中文引用格式:王喆,黄海辰,李瑞钦,等. 基于视觉语言多模态的建筑施工安全智能问答模型[J]. 中国安全科学学报,2025,35(10):106-114.

英文引用格式:WANG Zhe, HUANG Haichen, LI Ruiqin, et al. Intelligent question answering model for construction safety hazards based on vision-language multimodality[J]. China Safety Science Journal, 2025, 35(10): 106-114.

基于视觉语言多模态的建筑施工安全智能问答模型*

王 喆[1,2]副教授, 黄海辰[1,2], 李瑞钦[1,2], 魏永长[3]副教授

(1 武汉理工大学 安全科学与应急管理学院,湖北 武汉 430070;
2 武汉理工大学 中国应急管理研究中心,湖北 武汉 430070;
3 中南财经政法大学 工商管理学院,湖北 武汉 430073)

中图分类号:X948;TU714 **文献标志码**:A **DOI**: 10.16265/j.cnki.issn1003-3033.2025.10.1435

基金项目:教育部人文社会科学研究青年基金资助(21YJA630094, 20YJC630154);中央高校基本科研业务费专项资金项目(104972024DZB0003)。

【摘 要】 为提升建筑施工复杂环境下安全问题的智能化诊断水平,提出一种基于视觉语言多模态的建筑施工安全智能问答模型,构建建筑施工安全隐患图文对数据集,采用视觉编码器完成安全隐患图像的视觉编码,利用语言模型实现安全隐患问答文本的编码,通过多模态特征融合模块达成图像与文本信息的有效交互;构建适配建筑施工安全隐患场景视觉问答的特定提示模板,基于矩阵低秩分解对模型微调训练,并通过多轮提示词引导模型生成精确答案。结果表明:相较于现有对比模型,建筑施工安全智能问答模型在自动评估指标、GPT-4评价和专家评价中均表现更优,生成文本的流畅性与语义相关性显著提升;消融试验进一步验证了各子模块的有效性,证实矩阵低秩分解微调和多轮推理的协同作用是模型达成最优性能的关键,且合理设置低秩矩阵的秩参数可有效避免过拟合问题。

【关键词】 视觉语言; 多模态; 建筑施工安全; 安全隐患; 智能问答模型; 矩阵低秩分解

Intelligent question answering model for construction safety hazards based on vision-language multimodality

WANG Zhe[1,2], **HUANG Haichen**[1,2], **LI Ruiqin**[1,2], **WEI Yongchang**[3]

(1 School of Safety Science and Emergency Management, Wuhan University of Technology, Wuhan Hubei 430070, China; 2 China Emergency Management Research Center, Wuhan University of Technology, Wuhan Hubei 430070, China; 3 College of Business Administration, Zhongnan University of Economics and Law, Wuhan Hubei 430073, China)

Abstract: In order to enhance the intelligent diagnosis level of safety problems in complex construction environments, an intelligent question-answering model for construction safety hazards based on vision-language multimodality was proposed. A dataset of image-text pairs related to construction safety hazards was constructed. A visual encoder was used to complete the visual encoding of safety hazard images, and a

* 文章编号:1003-3033(2025)10-0106-09; 收稿日期:2025-06-10; 修稿日期:2025-08-13

language model was employed to encode the question-answering texts about safety hazards. A multimodal feature fusion module was adopted to achieve effective interaction between image and text information. A specific input template for visual question answering adapted to the scenario of construction safety hazards was constructed. The model was fine-tuned based on matrix low-rank decomposition, and multi-round prompts were used to guide the model in generating accurate answers. The results show that compared with existing contrastive models, the intelligent question-answering model for construction safety hazards performs better in automatic evaluation metrics, Generative Pre-trained Transformer(GPT)-4 evaluation, and expert evaluation, with significantly improved fluency and semantic relevance of the generated texts. Ablation experiments further verify the effectiveness of each sub-module, confirming that the synergistic effect of matrix low-rank decomposition fine-tuning and multi-round reasoning is the key for the model to achieve optimal performance, and that reasonably setting the rank parameter of the low-rank matrix can effectively avoid the overfitting problem.

Keywords: vision-language; multimodality; construction safety; safety hazard; intelligent question answering model; matrix low-rank decomposition

0 引　言

我国房屋建筑与市政基础设施工程建设规模持续扩大,安全生产面临严峻挑战。据统计,2020 年,全国房屋及市政工程领域共发生生产安全事故 689 起,累计造成 794 人死亡[1]。建筑施工具有工序繁琐、作业环境复杂、交叉作业多等特点[2],而人工巡检多依赖人员主观判断与经验积累,易因个体认知差异导致误判或漏检等。住房和城乡建设部发文明确提出,需加强人工智能等新技术在工程建造过程中的集成与应用[3]。因此,借助人工智能技术提升建筑施工复杂场景下安全隐患的智能化诊断水平,已成为推动建筑行业安全管理模式升级、满足行业高质量发展需求的关键路径。

近年来,深度学习技术的快速发展显著提升了机器视觉的特征提取能力,已广泛应用于建筑施工安全领域的外脚手架隐患图像识别[4]、安全帽佩戴检测[5]、现场危险区域识别[6]等任务,有效实现了施工安全隐患的自动化识别。与此同时,预训练语言模型通过大规模语料库的预训练过程,已具备基本文本语义理解能力。依据模型架构设计与训练任务设定的差异,经二次微调后,可较好适配下游文本理解或文本生成类任务,并在水电工程施工安全隐患文本分类[7]、消防领域应急决策支持[8]、煤矿安全领域智能问答[9]等场景中展现出优异的应用性能。然而,建筑施工安全管理需结合现场视觉信息与管理者知识经验制定安全处置措施,当前研究多局限于单一数据源,亟需开发可从多模态数据中理解复杂施工环境的智能技术。

在视觉语言多模态领域,现有研究多通过构建视觉语言模型解决图-文匹配、视觉问答等核心任务。其中,预训练方法因可在少量标注样本下提升模型性能,被广泛应用于视觉问答与推理任务[10]。CLIP(Contrastive Language-Image Pre-training)[11]作为代表性的预训练视觉语言模型,采用视觉编码器和文本编码器构成的双塔结构,通过余弦损失函数对比学习特征空间中的图像和文本,实现跨模态信息对齐,在图文检索等任务中表现出良好性能。此外,部分研究通过组合预训练视觉模型和语言模型,借助合适的多模态特征融合机制开展图-文联合训练,以实现多模态信息的交互和对齐。BLIP-2(Bootstrapping Language-Image Pre-training 2)[12]以预训练的视觉转换器(Vision Transformer,ViT)为视觉编码器提取图像特征,设计轻量级查询转换器(Querying Transformer,Q-Former)结构,以桥接冻结的视觉编码器和大语言模型,通过构建多模态损失函数完成图-文联合训练以实现跨模态信息交互和对齐,并进一步利用大语言模型的文本处理能力,完成视觉问答等下游任务。然而 CLIP 的判别式预训练仅能学习粗粒度图文表征,难以适配下游具体任务[13];基于掩码语言模型的文本编码器虽适用于理解任务,但受损失函数制约,难以用于开放式视觉问答。

鉴于此,笔者拟基于采集的建筑施工安全隐患图-文对数据集,构建基于视觉语言多模态的安全智能问答模型,设计针对建筑施工安全隐患场景视觉问答的特定提示模板,引入矩阵低秩分解的微调训练方法对模型微调训练。在此基础上,设计实例从多维度对比测试该模型与现有通用视觉语言模型

的生成答案,以全面评估模型性能,进而提升建筑施工复杂环境中安全问题的智能化诊断能力及安全管理的智能化水平。

1 图-文对数据集制作

1.1 数据采集

数据来源于武汉市某房屋建设工程施工项目日常人工巡检所拍摄的安全隐患照片和安全检查报告。其中,安全检查报告详细记录了施工现场的一项或多项安全隐患,每条隐患记录均包含工程项目基本信息、检查时间、检查内容、检查要求、整改措施等非结构化文本数据,以及直观呈现安全隐患的图像。

1.2 数据预处理

1) 数据筛选与提取。邀请专家人工筛选出安全管理人员在隐患排查治理环节中判定为需实施扣分处罚的经典记录,作为模型训练与验证的核心数据;同时,剔除可能对后续分析或模型训练产生负面影响的低质量数据,包括:①图像模糊、曝光异常(过度/不足)或存在障碍物遮挡,导致无法清晰呈现隐患细节的记录;②安全隐患描述或整改措施过于笼统、缺乏具体信息的记录,如"作业脚手架存在较大隐患"等。基于安全隐患图像,从检查内容和检查要求中人工提取与图像对应的安全隐患描述文本,从整改措施中人工提取对应的解决措施文本,并将二者分别作为2个文本标签。

2) 数据增强。为尽量降低拍摄设备、光线等因素导致的图像质量下降(如细节模糊)对模型的影响,增强模型鲁棒性,采用对原始图像随机添加高斯噪声或高斯模糊的策略,以丰富图像数据多样性。将处理后的图像与上述2个文本标签进行配对,构建结构化的图-文对数据集。

1.3 图-文对数据集概览

建筑施工安全隐患图-文对数据集包含282个数据样本,每个样本均由安全隐患图像、安全隐患描述文本、安全隐患解决措施文本3部分组成,如图1所示。其中,安全隐患图像呈现隐患的视觉信息,安全隐患描述文本为对图像中隐患的文字说明,安全隐患解决措施文本则是针对该隐患的具体整改方案。该数据集按照6∶2∶2的比例划分为训练集、验证集和测试集。

图1 建筑施工安全隐患图-文对数据示例(部分)

Fig. 1 Construction safety hazard image-text data example (partial)

建筑施工安全隐患图-文对数据集涵盖高处作业等多种隐患类型,具体的建筑施工安全隐患类型见表1。

表1 建筑施工安全隐患类型

Table 1 Construction safety hazard types

隐患类型	安全隐患
高处作业	坠落基准面超2 m的临空侧未设置临边防护等
作业脚手架	作业脚手架未按要求设置剪刀撑等
高处作业吊篮	吊篮未落地停放,作业人员未从地面进出吊篮等

续表1

隐患类型	安全隐患
施工用电	电气设备不带电的外露可导电部分未做保护接零等
模板支架	未按规范要求设置扫地杆等
基坑工程	基坑周边堆载不符合规范等
施工机具	圆盘锯未设置安全防护罩等
消防	部分消防栓内无水枪及水带等
文明施工	现场焚烧建筑垃圾等
其他	施工现场人员未佩戴安全帽等

2　建筑施工安全智能问答模型

2.1　安全智能问答模型整体框架

为有效融合建筑施工安全隐患的图像与文本信息、提升视觉问答任务答案的灵活性,解决专业领域视觉问答中多模态理解复杂、专业性强且样本较少的难题,提出一种基于视觉语言多模态的建筑施工安全智能问答模型(Construction Safety Intelligent Question-Answering Model,CSIQM),其整体框架如图 2 所示。

模型包含 3 个关键元素:视觉编码器、语言模型及多模态特征融合模块。其中,视觉编码器采用预训练的 ViT 提取安全隐患的视觉特征,采用 ChatGLM 语言模型提取文本特征,多模态特征融合模块通过一组可学习的查询向量,实现图像和文本特征的高效交互融合。为优化模型在特定任务上的性能,构建适配建筑施工安全隐患场景视觉问答的特定提示模板,基于矩阵低秩分解对模型微调训练;训练过程中,冻结视觉编码器和语言模型参数,仅更新多模态特征融合模块及低秩矩阵,在保留预训练模型特征提取能力的同时减少训练参数量;推理阶段采用多轮提示词推理方法,逐步引导模型生成更精确、全面的答案。

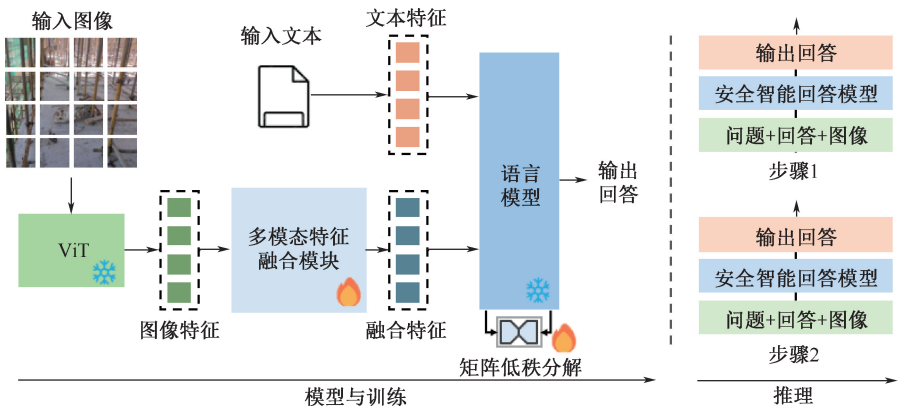

图 2　建筑施工安全智能问答模型整体框架

Fig. 2　Overall framework of construction safety intelligent question-answering model

2.2　ViT 视觉编码器

ViT 采用 Transformer 架构,通过自注意力机制在图像分类等任务上的性能已超越卷积神经网络[14]等深度学习模型。将 ViT 作为视觉编码器,借助自注意力机制高效捕捉图像各部分间的关联信息,实现全局特征的提取,不仅能精准识别安全隐患图像中的工人、机械设备、建筑材料等元素,还可捕捉这些元素间的空间关系和动态关系。ViT 视觉编码器对安全隐患图像的特征提取过程主要分为图像嵌入和多层特征提取 2 个步骤:

1) 图像嵌入。将给定的安全隐患图像分割为 N 个图像块,每个图像块 v 被可学习嵌入矩阵 $\boldsymbol{E}$ 线性投影成图像嵌入向量序列;随后在该序列前添加参数化的类别令牌(用于存储全局视觉特征),并为整个序列附加位置编码以保留图像块的位置信息,最终得到安全隐患图像的输入序列 z_0。计算过程见下式:

$$z_0 = (v_0; v_1^{\mathrm{p}}\boldsymbol{E}; v_2^{\mathrm{p}}\boldsymbol{E}; \cdots; v_N^{\mathrm{p}}\boldsymbol{E}) + \boldsymbol{E}_{\mathrm{p}} \tag{1}$$

式中:v_0 为参数化的类别令牌;$v_1^{\mathrm{p}}, v_2^{\mathrm{p}}, \cdots, v_N^{\mathrm{p}}$ 为 N 个图像块;$\boldsymbol{E}_{\mathrm{p}}$ 为位置编码矩阵。

2) 视觉特征提取。图像输入序列经多层 Transformer 编码器完成特征提取,每层编码器均由多头自注意力机制和多层感知机 2 个核心组件构成。在多头自注意力机制中,各注意力头为独立的自注意力机制,将输入向量与 1 个权重矩阵相乘,分别计算查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$,随后计算注意力分数 A:

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \sigma\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{2}$$

式中:$\sigma(x)$ 为 softmax 归一化指数激活函数,用于归一化处理注意力矩阵;d 为缩放因子。

所有注意力头的结果拼接起来,由线性变换得到最终的输出结果。

多层感知机用于视觉特征的非线性转换,由 2 个线性变换层组成,中间采用 Gelu 非线性激活函数。

视觉特征提取整体计算过程如下式:

$$z'_l = \theta(\varphi(z_{l-1})) + z_{l-1},\ l = 1,2,\cdots,L \tag{3}$$

$$z_l = \mu(\varphi(z_l')) + z_l',\ l = 1,2,\cdots,L \tag{4}$$

式中：z'_l 为 Transformer 编码器中间层输出的视觉特征；$\theta(x)$ 为多头自注意力机制；$\varphi(x)$ 为层归一化；z_{l-1} 为第 $l-1$ 层 Transformer 编码器输出的视觉特征；L 为 Transformer 编码器层数；z_l 为第 l 层 Transformer 编码器输出的视觉特征；$\mu(x)$ 为多层感知机。

2.3 ChatGLM

预训练语言模型能捕捉自然语言的深层特征（包括语法、语义及上下文信息等），将其引入建筑施工安全领域，有助于提升对安全隐患相关文本信息的理解能力。

ChatGLM 是基于通用语言模型（General Language Model，GLM）架构构建的大规模预训练语言模型，其网络结构由多层编码器块堆叠组成，通过自回归生成机制实现文本建模，并在中文自然语言处理任务中展现出卓越性能。该模型的核心建模机制依赖于多头自注意力计算，通过动态计算输入序列中各令牌位置的注意力权重分布，实现对文本全局依赖关系的有效捕获。

在预训练阶段，ChatGLM 引入自回归空白填空任务作为预训练目标，通过随机掩码操作构造不完整上下文，要求模型以自回归方式预测被掩码的文本片段，从而学习生成语义连贯且上下文一致的文本表征。经预训练后，ChatGLM 模型可通过在少量标签数据上进行有监督的微调，适配特定领域及任务需求。

2.4 多模态特征融合模块

在视觉语言多模态中，有效融合不同模态信息间的复杂非线性关系是难点。多模态特征融合模块（图 3）采用 Q-Former 结构连接视觉编码器和语言模型，其由 2 个共享自注意力层参数的 Transformer 子模块构成，分别用于与视觉特征交互、文本编码与解码。

图 3 中，构建 1 组可学习的查询向量（包含 M 个），每个查询向量通过共享子模块的自注意力层参数实现与文本的交互；同时，在交叉注意力层中与视觉特征进行交互，以从图像中提取视觉化的安全隐患信息。Q-Former 输出的视觉特征嵌入与文本高度相关，经 1 层全连接层影射至文本特征的语义空间后，得到融合特征 $\boldsymbol{z} = [z_1, z_2, \cdots, z_M]$（与文本特征的维度一致）。该融合特征与输入文本（含 K 个令牌）的文本特征 $\boldsymbol{t} = [t_1, t_2, \cdots, t_K]$ 拼接后，共同作

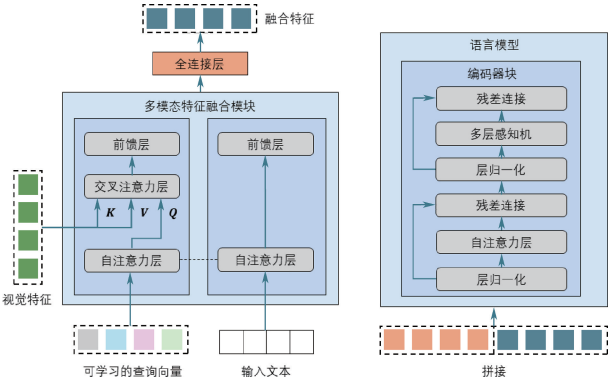

图 3 多模态特征融合模块

Fig. 3 Multi-modal feature fusion module

为语言模型编码器块的输入 $\boldsymbol{h}$：

$$\boldsymbol{h} = [z_1, z_2, \cdots, z_M, t_1, t_2, \cdots, t_K] \tag{5}$$

多模态特征融合模块在保留各模态特性的基础上，通过强化模态间特征的融合，实现图像与文本的深度语义关联，从而有效适配视觉问答等复杂应用场景。

2.5 基于矩阵低秩分解的微调训练

定义数据样本为三元组 $[I, H, S]$，其中，I 为安全隐患图像，H 为安全隐患描述文本，S 为安全隐患解决措施文本。构建适配建筑施工安全隐患场景视觉问答的特定提示模板，过程如下：

1）针对安全隐患描述文本 H 和安全隐患解决措施文本 S，构建包含多个语义相似问题的描述性问题集合 Q_H 和措施性问题集合 Q_S（表 2）。

表 2 问题集合

Table 2 Question sets

序号	Q_H	Q_S
1	图中包含什么安全隐患	应该采取什么措施消除图片中的安全隐患
2	图片展示了哪些潜在的安全隐患	为了保障施工安全，图片中的安全隐患应如何处理
⋮	⋮	⋮
k	通过这张图片你能发现什么安全隐患	图片中的安全隐患可以被什么措施消除

2）从 Q_H 中随机选取问题 $q_H \in Q_H$，与安全隐患图像 I 共同构成模型输入 $[I, q_H]$。微调过程中，以安全隐患描述文本 H 作为监督标签，指导模型学习基于图像和问题指令生成隐患描述 D。

3）从 Q_S 中随机选取问题 $q_S \in Q_S$，将第一阶段输入 $[I, q_H]$ 与生成的隐患描述 D 共同构成新输入 $[I, q_H, D, q_S]$。此阶段以安全隐患解决措施文本 S 作为监督标签，指导模型在给定图像、历史问题、历

史答案的基础上,生成针对性解决措施 J。

CSIQM 模型参数量庞大,若在少样本数据上对整个模型微调,易出现过拟合和灾难性遗忘现象,且计算资源消耗巨大。为此,采用基于矩阵低秩分解的微调方法,将模型中的部分权重矩阵近似为低秩形式,在显著减少训练参数量的同时,保留原始矩阵的主要信息,从而提升模型训练效率和泛化能力。具体操作中,选取 ChatGLM 中的部分自注意力层作为低秩分解对象,利用构建的建筑施工安全隐患图-文对数据集,对模型开展有监督的微调训练。

给定的权重矩阵 $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ 可分解为 2 个较小的矩阵 $\mathbf{U}$ 和 $\boldsymbol{Y}$:

$$\boldsymbol{W} \approx \boldsymbol{U}\boldsymbol{Y}^{\top} \quad \boldsymbol{U} \in \mathbb{R}^{m\times r}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}, \quad r \ll \min(m,n) \tag{6}$$

式中 r 为低秩矩阵的秩,是一个预设的小值。

微调训练目标融合采用 2 种损失函数,基于图像生成文本的语言模型自回归损失和视觉文本语义特征对齐的对比损失。其中,基于图像生成文本的语言模型自回归损失使模型能够结合图像和文本输入生成文本答案,采用交叉熵作为度量方式,最小化模型预测词分布与真实目标词分布间的差异,优化模型生成答案的准确性。自回归损失 L_{A} 可表示为:

$$L_{\mathrm{A}} = -\sum_{t=1}^{R} \ln p(y_t \mid y_{<t}, w, I) \tag{7}$$

式中:R 为由安全隐患图像 I 启发的文本序列长度;$p(y_t \mid y_{<t}, w, I)$ 为在给定先前所有词 $y_{<t}$、文本问题 w 及安全隐患图像 I 后预测的下一个词 y_t 的概率分布。

基于视觉-文本语义特征对齐的对比损失,使得安全隐患图像和模型生成文本在特征空间中对齐。具体而言,对比损失确保视觉特征向量与文本语义特征向量间具有较高相似度,同时降低与其他样本特征的相关性。对比损失 L_{C} 可表示为:

$$L_{\mathrm{C}} = -\ln \frac{\exp(\mathrm{sim}(\boldsymbol{b},\boldsymbol{g})/\tau)}{\sum_{i=1}^{n} \exp(\mathrm{sim}(\boldsymbol{b},\boldsymbol{g}_i)/\tau)} \tag{8}$$

式中:$\boldsymbol{b}$ 为安全隐患图像的视觉特征;$\boldsymbol{g}$ 为模型生成文本的语义特征;$\boldsymbol{g}_i$ 为不同样本的文本特征;τ 为调整对比损失的温度系数;$\mathrm{sim}(x)$ 为余弦相似度函数。

2.6　多轮提示词推理

推理阶段采用多轮提示词方法,通过迭代输入提示引导模型生成更精确、全面的答案,以深化问题理解并提升回答的准确性和完整性。

对于给定的含安全隐患图像及相应文本问题,首先执行第 1 轮推理以完成初步视觉识别与问题解析。此阶段将第 1 轮提示词设为“图中包含什么安全隐患?”,与图像共同输入模型,要求其识别并输出图像中的安全隐患,实现对复杂场景的初步认知。第 2 轮推理中,将提示词调整为“应该采取什么措施消除图片中的安全隐患?”,并与第 1 轮生成的隐患识别结果拼接为新的提示词,再结合图像重新输入模型。通过这种迭代机制,模型可基于已识别的隐患信息,进一步推理并提出针对性解决措施。该方法不仅增强了模型对复杂场景的全局理解能力,更提升了其实用性和可靠性。

3　安全智能问答模型实例验证

3.1　实例设置

采用显存容量为 40 G 的图形处理器有监督的微调训练模型,将学习率设为 exp(-4),迭代轮数设为 5 000 轮,序列输出最大长度设为 384。基线模型选择 ChatGLM-6B 和 VisualGLM-6B[15]。训练过程中冻结视觉编码器和语言模型参数,仅更新多模态特征融合模块及低秩矩阵。

为全面评估生成文本质量,确保其在准确性和语义一致性上达到较高水平,采用以下 3 种评估方法:

1) 自动评估指标。采用基于统计的文本评价方法,包括 BLEU(Bilingual Evaluation Understudy)、ROUGE (Recall-Oriented Understudy for Gisting Evaluation)、METEOR (Metric for Evaluation of Translation with Explicit Ordering)。BLEU 通过计算 n-gram 精确匹配率,并经几何平均值加权,衡量生成文本与参考文本的相似度,是文本生成任务中广泛使用的指标;ROUGE-L 是 ROUGE 的常用变体,基于最长公共子序列 (Longest Common Subsequence,LCS)计算召回率和准确率,重点评估参考文本内容的覆盖度及语义匹配度;METEOR 不仅考虑词的精确匹配,还纳入词形变化、同义词替换及词序等因素。

2) GPT-4 评价。将模型生成结果和真实答案输入到 GPT-4 进行评分,GPT-4 基于其强大的语义理解能力,根据生成文本与真实答案的相似度和质量给出评分(0~100),提供更细致的评估结果。

3) 专家评价。邀请建筑施工安全管理领域具备深厚专业背景和丰富实践经验的专家,从流畅度(文本表达的自然性、连贯性及易理解性)和相关性

(内容与建筑施工安全的关联度)2 个核心维度开展评分,取平均分作为最终得分。该方法可补充自动评估可能遗漏的细节信息和上下文语境,增强评估的专业性与全面性。

3.2 模型对比结果

为全面评估 CSIQM 模型在建筑施工安全智能问答中的性能,选取视觉语言多模态领域内具有广泛认可度的通用视觉语言模型作为对比基准。采用相同的测试集对 LLaVA(Large Language and Vision Assistant)[16]、CogVLM(Cognitive Vision-Language Model)[17]和 CSIQM 开展测试分析,通过对比输出结果,衡量 CSIQM 模型在特定任务上的表现。

图 4 为训练过程中不同迭代轮数对应的困惑度曲线。作为模型性能评估的常用指标,困惑度用于量化概率分布的不确定性,其值越低表明模型对数据的预测精度越高。由图 4 可知:随迭代轮数增加,模型困惑度呈逐步下降趋势,表明模型持续学习建筑施工安全领域知识;当迭代达到一定轮数后,曲线趋于平缓,表明模型性能已趋于稳定。

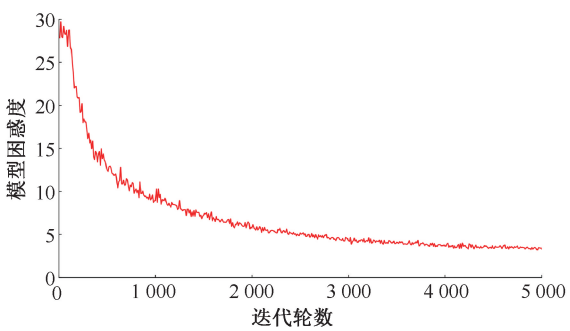

图 4 不同迭代轮数对应的困惑度曲线

Fig. 4 Changes in perplexity at different iteration numbers

CSIQM 模型的自动评估指标结果对比见表 3。由表 3 可知:CSIQM 模型在 BLEU、ROUGE-L 和 METEOR 这 3 个自动评估指标上,均较 LLaVA 和 CogVLM 模型有所提升,表明在建筑施工安全智能问答中,CSIQM 模型生成文本的质量、语法准确性及语义相似度等方面均优于对比模型。

表 3 CSIQM 模型的自动评估指标结果对比

Table 3 Comparison in automatic evaluation metrics of CSIQM model %

模型	BLEU	ROUGE-L	METEOR
LLaVA	0.4	10.0	9.7
CogVLM	4.2	19.5	21.3
CSIQM	24.5	38.1	44.7

CSIQM 模型的 GPT-4 评价和专家评价结果对比结果见表 4。由表 4 可知:CSIQM 在回答的流畅性和内容的相关性 2 个维度同样表现出色。

表 4 CSIQM 模型的 GPT-4 评价和专家评价结果对比

Table 4 Comparison in GPT-4 evaluation and expert evaluation results of CSIQM model

模型	GPT-4 评价	专家评价	
		流畅性	相关性
LLaVA	15	60	20
CogVLM	26	75	40
CSIQM	61	80	70

建筑施工安全智能问答的示例结果如图 5 所示。由图 5 可知:针对同一建筑施工安全隐患图像,不同模型生成的隐患解决措施存在显著差异,尤其在具体操作细节和技术规范性层面呈现明显分化。LLaVA 和 CogVLM 偏向输出一般性预防建议,而 CSIQM 则侧重于提供基于行业规范的精确技术指导。LLaVA 模型关注工作区域的整体整洁和常规安全管理;CogVLM 模型虽关注模板支架区域的杂物清理、杆件加固和防护设施增设,但缺乏可操作性细节描述;相比之下,CSIQM 模型提供的操作指南更为详尽,涵盖扫地杆设置的具体位置、连接可靠性要求等。将建筑施工安全智能问答模型作为辅助决策工具,可支持安全管理人员在现场巡检过程中,通过上传隐患图像和初步文本描述,即时获取针对具体隐患的高可行性、强可操作性解决措施,为整改方案制定提供数据支撑的决策依据。这一应用模式显著增强了现场管理的即时性和针对性,有效规避了传统人工决策中存在的主观性和不确定性。

3.3 消融试验

通过对比基线、多轮推理、低秩分解微调及其组合的效果,量化各部分对最终结果的贡献。模型架构各模块的消融试验结果见表 5,由表 5 可知:低秩分解微调在提升文本生成质量上表现出显著优势;多轮推理单独使用时效果有限,但与低秩分解微调结合可产生协同效应。相较于基线模型,单独应用多轮推理机制后,各项评价指标均有提升;引入矩阵低秩分解微调后,指标显著提高。其中,BLEU 从 1.0% 大幅增至 17.2%,ROUGE-L 从 14.0% 增至 29.1%,METEOR 从 15.3% 增至 32.7%,GPT-4 评分从 27 增至 42。这一结果表明:低秩分解微调对改善文本生成质量具有明确有效性;而多轮推理虽单独作用有限,但与低秩分解微调结合时,所有评价指标均达到最优水平。因此,低秩分解微调和多轮推理的协同作用是模型实现最佳表现的关键。

问题：应该采取什么措施消除图片中的安全隐患？

回答：为了消除图像中的安全危险，应采取几项措施。首先，在施工现场周围安装护栏或警示屏障，以防止未经许可的进入，特别是在危险区或正在进行的工作区。其次，确保所有工人都戴上适当的个人保护设备，如头盔、安全眼镜和适当的鞋子，以尽量减少受伤的风险。此外，工人应接受适当的安全培训，并应遵守已制定的安全规程　LLaVA

回答：1.清理现场：立即清理现场，将所有的垃圾和杂物清除干净，特别是地上的白色大块物体，以防止绊倒或受伤的风险。2.检查脚手架：对脚手架的各个部分进行检查，确保其牢固地固定在地面上，并且没有松动的部件。任何可见的损坏或磨损都应该得到修复。3. 在脚手架的顶部安装网状物，以防止工人掉落。这种网状物应该覆盖整个脚手架的顶部表面　CogVLM

回答：应严格遵循建筑施工模板安全技术规范进行整改，在模板支架的四周设置水平扫地杆，高度距离地面200 mm且与立杆垂直，并确保其牢固连接。扫地杆需连续布置于框架内外，以增强整体稳定性和承载能力。同时，应根据脚手架设计要求，对每一根立杆进行严格的规范设置，如使用符合标准的扣件式钢管作为支撑，以确保结构稳固可靠。整改完成后10日之内组织人员进行验收　CSIQM

图 5　建筑施工安全智能问答的示例结果

Fig. 5　Example results of visual question answering for safety in construction

表 5　模型架构各模块的消融试验结果

Table 5　Ablation study results of different modules in model architecture

方法	BLEU/%	ROUGE-L/%	METEOR/%	GPT-4
基线	1. 0	14. 0	15. 3	27
多轮推理	2. 6	17. 2	19. 3	30
低秩分解微调	17. 2	29. 1	32. 7	42
低秩分解微调+多轮推理	24. 5	38. 1	44. 7	61

矩阵低秩分解作为一种高效微调方法，其参数 r（低秩矩阵的秩）控制着微调参数量。不同矩阵低秩分解参数 r 的结果对比如图 6 所示，由图 6 可知：随微调参数量增加，2 种评估指标逐渐上升；但在 $r=32$ 后，指标趋于平缓且略有下降，表明存在过拟合风险。因此，选择合适的 r 值是在提升模型性能与避免过拟合间取得平衡的关键。

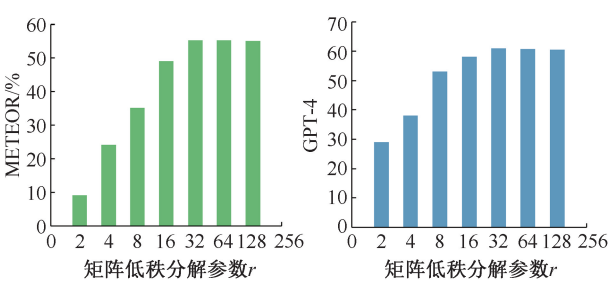

图 6　不同矩阵低秩分解参数 r 的结果对比

Fig. 6　Comparison of results with different low-rank decomposition r values

4　结　论

1）通过微调模型架构、训练策略和高质量专业数据的微调训练，建筑施工安全智能问答模型在 BLEU、ROUGE-L、METEOR 指标上分别达到 24. 5%、38. 1%、44. 7%，均优于 LLaVA 和 CogVLM，表明该模型生成文本不仅质量较高，且在语法准确度及语义相似度方面均表现优异。进一步结合 GPT-4 评价与专家评价结果可知：该模型生成的答案兼具良好流畅性与较强相关性，能够有效结合图像信息判别安全隐患并输出较详细的解决措施。

2）矩阵低秩分解微调和多轮推理机制的联合引入对模型性能提升贡献显著，且在微调训练过程中，当低秩矩阵的秩 $r=32$ 时，可有效避免过拟合问题，保障模型泛化能力。

3）建筑施工现场图像数据仅依赖人工巡检拍摄的静态照片，未纳入对动态环境的考量；模型未整合施工现场视频监控、无人机航拍、传感器感知等多源数据，导致多模态模型在复杂施工环境深层理解中的潜力尚未充分发掘，相关方向有待进一步研究。

参 考 文 献

[1]　中华人民共和国住房和城乡建设部. 住房和城乡建设部办公厅关于 2020 年房屋市政工程生产安全事故情况的通报[EB/OL]. (2022-10-27). https://www. mohurd. gov. cn/gongkai/zc/wjk/art/2022/art_17339_768565. html.

[2] 郭纯兵，阎卫东，李宇鹏，等. 基于系统动力学的智慧工地安全脆弱性研究[J]. 中国安全生产科学技术，2024，20(8)：42-50.
GUO Chunbing，YAN Weidong，LI Yupeng，et al. Study on safety vulnerability of smart construction sites based on system dynamics[J]. Journal of Safety Science and Technology，2024，20(8)：42-50.
[3] 中华人民共和国住房和城乡建设部，中华人民共和国国家发展和改革委员会，中华人民共和国科学技术部，等. 住房和城乡建设部等部门关于推动智能建造与建筑工业化协同发展的指导意见[EB/OL]. (2020-07-28). https://www.gov.cn/zhengce/zhengceku/2020-07/28/content_5530762.htm.
[4] 赵江平，刘星星，张想卓. 基于改进 YOLOv5s 的外脚手架隐患图像识别技术[J]. 中国安全科学学报，2023，33(12)：60-66.
ZHAO Jiangping，LIU Xingxing，ZHANG Xiangzhuo. Research on image recognition technology for external scaffold hidden danger based on improved YOLOv5s[J]. China Safety Science Journal，2023，33(12)：60-66.
[5] 郑楚伟，林辉. 基于 Swin Transformer 的 YOLOv5 安全帽佩戴检测方法[J]. 计算机测量与控制，2023，31(3)：15-21.
ZHENG Chuwei，LIN Hui. YOLOv5 helmet wearing detection method based on Swin Transformer[J]. Computer Measurement and Control，2023，31(3)：15-21.
[6] 石雪洁. 基于机器视觉技术的高层建筑施工现场危险区域识别方法[J]. 佳木斯大学学报:自然科学版，2023，41(3)：104-107.
SHI Xuejie. Research on the identification method of hazardous areas in high rise building construction sites based on machine vision technology[J]. Journal of Jiamusi University: Natural Science Edition，2023，41(3)：104-107.
[7] 王仁超，张毅伟，毛三军. 水电工程施工安全隐患文本智能分类与知识挖掘[J]. 水力发电学报，2022，41(11)：96-106.
WANG Renchao，ZHANG Yiwei，MAO Sanjun. Intelligent text classification and knowledge mining of hidden safety hazards in hydropower engineering construction[J]. Journal of Hydroelectric Engineering，2022，41(11)：96-106.
[8] 周倡弘，王聚全，杜浸，等. 基于大语言模型的消防应急决策支持技术研究[J]. 电信快报，2024(5)：19-27.
ZHOU Changhong，WANG Juquan，DU Wen，et al. Research on fire emergency decision support technology based on large language model[J]. Telecommunications Information，2024(5)：19-27.
[9] 洪亮，郭瑶，刘兴丽，等. 基于 RAG 的煤矿安全智能问答模型[J]. 黑龙江科技大学学报，2024，34(3)：487-492.
HONG Liang，GUO Yao，LIU Xingli，et al. Intelligent Q & A model of coal mine safety based on RAG[J]. Journal of Heilongjiang University of Science and Technology，2024，34(3)：487-492.
[10] 张飞飞，张建庆，屈思佳，等. 跨模态视觉问答与推理研究进展[J]. 数据采集与处理，2023，38(1)：1-20.
ZHANG Feifei，ZHANG Jianqing，QU Sijia，et al. Recent advances in visual question answering and reasoning[J]. Journal of Data Acquisition and Processing，2023，38(1)：1-20.
[11] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]. International Conference on Machine Learning，2021：8 748-8 763.
[12] LI Junnan，LI Dongxu，SAVARESE S，et al. BLIP-2：bootstrapping language-image pre-training with frozen image encoders and large language models[C]. International Conference on Machine Learning，2023：19 730-19 742.
[13] 廖宁，曹敏，严骏驰. 视觉提示学习综述[J]. 计算机学报，2024，47(4)：790-820.
LIAO Ning，CAO Min，YAN Junchi. Visual prompt learning：a survey[J]. Chinese Journal of Computers，2024，47(4)：790-820.
[14] 熊若鑫，宋元斌，王宇轩，等. 基于 CNN 的 3D 姿势估计在建筑工人行为分析中的应用[J]. 中国安全科学学报，2019，29(7)：64-69.
XIONG Ruoxin，SONG Yuanbin，WANG Yuxuan，et al. Application of convolutional neural network-based 3D posture estimation in behavioral analysis of construction workers[J]. China Safety Science Journal，2019，29(7)：64-69.
[15] DING Ming，YANG Zhuoyi，HONG Wenyi，et al. Cogview：mastering text-to-image generation via transformers[J]. Advances in Neural Information Processing Systems，2021，34：19 822-19 835.
[16] LIU Haotian，LI Chunyuan，WU Qingyang，et al. Visual instruction tuning[J]. Advances in Neural Information Processing Systems，2023，36：34 892-34 916.
[17] WANG Weihan，LYU Qingsong，YU Wenmeng，et al. CogVLM：visual expert for pretrained language models[J]. Advances in Neural Information Processing Systems，2024，37：121 475-121 499.

作者简介： 王 喆 (1980—)，男，湖北武汉人，博士，副教授，博士生导师，主要从事应急决策、人工智能和施工安全等方面的研究。E-mail：wangzhe@whut.edu.cn。