融合语音-视觉-手势多模态时空定位的车载兴趣点智能识别方法

李振龙　王洪磊　乔万顺　高泽磊　张耀

（一汽奔腾汽车股份有限公司，长春 130012）

【摘要】针对车载场景下单一语音交互模式存在的噪声敏感、意图模糊和方向信息缺失等问题，提出一种融合语音、视觉与手势的多模态车载兴趣点（POI）智能识别方法。构建了包含数据采集、预处理、核心算法与交互输出的系统架构：在语音意图识别环节，采用低秩适应（LoRA）方法微调通义千问-7B（Qwen-7B）模型，实现结构化意图输出；在时空定位环节，设计基于扇形区域的多策略检索机制，将语音方向词与视线方向、手势指向或导航路线相融合，使候选POI规模缩减74.47%；在综合预测环节，利用通义千问-VL（Qwen-VL）实现跨模态注意力融合，完成POI排序推荐。在自建车载多模态数据集上的试验结果表明，所提出方法的准确率达85.4%，F1值达84.3%，较最优基线方法分别提升8.6百分点和7.7百分点，且在信噪比0 dB条件下仍保持70.1%的准确率，验证了多模态融合对车载噪声环境的鲁棒性。

关键词：多模态融合　车载交互　兴趣点识别　大语言模型　时空定位

中图分类号：U495；TP391.41　　**文献标志码：**A　　**DOI：**10.20104/j.cnki.1674-6546.20260127

An Intelligent Recognition Method for In-Vehicle Points of Interest Fusing Audio-Visual-Gesture Multimodal Spatiotemporal Localization

Li Zhenlong, Wang Honglei, Qiao Wanshun, Gao Zelei, Zhang Yao

(FAW Car Co., Ltd., Changchun 130012)

【Abstract】To address the issues of noise sensitivity, ambiguous intent, and missing directional information in single-modality voice interaction for in-vehicle scenarios, the paper proposes a multimodal Point-of-Interest (POI) intelligent recognition method that fuses speech, vision, and gesture. The method establishes a four-layer system architecture comprising data collection, preprocessing, core algorithm, and interaction output. For speech intent recognition, Low-Rank Adaptation (LoRA) fine-tuned Qwen-7B is employed to generate structured intent output. For spatiotemporal localization, a fan-shaped region-based multi-strategy retrieval mechanism is designed, which integrates speech directional words with gaze direction, gesture pointing or navigation route, reducing the candidate POI scale by 74.47%. For comprehensive prediction, Qwen-VL is utilized to achieve cross-modal attention fusion for POI ranking. Experimental results on a self-built in-vehicle multimodal dataset demonstrate that the proposed method achieves an accuracy of 85.4% and an F1-score of 84.3%, improving by 8.6 and 7.7 percentage points respectively over the best baseline. Under 0 dB SNR conditions, the accuracy remains at 70.1%, confirming the robust effectiveness of multimodal fusion in vehicular noise environments.

Key words: Multimodal fusion, In-vehicle interaction, POI recognition, Large language model, Spatiotemporal localization

【引用格式】李振龙, 王洪磊, 乔万顺, 等. 融合语音-视觉-手势多模态时空定位的车载兴趣点智能识别方法[J]. 汽车工程师, 2026(10): 26-33.

LI Z L, WANG H L, QIAO W S, et al. An Intelligent Recognition Method for In-Vehicle Points of Interest Fusing Audio-Visual-Gesture Multimodal Spatiotemporal Localization[J]. Automotive Engineer, 2026(10): 26-33.

1　前言

随着智能网联汽车的快速发展，语音交互已成为车载信息系统的标配功能，但实际驾驶场景中的兴趣点（Point of Interest，POI）检索仍面临噪声敏感、方向信息缺失和意图模糊等挑战。Zhang等[1]的研

究表明：当信噪比低于10 dB时，主流语音识别（Automatic Speech Recognition，ASR）系统的词错率上升超过40%；驾驶员的POI需求往往伴随空间方向性（如“左边那个加油站”），纯语音通道无法传递此类空间语义，导致系统仅能返回全向POI列表，进一步加重认知负荷。

针对上述问题，学术界已开展多模态融合探索。Kunst等[2]通过实车眼动测试分析驾驶员在复杂路况下的视觉搜索策略，验证了眼动轨迹与驾驶员导航目标选择意图之间的强耦合性。陈启博等[3]提出基于深度相机的手势控制车载信息娱乐系统，在停车场景下取得了较好的识别效果。在POI推荐领域，Darapisut等[4]将位置服务（Location-Based Service，LBS）与协同过滤相结合，实现了基于位置的POI推荐，验证了地理位置在推荐精度提升方面的效果，但仍存在冷启动问题。在车载语音交互方面，陈勇等[5]将双向编码器表征法（Bidirectional Encoder Representations from Transformers，BERT）应用于车载指令理解，在固定指令集上取得了95%以上的准确率，但开放域泛化能力有限；Dong等[6]利用大语言模型（Large Language Model，LLM）进行交互式推荐，展示了LLM在意图理解方面的优势，但该方法的云端部署延迟难以满足车载实时性要求。在多模态融合方面，Li等[7]提出注意力引导的跨模态融合网络，然而现有工作多采用决策层简单投票方式，未能充分利用模态间的互补关系。在POI推荐方面，Zhang等[8]利用Transformer对用户长短期偏好进行建模，在序列推荐任务上达到最新技术水平（State Of The Art，SOTA），但这些方法依赖后端结构化数据，未考虑驾驶员的方向性需求。

综合来看，现有研究存在两个方面的关键不足：多模态信号的时空对齐与融合机制尚不完善；缺乏面向车载场景的端到端POI识别方案。为此，本文提出一种融合语音、视觉与手势的多模态车载POI智能识别方法，设计4层POI识别系统架构，实现端到端处理，提出基于扇形区域的多策略时空定位方法以大幅缩减候选POI数量，采用低秩适应（Low-Rank Adaptation，LoRA）微调通义千问-7B（Qwen-7B）实现结构化意图识别，并结合通义千问-VL（Qwen-VL）进行跨模态综合预测，从而提高识别准确率。

2 车载多模态POI识别方法

2.1 系统架构

如图1所示，本文提出的车载POI识别系统架构采用4层设计：数据采集层负责获取行驶数据、语音数据、视频数据和LBS数据等多源异构信息；数据预处理层完成语音识别、情绪分析、机器视觉处理（手势检测、视线追踪、表情识别）及数据对齐；核心算法层包含语音意图识别、多模态时空定位和多模态综合预测3个核心模块；交互输出层实现语音播报、人机交互（Human-Machine Interaction，HMI）界面展示、多轮对话和用户反馈。各层之间通过标准化接口传递结构化数据，确保系统的模块化结构和可扩展性。

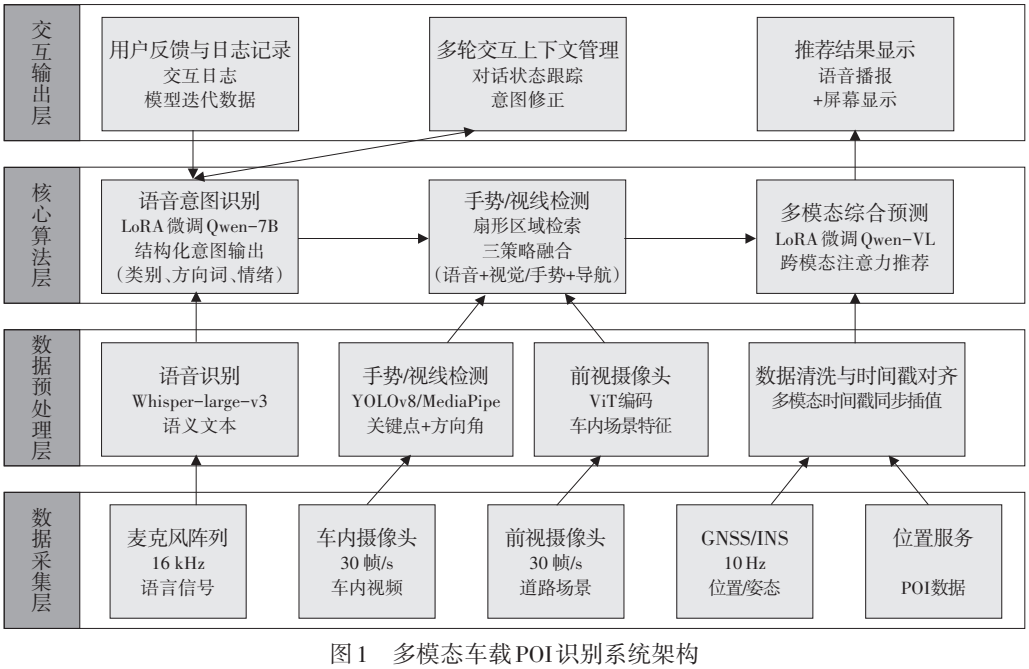


图1 多模态车载POI识别系统架构

在硬件部署方面，数据采集层依赖车载传感器：麦克风阵列采集16 kHz语音信号，车内摄像头以30帧/s的速率采集驾驶员面部和手势视频，前视摄像头采集道路场景，全球导航卫星系统（Global Navigation Satellite System，GNSS）/惯性导航系统（Inertial Navigation System，INS）模块提供定位和航向信息。预处理层和核心算法层部署于高通骁龙8295计算平台，该平台集成海克斯康（Hexagon）神经处理单元（Neural Processing Unit，NPU），可支持7 B参数模型的4位整数（4-bit Integer，INT4）量化推理，单次推理延迟约为1.2 s。需要指出的是，Qwen-7B的LoRA微调在云端英伟达（NVIDIA）A100图形处理单元（Graphics Processing Unit，GPU）上完成，微调后的适配器参数仅占原模型的0.6%，可便捷地通过空中下载（Over The Air，OTA）推送到车端。

2.2 语音意图识别

语音意图识别模块的目标是将驾驶员的自然语言输入转化为结构化意图表示。该模块的处理流程为：原始语音经Whisper-large-v3模型转写为文本，同时提取韵律特征（语速、音高变化、停顿模式）用于情绪状态判断。转写文本送入LoRA微调的Qwen-7B模型，输出包含5个字段的结构化意图：情绪状态（兴奋/好奇/焦虑/平静）、POI类别（停车场/加油站/美食/景点/服务/购物/酒店）、位置关键词（如“附近”“前方”“2公里”）、方向词（如“左边”“右侧”“前面”）、距离词（如“最近的”“3公里内”）。

LoRA微调的关键设计如下：在Qwen-7B的查询矩阵$\boldsymbol{Q}$和值矩阵$\boldsymbol{V}$上添加秩为16的低秩适配器，训练时冻结原模型参数，仅更新LoRA权重。训练数据为人工标注的2 400条车载POI对话语料，涵盖7个POI类别和4种情绪状态。训练超参数设置为：学习率为2×10^{-4}，批量大小为4，训练3个轮次（Epoch），学习率预热（Warmup）步数为100。微调后，在测试集上的意图分类准确率为89.3%，字段级F1值为86.7%。与全参数微调相比，LoRA方案在性能上仅低1.2百分点，但训练成本降低约70%。

2.3 多模态时空定位

多模态时空定位旨在解决“驾驶员指向某方向的POI”这一空间检索问题。该方法的基本思路是：根据语音意图中提取的方向词，结合视线或手势提供的空间方向信息，在驾驶员当前位置构建扇形检索区域，仅返回该区域内的候选POI。多模态时空定位流程如图2所示，根据可用辅助模态的不同，设计了3种融合策略。

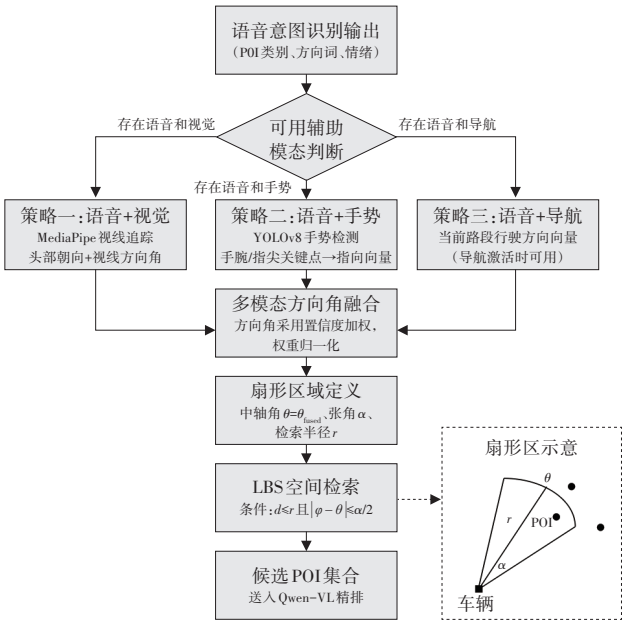


图2 多模态时空定位流程

策略1：语音+视线方向融合。当驾驶员说出包含方向词的语音指令（如“左边那个加油站”）时，系统同步提取媒体管道（MediaPipe）视线追踪模块输出的头部朝向和视线方向角。将语音方向词映射为基准角度（“左”对应车辆左侧90°，“右前”对应右前45°），再与视线方向角进行加权平均，权重根据视线置信度动态调整（置信度高于0.8时视线权重取0.7，否则取0.3）。融合后的方向角作为扇形中轴线，构建张角60°、半径200 m的扇形检索区域。

策略2：语音+手势指向融合。当检测到驾驶员手势指向动作时，YOLOv8手势检测模块输出手腕和指尖关键点坐标，通过向量计算得到手指延伸方向。该方向与语音方向词的融合方式与策略1类似，区别在于手势指向的空间分辨率通常高于视线方向，因此在置信度足够时（手势检测置信度>0.85），手势方向的权重可提升至0.8。

策略3：语音+导航路线融合。当车辆处于导航状态时，取当前路段的行驶方向向量作为辅助方向。该策略适用于驾驶员说出“前面”“前方”等与行驶方向一致的词，此时扇形中轴线与行驶方向对齐，检索车辆前方的POI。

当多种辅助模态同时可用时，系统按优先级选择，即手势指向>视线方向>导航路线，因为手势的空间指向性最为明确。

扇形区域内的POI通过LBS检索获得。设中轴角为$\theta=\theta_{\text{fused}}$（其中$\theta_{\text{fused}}$为融合方向角，由方向角经置

信度加权和权重归一化处理获得）、扇形张角为α、检索半径为r，则候选POI需满足：与当前位置的距离$d\leqslant r$，且POI相对方位角φ满足$|\varphi-\theta|\leqslant\alpha/2$。经过扇形过滤后，候选POI从平均47个缩减至12个，大幅降低了后续排序模块的计算负担。

2.4 多模态综合预测

多模态综合预测模块负责对候选POI进行精确排序，输出前3位推荐结果。该模块以Qwen-VL为核心，输入包括语音意图的结构化文本描述、候选POI的文本属性（名称、类别、评分、距离）和前视摄像头捕获的道路场景图像。Qwen-VL利用跨模态注意力机制将文本语义与视觉特征进行深度融合：当驾驶员说“那个看起来不错的餐厅”时，视觉分支提取道路场景中可见的招牌信息，与文本分支的POI属性进行交叉注意力计算，从而将视觉上显著匹配的POI排序提前。

具体实现上，将N个候选POI的属性信息拼接为文本序列，与语音意图文本共同作为文本输入，道路场景图像经视觉转换器（Vision Transformer，ViT）编码后作为视觉输入。Qwen-VL的跨模态注意力层允许文本词元（Token）关注视觉词元，从而实现视觉信息对文本排序的引导。模型输出为各候选POI的相关性评分，经归一化指数（Softmax）函数进行归一化处理后取前3位作为推荐结果。在实际部署中，Qwen-VL同样采用INT4量化推理，单次排序耗时约0.3 s。

Qwen-VL的微调实现如下：在视觉编码器ViT和语言模型Qwen-7B的查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$、值矩阵$\boldsymbol{V}$上添加秩为32的LoRA适配器，训练时冻结原模型参数，仅更新LoRA权重，可训练参数量约占原模型的1.2%。训练数据为人工标注的3 600条语音意图-候选POI列表-道路场景图像-正确POI标签四元组样本，覆盖7个POI类别和不同光照（白天/夜间/逆光）、天气（晴/雨/雾）条件下的道路场景。输入格式采用Qwen-VL的聊天标记语言（Chat Markup Language，ChatML）模板，文本输入为包含语音意图的结构化描述和候选POI属性列表，视觉输入为前视摄像头捕获的道路场景图像（统一缩放至分辨率448像素×448像素）。训练超参数设置为：学习率为1×10^{-4}，批量大小为8，训练5个轮次，Warmup步数为200，权重衰减为0.01，采用余弦退火学习率调度，梯度累积步数为2，LoRA随机失活（Dropout）概率为0.05。损失函数采用列表级最大似然损失（Listwise Maximum Likelihood Estimation，ListMLE）与信息噪声对比损失（Information Noise Contrastive Estimation，InfoNCE）的加权组合：

$$L=L_{\mathrm{LMLE}}+0.3L_{\mathrm{con}} \tag{1}$$

式中：L为损失率，L_{LMLE}为列表级最大似然损失，L_{con}为信息噪声对比损失。

ListMLE对候选POI的排序概率分布进行最大似然估计；对比损失通过InfoNCE形式拉近正确POI与视觉特征的语义距离，并增大错误POI与该特征的距离。训练硬件为8张NVIDIA A100 80 GB GPU，单次完整训练耗时约14 h。微调后，在测试集上的前3位命中率达92.1%，较未微调的Qwen-VL基线提升11.4百分点，验证了领域适配微调对车载POI排序任务的有效性。

2.5 多轮交互机制

考虑到驾驶员的POI需求可能在多轮对话中逐步明确，系统设计了多轮交互机制。当首轮推荐结果未被用户接受时，系统保留对话上下文（包括意图识别结果和已展示的POI），根据用户的反馈（如“太远了”“换一家”“有没有停车场”）更新意图表示，并重新触发检索和排序。具体而言：用户否定当前推荐时，系统在候选集中排除已展示POI后重新排序；用户追加约束时（如“便宜一点的”），将新约束作为补充条件叠加到意图表示中；用户切换类别时，重新执行意图识别和后续流程。对话上下文保留最近5轮完整对话历史。试验数据表明，85.4%的POI需求在3轮交互内得到满足，平均交互轮次为2.3轮。

3 车载多模态POI识别试验

3.1 试验设置

试验数据来自自建的车载多模态POI交互数据集。该数据集的采集过程如下：招募30名驾驶员（22名男性、8名女性，年龄为23~55岁）在北京市区及郊区道路进行真实驾驶测试，每人完成约40次POI检索任务，覆盖7个POI类别。采集设备包括：车载麦克风阵列（采样率16 kHz）、车内红外摄像头（帧速率30 帧/s，分辨率640像素×480像素）、前视摄像头（帧速率30 帧/s，分辨率1 920像素×1 080像素）及高精度GNSS/INS模块。数据集共包含1 186组多模态交互样本，按7∶1.5∶1.5的比例划分为训练集、验证集和测试集。数据标注由3名标注员独立完成，标注内容包括POI类别、方向信息、情绪状态

和目标POI，标注一致性卡帕系数（Kappa Coefficient）为0.83。

数据集中典型样本示例如下：

a. 样本1的语音指令为“左边那个加油站”，驾驶员视线方向角为-85°（车辆左侧），手势指向置信度为0.32（未检测到明确手势），目标POI为车辆左前方156 m处的加油站。

b. 样本2的语音指令为“前面那家餐厅”，车辆处于导航状态，行驶方向角为0°，目标POI为前方203 m处的某知名火锅店。

c. 样本3的语音指令为“找个停车场”，驾驶员视线方向角为45°（右前方），手势指向方向角为52°，目标POI为右前方178 m处的商场停车场。

d. 样本4的语音指令为“右边那个超市”，驾驶员视线方向角为80°（右侧），手势指向方向角为75°，目标POI为右前方95 m处的某大型连锁超市。

上述样本覆盖了语音+视线、语音+导航、语音+视线+手势等多种模态组合场景，体现了数据集在POI类别、方向信息和模态组合上的多样性。

对比方法包括：

a. 词频-逆向文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）+协同过滤[9]：基于文本特征和用户历史的传统推荐方法。

b. 文本卷积神经网络（Text Convolutional Neural Network，TextCNN）[10]+LBS检索：卷积网络提取文本特征后结合位置检索。

c. BERT[5]+多特征融合：采用BERT编码文本特征后与位置、时间特征拼接分类。

d. LLM对话推荐[6]+视觉辅助：利用Llama3-Instruct-8B进行对话式推荐，辅以视觉描述。

评估指标包括准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1值和平均倒数排名（Mean Reciprocal Rank，MRR）。试验环境为NVIDIA A100 GPU（训练）和高通骁龙8295（推理）。

3.2 消融试验

为验证各模态的贡献，设计了消融试验，结果如表1所示：仅使用语音模态时，准确率和F1值较低，这反映了单一语音通道在噪声和方向信息缺失条件下的局限性；加入视线模态后，准确率提升9.5百分点，主要得益于视线方向对POI检索范围的缩小；语音+手势组合模态的准确率为70.5%，略低于语音+视线组合模态，原因在于行驶过程中手势检测的稳定性不足；语音+导航组合模态的准确率为68.2%，提升有限，因为导航路线仅在车辆行驶方向上提供约束；语音+视线+手势组合模态的准确率达到78.6%；全模态融合（加入LBS属性信息）的准确率达到85.4%，F1值为84.3%，说明LBS属性信息（如评分、营业状态等）对最终排序有显著帮助。

表1 不同模态组合的消融试验结果 %

模态组合	准确率	精确率	召回率	F1值
仅语音	62.3	60.1	63.5	61.8
语音+视线	71.8	69.5	72.3	70.8
语音+手势	70.5	68.2	71.0	69.6
语音+导航	68.2	65.8	69.1	67.4
语音+视线+手势	78.6	76.5	79.2	77.8
全模态融合	85.4	84.7	82.5	84.3

3.3 对比试验

表2所示为本文方法与4种基线方法的对比结果。作为传统方法的代表，TF-IDF+协同过滤的性能主要受限于稀疏的用户行为数据，以及基于关键词匹配在语义理解方面的不足。TextCNN+LBS检索显著提升了准确率，表明卷积神经网络在文本特征提取上优于TF-IDF，但仍缺乏对方向性需求的有效支持。BERT+多特征融合进一步提高了准确率，BERT的上下文编码能力带来了明显提升，然而简单的多特征拼接方式未能充分挖掘模态信息间的互补关系。LLM对话推荐+视觉辅助在准确率上表现较优，大模型的强大理解能力带来了进一步提升，但云端推理延迟（平均2.87 s）和视觉信息依赖文本化描述的限制，制约了其实际应用效果。本文方法在准确率和MRR上均取得最优结果，分别达到85.4%和0.871，较LLM对话推荐+视觉辅助分别提升8.6百分点和0.079。该性能优势主要源自扇形区域检索有效缩小了候选集范围，以及Qwen-VL在跨模态注意力融合方面优于文本化视觉描述的方式。

表2 不同方法的对比试验结果

方法	准确率/%	精确率/%	召回率/%	F1值/%	MRR
TF-IDF+协同过滤	51.2	48.7	53.1	50.8	0.553
TextCNN+LBS检索	59.3	56.8	61.5	59.1	0.637
BERT+多特征融合	71.5	69.1	73.2	71.1	0.748
LLM对话推荐+视觉辅助	76.8	74.5	78.9	76.6	0.792
本文方法	85.4	84.7	82.5	84.3	0.871

3.4 多轮交互分析

图3所示为不同方法在多轮交互场景下的准确

率变化情况。本文方法的准确率在第1轮至第5轮交互中提升幅度逐轮递减，表明大部分POI需求可在3轮交互内得到满足。LLM对话推荐+视觉辅助方法呈现相似趋势，但整体准确率偏低，第5轮交互准确率仅为78.3%，该差距主要源于其视觉信息以文本描述形式输入，丢失了空间方位细节。BERT+多特征融合和TextCNN+LBS检索方法在多轮交互中的准确率提升幅度更为有限，原因在于这些方法缺乏有效的上下文管理机制，历史对话信息未能充分融入当前轮次的预测。

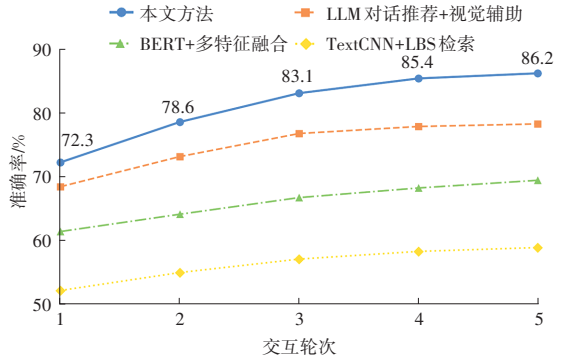


图3　多轮交互对POI识别准确率的影响

3.5　鲁棒性分析

车载环境的噪声水平随车速和路况变化显著，故鲁棒性是衡量方法实用性的关键指标，图4所示为不同噪声等级下的测试结果。在不同环境下，本文方法均获得了最高准确率，信噪比（Signal-to-Noise Ratio，SNR）≈0 dB的高噪声条件下准确率仍达到70.1%。相比之下，仅语音模态在高噪声条件下准确率骤降至31.5%，而本文方法在多模态融合下的降幅（15.3百分点）远小于单模态方法的降幅（30.8百分点），验证了视线和手势模态在语音通道退化时的补偿作用。具体而言，当ASR系统的词错率上升时，视线方向和手势指向仍可作为独立的空间线索缩小检索范围，即使语音意图识别出现偏差，扇形过滤仍能保留正确的候选POI。

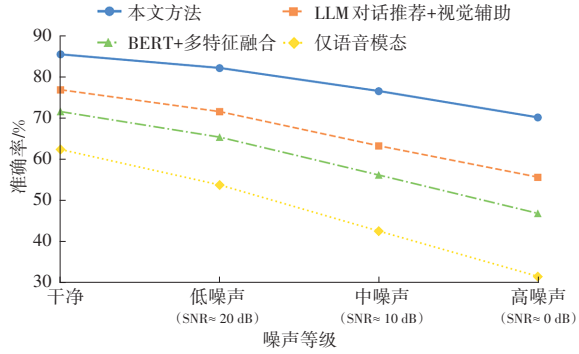


图4　不同噪声等级下的鲁棒性测试结果

3.6　细粒度分析

图5所示为不同POI类别的识别准确率。停车场和加油站的识别准确率最高（分别为93.1%和91.5%），原因在于这两类POI语义明确、视觉特征显著（加油站标志统一、停车场入口易于辨识）。美食类的识别准确率为85.2%，略低的原因在于餐厅种类繁多，“找个地方吃饭”的意图粒度较粗。酒店类的识别准确率最低（为76.8%），原因在于酒店需求在驾驶场景中出现频率较低，训练样本相对不足，且酒店建筑外观差异大，视觉匹配难度较高。

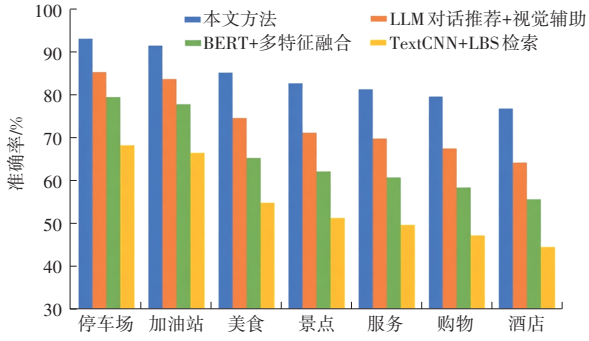


图5　不同POI类别的识别准确率对比

图6所示为不同车速区间下的准确率变化情况。低速（<20 km/h）行驶时准确率为89.2%，高速（100~120 km/h）行驶时准确率降至73.5%。车速提高主要带来两方面影响：一是手势检测稳定性下降（车身振动导致关键点抖动），二是POI相对位置的快速变化导致扇形检索的时效窗口缩短。本文方法在所有速度区间的识别准确率均优于对比方法，且降幅（15.7百分点）小于BERT+多特征融合方法（18.8百分点）和TextCNN+LBS检索方法（20.5百分点），表明多模态融合对速度变化的适应性更强。

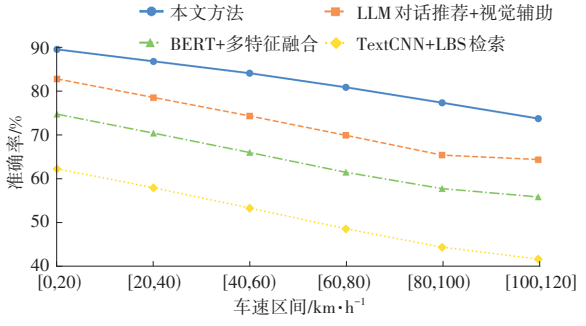
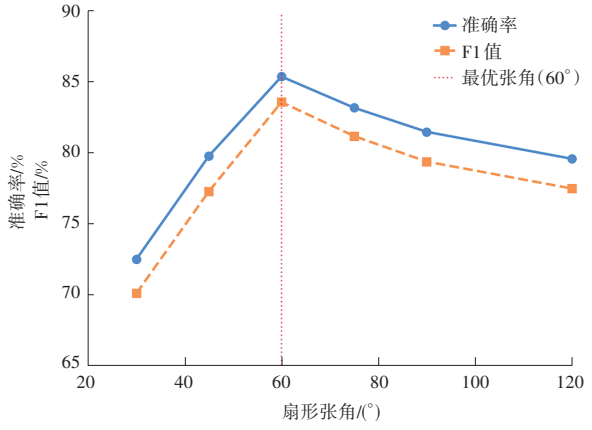


图6　不同车速区间下的POI识别准确率对比

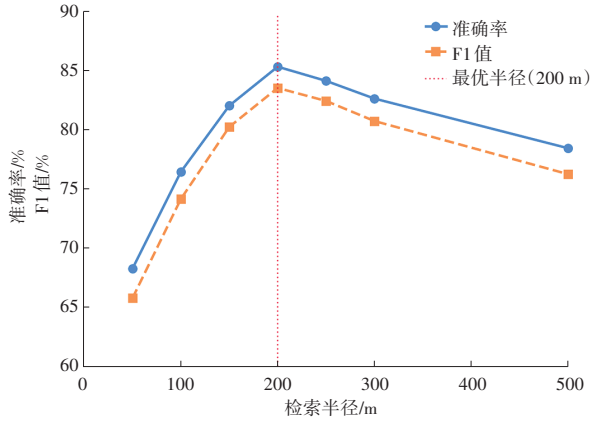
3.7　参数敏感性分析

扇形区域的张角和检索半径对系统性能有直接影响。图7a展示了张角从30°到120°的变化过程对系统性能的影响：张角过小（30°）时，目标POI可能落在检索区域外，准确率仅为72.5%；张角过大

（120°）时，过多无关POI被纳入候选集，排序难度增加，准确率降至79.6%；最优张角为60°，准确率达85.4%。图7b展示了检索半径对系统性能的影响：检索半径为50 m时准确率为68.3%（覆盖不足），检索半径为200 m时准确率达到最优（85.4%），检索半径为500 m时准确率降至78.5%（噪声过多）。



（a）扇形张角对性能的影响



（b）检索半径对性能的影响

图7　扇形区域参数敏感性分析

3.8　响应时间分析

车载交互对实时性有严格要求，一般认为2 s以内的响应时间可保证良好的用户体验。图8所示为各方法的响应时间对比结果。TF-IDF+协同过滤方法的平均响应时间最短（0.12 s），但准确率也最低。本文方法的平均响应时间为1.56 s，第95百分位数（95th Percentile，P95）为2.31 s，其中ASR系统推理时间约0.4 s，意图识别时间约0.5 s，扇形检索时间约0.06 s，综合预测时间约0.3 s，其余为数据传输和后处理开销。其中，P95略超2 s阈值，主要原因是候选POI数量较多时Qwen-VL的排序耗时增加。LLM对话推荐+视觉辅助方法的平均响应时间为2.87 s，P95达4.52 s，云端推理的网络延迟是其主要瓶颈。本文方法在准确率和实时性之间取得了较好的平衡，通过模型量化和计算图优化，P95响应时间有望进一步压缩至2 s以内。

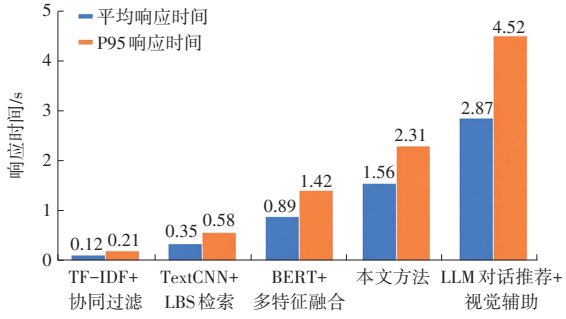


图8　各方法的系统响应时间对比

3.9　误差分析

图9所示为本文方法在5个主要POI类别上的混淆矩阵。对角线元素表明各类别的识别率均超过88%，其中加油站最高（97%）、停车场次之（93%）。误分类主要发生在语义相近的类别之间：购物被误分类为美食的比例为3%，景点被误分类为购物的比例为5%，这与实际场景中商业综合体兼具购物和餐饮功能有关。酒店被误分类为景点的比例为7%，部分历史建筑型酒店的外观与景点相似，视觉分支难以有效区分。

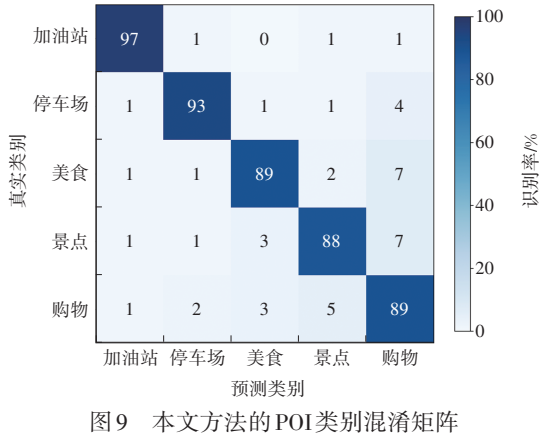


图9　本文方法的POI类别混淆矩阵

4　结束语

本文针对车载POI识别中单一语音模态的噪声敏感、方向信息缺失和意图模糊问题，提出了融合语音、视觉与手势的多模态智能识别方法。通过构建四层系统架构，实现了从多源数据采集到POI推荐的端到端处理。基于扇形区域的多策略时空定位方法有效融合了语音方向词与视线、手势或导航方向信息，其中，最优参数与城市道路的POI密度分布相关：在POI密集的商业区，较小的张角和半径可能更为合适；在郊区，则需要适当放大。此外，误差

分析中视觉混淆问题在多模融合中获得有效解决，进而将候选POI从平均47个缩减至12个。LoRA微调的Qwen-7B实现了高效的结构化意图识别，Qwen-VL的跨模态注意力机制完成了POI精确排序。

在自建车载多模态数据集上的试验验证了方法的有效性：准确率为85.4%，F1值为84.3%，较最优基线分别提升8.6百分点和7.7百分点；在0 dB信噪比下仍保持70.1%的准确率，验证了多模态融合对噪声的鲁棒性；85.4%的POI需求在3轮交互内得到满足。然而，本文方法仍存在以下局限性：高速行驶时手势检测稳定性下降，P95响应时间略超2 s阈值，酒店类POI的识别率有待提升。

后续工作将从以下方面展开：一是引入基于POI密度的自适应扇形参数调整策略，提升不同区域类型的适应性；二是集成轻量级光学字符识别（Optical Character Recognition，OCR）模块增强对招牌的文字识别能力，改善外观相似POI的区分度；三是探索端侧部署的进一步优化方案，包括模型蒸馏和计算图编译优化，以将P95响应时间压缩至1.5 s以内。

参 考 文 献

[1] ZHANG Z X, LI Y, REN W, et al. A Secure and Robust Multimodal Framework for In-Vehicle Voice Control: Integrating Bilingual Wake-up, Speaker Verification, and Fuzzy Command Understanding[J]. Eng, 2025, 6(11).

[2] KUNST K, HOFFMANN D, ERKAN A, et al. Analyzing Gaze During Driving: Should Eye Tracking Be Used to Design Automotive Lighting Functions?[J]. Journal of Eye Movement Research, 2025, 18(2).

[3] 陈启博, 闫坤, 陈杰伟, 等. 基于Kinect和改进DTW的交通指挥手势识别方法[J]. 桂林电子科技大学学报, 2026, 46(1): 59-66.
CHEN Q B, YAN K, CHEN J W, et al. Traffic Command Gesture Recognition Method Based on Kinect and Improved DTW[J]. Journal of Guilin University of Electronic Technology, 2026, 46(1): 59-66.

[4] DARAPISUT S, AMPHAWAN K, LEELATHAKUL N, et al. A Hybrid POI Recommendation System Combining Link Analysis and Collaborative Filtering Based on Various Visiting Behaviors[J]. ISPRS International Journal of Geo-Information, 2023, 12(10).

[5] 陈勇, 周宇翔, 王秀毅, 等. 一种基于BERT预训练语言模型的语音指令识别方法: CN119479646A[P]. 2025-02-18.
CHEN Y, ZHOU Y X, WANG X Y, et al. A Speech Command Recognition Method Based on BERT Pre-Trained Language Model: CN119479646A[P]. 2025-02-18.

[6] DONG H M, WANG W B, JIANG D L, et al. PC-LLMRec: Large Language Model for Personalized Interaction Recommendations in Intelligent Cockpit[J]. IET Intelligent Transport Systems, 2026, 20(1).

[7] LI K, XU L, ZHU C, et al. A Multimodal Graph Recommendation Method Based on Cross-Attention Fusion [J]. Mathematics, 2024, 12(15).

[8] ZHANG Q, ZENG S Z, CHEN Y R, et al. Adaptive Fusion of Dual-Scale Preference Disentanglement for Next POI Recommendation[J]. IEEE Access, 2026, 14: 5855-5868. DOI: 10.1109/ACCESS.2025.3647754.

[9] YE M, YIN P F, LEE W C, et al. Exploiting Geographical Influence for Collaborative Point-of-Interest Recommendation[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing China: Association for Computing Machinery, 2011: 325-334.

[10] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1746-1751.

（责任编辑　斛　畔）

修改稿收到日期为2026年6月26日。