面向6G自主通信的大型多模态模型辅助调度

Sunwoo Kim，IEEE会员；Byonghyo Shim，IEEE会士。首尔国立大学新媒体与通信研究所、电气与计算机工程系，韩国首尔。电子邮箱：swkim@islab.snu.ac.kr，bshim@snu.ac.kr

arXiv:2601.06211v1 [cs.IT] 8 Jan 2026

摘要——近年来，大语言模型（LLM）因能够针对给定查询快速生成准确答案而备受关注。这些模型已发展为大型多模态模型（LMM），能够理解和分析图像、文本等多模态输入。随着自主设备中的人工智能功能呈指数级增长，执行人工智能推理的数字处理单元——中央单元（CU）需要运行大型多模态模型，以有效控制这些设备。为确保命令顺利传递至设备，中央单元必须执行调度，根据信道条件分配资源块（RB）以进行数据传输，并选择调制与编码方案（MCS）索引。在6G的许多实际环境中，这项任务颇具挑战，因为用户即使稍有移动，也可能导致信道骤变。本文提出一种新颖的基于大型多模态模型的调度技术，以应对这一挑战。我们的核心思路是利用大型多模态模型分析视觉感知信息和导频信号，预测未来信道参数（例如距离、角度和路径增益）。通过利用大型多模态模型根据视觉感知信息预测可靠路径是否存在以及用户的几何信息，再结合从导频信号获得的历史信道状态，我们能够准确预测未来信道参数。利用这些预测结果，我们可以抢先做出信道感知调度决策。数值评估表明，与传统调度技术相比，所提出的技术可实现超过30%的吞吐量增益。

一、引言

近年来，生成式人工智能凭借其生成高质量内容的能力受到广泛关注，而这些内容甚至是人类也难以轻易生成的 [2]。在生成式人工智能的各种形式中，大语言模型（LLM）因其无需额外示例或提示即可回答问题（即零样本提示），并且仅凭少量输入—输出对即可解决问题（即少样本提示）的能力而备受关注 [3]。大语言模型最初主要专注于语言数据，但近年来已发展为大型多模态模型（LMM），能够处理和解读多种图像、视频和音频等模态。凭借灵活性和适应性，大型多模态模型如今已积极应用于自动化应用，例如无人驾驶车辆 [4],人形机器人 [5], 以及外科手术/农业制造 [6]。

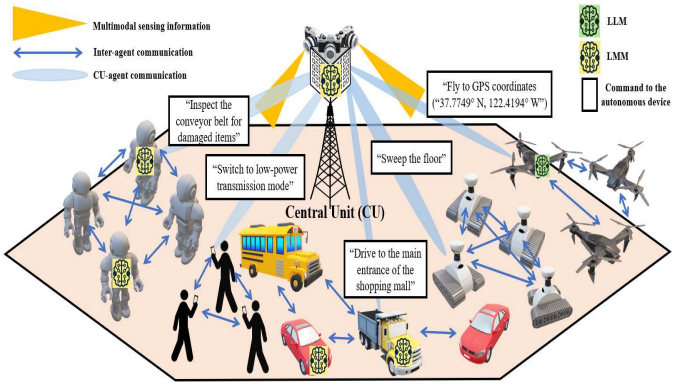


图1. 基于大语言模型/大型多模态模型的任务导向通信系统。

尽管人们已尝试将大语言模型/大型多模态模型用于自主任务（例如机器人、人形机器人和自动驾驶车辆），但其在未来无线通信系统中的应用仍相对有限 [7]。这主要是因为人工智能操作通常局限于单个智能体（例如人形机器人），因此无需与其他智能体及中央单元（CU）（例如云中心或数字单元）交换信息，或者至少这种交换并不那么重要。然而，未来随着自主设备执行的任务范围呈指数级增长，更多智能体将开展协作，使智能体间通信以及中央单元与智能体之间的通信都变得至关重要 [8]。在这类系统中，无线通信将在智能体之间交换状态信息以提供相互反馈方面发挥核心作用；同时，还将传输多模态感知信息（例如激光雷达和射频（RF）信号），以便利用中央单元中的大型多模态模型控制这些智能体（见图1）。

为了确保自主设备能够执行指定操作，例如汽车组装或远程手术，中央单元需要执行调度任务，即根据信道条件为数据传输分配资源块（RB），并选择适当的链路参数（即调制与编码方案（MCS））。例如，通过选择符合信道质量的调制与编码方案索引，中央单元可以满足符号错误率（SER）要求，同时确保向设备无缝传递命令。然而，在即将到来的6G环境中，调度将

本文的部分内容已发表于 IEEE 车辆技术会议（VTC），中国成都，2025 [1]年10月。本研究部分由韩国政府（科学技术信息通信部）资助的人工智能半导体研究生院（IITP-2025-RS-2023-00256081）项目提供支持；部分由韩国政府（科学技术信息通信部）资助的韩国国家研究基金会（NRF）项目（2022M3C1A3099336）提供支持；另有部分由韩国科学技术信息通信部通过 IITP（信息技术规划与评估研究院）监督的 ITRC（信息技术研究中心）支持项目（IITP-2023-2021-0-02048）提供支持。

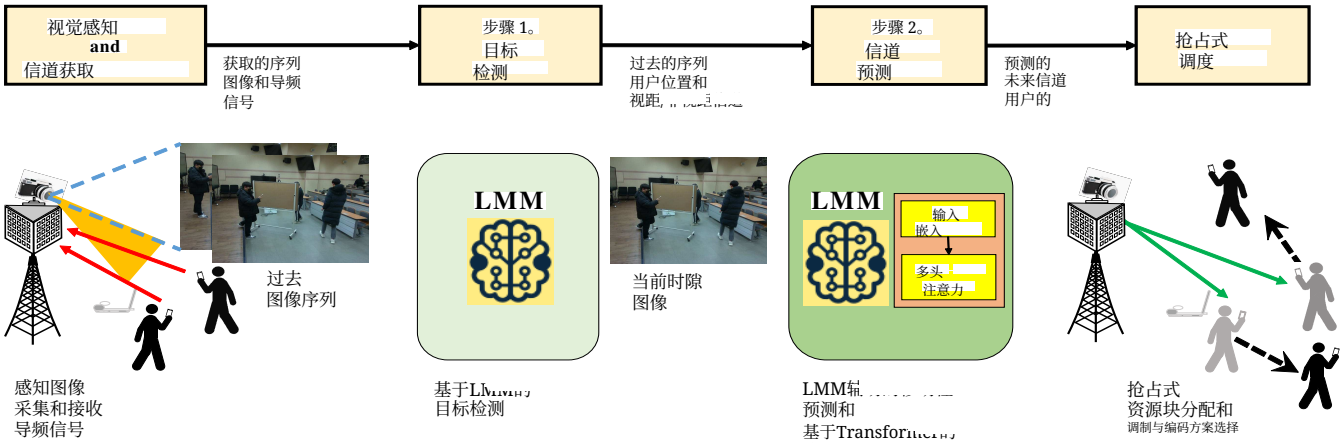


图2. 所提出的 LMM-PS 的概念示意图。

由于密集分布的障碍物、多种频谱以及自主设备的移动所导致的信道状态突然变化[1]，调度绝非易事。在这种情况下，依赖过时信道信息的传统调度技术可能无法满足用户的服务质量（QoS）要求，从而导致操作失败。

为了解决这一问题，可以考虑一种抢占式调度策略，根据预测的未来信道条件 [10]–[12] 做出调度决策。该策略的核心思想是捕捉历史信道条件之间的时间关系，并利用这一关系预测未来信道条件。该策略的有效性与信道预测质量密切相关，但仅依赖导频/控制信令和过去信道估计的传统抢占式调度机制，无法妥善处理由未来突发事件（例如阻塞或运动方向变化）引起的意外变化。

本文旨在提出一种利用大型多模态模型提升抢占式调度质量的新型调度技术。在所提出的技术中，我们将其称为基于大型多模态模型的抢占式调度（LMM-PS），并积极利用大型多模态模型预测用户未来的信道参数（例如距离、角度和路径增益）。与仅依赖导频信号的传统调度技术不同，LMM-PS 还将视觉感知信息作为 LMM 的输入，用于预测视距路径和强非视距路径的存在情况，以及用户的几何信息（距离和角度）。这样，LMM-PS 能够快速理解信道预测过程中的突发事件。由于 LMM 是一种强大的工具，能够从序列数据中提取相关特征并生成后续元素 [3]，，因此可根据视觉感知信息中捕捉到的用户历史移动情况，预测每个用户未来的信道参数。

所提出的 LMM-PS 的运行分为两个主要步骤（见图 2）。第一步中，利用专门用于目标检测（OD）的 LMM 提取用户和障碍物的位置与尺寸，该 LMM 旨在检测输入句子中指定的任意对象。第二步中，并行执行大型多模态模型辅助的移动性预测和基于 Transformer 的信道参数预测，分别预测每个用户未来的视距和非视距信道参数：

- LMM 辅助的视距信道参数预测：为了估计每个用户未来的视距信道参数（即视距路径的方位角/仰角和距离），我们利用 LMM 根据专门设计的文本提示来预测用户未来的位置。这些提示词是根据用户过去位置序列生成的语言指令。例如，为了根据视觉感知信息中的用户轨迹预测其未来位置，我们使用如下提示词：“根据以下过去的 x-y 坐标对序列：(x_1, y_1) $=(930, 380)$、(x_2, y_2) $=(932, 377)$、(x_3, y_3) $=(935, 374)$，预测下一组坐标 (x_4, y_4)”。预测出用户未来的位置后，即可通过简单的坐标变换获得未来的视距信道参数。

- 基于 Transformer 的 NLoS 信道参数预测：为了根据过去信道序列估计未来的 NLoS 信道参数，我们采用 Transformer （一种 DNN 模型），它通过注意力机制捕捉输入序列和输出序列之间的长期及短期相关性 [13]。通过为与输出值（即未来信道参数）高度相关的输入序列（即过去的信道）赋予更大的注意力权重，所提出的 LMM-PS 能够提取未来的 NLoS 信道参数，从而即使存在突发的信道状态变化，也能预测未来信道。

通过在动态变化的 6G环境中进行现实评估，我们证明，LMM-PS 能显著降低意外事件（例如遮挡）导致的信道预测误差，因此在系统总吞吐量方面优于传统调度方法。例如，我们发现，LMM-PS 相较于比例公平调度技术可实现 32% 的吞吐量提升。即使与基于深度强化学习（DRL）的抢占式调度技术相比，LMM-PS 也能实现 11% 的吞吐量增益。

我们简要总结本文使用的符号约定。我们分别用大写和小写粗体字母表示矩阵和

[1]即使用户仅发生轻微移动，也可能导致视距（LoS）路径消失，从而使信噪比（SNR）大幅下降（例如，在毫米波频段可下降高达 35dB [9]），这是因为信道从视距（LoS）转变为非视距（NLoS）。

表 I 符号汇总。

符号	说明
BS	基站
CU	中央单元
CQI	信道质量指示符
MCS	调制与编码方案
SER	符号错误率
DL	深度学习
DNN	深度神经网络
RB	资源块
毫米波	毫米波
THz	太赫兹
RI	秩指示符
PMI	预编码矩阵指示符
QoS	服务质量
LoS	视距
NLoS	非视距
OD	目标检测
SoC	片上系统
SDM	空分复用

向量。运算$(\cdot)^{\mathrm{T}}$、$(\cdot)^{\mathrm{H}}$和$(\cdot)^{\dagger}$ 分别表示转置、共轭转置和伪逆。此外，算子 $\otimes$ 和 $\odot$ 分别表示Hadamard积和Kronecker积。 $\mathbb{C}$ 和 $\mathbb{R}$ 分别表示复数域和实数。注意，用户设备（UE）的集合、资源块和MCS索引分别表示为 $\mathcal{K} = \{1, \cdots, K\}$、 $\mathcal{B} = \{1, \cdots, B\}$和 $\mathcal{M} = \{0, \cdots, M\}$。 $\|\cdot\|_p$表示 p-范数，$\mathbf{x}_i$ 表示矩阵X的第 i列， $x(i)$表示向量x的第i个元素。最后， $\mathcal{I}$、 $\mathcal{I}_d$、 $\mathcal{I}_c$和P分别表示RGB图像、深度图像、裁剪后的RGB图像和文本提示。在表 I 中，我们总结了专业术语。

本文其余部分安排如下：第二节讨论系统模型，并对抢占式调度问题进行建模。第三节详细介绍所提出的LMM-PS。第四节给出仿真结果，并在第五节总结全文。

II. 抢占式调度问题

本节简要讨论系统模型和传统调度技术。随后，我们将抢占式调度问题表述为一个约束优化任务。

A. 抢占式调度与系统模型

抢占式调度的主要目标是在未来时隙 T 中，通过为 K 用户设备分配资源块并为每个用户设备选择合适的调制与编码方案索引，使系统总吞吐量最大化。为表示在第 t 个时隙为每个用户设备分配资源块的情况，我们定义一个二进制向量 $\mathbf{x}_b^{(t)} = [x_b^{(t)}(1) \cdots x_b^{(t)}(K)] \in \mathbb{R}^K$:

$$x_b^{(t)}(k) = \begin{cases} 1 & \text{if the } b\text{-th RB is allocated to the } k\text{-th UE,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

需要确保每个用户设备在其所有已调度的资源块上都使用相同的调制与编码方案索引 m [14]。为表示分配给每个用户设备的调制与编码方案索引，我们定义二进制独热向量 $\mathbf{z}_k^{(t)} = [z_k^{(t)}(1) \cdots z_k^{(t)}(M)] \in \mathbb{R}^M$:

$$z_k^{(t)}(m) = \begin{cases} 1 & \text{if MCS } m \text{ is assigned to the } k\text{-th UE,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

根据 [15], 中的 5G NR资源定义， R_k 第 k 个用户设备的下行链路吞吐量可表示为

$$R_k^{(t)} = \sum_{m=1}^{M} \sum_{b=1}^{B} \left(1 - (1 - (1 - e_{\text{SER},k}^{(t)})^{N_s})\right) \cdot f(z_k^{(t)}(m)) x_b^{(t)}(k) N_{\text{SDM}} N_{\text{RE}}, \tag{3}$$

其中， $f(\cdot)$是调制与编码方案阶数表，用于将所选调制与编码方案索引 m 映射到单位频谱效率（U-SE）；单位频谱效率表示每个资源元素（RE）承载的比特数。 N_{SDM} 是空分复用（SDM）层数，即从基站（BS）向每个用户设备传输的数据流数量。此外，$e_{\text{SER},\ k}$ 是第 k 个用户设备的符号错误率（SER）， N_{RE} 是每个资源块包含的资源元素数量。注意，对于 K 个用户设备，系统总吞吐量为 $R^{(t)}\ \text{total} = \sum_{k=1}^{K} R_k^{(t)}$ 。

在本文中，我们考虑多输入单输出（MISO）系统，其中配备了由 $N_T = N_x \times N_y$ 根天线组成的均匀平面阵列（UPA）的基站为 K 个单天线用户设备提供服务。在该设置中，来自第 k 个用户设备的接收导频信号 $\mathbf{y}^{(t)} \in \mathbb{C}^{N_T \times 1}$ 表示为

$$\mathbf{y}_k^{(t)} = \mathbf{h}_k^{(t)} s + \mathbf{n}_k^{(t)}, \tag{4}$$

其中，$\mathbf{h}_k^{(t)} \in \mathbb{C}^{N_T \times 1}$ 是第 k 个用户设备的上行链路信道向量，$s \in \mathbb{C}$ 是上行链路导频符号，而 $\mathbf{n}_k^{(t)} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I})$ 是加性高斯噪声。

关于信道模型，我们采用块衰落多径信道模型，其中信道每隔一个块变化一次 [16]。需要指出的是，所提出的技术是一种基于 LMM 的技术，因此能够通过训练过程（例如微调和持续学习）灵活适应不同的任务目标和动态环境。这意味着其性能可能不会受到信道环境或干扰模式变化的显著影响。具体而言，第 k 个用户设备的上行链路信道向量 $\mathbf{h}_k^{(t)} \in \mathbb{C}^{M \times 1}$ 为

$$\mathbf{h}_k^{(t)} = \delta_k^{(t)} \mathbf{h}_{\text{LoS},k}^{(t)} + \mathbf{h}_{\text{NLoS},k}^{(t)} \tag{5}$$

$$= \delta_k^{(t)} \mathbf{h}_{\text{LoS},k}^{(t)} + \sum_{l=1}^{L} \mathbf{h}_{\text{NLoS},l,k}^{(t)}, \tag{6}$$

其中， L 为非视距路径的数量， $\delta_k^{(t)}$ 表示第 k个用户设备的视距链路状态:

$$\delta_k^{(t)} = \begin{cases} 1 & \text{if LoS link is available for the } k\text{-th UE,} \\ 0 & \text{else.} \end{cases} \tag{7}$$

请注意，视距和非视距信道分量表示为 [17]

$$\mathbf{h}_{\text{LoS},k}^{(t)} = \alpha_{\text{LoS},k}^{(t)} e^{-j2\pi f_c \frac{r_{\text{LoS},k}^{(t)}}{c}} \mathbf{a}(\theta_{\text{LoS},k}^{(t)}, \phi_{\text{LoS},k}^{(t)}), \tag{8}$$

$$\mathbf{h}_{\text{NLoS},l,k}^{(t)} = \alpha_{\text{NLoS},l,k}^{(t)} e^{-j2\pi f_c \frac{r_{\text{NLoS},l,k}^{(t)}}{c}} \mathbf{a}(\theta_{\text{NLoS},l,k}^{(t)}, \phi_{\text{NLoS},l,k}^{(t)}), \tag{9}$$

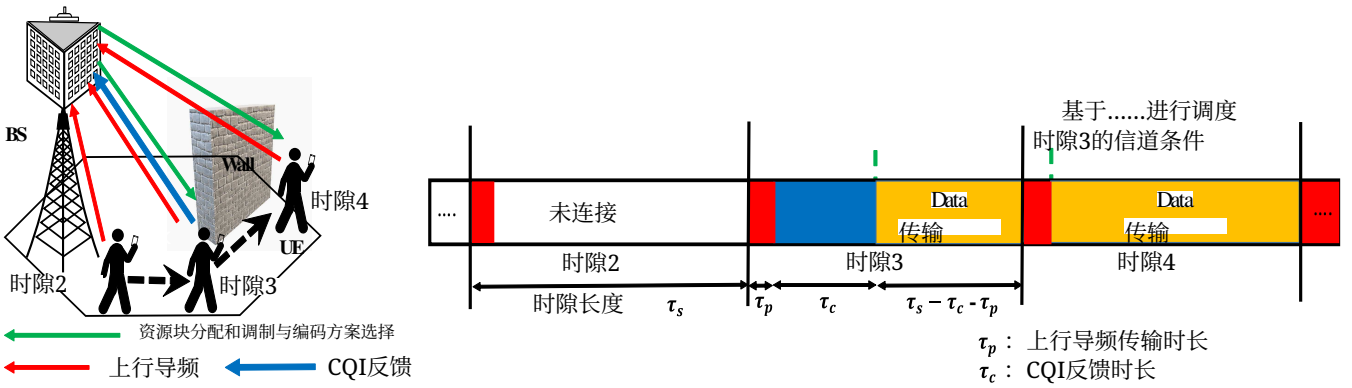


图3. 传统的基于CQI反馈的调度示意图。当基站基于过时的CQI反馈做出调度决策时，这些反馈可能已无法反映真实的信道条件，从而导致波束成形增益和传输速率下降。

其中，$\alpha^{(t)}$ LoS k 和 $\alpha^{(t)}$ NLoS $l,\ k$ 分别为视距路径和第 l 条非视距路径的复路径增益。此外，c 为光速，$\mathbf{a}(\theta,\phi) \in \mathbb{C}^{N_T\times 1}$ 为阵列响应向量，f_c 为载波频率，$r^{(t)}$ LoS k 和 $r^{(t)}$ NLoS $l,\ k$ 分别为基站与第 k 个用户设备之间的视距距离和第 l 条非视距路径的距离。最后，$(\theta^{(t)}$ LoS $k, \phi^{(t)}$ LoS $k)$ 和 $(\theta^{(t)}$ NLoS $l,\ k, \phi^{(t)}$ NLoS $l,\ k)$ 分别为视距路径和第 l 条非视距路径的方位角/仰角。

B. 传统调度技术

在传统调度技术中，用户设备会定期向基站传输信道质量指示值，以帮助基站为下行传输选择合适的调制与编码方案 [18], [19]。用户设备接收到来自基站的发射导频信号后，会计算下行信道响应，并得出秩指示符和预编码矩阵指示符 [20]。利用预定义的CQI映射表，可将预编码矩阵指示符映射为信道质量指示值 [21]，据此确定调制与编码方案索引 [22]。多年来，人们提出了各种基于深度学习（DL）的调度技术，以处理复杂的CQI到MCS映射过程 [18], [20], [23]。这些技术的关键特征是将CQI反馈和信道状态信息作为深度学习模型的输入。

在动态变化的6G环境中，由于障碍物密集分布、毫米波/太赫兹信道的信道相干时间较短以及频谱多样化，传统调度技术的性能可能会严重下降。这些因素会导致信道快速变化，使得可靠路径（可能是视距路径）即使在用户设备发生微小移动时也可能消失 [9], [24]。由于CQI反馈以相对较长的间隔上报（通常每8ms一次 [25]），当前时刻的实际信道条件可能与根据CQI反馈获得的估计信道条件存在显著差异（见图3）。显然，这种不匹配将导致波束成形增益和传输速率严重下降。

近年来，基于未来信道条件预测结果进行调度决策的抢占式调度技术受到了广泛关注 [10]–[12]。这些技术背后的普遍思路是捕捉历史CQI测量值中的时间依赖关系，以预测未来的CQI值。由于这些技术完全依赖于过去的信道信息，因此往往无法捕捉突发的信道变化，例如阻塞、手部颤动或移动方向的突然改变。相比之下，LMM-PS通过视觉感知图像检测信号传播环境，并利用LMM根据用户设备的移动推断突发事件（例如信号遮挡），从而能够快速理解和识别无线环境。

C. 抢先式调度问题建模

由于抢占式调度的主要目标是在未来时隙 T 中最大化系统总吞吐量，因此其目标函数可表示为

$$R_{\text{total}}^{(T)} = \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{b=1}^{B}\left(1-\left(1-\left(1-e_{\text{SER},k}^{(T)}\right)^{N_s}\right)\right)\cdot f(z_k^{(T)}(m))x_b^{(T)}(k)N_{\text{SDM}}N_{\text{RE}}. \tag{10}$$

除吞吐量最大化之外，还需要考虑两个主要约束：1）数据速率约束；2）SER约束。首先，数据速率约束要求分配给用户设备 $\sum_{b=1}^{B} x_b^{(T)}(k)$ 的资源块总数必须大于 $N^{(T-1)}$ RB k（满足用户设备速率要求所需的最小资源块数 $R_k^{(T-1)}$, min，位于第（$T-1$）个时隙）。在抢占式调度中，由于不知道用户设备在第 T 个（未来）时隙中的确切速率要求，基站需要在第（$T-1$）个时隙为第 T 个时隙分配资源块。为处理这一不确定性，调度器确保为第 T 个时隙分配的资源块满足第（$T-1$）个时隙中的速率要求。

此外，SER约束要求第 k个用户设备的 $e^{(T)}$ SER k 优于（低于）目标符号错误率 SER$_{k,\ \max}$ [26]。为此，必须基于信道估计 $\mathbf{h}_k$计算SER，该信道估计是利用传统的信道估计技术（例如线性最小均方误差（LMMSE））根据 $\mathbf{y}_k$得到的。根据 [27], , SER表示为

$$e_{\text{SER},k}^{(t)} = \begin{cases} 1-\left(1-Q\left(\sqrt{\frac{2|\hat{\mathbf{h}}_k^{(t)}|^2}{\sigma_n^2}}\right)\right)^2 & \text{if } Q_m=2, \\ 1-\left(1-\frac{2(\sqrt{Q_m}-1)}{\sqrt{Q_m}}Q\left(\sqrt{\frac{3|\hat{\mathbf{h}}_k^{(t)}|^2}{(Q_m-1)\sigma_n^2}}\right)\right)^2 & \text{else}, \end{cases} \tag{11}$$

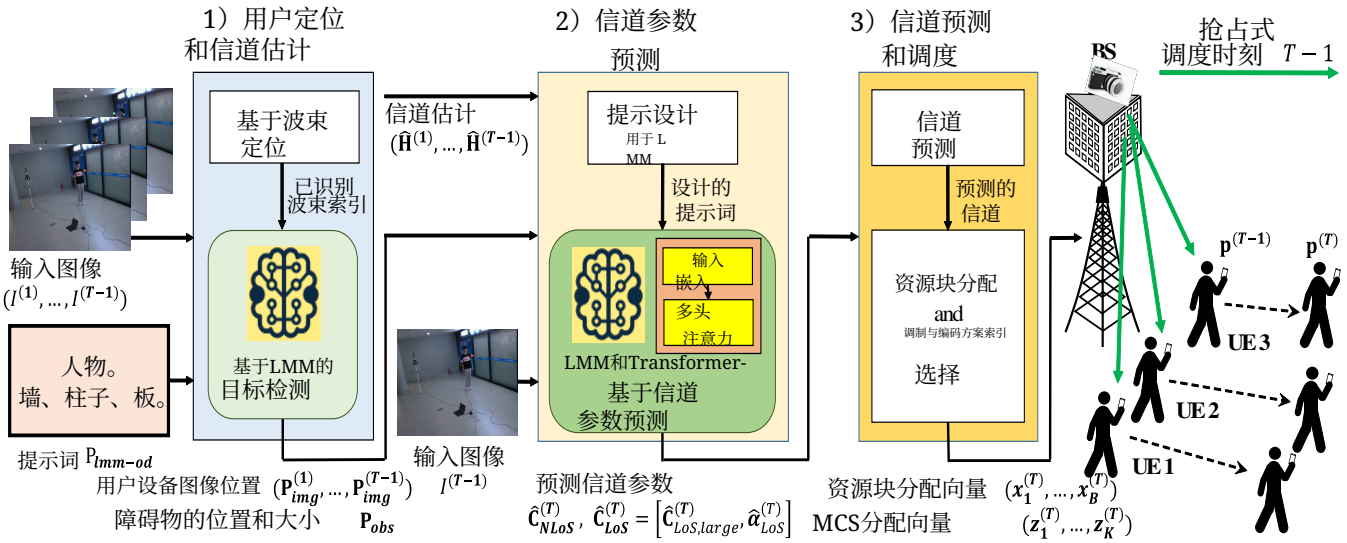


图 4. 所提出的 LMM-PS 的整体流程。

其中，$\frac{|\hat{\mathbf{h}}_k^{(t)}|^2}{\sigma_n^2}$ 是第 k个用户设备的信噪比， $Q(\cdot)$是Q函数，Q_m 是从阶数表[2]中选择的调制阶数。

通过纳入这些约束，抢占式调度问题 $\mathcal{P}$ 可表述为

$$\mathcal{P}: \max_{\{\mathbf{x}_1^{(T)},\cdots,\mathbf{x}_B^{(T)},\mathbf{z}_1^{(T)},\cdots,\mathbf{z}_K^{(T)}\}} R_{\text{total}}^{(T)} \tag{12a}$$

$$\text{s.t.} \quad \sum_{b=1}^{B} x_b^{(T)}(k) \geq N_{\text{RB},k}^{(T-1)}, \forall k \in \mathcal{K}, \tag{12b}$$

$$\sum_{k=1}^{K} x_b^{(T)}(k) \leq N_T, \quad \forall b \in \mathcal{B}, \tag{12c}$$

$$e_{\text{SER},k}^{(T)} \leq \text{SER}_{k,\max}, \quad \forall k \in \mathcal{K}, \tag{12d}$$

$$\sum_{m=1}^{M} z_k^{(T)}(m) = 1. \quad \forall k \in \mathcal{K}. \tag{12e}$$

在本工作中，我们假设空分复用层的数量 N_{SDM}固定为 1[3]。在我们的问题 $\mathcal{P}$中，采用资源约束(12c)来确保在一个资源块中调度的用户设备数量不超过基站天线数量 [28]。此外，我们使用 MCS分配约束 (12e)，以确保每个用户设备在所有已调度资源块中仅使用一种调制与编码方案。

[2]例如，如果使用 256阶正交幅度调制阶数表，那么当所选调制与编码方案索引为 $m_k \in \{0,1,2,3,4\}$时， Q_m 为 2；当 $m_k \in\{5,6,7,8,9,10\}$时为 4；当 $m_k \in \{11,12,13,14,15,16,17,18,19\}$时为 6；当 $m_k \in \{20,21,22,23,24,25,26,27\}$ [21]时为 8。

[3]这确保了 N_{SDM} 不超过用户设备接收天线的数量，符合 [28]中的要求。

代入式 (10) 后， $\mathcal{P}$ 可重新表示为

$$\mathcal{P}': \max_{\{\mathbf{x}_1^{(T)},\cdots,\mathbf{x}_B^{(T)},\mathbf{z}_1^{(T)},\cdots,\mathbf{z}_K^{(T)}\}} \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{b=1}^{B} \left(1-(1-(1-e_{\text{SER},k}^{(T)})^{N_s})\right) \cdot f(z_k^{(T)}(m)) x_b^{(T)}(k) N_{\text{RE}} \tag{13a}$$

$$\text{s.t.} \sum_{b=1}^{B} x_b^{(T)}(k) \geq B_{k,\min}, \forall k \in \mathcal{K}, \tag{13b}$$

$$\sum_{k=1}^{K} x_b^{(T)}(k) \leq N_T, \quad \forall b \in \mathcal{B}, \tag{13c}$$

$$e_{\text{SER},k}^{(T)} \leq \text{SER}_{k,\max}, \ \forall k \in \mathcal{K}, \tag{13d}$$

$$\sum_{m=1}^{M} z_k^{(T)}(m) = 1, \quad \forall k \in \mathcal{K}. \tag{13e}$$

遗憾的是，由于 $\mathcal{P}'$ 的目标函数是未来时隙 T的系统总吞吐量，因此只有在完全掌握用户设备的未来信道（即，$(\mathbf{h}_1^{(T)},\cdots,\mathbf{h}_K^{(T)})$）时，才能得到最优解 $\{\mathbf{x}^{(T)}\text{ opt }_b\}_{b=1}^{B}$ 和 $\{\mathbf{z}^{(T)}\text{ opt }_k\}_{k=1}^{K}$；而考虑到因果性问题，这显然不切实际。

III. 基于LMM的抢占式调度

LMM-PS的主要目标是借助视觉感知和大型多模态模型，通过预测用户设备的未来信道（$\mathbf{h}_1^{(T)},\cdots,\mathbf{h}_K^{(T)}$）来求解 $\mathcal{P}'$。简而言之，LMM-PS基于视觉感知序列和基于射频的历史信道估计序列来预测未来信道。然后，LMM-PS利用预测得到的未来信道，求出满足 $\mathcal{P}'$的约束条件 (13b)-(13e)的 $\{\mathbf{x}_b^{(T)}\}_{b=1}^{B}$ 和 $\{\mathbf{z}_k^{(T)}\}_{k=1}^{K}$。

具体而言，LMM-PS 包含三个主要阶段（见图 4）。首先，在用户定位和信道估计

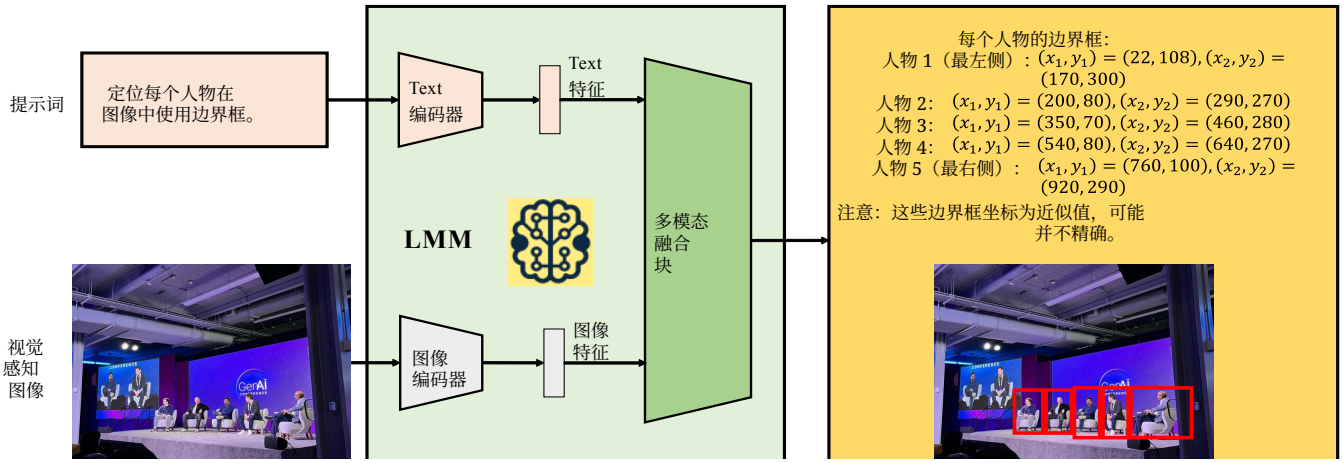


图 5. LMM 的基本结构。

阶段，基站使用 LMM-OD 识别用户设备的位置，以及障碍物的位置和大小。对于 LMM-OD，我们使用 GLIPv2，这是一种经过训练、可检测输入句子 [29] 中指定对象的现成 LMM。基站利用获取的用户设备位置和上行导频信号，在每个时隙中执行信道估计。

其次，在信道参数预测阶段，基站预测每个用户设备的未来信道参数（例如距离、角度和路径增益）。由于毫米波/太赫兹环境中的信道由少量传播路径（视距路径和 $1 \sim 4$ 非视距路径）建模而成，且每条路径仅包含少量信道参数，因此，LMM-PS 无需预测全维信道 $\mathbf{h}_k^{(T)}$，而是通过估计这些参数来有效预测未来信道。为了根据用户设备过去的图像位置预测其未来的大尺度视距信道参数（视距路径的距离/角度），基站使用 LLaVA。LLaVA 是一种 LMM，能够结合视觉和文本信息，捕捉多模态序列数据中的时空相关性。由于 LLaVA 同时使用图像和文本输入来生成后续元素，因此能够基于用户设备过去的轨迹 [30]有效预测其未来信道参数。此外，作为一种基于语言的模型，LLaVA 可以利用文本提示灵活适应新的任务目标（例如更改目标用户设备的类型）。通过相对简单的微调过程，LLaVA 还可以轻松适应各种部署环境。例如，使用基于低秩适配（LoRA）的微调技术 [31], ，我们只需训练新添加模块（称为适配器）的参数，即可帮助 LLaVA 在各种部署场景中快速执行稳健的用户设备轨迹预测。

与此同时，基站采用 Transformer，根据过去的信道估计来预测用户设备未来的非视距信道参数（以及视距路径增益）。由于 Transformer 会为与输出值（即未来信道参数）高度相关的输入数据（即过去的信道）分配相对较大的注意力权重，因此能够有效捕捉长期空间相关性；即使环境突然变化（例如出现新的散射体），这种相关性也往往能够保持，从而提高信道预测准确率。

最后，在信道预测与调度阶段，LMM-PS 使用预测的信道求解问题 $\mathcal{P}'$ in (13a) 未来信道。具体而言，基站将获取的信道参数代入信道模型 (8) 和 (9)，生成未来信道估计值。随后，利用这些估计值计算比例公平（PF）指标 [32]，以确定用户设备的调度优先级。接着，在满足约束 (13b)-(13c) 的资源块中，基站为 PF 值最高的用户设备分配最多的资源块，而为 PF 值较低的用户设备分配较少的资源块。同样，在满足约束 (13d)-(13e) 的 MCS 索引中，为每个用户设备选择最高的可用 MCS 索引。

A. LMM 在信道预测中的作用

大型多模态模型是大语言模型的扩展，旨在处理来自多种模态的输入（例如图像和文本）。大型多模态模型的关键组成部分包括针对每种模态的预训练编码器（例如，用于图像的 ViT [33] 和用于文本的 BERT [34]），以及一个融合网络，用于整合各模态的特征并生成特定任务的预测（见图 5） [35]。

与传统深度神经网络相比，在 LMM-PS 中使用 LMM 的主要优势可概括如下：

- 与只能识别预定义类别对象（如猫和笔记本电脑）的传统目标检测器（OD）相比，针对目标检测定制的 LMM（LMM-OD）能够轻松检测任意对象，无论这些对象属于预定义类别还是类别之外。这是因为传统目标检测器仅依赖图像输入，而 LMM-OD 同时利用图像和文本输入来理解输入句子，并识别句中提及对象在图像中的位置。这意味着 LMM-OD 能够灵活适应新的任务目标或环境变化。例如，如果新出现了会导致视距阻塞的障碍物（如机器人和摩托车），只需更新提示词，便可使 LMM-OD 不受其干扰，并专注于目标对象。

- 与仅依赖用户设备过去轨迹的传统轨迹预测技术不同，大型多模态模型能够捕捉场景中的时空关系，预测用户设备未来的像素级位置。借助这种增强的预测能力，可以检测意外事件

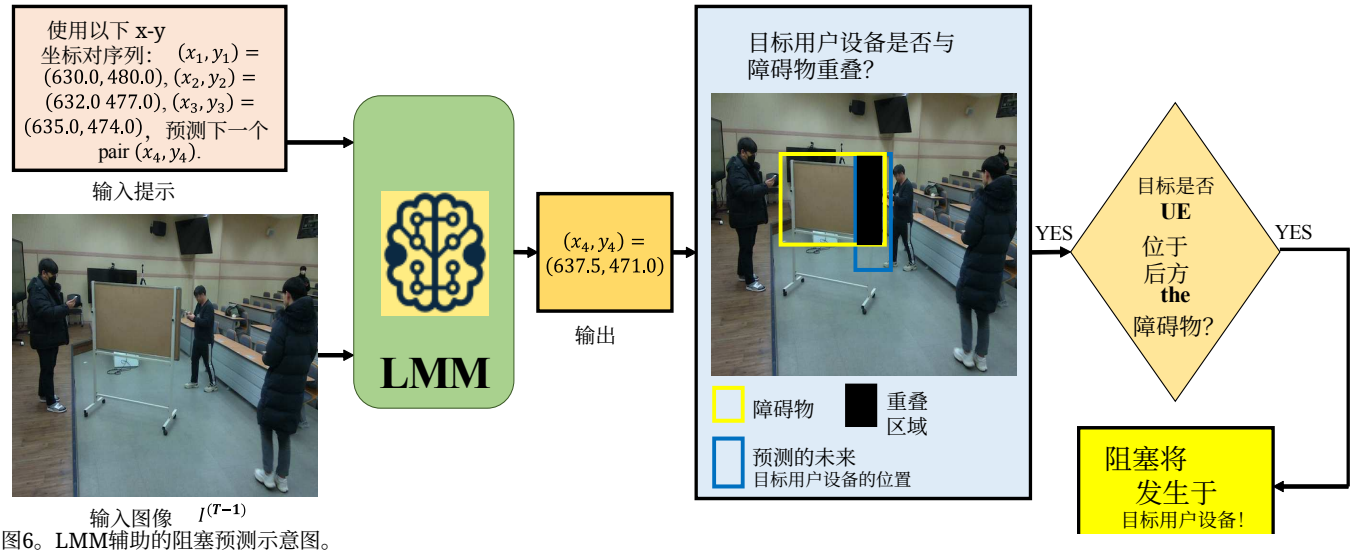


图6。LMM辅助的阻塞预测示意图。

（例如视距阻塞或移动方向突然改变），并在信道预测过程中将其纳入考虑。例如，大型多模态模型可以通过比较用户设备的预测位置与障碍物的位置和大小，识别可能与障碍物重叠的用户设备（见图6）；传统深度神经网络技术则无法做到这一点。需要注意的是，在探索超高频频谱（例如FR2频段和FR3中高频段）的6G信道环境中，大型多模态模型实现的高精度视距阻塞预测至关重要。在这类环境中，由于非视距路径上的信号衰减严重（最多只有少数几条路径能够保留），信道以视距为主；这意味着，即使仅发生一次视距阻塞漏检，也可能严重降低信道预测准确率。

- 每当环境或任务发生变化时就更新深度神经网络的权重，对于现实场景而言并不实际。相比之下，由于大型多模态模型是基于语言的模型，因此无需进行模型训练，借助提示工程即可轻松更新或修改任务。在各种提示技术中，我们的研究采用了思维链提示，该方法引导大型多模态模型通过中间步骤进行推理。例如，当输入类似“使用过去的x-y坐标对序列：$(x_1, y_1) = (630, 480)$、$(x_2, y_2) = (632, 477)$、$(x_3, y_3) = (635, 474)$，预测下一个坐标对(x_4, y_4)”的提示词时，如果模型生成了偏离既定模式的、不合理的输出（例如，$(x_4, y_4) = (941, 611)$），则可以使用额外的提示词来增强思维链推理（例如，“如果预测的坐标对在物理上不合理，请根据更长的序列预测下一个坐标对(x_6, y_6)：$(x_1, y_1) = (625, 485)$、$(x_2, y_2) = (628, 483)$、$(x_3, y_3) = (630, 480)$、$(x_4, y_4) = (632, 477)$、$(x_5, y_5) = (635, 474)$”）。

B. 基于大语言模型的用户定位与信道估计

LMM-PS的第一阶段称为用户定位与信道估计阶段，分为四个主要步骤：1）基于波束的定位，2）RGB图像中的目标检测，3）多用户ID识别，以及4）视距/非视距信道分量估计。

1) 基于波束的定位：在此步骤中，我们发射少量波束，以确定用户设备的大致方向。对于识别出的物理区域，我们使用安装在基站上的 RGB-D 相机采集 RGB 图像和深度图像。

2) RGB图像中的目标检测：对于采集到的 RGB 图像 $\mathcal{I}^{(t)}$，我们使用 LMM-OD $g_{\text{lmm-od}}$ 来识别用户设备的位置，以及障碍物的位置和大小。

- 用户定位：由于大多数用户设备都是小型移动设备（例如手机和笔记本电脑）[36], [37]，它们在 $\mathcal{I}(t)$ 中仅占据少数像素，因此$g$lmm-od难以识别这些设备。为克服这一不足，我们采用两阶段目标检测流程：$g$lmm-od先在 $\mathcal{I}(t)$ 中识别手持用户设备的人员，然后在所识别人员的裁剪图像 $\mathcal{I}(t)_c$ 中检测每个用户设备。例如，$g$lmm-od将文本提示“person”和 $\mathcal{I}(t)$ 作为输入，在 $\mathcal{I}(t)$ 中识别人员。随后，$g$lmm-od将 $\mathcal{I}(t)_c$ 和文本提示“cell phone”作为输入，以识别 K' 手机的图像位置 $\text{P}(t)\ \text{img} = \{(x(t)\ \text{img}_k, y(t)\ \text{img}_k)\}_{k=1}^{K'}$。结合通过像素到角度转换4得到的每个用户设备的角度（$\hat{\theta}(t)\ \text{LoS}_k$，$\hat{\phi}(t)\ \text{LoS}_k$）以及根据采集到的深度图像 $\mathcal{I}(t)_d$ 测得的距离 $\hat{r}(t)\ \text{LoS}_k$，我们得到 K' 组视距信道参数估计（见图 7）。

- 障碍物识别：为确定 $\mathcal{I}(t)$中的障碍物位置和尺寸，g LMM-OD使用指定待识别障碍物类型的文本提示。例如，使用输入文本提示“墙、板”，gLMM-OD返回 $\text{Pobs} = \{(x\text{img}_j, y\text{img}_j, w\text{img}_j, h\text{img}_j)\}_{j=1}^{J}$，其中包含 $\mathcal{I}(t)$中墙和板的中心坐标、宽度和高度。

3）多用户ID识别：由于LMM-OD能够在捕获图像 $\mathcal{I}^{(t)}$ 中检测到多个用户设备，因此需要通过基站分配的ID对每个检测到的用户设备进行识别。为此

[4]利用RGB-D相机的焦距（例如50 mm）、相机的方向角（θ_i, ϕ_i）以及每个检测对象的二维位置（$x_{\text{img}}, y_{\text{img}}$），即可按照 [9]进行像素到角度转换。

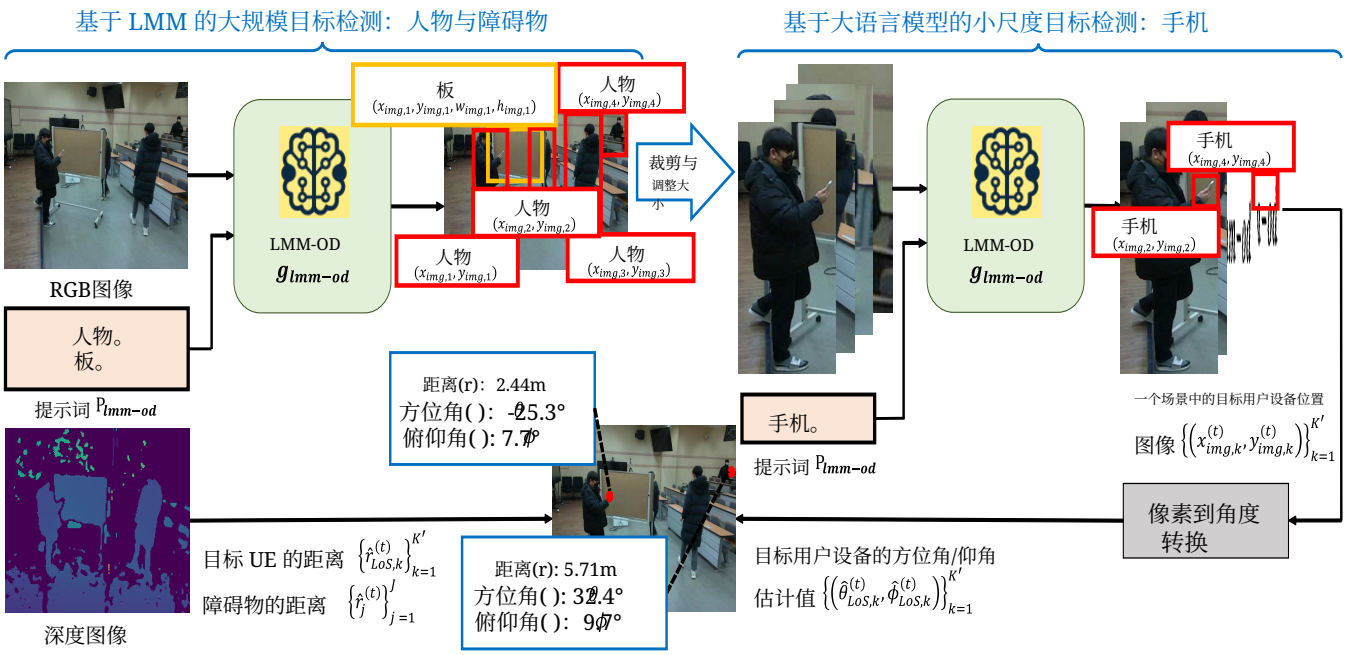


图7. 感知、LMM辅助用户定位与障碍物识别示意图。

我们通过以下步骤进行多用户ID识别（见图8）：

1）获取检测到的用户设备的方位角和俯仰角（$\hat{\theta}^{(t)}$ 视距 $_k$、$\hat{\phi}^{(t)}$ 视距 $_k$）后，我们为每个检测到的用户设备生成一个波束成形向量：$\mathbf{f}_k^{(t)}$ $= \mathbf{a}(\hat{\theta}^{(t)}$ 视距 $_k$、$\hat{\phi}^{(t)}$ 视距 $_k)$。

2) 由于通过传统信道估计技术可以获得所有 K UE（由基站服务）的信道估计，因此我们通过确定与 UE 的波束成形向量 $\mathbf{f}_k^{(t)}$ 相关性最大的信道估计 $\hat{\mathbf{h}}_k^{(t)}$，来识别每个检测到的 UE 的 ID。为此，我们计算每个 K' 波束成形向量与 K 信道估计之间的相关性 $|\hat{\mathbf{h}}_k^{(t)\mathrm{H}}\mathbf{f}_{(t)\,k}|$，生成相关矩阵 $\mathbf{R}^{(t)} \in \mathbb{R}^{K' \times K}$。3) 为了正确识别所有检测到的 UE 的 ID，必须在 K' 波束成形向量与 K 信道估计之间进行准确的一对一匹配。为此，我们利用 $\mathbf{R}^{(t)}$ 构建匹配问题，其中，将波束成形向量 $\mathbf{f}_{(t)\,k}$ 分配给信道估计 $\hat{\mathbf{h}}_k^{(t)}$ 的代价定义为相关性 $|\hat{\mathbf{h}}_k^{(t)\mathrm{H}}\mathbf{f}_k^{(t)}|$ 的倒数。随后，我们使用匈牙利算法 [38], 求解该问题，该算法通过迭代最小化总匹配代价，计算两个集合之间的最优一对一匹配。在我们的案例中，我们对波束成形向量和信道估计进行一对一匹配。

对于在 $\mathcal{I}^{(t)}$ 中未检测到的任何用户设备，我们从波束码本中选择使接收参考信号功率（RSRP）最大化的波束成形向量（即波束码字）。根据所选波束码字的索引，我们提取用户设备距离和角度的粗略估计值 [39]。在确定所有 K' 个已检测用户设备的 ID 后，我们通过将使用 LMM-OD 估计的每组距离和角度与已识别的用户设备进行匹配，确定所有 K 个用户设备的大尺度视距信道参数：

$$\hat{\mathbf{C}}^{(t)}_{\mathrm{LoS,\,large}} = [\, r^{(t)} \quad \theta^{(t)} \quad \phi^{(t)} \quad \cdots \quad r^{(t)}_{K} \; \theta^{(t)}_{K} \; \phi^{(t)}_{K} \,] \qquad (\hat{\mathrm{LoS}}_{,\,1},\ \hat{\mathrm{LoS}}_{,\,1},\ \hat{\mathrm{LoS}}_{,\,1})\,(\hat{\mathrm{LoS}}\,,\ \hat{\mathrm{LoS}}\,,\ \hat{\mathrm{LoS}}\,)。$$

4）视距/非视距信道分量估计：获取大尺度视距信道参数估计后，我们使用最小二乘（LS）技术，根据接收导频信号 $\mathbf{y}_k^{(t)}$ 估计小尺度视距路径增益参数：

$$\hat{\alpha}_{\mathrm{LoS},k}^{(t)} = e^{j2\pi f_c \frac{\hat{r}_{\mathrm{LoS},k}^{(t)}}{c}} (\mathbf{a}(\hat{\theta}_{\mathrm{LoS},k}^{(t)}, \hat{\phi}_{\mathrm{LoS},k}^{(t)}))^{\dagger} s\mathbf{y}_k^{(t)}. \qquad (14)$$

利用所获取的大尺度和小尺度信道参数，我们为每个用户设备估计全维视距信道分量，即 $\hat{\mathbf{h}}^{(t)}$ 视距 $_k = \hat{\alpha}^{(t)}$ 视距 $_k e^{-j2\pi f_c}$ 视距 $_k$ $\mathbf{a}(\hat{\theta}^{(t)}$ 视距 $_k$, $\hat{\phi}^{(t)}$ 视距 $_k)$。

由于信道 $\mathbf{h}_k^{(t)}$ 由视距和非视距分量组成（见式（5））, 我们通过从信道估计 $\hat{\mathbf{h}}_k^{(t)}$ 中减去视距分量估计 $\hat{\mathbf{h}}^{(t)}$ 视距 $_k$，来估计每个可见用户设备[5] 的非视距分量 $\hat{\mathbf{h}}^{(t)}$ 非视距 $_k$ （即，$\hat{\mathbf{h}}^{(t)}$ 非视距 $_k = \hat{\mathbf{h}}_k^{(t)} - \hat{\mathbf{h}}^{(t)}$ 视距 $_k$）。对于不可见用户设备，我们可以直接假设其视距路径被障碍物阻挡，因此其信道仅由非视距分量组成，令 $\hat{\mathbf{h}}^{(t)}$ 视距 $_k = \mathbf{0}$，且 $\hat{\mathbf{h}}^{(t)}$ 非视距 $_k = \hat{\mathbf{h}}_k^{(t)}$。

C. 基于 LMM 和 Transformer 的信道参数预测

在本小节中，我们讨论 LMM-PS 的第二阶段，即信道参数预测阶段。为了根据用户设备过去图像位置的时空关系预测未来的大尺度视距信道参数，LMM-PS 使用 LLaVA，这是一种能够捕捉这些关系并生成后续元素的 LMM。

[5]我们所说的可见和不可见用户设备，分别是指存在于和不存在于捕获图像 $\mathcal{I}^{(t)}$ 中的用户设备。

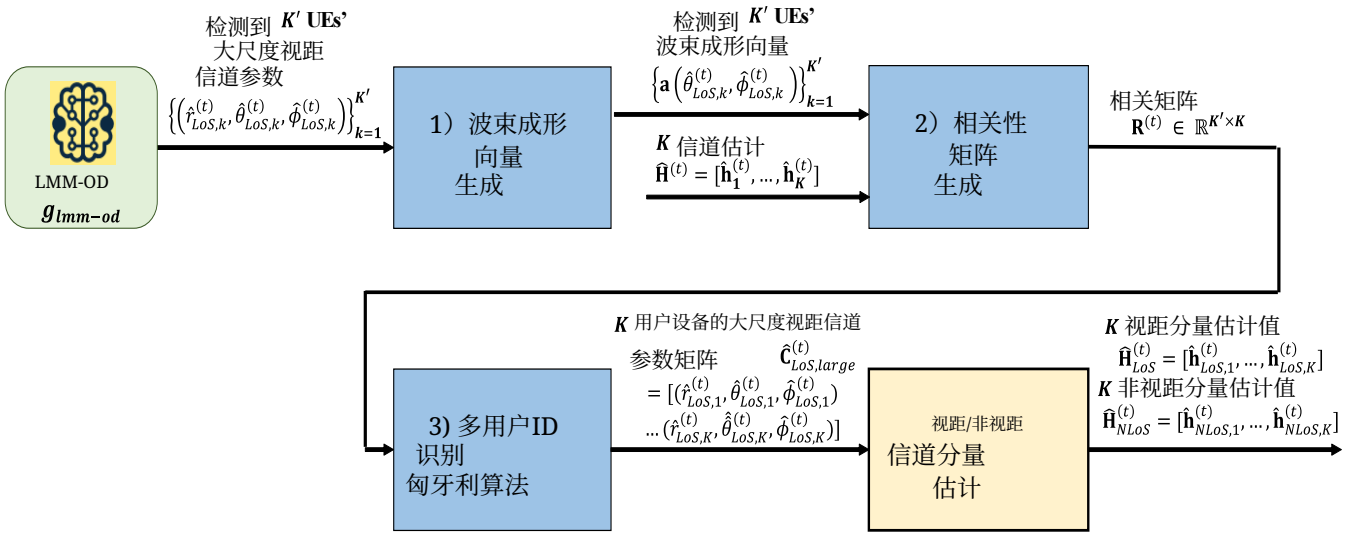


图8. LMM-PS 中多用户识别过程的逐步操作。

LMM-PS 还采用 Transformer，根据过去的视距/非视距分量预测未来的小尺度视距路径增益和非视距信道参数。Transformer 会对与未来参数高度相关的过去信道估计赋予更高注意力权重，从而有效预测这些值。

1) 使用 LMM 进行大尺度视距信道参数预测：由于 LLaVA 根据输入文本提示 P_{LLaVA}生成未来信道参数预测，因此精心设计 P_{LLaVA} 至关重要，这有助于提供清晰、准确的引导。具体而言，P_{LLaVA} 包含两类提示词：一类用于预测用户设备未来的角度，另一类用于预测其距离。

- 未来用户设备角度预测提示词：为了根据每个用户设备过去的图像位置 $\{(x(t)\text{ img }_k, y(t)\text{ img }_k)\}_{t=T-T_c}^{T-1}$预测其未来的方位角/俯仰角 $(\hat{}\theta(T)\text{ LoS }_k, \hat{}\phi(T)\text{ LoS }_k)$，我们使用如下提示词："根据以下过去的 x-y 坐标对序列：(x_1, y_1) = (740.0, 380.0), (x_2, y_2) = (742.5, 377.0), (x_3, y_3) = (745.0, 374.0)，预测下一个坐标对 (x_4, y_4)"。LLaVA 预测出下一个图像位置 $(x(T)\text{ img }_k, y(T)\text{ img }_k)$ 后，再通过像素到角度转换将其转换为角度估计值 $(\hat{}\theta(T)\text{ LoS }_k, \hat{}\phi(T)\text{ LoS }_k)$。
- 未来用户设备距离预测提示词：为了预测每个用户设备与基站之间的未来距离 $\hat{}r\ (T)$ 视距 $_k$，我们使用一系列过去的测量距离 $\{\hat{}r\ (t)$ 视距 $_k\}_{t=T-T_c}^{T-1}$。例如，我们使用如下提示词："使用过去的距离值序列：$r_1 = 12.3$、$r_2 = 11.2$、$r_3 = 10.3$，预测下一个距离值 r_4"（见图 9）。

LLaVA 预测出用户设备未来的图像位置和距离后，LMM-PS 会利用这些信息检测突发事件。具体而言，为了识别潜在的未来视距阻塞，LMM-PS 会将每个用户设备预测的图像位置与 $\mathcal{I}^{(T-1)}$ 中障碍物的位置和大小进行比较，检查其是否与任何障碍物重叠。如果检测到重叠，LMM-PS 会检查用户设备预测的、相对于基站的距离 $\hat{}\delta_k^{(T)}$ LoS $_k$ 是否短于重叠障碍物与基站之间的距离，从而确定该用户设备未来的视距链路状态。

与基站之间的距离 $\hat{}r_j$:

$$\hat{\delta}_k^{(T)} = \begin{cases} 1 & \text{if } \hat{r}_{\text{LoS},k}^{(T)} < \hat{r}_j, \forall (k,j) \in \mathcal{O}^{(T)} \\ 0 & \text{else,} \end{cases} \tag{15}$$

其中，$\mathcal{O}^{(T)}$ 是预测图像位置发生重叠的用户设备和障碍物的索引集合。

2）使用Transformer进行视距路径增益和非视距信道参数预测：由于视距和非视距信道分量取决于各自的信道参数，我们设计了两个独立的基于Transformer的信道预测器：一个用于视距（TCP-L），另一个用于非视距（TCP-N），以有效预测每个用户设备未来的视距路径增益和非视距信道参数。

在 TCP-L 和 TCP-N 中，我们使用过去估计的视距和非视距信道分量，分别构造输入矩阵 $\tilde{}$ $\tilde{}\mathrm{H}_{\text{LoS}}$ 和 H_{NLoS}。这些矩阵被嵌入到 $\mathrm{S}_{\text{embed}} \in \mathbb{R}^{K(T-1)\times NT}$ 中，然后输入 N_{attn} 个注意力块。每个注意力块依次执行多头注意力、层归一化、全连接（FC）层和另一个层归一化。此过程生成 TCP-L 和 TCP-N 的特征矩阵 $\mathrm{G}_{\text{LoS},\ B} \in \mathbb{R}^{K(T-1)\times N_T}$ 和 $\mathrm{G}_{\text{NLoS},\ B} \in \mathbb{R}^{K(T-1)\times N_T}$，然后分别通过多个全连接层将其转换为用户设备未来的视距路径增益 $\hat{}\boldsymbol{\alpha}^{(T)}$ LoS $\in \mathbb{R}^{K\times 2}$ 和初步预测的非视距信道参数 $\hat{}\mathrm{C}^{(T-1)}$ NLoS $\in \mathbb{R}^{K\times 5L_{\max}}$[6]（见图 10）。

为应对突发散射体导致的非视距路径数量变化 L，TCP-N 利用视觉感知和 Transformer。具体而言，TCP-N 使用图像特征提取器[7]，从捕获图像 $\mathcal{I}^{(T-1)}$ 中提取图像特征矩阵 $\mathrm{F} \in \mathbb{R}^{D\times 5L_{\max}}$，同时将 $\hat{}\mathrm{C}^{(T-1)}$ NLoS 嵌入到 $\mathrm{Z}_{\text{embed},\ F} \in \mathbb{R}^{K\times 5L_{\max}}$中。接下来，F 和 $\mathrm{Z}_{\text{embed},\ F}$ 经过 N_{attn} 个注意力块处理，生成特征矩阵 $\mathrm{G}_{\text{NLoS},\ F} \in \mathbb{R}^{K\times 5L_{\max}}$，然后通过多个全连接层转换为两个输出：（1）未来的非视距信道参数

[6]注意，$\hat{}\mathrm{C}^{(T)}$ NLoS 中的 5 来自路径增益的实部和虚部、方位角和俯仰角，以及距离。7

-图像特征提取器 [43]。在这项工作中，我们使用基于 Swin Transformer 的 以降低复杂度并销。

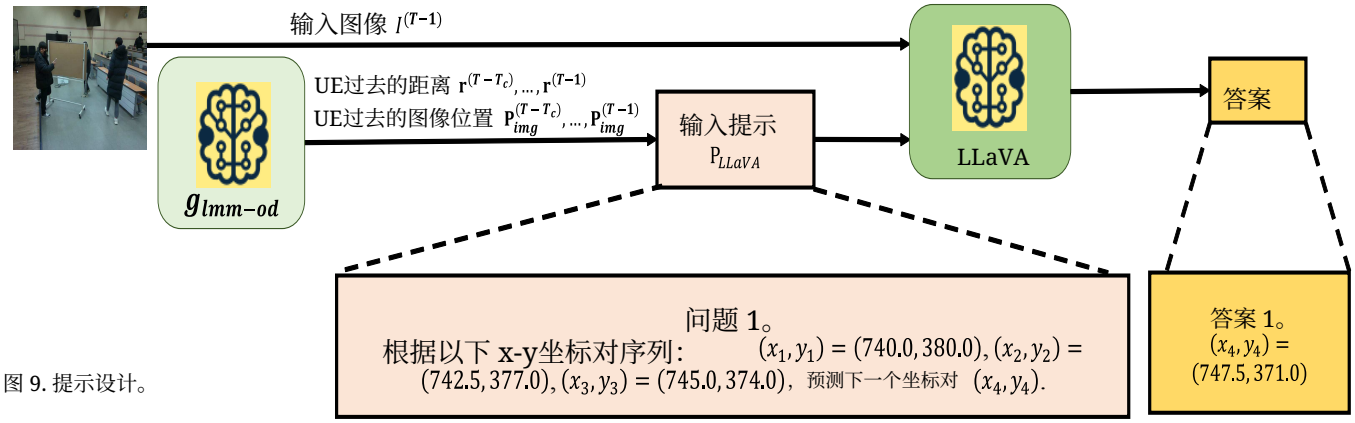


图 9. 提示设计。

预测值 $\hat{\mathbf{C}}^{(T)}$ 非视距 $\in \mathbb{R}^{K\times 5L_{\max}}$ 以及（2）非视距路径的存在概率（即分数值 $\hat{\mathbf{S}}^{(T)}$ 非视距 $\in \mathbb{R}^{K\times L_{\max}}$）。通过为与非视距信道参数高度相关的输入数据（即与路径散射体相关联的 RGB 像素）分配相对较大的注意力权重，TCP-N 能够从视觉感知图像中提取与信道相关的视觉特征，进而预测信道参数。最后，通过选择分数值高于预设阈值的预测信道参数集合，TCP-N 能够灵活调整预测的非视距路径数量。

利用获取的未来信道参数完成信道预测后，我们计算所有用户设备的比例公平度量，以确定其调度优先级。随后，基站在满足约束条件 (13b)-(13c) 的前提下，将最多数量的资源块分配给比例公平值最高的用户设备，并向比例公平值较低的用户设备分配较少的资源块。同样，从满足约束条件 (13d)-(13e) 的索引集合中选取最高的调制与编码方案索引。

D. TCP损失函数设计与训练过程

在训练阶段，我们通过最小化回归损失函数 L 来优化 TCP-L 和 TCP-N 的网络参数，该损失函数用于量化真实信道参数与预测信道参数之间的差异。为获取真实信道参数，我们使用 MATLAB 的射线追踪函数。在此过程中，我们可以识别信号传播路径及其对应的信道参数，然后应用逼真的路径损耗模型生成合成数据集。TCP-L 和 TCP-N 的损失函数表示为

$$L_{\text{TCP-L}} = \|\boldsymbol{\alpha}_{\triangle}\|_2^2 + \lambda\|\boldsymbol{\alpha}_{\triangle}\|_1,$$
$$\boldsymbol{\alpha}_{\triangle} = \boldsymbol{\alpha}_{\text{LoS}}^{(T)} - \hat{\boldsymbol{\alpha}}_{\text{LoS}}^{(T)}, \tag{16}$$

$$L_{\text{TCP-N}} = \|\mathbf{C}_{\triangle}\|_2^2 + \lambda\|\mathbf{C}_{\triangle}\|_1,$$
$$\mathbf{C}_{\triangle} = \mathbf{C}_{\text{NLoS}}^{(T)} - \hat{\mathbf{C}}_{\text{NLoS}}^{(T)}, \tag{17}$$

其中 $\|\boldsymbol{\alpha}_{\triangle}\|_2^2$ 和 $\|\mathbf{C}_{\triangle}\|_2^2$ 是均方误差（MSE）损失，$\|\boldsymbol{\alpha}_{\triangle}\|_1$ 和 $\|\mathbf{C}_{\triangle}\|_1$ 是平均绝对误差（MAE）损失，λ 是 MAE 损失项的正则化权重。请注意，在损失函数设计中，我们结合了 MAE 和 MSE，以缓解梯度消失问题。当 MSE $\|\boldsymbol{\alpha}_{\triangle}\|_2^2$ 和 $\|\mathbf{C}_{\triangle}\|_2^2$ 趋近于零时，其梯度也会变得非常小，从而减慢 TCP 的收敛速度。通过纳入梯度恒定的 MAE $\|\boldsymbol{\alpha}_{\triangle}\|_1$ 和 $\|\mathbf{C}_{\triangle}\|_1$，我们可以实现更稳定、一致的训练 [9]。

表 II 仿真参数

参数	数值
载波频率 f_c	28 吉赫兹
基站天线数量 N_T	64
时隙长度 τ_s	100 毫秒
带宽 W	100 MHz
资源块中的符号数 N_s	14
用户最小高度 $h_{\min}$	0.5 米
用户最大高度 $h_{\max}$	2 m
用户最大速度 $v_{\max}$	25 千米/小时
非视距路径最大数量 $L_{\max}$	3
训练样本数量	25, 600
验证样本数量	3, 000
测试样本数量	7, 400
隐藏单元数量 E	64
注意力块数量 N_{attn}	5
训练轮数	500
正则化系数 λ	1.0
时隙总数 T	11
资源块总数 $N_{\text{RB, total}}$	70
训练批次大小 B	128
Adam优化器的初始学习率 η	10^{-3}
时间窗大小 T_c	3

IV. 实验与讨论

A. 仿真设置

在我们的仿真中，基站为 $K=10$ 个单天线用户设备提供服务，这些设备随机分布在 $20\times 20\,m^2$ 服务区域内。请注意，v 和 h 分别表示用户设备的速度和高度，它们也分别从范围 $[h_{\min},h_{\max}]$ 和 $[0,v_{\max}]$ 中随机选取。每个用户设备沿直线移动[8]，到达服务区域边缘时随机改变方向。为生成第 p 条路径的路径增益 $\alpha_{k,\,p}=\sqrt{\beta_{k,\,p}^{\mathrm{L}}\beta_{k,\,p}^{\mathrm{S}}}=\sqrt{\rho(r_{k,\,p})}\cdot 10^{\frac{z_k\sigma_{\mathrm{sh}}}{10}}\beta_{k,\,p}^{\mathrm{S}}$，我们采用 3GPP TR 38.901 Rel. 中规定的路径损耗模型和阴影衰落，其中 $\rho(r)=-(31.84+21.5\log_{10}(r)+19\log_{10}(f_c))$

[8]尽管现实场景中的用户设备移动模式可能很复杂，但所提出的 LMM-PS 能够每 100 ms 预测一次用户设备轨迹，因此即使在类随机游走移动场景下，LMM-PS 仍能保持稳健的预测性能。

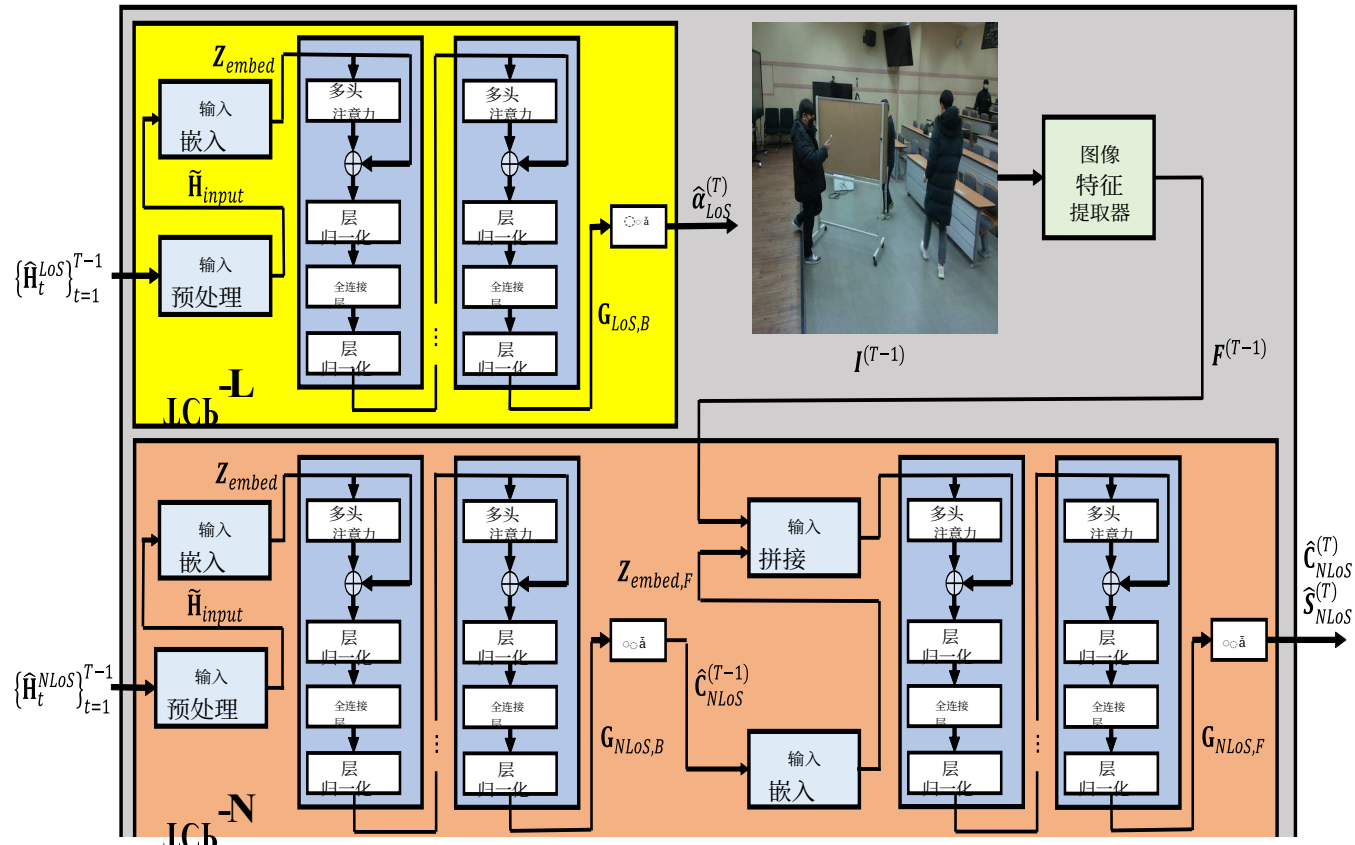


图 10. TCPin LMM-PS 的详细架构。



图 11. LMM-PS 在 VOBEM2 数据集上的目标检测结果。绿色和粉色边界框分别表示人员和障碍物。

表示路径损耗，$\sigma_{\text{sh}} = 4$，而 $z_k \sim \mathcal{N}(0,1)$。请注意，$\beta^{S}_{k,p} \sim \mathcal{CN}(0,1)$ 是小尺度衰落系数。系统参数汇总于表 II。

对于预训练的LLaVA模型，我们从最新的LLaVA模型系列中选择了LLaVA-v1.6-13B。该系列模型规模从v1.6-7B到v1.6-34B不等，以平衡推理速度和性能。在TCP-L和TCP-N的训练阶段，我们采用Adam优化器。Adam优化器被广泛使用，能够保持训练过程的稳健性 [41]。此外，为了帮助损失值在训练过程中平稳收敛，我们采用学习率衰减策略，每经过100个训练轮数后，将学习率缩小为原来的十分之一。

为评估 GLIPv2 的用户设备定位性能，我们使用专为无线通信设计的 VOBEM2 [42],a 感知数据集。VOBEM2 包含 104 对 RGB 图像和深度图像，其中有人员和手机，部分被障碍物遮挡。在我们的实验中，我们将墙壁、柱子或板定义为障碍物。为评估障碍物检测性能，我们对 VOBEM2 图像中的每个障碍物手动标注边界框（见 https://github.com/islab- github/VOBEM2）。随后，我们使用 GLIPv2 对这些新标注的图像进行目标检测（见图 11）。此外，为评估 LLaVA 的用户设备移动性和阻塞预测性能，我们使用 TCP 的合成测试样本。

在目标检测性能比较中，我们采用两种现成的目标检测器：1）Swin Transformer [43] 和 2）EfficientDet-D8 [44]。此外，在移动性（轨迹）预测性能比较中，我们使用两种基准方案：

1) 基于深度学习的轨迹预测 [45]：在此方案中，基站采用编码器-解码器架构，其中编码器根据历史轨迹提取特定区域内用户设备的共同移动模式，随后解码器基于这些模式预测未来轨迹。2) 基于Sage-Husa滤波器的轨迹预测 [46]：在此方案中，基站根据噪声统计估计用户设备移动测量值中未知且时变的噪声。之后，基站利用这些噪声估计值预测未来轨迹。

表 III 使用 GLIPv2 和 VOBEM2数据集进行目标检测与定位的性能。

	人/电话	定位误差（手机）		障碍物	定位误差	
	召回率 (%)	距离（厘米）	方位角/仰角（度）	精确率/召回率（%）	距离（厘米）	方位角/仰角（度）
GLIPv2	99.18 / 95.45	**0.84**	0.113 / 0.256	90.00 / 89.55	**25.9**	0.366 / 0.687
Swin Transformer	98.09 / 92.56	1.36	0.133 / 0.289	-	-	-
EfficientDet-D8	94.55 / 90.08	3.20	0.140 / 0.400	-	-	-

表 IV 使用 LLaVA-v1.6-13B 的移动性和阻塞预测性能。时延和功耗根据 Qualcomm AI 100 Edge 进行估算。

	定位误差（手机）			遮挡预测	资源使用	
	距离（cm）	方位角（度）	仰角（度）	准确率（%）	时延（ms）	功率（W）
LLaVA-v1.6-13B	12.78	0.27	0.1	91.7	25	15
基于深度学习的轨迹预测 [B3]	45.66	8.5	7.86	-	31	25
基于Sage-Husa滤波器的轨迹预测 [B4]	22.01	14.26	13.66	-	30	20
基于深度学习的阻塞预测	-	-	-	82.4	31	25

为比较调度性能，我们采用四种基准方案：

1) 基于深度强化学习的抢占式调度 [20]：基站利用深度强化学习，为第 T 个（未来）时隙确定每个用户设备的最优调制与编码方案索引和资源块分配。这些调度决策基于用户设备过去的估计信道和信噪比。2) 基于轮询的反应式调度 [47]：基站为每个用户设备分配相同数量的资源块，这意味着部分用户设备的速率要求可能无法满足。请注意，调制与编码方案索引根据第（$T-1$）个时隙的信噪比选取。3) 基于最大信噪比的反应式调度 [26]：基站优先向第（$T-1$）个时隙中信噪比较高的用户设备分配资源块。该过程持续进行，直到分配完所有资源块，且不考虑所有用户设备的公平性或服务质量保障。4) 基于比例公平的反应式调度 [48]：基站根据第（$T-1$）个时隙中观测到的用户设备信噪比做出调度决策，分配资源块以兼顾系统总吞吐量和服务质量需求满足度。

B. 仿真结果

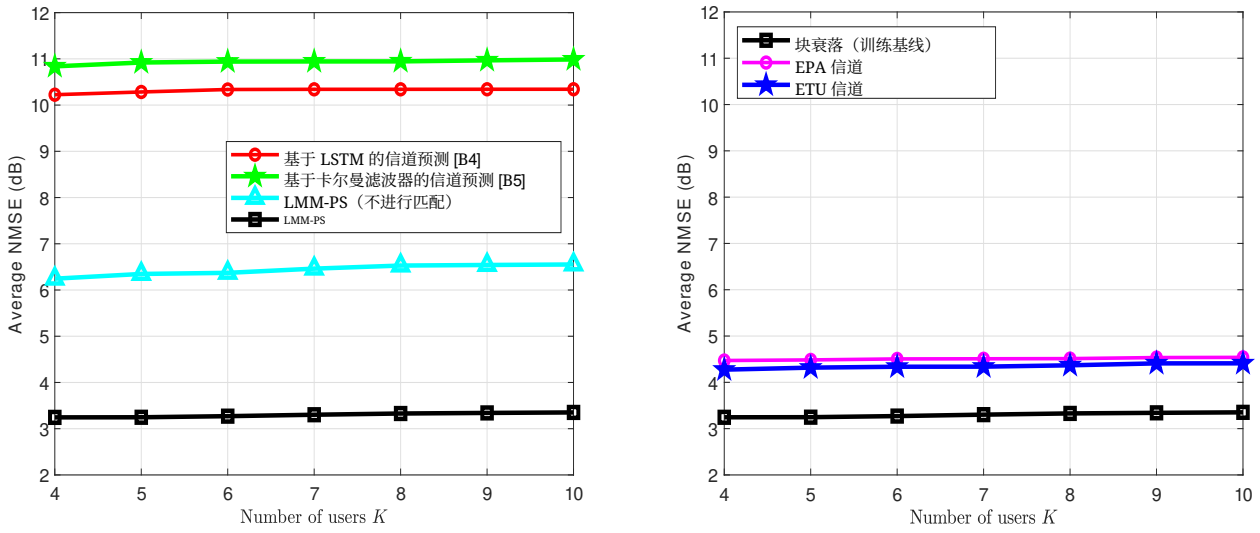
在表 III 中，我们从召回率（即在所有目标对象中被检测到的对象所占的百分比）方面展示了用户设备和手机检测性能。我们观察到，GLIPv2 优于 Swin Transformer 和 EfficientDet，在检测人物和手机时的平均召回率均超过 97%。通过利用图像和文本输入，GLIPv2 能够理解输入句子，并识别句中提及对象的图像位置，从而提升仅依赖图像输入的传统目标检测器的性能。

我们还在表 III 中展示了障碍物检测性能。结果表明，在障碍物检测方面，GLIPv2 的精确率（即正确检测到的目标占所有检测到目标的比例）和召回率均达到约 90%。由于传统目标检测器经过训练只能识别预定义的目标类别，因此难以检测训练数据中未被明确标注或未包含的障碍物。

在表 IV 中，我们评估了 LLaVA-v1.6-13B 的移动性预测性能。我们观察到，LLaVA 能够准确预测大尺度视距信道参数（距离和角度），平均距离误差小于 15 cm。此外，我们将所提出的 LMM-PS 的遮挡预测准确率与一种基于深度学习的遮挡预测方案进行比较，该方案使用长短期记忆（LSTM）网络，根据一系列过去的估计信道预测未来的遮挡状态。表 IV 中的结果表明，LMM-PS 利用 LLaVA 的移动性预测结果，其遮挡预测准确率相比基于 LSTM 的方案提高了 10% 以上，这得益于其能够通过视觉感知信息快速识别周围无线环境的能力。

我们还可以从表 IV 中看到，LMM-PS 的时延和功耗远低于传统技术。由于利用 LLaVA 从捕获图像中预测未来的用户设备位置，实质上，传统基于射频的轨迹预测技术的波束扫描时延被 LMM 推理时延取代，从而降低了 LMM-PS 的时延和功耗。例如，采用最新的面向人工智能的片上系统 Qualcomm AI 100 Edge 时，推理时间约为 25 ms，而传统基于射频技术的波束扫描时延约为 30 ms [50]。同样，在功耗方面，波束扫描阶段通常消耗 20 W [50]，而 Qualcomm AI 100 Edge 的功耗约为 15 W，这表明 LMM-PS 有望实现显著节能。

为验证所提出的 Transformer 辅助模型在视距路径增益和非视距信道参数估计方面的有效性，我们进一步考察了 LMM-PS 的信道预测性能，并以平均归一化均方误差衡量（见图12(a)）。我们还对两种近期方法进行了性能评估：[51] LSTM辅助信道预测和 [52] 基于卡尔曼滤波器的信道预测（见图12(a)中的 [51] 和 [52] ）。为确保公平比较，在所有测试方案中，我们都使用 LMM 预测的大尺度视距信道参数（视距路径的距离和角度）作为视距信道预测的输入。从图12(a)中可以看出，LMM-PS 的表现始终优于 [51] 和 [52]。由于 Transformer 会为与输出值（即未来信道参数）高度相关的输入数据（即过去的信道）分配相对较大的注意力权重



(a) (b)

图12。（a ）信道预测技术的平均 NMSE，以及（b）不同信道模式下 LMM-PS 的平均 NMSE ls.

因此，即使环境突然变化，LMM-PS 仍能有效捕捉往往得以保留的长期空间相关性，从而显著改善信道预测过程。

为展示 LMM-PS 在不同信道条件下的有效性，我们测试了其在扩展行人 A（EPA）和扩展典型城区（ETU）信道下的信道预测性能（见图12(b)）。我们观察到，LMM-PS 在这些模型下的整体性能与其在块衰落信道下的性能相似。例如，当 $K = 4$ 时，ETU 信道下的 LMM-PS 达到约 4.27 dB 的平均 NMSE，与块衰落信道下 LMM-PS 所达到的 NMSE（3.25 dB）相当。

在图13(a)中，我们绘制了 LMM-PS 的系统总吞吐量随 $\text{SER}_{\max}$ 变化的曲线。我们确认，LMM-PS 全面优于传统技术。例如，在 $\text{SER}_{\max} = 10^{-2}$ 时，LMM-PS 相较于基于深度强化学习的调度实现了 11% 的吞吐量增益。这主要是因为 LMM-PS 使用 LLaVA 预测用户设备的未来移动情况，并使用 Transformer 预测强非视距路径的信道参数，从而帮助 LMM-PS 预测未来的视距和非视距信道分量，用于计算未来的比例公平度量。

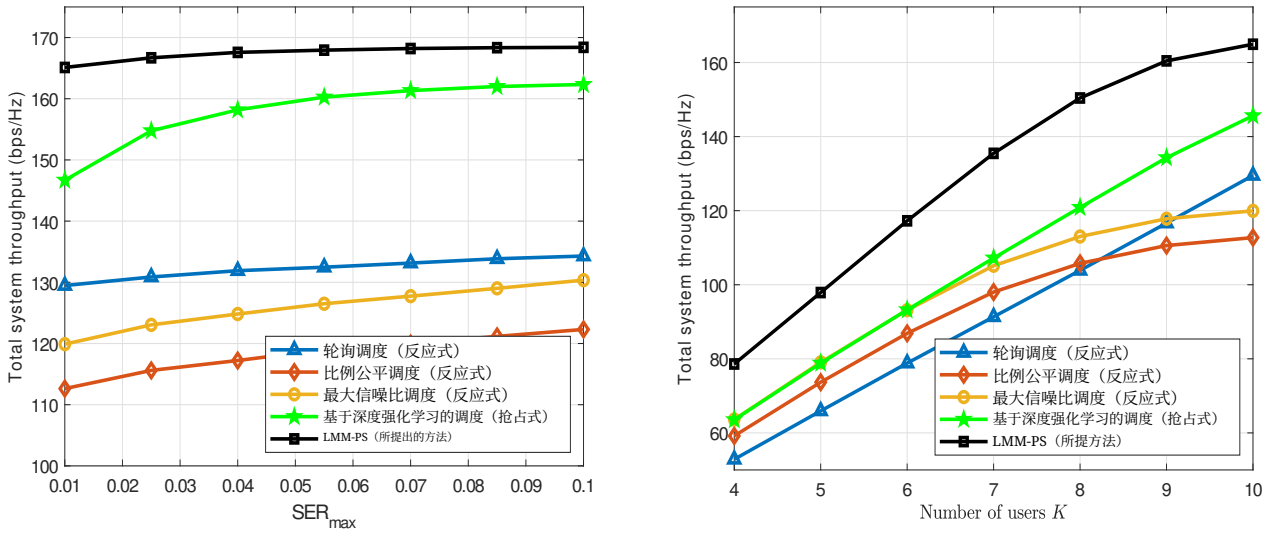
在图13(b)中，我们绘制了 LMM-PS 的系统总吞吐量随用户设备数量 K 变化的曲线。我们观察到，与传统技术相比，LMM-PS 实现了显著更高的吞吐量。仅依赖过去估计信道的传统技术通常无法捕捉无线环境中的突发信道变化（例如遮挡），而 LMM-PS 则能够利用视觉感知信息和大型多模态模型理解这些事件，从而显著提升系统总吞吐量。例如，当 $K = 7$ 时，LMM-PS 相较于轮询调度技术实现了 32% 的吞吐量增益。

为了评估 LMM-PS 的泛化能力，我们使用专为移动场景设计的视觉数据集 VOMM 数据集 [3]，重新评估了用户设备和障碍物的定位性能（见表V）。在障碍物检测评估中，我们为 VOMM 图像中的每个障碍物手动标注了边界框（见 https://github.com/islab-github/VOMM）。实验结果证实，LMM-PS 在 VOMM 中的定位性能与在 VOBEM2 中相当。根据表V中的定位结果，我们进一步评估了 LMM-PS 在各种移动场景下的吞吐量性能（见图 14(a)）。结果显示，LMM-PS 在 VOMM 中的吞吐量性能与在 VOBEM2 中相当。例如，当用户设备速度为 $v = 12$km/h 时，LMM-PS 在 VOMM 中可达到约 167.8 bps/Hz，与其在 VOBEM2 中实现的系统总吞吐量（168.4 bps/Hz）非常接近。

最后，为了评估 LMM-PS 在各种环境下的鲁棒性，我们通过部署密度各异的障碍物 ρ_o，设计了新的仿真空间（使用 AutoCAD 构建）。其中，XY 平面上障碍物所占面积与服务区域总面积之比即为障碍物密度。随后，我们在这些空间中测试了 LMM-PS 的吞吐量性能（见图 14(b)）。结果表明，与传统策略相比，LMM-PS 显著提升了系统总吞吐量。例如，当 $\rho_o = 20\%$ 时，LMM-PS 相较于基于轮询的技术，系统总吞吐量提升了 26%。

五、结论

本文提出了一种面向未来 6G 自主通信系统的 LMM 辅助抢占式调度技术，称为 LMM-PS。与仅依赖射频信号的传统方法不同，LMM-PS 利用视觉感知信息和大型多模态模型，快速检测并



(a) (b)

图13. (a) 系统总吞吐量随 $\text{SER}_{\max}$ ($K = 10$, $v = 12$km/h) 的变化，以及 (b) 系统总吞吐量随用户设备数量的变化（$\text{SER}_{\max} = 10^{-2}$, $v = 12$km/h）。

表V 使用 GLIPv2 和 VOMM 数据集进行目标检测和定位的性能。

	手机	定位误差（手机）		障碍物	定位误差	
	召回率（%）	距离（cm）	方位角/仰角（度）	精确率/召回率（%）	距离（厘米）	方位角/仰角（度）
GLIPv2	**99.57**	**0.67**	0.090 / 0.223	92.28 / 88.67	**46.8**	0.569 / 0.929
Swin Transformer	95.26	3.64	0.142 / 0.400	-	-	-

分析信号传播环境。通过预测移动设备的未来位置和突发事件，并将其与过去的基于射频的信道信息相结合，LMM-PS 能为抢占式调度生成准确的未来信道估计。数值实验表明，就系统总吞吐量而言，LMM-PS 显著优于传统的基于射频的调度技术。尽管当前版本的 LMM-PS 未采用最新的大型多模态模型（例如 GPT-4V 和 LLaMA-VID），但将现有骨干网络替换为能够处理新型输入（例如图像序列）的模型，将是一个很有前景的研究方向。我们认为，所提出的 LMM-PS 是一种有效工具，可支持新兴的基于 LLM/LMM 的自主通信范式；在该范式中，各项任务通过人工智能、多模态感知和无线通信的协同来执行。

参考文献

[1] S. Kim、J. Son 和 B. Shim，“面向自主通信系统的基于多模态大模型的调度”，即将在 IEEE Veh. Technol. Conf. (VTC) 发表，2025年。[2] Y. Chang 等，“大语言模型评估综述”，ACM Trans. Intell. Syst. Technol.，第 15 卷，第 3 期，第 1–45 页，2024年3月。[3] M. Xu 等，“资源高效型大语言模型与多模态基础模型综述”，arXiv 预印本 arXiv:2401.08092，2024年。[4] Z. Yang、X. Jia、H. Li 和 J. Yan，“LLM4Drive：面向自动驾驶的大语言模型综述”，arXiv 预印本 arXiv:2311 01043，2023年。[5] M. Ahn 等，“AutoRT：用于大规模编排机器人智能体的具身基础模型”，arXiv 预印本 arXiv:2401.12963，2024年。[6] G. Wang 等，“Surgical-LVLM：学习使大型视觉语言模型适应机器人手术中的落地式视觉问答”，arXiv 预印本 arXiv:2401.08092，2024年。[7] H. Zhou 等，“用于电信的大语言模型（LLM）：原理、关键技术与机遇综述”，IEEE 通信调查与教程，卷 27，第 3 期，页 1955–2005，2025年6月。[8] F. Jiang、Y. Peng、L. Dong、K. Wang、K. Yang、C. Pan、D. Niyato 和 O. A. Dobre，“大语言模型增强的 6G 通信多智能体系统”，IEEE 无线通信，卷 31，第 6 期，页 48-55，2024年12月。[9] Y. Ahn、J. Kim、S. Kim、S. Kim 和 B. Shim，“面向 6G 毫米波与太赫兹通信系统的感知与计算机视觉辅助移动性管理”，IEEE通信汇刊，卷 72，第 10 期，页 6044-6058，2024 年 10 月。[10] R. Zeng 、 T. Liu 、X. Yu 和 Z. Zhang，“高移动性环境下自适应调制与编码的新型信道质量指示预测方案”，IEEE Access，卷 7，页 11543-11553，2019年。[11] A. Balieiro、K. Dias 和 P. Guarda，“通过机器学习与进化计算解决 5G/6G网络中的 CQI反馈时延问题”，《智能与融合网络》，卷 3，第 3 期，页 271-281，2022年9月。[12] Q. He、G. Dan 和 G. P. Koudouridis，“基于短期流量预测的 5G下行链路半持续调度”，发表于 IEEE 全球通信大会（GLOBECOM）论文集，2020年。[13] A. Vaswani 等，“Attention Is All You Need”，《神经信息处理系统进展》，2017年。[14] 3GPP 技术规范 38.212，“5G；新空口；复用与信道编码”，v15.7.0，2019年9月。[15] A. Omri、M. Shaqfeh、A. Ali 和 H. Alnuweiri，“5G NR 系统中的同步流程”，IEEE Access，卷 7，页 41286–41295，2019年。[16] S. Kim、J. Wu 和 B. Shim，“面向毫米波可重构智能表面辅助通信的高效信道探测与相移控制”，IEEE无线通信汇刊，卷 23，第 1 期，页 231–246，2024年1月。[17] J. Son、I. Keum、H. Kim、H. Cho 和 B. Shim，“非视距场景下基于Transformer的环境感知定位”，发表于 IEEE 无线通信与网络会议（WCNC）论文集，2024年。[18] X. Ye、Y. Yu 和 L. Fu，“用于 LTE/新空口系统的基于深度强化学习的链路自适应技术”，IEEE车辆技术汇刊，卷 72，第 6 期，页 7364–7379，2023年6月。

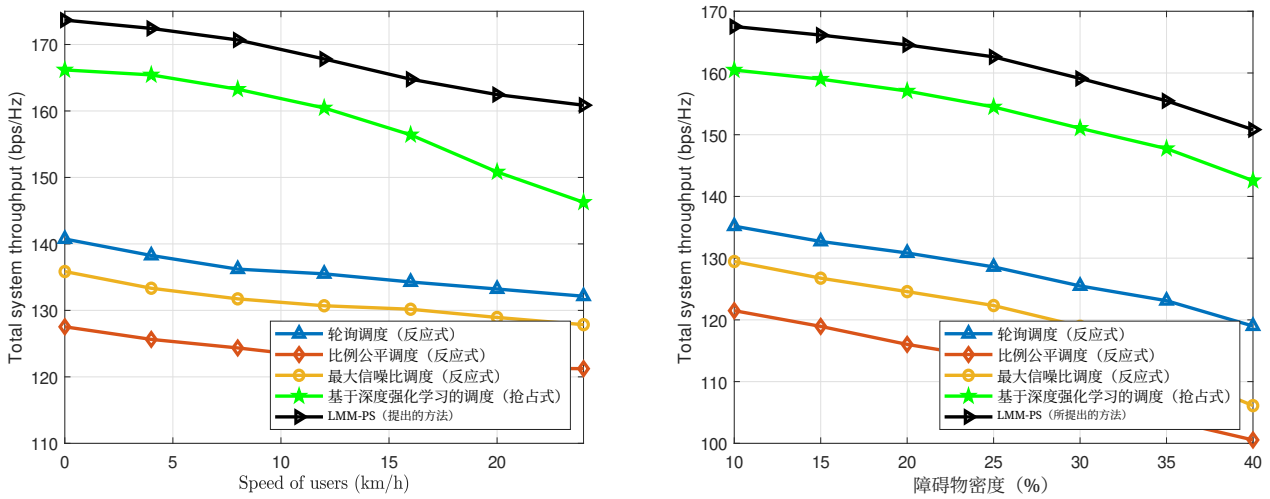


图 14. (a) 系统总吞吐量随用户设备速度的变化（$K = 10$，符号错误率$_{\max} = 10^{-1}$）；(b) 系统总吞吐量随障碍物密度的变化（$K = 10$，符号错误率$_{\max} = 10^{-1}$，$v = 12$km/h）。

[19] S. Jeong、H. Ju、S. Kim 和 B. Shim，“面向 6G 大规模 MIMO 系统的自动化环境感知信道反馈”，发表于 IEEE 全球通信研讨会（GC Wkshps）论文集，2023年。[20] Y. Liao、Z. Yang、Z. Yin 和 X. Shen，“面向 5G 大规模 MIMO-OFDM 下行链路的基于深度 Q 网络的自适应调制编码方案与空分复用”，IEEE通信快报，卷 27，第 1 期，页 185–189，2023年1月。[21] 3GPP 技术规范 38.214，“5G；新空口；控制信道物理层流程”，v15.6.0，2019年6月。[22] 3GPP 技术规范 38.214，“5G；新空口；数据信道物理层流程”，v16.2.0，2020年7月。[23] E. Bobrov、D. Kropotov、H. Lu 和 D. Zaev，“使用在线深度学习算法的大规模 MIMO 自适应调制与编码”，IEEE通信快报，卷 26，第 4 期，页 818–822，2022年4月。[24] Y. Ahn、J. Kim、S. Kim 和 B. Shim，“面向 6G 太赫兹通信的计算机视觉辅助主动移动性管理”，发表于 IEEE 全球通信大会（GLOBECOM）论文集，2023年。[25] 3GPP 技术规范 38.211，“5G；新空口；数据信道物理层流程”，v16.2.0，2018年6月。[26] H. Oh 和 H. Nam，“混合脉冲噪声与加性高斯白噪声环境下采用自适应调制的最大速率调度”，IEEE无线通信汇刊，卷 20，第 5 期，页 3308–3320，2021年5月。[27] A. J. Goldsmith，《无线通信》，英国剑桥：剑桥大学出版社，2005年。[28] Y. Chen、Y. Wu、Y. T. Hou 和 W. Lou，“mCore：为 5G 多用户 MIMO 系统实现亚毫秒级调度”，发表于 IEEE 计算机通信会议论文集，2021年。[29] H. Zhang、P. Zhang、X. Hu、Y-C. Chen、L. H. Li、X. Dai、L. Wang、L. Yuan、J-N. Hwang 和 J. Gao，“GLIPv2：统一定位与视觉语言理解”，发表于《神经信息处理系统进展》，卷 35，2022年。[30] H. Liu、C. Li、Y. Li、Y. Lee，“通过视觉指令微调改进基线”，发表于计算机视觉与模式识别会议（CVPR）论文集，页 26296–26306，2024年。[31] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang 和 W. Chen，“LoRA：大语言模型的低秩适配”，发表于国际学习表征会议（ICLR）论文集，2021年。[32] Y. Barayan 和 I. Kostanic，“LTE 中比例公平调度的性能评估”，发表于世界工程与计算机科学大会论文集，卷 2，页 712–717，2013年。[33] D. Alexey 等，“一幅图像胜过 16×16 个词：用于大规模图像识别的 Transformer”，发表于 arXiv 预印本 arXiv:2010.11929，2020年。[34] J. Devlin、M. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向 Transformer 预训练”，发表于 arXiv 预印本 arXiv:1810.04805，2019年。[35] S. Song、X. Li、S. Li、S. Zhao、J. Yu、J. Ma、X. Mao、W. Zhang 和 M. Wang，“如何弥合模态之间的差距：多模态综述大型语言模型，”IEEE知识与数据工程汇刊，第37卷，第9期，第5311–5329页，2025年9月。[36] S. Kim、Y. Ahn、D. Park和B. Shim，“VOMTC：用于毫米波与太赫兹通信的视觉目标，”IEEE认知通信与网络汇刊，第11卷，第1期，第243–257页，2025年2月。[37] S. Kim、J. Kim、J. Lee和B. Shim，“计算机视觉辅助毫米波与太赫兹通信波束成形，”韩国通信与信息科学学会期刊，第49卷，第12期，第1685–1694页，2024年12月。[38] H. W. Kuhn，“用于指派问题的匈牙利算法，”海军研究后勤（NRL），第52卷，第1期，第7–21页，2005年2月。[39] Y. Liu、J. Wu、S. Kim和B. Shim，“室内毫米波与太赫兹通信的视觉辅助阻塞预测与主动切换，”载于IEEE全球通信会议（GLOBECOM）论文集，2023年。[40] 3GPP TR 38.901，“0.5至100吉赫兹频率范围的信道模型研究”，v17.0.0，2022年。[41] D. P. Kingma和J. Ba，“Adam：一种随机优化方法，”arXiv预印本 arXiv:1412.6980，2014年。[42] S. Kim、J. Moon、J. Kim、Y. Ahn、D. Kim、S. Kim、K. Shim和B. Shim，“感知与计算机视觉在6G无线通信中的作用，”IEEE Wireless Commun.，第31卷，第5期，第264–271页，2024年10月。[43] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin和B. Guo，“Swin Transformer：使用移位窗口的层次化视觉变换器，”载于IEEE/CVF计算机视觉国际会议（ICCV）论文集，第10012–10022页，2021年。[44] M. Tan、R. Pang和Q. V. Le，“EfficientDet：可扩展且高效的目标检测，”载于计算机视觉与模式识别会议（CVPR）论文集，第10781-10790页，2020年。[45] C. Wang、L. Ma、R. Li、T. S. Durrani和H. Zhang，“通过机器学习方法探索轨迹预测，”IEEE Access，第7卷，第101441-101452页，2019年。[46] Y. Yang和W. Gao，“一种最优自适应卡尔曼滤波器，”《大地测量学杂志》，2006年。[47] H. M. Chaskar和U. Madhow，“具有可调时延的公平调度：一种轮询方法，”IEEE/ACM网络汇刊，第11卷，第4期，第592–601页，2003年8月。[48] R. Kwan、C. Leung和J. Zhang，“LTE中的比例公平多用户调度，”IEEE信号处理快报，第16卷，第6期，第461–464页，2009年6月。[49] M. Alrabeiah和A. Alkhateeb，“使用6 GHz以下信道进行毫米波波束与阻塞预测的深度学习，”IEEE通信汇刊，第68卷，第9期，第5504–5518页，2020年9月。[50] Y. Ahn、J. Kim、S. Kim、K. Shim、J. Kim、S. Kim和B. Shim，“迈向面向6G的智能毫米波与太赫兹通信：计算机视觉辅助波束成形，”IEEE Wireless Commun.，第30卷，第5期，第179-186页，2023年10月。[51] Y. Yang、D. B. Smith和S. Seneviratne，“深度学习信道

预测用于无线体域网中的发射功率控制，”载于IEEE国际通信会议论文集，2019年。[52] H. Kim、S. Kim、H. Lee、C. Jang、Y. Choi和J. Choi，“大规模MIMO信道预测：卡尔曼滤波与机器学习，”IEEE通信汇刊，第69卷，第1期，第518–528页，2021年1月。