Tailored Vision-Language Solutions for Comprehensive Hazard Identification on Construction Sites

Qihua Chen[1] and Xianfei Yin[1,*]

[1]Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong SAR
*Corresponding Author
qihua.chen@cityu.edu.hk, xianfyin@cityu.edu.hk

Abstract –

The identification of hazards is critical for mitigating accident risks on construction sites. Although current automatic recognition methods based on computer vision have demonstrated significant success, their performance can be limited by inadequate generalization capabilities and challenges in handling dynamic, complex scenes, leading to incomplete detection of various types of construction hazards. This study introduces a novel approach that leverages a tailored vision-language model (VLM) to enhance comprehensive hazard identification on construction sites. By capitalizing on the cross-modal understanding and robust generalization abilities of VLMs, this method aims to improve both the accuracy and comprehensiveness of construction hazard identification. To validate this approach, we constructed a specialized dataset comprising 1,139 images across 31 fine-grained categories of construction hazards for fine-tuning and evaluation. The experimental results revealed that, when fine-tuned on the Qwen2-VL-7B model, the VLM achieved a precision of 0.856, representing a substantial improvement from 0.495 precision in the non-fine-tuned state. Moreover, fine-tuning significantly reduced prediction errors and enhanced the model's capability to understand and identify construction hazards. In addition, advanced VLM models demonstrate greater application potential compared to traditional models such as CLIP. This study advances the accuracy and comprehensiveness of hazard identification on construction sites, offering a new technological perspective for integrated automatic hazard detection in construction environments.

Keywords –

Construction management; Hazard identification; Vision-language model; Fine-tuning

1 Introduction

There is a wide variety of hazards on construction sites, such as workers not wearing safety helmets and electrical wires submerged in water, which constitute major factors leading to construction safety accidents [1,2]. To mitigate these risks, traditional construction safety management has relied on the Behavior-Based Safety (BBS) approach, centered around human inspections and timely corrections [3]. However, with the rapid advancement of artificial intelligence technology, especially in computer vision, and the development of advanced construction site monitoring systems utilizing cameras [4,5], automated methods for detecting safety hazards have begun to emerge. These methods, for example, checking compliance with personal protective equipment (PPE) [6], monitoring workers' access to hazardous areas [7], and ensuring the orderly stacking of construction materials [8], have demonstrated efficacy. Despite this progress, most existing methods are limited to identifying a small number of hazards and fail to consider multiple hazard factors in an integrated manner. Moreover, because model training depends on labeled data, it limits the models' generalization capabilities and hinders their adaptation to unseen, complex dynamic scenes [6]. Consequently, there remains a critical need for a method that can accurately and comprehensively recognize various types of construction hazards through advanced scene understanding and perception, particularly within the complex and dynamic construction environments.

With the development of pre-trained vision-language model (VLMs), such as CLIP [9]，BLIP [10]，LLaVA [11], significant progress has been made in understanding and generating textual descriptions associated with images. These models provide richer and more detailed insights by accumulating robust image and text feature representation capabilities through pre-training on large-scale datasets. Compared to traditional computer vision methods, VLMs exhibit higher robustness and generalization capabilities when handling complex scenes [12]. Additionally, the self-supervised learning mechanism of VLMs allows them to acquire knowledge from unlabeled data, thereby reducing reliance on manually labeled data [13].

Therefore, the application of VLMs in construction

management holds great promise. In past studies, Yong et al. [14] utilized BLIP to identify ergonomics issues and their solutions with the aid of data augmentation. The study by Liang et al. [15] demonstrated the potential of using CLIP for few-shot learning to classify construction site objects, achieving an increase in classification accuracy from 58.33% to 83.33% with only one sample for training. For construction safety management, Tsai et al. [12] automatically generated construction safety observations by fine-tuning CLIP, attaining an average accuracy of 73.7% across nine types of violations. However, the proposed hazard identification methods generate caption content that requires professionals to judge the construction hazard categories, rather than directly outputting violation categories from the input images. Moreover, construction sites present a variety of complex hazards, such as unguarded rotating parts of machinery, necessitating more specific and fine-grained hazard identification capabilities. Traditional VLMs, like CLIP , show limited advanced language comprehension and understanding of complex scenarios [16]. It needs to be combined with other language models to generate the corresponding textual content [12]. Therefore, research gaps still exist in the end-to-end construction hazard identification and reporting process, which requires sophisticated visual reasoning capabilities. This highlights the need to explore and apply more advanced VLMs, such as the pre-trained Qwen2-VL [17], to address these challenges.

Therefore, to address the limitations and achieve more specific, fine-grained identification of construction hazard categories, this study proposes an end-to-end construction safety hazard recognition framework based on a fine-tuned VLM. This framework leverages the advanced image comprehension and reasoning capabilities of pre-trained VLMs to identify construction site safety hazards. The main contributions of this paper are as follows: (1) constructing a dataset with 31 fine-grained categories of construction hazards for fine-tuning and evaluating the hazard identification model; (2) demonstrating the effectiveness of advanced VLMs and fine-tuning methods in construction hazard identification; and (3) implementing an automated end-to-end process for construction site hazard inspections.

2 Methodology

Figure 1 illustrates the proposed framework, which comprises three principal components. Firstly, dataset construction: this stage involves collecting and annotating a series of image data related to construction hazards, serving as the foundation for fine-tuning the VLM. Secondly, VLM fine-tuning process: we utilize an advanced open-source pre-trained VLM and apply advanced fine-tuning techniques to optimize the model using the constructed construction hazard dataset. The performance of the fine-tuned model is subsequently evaluated using a test set. Lastly, deployment for construction hazard identification: the validated and optimized model is deployed at construction sites, leveraging real-time image data streams from CCTV systems to enable immediate monitoring and identification of construction hazards. Detailed descriptions of these three critical stages are provided in the following sections.

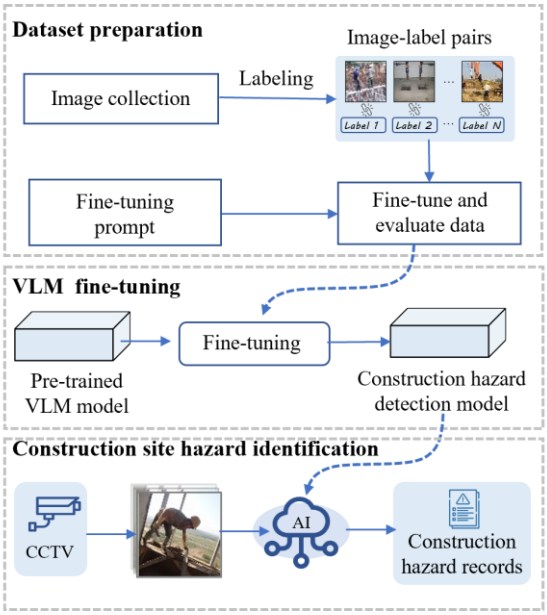

Figure 1. The framework of the proposed method

2.1 Dataset preparation

In BBS-based management, construction hazards encompass not only the use of workers' PPE but also other aspects, such as the utilization of construction tools and equipment. Based on the previous research by Chen et al. [3], Lee et al. [18] and practical experience from construction sites, this study has systematically identified common construction hazards, categorizing them into 31 fine-grained categories as presented in Table 1. Overall, Table 1 includes violations related to PPE wearing, the status of construction equipment, and the construction environment. Subsequently, based on the categories of construction safety hazards outlined in Table 1, we gathered images representing various types of construction hazards. The majority of these images were sourced from historical inspection records of construction projects. These hazard inspection logs contain images and detailed descriptions of the hazard categories. However, given the limited availability of historical inspection data, we supplemented images with publicly available datasets containing construction site images, such as CPPED [6] and image data collected

from the Internet. Subsequently, we conducted a filtering process for all images, which involved verifying their relevance to construction site scenarios and removing low-quality or blurry images. As a result, we assembled a dataset comprising 1,139 high-quality construction hazard images.

Table 1. 31 categories of construction hazards

No.	Construction hazards
H01	Not wearing safety helmets
H02	Wearing a safety helmet but not fastening the chin strap
H03	Not wearing safety vests
H04	Not wearing protective shoes
H05	Workers not wearing gloves during operations such as handling materials.
H06	Workers performing high-altitude work without wearing safety harnesses or improperly securing the harness hooks
H07	Illegally climbing over safety railings or barriers
H08	Debris piled on scaffolding
H09	Construction machinery illegally transporting or hoisting construction workers
H10	Workers smoking at the construction site
H11	Entering the operating range of an excavator
H12	Unauthorized use of open flames
H13	Rotating parts of machinery lack guards
H14	Crane operations without setting up safety warning lines around the crane
H15	Gas cylinders lacking protection
H16	Hooks missing safety pins
H17	Edge protection barriers not installed or having gaps
H18	Openings and shafts lack protective measures
H19	Guardrails not installed at open edges of staircases
H20	Distribution boxes left with doors open
H21	Distribution boxes not elevated (in contact with the ground)
H22	Electrical wires not elevated (laid directly on the ground) or poorly organized
H23	Unauthorized use of plug-in boards
H24	Electrical wires submerged in water
H25	Standing water present at the construction site
H26	Lack of clearance for construction waste
H27	Construction materials piled disorderly
H28	Temporary passageways at the construction site insufficiently stabilized or lacking safety measures
H29	Safety nets on exterior scaffolding damaged
H30	Dust prevention nets damaged
H31	Gaps existing in the perimeter fencing

According to the predefined categories, we categorized these images, ensuring that each image represented at least one category during the annotation process. The data annotation process is subject to standardized quality control measures, including cross-check inspections by annotators and final approval by experienced construction safety managers, to ensure the reliability of the dataset. Samples of hazard images and corresponding labels are illustrated in Figure 2.

In order to verify the performance of fine-tuned VLMs, a portion of the data from the dataset is extracted for testing, ensuring that the number of labels in each category within the test set is not less than 8. The stratified sampling method is used to divide the dataset into a test set and a training set in a 3:7 ratio. The final proportion of each label in both the training set and the test set is shown in Fig. 3. Subsequently, we generate the final datasets for both the training and test sets using designed VLM fine-tuning prompt as illustrated in Fig. 4

Figure 2. Examples of construction hazards and corresponding labels

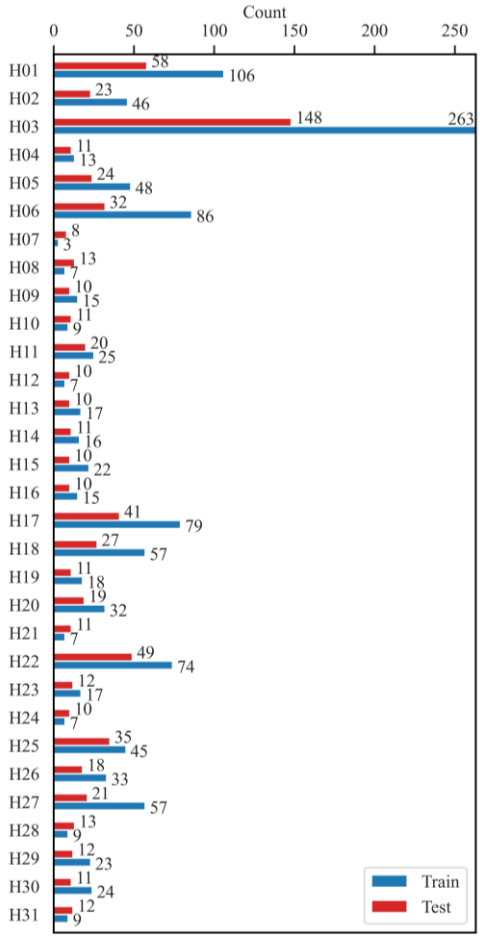

Figure 3. Number of labels in the test and train sets

User: You are an expert in identifying construction site hazards. Please list all categories of hazards present in the image.

ChatCH: The construction site hazard categories identified in the image are as follows: {*Hazard Categories Included*}

Figure 4. Prompt for VLM fine-tuning

2.2 VLM fine-tuning

Fine-tuning a VLM is a process that adapts a pre-trained VLM to a specific task or domain. Initially, the pre-trained model undergoes extensive training on a large-scale dataset, enabling it to learn general feature representations. Subsequently, during fine-tuning, the model leverages task-specific data to enhance its performance on specialized tasks and improve its generalization capabilities in targeted applications. Through this adaptation, the fine-tuned VLM can better address the unique challenges of specific domains. To enhance the model's performance in construction hazard identification, further fine-tuning is necessary to improve the VLM model's understanding of construction-specific scenarios. Low-rank adaptation (LoRA) [19] stands out as an effective fine-tuning strategy that addresses challenges posed by traditional computer vision methods, such as overfitting. Figure 5 compares the direct use of pre-trained weights with the LoRA-based fine-tuning process. As illustrated in Figure 5, LoRA's core principle involves updating only a select subset of parameters deemed most crucial for the new task. Specifically, LoRA achieves parameter updates by incorporating low-rank matrices into the pre-trained model, rather than directly altering a vast number of original parameters. This method not only reduces the number of parameters requiring training, thereby lowering computational expense, but also aids in preserving the pre-trained model's knowledge and mitigating catastrophic forgetting.

In LoRA, the original weight matrix W is updated to $W = W_0 + AB$, where W_0 denotes the pre-trained VLM weights, and AB represents the product of matrices A and B. Here, A and B are designed to be low-rank matrices with rank r, which allows for minimal changes to W_0 while enabling efficient fine-tuning. The result of this multiplication is a low-rank matrix that is added to the original weights to form the new weight matrix. During the fine-tuning process, the objective is to adjust these parameters by minimizing the loss function $L(\theta, A, B)$ on the target task, which is formulated as follows:

$$L(\theta, A, B) = \sum_i \ell(f(x_i; W_0 + AB), y_i), \quad (1)$$

where θ represents the other parameters, x_i denotes the i-th sample of the input data, and y_i denotes the true label of the i-th sample. By precisely tuning the model parameters, LoRA enables the VLM to more effectively understand and predict various safety hazards in construction sites, thereby providing robust technical support for safety management and risk prevention.

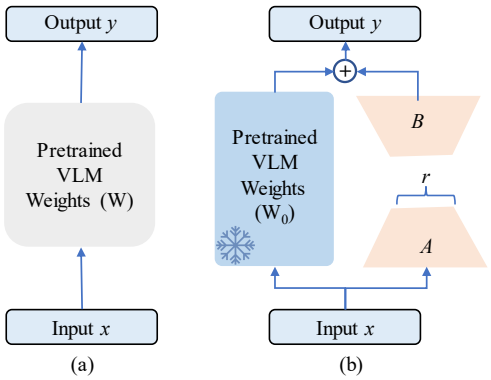

Figure 5. Comparison of two data processing workflows: (a) direct use of pre-trained weights, (b) LoRA-based fine-tuning process

2.3 Construction site hazard identification

The fine-tuned VLM is deployed on the server side, as illustrated in Fig. 1. The system interfaces with a CCTV monitoring system at the construction site to acquire live or stored video streams. To optimize computational efficiency and reduce processing load, key frames are selected at fixed time intervals. For each extracted frame, the VLM extracts rich visual features and semantic information relevant to the construction scene.

Based on these extracted features, the VLM classifies images to identify whether they contain categories of construction hazards that require inspection, as outlined in Table 1. Identified hazards, along with associated metadata such as location, type, and timestamp, are aggregated into a hazard identification dataset. Leveraging an LLM, the system can automatically generate a structured report on construction hazards based on this aggregated data. The report includes detailed textual descriptions, and relevant image evidence may be embedded to facilitate a more intuitive understanding for site managers of the conditions. Through this process, the fine-tuned VLM significantly contributes to real-time monitoring and alerting of potential hazards on the construction site, thereby enhancing the smart level of safety management.

3 Experiment and Result

3.1 Experiment setup

Based on the constructed construction hazard dataset, this study employs the LoRA technique to fine-tune the

VLM. Specifically, two different parameter-size versions of Qwen2-VL are selected for this purpose: a 2-billion-parameter model and a 7-billion-parameter model. Specialized fine-tuning and validation experiments are conducted for both versions. During the testing phase, Qwen2.5-32B LLM is utilized to assist in evaluating the match between the fine-tuned VLM predictions and the actual labels. This evaluation ensures that the VLM's performance is accurately assessed and optimized for real-world applications.

To establish an effective benchmark for comparison, the CLIP model with ViT-B/32 as the image encoder is used as a baseline for classifying construction hazard images. The process involves fine-tuning the CLIP model using image-label pairs from the training dataset. During the validation phase, similarity scores are calculated between the images in the test set and all possible hazard labels. If the similarity score exceeds a predefined threshold d, the corresponding label is considered a prediction, thereby evaluating model performance. This approach aligns with the methodology proposed by Chen et al. [20].

For evaluating the models' performance, a series of widely accepted evaluation metrics are applied. Precision (the proportion of correctly identified positive samples among those predicted as positive), recall (the proportion of actual positives which are correctly identified as such), and F1-score (the harmonic mean of precision and recall) are calculated for each category using the following formulas:

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{4}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives. Additionally, the F2 score (β = 2) is introduced to emphasize recall, focusing more on minimizing missed identifications. It is calculated using the following formula:

$$F2\ score = (1+\beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}. \tag{5}$$

This is critical given the serious consequences of safety accidents at construction sites. The F2 score places greater weight on recall compared to precision, ensuring that fewer potential hazards are overlooked. Moreover, Cohen's Kappa coefficient is used to measure the agreement between observed and expected consistency, applying the formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{6}$$

where p_o is the observed proportion of agreement and p_e is the expected probability of random agreement [21]. This coefficient verifies whether the model predictions significantly outperform random guessing, thereby increasing confidence in the model's reliability. A Cohen's Kappa coefficient value close to 1 indicates almost perfect agreement, suggesting that the model significantly outperforms random chance. In contrast, values near or below 0 indicate agreement no better than, or even worse than, what would be expected by random chance, highlighting potential limitations in the model's ability.

3.2 Result

As illustrated in Figure 6, the performance of the fine-tuned Qwen2-VL-2B and Qwen2-VL-7B models in identifying construction hazards as a function of Epoch is depicted. The figure demonstrates that, at the outset (Epoch = 0), prior to fine-tuning, both models exhibit lower precision, recall, and F1 scores. As the number of Epochs increases, there is a marked improvement in model performance. Specifically, the precision for both the Qwen2-VL-7B and Qwen2-VL-2B models rises from approximately 0.4 to around 0.8 following fine-tuning. The comparison with their no fine-tuning performance can be seen in Table 2. Notably, the fine-tuned Qwen2-VL-7B model achieves a precision of 0.856 by the 9th Epoch, while the fine-tuned Qwen2-VL-2B model reaches a precision of 0.818 by the 27th Epoch. Similarly, recall values for both models increase from about 0.2 to approximately 0.6, and F1 scores rise from roughly 0.2 to nearly 0.7. These observations indicate that fine-tuning significantly enhances the models' ability to identify construction hazards. Furthermore, it is evident from the figure that, in their initial state, the Qwen2-VL-7B model in recall and F1 score outperform the Qwen2-VL-2B model. However, as the number of fine-tuning Epochs increases, the performance gap between the two models narrows, suggesting that with sufficient fine-tuning, the smaller model can achieve performance comparable to that of the larger model. This finding underscores the importance of fine-tuning strategies in optimizing model performance, highlighting that even smaller-scale models can attain significant improvements when subjected to effective fine-tuning procedures.

The loss curve depicted in Figure 7 illustrates the trend of the loss value as a function of epochs during the training process of the CLIP model. Initially, the loss value is relatively high but gradually decreases over time before eventually plateauing. Meanwhile, the recall curve remains close to 1 for most of the training period, indicating near-perfect identification of positive samples

by the model. However, it exhibits a decline in later stages, suggesting that the model's ability to accurately identify positive samples diminishes over extended training. Despite undergoing fine-tuning, the CLIP model does not demonstrate significant improvements in precision or F1 score, as shown in Table 2. The consistently low precision indicates that the model struggles to effectively differentiate construction hazards, while the low F1 score reflects an overall deficiency in balancing precision and recall. This suggests that the model fails to fully comprehend and accurately distinguish the various features associated with construction hazards. Therefore, when applied to the task of recognizing construction hazards, the pre-trained CLIP model, despite achieving high recall, exhibits limitations in enhancing precision and the F1 score.

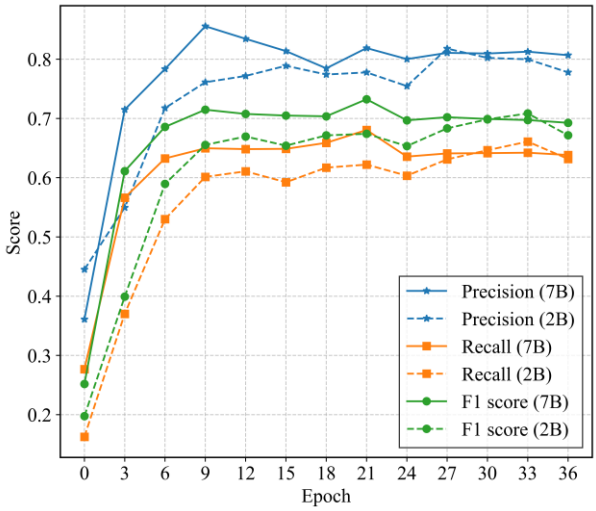

Figure 6. Comparison of Qwen2-VL-2B and 7B's construction hazard identification performance as a function of Epoch

Table 2 also presents the F2 scores and Cohen's Kappa coefficients for various models. The table compares two primary models: CLIP (ViT-B/32), and Qwen2-VL in both 2B and 7B, each evaluated in terms of their pre-trained and fine-tuned performance on the test dataset. For the CLIP model using the ViT-B/32 as the image encoder, both the F2 score and Cohen's Kappa coefficient remain nearly unchanged, even after fine-tuning, indicating that the fine-tuning process did not yield significant improvements for this model. In contrast, the Qwen2-VL models exhibit substantial gains following fine-tuning. Initially, the Qwen2-VL-2B has an F2 score of 0.186 and a Cohen's Kappa of 0.118, which significantly improves to an F2 score of 0.661 and a Cohen's Kappa of 0.613 after fine-tuning. Similarly, the Qwen2-VL-7B starts with an F2 score of 0.290 and a Cohen's Kappa coefficient of 0.194, which then increases to an F2 score of 0.683 and a Cohen's Kappa of 0.600 after fine-tuning, with respective increases of 0.393 and 0.406. These results underscore the effectiveness of fine-tuning in enhancing the performance of the Qwen2-VL models, particularly in terms of improving both precision-recall balance.

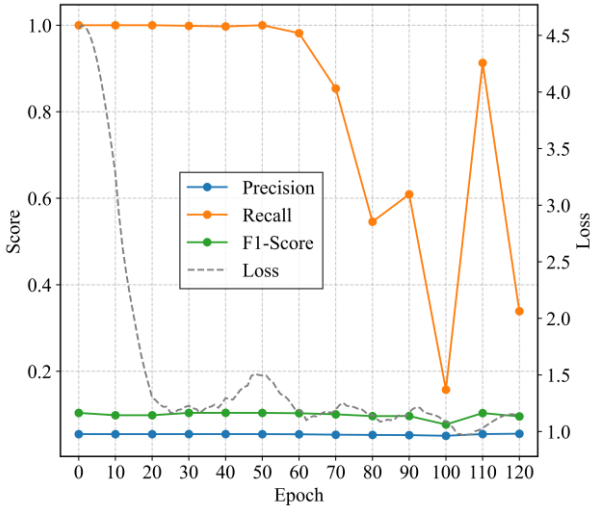

Figure 7. Performance and loss curves for construction hazard identification based on fine-tuned CLIP as a function of Epoch

Table 2. Performance comparison of different models before and after fine-tuning

Models		Precision	Recall	F1 score	F2 score	Cohen's Kappa coefficient
CLIP	ViT-B/32	0.055	**1.0**	0.104	0.225	0
	ViT-B/32 (Fine-tuned)	0.055 (+0.0)	0.999(-0.001)	0.104(+0.0)	0.225(+0.0)	0(+0.0)
Qwen 2-VL	2B	0.445	0.163	0.198	0.186	0.118
	2B (Fine-tuned)	0.818 (+0.373)	0.631 (+0.468)	0.683 (+0.485)	0.661 (+0.475)	**0.613** (0.495)
	7B	0.361	0.277	0.252	0.290	0.194
	7B (Fine-tuned)	**0.856** (+0.495)	0.650 (+0.373)	**0.715** (+0.463)	**0.683** (+0.393)	0.600 (+0.406)

Note: Values in parentheses indicate the change after model fine-tuning compared to before.

The effectiveness of the proposed method in identifying construction safety hazards is illustrated using real-world images extracted from an actual construction site video, with one of the extracted images shown in Figure 8. The selected construction site is a stockpile area, where our approach demonstrates its capability to accurately detect potential construction hazards present in the images and generate corresponding hazard reports for further documentation. This application highlights the practical utility of our method in a real-world setting, showcasing its ability to provide reliable hazard identification. Such improvements are crucial for preventing accidents and ensuring worker safety, thereby underscoring the significant impact and value of this method in advancing construction site management.

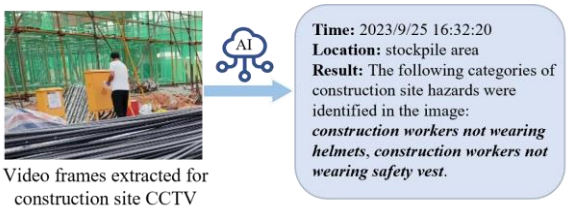

Figure 8. A case of identifying construction hazards

4 Discussion and Conclusion

This study introduces a fine-tuned VLM designed to comprehensively identify safety hazards on construction sites. Specifically, our work encompasses the following aspects: (1) we developed a specialized dataset for construction hazards, classifying 31 types of common hazards at a fine-grained level. (2) we demonstrated the effectiveness of using pre-trained VLMs and fine-tuning techniques for identifying construction hazards. Additionally, we compared the performance differences among various pre-trained VLMs in the context of construction hazard identification. (3) we implemented an automated end-to-end process for construction site hazard inspections, ranging from image acquisition to report generation, thereby enhancing the automation level of construction management.

The experimental results indicate that the fine-tuning process significantly enhances the VLM model's ability to handle complex and dynamic scenarios, improving its generalization capability and robustness. Through enhanced cross-modal understanding, the precision of the fine-tuned Qwen2-VL-7B reaches 0.856, marking a substantial improvement from 0.495 precision in the non-fine-tuned state. The use case of real-world site videos further demonstrates its future application potential in construction management, such as being deployed on robotic dogs or drones to facilitate the development of automated construction hazard inspections.

Moreover, the limitations of traditional VLMs in comprehending construction scenarios underscore the necessity for advanced VLM modeling applications. Our study also revealed that even VLM models with fewer parameters can achieve recognition results comparable to those of larger VLM models when appropriately fine-tuned. This insight suggests that future research should focus on developing specialized, compact models to achieve high-precision construction hazard recognition while using minimal computational resources.

Although the proposed method has shown promising performance, there are two limitations. First, we conducted a preliminary application on a real construction site to assess its usability. However, this case study was limited in scale and environmental control, warranting future systematic validation. Moreover, the study not analyse the model's performance for each category, further investigation, particularly in exploring the model's robustness, few-shot learning capabilities, and speed comparisons will conduct in future work.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 72404233), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515010190) and the New Faculty Start-up Grant from the City University of Hong Kong (Project No. 9610701).

The authors also thank Kaisheng Liu, a construction safety manager, for his assistance in processing and categorizing the raw construction hazard images.

References

[1] B. Choi, S. Lee. An empirically based agent-based model of the sociocognitive process of construction workers' safety behavior. *Journal of Construction Engineering and Management*, 144 (2): 04017102, 2018. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001421

[2] B. Yuan, S. Xu, M. Niu, K. Guo. Will improved safety attitudes necessarily curb unsafe behavior? Hybrid method based on NCA and SEM. *Journal of Environmental and Public Health*, 2022 (1): 9271690, 2022. https://doi.org/10.1155/2022/9271690

[3] Q. Chen, D. Long, C. Yang, H. Xu. Knowledge graph improved dynamic risk analysis method for behavior-based safety management on a construction site. *Journal of Management in Engineering*, 39 (4): 04023023, 2023. https://doi.org/10.1061/JMENEA.MEENG-5306

[4] B. Xiao, Y. Zhang, Y. Chen, X. Yin. A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation, *Advanced Engineering Informatics*. 50 101372, 2021. https://doi.org/10.1016/j.aei.2021.101372.

[5] B. Xiao, X. Yin, S.-C. Kang. Vision-based method of automatically detecting construction video highlights by integrating machine tracking and CNN feature extraction. *Automation in Construction*, 129 103817, 2021. https://doi.org/10.1016/j.autcon.2021.103817.

[6] Q. Chen, D. Long, S. Wang, Q. Chen, B. Yuan. Real-Time Detection of Personal Protective Equipment Violations for Construction Workers Using Semi-Supervised Learning and Video Clips. *Journal of Construction Engineering and Management*, 2025. https://doi.org/10.1061/JCEMD4/COENG-15310.

[7] X. Mei, X. Zhou, F. Xu, Z. Zhang. Human intrusion detection in static hazardous areas at construction sites: Deep learning–based method. *Journal of Construction Engineering and Management*, 149 (1): 04022142, 2023. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002409

[8] Y.G. Lim, J. Wu, Y.M. Goh, J. Tian, V. Gan. Automated classification of "cluttered" construction housekeeping images through supervised and self-supervised feature representation learning. *Automation in Construction*, 156 105095, 2023. https://doi.org/10.1016/j.autcon.2023.105095.

[9] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark. Learning transferable visual models from natural language supervision. International conference on machine learning, PMLR, 2021, pp. 8748-8763. https://proceedings.mlr.press/v139/radford21a.

[10] J. Li, D. Li, C. Xiong, S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International conference on machine learning, PMLR, 2022, pp. 12888-12900. https://proceedings.mlr.press/v162/li22n.html.

[11] H. Liu, C. Li, Q. Wu, Y.J. Lee. Visual instruction tuning, *Advances in neural information processing systems*. 36, 2024.

[12] W.-L. Tsai, P.-L. Le, W.-F. Ho, N.-W. Chi, J.J. Lin, S. Tang, S.-H.J.A.i.C. Hsieh. Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention. *Automation in Construction*, 169 105863, 2025. https://doi.org/10.1016/j.autcon.2024.105863

[13] D. Gil, G. Lee. Zero-shot monitoring of construction workers' personal protective equipment based on image captioning. *Automation in Construction*, 164 105470, 2024. https://doi.org/10.1016/j.autcon.2024.105470.

[14] G. Yong, M. Liu, S. Lee. Explainable Image Captioning to Identify Ergonomic Problems and Solutions for Construction Workers. *Journal of Computing in Civil Engineering*, 38 (4): 04024022, 2024. https://doi.org/10.1061/JCCEE5.CPENG-5744

[15] Y. Liang, P. Vadakkepat, D.K.H. Chua, S. Wang, Z. Li, S. Zhang. Recognizing temporary construction site objects using CLIP-based few-shot learning and multi-modal prototypes. *Automation in Construction*, 165 105542, 2024. https://doi.org/10.1016/j.autcon.2024.105542.

[16] Y. Ouali, A. Bulat, A. Xenos, A. Zaganidis, I.M. Metaxas, G. Tzimiropoulos, B.J.a.p.a. Martinez. Discriminative Fine-tuning of LVLMs. *arXiv preprint arXiv:2412.04378*, 2024. https://doi.org/10.48550/arXiv.2412.04378

[17] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:.12191*, 2024. https://doi.org/10.48550/arXiv.2409.12191

[18] P.-C. Lee, J. Wei, H.-I. Ting, T.-P. Lo, D. Long, L.-M. Chang. Dynamic analysis of construction safety risk and visual tracking of key factors based on behavior-based safety and building information modeling. *KSCE Journal of Civil Engineering*. 23 (10): 4155-4167, 2019. https://doi.org/10.1007/s12205-019-0283-z.

[19] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:.09685*, 2021. https://arxiv.org/pdf/2106.09685v1/1000

[20] Z. Chen, H. Chen, M. Imani, R. Chen, F. Imani. Vision language model for interpretable and fine-grained detection of safety compliance in diverse workplaces. *Expert Systems with Applications*. 265 125769, 2025. https://doi.org/10.1016/j.eswa.2024.125769.

[21] M. Ferdous, S.M.M. Ahsan. PPE detector: a YOLO-based architecture to detect personal protective equipment (PPE) for construction sites. *PeerJ Computer Science*, 8 e999, 2022. https://doi.org/10.7717/peerj-cs.999.