

Large Multimodal Model for Context-aware Construction Safety Monitoring

Taegeon Kim¹, Seokhwan Kim², Minkyu Koo³, Minwoo Jeong⁴, and Hongjo Kim^{5*}

¹ Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea, E-mail address: ktg9655@yonsei.ac.kr

² Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea, E-mail address: yyksh2019@yonsei.ac.kr

³ Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea, E-mail address: kmk0119804@yonsei.ac.kr

⁴ Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea, E-mail address: minuj9855@yonsei.ac.kr

⁵ Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea, E-mail address: hongjo@yonsei.ac.kr (corresponding author)

Abstract: Recent advances in construction automation have led to increased use of deep learning-based computer vision technology for construction monitoring. However, monitoring systems based on supervised learning struggle with recognizing complex risk factors in construction environments, highlighting the need for adaptable solutions. Large multimodal models, pretrained on extensive image-text datasets, present a promising solution with their capability to recognize diverse objects and extract semantic information. This paper proposes a methodology that generates training data for multimodal models, including safety-centric descriptions using GPT-4V, and fine-tunes the LLaVA model using the LoRA method. Experimental results from seven construction site hazard scenarios show that the fine-tuned model accurately assesses safety status in images. These findings underscore the proposed approach's effectiveness in enhancing construction site safety monitoring and illustrate the potential of large multimodal models to tackle domain-specific challenges.

Key words: Construction automation, Large multimodal model, Context-aware safety assessment, Data generation

1. INTRODUCTION

Recent research in construction automation increasingly adopts deep learning-based computer vision technologies [1]. The real-time object detection [2] and segmentation model [3], capable of identifying personal protective equipment (PPE), facilitate the assessment of worker safety status. [4]. Moreover, the use of action recognition for worker behavior enables the determination of unsafe actions, further enhancing safety measures on construction sites [5–7].

However, the deployment of supervised learning-based models, limited to recognizing pre-defined classes, faces challenges in construction environments due to the coexistence of multiple, complex risk factors. This challenge highlights the need for adaptable models to the diverse and unpredictable nature of construction sites.

Large multimodal models [8,9], pre-trained on extensive image-text datasets, have emerged as a promising solution. Unlike supervised learning-based models restricted to specific tasks, large multimodal models excel in recognizing various objects and extracting semantic information in given image.

Chen et al. [10] employed a large multimodal model to evaluate the safety status of construction site images. However, Morris et al. [11] noted that performance of model [12] in vision-language tasks is

on par with the top 50% of skilled adults. Additionally, the pre-trained datasets lack construction domain-specific knowledge, requiring improvements in model performance for effective safety monitoring.

To address this issue, this study proposes a framework that utilizes existing computer vision model training data to 1) generate training data containing safety-centric descriptions for large multimodal models, and 2) fine-tune these models to optimize into the construction domain. An experimental comparison of the safety status assessment capabilities of a GPT-4V model [13] and the proposed fine-tuned model across seven construction site hazard scenarios was conducted. The results demonstrate that the fine-tuned model outperformed the GPT-4V model by an average of 30.71% across the scenarios, indicating a significant advancement in recognizing various situations on construction sites.

2. RESEARCH BACKGROUND

2.1. Challenges in using Large Multimodal Model for Construction Site Monitoring.

Effective deployment of large multimodal models in targeted applications necessitates precise assignment of downstream tasks, detailing the specific tasks these models need to undertake.

Crafting prompts, a process termed prompt engineering, is crucial for adapting models to new downstream tasks [14,15]. This method involves crafting prompts that guide the models in understanding and executing specific tasks. However, maintaining consistency and accuracy in model responses presents a challenge. Specifically, the complexity and variability of construction sites demand detailed and clear prompts for accurate scenario recognition and safety status assessment.

2.2. Multimodal Train Data Preparation

In vision-language tasks, training data for large multimodal models typically comprises images paired with corresponding textual descriptions. Large-scale caption dataset such as COCO [16], CC12M [17], and LAION [18] can be utilized for training large multimodal models. However, for higher performance in vision-language tasks, it is effective to construct training datasets as instruction-following data, which trains the model to understand user instructions and generate responses. Developing these datasets demands extensive domain-specific knowledge and significant effort. To address this challenge, Liu et al. [19] has proposed methods of constructing multimodal vision-language instruction-following data based on research [20] on the effectiveness of the text-annotation task using the Generative Pre-Training (GPT) model. By utilizing these instruction-following datasets, the large multimodal model named LLaVA (Large Language and Vision Assistant) was developed, demonstrating high performance in vision-language tasks.

3. METHODOLOGY

This paper proposes method enabling large multimodal models to identify risk factors in construction site images. As illustrated in Figure 1, the method features a two-step process: 1) generating instruction-following data including safety-centric detailed descriptions of hazardous conditions in construction site images using GPT-4V, and 2) fine-tuning the LLaVA model using the LoRA(Low-Rank Adaptation) [21] method with the generated data to develop a large multimodal model capable of assessing risk situations in construction site images.

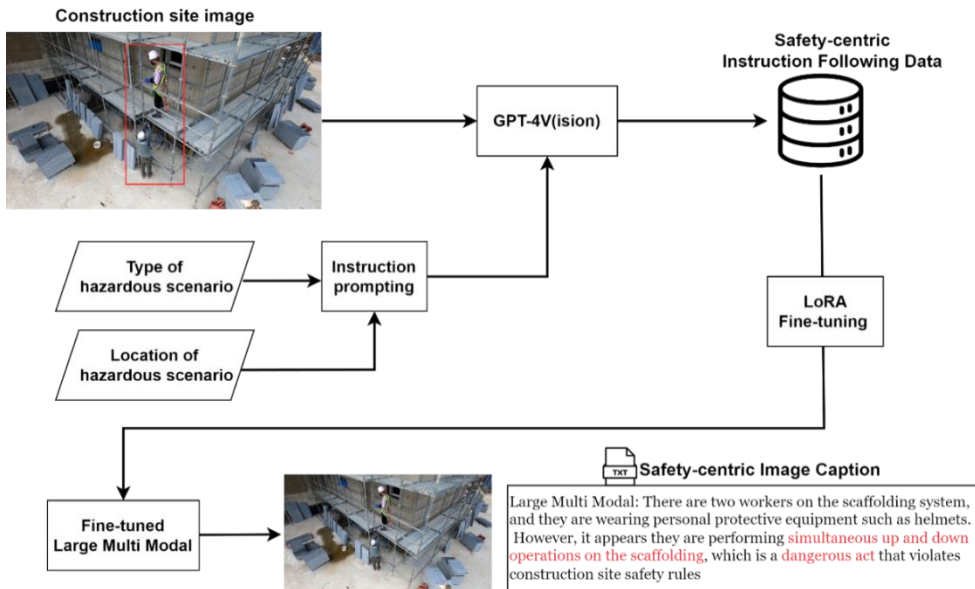


Figure 1. The overall framework of the proposed method

3.1. GPT-4V(ision)-assisted Instruction-following Data Generation in the Construction Domain

To generate instruction-following data, the study utilizes a construction site hazardous conditions dataset from ‘The Open AI Dataset Project (AI-Hub, S. Korea)’. This dataset comprises images, types of hazardous conditions (e.g., simultaneous upper and lower scaffolding work, lack of signal man during excavator operations), mask labels for construction objects, and bounding boxes indicating areas to identify hazardous situations.

Utilizing the types of hazardous conditions and the bounding box, input text prompts for GPT-4V are generated like instruction prompting content in Figure 2. The prompts include instructions to provide safety-centric descriptions, thereby focusing GPT-4V on generating responses that highlight risk factors.

Contrary to the method used for generating instruction-following data for LLaVA [19], which does not use images as input data, this study utilizes images as input data. The objective is to enrich the instruction-following data with not only information about the hazardous situations but also visual details present in the images. By inserting both images and the crafted text prompts into GPT-4V, safety-centric descriptions that include visual information from the images are generated, like in the output response in Figure 2. These output texts, paired with the images, form the instruction-following data for developing construction domain-specific large multimodal models

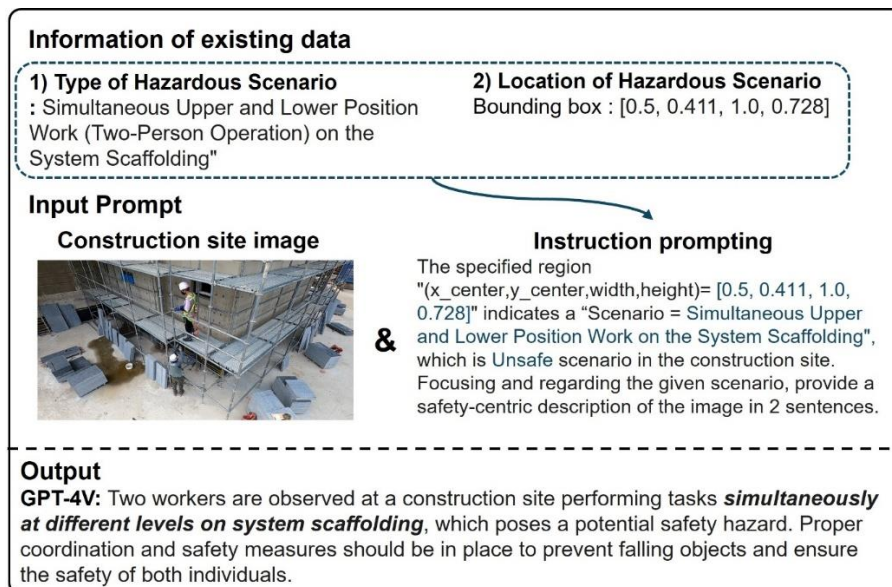


Figure 2. Example of generating instruction-following data in construction domain

3.2. Fine-tuning Large Multimodal Using LoRA method

Directly updating all parameters of a large multimodal model with the generated domain-specific data could lead to overfitting and a decrease in the pre-trained knowledge and performance of the model. To mitigate these issues, the LoRA method is employed instead of updating all the weights in the weight matrix of model. LoRA fine-tunes two smaller matrices that approximate the original weight matrix, enabling the model to maintain its pre-trained knowledge and performance.

The LoRA method not only reduces training time but also ensures the model adapts to the construction domain while maintaining its generalizability and robustness, addressing the challenge of applying large multimodal models to domain-specific tasks such as safety monitoring on construction sites.

4. EXPERIMENT

4.1. Experimental Setup

The experiment utilized a workstation equipped with four NVIDIA RTX 4090 GPUs. The dataset utilized for this study includes seven hazardous scenarios of construction sites, as shown in Table 1. For each scenario, 200 images and labels were applied to the proposed method for generating instruction-following data, which were then employed as training data. The test dataset consisted of images from various construction sites, excluded from the training set, with 40 images per scenario. Figure 3 displays sample images from the dataset used.

The experiment employed the Vicuna-13B-based LLaVA model as the large language model. The model was fine-tuned with the generated train data using the LoRA method during 20 epochs. This fine-tuning process took 2 hours and 51 minutes, with the training loss reducing from an initial measurement of 1.6393 to 0.0003.

Table 1. Scenarior classes for the instruction following data generation

No.	Scenario	Safe/Unsafe
1	Not working at the very top of the ladder	Safe
	Working at the very top of the ladder	Unsafe
2	No fall hazard thing present on the scaffolding	Safe
	Placement of fall hazard thing on the scaffolding	Unsafe
3	No simultaneous work at upper and lower positions on the scaffolding	Safe
	Simultaneous upper and lower positions work on the scaffolding	Unsafe
4	Signalman placement during excavator operations	Safe
	No signalman placement during excavator operations	Unsafe
5	Fire extinguisher placement next to the welding machine	Safe
	No fire extinguisher positioned next to the welding machine	Unsafe
6	Fire extinguisher and welding blanket placement during circular saw operations	Safe
	No fire extinguisher and welding blanket placement during circular saw operation	Unsafe
7	Material loading below the driver's line of sight during forklift operation	Safe
	Material loading above the driver's line of sight during forklift operation	Unsafe



Figure 3. Examples of used scenario image

4.2. Experimental results and discussion

Table 2 shows the safety status assessment accuracy for each scenario, comparing the performance between the fine-tuned LLaVA model and GPT-4V. The results demonstrate that the fine-tuned LLaVA model exceeded the performance of GPT-4V in all scenarios.

Table 2. Comparative performance of the LLaVA fine-tuned by proposed method and GPT-4V in safety status assessment accuracy

	Accuracy of Safety Status Assessment	
	Fine-tuned LLaVA in construcion domain	GPT-4V
Scenario 1	97.5%	47.5%
Scenario 2	95%	77.5%
Scenario 3	95%	62.5%
Scenario 4	97.5%	17.5%
Scenario 5	70%	60%
Scenario 6	75%	62.5%
Scenario 7	87.5%	75%
Mean Accuracy	88.21%	57.5%

Figure 4 illustrates examples of the outcomes from both models. The fine-tuned LLaVA model provided safety-centric descriptions for the images, presenting evidence for the safety status assessment that was appropriately aligned with the hazard scenarios. While GPT-4V was capable of generating accurate and detailed descriptions of the images, its performance in assessing safety status lacked consistency and accuracy in recognizing hazardous elements. The higher performance of the fine-tuned LLaVA model is attributed to training on instruction-following data that concentrates on risk factors, thus improving the accuracy of safety status assessments.

The simplicity of the prompts used to test the models may also be a contributing factor. GPT-4V faced difficulties in accurately assessing safety status due to these simple prompts. Although the performance of GPT-4V might improve with more specific prompts, crafting prompts that include safety judgment criteria specific to various construction site hazard scenarios poses a challenge.

These findings underscore the effectiveness of the proposed framework in enhancing the ability of large multimodal models to accurately assess safety conditions in construction site images.

Examples of Context-aware Construction safety Monitoring



Look at the image in detail
Answer for following 2 requirements

1. Provide the description of the image
2. Please respond with either 'Safe' or 'Unsafe' to indicate whether the situation in the photo is dangerous or not

Proposed fine-tuned model :

1. A forklift is depicted in the image, carrying a stack of materials that **appears to obstruct the driver's forward visibility**, posing a potential safety hazard. A worker is standing nearby, wearing a safety vest and hard hat, indicating a commitment to safety protocols despite the unsafe material loading practice.
2. **Unsafe**. The materials loaded on the forklift **obstruct the driver's line of sight**, creating a risk of accidents due to reduced visibility.

GPT-4V:

1. The image shows an outdoor setting where a person in a high-visibility vest and hardhat is standing on a paved area, observing a forklift. The forklift is transporting a large stack of white material which appears to be securely placed on its forks. Behind the scene is a fence and beyond that, some dense greenery. The worker is standing a safe distance away from the forklift, which is currently stationary.
2. **Safe**, The situation seems to be under control, with safety measures being followed as the worker maintains a safe distance from the operational forklift. The area around the forklift is clear, without evident hazards, and the load being carried by the forklift appears to be stable and properly secured.



Look at the image in detail
Answer for following 2 requirements

1. Provide the description of the image
2. Please respond with either 'Safe' or 'Unsafe' to indicate whether the situation in the photo is dangerous or not

Proposed fine-tuned model :

1. **A signalman is positioned at a safe distance** from an excavator, using hand signals to guide the operator during excavation operations. The scene takes place in a construction area near a modern building, with clear ground markings and proper safety gear ensuring a controlled and safe work environment.
2. **Safe**, The situation appears to be safe, **as the signalman is at a safe distance from the excavator and is wearing appropriate visibility gear**. The excavator is operating in a construction zone, where such equipment is commonly used, and the signalman seems to be following proper safety protocols by maintaining a clear line of sight with the excavator operator.

GPT-4V:

1. The image shows a heavy orange colored wheeled excavator with its arm extended and bucket raised above the roofline of a dark red and gray paneled building in the background. In front of the building, slightly to the right, there is a person wearing a red jacket with a hood, jeans, and holding something that looks like a helmet or hat close to their head with their right hand, as if they might be taking a photograph of themselves with the excavator in the background. The environment looks like an industrial or commercial area with clear skies.
2. **Unsafe**, The situation can be considered unsafe since the person is standing relatively close to a heavy piece of construction equipment that is in operation. There is potential danger from the excavator moving unexpectedly or from the bucket dropping, which could result in injury to the person nearby. Safety best practices generally require maintaining a safe distance from operational heavy machinery.

Figure 4. Examples of the outcomes from fine-tuned LLaVA and GPT-4V

5. CONCLUSION

This paper introduced a method to enhance safety monitoring technologies for identifying hazardous situations on construction sites. The proposed approach addresses challenges in creating high-quality multimodal training data within the construction domain, demonstrating the feasibility of optimizing large multimodal models for such applications. Through developing and fine-tuning these models,

especially by using construction domain-specific instruction-following data, this study has shown significant improvements in the ability of model to accurately assess safety conditions in construction site images.

The experiments and results presented confirm that the fine-tuned large multimodal model, specifically the LLaVA model enhanced through LoRA method training, outperforms GPT-4V in safety status assessment tasks. This result is due to the training model with data that not only includes visual information but also safety-centric descriptions, enabling a more focused and accurate identification of risk factors within construction environments.

Future research aims to advance this method in several ways. One method involves adding various kinds of hazardous scenarios, thereby enhancing the versatility and robustness in safety monitoring applications. Additionally, techniques like Retrieval-Augmented Generation (RAG) can improve the reliability of response refinement. The ultimate goal is to develop more sophisticated context-aware safety monitoring technologies that can adapt to the diverse and dynamic nature of construction sites.

ACKNOWLEDGMENTS

This research was conducted with the support of the “2023 Yonsei University Future-Leading Research Initiative (No. 2023-22-0114)” and the “National R&D Project for Smart Construction Technology (No. RS-2020-KA156488)” funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation. And this research used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

REFERENCES

- [1] O. Maali, C.-H. Ko, and P. H. D. Nguyen, “Applications of existing and emerging construction safety technologies,” *Automation in Construction*, vol. 158, p. 105231, Feb. 2024
- [2] G. Jocher et al., “ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation.” Zenodo, Nov. 22, 2022.
- [3] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT: Real-time Instance Segmentation.” arXiv, Oct. 24, 2019.
- [4] W.-C. Chern, J. Hyeon, T. V. Nguyen, V. K. Asari, and H. Kim, “Context-aware safety assessment system for far-field monitoring,” *Automation in Construction*, vol. 149, p. 104779, May 2023.
- [5] H. Guo, Z. Zhang, R. Yu, Y. Sun, and H. Li, “Action Recognition Based on 3D Skeleton and LSTM for the Monitoring of Construction Workers’ Safety Harness Usage,” *Journal of Construction Engineering and Management*, vol. 149, no. 4, p. 04023015, Apr. 2023.
- [6] X. Luo, H. Li, X. Yang, Y. Yu, and D. Cao, “Capturing and Understanding Workers’ Activities in Far-Field Surveillance Videos with Deep Action Recognition and Bayesian Nonparametric Learning,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 4, pp. 333–351, 2019.
- [7] P. Zhai, J. Wang, and L. Zhang, “Extracting Worker Unsafe Behaviors from Construction Images Using Image Captioning with Deep Learning–Based Attention Mechanism,” *Journal of Construction Engineering and Management*, vol. 149, no. 2, p. 04022164, Feb. 2023.
- [8] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision.” arXiv, Feb. 26, 2021.
- [9] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.” arXiv, Feb. 15, 2022.
- [10] H. Chen et al., “Augmented reality, deep learning and vision-language query system for construction worker safety,” *Automation in Construction*, vol. 157, p. 105158, Jan. 2024.
- [11] M. R. Morris et al., “Levels of AGI: Operationalizing Progress on the Path to AGI.” arXiv, Jan. 05, 2024.
- [12] X. Chen et al., “PaLI: A Jointly-Scaled Multilingual Language-Image Model.” arXiv, Jun. 05, 2023.
- [13] OpenAI et al., “GPT-4 Technical Report.” arXiv, Dec. 18, 2023.
- [14] J. Gu et al., “A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models.” arXiv, Jul. 24, 2023.

- [15] H. Strobelt et al., “Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 01, pp. 1146–1156, Jan. 2023
- [16] X. Chen et al., “Microsoft COCO Captions: Data Collection and Evaluation Server.” arXiv, Apr. 03, 2015.
- [17] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, “Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts.” arXiv, Mar. 30, 2021.
- [18] C. Schuhmann et al., “LAION-5B: An open large-scale dataset for training next generation image-text models.” arXiv, Oct. 15, 2022.
- [19] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning.” arXiv, Dec. 11, 2023.
- [20] F. Gilardi, M. Alizadeh, and M. Kubli, “ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 120, no. 30, p. e2305016120, Jul. 2023.
- [21] E. J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models.” arXiv, Oct. 16, 2021.