2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

ENHANCING VISUAL-LLM THROUGH PROMPT ENGINEERING AND HYBRID RETRIEVAL-AUGMENTED GENERATION FOR SITE SAFETY COMPLIANCE CHECKING

Koi Xiaowen Guo[1], Peter Kok-Yiu Wong[1], Jack C.P. Cheng[1], Xingyu Tao[1], and Pak-Him Leung[2]
[1]The Hong Kong University of Science and Technology, Hong Kong SAR
[2] AutoSafe Limited, Hong Kong SAR

Abstract

The increasing prevalence of safety incidents on construction sites states the urgent need for enhanced monitoring. This study proposes an innovative hybrid Retrieval-Augmented Generation (RAG) algorithm to compliance check accuracy for site images. By integrating the Visual Language Model (VLM), we developed an algorithm capable of mastering domain knowledge without fine-tuning and addressing the limitation of interpreting RAG technology with visual information. A three-phased prompting framework was designed to enhance the VLM's compliance analysis abilities. Experiments based on actual construction site in Hong Kong demonstrated 21.98% increase in retrieval accuracy.

Introduction

The construction industry is inherently a high risk undertaking and prone to fatal accidents. For instance, the number of fatalities in the construction industry in the United States in 2022 was 1,069, an increase of 83 compared to the previous year (U.S. Department of Labor, 2023). Despite the implementation of a variety of safety management measures, accidents continue to occur frequently. With the continuous advancement of AI technology, the industry is beginning to recognise the important role and potential of AI in enhancing the effectiveness of construction safety management.

Computer vision (CV) techniques have been widely explored and implemented in the domain of construction safety. These methods analyse site images or videos to identify potential safety hazards and perform comprehensive safety assessments (Cheng, JackC. P. et al., 2022; Luo, H. et al., 2020). However, the current CV-based safety analysis models still exhibit certain limitations. Such models are often tailored for specific tasks, which results in their limited generalisability when faced with variability in the built environment. As an illustration, a model designed to detect personal protective equipment (PPE) may fail to identify the risk of a worker falling. Furthermore, these models are typically constrained to process a limited range of semantic information, such as the label and location of an object. They are unable to provide a comprehensive understanding of complex semantic relationships, such as the interactions between diverse construction elements, which are essential for conducting thorough safety compliance inspections.

Large language models (LLMs) based on pre-trained transformers, such as the GPT family, have demonstrated the potential to process multimodal data and deal with complex sematic reasoning, including images, text, and audio, for example, GPT-4o developed by OpenAI (OpenAI, 2023). The advances in these models have opened up new avenues for extracting deep semantic information from images and videos of construction sites.

While Visual Language Models (VLMs) pre-trained in the general domain show potential for common tasks, they may not be directly usable for tasks in specialised domains. As several studies have pointed out, the direct use of large language models (LLMs) in specific domains often results in poor performance due to the scarcity of the specific domain training data and challenges in the training process (Ouyang et al., 2022; Wei et al., 2022). Fine-tune the model using datasets from special domains to gain knowledge of the relevant domain requires a large amount of data, and insufficient data can lead to model overfitting or even performance degradation (Kirkpatrick et al., 2017). In the site safety area, the limited data set of unsafe behaviours restricts the ability to fine-tune the VLM for complex tasks.

An application of knowledge enhancement in LLMs is Retrieval-Augmented Generation (RAG), an approach to enhance LLMs' comprehension and generative capabilities by integrating external knowledge. RAG generally involves the retrieval of relevant information from a knowledge database based on user queries. The retrieved information then augments the prompting of an LLM to generate a more context-specific response. RAG can provide a higher degree of customized and domain-tailored application of LLMs, through injecting grounded information to guide LLMs' reasoning process.

Considering the aforementioned challenges, we propose an RAG-based framework to augment the capabilities of VLMs in the construction safety domain, which serves to enhance the understanding and application of domain-

specific knowledge by integrating safety regulations into an external knowledge base (Gao, L. et al., 2022; Ma et al., 2023), thus obviating the necessity to fine-tune the VLM itself. A hybrid searching and three-phase prompting framework is developed to assist VLMs in the mapping of image features to textual features. This strategy enhances the model's accuracy in processing domain knowledge tasks and reduces hallucinations phenomenon.

The proposed framework is validated with images from an actual construction site in Hong Kong. The analytical system offers significant time and cost savings in data collection and training, while enabling VLMs to acquire domain-specific knowledge and automatically extract complex semantic information from safety regulations to facilitate compliance checking based on construction site images. Furthermore, our research illustrates the adaptability and scalability of VLM in responding to alterations in safety requirements, modifications in monitoring tasks, and changes in construction scenarios. This considerably reduces the time devoted to safety inspection and provides robust adaptability to changes in construction activities.

The remainder of this paper is organized as follows. Section 2 details our methodology, covering the hybrid retrieval mechanism, knowledge base construction, and the three-stage prompting strategy. Section 3 presents the experimental design and results, including dataset construction, evaluation metrics, and comparative analysis with baseline methods. It also discusses the robustness of the model, limitations of the data, and potential integration with existing systems. Finally, Section 4 concludes the paper and outlines future research directions.

Methodology

Our overall workflow is shown in *Figure 1*, and involves employing the VLM to process construction site images by extracting image features to text features. The RAG technique and our proposed three-phase prompting strategy are then utilised to improve the stability and scalability of the VLM towards domain-specific knowledge. The hybrid retrieval mechanism further enhances the RAG technique.

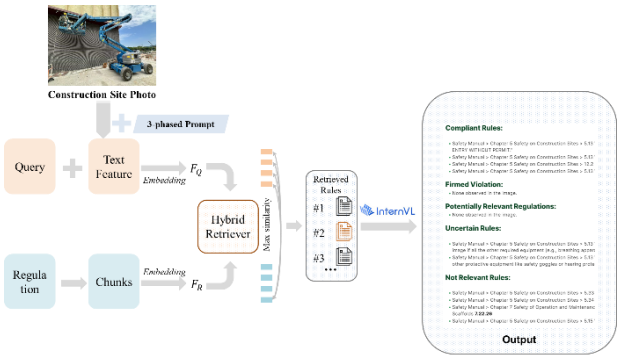

Figure 1: Overall framework

Constitution of the domain-specific knowledge base for safety behaviours

Data storage constitutes a pivotal component in the framework's efficacy. Before ingesting construction safety regulation texts into the knowledge base, subjecting these texts to a vectorisation process is imperative. Safety documents need to be split into text blocks containing less information and having similar attributes to improve the effectiveness of semantic retrieval. The size of chunking is crucial for retrieval performance. Unsuitable chunking approaches are prone to knowledge confusion and information loss. For example, if the chunks are overly large, splitting the end and beginning of rule 5.20 on lifting and rule 5.21 on welding into the same chunk will result in semantic confusion and affect the search accuracy. Therefore, in construction safety, dividing text blocks according to the content of each section, especially the regulation codes, is the optimal choice.

When dealing with safety regulations, key information lies beyond the specific content of the regulation: the regulation title, the chapter name, the source, etc., are also important. Therefore, we include the title and chapter name in each chunk and stored source in metadata to record the provenance of each rule and assigned a unique ID to each chunk (*Figure 2*). This approach enables rapid locating and modification in case of future updates or changes to the safety regulations, without having to reprocess and vectorise the entire document. The chunk is converted into an n-dimensional floating-point vector using a text vectoriser, and these vectorised chunks and their associated metadata are stored in a knowledge base for subsequent semantic retrieval and analysis.

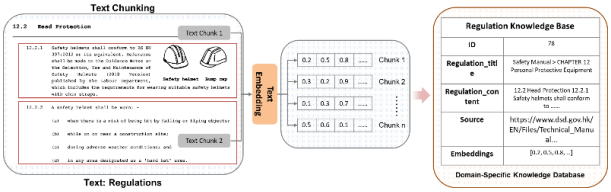

Figure 2: Knowledge Database Construction

Retrieval Strategy

Construction safety regulation retrieval is aims to efficiently retrieve safety regulations related to a specific site image from the knowledge base. In traditional manual compliance inspections, managers may miss important details when key information is scattered in different sections of the safety manual. For example, the rules for working at height lie in Chapter 5, while for PPE lie in Chapter 12. The automated semantic search allows separate relevant rules to be simultaneously considered, thus improving the comprehensiveness and accuracy of safety compliance checks.

Semantic Searching

The mainstream approach is using dense vector retrieval, matching based on semantic relevance. The basic principle of this technique is to embed text chunks in computer-understandable multidimensional vectors. At

the same time, a user's question is similarly vectorised. By comparing the similarity score, subtle semantic connections between the user's question and the documents can be found. The RAG system will extract the most relevant content as context information and feed it to the LLM with the user's question, helping the model to answer the user's question more precisely (Gao, Y. et al., 2024).

Since this study concerns multimodal data, after vectorisation, additional processing is necessary to ascertain the similarity between image and textual content. We propose an innovative architecture enabling the extraction and conversion of image features into text features. The converted site monitoring text features can then be combined with user queries, and the similarity scores are calculated with the regulatory text chunks. Similarity calculation is normally based on cosine similarity (Eq. 1) or Euclidean distance, both give similar results. Based on the similarity score ranking, the *top n* regulations can be incorporated into the VLM's input for subsequent compliance assessment. The specific conversion method is described in prompt engineering section, and detailed model used is provided in experiment design section.

The cosine similarity is used in this paper, and the calculation equation is shown below, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are two corresponding vectors to be compared.

$$\text{similarity} = \frac{A \cdot B}{|A||B|} \tag{1}$$

Hybrid Searching

Semantic-based dense vector retrieval performs well in handling complex text searches, especially in understanding similar semantics, multilingual understanding, and fault tolerance (Zhang, C. et al., 2024). However, for acronyms, phrases, or keyword searches it is less effective. In contrast, sparse retrieval is more effective for exact matches. Considering safety rules contain many specialised terms and acronyms, such as PPE and Powered Elevating Work Platforms (PEWP), sparse retrieval improves the chances of obtaining such rules.

A two-stage hybrid retrieval algorithm is deployed for this purpose, combining the advantages of dense and sparse retrieval while compensating for their respective shortcomings. The first stage utilises both the dense and sparse search models to recall the relevant regulations comprehensively. The second stage employs a reranker to perform fine-grained ranking and low-quality filtering of the results from the first stage, thereby ensuring the accuracy and relevance of the results. Figure 3 shows the framework of hybrid retrieval.

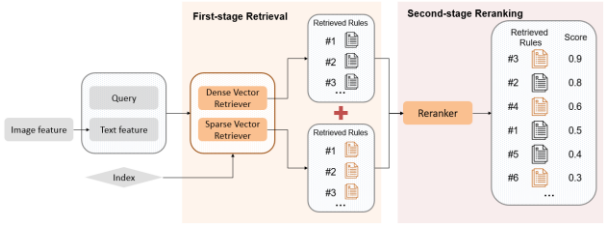

Figure 3: Framework of Hybrid Retrieval

This study employs the BM25 (Best Matching 25) algorithm (Trotman et al., 2014) as a sparse retrieval model for similarity retrieval.The algorithm's core idea is the combination of term frequency (TF) and inverse document frequency (IDF), with consideration of document length, is used to calculate the similarity score. The BM25 algorithm's formula is provided below:

$$BM25(D,Q) = \sum_{i=1}^{n} \frac{(k+1)\cdot f_i}{f_i + k\cdot\left(1 - b + b\cdot\frac{|D|}{avgdl}\right)} \cdot log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right) \tag{2}$$

where: D is the document, Q is the query, N is the number of documents in the document set, f_i is the frequency of term i in documents, n_i is the number of documents containing term i, $|D|$ is document length, avgdl is the average length of the document set, and k and b are adjustable parameters.

In the first stage, the model filters the $top\ n$ and $top\ m$ most relevant regulatory text chunks by vector retrieval and BM25 retrieval, respectively. In the second stage, these $m + n$ text blocks are merged into a list and combined with user questions and image features to form document-query pairs, which serve as inputs to the ranking model to determine the final $top\ i$ chunks.

The underlying rationale is that an VLM's capacity to access information is diminished when the context window has excessive information. The model's compliance with instructions is adversely affected when the context window is overloaded (Gao, Y. et al., 2024). Consequently, these text blocks require further filtering through the reranking model before feeding them into the VLM.

Following the filtering of the most relevant rules based on semantic similarity, the resulting $top\ i$ text is to integrate with the VLM prompt to determine whether a violation has occurred.

Prompt Engineering

Three-phase Prompting Strategy for feature mapping

We propose a three-phase prompting strategy to convert image features into text features to match rule features. Several studies have improved the performance of the LLM after decomposing complex tasks into simple subtasks (Wei et al., 2024; Zhou et al., 2022). Therefore, we adopt a phased guidance approach to reduce the cognitive load at each step, improve the comprehension and reasoning ability of the VLM. For the mapping site monitoring image features task, a three-phase prompting strategy is designed based on human cognitive habits,

which is processing the prioritising global information over local detail (Rezvani et al., 2020).

The following section presents the detailed design ideas for the strategy (Figure 4).

1. **Macro-level Scene Comprehension:** The initial phase concentrates on the comprehensive layout of the site and the construction environment, analogous to a "macro-scan" conducted by a human observer.
2. **Activity and Object Recognition:** The second phase concerns the recognition and classification of activities and objects on the construction site. In this phase, the model identifies the key activities and objects in the image without delving into the minutiae.
3. **Behavior and Detail Analysis:** The final phase focuses on finer details, including the specific actions of the worker, the tools used, and their interactions, which emphasises the processing of local features depending on the context provided by the previous phases.

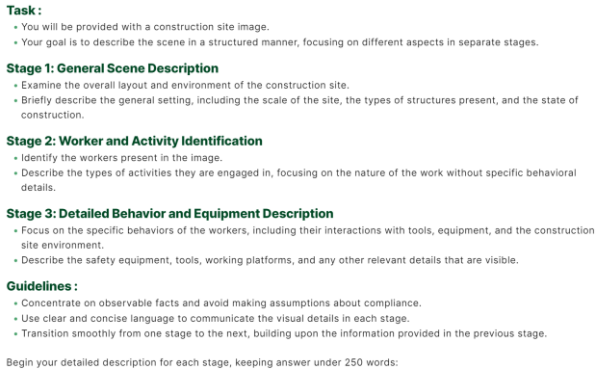

Figure 4: image feature mapping prompt

Experiment Design and Results

This section assesses the viability of the proposed RAG algorithm in comparison to conventional vector semantic retrieval, and its capacity to facilitate the safety regulations checking through our prompt-augmented VLM. All responses are generated using the InternVL-2 Pro model (Chen, Z. et al., 2024).

The proposed method was executed on the Ubuntu 22.04 64-bit system environment with an Intel Core i7 13700H CPU, a GeForce RTX 4060 Laptop 16G GPU.

Experimental design

Knowledge database preparation

Two Hong Kong construction safety codes were selected for the construction of the code knowledge base: the Safety Manual developed by the Drainage Services Department (DSD) and the Work-At-Height Safety Handbook published by the Construction Industry Council (CIC). 98 building safety regulations were extracted for evaluation based on different scenarios. All clauses were saved in JSON format according to the data schema in Section 3.1.

Collection of construction site images

The existing large-scale datasets comprise primarily life scenes and lack construction scenes. To address this issue, we collected and constructed our own test dataset to validate the InternVL's ability to judge safety compatibility in construction site images. The dataset contains 150 construction site cases across scenarios with different complexity and layouts, obtained from the internet and an actual DSD construction site in Sheung Shui, Hong Kong, with an even distribution of images across the different scenario categories, as shown in Figure 5.

Figure 5: Image Dataset

Data storage and retrieval

ChromaDb (Max Isom, 2024), an open-source, highly scalable vector database that supports vector embedding, was adopted to facilitate the storage and indexing of extensive vector data. In this study, a bce-embedding-base_v1 embedding model (NetEase Youdao, 2023) is employed, an Opaki_bm25 Python library is used for sparse retrieval, and a bce-reranker-base_v1 (NetEase Youdao, 2023) reranking model is adopted. The top 15 terms from the vector and the BM25 search were selected respectively based on the Euclidean distance score as inputs to the reranker. The reranker's score was calculated, and the top 10 rules were included as inputs for the VLM.

Evaluation metrics

To assess the efficacy of the hybrid retrieval method, we established a baseline model that employs the vector model only for retrieval and selects the top 10 rules with the highest similarity scores as the final retrieval results. This comparison enables an evaluation of the hybrid retrieval approach in improving retrieval accuracy. The retrieval performance was evaluated through accuracy.

For the retrieval accuracy calculation, R_i is the set of relevant rules retrieved for the i-th image, and T_i is the total set of rules retrieved for the i-th image. The function $f_r(i)$ can be defined as:

$$f_r(i) = \frac{\sum_{j=1}^{|T_i|} I(r_j \in R_i)}{|T_i|} \quad (3)$$

where r_j is the j-th rule in the set T_i, and $I(r_j \in R_i)$ is an indicator function that equals 1 if the j-th rule r_j is in the set of relevant rules R_i, and 0 otherwise.

To obtain the overall retrieval accuracy P_r, we calculate the average of $f_r(i)$ across all images:

$$P_r = \frac{1}{N}\sum_{i=1}^{N} f_r(i) \quad (4)$$

where N is the total number of images in the dataset being evaluated.

Results

The baseline approach is defined as the method that employs only the dense vector retrieval for semantic similarity retrieval, rather than hybrid retrieval, and the comparative accuracy is presented in Table 1.

Table 1: Accuracy Comparison between Hybrid and Dense Vector Only Retrieval

Dense vector only retrieval accuracy	Hybrid retrieval accuracy	Improvement
53.69%	75.67%	+21.98%

It can be observed that employing a hybrid retrieval strategy significantly enhances the accuracy of the search process. The reason is that the regulations contain many fixed and specialised terms, such as Powered Elevating Work Platform (PEWP), Personal Protective Equipment (PPE), etc. The BM25 algorithm can efficiently retrieve such terms from the image feature extraction results. Vector models, while effective in other contexts, are unsuitable for performing the exact matching of such specialized terms. The combination of the BM25 algorithm and the vector model enables a more accurate retrieval of regulations related to construction safety.

Discussion

In this section, we will discuss the different aspects of the system's performance, including the image feature mapping and compliance analysis performance.

1. Image Feature Mapping Performance

The efficacy of a three-phase prompting strategy was evaluated compared to a typical prompting approach in directing VLMs in image feature transformation. The typical prompting approach requires the VLM to recognize image details comprehensively, by the prompt "*Please describe the construction site image comprehensively.*" However, the output lacked structure, was cumbersome and unfocused, and failed to order the detail checking. For instance, the approach fails to recognise welding on wet ground, which is essential for ensuring the safety of welding operations on construction sites (Figure 7).

In contrast, the three-phase prompting strategy allows the VLM to emulate the human cognitive process of recognising the features of construction site images in a layered manner. The staged focus allows the VLM to rely on the information from the previous steps for subsequent feature detection, thereby enabling a more hierarchical, focused and structured conversion of image features to text features. It reduces the impact of local information bias on the overall judgment through staged feature extraction. In comparison, the VLM easily identified the key element of the wet floor in the same case with the proposed prompt, simultaneously improving the structure and conciseness of the textual features and enhancing the subsequent retrieval task efficiency (Figure 8).

Figure 6: Case 1: Welding on Wet Floor (Generated by Microsoft Copilot)

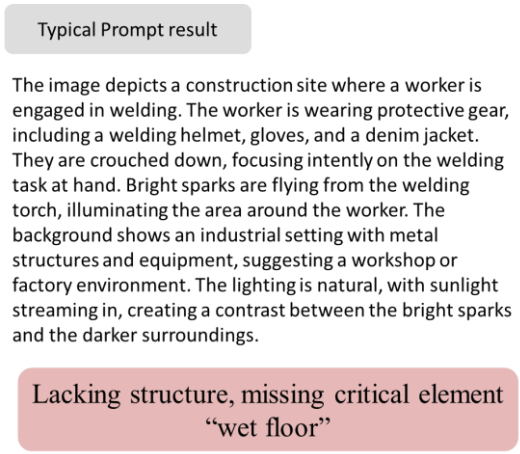

Figure 7: Typical Prompt Result of Image Feature Extraction

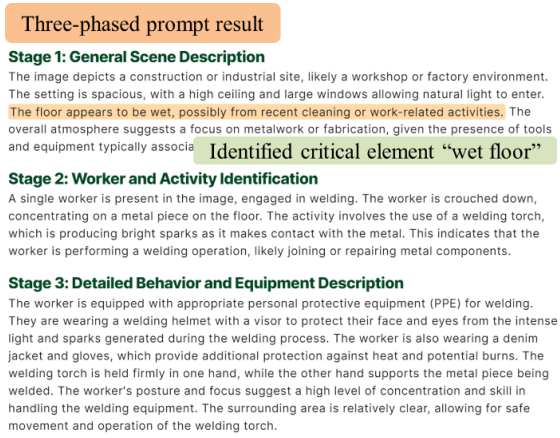

Figure 8: Three-phased Prompt Result of Image Feature Extraction

2. Compliance Analysis Performance

During the compliance check process by the VLM, we observed that the presence of irrelevant regulations could interfere with the VLM's ability to make correct judgments, leading more readily to hallucination phenomena. For instance, in Case 2, six correct regulations were retrieved (Figure 9), but the VLM's output included three incorrect judgments (Figure 10). In contrast, the correct retrievals increased to eight in thefor both Case 3 (Figure 11), and each rule returned a correct judgment by the VLM (Figure 12).

Figure 9: Retrieval Result for Case 2

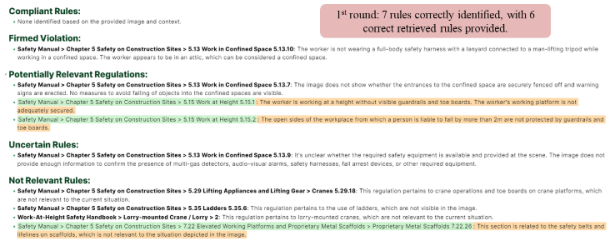

Figure 10: Compliance Analysis Results for Case 2

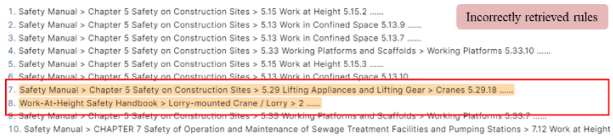

Figure 11: Retrieval Result for Case 3

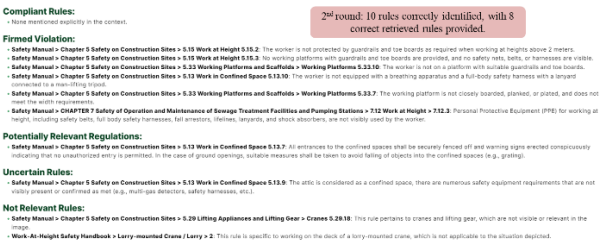

Figure 12: Compliance Analysis Results for Case 3

We constructed a graph (Figure 13) with the x-axis representing the number of correctly retrieved rules by RAG and the y-axis representing the number of correct judgments made by the VLM. The darker colour indicates a higher density of points. It can be observed that the points are predominantly concentrated in the upper right corner of the graph, indicating that the probability of the VLM making correct judgments increases with the number of correctly retrieved rules, a phenomenon ubiquitous in the experimental results. Thus, we posit that an excess of irrelevant information may interfere with the VLM's judgement.

To address the aforementioned challenges, the following optimization strategies are proposed: 1. Workflow Optimization for Individual Regulation Processing: A new workflow is recommended for implementation, wherein each regulation is processed independently. The VLM will assess each regulation separately rather than concurrently handling multiple regulations. 2. Enhancement of VLM's Feature Recognition Capability: To reduce the interference of incorrect retrieval content from the RAG system on the VLM's accuracy, we suggests enhancing the VLM's feature recognition capabilities by training the VLM with more construction domain features and patterns.

Through implementing these strategies, it is expected that the performance of the VLM in safety compliance checks on construction sites will be significantly enhanced, providing more reliable technical support for site safety management.

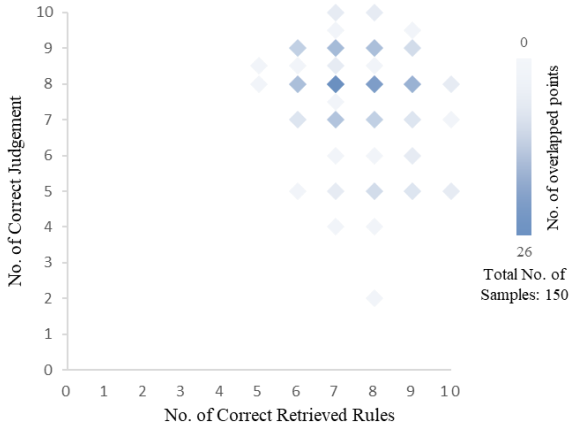

Figure 13: Correlation between 1st round Correct Retrieved Rules and Judgement

While the current framework has enhanced the model's adaptability to diverse scenarios through hybrid retrieval and phased prompting strategies, the sensitivity of VLMs to input order variations and the challenges associated with complex localization still require further investigation. Future work will integrate dynamic feature fusion and adversarial testing to systematically optimize the model's robustness in edge cases. Additionally, although the dataset used in this study covers a variety of construction scenarios, systematic analysis of challenging visual conditions such as extreme lighting, dense occlusions, or adverse weather remains to be explored. Future work will expand the dataset and introduce synthetic data augmentation techniques to quantify the model's performance boundaries in such complex environments, thereby further enhancing the reliability of practical deployment.

The current framework has achieved automated and efficient processing of compliance checks. Future research will further explore lightweight deployment strategies, such as model pruning and edge computing, to accommodate resource-constrained device environments on actual construction sites. The modular design and open interfaces of the framework support flexible integration into existing safety management systems. By adapting to

conventional data formats, such as image input and structured report output, and incorporating a dynamic knowledge base update mechanism, the framework can be seamlessly embedded into manual inspection or automated monitoring processes to assist in rapid decision-making. Future work will explore in-depth collaboration with Building Information Modeling (BIM) platforms to further enhance the practical value of the framework.

Conclusion

This research delves into the application of VLMs in the field of construction site safety monitoring and a retrieval-augmented generation algorithm aimed at enhancing the accuracy and efficiency of safety compliance checks is proposed. By integrating multimodal large language models, hybrid RAG technology, and a meticulously designed three-phase prompting strategy, the integration of multimodal data, automatically retrieving relevant information from a vast repository of regulations based on the content of construction site images, and conducting compliance analysis is achieved. This approach transcends the limitations of traditional methods that can only retrieve textual information. The method avoids fine-tuning the VLM, allowing it to flexibly handle a vast domain knowledge while maintaining high scalability and flexibility, thus saving significant resources and time costs. The proposed framework is validated with images from an actual construction site in Hong Kong and achieving a promising outcome. Experimental validation indicates that compared to traditional pure dense vector retrieval, the hybrid retrieval model is more suitable for the construction domain regulations and increases retrieval accuracy by 21.98%.

While our proposed strategy has demonstrated efficacy and achieved notable results, there are still some limitations. Firstly, although our three-phase prompting strategy effectively guides the VLM in feature extraction and transformation, it may still fall short in capturing all subtle safety violations in certain scenarios. This limitation may stem from the current VLM technology's challenges in comprehending and reasoning complex visual contexts. In forthcoming endeavours, we intend to design more potent prompt strategies to further stimulate the VLM's capabilities and integrate ontologies to enhance the system's retrieval accuracy and judgment capabilities. Secondly, due to computational resource constraints, our model currently achieves efficient automated monitoring, which may impact the model's response speed and efficiency in certain situations. To address this issue, future research could explore more efficient search strategies, such as optimizing the retrieval process through the integration of knowledge graphs. Knowledge graphs offer rich semantic information and entity relationships, facilitating faster location of relevant information and improving the efficiency and accuracy of retrieval.

Acknowledgments

The authors would like to acknowledge the project (No. DEMP/2023/25) funded by Drainage Services Department, the Government of the Hong Kong Special Administrative Region, Wanchai, Hong Kong Island, Hong Kong, SAR, for providing support to this research. The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y. and Dai, J. (2024) InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks [Online]. arXiv. Available from: <http://arxiv.org/abs/2312.14238> [Accessed 26 September 2024].

Cheng, JackC. P., Wong, P. K.-Y., Luo, H., Wang, M. and Leung, P. H. (2022) Vision-Based Monitoring of Site Safety Compliance Based on Worker *Re*-Identification and Personal Protective Equipment Classification. *Automation in Construction* [Online], 139 July, p. 104312. Available from: <https://www.sciencedirect.com/science/article/pii/S0926580522001856> [Accessed 25 October 2024].

Gao, L., Ma, X., Lin, J. and Callan, J. (2022) Precise Zero-Shot Dense Retrieval without Relevance Labels [Online]. arXiv. Available from: <http://arxiv.org/abs/2212.10496> [Accessed 26 September 2024].

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. and Wang, H. (2024) Retrieval-Augmented Generation for Large Language Models: A Survey [Online]. arXiv. Available from: <http://arxiv.org/abs/2312.10997> [Accessed 26 September 2024].

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D. and Hadsell, R. (2017) Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences* [Online], 114 (13) March, pp. 3521–3526. Available from: <https://www.pnas.org/doi/full/10.1073/pnas.1611835114> [Accessed 26 November 2024].

Luo, H., Wang, M., Wong, P. K.-Y. and Cheng, J. C. P. (2020) Full Body Pose Estimation of Construction Equipment Using Computer Vision and Deep Learning Techniques. *Automation in Construction* [Online], 110 February, p. 103016. Available from: <https://www.sciencedirect.com/science/article/pii/S09265805193063 4X> [Accessed 17 December 2024].

Ma, X., Gong, Y., He, P., Zhao, H. and Duan, N. (2023) Query Rewriting for Retrieval-Augmented Large Language Models [Online]. arXiv. Available from: <http://arxiv.org/abs/2305.14283> [Accessed 26 September 2024].

Max Isom (2024) Chroma [Online]. Chroma-core. Available from: <https://github.com/chroma-core/chroma> [Accessed 1 December 2024].

NetEase Youdao, Inc. (2023) BCEmbedding: Bilingual and Crosslingual Embedding for RAG [Online]. Available from: <https://github.com/netease-youdao/BCEmbedding>.

OpenAI (2023) GPT-4V(Ision) System Card [Online]. Available from: <https://cdn.openai.com/papers/GPTV_System_Card.pdf> [Accessed 17 December 2024].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. and Lowe, R. (2022) Training Language Models to Follow Instructions with Human Feedback [Online]. arXiv. Available from: <http://arxiv.org/abs/2203.02155> [Accessed 26 November 2024].

Rezvani, Z., Katanforoush, A. and Pouretemad, H. (2020) Global Precedence Changes by Environment: A Systematic Review and Meta-Analysis on Effect of Perceptual Field Variables on Global-Local Visual Processing. *Attention, Perception, & Psychophysics* [Online], 82 (5) July, pp. 2348–2359. Available from: <https://doi.org/10.3758/s13414-020-01997-1> [Accessed 2 December 2024].

Trotman, A., Puurula, A. and Burgess, B. (2014) Improvements to BM25 and Language Models Examined [Online]. In: *Proceedings of the 19th Australasian Document Computing Symposium*, *November 26, 2014*. New York, NY, USA: Association for Computing Machinery, pp. 58–65. Available from: <https://dl.acm.org/doi/10.1145/2682862.2682863> [Accessed 2 December 2024].

U.S. Department of Labor (2023) *Injuries, Illnesses, and Fatalities* [Online]. Bureau of Labor Statistics. Available from: <https://www.bls.gov/iif/fatal-injuries-tables/fatal-occupational-injuries-table-a-1-2022.htm> [Accessed 25 November 2024].

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M. and Le, Q. V. (2022) Finetuned Language Models Are Zero-Shot Learners [Online]. arXiv. Available from: <http://arxiv.org/abs/2109.01652> [Accessed 26 November 2024].

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V. and Zhou, D. (2024) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*, *April 3, 2024*. Red Hook, NY, USA: Curran Associates Inc., pp. 24824–24837.

Zhang, C., Peng, B., Sun, X., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Liu, M., Zhang, Y., Fei, C., Yin, C. H., Yan, L. K. and Wang, T. (2024) From Word Vectors to Multimodal Embeddings: Techniques, Applications, and Future Directions For Large Language Models [Online]. arXiv. Available from: <http://arxiv.org/abs/2411.05036> [Accessed 1 December 2024].

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. V. and Chi, E. H. (2022) Least-to-Most Prompting Enables Complex Reasoning in Large Language Models [Online]. Available from: <https://openreview.net/forum?id=WZH7099tgfM> [Accessed 3 December 2024].