中图分类号:U412.36+6;U495 文献标志码:A 文章编号:2097-017X(2026)04-0033-10
DOI:10.3969/j.issn.2097-017X.2026.04.004

基于多模态大模型的高速公路异常事件检测方法研究

刘化学[1],周 洁[2],崔 亮[1],莫洪韵[2],陆如洋[2],闫宇凤[3]

(1. 苏交科集团股份有限公司,江苏 南京 210019;2. 江苏交通控股有限公司,江苏 南京 210019;
3. 河海大学土木与交通学院,江苏 南京 210024)

摘要:高速公路异常事件处置的关键难点在于,现有方法往往只能基于单一感知信号识别局部异常,难以同时完成事件判别、诱因解释与风险提示。针对这一问题,提出了一种基于多模态大模型的高速公路异常事件检测方法。该方法融合视频监控、毫米波雷达、交通流量及气象信息等多源异构数据,通过"模态扩展-场景微调-跨模态注意力Transformer融合"将不同特征映射至CLIP共享语义空间,实现异构数据的统一表达;结合交通知识图谱与Prompt工程,引导大模型深度解析异常事件的类型、诱因及潜在风险;基于DAIR-V2X数据集的实验结果表明,所提方法在异常事件检测的准确性、语义理解的全面性方面均显著优于传统方法,为高速公路的多维感知提供了新的技术路径和理论支撑。

关键词:高速公路;异常事件检测;多模态大模型;跨模态对齐;语义理解;知识图谱

Research on expressway anomaly event detection methods based on multimodal large models

LIU Huaxue[1], ZHOU Jie[2], CUI Liang[1], MO Hongyun[2], LU Ruyang[2], YAN Yufeng[3]

(1. *JSTI Group Co. ,Ltd. , Nanjing* 210019, *China*; 2. *Jiangsu Communications Holding Co*, , *Ltd.* , *Nanjing* 210019, *China*; 3. *Collage of Civil and Transportation Engineering*, *Hohai University.* , *Nanjing* 210024, *China*)

Abstract: A key challenge in expressway anomaly handling is that existing methods usually identify local abnormalities from a single sensing signal, but struggle to jointly determine event type, cause, and risk. To address this issue, this paper proposes a multimodal large-model-based method for expressway anomaly event detection. This paper addresses the demand for intelligent expressway management by proposing a multimodal large-model-based anomaly detection method. It integrates heterogeneous data from multiple sources, including video surveillance, millimeter-wave radar, traffic flow, and meteorological information, through a "modality expansion-scenario fine-tuning-cross-modal attention Transformer fusion" approach. This process maps diverse features into the CLIP shared semantic space, enabling unified representation of heterogeneous data. Combining traffic knowledge graphs with

收稿日期:2026-04-27
基金项目:江苏省自然科学基金资助项目(BK20230964)
第一作者简介:刘化学(1986—),男,硕士,高级工程师。研究方向:交通智慧化治理、公路智慧扩容改造。
通信作者简介:崔 亮(1997—),男,硕士,工程师。研究方向:高速公路智慧扩容改造、交通基础设施智能化升级。

prompt engineering guides the large model to deeply analyze the types, causes, and potential risks of abnormal events. Experimental results based on the DAIR-V2X dataset demonstrate that the proposed method significantly outperforms traditional approaches in both detection accuracy and comprehensive semantic understanding. This provides a new technical pathway and theoretical foundation for multidimensional perception on expressways.

Keywords: expressway; anomaly event detection; multimodal large model; cross-modal alignment; semantic understanding; knowledge graph

高速公路作为交通运输体系的核心枢纽,其通行效率与交通安全直接影响区域经济发展与公众出行体验[1]。近年来,随着机动车保有量的快速增长,高速公路日均车流量屡创新高,叠加恶劣天气、突发事故等不确定因素,异常事件发生频率显著上升。在智慧交通发展背景下,多源异构数据融合与人工智能技术的深度应用为解决上述问题提供了可能。以 CLIP(Contrastive Language-Image Pre-training)为代表的多模态大模型凭借其强大的跨模态特征提取与语义理解能力,在基础视觉与语言任务中展现出卓越性能,为交通异常事件的全方位感知提供了潜在的技术支撑。

传统异常事件检测方法主要分为基于交通流参数与规则统计判别、基于机器学习、基于视频图像与目标轨迹分析三类。吴晓佩[2]利用多SVM分类器对高速公路复杂交通信息进行融合,提升了异常事件检测的准确性和可靠性,但融合层次仍主要停留在特征和决策层。隋靓等[3]基于车辆运动轨迹构建高速公路异常事件检测算法,可识别逆行、停车等事件,但对视频目标提取和轨迹连续性依赖较强。李旭等[4]采用SSD目标检测与卡尔曼滤波跟踪方法开展基于视频分析的高速公路异常事件检测,能够对违停、逆行等行为进行判别,但仍缺少对事件诱因和风险语义的表达。孙倩等[5]提出基于离散时间信号相关性的自动交通事件检测算法,增强了交通流时序变化的可解释性,但对多模态感知数据的联合建模仍有不足。WEIL等[6]系统梳理了交通事件检测中的传感器与算法,指出传统检测方法高度依赖检测器布设和数据质量。TENG等[7]提出基于检测延迟的高速公路事件检测算法,JIN等[8]基于交通流基本图构建自动事件检测方法,上述方法能够利用流量、速度等突变特征识别异常,但规则阈值的场景适应性有限。随着人工智能技术的发展,机器学习方法被广泛应用于异常检测。RITCHIE等[9]利用人工神经网络开展高速公路事件检测仿真,DIA等[10]进一步基于实测数据构建神经网络检测模型,SRINIVASAN等[11]提出面向交通网络事件检测的智能识别方法,上述方法增强了非线性特征建模能力,但对训练样本规模、特征选择和跨场景泛化能力仍较敏感。在视频感知方面,SHEHATA等[12]研究了智能道路场景下的视频事件自动检测方法,BUCH等[13]系统梳理了计算机视觉在交通分析中的应用,此类方法能够提取更直观的空间目标与轨迹信息,但在光照变化、恶劣天气、遮挡及复杂交通环境下易出现误检或漏检。CHEN等[14]采用决策树学习方法进行高速公路自动事件检测,提高了规则判别方法的自适应能力,但对特征构造和样本分布仍具有较强依赖。MA等[15]将交通状态映射为图像形式,利用深度卷积神经网络进行大规模路网速度预测,为交通异常识别中的时空特征表达提供了参考。YAO等[16]面向第一视角交通视频提出无监督交通事故检测方法,降低了对人工标注样本的依赖,但对固定监控视角和多源传感器融合的适配仍有限。XU等[17]构建了面向视频监控的交通事故检测大规模基准数据集,推动了复杂场景下事故检测模型评测,但仍主要聚焦视觉模态,尚未充分融合雷达、流量和气象等异构信息。总体来看,传统方法多依赖单一数据源或浅层特征判别,难以充分融合视频、雷达、流量、气象等多源异构信息,也缺乏对异常事件类型、诱因和风险语义的深层解析。

针对上述问题,本文立足高速公路智慧管控实际需求,提出一种基于多模态大模型的高速公路异常事件检测方法。该方法以“跨模态融合-领域适配-语义深化-实时检测”为核心逻辑,融合视频监控、毫米波雷达、交通流量及气象信息等多源异构数据,构建“模态扩展-场景微调-跨模态注意力 Transformer 融合”的技术框架,并结合交通知识图谱与结构化 Prompt,对异常事件的类型、诱因及潜在风险进行解析,为后续人工研判和交通管控提供更具解释性的依据。

本文的主要工作与贡献如下:

(1) 在CLIP框架基础上扩展雷达、交通流量和气象编码分支,并结合场景微调与跨模态注意力融合,构建适用于高速公路场景的四模态特征表征方

法。该方法重点解决专业传感器模态与视觉语义空间难以对齐的问题，提高了不同异常场景下的特征表达一致性。

（2）结合交通知识图谱与结构化Prompt，将融合特征进一步映射为事件类型、诱因和风险等级等结构化语义结果。该设计侧重提升异常事件识别结果的可解释性，为后续人工研判和管控决策提供更清晰的信息支撑。

1　数据处理与多模态大模型特征融合

1.1　多源数据预处理

本文选取由清华大学智能产业研究院（AIR）联合北京市高级别自动驾驶示范区、百度Apollo、北京智源人工智能研究院共同发布的DAIR-V2X数据集进行实验。数据集涵盖10 km高速公路全时段、多气象场景数据。

1.1.1　视频数据预处理

首先进行图像增强，针对图像噪声，采用高斯滤波平滑噪声。接着采用RetinaNet目标检测器，检测图像中的“车辆、护栏、路面标线”三类目标，采用KLT光流法跟踪车辆目标。然后对感兴趣区域（ROI）进行提取，最后将ROI图像缩放至224×224，采用Z-score标准化处理，使ROI图像的数据分布与CLIP预训练数据分布保持一致。

1.1.2　雷达数据预处理

首先进行噪声剔除。雷达散射截面（Radar Cross Section，RCS）是描述目标对雷达电磁波散射能力的物理量，其数值与目标大小、形状、材质及雷达入射角度直接相关，单位为dBm^2。基于DAIR-V2X数据集的雷达数据统计特性，对10 km高速公路路段的雷达回波数据进行系统分析后发现，车辆目标的RCS值普遍高于$-15\ dBm^2$。因此，设定RCS阈值为$-15\ dBm^2$，剔除所有RCS值低于该阈值的雷达点，实现噪声点的高效筛选。接着采用线性卡尔曼滤波对车辆的速度与距离数据进行平滑处理。最后采用时间戳对齐法将视频数据与雷达数据进行时空同步。

1.1.3　交通流量与气象数据预处理

分别采用滑动窗口插值、3σ准则处理缺失/异常值，通过参数筛选与指数移动平均优化数据质量。

1.2　多模态大模型选型与对比分析

多模态大模型的核心价值在于“跨模态语义关联”，通过“模型选型-模态扩展-场景微调-跨模态融合”四步流程，构建适配高速公路场景的多模态大模型，实现“视频-雷达-流量-气象”四模态特征的深度融合，为后续语义理解提供统一的特征表示。

当前主流的多模态大模型包括CLIP（OpenAI）、FLAVA（Google）、ALBEF（Microsoft）、BLIP-2（Salesforce），需从“模态兼容性、迁移性、计算效率”三个维度进行对比，选择最适配高速公路场景的基础模型。

表1从模态架构、预训练方式、跨模态迁移能力、参数规模与实时性等方面，对主流多模态大模型进行了对比。对比结果表明：CLIP的“图像-文本”双模态架构可灵活扩展至传感器模态，通过对比学习预训练，具备强大的零样本跨模态迁移能力，无需大量交通领域文本标注即可适配；CLIP原始参数规模约为1.2 B，经剪枝后可降至300 M，单样本推理时间<50 ms，满足高速公路管控的实时性要求。因此，本研究选择CLIP（ViT-B/32）作为基础多模态大模型，后续进行“模态扩展”与“场景微调”，使其支持“视频-雷达-流量-气象”四模态融合。

表1　多模态大模型对比

Tab. 1　Multimodal large model comparison

模型名称	核心模态	预训练数据量	跨模态迁移性	计算复杂度	交通场景适配难度
CLIP（ViT-B/32）	图像-文本	4亿图像-文本对	优（零样本迁移）	低（1.2 B）	低（开源可修改）
FLAVA	图像-文本-音频	35亿多模态样本	优（多模态统一建模）	中（3.5 B）	中（音频模态冗余）
ALBEF	图像-文本	1.4亿图像-文本对	良（需文本数据微调）	中（1.8 B）	高（文本标注需求大）
BLIP-2	图像-文本	1.2亿图像-文本对	优（语义理解强）	高（13 B）	高（计算量大）

CLIP模型的核心思想是通过“对比学习＋跨模态对齐”构建通用语义空间，打破图像与文本的模态壁垒。其架构由双分支编码器组成：

（1）图像编码器：采用Vision Transformer（ViT-B/32），将输入图像分割为32×32的图像块，通过位置编码与多头自注意力机制提取全局图像特征，输出512维图像特征向量；

（2）文本编码器：采用Transformer架构，将文

本序列转换为词嵌入后，通过自注意力机制与层归一化提取文本语义特征，输出512维文本特征向量。

训练阶段，CLIP通过对比学习优化“图像-文本”匹配损失，使匹配的图像-文本对在语义空间中的余弦相似度最大化，不匹配的相似度最小化，最终实现图像与文本模态的深度语义对齐，具备强大的零样本跨模态迁移能力。

1.3 多模态大模型的模态扩展

针对视频、雷达及流量数据的异构性，为各模态设计专用编码分支，每个分支的输入为预处理后的模态特征块，输出为512维特征向量，与CLIP的图像特征对齐。具体编码分支设计如下：

雷达模态：采用3层MLP编码器，输入为256维雷达特征块，隐藏层维度分别为512、256，激活函数采用ReLU，输出512维特征向量；

交通流量模态：针对交通流量的强时序依赖性，设计“轻量级1D-CNN＋MLP”编码分支。1D-CNN层包含1个卷积核大小为3、步长为1、输出通道数为64的卷积层，后跟1个全局平均池化层；MLP层包含2个全连接层，第一层输入维度为64、输出维度为256，第二层输入维度为256、输出维度为512，通过1D-CNN捕捉流量数据的短期时序依赖关系，弥补单一MLP对时序特征建模能力不足的问题。

为解决多模态特征的异构性问题，设计“模态内标准化＋模态间对比学习”的对齐机制，确保四模态特征在统一语义空间中可比较。构建四模态对比损失函数，使同一事件的四模态特征距离更近，不同事件的特征距离更远，损失函数公式如下：

$$L_{\text{total}}=\frac{1}{C_4^2}\sum_{i<j}L_{\text{contrast}}(F_i,F_j) \tag{1}$$

式中，$C_4^2=6$为四模态中选2个的组合数；F_i、F_j为任意两个模态的特征向量；L_{contrast}为双模态对比损失，公式为：

$$L_{\text{contrast}}(F_i,F_j)=-\lg\frac{\exp(s(F_i,F_j)/\tau)}{\sum_{k\neq m}\exp(s(F_i^k,F_j^m)/\tau)} \tag{2}$$

式中，$s(F_i,F_j)=F_i\cdot F_j^{\mathrm{T}}$为余弦相似度；$\tau$为温度参数；$F_i^k$为第$k$个正样本的$i$模态特征；$F_j^m$为第$m$个负样本的$j$模态特征。

通过该对齐机制，四模态特征被映射至统一语义空间，确保后续融合特征的一致性。

1.4 多模态大模型的场景微调

CLIP的预训练数据以通用场景为主，需在高速公路异常事件数据集上进行场景微调，使模型学习交通领域的跨模态关联。

（1）微调数据集构建。构建“高速公路多模态微调数据集（HE-MF Dataset）”，每条样本为“四模态特征＋事件文本描述”的组合。四模态特征包括视频图像（224×224）、雷达特征块（256维）、流量特征块（128维）、气象特征块（8维），均来自预处理后的高质量数据。

（2）微调策略与超参数设置。冻结CLIP图像分支与文本分支的前10层参数，仅微调后5层与传感器编码分支的参数，采用AdamW优化器。初始学习率为1×10^{-5}，采用余弦退火衰减，迭代轮次为50轮，若验证集准确率连续5轮无提升，则停止训练，保存最优模型。模型训练过程损失函数收敛趋势如图1所示。

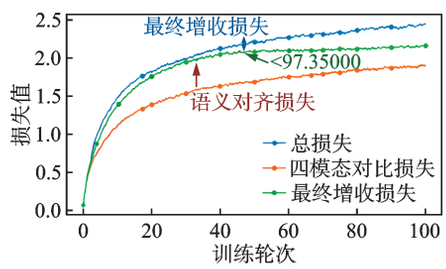

图1 模型训练过程损失函数收敛趋势

Fig.1 Convergence trend of loss function during model training

（3）微调效果验证。为覆盖核心模态关联场景及涵盖“视觉-传感器”“视觉-语义”“传感器-交通流”“环境-视觉”四类典型模态组合，本文选取图像→雷达、图像→文本、雷达→流量、气象→图像四类跨模态检索任务，对场景微调前后的模型性能进行验证，具体结果如表2所示。

表2 微调前后跨模态检索准确率对比

Tab. 2 Comparison of cross-modal retrieval accuracy before and after fine-tuning

检索任务	检索准确率/%		
	微调前	微调后	提升幅度
图像→雷达	62.3	89.5	27.2
图像→文本	75.1	94.2	19.1
雷达→流量	58.7	87.3	28.6
气象→图像	65.2	90.1	24.9
四模态交叉检索平均	65.3	90.3	25.0

结果表明：微调后四模态交叉检索平均准确率从65.3%提升至90.3%，证明模型已学习到高速公路场景的跨模态关联，为后续语义理解奠定基础。

1.5　多模态特征融合算法

经过模态扩展与场景微调后，四模态特征已实现语义对齐，需进一步融合为“单模态融合特征向量”，为语义理解层提供统一输入。本研究采用“跨模态注意力 Transformer”实现融合，该算法可自适应学习各模态的权重，突出关键模态的贡献。多模态特征融合架构如图2所示。

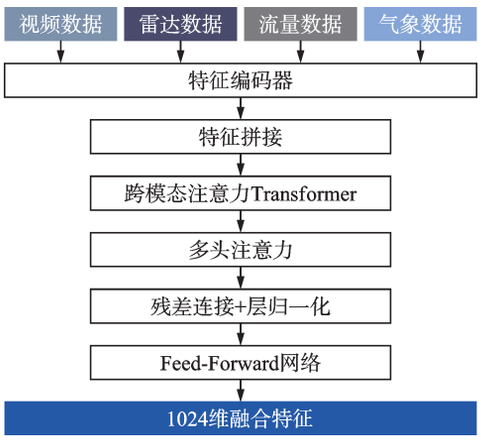

图2　多模态特征融合架构

Fig.2　Multimodal feature fusion architecture

（1）跨模态注意力 Transformer 架构。架构分为“特征拼接-多头注意力-残差连接-层归一化”四部分，具体见图2。

特征拼接：将四模态的标准化特征（512维）拼接为2048维特征向量：

$$F_{\text{concat}}=[\,F_{\text{img}}\quad F_{\text{radar}}\quad F_{\text{flow}}\quad F_{\text{weather}}\,]$$

多头注意力：采用8头注意力机制，将 F_{concat} 分为8个256维的子向量，每个头计算不同模态间的注意力权重，公式为：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}\tag{3}$$

式中，$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=F_{\text{concat}}$；$d_k=256$ 为每个头的特征维度；将 Softmax 函数计算得到的权重记为注意力权重矩阵 $\boldsymbol{A}\in\mathbf{R}^{2048\times2048}$，矩阵元素 $A_{i,j}$ 表示第 j 个特征对第 i 个特征的贡献度。

残差连接与层归一化：注意力输出与原始 F_{concat} 进行残差连接，然后通过层归一化消除特征分布偏移，公式为：

$$F_{\text{attn}}=\text{LayerNorm}(F_{\text{concat}}+\text{MultiHeadAttention}(F_{\text{concat}}))\tag{4}$$

Feed-Forward 网络：通过2层全连接网络（隐藏层维度1024，激活函数 GELU）对 F_{attn} 进行非线性变换，最终输出1024维的多模态融合特征 F_{fusion}。

（2）模态权重分析。通过可视化注意力权重矩阵 $\boldsymbol{A}$，分析不同场景下各模态的贡献度。不同场景下各模态的注意力权重占比如表3所示。

表3　不同场景下各模态的注意力权重占比

Tab. 3　Attention weight distribution across modalities in different scenarios

场景类型	注意力权重占比/%				权重标准差/%
	视频图像	雷达	交通流量	气象	
追尾事故	42.3	35.1	12.5	10.1	14.2
车辆抛洒物	45.7	28.3	15.2	10.8	15.1
道路施工	38.5	18.2	25.3	18.0	9.8
交通拥堵	15.6	12.3	58.7	13.4	22.5
恶劣天气	22.4	18.5	25.1	34.0	7.8

结果表明：模型可自适应调整模态权重。事故类场景（追尾、抛洒物）中视频与雷达权重占比超70%，拥堵场景中流量权重占比近60%，恶劣天气场景中气象权重占比超30%，证明融合算法能够捕捉场景与模态的关联，突出关键信息。

（3）融合特征的有效性验证。通过“异常事件分类准确率”验证 F_{fusion} 的有效性：将 F_{fusion} 输入到 SoftMax 分类器，对5类异常事件进行分类，与单一模态特征、简单拼接特征对比，结果如表4所示。

表4　不同特征的异常事件分类准确率对比

Tab. 4　Comparison of classification accuracy for anomalous events with different features

特征类型	准确率/%	精确率/%	召回率/%	F1分数/%
视频图像特征	82.3	80.5	78.6	79.5
雷达特征	79.1	77.3	75.2	76.2
交通流量特征	85.6	83.2	81.5	82.3
四模态简单拼接特征	89.5	87.7	86.3	87.0
本文融合特征 F_{fusion}	96.2	94.7	94.3	94.5

结果表明：F_{fusion} 的分类准确率达96.2%，较单一模态最高提升17.1个百分点，较简单拼接提升6.7个百分点，证明融合特征包含更全面的场景信息，可有效支撑后续语义理解。

2　异常事件语义理解

异常事件语义理解的核心是“从特征到语义”的转化，即通过多模态融合特征 F_{fusion}，解析出事件的类型、诱因及风险的结构化语义信息。本研究通过“多模态大模型语义解码＋Prompt工程引导＋交通知识图谱约束”的协同机制，实现深层语义理解，区别于传统方法的“特征匹配式”识别。

2.1 交通知识图谱构建

多模态大模型虽具备强大的泛化能力，但在专业领域易产生“语义漂移”，需依托交通知识图谱注入领域约束，确保语义理解的准确性。交通知识图谱的构建，参考《道路交通信息服务 交通事件分类与编码》(GB/T 29100—2012)和《道路交通信息采集 事件信息集》(GB/T 20134—2025)，采用“实体-关系-属性”三元组结构，系统整合异常事件、环境因素、路网元素、管控措施等核心信息，形成大模型语义推理的“领域知识库”。

2.2 Prompt构建模板设计

Prompt模板遵循“事件核心要素＋场景约束＋知识图谱关联”的设计原则，确保每个模板既包含结构化数据的关键字段，又符合自然语言表达习惯，同时嵌入交通知识图谱中的实体、关系约束(如事件类型规范、诱因分类、风险等级定义)。具体模板及示例如表5所示。

表5 Prompt模板构建

Tab. 5 Prompt template construction

Prompt模板类型	模板结构	具体示例	对应结构化数据来源
交通事故类	"A [事件亚型] incident on [路段信息], induced by [核心诱因] ([辅助诱因]), [环境属性], visibility [数值] meters, risk level[等级], involving [车辆类型 /数量]."	"A traffic rear-end collision incident on Highway G15 Section K230, induced by wet road surface (light rain), overcast weather, visibility 80 meters, risk level medium, involving 2 passenger cars and 1 truck."	事件类型＝追尾事故(知识图谱实体)；路段＝G15高速K230段(路网元素)；诱因＝雨天路滑(环境－事件关系)；能见度＝8 m(环境属性)；风险等级＝中等(风险参数)；涉及车辆＝2客1货(事件属性)
道路施工类	"A [施工类型] event on [车道信息] of [公路编号], construction length [数值] meters, speed limit [数值] km/h, [环境条件], risk level [等级], impact scope [范围描述]."	"A road maintenance construction event on the right lane of Highway G40, construction length 60 meters, speed limit 40 km/h, no adverse weather, risk level low, impact scope partial lane occupation."	事件类型＝道路养护施工(知识图谱实体)；车道＝右侧车道(路网元素)；施工长度＝60 m(施工属性)；限速＝40 km/h(管控措施)；环境＝无不利天气(环境属性)；风险等级＝低(风险参数)
车辆抛洒物类	"A [抛洒物类型] incident on [车道信息] of [公路编号], induced by [核心诱因] ([辅助诱因]), [环境条件], visibility [数值] meters, risk level [等级], causing [影响结果]."	"A cargo spillage incident on the middle lane of Highway G42, induced by unsecured freight (sudden braking), cloudy weather, visibility 120 meters, risk level medium, causing temporary lane blockage and traffic slowdown."	事件类型＝车辆抛洒物(知识图谱实体)；抛洒物类型＝货物散落(事件属性)；车道＝中间车道(路网元素)；公路编号＝G42高速(路网属性)；诱因＝货物捆扎不牢(事件-诱因关系)；辅助诱因＝紧急制动(事件-诱因关系)；环境＝阴天(环境属性)；能见度＝120 m(环境属性)；风险等级＝中等(风险参数)；影响结果＝车道短时受阻、车辆减速通行(事件影响属性)
车辆故障类	"A [故障类型] fault on [车道信息] of [公路编号], location near [标识点], [天气条件], visibility [数值] meters, risk level [等级], causing [影响结果]."	"A vehicle tire blowout fault on the fast lane of Highway S20, location near exit 35, heavy fog weather, visibility 30 meters, risk level high, causing temporary traffic congestion."	事件类型＝轮胎爆胎故障(知识图谱实体)；车道 ＝ 快车道(路网元素)；位置 ＝ 35号出口附近(路网属性)；天气＝大雾(环境因素)；能见度＝ 30 m(环境属性)；风险等级＝高(风险参数)
交通拥堵类	"A traffic jam incident on [公路类型], induced by [核心诱因], vehicle density [数值] pcu/km, [环境条件], risk level [等级], congestion length [数值] meters."	"A traffic jam incident on Urban Ring Highway, induced by peak-hour flow concentration, vehicle density 130 pcu/km, sunny weather, risk level medium, congestion length 800 meters."	事件类型＝交通拥堵(知识图谱实体)；诱因＝高峰流量集中(事件-诱因关系)；车辆密度＝ 130 pcu/km(交通流属性)；环境＝晴天(环境属性)；风险等级＝中等(风险参数)；拥堵长度＝800 m(事件属性)

2.3 结构化数据

Textualization是将离散的结构化数据转化为连贯、规范自然语言文本的过程，核心是字段映射＋逻辑串联＋领域约束，确保转化后的文本既包含完整结构化信息，又适配CLIP文本分支的语义理解习惯。具体步骤如下：

(1) 结构化数据字段提取与标准化。从交通知识图谱(实体-关系-属性三元组)和多模态融合特征对应的标注数据中，提取核心字段，包括5大类：

事件核心要素：事件类型(如追尾事故、道路施工)、事件亚型(如轮胎爆胎、养护施工)、发生地点

(公路编号、路段里程、车道类型、标识点);

诱因信息:核心诱因(如路滑、流量集中)、辅助诱因(如雨天、施工占道);

环境属性:天气类型(小雨/大雾/晴天)、能见度(数值+单位)、路面状况(干燥/湿滑);

路网与管控信息:施工长度、限速值、车道占用情况;

风险与影响参数:风险等级(低/中/高)、涉及车辆类型/数量、拥堵长度、车辆密度。对字段进行标准化处理,例如:公路编号统一为"Highway[字母+数字]"(如G15高速)、风险等级统一为"low/medium/high"、能见度统一为"[数值] meters",避免表述歧义。

(2) 逻辑串联与文本生成。按照"事件主体→发生场景→诱因说明→属性补充→风险标注"的逻辑语序,将标准化字段代入Prompt模板,生成自然语言文本。关键原则:

事件主体前置:优先明确事件类型(如"A traffic rear-end collision incident"),让模型快速定位核心语义;

诱因显性化:用"induced by"关联核心诱因与辅助诱因,体现因果关系(如"induced by wet road surface (light rain)");

属性结构化罗列:环境、路网、风险等属性按"重要性排序"(环境→路网→影响),避免信息混乱;

领域术语准确:严格沿用知识图谱中的规范表述(如"pcu/km""partial lane occupation"),确保术语表述与领域知识库保持一致。

2.4　数据序列化过程

数据序列化是将Textualization后的自然语言文本转化为CLIP模型可解析的向量形式的关键步骤,实现"自然语言→Token序列→语义向量"的转化,与多模态融合特征F_{fusion}形成协同输入。具体流程如下:

(1) 文本预处理。对生成的自然语言文本进行清洗优化。

格式标准化:首字母大写,其余小写;统一使用英文标点,去除冗余空格;

拼写与语法修正:例如修正"visbility"为"visibility"、"lane ocupation"为"lane occupation";

长度适配:CLIP文本分支默认最大输入长度为77个Token,对超长文本(如复杂多诱因事件)进行核心信息保留截断,对短文本用Padding Token([PAD])填充至固定长度,避免模型输入维度不匹配。

(2) 分词与编码。将预处理后的文本输入CLIP的文本编码器(Text Encoder),执行分词与编码操作。

子词拆分:基于CLIP预训练的BPE(Byte Pair Encoding)分词算法,将文本拆分为模型词典中的子词(Subword),例如"rear-end collision"拆分为"rear-end""collision","partial lane occupation"拆分为"partial""lane""occupation";

Token ID映射:将每个子词映射为词典中对应的唯一整数ID(Token ID),同时添加特殊Token:句首添加"[CLS]"(用于聚合全局语义)、句尾添加"[SEP]"(标识句子结束),形成Token ID序列。

(3) 文本语义向量编码。CLIP的文本编码器为Transformer结构(含12层编码器、多头注意力机制),对Token ID序列进行编码。

嵌入层处理:将Token ID序列转化为低维稠密的词嵌入向量(维度为512),同时添加位置编码(Positional Encoding),体现文本中词的顺序关系;

Transformer编码:通过多头注意力机制捕捉子词间的语义关联(如"wet road surface"与"light rain"的因果关联),通过前馈神经网络进行特征转换,最终输出"[CLS]"位置对应的向量作为文本语义向量F_{text}(维度为512),该向量与多模态融合特征转化后的语义向量F_{semantic}处于同一语义空间。

(4) 特征融合与输入。将文本语义向量F_{text}与参与融合转换后的512维F_{semantic}进行加权融合(权重通过模型微调学习),得到增强型融合特征F_{enhanced},公式为:

$$F_{\text{enhanced}} = \alpha \cdot F_{\text{semantic}} + (1-\alpha) \cdot F_{\text{text}} \tag{5}$$

式中,$\alpha \in [0,1]$为融合权重,确保多模态感知特征与文本语义特征的协同作用。最终将F_{enhanced}输入到CLIP文本分支,进行后续语义解码,具体语义推理过程详见2.6。

2.5　Prompt优化效果验证

通过"语义理解准确率"验证不同Prompt的性能,对比"无约束Prompt""简单约束Prompt""本文Prompt(含KG约束+数据序列化)"的效果,结果如表6所示。

结果表明:本文Prompt通过结构化模板设计、知识图谱约束注入及完整的数据序列化流程,综合语义准确率达93.2%,较无约束Prompt提升21.2个百分点,证明该设计能有效引导大模型进行精准语义解码。

表6　不同Prompt的语义理解准确率对比

Tab. 6　Comparison of semantic understanding accuracy across different prompts

Prompt类型	语义理解准确率/%			综合准确率/%
	事件类型识别	诱因分析	风险评估	
无约束Prompt	85.2	68.3	62.5	72.0
简单约束Prompt	91.5	78.7	71.2	80.5
本文Prompt（含KG约束）	97.8	92.1	89.6	93.2

2.6　大模型语义推理过程

图3展示了语义理解流程中由融合特征输入、特征解码、知识图谱查询到结构化语义输出的整体过程。语义推理是“从融合特征到结构化语义”的核心步骤，基于“CLIP文本分支＋知识图谱查询”实现，分为以下四步：

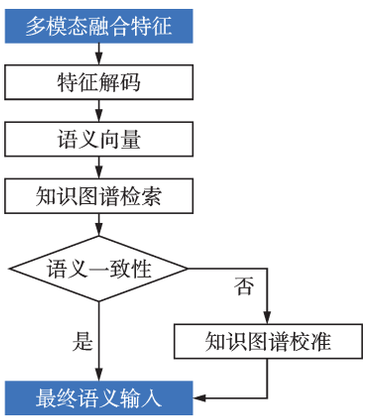

图3　语义理解流程图

Fig.3　Semantic Understanding Flowchart

（1）特征解码：融合特征→语义向量。

将多模态融合特征 F_{fusion}（1024维）输入CLIP的文本分支，通过全连接层将其映射为512维的语义向量 F_{semantic}，该向量与文本描述的语义向量在同一空间中，公式为：

$$F_{\text{semantic}} = \text{ReLU}(\boldsymbol{W} \cdot F_{\text{fusion}} + b) \tag{6}$$

式中，$\boldsymbol{W} \in \mathbf{R}^{512 \times 1024}$ 为权重矩阵；$b \in \mathbf{R}^{512}$ 为偏置项，均为微调过程中学习的参数。

（2）语义生成：语义向量→初步语义结果。基于向量相似度的知识检索与推理机制，计算 F_{semantic} 与知识图谱中“事件类型-诱因-风险”文本向量的余弦相似度，选择相似度最高的文本作为初步语义结果。

（3）知识校验：语义结果→事实一致性验证。通过知识图谱的SPARQL查询，验证初步语义结果的事实一致性，避免“矛盾语义”。

（4）结果校准：校验后→最终语义结果。基于知识校验的置信度调整，对初步语义结果进行校准，输出结构化的最终语义结果。

为适配后续强化学习管控决策层的输入要求，将语义理解结果标准化为“语义特征向量”，包含事件类型向量、诱因向量和风险参数向量3类子向量。

3　实验与结果分析

3.1　实验环境

硬件：Intel Core i9-13900K处理器，64 GB DDR5内存，NVIDIA RTX 4090显卡；软件：Python 3.9，PyTorch 2.0，CLIP官方开源库，Neo4j 5.10，OpenCV 4.8。

3.2　评价指标

3.2.1　异常事件检测指标

采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1分数评估检测性能：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

式中，TP 为真阳性（正确检测异常）；TN 为真阴性（正确检测正常）；FP 为假阳性（正常误判异常）；FN 为假阴性（异常误判正常）。

3.2.2　语义理解指标

采用语义理解准确率（Semantic Accuracy），包括事件类型识别准确率（Acc_{type}）、诱因分析准确率（Acc_{cause}）、风险评估准确率（Acc_{risk}）。综合语义准确率为三者平均值：

$$Acc_{\text{semantic}} = \frac{Acc_{\text{type}} + Acc_{\text{cause}} + Acc_{\text{risk}}}{3} \tag{11}$$

本文按照“整体性能验证—分类别效果分析—语义理解评估—模块贡献分析”的顺序开展实验。首先比较不同模型的总体检测能力；其次统计追尾事故、车辆抛洒物、道路施工、交通拥堵和恶劣天气5类典型事件的识别效果；再评估知识图谱约束与Prompt设计对语义理解的增益；最后通过消融实验分析各组成模块的作用。考虑到不同事件样本规模

与识别难度存在差异，除总体准确率外，同时报告精确率、召回率和F1分数，以避免单一准确率掩盖小样本事件的检测短板。

3.3　异常事件检测性能对比

采用时间序列分层划分，按8∶2的比例将DAIR-V2X数据集的四模态数据划分为训练集与测试集。测试集覆盖追尾事故、车辆抛洒物、道路施工、交通拥堵和恶劣天气5类典型场景，其中训练集用于多模态模型的模态扩展、场景微调及跨模态融合模型训练，测试集为独立数据子集，未参与任何模型训练过程，仅用于异常事件检测性能的对比验证。将本文方法与3类对比方法（传统单模态、无大模型多模态、基础CLIP）进行对比，结果如表7所示。

表7　异常事件检测性能对比

Tab. 7　Performance comparison of anomaly event detection

方法	准确率/%	精确率/%	召回率/%	F1分数/%
视频CNN	82.3	80.5	78.6	79.5
多模态CNN＋MLP	92.1	90.3	89.7	90.0
基础CLIP（未适配）	93.5	91.8	91.2	91.5
本文方法（适配CLIP＋对比学习）	96.2	94.7	94.3	94.5

由表7可知：本文方法的各项检测指标均最优，准确率达96.2%，较视频-CNN提升13.9个百分点。说明经过场景适配后的CLIP能够更充分地利用视频、雷达、流量与气象信息之间的互补关系，从而降低单一模态在复杂场景下的漏检和误检。

3.4　不同类型异常事件检测效果分析

为进一步分析本文方法在不同异常事件类别下的适应性，本文选取追尾事故、车辆抛洒物、道路施工、交通拥堵和恶劣天气5类典型异常场景，分别统计基础CLIP（未适配）与本文方法的F1分数，并计算本文方法相对基础CLIP的提升幅度，结果如表8所示。

由表8可知，本文方法在5类典型异常场景中均取得了高于基础CLIP的F1分数。其中，交通拥堵和追尾事故的识别效果最好，这是因为两类场景在视频形态、流量变化和雷达回波上均具有较稳定的模式；车辆抛洒物的F1分数相对较低，但仍达到92.0%，且较基础CLIP提升4.8个百分点，说明抛洒物确属高速公路异常事件中的高难度类别，其主要难点在于目标尺度小、出现持续时间短、易被前车遮挡，同时对雷达回波和流量扰动的影响较弱。该结果进一步表明，本文方法对不同类型异常事件具有较好的适应性，尤其在小目标、短时异常场景下表现出更明显的增益。

表8　不同类型异常事件检测效果

Tab. 8　Detection performance for different types of anomalous events

事件类型	F1分数/%	
	基础CLIP	本文方法
追尾事故	93.1	96.3
车辆抛洒物	87.2	92.0
道路施工	91.5	94.8
交通拥堵	94.6	97.1
恶劣天气	91.1	93.8
平均值	91.5	94.8

3.5　语义理解性能对比

将本文方法与规则驱动、基础CLIP＋简单Prompt方法对比，结果如表9所示。

表9　语义理解性能对比

Tab. 9　Semantic understanding performance comparison

方法	Acc_{type}/%	Acc_{cause}/%	Acc_{risk}/%	$Acc_{semantic}$/%
规则驱动	85.2	72.3	68.5	75.3
基础CLIP＋简单Prompt	91.5	82.7	79.3	84.5
本文方法（CLIP＋知识图谱Prompt）	97.8	92.1	89.6	93.2

由表9可知，本文方法的综合语义准确率达93.2%，较规则驱动提17.9个百分点。关键在于：知识图谱为大模型提供了领域语义约束，结构化Prompt引导大模型聚焦事件-诱因-风险的关联解析，避免了语义理解的模糊性。

3.6　消融实验

为验证多模态大模型各核心模块的作用，设计消融实验，结果如表10所示。

表10　消融实验结果

Tab. 10　Melting experiment results

方法	检测准确率/%	语义理解准确率/%
本文方法（完整）	96.2	93.2
去除传感器模态扩展	93.8	88.5
去除知识图谱Prompt	95.1	85.7
去除语义特征驱动DRL	96.0	93.0

由表10可知，各核心模块均对性能有贡献。其中传感器模态扩展和知识图谱Prompt的作用最为明显。传感器模态扩展增强了多源数据互补性，知识图谱Prompt则对语义理解深度有显著提升作用。

4 结束语

本文围绕高速公路异常事件监测中多源信息难对齐、识别结果解释性不足的问题，构建了基于适配CLIP的四模态检测方法，并在DAIR-V2X数据集上完成验证。结果表明，适配后的CLIP在整体检测准确率和语义理解准确率上均优于视频CNN、多模态CNN+MLP及基础CLIP等对比方法。分类别结果进一步说明，该方法对追尾、施工、拥堵和恶劣天气等场景具有较稳定的识别能力，对车辆抛洒物这类小目标、高瞬时事件也表现出一定优势。

从方法层面看，本文聚焦两个较为具体的问题：一是通过模态扩展、场景微调与跨模态注意力融合，改善视频、雷达、流量和气象信息的统一表征；二是借助知识图谱约束和结构化Prompt，将检测结果进一步组织为事件类型、诱因和风险等级等可解释输出。上述结果表明，CLIP类视觉-语言模型经过交通场景适配后，具备作为高速公路异常事件监测基础模型的应用潜力。

面向中国未来高速公路场景的推广应用，仍需重点开展以下工作：构建覆盖不同区域、天气、路型和设备条件的本土化多模态数据集，尤其补充抛洒物、次生事故等小样本事件；研究面向边缘侧部署的轻量化压缩和实时推理方案；完善知识图谱持续更新、人工复核和安全校验机制，降低模型误报、漏报与“幻觉”风险；推动异常事件分类、标注和评测规则标准化。只有将模型能力、工程部署和业务规则协同起来，CLIP在高速公路异常事件监测中的应用价值才能更加稳定地转化为实际管控效能。

参考文献：

[1] 2023年交通运输行业发展统计公报[J]. 中国水运，2024(8)：31-35.

[2] 吴晓佩. 基于多SVM分类器融合的高速公路异常事件检测方法[J]. 现代交通技术，2014，11(4)：63-67.

[3] 隋靓，党建武. 基于运动目标轨迹的高速公路异常事件检测算法研究[J]. 计算机应用与软件，2018，35(1)：246-252.

[4] 李旭，张秀杰. 基于视频分析的高速公路异常事件检测[J]. 公路交通科技(应用技术版)，2019(7)：299-302.

[5] 孙倩，郭忠印. 基于离散时间信号相关性的交通事件检测算法[J]. 同济大学学报(自然科学版)，2018，46(11)：1508-1513.

[6] WEIL R，WOOTTON J，GARCÍA-ORTIZ A. Traffic incident detection：sensors and algorithms[J]. Mathematical and Computer Modelling，1998，27(9-11)：257-291.

[7] TENG H L，QI Y. Detection-delay-based freeway incident detection algorithms[J]. Transportation Research Part C：Emerging Technologies，2003，11(3-4)：265-287.

[8] JIN J，RAN B. Automatic freeway incident detection based on fundamental diagrams of traffic flow[J]. Transportation Research Record：Journal of the Transportation Research Board，2009，2099(1)：65-75.

[9] RITCHIE S G，CHEU R L. Simulation of freeway incident detection using artificial neural networks[J]. Transportation Research Part C：Emerging Technologies，1993，1(3)：203-217.

[10] DIA H，ROSE G. Development and evaluation of neural network freeway incident detection models using field data[J]. Transportation Research Part C：Emerging Technologies，1997，5(5)：313-331.

[11] SRINIVASAN D，CHEU R L，POH Y P，et al. Development of an intelligent technique for traffic network incident detection[J]. Engineering Applications of Artificial Intelligence，2000，13(3)：311-322.

[12] SHEHATA M S，CAI J，BADAWY W M，et al. Video-based automatic incident detection for smart roads：the outdoor environmental challenges regarding false alarms[J]. IEEE Transactions on Intelligent Transportation Systems，2008，9(2)：349-360.

[13] BUCH N，VELASTIN S A，ORWELL J. A review of computer vision techniques for the analysis of urban traffic[J]. IEEE Transactions on Intelligent Transportation Systems，2011，12(3)：920-939.

[14] CHEN S Y，WANG W. Decision tree learning for freeway automatic incident detection[J]. Expert Systems with Applications，2009，36(2)：4101-4105.

[15] MA X L，DAI Z，HE Z B，et al. Learning traffic as images：a deep convolutional neural network for large-scale transportation network speed prediction[J]. Sensors，2017，17(4)：818.

[16] YAO Y，XU M Z，WANG Y C，et al. Unsupervised traffic accident detection in first-person videos[C]// Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE，2020：273-280.

[17] XU Y J，HU H，HUANG C W，et al. TAD：a large-scale benchmark for traffic accidents detection from video surveillance[J]. IEEE Access，2025，13：2018-2033.