中文引用格式:安思齐,蔡昂林,马子程,等. 集成多模态大模型的施工安全隐患识别[J]. 中国安全科学学报,2025,35(9):185-192.
英文引用格式:AN Siqi, CAI Anglin, MA Zicheng, et al. Multimodal large model-based approach for construction safety hazard recognition [J]. China Safety Science Journal,2025,35(9):185-192.

集成多模态大模型的施工安全隐患识别*

安思齐[1], 蔡昂林[2], 马子程[2], 朱宝岩**[1]教授
(1 辽宁工程技术大学 安全科学与工程学院,辽宁 葫芦岛 125105;
2 中国矿业大学(北京) 理学院,北京 100083)

中图分类号:X948 **文献标志码**:A **DOI**: 10.16265/j.cnki.issn1003-3033.2025.09.1298

【摘 要】 为提升施工场景中安全隐患的自动识别和安全管理水平,构建一个集成多模态大模型的施工安全隐患识别模型,进而构成其核心组件——多模态安全隐患识别模型 LLaVA-CS(用于施工场景(Construction Site,CS)下的多模态视觉-文本大语言模型(LLaVA));该系统将图像(施工现场照片)与安全操作规程(工人行为描述)相结合,利用多模态学习和深度学习技术,实时监控和分析施工现场;为支持系统的有效运行,构建一个涵盖不同光照、遮挡和多人场景等复杂条件的多模态数据集,弥补现有公开数据集的空白。结果表明:通过对 LLaVA-1.5 模型进行提示调优,LLaVA-CS 模型能有效融合视觉与文本信息,提升安全隐患识别的精度和可解释性。集成该模型的施工安全隐患识别方法在多个实际施工项目中识别准确率达到 0.722 2,能够实时生成详细的解释文本,帮助管理人员快速理解安全隐患的具体情境,增强安全管理的决策支持。将多模态大模型应用于施工安全管理系统,有助于提供实时、可解释的安全监控解决方案。

【关键词】 多模态大模型; 施工安全隐患; 复杂施工场景; 安全管理; 提示调优

Multimodal large model-based approach for construction safety hazard recognition

AN Siqi[1], **CAI Anglin**[2], **MA Zicheng**[2], **ZHU Baoyan**[1]
(1 School of Safety Science and Engineering, Liaoning Technical University, Huludao Liaoning 125105, China; 2 School of Science, China University of Mining and Technology-Beijing, Beijing 100083, China)

Abstract: In order to enhance the automatic recognition of safety hazards and improve safety management in construction scenarios, a multimodal large-model-based method for construction safety hazard recognition was proposed and its core component—the multimodal safety hazard recognition model, LLaVA(Large Language and Vision Assistant)-CS(Construction Site), was implemented. The system integrated images (construction site photos) with safety operating procedures (worker behavior descriptions), leveraging multimodal learning and deep learning technologies to perform real-time monitoring and analysis of construction sites. To support the system´s effective operation, a multimodal

* 文章编号:1003-3033(2025)09-0185-08; 收稿日期:2025-04-11; 修稿日期:2025-06-15
** 通信作者:朱宝岩(1963—),男,辽宁建昌人,博士,教授,主要从事风险预控、安全行为、安全经济、职业健康管理体系、安全与应急管理等方面的研究。E-mail:361232175@qq.com。

dataset covering complex conditions such as varying lighting, occlusions, and multi-person scenarios was constructed, addressing the gaps in existing public datasets. Through prompt tuning of the LLaVA-1.5 model, the LLaVA-CS model effectively integrated visual and textual information, enhancing the accuracy and interpretability of safety hazard recognition. Experimental results show that this method achieves an accuracy of 0.722 2 in multiple real-world construction projects, generating detailed explanatory texts in real time to help managers quickly understand specific safety hazard contexts, thereby improving decision-making in safety management. This study innovatively applies multimodal large models to construction safety management systems, providing real-time, interpretable safety monitoring solutions and offering new technical support and optimization directions for construction safety management.

Keywords: multimodal large model; construction safety hazards; complex construction scenarios; accident prevention; tips for tuning

0 引 言

随着我国城镇化进程的加快,基础设施建设持续扩张。然而,施工人员流动性高、安全培训不足、监管不到位,这些因素增加了施工现场发生安全事故的风险,导致安全隐患频发。复杂的施工环境,如光照变化、遮挡物、背景干扰等,使得对安全隐患的自动识别变得尤为困难。因此,如何在这种动态、复杂的环境下有效识别安全隐患,成为施工安全管理中的一大挑战。

现有的安全隐患识别大多依赖单模态方法。如郁润[1]利用掩膜区域卷积神经网络识别工人和梯子的关系,结合长短期记忆网络分析工人的动作轨迹;范冰倩等[2]通过引入高效通道注意力模块,优化了基于卷积神经网络的不安全行为识别;左明成等[3]利用多人姿态估计和目标识别技术,提升了煤矿井下场景中对安全帽佩戴的识别能力。然而,这些基于图像的单模态方法存在诸多不足。首先,单模态模型通常是“黑盒”,虽然能做出判断,但缺乏透明性和可解释性,使管理人员难以理解具体的不安全行为原因。此外,单模态模型通常依赖单一数据源,无法结合其他数据模态(如文本描述或操作规范)来理解工人行为的背景和意图,从而容易错判行为的安全性。

针对上述不足,已有研究逐步探索多模态方法以增强不安全行为识别的精度与鲁棒性。如李健等[4]构建了结合 2D 和 3D 目标检测的多模态起重伤害预警模型,实现了起重现场危险因素的全面监测。谢定坤等[5]提出基于深度学习的多模态融合方法,通过堆叠式交叉注意力机制提取图像和文本特征,实现多种不安全行为的识别。孙昕璐等[6]采用生理-心理多模态监测技术,通过眼动和脑部活动数据构建隐患识别能力指数,提升了隐患识别的认知理解能力。然而,现有多模态方法基于传统深度学习框架或简单的多模态融合技术,尚未充分优化多模态融合机制,对隐患识别中的深层次行为语义和上下文语义的理解能力依然不足,难以在复杂场景中实现更加精准的识别。

多模态大模型(生成型预训练变换模型 4(Generative Pre-trained Transformer 4, GPT-4)[7]和多模态视觉-文本大语言模型(Large Language and Vision Assistant, LLaVA)[8-9])在图像生成[10]、自然语言处理[7]、跨模态任务[11]中展现出显著的性能优势。GPT-4 是 OpenAI 开发的基于 Transformer 架构的大规模多模态模型,通过自注意力机制捕捉文本中长距离依赖关系。GPT-4 在大规模文本数据上进行预训练,具备强大的自然语言生成和理解能力,同时,部分版本支持多模态输入(文本与图像)。LLaVA-1.5 专注于有效融合视觉与语言信息,其架构结合了视觉编码器与语言模型,通过交叉注意力等机制实现图像与文本特征的深度融合,从而适用于图像描述、视觉问答等跨模态任务。相比 GPT-4,LLaVA-1.5 更加侧重于利用图像信息生成与视觉相关的文本输出。

相比于单模态模型,多模态可为模型提供更多元的信息[12-13]。然而,这些模型尚未在施工场景中进行专门训练,无法有效识别施工现场特有的安全隐患。此外,多模态大模型在面对复杂的施工环境(如光照变化、遮挡、多人协作等)时,仍存在鲁棒性不足的问题,难以及时、准确地识别安全隐患。

鉴于此,笔者拟构建一个集成多模态大模型的施工安全管理系统,并建成其核心组件——多模态安全隐患识别模型 LLaVA-CS,以期为施工安全管理提供新的技术支持和优化方向。

1　多模态大模型的施工安全系统

1.1　收集和整理规则文本

施工安全管理系统收集并整理施工现场的安全规范和操作规程,形成一个系统的规则文本库。这些规则文本来源于国家标准、行业规范以及企业内部的安全作业指导书,涵盖安全操作流程、设备使用规范和个人防护要求等内容。专业的安全管理人员负责定期更新和维护规则文本库,以确保其时效性和完整性。

1.2　收集处理实时图像和视频数据

为持续收集实时的图像和视频数据,在施工现场的关键位置安装固定摄像头,每个工作小组负责人佩戴智能安全帽和工作记录仪等智能穿戴设备。固定摄像头布置在高风险区域、人员密集区和关键作业点。例如:在高空作业区域(如脚手架、吊篮、塔吊),摄像头用于监控工人是否正确佩戴安全带和安全绳,检测违规攀爬、跨越等不安全行为;在大型机械设备操作区(如起重机、挖掘机、打桩机周围),摄像头监控设备操作的规范性,防止人员违规进入危险作业区域;在危险品存放区(如油料、化学品、易燃易爆物品的储存区域),摄像头监督人员的进出和操作,防止未经授权的人员接近或违规操作;在主要出入口、通道和楼梯口,摄像头用于监控人员是否佩戴安全帽、反光衣等安全装备。这些摄像头具备高清晰度、广角镜头、夜视功能和防尘防水等性能,所采集的视频数据通过有线或无线网络实时传输至中央服务器。同时,工人佩戴的智能穿戴设备提供第一视角的数据。智能安全帽能够实时记录工人视野内的图像和视频,捕捉操作细节和周围环境的情况。工作记录仪通常佩戴在胸前或肩部,具备摄像、录音、定位等功能,为事后分析提供依据。

1.3　数据分析与施工安全隐患识别

所有采集到的数据在经过初步处理后,被传输至中央服务器,供模型进行实时分析。初步处理过程包括数据预处理、事件过滤和数据标记等步骤。数据预处理涉及对图像和视频数据进行去噪、增强和压缩等操作,旨在提升数据质量并减少存储和传输压力;事件过滤利用简单的规则或轻量级模型,筛除正常行为和明显的安全行为,集中资源分析潜在的安全隐患;数据标记根据摄像头或设备的编号、位置、时间戳等信息,标注数据,便于模型关联和溯源。

在中央服务器上,系统整合规则文本库与实时图像数据。与单模态模型不同,多模态模型(图 1)在接收到图像数据后,先执行预处理和特征提取,再结合规则文本库中的安全标准,深入分析施工现场的实际情况。通过多模态特征融合技术,模型能够有效匹配视觉信息与文本规则。当模型识别出安全隐患时,会生成详细的分析报告,包括安全隐患的类型、发生位置、涉及人员以及可能的风险等级。

施工安全管理框架如图 2 所示,依托规则文本、图像视频数据与多模态融合模型,实现施工现场安全隐患的实时识别与响应。

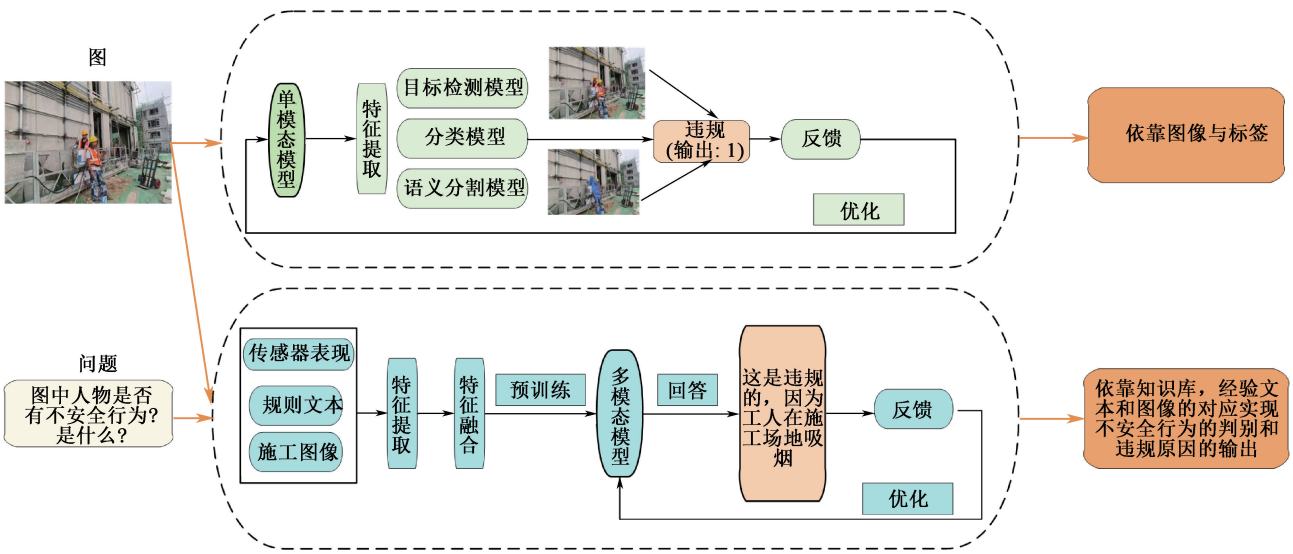

图 1　多模态模型与单模态模型的原理差异

Fig. 1　Principle differences between multimodal and unimodal models

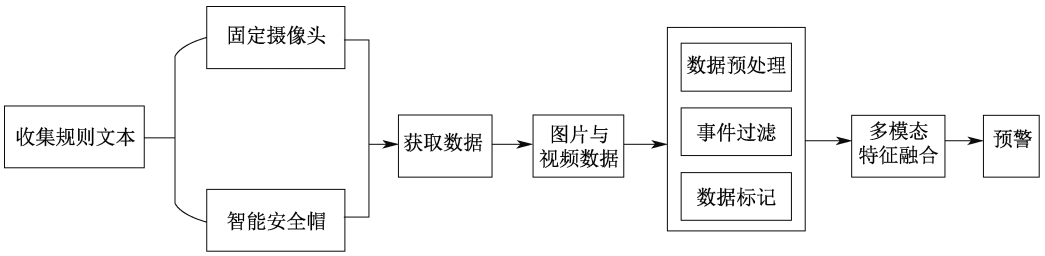

图 2　施工安全管理框架

Fig. 2　Construction safety management framework

1.4　输出警告和预警机制

当检测到安全隐患时，系统会自动输出警告。警告方式包括通过现场的警报器、警示灯等装置，立即提醒现场人员注意安全；向安全管理人员的移动设备发送预警信息和详细描述，便于其及时采取措施；根据安全隐患的严重程度，自动生成处理建议，辅助管理人员进行决策。因此，系统能够在第一时间响应潜在的安全风险。

2　多模态施工安全隐患识别模型

多模态安全隐患识别模型是施工安全管理系统的核心组成部分，基于 LLaVA-1.5[11]，结合多模态融合技术和提示调优，提出用于施工场景（Construction Site，CS）下的 LLaVA，即 LLaVA-CS 模型，如图 3 所示。首先，利用视觉编码器提取图像特征，接着，通过文本编码器获取安全隐患的文本描述；然后，进行多模态特征融合；最后，通过分类器输出判别结果和具体文本描述。

2.1　视觉特征提取

视觉特征提取模块使用 CLIP 视觉编码器，该编码器基于视觉转换器（Vision Transformer，ViT）架构[14]。然而，针对施工场景中的复杂条件（如光照变化、遮挡、背景噪声等），在视觉特征提取过程中引入一系列预处理步骤，以确保从不同环境下的图像中提取到有效特征。首先，在图像进入 ViT 之前，模型会根据当前光照条件（通过图像亮度分析动态调整）进行预处理操作，采用自适应直方图均衡化增强图像对比度，从而确保在光线不足或过亮的情况下，施工人员的动作依然清晰可见。其次，为应对场景中的遮挡问题，采用多视角数据增强技术，通过构建从不同角度采集的图像样本，增强模型对遮挡区域信息的补全能力。在特征提取阶段，ViT 将施工场景图像划分为固定大小的图像块（patch），并将其转化为高维特征向量。最后，通过这种多重预处理和特征融合方法，模型能够更好地完成复杂环境中的视觉信息提取需求。

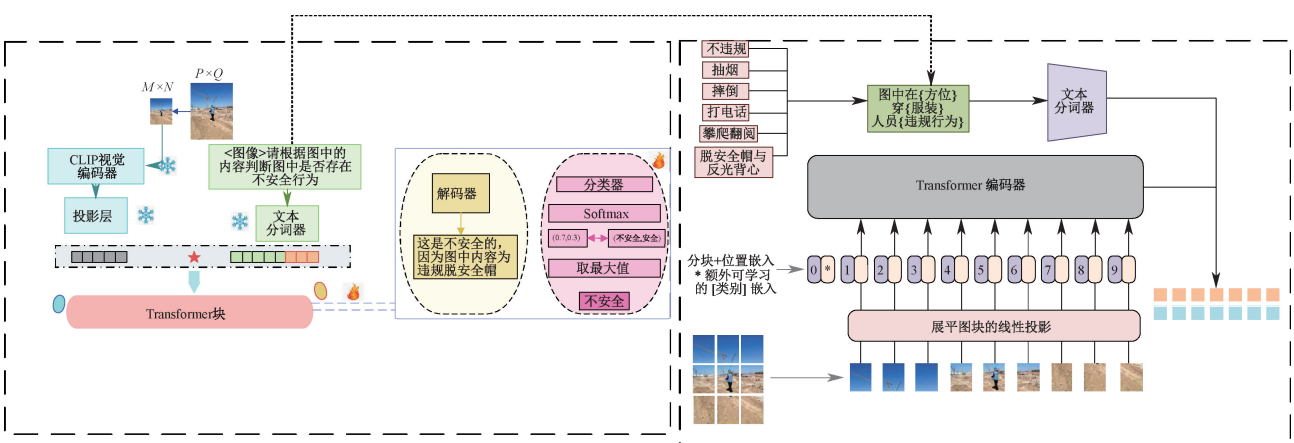

图 3　安全隐患识别系统框架

Fig. 3　Frame diagram of unsafe behavior recognition system

2.2　文本特征提取

与多模态学习中的语言模型（如 GPT[15] 和双向变换器模型（Bidirectional Encoder Representations from Transformers，BERT）[16]）的应用理念类似，文本特征提取模块使用经过预训练的文本分词器，将施工场景的文本信息（如违规行为描述）转化为模

型可处理的 token 序列。通过文本嵌入将这些 token 转换为固定维度的向量,与图像特征对齐,形成共同的嵌入空间。通过预训练的语言模型,使文本嵌入能有效地捕捉文本含义并与图像特征进行整合。

2.3　多模态特征融合

多模态特征融合模块通过自注意力机制[17]实现视觉和文本特征的深度结合。在施工场景的应用中,由于图像和文本信息的多样性和复杂性,提出一种针对多模态融合的动态权重调整机制。

在动态权重调整中,结合施工场景的环境条件(如光线、视角、遮挡等),根据这些条件动态调整视觉特征 X_v 和文本特征 X_t 的权重。权重函数为 $\alpha_v(c)$ 和 $\alpha_t(c)$,其中,c 表示当前的环境条件。c 通过输入图像的特定指标(亮度、遮挡程度等)计算得到,并将其作为调整权重的依据。根据环境条件,定义视觉和文本特征的动态权重调整公式:

$$X_f = X_v\alpha_v(c) + X_t\alpha_t(c) \tag{1}$$

其中,权重满足:

$$\alpha_v(c) + \alpha_t(c) = 1 \tag{2}$$

在传统自注意力机制的基础上,模型根据施工现场环境条件(如光线、视角、遮挡)动态调整视觉和文本特征的权重。当图像中遮挡严重时,$\alpha_v(c)$ 降低,$\alpha_v(c)$ 增加;当图像质量良好且文本信息不完整时,$\alpha_t(c)$ 降低,$\alpha_v(c)$ 增加。图像遮挡程度通过检测某些区域的像素变化率 O 来表示,那么权重可定义为:

$$\alpha_v(c) = \frac{1}{1 + \exp(\beta(O - K_0))}, \alpha_t(c) = 1 - \alpha_v(c) \tag{3}$$

式中:β 为控制权重变化的敏感度参数;K_0 为遮挡程度的阈值。当 K 较高时(即遮挡严重),$\alpha_v(c)$ 会趋近于 0,而 $\alpha_t(c)$ 会接近 1,使得模型更多依赖于文本特征。

2.4　输出判别结果与文本描述

在输出模块中,优化传统多层感知机分类器,结合多模态特征的动态加权结果进行分类,并生成对应的违规行为描述。与传统方法不同的是,设计一个基于情景推理的多模态生成机制,能够根据图像场景的复杂性生成更加细致的违规行为描述。具体来说,当模型检测到多个安全隐患时,输出模块会结合上下文信息,生成逻辑上连贯的多轮对话的描述,确保最终的违规理由能够详细覆盖所有被发现的安全隐患。

此外,为提高分类器在复杂施工场景中的鲁棒性,引入基于注意力机制的多层自校验过程,模型在生成文本描述的同时,实时检查生成的内容是否与视觉特征和分类结果一致,从而避免因模态之间的信息不对称而导致的错误分类。最终,通过 Softmax 函数生成类别概率分布,并通过 Argmax 输出最可能的安全隐患类别,同时生成具有逻辑一致性的违规行为描述。

3　多模态施工安全隐患识别模型训练与分析

3.1　训练数据集

依据“2-4”模型(第 6 版)[18]当中对个体动作的描述,以及安全信用模型[19]中人的安全隐患的分析;同时,根据典型安全隐患问题与正确做法图集[20]以及历年住房和城乡建设部办公厅房屋市政工程生产安全事故情况的通报,确定 5 种常见且易引发安全事故的安全隐患作为识别对象,即施工场地违规抽烟、攀爬翻越安全护栏、施工现场未佩戴安全帽与反光背心、施工场地违规接打电话和施工场地摔倒。与传统分类方法相比,所选择的安全隐患在事故致因中占据显著比例[21],且在施工现场中危险程度较大,导致事故发生的风险较大,与多模态识别技术结合,预防潜力巨大。

为获取安全隐患图像数据,结合网络搜索、施工现场模拟和实地拍摄等方法收集样本。在收集数据的过程中,模拟多个施工现场的安全隐患场景,并通过不同视角、多角度拍摄图像,确保数据样本的多样性和全面性。构建施工现场安全隐患数据的多模态数据集包括 2 个模态:①图像模态是指在施工场景下所采集到的照片;②文本模态是对图片中工人行为的文字描述,包含对其动作、姿势以及潜在违规行为的解释。

为提升模型在复杂施工场景中的适应性,构建的多模态数据集涵盖多种复杂条件,包括不同光照条件、遮挡情况和多人场景。光照条件部分,数据采集覆盖强光(如晴天正午的强烈直射光)、弱光(如清晨或傍晚低光照环境)、背光(光源位于主体背后形成明暗对比)以及阴影区域(由建筑物或设备投射形成的局部暗区),以全面模拟施工现场光线条件的多变特性。遮挡情况通过 2 种方式实现:①使用施工材料(如脚手架、钢板、工具等)人为遮挡工人身体或操作区域;②选择实际施工场景中设备、障碍物遮挡的自然情况,并结合多视角采集增强模型对遮挡信息的补全能力。多人场景数据涵盖 2~

5 人的多工种协作与行为交互,设计搬运、指挥、操作等典型工种任务场景,捕捉工人之间的协作关系和行为模式。最终,收集并处理的数据样本共计 1 897 张,涵盖 5 个违规类别,其中,约 20%为弱光或背光场景,15%涉及遮挡样本,25%涵盖多人协作场景。施工安全隐患标签数量汇总见表 1。

表 1　施工安全隐患标签数量汇总

Table 1　A summary of the number of unsafe behavior labels

标签类别	数量	比例/%
不违规	211	8.42
抽烟	286	11.41
攀爬翻越安全护栏	374	14.92
未佩戴安全帽与反光背心	991	39.55
接打电话	317	12.65
摔倒	327	13.05
合计	2506	100

3.2　多模态施工安全隐患识别试验设计

使用 Pytorch 实现代码,利用 8 张 H800 图形处理器对模型进行高效的提示调优训练。安全隐患的图像数据集按照 7∶3 随机划分为训练集与测试集,以确保数据分布的代表性与测试结果的可靠性。训练过程中,采用准确率作为核心评估指标,深入分析模型在不同训练步骤中的表现。

在初始阶段,每张图像 X_v(施工现场的照片)生成对应的多轮对话数据 $(X_q^1,X_a^1),\cdots,(X_q^T,X_a^T)$,其中,$T$ 为总对话轮数。问题 X_q 涵盖对施工场景中安全隐患的检测需求(如请判断图中是否存在安全隐患),答案 X_a 则提供具体的判断和解释。在生成过程中,对于第 1 轮对话 $T=1$,模型通过不同输入形式组合 $[X_q^1,X_v]$ 或 $[X_v,X_q^1]$ 引导模型回答,后续轮次 $T>1$ 则直接使用当前问题 X_q^t 生成相应的回答。在此基础上,通过提示调优调整大型语言模型[15],特别是对于安全隐患描述的不确定性,通过优化提示设计,使得模型在提示的指引下关注图像中需要重点关注的区域,确保模型能够有效利用视觉和文本信息,生成准确且合乎场景的判断。对于一个长度为 L 的序列,目标答案的条件概率计算如下:

$$p(X_a \mid X_v,X_s)=\prod_{i=1}^{L} p_\theta(x_i X_v,X_s^{<i},X_a^{<i}) \tag{4}$$

式中:θ 为可训练参数;$X_s^{<i}$ 和 $X_a^{<i}$ 分别为当前预测标记 x_i 之前的所有指令和回答标记。式(1)特别强调引入图像 X_v 作为生成所有答案的基础,以确保视觉信息的充分利用。

在提示调优之后,模型进入端到端微调阶段。此阶段固定视觉编码器的权重,重点微调投影矩阵 $\boldsymbol{W}$ 和语言模型的权重 ϕ,进一步优化模型性能,最大化目标答案的条件概率。

3.3　施工安全隐患判定能力

模型训练准确率随训练轮数的变化如图 4 所示。由图 4 可知:LLaVA-CS 模型的准确率从训练初期的 0.513 9 逐步提升至 0.722 2,表明模型的不断学习和优化显著增强了其对施工现场安全隐患的识别能力。在第 8—第 12 个训练周期,模型的准确率从 0.652 8 升至 0.708 3,展示出模型对安全隐患特征的逐渐掌握。最终,在第 19 个周期,模型达到 0.722 2 的最高准确率,并保持了稳定性。

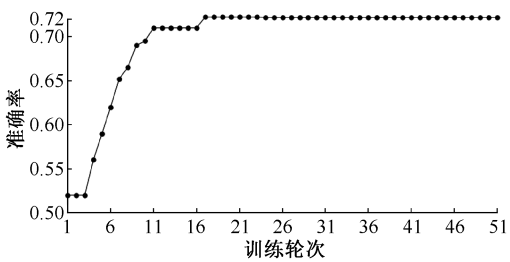

图 4　模型训练准确率随训练轮数的变化

Fig. 4　Training accuracy over epochs

各模型在安全隐患识别任务中的精确率对比见表 2,从表 2 可以看出,LLaVA-CS 模型在精确率方面表现最为优异,达到 0.722 2,显著领先于其他对比方法。这表明:LLaVA-CS 在施工场景中的安全隐患识别任务上具有明显的优势。相比之下,EfficientNet 紧随其后,精确率为 0.715 6,Swin Transformer 的精确率为 0.637 5,表明其对安全隐患的识别能力有限。ResNet 的精确率仅为 0.362 4,显示出该模型在复杂施工场景下存在较大的局限性,难以捕捉关键的安全隐患特征。

表 2　各模型在施工安全隐患识别任务中的精确率对比

Table 2　Comparison of model precision in hazard recognition tasks

模型	精确率
LLaVA-CS	0.722 2
EfficientNet	0.705 6
Swin Transformer	0.637 5
ResNet	0.362 4

通过对比分析发现,LLaVA-CS 模型在训练过程中有效融合了多模态信息,并利用其深度学习架构中的改进机制,充分挖掘了施工场景中安全隐患

的语义特征和视觉特征,进而在整体表现上显著超越了其他方法。

3.4　模型多模态输出能力分析

为验证模型在不同施工场景下的表现,设计一系列基于多模态输出的试验,并与 GPT-4 和 LLaVA-1.5 进行对比,模型之间的共性与差异见表 3,试验结果见表 4。结果表明:LLaVA-CS 在图像理解、语义准确性、上下文推理等方面表现优异,特别是在复杂场景中,能够通过图像和文本的有效结合,生成合理的解释。相比之下,GPT-4 和 LLaVA-1.5 在不具备施工领域专业知识的情况下,虽然能够识别部分安全隐患,但常常在细节上出现误判,尤其是在处理复杂或模糊场景时,二者均表现出局限性。

表 3　模型共性与差异对比

Table 3　Comparative analysis of model commonalities and differences

特性	GPT-4	LLaVA-1.5	LLaVA-CS
模型类型	通用大规模语言模型,具备多模态能力(文本/图像)	多模态模型(文本/图像)	专用多模态模型,针对建筑安全隐患识别进行领域优化
架构	基于 Transformer 架构	基于 Transformer 结合视觉编码器	在 LLaVA-1.5 基础上引入动态特征融合与上下文适应机制
输入数据	文本输入为主,但支持图像输入	同时处理图像与文本数据	建筑工地图像与文本数据
主要应用	文本生成、问答、对话等通用自然语言处理任务	图像描述、视觉问答等跨模态任务	建筑安全隐患的自动识别与解释,为施工安全监控提供决策支持
训练数据	多领域大规模文本数据,部分版本包含图像数据	综合图像与文本数据集,专注于跨模态信息融合任务	定制化多模态数据集,专注于建筑工地复杂场景中的安全隐患识别

表 4　试验案例列举

Table 4　Examples in the experiment

请根据图中的内容判断图中是否存在安全隐患,并给出判断				
—				
LLaVA-CS	这是不安全的,因为图中人物在吸烟且缺少防护措施	这是不安全的,因为图中内容为缺少防护措施	图中不存在安全隐患,因为图中内容为违规脱安全帽	这是安全的
GPT-4	这是不安全的,因为图中工人在高处工作时没有使用安全带,并且施工平台可能存在坠落风险	图中的工人也戴着安全帽和反光背心,看起来他在施工现场。这是安全的行为,没有明显的安全隐患可见	基于图中内容的判断,没有发现明显的安全隐患	这是不安全的,因为图中的人物正在做出不雅手势
LLaVA-1.5	这是不安全的,因为图中人物在没有使用安全带的情况下站在高处	这是不安全的,因为图中人物在没有使用任何个人保护设备的情况下,进入了一个正在施工的区域	这是不安全的。一个工人在高处作业时没有使用安全带	这是不安全的,因为图中人物在高处作业,没有使用安全带

LLaVA-CS 在一些复杂或模糊场景中仍具有局限性,如未能准确识别安全隐患或生成的描述与图像内容不符。这可能是由于训练数据分布不均衡、安全隐患样本不足,以及安全隐患类别有限,导致模型在这些情况下的表现不佳。

4　结　论

1) 集成多模态大模型的视频监控安全管理系统能够精准识别施工现场常见的安全隐患,并根据用户的提问,生成工人安全隐患的具体描

述,为施工现场的智能化安全监控奠定坚实的基础。

2) 构建一个专门用于复杂施工场景的多模态安全隐患数据集,涵盖施工现场常见的安全隐患类型,并包含不同光照条件、遮挡情况和多人场景的数据,填补现有公开数据集的空白。

3)受样本覆盖不足与类别分布不均衡影响,模型在个别行为识别上仍存在不稳定性。后续将重点扩充并均衡更大规模、高质量的多模态隐患数据集,以进一步提升识别精度与泛化能力。

参 考 文 献

[1] 郁润. 基于计算机视觉的施工现场工人不安全行为识别方法研究[D]. 北京:清华大学,2019.
YU Run. Computer-vision-based method for the recognition of construction workers´ unsafe behaviors [D]. Beijing: Tsinghua University, 2019.
[2] 范冰倩,董秉聿,王彪,等. 基于深度学习的地铁施工作业人员不安全行为识别与应用[J]. 中国安全科学学报,2023,33(1):41-47.
FAN Bingqian, DONG Bingyu, WANG Biao, et al. Identification and application of unsafe behaviors of subway construction workers based on deep learning [J]. China Safety Science Journal, 2023, 33(1): 41-47.
[3] 左明成,焦文华. 面向煤矿井下作业场景的安全帽佩戴识别算法[J]. 中国安全科学学报,2024,34(3):237-246.
ZUO Mingcheng, JIAO Wenhua. Helmet-wearing recognition algorithm for coal mine underground operation scenarios [J]. China Safety Science Journal, 2024, 34(3): 237-246.
[4] 李健,奥帅,张在成,等. 基于多模态的装配式建筑起重伤害预警模型[C]. 2022 年工业建筑学术交流会论文集(上册),2022:440-444.
[5] 谢定坤. 多模态融合的施工现场工人不安全行为识别方法研究[D]. 武汉:华中科技大学,2020.
XIE Dingkun. A multimodal fusion approach for identifying unsafe behavior in Construction [D]. Wuhan: Huazhong University of Science and Technology, 2020.
[6] 孙昕璐. 基于生理心理多模态监测的施工现场隐患识别能力评估[D]. 北京:清华大学,2020.
SUN Xinlu. A multimodal study to assess hazard recognition ability on construction site [D]. Beijing: Tsinghua University, 2020.
[7] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[R]. OpenAI,2023.
[8] LIU Haotian,LI Chunyuan,WU Qingyang,et al. Visual instruction tuning[J]. Advances In Neural Information Processing Systems, 2023, 36: 34 892-34 916.
[9] LIU Haotian, LI Chunyuan, WU Qingyang, et al. Llava-plus: learning to use tools for creating multimodal agents[C]. European Conference on Computer Vision, 2024: 126-142.
[10] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]. International Conference on Machine Learning, 2021: 8 821-8 831.
[11] LIU Haotian,LI Chunyuan,WU Qingyang,et al. Improved baselines with visual instruction tuning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2024: 26 296-26 306.
[12] JIA Chao,YANG Yinfei, XIA Ye, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]. International Conference on Machine Learning, 2021: 4 904-4 916.
[13] RADFORD A, KIM J, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]. International Conference On Machine Learning, 2021: 8 748-8 763.
[14] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale [C]. International Conference on Learning Representations,2020: 1-20.
[15] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances In Neural Information Processing Systems,2020,33:1 877-1 901.
[16] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 :Long and Short Papers,2019: 4 171-4 186.
[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 6 000-6 010.
[18] 傅贵,陈奕燃,许素睿,等. 事故致因"2-4"模型的内涵解析及第 6 版的研究[J]. 中国安全科学学报,2022,32(1):12-19.
FU Gui, CHEN Yiran, XU Surui, et al. Detailed explanations of 24Model and development of its 6th version [J]. China Safety Science Journal, 2022, 32(1):12-19.
[19] 安思齐. 建筑业安全信用评价研究[D]. 葫芦岛:辽宁工程技术大学,2023.
AN Siqi. Research on safety credit evaluation of construction industry[D]. Huludao: Liaoning Technical University, 2023.
[20] 广东省住房和城乡建设厅. 关于推广使用《广东省建筑施工安全生产隐患识别图集(三)》的通知[EB/OL]. (2023-05-06). https://zfcxjst.gd.gov.cn/gkmlpt/content/4/4180/post_4180540.html#1422.
[21] 温国锋,房颖,张帆帆. 复杂工程项目施工阶段行为风险评价模型[J]. 中国安全科学学报,2017,27(8):162-168.
WEN Guofeng, FANG Ying, ZHANG Fanfan. Model for evaluating behavior risk in construction stage of complex construction project [J]. China Safety Science Journal, 2017, 27(8): 162-168.

作者简介: 安思齐 (1998—),男,黑龙江大庆人,硕士,主要从事安全管理、安全信用评价等方面的工作。E-mail:ansi.qi@163.com。