Data-Centric Engineering (2026), 7: e11
doi:10.1017/dce.2026.10044

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Are large pre-trained vision language models effective construction safety inspectors

Xuezheng Chen[1] and Zhengbo Zou[2]

[1]The University of British Columbia – Vancouver Campus, Canada
[2]Civil Engineering and Engineering Mechanics, Columbia University, USA
Corresponding author: Zhengbo Zou; Email: zhengbo.zou@columbia.edu

Received: 24 June 2025; **Revised:** 23 December 2025; **Accepted:** 18 February 2026

Keywords: computer vision; construction safety inspection; multi-modal dataset; natural language processing; vision language models

Abstract

Construction safety inspections typically involve a human inspector identifying safety concerns on-site. With the rise of powerful vision language models (VLMs), researchers are exploring their use for tasks such as detecting safety rule violations from on-site images. However, there is a lack of open datasets to comprehensively evaluate and further fine-tune VLMs in construction safety inspection. Current applications of VLMs use small, supervised datasets, limiting their applicability in tasks they are not directly trained for. In this article, we propose the *ConstructionSite 10 k*, featuring 10,000 construction site images with annotations for three inter-connected tasks, including image captioning, safety rule violation visual question answering (VQA), and construction element visual grounding. Our subsequent evaluation of current state-of-the-art large pre-trained VLMs shows notable generalization abilities in zero-shot and few-shot settings, while additional training is needed to make them applicable to actual construction sites. This dataset allows researchers to train and evaluate their own VLMs with new architectures and techniques, providing a valuable benchmark for construction safety inspection.

Impact statement

Vision-language models (VLMs) are a type of AI model that processes images as input and generates textual output. Large pre-trained VLMs, equipped with billions of parameters and trained on vast image-text pairs from the internet, are capable of performing tasks they were never explicitly trained on, using only a few prompts. This flexibility allows such models to be applied to a wide range of construction safety-related tasks without requiring changes to their architecture or additional training. This article introduces an evaluation framework for assessing how well large pre-trained VLMs can perform construction safety tasks, along with a comprehensive dataset specifically designed for this purpose. Practitioners can use this dataset to fine-tune their own VLM to suit specific needs for safety inspection.

1. Introduction

Hazardous working conditions and risky behaviors at construction sites accounted for a concerning 21% of all workplace fatalities in the United States in 2022 (US-BUREAU-OF-LABOR-STATISTICS, n.d.). To ensure compliance with safety regulations, such as wearing proper personal protective equipment

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

(PPE), proper marking of danger zones, and maintaining exclusion zones, construction site inspectors must walk around sites to identify hazards, document hazards (e.g., take photos), propose solutions, and write reports. This traditional yet prevalent practice is time-consuming, labor-intensive, and does not guarantee comprehensive coverage of all safety violations due to individual judgments.

To reduce the demand for time and manpower, the construction industry has been actively employing artificial intelligence (AI) models based on deep learning, a subset of machine learning that uses multi-layered neural networks to automatically learn hierarchical patterns from data. Early implementations of deep learning focused on specialized vision-only architectures such as convolutional neural networks (CNN) for visual recognition in safety inspection tasks, such as PPE detection or pose estimation, through on-site images and videos (An et al., 2021; Chen and Demachi, 2021; Xiong and Tang, 2021). The novel transformer architecture makes it possible for AI models to effectively understand long paragraphs of text. Derived from this advancement, VLMs are a type of deep learning model that fuse the vision and language modalities to understand both image and text inputs. The construction industry has been actively incorporating VLMs with supervised training into various safety inspection tasks, such as PPE detection (Gil and Lee, 2024), construction scene captioning (Liu et al., 2020; Xiao et al., 2022; Jung et al., 2024; Kim et al., 2025), safety rule violation detection (Liu et al., 2022; Fang et al., 2023a; Chan et al., 2025; Kim et al., 2025), and safety report generation (Tsai et al., 2023).

Although previous studies have demonstrated the feasibility of training VLMs with construction site images for specialized tasks, the small model size and limited training data, both in size and task type, hinder the models' ability to generalize to further downstream tasks for which they were not explicitly trained, creating a *modeling challenge* (Liu et al., 2020, 2022; Fang et al., 2023a). Moreover, the lack of a comprehensive dataset means researchers have to take photos of individual construction sites, design their own tasks, and annotate their own data, constituting a *data challenge* (Liu et al., 2020; Xiao et al., 2022; Fang et al., 2023a; Tsai et al., 2023; Gil and Lee, 2024; Chan et al., 2025; Kim et al., 2025). Data augmentation techniques have been adopted to mitigate data scarcity in prior works. Tsai et al. (2023) fine-tuned a Contrastive Language-Image Pre-training (CLIP) model for safety violation report generation and addressed class imbalance of violation data by resampling minority classes to ensure balanced fine-tuning. Similarly, Liu et al. (2020) augmented their dataset by generating five synonymous descriptions for each construction site image, thereby increasing the diversity of textual captions. However, these methods often fail to provide more training opportunities about domain-specific challenges in construction safety tasks, such as new spatial relationships in cluttered environments. Some researchers have used VLMs to generate training data (Jung et al., 2024; Chan et al., 2025; Kim et al., 2025). However, such data are often either small in scale, limited in task diversity, or lacks human verification. This insular landscape hinders the development of new models and their application in construction safety tasks. In this work, we attempt to tackle both the modeling and data challenges.

With recent breakthroughs of large pre-trained VLMs (e.g., GPT family and LLaVA) in a variety of everyday language and vision tasks, it is natural to ask: *Can pre-trained VLMs be directly deployed to improve construction safety through automated inspections?* The rationale behind this query is the flexibility provided by large pre-trained VLMs such as Gemini. For instance, in everyday object detection tasks, there is no need to restructure or retrain the model to detect a new object category—a simple prompt adjustment can address the issue due to the powerful generalization capabilities of large models. Moreover, the in-context learning capacity means that models can quickly output in the format desired given few examples, also called a few-shot setting. Additionally, large pre-trained VLMs are able to generate language output, which can be later parsed into other formats (e.g., bounding boxes). Therefore, these models have the potential to be applied to a broader range of construction tasks, such as generating safety violation reports, as compared to other specialized models.

To make better use of these models, it is crucial to understand the perception and decision-making capacities of large pre-trained VLMs. With this crucial task in mind, we designed three experiments to benchmark large VLMs' capabilities for construction safety inspection. These experiments include: (1) describing images in natural language to demonstrate image understanding, (2) safety rule violation VQA, i.e., giving a model an image-text pair and query the model if there is any safety rule violations, and

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

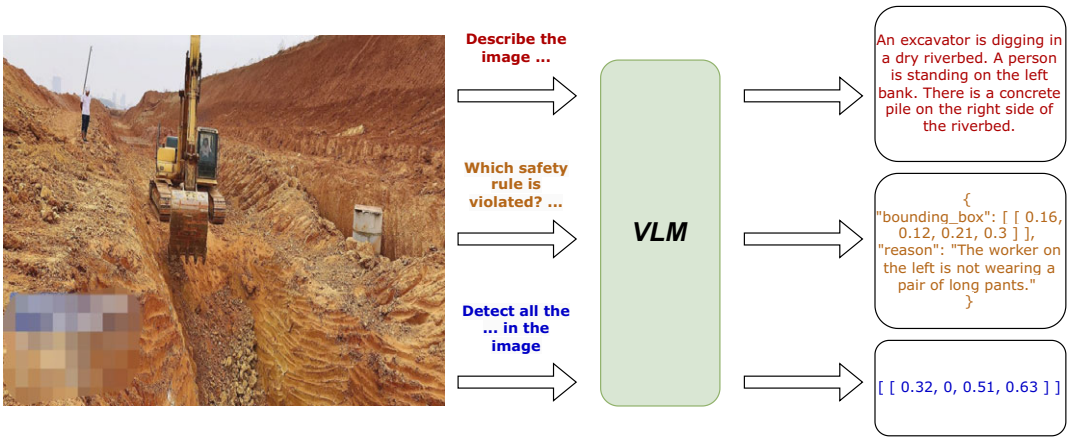


Figure 1. *The high-level schematic diagram shows the three tasks employed in this article. The VLM will receive the construction site image and prompts for three tasks: image captioning, safety rule violation VQA, and construction element visual grounding.*

(3) construction element visual grounding, i.e., locating construction elements in an image based on a natural language query. A high-level schematic diagram showing the nature of three tasks are presented in Figure 1. To the best of our knowledge, there has been no research focused on evaluating large pre-trained VLMs in the context of understanding construction site images. Specifically, it remains to be determined whether these models can be effectively utilized for downstream tasks without further supervised training.

To implement these tasks, we first curated a dataset we named *ConstructionSite 10 k*, which consists of 10,013 construction site images of varying quality and topics, accompanied by multiple types of human-labeled annotations. These annotations include image captions; safety rule violation, reasoning, and grounding; and construction site element visual grounding. *ConstructionSite 10 k* addresses the lack of image captioning datasets with comprehensive description of site images. The dataset also introduces downstream tasks such as VQA and visual grounding tailored specifically for construction applications. *ConstructionSite 10 k* can serve as a pre-training dataset for VLMs to understand construction sites and construction related tasks or as a fine-tuning dataset for VQA or visual grounding purposes. We tested several popular large pre-trained VLMs, including the state-of-the-art (SOTA) Gemini series and the open-source LLaVA series, in both zero-shot and few-shot settings to determine whether these high-performing pre-trained models already possess the ability to generalize their knowledge and apply it to construction safety inspection.

Testing results show that the latest large proprietary models such as GPT and Gemini can understand construction related objects in images, including their attributes, interactions, and background, underscored by thoroughly generated image captions, and setting up the SOTA in the image captioning task. Smaller open-source VLMs do not perform as well as proprietary models. In the safety violation VQA section, VLMs score high in recall but low in precision, as compared to ground truth labeled by humans. Nonetheless, none of the VLMs achieved desirable results in any of the visual grounding tasks. Performance degrades to an Intersection over Union (IoU) below 20% for abstract regions (e.g., safety violation regions), objects with irregular shapes, and specific types of objects within a category (e.g., workers *with white hard hat* rather than any workers). While VLMs, including the SOTA Gemini series, CURRENTLY still require improvements to be fully competent in construction inspection tasks, *ConstructionSite 10 k* and the evaluation framework proposed in this article lay the groundwork for further training and deployment of large pre-trained VLMs in construction-related applications.

We summarize our contribution as the following:

1. We propose *ConstructionSite 10 k*, which contains 10,013 images taken at different construction sites and annotations containing image caption, safety violations VQA, and object visual grounding.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

2. We propose a framework using image captioning, three-stage safety rule violation VQA, and construction element visual grounding to probe VLMs' understanding of construction site images and their capacity to follow the prompt and perform construction safety inspection tasks.
3. We perform comprehensive experiments to evaluate state-of-the-art, off-the-shelf large pre-trained generative VLMs on their ability to generalize their knowledge by performing construction-related tasks in a zero-shot and five-shot settings.

2. Background

2.1. Vision language models

VLMs have become prominent due to their remarkable capabilities in understanding both images and human languages. Among various efforts to connect the visual and linguistic worlds, Radford et al. (2021) stand out as pioneers with their creation of CLIP. The CLIP model, which utilizes both an image encoder and a text encoder, was pre-trained on image-text pairs to understand images, language, and their relationships. Building on the pre-trained image encoder from CLIP and the open-source LLaMA language model, Liu et al. (2023a, 2024) fine-tuned their LLaVA model using instruction-following data. This visual instruction tuning approach transforms a VLM from a next-word predictor to a visual assistant capable of following instructions and responding to requests. Their understanding of images and text, along with their ability to follow prompts, makes them promising tools for performing construction site inspections.

Moreover, VLMs' in-context learning ability enables them to be quickly applied to different construction inspection tasks in zero-shot (i.e., instruction only) and few-shot (i.e., instruction and some examples in human language) settings without additional supervised training or architectural change. The models quickly understand the task with merely a description of the task. GPT-2, notable for its strong unsupervised task learning capabilities, showcased SOTA performance across various benchmarks despite being under-fitted to the WebText dataset (Radford et al., 2019). Building on this, GPT-3 (Brown et al., 2020) further illustrates that large pre-trained models can achieve performance improvements in few-shot settings. The success of zero-shot and few-shot inferences underscores the superior generalization capacity of large pre-trained models that excel in domains they are not trained for without the need for human-labeled datasets or modifications to their architecture.

To improve the reasoning capability of large models, Chain-of-thought (CoT) prompting was proposed by Wei et al. (2023). CoT introduces intermediate reasoning steps when querying large models with complex questions. This approach encourages models to break down a complex task into smaller, more manageable parts and think through each segment in detail.

In this article, we investigate the core question of whether current VLMs are capable of understanding construction sites and conduct construction safety inspection through in-context learning. We achieve this by evaluating the capabilities of both larger and smaller VLMs through image captioning tasks and safety rule violation tasks across zero-shot, few-shot settings, and CoT scenarios. Our objective is to verify whether the idea of leveraging in-context learning can inform strategies for future training and inference of large pre-trained models. This exploration not only tests the robustness of these models but also explores avenues for optimizing their application in real-world scenarios.

2.2. Applications of VLMs in construction

A summary of recent works of applying VLMs in the construction industry, as well as our own work, is presented Table 1. The ability of VLMs to understand both images and human language facilitates their deployment in inspection tasks within construction sites. For instance, in enhancing PPE compliance detection, Gil and Lee (2024) leveraged the image captioning capability of the CLIP model. This involves describing cropped regions of construction site images containing human bodies with a image captioning model derived from CLIP. The resulting text embeddings were then used to calculate cosine similarity with prompts indicating compliance or violation to determine rule adherence. Remarkably, their proposed zero-shot method outperformed traditional YOLO methods in this context.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Table 1. *Recent works of applying VLM for construction site inspection*

Reference	Task	Training data	Dataset open source or not
Gil and Lee (2024)	Image captioning, PPE detection	None (zero-shot)	Yes, test set is available upon request
Fang et al. (2023a)	Safety rule violation detection	700 images	No
Liu et al. (2022)	Safety rule violation detection and grounding	MSCOCO object detection dataset as pre-training data, 860 images with augmentation as fine-tuning data	Yes, fine-tuning data are available upon request
Liu et al. (2020)	Image captioning	7,400 images and 36,900 captions	Yes, on GitHub
Xiao et al. (2022)	Image captioning	4,000 images and 8,000 captions	Yes, data are available upon request
Jung et al. (2024)	Image captioning, Safety report generation	6,400 images and 13,000 captions	Yes, on GitHub
Kim et al. (2025)	Image captioning, safety rule violation detection	2,000 images, 2,000 cations, and 2,000 safety rule VQA	Yes, data are available upon request
Chan et al. (2025)	Video-based safety rule violation detection	900 images, 900 captions, and 900 safety rule VQA	Yes, data are available upon request
Tsai et al. (2023)	Safety report generation	806 image-text pairs with augmentation as fine-tuning data	No
Ours	image captioning, multiple choice, question answering, visual grounding	None (zero-shot and few-shot)	Yes, both training set and test set is available online

To broaden the application of VLMs beyond PPE detection, Fang et al. (2023a) proposed a VLM architecture comprising a Faster R-CNN as the image encoder and a GRU network as the text encoder. Their model was trained to detect eight different types of safety rule violations in construction site images. Expanding on this work, Liu et al. (2022) further developed the research by incorporating a model capable of grounding these violations with bounding boxes. They employed DarkNet as the image encoder and BERT as the text encoder.

Generative VLMs can be used to describe the construction scene and aid in construction management tasks. Both Liu et al. (2020) and Xiao et al. (2022) developed generative VLMs for this purpose. Liu et al. (2020) trained a model consisting of a CNN image encoder and an LSTM text decoder. This model generated structured descriptions by selecting words from a pre-defined descriptive sentence bank. On the other hand, Xiao et al. (2022) employed attention mechanisms in their model, focusing specifically on describing construction vehicles. Additionally, Tsai et al. (2023) aimed to use VLMs to automate the generation of safety violation reports for superior efficiency and consistency. To achieve this, they fine-tuned a variant of the CLIP model with two separate text encoders: one encoding "violation type" and the other "caption type." These encoded choices were used to calculate cosine similarity with the encoded image to determine the correct attributes. Subsequently, the encoded attributes, image, and image caption embeddings were fed into a GPT-2 model to generate safety violation reports with a standardized format.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

To further advance VLM applications in construction, several studies have explored domain-specific fine-tuning. Both Jung et al. (2024) and Kim et al. (2025) adapted pre-trained VLMs using construction site image caption datasets. Jung et al. (2024) introduced the first vision transformer–based image captioning model for construction, fine-tuning an mPLUG–BERT architecture (Devlin et al., 2019; Li et al., 2022) to generate domain-specific captions that support downstream tasks such as report generation and image retrieval. Kim et al. (2025) fine-tuned the LLaVA-1.5 image encoder to emphasize construction activities, enabling the model to produce safety-aware captions that describe the scene, assess safety compliance, and provide justifications. Similarly, Chan et al. (2025) fine-tuned CogAgent (Hong et al., 2024) using construction object labels and safety ontologies, transforming it into a video-based analyzer for detecting working-at-height violations.

In this article, we aim to overcome the limitations of task-specific models and supervised training by exploring the capabilities of large pre-trained generative VLMs in the context of construction site images. Our approach focuses on assessing these models' ability to comprehensively understand construction site images and accurately express this understanding through unconstrained image captions. Additionally, we leverage their capacity for in-context learning to tackle tasks such as VQA and visual grounding.

2.3. Multi-modal datasets

There have been extensive efforts to compile multi-modal datasets aimed at training VLMs to effectively integrate vision and language modalities. Among the popular datasets are Flickr (Hodosh et al., 2013), MSCOCO (Chen et al., 2015), and Abstract Scene (Zitnick and Parikh, 2013) datasets. These datasets contain large volumes of images, each accompanied by multiple captions, and have been widely utilized as pre-training datasets for VLMs. RefCOCO (Kazemzadeh et al., 2014), pioneering in visual grounding, plays a crucial role in this domain by enabling the detection of objects through language prompts provided to VLMs. Visual grounding, distinct from traditional object detection, allows for additional constraints (e.g., color, location) to be imposed on the detection task. Furthermore, VQA datasets such as VizWiz (Gurari et al., 2018), OKVQA (Marino et al., 2019), and ScienceQA (Lu et al., 2022) significantly expand the utility of VLMs beyond image description to answering questions related to images, which broadens the scope of applications for these models.

Datasets containing images with alt-text style captions tend to be straightforward. These images often have excellent illumination, fewer objects, and simple layouts. In contrast, images captured at construction sites present significant challenges due to poor weather conditions, varying illumination, and the complex perspectives from UAVs or tower cranes, often at a distance. Moreover, these images contain rich content with multiple objects from diverse classes and various interactions, posing considerable confusion for machine learning models. The differences between these types of datasets render VLMs trained on common object-centric datasets less effective in construction-related tasks.

While there are publicly available datasets containing construction site images, such as SODA (Duan et al., 2022), MOCS (An et al., 2021), and ACID (Xiao and Kang, 2021), their primary focus is typically on object detection tasks. The CASIC[1] dataset (Liu et al., 2020) aims to address the gap by providing captioned construction site images. However, it covers only a limited range of activities with an average of merely 12 words per caption. The VSD dataset (Jung et al., 2024) tried to create an image captioning dataset with natural language but relied on uncorrected machine-generated captions. The captions also only provide a summary of the entire scene without diving into details. Table 2 compares our dataset with the construction-related computer vision datasets aforementioned to better visualize the similarities and differences between the datasets.

This deficiency underscores the need for specialized datasets that comprehensively capture the complexities of construction site environments and capture as many semantic features as possible from images. Such datasets would enable VLMs to better understand the context surrounding the main scene

[1] Unofficial abbreviation for "Construction Activity Scene Image Caption."

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Table 2. *Comparisons between our dataset with popular construction-related computer vision datasets*

Dataset	No. of images	Type of annotations	Avg. caption length	No. of question-answer pairs per image	Object category for detection	Features
SODA (Duan et al., 2022)	20,000	Bounding box	-	-	15	1. Object detection at construction sites
MOCS (An et al., 2021)	42,000	Bounding box, segmentation mask	-	-	13	1. Object detection and instance segmentation at construction sites 2. Collected from a variety of construction projects
ACID (Xiao and Kang, 2021)	10,000	Bounding box	-	-	10	1. Focusing on construction machines
CASIC (Liu et al., 2020)	7,000	Caption	11.8	-	-	1. Each image focuses on only one construction activity 2. Each image is described with five similar short sentences
VSD (Jung et al., 2024)	7,800	Caption	10.3	-	-	1. Training and validation data include diverse actions and job-site environments to prevent overfitting 2. Some captions are generated using pre-trained VLM and validated against annotated object existence
ConstructionSite 10 k	10,000	Caption, bounding box, multiple choices, one or two sentences for safety rule VQA, bounding box	53.0	3	3	1. Detailed image captions for construction site images 2. Captions contain predicates indicating relationships between instances 3. Relative and absolute positional information contained 4. Three-stage VQA 5. visual grounding under constraints

and establish more nuanced relationships between objects. Moreover, datasets focusing on downstream tasks like safety rule violation VQA and visual grounding in construction site inspection are also lacking.

To address these gaps, this article proposes the creation of an image captioning dataset aimed at enhancing VLM training. Our dataset also introduces VQA and visual grounding tasks specific to construction site inspection. This initiative seeks to foster advancements in VLMs' capabilities within the challenging contexts of construction environments.

2.4. Research objective

We hypothesize that if VLMs can grasp every detail and relationship within a construction site image, they can be applied to various tasks simply by adjusting the input prompt. To test our hypothesis, we have curated a comprehensive construction image caption dataset, *ConstructionSite 10 k*, and developed a framework to evaluate the VLMs' capacity to perform multiple construction site-related tasks under zero-shot and few-shot settings. The proposed framework includes three tasks: image captioning, safety rule violation VQA, and visual grounding tasks. It also includes metrics used to evaluate models' performance on each task, as well as prompts and hyperparameters used.

3. ConstructionSite 10 k

This section presents a comprehensive analysis and visualization of the *ConstructionSite 10 k* dataset. The dataset is publicly available at: https://huggingface.co/datasets/LouisChen15/ConstructionSite

3.1. Images

The construction site images utilized in this dataset were collected and made publicly available by An et al. (2021). We selected a total of 10,013 images that are clear, have acceptable illumination, and contain distinguishable construction elements that are not occluded by mosaics.

Images captured across various construction sites, times, and with different equipment will exhibit distinct characteristics. Building on the groundwork laid by An et al. (2021), we assign four key features to each image to reflect its condition: camera distance, illumination, plan view, and quality of information. These labels are intended to aid users in selecting images suitable for specific purposes.

Within this framework, camera distance refers to the proximity of the camera to the majority of construction elements in the image. Illumination describes the lighting conditions present in the image. An image is classified as a plan view if the angle of view exceeds approximately 45 degrees from the ground. Images with sparse information typically depict isolated instances or small groups of construction elements such as humans, equipment, and material stockpiles, each clearly distinguishable without significant occlusion. In contrast, images labeled as rich information portray a diverse array of construction elements with complex interactions and occlusions between them.

The assignment of these labels follows predefined criteria but also incorporates annotator judgment. Figure 2 provides visual examples showcasing images with different feature labels. Figure 3 illustrates the distribution of these features across the dataset. These features offer useful references for researchers who want to divide the dataset to create challenging subsets for model training. This study represents one of the earliest attempts to systematically describe image conditions through assigned labels.

3.2. Image annotations and corresponding tasks

The images were manually annotated by the authors, with various types of annotations corresponding to different tasks. All annotations were reviewed and validated by two domain experts- construction engineers with experience in both construction site safety inspections and academic research. The following subsections are categorized by tasks used to evaluate VLMs. The time used to create the dataset and prepare the annotations are presented in Figure 4.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press



Figure 2. *The figure presents construction site images from the dataset, each annotated with three labels arranged in a grid structure. For example, the bottom-left image is labeled as "sparse info," "night," and "short distance." The images are uniformly scaled for demonstration purposes, resulting in slightly altered appearances. These images and labels aim to showcase representative samples and do not depict the actual distribution of the dataset.*

The dataset was split into a training set comprising 70% of the images and corresponding annotations, and a test set comprising the remaining 30%. A more detailed split of annotations is available in Figure 5 and Table 3. The training set is intended for model training and fine-tuning, while the test set is exclusively used in this study to evaluate the performance of large pre-trained VLMs.

3.2.1. Image captioning

The annotations include detailed and objective descriptions in English with a focus on construction elements such as workers, construction equipment, and materials. Attributes of objects such as color, location, number, and activities are included in the image captions. We first manually annotated around 2,000 images. To increase the annotation speed, we utilized GPT-4 V in a five-shot setting to create the structure of the captions for the remaining images. Human annotators then adapted the machine-generated descriptions to ensure their fidelity and adequacy. The GPT-aided approach results in an approximately 20% reduction of annotation time for each caption. A workflow of the GPT-assisted image caption

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

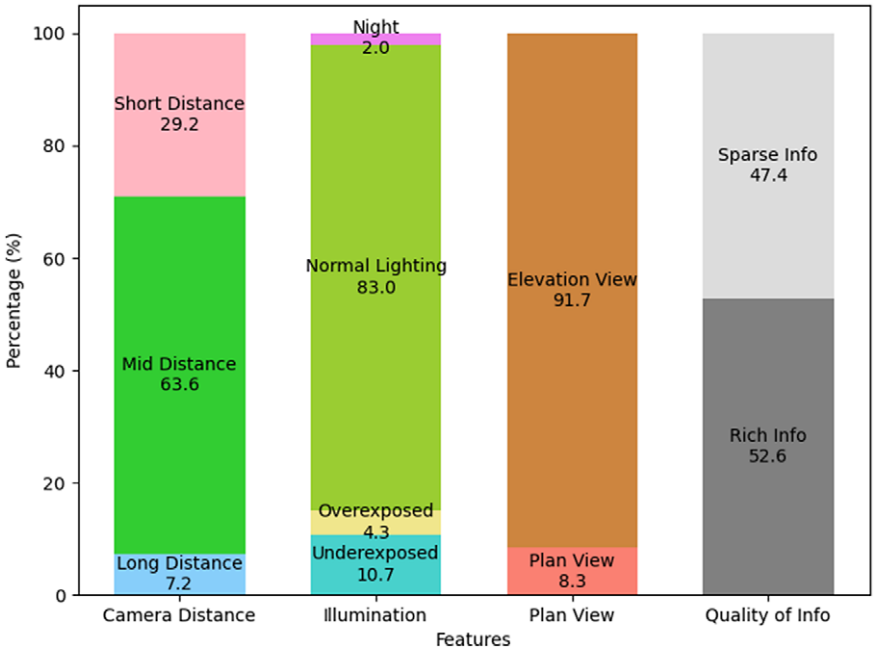


Figure 3. *Distribution of images in the dataset across four features. These statistics unveil the diversity of construction site imagery.*

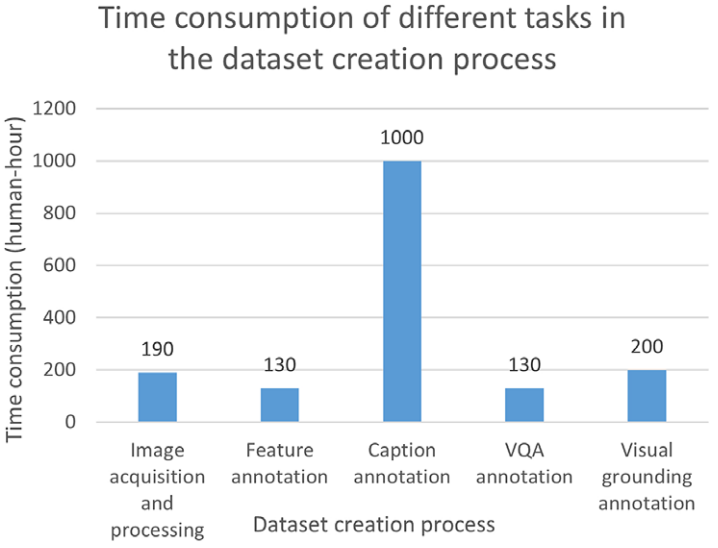


Figure 4. *The time consumption breakdown to prepare the dataset. The time shown in the figure approximates and includes the time to create annotation software.*

annotation is shown in Figure 6. The process involves three key steps: (1) inputting task-specific prompts, few-shot examples, and a query image into the GPT-4 V annotator; (2) obtaining an initial AI-generated caption; and (3) refining this caption through human annotation to ensure accuracy and relevance. This hybrid approach leverages the efficiency of GPT-4 V while prioritizing human oversight to produce high-quality, domain-specific annotations.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

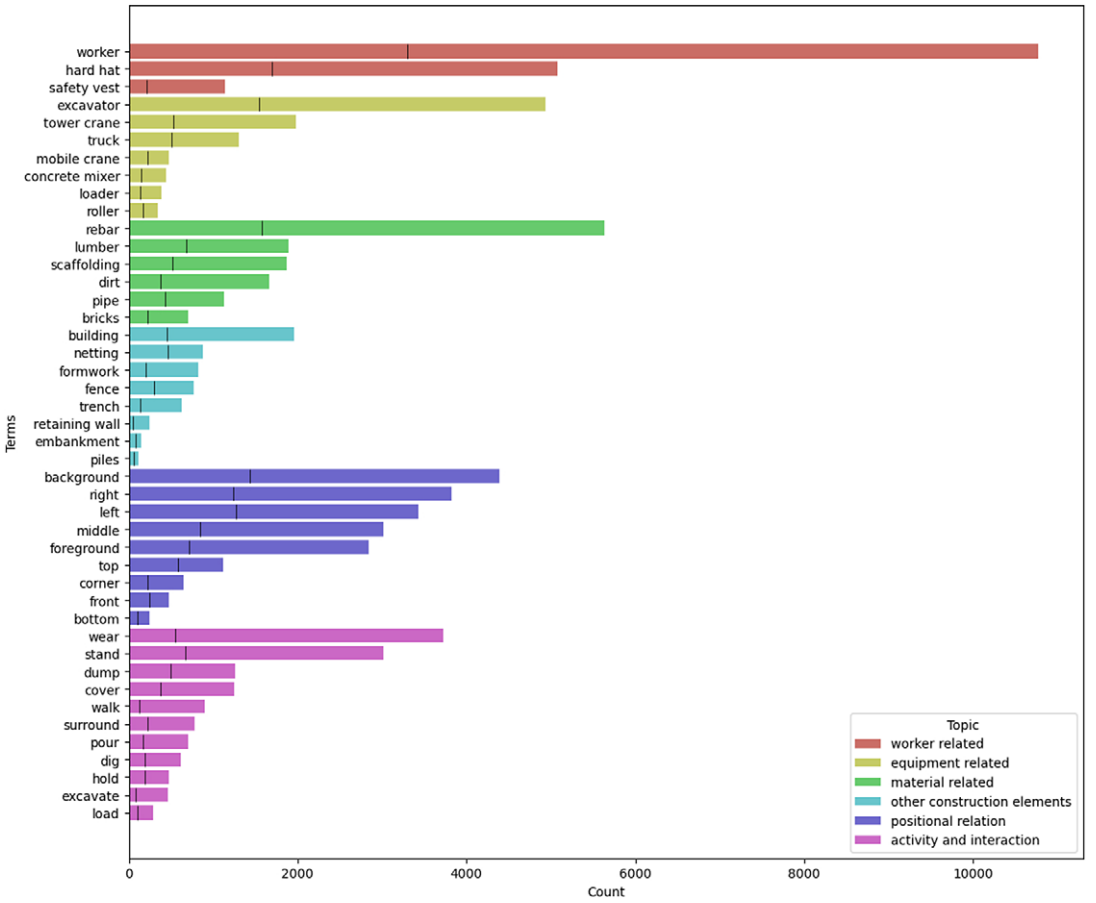


Figure 5. *Word frequency of prevalent terms across topic categories in the reference image captions. Synonymous and plural noun forms (e.g., concrete truck and concrete mixers) are consolidated under a single canonical term. Verb occurrences are aggregated at the lemma level. For each term, the bar shows the total frequency, with occurrences from the test split on the left and training split on the right, separated by a black vertical marker.*

Unlike traditional image captioning datasets, our annotations include not only the most visible objects in the foreground, but also construction elements scattered throughout the image, including those in the corners. Background details are also included as an attempt to capture every possible detail. A summary of frequently used words in the image captions is presented in Figure 5, showcasing the diverse types of objects and activities described in the dataset.

3.2.2. Safety rule violation VQA and object visual grounding

To further evaluate VLMs' understanding of construction sites and facilitate the training of VLMs for downstream applications, we introduce safety rule violation VQA annotations. This task plays a pivotal part in training and evaluating proficiency in processing an overwhelming amount of visual and textual information while still strictly following the prompt and generating human-readable responses. Prior works (Liu et al., 2022; Fang et al., 2023a) have identified about 10 common safety concerns at Chinese construction sites. However, differences between the datasets and the AI models used in their works and our article necessitated adjustments of safety rules to ensure relevance. We adopted three safety rules from their works: "edge protection," "usage of safety harnesses," and "workers within excavator operating

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Table 3. *Statistics of the annotations*

Image captioning	Test set	Training set
Image count	3,004	7,009
Sentence count	10,754	21,189
Word count	145,674	385,504
Safety rule violation VQA		
PPE	323	677
Safety harness	25	59
Edge protection	63	109
Blind spot	24	46
Object detection		
Excavator	1,080	2,415
Rebars	327	846
Workers with white hard hats	314	680

Note. For image captioning, the table reports the number of images, sentences, and words. For safety rule violation VQA and object visual grounding, it shows the number of ground truth annotations (occurrences)

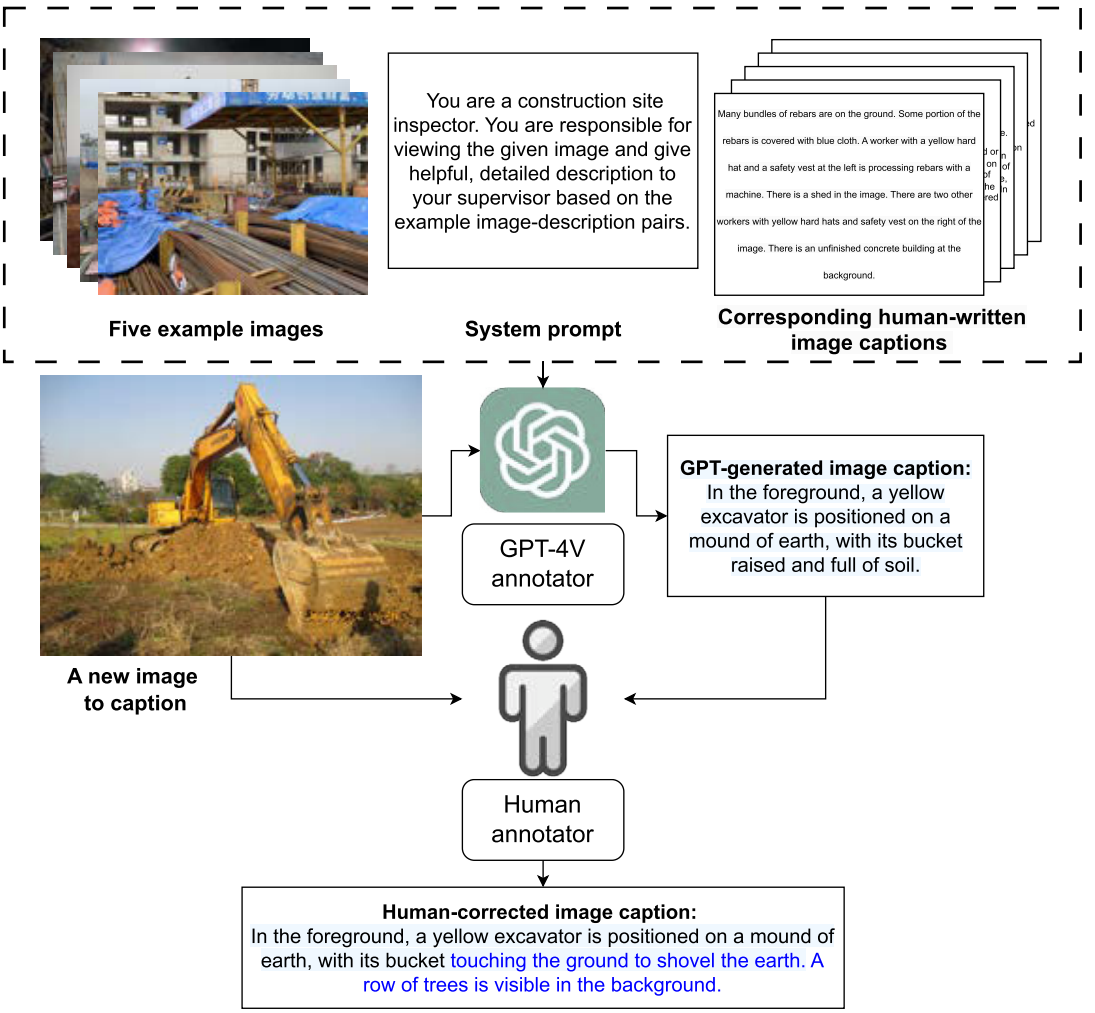


Figure 6. *The overall workflow of GPT-assisted image caption annotation.*

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Table 4. *Four safety rules used in the dataset*

Safety rule ID	Description of the rule
Rule 1	Use of basic PPE when on foot at construction sites. (hard hats, properly worn clothes covering shoulders and legs, shoes that can cover toes, high-visibility retroreflective vests at night, face shield or safety glasses when cutting, welding, grinding, or drilling).
Rule 2	Use of a safety harness when working from a height of three meters and the edges are without any edge protection.
Rule 3	Adoption of edge protection or edge warning, including guardrails, fences, for underground projects three meters in depth with a steep retaining wall and for humans to stand.
Rule 4	Appearance of workers in the blind spots of the operator and within the operation radius of excavators in operation, or excavators with operators inside.

zones," which align more with our dataset's context. To address regional specificity, we incorporated guidelines from WorkSafeBC (2024), which emphasize rigorous PPE requirements (e.g., full-body coverage for shoulders, legs, and toes). The four rules span PPE adherence, safety harness usage, edge protection, and human-machine interaction hazards. These categories collectively address the majority of *visually identifiable* safety risks in our dataset, balancing breadth and annotation feasibility. The rules are presented in Table 4. Each image annotation includes all observed violations of the safety rules, accompanied by one to two sentences explaining who and where the rule was violated, along with bounding boxes that ground the violators. This visual grounding task differs from traditional visual grounding since the object to ground is not given in the prompt, but needs to be determined by the VLMs based on the image, safety rules, and their prior answers.

3.2.3. Visual grounding

Given the dominant text-based annotations, we consider VLMs to be only employed at construction sites if they are competent enough to ground the required object precisely. Thus, we include the construction element visual grounding section in this dataset. The visual grounding annotations include the bounding boxes of three types of construction elements: excavators, rebars, and workers with white hard hats. These elements were chosen because they are commonly seen at construction sites and represent the three crucial construction components. These three objects present increasing levels of difficulty for VLMs. Excavators are regular in shape and large in size, while workers with white hard hats introduce a specific constraint within the broader category of workers. The statistics of these annotations are shown in Table 3.

4. Model selection

We selected six large pre-trained VLMs for testing, grouped into two categories: (a) larger proprietary models that represent state-of-the-art performance, and (b) smaller open-source models that are freely available, flexible, and locally deployable. The LLaVA series, MiniGPT-4, and GPT-4 series were considered state-of-the-art in 2024, while Gemini 2.5 and Qwen 2.5 represent the state-of-the-art in 2025.

LLaVA (Liu et al., 2024) The LLaVA variants employed in this article are LLaVA v1.5 13B and LLaVA v1.5 7B. LLaVA uses Vicuna v1.5 (Chiang et al., 2023) as the language encoder; for the vision encoder, LLaVA uses CLIP (Radford et al., 2021). The LLaVA v1.6 34B variant is used for the safety rule violation VQA test. It is accessed via the Gradio server API, where we cannot modify the system prompt.[2]

MiniGPT-4 (Chen et al., 2023; Zhu et al., 2023) The MiniGPT-4 variant employed in this article is MiniGPT-4 v2. The model uses an EVA (Fang et al., 2023b) as the vision encoder while the language

[2] The Gradio server is accessed through: https://llava.hliu.cc/.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

encoder is Llama2-chat-7B. The vision encoder EVA can be considered as a scaled-up and more powerful version of CLIP.

Qwen-2.5 (Bai et al., 2025) The Qwen variant used in this study is Qwen-2.5-VL (7B), selected for its state-of-the-art performance on multiple benchmarks in document understanding and VQA among open-source VLMs, as well as its strong video grounding capabilities. The model integrates a vision transformer as the vision encoder and the Qwen-2.5 language model (LLM) (Yang et al., 2025) as the language backbone.

GPT-4 (OpenAI et al., 2024) The GPT-4 variants employed in this article are GPT-4-1106-vision-preview (GPT4V) and GPT-4o-2024-05-13 (GPT4o). The model is chosen for its outstanding performance in image captioning and reasoning, however, we know little about its architecture, dataset, and training method.

Gemini-2.5 (Comanici et al., 2025) The Gemini-2.5 variant used in this study is Gemini-2.5-Flash, selected for its state-of-the-art performance in image captioning and reasoning. In addition, the model is reported to be further trained to convert images to structural representation and then perform image reconstructions, which may enpower it with enhanced spatial understanding capabilities.

Grounding DINO (Liu et al., 2023b) The Grounding DINO variant used in this article is GroundingDINO-B with Swin-B as backbone. The model is a specialized VLM designed for open-set object detection. Although it does not generate text outputs like other VLMs, it is expected to achieve superior performance on tasks related specifically to object detection. The model serves as the upper bound for the automated construction site object visual grounding task.

Traditional Computer Vision (CV) Baseline (Faster R-CNN) To provide a meaningful comparison with large pre-trained VLMs in the Object Visual Grounding task, we include a traditional computer vision baseline using Faster R-CNN (Ren et al., 2015) with a ResNet-50 backbone and Feature Pyramid Network (FPN). The ResNet-50 backbone was pre-trained on ImageNet, providing robust visual feature extraction. Since existing open-source models are not trained to detect challenging construction site objects (e.g., rebars) or specifically constrained objects (e.g., workers with white hard hats), we fine-tuned the model to enable detection of the categories required for our task. We trained the detection head (region proposal network and box/class predictors) from scratch on the training split of our Object Visual Grounding task. Fine-tuning consisted of 30 epochs using stochastic gradient descent (SGD) with a learning rate of 5×10^{-4}, momentum of 0.9, and weight decay of 1×10^{-4}. Data augmentation included resizing, padding, horizontal flips, color jittering, and small rotations. The resulting model serves as a quantitative baseline to evaluate the performance difference between large pre-trained VLMs and traditional computer vision models. The training procedures followed those in An et al. (2021). Since fine-tuning a CNN-based traditional model is not the primary focus of this article, we do not provide further details.

5. Model evaluation

5.1. Image captioning

We use one system prompt and one user prompt for image captioning. The system prompt is established prior to initiating each conversation. The system prompt is: "You are a construction site inspector. You are responsible for viewing the given image and give helpful and polite answers to your supervisor."

Each model then receives an image and a user prompt as input. The user prompt is: "Please describe the image. Your description should include information about people, construction equipment, and material stockpiles. Do not make assumptions, be concise, and describe facts only. Please describe with only one paragraph."

We conduct queries with GPT4V, GPT4o, Gemini-2.5-VL, and LLaVA 13B in a 5-shot setting. As demonstrated by previous work on in-context learning Min et al. (2022), few-shot prompting has a limited impact on the underlying knowledge of large pre-trained models. Nevertheless, few-shot examples can improve performance by encouraging more consistent and well-structured outputs. Accordingly, in this study, few-shot examples are not used as teaching signals for task adaptation or performance enhancement. Instead, they serve as format-conditioning templates to guide the models toward consistent output

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

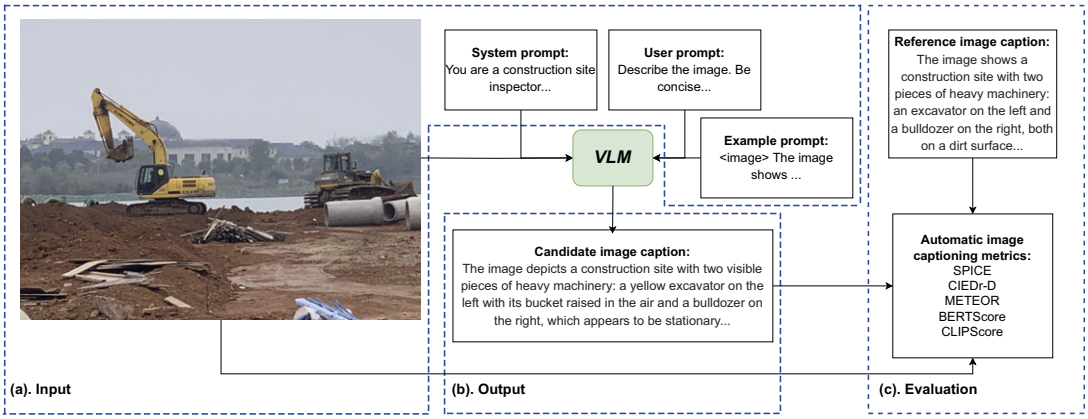


Figure 7. *Workflow of the image captioning task. The image, system and user prompts are given to the VLMs as inputs, the models generate an image caption as an output. The example prompts will be used to give examples to the VLMs in few-shot settings. The generated candidate caption, human-labeled reference caption, or the image is then evaluated with automatic metrics.*

structures, thereby boosting performance with more homogeneous response styles. The five few-shot examples are randomly sampled from the training set to avoid information leakage from the test set. For GPT-4 and Gemini models, which support inputting multiple images within a single conversation, we provide five images along with their corresponding human-annotated captions as example prompts before testing the model with the testing image. For LLaVA models, however, only one image can be input per conversation. Therefore, we input five captions without corresponding images for in-context learning, inspired by Alayrac et al. (2022).

The workflow of the task and the evaluation scheme is presented in Figure 7. For GPT and Gemini models, we use a single seed due to inference cost. For open-source models, we employ five different seeds, and all reported results represent the average across these five seeds. The model's hyperparameters are set as follows: temperature = 0.2, top_p = 1.0, max_token = 1024, chosen because they are either default values or closely resemble them for the employed models.

The evaluation metrics used for the image captioning task are: SPICE, CIDEr-D, METEOR, BERT-Score, and CLIPScore, for their wide usage in evaluating machine-generated image captions and superior correlation with human judgment[3].

SPICE (Anderson et al., 2016) This metric parses the caption to scene graphs, which are represented by tuples, and then calculates the f1 score of candidate and reference tuples. The types of tuples employed by this metric include objects, attributes, and relationships. By converting captions into tuples, the metric focuses exclusively on the semantic meaning while ignoring the quality of text.

CIDEr-D (Vedantam et al., 2015) The metric calculates the similarity of the candidate caption and the consensus of one or multiple reference caption(s). It is achieved by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram. The equation checks the occurrence of an n-gram in a reference caption or in the candidate caption. The "TF" term places higher weights on frequently occurring n-grams describing one image, while the "IDF" term reduces the weights of popular words across the entire dataset since those words usually contain less unique information.

METEOR (Denkowski and Lavie, 2014) This metric exhaustively maps words in the candidate caption to the reference caption according to four matchers, one after another: exact token matching, stem tokens, synonyms, and then paraphrases, followed by yielding a score. The total number of matched words can be obtained.

[3] Results of rule-based metrics SPICE, CIDEr-D, and METEOR are obtained using the official tools provided by MSCOCO image caption competition for comparison purposes: https://github.com/tylin/coco-caption.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

BERTScore (Zhang et al., 2020) The metric makes use of an LLM to encode text to contextual embeddings and calculate the cosine similarity between candidate and reference embeddings. For this article, we used RoBERTa-large as the embedding model. The results are rescaled with the baseline. For a candidate caption and a reference caption, the pairwise cosine similarity is calculated between the embedding of their tokens. The maximum cosine similarity scores are then used to calculate the precision, recall, and f1 score.

CLIPScore (Hessel et al., 2021) The metric employs CLIP (Radford et al., 2021) to encode the images and caption texts to visual embedding and text embedding, and a cosine similarity is calculated between embedding from the two modalities followed by multiplying the result with a weight $w = 2.5$. The metric enables evaluating candidate captions without references.

5.2. Safety rule violation VQA

This section consists of three tasks: multi-label classification, reasoning, and visual grounding, each detailed in Section 3. For the model evaluation, we derive a three-stage framework: the images and prompts are given to the VLMs at Stage 1; the violated rules selected by the VLMs will be checked at Stage 2; for correctly selected violations, the reasoning and bounding boxes will be evaluated at Stage 3. The workflow of the task and evaluation scheme is presented in Figure 8.

Prior to beginning these tasks, we set the hyperparameters of the VLMs identical to those in the image caption task. The new system prompt is formulated as follows to ensure a structured response: "You are a construction site safety inspector. Your responsibility is to review the provided image and provide helpful, detailed, and polite responses to your supervisor. You should only answer questions posed by the supervisor in the exact manner requested."

The model then receives a user prompt detailing the four safety rules, requesting a choice for the ID of the violated rule, an explanation for the choice, and a bounding box to ground the explanation. Additionally, we experimented with the CoT technique by querying the model independently for each of the four rules.

Due to the complexity of the proposed VQA task—which involves multi-label classification, reasoning, and bounding box generation—the correctness and consistency of answer formatting are critical for reliable evaluation. For this reason, and consistent with the rationale described in Section 5, we adopt a few-shot setting for the models of GPT4, Gemini, and LLaVA 34B as these models have sufficient context length to accommodate complex example prompts. As explained in Section 5, few-shot examples are randomly selected from the training set. In the five-shot setting, one example contains no safety rule violation, while each of the remaining examples corresponds to a different safety rule violation. This design ensures coverage of the possible input–output structures encountered during model inference. In the one-shot setting, we select an image that violates the largest number of safety rules, providing a compact yet comprehensive template for output formatting.

For multi-label classification (i.e. choosing the violated rules), the evaluation metrics are precision and recall for each rule. For the visual grounding, we use IoU. Their equations are listed in Eqs (1), (2), (3):

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}, \tag{1}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}, \tag{2}$$

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ union}. \tag{3}$$

For evaluating the quality of reasoning, we used three criteria:

- Relevance: Whether the explanation adheres to the specific safety rule.
- Equivalence: Whether the explanation is talking about the same violation, in terms of object, and reason, as the ground truth.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

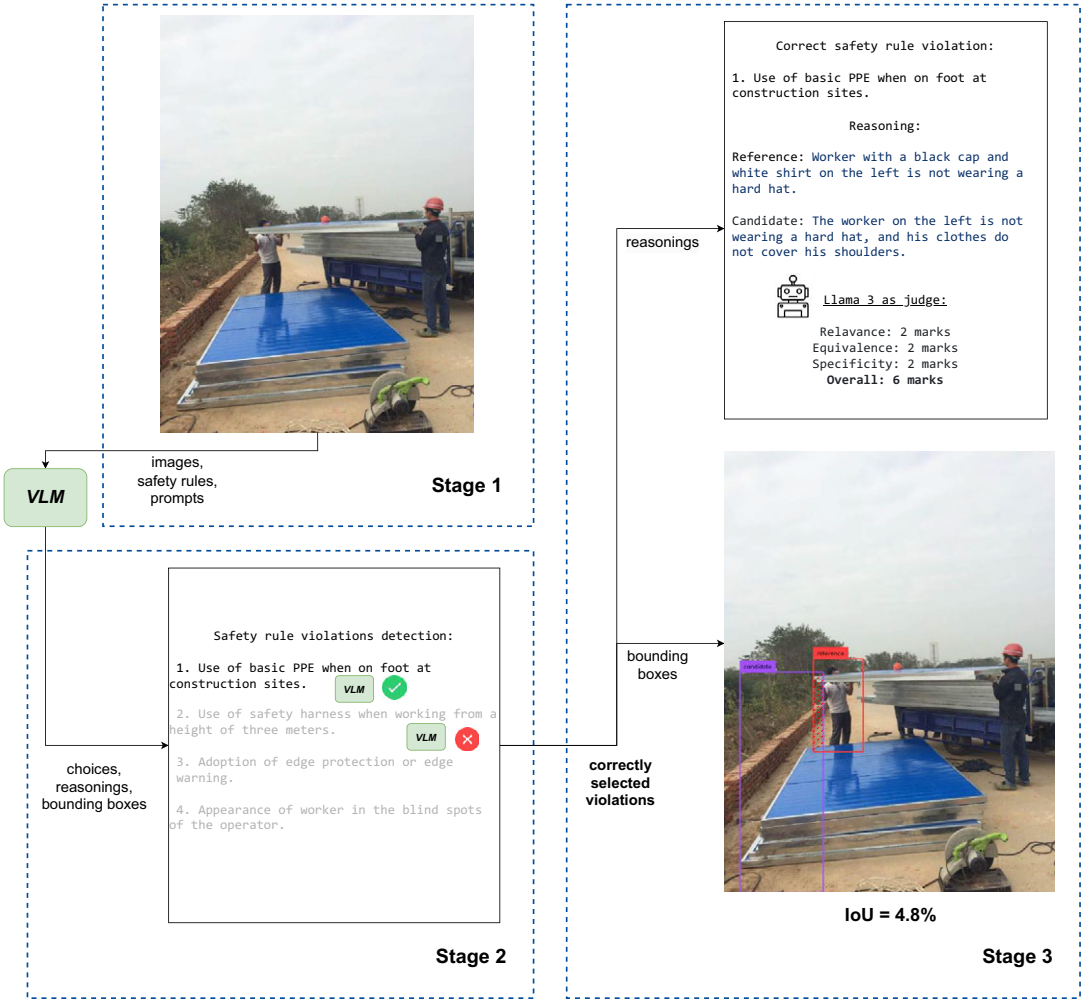


Figure 8. *A three-stage query and evaluation workflow for the safety rule VQA test. The images and prompts are given to VLM in* Stage 1. *The VLMs generate choices, reasoning, and bounding boxes as outputs. If the VLM selects the correct violations in* Stage 2*, the reasoning and bounding boxes will be evaluated in* Stage 3. *The safety rules depicted in the figure are simplified for clarity, with solid font indicating rules relevant to the image and grayed-out font indicating irrelevant rules. In this example, the VLM selects Rule 1 and Rule 2, but only Rule 1 is violated in the image. The reasoning for the violation of the correctly chosen safety rule (Rule 1) is then provided to Llama 3 as the candidate reasoning, alongside the reference reasoning. For the visual grounding, the candidate bounding box, marking the location of the violation, is compared with the reference bounding box. This comparison yields an IoU score.*

- Specificity: Whether the explanation pinpoints the specific violator by describing the location or attribute.

Drawing inspiration from (Hodosh et al., 2013; Zheng et al., 2023; Liu et al., 2023a) and to expedite the evaluation process, we employ another LLM, Meta Llama 3 8B Instruct (AI@Meta, 2024), as a judge. The judge assigns three integer marks ranging from 0 to 2 for each of three criteria (i.e., relevance, equivalence, and specificity): 0 indicates incapability for the criterion, 1 suggests attempting but not succeeding well, and 2 signifies acceptability for the criterion. Therefore, the maximum total score across

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

three criteria is 6. For a meaningful evaluation, we only assess reasoning and visual grounding for correctly selected violations.

Evaluation of reasoning for each rule was conducted separately. We provided the judge with three human judgment examples for the evaluation of each rule to create a three-shot setting (Zheng et al., 2023), and asked it to evaluate all the VLM reasonings for each rule. Spearman's rank correlation test and Pearson correlation test were employed to measure the correlations between human evaluation and judge evaluation (Macháček and Bojar, 2014). A resulting Spearman's coefficient $\rho = 0.83$, and a Pearson's coefficient $r = 0.91$ proves a strong monotonic and linear relationship between human evaluation and judge evaluation. To ensure consistent and deterministic model behavior, we utilize the beam search method with a fixed random seed of 20, set the number of beams to 5, and consistently use the same three-shot examples for each rule. The beam search method generates a sentence word by word, resembling a tree structure. At each step, a fixed number k (5 in our case) of candidate sentences (beams) are considered when selecting the next word, and the new k candidate beams are retained for the following word.

5.3. Visual grounding

Each model will be tasked with grounding three types of objects: excavators, rebars, and workers with white hard hats. These tasks will be presented to the models individually. The difficulty of these tasks increases progressively, because:

- Excavators are relatively large and have distinct features.
- Rebars vary in shape and can be easily confused with other construction materials.
- Workers with white hard hats restrict the detection target to a specific type of worker.

This progression in difficulty aims to challenge the models' capabilities in visual grounding across different object types.

The evaluation metric is the IoU of reference and candidate bounding boxes, which is explained in Section 5, and the equation is shown in Eq. (3). The overall workflow mirrors the visual grounding branch depicted in Figure 8.

6. Results and discussion

6.1. Resource usage

Experiments for open-source VLMs are conducted on an NVIDIA RTX4080 (16GB) GPU or NVIDIA Tesla V100 High Performance Computer. Proprietary models are running with the respective official API with a 900 Mbps download and 200 Mbps upload internet connection. Due to GPU memory constraints, open-source VLMs with more than 7B parameters need to be 4-bit quantized when running on RTX4080. We summarized the resource usage in Table 5. As the resource usage differences within large proprietary models and within smaller open-source models are minimal, we report results only for Qwen-2.5-VL and Gemini-2.5-Flash, which serve as up-to-date representative examples for 2025. While larger proprietary models provide stronger reasoning ability, smaller open-source VLMs offer a favorable trade-off between accuracy and scalability, making them more suitable for resource-constrained industrial adoption.

6.2. Image captioning

Examples of the image captioning task is presented in Figure 9. The result for the image captioning task is summarized in Table 6. For the three rule-based metrics (i.e. SPICE, CIDEr-D, METEOR), we present the SOTA scores from the MSCOCO dataset benchmarks for comparison, while for the CLIPScore, we present a human baseline. It should be noted that MSCOCO SOTA provide insight into the performance of models pretrained or fine-tuned with the MSCOCO dataset, making this a slightly unfavorable comparison for models trained using the *ConstructionSite 10 k* dataset. On one hand, captions in MSCOCO

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Table 5. *The reported resource usage reflects the consumption per image after completing one round of image captioning (5-shot), VQA (5-shot), and object grounding (for three object as per our test)*

	Inference latency per image (seconds)			Average resource consumption per image		
	Image captioning (5-shot)	VQA (5-shot for proprietary models)	Object grounding	Token usage	Cost (USD)	GPU usage (seconds)
Proprietary models	1.7	2.1	1.9	5651	0.0015	-
Open-source models	4.5	1.9	0.8	-	-	8.8

Note. The cost estimates are derived from the total number of tokens processed, with the note that proprietary models typically apply different pricing schemes depending on the type of token.

typically contain significantly fewer words on average compared to those in *ConstructionSite 10 k.* On the other hand, images in the *ConstructionSite 10 k* dataset include more semantic complexity, making it more challenging for models to capture all details accurately.

The results demonstrate that powerful models like Gemini-2.5-VL and GPT4o, tested under a 5-shot setting, achieve comparable or higher scores than the SOTA benchmarks from MSCOCO for the three rule-based metrics. Even without few-shot examples, the GPT4V model shows a strong ability to understand construction site images and effectively express its understanding, outperforming LLaVA models.

The outstanding performance of GPT4V 5-shot compared to the zero-shot setting demonstrates the undeniable in-context learning capacity of GPT models. Despite no changes to the image, text encoders, or decoder, these VLMs achieve higher scores by aligning their captions closely with human-annotated versions in terms of terminology, length, and format. This suggests that large pre-trained VLMs can quickly adapt to new tasks with minimal demonstrations.

In the *ConstructionSite 10 k* dataset, the Gemini model set a new SOTA record in the image captioning task. MiniGPT-4, LLaVA, and Qwen models, constrained by their smaller size and less extensive training, do not achieve comparable results to large proprietary models across any metric. Additionally, the LLaVA 13B model does not benefit significantly from the few-shot setting, possibly due to the lack of prompt images.

The CLIPScore, a model-based metric, offers the opportunity to compare image captions directly with images without relying on an intermediary such as a human-written caption. When evaluating images without any reference, all models surprisingly received high scores compared to human annotations. This suggests that all models generate high-quality image captions that exhibit high cosine similarities with the images. However, the limitations of text encoders in encoding long sentences, spatial relationships between objects, and numerical details (Kamath et al., 2023), coupled with potential gaps in training data related to construction sites, may affect the fidelity of CLIPScore.

VLMs tend to generate verbose descriptions even when prompted to be concise. The word count analysis reveals that models typically use twice as many words to describe an image compared to human annotators. Interestingly, GPT4o achieves the best overall performance with the fewest words among the models. This phenomenon may stem from harsh penalties imposed on candidate sentences that significantly exceed the length of reference captions.

6.3. Safety rule violation VQA

Before evaluation, we conduct post-processing on candidate captions to ensure they are readable, free of syntax errors, and uniformly formatted for fair automatic evaluation. We did not include testing results of

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

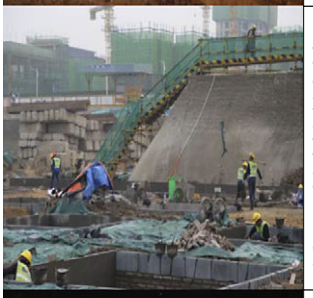
	Candidate: The image depicts a busy construction site with various activities in progress. There is a large crane in the center, several excavators, and trucks visible in the background. A dump truck is in the foreground with its bed raised, suggesting it is either depositing or about to deposit its load. There are large pipes laid out on the ground towards the right side of the image, indicating potential infrastructure work. One individual is visible walking near the center of the image, and the background shows construction cranes and partially constructed buildings, suggesting a large-scale development. The ground is uneven and covered with dirt, indicative of earthmoving operations.	Reference: The image shows a construction site with a variety of heavy machinery and equipment. In the center, a dump truck has its box tilted. A single worker is visible on foot near a water truck on the left. There are two excavators, two water trucks, and one blue crawler crane on the left. Large black pipes are laid out on the ground. A stack of green pipes is also visible. To the right, there is an orange bulldozer. In the background, there are two tower cranes visible against the skyline of residential buildings under construction. The ground is uneven with piles of dirt in the foreground.	BERTScore: 43.9 SPICE: 27.8 CIDEr-D: 187.7 METEOR: 37.3 CLIPScore: 85.0 Reference CLIPScore: 85.1
	Candidate: The image depicts an active construction site with several workers engaged in various tasks. In the foreground, there are at least four workers: one is bending over a pile of bricks, another is handling a wheelbarrow, and two others are working on or near a section of the foundation that is covered with a green tarp. There is construction material scattered around, including bricks, wheelbarrows, and buckets. In the midground, a worker wearing a yellow vest is walking across the site, and another is operating machinery near a large pile of sand or gravel, which is being moved via a conveyor belt system. In the background, there are scaffolding structures and a partially constructed building with green netting. A worker is visible on the elevated walkway of the scaffolding. The environment is overcast, and the ground appears muddy and uneven.	Reference: The image shows a construction site with at least seven workers visible, all wearing yellow hard hats and safety vests. In the foreground, two workers are in different foundation pits with concrete blocks and brick walls. There are piles of bricks, sand, and covered sandbags scattered around the site. In the background, a huge stack of precast concrete piles is on the left. Some wooden beams are on top of the piles. On the right, a staircase made of yellow and black metal poles and covered with green fabric connects the bottom of the excavation pit with the top of the concrete platform. A worker is standing on top of the concrete platform, with a roll of black ropes on his shoulder. There are construction buildings and three tower cranes in the far background.	BERTScore: 32.4 SPICE: 17.7 CIDEr-D: 130.1 METEOR: 33.9 CLIPScore: 80.4 Reference CLIPScore: 74.5
	Candidate: The image depicts a nighttime scene at a construction or demolition site with a large pile of debris in the foreground, consisting mainly of broken bricks and rubble. There is a Hyundai excavator to the right side of the image, with its arm extended over the debris pile, suggesting it is engaged in work. No people are visible in the image, indicating that the operator may be inside the excavator or out of the frame. The lighting is poor, and the visibility is limited due to the darkness. There are no other pieces of equipment or material stockpiles visible in the image.	Reference: The image shows a nighttime scene at a construction site with a Hyundai excavator positioned to the right, its arm extended over a pile of debris consisting mainly of broken bricks and rubble. The lighting conditions are low, with only the excavator's lights illuminating the area.	BERTScore: 44.5 SPICE: 41.5 CIDEr-D: 320.5 METEOR: 58.9 CLIPScore: 74.7 Reference CLIPScore: 74.3
	Candidate: The image depicts a construction site with a group of workers engaged in various activities. There are at least twelve individuals visible, most of whom are wearing hard hats. The workers are scattered across a rocky terrain with large stone blocks, some of which are being actively moved or positioned by the workers. In the background, there is a steep rock face, suggesting that the site may be in a mountainous or hilly area. There are no visible construction machines in the image, but there are steel reinforcing bars protruding from a concrete footing or foundation, indicating ongoing construction work. The workers appear to be manually handling the materials, and there is no clear indication of the specific construction project they are undertaking.	Reference: The image shows a group of workers at a rocky construction site. There are approximately 20 individuals visible, most of whom are wearing hard hats in various colors such as yellow and red. Some workers are standing, observing, or conversing, while others are actively engaged in manual labor, handling and positioning large rocks. The background features a steep, rocky hillside with trees on the hillside. An orange pipe can be seen among the large group of workers.	BERTScore: 41.7 SPICE: 24.1 CIDEr-D: 205.9 METEOR: 37.5 CLIPScore: 91.2 Reference CLIPScore: 93.1

Figure 9. *The figure displays examples of candidate captions generated by the GPT-4 model, alongside reference captions and their evaluations (all scores are in %) for the image captioning task. For all evaluation metrics except the one assessing the reference caption, higher scores indicate a "better" candidate caption. While the five types of image captioning metrics assign different scores to the same caption, they generally exhibit a monotonic relationship. This means that if one metric assigns a higher score to a specific caption, it is likely that the others will also assign higher scores.*

MiniGPT-4 and LLaVA 7B due to their inconsistent ability to generate useful responses in VQA and object grounding tasks. Similarly, smaller open-source VLMs like LLaVA 13B and Qwen 7B with around 10B parameters were not tested in few-shot scenarios, as they struggle to leverage in-context learning and sometimes repeat examples in their answers.

Additionally, we conducted the same test with human participants who are graduate students in civil engineering, who received the exact same image-prompt pairs as the VLMs.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Table 6. *The evaluation results (in %) for image description with automatic metrics*

Model	SPICE	CIDEr-D	METEOR	BERTScore	CLIPScore	Average words per caption
GPT4V	18.2	120.5	33.7	28.0	82.2	119
GPT4V 5-shot	23.1	152.1	38.7	33.5	82.1	101
GPT4o 5-shot	24.9	169.9	39.4	37.7	81.8	80
Gemini–2.5 5-shot	26.2	183.0	38.6	38.9	81.4	69
MiniGPT4	15.5	73.2	28.3	31.1	77.9	60
LLaVA 7B	12.8	62.1	28.5	26.4	78.6	82
LLaVA 13B	14.6	69.2	29.8	27.8	79.6	84
LLaVA 13B 5-shot	15.1	70.5	30.8	27.2	76.5	101
Qwen 7B 5-shot	20.2	119.4	32.9	33.1	82.1	72
Average	19.0	113.3	33.4	31.5	80.2	85
Median	18.2	119.4	32.9	31.1	81.4	82
Human annotation	-	-	-	-	80.0	49
MSCOCO SOTA[4]	27.0	155.1	33.9	-	-	-

Note. Higher scores in SPICE, CIDEr-D, METEOR, and BERTScore indicate greater similarity between human-annotated and VLM-generated descriptions. In CLIPScore, higher scores indicate greater similarity between machine-generated descriptions and images. MSCOCO SOTA scores for the three rule-based metrics are included to provide context but are not directly comparable with other scores in this table. The best-performing VLM results are underlined.

The testing results are presented in Tables 7 and 8. To provide a clearer visualization of the results and evaluations, examples of safety rule violation reasoning are presented in Figure 10.

High precision scores across all rules for human performance validate the ground truth. However, relatively lower recall rates indicate that construction engineers may overlook some dangers at construction sites, even when violations of safety protocols are obvious.

Table 7. *The result (in %) of the safety rule violation detection*

	0		1		2		3		4	
Rule ID	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall
GPT4V	97.0	24.6	20.4	76.4	2.7	85.7	5.3	87.7	3.7	84.0
GPT4V 5-shot	98.9	32.0	18.2	89.4	5.4	85.7	4.3	89.2	4.4	68.0
Gemini–2.5 5-shot	97.5	62.1	25.6	81.7	8.8	76.0	7.0	39.7	5.0	70.8
LLaVA 13B	88.0	43.4	12.0	54.0	1.0	42.9	2.2	33.8	0.9	44.0
LLaVA 13B CoT	91.0	44.0	12.0	55.0	3.0	52.0	6.0	18.0	2.0	32.0
LLaVA 34B 1-shot	87.1	88.7	14.3	13.0	3.0	33.3	4.5	21.5	5.2	20.0
Qwen 7B	91.2	42.3	12.9	71.5	2.3	48.0	3.4	39.7	1.7	54.2
Average	93.0	48.2	16.5	63.0	3.7	60.5	4.7	47.1	3.3	53.3
Median	91.2	43.4	14.3	71.5	3.0	52.0	4.5	39.7	3.7	54.2
Human upper bound	95.3	98.3	95.6	66.6	92.0	92.0	59.4	60.3	81.0	65.4

Note. Rule 0 indicates no violation and therefore cannot co-exist with other rules, whereas other rules can co-exist within the same image. The models answer the question in a multi-label classification setting. The best-performing VLM results are underlined.

[4] MSCOCO SOTA results are obtained from website: https://paperswithcode.com/sota/image-captioning-on-coco-captions.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Table 8. *The table presents reasoning results for safety rule violation VQA*

	1			2		
Rule ID	Violations detected (out of 323)	Average score from LLM judge	IoU	Violations detected (out of 25)	Average score from LLM judge	IoU
GPT4V	246	4.5	11.9	18	5.7	12.1
GPT4V 5-shot	288	4.7	14.0	18	5.7	13.1
Gemini–2.5 5-shot	264	5.0	3.7	19	5.7	2.5
LLaVA 13B	174	2.9	20.0	9	5.0	17.9
LLaVA 13B CoT	177	3.1	6.9	11	1.0	10.7
LLaVA 34B 1-shot	42	3.9	-	7	5.9	-
Qwen 7B	231	3.1	23.1	12	5.9	40.1
Average	203	3.9	13.3	13	5.0	16.1
Median	231	3.9	13.0	12	5.7	12.6
Human upper bound	215	5.2	-	23	5.9	-

	3			4		
	Violations detected (out of 63)	Average score from LLM judge	IoU	Violations detected (out of 24)	Average score from LLM judge	IoU
-	57	5.6	30.6	21	5.7	10.0
-	58	5.7	32.6	17	5.9	22.5
-	25	5.7	4.9	17	5.9	9.4
-	22	3.7	8.3	11	4.6	21.5
-	12	4.8	0.5	8	1.1	0.0
-	14	5.3	-	5	5.6	-
-	25	5.1	25.0	13	5.8	25.8
-	30	5.1	17.0	13	5.0	14.9
-	25	5.3	16.7	13	5.7	15.8
-	38	5.8	-	17	6.0	-

Note. "Violations detected" denotes the number of images in which the respective safety rule violations were correctly identified. "Average score from LLM Judge" ranges from 0 to 6. Intersection over Union (IoU) is reported as a percentage (%). The best-performing VLM results are underlined.

For the violation detection, the high recall rates and significantly lower precision of all VLMs suggest that the models tend to be conservative and may overshoot in identifying safety violations. Additionally, the results indicate that the chain-of-thought method does not significantly improve model performance when tests do not involve strong sequential dependencies, as demonstrated in previous studies (Hessel et al., 2023; Liu et al., 2023a). A closer examination of numbers reveals that the models perform relatively worse on safety rules 2 and 4 compared to human upper bounds. Safety rule 2 requires workers working at heights to wear safety harnesses, while safety rule 4 requires workers to stay clear of an excavator's operating radius. Both rules demand an understanding of spatial relationships within the image. Specifically, safety rule 2 requires the VLM to determine whether a worker is working at height, whereas safety rule 4 requires assessing whether a person is near the excavator and reasoning about the potential risk of being hit. Humans—especially expert civil engineers—can easily arrive at the correct conclusion using

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

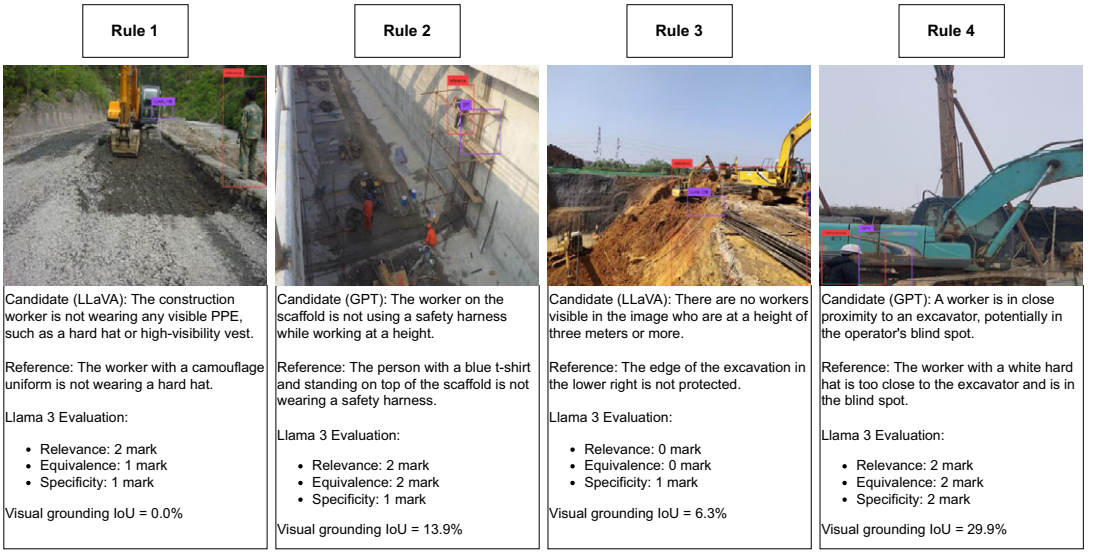


Figure 10. *Examples of safety rule violation reasoning. The figure includes the candidate, reference reasoning, and the evaluations.*

spatial reasoning and common sense. The results, however, indicate that the VLM still struggles with these types of spatial and logical reasoning tasks.

In terms of reasoning, only GPT and Gemini models approach the upper bounds set by human-level textual explanations. Open-source LLaVA and Qwen models, utilizing smaller language models, struggle to produce high-quality textual explanations. Regarding visual grounding, none of the models accurately pinpoint the exact location of violations, as evidenced by notably low IoU scores. The low IoU scores further reflect the limited spatial reasoning capability discussed above, particularly for the LLaVA and Gemini models, which are either smaller in model size or are not explicitly reported to be trained on object localization or bounding box generation tasks. These findings are visualized in Figure 10. In the worst cases, some models even fail to provide valid bounding box coordinates for evaluation.

Contrary to the intuition that "larger models perform better on everything," Qwen 7B demonstrates impressive visual grounding capacity in safety rule violation detection. This behavior may stem from Qwen's pre-training strategy, which uses absolute pixel coordinates for object localization (Bai et al., 2025). This approach likely enhances its visual grounding capability compared to models trained on normalized coordinates, as is common in visual instruction tuning (Liu et al., 2023a). We blame the poor grounding capacity of large proprietary models- Gemini and GPT series- on the absence of specific visual grounding training (OpenAI et al., 2024; Comanici et al., 2025).

We also acknowledge the limitations of the safety rule violation VQA:

- The four safety rules do not exhaustively cover all potential safety violations at construction sites.
- While humans can readily estimate distances in 2D images using visual comparisons and common sense, quantitative estimations (e.g., "three meters" in our work) remain challenging for VLMs, as their training data often lack precise and reliable spatial annotations.
- Localizing safety rule violations using bounding boxes may be imprecise, since a VLM can achieve substantial overlap with the ground-truth bounding box while still failing to accurately capture the true region or underlying nature of the violation. In particular, for geometrically complex cases—such as inclined or partially occluded workers—bounding boxes may overshoot the relevant area. More precise annotation formats, such as segmentation masks, could serve as a promising alternative in future work.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Table 9. *The IoU result (in %) of object detection results*

	Excavators		Rebars		Workers with white hard hats	
	Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged
IoU-Object Exist						
GPT4V	28.6	35.8	10.9	18.2	5.9	10.1
Gemini–2.5	60.8	60.0	3.2	4.8	8.6	5.9
LLaVA 7B	31.9	34.7	8.4	13.4	14.0	15.0
LLaVA 13B	47.7	54.5	14.9	19.0	19.5	25.4
Qwen 7B	64.7	75.4	4.0	10.7	32.1	31.5
Grounding DINO	63.9	71.0	15.6	15.3	12.9	11.2
Average	49.6	55.2	9.5	13.6	15.5	16.5
Median	54.3	57.3	9.7	14.4	13.5	13.1
Traditional CV model, Confidence threshold = 0.50	65.9	67.3	45.5	50.1	41.5	37.9
Traditional CV model, Confidence threshold = 0.75	69.9	69.7	34.6	32.4	37.8	32.6
IoU-Total						
GPT4V	22.3	29.3	5.7	11.6	3.1	9.1
Gemini–2.5	55.7	58.7	2.0	2.6	5.4	5.6
LLaVA 7B	11.5	21.0	1.0	3.1	1.5	3.8
LLaVA 13B	17.4	32.6	1.7	3.8	2.1	3.2
Qwen 7B	56.5	67.2	3.5	7.7	11.6	21.4
Grounding DINO	42.8	45.4	2.6	2.0	3.0	2.1
Average	34.4	42.4	2.7	5.1	4.4	7.5
Median	32.6	39.0	2.3	3.5	3.0	4.7
Traditional CV model, Confidence threshold = 0.50	51.1	53.9	27.5	43.5	18.4	25.1
Traditional CV model, Confidence threshold = 0.75	62.8	63.5	28.2	31.1	26.6	27.0

Note. The "IoU-Total" result is calculated across all images, while the "IoU-Object Exist" result is calculated on images where the object to ground exists. Macro averaging refers to calculating IoU for individual images first and then taking the average, micro averaging refers to dividing the summed intersection by the summed union. The macro averaging treats each image equally, while micro averaging awards correct predictions when the object is larger in the image.

6.4. Visual grounding

The result for visual grounding are presented in Table 9. The results consist of two sections. One section shows the result of all images, no matter the object to ground exist in the image or not, which means false positive cases (no object in the image but bounding box coordinates are given) and false negative cases (object presents in the image but bounding box coordinates are not given) will get an IoU of 0, we call it "IoU-Total." The other section only account for cases where the object to ground exists in the image, which we call "IoU-Object Exist." The equations showing the two types of IoU are shown in Eqs (4) and (5):

$$IoU - Object\ Exist = \frac{intersection\left(\widehat{X}, X \mid 1 \in X\right)}{union\left(\widehat{X}, X \mid 1 \in X\right)}, \tag{4}$$

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

$$IoU - Total = \frac{intersection\left(\widehat{X}, X\right)}{union\left(\widehat{X}, X\right)}, \tag{5}$$

where $\widehat{X}$ is the candidate binary map and X is the reference binary map. "IoU-Object Exist" demonstrates the ability of VLMs to precisely locate an object in an image by outputting bounding box coordinates, while "IoU-Total" also reflects their capability to determine whether the object exists. We will primarily examine "IoU-Object Exist" as our focus is on evaluating the current precision of the VLMs.

In the object visual grounding task, LLaVA and Qwen models also show comparable or higher performance than the GPT and Gemini model, which echoes the analysis of more spatial training data mentioned in Section 6. The lower accuracy of smaller open-source models on rebars is likely due to their limited exposure to such objects during pre-training; rebars are relatively uncommon in general image datasets and are predominantly found in construction-specific environments.

A closer examination of the results reveals clear trends. For unconstrained object categories such as excavators and rebars, earlier generative VLMs like GPT-4 V and LLaVA perform substantially worse than Grounding DINO, which is optimized for visual grounding. However, Grounding DINO's advantage diminishes when the task requires grounding more specific object types, where its performance begins to deteriorate. Notably, newer VLMs released in 2025—such as Gemini-2.5-Flash and Qwen-2.5-VL—show substantial improvements over their predecessors, with Qwen-2.5-VL (7B) achieving the best performance in our visual grounding evaluation. We also noticed that in most cases, the micro-averaged IoU equals or exceeds the macro-averaged IoU, indicating that VLMs typically achieve better results when objects are large in the images.

Overall, due to inferior performance in grounding rebars and workers wearing white hard hats, there remains substantial room for improvement for generative VLMs. Presently, generative VLMs cannot reliably function as visual grounding models at construction sites, particularly under realistic conditions where tasks impose additional constraints on attributes.

Examining the "IoU-Total" section reveals that when false positive cases are considered, the performance of LLaVA and Qwen models is significantly impacted. Specifically, LLaVA's performance in grounding excavators is halved, and there is a 90% decrease in grounding rebars and workers with hard hats compared to "IoU-Object Exist." Qwen also experienced a 60% drop in detecting workers with white hard hats. Since in the "IoU-Object Exist" cases, we assist the models in eliminating false positive predictions, this difference in results indicates the two open-source models tend to make bold guesses, resulting in the accidental capture of a substantial portion of the object. Thus, the high IoU rates observed for the LLaVA and Qwen models may be attributed to chance predictions, as they tend to provide predictions for all images. Conversely, the modest decline in the performance of GPT and Gemini demonstrates its cautious approach in making decisions.

For this task, we include a Faster R-CNN model as a baseline to compare the out-of-context grounding capabilities of large pre-trained VLMs—which are stronger in in-context text-based reasoning—with a CNN-based model specialized in object detection. We report IoU-based performance at two confidence thresholds: 0.5 and 0.75, meaning a predicted bounding box is considered valid if its confidence score exceeds the threshold. The 0.5 threshold provides an overall view of coverage, while the 0.75 threshold emphasizes higher-confidence predictions with more precise localization. As shown in Table 9, the traditional CV model achieves IoU scores approximately twice the average VLM IoU for excavator detection, and several times higher for rebars and workers with white hard hats, demonstrating significantly better bounding box grounding capability. However, this comparison may be somewhat biased due to the availability of task-specific training data for the Faster R-CNN model, while all large pre-trained VLMs are not fine-tuned on task-specific data. The traditional CV model's limited generalization to new tasks and unseen object types highlights the trade-off between specialized detection and the multi-modal reasoning capabilities of VLMs in construction site safety inspection.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press



Figure 11. *The image displays visual grounding examples from GPT model. Each row corresponds to a different object category: the first row shows excavators, the second row shows rebar detection, and the third row shows workers with white hard hats. The model excels in detecting larger objects but struggles with irregular shapes, such as rebar piles, and specific constraints, such as workers with white hard hats.*

To better visualize the performance, we presents some visual grounding results in Figure 11.

6.5. Implications for VLM design and selection in construction safety inspection

The evaluated models exhibit substantial performance variation across different tasks. Across experiments, we observe consistent distinctions between (i) large proprietary, instruction-tuned VLMs and (ii) open-source VLMs constructed by aligning vision encoders with large language models, with Qwen-2.5-VL as a notable exception.

Models with extensive multimodal pre-training and strong instruction tuning consistently perform better on language-intensive tasks. This trend is evidenced by the generally above-average and above-median performance of GPT and Gemini models in image captioning and safety rule violation detection and reasoning. Their strong performance suggests that semantic alignment, instruction-following capability, and robustness to long and complex prompts are critical for high-level safety understanding, coherent textual reasoning, and consistent output formatting. While these proprietary models are closed-source and their full architectural details are unavailable, their reported designs OpenAI et al. (2024) and Comanici et al. (2025) indicate that large model capacity, transformer-based vision–language encoders, massive multimodal pre-training, and subsequent instruction tuning jointly contribute to superior language-centric performance.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

In contrast, spatial grounding tasks—such as localizing safety rule violations—remain challenging for all evaluated VLMs. Average and median IoU scores are generally below 20%, with the exception of grounding excavators, which are visually salient objects with relatively simple geometry. Notably, Qwen-2.5-VL, which explicitly incorporates detection-oriented pre-training with accurate coordinates Bai et al. (2025), consistently outperforms other open-source models in IoU metrics and achieves grounding performance comparable to GPT-4 V and Gemini-2.5. This observation indicates that strong language reasoning alone does not guarantee accurate visual grounding. Instead, targeted spatial supervision can substantially enhance grounding performance, enabling smaller open-source models to match or exceed the grounding capability of much larger proprietary models.

These findings lead to a practical question for construction safety inspectors and researchers: *Which type of VLM is most suitable for a given inspection task?* Since all evaluated VLMs are transformer-based, differences in task performance are better explained by model scale, pre-training objectives, and the inclusion of spatial supervision rather than by architectural variations within the transformer family. Our results suggest that instruction-tuned multimodal models are preferable for reasoning-heavy, open-ended inspection tasks that require semantic understanding and natural language explanation. Conversely, tasks demanding precise geometric localization benefit more from models trained with explicit visual grounding data. Given the consistently limited grounding accuracy across VLMs, closed-set detection tasks may still be better served by traditional computer vision models, which remain more reliable for narrowly defined, single-purpose detection problems.

As new VLMs continue to emerge, their suitability for specific construction safety inspection scenarios can be anticipated based on their training objectives, degree of spatial supervision, and balance between vision and language capacity. We emphasize that these insights are derived from empirical trends observed in this benchmark rather than from exhaustive architectural ablation studies.

7. Conclusion

This article introduces *ConstructionSite 10 k*, a multi-modal dataset comprising 10,013 diverse construction site images and three different tasks designed for pre-training or fine-tuning purposes. The image captioning task provides detailed descriptions of construction scenes, encompassing background details and inter-object relationships. This task aims to train and evaluate VLMs on their understanding of construction site images. The VQA task features annotations identifying safety rule violations along with corresponding bounding boxes indicating the violation locations. The visual grounding task includes bounding boxes for three key construction elements: excavators, rebars, and workers wearing white hard hats. These annotations are intended to train and assess VLMs in tasks related to construction site inspection.

We evaluate the capability of popular off-the-shelf cloud-based and open-source large pre-trained VLMs to generate captions for construction site images without specific construction-related training. Performance assessment employs both rule-based and model-based metrics. In a five-shot setting, the top-performing Gemini model demonstrates superior image comprehension and in-context learning capacity compared to other smaller models.

In the safety violation VQA testing section, we devised a three-stage framework to assess whether models, without supervised training, can understand safety rules, identify violations, justify their answers with reasoning, and accurately draw bounding boxes to localize violation behaviors. We utilized Llama 3 as an evaluator to assess the quality of reasoning. Our findings indicate that for violation detection, most models exhibit higher recall than precision, whereas human experts typically show higher precision than recall. However, nearly all models struggle with accurately grounding violations using bounding boxes. In contrast, in the textual reasoning section, GPT models consistently scored above five marks, while LLaVA models can average around four marks.

For object detection, VLMs can perform well for simple objects such as "excavator" with an IoU above 30%. But the performance declines notably to below 20% for objects with irregular shapes or additional constraints, such as workers wearing white hard hats.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

The dataset and testing framework serve as a foundational resource for further training, fine-tuning, and application of large pre-trained VLMs in understanding construction site images and conducting inspection tasks. This infrastructure can effectively benchmark model performance, thereby enhancing model architecture and training strategies. Researchers in the future can benefit from avoiding the manual collection and annotation of datasets, while also being able to benchmark and compare their models against others. This article pioneers the use of large generative pre-trained VLMs to tackle multiple construction site inspection tasks in zero-shot and few-shot scenarios. It demonstrates the promising capability of these models, even without additional training, to handle construction-related tasks. However, they are still behind their human counterparts in terms of quality and consistency of performance. This research lays the groundwork for the widespread adoption of large pre-trained VLMs or their architectures as the preferred choice for construction inspection tasks.

This article acknowledges its limitations. We recognize a lack of diversity in the Safety Rule Violation VQA and Construction Element Visual Grounding sections of our dataset. The four safety rules included do not represent all possible safety violations that can occur at construction sites and other visual grounding data types, such as segmentation can be a more promising alternative for geometrically complex cases. Similarly, the number of constrained construction elements in our dataset may not exhaustively cover all types of construction elements. Moreover, we did not fine-tune (i.e., retrain) the large pre-trained models utilized here in our article. We anticipate that fine-tuning these models could potentially yield superior performance compared to the zero-shot or few-shot approaches employed. Additional fine-tuning on the image captioning task enhances the ability of VLMs to capture the semantic content of images and align them more effectively with corresponding texts. Similarly, fine-tuning on VQA and visual grounding tasks improves the models' performance on those respective tasks. Additionally, considering the substantial performance gaps between SOTA Gemini model and smaller-scale open-source models like LLaVA or MiniGPT-4, exploring methods to train smaller models to achieve comparable performance to Gemini models hold promise. This direction could prove cost-effective for applications in construction sites, where smaller models may offer practical advantages.

Abbreviations

CLIP	contrastive language-image pre-training
CNN	convolutional neural network
CoT	chain-of-thought
IoU	intersection over union
PPE	personal protective equipment
SOTA	state-of-the-art
VLM	vision-language model
VQA	visual question answering

Author contribution. Conceptualization: Z.Z.; Data curation: X.C.; Formal analysis: X.C.; Project administration: Z.Z.; Supervision: Z.Z.; Validation: X.C.; Visualization: X.C.; Writing—original draft: X.C.; Writing—review and editing: Z.Z.

Data availability statement. The dataset is publicly available at: https://huggingface.co/datasets/LouisChen15/ConstructionSite Chen (2025).

Funding statement. This work received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The author declare none.

AI statement. This work uses AI tools as part of the research method. Specifically, we benchmark current state-of-the-art AI systems' capabilities to understand safety as a concept in the context of construction. The specific AI models used include the GPT family and LLaVA.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

References

AI@Meta (2024) Llama 3 model card. Available at https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (accessed 25 March 2026).

Alayrac J-B, **Donahue J**, **Luc P**, **Miech A**, **Barr I**, **Hasson Y**, **Lenc K**, **Mensch A**, **Millican K**, **Reynolds M**, **Ring R**, **Rutherford E**, **Cabi S**, **Han T**, **Gong Z**, **Samangooei S**, **Monteiro M**, **Menick J**, **Brock A**, **Nematzadeh A**, **Sharifzadeh S**, **Binkowski M**, **Barreira R**, **Vinyals O**, **Zisserman A**, **Simonyan K** (2022) Flamingo: A visual language model for few-shot learning. In *NeurIPS*. https://openreview.net/forum?id=EbMuimAbPbs.

An X, **Zhou L**, **Liu Z**, **Wang C**, **Li P and Li Z** (2021) Dataset and benchmark for detecting moving objects in construction sites. *Automation in Construction 122*, 103482. https://doi.org/10.1016/j.autcon.2020.103482.

Anderson P, **Fernando B**, **Johnson M and Gould S** (2016) Spice: Semantic propositional image caption evaluation. In *ECCV*, Vol. 9909, pp. 382–398. https://doi.org/10.1007/978-3-319-46454-1_24.

Bai S, **Chen K**, **Liu X**, **Wang J**, **Ge W**, **Song S**, **Dang K**, **Wang P**, **Wang S**, **Tang J**, **Zhong H**, **Zhu Y**, **Yang M**, **Li Z**, **Wan J**, **Wang P**, **Ding W**, **Fu Z**, **Xu Y**, **Ye J**, **Zhang X**, **Xie T**, **Cheng Z**, **Zhang H**, **Yang Z**, **Xu H and Lin J** (2025) Qwen2.5-vl technical report. Preprint. https://arxiv.org/abs/2502.13923.

Brown T, **Mann B**, **Ryder N**, **Subbiah M**, **Kaplan JD**, **Dhariwal P**, **Neelakantan**, *et al.* (2020) Language models are few-shot learners. In *NeurIPS*, pp. 1877–1901, Vol. *33*. Red Hook, NY: Curran Associates, Inc.

Chan C-F, **Wong PK-Y**, **Guo X**, **Jack CPC**, **Chan JP-C**, **Leung P-H and 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.

Chen X (2025) Data for "Are large pre-trained vision language models effective construction safety inspectors". August 2025. Hugging Face, [dataset]. Available at https://huggingface.co/datasets/LouisChen15/ConstructionSite (accessed 25 March 2026).

Chen S and Demachi K (2021) Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. *Automation in Construction 125*, 103619. https://doi.org/10.1016/j.autcon.2021.103619.

Chen X, **Fang H**, **Lin T-Y**, **Vedantam R**, **Gupta S**, **Dollar P and Zitnick CL** (2015) Microsoft coco captions: Data collection and evaluation server. Preprint, arXiv:1504.00325.

Chen J, **Zhu D**, **Shen X**, **Li X**, **Liu Z**, **Zhang P**, **Krishnamoorthi R**, **Chandra V**, **Xiong Y and Elhoseiny M** (2023) Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. Preprint, arXiv:2310.09478.

Chiang W-L, **Li Z**, **Lin Z**, **Sheng Y**, **Wu Z**, **Zhang H**, **Zheng L**, **Zhuang S**, **Zhuang Y**, **Gonzalez JE**, **Stoica I and Xing EP** (2023) Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. Available at https://lmsys.org/blog/2023-03-30-vicuna/ (accessed 25 March 2026).

Comanici G, **Bieber E**, **Schaekermann M**, **Pasupat I**, **Sachdeva N**, **Dhillon I**, **Blistein M**, **Ram O**, **Zhang D**, *et al.* (2025) Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint, https://arxiv.org/abs/2507.06261.

Denkowski M and Lavie A (2014) Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Stroudsburg, PA: Association for Computational Linguistics.

Devlin J, **Chang M-W**, **Lee K and Toutanova K** (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, Stroudsburg, PA: Association for Computational Linguistics. Preprint. https://arxiv.org/abs/1810.04805; https://aclanthology.org/N19-1423/.

Duan R, **Deng H**, **Tian M**, **Deng Y and Lin J** (2022) Soda: A large-scale open site object detection dataset for deep learning in construction. *Automation in Construction 142*, 104499. https://doi.org/10.1016/j.autcon.2022.104499.

Fang W, **Peter EDL**, **Ding L**, **Xu S**, **Kong T and Li H** (2023a) Computer vision and deep learning to manage safety in construction: Matching images of unsafe behavior and semantic rules. *IEEE Transactions on Engineering Management 70*(12), 4120–4132. https://doi.org/10.1109/TEM.2021.3093166.

Fang Y, **Wang W**, **Xie B**, **Sun Q**, **Wu L**, **Wang X**, **Huang T**, **Wang X and Cao Y** (2023b) Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, Los Alamitos, CA: IEEE Computer Society. Preprint, arXiv:2211.07636; https://openaccess.thecvf.com/content/CVPR2023/papers/Fang_EVA_Exploring_the_Limits780_of_Masked_Visual_Representation_Learning_at_CVPR_2023_paper.pdf.

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

Gurari D, **Li Q**, **Stangl A**, **Guo A**, **Lin C**, **Grauman K**, **Luo J and Bigham JP** (2018) Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pp. 3608–3617, Los Alamitos, CA: IEEE Computer Society.

Hessel J, **Holtzman A**, **Forbes M**, **Le Bras R and Choi Y** (2021) 'Clips'core: A reference-free evaluation metric for image captioning. In *EMNLP*. Stroudsburg, PA: Association for Computational Linguistics.

Hessel J, **Marasović A**, **Hwang JD**, **Lee L**, **Da J**, **Zellers R**, **Mankoff R and Choi Y** (2023) Do androids laugh at electric sheep? Humor 'understanding' benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.41/.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Hodosh M, **Young P and Hockenmaier J** (2013) Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research 47*, 853–899. https://doi.org/10.1613/jair.3994.

Hong W, **Wang W**, **Lv Q**, **Xu J**, **Yu W**, **Ji J**, **Wang Y**, **Wang Z**, **Zhang Y**, **Li J**, **Xu B**, **Dong Y**, **Ding M and Tang J** (2024) CogAgent: A visual language model for GUI agents. In *CVPR*. Los Alamitos, CA: IEEE Computer Society. Preprint. https://arxiv.org/abs/2312.08914; https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10655402.

Jung Y, **Cho I**, **Hsu S-H and Golparvar-Fard M** (2024) Visualsitediary: A detector-free vision-language transformer model for captioning photologs for daily construction reporting and image retrievals. *Automation in Construction 165*, 105483. https://doi.org/10.1016/j.autcon.2024.105483.

Kamath A, **Hessel J and Chang K-W** (2023) Text encoders bottleneck compositionality in contrastive vision-language models. In *EMNLP*. Stroudsburg, PA: Association for Computational Linguistics.

Kazemzadeh S, **Ordonez V**, **Matten M and Berg T** (2014) ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*. Association for Computational Linguistics, pp. 787–798. https://doi.org/10.3115/v1/D14-1086.

Kim T, **Kim S**, **Chern W-C**, **Park S**, **Kim D and Kim H** (2025) Optimizing large vision-language models for context-aware construction safety assessment. *Automation in Construction 180*, 106510. https://doi.org/10.1016/j.autcon.2025.106510.

Li C, **Xu H**, **Tian J**, **Wang W**, **Yan M**, **Bi B**, **Ye J**, **Chen H**, **Xu G**, **Cao Z**, **Zhang J**, **Huang S**, **Huang F**, **Zhou J and Si L** (2022) mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *EMNLP*. Stroudsburg, PA: Association for Computational Linguistics. Preprint. https://arxiv.org/abs/2205.12005; https://aclanthology.org/2022.emnlp-main.488/.

Liu J, **Fang W**, **Peter EDL**, **Hartmann T**, **Luo H and Wang L** (2022) Detection and location of unsafe behaviour in digital images: A visual grounding approach. *Advanced Engineering Informatics 53*, 101688. https://doi.org/10.1016/j.aei.2022.101688.

Liu H, **Li C**, **Li Y and Lee YJ** (2024) Improved baselines with visual instruction tuning. In *CVPR*. Los Alamitos, CA: IEEE Computer Society. https://ieeexplore.ieee.org/document/10655294.

Liu H, **Li C**, **Wu Q and Lee YJ** (2023a) Visual instruction tuning. In *NeurIPS*, Vol. *36*. Red Hook, NY: Curran Associates, Inc.

Liu H, **Wang G**, **Huang T**, **He P**, **Skitmore M and Luo X** (2020) Manifesting construction activity scenes via image captioning. *Automation in Construction 119*, 103334. https://doi.org/10.1016/j.autcon.2020.103334.

Liu S, **Zeng Z**, **Ren T**, **Li F**, **Zhang H**, **Yang J**, **Li C**, **Yang J**, **Su H**, **Zhu J and Zhang L** (2023b) Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*. Cham, Switzerland: Springer. https://link.springer.com/chapter/10.1007/978-3-031-72970-6_3.

Lu P, **Mishra S**, **Xia T**, **Qiu L**, **Chang K-W**, **Zhu S-C**, **Tafjord O**, **Clark P and Kalyan A** (2022) Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*. Red Hook, NY: Curran Associates, Inc.

Macháček M and Bojar O (2014) Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, MD: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3336.

Marino K, **Rastegari M**, **Farhadi A and Mottaghi R** (2019) Ok-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*. Los Alamitos, CA: IEEE Computer Society.

Min S, **Lyu X**, **Holtzman A**, **Artetxe M**, **Lewis M**, **Hajishirzi H and Zettlemoyer L** (2022) Rethinking the role of demonstrations: What makes in-context learning work?, pp. 11048–11064. https://doi.org/10.18653/v1/2022.emnlp-main.759.

OpenAI JA, **Adler S**, **Agarwal S**, **Ahmad L**, **Akkaya I**, **Aleman FL**, **Almeida D**, **Altenschmidt J**, *et al.* (2024) GPT-4 technical report. Preprint, arXiv:2303.08774.

Radford A, **Kim JW**, **Hallacy C**, **Ramesh A**, **Goh G**, **Agarwal S**, **Sastry G**, **Askell A**, **Mishkin P**, **Clark J**, **Krueger G and Sutskever I** (2021) Learning transferable visual models from natural language supervision. In *ICML*. PMLR. https://proceedings.mlr.press/v139/radford21a.html.

Radford A, **Wu J**, **Child R**, **Luan D**, **Amodei D and Sutskever I** (2019) Language models are unsupervised multitask learners. Available at https://api.semanticscholar.org/CorpusID:160025533 (accessed 25 March 2026).

Ren S, **He K**, **Girshick R and Sun J** (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In CortesC, LawrenceN, LeeD, SugiyamaM and GarnettR (eds), *Advances in Neural Information Processing Systems*, Vol. *28*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

Tsai WL, **Lin JJ and Hsieh S-H** (2023) Generating construction safety observations via CLIP-based image-language embedding. In KarlinskyL, MichaeliT and NishinoK (eds), *Computer Vision – ECCV 2022 Workshops*. Cham, Switzerland: Springer Nature, pp. 366–381.

US-BUREAU-OF-LABOR-STATISTICS (n.d.) Industries at a glance: Construction: Naics 23. Available at https://www.bls.gov/iag/tgs/iag23.htm (accessed 25 March 2026).

Vedantam R, **Zitnick C and Parikh D** (2015) CIDEr: Consensus-based image description evaluation, pp. 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087.

Wei J, **Wang X**, **Schuurmans D**, **Bosma M**, **Ichter B**, **Xia F**, **Chi E**, **Le Q and Zhou D** (2023) Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*. Red Hook, NY: Curran Associates, Inc. https://arxiv.org/abs/2201.11903; https://dl.acm.org/doi/10.5555/3600270.3602070.

WorkSafeBC (2024) Worksafebc. https://www.worksafebc.com/en/law-policy/occupational-health-safety

Xiao B and Kang S-C (2021) Development of an image data set of construction machines for deep learning object detection. *Journal of Computing in Civil Engineering 35*(2), 05020005.

Xiao B, **Wang Y and Kang S-C** (2022) Deep learning image captioning in construction management: A feasibility study. *Journal of Construction Engineering and Management 148*(7), 04022049. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002297.

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press

Xiong R and Tang P (2021) Pose guided anchoring for detecting proper use of personal protective equipment. *Automation in Construction 130*, 103828. https://doi.org/10.1016/j.autcon.2021.103828.

Yang A, **Yang B**, **Zhang B**, **Hui B**, **Zheng B**, **Yu B**, **Li C**, **Liu D**, **Huang F**, **Wei H**, **Lin H**, **Yang J**, **Tu J**, **Zhang J**, **Yang J**, **Yang J**, **Zhou J**, **Lin J**, **Dang K**, **Lu K**, **Bao K**, **Yang K**, **Le Yu ML**, **Xue M**, **Zhang P**, **Zhu Q**, **Men R**, **Lin R**, **Li T**, **Tang T**, **Xia T**, **Ren X**, **Ren X**, **Fan Y**, **Su Y**, **Zhang Y**, **Wan Y**, **Liu Y**, **Cui Z**, **Zhang Z and Qiu Z** (2025) Qwen2.5 technical report. Preprint. https://arxiv.org/abs/2412.15115.

Zhang T, **Kishore V**, **Wu F**, **Weinberger KQ and Artzi Y** (2020) BERTScore: Evaluating text generation with BERT. In *ICLR*. https://openreview.net/forum?id=SkeHuCVFDr.

Zheng L, **Chiang W-L**, **Sheng Y**, **Zhuang S**, **Wu Z**, **Zhuang Y**, **Lin Z**, **Li Z**, **Li D**, **Xing EP**, **Zhang H**, **Gonzalez JE and Stoica I** (2023) Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS*. Red Hook, NY: Curran Associates, Inc. Preprint. https://arxiv.org/abs/2306.05685; https://dl.acm.org/doi/10.5555/3666122.3668142.

Zhu D, **Chen J**, **Shen X**, **Li X and Elhoseiny M** (2023) MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*. https://arxiv.org/abs/2304.10592.

Zitnick CL and Parikh D (2013) Bringing semantics into focus using visual abstraction. In *CVPR*. Los Alamitos, CA: IEEE Computer Society.

Cite this article: Chen X and Zou Z (2026). Are large pre-trained vision language models effective construction safety inspectors. *Data-Centric Engineering*, 7, e11. doi:10.1017/dce.2026.10044

https://doi.org/10.1017/dce.2026.10044 Published online by Cambridge University Press