Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.39455

Multimodal Large Language Models for Construction Safety Inspection

Yinan Zhu

Department of Electrical and Computer Engineering, University of California, Davis, USA
zynzhu@ucdavis.edu

Abstract. Multimodal Large Language Models (MLLMs) have drawn more research interest in construction safety inspection recently due to their capacity to integrate visual understanding, language understanding, and reasoning. In this paper, a survey is conducted to review the recent research on MLLMs in construction safety inspection, covering techniques, datasets, evaluation, applications, and limitations. The current studies primarily utilize prompt-based adaptation and in-context learning, domain-specific fine-tuning, and knowledge-augmented reasoning to enhance the performance of MLLMs in the construction domain. Applications of MLLMs in construction safety inspection include hazard detection and localization, safety inspection, visual question answering, assistance, and report generation. From the current studies, it can be found that MLLMs can be used to perform flexible and interpretable analysis for construction safety inspection with the assistance of domain knowledge and prompts. Nevertheless, the limitations of MLLMs in construction safety inspection are still present, such as poor visual grounding and spatial reasoning, lack of domain knowledge, lack of large and balanced datasets, and uncertainty of application in real-world scenarios. For the future work, it is recommended to enhance multimodal grounding, construct a larger and more balanced dataset in the construction domain, integrate safety regulations in a trustworthy manner, and perform human validation systematically. MLLMs have great potential to be applied in construction safety inspection; however, they can be used as decision aids rather than replacements for safety inspectors at present.

Keywords: MLLM, construction safety inspection, hazard identification, human oversight

1. Introduction

Safety inspection is a necessary part in the construction field. It requires an understanding of both site conditions and applicable safety rules. Inspectors must identify safety hazards, assess observed activities with relevant rules, and report findings to communicate corresponding actions. These required tasks can be challenging, especially in complicated and always changing construction environments. Manual inspections mainly depend on inspectors' personal experience, attention and judgement, while many existing computer vision approaches focus on predefined hazard categories and lack flexibility. Although the two approaches mentioned support automated monitoring,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.39455

interpreting broader scene context and connecting visual observations with safety requirements remain important challenges [1].

Multimodal large language models (MLLMs) can combine visual information with language understanding and generation [2], including text, images and videos to support scene understanding and language generation to help address these challenges. Recent studies have explored MLLM use in fields including identifying hazards, assessing compliance, answering questions about site images, and generating inspection reports [1, 3-5]. Researchers have also investigated prompt engineering, domain-specific fine-tuning, and external knowledge integration to improve performance under construction setting circumstances [3, 6-8]. The key findings suggest that MLLMs are able to support different stages of the construction safety inspection process, which includes identifying and locating hazards, assessing compliance with safety regulations, answering questions about site conditions, and generating inspection reports and corrective recommendations.

An existing review has examined MLLM applications across the construction industry. It includes adaptation strategies and evaluation practices [9]. Building on this broader research, a highly focused review of construction safety inspection will help clarify how technical approaches including MLLMs can relate to specific safety tasks and how their effectiveness is assessed. The studies and research considered in this review are different in their task definitions, data sources, model configurations and evaluation baselines. The reported performance needs to be interpreted in relation to task and experimental setting. To be specific, hazard identification, compliance assessment, visual question answering, and report generation require different forms of evidence to build their usefulness. However, their ability to produce detailed explanations does not guarantee the observations of the construction site or safety judgements are correct. Therefore, MLLMs' reliability needs to be carefully examined [10].

This review examines MLLM applications in construction safety inspection through three perspectives: safety inspection application tasks, implementation with corresponding evaluation, and remaining challenges. First, it organizes existing studies around hazard identification and localization, safety compliance assessment, visual question answering and interactive assistance, safety reporting and corrective recommendations. Second, it compares methods, dataset characteristic, and task-specific evaluations. Third, it discusses limitations in visual and spatial understanding, domain knowledge and reasoning reliability, data quality and verification, practical deployment and human oversight. Throughout this analysis, the review aims to clarify current capabilities, identify gaps in the supporting evidence, and outline directions for developing more reliable tools to assist construction safety inspectors.

2. Cross-study comparison: methods, data and evaluation

2.1. Data

The reviewed studies primarily use construction site images and text information. Site images provide visual information about worker conditions, safety gear, surrounding environment and equipment status, while text includes user questions, safety regulations and hazard descriptions [1, 5]. In addition, some systems obtain visual inputs by extracting frames at a fixed time from construction site videos [6]. Approaches such as VS-BIM, which incorporate panoramic video and building information models, are also used to better support spatial understanding [11].

The datasets collected in these studies are different in their sources and task-oriented annotation formats. ConstructionSite10k builds on previously published construction site images and contains 10,013 images, in which approximately 2,000 images were manually annotated while the remaining

DOI: 10.54254/2755-2721/2027.39455

annotations were first drafted with the assistance of GPT-4V and then revised by human annotators. All annotations were reviewed by two domain experts in the end. The annotations mainly focused on three tasks: image captioning, safety violation question answering, and visual grounding [5]. Construction Hazard Dataset (CHD) combines historical inspection images with published datasets' images and other online resources. It contains 1,308 images across 32 common and specific construction hazard categories. The images were paired with hazard labels, which were checked through cross-review among annotators and final review by construction managers. The resulting image-text pairs were used to adjust and evaluate the construction hazard identification model, ChatCH [3]. ConHazard-VQA collects images from 12 construction projects in China, online resources and Site Object Detection Dataset (SODA). It contains 226 images and 1,034 image-question-answer pairs with 1,335 annotations. The images and annotations cover binary classification, counting, and open-ended questions. Reference answers were provided through independent expert annotation and cross-validation [5].

Overall, different aspects of multimodal construction safety analysis were considered in existing datasets. However, most current datasets primarily rely on static construction images combined with textual annotations. Other modalities, such as video, audio and sensor data, are still rarely concluded, which indicates a need for larger and more diverse multimodal datasets for real-world construction safety applications.

2.2. Method

Existing studies have adopted different methods to adapt MLLMs to construction safety inspection field. The approaches mainly focus on improving task instructions, incorporating construction knowledge and enhancing the models' ability to interpret and annotate visual information. Existing methods can be grouped into three categories: prompt-based adaptation and in-context learning, domain-specific fine-tuning and knowledge-augmented reasoning.

2.2.1. Prompt-based adaptation and in-context learning

Prompt-based adaptation and in-context learning are used for adapting general purpose MLLMs to specific construction safety tasks without changing the underlying model parameters. Existing studies mainly apply zero-shot, few-shot, chain-of-thought and structured prompting strategies to improve hazard recognition and safety inspection reasoning. For instance, Chan and Zou evaluated large, pre-trained vision language models under zero-shot and five-shot settings for safety rule violation visual question answering, image captioning and visual grounding. It demonstrates the potential of in-text examples to guide models toward specific construction safety inspection tasks without additional training [10]. Similarly, Chaudhary et al. compared zero-shot, few-shot and chain-of-thought prompting across several MLLMs for hazard identification in real-world construction site images, showing that model performance varied substantially with prompt design [8]. More structured prompt designs have been explored. Chan et al. incorporated construction safety ontology knowledge into a hierarchical prompting framework to support more domain-specific tasks, while Guo et al. combined a vision-language model with a bi-stage retrieval-augmented generation framework and developed a TriPhased Prompt and a decision-tree-based compliance judgement prompt to improve the interpretation of worker behavior and safety compliance [6, 7].

DOI: 10.54254/2755-2721/2027.39455

2.2.2. Domain-specific fine-tuning

Domain-specific fine-tuning is another major strategy used for adapting general-purpose MLLMs to construction safety inspection, as shown in Figure 1. Different from prompt-based methods, fine-tuning modifies model parameters by construction-specific image-text pairs or instruction-following data. This allows the model to learn domain terminology, hazard patterns and task requirements more directly. In the ChatCH framework, a pretrained Qwen2-VL model was adapted using construction hazard image-text pairs through parameter-efficient fine-tuning (PEFT), enabling more specialized hazard identification while retaining the general capabilities of the original model [3]. Other studies use staged fine-tuning strategies or combine fine-tuned multimodal models with domain knowledge for safety inspection. Fine-tuning method provides stronger task alignment compared with prompt-only adaptation, but additional labeled data and computational resources are required.

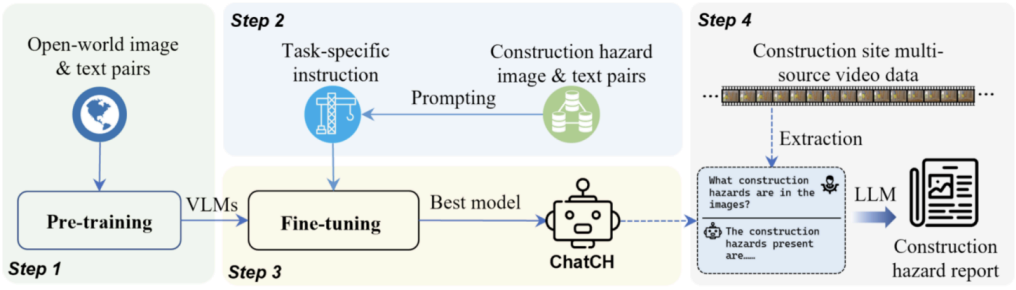

Figure 1. Framework of ChatCH for construction hazard identification and report generation [3]

2.2.3. Knowledge-augmented reasoning

Knowledge-augmented reasoning incorporates external construction safety knowledge to improve the performance of MLLMs. Liu et al. developed a retrieval-augmented generation framework along with a construction occupational health and safety knowledge base for hazard-related visual question answering [5]. Guo et al. similarly proposed a bi-stage retrieval-augmented generation framework to retrieve relevant safety regulations from construction images before conducting compliance analysis [7]. Wang et al. combined a domain knowledge base with a multimodal LLM and multi-step reasoning engine to support interpretation of evolving safety standards [1]. These approaches can reduce the reliance on the model's internal knowledge and help connect visual observations with explicit safety regulations.

2.3. Evaluation

Evaluation methods vary among studies according to the specific safety inspection tasks. For hazard identification, quantitative metrics such as precision, recall, F1-score and accuracy are widely used [8]. Precision measures the proportion of predicted hazards that are actually correct, reflecting the model's ability to avoid false alarms. Recall measures the proportion of actual hazards that are successfully identified, indicating how effectively the model avoids missed detections. F1-score combines precision and recall into a single metric and is especially useful when both false positives and false negatives are important. Accuracy measures the overall proportion of correct predictions among all samples and provides a general indication of classification performance. VQA and visual grounding studies further evaluate answer correctness, reasoning quality and localization

DOI: 10.54254/2755-2721/2027.39455

performance by metrics such as Intersection over Union (IoU) [10]. For more complicated safety reasoning tasks, some studies also assess response relevance and faithfulness to domain knowledge or safety regulations [1]. Additionally, certain studies compared different MLLMs, prompting strategies and retrieval frameworks against baseline methods to evaluate relative performance [5].

3. Applications across construction safety inspection tasks

3.1. Hazard identification and localization

Hazard identification and localization is one of the most important and common applications of MLLMs in construction safety inspection. Existing studies use construction images and video frames to identify unsafe conditions, hazardous objects and worker behaviors. Some systems also attempt to localize in-scene detected hazards. For example, ChatCH was developed to recognize multiple categories of construction hazards from site images and support automated hazard reporting [3]. Chen and Zou's studies evaluate large pre-trained VLMs on safety violation detection and visual grounding tasks, in which models are required to identify the existence of a hazard and its location [10]. Other comparative studies further show that prompting strategy can influence hazard recognition performance, indicating that model capability depends on both visual understanding and task formulation [8].

3.2. Safety compliance assessment

Safety compliance assessment differs from hazard identification, in that it extends hazard detection by evaluating whether observed site conditions and worker behaviors comply with safety regulations. Analyzed studies combined visual understanding with explicit safety rules to support this application. Guo et al., for example, proposed a visual-LLM framework that retrieves relevant regulations and evaluates construction site images through structured compliance prompts [7]. Similarly, Wang et al. integrated a MLLM with a domain knowledge base and multi-step reasoning to assess site conditions according to evolving safety standards [1]. These studies show that compliance assessment requires not only recognizing visual hazards but also connecting those observations with domain-specific rules and regulatory knowledge.

3.3. Visual question answering and interactive assistance

Visual question answering and interactive assistance provide MLLM users with a more flexible way to interact with construction safety information. Instead of only producing predefined hazard labels, MLLMs can answer users' questions about construction environment, safety violations and potential risks based on visual inputs. Liu et al. developed a construction OHS VQA framework that combines site images with retrieved safety knowledge to answer classification, counting and open-ended questions [5]. Similarly, ConstructionSite10k evaluates large pre-trained VLMs on safety-rule violation VPQ, showing their potential to support question-based inspection tasks [10]. Such interactive systems can allow inspectors or site managers to obtain more targeted explanations and safety information from construction scenes.

DOI: 10.54254/2755-2721/2027.39455

3.4. Safety reporting and corrective recommendations

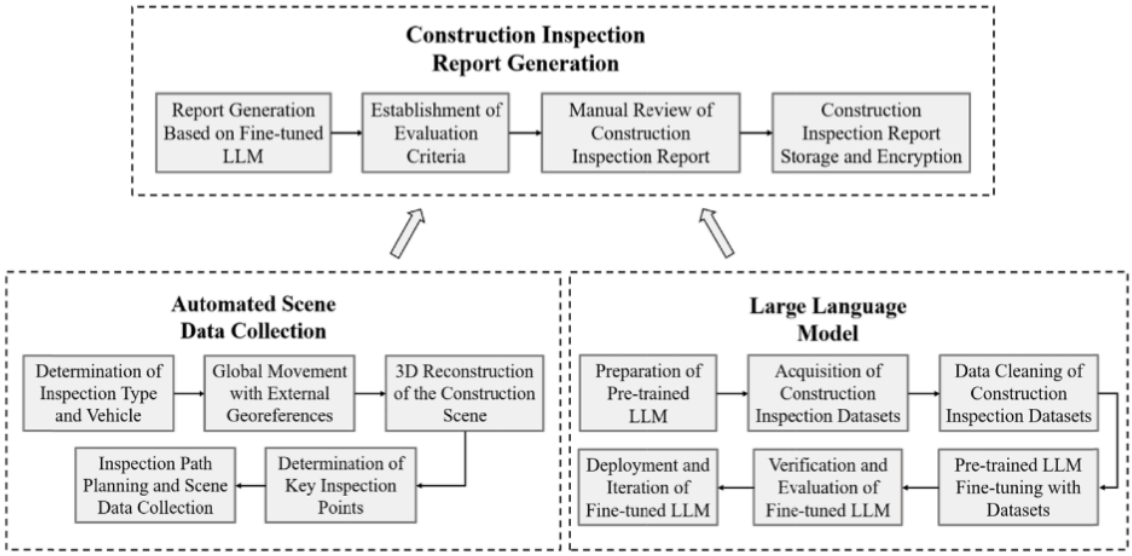

Figure 2. Framework of AutoRepo for automated construction inspection reporting [4]

Safety reporting and corrective recommendations is a further application of MLLMs beyond hazard detection and compliance assessment. As presented in Figure 2, AutoRepo combines scene information collected by unmanned vehicles with a MLLM, aiming to automatically generate standardized construction inspection reports [4]. ChatCH also supports end-to-end construction hazard report generation by converting detected hazards from site images or video frames into structured textual reports through prompt engineering [3]. These applications demonstrate the abilities of MLLMs to transform visual inspection results into more interpretable and communicable outputs and reduce the manual effort required for documentation and follow-up actions.

4. Challenges and future directions

4.1. Visual and spatial understanding

Visual and spatial understanding remains a major challenge for MLLMs in construction safety inspection. Construction sites usually contain complex object relationships, changing viewpoints, spatial effects and various worker activities, all of which can make it difficult for models to correctly interpret spatial context from individual images. Chen and Zou's studies have shown that VLMs still struggle with precise visual grounding and localization of safety violations, as performance generally remains limited in evaluations [10]. Recent approaches such as VS-BIM attempt to address this problem by combining panoramic videos, building information modeling (BIM) and 3D cognitive maps to provide stronger spatial and spatiotemporal context for MLLMs, and the detailed process is shown in Figure 3 [11]. Future research may further improve multimodal scene modeling and robustness under challenging visual conditions, including low-light conditions, occlusion and limited image input.

DOI: 10.54254/2755-2721/2027.39455

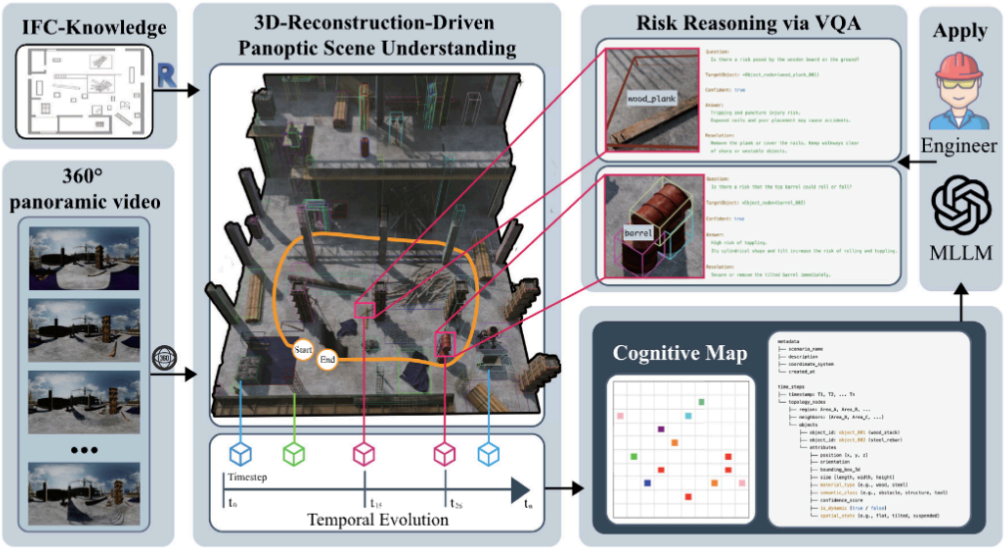

Figure 3. VS-BIM framework for spatial reasoning in construction safety inspection [11]

4.2. Domain knowledge and reasoning reliability

Domain knowledge and reasoning reliability are critical challenges for MLLMs in construction safety inspection. General-purpose models are primarily trained on broad datasets covering various aspects and may lack sufficient understanding of construction-specific terminology, hazard categories, real-life practices and evolving safety regulations [1]. Existing studies introduce domain knowledge bases, safety ontologies and retrieval mechanisms to improve faithfulness and support more reliable reasoning [1, 6]. However, accurate reasoning still depends largely on whether the model is able to retrieve and interpret relevant knowledge correctly. Future research therefore should focus on stronger grounding between visual evidence and regulatory knowledge, as well as more transparent reasoning processes that can be inspected by human professionals.

4.3. Data quality and verification

Data quality and verification remain important limitations for MLLM-based construction safety inspection. Construction datasets often include imbalanced hazard categories, limited real-site coverage and variations in image quality, which may reduce model generalizability [5]. Recent datasets, such as ConstructionSite10k and ConHazard-VQA provide broader annotations and more diverse task settings, but accurate labeling and reliable sampling are still essential for meaningful evaluation [5, 10]. Future studies should expand real-world datasets, improve annotation consistency and strengthen human verification of both training data and model-generated safety judgements.

4.4. Practical deployment and human oversight

Practical deployment and human oversight are essential considerations for MLLM-based safety inspection. Several studies have demonstrated real-world deployment, including automated report generation with AutoRepo and VLM-based safety monitoring on active construction sites [4, 6]. However, field conditions vary much more than controlled experiments and model errors can directly affect safety decisions. Therefore, MLLMs should be appropriately used as decision-support tools instead of fully autonomous inspectors. Future systems must include complete human

DOI: 10.54254/2755-2721/2027.39455

verification standards, clear responsibility for final safety judgements and continued testing under diverse real-life on-site conditions.

5. Conclusion

Multimodal large language models have promising capabilities of aiding safety inspection in construction through visual and language reasoning. They have been applied to hazard identification and localization, safety compliance judgement, visual question answering, and safety report generation. Prompt-based adaptation, domain fine-tuning, and knowledge-augmented reasoning are commonly used to adapt multimodal large language models to construction tasks. However, there are still challenges in visual and spatial understanding, domain reasoning, data, and real-world usage. Specifically, for safety inspection, models need to not only detect hazards but also understand spatial relationships and safety regulations to provide human safety inspectors with reference. Hence, future work should aim at improving multimodal grounding, real-world datasets, domain knowledge integration, and human validation. Multimodal large language models are more probably used as aids for human safety inspectors rather than replacing human safety judgment.

References

[1] Wang, Y., Luo, H., & Fang, W. (2025). An integrated approach for automatic safety inspection in construction: Domain knowledge with multimodal large language model. *Advanced Engineering Informatics, 65*, 103246. https://doi.org/10.1016/j.aei.2025.103246

[2] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review, 11*(12), nwae403. https://doi.org/10.1093/nsr/nwae403

[3] Chen, Q., & Yin, X. (2025). Tailored vision-language framework for automated hazard identification and report generation in construction sites. *Advanced Engineering Informatics, 66*, 103478. https://doi.org/10.1016/j.aei.2025.103478

[4] Pu, H., Yang, X., Li, J., & Guo, R. (2024). AutoRepo: A general framework for multimodal LLM-based automated construction reporting. *Expert Systems with Applications, 255*, 124601. https://doi.org/10.1016/j.eswa.2024.124601

[5] Liu, Y., Li, L., Su, X., Lin, H., Zhang, L., Guan, D., & Cai, H. (2026). Retrieval-augmented multimodal large language models for visual question answering of construction occupational health and safety hazards. *Journal of Management in Engineering, 42*(6), 04026044. https://doi.org/10.1061/JMENEA.MEENG-7452

[6] Chan, C. F., Wong, P. K.-Y., Guo, X., Cheng, J. C. P., Chan, J. P. C., Leung, P.-H., & Tao, X. (2025). Context-aware vision-language model agent enriched with domain-specific ontology for construction site safety monitoring. *Automation in Construction, 177*, 106305. https://doi.org/10.1016/j.autcon.2025.106305

[7] Guo, K. X., Wong, P. K.-Y., Cheng, J. C. P., Chan, C.-F., Leung, P.-H., & Tao, X. (2025). Enhancing visual-LLM for construction site safety compliance via prompt engineering and bi-stage retrieval-augmented generation. *Automation in Construction, 179*, 106490. https://doi.org/10.1016/j.autcon.2025.106490

[8] Chaudhary, N., Uddin, S. M. J., Chandra, S. S., Ovid, A., & Albert, A. (2026). Prompt to protection: A comparative study of multimodal LLMs in construction hazard recognition. *IEEE Access, 14*, 70565–70580. https://doi.org/10.1109/ACCESS.2026.3691685

[9] Erfani, A., & Mansouri, A. (2026). Applications of multimodal large language models in construction industry. *Advanced Engineering Informatics, 69*, 103909. https://doi.org/10.1016/j.aei.2025.103909

[10] Chen, X., & Zou, Z. (2026). Are large pre-trained vision language models effective construction safety inspectors. *Data-Centric Engineering, 7*, e11. https://doi.org/10.1017/dce.2026.10044

[11] ssssssWang, L., Liu, Y., Wang, C., An, H., & Li, Y. (2026). VS-BIM: A cognitive map-driven framework enhancing MLLM for automatic safety inspection in construction. *Advanced Engineering Informatics, 69*, 103985. https://doi.org/10.1016/j.aei.2025.103985