Large Multimodal Model-Based Environment-Aware Mobility Management

Seokhyun Jeong, *Student Member, IEEE*, Sangmok Shin, *Student Member, IEEE*, Seungnyun Kim, *Member, IEEE*, Jiao Wu, *Member, IEEE*, and Byonghyo Shim, *Fellow, IEEE*

***Abstract*—Recently, large language models (LLMs) have been successfully adopted in various fields, including wireless communications, robotics, and autonomous vehicles, owing to their outstanding adaptability and reasoning abilities. Despite their huge potential, the application of LLMs for mobility management is relatively scarce since it requires not only analyzing wireless measurements but also predicting dynamic user trajectories and making real-time handover decisions across densely deployed small base stations (SBSs). In this paper, we propose an environment-aware mobility management scheme based on large multimodal models (LMMs), which extend capabilities of LLMs to process multimodal sensing data. By leveraging LMMs, the proposed scheme extracts contextual information on the surrounding environments from RGB-D images to capture user equipment (UE) mobility patterns and identify signal reflections and blockages caused by static reflectors and dynamic obstacles. Using the extracted environmental information, the proposed scheme learns the intrinsic mapping from UE and SBS positions to channel capacity, referred to as channel capacity map (CCM), from which future channel capacities along UE trajectories are predicted. Based on the predicted channel capacities, we determine proactive handover decisions maximizing the cumulative channel capacities. Simulation results demonstrate that the proposed scheme achieves substantial channel capacity improvements over conventional deep learning (DL)-based approaches.**

***Index Terms*—Large multimodal model, mobility management, handover, ultra-dense network, sensing.**

I. Introduction

Large language models (LLMs), such as ChatGPT and Gemini, are attracting great attention these days for their powerful capabilities to solve unlimited complex tasks [1]. Indeed, owing to the vast network size and extensive pretraining on diverse datasets, LLMs can solve a bewildering variety of problems with minimal examples (i.e., few-shot learning), or even without any example at all (i.e., zero-shot learning). Benefited by advanced natural language processing (NLP) capabilities, LLMs can handle task instructions in the form of text prompts. Recently, functionalities of LLMs have been extended to simultaneously process multimodal data, including images, audio, and video [2]. This extended version, called large multimodal model (LMM), can extract rich contextual information on the surrounding environment and thus can handle complex real-world tasks. For this reason, LMMs can be used for a variety of applications, including smart factories, robotics, autonomous driving, and personal assistance [3].

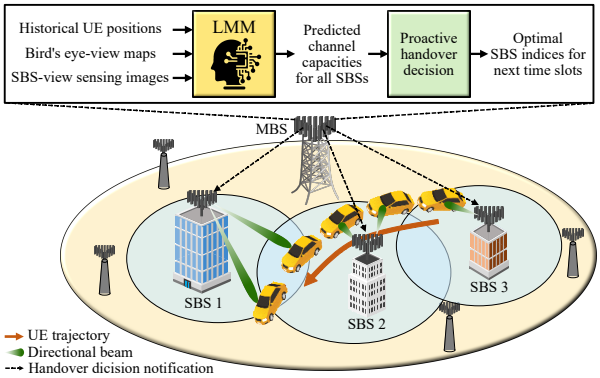


Fig. 1: Visualization of mobility management in UDN systems.

While there are recent research efforts integrating LMM into wireless systems, papers addressing LMM-based mobility management are relatively scarce [4], [5]. This is because mobility management involves far more diversified functionalities than those required by resource allocation, such as predicting the dynamic movements of user equipments (UEs) and solving intricate cell association problems. In fact, the primary goal of mobility management is to associate each UE with an appropriate base station (BS) to ensure a reliable connection (i.e., cell association) [6]. In particular, in ultra-dense networks (UDNs), wireless channels between a UE and a small base station (SBS) fluctuate rapidly as the UE moves through densely deployed SBSs (e.g., pico and femto cells) and thus a delay in the cell association process will affect the reliable connection considerably (see Fig. 1) [7], [8], [9].

To guarantee seamless connectivity in UDNs, a handover technique transferring the ongoing cell association to a nearby SBS is needed. In 4G Long-Term Evolution (LTE) and 5G New Radio (NR), handovers are triggered based on UE-

Received 23 September, 2025; revised 13 March, 2026 and 17 May, 2026; accepted 3 July, 2026. This work was supported in part by the National Research Foundation (NRF) of Korea under Grant RS-2022-NR070834 and 2022M3C1A3099336, the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT)(IITP-2026-2021-0-02048) and the National Research Foundation, Singapore and Infocomm Media Development Authority under its Communications and Connectivity Bridging Funding Initiative. An earlier version of this paper was presented in part at the IEEE Global Communications Conference (GLOBECOM), Taipei, Taiwan, December 2025. (*Corresponding author: Byonghyo Shim.*)

Seokhyun Jeong, Sangmok Shin, 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; smshin@islab.snu.ac.kr; bshim@snu.ac.kr).

Seungnyun Kim is with the Information Systems Technology and Design (ISTD) pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: snkim94@mit.edu).

Jiao Wu is with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900 Saudi Arabia (e-mail: jiao.wu@kaust.edu.sa).

reported reference signal received power (RSRP) measurements when the target SBS (T-SBS) provides a higher RSRP than the serving SBS (S-SBS) over four consecutive reports [10]. While this approach has been used for many years, it is not quite effective for 5G and upcoming 6G networks, since it induces a considerable handover delay [6][1]. To mitigate this issue, several approaches for reducing the handover delay by proactively performing handover have been suggested. In [12], [13], [14], proactive mobility management techniques that utilize machine learning (ML) and deep learning (DL) models have been studied. In [15], [16], [17], deep reinforcement learning (DRL)-based mobility management techniques that use policy optimization for the handover control have been proposed. The potential weakness of these techniques is that they rely exclusively on wireless measurements (e.g., RSRP), which hinders rapid adaptation to abrupt channel variations caused by the dynamic environment (e.g., human/vehicle movements and sudden environmental changes) [18]. Recently, integrated sensing and communications (ISAC)-based handover techniques that leverage the sensing data (e.g., images, radar signals, and LiDAR point clouds) to predict potential link blockages have gained attention [19], [20], [21], [22], [23]. While the ISAC-based approaches can incorporate environmental features to some extent, these techniques focus primarily on detecting line-of-sight (LoS) path and overlook the impact of multipath propagation (e.g., reflection and scattering), limiting their effectiveness in rich scattering environments [24].

An aim of this paper is to put forth an environment-aware mobility management framework empowered by the LMM technology. The proposed framework, referred to as *large multimodal model-based environment-aware mobility management* (LMM-EMM), preemptively associates the UE with a proper SBS by leveraging environmental context and UE mobility patterns extracted from multimodal data (e.g., RGB-D images and wireless measurements). Key ingredient in achieving this mission is LMM, a generative artificial intelligence (AI) model specialized in extracting correlated features across multiple modalities and performing high-level reasoning. In the proposed framework, LMM learns the UE movement pattern and the reflection geometry between the UE and SBSs by analyzing the correlations between the physical environments (e.g., buildings, road structures, and dynamic obstacles) obtained from bird's eye-view (BEV) maps, SBS-view sensing images, and also the wireless measurements. These features are then exploited to construct a fundamental mapping between the UE position and the channel capacity, henceforth dubbed as *channel capacity map* (CCM). Using CCM along with the predicted UE trajectory, the channel capacity can be estimated without requiring any real-time measurements, thereby facilitating fast and accurate environment-aware handovers. The main contributions of this paper are summarized as follows:

- We propose an LMM-based environment-aware mobility management framework that ensures fast and reliable connectivity. Unlike conventional schemes that rely primarily on wireless signal measurements, LMM-EMM exploits the multimodal reasoning capability of LMM to incorporate a richer contextual understanding of the surrounding environment, thereby supporting more predictive and robust mobility decisions.
- We develop an environment-aware channel capacity estimation technique that captures the reflection geometry of the propagation channel. To this end, we construct the CCM, an end-to-end mapping from the UE position to channel capacity. A main step of the CCM construction is analyzing the environmental context and inferring the underlying propagation characteristics, and we use LMM for this purpose. By leveraging CCM, we can accurately predict the future channel capacity for each SBS and make proactive handover decisions that reduce latency and enhance link reliability.
- From extensive simulations on realistic UDN environments, we show that the proposed LMM-EMM achieves a significant capacity gain over the conventional mobility management techniques. Specifically, LMM-EMM achieves about 45% improvement in channel capacity compared to the 5G NR mobility management scheme. Even when compared with the long short-term memory (LSTM)-based and DRL-based approaches, LMM-EMM yields 21% and 15% capacity gains, respectively.

The rest of this paper is organized as follows: In Section II, we explain the UDN system model and briefly introduce conventional mobility management techniques. In Section III, we present the overall architecture of LMM-EMM. In Section IV, we discuss practical issues. In Section V, we demonstrate the numerical results and conclude the paper in Section VI.

Notations: Bold upper and lower case symbols denote matrices and vectors, respectively. Sets are denoted by calligraphic font, for example, $\mathcal{X}$. Superscript $(\cdot)^{\mathrm{T}}$ denotes the transpose. $x^{(t)}$ and $x^{(t_1:t_2)}$ denote the value of x at time slot t and $\{x^{(t_1)}, x^{(t_1+1)}, \cdots, x^{(t_2)}\}$ where $t_1 < t_2$, respectively. $\mathbf{X}_1 \otimes \mathbf{X}_2$ denotes the Kronecker product of matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. $||\mathbf{x}||_2$ and $[\mathbf{x}]_n$ denote the Euclidean norm and the nth element of a vector $\mathbf{x}$, respectively. $\mathbb{E}[\cdot]$ denotes an expectation of a random variable. $|\mathcal{X}|$ denotes the number of elements in a set $\mathcal{X}$. $\bigcup_i \mathcal{X}_i$ denotes the union of the sets $\mathcal{X}_i$ over all i.

II. Ultra-dense Network System

In this section, we provide a brief overview of UDN systems and discuss conventional mobility management techniques.

A. Downlink UDN System Model

We consider a millimeter wave (mmWave) multiple-input single-output (MISO) UDN system consisting of M SBSs equipped with $N = N_x \times N_y$ uniform planar array (UPA) antennas and a UE equipped with a single antenna. The set of SBS indices is denoted as $\mathcal{M} = \{1, 2, \ldots, M\}$. The SBSs are

[1]Note that the S-SBS can initiate the handover only after receiving a sequence of RSRP reports from the UE. In 5G NR, the time interval between adjacent RSRP reports is 40-120 ms, and thus the total handover latency can reach up to 360 ms. Clearly, this time far exceeds the channel coherence time of 5G NR. For example, in 5G NR with carrier frequency of $f_c = 3.5$ GHz and UE speed of $v = 10$ m/s, the large-scale fading channel coherence time is $40\sqrt{\frac{9}{16\pi}}\frac{c}{vf_c} \approx 145.1$ ms [11].

connected to the macro base station (MBS) through backhaul links to share transmit data and control signals. We consider a global coordinate system (GCS) where the position vectors of the mth SBS and the UE at the tth time slot are $\mathbf{p}_{\mathrm{sbs},m}^{(t)} = [x_{\mathrm{sbs},m}\ y_{\mathrm{sbs},m}\ z_{\mathrm{sbs},m}]^{\mathrm{T}}$ and $\mathbf{p}_{\mathrm{ue}}^{(t)} = \left[x_{\mathrm{ue}}^{(t)}\ y_{\mathrm{ue}}^{(t)}\ z_{\mathrm{ue}}^{(t)}\right]^{\mathrm{T}}$, respectively. We also consider an orthogonal frequency division multiplexing (OFDM) system with S subcarriers, a carrier frequency of f_{c}, and a system bandwidth of B. The frequency of the sth subcarrier is $f_s = f_{\mathrm{c}} + (s - \frac{S}{2})\frac{B}{S}$ for $s = 1, \cdots, S$. Assuming the equal power allocation among the data streams, the MISO channel capacity $R^{(t)}(m)$ between the mth SBS and the UE at the time slot t is

$$R^{(t)}(m) = B\sum_{s=1}^{S}\log_2\left(1 + \frac{P_{\mathrm{t}}}{N\sigma_{\mathrm{n}}^2}\left\|\mathbf{h}_m^{(t)}[s]\right\|_2^2\right) \tag{1}$$

where $\mathbf{h}_m^{(t)}[s] \in \mathbb{C}^N$ is the downlink channel vector from the mth SBS to the UE at the sth subcarrier and the tth time slot, P_{t} is the SBS transmission power, and σ_{n}^2 is the noise power [25]. Since $R^{(t)}(m)$ depends on $\{\mathbf{h}_m^{(t)}[s]\}_{s=1}^S$, it varies dynamically as the UE moves [26]. Thus, if the link quality of S-SBS is lower than the predefined threshold due to the movement of the UE, the UE must search for a T-SBS and initiate a handover process.

In our work, we adopt a block-fading geometric multipath channel model where the channel remains constant within a time slot with duration τ_s. Each time slot is divided into the channel estimation (CE) period τ_{ce}, during which resource blocks (RBs) are dedicated to transmitting reference signals (e.g., CSI-RS), and the data transmission period τ_{dt}. Under this channel model, the downlink channel vector $\mathbf{h}_m^{(t)}[s]$ is expressed as the sum of an LoS path (denoted by $l = 0$) and $L-1$ non-line-of-sight (NLoS) paths:

$$\begin{aligned}\mathbf{h}_m^{(t)}[s] = &\sqrt{\beta_{m,0}^{(t)}}\alpha_{m,0}^{(t)}e^{-j2\pi f_s\tau_{m,0}^{(t)}}\mathbf{a}(\theta_{m,0}^{(t)},\phi_{m,0}^{(t)})\\ &+\sum_{l=1}^{L-1}\sqrt{\beta_{m,l}^{(t)}}\alpha_{m,l}^{(t)}e^{-j2\pi f_s\tau_{m,l}^{(t)}}\mathbf{a}(\theta_{m,l}^{(t)},\phi_{m,l}^{(t)})\end{aligned} \tag{2}$$

where $\beta_{m,l}^{(t)}$ is the large-scale fading coefficient (LSFC), $\alpha_{m,l}^{(t)} \sim \mathcal{CN}(0,1)$ is the small-scale fading coefficient (SSFC), $\tau_{m,l}^{(t)}$ is the time delay, and $\theta_{m,l}^{(t)} \in [0, 2\pi)$ and $\phi_{m,l}^{(t)} \in [0, \pi]$ are the azimuth angle of departure (AoD) and zenith angle of departure (ZoD) of the lth path between the mth SBS and the UE, respectively. Also, $\mathbf{a}\left(\theta_{m,l}^{(t)}, \phi_{m,l}^{(t)}\right) \in \mathbb{C}^N$ is the SBS array steering vector given by

$$\begin{aligned}\mathbf{a}\left(\theta_{m,l}^{(t)}, \phi_{m,l}^{(t)}\right) = &\left[1 \cdots e^{j\frac{2\pi d}{\lambda}(N_x-1)\cos\theta_{m,l}^{(t)}\sin\phi_{m,l}^{(t)}}\right]\\ &\otimes\left[1 \cdots e^{j\frac{2\pi d}{\lambda}(N_y-1)\sin\theta_{m,l}^{(t)}\sin\phi_{m,l}^{(t)}}\right]\end{aligned} \tag{3}$$

where d and λ are the antenna spacing and signal wavelength, respectively.

Specifically, $\beta_{m,l}[s]$ is expressed as

$$\beta_{m,l}[s] = \frac{\Gamma_{m,l}}{4\pi f_s\tau_{m,l}}e^{-\left(\frac{8\pi^2\sigma_{m,l}^2 f_s^2\cos^2(\psi_{m,l})}{c^2} + \frac{c\tau_{m,l}k_{\mathrm{abs}}(f_s)}{2}\right)} \tag{4}$$

where $k_{\mathrm{abs}}(f_s)$, c, $\Gamma_{m,l}$, $\sigma_{m,l}$, and $\psi_{m,l}$ are the molecular absorption coefficient, the speed of light, the Fresnel coefficient, the roughness coefficient, and the angle of incidence, respectively [27]. These parameters are determined by the UE and SBS positions $\mathbf{p}_{\mathrm{ue}}$ and $\mathbf{p}_{\mathrm{sbs},m}$, as well as the reflection surface $\mathcal{E}_{m,l} = \{\mathbf{x} \mid \mathbf{n}_{m,l}^{\mathrm{T}}\mathbf{x} = b_{m,l}\}$ ($\|\mathbf{n}_{m,l}\|_2 = 1$) [28]:

$$\tau_{m,l} = \frac{1}{c}\|\mathbf{d}_m - 2(\gamma_{m,l} - b_{m,l})\mathbf{n}_{m,l}\|_2 \tag{5}$$

$$\psi_{m,l} = \arccos\left(\frac{|\gamma_{m,l} + \delta_{m,l}|}{\sqrt{\|\mathbf{d}_m\|_2^2 + 4\gamma_{m,l}\delta_{m,l}}}\right) \tag{6}$$

$$\Gamma_{m,l} = \frac{\cos\psi_{m,l} - \sqrt{\epsilon - \sin^2\psi_{m,l}}}{\cos\psi_{m,l} - \sqrt{\epsilon + \sin^2\psi_{m,l}}} \tag{7}$$

where $\mathbf{d}_m = \mathbf{p}_{\mathrm{ue}} - \mathbf{p}_{\mathrm{sbs},m}$, $\gamma_{m,l} = \mathbf{n}_{m,l}^{\mathrm{T}}\mathbf{p}_{\mathrm{ue}} - b_{m,l}$, $\delta_{m,l} = \mathbf{n}_{m,l}^{\mathrm{T}}\mathbf{p}_{\mathrm{sbs},m} - b_{m,l}$, and ϵ is the dielectric permittivity of the reflector.

By concatenating $\mathbf{h}_m^{(t)}[s]$ for all subcarriers, we obtain the frequency domain channel matrix $\mathbf{H}_m^{(t)} \in \mathbb{C}^{N\times S}$:

$$\mathbf{H}_m^{(t)} = \left[\mathbf{h}_m^{(t)}[1]\,\mathbf{h}_m^{(t)}[2]\cdots\mathbf{h}_m^{(t)}[S]\right] \tag{8}$$

$$= \sum_{l=0}^{L-1}\sqrt{\beta_{m,l}^{(t)}}\alpha_{m,l}^{(t)}\mathbf{a}\left(\theta_{m,l}^{(t)},\phi_{m,l}^{(t)}\right)\mathbf{b}^{\mathrm{T}}\left(\tau_{m,l}^{(t)}\right) \tag{9}$$

where $\mathbf{b}(\tau_{m,l}^{(t)}) \in \mathbb{C}^S$ is the phase shift vector of the OFDM subcarriers given by

$$\mathbf{b}(\tau_{m,l}^{(t)}) = \left[e^{-j2\pi f_1\tau_{m,l}^{(t)}}\ e^{-j2\pi f_2\tau_{m,l}^{(t)}}\ \cdots\ e^{-j2\pi f_S\tau_{m,l}^{(t)}}\right]^{\mathrm{T}}. \tag{10}$$

For brevity, we define the collection of the geometric channel parameters $\mathcal{G}_{m,l}^{(t)}$ of the lth path between the mth SBS and UE at the tth time slot as

$$\mathcal{G}_{m,l}^{(t)} = \left\{\beta_{m,l}^{(t)}, \alpha_{m,l}^{(t)}, \tau_{m,l}^{(t)}, \vartheta_{m,l}^{(t)}, \varphi_{m,l}^{(t)}, \theta_{m,l}^{(t)}, \phi_{m,l}^{(t)}\right\}. \tag{11}$$

Then, we can express $\mathbf{H}_m^{(t)}$ as a function of $\{\mathcal{G}_{m,l}^{(t)}\}_{l=1}^L$ as

$$\mathbf{H}_m^{(t)} = \mathbf{P}\left(\mathcal{G}_{m,0}^{(t)}\right) + \sum_{l=1}^{L-1}\mathbf{P}\left(\mathcal{G}_{m,l}^{(t)}\right) \tag{12}$$

where $\mathbf{P}(\mathcal{G}_{m,l}^{(t)}) \in \mathbb{C}^{N\times S}$ denotes the lth multipath component of $\mathbf{H}_m^{(t)}$ corresponding to $\mathcal{G}_{m,l}^{(t)}$:

$$\mathbf{P}(\mathcal{G}_{m,l}^{(t)}) = \sqrt{\beta_{m,l}^{(t)}}\alpha_{m,l}^{(t)}\mathbf{a}\left(\theta_{m,l}^{(t)},\phi_{m,l}^{(t)}\right)\mathbf{b}^{\mathrm{T}}\left(\tau_{m,l}^{(t)}\right). \tag{13}$$

Note that geometric channel parameters vary continuously when reflectors remain unchanged. However, changes in reflectors can cause abrupt transitions, which lead to discontinuities in channel capacity [29].

Remark 1. *As shown in Fig. 2, the channel capacity $R^{(t)}(m)$ for each SBS is piecewise continuous over time. The discontinuity arises from the discontinuity of reflection points on different reflectors.*

B. Channel Capacity Map

In this subsection, we introduce the concept of the CCM, which characterizes the achievable channel capacity as a function of the UE position, the SBS position, and environmental factors (e.g., reflectors). For notational simplicity, the time index t is omitted without loss of generality.

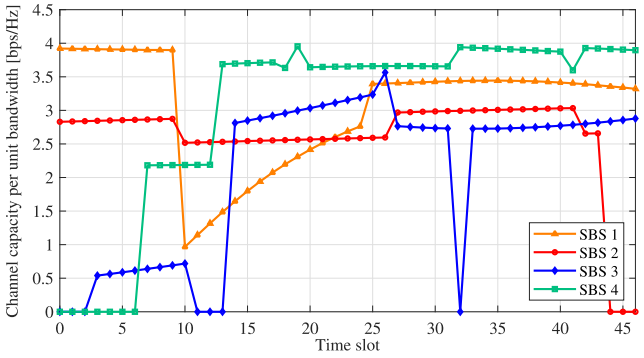


Fig. 2: Piecewise continuity of channel capacity with UE movement.

Lemma 1. *When the number of transmit antennas N is sufficiently large, the channel capacity $R_{\text{ideal}}(m)$ can be expressed as a function of LSFCs $\{\beta_{m,l}[s]\}_{l=0}^{L-1}$. That is,*

$$R_{\text{ideal}}(m) = B\sum_{s=1}^{S}\log_2\left(1+\frac{P_{\text{t}}}{\sigma_{\text{n}}^2}\left(\beta_{m,0}[s]+\sum_{l=1}^{L-1}\beta_{m,l}[s]\right)\right). \tag{14}$$

Proof. See Appendix A. □

In practice, moving obstacles (e.g., vehicles and pedestrians) can intermittently block the LoS link between UE and SBS. Such dynamic blockages significantly impact the achievable channel capacity due to the pronounced power disparity between the LoS and NLoS components [27].

Remark 2. *The achievable channel capacity is a function of static channel capacity and LoS indicator as*

$$R(m) = c_{\text{los}}(m)R_{\text{ideal}}(m) + (1-c_{\text{los}}(m))R_{\text{nlos}}(m) \tag{15}$$

where $R_{\text{nlos}}(m)$ is the capacity of NLoS channel given by

$$R_{\text{nlos}}(m) = B\sum_{s=1}^{S}\log_2\left(1+\frac{P_{\text{t}}}{\sigma_{\text{n}}^2}\sum_{l=1}^{L-1}\beta_{m,l}[s]\right). \tag{16}$$

Also, $c_{\text{los}}(m)$ is the LoS indicator defined as

$$c_{\text{los}}(m) = \begin{cases} 1 & \text{if the LoS for SBS } m \text{ exists} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

From Equation (14) and (16), one can see that $R_{\text{ideal}}(m)$ and $R_{\text{nlos}}(m)$ are functions of LSFCs $\{\beta_{m,l}[s]\}_{l=0}^{L-1}$. Moreover, as shown in Equation (4)-(7), the LSFC $\beta_{m,l}[s]$ is determined by the UE position $\mathbf{p}_{\text{ue}}$, SBS position $\mathbf{p}_{\text{sbs},m}$, and the reflection surface $\mathcal{E}_{m,l} = \{\mathbf{x} \mid \mathbf{n}_{m,l}^{\text{T}}\mathbf{x} = b_{m,l}\}$. Therefore, $R_{\text{ideal}}(m)$ and $R_{\text{nlos}}(m)$ can be directly obtained from $\mathbf{p}_{\text{ue}}$, $\mathbf{p}_{\text{sbs},m}$, and $\mathcal{E} = \cup_{m,l}\mathcal{E}_{m,l}$.

Remark 3. *The static channel capacities $R_{\text{ideal}}(m)$ and $R_{\text{nlos}}(m)$ are fully characterized by three components: the UE position $\mathbf{p}_{\text{ue}}$, the SBS position $\mathbf{p}_{\text{sbs},m}$, and the reflector information $\mathcal{E}$, through CCM f_{ccm}:*

$$\left(R_{\text{ideal}}(m), R_{\text{nlos}}(m)\right) = f_{\text{ccm}}(\mathbf{p}_{\text{ue}}, \mathbf{p}_{\text{sbs},m}, \mathcal{E}). \tag{18}$$

Remarks 2 and 3 show that predicting $R(m)$ requires the UE position $\mathbf{p}_{\text{ue}}$, the SBS positions $\{\mathbf{p}_{\text{sbs},m} \mid m \in \mathcal{M}\}$, the reflector information $\mathcal{E}$, and the LoS indicators $\{c_{\text{los}}(m) \mid m \in \mathcal{M}\}$.

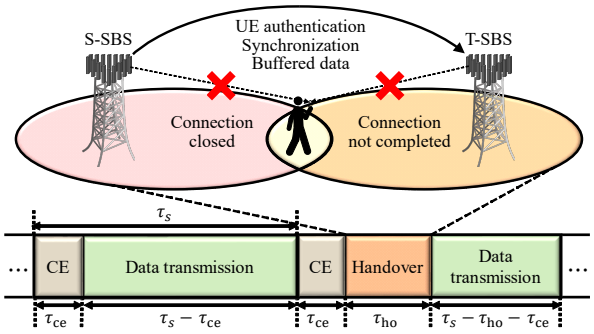


Fig. 3: Illustration of service interruption during the handover.

C. Handover Process

The handover process in 5G NR consists of three main phases: preparation, execution, and completion [30]. The preparation phase begins when the RSRP of T-SBS exceeds that of S-SBS by a configured offset (i.e., event A3 [30]). At this point, the UE initiates the handover process by sending the radio resource control (RRC) measurement report to S-SBS. If S-SBS continues to receive these reports throughout the designated period (i.e., time-to-trigger), it sends the handover request to T-SBS, followed by a handover acknowledgment from T-SBS to S-SBS. Next, in the execution phase, S-SBS transfers the UE identification and security information (e.g., encryption algorithms) to T-SBS, along with any downlink data that has not yet been delivered to the UE. Then, the UE establishes a new connection to T-SBS via the random access. Lastly, in the completion phase, the core network finalizes the handover process by redirecting the UE data path to T-SBS.

Note that during the handover process, data transmission is temporarily interrupted since the UE cannot receive data from T-SBS until UE authentication and synchronization are completed (see Fig. 3). Thus, the effective channel capacity $R_{\text{eff}}^{(t)}(m^{(t-1:t)})$ at the tth time slot is expressed as a weighted sum of channel capacities during and after the handover process (i.e., $R^{(t)}(m^{(t)})(1-\mathbb{1}_{\text{ho}}(m^{(t-1:t)})$ and $R^{(t)}(m^{(t)})$):

$$\begin{aligned} R_{\text{eff}}^{(t)}(m^{(t-1:t)}) &= \frac{1}{\tau_s}\big(\tau_{\text{ho}}R^{(t)}(m^{(t)})(1-\mathbb{1}_{\text{ho}}(m^{(t-1:t)})) \\ &\quad +(\tau_s-\tau_{\text{ho}}-\tau_{\text{ce}})R^{(t)}(m^{(t)})\big) \qquad (19) \\ &= \frac{\tau_s-\tau_{\text{ce}}}{\tau_s}(1-\mu\mathbb{1}_{\text{ho}}(m^{(t-1:t)}))R^{(t)}(m^{(t)}) \qquad (20) \end{aligned}$$

where τ_{ho} is the handover process time, and μ is the handover cost coefficient defined as

$$\mu = \frac{\tau_{\text{ho}}}{\tau_s-\tau_{\text{ce}}}. \tag{21}$$

Also, $\mathbb{1}_{\text{ho}}(m^{(t-1:t)}) \in \{0,1\}$ is the binary handover indicator at the tth time slot defined as

$$\mathbb{1}_{\text{ho}}(m^{(t-1:t)}) = \begin{cases} 1 & \text{if } m^{(t-1)} \neq m^{(t)} \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Note that in UDN environments, UEs may undergo frequent handovers due to the dense deployment of SBSs, which consumes a considerable portion of the total transmission time and thus degrades throughput significantly.

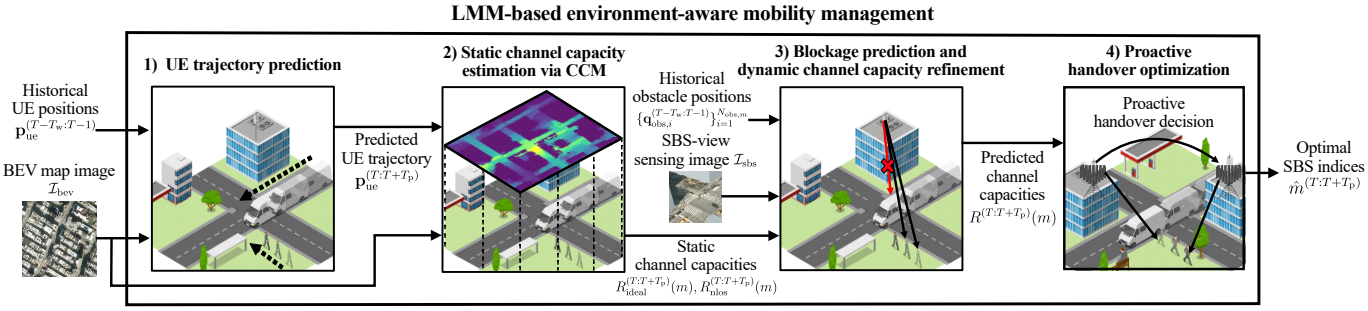


Fig. 4: Overall procedure of the proposed LMM-EMM.

D. Conventional Mobility Management Techniques

A major drawback of the 5G NR handover is that the signal quality deterioration and decision latency are considerable since handover is triggered only after the RSRP drop has been reported. This issue is even more pronounced in UDN scenarios where handover occurs frequently due to the reduced cell coverage and increased number of SBSs. To overcome the limitations of the reactive handover mechanism, proactive handover techniques that trigger handover before the signal quality degradation have been proposed [14]. Essence of this approach is to forecast the future RSRP for all candidate SBSs using historical RSRP measurements and then pick the SBS that maximizes the throughput. For the RSRP prediction, recurrent neural network (RNN) architectures such as LSTM and gated recurrent unit (GRU) have been widely employed [13], [31].

While conventional proactive handover techniques are effective to some extent, they exclusively rely on radio-frequency (RF) signals so that they fall short in capturing abrupt changes in the channel (e.g., blockages caused by dynamic obstacles). Note that RF signals capture gradual variations in signal quality, so it is very difficult to distinguish whether signal quality variations are transient (e.g., sudden blockages) or persistent (e.g., continuous path loss changes due to UE movement) by merely checking the received signals [18]. To address this issue, approaches that leverage sensing data (e.g., RGB images, radar signals, and LiDAR point clouds) obtained from various sensors (e.g., camera, radar, and LiDAR) have been proposed recently [19], [20], [21], [22], [23]. These techniques focus primarily on identifying the LoS component of the channel so that they suffer a degradation of the handover quality in NLoS scenarios.

E. Proactive Handover Decision Problem Formulation

Effective mobility management must ensure seamless execution of handovers while avoiding unnecessary ones (e.g., ping-pong effects) caused by channel variations [32]. To this end, we formulate a long-term proactive handover problem that not only responds to instantaneous degradations but also seeks to maximize channel capacity over extended periods. Essence of the proactive handover problem $\mathscr{P}$ is to determine the SBS indices $m^{(T:T+T_\mathrm{p})}$ for the time slots T to $T+T_\mathrm{p}$, with an objective to maximize the cumulative channel capacity:

$$\mathscr{P}: \max_{m^{(T:T+T_\mathrm{p})}} \sum_{t=T}^{T+T_\mathrm{p}} R_\mathrm{eff}^{(t)}(m^{(t-1:t)}) \tag{23a}$$

$$\text{s.t. } R_\mathrm{eff}^{(t)}\left(m^{(t-1:t)}\right) \geq R_\mathrm{min}^{(t)} \quad \forall t = T, \cdots, T+T_\mathrm{p} \tag{23b}$$

where $R_\mathrm{min}^{(t)}$ is the minimum capacity requirement to ensure the quality of service (QoS) of the UE at the tth time slot. The salient feature of $\mathscr{P}$ is that the objective function accounts for the cumulative channel capacity over $T_\mathrm{p}+1$ future time slots, which is clearly distinct from conventional approaches maximizing the instantaneous capacity (i.e., $R^{(T)}(m^{(T)})$) exclusively. Using the cumulative channel capacity as a performance metric, we can reduce redundant handovers caused by instantaneous channel fluctuations.

To determine the optimal SBS indices for future time slots, we need to know the future channel capacities (i.e., $\{\{R^{(T+t)}(m^{(T+t)})\}_{t=0}^{T_\mathrm{p}} \mid m \in \mathcal{M}\}$). Obviously, due to the causality issue, future channel capacities cannot be directly observed and need to be predicted instead. Unfortunately, conventional approaches predicting future channel capacities based on temporal correlation (e.g., Kalman filter and RNNs) are not so effective, in particular for urban environments, due to the discontinuity of channel capacity (see Remark 1).

III. Large Multimodal Model-Based Environment-Aware Mobility Management

The primary goal of LMM-EMM is to proactively associate the UE with proper SBSs that maximize channel capacity. To this end, we leverage the multimodal reasoning capabilities of LMM. Specifically, we first extract the UE mobility patterns and scattering geometry from the multimodal sensing data. Unlike conventional DL-based approaches that only capture spatial correlations of the channel without any contextual information, LMM-EMM exploits the road geometry and surrounding environment to infer the behavioral intention of the UE (e.g., turning or accelerating) and perceive signal reflection patterns. Next, using extracted features, we construct CCM, a spatial mapping between the UE position and the channel capacity, used for the estimation of the future channel capacity along the predicted UE trajectory. Furthermore, by analyzing RGB-D images with LMM, we predict potential path blockages and then refine the channel capacity, thereby facilitating optimal SBS selection even under high mobilities.

A key feature of LMM-EMM is that LMM extracts environmental knowledge (e.g., the spatial layout of roads and buildings) from multimodal data and reuses this shared understanding across trajectory prediction, channel capacity estimation, and blockage prediction. Unlike conventional DL approaches that learn independent input-output mappings for each task without sharing physical knowledge, LMM-EMM leverages a common environmental interpretation to produce

mutually coherent predictions, thereby achieving more reliable mobility management performance. The procedure of LMM-EMM consists of four steps (see Fig. 4).

1) **UE trajectory prediction**: We predict the future UE trajectory $\mathbf{p}_{\text{ue}}^{(T:T+T_\text{p})}$ using historical UE positions $\mathbf{p}_{\text{ue}}^{(T-T_\text{w}:T-1)}$ and the BEV map $\mathcal{I}_{\text{bev}}$.
2) **Static channel capacity estimation via CCM**: By utilizing CCM, we estimate the static channel capacities $\{R_{\text{ideal}}^{(T:T+T_\text{p})}(m), R_{\text{nlos}}^{(T:T+T_\text{p})}(m) \mid m \in \mathcal{M}\}$ from the predicted UE trajectories.
3) **Blockage prediction and dynamic channel capacity refinement**: We predict path blockages from dynamic obstacles by tracking obstacle states from SBS-view sensing images, and then derive the future achievable channel capacities $\{R^{(T:T+T_\text{p})}(m) \mid m \in \mathcal{M}\}$.
4) **Proactive handover optimization**: We determine future SBS indices $m^{(T:T+T_\text{p})}$ by solving $\mathscr{P}$ in (23) using dynamic programming (DP). $\{R^{(T:T+T_\text{p})}(m) \mid m \in \mathcal{M}\}$ obtained in step 3 is used in this process.

A. LMM-Based UE Trajectory Prediction

In the UE trajectory prediction step, we estimate the future UE trajectory using historical UE positions and the BEV map:[2]

$$\mathbf{p}_{\text{ue}}^{(T:T+T_\text{p})} = f_{\text{traj}}\big(\mathbf{p}_{\text{ue}}^{(T-T_\text{w}:T-1)}, \mathcal{I}_{\text{bev}}\big) \tag{24}$$

where f_{traj} is the UE trajectory prediction function and T_w is the observation time window. Also, $\mathcal{I}_{\text{bev}}$ is the BEV map image representing the static road topology, lane structures, and surrounding environments, which can be captured at the rooftop of a building or a satellite [35]. It is important to note that the UE movement is constrained by physical environments, such as road and building areas, which are denoted by $\mathcal{A}_\text{R}$ and $\mathcal{A}_\text{B}$, respectively [36]. For example, vehicles move along lanes with their trajectories confined within road boundaries (i.e., $\text{Prob}(\mathbf{p}_{\text{ue}}^{(t)} = \mathbf{p} \mid \mathbf{p} \notin \mathcal{A}_\text{R}) = 0$), and pedestrians cannot traverse obstacles like buildings and walls (i.e., $\text{Prob}(\mathbf{p}_{\text{ue}}^{(t)} = \mathbf{p} \mid \mathbf{p} \in \mathcal{A}_\text{B}) = 0$) (see Fig. 5). In LMM-EMM, we exploit BEV images to incorporate $\mathcal{A}_\text{R}$ and $\mathcal{A}_\text{B}$ in UE trajectory prediction. This approach filters out impossible position predictions and also provides contextual cues for recognizing mobility patterns (e.g., turning at intersections) [37]. Specifically, we learn the conditional probability of the future UE positions given past positions using the next token prediction capability of LMM in an autoregressive manner:

$$\begin{aligned}&\text{Prob}\big(\mathbf{p}_{\text{ue}}^{(T:T+T_\text{p})} \mid \mathbf{p}_{\text{ue}}^{(T-T_\text{w}:T-1)}, \mathcal{A}_\text{R}, \mathcal{A}_\text{B}\big)\\ &\quad= \prod_{t=0}^{T_\text{p}} \text{Prob}\big(\mathbf{p}_{\text{ue}}^{(T+t)} \mid \mathbf{p}_{\text{ue}}^{(T-T_\text{w}:T+t-1)}, \mathcal{A}_\text{R}, \mathcal{A}_\text{B}\big).\end{aligned} \tag{25}$$

Note that LMM is trained with the autoregressive language modeling objective (i.e., next token prediction), where the model learns to predict the most likely subsequent token given all preceding tokens in a sequence [38]. This next token prediction mechanism of LMM is particularly beneficial in LMM-EMM since the next token prediction can be interpreted as predicting the future UE position based on the historical UE trajectory.

[2]UE position can be acquired from various positioning techniques leveraging global navigation satellite system (GNSS), sensing, and wireless measurements [33], [34].

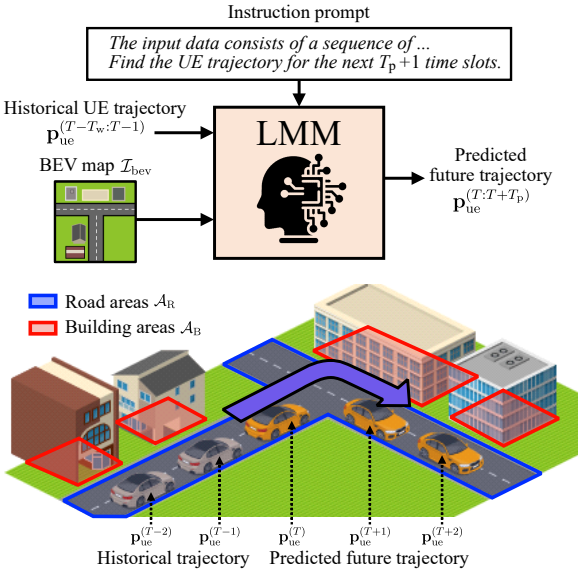


Fig. 5: Illustration of LMM-based trajectory prediction using BEV images.

An intriguing feature of LMM-EMM is the instruction learning strategy that trains the model using natural language instructions specifying not only the input, task, and desired output, but also the task-relevant context. In our case, the instruction prompts explicitly specify which environmental information should be emphasized in the multimodal data, allowing LMM to focus on the features most relevant to each task (see Fig. 6). For the UE trajectory prediction task, the instruction prompts are structured as follows:

- **Input description**: *"The input is a sequence of historical UE positions from $(T-T_\text{w})$th time slot to $(T-1)$th time slot $\mathbf{p}_{\text{ue}}^{(T-T_\text{w}:T-1)}$ and the BEV map image $\mathcal{I}_{\text{bev}}$."*
- **Environment description**: *"The UE movement is constrained by physical environments such as road boundaries $\mathcal{A}_\text{R}$ and buildings $\mathcal{A}_\text{B}$. Ensure the UE trajectory is in feasible regions within the given environment."*
- **Task instruction**: *"Find the UE trajectory for the next $T_\text{p}+1$ time slots $\mathbf{p}_{\text{ue}}^{(T:T+T_\text{p})}$."*

Given the instruction prompts, the LMM generates the response prompt as *"The predicted UE trajectory for the next $T_\text{p}+1$ time slots is $\mathbf{p}_{\text{ue}}^{(T:T+T_\text{p})}$."*

B. LMM-Based Static Channel Capacity Estimation via CCM

In the channel capacity estimation step, we learn the CCM f_{ccm} in (18), which maps the UE and SBS positions $(\mathbf{p}_{\text{ue}}^{(t)}, \mathbf{p}_{\text{sbs},m})$ to the static channel capacities $(R_{\text{ideal}}^{(t)}(m), R_{\text{nlos}}^{(t)}(m))$, parametrized by reflector information $\mathcal{E}$. Two main challenges in learning f_{ccm} are: 1) the discontinuous variation of static channel capacities with UE mobility (see Remark 1), and 2) the difficulty of directly acquiring reflector information $\mathcal{E}$, as accurately modeling and extracting $\mathcal{E}$ in

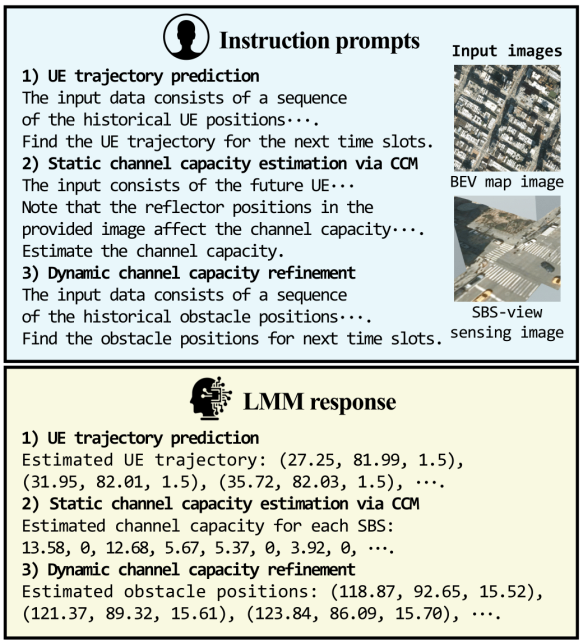


Fig. 6: Examples of instruction prompt and LMM response.

real environments is highly complex. To address this issue, we learn a surrogate function $\tilde{f}_{\mathrm{ccm}}$ that takes the BEV map image, which implicitly encodes $\mathcal{E}$:

$$\big(R^{(t)}_{\mathrm{ideal}}(m), R^{(t)}_{\mathrm{nlos}}(m)\big) = \tilde{f}_{\mathrm{ccm}}\big(\mathbf{p}^{(t)}_{\mathrm{ue}}, \mathbf{p}_{\mathrm{sbs},m}, \mathcal{I}_{\mathrm{bev}}\big). \tag{26}$$

Furthermore, we exploit the multimodal reasoning capability of LMMs to interpret reflector information from the BEV map and to pinpoint locations where channel capacity discontinuities may occur. In doing so, LMM-EMM integrates environmental context into the surrogate model, facilitating reliable approximation of f_{ccm} and precise estimation of static channel capacities at arbitrary UE and SBS locations. The instruction prompts for the LMM-based channel capacity estimation are structured as follows:

- **Input description**: *"The input is the future UE position* $\mathbf{p}^{(t)}_{\mathrm{ue}}$ *and the BEV map* $\mathcal{I}_{\mathrm{bev}}$*."*
- **Environment description**: *"Note that the reflector positions in the provided image affect the channel capacity by reflecting or blocking the propagation paths."*
- **Task instruction**: *"Estimate the static channel capacities* $\{(R^{(t)}_{\mathrm{ideal}}(m), R^{(t)}_{\mathrm{nlos}}(m)) \mid m \in \mathcal{M}\}$ *from the UE position."*

Given the instruction prompts, we obtain the estimated channel capacities from the LMM response: *"The estimated static channel capacities are* $\{(R^{(t)}_{\mathrm{ideal}}(m), R^{(t)}_{\mathrm{nlos}}(m)) \mid m \in \mathcal{M}\}$*."*

C. LMM-Based Blockage Prediction and Dynamic Channel Capacity Refinement

In the blockage prediction and dynamic channel capacity refinement step, we first estimate the LoS indicators in (17) $\{c^{(T:T+T_{\mathrm{p}})}_{\mathrm{los}}(m) \mid m \in \mathcal{M}\}$ from SBS-view sensing images. Using the estimated LoS indicators, together with the static channel capacity estimates $\{R^{(T:T+T_{\mathrm{p}})}_{\mathrm{ideal}}(m), R^{(T:T+T_{\mathrm{p}})}_{\mathrm{nlos}}(m) \mid m \in \mathcal{M}\}$, we then compute the achievable channel capacity $\{R^{(T:T+T_{\mathrm{p}})}(m) \mid m \in \mathcal{M}\}$ according to (15).

Specifically, let $\mathcal{N}_{\mathrm{obs},m} = \{1, \ldots, N_{\mathrm{obs},m}\}$ be the set of objects near the SBS m. For each object $i \in \mathcal{N}_{\mathrm{obs},m}$, we denote the center position of the object i in the camera coordinate system as

$$\mathbf{q}^{(t)}_{\mathrm{obs},i} = \big[x^{(t)}_{\mathrm{obs},i}\ y^{(t)}_{\mathrm{obs},i}\ d^{(t)}_{\mathrm{obs},i}\big] \tag{27}$$

where $x^{(t)}_{\mathrm{obs},i}$, $y^{(t)}_{\mathrm{obs},i}$, and $d^{(t)}_{\mathrm{obs},i}$ are x-coordinate, y-coordinate, and depth of the object centroid, respectively. Then, the 2D rectangular region $\mathcal{A}^{(t)}_i \subset \mathbb{R}^2$ in the image occupied by the object i at time t is given by

$$\begin{aligned}\mathcal{A}^{(t)}_i = &\Big[x^{(t)}_{\mathrm{obs},i} - \frac{1}{2}w^{(t)}_{\mathrm{obs},i}, x^{(t)}_{\mathrm{obs},i} + \frac{1}{2}w^{(t)}_{\mathrm{obs},i}\Big] \\ &\times \Big[y^{(t)}_{\mathrm{obs},i} - \frac{1}{2}h^{(t)}_{\mathrm{obs},i}, y^{(t)}_{\mathrm{obs},i} + \frac{1}{2}h^{(t)}_{\mathrm{obs},i}\Big]\end{aligned} \tag{28}$$

where $w^{(t)}_{\mathrm{obs},i}$ and $h^{(t)}_{\mathrm{obs},i}$ are the width and height of the bounding box of the object i in the camera coordinate system, respectively. Using this, $c^{(t)}_{\mathrm{los}}(m)$ is defined as

$$c^{(t)}_{\mathrm{los}}(m) = \begin{cases} 0 \text{ if } (\bar{x}^{(t)}_{\mathrm{ue}}, \bar{y}^{(t)}_{\mathrm{ue}}) \in \mathcal{A}^{(t)}_i \text{ and } d^{(t)}_{\mathrm{obs},i} < d^{(t)}_{\mathrm{ue}}\ \exists i \in \mathcal{N}_{\mathrm{obs},m} \\ 1 \text{ otherwise} \end{cases} \tag{29}$$

where $d^{(t)}_{\mathrm{ue}} = \|\mathbf{p}^{(t)}_{\mathrm{ue}} - \mathbf{p}^{(t)}_{\mathrm{sbs},m}\|_2$ and $(\bar{x}^{(t)}_{\mathrm{ue}}, \bar{y}^{(t)}_{\mathrm{ue}})$ is the x- and y-coordinates of UE in the camera coordinate system at time slot t (see Fig. 7). Note that $(\bar{x}^{(t)}_{\mathrm{ue}}, \bar{y}^{(t)}_{\mathrm{ue}})$ can be obtained from the GCS UE position $\mathbf{p}^{(t)}_{\mathrm{ue}}$ through rotation as $(\bar{x}^{(t)}_{\mathrm{ue}}, \bar{y}^{(t)}_{\mathrm{ue}}) = \big(\frac{x^{(t)}_r}{z^{(t)}_r}, \frac{y^{(t)}_r}{z^{(t)}_r}\big)$ where

$$\big[x^{(t)}_r\ y^{(t)}_r\ z^{(t)}_r\big]^{\mathrm{T}} = \mathbf{R}\big(\mathbf{p}^{(t)}_{\mathrm{ue}} - \mathbf{p}_{\mathrm{sbs},m}\big). \tag{30}$$

Here, $\mathbf{R}$ is the camera rotation matrix given by [6]

$$\mathbf{R} = \begin{bmatrix} \sin\theta & -\cos\theta & 0 \\ -\sin\phi\cos\theta & -\sin\phi\sin\theta & \cos\phi \\ \cos\phi\cos\theta & \cos\phi\sin\theta & \sin\phi \end{bmatrix} \tag{31}$$

where θ and ϕ denote the yaw and pitch angles of the camera.

To determine $c^{(t)}_{\mathrm{los}}(m)$, the SBS m needs to know the future positions of objects $\mathbf{q}^{(T:T+T_{\mathrm{p}})}_{\mathrm{obs},i}$ for $i \in \mathcal{N}_{\mathrm{obs},m}$. Analogous to UE trajectory prediction, these positions are inferred from historical positions $\mathbf{q}^{(T-T_{\mathrm{w}}:T-1)}_{\mathrm{obs},i}$. Unlike UE position data, which can be reported from UE via uplink channel, historical object positions are difficult to obtain since no communication links exist between the SBS and the objects [39]. As a remedy, we utilize object detection (OD) techniques to accurately extract $\mathbf{q}^{(T-T_{\mathrm{w}}:T-1)}_{\mathrm{obs},i}$ from SBS-view RGB-D images $\mathcal{I}_{\mathrm{sbs},m}$, and then use LMM to predict the future object trajectories:

$$\mathbf{q}^{(T:T+T_{\mathrm{p}})}_{\mathrm{obs},i} = f_{\mathrm{obs}}\big(\mathbf{q}^{(T-T_{\mathrm{w}}:T-1)}_{\mathrm{obs},i}, \mathcal{I}_{\mathrm{sbs},m}\big) \tag{32}$$

where f_{obs} is the obstacle trajectory prediction function. The instruction prompts for the LMM-based obstacle position prediction are designed as follows:

- **Input description**: *"The input is a sequence of the historical obstacle positions in the camera coordinate system* $\mathbf{q}^{(T-T_{\mathrm{w}}:T-1)}_{\mathrm{obs},i}$ *and the SBS-view sensing image* $\mathcal{I}_{\mathrm{sbs}}$*."*
- **Environment description**: *"In the camera coordinate system, obstacle movement is constrained by environmental elements including road boundaries for vehicles and*

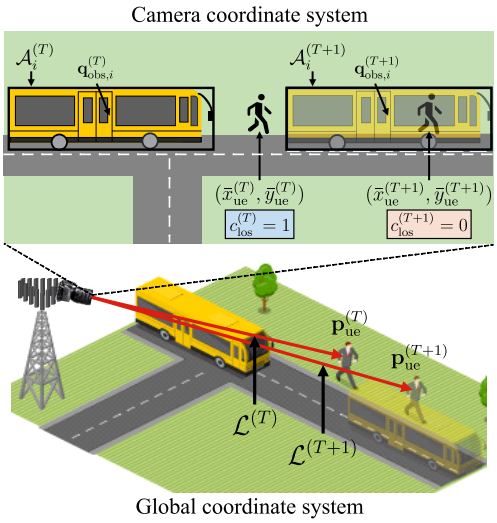


Fig. 7: Illustration of LMM-based dynamic blockage prediction with SBS-view sensing images.

immobile structures such as buildings that define non-navigable regions observed in $\mathcal{I}_{\text{sbs}}$*."*

- **Task instruction**: *"Find the positions of dynamic obstacles for the next* $(T_{\text{p}}+1)$ *time slots* $\mathbf{q}_{\text{obs},i}^{(T:T+T_{\text{p}})}$*."*

The LMM response prompts generated from the instruction prompts are *"The positions of dynamic obstacles for the next* $T_{\text{p}}+1$ *time slots are* $\mathbf{q}_{\text{obs},i}^{(T:T+T_{\text{p}})}$*."*

Once we obtain the predicted static channel capacities and LoS indicators $\{R_{\text{ideal}}^{(T:T+T_{\text{p}})}(m), R_{\text{nlos}}^{(T:T+T_{\text{p}})}(m), c_{\text{los}}^{(T:T+T_{\text{p}})}(m) \mid m \in \mathcal{M}\}$, for all SBSs, we acquire the achievable channel capacity $\{R^{(T:T+T_{\text{p}})}(m) \mid m \in \mathcal{M}\}$ using (15).

D. Proactive Handover Optimization

In the proactive handover optimization step, we find the optimal SBS indices $\hat{m}^{(T:T+T_{\text{p}})}$ in future time slots by solving the optimization problem $\mathscr{P}$ in (23) based on the estimated channel capacities $\{R^{(T:T+T_{\text{p}})}(m) \mid m \in \mathcal{M}\}$. One intuitive yet naive approach to solve $\mathscr{P}$ is an exhaustive search, which evaluates all possible sequences of SBSs. Although exhaustive search guarantees the optimal solution, it becomes computationally prohibitive in UDN because the number of candidates grows exponentially as $M^{T_{\text{p}}+1}$.

To obtain a tractable solution of $\mathscr{P}$, we adopt a DP-based approach, which solves complex problems by decomposing them into a series of simpler subproblems [40]. Owing to the cumulative structure of the capacity, the optimal solution at each time slot can be recursively derived from the solutions of previous time slots. In fact, the computational complexity of the proposed DP-based method is $T_{\text{p}}M^2$, which is significantly lower than that of the exhaustive search. Specifically, let $g(T+t, m^{(T+t)})$ be the maximum cumulative capacity when the UE is connected to the SBS $m^{(T+t)}$ at time slot $T+t$:

$$g\big(T+t, m^{(T+t)}\big) = \max_{m^{(T:T+t-1)}} \Bigg[\sum_{t'=0}^{t} R_{\text{eff}}^{(T+t')}\big(m^{(T+t'-1:T+t')}\big) \times \mathbb{1}_{\min}\big(m^{(T+t'-1:T+t')}\big)\Bigg] \quad (33)$$

where $\mathbb{1}_{\min}(m^{(t-1:t)})$ is the capacity compliance indicator defined as

$$\mathbb{1}_{\min}(m^{(t-1:t)}) = \begin{cases} 1 & \text{if } R_{\text{eff}}^{(t)}(m^{(t-1:t)}) \geq R_{\min}^{(t)} \\ -\infty & \text{if } R_{\text{eff}}^{(t)}(m^{(t-1:t)}) < R_{\min}^{(t)}. \end{cases} \quad (34)$$

Note that $g(T+t, m^{(T+t)})$ can be decomposed as

$$g(T+t, m^{(T+t)}) = \max_{m^{(T+t-1)}} \Big[g\big(T+t-1, m^{(T+t-1)}\big) + R_{\text{eff}}^{(T+t)}\big(m^{(T+t-1:T+t)}\big)\mathbb{1}_{\min}\big(m^{(T+t-1:T+t)}\big)\Big]. \quad (35)$$

Based on this recurrence formula, starting from initial $g(T, m^{(T)}) = R_{\text{eff}}^{(T)}(m^{(T-1:T)})\mathbb{1}_{\min}(m^{(T-1:T)})$, we sequentially compute $\big\{\{g(T+t, m^{(T+t)})\}_{t=1}^{T_{\text{p}}} \mid m \in \mathcal{M}\big\}$, which are subsequently stored in the DP table. After that, $\hat{m}^{(T:T+T_{\text{p}})}$ are retrieved in reverse chronological order from the computed DP table. First, $\hat{m}^{(T+T_{\text{p}})}$ is obtained as

$$\hat{m}^{(T+T_{\text{p}})} = \underset{m^{(T+T_{\text{p}})}}{\operatorname{argmax}}\ g\big(T+T_{\text{p}}, m^{(T+T_{\text{p}})}\big). \quad (36)$$

Then, $\hat{m}^{(t)}$ are determined recursively for $t = T_{\text{p}}-1, \cdots, 0$:

$$\hat{m}^{(T+t)} = \underset{m^{(T+t)}}{\operatorname{argmax}} \Big[g(T+t, m^{(T+t)}) + R_{\text{eff}}^{(T+t+1)}(m^{(T+t:T+t+1)})\mathbb{1}_{\min}(m^{(T+t:T+t+1)})\Big]. \quad (37)$$

Finally, the SBSs execute handovers according to the determined indices $\hat{m}^{(T:T+T_{\text{p}})}$.

IV. Practical Issues

In this section, we discuss several practical issues related to the implementation of LMM-EMM, focusing on end-to-end latency and challenges associated with its deployment.

A. End-to-end Latency

The key requirement for reliable handover decision-making is that the total end-to-end latency of LMM-EMM remain shorter than the timescale over which the channel varies sufficiently to affect the ergodic channel capacity. As shown in Equation (38), the ergodic channel capacity is determined by slow-varying channel parameters, such as angles and path losses, whose coherence time T_{c} (i.e., beam coherence time) is given by [41]

$$T_{\text{c}} = \frac{D}{v\cos\theta}\cos^{-1}(\psi^2\log\zeta + 1) \quad (38)$$

where D is the communication distance, θ is the beam angle, v is the speed of the UE, ψ is the beamwidth, and $\zeta \in [0,1]$ is the threshold ratio of the received power to its peak value. For example, when a UE moves at 25 km/h with $D = 70$ m, $\theta = 60°$, and $\zeta = 0.8$, and the SBS is equipped with $N = 32$ antennas such that $\psi \approx \frac{4}{N} = 0.125$ rad, then the beam coherence time is $T_{\text{c}} \approx 840$ ms.

In our implementation, the latency of each LMM-EMM module on an NVIDIA L40S GPU is summarized as follows:

- **UE trajectory prediction**: $T_1 = 215$ ms (20 ms for UE location report and 195 ms for LMM inference)

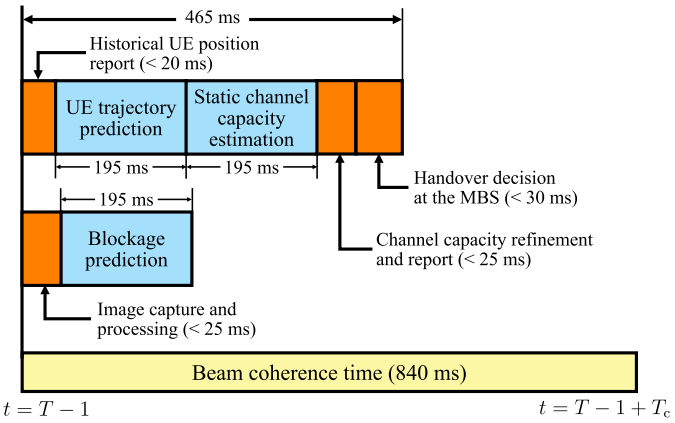


Fig. 8: End-to-end latency vs. beam coherence time.

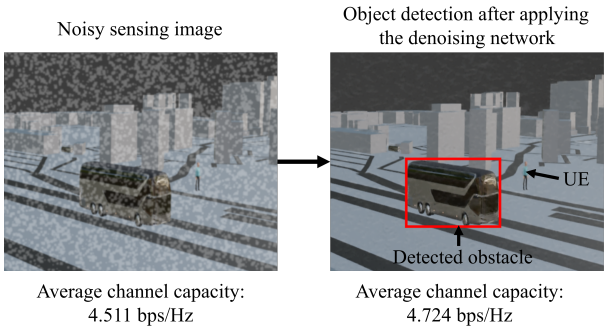
- **Static channel capacity estimation**: $T_2 = 195\,\text{ms}$ (195 ms for LMM inference)
- **Blockage prediction**: $T_3 = 220\,\text{ms}$ (25 ms for multi-modal image capture and processing, and 195 ms for LMM inference)
- **Channel capacity refinement, report, and handover**: $T_4 = 55\,\text{ms}$ (25 ms for channel capacity refinement and report, and 30 ms for DP-based handover decision)

Since blockage prediction can be performed in parallel with UE trajectory prediction and static channel capacity estimation, as it does not depend on others' outputs, these three modules are completed within $\max(T_1 + T_2, T_3) = 410\,\text{ms}$. Therefore, the total end-to-end latency of LMM-EMM is $T_{\text{tot}} = \max(T_1+T_2, T_3)+T_4 = 465\,\text{ms}$, which is substantially shorter than the beam coherence time $T_c \approx 840\,\text{ms}$ (see Fig. 8). Moreover, this latency is expected to decrease further as GPU hardware continues to evolve, leaving considerable room for further optimization in future implementations.

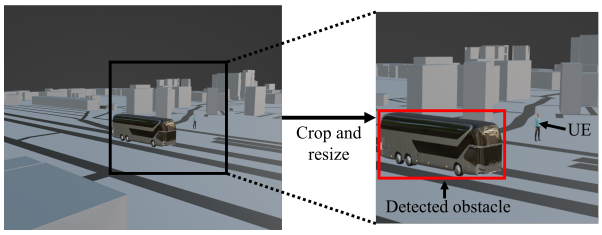
B. Deployment Issues

1) Noisy Sensing: In real-world settings, sensing images are often degraded by environmental factors such as rain, snow, fog, or dust. To demonstrate the robustness of the proposed LMM-EMM framework to imperfections in sensing data, we evaluate the handover performance of LMM-EMM under noisy sensing conditions caused by environmental factors. Under this setting, LMM-EMM achieves the average channel capacity of 4.511 bps/Hz, which corresponds to only a 4% reduction from the noise-free case. Furthermore, when the denoising and restoration techniques for adverse visual conditions are applied, LMM-EMM achieves the average channel capacity of 4.724 bps/Hz (see Fig. 9a) [42]. This corresponds to only a 1% decrease compared to 4.770 bps/Hz in the noise-free scenario (see Fig. 16).

2) FoV and Resolution Issues: One practical issue of LMM-EMM is that the SBS-view camera has a limited field of view (FoV) and resolution, which may restrict its ability to capture the entire scene. While it is true that a single camera has a limited FoV, the entire coverage area can be monitored by deploying multiple cameras, each covering a different sectorized region, similar to antenna sectorization. For example, three RGB-D cameras, each with a FoV of 120°, can cover the entire area surrounding the SBS.



(a) Illustration of the noisy sensing image and object detection after applying the denoising network.



(b) Illustration of the cropped image region around the predicted UE position.

Fig. 9: Examples of sensing images used for blockage prediction under noisy and low-resolution scenarios.

Regarding the image resolution issue, objects that contribute to LoS blockage (e.g., car, bus, and truck) can still be effectively identified even in low-resolution images because they are typically large enough to occupy multiple pixels in the captured images. Furthermore, in our work, we crop and resize the image region around the predicted UE position to magnify relevant objects and improve detection performance (see Fig. 9b).

3) Availability of Cameras and BEV Maps: One might be concerned that LMM-EMM can operate only when RGB-D cameras are installed and BEV maps are available at the SBSs. However, with ISAC emerging as a key component of upcoming 6G, SBSs are expected to be equipped with diverse sensing modalities including RGB-D cameras. Moreover, BEV maps can be readily obtained from publicly available sources such as OpenStreetMap [43] and Google Maps.

If the serving SBS cannot collect the sensing data due to the absence of sensors, it can request a neighboring SBS that can observe the UE and surrounding obstacles to estimate the blockage state. Specifically, the serving SBS transmits the estimated UE trajectory to the neighboring SBS. The neighboring SBS then constructs the LoS line segment between the serving SBS and the UE, and uses its own RGB-D camera to detect obstacles and represent them as 3D boxes. By checking whether the LoS line segment intersects the 3D boxes, the blockage state can be determined [39]. If no such neighbor exists, the blockage state can be inferred from past channel measurements (e.g., signal-to-interference-plus-noise ratio (SINR)), and the inferred past state can be used as a proxy for the future LoS state. Such temporal consistency is reasonable because obstacles that cause blockage (e.g., buses, vehicles) typically have non-negligible volume, so the blockage state does not change significantly over a short time

interval.

4) Robustness to Reflector Material Mismatch: A possible issue in LMM-EMM is that incorrect knowledge about the material properties may introduce errors in the channel capacity estimation. However, since material properties generally do not change significantly over time, the associated errors can be controlled via fine-tuning with real datasets. Even if the reflector material is incorrectly characterized, LMM-EMM can still maintain reasonable performance because the propagation geometry (e.g., LoS paths, reflection angles) remains unchanged despite the material mismatch. To validate this, we conduct an ablation study in which the reflector material in the validation dataset is changed from glass to concrete. We observe a performance degradation of 5.2% compared with the original case, which demonstrates that the proposed technique still maintains reasonable performance under incorrect knowledge of reflector materials (see Fig. 10).

5) Synchronization: According to the IEEE 1588-v2 protocol, inter-cell time synchronization errors within $0.1\,\mu$s may exist [44]. To evaluate the impact of these synchronization errors, we measure the average channel capacity under such timing offsets. We observe that the average channel capacity remains $4.770\,\mathrm{bps/Hz}$, identical to the perfectly synchronized case of $4.770\,\mathrm{bps/Hz}$, indicating no performance degradation. This is because the synchronization error is sufficiently small that the UE displacement during this interval is negligible, and thus the corresponding channel can be considered as unchanged.

6) SBS Position Error: Inaccurate SBS positions may affect blockage detection performance because the estimated LoS path can deviate from the actual one. To evaluate the impact of SBS position error, we plot the average channel capacity of the proposed LMM-EMM as a function of the SBS position error. We observe that under an SBS position error of 1 m, the average channel capacity reaches $4.758\,\mathrm{bps/Hz}$, which is only 0.3% lower than that under perfect SBS position knowledge (see Fig. 11). Moreover, since SBS positions are typically fixed after deployment, they can be calibrated over time, so persistent location errors are unlikely in practical systems.

7) Soft Blockage: In practical environments, a soft blockage model may arise due to partial blockage and diffraction, which could introduce discrepancies in the channel capacity estimation. However, in the considered mmWave scenario, the impact of such effects is limited for two main reasons.

First, diffraction and refraction are generally weak at mmWave frequencies due to the short wavelength and high penetration loss [45]. As a result, the corresponding path components contribute much less to the received power than the reflected paths. Their effect on channel capacity variation is thus limited, and neglecting them is a reasonable approximation in the considered mmWave scenario.

Second, the impact of partial blockage on the channel capacity variation is expected to be limited in the considered mmWave scenario. This is because objects that induce partial blockage (e.g., pedestrians) typically yield a much smaller path gain $\beta_{m,l}[s]$ than specular reflectors (e.g., buildings), as they usually have a significantly larger roughness coefficient $\sigma_{m,l}$ (see Equation (4))[3]. Consequently, the channel capacity variation caused by soft blockages is relatively minor compared with the dominant effects of LoS and reflected paths.

To validate the adopted LoS/NLoS approximation combined with the reflection model, we computed the cosine similarity between the approximated and actual channels at various roughness coefficients of the objects. We observe that the cosine similarity exceeds 0.93 even when the roughness coefficient is 3 mm, which confirms that the approximated model is sufficiently accurate for channel capacity comparison and handover decision-making (see Fig. 12).

V. Experimental Results

A. Simulation Setup

We consider the UDN system where $M = 14$ SBSs equipped with $N_x \times N_y = 8 \times 4$ UPA antennas serve a UE equipped with a single antenna. The antenna orientations of SBS and UE are $(0°, -10°, 0°)$ and $(0°, 0°, 0°)$, respectively. The UE moves along the road at a speed of $v = 25\,\mathrm{km/h}$ and it turns left, turns right, or continues straight ahead at each intersection with probabilities of 25%, 25%, and 50%, respectively. The carrier frequency and bandwidth are $f_\mathrm{c} = 28\,\mathrm{GHz}$ and $B = 100\,\mathrm{MHz}$, respectively. The observation time window T_w and the prediction length T_p are set to 5. We consider the signal-to-noise ratio (SNR) of 15 dB and the handover interruption time of $\tau_\mathrm{ho} = 36\,\mathrm{ms}$. Also, an RGB-D camera having a resolution of 1024×768 and a FoV of $120°$ is mounted on the top side of the SBS. For the pretrained LMM, we adopt the LLaVA-1.5-7B model [2], which is one of the most prevalent open-source LMMs. We conduct experiments on an Intel Xeon Gold 6326 CPU server equipped with a 16-core CPU and an NVIDIA L40S GPU with 48 GB of memory.

We compare the mobility management performance of the proposed LMM-EMM with four conventional techniques and the ideal scheme with perfect channel state information (CSI):

1) *Perfect CSI-based scheme*: an upper bound of average capacity assuming all channel capacities and blockage indicators are perfectly known.
2) *DRL-based scheme* [17]: A proactive handover scheme that uses deep Q-learning (DQN) to optimize the handover parameters (i.e., handover offset and time-to-trigger (TTT)) with handover-type prediction.
   - **State**: RSRP and SINR from both the serving and target SBSs, and the UE's distance to the target SBS, speed, and moving direction.
   - **Action**: handover margin and time-to-trigger.
   - **Reward**: negative weighted sum of the ping-pong, too-early, and too-late handover occurrence rates.
3) *Vision-aided scheme* [19]: A proactive handover scheme that employs SBS-view camera images and a YOLOv3 object detection model to predict the LoS path blockages and proactively trigger handover.
4) *LSTM-based scheme* [13]: A proactive handover scheme that uses LSTM to predict the future channel capacity in

[3]For instance, a path partially blocked by a pedestrian with a roughness coefficient of $\sigma_{m,l} = 3\,\mathrm{mm}$ yields a $\beta_{m,l}[s]$ that is approximately $1/467$ of that for a typical glass building with $\sigma_{m,l} = 1\,\mathrm{mm}$.

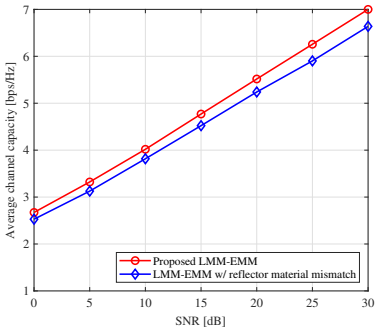


Fig. 10: Average channel capacity of LMM-EMM with reflector material mismatch.

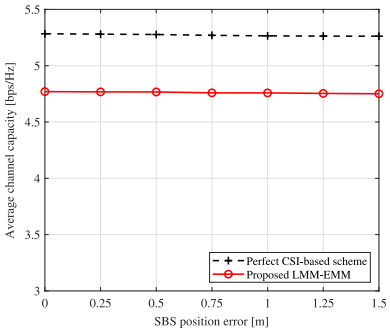


Fig. 11: Average channel capacity vs. SBS position error.

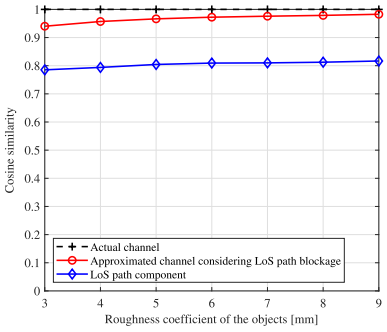


Fig. 12: Cosine similarity between approximated channel and actual channel.

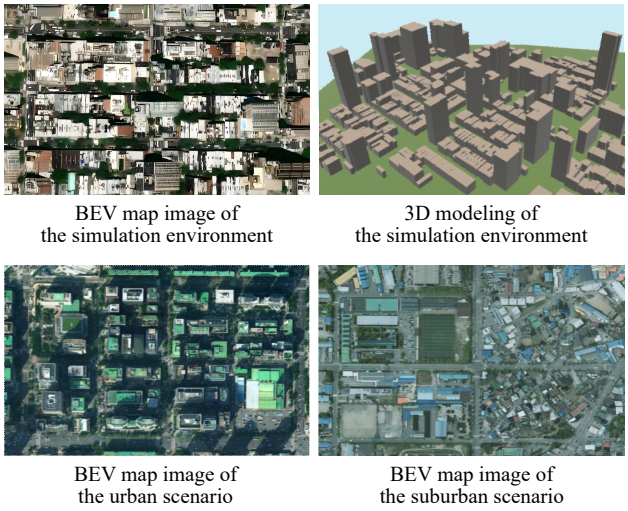


Fig. 13: Visualizations of the simulation environment on real-world scenarios.

an autoregressive manner and performs handover based on the predicted capacity.

5) *5G NR handover* [10]: A reactive handover scheme where the SBS determines the handover decisions based on consecutive RSRP reports from the UE.

B. Multimodal Dataset Generation and LMM Fine-Tuning

1) Multimodal Dataset Generation: To construct a realistic dataset for UDN scenarios, we develop a comprehensive pipeline that integrates 3D environmental modeling, wireless channel generation, and sensory data acquisition. First, we reconstructed a 3D urban environment resembling a real-world deployment using building layouts extracted from OpenStreetMap, which has been reported to provide sub-meter positional accuracy [43]. Based on this 3D environment, we performed ray tracing and wireless channel generation using NVIDIA Sionna RT [46], which provides physically consistent channel modeling based on ray tracing and 3GPP TR 38.901 [47]. Since 3GPP TR 38.901 is widely adopted to emulate realistic propagation characteristics (e.g., reflection and blockage), the generated wireless channels closely resemble practical deployments. Therefore, although the data are simulated, the dataset maintains high physical fidelity and supports reliable performance evaluation for practical deployment scenarios. To validate performance in general wireless environments, we generate two additional real-world scenarios: 1) an urban environment characterized by high building density and tall structures, and 2) a suburban environment with moderate building density and low heights (less than 10 m) (see Fig. 13).

2) LMM Fine-Tuning: For LMM fine-tuning, we employ supervised fine-tuning (SFT) in which the LMM is trained on input–output pairs generated by a real-world wireless simulator (NVIDIA Sionna RT [46]) to minimize the negative log-likelihood (NLL) loss $\mathcal{J}$ for the response tokens $\boldsymbol{w}^*$:

$$\mathcal{J} = -\frac{1}{N_{\text{token}}} \sum_{i=1}^{N_{\text{token}}} \log\left([\boldsymbol{\pi}_i]_{w_i^*}\right). \tag{39}$$

By adjusting the model parameters through SFT, LMM-EMM can capture fine-grained mobility behaviors and channel variations, thereby enhancing mobility management performance. One major issue in SFT is the excessive computational burden of updating a massive number of network parameters (e.g., 7 billion parameters in LLaVA-1.5-7B), which results in extended training times and considerable resource demands. We address this issue by adopting the low-rank adaptation (LoRA) technique, which adjusts only a small fraction of parameters relevant to the task [48]. Specifically, we select a subset of parameters $\mathbf{W} \in \mathbb{R}^{X\times Y}$ and update it with the product of two low-rank matrices $\mathbf{A} \in \mathbb{R}^{X\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times Y}$:

$$\mathbf{W}' = \mathbf{W} + \mathbf{AB} \tag{40}$$

where $\mathbf{W}' \in \mathbb{R}^{X\times Y}$ and r denote the updated model parameters and the rank of low-rank matrices, respectively. We use a LoRA rank of $r = 16$ and a learning rate of 1×10^{-4}, which is decayed by a factor of 0.1 every 10 steps. The generated dataset for fine-tuning contains $N_{\text{train}} = 16\,000$, $N_{\text{val}} = 2\,000$, and $N_{\text{test}} = 2\,000$ samples for training, validation, and test, respectively.

C. Simulation Results

In Fig. 14, we visualize CCM generated by LMM-EMM and the conventional DL-based scheme [49]. To assess the quality of the generated CCM, we evaluate the normalized mean square error (NMSE) against the ground-truth CCM as shown in Appendix B. The results indicate that the proposed

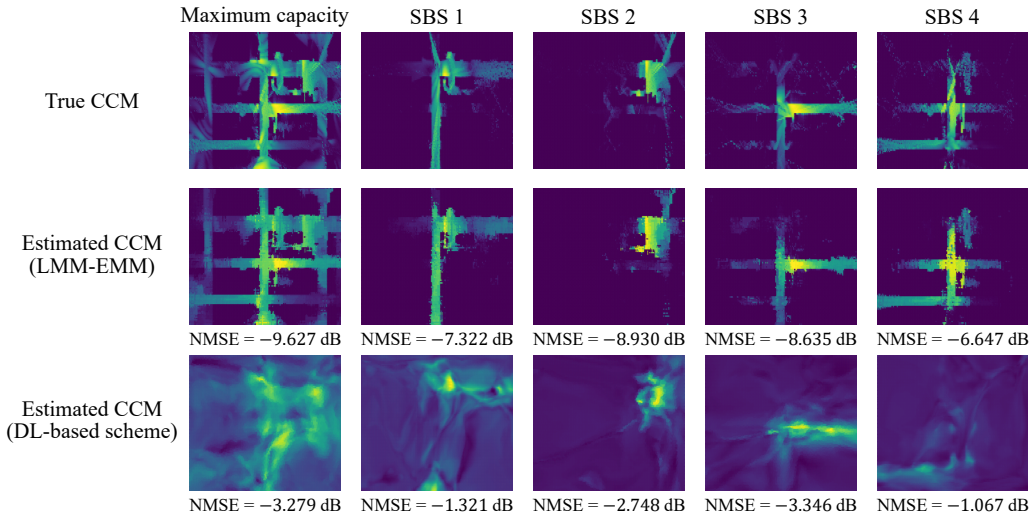


Fig. 14: Comparison of CCMs generated by LMM-EMM and the conventional DL-based scheme.

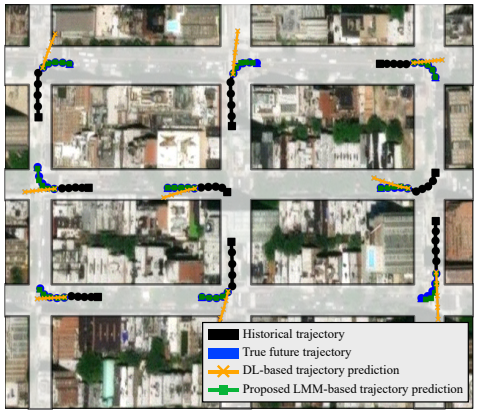


Fig. 15: Visual comparison of predicted UE trajectories for LMM-EMM and the DL-based trajectory prediction.

LMM-EMM generates a more accurate CCM than the DL-based scheme, achieving a 6.3 dB improvement in NMSE. Additionally, in Fig. 15, we compare the predicted UE trajectory of LMM-EMM and the DL-based trajectory prediction. We observe that the trajectory predicted by LMM-EMM closely follows the ground-truth path and accurately captures direction changes and UE movement. Overall, these findings indicate that LMM-EMM understands environmental features well, including intersections and obstacle locations, so that it can facilitate reliable handover decisions.

In Fig. 16, we evaluate the channel capacity as a function of SNR. We see that the proposed LMM-EMM achieves a significant improvement in the channel capacity over the conventional mobility management techniques. For instance, at an SNR of 15 dB, LMM-EMM achieves more than 45% gain in channel capacity compared to the 5G NR handover scheme. Even when compared to the LSTM-based technique, LMM-EMM achieves a 21% gain in channel capacity. This is because the proposed LMM-EMM can properly comprehend the environmental information to predict the accurate channel capacity from multiple SBSs.

In Fig. 17, we evaluate the handover performance as a function of the UE speed. Due to the accurate estimation of the future trajectory of the UE and corresponding channels from the environment, the proposed LMM-EMM outperforms conventional mobility management techniques in terms of channel capacity. For example, when the speed of the UE is 25 km/h, LMM-EMM achieves more than 50% and 23% higher channel capacity compared to the 5G NR handover and LSTM-based technique, respectively. Since LMM-EMM intelligently analyzes road geometry to accurately estimate the UE trajectory and corresponding channel capacities, it achieves robust performance across diverse mobility scenarios.

To evaluate the performance of LMM-EMM in various deployment scenarios, we plot the channel capacity of the UE as a function of the number of SBSs in Fig. 18. Since each UDN system contains a different number of SBSs, validation across diverse numbers of SBSs is crucial to demonstrate generalization performance. We see that the channel capacity of the proposed LMM-EMM is much higher than that of conventional techniques. For example, when the number of SBSs is 12, LMM-EMM achieves more than 40% higher capacity compared to the 5G NR handover. By exploiting CCM, LMM-EMM can accurately estimate the channel capacity of each SBS and associate the appropriate SBS with the UE, regardless of the number of SBSs.

In Fig. 19, we evaluate the handover performance with various handover interruption times. The results demonstrate that the proposed LMM-EMM significantly outperforms conventional mobility management techniques. For instance, when $\tau_{\mathrm{ho}} = 54$ ms, LMM-EMM achieves more than 38% capacity gain over the 5G NR handover. Even when compared to the vision-based technique, LMM-EMM achieves a 17% increase in channel capacity. Notably, LMM-EMM exhibits minimal degradation in average capacity as the handover interruption time increases. This indicates that LMM-EMM can avoid redundant handovers through accurate recognition of transient channel variations caused by dynamic obstacles.

To assess the generalization performance of LMM-EMM, we evaluate the channel capacity across diverse scenarios, including urban and suburban areas, in Fig. 20. We observe that

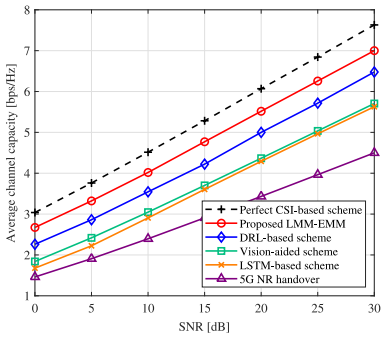


Fig. 16: Average channel capacity as a function of SNR.

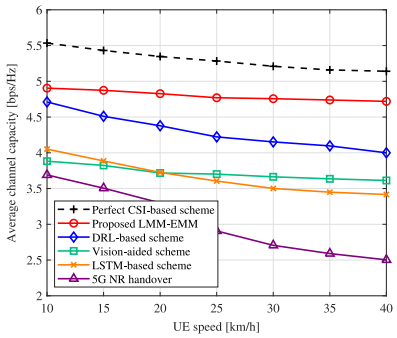


Fig. 17: Average channel capacity as a function of the UE speed.

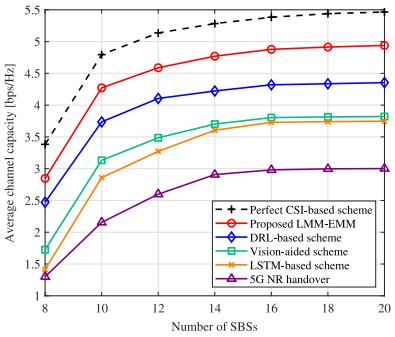


Fig. 18: Average channel capacity as a function of the number of SBSs.

LMM-EMM exhibits superior performance against conventional techniques across all wireless scenarios. For example, LMM-EMM achieves 30% and 26% higher channel capacity than the DRL-based technique in urban and suburban settings, respectively. This generalization performance results from the multimodal reasoning capability of LMM, which accurately interprets the spatial distribution of reflectors and the UE mobility patterns in diverse environments.

To verify that LMM-EMM remains effective even with smaller antenna arrays, in Fig. 21, we evaluate the channel capacity under various antenna array sizes. We observe that although the instantaneous channel experiences fluctuations due to small-scale fading, these fluctuations rarely change the handover decision. For example, even when the antenna array size is 8, LMM-EMM achieves 4.017 bps/Hz, which corresponds to 90% of the channel capacity achieved by the optimal handover with perfect CSI. These results indicate that the proposed LMM-EMM retains most of the ideal handover gain even when the number of antennas is small.

In Fig. 22, we evaluate the average channel capacity performance of LMM-EMM across different architectural layouts, including highway and indoor scenarios. To this end, we perform few-shot fine-tuning using only 600 samples collected from those environments. We observe that at an SNR of 15 dB, LMM-EMM still achieves 10% and 7% channel capacity gain over the conventional DRL-based approach in highway and indoor scenarios, respectively (see Fig. 22). This strong adaptability of LMM-EMM stems from the ability of LMM to capture the scenario-invariant propagation mechanisms (e.g., reflection and blockage) that are shared across different environments. While architectural layouts vary across environments, these variations primarily affect site-specific geometry, not the fundamental propagation physics. This means that LMM can quickly understand the propagation characteristics in new environments and rapidly adapt to unseen scenarios.

Fig. 23 shows an ablation study where each module is replaced with a CNN+LSTM model. Replacing the trajectory prediction, channel capacity estimation, and blockage prediction modules with CNN+LSTM leads to data-rate degradations of 6.4%, 24.4%, and 6.1%, respectively, since a simple discriminative model lacks the multimodal reasoning provided by LMM. The degradation is particularly noticeable for the channel capacity estimation module, because the channel capacity is more sensitive to variations and piecewise discontinuities in the propagation environment than the UE and blockage trajectories. Also, in Fig. 24, we conduct an experiment where the backbone LMM is replaced with lightweight models: 1) MobileVLM-1.7B [50] and 2) TinyLLaVA-1.5B [51]. We observe that MobileVLM-1.7B and TinyLLaVA-1.5B achieve 4.263 bps/Hz and 4.305 bps/Hz, respectively. Despite reducing the parameter size by more than 75% compared with LLaVA-1.5-7B, the performance degradation remains marginal, which demonstrates the feasibility of lightweight deployment.

To investigate how the accumulation of errors affects overall system performance, we conduct an error propagation experiment across the entire framework in Table I. Specifically, we first inject artificial noise into the input UE trajectory and then evaluate its impact on various metrics, including the UE trajectory prediction root mean square error (RMSE), static channel capacity estimation NMSE, average channel capacity, and the worst 10% and 20% case channel capacity. The definitions of RMSE and NMSE are provided in Appendix B. When noise with variance 1 m is added to the input UE trajectory, the UE trajectory prediction RMSE is 0.675 m and the average channel capacity is 4.657 bps/Hz. Compared with the noise-free case, the channel capacity decreases by only 2.4%, indicating that the performance degradation is limited.

Table II presents the robustness of LMM-EMM to the error propagation in real-world scenarios. The error level for each module was chosen to reflect practically plausible conditions, motivated by prior studies [33], [52], [53]. As shown in Table II, the performance is most sensitive to errors in the CCM estimation module. When the noise with a variance of 3 bps/Hz is added to the CCM estimation, the cumulative capacity per unit bandwidth decreases to 4.554 bps/Hz, which is the lowest among all error sources. This is because the noise in the CCM makes it difficult for LMM to learn the underlying physical relationship between the propagation environment (e.g., reflection) and estimate the resulting channel capacity.

In real-world scenarios, small-scale elements (e.g., trees, small objects) can cause temporary blockages and scattering effects, leading to some minor errors in blockage detection and channel capacity estimation. To account for such real-world variability, in Table III, we evaluate the average channel capacity under additional blockage detection errors and CCM estimation noise. We observe that even with an additional

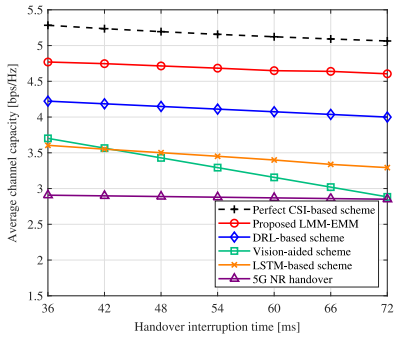


Fig. 19: Average channel capacity as a function of handover interruption time.

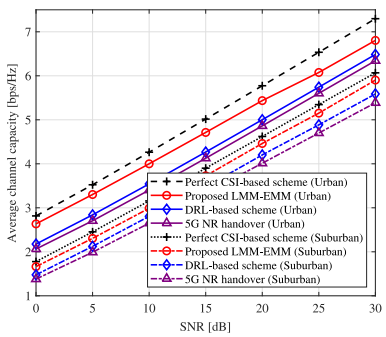


Fig. 20: Average channel capacity as a function of SNR in various wireless scenarios.

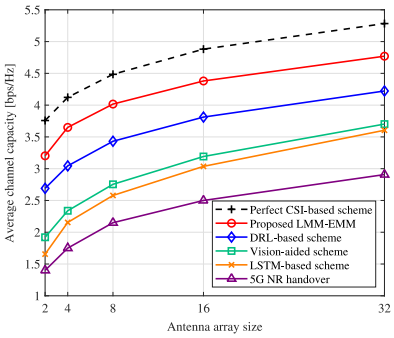


Fig. 21: Average channel capacity as a function of antenna array size.

TABLE I: Impact of input UE trajectory errors on prediction accuracy and communication performance.

Noise variance	**0 m**	**0.5 m**	**1 m**	**1.5 m**	**2 m**
UE trajectory prediction RMSE	0.252 m	0.490 m	0.675 m	0.844 m	0.964 m
Static channel capacity estimation NMSE	−9.627 dB	−9.327 dB	−8.549 dB	−7.542 dB	−6.379 dB
Blockage prediction accuracy	95.5%	95.4%	94.9%	93.7%	92.2%
Average channel capacity	4.770 bps/Hz	4.729 bps/Hz	4.657 bps/Hz	4.556 bps/Hz	4.371 bps/Hz
10% worst case channel capacity	2.490 bps/Hz	2.479 bps/Hz	2.440 bps/Hz	1.987 bps/Hz	1.798 bps/Hz
20% worst case channel capacity	2.744 bps/Hz	2.736 bps/Hz	2.670 bps/Hz	2.363 bps/Hz	2.181 bps/Hz
Handover failure rate	4.89 %	5.98 %	6.52 %	8.42 %	9.24 %
Service interruption probability	6.61 %	6.81 %	6.90 %	7.25 %	7.82 %

TABLE II: Ablation study on the impact of UE trajectory prediction, CCM estimation, and blockage prediction error

Error source	**Amount of error**	**UE trajectory prediction RMSE**	**Channel capacity estimation NMSE**	**Average channel capacity**
No noise addition	-	0.252 m	−9.627 dB	4.770 bps/Hz
Input UE trajectory	Noise variance = 1 m	0.675 m	−9.061 dB	4.659 bps/Hz
UE trajectory prediction	Noise variance = 1 m	1.083 m	−5.853 dB	4.632 bps/Hz
CCM estimation	Noise variance = 3 bps/Hz	-	−4.680 dB	4.554 bps/Hz
Blockage prediction	Error probability +10%p	-	-	4.747 bps/Hz

blockage error of 20% and a CCM noise standard deviation of 2 bps/Hz, LMM-EMM achieves an average channel capacity of 4.659 bps/Hz, corresponding to only a 2.4% performance degradation. This robustness is because handover decisions are often dominated by user location and major blockage events caused by large reflectors or obstacles, rather than by minor dispersion effects from small-scale objects.

In Table IV, we present the memory and computational complexities of LMM-EMM compared to conventional mobility management techniques. Although LMM-EMM exhibits the highest memory usage and longest inference time, its inference time remains shorter than the maximum handover latency (i.e., 360 ms) specified in 5G NR. Furthermore, the results can be pre-computed in advance of actual handover events by predicting future channel capacities, which further relaxes the practical constraints on inference time. This indicates that LMM-EMM has a feasible time complexity and is therefore compatible with the 5G NR standard. Moreover, recent progress in lightweight LMM and GPU hardware is expected to further alleviate memory and time complexities, so these issues are unlikely to pose practical concerns [54].

VI. Conclusion

In this paper, we proposed an environment-aware mobility management scheme for mmWave UDN systems. The key idea of the proposed LMM-EMM scheme is to leverage the CCM, an intrinsic mapping from UE and SBS positions to channel capacity. By harnessing the multimodal reasoning capability of LMMs, LMM-EMM captures the mobility pattern of the UE as well as reflection geometry and transient blockages caused by obstacles. Using the trained CCM with the predicted UE trajectory and blockage indicators, LMM-EMM estimates future channel capacities and employs DP to proactively determine handover decisions that maximize cumulative channel capacity. From numerical evaluations on various environments, we demonstrated that LMM-EMM achieves substantial channel capacity gains over the conventional DL-based methods. In this paper, we restricted our attention to mobility management but we believe that there are many research extensions of LMM-EMM such as user scheduling and random access.

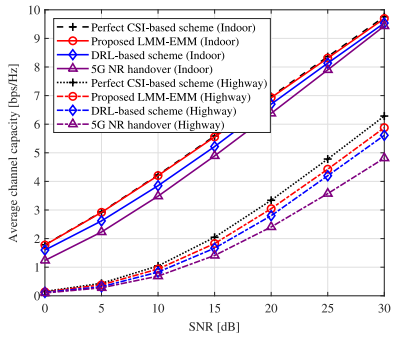


Fig. 22: Average channel capacity as a function of SNR in different architectural layouts after few-shot adaptation.

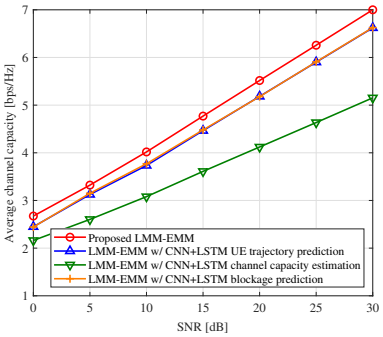


Fig. 23: Average channel capacity vs. SNR, where each module is replaced by a CNN+LSTM.

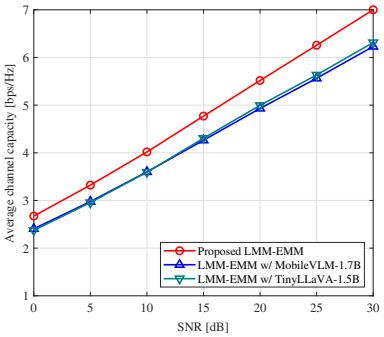


Fig. 24: Average channel capacity vs. SNR with the backbone replaced by a lightweight LMM.

TABLE III: Average channel capacity (bps/Hz) at SNR $= 15$ dB, with blockage detection error rate p_{block} and CCM noise standard deviation σ_n.

$\sigma_n \backslash p_{\text{block}}$	**10%**	**15%**	**20%**	**25%**	**30%**
0.5 bps/Hz	4.754	4.754	4.743	4.738	4.727
1 bps/Hz	4.736	4.735	4.726	4.721	4.706
1.5 bps/Hz	4.706	4.704	4.697	4.695	4.678
2 bps/Hz	4.665	4.668	4.659	4.662	4.641
2.5 bps/Hz	4.642	4.647	4.647	4.653	4.616

TABLE IV: Memory usage and inference time

Methods	**Memory usage**	**Inference time**
Proposed LMM-EMM	14.16 GB	195.72 ms
DRL-based scheme	1.24 GB	45.68 ms
Vision-aided scheme	0.82 GB	14.91 ms
LSTM-based scheme	0.27 GB	6.69 ms

APPENDIX A
PROOF OF LEMMA 1

In MISO systems, the channel capacity is expressed as

$$R^{(t)}(m) = B\sum_{s=1}^{S}\log_2\left(1+\frac{P_{\text{t}}}{N\sigma_{\text{n}}^2}\sum_{i=1}^{N}|h_m^{(t)}[s,i]|^2\right) \quad (41)$$

where $h_m^{(t)}[s,i]$ is the ith element of $\mathbf{h}_m^{(t)}[s]$. When N is sufficiently large, by applying the law of large numbers, we obtain

$$\frac{1}{N}\sum_{i=1}^{N}|h_m^{(t)}[s,i]|^2 \approx \mathbb{E}\big[|h_m^{(t)}[s,i]|^2\big] \quad (42)$$

$$= \sum_{l=0}^{L-1}\beta_{m,l}^{(t)}\mathbb{E}\big[|\alpha_{m,l}^{(t)}|^2\big] + \sum_{p\neq l}\sqrt{\beta_{m,p}^{(t)}\beta_{m,l}^{(t)}}\mathbb{E}\big[\alpha_{m,p}^{(t)}\bar{\alpha}_{m,l}^{(t)}\big]e^{j\vartheta} \quad (43)$$

$$= \beta_{m,0}^{(t)} + \sum_{l=1}^{L-1}\beta_{m,l}^{(t)}. \quad (44)$$

This is from $\mathbb{E}\big[|\alpha_{m,l}^{(t)}|^2\big]=1$ and $\mathbb{E}\big[\alpha_{m,p}^{(t)}\bar{\alpha}_{m,l}^{(t)}\big]=0$ for $p\neq l$, where $\bar{\alpha}_{m,l}^{(t)}$ is the conjugate of $\alpha_{m,l}^{(t)}$ and ϑ is some real value between 0 and 2π. By plugging (44) into (41), we finally obtain

$$R^{(t)}(m) = B\sum_{s=1}^{S}\log_2\left(1+\frac{P_{\text{t}}}{\sigma_{\text{n}}^2}\left(\beta_{m,0}^{(t)}+\sum_{l=1}^{L-1}\beta_{m,l}^{(t)}\right)\right). \quad (45)$$

APPENDIX B
EVALUATION METRICS

This appendix defines the evaluation metrics used in this paper, namely the RMSE for UE trajectory prediction and the NMSE for channel capacity estimation.

1) *UE trajectory prediction RMSE*: The RMSE for UE trajectory prediction is defined as

$$\text{RMSE} = \sqrt{\mathbb{E}\Big[\Big(\hat{\mathbf{p}}_{\text{ue}}^{(T:T+T_p)} - \mathbf{p}_{\text{ue}}^{(T:T+T_p)}\Big)^2\Big]} \quad (46)$$

where $\hat{\mathbf{p}}_{\text{ue}}^{(T:T+T_p)}$ is the estimated future UE trajectory over the prediction horizon from T to $T+T_p$.

2) *Channel capacity estimation NMSE*: The NMSE for channel capacity estimation is defined as

$$\text{NMSE} = \mathbb{E}\left[\frac{\big(\hat{R}^{(t)}(m) - R^{(t)}(m)\big)^2}{R^{(t)}(m)^2}\right] \quad (47)$$

where $\hat{R}^{(t)}(m)$ is the estimated channel capacity at the tth time slot for the mth SBS.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 34 892–34 916.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 24 824–24 837.

[4] A. Maatouk, N. Piovesan, F. Ayed, A. De Domenico, and M. Debbah, "Large language models for telecom: Forthcoming impact on the industry," *IEEE Commun. Mag.*, vol. 63, no. 1, pp. 62–68, 2025.

[5] H. J. Yang, H. Kim, H. Noh, S. Kim, and B. Shim, "Large multimodal model-empowered task-oriented autonomous communications: Design methodology and implementation challenges," *IEEE Veh. Technol. Mag.*, 2026.

[6] 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, 2024.

[7] J. Moon, S. Kim, H. Ju, and B. Shim, "Energy-efficient user association in mmWave/THz ultra-dense network via multi-agent deep reinforcement learning," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 2, pp. 692–706, 2023.

[8] S. Kim, J. Wu, and B. Shim, "Efficient channel probing and phase shift control for mmWave reconfigurable intelligent surface-aided communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 231–246, 2023.

[9] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 2016.

[10] *Radio Resource Control (RRC)*, 3rd Generation Partnership Project 3GPP™ TS 38.331 V18.4.0, Dec. 2024, Release 18.

[11] R. H. Clarke, "A statistical theory of mobile-radio reception," *Bell Sys. Technol. J.*, vol. 47, no. 6, pp. 957–1000, 1968.

[12] L. Yan, H. Ding, L. Zhang, J. Liu, X. Fang, Y. Fang, M. Xiao, and X. Huang, "Machine learning-based handovers for sub-6 GHz and mmWave integrated vehicular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4873–4885, 2019.

[13] S. H. A. Shah and S. Rangan, "Multi-cell multi-beam prediction using auto-encoder LSTM for mmWave systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10 366–10 380, 2022.

[14] H.-S. Park, H. Kim, C. Lee, and H. Lee, "Mobility management paradigm shift: from reactive to proactive handover using AI/ML," *IEEE Netw.*, vol. 38, no. 2, pp. 18–25, 2024.

[15] L. Jiao, P. Wang, A. Alipour-Fanid, H. Zeng, and K. Zeng, "Enabling efficient blockage-aware handover in RIS-assisted mmWave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2243–2257, 2021.

[16] W. Huang, M. Wu, Z. Yang, K. Sun, H. Zhang, and A. Nallanathan, "Self-adapting handover parameters optimization for SDN-enabled UDN," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6434–6447, 2022.

[17] K. Sun, Q. Han, Z. Yang, W. Huang, H. Zhang, and V. C. Leung, "Proactive handover type prediction and parameter optimization based on machine learning," *IEEE Trans. Wireless Commun.*, vol. 24, no. 4, pp. 3515–3528, 2025.

[18] S. Kim, J. Moon, J. Kim, Y. Ahn, D. Kim, S. Kim, K. Shim, and B. Shim, "Role of sensing and computer vision in 6G wireless communications," *IEEE Wireless Commun.*, vol. 31, no. 5, pp. 264–271, 2024.

[19] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided 6G wireless communications: Blockage prediction and proactive handoff," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10 193–10 208, 2021.

[20] U. Demirhan and A. Alkhateeb, "Radar aided proactive blockage prediction in real-world millimeter wave systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2022, pp. 4547–4552.

[21] Y. Liu, J. Wu, S. Kim, and B. Shim, "Vision-aided blockage prediction and proactive handover for indoor mmWave and terahertz communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2023, pp. 7411–7416.

[22] C. Chaccour, W. Saad, M. Debbah, and H. V. Poor, "Joint sensing, communication, and AI: A trifecta for resilient THz user experiences," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11 444–11 460, 2024.

[23] Y. Feng, C. Zhao, H. Luo, F. Gao, F. Liu, and S. Jin, "Networked ISAC based UAV tracking and handover towards low-altitude economy," *IEEE Trans. Wireless Commun.*, vol. 24, no. 9, pp. 7670–7685, 2025.

[24] C. Shen, C. Tekin, and M. van der Schaar, "A non-stochastic learning approach to energy efficient mobility management," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3854–3868, 2016.

[25] A. F. Molisch, M. Steinbauer, M. Toeltsch, E. Bonek, and R. S. Thoma, "Capacity of MIMO systems based on measured wireless channels," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 561–569, 2002.

[26] J. Tang, F. Tang, S. Long, M. Zhao, and N. Kato, "Utilizing large language models for advanced optimization and intelligent management in space-air-ground integrated networks," *IEEE Netw.*, vol. 39, no. 5, pp. 173–181, 2024.

[27] D. Solomitckii, Q. C. Li, T. Balercia, C. R. Da Silva, S. Talwar, S. Andreev, and Y. Koucheryavy, "Characterizing the impact of diffuse scattering in urban millimeter-wave deployments," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 432–435, 2016.

[28] S. Kim, S. Jeong, J. Wu, B. Shim, and M. Z. Win, "Large multimodal model-based environment-aware channel estimation," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 12, pp. 4059–4075, 2025.

[29] H. Ju, S. Jeong, S. Kim, B. Lee, and B. Shim, "Transformer-assisted parametric CSI feedback for mmWave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18 774–18 787, 2024.

[30] *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP)*, 3rd Generation Partnership Project 3GPP™ TS 36.423 V18.4.0, Mar. 2025, Release 18.

[31] X. Hu, Y. Huo, X. Dong, F.-Y. Wu, and A. Huang, "Channel prediction using adaptive bidirectional GRU for underwater MIMO communications," *IEEE Internet Things J.*, vol. 11, no. 2, pp. 3250–3263, 2023.

[32] Q. Liu, G. Chuai, J. Wang, and J. Pan, "Proactive mobility management with trajectory prediction based on virtual cells in ultra-dense networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8832–8842, 2020.

[33] L. Italiano, B. C. Tedeschini, M. Brambilla, H. Huang, M. Nicoli, and H. Wymeersch, "A tutorial on 5G positioning," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 3, pp. 1488–1535, 2024.

[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, 2024.

[35] Y. Zeng, J. Chen, J. Xu, D. Wu, X. Xu, S. Jin, X. Gao, D. Gesbert, S. Cui, and R. Zhang, "A tutorial on environment-aware communications via channel knowledge map for 6G," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 1478–1519, 2024.

[36] 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. Veh.*, vol. 10, no. 2, pp. 794–807, 2025.

[37] I. Keum, J. Son, H. Kim, J. Moon, and B. Shim, "Deep learning-based NLoS localization using geometric map image," *IEEE Trans. Veh. Technol.*, 2025.

[38] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2023.

[39] S. Kim, S. Saha, S. Jeong, B. Shim, and M. Z. Win, "Large multimodal model-based environment-aware beam management," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 991–1007, 2026.

[40] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[41] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5014–5029, 2016.

[42] P. W. Patil, S. Gupta, S. Rana, S. Venkatesh, and S. Murala, "Multi-weather image restoration via domain translation," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 21 696–21 705.

[43] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, 2008.

[44] *SMPTE Engineering Guideline - SD-SDI and HD-SDI Standards Roadmap*, IEEE Standard EG 2111-1:2021, 2021.

[45] G. R. MacCartney, S. Deng, S. Sun, and T. S. Rappaport, "Millimeter-wave human blockage at 73 ghz with a simple double knife-edge diffraction model and extension for directional antennas," in *IEEE Veh. Technol. Conf. (VTC-Fall)*, 2016.

[46] J. Hoydis, F. Aït Aoudia, S. Cammerer, M. Nimier-David, N. Binder, G. Marcus, and A. Keller, "Sionna RT: Differentiable ray tracing for radio propagation modeling," in *Proc. IEEE Global Commun. Conf. Workshops (GC Wkshps)*, 2023, pp. 317–321.

[47] *Study on channel model for frequencies from 0.5 to 100 GHz*, 3rd Generation Partnership Project 3GPP™ TR 38.901 V18.0.0, Sep. 2020, Release 18.

[48] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.

[49] Y. Zeng and X. Xu, "Toward environment-aware 6G communications via channel knowledge map," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 84–91, 2021.

[50] X. Chu *et al.*, "Mobilevlm: A fast, strong and open vision language assistant for mobile devices," *arXiv preprint arXiv:2312.16886*, 2023.

[51] B. Zhou *et al.*, "Tinyllava: A framework of small-scale large multimodal models," *arXiv preprint arXiv:2402.14289*, 2024.

[52] T. Nishio *et al.*, "Proactive received power prediction using machine learning and depth images for mmWave networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2413–2427, 2019.

[53] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, 2020.

[54] J. Chen, L. Ye, J. He, Z.-Y. Wang, D. Khashabi, and A. Yuille, "Efficient large multi-modal models via visual context compression," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024, pp. 73 986–74 007.

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 artificial intelligence for wireless communications.

Sangmok Shin (Student Member, IEEE) received the B.S. degree from Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea, in 2023. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering at Seoul National University, Seoul, South Korea. His research interests include AI-assisted physical-layer wireless communications, with particular emphasis on beam management, channel estimation, and channel prediction. Mr. Shin was a recipient of the Samsung Humantech Paper Award Bronze Prize and the Graduate Student of the Year Award from the Department of Electrical and Computer Engineering in 2025.

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

He is an Assistant Professor at the Singapore University of Technology and Design (SUTD). Prior to joining SUTD, he was 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.

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, deep 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, the JCN Best Paper Award in 2024, and the SNU Academic Research Award in 2025. Dr. Shim was an Elected Member of the Signal Processing for Communications and Networking (SPCOM) Technical Committee of the IEEE Signal Processing Society. He has 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).