the class c_i of each object i. The complete environmental information $\mathcal{E}_{\text{abs}} = \{(\partial\mathcal{M}_i, c_i) : i \in \mathcal{I}_{\text{obj}}\}$ can be determined as $\mathcal{E}_{\text{abs}} = \bigcup_{n=1}^{N_{\text{samp}}} \mathcal{E}_{\text{abs}}^{(n)}$.

To perform semantic segmentation on $\mathcal{D}_{\text{sens}}^{(n)}$, we use an LMM. Specifically, we employ a chain-of-thought (CoT) approach, a prompt engineering technique that guides the LMM to solve complex tasks by decomposing them into simpler subtasks and solving each subtask step-by-step [66]. In our case, the semantic segmentation task is divided into three subtasks: identifying the pixel regions in $\mathcal{D}_{\text{sens}}^{(n)}$ corresponding to $\hat{\boldsymbol{P}}_{\text{refl}}^{(n)}$, performing semantic segmentation on these regions, and extracting $\hat{\mathcal{I}}_{\text{sca}}^{(n)}$ and $\mathcal{E}_{\text{abs}}^{(n)}$. The CoT prompt $\mathcal{P}_{\text{seg}}^{(n)}$ for semantic segmentation at the nth data sample contains input description prompt, scenario specification prompt, and task instruction prompt.

- **Input description prompt**: The input description prompt provides detailed explanation on $\mathcal{D}_{\text{sens}}^{(n)}$. For example, "*The input $\mathcal{D}_{\text{sens}}^{(n)}$ consists of an RGB-D image captured by a camera mounted at the BS. Each image contains RGB channels for color information and a depth channel.*"
- **Scenario specification prompt**: The scenario specification prompt describes the task's objective. For example, "*The BS is located at $\boldsymbol{p}_{\text{bs}}$, and the environment contains $L-1$ scatterers identified as $\hat{\mathcal{I}}_{\text{sca}}^{(n)}$. Each scatterer i is characterized by a boundary $\partial\mathcal{M}_i$ and a class label c_i, collectively forming $\mathcal{E}_{\text{abs}}^{(n)} = \{(\partial\mathcal{M}_i, c_i) : i \in \hat{\mathcal{I}}_{\text{sca}}^{(n)}\}$.*"
- **Task instruction prompt**: The task instruction prompt provides detailed steps for semantic segmentation. For example, "*Perform semantic segmentation on $\mathcal{D}_{\text{sens}}^{(n)}$ to extract scatterer indices $\hat{\mathcal{I}}_{\text{sca}}^{(n)}$ and abstract environmental information $\mathcal{E}_{\text{abs}}^{(n)}$. First, use $(\hat{\boldsymbol{\theta}}^{(n)}, \hat{\boldsymbol{\phi}}^{(n)})$ to locate regions corresponding to $\hat{\boldsymbol{P}}_{\text{refl}}^{(n)}$. Next, classify these regions into object classes and delineate their boundaries. Finally, associate each $\hat{\boldsymbol{p}}_{\text{refl},l}^{(n)}$ with its object index $\hat{i}_{\text{sca},l}^{(n)}$.*"

3) Visual Channel Parameter Refinement: Once the initial VCPs $(\hat{\boldsymbol{P}}_{\text{refl}}^{(n)}, \hat{\mathcal{I}}_{\text{sca}}^{(n)})$ are obtained, they are refined to $(\boldsymbol{P}_{\text{refl}}^{(n)}, \mathcal{I}_{\text{sca}}^{(n)})$ using the pilot measurements $\mathbf{Y}_{\text{pilot}}^{(n)}$. Given that GCP estimation techniques provide a certain level of accuracy, we can readily assume that the scatterer indices $\hat{\mathcal{I}}_{\text{sca}}^{(n)}$ are correct, i.e., $\hat{\mathcal{I}}_{\text{sca}}^{(n)} = \mathcal{I}_{\text{sca}}^{(n)}$. Thus, the refinement process focuses on adjusting $\hat{\boldsymbol{P}}_{\text{refl}}^{(n)}$. Specifically, based on the boundary information $\partial\mathcal{M}_{i_{\text{sca},l}}^{(n)}$, a potential reflection region $\mathcal{R}_l^{(n)}$ is defined for each scatterer. Then $\hat{\boldsymbol{p}}_{\text{refl},l}^{(n)}$ is adjusted within $\mathcal{R}_l^{(n)}$ to minimize the discrepancy between the channel reconstructed from the refined points and the actual channel in $\mathbf{Y}_{\text{pilot}}^{(n)}$. For the reflection point adjustments, a simple grid search or advanced optimization techniques such as particle swarm optimization (PSO) can be used.

D. Scatterer Index Classification

In this stage, the LMM learns the mapping function between the UE position $\boldsymbol{p}_{\text{ue}}$ and the scatterer indices $\mathcal{I}_{\text{sca}}$. Specifically, for each object $i \in \mathcal{I}_{\text{obj}}$, the LMM models the conditional probability $f(S_i \,|\, \boldsymbol{p}_{\text{ue}})$ of the binary indicator S_i as

$$f(S_i \,|\, \boldsymbol{p}_{\text{ue}}) \approx \hat{f}_{\text{sca},i}(S_i \,|\, \boldsymbol{p}_{\text{ue}}; \mathcal{W}_{\text{sca},i}) \tag{49}$$

where $\hat{f}_{\text{sca},i}$ is the approximated probability distribution and $\mathcal{W}_{\text{sca},i}$ represent the network parameters. To do so, labeled samples $\mathcal{X}_{\text{sca},i} = \{(\boldsymbol{p}_{\text{ue}}^{(n)}, S_i^{(n)})\}_{n=1}^{N_{\text{samp}}}$ are extracted from $\mathcal{X}_{\text{samp}}$ as

$$S_i^{(n)} = \begin{cases} 1 & \text{if } i \in \mathcal{I}_{\text{sca}}^{(n)} \\ 0 & \text{otherwise.} \end{cases} \tag{50}$$

Note that determining $\hat{f}_{\text{sca},i}$ is a binary classification task. In the conventional DL models, $\hat{f}_{\text{sca},i}$ is learned by directly

training $\mathcal{W}_{\rm sca}$ using $\mathcal{X}_{{\rm sca},i}$. While this approach is effective to some extent, it might not work well when the amount of training data is limited. Moreover, the network parameters must be retrained for each new task, which significantly increases computational overhead and limits adaptability.

To ensure versatility across diverse tasks, the LMM leverages in-context learning (also known as few-shot learning) [5]. In this approach, rather than directly learning $\hat{f}_{{\rm sca},i}$, the LMM learns an algorithm $\mathcal{A}$ that selects the most suitable function from a family of pretrained binary classifiers $\mathcal{F}$ as

$$\hat{f}_{{\rm sca},i} = \mathcal{A}\big(\big\{(\boldsymbol{p}_{\rm ue}^{(n)}, S_i^{(n)})\big\}_{n=1}^{N_{\rm samp}}; \mathcal{P}_{{\rm sca},i}\big) \tag{51}$$

where $\mathcal{P}_{{\rm sca},i}$ is the scatterer index classification prompt. Using this approach, the LMM can effectively perform the scatterer index classification without retraining the entire network. Since the selected function $\hat{f}_{{\rm sca},i}$ is determined by $\mathcal{P}_{\rm sca}$, it is crucial to design $\mathcal{P}_{\rm sca}$ carefully so that it provides clear guidance for the function selection process. Specifically, $\mathcal{P}_{\rm sca}$ includes three elements: 1) input description prompt; 2) scenario specification prompt; and 3) task instruction prompt.

- **Input description prompt**: The input description prompt outlines the structure and contextual meaning of the input $\mathcal{X}_{{\rm sca},i}$. For example, "*The input $\mathcal{X}_{{\rm sca},i}$ consists of $N_{\rm samp}$ samples, where each sample is a pair of a three-dimensional UE position $\boldsymbol{p}_{\rm ue}^{(n)}$ and a corresponding binary label $S_i^{(n)}$. The label $S_i^{(n)}$ takes the value 1 if the object i is used as a scatterer, and 0 otherwise.*"
- **Scenario specification prompt**: The scenario specification prompt explains the abstract environmental information $\mathcal{E}_{\rm abs}$ (i.e., boundaries and classes of the objects). For instance, "*The BS is located at $\boldsymbol{p}_{\rm bs}$ and there are $N_{\rm obj}$ objects within the wireless environments, represented by $\mathcal{I}_{\rm obj}$. The boundaries and classes of these objects are $\mathcal{E}_{\rm abs} = \big\{(\partial\mathcal{M}_i, c_i) : i \in \mathcal{I}_{\rm obj}\big\}$.*"
- **Task instruction prompt**: The task instruction prompt provides detailed instructions on the scatterer index classification task. Note that an object is qualified as a scatterer if and only if there is a reflection point on the object that satisfies the law of reflection with respect to the BS position $\boldsymbol{p}_{\rm bs}$ and the UE position $\boldsymbol{p}_{\rm ue}$. Thus, these geometric constraints of signal reflection should be incorporated into the task instruction prompt. For example, "*Perform a binary classification to determine whether the object i is used as a scatterer for a given UE position. The decision criterion is that an object acts as a scatterer if there is a reflection point on the object i that satisfies the law of reflection with respect to the BS and UE.*"

Once $\hat{f}_{{\rm sca},i}$ is determined for all $i \in \mathcal{I}_{\rm obj}$, we can obtain $f(\mathcal{I}_{\rm sca} \,|\, \boldsymbol{p}_{\rm ue})$ using (47), thereby establishing the relationship between the UE position $\boldsymbol{p}_{\rm ue}$ and the scatterer indices $\mathcal{I}_{\rm sca}$. Furthermore, the scatterer index set $\mathcal{I}_{\rm sca}$ is obtained as

$$\mathcal{I}_{\rm sca} = \big\{i \in \mathcal{I}_{\rm obj} \,:\, S_i = 1\big\}. \tag{52}$$

E. Reflection Point Regression

After the scatterer index classification, the LMM identifies the reflection points $\boldsymbol{P}_{\rm refl}$ on the scatterers. While $\boldsymbol{p}_{{\rm refl},l}$ can be explicitly determined if the detailed environmental information $\mathcal{E}_{\rm det}$, including comprehensive 3D shapes of the scatterers, is available (see (25) and (26)), obtaining such detailed information is highly challenging in practice. As a remedy, for each object $i \in \mathcal{I}_{\rm obj}$, the LMM estimates $f(\boldsymbol{p}_{{\rm refl},i} \,|\, S_i, \boldsymbol{p}_{\rm ue})$:

$$f(\boldsymbol{p}_{{\rm refl},i} \,|\, S_i, \boldsymbol{p}_{\rm ue}) \approx \hat{f}_{{\rm refl},i}(\boldsymbol{p}_{{\rm refl},i} \,|\, S_i, \boldsymbol{p}_{\rm ue}; \mathcal{W}_{{\rm refl},i}) \tag{53}$$

where $\hat{f}_{{\rm refl},i}$ is the approximated probability distribution, parametrized by the network parameters $\mathcal{W}_{{\rm refl},i}$. To facilitate this task, $\mathcal{X}_{\rm samp}$ is partitioned into the labeled subset $\mathcal{X}_{{\rm refl},i} = \big\{(\boldsymbol{p}_{\rm ue}^{(n)}, \boldsymbol{p}_{\rm refl}^{(n)})\big\}_{n\in\mathcal{N}_i}$ for each object $i \in \mathcal{I}_{\rm obj}$ where $\mathcal{N}_i = \big\{n : S_i^{(n)} = 1\big\}$. One can see that determining $\hat{f}_{{\rm refl},i}$ is a multivariate regression task. Since we need to map two three-dimensional vectors, i.e., $\boldsymbol{p}_{\rm ue}$ and $\boldsymbol{p}_{\rm refl}$, it requires a much more precise and specialized network compared to the scatterer index classification. Therefore, the in-context learning approach utilized in the scatterer index classification may not achieve the necessary level of accuracy for this task.

To enhance the LMM's performance and capability for reflection point regression, we employ the low-rank adaptation (LoRA) method which fine-tunes the network by updating only a small set of task-specific parameters [67]. In this approach, a subset of parameters $\boldsymbol{W} \in \mathbb{R}^{X\times Y}$ (e.g., query, key, value weight matrices in attention layers) called a low-rank adaptor is chosen based on its relevance with the task. These parameters are augmented with a multiplication of two learnable matrices $\boldsymbol{A} \in \mathbb{R}^{X\times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r\times Y}$ where r ($\leq \min\{X, Y\}$) is the rank of these matrices (i.e., LoRA rank):

$$\tilde{\boldsymbol{W}} = \boldsymbol{W} + \boldsymbol{A}\boldsymbol{B} \tag{54}$$

where $\tilde{\boldsymbol{W}}$ is the updated parameters. Then $\boldsymbol{A}$ and $\boldsymbol{B}$ are trained using the task-specific samples $\mathcal{X}_{{\rm refl},i}$. In doing so, the LMM can be effectively tailored for the reflection regression task while significantly reducing computational and memory overhead. Similar to $\mathcal{P}_{\rm sca}$, the reflection point regression prompt $\mathcal{P}_{\rm refl}$ consists of three types of prompts for input description, scenario specification, and task instruction.

- **Input description prompt**: The input description prompt provides details about the input samples $\mathcal{X}_{{\rm refl},i}$. For example, *"The input $\mathcal{X}_{{\rm refl},i}$ consists of $|\mathcal{N}_i|$ pairs of three-dimensional UE positions $\boldsymbol{p}_{\rm ue}^{(n)}$ and corresponding reflection points $\boldsymbol{p}_{\rm refl}^{(n)}$ on the surface of object i."*
- **Scenario specification prompt**: The scenario specification prompt describes the abstract environmental information $\mathcal{E}_{\rm abs}$ required for the reflection point regression. For instance, *"The BS is located at $\boldsymbol{p}_{\rm bs}$, and there are $N_{\rm obj}$ objects within the wireless environment, represented by $\mathcal{I}_{\rm obj}$. The boundaries and classes of these objects are described by $\mathcal{E}_{\rm abs} = \big\{(\partial\mathcal{M}_i, c_i) : i \in \mathcal{I}_{\rm obj}\big\}$."*
- **Task instruction prompt**: The task instruction prompt provides instructions for the reflection point regression task. For example, *"Perform a regression task to determine the reflection point $\boldsymbol{p}_{\rm refl}$ on the surface of object i given the UE position $\boldsymbol{p}_{\rm ue}$. The reflection point must satisfy the law of reflection with respect to $\boldsymbol{p}_{\rm bs}$ and $\boldsymbol{p}_{\rm ue}$."*

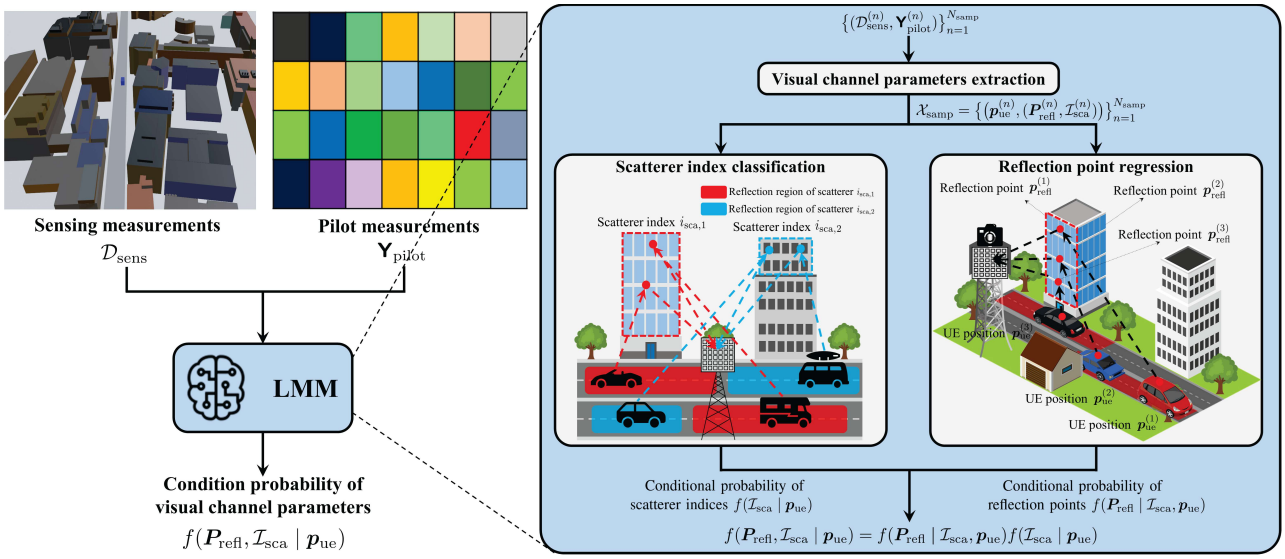


Fig. 3. Illustration of the reflection learning of LMM-CE.

F. Robustness to Environmental Factors

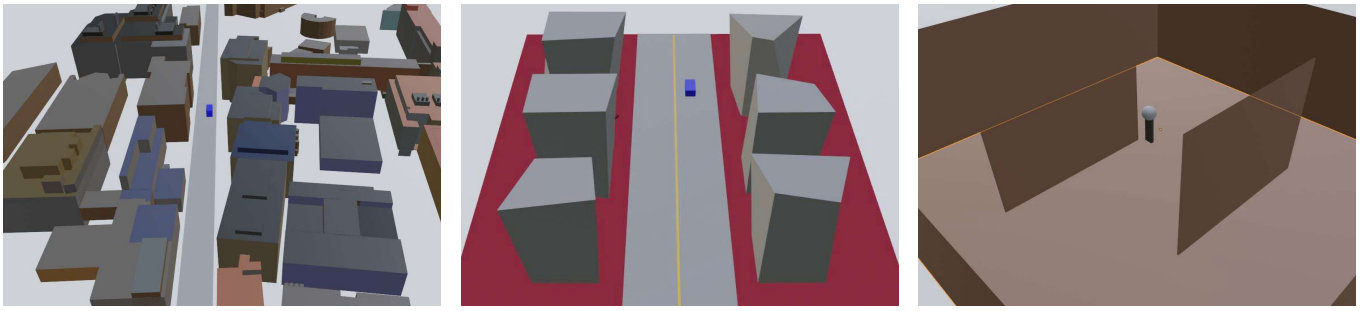
In LMM-CE, the multimodal data (i.e., pilot and sensing measurements) is used solely during the initial CKM construction to generate the VCP dataset. Once the CKM is established, channel information is inferred directly from the UE position using the learned CKM. This approach not only reduces the computational complexity associated with processing multimodal data but also improves robustness to environmental factors such as noise, occlusions, and misalignments. While these environmental factors may not affect the inference stage, they can influence the VCP dataset generation during the initial CKM construction stage. In this subsection, we discuss strategies for ensuring robustness against these challenges.

- **Noise**: Noise in the RGB-D sensor data can affect the performance of VCP extraction, especially in low-light or rainy conditions. However, since the RGB-D images are used only for the initial construction of the CKM, we can selectively choose RGB-D training data captured under conditions with lower noise to maximize the VCP extraction performance.
- **Occlusion**: While occlusion can impact the multipath propagation environment, it has minimal effect on the performance of VCP extraction for the following reasons. First, occlusion of dominant scatterers, such as buildings, from the BS's view is rare due to their high elevation. Second, in cases of persistent occlusion, a blocked scatterer no longer contributes to signal propagation, so its absence does not impact performance. Finally, for temporary occlusions, we can exclude affected images during CKM dataset construction, ensuring only reliable data is used for training.
- **Misalignment**: If the camera is not accurately directed toward the UE direction, it can lead to missing or incorrect data. To circumvent such events, we can leverage the synchronization signal block (SSB) beam index obtained at the initial access. During the initial access, the BS sequentially transmits a series of beam codewords carrying SSBs, and then the UE reports the index of the SSB beam corresponding to the highest received signal power to the BS. Since these SSB beams are directed toward distinct angles, the BS can obtain a rough estimate of the UE direction from the reported SSB beam index. Furthermore, considering that typical RGB-D camera can provide horizontal and vertical FoVs of 70° and 40°, respectively, while the angular spreads of AOA and ZOA in mmWave/THz bands are approximately 10–15° and 7–11°, respectively, each RGB-D camera is well-suited to capture the signal propagation [38].

V. Numerical Results

A. Simulation Setup

We consider a mmWave MISO system where a BS equipped with a $N_{\text{h}} \times N_{\text{v}} = 8 \times 4$ uniform planar array (UPA) antenna serves a single antenna UE. The BS is located at $\boldsymbol{p}_{\text{bs}} = [0 \ 0 \ 15]^{\text{T}}$ and the orientation of the BS antenna array is $(\alpha, \beta, \gamma) = (0°, 100°, 0°)$. Accordingly, the position vector of the (i,j)th BS antenna is given by $\boldsymbol{p}_{\text{bs},(i-1)N_{\text{h}}+j} = \boldsymbol{p}_{\text{bs}} + \boldsymbol{R}\left[\left(i - \frac{N_{\text{h}}-1}{2}\right) d_{\text{h}} \ \left(j - \frac{N_{\text{v}}-1}{2}\right) d_{\text{v}} \ 0\right]^{\text{T}}$ where $\boldsymbol{R} \triangleq \boldsymbol{R}_x(\alpha)\boldsymbol{R}_y(\beta)\boldsymbol{R}_z(\gamma)$ is the rotation matrix and $d_{\text{h}} = d_{\text{v}} = \frac{c}{2f_{\text{c}}}$ are the horizontal and vertical antenna spacings. For the analog-digital hybrid architecture, we set the number of RF chains to $N_{\text{RF}} = 8$. We employ the geometric frequency-selective channel model operating in the mmWave FR2 band with carrier frequency $f_{\text{c}} = 28\,\text{GHz}$ and system bandwidth $B = 400\,\text{MHz}$ [14], [38]. The pilot transmission spans $T = 14$ OFDM symbols across 64 resource blocks (RBs), resulting in a total of $S = 64 \times 12 = 768$ pilot subcarriers. We set the signal-to-noise ratio (SNR) to 20 dB. For the proposed LMM-CE, we employ LLaVA-NeXT-Interleave 7B, a pre-trained LMM model based on the Qwen-1.5-7B-Chat architecture with 7 billion parameters and specialized in handling multi-image, video, and 3D data [68]. We fine-tune the attention



(a) UMi scenario (MIT campus) (b) UC scenario (c) InH scenario

Fig. 4. Examples of RGB-D images for various wireless environments.

layer of this model using the LoRA technique with the LoRA rank of $r = 8$. We use the cross-entropy between the predicted and true token probability distributions as a loss function (see (43)). The fine-tuning process is performed on a system equipped with an Intel Xeon 6326 CPU with 16-core processors and an NVIDIA L40S GPU with 48 GB of VRAM.

As a performance metric, we evaluate the normalized mean square error (NMSE) of channel prediction defined as $\text{NMSE} = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \frac{\|\boldsymbol{H}(\boldsymbol{p}_{\text{ue}}^{(n)}) - \hat{\boldsymbol{H}}(\boldsymbol{p}_{\text{ue}}^{(n)})\|_{\text{F}}^2}{\|\boldsymbol{H}(\boldsymbol{p}_{\text{ue}}^{(n)})\|_{\text{F}}^2}$ where $\hat{\boldsymbol{H}}(\boldsymbol{p}_{\text{ue}}^{(n)})$ is the predicted channel matrix for a given UE position $\boldsymbol{p}_{\text{ue}}^{(n)}$. For comparison, we use six channel mapping schemes and a RF-based channel estimation scheme: 1) vision transformer (ViT)-based VCP regression scheme; Transformer-based VCP regression scheme [69]; 3) CNN-based VCP regression scheme; 4) linear VCP regression scheme; 5) LMM-based GCP regression scheme; 6) Transformer-based GCP regression scheme; and 7) CS-based channel estimation scheme [15].

B. Multimodal Dataset Generation

We employ the NVIDIA Sionna™ platform to generate the sensing and pilot measurements dataset. Initially, we construct three wireless environments using Blender: 1) urban micro (UMi) on the Massachusetts Institute of Technology (MIT) campus; 2) urban canyon (UC); and 3) indoor hotspot (InH) (see Fig. 4). Specifically, the UC scenario includes up to 8 buildings positioned along the sides of a four-lane road aligned parallel to the x-axis, with the UE moving the road. The width of each building and lane is 6 m and 2 m, respectively. The InH scenario features four side walls and two obstacles within the environment. Throughout the simulations, the UC scenario is used unless specified otherwise. Next, we perform ray tracing on the generated environments to determine the multipath components (i.e., angles, delays, and path gains) and corresponding VCPs. These parameters are then used to generate pilot measurements. Lastly, using the rendering function of Sionna, we capture images of the wireless environments using an RGB-D camera with a pixel resolution of 1024×768 and a FoV of 120° to generate the sensing measurements. Finally, we generate $N_{\text{data}} = 8{,}000$, $N_{\text{val}} = 1{,}000$, and $N_{\text{test}} = 1{,}000$ samples for training, validation, and test, respectively.

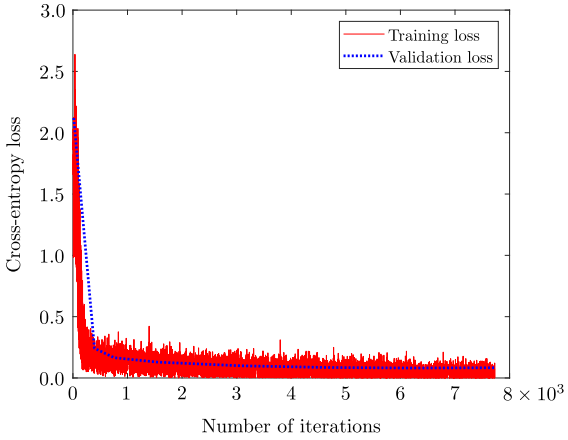


Fig. 5. Reflection point MSE as a function of the number of iterations.

C. Simulation Results

In Fig. 5, we plot the cross-entropy losses during training and validation in the fine-tuning process for reflection point regression. We observe that as the number of iterations increases, both training and validation losses decrease substantially. This implies that the LoRA-based fine-tuning process is working properly and the reflection point regression performance can be enhanced with only a small number of parameter updates.

In Table I, we compare the channel NMSE, inference time, and memory usage of LMM-CE with conventional DL models, including ViT, Transformer, and CNN. We see that LMM-CE achieves significantly higher channel prediction accuracy compared to the conventional DL models; however, the inference time and memory usage of LMM-CE are also the highest. Nevertheless, considering ongoing advancements in fast and low-power LMM models and GPU architectures, together with the model compression techniques (e.g., pruning and quantization) and edge learning techniques (e.g., federated learning (FL) and split learning), we expect that the inference time of LMM-CE can be further reduced down the road [70]. We also compare the LMMs with different numbers of parameters. We observe that the LMM with a larger number of parameters performs slightly better but also has longer inference time and higher memory usage. This demonstrates

TABLE I
COMPARISON OF CHANNEL NMSE, INFERENCE TIME, AND MEMORY USAGE WITH CONVENTIONAL DL-BASED METHODS

	Proposed LMM-CE (13B)	Proposed LMM-CE (7B)	ViT-based VCP regression scheme	Transformer-based VCP regression scheme	CNN-based VCP regression scheme
Channel NMSE	−14.03 dB	−13.88 dB	−10.58 dB	−9.39 dB	−0.58 dB
Inference time	306.0 ms	191.8 ms	62.5 ms	43.4 ms	0.5 ms
Memory usage	26.83 GB	14.16 GB	109.4 MB	33.8 MB	8.5 MB

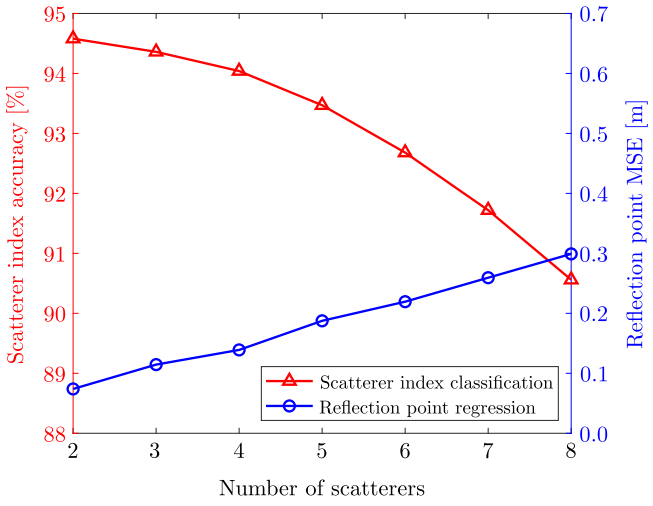


Fig. 6. Scatterer index accuracy and reflection point MSE as functions of the number of scatterers.

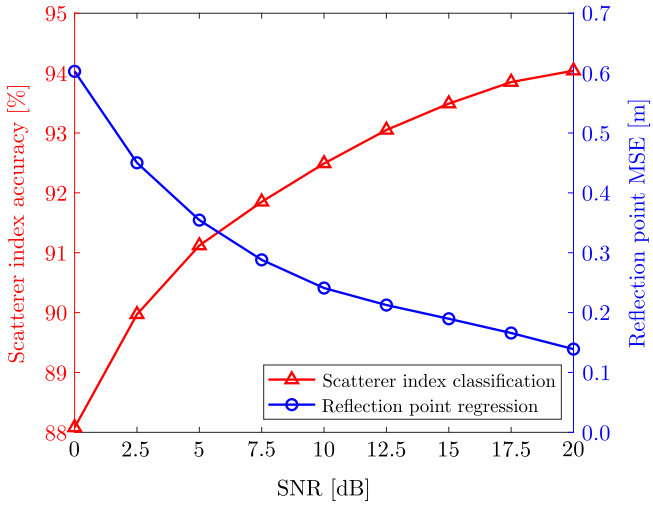


Fig. 7. Scatterer index accuracy and reflection point MSE as functions of the number of scatterers.

that LMM-CE can effectively be implemented with lighter LMMs, reducing inference time and resource usage.

In Fig. 6, we plot the reflection learning performances in terms of scatterer index accuracy and reflection point mean squared error (MSE) as functions of the number of scatterers. The scatterer index accuracy is defined as the ratio of the number of samples with correctly classified scatterer sets to the total number of samples. Similarly, the reflection point MSE is defined as the average MSE of the predicted reflection points. We observe that while the reflection learning performance shows a slight decrease as the number of scatterers increases, the degradation is marginal. Considering that the number of propagation paths in mmWave/THz band is typically lower than 6, this demonstrates that the proposed reflection learning can be an effective solution for identifying signal propagation environments in mmWave/THz systems.

In Fig. 7, we plot the reflection learning performance in terms of scatterer index accuracy and reflection point MSE as functions of SNR. Recall that the training dataset $\mathcal{X}_{\text{samp}} = \{\boldsymbol{P}_{\text{refl}}^{(n)}, \mathcal{I}_{\text{sca}}^{(n)}\}_{n=1}^{N_{\text{samp}}}$ for the reflection learning are extracted from the sensing and pilot measurements $\{\mathcal{D}_{\text{sens}}^{(n)}, \mathbf{Y}_{\text{pilot}}^{(n)}\}_{n=1}^{N_{\text{samp}}}$. Thus, the accuracies of the extracted VCPs are affected by the SNR. We observe that the proposed LMM-based reflection learning technique accurately determines both scatterer indices and reflection points. For example, when SNR = 20 dB, the proposed technique achieves 94% scatterer index classification accuracy and 0.13 m reflection point regression MSE. Even in the low SNR regime, it achieves 88% scatterer index accuracy and a reflection point MSE of less than a meter.

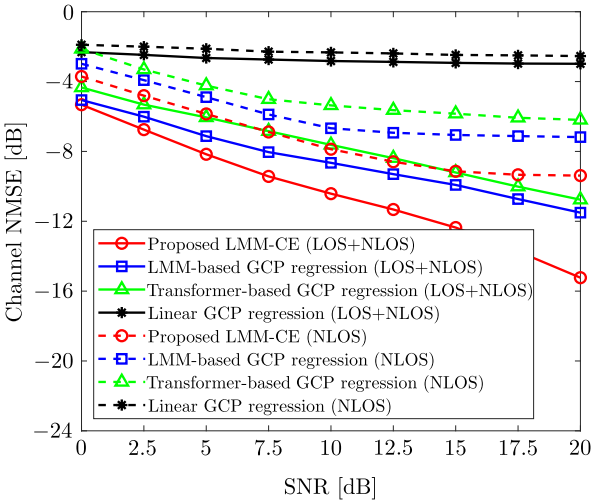


Fig. 8. Channel NMSE as a function of the SNR.

In Fig. 8, we investigate the channel NMSE as a function of SNR under two scenarios: combined LOS and NLOS propagation scenario, and NLOS-only propagation scenario. We observe that the proposed LMM-CE outperforms the benchmark schemes by a large margin. For example, in the combined LOS and NLOS scenario at SNR = 20 dB, LMM-CE achieves around 12.7 dB NMSE gain over the linear

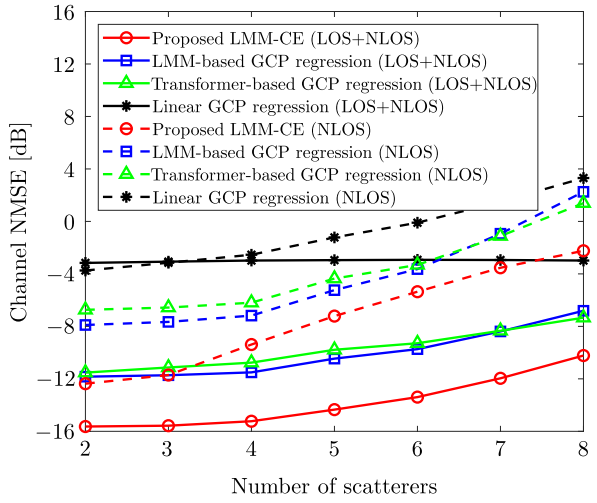


Fig. 9. Channel NMSE as a function of the number of scatterers.

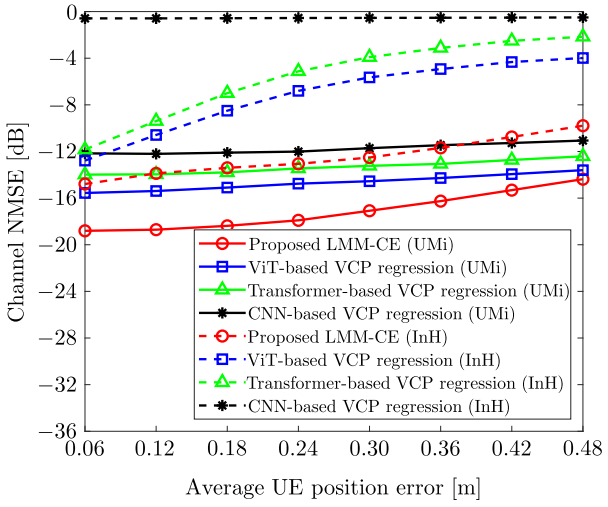


Fig. 10. Channel NMSE in UMi and InH scenarios.

GCP regression schemes. Even when compared to the LMM-based GCP regression scheme, the NMSE gain is more than 4 dB. This is because the piecewise continuity of GCPs, induced by the geometric distributions of scatterers and reflection points, is not adequately captured in the GCP regression schemes. In contrast, by sequentially determining the scatterer indices and reflection points, LMM-CE can effectively identify local regions where reflection points exhibit continuity, thereby enhancing the regression performance significantly.

In Fig. 9, we plot the channel NMSE as a function of the number of scatterers. We observe that the proposed LMM-CE significantly reduces the channel NMSE over the benchmark schemes. For instance, in the combined LOS and NLOS scenario with 6 scatterers, the channel NMSE of LMM-CE is 4 dB lower than that of the DL-based GCP regression scheme. One of the key advantages of LMM over conventional DL models is that it can comprehend task-specific instructions, even when these instructions are not mathematically defined, and leverage this understanding to enhance task performance. In our task, the LMM can interpret the underlying physical law (i.e., the law of reflection) that governs the relationship between the UE position and the VCPs based on the prompts. In contrast, the DL-based scheme has no such mechanism to incorporate such physical principles into the regression task.

In Fig. 10, we present the channel NMSE performance across various scenarios, including UMi and InH. We observe that the proposed LMM-CE scheme performs effectively in all scenarios tested. Furthermore, to demonstrate the superiority of our LMM-based approach, we compare the channel NMSE performance of LMM-CE with that of VCP regression schemes based on ViT, Transformer, and CNN. We observe that our LMM-based approach outperforms the conventional DL-based approaches. Furthermore, the channel NMSE gain of the LMM-based approach increases as the UE location error grows. This robustness of LMMs arises from the large network size and extensive training on diverse and ambiguous data, which enables them to effectively adapt to unseen noisy inputs.

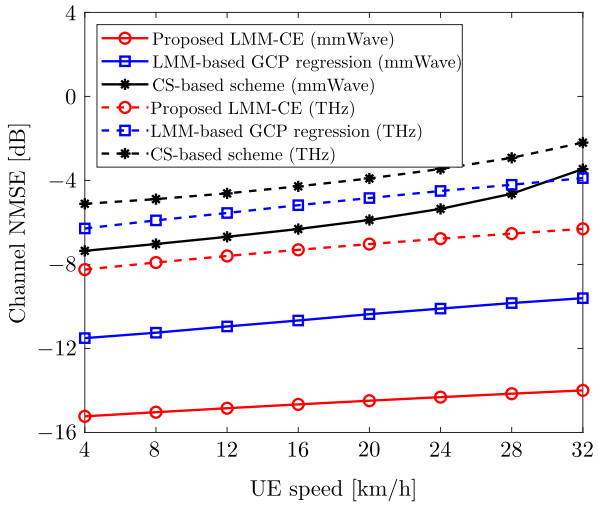


Fig. 11. Channel NMSE as a function of the speed of UE.

In Fig. 11, we plot the channel NMSE as a function of the speed of UE. We observe that the performance of the CS-based channel estimation scheme decreases sharply as the speed of UE increases, while those of channel mapping schemes show only marginal changes. This is because the channel mapping schemes learn the CKM beforehand and directly infer the channel from the UE positions during inference, without using pilot measurements. In contrast, the CS-based scheme relies heavily on pilot measurements for channel estimation. As a result, it experiences significant performance degradation in high-mobility scenarios where acquiring reliable and sufficient pilot measurements is challenging. Furthermore, to show the efficacy of LMM-CE in THz band, we present the channel NMSE performance in 0.1 THz band in Fig. 11. We observed that while LMM-CE remains effective in estimating the channel in the THz regime, the channel NMSE gain is slightly lower compared to that in the mmWave band. This is because, due to the extremely high path loss, the number of paths is significantly reduced in THz band. Consequently, the benefit

TABLE II
CHANNEL NMSE PERFORMANCE OF LMM-CE ACROSS DISTINCT UNTRAINED ENVIRONMENTS

Environment	Channel NMSE
Trained environment	−13.88 dB
Environments with different scatterer shapes	−6.54 dB
Environments with different scatterer positions	−6.68 dB
Environments with different BS positions	−8.17 dB

TABLE III
NMSE PERFORMANCE OF LMM-CE FOR DIFFERENT LANE NUMBERS

Number of lanes	Proposed LMM-CE	LMM-based GCP regression
2 lanes (LOS+NLOS)	−22.07 dB	−16.39 dB
4 lanes (LOS+NLOS)	−13.39 dB	−9.74 dB
6 lanes (LOS+NLOS)	−12.35 dB	−9.02 dB
2 lanes (NLOS)	−14.04 dB	−9.40 dB
4 lanes (NLOS)	−5.36 dB	−3.63 dB
6 lanes (NLOS)	−4.15 dB	−2.90 dB

of the proposed reflection learning technique is somewhat diminished.

To demonstrate the generalization capability of LMM-CE, in Table II, we show the channel NMSE performance in three distinct untrained environments. In the first environment, we employ scatterers with various shapes, including cylinders and pentagons. In the second environment, we randomly change the positions of the scatterers. In the third environment, we change the positions of the BS. We observe that LMM-CE can effectively predict the channel even in these unseen environments. Furthermore, in Table III, we present the channel NMSE performance of the proposed LMM-CE scheme across different numbers of lanes. In this figure, we fine-tune the LMM on samples collected from the six-lane environment and then test its performance on samples collected from environments with different numbers of lanes. We observe that while the LMM-based GCP regression scheme performs well only in the two-lane environment, LMM-CE demonstrates robust performance even in the six-lane environment. In the two-lane environment, the UE moves exclusively along the road so the reflection point changes linearly. However, in the six-lane environment, the UE moves both along and across the road, leading to nonlinear variations in the reflection points. By exploiting the physical principles that determine the VCPs, LMM-CE can effectively handle these nonlinear variations and accurately predict the VCPs.

VI. CONCLUSION

In this paper, we proposed an LMM-based environment-aware channel estimation scheme called LMM-CE that integrates sensing and pilot measurements to extract contextual channel information. To do so, we introduced the notion of VCPs, including the positions of UE, reflection points, and scatterers. Unlike the GCPs (i.e., angles, delays, and path gains) which provide only analytic descriptions of the signal propagation, the VCPs directly visualize the propagation environment. Using the VCPs, the physical interactions (e.g., reflections and blockages) of transmitted signals with the surrounding environments can be effectively identified. To uncover the intrinsic relationship between the UE position and the VCPs, we developed a reflection learning technique that extracts the VCPs and learns their probability distributions using the LMM. Finally, we established a CKM between the UE position and the channel. From the numerical evaluations, we demonstrated that LMM-CE can accurately predict the channel information across entire wireless environments. In this work, we focused on the channel estimation task; however, we believe that LMM-CE holds significant potential for broader applications such as beam tracking and handover.

REFERENCES

[1] Y. Chang et al., "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Mar. 2024.

[2] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds Mach.*, vol. 30, no. 4, pp. 681–694, Dec. 2020.

[3] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, Jun. 2020, pp. 1877–1901.

[4] C. Cui et al., "A survey on multimodal large language models for autonomous driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 958–979.

[5] H. Zhou et al., "Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 3, pp. 1955–2005, 3rd Quart., 2025.

[6] M. Xu et al., "When large language model agents meet 6G networks: Perception, grounding, and alignment," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 63–71, Dec. 2024.

[7] F. Jiang et al., "Large language model enhanced multi-agent systems for 6G communications," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 48–55, Dec. 2024.

[8] Y. He, J. Fang, F. R. Yu, and V. C. Leung, "Large language models (LLMs) inference offloading and resource allocation in cloud-edge computing: An active inference approach," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 11253–11264, Dec. 2024.

[9] W. Lee and J. Park, "LLM-empowered resource allocation in wireless communications systems," 2024, *arXiv:2408.02944*.

[10] L. Qian and J. Zhao, "User association and resource allocation in large language model based mobile edge computing system over 6G wireless communications," in *Proc. IEEE 99th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2024, pp. 1–7.

[11] E. Nayebi and B. D. Rao, "Semi-blind channel estimation for multiuser massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 540–553, Jan. 2018.

[12] C. Huang, J. Xu, W. Zhang, W. Xu, and D. W. K. Ng, "Semi-blind channel estimation for RIS-assisted MISO systems using expectation maximization," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 10173–10178, Sep. 2022.

[13] C. R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 164–174, Nov. 2010.

[14] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[15] K. Venugopal, A. Alkhateeb, N. G. Prelcic, and R. W. Heath Jr., "Channel estimation for hybrid architecture-based wideband millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1996–2009, Sep. 2017.

[16] K. Hassan, M. Masarra, M. Zwingelstein, and I. Dayoub, "Channel estimation techniques for millimeter-wave communication systems: Achievements and challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1336–1363, 2020.

[17] K. Dovelos, M. Matthaiou, H. Q. Ngo, and B. Bellalta, "Channel estimation and hybrid combining for wideband terahertz massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1604–1620, Jun. 2021.

[18] J. Wu, S. Kim, and B. Shim, "Parametric sparse channel estimation for RIS-assisted terahertz systems," *IEEE Trans. Commun.*, vol. 71, no. 9, pp. 5503–5518, Sep. 2023.

[19] P. Dong, H. Zhang, G. Y. Li, I. S. Gaspar, and N. NaderiAlizadeh, "Deep CNN-based channel estimation for mmWave massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 989–1000, Sep. 2019.

[20] W. Ma, C. Qi, Z. Zhang, and J. Cheng, "Sparse channel estimation and hybrid precoding using deep learning for millimeter wave massive MIMO," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2838–2849, May 2020.

[21] S. Kim, A. Lee, H. Ju, K. A. Ngo, J. Moon, and B. Shim, "Transformer-based channel parameter acquisition for terahertz ultra-massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 11, pp. 15127–15132, Nov. 2023.

[22] W. Yu et al., "An adaptive and robust deep learning framework for THz ultra-massive MIMO channel estimation," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 4, pp. 761–776, Jul. 2023.

[23] D. Liu et al., "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart. 2016.

[24] Y. Ahn, J. Kim, S. Kim, S. Kim, and B. Shim, "Sensing and computer vision-aided mobility management for 6G millimeter and terahertz communication systems," *IEEE Trans. Commun.*, vol. 72, no. 10, pp. 6044–6058, Oct. 2024.

[25] S. Kim, J. Wu, B. Shim, and M. Z. Win, "Cell-free massive non-terrestrial networks," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 1, pp. 201–217, Jan. 2025.

[26] Y. Zeng et al., "A tutorial on environment-aware communications via channel knowledge map for 6G," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 1478–1519, 3rd Quart., 2024.

[27] E. Dall'Anese, S.-J. Kim, and G. B. Giannakis, "Channel gain map tracking via distributed kriging," *IEEE Trans. Veh. Technol.*, vol. 60, no. 3, pp. 1205–1211, Mar. 2011.

[28] S.-J. Kim, E. Dall'Anese, and G. B. Giannakis, "Cooperative spectrum sensing for cognitive radios using Kriged Kalman filtering," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 1, pp. 24–36, Feb. 2011.

[29] S. Shrestha, X. Fu, and M. Hong, "Deep spectrum cartography: Completing radio map tensors using learned neural models," *IEEE Trans. Signal Process.*, vol. 70, pp. 1170–1184, 2022.

[30] C. Ding and I. W.-H. Ho, "Digital-twin-enabled city-model-aware deep learning for dynamic channel estimation in urban vehicular environments," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 3, pp. 1604–1612, Sep. 2022.

[31] L. U. Khan, Z. Han, W. Saad, E. Hossain, M. Guizani, and C. S. Hong, "Digital twin of wireless systems: Overview, taxonomy, challenges, and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2230–2254, 4th Quart., 2022.

[32] G. Liang et al., "Data augmentation for predictive digital twin channel: Learning multi-domain correlations by convolutional TimeGAN," *IEEE J. Sel. Topics Signal Process.*, vol. 18, no. 1, pp. 18–33, Jan. 2024.

[33] J. Zhang, S. Xu, Z. Zhang, C. Li, and L. Yang, "A denoising diffusion probabilistic model-based digital twinning of ISAC MIMO channel," *IEEE Internet Things J.*, vol. 12, no. 15, pp. 29121–29134, Aug. 2025.

[34] S. Kim, J. Moon, J. Wu, B. Shim, and M. Z. Win, "Vision-aided positioning and beam focusing for 6G terahertz communications," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 9, pp. 2503–2519, Sep. 2024.

[35] T. T. Nguyen, K. Elbassioni, N. C. Luong, D. Niyato, and D. I. Kim, "Access management in joint sensing and communication systems: Efficiency versus fairness," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 5128–5142, May 2022.

[36] S. Kim et al., "Role of sensing and computer vision in 6G wireless communications," *IEEE Wireless Commun.*, vol. 31, no. 5, pp. 264–271, Oct. 2024.

[37] A. Conti, S. Mazuelas, S. Bartoletti, W. C. Lindsey, and M. Z. Win, "Soft information for Localization-of-Things," *Proc. IEEE*, vol. 107, no. 11, pp. 2240–2264, Nov. 2019.

[38] *Study Channel Model for Frequencies From 0.5 to 100 GHz*, document TR 38.901, 3GPP, Sep. 2024.

[39] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2nd Quart., 2017.

[40] C. Han et al., "Terahertz wireless channels: A holistic survey on measurement, modeling, and analysis," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 3, pp. 1670–1707, 3rd Quart., 2022.

[41] H. Poddar, S. Ju, D. Shakya, and T. S. Rappaport, "A tutorial on NYUSIM: Sub-terahertz and millimeter-wave channel simulator for 5G, 6G, and beyond," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 824–857, 2nd Quart., 2024.

[42] C. Lin and G. Y. Li, "Adaptive beamforming with resource allocation for distance-aware multi-user indoor terahertz communications," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2985–2995, Aug. 2015.

[43] I. E. Gordon et al., "The HITRAN2020 molecular spectroscopic database," *J. Quantum Spectrosc. Radiat. Transf.*, vol. 277, Jan. 2021, Art. no. 107949.

[44] R. Piesiewicz, C. Jansen, D. Mittleman, T. Kleine-Ostmann, M. Koch, and T. Kurner, "Scattering analysis for the modeling of THz communication systems," *IEEE Trans. Antennas Propag.*, vol. 55, no. 11, pp. 3002–3009, Nov. 2007.

[45] A. Conti, M. Z. Win, and M. Chiani, "Slow adaptive M-QAM with diversity in fast fading and shadowing," *IEEE Trans. Commun.*, vol. 55, no. 5, pp. 895–905, May 2007.

[46] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999.

[47] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block coding for wireless communications: Performance results," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 3, pp. 451–460, Mar. 1999.

[48] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

[49] A. Conti, W. M. Gifford, M. Z. Win, and M. Chiani, "Optimized simple bounds for diversity systems," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2674–2685, Sep. 2009.

[50] J. H. Winters, "The diversity gain of transmit diversity in wireless systems with Rayleigh fading," *IEEE Trans. Veh. Technol.*, vol. 47, no. 1, pp. 119–123, Feb. 1998.

[51] W. M. Gifford, A. Conti, M. Chiani, and M. Z. Win, "On the SNR penalties of ideal and non-ideal subset diversity systems," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3708–3724, Jun. 2012.

[52] A. M. Sayeed and B. Aazhang, "Joint multipath-Doppler diversity in mobile wireless communications," *IEEE Trans. Commun.*, vol. 47, no. 1, pp. 123–132, Jan. 1999.

[53] Z. Chen, F. Li, Y. Quan, Y. Xu, and H. Ji, "Deep texture recognition via exploiting cross-layer statistical self-similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5231–5240.

[54] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[55] M. Z. Win et al., "Network localization and navigation via cooperation," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 56–62, May 2011.

[56] J. A. del Peral-Rosado, R. Raulefs, J. A. López-Salcedo, and G. Seco-Granados, "Survey of cellular mobile radio localization methods: From 1G to 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 1124–1148, 2nd Quart., 2018.

[57] A. Conti, G. Torsoli, C. A. Gómez-Vega, A. Vaccari, G. Mazzini, and M. Z. Win, "3GPP-compliant datasets for xG location-aware networks," *IEEE Open J. Veh. Technol.*, vol. 5, pp. 473–484, 2024.

[58] A. Conti, G. Torsoli, C. A. Gómez-Vega, A. Vaccari, and M. Z. Win, "xG-Loc: 3GPP-compliant datasets for xG location-aware networks," IEEE Dataport, Tech. Rep., Dec. 2023, doi: 10.21227/rper-vc03.

[59] H. K. Dureppagari, C. Saha, H. S. Dhillon, and R. M. Buehrer, "NTN-based 6G localization: Vision, role of LEOs, and open problems," *IEEE Wireless Commun.*, vol. 30, no. 6, pp. 44–51, Dec. 2023.

[60] F. Morselli, S. M. Razavi, M. Z. Win, and A. Conti, "Soft information-based localization for 5G networks and beyond," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9923–9938, Dec. 2023.

[61] Z. Lan, L. Liu, B. Fan, Y. Lv, Y. Ren, and Z. Cui, "Traj-LLM: A new exploration for empowering trajectory prediction with pre-trained large language models," *IEEE Trans. Intell. Vehicles*, vol. 10, no. 2, pp. 794–807, Feb. 2025.

[62] M. Z. Win, Y. Shen, and W. Dai, "A theoretical foundation of network localization and navigation," *Proc. IEEE*, vol. 106, no. 7, pp. 1136–1165, Jul. 2018.

[63] H. Chen, H. Sarieddeen, T. Ballal, H. Wymeersch, M. Alouini, and T. Y. Al-Naffouri, "A tutorial on terahertz-band localization for 6G communication systems," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 3, pp. 1780–1815, 3rd Quart., 2022.

[64] M. Z. Win, W. Dai, Y. Shen, G. Chrisikos, and H. V. Poor, "Network operation strategies for efficient localization and navigation," *Proc. IEEE*, vol. 106, no. 7, pp. 1224–1254, Jul. 2018.
[65] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.
[66] Z. Ling et al., "Deductive verification of chain-of-thought reasoning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 36407–36433.
[67] J. E. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent (ICLR)*, Jan. 2021, pp. 10173–10178.
[68] F. Li et al., "LLaVA-next-interleave: Tackling multi-image, video, and 3D in large multimodal models," 2024, *arXiv:2407.07895*.
[69] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[70] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, "Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 1, pp. 9–39, Jan. 2023.

Seungnyun Kim (Member, IEEE) received the B.S. (with honor) and Ph.D. degrees in electrical and computer engineering from Seoul National University (SNU), Seoul, South Korea, in 2016 and 2023, respectively.

He is currently a Postdoctoral Fellow with the Wireless Information and Network Sciences Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. His research interests include information theory, optimization methods, and machine learning with applications to real-world problems, including wireless communications, network localization and navigation, and non-terrestrial networks.

Dr. Kim was a recipient of the Sejong Science Fellowship from Korean Government in 2023, the Best Ph.D. Dissertation Award from SNU in 2023, the Qualcomm Innovation Fellowship Finalist in 2021, and the Samsung Humantech Paper Award Gold Prize in 2019.

Seokhyun Jeong (Student Member, IEEE) received the B.S. degree from the Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea, in 2022, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His main areas of research are in the artificial intelligence for the wireless communications.

Jiao Wu (Member, IEEE) received the B.S. degree in communication engineering from North China Electric Power University (NCEPU), Beijing, China, in 2015, the M.S. degree in electronics and communication engineering from University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018, and the Ph.D. degree in electrical and computer engineering from Seoul National University (SNU), Seoul, South Korea, in 2023.

She is currently a Postdoctoral Fellow with the Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. Her research interests include signal processing, optimization techniques, and machine learning with the applications to reconfigurable intelligent surfaces-assisted communications, extremely large-scale antenna systems, integrated sensing and communications, and non-terrestrial networks.

Byonghyo Shim (Fellow, IEEE) received the B.S. and M.S. degrees in Control and Instrumentation Engineering from Seoul National University, South Korea, in 1995 and 1997, respectively, and the M.S. degree in mathematics and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2004 and 2005, respectively. From 1997 to 2000, he was an Officer (First Lieutenant) and an Academic full-time Instructor in the Department of Electronics Engineering, Korean Air Force Academy. From 2005 to 2007, he was a Staff Engineer with Qualcomm Inc., San Diego, CA, USA. From 2007 to 2014, he was an Associate Professor with the School of Information and Communication, Korea University, Seoul. Since 2014, he has been with Seoul National University (SNU), where he is currently a Professor of the Department of Electrical and Computer Engineering and Vice Dean of Engineering College.

His research interests include wireless communications, machine learning, and statistical signal processing. Dr. Shim was a recipient of the M. E. Van Valkenburg Research Award from the ECE Department, University of Illinois, in 2005, the Haedong Young Engineer Award from IEIE in 2010, the Irwin Jacobs Award from Qualcomm and KICS in 2016, the Shinyang Research Award from the Engineering College of SNU in 2017, the Okawa Foundation Research Award in 2020, the IEEE Comsoc AP Outstanding Paper Award in 2021, and the JCN Best Paper Award in 2024. He was an Elected Member of the Signal Processing for Communications and Networking (SPCOM) Technical Committee of the IEEE Signal Processing Society. He has been served as an Associate Editor for IEEE Transactions on Wireless Communications (TWC), IEEE Transactions on Communications (TCOM), IEEE Transactions on Vehicular Technology (TVT), IEEE Transactions on Signal Processing (TSP), IEEE Wireless Communications Letters (WCL), and Journal of Communications and Networks (JCN) and a Guest Editor for IEEE Journal on Selected Areas in Communications (JSAC).

Moe Z. Win (Fellow, IEEE) is the Robert R. Taylor Professor at the Massachusetts Institute of Technology (MIT) and the founding director of the Wireless Information and Network Sciences Laboratory. Prior to joining MIT, he was with AT&T Research Laboratories and with the NASA Jet Propulsion Laboratory.

His research focuses on decision science and quantum information—encompassing fundamental theory, algorithm design, and network experimentation—for a broad range of real-world problems. His current research topics include AI-native xG networks; space-air-ground networks; network localization and navigation; networked control; ultra-wideband systems; and quantum sensing, communications, and control. He has served the IEEE Communications Society as an elected Member-at-Large on the Board of Governors, as elected Chair of the Radio Communications Committee, and as an IEEE Distinguished Lecturer. Over the last two decades, he has held various editorial positions for IEEE journals and organized numerous international conferences. He has served on the SIAM Diversity Advisory Committee.

Dr. Win is an elected Fellow of the AAAS, the EURASIP, the IEEE, and the IET. He was honored with two IEEE Technical Field Awards: the IEEE Kiyo Tomiyasu Award (2011) and the IEEE Eric E. Sumner Award (2006, jointly with R. A. Scholtz). His publications, co-authored with students and colleagues, have received several awards. Other recognitions include the MIT Frank E. Perkins Award (2024), the MIT Everett Moore Baker Award (2022), the IEEE Vehicular Technology Society James Evans Avant Garde Award (2022), the IEEE Communications Society Edwin H. Armstrong Achievement Award (2016), the Cristoforo Colombo International Prize for Communications (2013), the Copernicus Fellowship (2011) and the *Laurea Honoris Causa* (2008) from the Università degli Studi di Ferrara, and the U.S. Presidential Early Career Award for Scientists and Engineers (2004).