A Multi-Embodied-Agent-Based Dual-Level VLM Framework for Construction Safety Inspection

Chak Fu Chan[1], Du Shuang[1], Goh Yang Miang[1,*]

[1] Department of the Built Environment, National University of Singapore, Singapore
chak.fu.chan@u.nus.edu, duduss_08@u.nus.edu, bdggym@nus.edu.sg

Abstract –

Construction safety inspection remains a critical yet challenging task due to complex site environments, dynamic worker activities, and frequent visual occlusions. While recent advances in computer vision and vision-language models (VLMs) have enabled automated safety reasoning, most existing systems rely on single-agent, single-view observations, which are inherently vulnerable to occlusion and viewpoint bias. There is a lack of mechanism for consistent, reliable multi-agent interpretation and visual grounding for VLMs. These limitations restrict the applicability in complex, safety-critical construction scenarios.

To address these challenges, this paper proposes a multi-embodied-agent, dual-level VLM framework for construction safety inspection. The framework is structured as a four-layer architecture comprising coordinated robotic data acquisition, view-level intelligence, view-to-global object mapping, and global-level intelligence. At the sensing level, multiple autonomous agents collaboratively collect synchronized, risk-aware, and diversity-driven multi-view visual data. At the perception level, conventional CNN-based object detection and tracking modules provide accurate localization and identity consistency, while lightweight view-level VLM perform localized safety assessments. At the reasoning level, cross-view object observations are aligned into globally consistent identities, enabling global-level VLM to conduct cross-time and cross-view consistency reasoning and generate reliable, explainable site-level safety decisions.

To validate the feasibility of the proposed approach, experiments are conducted on a public dataset to evaluate the view-level VLM component. Quantitative and qualitative results demonstrate that the trained model achieves strong safety compliance reasoning and interpretable outputs even with limited training data. Overall, the proposed framework provides a scalable and robust foundation for multi-agent, VLM-enabled construction safety inspection.

Keywords –
Construction safety inspection, Embodied artificial intelligence, Unmanned aerial vehicles (UAVs), Vision-Language Models (VLMs), Multi-agent system

1 Introduction

Ensuring worker safety on construction sites is paramount, as safety incidents can often lead to severe injuries, project delays, and financial losses [1]. Modern automation offers tools like computer-vision-based detection systems to assist inspections [2, 3], but these solutions based on traditional CNN models are typically confined to specific, pre-defined tasks [4]. Such approaches lack broader contextual understanding, for example, a purely vision-based system may detect missing helmets but cannot interpret how a missing guardrail and a worker's proximity to the floor edge together constitute a potential hazard scenario. Recently, multi-modal large language models (MLLM) have been widely adopted for multiple tasks upon their multi-modal contextual understanding, upon which vision-language models (VLM) can jointly reason over visual observations and textual knowledge to perform multiple types of tasks such as image captioning, visual grounding, and visual question answering (VQA) [5]. Compared to a traditional vision system, the VLM system can enable higher-level semantic understanding of complex scenes and safety regulations [4, 6, 7]. However, VLMs are less reliable for precise localization and tracking [4], making them unsuitable as standalone solutions for object-centric inspection.

In the meantime, unmanned aerial vehicle (UAV) based visual inspection has become a transformative approach in the monitoring of construction and infrastructure systems [8, 9]. UAVs can rapidly access hard-to-reach areas and, when equipped with cameras and sensors, provide real-time data to safety managers, often replacing manned vehicles with greater efficiency, safety, and lower cost. Based on UAV visual sensing, vision–language models (VLMs) can be embedded within the pipeline to perform rule-based safety reasoning on the captured visual data. However, traditional vision systems using only a single sensor view suffer significantly from occlusion and blind spots, which reduce detection and perception accuracy [10, 11]. The same applies to drone-

based inspection in construction and infrastructure contexts [12, 13], where occlusion, motion blur and blind spots can degrade detection performance unless complemented by additional viewpoints. This limitation motivates multi-view and coordinated UAV approaches to maximize coverage and detectability of objects and activities on the construction site [14, 15]. Nonetheless, most multi-UAV inspection strategies remain navigation-centric rather than data-centric, with limited consideration of the quality, diversity, or information gain of the visual data collected [16, 17]. Moreover, a systematic consistency mechanism is required to align multi-view observations when performing safety analysis.

To address these challenges in developing a multi-UAV-based VLM system for safety inspection, namely lack of data-centric inspection strategy, lack of view consistency across agents and lack of visual grounding in VLM, we propose the following three objectives in this paper:

(1) **Design a data-centric multi-agent inspection framework** that incorporates global path optimization and local viewpoint selection to cover large-scale, complex construction sites. It will adopt active vision mechanism [18] to adapt to scene complexity, worker density, and risk level, improving information gain and reducing perception uncertainty of the captured data;

(2) **Develop a dual-level consistency-aware VLM framework** that enables view-level safety assessment at individual robotic viewpoints while coordinating multi-view observations through a global-level intelligence to generate consistent, site-level safety decisions;

(3) **Integrate conventional computer vision modules with vision–language models** to compensate for the limited visual-grounding capability of VLMs, where CNN-based object detection and tracking provide accurate localization and identity consistency, allowing VLMs to perform object-centric safety reasoning.

By jointly addressing these objectives, this study combines the strengths of multi-view robotic sensing, object-level visual grounding, and language-based contextual reasoning. The proposed approach aims to enhance inspection reliability and support safety managers with interpretable and consistent decision-making across dynamic construction environments. Empirical work has been focused on the view level as a feasibility study in this paper. While UAVs are used in this study as an illustrative example, the framework can be further extended to accommodate a variety of field robots, including UGVs, for diverse construction inspection scenarios.

2 Proposed Framework

This study will propose a multi-embodied-agent, dual-level VLM framework for construction safety inspection, aimed at addressing the limitations of single-agent, single-view systems. By integrating multiple field robots (e.g., UAVs as an example in this study) with hierarchical perception and reasoning, the framework supports both real-time safety alerting and robust multi-view decision support under a human-in-the-loop paradigm.

The methodology is structured as a four-layer framework that separates data acquisition, local perception, cross-view integration, and global safety reasoning:

Level 1 - Embodied Robotic Layer focuses on collaborative site inspection, where multiple robots navigate predefined checkpoints and collect visual data from diverse viewpoints.

Level 2 - View-Level Intelligence performs object-centric perception and preliminary safety reasoning on individual views, enabling the detection of immediate violations and on-site warnings.

Level 3 - View-to-Global Mapping aligns object observations across viewpoints, establishing consistent global identities for objects in cross-view integration.

Level 4 - Global-Level Intelligence conducts multi-view, object-centric, and consistent reasoning at each checkpoint to generate final safety compliance decisions and recommendations for safety managers.

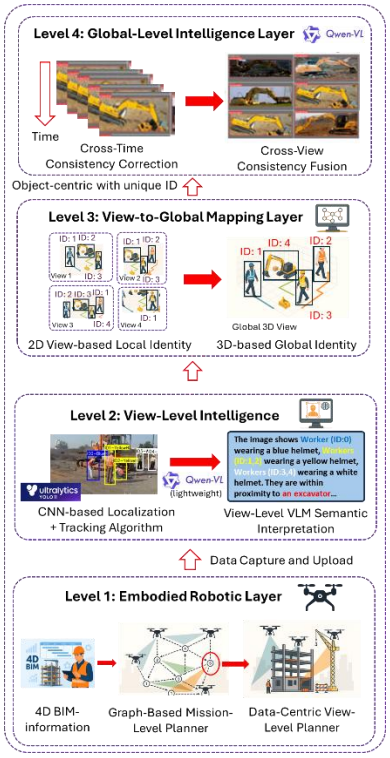

Figure 1 Overview of the four-layer, multi-embodied-agent, dual-level VLM framework

2.1 Embodied Robotic Layer

The embodied robotic layer focuses on coordinated

site inspection using multiple UAVs. A hierarchical path planning strategy is adopted to decouple mission-level routing from view-level sensing, enabling both efficient site coverage and data-centric visual acquisition for safety inspection. To ensure consistency across multiple UAV observations, the inspection process follows a checkpoint-based synchronization strategy rather than continuous real-time inter-UAV localization.

2.1.1 Mission-Level Planner: Graph-Based Checkpoint Routing

At the mission level, the construction site is represented as a graph: $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where nodes $\mathcal{N}$ correspond to predefined inspection checkpoints derived from the building information model (BIM) and site layouts, and edges $\mathcal{E}$ represent feasible travel paths. Each edge is weighted by the travel distance d_{ij} between checkpoints n_i and n_j.

The mission-level planner determines an inspection sequence that minimizes total travel distance:

$$\min_{\pi} \sum_{k=1}^{N-1} d_{\pi_k \pi_{k+1}} \tag{1}$$

Where $\pi = (n_{\pi_1}, n_{\pi_2}, \dots, n_{\pi_N})$ denotes the ordered checkpoint sequence. All UAVs are assumed to move synchronously between checkpoints, ensuring temporally aligned multi-view observations at each inspection location.

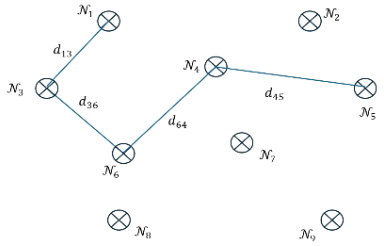

Figure 2. Graph-based mission-level path planning problem

2.1.2 View-Level Planner: Active Viewpoint Selection

At each checkpoint, a view-level planner selects a set of camera viewpoints to maximize visual coverage, visibility, and multi-view complementarity. Using prior coarse 3D reconstruction and 4D-BIM information, the local inspection region Ω_i around a checkpoint n_i is divided into a set of coarse cells:

$$\Omega_i = \{c_1, c_2, \dots, c_M\} \tag{2}$$

Each cell c_m is associated with **Risk density** $R(c_m)$ **and Object density** $O(c_m)$,estimated from historical data, BIM annotations, or real-time information.

Each cell c_m contains a set of points $P_m = \{\, p_j \in \mathbb{R}^3 \mid j = 1, \dots, \mathrm{n}_m \,\}$ that are used for fine-grained line-of-sight analysis. A viewpoint v_k is defined as:

$$v_k = \{\mathrm{x}_k,\ \mathrm{q}_k\} \tag{3}$$

Where $\mathrm{x}_k \in \mathbb{R}^3$is the 3D position of the UAV camera, $\mathrm{q}_k$denotes the camera orientation (e.g., yaw-pitch-roll or unit quaternion).

For each candidate viewpoint v_k and point p_j, visibility is defined as:

$$V(v_k, p_j) = \begin{cases} 1, & \text{if } \begin{array}{l}\text{(i) line-of-sight exists} \\ \text{(ii) } p_j \text{ lies within FOV} \\ \text{(iii) } \| p_j - \mathrm{x}_k \| \le d_{\max} \\ \text{(iv) Otherwise}\end{array} \\ 0, & \end{cases} \tag{4}$$

Where $d_{\max}$is the maximum effective sensing distance of the onboard camera.

To avoid double-counting overlapping observations, each point contributes to information gain only once, regardless of how many viewpoints observe it:

$$C(p_j) = \mathbb{I}\left(\sum_{v_k \in \mathcal{S}_i} V\,(v_k, p_j) \ge 1 \right) \tag{5}$$

The total visual information gain at a checkpoint n_i is then defined as:

$$IG(\mathcal{S}_i) = \sum_{c_m \in \Omega_i} \sum_{p_j \in P_m} C\,(p_j) \cdot R(c_m) \cdot O(c_m) \tag{6}$$

Where $\mathcal{S}_i$is the selected set of viewpoints. This formulation can operate at point-level resolution for accurate occlusion handling, while prioritizing high-risk and high-object-density regions.

Point-Based Angular Diversity: To encourage complementary multi-view observations, angular diversity is also defined at the point level. For a point p_j jointly visible from two viewpoints v_k, v_l, the viewing angle θ between the two sight lines is computed:

$$\theta(p_j; v_k, v_l) = \arccos \frac{(\mathrm{p}_j - \mathrm{x}_k) \cdot (\mathrm{p}_j - \mathrm{x}_l)}{\| \mathrm{p}_j - \mathrm{x}_k \| \ \ \| \mathrm{p}_j - \mathrm{x}_l \|} \tag{7}$$

and diversity is measured as:

$$\theta \sin\left(\frac{\theta}{2}\right) \tag{8}$$

Where $\theta \in [0, 180°]$. It ensures angular diversity increases smoothly from 0 to 1 as viewpoints become more widely separated.

The total angular diversity term sums contributions over all jointly visible points and viewpoint pairs:

$$D(S_i) = \sum_{c_m \in\, \Omega_i} \sum_{p_j \in\, P_m} \sum_{v_k, v_l \in S_i, k<l} V\big(v_k, p_j\big) \cdot \tag{9}$$

$$V(v_l, p_j) \cdot \sin\left(\frac{\theta(p_j;\, v_k, v_l)}{2}\right)$$

Viewpoint Selection Objective: The view-level planner selects a fixed number of viewpoints by maximizing:

$$\max_{\mathcal{S}_i} \; IG(\mathcal{S}_i) + \lambda D(\mathcal{S}_i) \quad \text{s.t.} \mid \mathcal{S}_i \mid = N_i \qquad (10)$$

Where λ balances information gain and viewpoint diversity, and N_i is proportional to the estimated risk and object density at the checkpoint:

$$N_i = \left\lceil \alpha \sum_{c_m \in \Omega_i} R(c_m) \cdot O(c_m) \right\rceil \qquad (11)$$

Where α is the scaling factor.

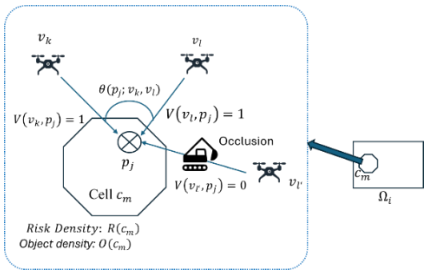

Figure 3. Viewpoint selection considering both information gain and viewpoint diversity

By combining mission-level graph-based routing with point-based, risk-aware, and diversity-driven view selection, the embodied robotic layer generates synchronized multi-UAV inspection trajectories that prioritize high-risk areas and produce high-quality, complementary visual evidence, which are uploaded to cloud server for downstream perception and vision–language reasoning.

2.2 View-Level Intelligence

The view-level intelligence layer operates independently on each UAV's sensory input. It is responsible for localized perception and preliminary safety reasoning at each robotic viewpoint, while preserving inter-view traceability of the observations. Its primary role is to transform raw visual data into object-centric, semantically enriched representations that can support both immediate safety alerting and consistent cross-view integration at higher levels of the framework.

2.2.1 CNN-based modules for object-centric representations

For data at each viewpoint, conventional computer vision modules are first employed to perform reliable object detection and tracking. CNN-based detectors (e.g., YOLO-family models) are used to identify safety-relevant entities such as workers, equipment, and hazardous site elements. Multi-object tracking (via algorithms like DeepSORT or unique identifiers by RFID) is then applied to maintain the temporal continuity of detected objects within each view. Each detected object o_i at each view is assigned a temporary local identifier and represented as an object-centric state:

$$o_i = \{c_i, b_i, \tau_i\} \qquad (12)$$

Where c_i denotes the object category, b_i the bounding box, and τ_i the temporal track ID. These object-centric representations provide precise localization and temporal consistency, compensating for the limited grounding capabilities of VLM.

2.2.2 View-Level Semantic Interpretation by Lightweight VLM

Building upon object-centric perception, a view-level lightweight VLM performs high-level safety reasoning for each viewpoint. Instead of processing raw images alone, the VLM is prompted with both the visual input and structured object information, enabling object-aware reasoning grounded in detected entities. For each object or object interaction at view v, the VLM generates a per-view semantic assessment:

$$s_i^v = \{\ell_i^v, e_i^v\} \qquad (13)$$

Where ℓ_i^v denotes the per-view inferred safety label (e.g., compliance/violation with certain rule), e_i^v represents supporting textual explanations or rule references. This structured output enhances transparency and facilitates downstream multi-view aggregation.

Based on the VLM's safety assessments, the view-level intelligence can issue real-time safety alerts when violations are detected, without waiting for global-level aggregation and validation. It will also output the object-centric records to the view-to-global mapping layer.

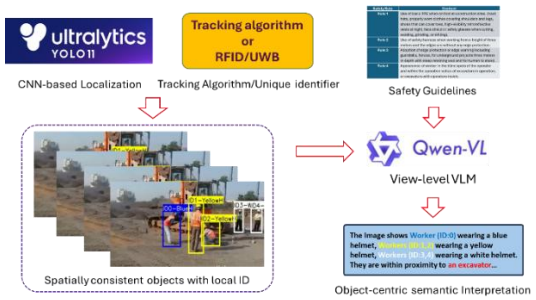

Figure 4 View-level intelligence layer

2.3 View-to-Global Mapping Layer

The view-to-global mapping layer aligns object observations from multiple viewpoints and assigns consistent global identities for object-centric reasoning. Let $\mathcal{O}_i^v(t)$ denote an object with a local identifier i detected in view $v \in \mathcal{V}$ at time t. The goal is to establish a global

mapping:

$$g: \mathcal{O}_i^v(t) \mapsto o_j,\ o_j \in \mathcal{O} \tag{14}$$

Where $\mathcal{O}$ is the global object set shared across all views.

Two complementary identity association methods can be introduced.

2.3.1 Method 1 - Spatial Data-Based Identity Association

For each locally tracked object $\mathcal{O}_i^v(t)$, a 2D object coordinate $\mathbf{p}_i^v(t) \in \mathbb{R}^2$ derived from the bounding box center is first back-projected to the global 3D world using UAV pose and depth estimate (e.g. from LIDAR):

$$\mathbf{X}_i^v(t) = \Pi^{-1}(\mathbf{p}_i^v(t), \mathbf{T}^v(t)) \tag{15}$$

where $\Pi^{-1}(\cdot)$is the camera back-projection function and $\mathbf{T}^v(t)$is the UAV pose.

Across time, each object yields a reconstructed 3D trajectory:

$$\mathcal{T}_i^v = \{\mathbf{X}_i^v(t_1), \mathbf{X}_i^v(t_2), \dots, \mathbf{X}_i^v(t_K)\} \tag{16}$$

Trajectories from different views are merged and clustered via spatial similarity. A pairwise trajectory distance is defined as:

$$d(\mathcal{T}_i^v, \mathcal{T}_{i'}^{v'}) = \frac{1}{K}\sum_{t=t_1}^{t_K} \| \mathbf{X}_i^v(t) - \mathbf{X}_{i'}^{v'}(t) \|_2 \tag{17}$$

Clustering assigns a **global identity**:

$$o_j \leftarrow \arg\min_j \ d(\mathcal{T}_i^v, \mathcal{C}_j) \tag{18}$$

where $\mathcal{C}_j$ is the cluster centroid representing the global object o_j. This method enables sensor-independent global identity assignment using only spatial geometry and UAV pose information.

2.3.2 Sensor-Assisted Identity Association

When workers or assets carry RFID/UWB tags, each detected local object directly inherits a global identifier:

$$g(\mathcal{O}_i^v(t)) = \text{TagID}(\mathcal{O}_i^v(t)) = o_j,\ o_j \in \mathcal{O} \tag{19}$$

Identity assignment becomes deterministic and eliminates cross-view ambiguity and bypasses clustering.

2.3.3 Global Object Record

After identity mapping, each global object o_j aggregates all connected view-level 2D observations:

$$\mathcal{R}(o_j) = \{(\mathbf{b}_i^v(t), \ell_i^v, t) \mid g(\mathcal{O}_i^v(t)) = o_j\} \tag{20}$$

where $\mathbf{b}_i^v(t)$ is the bounding box of the object with local identifier i from view v at time t, and ℓ_i^v is the object semantic label from view v. The resulting record is object-centric and globally consistent, forming the basis for downstream multi-view reasoning.

2.4 Global-Level Intelligence Layer

The global-level intelligence layer conducts object-centric, multi-view reasoning to produce consistent and explainable safety assessments at the site level. While the view-level intelligence provides an initial semantic label for each object trajectory, global-level reasoning should further improve reliability by (1) verifying or correcting the label with a cross-time consistency check for each view, and (2) reconciling multi-view semantic observations through cross-view consistency fusion. After that, the VLM should output final rule-based safety compliance decisions and recommendations for safety managers and support a human-in-the-loop process to inspect supporting visual evidence and intervene when necessary.

2.4.1 Cross-Time Consistency Correction

Given a global object o_j, each view provides a per-view semantic prediction regarding a specific rule:

$$\ell_j^v \in \{\text{Compliant, Non-compliant, Uncertain}\} \tag{21}$$

View-level VLM already processes per-view short video clips or multi-frame input and therefore embeds local temporal reasoning. However, transient visual artifacts may still cause misclassification at the view level. The view-level VLM is a lightweight one designed to enable efficient batched inference of multi-view inputs, but it is not suitable for more complex consistency checks. To address this, the global VLM performs an additional cross-time consistency check over the sequence of frames for each global object within a single view. The consistency check should enforce the following principles via prompting: **1) Persistent evidence dominates:** violations observed consistently across frames override brief compliant or unknown states**; 2) Momentary ambiguity does not negate compliance:** short ambiguous intervals do not alter otherwise consistent compliant states**; and 3) Unknown does not imply compliance:** missing evidence is treated as uncertainty rather than compliance. Based on the cross-time consistency check, the global-level VLM will output a corrected per-view semantic label $\tilde{\ell}_j^v$ for object o_j.

2.4.2 Cross-View Consistency Fusion

After temporal correction, multiple views $v \in \mathcal{V}_j$ produce labels $\tilde{\ell}_j^v$ for the same global object o_j. However, they may still exhibit inconsistency due to viewpoint-dependent occlusion, perspective distortion or uncertainty. Multi-view fusion aims to reconcile discrepancies with the following reasoning rules via prompting: **1) Agreement Rule:** consistent observations across views reinforce the decision**; 2) Override Rule:** explicit violation overrides compliant observations, with a confidence threshold for the override**; 3) Occlusion Rule:** missing or "occluded" views are treated as insufficient evidence

rather than compliance/violation**; and 4)Relevancy Rule**: prioritizes referencing the views that provide highest relevancy or strongest evidence, while discarding the views that are of low relevancy or quality. Based on the cross-view consistency mechanism, the global-level VLM can output the single and most reliable interpretation $\hat{\ell}(o_j)$ for each object's safety state.

2.4.3 Site-Level Safety Assessment with Human-in-the-Loop

Object-level assessments are aggregated to form a site-level view at the inspection checkpoint:

$$\mathcal{S}_{\text{site}} = \{\hat{\ell}(o_j)\}_{j=1}^{|\mathcal{O}|} \tag{22}$$

The global VLM outputs both: **1) Decision** (violation/compliance/uncertainty)**; and 2) Explanation** (domain rule + multi-view justification) enabling traceability and human-in-the-loop verification, which are critical for safety-critical applications. Human feedback can also be used to further refine and enhance the VLM via training methods like reinforcement learning.

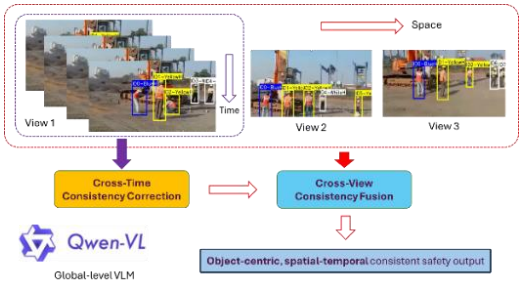

Figure 5. Global-level intelligence layer

3 Experiment and Validation

To empirically validate the effectiveness of the view-level VLM component within our framework, we conducted experiments using a subset of the *ConstructionSite* dataset developed by Chen et al [19], a multimodal benchmark for construction image understanding and inspection tasks. It comprises approximately 10,013 real construction site images, each richly annotated with captions, safety rule violation labels, and bounding boxes for key objects such as excavators, rebars, and workers with hard hats. The dataset also exhibits diverse visual attributes (e.g., illumination conditions, camera distance, and view types), many of which resemble those encountered in UAV-based inspection scenarios.

We mainly validate the safety-checking performance of our developed VLM model via visual question answering (VQA) with quantitative and qualitative analysis. Four safety rules associated with the dataset are described in Table 1. These rules cover fundamental PPE compliance, fall protection during work at height, edge protection in excavation and underground works, and hazardous proximity between workers and operating excavators. Each image–question pair requires the model to assess visual evidence from a single viewpoint and determine whether the safety rule is complied with or violated.

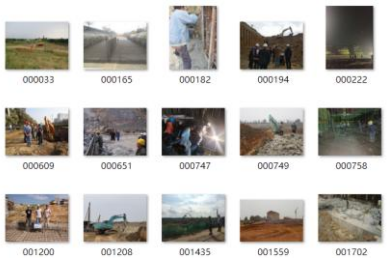

Figure 6. A snapshot of the *ConstructionSite*

Table 1. Description of safety rules for VQA

Safety Rules	Content
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 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 steep retaining wall and for human to stand.
Rule 4	Appearance of worker in the blind spots of the operator and within the operation radius of excavators in operation, or excavators with operators inside.

To evaluate the feasibility of developing a view-level VLM under limited data availability, 439 samples were selected for training, and 739 samples were selected for testing. This subset was chosen to maximize diversity in visual conditions and safety rule coverage, while the majority of remaining dataset samples contain repeated instances of same PPE compliance rules. Qwen3-VL-8B was selected as the base model, due to its decent multimodal reasoning and modest parameter number that enables batched inference of concurrent view-level inputs. All samples were treated as single-view inputs, consistent with the role of the view-level module in the proposed framework. The view-level VLM was trained using a two-stage learning strategy consisting of supervised fine-tuning (SFT) followed by Group Relative Policy Optimization (GRPO) -based reinforcement learning (RL), improving the model's ability in safety-related VQA.

3.1 Quantitative Analysis

To quantitatively assess the performance of the VLM in safety compliance checking, two metrics are defined:

Violation Recall: This metric measures the model's ability to correctly identify actual safety violations. It is computed on a per-rule basis as:

$$\boldsymbol{Recall_{vio}} = \frac{TP}{TP + FN} \quad (23)$$

Where TP = True Positive, number of rule violation cases correctly detected; FN = False negative, number of rule violation cases not detected.

Overall Accuracy: This metric reflects the model's general reliability in producing accurate compliance assessments and reducing false alarms, and it evaluates the correctness of all predictions under both rule "violation" and "no violation" cases. It is defined as:

$$\boldsymbol{Acc_{ovr}} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (24)$$

Table 2. Quantitative results of Safety VQA before and after the training

Metric	Before Training	After SFT + GRPO RL
Violation Recall	0.789	0.820
Overall Accuracy	0.702	0.906

The quantitative results in Table 2 demonstrate the effectiveness of domain-specific training for improving view-level safety reasoning. Before fine-tuning, the base model achieved an overall accuracy of **0.702**, indicating limited reliability in construction safety VQA. After training, the overall accuracy increased substantially to **0.906**, showing that the model is significantly better at producing correct safety compliance assessments.

Violation recall also improved from **0.789** to **0.820** after training, indicating enhanced sensitivity to actual safety violations, which is critical for safety inspection tasks. The relatively smaller increase in recall compared to overall accuracy suggests the training samples are still limited in number to capture the full complexity of violation scenarios. Moreover, certain violation cases may remain challenging to identify from a single viewpoint (e.g., spatial proximity between workers and nearby excavators may be hard to determine from several views). This reinforces the need for multi-view reasoning.

Nonetheless, the overall results still suggest that even with limited training data, targeted SFT and RL can effectively adapt a VLM to construction safety scenarios, supporting its feasibility for localized view-level safety assessment within the proposed framework.

3.2 Qualitative Analysis

Qualitative analysis is conducted to further examine the reasoning behavior of the developed view-level VLM in VQA of the safety rules. Figure 7 present a representative example in which the model is provided with a single construction site image and required to assess compliance with the four predefined safety rules. The model is prompted to explicitly identify relevant visual entities, associate them with corresponding safety rules, and produce structured rule-level compliance decisions.

For **worker1 [380, 760, 405, 920]:** He is wearing a yellow hard hat and long-sleeve shirt and pants; no safety harness visible but he is not at significant height without edge protection — Rule 2 not violated. Rule 1: PPE appears compliant.
For **worker2 [405, 770, 435, 930]:** Wearing blue shirt and pants, no hard hat visible — Rule 1 violation.
...
For **worker8 [505, 510, 535, 590]:** On scaffolding without visible guardrail or harness — Rule 2 violation.
...

Rule 1: Multiple workers lack hard hats (e.g., worker2 [405, 770, 435, 930], worker3 [440, 760, 470, 930], worker5 [580, 860, 605, 930]) — violation confirmed.
Rule 2: Multiple workers are at height without edge protection and no visible safety harness (e.g., worker6 [610, 440, 645, 550], worker7 [730, 330, 755, 400], worker8 [505, 510, 535, 590]) — violation confirmed.
Rule 3: No guardrails or edge protection visible along elevated edges where workers are present — but since workers are on scaffolding and not directly at unprotected floor edges, and no open excavation or unprotected floor openings are visible near them, Rule 3 is not violated.
Rule 4: No excavator visible in the image — no worker is within its operation radius — Rule 4 not violated."

Figure 7. Compliance output (partial) with object visual grounding by the trained VLM

As illustrated in the example, the VLM first identifies workers as key entities and grounds its reasoning on explicit bounding boxes for each detected worker. This object-centric representation enables the model to systematically inspect workers one by one, assess PPE usage, working-at-height conditions, and environmental context. The model demonstrates consistent reasoning by correctly identifying PPE violations (Rule 1) due to missing hard hats, as well as fall-protection violations (Rule 2) for workers operating at scaffolding without visible safety harnesses. At the same time, the model appropriately refrains from reporting violations for Rule 3 and Rule 4 when corresponding visual evidence such as unprotected excavation edges or excavators in operation is absent.

This example highlights several strengths of the

trained view-level VLM. First, it shows improved grounding between textual safety rules and visual evidence, supported by explicit localization of workers. Second, the step-by-step inspection process enhances transparency and interpretability, allowing safety managers to trace rule violations back to specific workers and visual cues. However, limitations still exist. In this example, pure VLM-based inference without explicit CNN module support fails to correctly ground one worker (highlighted by a red oval) located at the leftmost region of the image. This further justifies the integration of CNN-based detection and tracking modules in the proposed framework for accurate object localization and identity consistency before high-level vision-language reasoning is performed.

4 Conclusion and Future Work

This study proposes a multi-embodied-agent, dual-level VLM framework for construction safety inspection, addressing key limitations of existing systems, including single-agent perception constraints, multi-view inconsistency, and limited visual grounding of VLMs. Motivated by these gaps, the framework is designed around three research objectives and implemented through a four-layer hierarchical architecture.

At the data acquisition level, the Embodied Robotic Layer enables coordinated multi-agent inspection through risk-aware, diversity-driven active viewpoint planning. By synchronizing multiple agents at predefined checkpoints and prioritizing high-risk regions, the framework improves information gain, mitigates occlusions, and reduces perception uncertainty. At the cross-view integration and reasoning level, view inconsistency is resolved through the integration of View-Level Intelligence, View-to-Global Mapping, and Global-Level Intelligence, enabling localized object-centric perception and globally consistent object identities for cross-view and cross-time reasoning.

At the perception level, the limited visual grounding is mitigated by explicitly coupling CNN-based detection and tracking with VLM-based semantic reasoning, combining accurate localization and temporal continuity with contextual safety interpretation.Experimental results on a public construction safety dataset demonstrate that the view-level VLM component achieves effective safety compliance reasoning and interpretable outputs with limited training data, and provides a reliable foundation for the entire multi-view, multi-agent framework. In this work, only the view-level component is validated. Future work will focus on full-system validation using synchronized multi-agent datasets and real-world deployments, tighter human-in-the-loop integration, and extension to other construction monitoring tasks such as progress tracking and quality inspection.

References

[1] Zhou, Z., Y.M. Goh and Q. Li, Overview and analysis of safety management studies in the construction industry. *Safety Science*, 2015. **72**: p. 337-50.

[2] Guo, B.H.W., Y. Zou, Y. Fang, Y.M. Goh and P.X.W. Zou, Computer vision technologies for safety science and management in construction: A critical review and future research directions. *Safety Science*, 2021. **135**: p. 105130.

[3] Cai, R., X. Liu, Y. Tan, J. Li, J. Tang, D. Tang, and X. Chen, Vision-based multi-site safety inspection path planning using multi-agent system in large-scale work environment. *Expert Systems with Applications*, 2026. **303**: p. 130663.

[4] Chan, C.-F., P.K.-Y. Wong, X. Guo, J.C. Cheng, J.P.-C. Chan, P.-H. Leung, and X. Tao, *Context-aware vision-language model agent enriched with domain-specific ontology for construction site safety monitoring. Automation in Construction*, 2025. **177**: p. 106305. In Proceedings of *The Sixth International Conference on Civil and Building Engineering Informatics*. 2025.

[5] Chen, J., D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, Minigpt-v2 large language model as a unified interface for vision-language multi-task learning. *arXiv preprint* arXiv2310.09478, 2023.

[6] Chan, C.-F., X. Guo, P.K.-Y. Wong, J.P.-C. Chan, J.C. Cheng, P.-H. Leung, and X. Tao, A Domain Knowledge-Enhanced Large Vision-Language Model for Construction Site Safety Monitoring.

[7] Guo, K.X., P.K.-Y. Wong, J.C.P. Cheng, C.-F. Chan, P.-H. Leung and X. Tao, Enhancing visual-LLM for construction site safety compliance via prompt engineering and Bi-stage retrieval-augmented generation. *Automation in Construction*, 2025. **179**: p. 106490.

[8] Liu, Y., J.K.W. Yeoh and D.K.H. Chua, Deep Learning–Based Enhancement of Motion Blurred UAV Concrete Crack Images. *Journal of Computing in Civil Engineering*, 2020. **34**(5): p. 04020028.

[9] Liu, Y., D.K.H. Chua and J.K.W. Yeoh, Automated engineering analysis of crack mechanisms on building facades using UAVs. *Journal of Building Engineering*, 2025. **103**: p. 112176.

[10] Deng, H., Z. Ou and Y. Deng, Multi-Angle Fusion-Based Safety Status Analysis of Construction Workers. *Int J Environ Res Public Health*, 2021. **18**(22).

[11] Shao, Z., Y.M. Goh, J. Tian, Y.G. Lim and V.J.L. Gan, Computer Vision-Based Monitoring of Construction Site Housekeeping: An Evaluation of CNN and Transformer-Based Models. In Proceedings of *Computing in Civil Engineering*. 2023: p. 508-15.

[12] Panigati, T., M. Zini, D. Striccoli, P.F. Giordano, D. Tonelli, M.P. Limongelli, and D. Zonta, Drone-based bridge inspections: Current practices and future directions. *Automation in Construction*, 2025. **173**: p. 106101.

[13] Liang, Z., J.C.F. Chan, J. Zhang, Z. Liang, B. Wang, M. Wang, and J.C.P. Cheng, Multimodal LLM-driven language-embedded 3D gaussian splatting for semantic and realistic digitization of historical buildings. *Automation in Construction*, 2026. **181**: p. 106628.

[14] Shakeri, R., M.A. Al-Garadi, A. Badawy, A. Mohamed, T. Khattab, A.K. Al-Ali, K.A. Harras, and M. Guizani, Design Challenges of Multi-UAV Systems in Cyber-Physical Applications: A Comprehensive Survey and Future Directions. *IEEE Communications Surveys & Tutorials*, 2019. **21**(4): p. 3340-85.

[15] Dutta, A., S. Das, J. Nielsen, R. Chakraborty and M. Shah. Multiview Aerial Visual RECognition (MAVREC): Can Multi-view Improve Aerial Visual Perception? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[16] Hu, D., V.J.L. Gan and C. Yin, Robot-assisted mobile scanning for automated 3D reconstruction and point cloud semantic segmentation of building interiors. *Automation in Construction*, 2023. **152**: p. 104949.

[17] Alirezazadeh, S. and L.A. Alexandre, A Survey on Task Allocation and Scheduling in Robotic Network Systems. *IEEE Internet of Things Journal*, 2025. **12**(2): p. 1484-508.

[18] Burusa, A.K., E.J. van Henten and G. Kootstra. Gradient-based Local Next-best-view Planning for Improved Perception of Targeted Plant Nodes. In *Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA).* 2024.

[19] Chen, X. and Z. Zou, Are Large Pre-trained Vision Language Models Effective Construction Safety Inspectors? *arXiv preprint* arXiv:2508.11011, 2025.