

Computation Efficient Beam Management via Early-Exit Large Multimodal Models

Junhong Lee, Seonghoon Kim, Hokyung Kim, and Byonghyo Shim

Institute of New Media and Communications, Department of Electrical and Computer Engineering,
Seoul National University, Republic of Korea

Email: {jhlee, shkim, hkim}@islab.snu.ac.kr, bshim@snu.ac.kr

Abstract—Future 6G wireless systems are expected to exploit the upper mid-band with large-scale MIMO to support high-capacity and reliable communication. To fully leverage the advantages of upper mid-band systems, timely and accurate beam management is essential to maintain robust links under user mobility and dynamic blockage. Conventional RF-based beam management techniques mainly rely on instantaneous channel observations. As a result, they have limited capability to anticipate abrupt channel variations caused by environmental changes. Large multimodal models (LMMs) provide a promising solution by jointly interpreting RF measurements and contextual information from visual or spatial observations, enabling environment-aware beam tracking. However, performing full LMM inference for every beam update introduces non-negligible processing latency, which can make beam decisions outdated in rapidly varying channels. To address this issue, we propose an early-exit LMM-based beam management framework that adaptively terminates inference at intermediate layers using prediction heads and confidence-aware exit decisions. By avoiding unnecessary computation for easy beam management instances, the proposed framework reduces beam decision latency while preserving the context-aware reasoning capability of LMMs. Numerical results show that the proposed scheme improves beam tracking efficiency and achieves a 28% improvement in data rate over 5G NR beam management at an SNR of 10 dB.

I. INTRODUCTION

To support emerging data-demanding and delay-sensitive applications such as physical AI and autonomous driving, millimeter wave (mmWave) bands have been adopted in 5G to provide wide bandwidth. However, due to severe path loss and limited coverage, the practical deployment of mmWave systems remains challenging [1]. Recently, the upper mid-band, also referred to as frequency range 3 (FR3), has attracted significant attention as a promising spectrum for future wireless networks since it can strike a balance between bandwidth and coverage. In the upper mid-band, improved propagation characteristics make non line-of-sight (NLoS) components more critical, leading to richer multipath propagation. Therefore, beamforming in upper-mid-band systems should consider propagation loss and the time-varying nature of multipath channels.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022M3C1A3099336), and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Semiconductor Support Program to Nurture the Best Talents (IITP-2023-RS-2023-00256081).

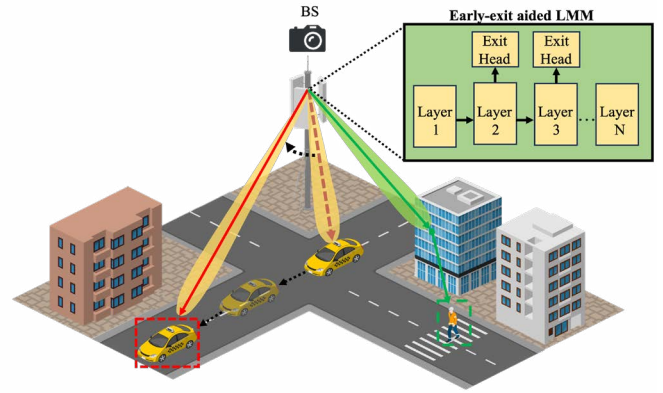


Fig. 1: Conceptual diagram of the early-exit LMM-aided beam management.

Large-scale multiple-input multiple-output (MIMO) beamforming is a key technology for improving network capacity and compensating for propagation loss. Beamforming concentrates the signal energy from multiple antennas toward the desired propagation directions, thereby improving the signal-to-noise ratio (SNR), data rate, and communication reliability. Since these benefits strongly depend on accurate alignment between directional beams and signal propagation paths, beam management is of great importance. However, in FR3 environments, the optimal beam direction varies rapidly due to user mobility, multipath propagation, and dynamic obstacles. Therefore, beam management should promptly and accurately adapt to channel variations to sustain reliable directional links in upper mid-band MIMO systems.

Conventional RF-based beam management methods rely on pilot signals and previous beam indices to estimate the optimal beam direction [2]. Although these methods directly reflect instantaneous channel conditions, the acquired channel information may quickly become outdated due to environmental dynamics. Consequently, their ability to understand the surrounding propagation environment remains limited. As a result, RF-based methods may suffer from delayed beam recovery and increased training overhead under discontinuous channel variations in dynamic environments [3].

In upper mid-band systems, the propagation environment is more complex than that in conventional sub-6 GHz systems, while NLoS and reflected paths become more significant than

in highly directional mmWave systems. Accordingly, beam management should not be limited to tracking a dominant propagation path but should accurately capture multiple propagation components induced by user mobility, blockage, and surrounding scatterers. This necessitates environment-aware channel estimation that can exploit both RF measurements and contextual information from the surrounding environment. To address this, large multimodal models (LMMs) have emerged as a promising solution for context-aware wireless communication by jointly processing RF measurements and semantic environmental information [4]–[6]. By exploiting multimodal context, LMM-aided beam management can infer potential blockages and mobility patterns that are difficult to capture using traditional RF-based beam management [7]. However, directly applying full-depth LMM inference to beam management introduces substantial processing latency, leading to stale beam decisions.

In this paper, we propose a latency-efficient LMM-aided beam management technique. We adopt an early-exit strategy [8] by adding one or more exit branches, which enable inference to exit at an intermediate layer when sufficient confidence is reached. The proposed technique, termed *early-exit LMM-aided beam management* (EE-LMMBM), supports latency-efficient beam management by performing early exits for high-confidence beam predictions while using deeper layers only for uncertain cases. Numerical results show that EE-LMMBM improves the data rate by 28% over 5G NR beam management at an SNR of 10 dB and achieves 7.5 bps/Hz at a user speed of 90 km/h, compared with 3.3 bps/Hz for full-depth LMM-based beam management.

II. SYSTEM MODEL

A. UM-MISO Communication System Model

We consider a downlink UM-MISO communication system, where a base station (BS) is equipped with $N = N_x \times N_y$ uniform planar array (UPA) antennas. At time step t , the BS transmits a symbol $s^{(t)}$ to the mobile user using a transmit beamforming vector $\mathbf{f}^{(t)} \in \mathbb{C}^{N \times 1}$. The received signal at the user at time step t is given by

$$y^{(t)} = \left(\mathbf{h}^{(t)} \right)^H \mathbf{f}^{(t)} s^{(t)} + n_k^{(t)}, \quad (1)$$

where $\mathbf{h}^{(t)} \in \mathbb{C}^N$ denotes the downlink channel vector from the BS to the mobile user, $\mathbf{f}^{(t)} \in \mathbb{C}^N$ is the transmit beamforming vector for the user, and $n_k^{(t)} \sim \mathcal{CN}(0, \sigma^2)$ is additive white Gaussian noise. Then, the signal-to-noise ratio (SNR) of the user at time step t is given by $\gamma^{(t)} = \frac{|\mathbf{h}^{(t)}|^H \mathbf{f}^{(t)}|^2}{\sigma^2}$. Then, the achievable data rate of the k -th mobile at time step t is given by

$$R^{(t)} = \log_2 \left(1 + \gamma^{(t)} \right). \quad (2)$$

To capture the directional nature of the high carrier frequency wireless channel, we consider a geometric channel model. The channel vector at time step t is represented as

$$\mathbf{h}^{(t)} = \sum_{\ell=1}^{L^{(t)}} \beta_{\ell}^{(t)} e^{j\alpha_{\ell}^{(t)}} \mathbf{a}(\theta_{\ell}^{(t)}, \phi_{\ell}^{(t)}) \in \mathbb{C}^{N_{\text{tx}}}, \quad (3)$$

where $L^{(t)}$ is the number of propagation paths, $\beta_{\ell}^{(t)}$ and $\alpha_{\ell}^{(t)}$ are the complex path gain and phase shift of the ℓ -th path, respectively. $\theta_{\ell}^{(t)} \in [-\pi, \pi)$ and $\phi_{\ell}^{(t)} \in [0, \pi]$ are the azimuth angle of departure (AoD) and zenith angle of departure (ZoD), respectively. Also, $\mathbf{a}(\theta, \phi) = \mathbf{a}_x(\theta, \phi) \otimes \mathbf{a}_y(\theta, \phi)$ is the array steering vector of the UPA, where $\mathbf{a}_x(\theta, \phi)$, $\mathbf{a}_y(\theta, \phi)$ are the steering vectors for the horizontal and vertical dimensions of the UPA, given by

$$\mathbf{a}_x(\theta, \phi) = \frac{1}{\sqrt{N_x}} \left[1, e^{j\pi \sin \theta \cos \phi}, \dots, e^{j\pi (N_x-1) \sin \theta \cos \phi} \right]^T \quad (4)$$

$$\mathbf{a}_y(\theta, \phi) = \frac{1}{\sqrt{N_y}} \left[1, e^{j\pi \sin \theta \sin \phi}, \dots, e^{j\pi (N_y-1) \sin \theta \sin \phi} \right]^T \quad (5)$$

assuming a half-wavelength spacing between antenna element.

B. Beam Management for MISO Systems

In this subsection, we discuss the basic beam management process in the considered ultra-massive multiple-input single-output (UM-MISO) system. The objective of beam management is to properly align the beamforming vector $\mathbf{f}$ with the downlink channel vector $\mathbf{h}$. At time slot t , the BS acquires RF pilot-based information $\mathbf{R}^{(t)}$, which can be obtained from SRS or CSI-RS. Based on $\mathbf{R}^{(t)}$, the BS can estimate downlink channel vector as $\hat{\mathbf{h}}^{(t)}$ using various channel estimation schemes. Given a beam codebook $\mathcal{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_I\}$, the BS selects the beam that maximizes the estimated beamforming gain as

$$\hat{i}^{(t)} = \arg \max_{i \in \{1, \dots, I\}} \left| \mathbf{f}_i^H \hat{\mathbf{h}}^{(t)} \right|^2. \quad (6)$$

The selected beamforming vector is then given by $\mathbf{f}^{(t)} = \mathbf{f}_{\hat{i}^{(t)}}$.

However, in UM-MISO systems, channel estimation and beam management can become computationally demanding due to the large antenna dimension. As a result, the beam management latency, denoted by t_{BM} , may become non-negligible. During this latency, the wireless channel can vary due to user mobility and environmental dynamics. Consequently, the beam selected based on $\hat{\mathbf{h}}^{(t)}$ is actually applied to the aged channel $\mathbf{h}^{(t+t_{\text{BM}})}$, and the effective beamforming gain becomes

$$G_{\text{eff}}^{(t, t_{\text{BM}})} = \left| \left(\mathbf{h}^{(t+t_{\text{BM}})} \right)^H \mathbf{f}_{\hat{i}^{(t)}} \right|^2. \quad (7)$$

Advanced channel estimators, such as LMM-based channel estimation, can reduce the instantaneous estimation error $\|\mathbf{h}^{(t)} - \hat{\mathbf{h}}^{(t)}\|$ by exploiting contextual multimodal information. However, such estimators may also increase t_{BM} , resulting in a larger channel aging error $\|\mathbf{h}^{(t+t_{\text{BM}})} - \mathbf{h}^{(t)}\|$. Therefore, practical beam management should account for the trade-off between channel estimation accuracy and channel variation induced by beam management latency.

III. EARLY-EXIT LMM-AIDED BEAM MANAGEMENT

In this section, we present the proposed early-exit LMM-aided beam management framework. The key idea is to exploit multimodal context for environment-aware beam prediction while reducing the inference latency of the LMM using early exits.

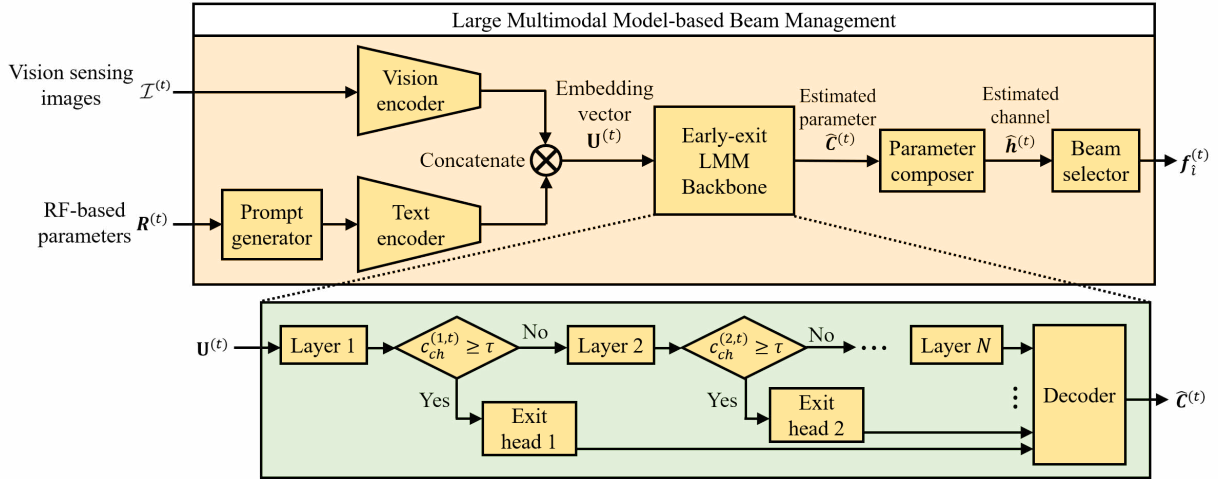


Fig. 2: Illustration of the proposed EE-BM.

A. LMM-aided Channel Estimation

Conventional RF pilot-based beam management techniques mainly rely on historical channel measurements to estimate the beam direction, and this fails to capture abrupt channel variations caused by user mobility and the surrounding environments. To overcome this limitation, LMM-aided beam management has emerged as a promising approach by incorporating multimodal sensing information into the beam decision process. In this context, LMM exploits RF measurements and sensing information to obtain a semantic understanding of the propagation environment. Specifically, the LMM first detects mobile users from the visual observation and extracts contextual information about the surrounding objects, such as their positions, motions, and reflective surfaces. In particular, even in NLoS or partially blocked scenarios, the LMM can infer signal reflection paths by understanding the spatial relationships among the BS, mobile users, and their surroundings, thereby achieving a higher beamforming gain and more reliable beam alignment.

To formally describe the multimodal encoding process, let the input to the LMM at time t be defined as

$$\mathcal{X}^{(t)} = \{\mathbf{I}^{(t)}, \mathbf{R}^{(t)}\},$$

where $\mathbf{I}^{(t)}$ denotes the visual sensing information and $\mathbf{R}^{(t)}$ denotes the RF-based estimated channel parameters.

The visual sensing information is encoded by the vision encoder and projected into the LMM embedding space as

$$\mathbf{E}_{\text{vision}}^{(t)} = \mathbf{W}_{\text{vision}} f_{\text{vision}}(\mathbf{I}^{(t)}),$$

where $f_{\text{vision}}(\cdot)$ is the vision encoder and $\mathbf{W}_{\text{vision}}$ is the visual projection layer. The RF-based estimated channel parameters are converted into a structured textual representation and embedded by the language embedding layer as

$$\mathbf{E}_{\text{text}}^{(t)} = f_{\text{text}}(\mathbf{R}^{(t)}),$$

where $f_{\text{text}}(\cdot)$ denotes the text embedding layer of the LMM.

The multimodal token representation is then constructed by concatenating the visual and textual embeddings:

$$\mathbf{U}^{(t)} = \text{concat}(\mathbf{E}_{\text{vision}}^{(t)}, \mathbf{E}_{\text{text}}^{(t)}).$$

The Transformer backbone of the LMM consists of N sequential Transformer layers and processes the unified token representation as

$$\mathbf{Z}^{(t)} = \mathcal{B}_{\text{LMM}}(\mathbf{U}^{(t)}) = \mathcal{B}_N \circ \mathcal{B}_{N-1} \circ \dots \circ \mathcal{B}_1(\mathbf{U}^{(t)}),$$

where $\mathcal{B}_{\text{LMM}}(\cdot)$ denotes the Transformer backbone of the LMM and $\mathcal{B}_n(\cdot)$ denotes the n -th Transformer layer. Based on the hidden representation $\mathbf{Z}^{(t)}$, the LMM decoder $\mathcal{D}_{\text{LMM}}(\cdot)$ generates the estimated channel parameter matrix $\hat{\mathbf{C}}^{(t)}$ as

$$\hat{\mathbf{C}}^{(t)} = \mathcal{D}_{\text{LMM}}(\mathbf{Z}^{(t)}).$$

The estimated channel parameter matrix is $\hat{\mathbf{C}}^{(t)} = [\hat{\beta}^{(t)}, \hat{\alpha}^{(t)}, \hat{\phi}^{(t)}, \hat{\theta}^{(t)}]^T$. Based on the parameter composer, the estimated channel is given by

$$\hat{\mathbf{h}}^{(t)} = \sum_{\ell=1}^{\hat{L}} \hat{\beta}_{k,\ell}^{(t)} e^{j\hat{\alpha}_{k,\ell}^{(t)}} \mathbf{a}(\hat{\phi}_{k,\ell}^{(t)}, \hat{\theta}_{k,\ell}^{(t)}) \in \mathbb{C}^N, \quad (8)$$

where $\hat{L}$ denotes the estimated number of propagation paths. Finally, based on $\hat{\mathbf{h}}^{(t)}$, the beam selector determines the beam index $\hat{i}^{(t)}$ that maximizes the beamforming gain, as defined in (6), and selects the corresponding beamforming vector $\mathbf{f}_i^{(t)}$.

B. Exit head training

LMM-based beam management introduces non-negligible inference latency, leading to outdated beam decisions. To reduce the inference latency, we introduce intermediate exit heads into the LMM backbone. The key idea is to allow the model to produce an early channel reconstruction before completing the full-depth LMM inference when the intermediate representation is already sufficiently informative. The newly attached exit head is randomly initialized, so its output token

does not reflect the semantic reliability of the intermediate LMM representation. Therefore, we train lightweight exit heads while keeping the LMM backbone frozen. During exit-head training, multimodal inputs, including RF measurements and visual sensing information, are first encoded by the frozen LMM backbone into intermediate hidden representations. Then, a lightweight MLP head $g_{\text{exit}}^{(\ell)}(\cdot)$ at layer ℓ maps the intermediate hidden state to an estimated channel parameter matrix, which is used to estimate the channel state.

$$\hat{\mathbf{C}}_{\text{exit}}^{(\ell,t)} = \mathcal{D}_{\text{LMM}} \left(g_{\text{exit}}^{(\ell)} \left(\mathbf{Z}^{(\ell,t)} \right) \right)$$

The training objective $\mathcal{L}_{\text{exit}}$ is defined to minimize the prediction error between the intermediate channel parameter matrix and the ground-truth channel parameter matrix. By optimizing only the lightweight exit heads, the proposed training procedure enables reliable intermediate channel estimation without the high cost of full LMM fine-tuning. Specifically, for the exit head attached to the ℓ -th LMM layer, the exit loss is defined as

$$\mathcal{L}_{\text{exit}}^{(\ell)} = \left\| \hat{\mathbf{C}}_{\text{exit}}^{(\ell,t)} - \mathbf{C}^{(t)} \right\|_F^2,$$

where $\hat{\mathbf{C}}_{\text{exit}}^{(\ell,t)}$ is the channel parameter matrix estimated by the ℓ -th exit head, and $\mathbf{C}^{(t)}$ is the ground-truth channel parameter matrix at time t . When multiple exit heads are attached to different LMM layers, the total exit-head training loss is given by

$$\mathcal{L}_{\text{exit}} = \sum_{\ell \in \mathcal{E}} \lambda_{\ell} \mathcal{L}_{\text{exit}}^{(\ell)},$$

where $\mathcal{E}$ denotes the set of exit layers and λ_{ℓ} is the loss weight for the ℓ -th exit head.

C. Confidence-based Early Exit Mechanism

After training the exit heads, the LMM performs inference in a layer-wise manner. At each exit layer, the corresponding exit head estimates the channel parameter matrix from the intermediate hidden representation. The ℓ -th exit head produces the intermediate channel parameter estimate and its reliability logit as

$$\left[\hat{\mathbf{C}}_{\text{exit}}^{(\ell,t)}, s_{\text{exit}}^{(\ell,t)} \right] = \mathcal{D}_{\text{LMM}} \left(g_{\text{exit}}^{(\ell)} \left(\mathbf{Z}^{(\ell,t)} \right) \right), \quad (9)$$

where $\hat{\mathbf{C}}_{\text{exit}}^{(\ell,t)}$ denotes the channel parameter matrix estimated by the ℓ -th exit head, and $s_{\text{exit}}^{(\ell,t)}$ is a scalar reliability logit for the intermediate channel estimation.

To obtain a normalized channel-level confidence score, the reliability logit is converted by the sigmoid function as

$$c_{\text{ch}}^{(\ell,t)} = \sigma \left(s_{\text{exit}}^{(\ell,t)} \right), \quad (10)$$

where $\sigma(\cdot)$ denotes the sigmoid function. The LMM exits at the first layer ℓ^* whose channel-level confidence exceeds a predefined threshold:

$$\ell^* = \min \left\{ \ell \in \mathcal{E} \mid c_{\text{ch}}^{(\ell,t)} \geq \tau \right\}, \quad (11)$$

where $\mathcal{E}$ is the set of layers equipped with exit heads, and τ is the confidence threshold.

If the early-exit condition is satisfied, the intermediate channel parameter matrix $\hat{\mathbf{C}}_{\text{exit}}^{(\ell^*,t)}$ is directly used for channel reconstruction and beam management, and the remaining LMM layers are skipped. Otherwise, the model continues forward propagation until the next exit layer or the final layer. When no intermediate exit satisfies the confidence condition, the final-layer output is used for channel estimation.

The threshold τ controls the tradeoff between inference latency and channel estimation reliability. A smaller value of τ allows more samples to exit early and reduces inference latency, while a larger value makes the model more conservative and improves the reliability of the estimated channel. In practice, τ is selected as a validation hyperparameter by sweeping candidate values and choosing the operating point that provides a desirable latency–performance trade-off.

However, the confidence score obtained from the reliability logit can be overconfident and may not accurately reflect the true reliability of the intermediate channel estimation. To alleviate this issue, we calibrate the confidence of each exit head using temperature scaling on a validation set [9]. The calibrated channel-level confidence is computed as

$$\tilde{c}_{\text{ch}}^{(\ell,t)} = \sigma \left(\frac{s_{\text{exit}}^{(\ell,t)}}{T_{\ell}} \right), \quad (12)$$

where T_{ℓ} is the temperature parameter optimized on a validation set while keeping the trained exit head fixed. The final early-exit decision is made using the calibrated confidence score:

$$\tilde{c}_{\text{ch}}^{(\ell,t)} \geq \tau. \quad (13)$$

This calibration step reduces unreliable overconfident predictions and prevents premature exits.

IV. NUMERICAL RESULTS

A. Simulation Setup

We consider a UM-MISO system in which the BS, equipped with $N = 8 \times 8$ UPA antennas, serves a single-antenna mobile user at a carrier frequency of 10GHz . To construct the dataset for training the LMM and the exit head, we utilize CARLA, which provides a 3D simulation environment with realistic visual observations. In the CARLA simulation, we assume that the BS is located at the center of the simulated area, and the RGB cameras mounted at the BS continuously capture the surrounding environment. The wireless channel is generated by performing ray tracing in MATLAB based on the CARLA simulation geometry. For the pretrained LMM, we use LLaVA-7B [10]. To emulate a practical beam management processing time, the server-side computation is implemented on an NVIDIA L40S GPU.

B. Simulation Results

In Fig. 3, we plot the data rate as a function of the transmit SNR. We observe that EE-LMMBM achieves a higher data rate than the conventional beam management scheme [11], while maintaining performance comparable to LMM-based beam management. This improvement stems from the ability

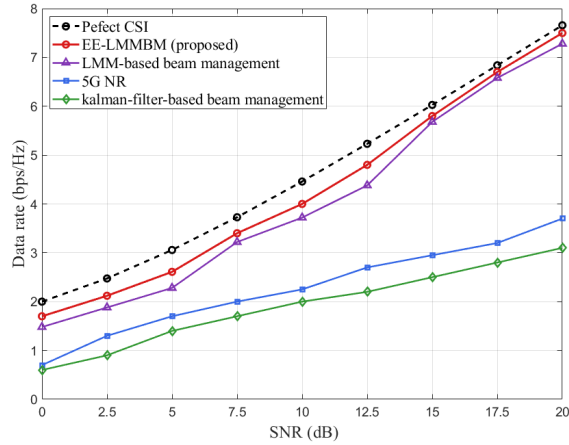


Fig. 3: Data rate vs. transmit SNR

of EE-LMMBM to exploit multimodal information collected at the BS. By using visual and RF-domain observations, EE-LMMBM can accurately infer complex NLoS channel conditions and select an appropriate beam with reduced computational overhead and beam management latency. At an SNR of 10 dB, EE-LMMBM improves the data rate by 28% compared with 5G NR beam management.

In Fig. 4, we evaluate the data rate with respect to the average user speed. At low mobility, both LMM-based methods achieve relatively high data rates since the channel variation is moderate and the beam decision remains valid during the inference process. However, as the user speed increases, the performance gap between the proposed EE-LMMBM and conventional LMM-based beam management becomes significant. For example, at 90 km/h, the proposed EE-LMMBM achieves about 7.5 bps/Hz, whereas conventional LMM-based beam management drops to about 3.3 bps/Hz. This represents a data rate improvement over full-depth LMM-based beam management. The main reason is that full-depth LMM inference incurs non-negligible processing latency, which can make the predicted beam direction outdated in high-mobility scenarios. This result demonstrates that the proposed EE-LMMBM can update the beam direction in a more timely manner, effectively mitigating latency-induced beam misalignment.

V. CONCLUSION

In this paper, we proposed an early-exit-aided LMM-based beam management method for UM-MISO systems. While conventional RF-pilot-based beam management depends on instantaneous channel measurements, the proposed framework leverages contextual information obtained from the wireless environment to perform environment-aware beam management. Specifically, the LMM captures visual and semantic cues related to UE movement, blockage, and reflections. This enables more reliable beam tracking under dynamic channel conditions. We introduced an early-exit mechanism that adaptively terminates LMM inference when the intermediate beam prediction is sufficiently confident. This reduces

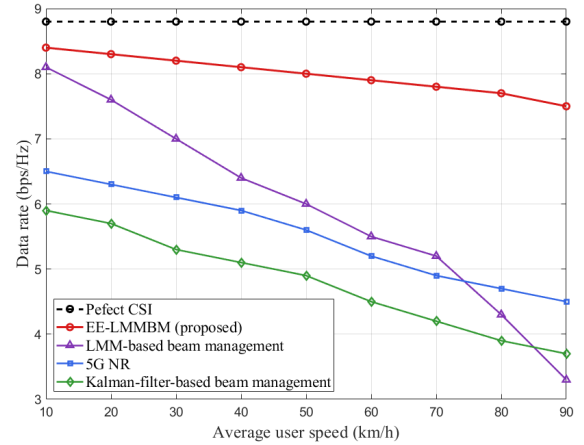


Fig. 4: Data rate vs. average user speed

inference latency and computational cost, thereby alleviating stale-beam-induced misalignment. Numerical results verify that the proposed method improves the effective sum rate and reduces the beam misalignment probability while requiring lower inference latency than full LMM inference, showing its potential for efficient and reliable beam management in future UM-MISO systems.

REFERENCES

- [1] H. J. Yang *et al.*, "Advancing multi-robot networks via mllm-driven sensing, communication, and computation: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 5833–5871, Apr. 2026.
- [2] G. Liu, A. Liu, R. Zhang, and M.-J. Zhao, "Angular-domain selective channel tracking and doppler compensation for high-mobility mmwave massive mimo," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2902–2916, May 2021.
- [3] H. Lee *et al.*, "Towards 6G hyper-connectivity: Vision, challenges, and key enabling technologies," *J. Commun. Netw.*, vol. 25, no. 3, pp. 344–354, Jun. 2023.
- [4] S. Kim, S. Saha, S. Jeong, B. Shim, and M. Z. Win, "Large multimodal model-based environment-aware beam management," *IEEE Journal on Selected Areas in Communications*, vol. 44, pp. 991–1007, 2026.
- [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.*, 2025, early access.
- [6] H. Noh, B. Shim, and H. J. Yang, "Adaptive resource allocation optimization using large language models in dynamic wireless environments," *IEEE Trans. Veh. Technol.*, vol. 74, no. 10, pp. 16 630–16 635, Oct. 2025.
- [7] J. Moon, J. Jung, and B. Shim, "Large multimodal model-aided environment-aware beam tracking for upper mid-band," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2025, pp. 1137–1142.
- [8] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *2016 23rd international conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 2464–2469.
- [9] Z. Chen, W. Hu, G. He, Z. Deng, Z. Zhang, and R. Hong, "Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models," in *Proc. 31st Int. Conf. Comput. Linguistics (COLING)*, 2025, pp. 3095–3109.
- [10] F. Li *et al.*, "LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3D in large multimodal models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.
- [11] H. Kim, S. Kim, H. Lee, C. Jang, Y. Choi, and J. Choi, "Massive mimo channel prediction: Kalman filtering vs. machine learning," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 518–528, Jan. 2021.