E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)04-1216-11

论文引用格式:Li J W, Yang C Y, Zhang Y C, Sun W L, Meng L and Meng X X. 2026. Large model retrieval enhancement framework for construction site risk identification. Journal of Image and Graphics, 31(4):1216-1226(李嘉威, 杨成业, 张尧臣, 孙玮琳, 孟雷, 孟祥旭. 2026. 面向工地风险隐患识别的大模型检索增强框架. 中国图象图形学报, 31(4):1216-1226)[DOI:10.11834/jig.250333]

面向工地风险隐患识别的大模型检索增强框架

李嘉威[1],杨成业[1],张尧臣[2],孙玮琳[1],孟雷[1*],孟祥旭[1]

1. 山东大学软件学院,济南　250101; 2. 浪潮软件科技有限公司,济南　250100

摘　要:**目的**　工地风险隐患识别旨在通过自动化技术提升施工现场安全管理水平。现有基于大语言模型的研究分为两类:一是利用图文匹配进行协同推理,但对复杂隐患特征捕捉不足;二是通过专业数据集进行指令微调或多轮对话引导,但存在训练成本高、泛化能力差的问题。**方法**　提出一种基于相似案例检索增强的隐患识别方法,通过提示微调技术动态融合外部知识库与检索案例上下文,设计有效的图像检索策略,解决了大模型因领域知识缺失与特征关联弱化导致的误判问题。该方法包括检索库、图像相似度检索和大模型检索增强3个模块,实现了无训练优化下的高效识别。**结果**　实验基于真实施工数据,在多种大模型上测试该方法并进行系统性评估,其中GLM-4V(general language modelvision-4)的识别正确率提升至50%,较基线方法提高35.49%,且在多数隐患类别上均表现出性能增益;进一步通过消融实验验证关键模块的有效性,引入学习感知图像块相似度(learned perceptual image patch similarity,LPIPS)算法与图像检索模块中的对比语言—图像预训练(contrastive language-image pretraining,CLIP)算法进行对比。结果表明,所构建的图像计算检索策略具备优越性。**结论**　本文提出的基于相似案例检索增强方法显著提升了大模型对工地风险隐患识别的准确率与上下文理解能力,在多类别隐患场景下均表现出良好的泛化性能,为施工现场安全风险智能检测提供了新的理论支撑与技术路径。

关键词:大语言模型(LLM);风险隐患检测;多模态识别;检索增强生成;提示微调

Large model retrieval enhancement framework for construction site risk identification

Li Jiawei[1], Yang Chengye[1], Zhang Yaochen[2], Sun Weilin[1], Meng Lei[1*], Meng Xiangxu[1]

1. School of Software, Shandong University, Jinan　250101, China; 2. Inspur Software Technology Co., Ltd., Jinan　250100, China

Abstract: **Objective** The primary objective of construction site hazard identification is to elevate the safety management standards within operational construction environments substantially by leveraging advanced automation technologies. In recent years, the pervasive adoption of large language models (LLMs) has opened new avenues for research in this critical domain. The overarching goal is to mitigate human error, enhance detection efficiency, and proactively prevent accidents through intelligent systems capable of interpreting complex visual and textual data from construction sites. A meticulous

收稿日期:2025-07-25;**修回日期**:2025-10-02;**预印本日期**:2025-10-09

*通信作者:孟雷　lmeng@sdu.edu.cn

基金项目:山东省重点研发计划资助(2024TSGC0667);教育部中国高校产学研创新基金项目(2024LC030);山东省基础软件重点实验室项目(11150004040955);数字媒体技术教育部工程研究中心项目

Supported by: Shandong Provincial Key R&D Program, China(2024TSGC0667); China Higher Education Industry-Academia-Research Innovation Fund(2024LC030); Shandong Key Laboratory of Foundational Software(11150004040955); Engineering Research Center of Digital Media Technology, Ministry of Education

analysis of the current research landscape, based on LLMs, reveals that existing methodologies can be broadly classified into two distinct categories, each presenting its respective set of advantages and limitations. The first approach leverages the capability of image-text matching to perform collaborative reasoning by integrating visual input with textual hazard descriptions. The second method involves constructing domain-specific datasets to fine-tune large models through instruction tuning or guide them via multiturn dialogues. The former enhances the alignment between images and semantic representations through multimodal fusion yet exhibits limitations in capturing complex hazard characteristics. The latter strengthens the model's analytical depth with domain knowledge infusion but suffers from high training costs and poor generalizability. **Method** This study addresses these limitations by proposing a risk-detection, retrieval-augmented generation hazard identification method that dynamically integrates external knowledge bases with retrieved case contexts through prompt tuning, thereby resolving misjudgments caused by LLMs' lack of domain knowledge and weakened feature associations. The proposed architecture is systematically structured into three cohesive and interdependent core modules, each serving a distinct and vital function: The first module is the retrieval database module. It serves as the external knowledge repository, populated with a comprehensive collection of historical construction hazard cases. Each entry within this database is a rich, multimodal data object comprising visual data (images) and its corresponding textual annotation, which includes a detailed description of the hazard type, its location, and contextual information. The integrity, diversity, and relevance of this database are paramount because it forms the foundational knowledge source for the entire system. The second module is the image similarity retrieval module. This component is responsible for the efficient and accurate retrieval of the most relevant cases from the database, given a new query image from a construction site. At the heart of this module is a powerful vision-language model, specifically the contrastive language-image pretraining (CLIP) model. CLIP excels at mapping images and text into a shared, high-dimensional semantic embedding space. When a new query image is processed, it is encoded into an embedding vector. Then, this vector is compared against the precomputed embeddings of all images in the retrieval database using a similarity metric. The top-K most semantically similar cases are retrieved, thereby ensuring that the subsequent reasoning steps are informed by visually and contextually analogous examples. The third module is the LLM retrieval-augmentation reasoning module, which is the central reasoning engine. The retrieved similar cases (both their images and text) are formatted into a structured prompt, thereby providing the LLM (e. g. , GLM-4V) with a critical few-shot learning context. This prompt, which also includes the query image, guides the LLM to perform in-context learning. This entire framework operates in a training-free manner for the LLM itself. Thus, no additional fine-tuning is required, thereby guaranteeing efficiency, reducing computational overhead, and enhancing scalability and ease of deployment. **Result** A rigorous empirical evaluation was conducted to validate the efficacy of the proposed framework. Experiments were systematically performed using authentic, real-world construction site data, encompassing various hazard scenarios and environmental conditions. The framework was subjected to comprehensive testing and systematic assessment across multiple state-of-the-art large language models to ensure the robustness and general applicability of the approach. The results were highly promising and demonstrated a substantial quantitative improvement. When integrated with the GLM-4V model, the retrieval-augmentation framework achieved a recognition accuracy of 50%. This value represents a significant and remarkable improvement of 35. 49% over the baseline performance of the vanilla GLM-4V model without retrieval augmentation. Beyond this aggregate metric, a detailed category-wise analysis revealed consistent performance gains across a majority of individual hazard types. This finding indicates that the method enhances the model's capability universally rather than being biased toward a specific hazard category. Furthermore, ablation studies were designed, and the LPIPS algorithm was introduced to compare it with the CLIP algorithm used in the image similarity retrieval module. Results demonstrated the clear superiority of the CLIP-based semantic retrieval strategy over the LPIPS-based perceptual strategy in this specific task. **Conclusion** As proposed in this paper, the retrieval-augmented method based on similar cases delivers a significant breakthrough in automated construction site safety monitoring. It tangibly and markedly enhances the key performance indicators of LLMs——namely, accuracy and contextual understanding ability——in the complex task of hazard identification. The framework demonstrates robust generalization performance across a wide spectrum of multicategory hazard scenarios, thereby effectively addressing the core limitations of previous approaches related to training cost and adaptability. The training-free nature of the approach makes it particularly attractive for real-world deploy-

ment, thereby offering a scalable and sustainable solution for enhancing on-site safety protocols.

Key words: large language model (LLM); risk detection; multimodal recognition; retrieval enhancement generation; prompt fine-tuning

0 引 言

工地风险隐患识别(Ballal等,2024)旨在通过自动化手段替代传统人工巡检,提升施工安全。当前主流方法基于计算机视觉技术,如目标检测与图像分类(Soumya等,2024),但存在泛化能力不足、识别精度低等问题(Fan等,2024)。多模态大模型的发展为隐患识别提供了新思路,然而,由于缺乏工地场景专业知识,模型在复杂环境中易出现误判(郭园方等,2025)。因此,增强模型对隐患场景的理解能力成为核心挑战。

现有研究主要分为两类:一是利用图文匹配(Wang等,2024c)能力,结合图像与隐患描述进行协同推理;二是通过构建专业数据集(Yu等,2024),对大模型进行指令微调或多轮对话引导。前者通过多模态对齐提升图像与语义匹配,但对复杂隐患特征把握有限;后者通过领域知识增强模型分析深度,但存在训练成本高、通用性差的问题。因此,现有方法在领域知识适配性与上下文关联性方面仍存在亟待突破的技术瓶颈。

为在避免高成本微调的同时增强大模型对复杂隐患识别的准确性与领域适应性,引入外部知识源并实施检索增强已成为一种值得探索的解决路径(Arslan等,2024)。该方法能够有效结合已有知识库与实时检索机制,在不显著增加训练负担的前提下有效提升模型的上下文感知能力。

本文提出一种基于相似案例检索增强的风险隐患检测识别方法,通过提示微调(Kim等,2025)技术动态融合外部知识库(Zhu等,2024)与检索案例上下文,设计有效的图像检索策略,缓解现有多模态大模型因领域知识缺失与特征关联弱化导致的误判问题,如图1所示。方法包括3个核心模块:1)检索库模块,构建结构化隐患案例数据库;2)图像相似度检索模块,基于CLIP(contrastive language-image pre-training)(Ghosh等,2024)模型定位最相关案例;3)大模型检索增强模块,通过提示微调生成准确隐患类别与描述。

为了验证模型的有效性,本文基于真实施工工地采集数据构建了测试集,选取GLM-4V(general language model vision-4)、GPT-4o(chat generative pre-trained Transformer 4 omni)(Lewandowski等,2024)以及DeepSeek-VL2这3种主流多模态大模型进行对比评估。实验从识别准确率、误判率和上下文理解能力等方面展开验证。实验结果表明,该方法在多种场景下均显著提升了模型对复杂隐患的识别能力,尤其在难以直接判断的隐患图像中表现出更高的理解深度与判别稳定性。

具体而言,本文的主要贡献如下:1)提出了一种基于相似案例检索增强的风险隐患识别框架,创新性地融合大模型提示学习与实例检索机制,为提升大模型在隐患识别任务中的准确性提供了新路径;2)在算法设计上构建了即插即用的检索增强模块,通过提示微调策略实现大模型的无训练优化,并使其能够在无需额外训练的前提下,快速适应风险隐患识别任务;3)实验部分系统评估了不同大模型在实际场景的识别表现,明确了检索增强在提升模型泛化能力与解释能力方面的优势,为后续多模态大模型工业安全领域的应用提供了理论支持和实践参考。

1 相关工作

1.1 传统隐患识别方法

工地风险隐患识别任务旨在通过对施工现场潜在危险与隐患的排查、识别与评估,及时采取防范措施,确保工地安全并预防事故发生。传统方法主要依赖安全管理人员的经验判断,通过定期巡查发现隐患并进行整改,然而该方法存在人工疏漏、重复性高以及无法实时监控等局限性。

近年来,基于物联网技术(Wang等,2022)与计算机视觉技术(Hou等,2023)的隐患识别方法成为研究热点。基于物联网技术的方法通过部署传感器网络,实时监测环境、设备状态及人员行为等数据,结合数据分析技术进行隐患识别。该方法能够动态捕捉施工现场信息,迅速发现潜在风险,但存在成本高、设备需求量大及对数据管理与处理能力要求较

高等问题。基于计算机视觉技术的方法则依托图像或视频数据，利用图像处理、目标检测、语义分割及行为识别等技术自动识别安全隐患。其核心在于特征提取，包括传统方法（如尺度不变特征转换（scale invariant feature transform，SIFT）（Arooj 等，2024）、方向梯度直方图（histogram of oriented gradient，HOG）（Zhang 等，2024））与深度学习方法（如卷积神经网络（convolutional neural network，CNN）（侯志强 等，2025）、Transformer（Yao 等，2024））。尽管该方法提升了隐患识别的自动化程度，但其性能受限于图像质量与算法精度，且对计算资源需求较高。

1.2 基于大语言模型的隐患识别方法

大语言模型（large language model，LLM）在工业隐患识别中的应用逐渐受到关注。LLM 凭借其强大的语义理解与上下文推理能力（Cui 等，2024），能够从文本、图像和传感器数据等多模态信息中提取关键特征，实现对潜在隐患的精准识别。然而，数据稀缺、领域适应性不足以及多模态融合复杂性等问题仍制约其广泛应用。现有研究通过少样本学习（Zeng 和 Xiao，2024）、多模态对齐（Wang 等，2024a）等技术部分缓解了这些挑战，但仍需要进一步进行优化以提升模型性能。

多模态对齐通过将视觉、文本和传感器等多源数据映射到统一语义空间，实现跨模态特征的协同优化，提升风险识别的准确性与鲁棒性。常用方法包括对比学习（Wei 等，2024）、跨模态注意力机制（Luo 等，2024）和语义增强表示。例如，Myriad 结合视觉专家模型与对比学习优化工业场景中的风险识别，大语言模型增强 CLIP 框架（large language model to CLIP，LLM2CLIP）通过语义增强表示与对比学习融合视觉和语言特征。尽管多模态对齐提升了性能，但其计算复杂度较高，对硬件资源需求较大。少样本学习利用预训练模型的泛化能力，在有限标注数据下快速适应新任务，减少对大规模标注数据的依赖。常见方法包括元学习、提示学习与迁移学习。AnomalyGPT（Gu 等，2024）与 FedITD（Wang 等，2024d）通过提示学习与迁移学习，在少样本条件下实现精准的工业异常检测。然而，少样本学习可能导致模型在新场景中的泛化能力下降，需针对性调整以提高性能。

如图 1 所示，本文提出了一种基于相似案例检索的增强大语言模型性能的方法，通过构建检索库并匹配与目标相似度最高的案例作为参考，在计算资源消耗较低的情况下实现了隐患检测任务中的少样本学习。该方法的核心在于利用检索机制从小规模案例库中筛选出与当前任务最相关的样本，作为

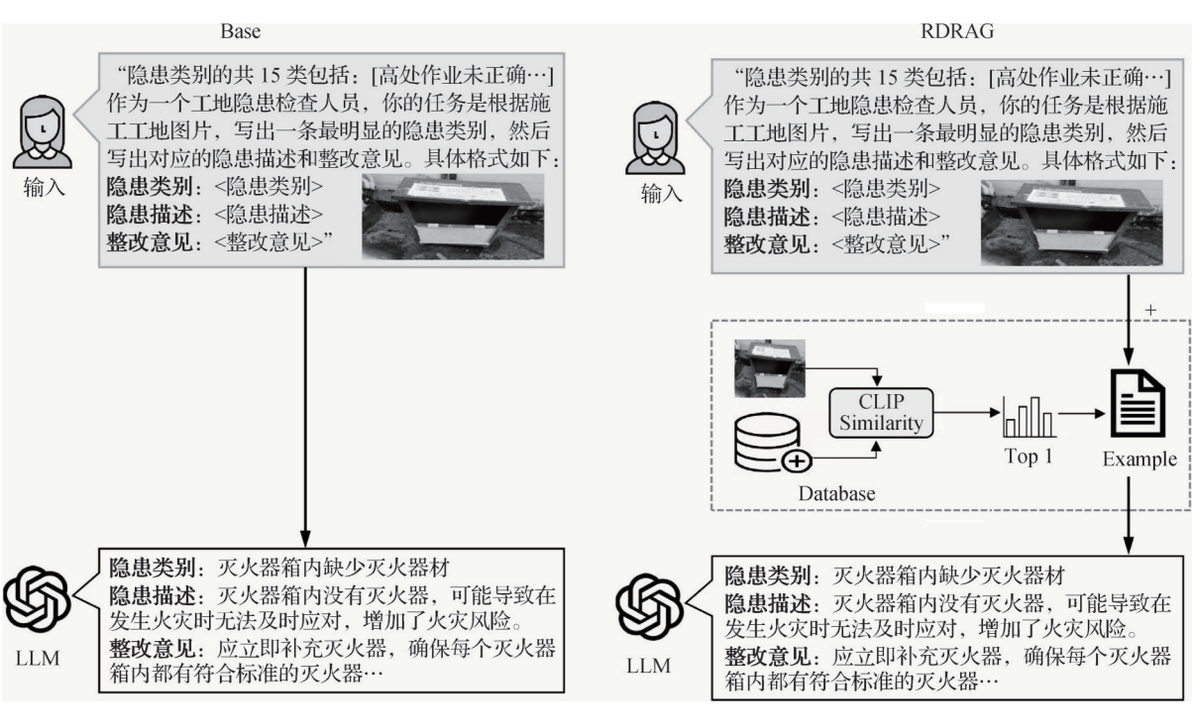

图 1 基于相似案例的检索增强方法

Fig. 1 Retrieval-augmented method based on similar cases

上下文信息输入模型，从而提升模型在低资源场景下的泛化能力和任务适应性。通过引入案例检索增强策略，使模型能够有效利用外部知识库中的结构化信息，减少对大规模标注数据的依赖，同时显著降低计算复杂度，为资源受限环境下的高效部署提供了可行方案。

2 问题描述

在一个多模态隐患识别系统中，数据集 $\boldsymbol{D}=(\boldsymbol{I}_1,C_1,L_1),(\boldsymbol{I}_2,C_2,L_2),\cdots,(\boldsymbol{I}_N,C_N,L_N)$ 包含了 N 对多模态数据，其中，$\boldsymbol{I}_i$ 表示工地施工图像，C_i 表示隐患描述的文本信息，L_i 表示该隐患类别。由于数据集中包含了图像和文本信息，分别使用 $\boldsymbol{I}_i^S$ 和 C_i^S 表示图像和隐患描述的文本部分。

与传统的需要大规模训练数据的多模态模型不同，本文提出了基于检索增强生成的无训练优化框架，依赖于从数据集中检索出与当前图像 $\boldsymbol{I}_i$ 最相似的历史案例，然后将这些检索到的案例的隐患描述与当前图像共同作为 prompt 输入给大语言模型，生成最终的隐患类别 L_i' 和描述 C_i'。这个过程不涉及任何训练步骤，因此大语言模型在此过程中并不需要进行参数优化。

具体而言，对于每个输入样本 $(\boldsymbol{I}_i,C_i,L_i)$，先从数据集中检索出 K 个与当前图像 $\boldsymbol{I}_i$ 最相似的历史样本 $\{(\boldsymbol{I}_j,C_j,L_j)\,|\,j\in 1,\cdots,K\}$，然后将这些检索到的隐患描述片段 $\{C_j\,|\,j\in 1,\cdots,K\}$ 与当前图像 $\boldsymbol{I}_i$ 一同输入传递给大语言模型，生成模型依据输入生成当前图像的隐患类别 L_i' 和隐患描述 C_i'。

用 L_i' 和 C_i' 与实际隐患类别 L_i 和描述 C_i 之间的差异评估结果生成质量。具体而言，模型的目标是通过检索得到最相似的案例，并通过这些相似案例辅助生成隐患类别和描述，从而提高识别的准确性和实用性。可以表示为

$$L_i',C_i'=f(\boldsymbol{I}_i,A\tilde{C}_j\,|\,j\in 1,\cdots,K) \tag{1}$$

式中，A 表示聚合操作，将检索到的 C_j 合并为一个统一的上下文向量，$\tilde{C}_j$ 表示从相似样本中检索到的隐患描述片段。最终，生成的隐患类别 L_i' 和描述将 C_i' 与实际类别 L_i 和描述 C_i 进行比较，从而评估该检索增强生成过程的效果，确保生成的结果在每个实例中尽可能接近真实标注。

3 方 法

本文提出一种基于检索增强生成的无训练优化框架（risk-detection retrieval-augmented generation，RDRAG），其核心结构如图2所示。该框架旨在应对现有多模态方法在隐患识别中存在的领域知识缺失与特征关联弱化问题，同时避免高成本微调与泛化性弱的局限。RDRAG 不依赖额外训练，而是通过引入外部知识库与相似案例检索机制动态构建提示信息，从而增强模型对复杂隐患的判别能力与语义关联性。该方法既保持了模型原有的推理泛化能力，又显著提升了在专业场景中输出的准确性与可靠性。RDRAG 包含3个主要模块，分别是提示词（prompt）设计（Cao等，2025）、大模型检索增强和相似案例检索。通过以上3个模块紧密结合，RDRAG 框架能够在无需大规模训练的情况下，实现对多模态数据的有效理解与生成，极大提升了隐患识别任务的准确性与可靠性。

3.1 提示词（prompt）设计

在 RDRAG 框架中的 prompt 设计部分，本文致力于通过精心设计不同的提示格式，控制大模型的输出结果。通过向提示中添加不同类型的信息，能够引导模型更准确地生成与工地隐患相关的类别、描述和整改意见。为了实现这一目标，提出4种提示设计方式，并通过实验定量数据确立最适合任务需求的设计。4种提示设计方式如下：

1）基础指令（type1）。不添加任何信息，只提出要求。只提供了对模型的基本任务要求："作为一个工地隐患检查人员，你的任务是根据施工工地图像，写出一条最明显的隐患类别，然后写出对应的隐患描述和整改意见"。

2）类别引导（type2）。添加类别信息。在这种设计中，将隐患类别明确列出，要求模型从中选择最相关的类别。隐患类别共15类，包括：高处作业未正确使用安全带、基坑支护措施不到位、灭火器未按规定要求放置、配电箱未及时锁闭、起重吊装设备钢丝绳磨损/断丝严重/搭接长度不足、汽车吊/随车吊/泵车支腿未全部伸出、未垫枕木进行作业、设备安全防护设施/装置缺失或失效、未按规定穿戴反光安全服、未按规定配置灭火器/消防设施等、未按规定设置接地线或接地不良、现场防护栏等安全防护设施

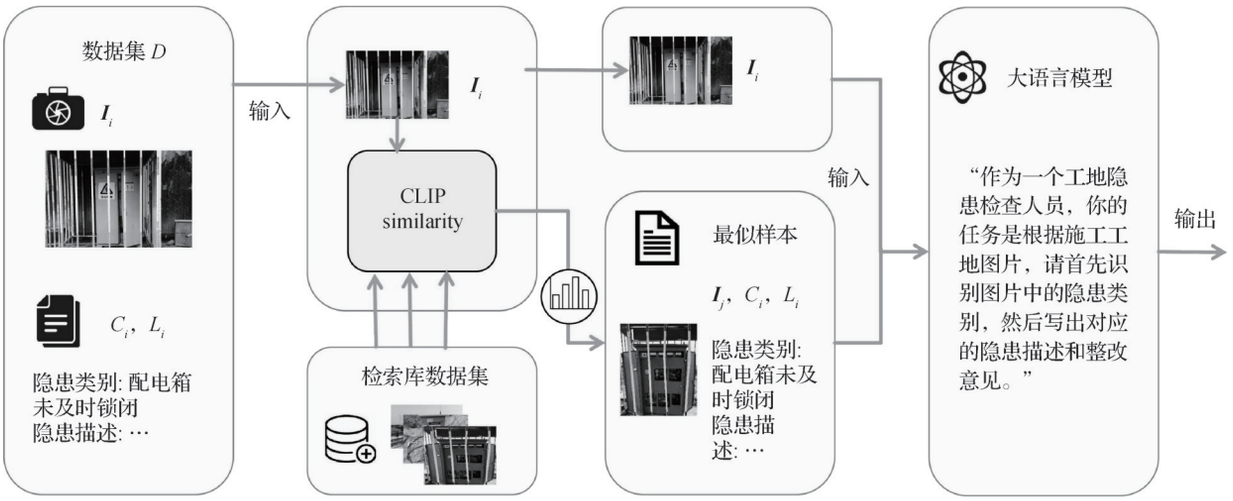

图2　基于检索增强生成的无训练优化框架

Fig. 2　Training-free optimization framework for retrieval-augmented generation

缺失/破损或设置不规范。

3)格式规范(type3)。添加输出格式信息。这种设计进一步优化了输出结构,提供了明确的格式要求。具体格式如下:隐患类别:<隐患类别>;隐患描述:<隐患描述>;整改意见:<整改意见>。通过这种格式化设计,能确保生成的隐患类别、描述和整改意见按标准格式呈现,提高了输出的可读性和一致性。

4)复合增强(type4)。同时添加类别和格式信息。隐患类别共15类(同type2),格式信息包括隐患类别、隐患描述和整改意见(同type3)。这种设计结合了类别信息和格式要求,既能限制模型的输出类别,又能确保生成结果的格式规范,但同时也需要实验验证过于复杂的信息是否会影响大模型的输出结果。

3.2　相似案例检索

相似案例检索模块的核心目标是通过CLIP模型从数据集中检索出与当前图像$\boldsymbol{I}_i$最相似的历史案例。CLIP是一个将图像与文本映射到共同嵌入空间的模型,通过计算图像和文本之间的相似度实现跨模态匹配。以下是该模块的详细算法描述。

首先,通过CLIP模型对输入图像$\boldsymbol{I}_i$和历史图像$\boldsymbol{I}_j$进行特征提取,CLIP模型会将每幅图像映射到一个嵌入空间$f(\boldsymbol{I}_i)$和$f(\boldsymbol{I}_j)$,并将图像和对应的文本描述映射到同一共享空间,具体为

$$\begin{cases} f(\boldsymbol{I}_i) = CLIP(\boldsymbol{I}_i) \\ f(\boldsymbol{I}_j) = CLIP(\boldsymbol{I}_j) \end{cases} \tag{2}$$

这里,$f(\boldsymbol{I}_i)$和$f(\boldsymbol{I}_j)$为图像$\boldsymbol{I}_i$和$\boldsymbol{I}_j$在CLIP嵌入空间中的向量表示。接下来,通过计算图像特征向量之间的余弦相似度衡量当前图像$\boldsymbol{I}_i$和历史图像$\boldsymbol{I}_j$之间的相似性。具体为

$$Sim(\boldsymbol{I}_i, \boldsymbol{I}_j) = \frac{f(\boldsymbol{I}_i) \cdot f(\boldsymbol{I}_j)}{\| f(\boldsymbol{I}_i) \| \cdot \| f(\boldsymbol{I}_j) \|} \tag{3}$$

式中,$Sim(\boldsymbol{I}_i, \boldsymbol{I}_j)$表示当前图像$\boldsymbol{I}_i$与历史图像$\boldsymbol{I}_j$之间的余弦相似度,$f(\boldsymbol{I}_i)$和$f(\boldsymbol{I}_j)$分别是图像的特征向量,$\|\cdot\|$表示向量的L2范数。

通过计算每个历史图像相似度后,对所有历史样本进行排序,选择与当前图像最相似的K个历史案例$(\boldsymbol{I}_j, C_j, L_j) \big| j \in \{1, 2, \cdots, K\}$作为候选样本。具体为

$$\{(\boldsymbol{I}_j, C_j, L_j) \big| j \in \{1, 2, \cdots, K\}\} = Top\text{-}K(Sim(\boldsymbol{I}_i, \boldsymbol{I}_j)) \tag{4}$$

式中,$Top\text{-}K$操作表示从所有历史样本中选择前K个相似度最高的样本。最后,返回检索到的K个历史案例,包含图像$\boldsymbol{I}_j$、隐患描述C_j和隐患类别L_j,供下游模块使用。

3.3　大模型检索增强算法

大模型检索增强模块的目标是将从相似案例检索模块中获得的K个隐患描述$C_j, j \in 1, 2, \cdots, K$和当前图像$\boldsymbol{I}_i$一同输入到多模态大语言模型中,以生成最终的隐患类别L_i'和隐患描述C_i'。

首先,将当前图像$\boldsymbol{I}_i$和从相似案例检索到的K个隐患描述$C_j, j \in 1, 2, \cdots, K$组成一个输入提示。该提示结合了图像和与之相关的历史隐患描述,以帮助多模态大语言模型生成准确的输出,即

$$Prompt_i = Concat(\boldsymbol{I}_i, \{C_j \big| j \in \{1, 2, \cdots, K\}\}) \tag{5}$$

式中,$Concat$表示将图像和隐患描述拼接成一个完

整的输入。接着将提示$Prompt_i$输入多模态大语言模型(例如GPT或其他类似语言模型)中,模型根据输入图像和文本描述生成当前图像的隐患类别L_i'和隐患描述C_i'。具体为

$$L_i', C_i' = LM(Prompt_i) \tag{6}$$

式中,LM表示多模态大语言模型,L_i'和C_i'分别表示模型生成的隐患类别和描述。

RDRAG方法的各个关键模块,旨在结合相似案例检索和大模型检索增强,优化多模态隐患识别系统的准确性和可靠性。在无训练优化框架下,RDRAG方法不依赖于大规模训练数据,通过动态调整提示(prompt)和借助相似案例增强大语言模型的表现。

4 实 验

4.1 数据集

为了验证添加相似案例检索对大语言模型在隐患识别任务中性能提升的有效性,选取了山东省高速施工真实图像,并以此构建了Rwecd数据集进行实验研究。该数据集包含325幅隐患图像样本,涵盖15种不同的隐患类别,确保数据覆盖了目标工业场景的主要隐患类别和典型环境,每幅图像均标注了其对应的隐患类别及隐患描述,用于与大模型的输出结果进行对比,以评估识别准确性。在实验设计中,采用分层抽样策略,从数据集中抽取105幅图像样本构建案例检索库,确保各类别的样本分布均衡;其余220幅图像样本作为测试集,用于评估模型在未见数据上的泛化能力。

4.2 评估指标

为了更精确地衡量大模型在隐患识别任务中的表现,引入以下3种评估指标对生成结果进行评估。

1)Category Accuracy。分类准确率指标用于衡量大模型所预测的隐患类别L_i'与真实类别L_i之间的一致性,反映了模型在隐患分类任务上的判别能力。定义为

$$f_{\text{CategoryAccuracy}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{k}(L_i' = L_i) \tag{7}$$

式中,$\mathbb{k}(\cdot)$是指示函数,当$L_i' = L_i$时取值为1,否则为0,N表示总样本数量。

2)BERT Similarity(Wang等,2024b)。该指标用于衡量模型生成的隐患描述C_i'与真实隐患描述C_i之间的语义相似性,使用预训练的BERT(bidirectional encoder representations from Transformers)模型对两个文本进行编码并计算余弦相似度,评估生成内容在深层语义层面的贴合程度。定义为

$$f_{\text{BERTSim}}(C_i', C_i) = \frac{f_{\text{BERT}}(C_i') \cdot f_{\text{BERT}}(C_i)}{\| f_{\text{BERT}}(C_i') \| \cdot \| f_{\text{BERT}}(C_i) \|} \tag{8}$$

式中,$f_{\text{BERT}}(\cdot)$为使用BERT提取的句向量表示。

3)TF-IDF Similarity(Jain等,2024)。该指标用于衡量生成文本C_i'和真实描述C_i在关键词层面的重合程度。利用TF-IDF(term frequency-inverse document frequency)向量表示两个文本,并计算余弦相似度,评估二者在关键信息覆盖方面的匹配度。其定义为

$$f_{\text{TFIDFSim}}(C_i', C_i) = \frac{f_{\text{TFIDF}}(C_i') \cdot f_{\text{TFIDF}}(C_i)}{\| f_{\text{TFIDF}}(C_i') \| \cdot \| f_{\text{TFIDF}}(C_i) \|} \tag{9}$$

式中,$f_{\text{TFIDF}}(\cdot)$为文本的TF-IDF向量表示:最终模型生成效果可以通过以上3个指标综合评估,分别对应类别准确性、语义匹配程度与关键词匹配程度,为无训练优化框架的有效性提供客观度量。

4.3 对比模型

本文在Rwecd数据集上验证了RDRAG的有效性,并与目前主流的多模态大模型进行对比。

1)GLM-4V是通用语言模型智谱ChatGLM系列的多模态扩展版本,专注于结合视觉和语言信息进行任务处理。

2)GPT-4o是OpenAI推出的多模态大语言模型,是ChatGPT系列的升级版本,以强大的语言能力为核心,扩展至多模态交互。

3)DeepSeek-VL2是DeepSeek团队开发的多模态大语言模型,专注于视觉与语言的深度融合,适合复杂检索与生成任务。

4.4 prompt对比分析

为系统评估提示模板设计对多模态大模型性能的影响,基于GLM-4V设计并展开4组对比实验,分别添加4种不同的提示方式(type1—type4),在安全隐患图像识别任务上开展对比研究,并通过计算BERT Similarity和TF-IDF Similarity评估不同提示模板下识别结果的语义匹配程度与关键词匹配程度,并以此选择最优的prompt方案。4个实验组评估数据如图3所示,得到分析结果如下:

1)type1提示模板在两项关键评估指标上均表现欠佳,反映出其生成结果的一致性与判别准确性

存在显著不足。该模板采用最简结构，未引入任何约束条件，适用于无需额外上下文引导的任务场景。然而，由于其缺乏明确的语义引导，模型输出呈现较高的随机性，甚至无法稳定判别目标类别。这一现象表明，在复杂视觉理解任务中，过于简化的提示设计可能导致模型性能的显著退化。

2）type2 模板通过添加类别信息显著提升了关键词匹配度，但语义相似性却出现下降。这一矛盾性结果表明，类别约束虽然能有效缩小模型输出空间，减少类别误判，但由于缺乏结构化输出要求，生成内容的语义连贯性与格式一致性仍存在缺陷。该发现印证了提示需在约束强度与表达自由度间寻求平衡。

3）type3 模板通过规范化输出格式，实现了 4 组中最优的语义相似性，但其关键词匹配度较 type2 有所降低。这一差异凸显了结构化提示的双重效应：一方面，固定输出模板可显著提升结果的可读性与逻辑一致性；另一方面，未明确限定类别范围可能导致模型在高层语义推理时出现偏差。

4）type4 实验组同时添加类别和格式信息，在保持语义相似性与 type3 组结果接近的同时，进一步提高了结果的关键词匹配程度。这一设计结合了类别和格式信息，确保了模型输出的类别选择准确且格式规范。通过明确类别范围，减少了错误分类的风险，而格式化的输出增强了结果的结构化和一致性。这对于实际的工地隐患检测任务尤为重要，能够提升生成结果的可用性和后续处理的便利性。因此最终选择 type4 作为最终实验的 prompt 方案。

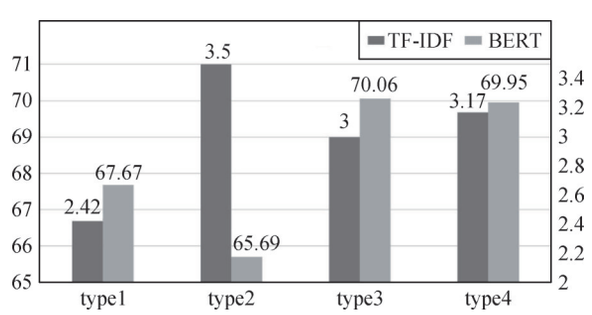

图 3　prompt 对比实验评估数据

Fig. 3　Evaluation data for prompt comparison

4.5　性能评估

在 GLM-4V、GPT-4o 和 DeepSeek-VL2 这 3 种多模态大语言模型上，本文设计并部署了 3 组方案进行实验。首先，采用设计好的 prompt，不引入任何辅助方法，作为 Base 对照组。其次，作为传统方法的对照组，在原有 prompt 设计中添加思维链（chain-of-thought，COT）（Miao 等，2024）进行引导，即要求大模型首先在样本图像中定位关键隐患目标物，再基于目标物对隐患类别进行判断，以测试思维链引导对模型推理能力的提升效果。最后，采用 RDRAG 框架，从案例库中提取与样本图像相似度最高的案例，并将其作为附加上下文信息输入模型，以评估检索增强方法对模型识别性能的优化作用。实验结果如表 1 所示。可以看出，在引入 COT 方法后，3 种大模型的各项数据均未得到明显提升，DeepSeek-VL2 甚至出现了下降。相比之下，采用 RDRAG 后，GLM-4V 和 DeepSeek-VL2 的识别准确性有较大幅度提升，同时 3 种大模型的语义匹配程度和关键词匹配程度也均得到较明显提升。实验结果表明，传统思维链（COT）引导方法在提升大模型隐患识别性能方面效果有限，甚至某些情况下会干扰大模型识别。这一现象可能源于 COT 方法在复杂场景中未能有效解决背景干扰问题，导致模型在定位关键隐患目标物时出现偏差，进而影响后续隐患类别的判断。相比之下，RDRAG 框架通过从案例库中检索相似案例作为附加上下文信息，显著提升了模型的识别性能，该结果验证了 RDRAG 框架在增强模型对复杂场景适应能力方面的有效性，表明检索增强方法能更好地辅助模型理解上下文信息，从而提升隐患识别的精度和鲁棒性。

表 1　实验结果数据

Table 1　Experimental results data

方法	模型	准确率/%	BERT	TF-IDF
Base	GLM-4V	14.51	69.95	3.17
	GPT-4o	53.54	71.67	5.75
	DeepSeek-VL2	14.91	68.15	2.34
COT	GLM-4V	17.28	70.09	3.68
	GPT-4o	55.08	71.30	4.64
	DeepSeek-VL2	12.11	66.87	2.33
RDRAG	GLM-4V	50.00	77.51	11.83
	GPT-4o	59.09	73.81	6.40
	DeepSeek-VL2	36.53	72.25	6.86

4.6　消融实验

为了评估图像相似度检索模块的有效性，以及选取 CLIP 作为计算图像相似度算法的合理性，设计

实施了消融实验。实验设计中，首先移除图像相似度检索模块，并采用随机检索策略，即从检索库中随机抽取案例作为附加上下文信息输入模型，以此构建Base组；接着，用学习感知图像块相似度（learned perceptual image patch similarity，LPIPS）算法替换图像相似度检索模块中的CLIP算法，以此计算并搜索相似度最高的图像，构建LPIPS组。LPIPS是一种基于深度学习的衡量图像相似性的指标，强调局部感知相似性，关注细节结构，其应用在图像修复等任务中有较好的表现，但不同于CLIP，LPIPS无跨模态能力，因此缺乏图像语义内容的理解能力。实验结果如表2所示。

通过对比分析，得到如下结果：1）检索库机制对模型性能的影响：3种多模态大语言模型引入RDRAG后评估指标有所提升，其中，GLM-4V和DeepSeek-VL2提升效果较为显著，无论是LPIPS还是CLIP应用在检索模块中都提高了多模态大语言模型的识别能力，说明提示学习通过优化模型的上下文理解能力，显著增强了隐患识别的准确性与鲁棒性。2）CLIP方法的有效性：相比于应用CLIP的RDRAG组，应用LPIPS的实验组对多模态大语言模型的提高效果不够明显，甚至在GPT-4o模型上准确率出现了下降，说明LPIPS虽然在感知相似性任务中具有较好的表现，但由于缺少跨模态的识别能力，在理解图像内容相似性上，CLIP是更加有效的选择。

表2　消融实验结果

Table 2　Ablation analysis results

模型	方法	准确率/%	BERT	TF-IDF
GLM-4V	RDRAG	50.00	77.51	11.83
	LPIPS	43.64	77.11	9.63
	Base	37.73	76.49	6.66
GPT-4o	RDRAG	59.09	73.81	6.40
	LPIPS	42.92	74.18	7.73
	Base	59.09	73.01	6.26
DeepSeek-VL2	RDRAG	36.53	72.25	6.86
	LPIPS	27.85	70.44	4.85
	Base	24.20	70.17	3.31

4.7　深入分析

为探究本文提出的RDRAG方法对多模态大语言模型在隐患识别任务中性能提升的有效性，以GLM-4V模型为例，分别对数据集中15个隐患类别进行了系统评估。本文统计了模型在引入RDRAG前后不同类别Category Accuracy的数据，通过定性分析，探究影响RDRAG方法在少样本类别中表现的因素，实验数据如表3所示。通过深入分析不同

表3　不同类别间准确率对比

Table 3　Per-class accuracy benchmarking

类别编号	隐患内容	数量	Base/%	RDRAG/%
1	未按规定穿戴反光安全服	4	0.00	33.33
2	高处作业未正确使用安全带	15	46.00	33.33
3	配电箱未及时锁闭	30	26.00	60.00
4	未按规定配置灭火器、消防设施等	20	0.00	50.00
5	现场防护栏等安全防护设施缺失、破损或设置不规范	25	32.00	23.53
6	设备安全防护设施、装置缺失或失效	25	12.00	64.71
7	起重吊装设备钢丝绳磨损、断丝严重，搭接长度不足	25	0.00	58.82
8	汽车吊、随车吊、泵车支腿未全部伸出、未垫枕木进行作业	30	0.00	70.00
9	基坑支护措施不到位	12	66.67	12.50
10	灭火器未按规定要求放置	6	33.33	0.00
11	未按规定设置接地线或接地不良	28	0.00	31.58
12	安全警示标志标识缺失或设置不规范	20	55.00	35.71
13	灭火器压力不足，灭火器、消防设施等未按规定进行检查、维护	25	0.00	23.53
14	不符合“三级配电两级漏电保护、一机一闸一漏一箱”要求	30	0.00	60.00
15	电缆外皮破损或敷设不规范	30	0.00	65.00

类别的准确率变化，得到以下结果：

1）RDRAG方法的优化效果。在引入RDRAG方法后，GLM-4V模型在大部分隐患类别中的评估指标均呈现显著性提升，尤其是在数据样本较多（如第3、8类）或关键目标物相似（如第4、13类）的类别识别上有着正向提升效果。

2）少样本类别的表现。对于一些样本数量极少的类别（如第1、10类），RDRAG方法的优化效果并不稳定，这是由于模型本身难以学习到判别性特征，此时性能上限受到数据制约。

3）小目标感知问题。大模型在图像细节及上下文环境捕捉方面表现优异，但易受复杂背景干扰，导致在复杂场景中难以精准识别隐患点，而RDRAG方法在场景较为复杂的类别（如第6、8、15类）识别上，展现出了较为显著的优化效果。

5 结 论

本文提出一种基于检索增强生成的无训练优化框架RDRAG。该框架通过结合提示微调和相似案例检索，解决多模态领域中因特征关联弱化而导致的误判问题。实验表明，提出的RDRAG可以通过检索增强优化大模型的输出，显著提升大模型对隐患的上下文理解能力，从而增强其识别与泛化的检测性能。未来将尝试引入更精细的RAG（Zhu等，2024）提示增强技术以提升模型的推理能力，弥补模型在复杂场景下隐患点捕捉能力的不足并强化其在小目标感知类隐患识别任务中的表现。

参考文献（References）

Arooj S, Altaf S, Ahmad S, Mahmoud H and Mohamed A S N. 2024. Enhancing sign language recognition using CNN and SIFT: a case study on Pakistan sign language. Journal of King Saud University-Computer and Information Sciences, 36(2): #101934 [DOI: 10.1016/j.jksuci.2024.101934]

Arslan M, Munawar S and Cruz C. 2024. Business insights using RAG-LLMs: a review and case study. Journal of Decision Systems: #2410040 [DOI: 10.1080/12460125.2024.2410040]

Ballal S, Patel K A and Patel D A. 2024. Enhancing construction site safety: natural language processing for hazards identification and prevention. Journal of Engineering, Project, and Production Management, 14(2): #0014 [DOI: 10.32738/jeppm-2024-0014]

Cao J L, Li M Z N, Wen M and Cheung S C. 2025. A study on prompt design, advantages and limitations of ChatGPT for deep learning program repair. Automated Software Engineering, 32(1): #30 [DOI: 10.1007/s10515-025-00492-x]

Cui Y N, Sun Z Q and Hu W. 2024. A prompt-based knowledge graph foundation model for universal in-context reasoning //Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #227 [DOI: 10.52202/079017-0227]

Fan C, Wu Q T, Zhao Y and Mo L K. 2024. Integrating active learning and semi-supervised learning for improved data-driven HVAC fault diagnosis performance. Applied Energy, 356(2): #122356 [DOI: 10.1016/j.apenergy.2023.122356]

Ghosh A, Acharya A, Jain R, Saha S, Chadha A and Sinha S. 2024. CLIPSyntel: CLIP and LLM synergy for multimodal question summarization in healthcare//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: #2458 [DOI: 10.1609/aaai.v38i20.30206]

Gu Z P, Zhu B K, Zhu G B, Chen Y Y, Tang M and Wang J Q. 2024. AnomalyGPT: detecting industrial anomalies using large vision-language models//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: #215 [DOI: 10.1609/aaai.v38i3.27963]

Guo Y F, Yu Z T, Liu A S, Zhou W B, Qiao T, Li B, et al. 2025. Recent progress of the security research for multimodal large models. Journal of Image and Graphics, 30(6): 2051-2081 (郭园方, 余梓彤, 刘艾杉, 周文柏, 乔通, 李斌, 等. 2025. 多模态大模型安全研究进展. 中国图象图形学报, 30(6): 2051-2081) [DOI: 10.11834/jig.250067]

Hou X Y, Li C Q and Fang Q. 2023. Computer vision-based safety risk computing and visualization on construction sites. Automation in Construction, 156: #105129 [DOI: 10.1016/j. autcon. 2023. 105129]

Hou Z Q, Qu M J, Li J G, Ma S G, Wang Y C and Yang X B. 2025. Lightweight CNN-Transformer combined network for real-time semantic segmentation. Journal of Image and Graphics, 30(7): 2437-2450 (侯志强, 屈敏杰, 李俊歌, 马素刚, 王昀琛, 杨小宝. 2025. 轻量级CNN-Transformer相结合的实时语义分割网络. 中国图象图形学报, 30(7): 2437-2450) [DOI: 10.11834/jig.240527]

Jain S, Jain S K and Vasal S. 2024. An effective TF-IDF model to improve the text classification performance//Proceedings of the 13th IEEE International Conference on Communication Systems and Network Technologies (CSNT). Jabalpur, India: IEEE: 1-4 [DOI: 10.1109/CSNT60213.2024.10545818]

Kim T T, Makutonin M, Sirous R and Javan R. 2025. Optimizing large language models in radiology and mitigating pitfalls: prompt engineering and fine-tuning. RadioGraphics, 45(4): #240073 [DOI: 10.1148/rg.240073]

Lewandowski M, Łukowicz P, Świetlik D and Barańska-Rybak W.

2024. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the specialty certificate examination in dermatology. Clinical and Experimental Dermatology, 49(7): 686-691 [DOI: 10.1093/ced/llad255]

Luo Y M, Wang R M, Zhang F W, Zhou F, Liu M Y and Feng J W. 2024. Video Q and A based on two-stage deep exploration of temporally-evolving features with enhanced cross-modal attention mechanism. Neural Computing and Applications, 36(14): 8055-8071 [DOI: 10.1007/s00521-024-09482-8]

Miao J, Thongprayoon C, Suppadungsuk S, Krisanapan P, Radhakrishnan Y and Cheungpasitporn W. 2024. Chain of thought utilization in large language models and application in nephrology. Medicina, 60(1): #148 [DOI: 10.3390/medicina60010148]

Soumya A, Linga Reddy C, Vishnu C and Krishna Mohan C. 2024. Multi-class object classification using deep learning models in automotive object detection scenarios//Proceedings of the 16th International Conference on Machine Vision. Yerevan, Armenia: SPIE: #1307206 [DOI: 10.1117/12.3023463]

Wang F, Ding L, Rao J, Liu Y, Shen L and Ding C X. 2024a. Can linguistic knowledge improve multimodal alignment in vision-language pretraining? ACM Transactions on Multimedia Computing, Communications and Applications, 20(12): #364 [DOI: 10.1145/3690640]

Wang J J, Huang J X, Tu X H, Wang J M, Huang A J, Laskar M T R, et al. 2024b. Utilizing BERT for information retrieval: survey, applications, resources, and challenges. ACM Computing Surveys, 56(7): #185 [DOI: 10.1145/3648471]

Wang J Y, Zhang H J, Zhong Y H, Liang Y B, Ji R W and Cang Y R. 2024c. Advanced multimodal deep learning architecture for image-text matching//Proceedings of the 4th IEEE International Conference on Electronic Technology, Communication and Information (ICETCI). Changchun, China: IEEE: 1185-1191 [DOI: 10.1109/ICETCI61221.2024.10594167]

Wang X T, Liu C Y, Song X N and Cui X Z. 2022. Development of an internet-of-things-based technology system for construction safety hazard prevention. Journal of Management in Engineering, 38(3): #04022009 [DOI: 10.1061/(ASCE)ME.1943-5479.0001035]

Wang Z Q, Wang H P and El Saddik A. 2024d. FedITD: a federated parameter-efficient tuning with pre-trained large language models and transfer learning framework for insider threat detection. IEEE Access, 12: 160396-160417 [DOI: 10.1109/ACCESS.2024.3482988]

Wei Y B, Xu Y, Zhu L, Ma J W and Peng C M. 2024. Multi-level cross-modal contrastive learning for review-aware recommendation. Expert Systems with Applications, 247: #123341 [DOI: 10.1016/j.eswa.2024.123341]

Yao W J, Bai J J, Liao W, Chen Y H, Liu M J and Xie Y. 2024. From CNN to transformer: a review of medical image segmentation models. Journal of Imaging Informatics in Medicine, 37(4): 1529-1547 [DOI: 10.1007/s10278-024-00981-7]

Yu X, Zhang Z X, Niu F F, Hu X, Xia X and Grundy J. 2024. What makes a high-quality training dataset for large language models: a practitioners' perspective//Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering. Sacramento, USA: Association for Computing Machinery: 656-668 [DOI: 10.1145/3691620.3695061]

Zeng W and Xiao Z Y. 2024. Few-shot learning based on deep learning: a survey. Mathematical Biosciences and Engineering, 21(1): 679-711 [DOI: 10.3934/mbe.2024029]

Zhang L, Xu W Y, Shen C and Huang Y P. 2024. Vision-based on-road nighttime vehicle detection and tracking using improved HOG features. Sensors, 24(5): #1590 [DOI: 10.3390/s24051590]

Zhu L B, Rong Y, McGee L A, Rwigema J C M and Patel S H. 2024. Testing and validation of a custom retrained large language model for the supportive care of HN patients with external knowledge base. Cancers, 16(13): #2311 [DOI: 10.3390/cancers16132311]

作者简介

李嘉威，男，硕士研究生，主要研究方向为大语言模型。
E-mail：lijiawei011124@163.com
孟雷，通信作者，男，教授，主要研究方向为多模态深度学习及应用。E-mail：lmeng@sdu.edu.cn
杨成业，女，硕士研究生，主要研究方向为对话系统和大语言模型。E-mail：yangchengye2000@gmail.com
张尧臣，男，正高级工程师，主要研究方向为人工智能、大语言模型、隐私计算及网络安全。E-mail：zyc@inspur.com
孙玮琳，女，博士研究生，主要研究方向为大语言模型驱动的三维场景生成。E-mail：sunweilin@mail.sdu.edu.cn
孟祥旭，男，教授，主要研究方向为人机交互与虚拟现实。
E-mail：mxx@sdu.edu.cn