www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

TOWARD TRUSTWORTHY CONSTRUCTION SAFETY HAZARD DETECTION WITH VISUAL-LANGUAGE MODELS

SUBMITTED: May 2026
PUBLISHED: September 2026
EDITOR: Frédéric Bosché
DOI: 10.36680/j.itcon.2026.045

Mina Sadat Orooje, PhD candidate
Department of Architecture, Construction Engineering and Built Environment, Politecnico di Milano, Italy
https://orcid.org/0009-0002-4131-982X
minasadat.orooje@polimi.it

Fulvio Re Cecconi, Associate Professor
Department of Architecture, Construction Engineering and Built Environment, Politecnico di Milano, Italy
https://orcid.org/0000-0001-7716-8854
fulvio.rececconi@polimi.it

Gaang Lee, Dr.-Ing., Assistant Professor
Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada
https://orcid.org/0000-0002-6341-2585
gaang@ualberta.ca

Qipei (Gavin) Mei, PhD, PEng, Assistant Professor
Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada
https://orcid.org/0000-0003-1409-3562
qipei@ualberta.ca

Muhammad Adil, M.Sc., Research Assistant
Department of Civil and Environmental Engineering, University of Alberta, Canada
https://orcid.org/0009-0009-1346-6479
madil2@ualberta.ca

SUMMARY: *The construction industry continues to experience high rates of accidents and fatalities, underscoring the need for proactive and reliable safety management. Accurate hazard identification is essential for effective image-based monitoring and decision support on construction sites. Vision–language models (VLMs) have shown strong potential for interpreting complex visual environments; however, their deployment in safety-critical applications is limited by hallucination, where hazards are inferred without sufficient evidence. To address this limitation, this study proposes a retrieval-augmented hazard detection framework that grounds VLM outputs in visual evidence through image-to-image retrieval of visually similar hazard scenarios and structured domain knowledge within a retrieval-augmented generation (RAG) pipeline. The framework integrates image-to-image retrieval of real-world hazard scenarios, supported by a curated knowledge base of 1,040 annotated construction site images, structured hazard taxonomies, and regulation-aware verification to enforce condition-constrained hazard reasoning before compliance assessment. The proposed approach is evaluated on a separate set of 260 construction site images spanning ten construction-safety hazard categories aligned with OSHA safety domains and the construction-safety literature. At the end-to-end report level, the complete pipeline achieves an F1-score of 0.917 (precision 0.939, recall 0.896) and a hallucination rate of 6.1%, substantially lower than the VLM-only baseline. At the hazard-verification gate, adding taxonomy-constrained reasoning to RAG-1 yields a precision of 0.987 and an F1-score of 0.941, while RAG-1 alone achieves an F1-score of 0.923 with 89.8% retrieval coverage. These results indicate that visual grounding provides the largest reliability gain, with structured condition verification adding further precision and gated regulatory retrieval providing traceable compliance support after hazard confirmation.*

KEYWORDS: *Vision–Language Models (VLMs), Retrieval-Augmented Generation (RAG), Construction Safety, Hallucination Mitigation, Hazard Detection.*

REFERENCE: *Orooje, M. S., Re Cecconi, F., Lee, G., Mei, Q., & Adil, M. (2026). Toward trustworthy construction safety hazard detection with visual-language models. Journal of Information Technology in Construction (ITcon), 31, 1067-1095. https://doi.org/10.36680/j.itcon.2026.045*

1. INTRODUCTION

Construction remains one of the most hazardous industries in the world. In the United States alone, the construction industry recorded 1,075 occupational fatalities in 2023, more than any other sector, with falls, slips, and trips accounting for nearly 40% of those deaths (U.S. Bureau of Labor Statistics, 2024). Globally, International Labour Organization estimates indicate approximately 60,000 fatal accidents on construction sites each year, equivalent to about one fatal accident every ten minutes (International Labour Organization, 2016). These figures are not simply the cost of a dangerous trade. In this study, a hazard is treated as an unsafe condition or situation with the potential to cause harm, such as an unguarded edge, unstable materials, or missing personal protective equipment (PPE), whereas an accident is the event in which harm occurs. This distinction is actionable: if hazards can be identified reliably and early, most accidents are preventable (Xu et al., 2023). Traditional safety management has been predominantly reactive, relying on incident reporting, accident investigation, and post-event analysis. While these practices remain essential for understanding past failures, they offer limited ability to prevent exposure before harm occurs (Phinias, 2023). The field has consequently shifted toward proactive approaches that identify and control hazards before any incident takes place. In 2024, 89% of companies were using proactive safety metrics such as audits, risk assessments, and site inspections to monitor and improve their safety management systems (DEKRA, 2024), a clear signal that the industry recognizes early hazard identification as an important lever for reducing fatalities. This shift has been accelerated by the growing volume of visual data now produced on construction sites. Photographs, surveillance cameras, inspection imagery, and mobile devices collectively generate a continuous record of site conditions, worker activity, and equipment state (Fang et al., 2018; Nath & Behzadan, 2020). If analysed effectively, these data create opportunities for automated and near-real-time hazard monitoring and early intervention (Raman & Mitra, 2023).

Recent advances in artificial intelligence have significantly improved automated construction safety monitoring. Early computer-vision approaches, particularly those based on convolutional neural networks and object detection, can reliably identify safety-relevant elements such as workers, personal protective equipment (PPE), and machinery (Fang et al., 2018; Nath & Behzadan, 2020). However, these approaches are less suited to relational hazard reasoning, where risk depends on interactions among workers, equipment, environmental conditions, and protective controls rather than on isolated object presence alone (Hallowell & Gambatese, 2009; Seo et al., 2015).

To address this limitation, vision-language models (VLMs) have been introduced for more context-aware construction-safety interpretation (Adil et al., 2025; Guo et al., 2025). By jointly processing visual and textual information, VLMs and related vision–language encoders such as Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021), Bootstrapping Language-Image Pretraining (BLIP) (Li et al., 2022), and GPT-4V can support richer semantic interpretation of construction scenes and natural-language safety reporting. Recent construction-safety systems have also incorporated regulatory retrieval, domain ontologies, and structured prompting to strengthen traceability and compliance reasoning (Chan et al., 2025; Guo et al., 2025; Tran et al., 2024).

Despite these advancements, several challenges remain in AI-based construction safety monitoring. First, hazard detection is inherently relational and requires understanding interactions among workers, equipment, environmental conditions, and controls rather than isolated object recognition (Adil et al., 2025; Chan et al., 2025). Second, among the construction-safety systems reviewed in Section 2.4, runtime retrieval is still dominated by regulations, textual knowledge, ontologies, or generated representations rather than annotated visual precedent cases (Chan et al., 2025; Guo et al., 2025; Tran et al., 2024). Third, compliance-oriented retrieval can be invoked without an explicit visual-evidence gate, creating a risk that an unsupported hazard interpretation propagates into the compliance stage (Guo et al., 2025; Tran et al., 2024).

Among these challenges, this study specifically focuses on hallucination as a critical reliability issue in vision–language models (Adil et al., 2025; Alansari & Luqman, 2025), where models generate hazard descriptions that are not supported by visual evidence. A model may flag missing PPE that is clearly visible, cite a fall risk where no elevation exists, or invoke a regulatory violation based on learned statistical associations rather than observable site conditions. This reliability gap poses substantial risks in safety-critical AI applications (Amodei et al., 2016). Hallucination remains documented across contemporary large vision-language model families, so it should not be treated as a problem solved by model scale alone (Leng et al., 2024; Qu et al., 2024; R. Zhang et al., 2024).

In construction safety, the consequences are operational rather than abstract. Repeated unsupported warnings can erode practitioner trust and create alert fatigue, while unsupported hazard claims can also propagate into irrelevant compliance retrieval and divert attention from genuine risks. The core reliability problem addressed in this study is therefore evidential: a fluent hazard description is not sufficient unless the claim is supported by observable scene evidence. This motivates an architecture that grounds the perceptual decision before regulatory information is introduced.

Recent advances in VLMs have significantly improved the ability of AI systems to interpret complex construction scenes and support safety analysis (Adil et al., 2025; Guo et al., 2025). However, despite these advancements, a fundamental reliability challenge remains in current VLM-based hazard-detection systems. The hallucination issue mainly results from three technical caveats of existing VLM-based hazard-detection pipelines (Rohrbach et al., 2018; Li et al., 2023; Qu et al., 2024). First, existing systems lack condition-grounded hazard reasoning. They can identify objects associated with risk but cannot determine whether a hazardous situation exists because they do not model the relationships between scene elements. In construction environments, hazards arise from the interaction of multiple factors such as worker position, site conditions, and the presence or absence of protective measures rather than from any single element in isolation (Larouzée & Guarnieri, 2015; Ceylan et al., 2021).

Another limitation is the lack of runtime visual grounding in real-world precedent cases (Lewis et al., 2020; Guo et al., 2025). While retrieval-augmented generation (RAG) offers a promising solution by integrating information retrieval with generative models (Lewis et al., 2020), its construction-safety applications have primarily emphasized textual resources such as regulations and safety documents (Guo et al., 2025; Tran et al., 2024). In the systems reviewed in Section 2.4, retrieval of annotated construction images as visual precedent for the query scene remains comparatively limited. This gap motivates the RAG-1 stage proposed here.

A further limitation is the weak integration between hazard identification and compliance assessment in existing construction-safety pipelines. Although regulations and safety documents can be incorporated into retrieval-augmented systems, recent work has largely focused on regulatory access and compliance reasoning rather than on an explicit rule that withholds compliance retrieval until the hazard is visually confirmed (Guo et al., 2025; Tran et al., 2024). This sequencing matters in safety-critical settings because an unsupported hazard inference can otherwise propagate into an unsupported violation report.

To address these limitations, this study proposes a retrieval-augmented vision-language framework for construction hazard detection. The framework integrates image-to-image retrieval of visually similar construction scenarios, taxonomy-guided hazard reasoning, and regulation-aware verification into a unified pipeline. Rather than relying on free-form generation, hazard identification is grounded in retrieved visual precedent and structured condition-based reasoning before compliance assessment is performed. This study makes four contributions:

1. A retrieval-grounded hazard detection framework that incorporates image-to-image retrieval to support visual reasoning with similar construction scenarios.
2. A taxonomy-based hazard representation that enforces condition-based verification across human, environmental, and organizational factors.
3. A staged hazard detection pipeline that separates hazard confirmation from compliance assessment to reduce unsupported regulatory claims.
4. A component-level ablation across nine configurations and multiple vision-language backbones that attributes the reduction in hallucination to specific design choices rather than to model capacity, reported with bootstrap confidence intervals and paired significance tests.

This study is positioned as applied systems research rather than as a contribution to model architecture. The novelty claimed here is not a new encoder, training objective, or fine-tuning scheme: the CLIP visual encoder, the VLM reasoning backbone, and the sentence-transformer regulation encoders are used off the shelf and are left unmodified. The contribution is fourfold: visual precedent retrieval, an explicit and auditable three-level hazard representation with AND-logic activation conditions, a staged hazard-confirmation-before-compliance architecture, and component-level empirical validation across multiple configurations and VLM backbones.

The first two contributions convert hazard interpretation from unconstrained pattern completion into evidence checking: retrieved visual cases provide scene-level precedent, while the taxonomy converts domain knowledge

into observable conditions that can be audited. The staged architecture then gates regulation retrieval behind hazard confirmation so that unconfirmed hazards cannot trigger the compliance stage.

The fourth contribution is empirical rather than architectural: the ablation, backbone-generalisation, sensitivity, statistical, robustness, and deployment analyses show which components account for the observed reliability gains. The study is therefore evaluated against criteria appropriate to applied systems research, including reproducibility, component-level attribution of performance, deployment latency and cost transparency, and practitioner-assessed usefulness of the generated reports. The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed framework. Section 4 describes the experimental design. Section 5 reports results. Section 6 discusses findings and practical implications. Section 7 states the limitations of the study, and Section 8 concludes.

2. RELATED WORK

This section reviews three strands of literature that motivate the proposed framework: vision–language models in construction safety, existing hallucination mitigation strategies for vision-language models, and retrieval-augmented generation (RAG) for safety-related reasoning. Together, these studies clarify both the promise and limitations of current approaches, and they motivate the need for a framework that combines visual grounding, structured hazard reasoning, and regulation-aware verification.

2.1 Vision–language models in construction safety

Vision–language models (VLMs) have become increasingly relevant in construction safety research because they can jointly process visual and textual information and support context-aware interpretation of complex site conditions (Adil et al., 2025; Guo et al., 2025). Models such as CLIP, BLIP, and GPT-4V enable richer semantic understanding of construction scenes and allow safety-related findings to be expressed in natural language. In this context, two applications have received particular attention: image captioning and visual question answering (VQA) (Tran et al., 2024). Captioning approaches generate textual descriptions of site conditions, while VQA systems allow inspectors to ask targeted questions, such as whether a worker is wearing personal protective equipment or whether a protective barrier is present (Adil et al., 2025). These capabilities make VLMs attractive for construction safety analysis because they support more flexible and interpretable scene understanding than conventional visual recognition pipelines (Guo et al., 2025).

Despite these advantages, VLMs remain vulnerable to hallucination in safety-critical applications. They may generate plausible but unsupported hazard descriptions by relying on learned statistical associations rather than on the actual visual evidence present in the image (Alansari & Luqman, 2025; R. Zhang et al., 2024). For example, the presence of scaffolding may lead to a model reporting a fall hazard even when the platform is properly guarded, simply because scaffolding and fall risk frequently co-occur in training data (Leng et al., 2024; Zhu et al., 2025). This problem is particularly concerning in construction safety, where accurate interpretation often depends on spatial relationships and interactions among scene elements rather than object presence alone (Adil et al., 2025; Chan et al., 2025). These limitations have motivated increasing research into methods that improve the reliability and grounding of VLM-based hazard analysis.

2.2 Existing VLM hallucination mitigation strategies and their limitations

Hallucination in VLMs has led to a growing body of research on methods designed to improve reliability (Alansari & Luqman, 2025; Nakharu & Kumar, 2025). Existing strategies generally fall into three broad categories: prompting-based control, self-consistency methods, and post-hoc verification. Prompting-based approaches instruct the model to focus on visible evidence, avoid speculation, or produce responses in a structured format (Adil et al., 2025). Self-consistency methods generate multiple candidate responses and select the most stable or frequently supported result, thereby reducing unstable or outlier predictions. Post-hoc verification methods evaluate the generated output after inference using a secondary model or a rule-based checker (D. Zhang et al., 2025). These strategies can improve output quality, but their value in construction hazard detection remains limited because the central problem is not only whether the response is fluent or stable, but whether the reported hazard is supported by sufficient visual and contextual evidence (Leng et al., 2024; Zhu et al., 2025). Prompting can guide how the model responds, self-consistency can reduce random variation, and post-hoc verification can identify some errors after generation. However, these methods do not by themselves ensure that hazard predictions are

grounded in visually similar precedent cases or in an explicit condition-based reasoning structure (Guo et al., 2025). In safety-critical settings, this distinction is important because a response may appear coherent and internally consistent while still describing a hazard that is not actually present in the scene (Adil et al., 2025).

Accordingly, existing mitigation strategies provide partial improvements in reliability, but they do not fully resolve the evidence-grounding problem in construction hazard analysis. This limitation points to the need for methods that incorporate external knowledge and visual grounding directly into the hazard reasoning process (Lewis et al., 2020), rather than relying only on output-level correction.

2.3 Retrieval-augmented generation for hallucination mitigation in construction safety

Retrieval-augmented generation (RAG) addresses hallucination by grounding model outputs in retrieved external information rather than relying only on parametric knowledge (Lewis et al., 2020). In construction safety research, RAG has mainly been used to retrieve textual resources such as regulations, safety manuals, and incident records, which are then incorporated into the model's reasoning process (Guo et al., 2025; Tran et al., 2024). This improves traceability and helps connect generated safety judgments to verifiable external sources. However, current text-based RAG approaches remain limited for visual hazard analysis. Construction hazards are often defined by visual relationships, including worker position, equipment arrangement, proximity to danger zones, and the presence or absence of protective controls. These are properties that textual retrieval alone may not represent adequately. As a result, a system may retrieve relevant regulations or safety guidance without first establishing whether the hazard is visible in the image under analysis (Guo et al., 2025).

This limitation suggests that effective hallucination mitigation in construction safety requires more than text-grounded retrieval. It requires a framework that also grounds hazard reasoning in visually similar precedent cases and evaluates hazards through explicit, evidence-based conditions before linking them to compliance-related guidance. This need motivates the framework proposed in this study.

Table 1: Comparative characteristics of recent VLM-based construction safety systems.

Study / system	Visual grounding	Knowledge and retrieval	Reliability / hallucination control	Dedicated hallucination evaluation
Adil et al. (2025)	Whole-image analysis; no visual precedent	Safety guidelines embedded in prompts; no retrieval	Structured prompt engineering	Not reported; BERTScore used
Guo et al. (2025)	Direct interpretation of the query image	Regulation corpus with bi-stage RAG (BiRAG)	Hierarchical prompting and compliance reasoning	Not reported; retrieval and compliance accuracy assessed
Chan et al. (2025)	Image/video analysis without retrieved visual cases	Domain-specific safety ontology	Ontology-guided prompting and curriculum fine-tuning	Not reported; violation-detection accuracy assessed
Wang et al. (2025)	Multimodal visual and audio grounding	RAG-supported safety inspection knowledge	Retrieval-grounded report generation	No dedicated metric; standard detection metrics reported
Li et al. (2026a)	Domain-specific construction-image grounding	Knowledge learned from annotated image-text data; no runtime retrieval	LoRA-based domain adaptation	Not reported; F1 and captioning metrics used
Wang et al. (2026)	Construction-element visual grounding	Retrieval-augmented caption re-mapping	Cross-verification between two VLMs	No dedicated hallucination metric
Adil et al. (2026)	Object-level grounding using YOLOv11n	Detected objects and safety criteria encoded in prompts	Detection-guided VLM reasoning	No dedicated metric; F1 and BERTScore reported
Li et al. (2026b)	Pixel-level hazard grounding	Regulatory knowledge embedded in domain-specific training data	Visual-semantic grounding and multimodal decoding	Not reported as a dedicated outcome
Chan et al. (2026)	Object-centric visual grounding	Visual-grounding dataset with domain safety rules	Grounded reasoning with reinforcement learning	Not reported as a dedicated metric

2.4 Context with recent VLM-based approaches to construction safety

Recent VLM-based approaches to construction safety have incorporated prompt engineering, regulatory retrieval, domain ontologies, fine-tuning, visual grounding, and cross-model verification. Table 1 compares these systems across four dimensions relevant to model reliability: visual grounding, knowledge retrieval, hallucination control,

and explicit hallucination evaluation. Evidence-based gating is discussed separately below because it is a sequencing property of the proposed architecture rather than a consistently reported feature of prior systems.

Against this background, three distinctions characterize the proposed framework. *First*, although recent studies increasingly incorporate object-level, pixel-level, or semantic visual grounding, retrieval remains largely cantered on textual regulations, ontological knowledge, or model-generated representations. The proposed framework instead retrieves annotated visual cases as contextual evidence for interpreting the query scene, thereby grounding hazard assessment in visual precedent.

Second, hallucination is treated as an explicit evaluation outcome rather than solely as a problem addressed through prompting, fine-tuning, grounding, or cross-model verification. Hallucination is operationally defined and directly measured, together with inter-annotator agreement.

Third, the framework introduces evidence-based gating between hazard identification and regulatory assessment. A hazard is retained only when predefined observable conditions are supported by the image; otherwise, the label is withheld and the downstream compliance stage is not initiated. Among the reviewed systems, Wang et al. (2026) is the closest methodological comparison because it uses cross-model verification with retrieval-augmented caption re-mapping. However, its verification operates primarily on generated representations, whereas the proposed framework conditions subsequent reasoning on the sufficiency of the visual evidence itself.

3. THE PROPOSED FRAMEWORK

Based on the limitations identified in Section 2, the proposed framework addresses three key design gaps: limited visual grounding, lack of structured hazard reasoning, and insufficient evidence linkage between hazard detection and compliance assessment. Table 2 summarizes these gaps and their corresponding design components in the framework.

Table 2: Literature gaps, contributing studies, and proposed solutions.

LITERATURE GAP	ROOT CAUSE	CONTRIBUTING STUDIES	PROPOSED SOLUTION
No visual grounding	Single-image analysis; no precedent case consulted	CNN/VLM studies process each image independently	Image-to-image retrieval: retrieve visually similar hazard precedents
NO CONDITION-BASED VERIFICATION	Pattern matching, not condition checking	Many VLM-based systems rely primarily on learned correlations rather than explicit hazard conditions	Taxonomy-based condition verification: hazard reported only when all activation conditions are met
LIMITED EVIDENCE GROUNDING BEFORE COMPLIANCE	Hazard confirmation is not explicitly tied to sufficient visual evidence	RAG safety systems retrieve regulations early, producing unsupported violation reports	Regulation-aware staged verification: compliance assessed only after visual grounding

These gaps share a common structural root: existing approaches treat hazard detection as a single-step generative task rather than a staged inference problem. The proposed framework addresses this limitation by decomposing hazard detection into a sequence of structured reasoning stages, ensuring that each stage is grounded in sufficient visual and contextual evidence before proceeding to the next. The overall design of the framework is illustrated in Figure 1.

As shown in Figure 1, the proposed framework is operationalized as a four-stage pipeline corresponding to the four labelled components: RAG-1, VLM, RAG-2, and LLM. It accepts a construction site image as input from inspection photographs, surveillance cameras, or mobile devices and encodes it into a high-dimensional vector representation using a CLIP-based visual embedding model.

In the first stage, RAG-1 (visual retrieval) queries a curated vector database of annotated hazard scenes via FAISS approximate nearest-neighbour search, retrieving the top-k most visually similar cases along with their structured hazard metadata. In the second stage, the VLM analyses the input image jointly with the retrieved context, evaluating whether the activation conditions defined in the hazard taxonomy are satisfied before any hazard is

reported. This step ensures that hazard identification is based on explicit condition verification rather than implicit statistical associations. Only after a hazard is visually and evidentially confirmed does the framework trigger the third stage, RAG-2 regulatory retrieval, in which relevant safety standards are ranked based on semantic similarity to the detected hazard type and jurisdictional context.

Finally, the LLM synthesizes the verified hazard, its visual evidence, and the retrieved regulations into a structured report linking each identified hazard to its supporting visual evidence, condition-based explanation, and retrieved regulatory context. The sequencing principal grounding before compliance is a key design feature of the framework and the primary mechanism used to prevent unsupported hazards from propagating directly into the compliance stage.

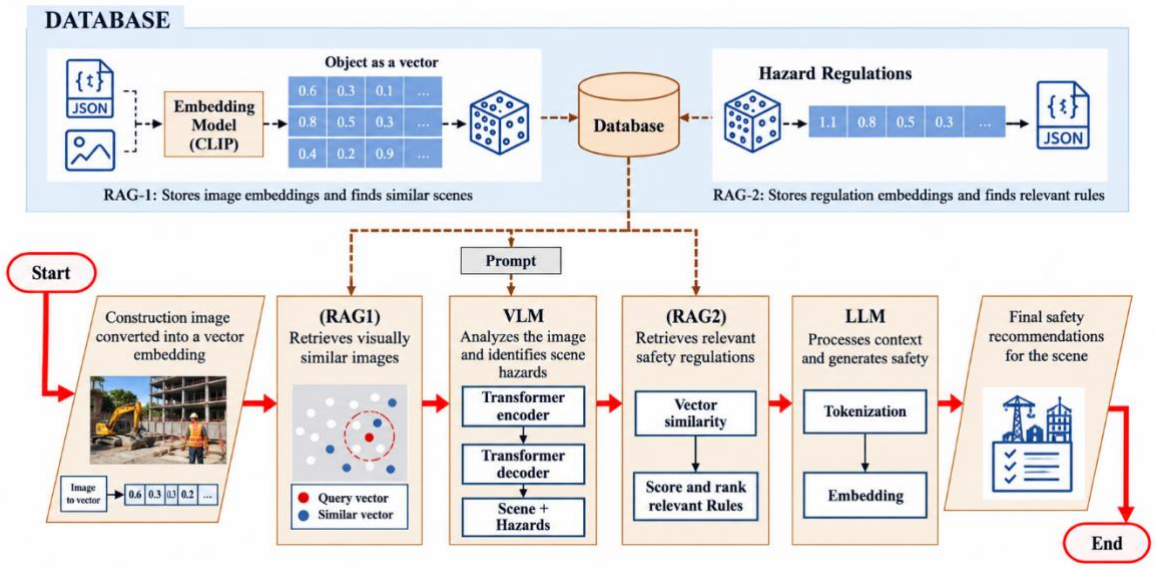

Figure 1: Multimodal pipeline for hazard detection and safety recommendations. RAG-1 retrieves similar scenes; RAG-2 retrieves relevant regulations. The VLM performs condition-constrained reasoning before regulatory assessment.

3.1 Hazard taxonomy and knowledge representation

In construction environments, many hazards cannot be inferred reliably from the presence of a single object alone; risk often depends on the relationship among worker exposure, physical conditions, and protective or organizational controls (Hallowell & Gambatese, 2009; Larouzée & Guarnieri, 2015). Systems that reason primarily from object presence therefore risk over-interpreting safety significance when the relevant relational context is missing (Adil et al., 2025; Chan et al., 2025). To address this, the proposed framework encodes hazard knowledge as a three-level taxonomy, illustrated in Figure 2. A note on terminology is warranted at this point. Throughout this paper, condition-constrained reasoning refers to the verification of a predefined set of observable conditions derived from accident-causation theory (Ceylan et al., 2021; Larouzée & Guarnieri, 2015).

This term is used deliberately instead of causal reasoning because the framework does not perform causal inference in the statistical or interventionist sense. Rather, it assesses whether the conditions associated with a given hazard are simultaneously observable in an image. It does not construct causal graphs, estimate interventional effects, or evaluate counterfactual scenarios. Instead, the taxonomy translates established causal mechanisms from the safety literature into observable verification conditions. The relationship between this rule-based operationalization and formal causal modelling is discussed in Section 7.

As shown in Figure 2, the taxonomy is organized across three levels. Level 1 defines ten construction-safety hazard categories aligned with OSHA safety domains and the construction-safety literature: fall, struck-by, slip/trip, electrocution, caught-in-between, chemical exposure, collapse, fire/explosion, fatigue, and noise. These categories define the outcome space that the framework is designed to detect. Level 2 organizes contributing factors into three classes: human behaviours (e.g., failure to wear PPE, unsafe climbing practices, ignoring lockout/tagout

procedures), unsafe physical conditions (e.g., missing guardrails, poor lighting, defective machinery), and organizational or system failures (e.g., absent safety barriers, poor supervision, lack of hazard reporting). Level 3 specifies observable sub-factors within each class the concrete, visually verifiable indicators that the model is instructed to look for in the input image.

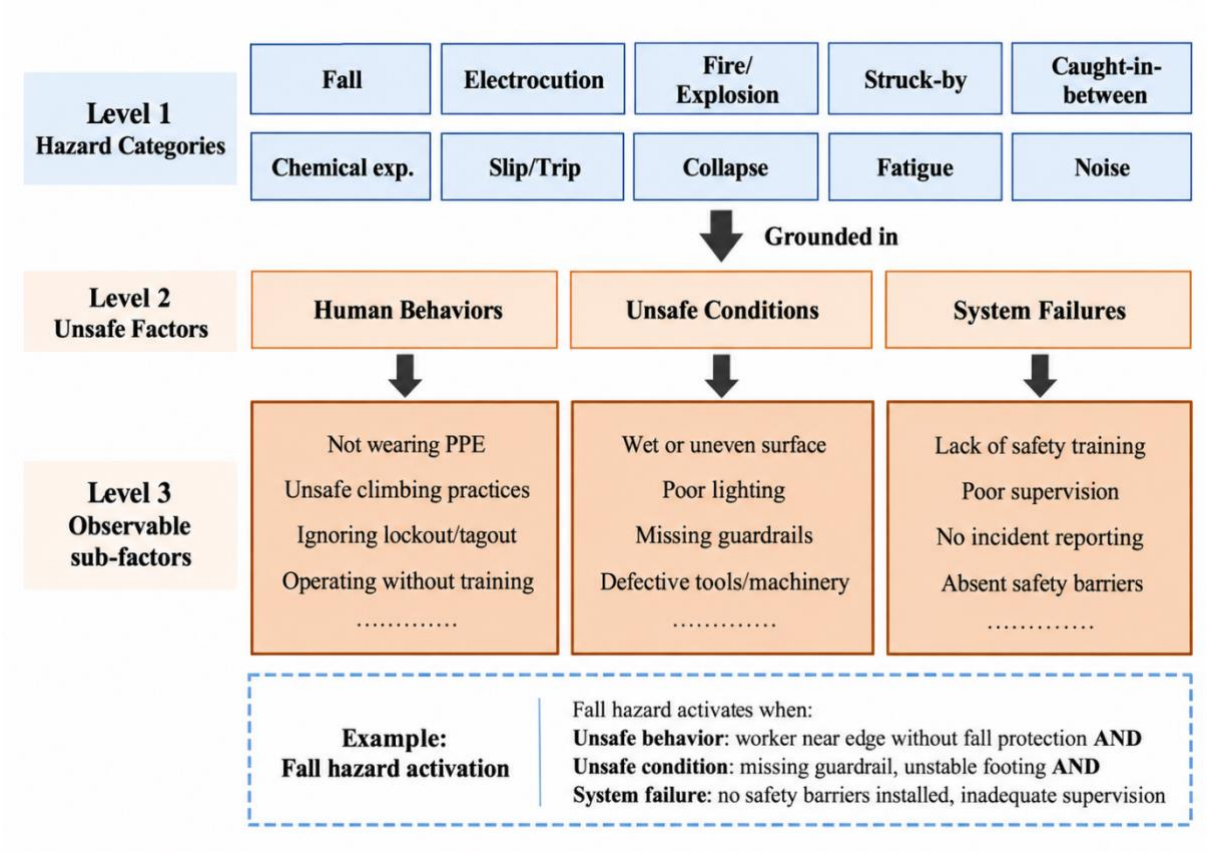

Figure 2: Three-level hazard taxonomy. Level 1 defines the ten study hazard outcomes; Level 2 groups contributing-factor classes; Level 3 specifies observable conditions.

The hazard representation is operationalized as a condition-based taxonomy with an AND-logic constraint: a Level 1 hazard is reported only when at least one contributing factor from each of the three Level 2 classes is simultaneously satisfied. For a fall hazard, this means: a worker near an unprotected edge (human behaviour) AND a missing guardrail or unstable footing (unsafe condition) AND an absent safety barrier or inadequate supervision (system failure). This constraint prevents the framework from flagging a hazard because of a single detected object, the most common source of hallucination in prior VLM-based approaches (Rohrbach et al., 2018; Li et al., 2023). Formally, let B, C, and S denote the sets of observable indicators from human behaviour, unsafe condition, and system failure classes respectively. A Level 1 hazard h is activated if and only if:

$$Activate(h) = 1 \text{ if and only if } |B \cap Obs| \geq 1 \text{ AND } |C \cap Obs| \geq 1 \text{ AND } |S \cap Obs| \geq 1 \quad (1)$$

where Obs denotes the set of observable indicators confirmed in the input image. In other words, B, C, and S represent the sets of Level 3 indicators defined for the human behaviour, unsafe condition, and system failure classes respectively; B ∩ Obs denotes the subset of behaviour indicators that are confirmed in the image; and the condition |B ∩ Obs| ≥ 1 requires at least one such indicator to be confirmed. A hazard label h is assigned if and only if all three classes simultaneously contribute at least one confirmed observable indicator.

The AND-logic formulation is grounded in established accident-causation theory. Reason's Swiss Cheese Model (Larouzée & Guarnieri, 2015) and the Systems-Theoretic Accident Model and Processes (STAMP) framework (Ceylan et al., 2021) both conceptualize accident causation as the alignment of multiple failure conditions across

behavioural, environmental, and organizational layers, providing a conceptual parallel to the three-class condition structure used here. The AND-logic constraint enforces a multi-factor verification standard: a fall hazard requires not merely elevation, but co-occurring evidence of exposure, unsafe physical conditions, and missing or inadequate controls. The trade-off is acknowledged: strict AND-logic may reduce recall where a hazard is only partly visible or where one factor class cannot be resolved from a single image. Section 7 discusses this limitation, and Table 5 shows the associated precision-recall trade-off. All hazard knowledge is encoded in a structured JSON format. Each image in the retrieval database carries metadata describing scene context, worker activities, equipment status, identified hazard types, and the Level 3 sub-factors observed. This representation supports both vector-based retrieval and the structured reasoning prompt passed to the VLM in Section 3.3.

3.2 Hazard-oriented image-to-image retrieval (RAG-1)

The first retrieval stage, RAG-1, shown in the left portion of Figure 1, grounds hazard inference in visually similar real-world construction scenarios rather than relying only on statistical patterns learned during model training. Retrieving annotated precedent cases provides concrete evidence of how hazard configurations appear across different scenes and reduces the need for the model to infer unsupported hazards from scene stereotypes.

Each image in the retrieval database is encoded using a CLIP visual encoder, producing a 512-dimensional embedding that captures semantic visual features including spatial layout, worker positioning, equipment configuration, and environmental context. These embeddings are indexed using FAISS with an IVF flat index, enabling sub-linear approximate nearest-neighbour search across the image database. At inference time, the input image is encoded using the same model and used as a query; the system returns the top-k most similar images (k = 5 in all reported experiments) along with their JSON metadata records. The retrieved cases serve two functions in the downstream reasoning stage. First, they provide visual precedent concrete examples of what a hazard of a given type looks like in a real construction scene constraining the VLM's interpretation of the input image. Second, the associated metadata supplies the structured condition context hazard type, contributing Level 2 and Level 3 factors, and site conditions that the VLM uses to evaluate whether the same activation conditions are present in the query image. This sequential design ensures that regulatory claims are never made ahead of visual confirmation, which is a key architectural distinction from prior systems.

3.3 VLM-based condition-constrained hazard reasoning

With retrieved visual precedents and their structured metadata available, the VLM performs condition-based hazard reasoning over the input image. As illustrated in the central stage of Figure 1, GPT-4V is used for its strong capability in joint visual and textual reasoning across complex, multi-element scenes. The reasoning process is driven by a structured prompt with three components: (1) the input construction image; (2) the JSON metadata from the top-k retrieved cases, summarizing the hazard types and activation conditions observed in similar scenes; and (3) explicit analytical instructions directing the model to evaluate observable evidence for each Level 2 factor category as defined in Figure 2 before reporting any hazard.

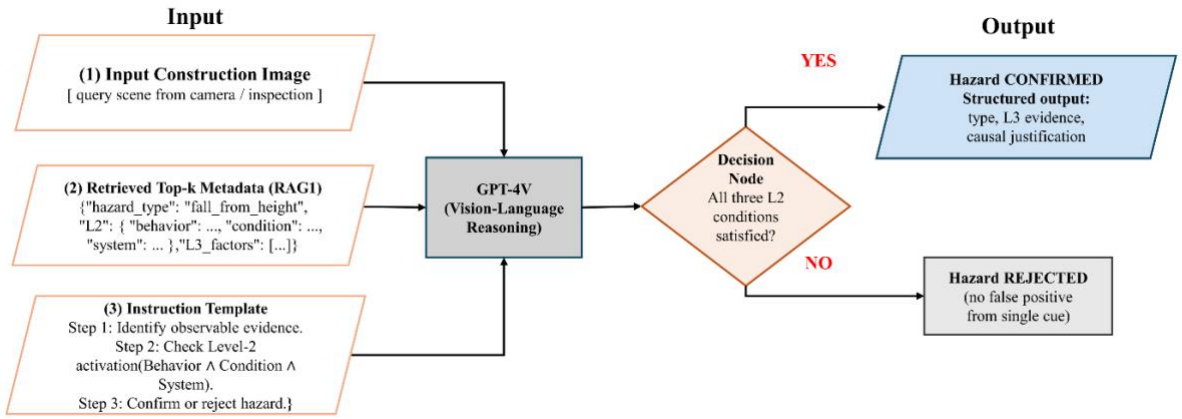

Figure 3: VLM-based condition-constrained hazard reasoning module. A structured prompt (image, RAG metadata, and instructions) feeds GPT-4V, which confirms hazards only when all three Level 2 conditions are satisfied.

The prompt instructs the model to reason sequentially: first identify what is visually observable, then check whether the activation conditions from the taxonomy are met, and only then assign a hazard label. Free-form hazard description without condition-based justification is explicitly prohibited in the prompt design. Figure 3 illustrates this end-to-end reasoning structure: the three prompt components feed into GPT-4V, whose output passes through a decision node that confirms a hazard only when all three Level 2 conditions (behaviour, condition, system) are simultaneously satisfied.

This constraint is the key mechanism for hallucination reduction. A standard VLM prompt asks "What hazards are present?" an open generative question that invites statistical pattern completion. The structured prompt in this framework asks instead: "Given these retrieved cases and these activation conditions, is the evidence sufficient to confirm this hazard?" The shift from generation to verification fundamentally changes how the model reasons. If the required conditions are not observable in the image, the hazard is not reported, regardless of how statistically plausible it appears given the scene context.

3.4 Regulation-aware retrieval and verification (RAG-2)

Regulatory compliance assessment is triggered only after hazards have been confirmed through condition-constrained reasoning. As shown in the right portion of Figure 1, RAG-2 connects confirmed hazard labels to jurisdiction-specific safety regulations through semantic retrieval. This sequencing is a deliberate architectural choice that distinguishes the proposed framework from prior RAG-based safety systems in which regulations may be retrieved before sufficient visual confirmation. The gate prevents unconfirmed hazards from initiating regulatory retrieval; it does not, by itself, guarantee that every retrieved provision is applicable, which is why RAG-2 is evaluated separately in Section 5.3. Once a hazard is confirmed, its Level 1 category label is used to query a semantic retrieval database of encoded regulatory documents. The database incorporates construction safety standards from two jurisdictions, with each entry tagged by hazard category and jurisdiction. Regulatory embeddings are generated using a sentence-transformer model and retrieved by cosine-similarity ranking. The top three candidate provisions are passed to the output stage after filtering by hazard category and jurisdictional context. Figure 4 illustrates this gated flow: a confirmed hazard passes through the hazard-confirmation gate and is then matched against the regulation vector database; if no hazard is confirmed, the pipeline halts and regulatory retrieval is not initiated. Compliance statements in the final report are therefore traceable to a visually confirmed hazard, while the relevance of the retrieved provision remains an independently evaluated retrieval question.

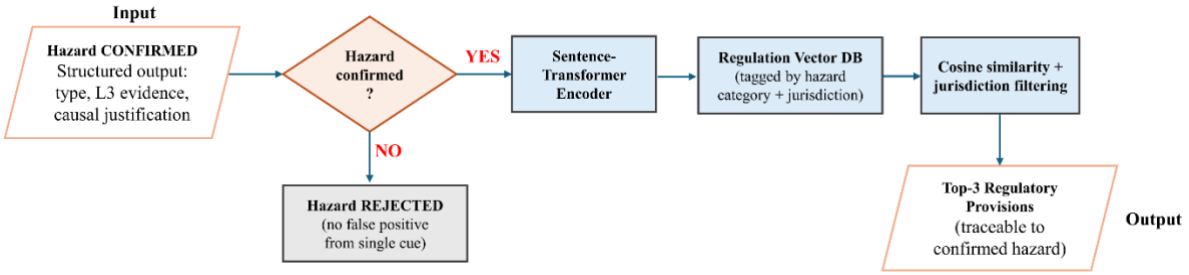

Figure 4: Regulatory retrieval with hazard-confirmation gate (RAG-2). Only confirmed hazards are matched against the regulation vector database with jurisdiction filtering to return the top three candidate provisions.

3.5 Hazard report and safety recommendation generation

The final stage integrates the outputs of the condition-constrained reasoning module and the regulatory retrieval module into a structured hazard report. Each report entry specifies the detected hazard type, the Level 3 sub-factors observed in the image that satisfy the activation conditions, the regulatory provisions retrieved as relevant candidates, and corrective actions supported by those provisions. This structure serves two purposes beyond documentation. First, it makes the system's reasoning transparent: a safety manager can inspect why a hazard was flagged, which observable conditions triggered it, and whether the claim is supported by the image. Second, it creates an auditable record that can support digital safety-management workflows and compliance review. Retrieved regulatory provisions are presented as evidence-linked candidates rather than infallible compliance determinations, consistent with the independent RAG-2 error analysis in Section 5.3.

4. TESTING AND EVALUATION

This section describes the dataset, model configurations, and evaluation metrics used to assess the proposed framework.

4.1 Dataset

A dedicated construction hazard image dataset was compiled to support both the retrieval knowledge base and the performance evaluation of the proposed framework. To prevent data leakage and ensure that reported results reflect genuine generalization to unseen construction scenes, the dataset is divided into two strictly separated subsets with no overlap. The first subset forms the RAG-1 retrieval database, comprising 1,040 annotated construction site images compiled from three publicly available sources: (i) the SODA dataset (Duan et al., 2022) for PPE and fall-related hazards; (ii) OSHA photographic inspection records (Occupational Safety and Health Administration, 2024) for chemical exposure, collapse, fire/explosion, and noise hazards; and (iii) construction safety images compiled in prior computer-vision studies (Fang et al., 2018; Nath & Behzadan, 2020; Seo et al., 2015). Together, these sources span building, road, bridge, and industrial construction environments under diverse conditions. Visually duplicate images were removed during curation, and each retained image was annotated according to the three-level hazard taxonomy defined in Section 3.1, with labels reviewed by construction-safety domain experts. For the analyses reported in this paper, category labels are restricted to the ten Level 1 hazard groups defined in Section 3.1. Across the 1,040 retrieval images, 1,674 hazard instances are assigned to the ten analytical hazard categories. Figure 5 reports the absolute instance counts: fall (591), struck-by (432), slip/trip (287), collapse (170), caught-in-between (71), chemical exposure (53), noise (22), fatigue-related (19), fire/explosion (16), and electrocution (13). These are hazard-instance counts rather than counts of unique images, because one construction image may contain more than one co-occurring hazard. Figure 6 presents the same 1,674 annotated instances as percentage shares of a single common denominator: fall (35.30%), struck-by (25.81%), slip/trip (17.14%), collapse (10.16%), caught-in-between (4.24%), chemical exposure (3.17%), noise (1.31%), fatigue-related (1.14%), fire/explosion (0.96%), and electrocution (0.78%). Reporting both counts and shares avoids mixing image-level and instance-level denominators and makes the underrepresentation of the rare categories explicit.

Together, Figures 5 and 6 characterize the RAG-1 retrieval knowledge base using one consistent unit of analysis: annotated hazard instances. Figure 5 shows absolute support, whereas Figure 6 shows the normalized class distribution. The four rarest retrieval categories noise, fatigue-related, fire/explosion, and electrocution have relatively limited support and are treated cautiously in category-specific interpretation.

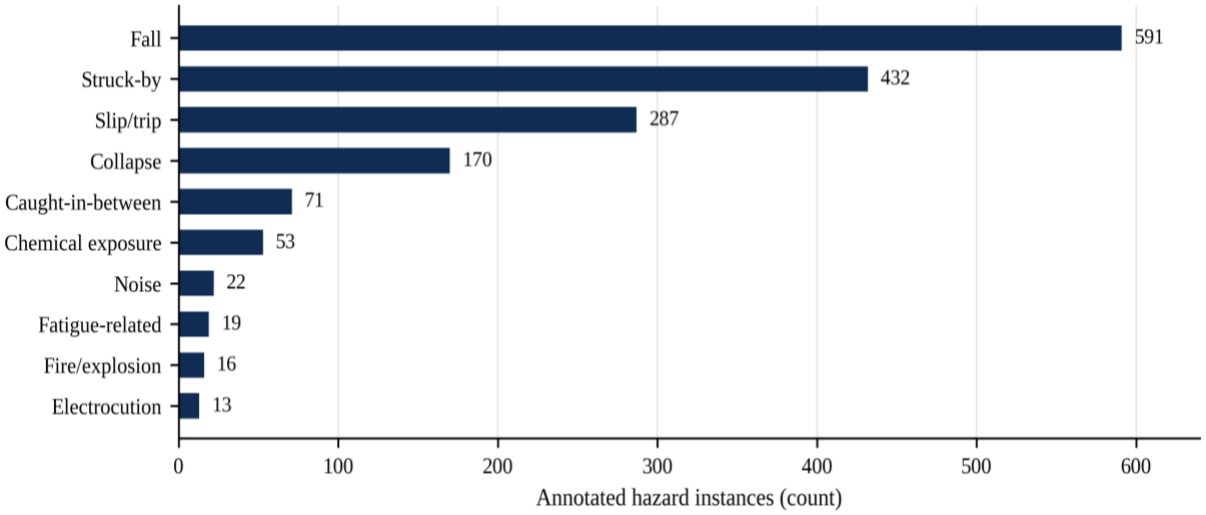

Figure 5: Absolute counts of annotated hazard instances across the ten analytical categories in the RAG-1 retrieval database (n = 1,674 instances from 1,040 images).

The second subset consists of 260 previously unseen test images used exclusively for performance evaluation. To prevent data leakage, a strict scene-level separation was enforced between the retrieval database and the test set. Specifically, images captured from the same construction site, project, or hazard instance were not split across the

two subsets, even if they differed in viewpoint, time, or camera angle. This ensures that no semantically equivalent or near-duplicate scenes appear in both the retrieval and testing datasets, thereby providing a more reliable assessment of generalization performance. Images were drawn from the same three public collections used to construct the retrieval knowledge base. To ensure comparable environmental coverage across subsets, data separation was enforced within each source at both the scene and project levels, rather than by assigning entire sources exclusively to a single subset. First, the SODA dataset (Safety Observation Dataset for construction sites) (Duan et al., 2022), which provides annotated PPE and fall-related hazard images across building and civil construction contexts. Second, supplementary construction safety images drawn from publicly available OSHA photographic inspection records (Occupational Safety and Health Administration, 2024), providing coverage for chemical exposure, collapse, fire/explosion, and noise-related hazard categories. Third, near-miss scenario images collected from open-access construction safety repositories including those documented in prior computer vision safety studies (Fang et al., 2018; Nath & Behzadan, 2020).

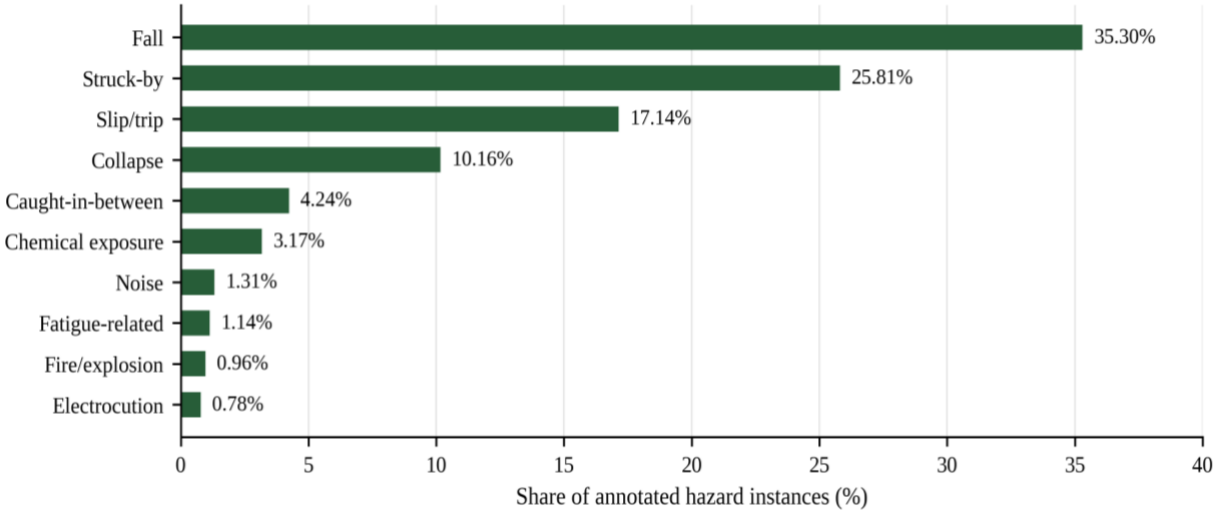

Figure 6: Percentage distribution of the same 1,674 annotated hazard instances across the ten analytical categories in the RAG-1 retrieval database. Percentages use a single instance-level denominator and sum to approximately 100% because of rounding.

Together, these sources span building construction, roadwork, bridge construction, and industrial environments across diverse lighting, camera angles, and site conditions. Where dataset sources overlap in hazard coverage, images were selected to maximize category diversity and minimize redundancy across the retrieval knowledge base.

4.2 Annotation protocol and reliability

Because both the reference hazard labels and hallucination judgements rely on human annotation, the annotation procedure was defined explicitly. For each of the ten Level 1 hazard categories, the annotation guideline specified the Level 3 observable indicators required for assigning a label, together with positive and negative examples and rules for ambiguous cases, including partial occlusion, distant subjects, and situations in which the presence or adequacy of a protective control could not be determined from the image. Two annotators with 3 and 4 years of construction-safety experience, respectively, annotated the dataset.

A randomly selected 20% of the held-out test set (52 images) was independently double-annotated to assess reliability. Agreement on Level 1 hazard-category assignment yielded Cohen's $\kappa = 0.86$, while agreement on the multi-label hazard-instance set yielded Krippendorff's $\alpha = 0.82$, indicating strong agreement. Disagreements were first reviewed jointly by the two annotators; six cases that remained unresolved were adjudicated by a senior construction-safety professional. Class support was considered explicitly when interpreting category-level results. The 260-image test set contained 421 annotated hazard instances across the ten analytical categories: slip/trip (109), fall (105), struck-by (82), electrocution (49), caught-in-between (30), chemical exposure (24), fatigue-related (10), noise (6), collapse (5), and fire/explosion (1).

These counts are shown in Figure 7. The four categories with 10 or fewer test instances fatigue-related, noise, collapse, and fire/explosion are treated as low-support categories; their category-level results are shown descriptively and are not used for inferential claims. The main conclusions are based on aggregate results and on the six categories with greater support.

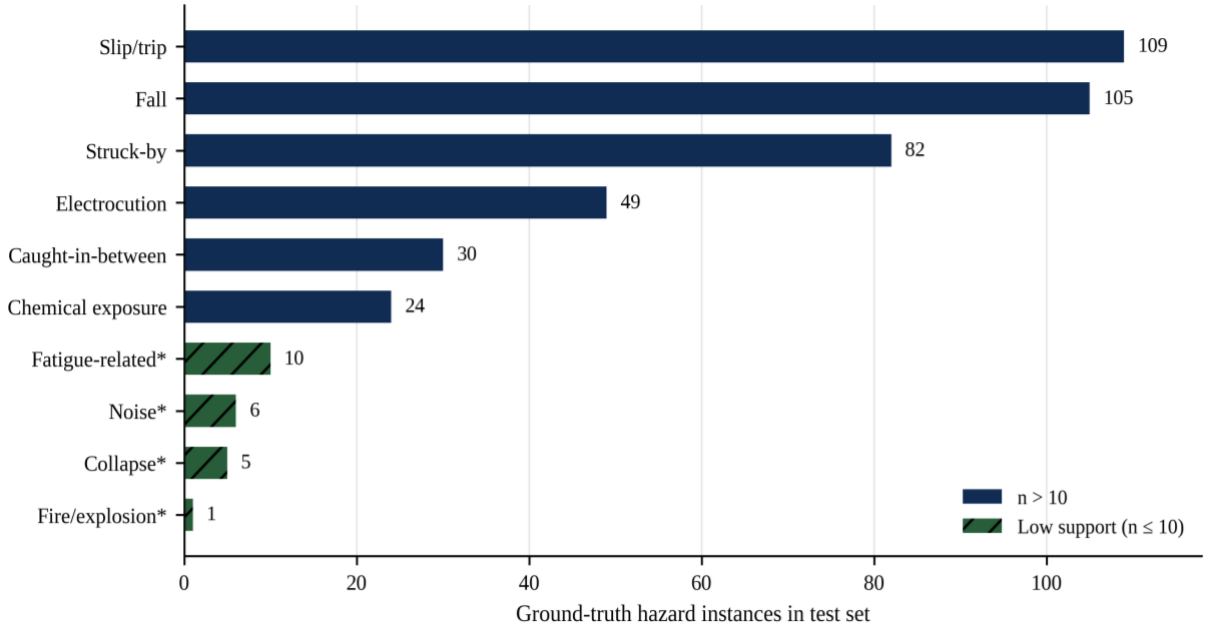

Figure 7: Ground-truth hazard-instance support in the held-out test set (n = 421). Categories with 10 or fewer instances are designated low-support and are interpreted descriptively only.

Leakage control. Separation between the retrieval knowledge base and the test set was enforced at multiple levels. *First,* images were partitioned within each source using the original dataset identifiers and, where available, project, inspection-case, or site-folder information. Images belonging to the same scene, project, activity sequence, or hazard instance were assigned to only one subset.

Second, automated near-duplicate screening was applied before the final dataset was fixed. A 64-bit perceptual hash (pHash) was used to flag cross-subset image pairs with a Hamming distance of ≤5, while CLIP embedding similarity was used to flag pairs with cosine similarity >0.95. The screening identified 17 candidate cross-subset pairs for manual review. Of these, 5 test images were judged to represent the same or substantially overlapping scene as retrieval images and were excluded before formation of the final test set.

Following this procedure, the final dataset comprised 1,300 images: 1,040 retrieval images and 260 held-out test images. The maximum observed cross-subset CLIP cosine similarity after duplicate removal was 0.947, and the minimum cross-subset pHash Hamming distance was 6. Together with manual scene-level review, these checks were used to reduce the risk that near-duplicate visual precedents could inflate retrieval or hazard-detection performance. Table 3 reports the resulting allocation by source together with the separation rule applied to each.

Table 3: Allocation of images to the retrieval knowledge base and the test set, by source.

Source	Retrieval images	Test images	Total	Separation rule applied
SODA (Duan et al., 2022)	480	120	600	Split by original image and scene identifiers; no scene shared
OSHA inspection records (Occupational Safety and Health Administration, 2024)	320	80	400	Split by inspection case and project identifier
Prior computer-vision safety studies	240	60	300	Split by source publication, project, and site folder
Total	1,040	260	1,300	—

4.3 Model configurations and testing setup

Models and configurations. The primary comparison contrasts the unconstrained GPT-4V baseline with the complete proposed framework, while the component-level evaluation comprises nine configurations on the same held-out test set. The baseline receives only the input construction image and a general hazard-identification prompt, with no retrieved context, taxonomy constraints, or structured reasoning instructions. GPT-4V is retained as the reference baseline because it is widely benchmarked in recent construction-safety studies.

The claim that hallucination is not resolved simply by upgrading the backbone is tested rather than assumed: Section 5.1 reports the baseline and the complete pipeline across four additional backbones, including two open-weight models. The proposed framework extends the baseline with CLIP image encoding, FAISS-based image-to-image retrieval of the top five similar hazard cases, taxonomy-constrained condition reasoning, gated sentence-transformer-based regulation retrieval, and structured report generation.

Configurations evaluated. To separate the contribution of each component, nine configurations are evaluated on the same 260-image test set: (1) GPT-4V with a generic hazard-identification prompt, the unconstrained baseline; (2) GPT-4V with the structured prompt only, without retrieval or taxonomy, isolating the effect of prompt structure; (3) GPT-4V with self-consistency, generating five responses and taking the majority label, isolating the effect of output stability; (4) GPT-4V with a post-hoc verifier applied to the baseline output, representing output-level correction; (5) GPT-4V with text-only retrieval, in which regulations are retrieved before visual confirmation, representing the arrangement adopted in prior RAG-based safety systems; (6) GPT-4V with image-to-image retrieval and no taxonomy, isolating visual grounding; (7) GPT-4V with the taxonomy and no retrieval, isolating structured constraints; (8) GPT-4V with both retrieval and taxonomy; and (9) the full pipeline, which adds gated regulatory retrieval. Configurations 6 to 9 correspond to the pipeline stages described in Section 3; configurations 2 to 5 correspond to the mitigation strategies reviewed in Section 2.2 and are included so that the proposed framework is compared against the alternatives available to a practitioner rather than only against an unconstrained model.

Backbone models. Because the framework constrains inference rather than modifying the model, its behaviour should not depend on a particular backbone. To test this, the baseline and the full pipeline were additionally evaluated with GPT-4o, Gemini 2.5 Pro, Qwen2.5-VL-7B and LLaVA-1.6, the last two of which are open-weights models that can be run without commercial API access. Results are reported in Section 5.1.

Implementation details. Images are encoded with CLIP ViT-B/32 (openai/clip-vit-base-patch32), producing 512-dimensional vectors indexed in FAISS using an IVF-Flat index with nlist = 32, nprobe = 8, and inner-product similarity on L2-normalised vectors. Regulatory passages are encoded with sentence-transformers/all-mpnet-base-v2, producing 768-dimensional embeddings ranked by cosine similarity. The structured reasoning prompt passes the input image, the top five retrieved metadata records, and explicit condition-evaluation instructions to the VLM. The configuration-level instruction templates are reported directly below so that the nine experimental conditions and their runtime inputs are auditable in the main manuscript. The top three candidate regulatory provisions per confirmed hazard are returned after filtering by hazard category and jurisdictional context. Sensitivity to retrieval depth and encoder choice is reported in Section 5.1.

Prompt specification and runtime placeholders. Table 4 reports the instruction templates for the nine configurations. Bracketed terms denote runtime inputs: [QUERY_IMAGE], retrieved RAG-1 cases, the hazard taxonomy, jurisdiction, regulatory passages, baseline output, and Stage-A confirmed hazards. These placeholders separate fixed prompt instructions from case-specific content supplied during inference.

Prompt reproducibility note. Configurations 1-9 differ only as specified above with respect to prompt structure, repeated generation/verification procedure, retrieved visual context, taxonomy constraints, and regulatory context. For the additional backbone evaluations in Table 6, the same configuration-level instruction content is used, with provider-specific message and image wrappers adapted to the corresponding API. Provider-specific transport wrappers are not treated as experimental instruction content and are not reproduced here.

Table 4: Configuration-level prompt templates for the nine evaluated configurations. Bracketed fields denote runtime content rather than literal prompt text.

Configuration	Prompt / procedure
1. Generic-prompt baseline	Input: [QUERY_IMAGE] Analyse this construction-site image. What safety hazards are present? List each hazard you identify and briefly explain it based on the image.
2. Structured prompt only	Input: [QUERY_IMAGE] Analyse the construction-site image in a structured sequence. 1. State only the safety-relevant scene elements that are visually observable. 2. Identify any construction-safety hazard categories supported by those observations. 3. For each reported hazard, state the visible evidence supporting the claim. 4. If the image does not provide sufficient visual evidence for a hazard, do not report that hazard. Return a concise hazard list with the supporting visual evidence for each item.
3. Self-consistency (n = 5)	Use the Configuration 2 prompt. Execute it independently five times for the same [QUERY_IMAGE]. Convert each response to the ten Level 1 hazard labels, then retain the majority label decision for each category. This configuration changes the sampling/aggregation procedure rather than adding retrieval, taxonomy, or regulatory context.
4. Post-hoc verifier	Input image: [QUERY_IMAGE] Initial hazard output (Configuration 1 output): [BASELINE_OUTPUT] Verify each hazard claim in the initial output against the image. For each claim: (i) identify the visual evidence that supports it; (ii) mark the claim as SUPPORTED only if the required evidence is visible; and (iii) otherwise mark it as UNSUPPORTED and remove it from the final hazard list. Do not add a new hazard that was not present in the initial output. Return only the verified hazard list and a brief evidence statement for each retained claim.
5. Text-RAG only (regulations first)	Input image: [QUERY_IMAGE] Jurisdiction: [JURISDICTION] pre-confirmation regulatory context: [PRECONFIRMATION_REGULATORY_CONTEXT] Using the construction-site image and the regulatory passages retrieved before visual-precedent/taxonomy verification, identify the safety hazards that are present. For each reported hazard, provide a brief image-based explanation and identify any retrieved provision that appears relevant. Do not use RAG-1 visual precedent retrieval or the three-level taxonomy in this configuration.
6. RAG-1 image retrieval, no taxonomy	Query image: [QUERY_IMAGE] Retrieved visual precedents and metadata: [TOP5_RETRIEVED_CASES_JSON] Analyse the query image using the retrieved construction cases as visual precedent. Identify the safety hazards supported by the query image and explain briefly how the visible scene is consistent with the retrieved cases. Base the final hazard claims on the query image; do not report a hazard solely because it appears in the retrieved metadata.
7. Taxonomy only, no retrieval	Query image: [QUERY_IMAGE] Hazard taxonomy and observable indicators: [HAZARD_TAXONOMY_JSON] Evaluate the ten hazard categories using the supplied three-level taxonomy. For each candidate hazard: (1) identify at least one observable human-behaviour indicator; (2) identify at least one observable unsafe physical-condition indicator; and (3) identify at least one observable organizational/system-control indicator. Confirm the Level 1 hazard only if all three Level 2 classes have at least one observable supporting indicator in the query image. If any required class is absent or cannot be resolved from the image, withhold the hazard label. Return the confirmed hazard(s) and the specific observable indicators supporting each class.
8. RAG-1 + taxonomy verification gate	Query image: [QUERY_IMAGE] Retrieved visual precedents and metadata: [TOP5_RETRIEVED_CASES_JSON] Hazard taxonomy and observable indicators: [HAZARD_TAXONOMY_JSON] Reason in the following order. 1. Identify the safety-relevant elements that are directly observable in the query image. 2. Use the retrieved cases only as contextual visual precedent; do not copy a hazard label from retrieval unless the query image supports it. 3. For each candidate Level 1 hazard, evaluate the three Level 2 classes: human behaviour, unsafe physical condition, and organizational/system failure.

Configuration	Prompt / procedure
	4. Confirm a hazard only when at least one required Level 3 indicator from each of the three classes is observable in the query image. 5. If one or more required classes cannot be verified, withhold the hazard label. Return, for each confirmed hazard: the Level 1 category, the observed human-behaviour indicator(s), unsafe-condition indicator(s), system-failure/control indicator(s), and a concise evidence-based explanation.
9. End-to-end proposed pipeline	STAGE A - HAZARD CONFIRMATION Use the Configuration 8 prompt with [QUERY_IMAGE], [TOP5_RETRIEVED_CASES_JSON], and [HAZARD_TAXONOMY_JSON]. If no hazard satisfies the three-class activation rule, return NO CONFIRMED HAZARD and stop; do not initiate regulatory retrieval. STAGE B - REGULATORY QUERY (only for confirmed hazards) For each confirmed Level 1 hazard, query the regulation vector database using the confirmed hazard category and [JURISDICTION]. Return the top three candidate provisions after hazard-category and jurisdiction filtering. STAGE C - FINAL REPORT GENERATION Confirmed hazards and visual evidence: [STAGE_A_OUTPUT] Jurisdiction: [JURISDICTION] Candidate regulatory provisions: [TOP3_REGULATORY_PROVISIONS] Generate a structured construction-safety report. For each confirmed hazard include: the confirmed hazard category; visual evidence and the condition-based explanation that satisfied the activation rule; the retrieved regulatory provision(s) as candidate compliance support; and a concise corrective action supported by the available evidence and retrieved provision(s). Do not introduce a hazard that was not confirmed in Stage A. Do not present a retrieved provision as an infallible compliance determination; it is an evidence-linked candidate that may require professional verification.

4.4 Evaluation metrics

Four metrics are used to assess performance. Precision, recall, and F1-score measure the correctness and completeness of reported hazard claims; hallucination rate measures the safety-specific reliability of those claims, a dimension that standard object-detection benchmarks do not capture. For configurations 1-8, the scored output is the hazard-claim set emitted at that configuration's final decision stage. For configuration 9, the end-to-end score is computed on hazard claims appearing in the final regulation-conditioned report. The hazard-confirmation gate itself is not changed by RAG-2; therefore, any difference between configurations 8 and 9 reflects downstream report-generation claims rather than a change in the confirmed gate label.

Precision measures the proportion of reported hazards that are genuinely present in the image:

$$Precision = TP / (TP + FP) \tag{2}$$

High precision means the system generates few false hazard reports critical for maintaining practitioner trust in automated safety monitoring.

Recall measures the proportion of actual hazards that the system successfully detects:

$$Recall = TP / (TP + FN) \tag{3}$$

High recall ensures that real safety risks are not missed during automated inspection.

F1-score balances precision and recall into a single performance measure:

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \tag{4}$$

F1-score balances precision and recall at the configuration level. Because category support varies substantially, per-category F1 is not used as a primary inferential result; Figure 7 instead reports category support, while Figure 9 reports per-category recall descriptively.

Hallucination rate measures the proportion of hazards that are not supported by sufficient visual evidence in the input image. In construction safety analysis, hallucinated hazards can lead to incorrect safety assessments and reduce confidence in automated monitoring systems.

In this context, hallucination represents a specific subset of false positive errors. While all hallucinated predictions are false positives, not all false positives are hallucinations. For example, a system may incorrectly detect a hazard due to poor image quality, occlusion, or visual ambiguity, even when following correct reasoning procedures. In contrast, hallucination refers specifically to cases where hazards are inferred without sufficient supporting visual evidence. Accordingly, the hallucination rate is defined as:

$$Hallucination\ Rate = H / (TP+FP) \tag{5}$$

where H denotes the number of hallucinated hazard predictions, identified because the reported hazard lacks sufficient supporting visual evidence in the input image. This metric is reported separately from precision to emphasize the safety-specific interpretation of unsupported hazard claims. Although related to the false-discovery rate, hallucination is defined by the annotation rule rather than by precision alone. Each false positive was classified explicitly: annotators inspected the image at full resolution and determined whether at least one Level 3 indicator required by the reported hazard category was identifiable. A false positive was labelled a hallucination if no required supporting indicator was identifiable. False positives in which a relevant indicator was visible, but the overall configuration was safe, or in which image degradation made the evidence unresolvable, were recorded separately. Two annotators independently applied this rule to all false positives; agreement on the hallucination judgement was Cohen's $\kappa = 0.89$, and 12 disagreements were adjudicated by a third reviewer. Because precision is displayed to three decimals and hallucination to one decimal percentage point, arithmetic on the printed summaries can create apparent boundary effects. For configuration 1, the displayed hallucination rate is 55.0%, whereas 1 - the displayed precision (0.451) gives a false-discovery rate of 54.9%; for configuration 9, the displayed hallucination rate (6.1%) and 1 - 0.939 = 6.1% coincide. These printed values are independently rounded summaries and should not be read as exact complements or as exact subset bounds; the definition of hallucination as a subset of false positives applies to the underlying unrounded claim-level quantities.

Statistical treatment. F1 scores reported in Table 5 and pairwise F1 differences are accompanied by 95% bootstrap confidence intervals obtained by resampling the 260 images 1,000 times using the percentile method. Because every configuration is evaluated on the same images, configurations are compared pairwise rather than treated as independent samples: differences in per-image correctness are tested with McNemar's test, and F1 differences with a paired bootstrap test over the same resamples. Reported p-values are adjusted for multiple comparisons using the Holm-Bonferroni procedure across the 36 planned comparisons. Effect sizes are reported as absolute differences together with their confidence intervals where applicable.

4.5 Practitioner observations

To complement the quantitative evaluation, expert observations were gathered to examine the practical usefulness and interpretability of the generated hazard reports. Two construction-safety professionals independently reviewed a sample of outputs from the baseline and the proposed framework for the same images: a site safety manager with three years of inspection experience and an HSE coordinator with three years of relevant experience. These practitioners were not members of the annotation team and did not assign ground-truth labels or hallucination judgements; the practitioner review and annotation processes therefore served separate purposes. Their professional roles align with the intended user group for site-level hazard reporting, which is why practitioner review was used to assess report legibility and practical interpretability. The review was guided by three questions: (i) whether the identified hazard was realistic given the visible scene conditions, (ii) whether the explanation described the hazard in condition-based terms, and (iii) whether the output would support practical safety decision-making on site.

The practitioner assessment was exploratory rather than controlled, and judgements were recorded qualitatively rather than on a numerical rating scale; no inter-rater statistic is therefore reported for this exercise. Its purpose is deliberately narrower than the quantitative evaluation. The quantitative analysis establishes detection and hallucination performance, whereas the practitioner observations address whether the resulting structured reports are understandable and operationally interpretable to the intended user group. These observations are not used to support quantitative claims about adoption, inspection speed, or decision quality. The limited practitioner sample is acknowledged in Section 7.

4.6 Reproducibility

Because this study is positioned as applied systems research, the components that determine its behaviour are specified rather than summarised. The hazard-taxonomy structure and AND-logic activation rule are specified in Figure 2 and Section 3.1; the category-specific taxonomy JSON used at runtime is archived by the authors and is available from the corresponding author on request. The configuration-level prompt templates and runtime-input placeholders are reported directly in Section 4.3, and the embedding-model identifiers together with the FAISS index parameters are stated there in full.

The image data are drawn entirely from public sources: the SODA dataset (Duan et al., 2022), OSHA photographic inspection records (Occupational Safety and Health Administration, 2024), and images compiled in prior computer-vision safety studies and are identified by source in Section 4.1 rather than redistributed, in accordance with the licences under which they were released. These materials provide sufficient detail to reproduce the reported configuration logic, retrieval settings, and evaluation protocol; exact replication of the category-specific taxonomy contents requires the archived taxonomy JSON.

5. RESULTS AND ANALYSIS

The evaluation examines the proposed framework from two complementary perspectives. The quantitative analysis measures hazard-claim accuracy and the frequency of unsupported claims. The qualitative analysis examines whether the structured outputs are understandable and useful to construction-safety practitioners. These perspectives address different questions and are interpreted separately throughout the results.

5.1 Quantitative results

Table 5: Component-level ablation on the same 260-image test set. Values in brackets are 95% bootstrap confidence intervals over 1,000 image-level resamples. Configurations 1–8 are scored on the hazard-claim set emitted at their decision stage; configuration 9 is scored end-to-end on hazard claims in the final regulation-conditioned report. Retrieval coverage is not applicable (n/a) where image retrieval is not used.

No. / Configuration		Precision	Recall	F1 [95% CI]	Hallucination rate	Coverage
1	GPT-4V, generic prompt (baseline)	0.451	0.658	0.535 [0.487, 0.581]	55.0%	n/a
2	+ structured prompt only	0.612	0.703	0.654 [0.610, 0.695]	34.2%	n/a
3	+ self-consistency (n = 5)	0.674	0.724	0.698 [0.658, 0.736]	27.8%	n/a
4	+ post-hoc verifier	0.758	0.706	0.731 [0.694, 0.765]	18.5%	n/a
5	+ text-RAG only (regulations first)	0.721	0.746	0.733 [0.696, 0.768]	24.1%	n/a
6	+ RAG-1 image retrieval, no taxonomy	0.950	0.898	0.923 [0.899, 0.944]	4.3%	89.8%
7	+ taxonomy only, no retrieval	0.902	0.861	0.881 [0.850, 0.907]	8.1%	n/a
8	+ RAG-1 + taxonomy	0.987	0.899	0.941 [0.922, 0.957]	1.0%	89.8%
9	+ RAG-2 + report generation (end-to-end proposed)	0.939	0.896	0.917 [0.892, 0.938]	6.1%	89.8%

The central claim of this paper is that grounding hazard reasoning in visual evidence and explicit condition structure reduces unsupported hazard claims while maintaining strong detection performance. Table 5 shows that the end-to-end report-level hallucination rate decreases from 55.0% for the unconstrained VLM baseline to 6.1% for the complete pipeline, an 88.9% relative reduction, while report-level F1 increases from 0.535 to 0.917 and precision from 0.451 to 0.939. These rates are claim-level quantities: 55.0% and 6.1% refer to reported hazard claims, not to the proportion of images containing at least one hallucination. The RAG-1 image-retrieval configuration alone achieves a precision of 0.950, recall of 0.898, F1-score of 0.923, and hallucination rate of 4.3%. This confirms that visual precedent retrieval is itself a strong grounding mechanism. RAG-1 achieves 89.8%

coverage, meaning that at least one relevant precedent is retrieved for 89.8% of the 421 ground-truth hazard instances. The taxonomy-constrained gate then reaches the strongest hazard-decision performance, with precision 0.987 and F1 0.941.

All pairwise comparisons were conducted on the same 260 images and are reported with Holm–Bonferroni adjustment across the 36 planned comparisons. The improvement of the full pipeline over the unconstrained baseline is substantial: $\Delta F1 = 0.382$ (95% CI [0.334, 0.427]). The paired image-level correctness comparison also favours the pipeline (McNemar $p = 3.21 \times 10^{-16}$), with 112 images corrected by the pipeline against 18 degraded by it. Component ablations show the same pattern: adding taxonomy-constrained verification to retrieval alone (configuration 6 → 8) yields $\Delta F1 = 0.018$, with McNemar $p = 0.012$ for paired image-level correctness and adding retrieval to the taxonomy alone (configuration 7 → 8) yields $\Delta F1 = 0.060$, with McNemar $p = 9.12 \times 10^{-4}$ for paired image-level correctness.

Table 5 allows the contribution of each component to be read directly, and Figure 8 presents the same decomposition graphically. Configurations 2 to 5 represent the mitigation strategies reviewed in Section 2.2. Configuration 7 applies the taxonomy without retrieval so that the contributions of visual grounding and structured constraints can be separated rather than inferred.

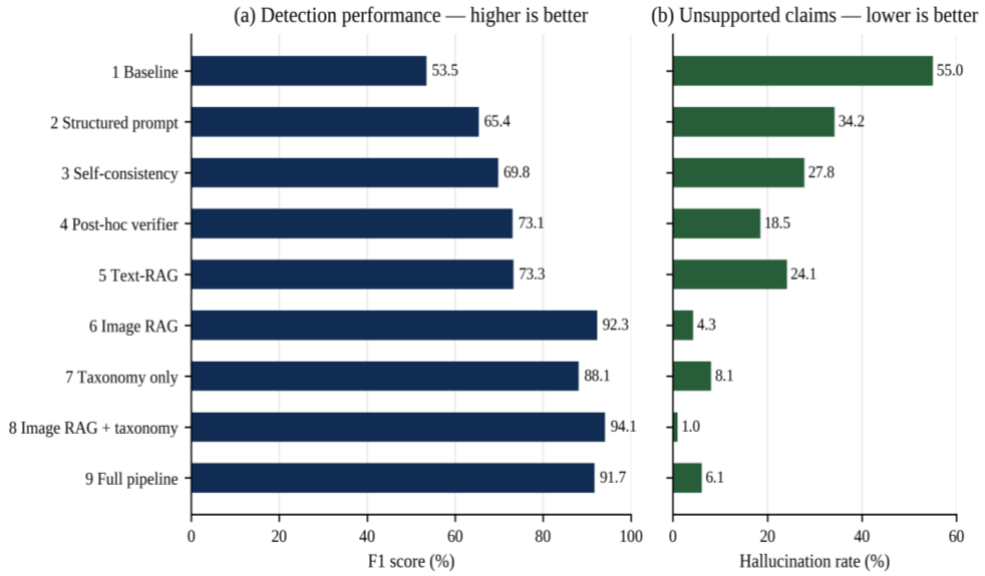

Figure 8: Component-level ablation across all nine configurations. Panel (a) reports F1 score and panel (b) reports claim-level hallucination rate for the same 260-image test set. Values correspond directly to Table 5 and are printed at the end of each bar; higher F1 and lower hallucination are better.

Table 6: Backbone generalisation. Baseline and full pipeline evaluated with several vision-language models on the same test set.

Backbone	Baseline F1	Baseline hallucination	End-to-end report F1	End-to-end report hallucination	Absolute reduction
GPT-4V	0.535	55.0%	0.917	6.1%	48.9 pp
GPT-4o	0.602	47.8%	0.925	5.4%	42.4 pp
Gemini 2.5 Pro	0.621	44.9%	0.928	4.8%	40.1 pp
Qwen2.5-VL-7B (open weights)	0.487	58.6%	0.889	8.7%	49.9 pp
LLaVA-1.6 (open weights)	0.441	62.3%	0.861	11.2%	51.1 pp

Table 6 indicates that the framework's reliability benefit is observed across all tested VLM backbones. Every model exhibits substantial hallucination in the unconstrained configuration and a markedly lower end-to-end report hallucination rate under the proposed pipeline. This pattern supports the conclusion that the reliability gain is not specific to GPT-4V. Because the number and architecture of tested backbones remain limited, we interpret the result as evidence of cross-backbone transfer within the evaluated models rather than as proof of universal backbone independence.

Adding taxonomy-constrained condition verification produces the strongest hazard-decision configuration, with precision 0.987, recall 0.899, F1-score 0.941, and a 1.0% hallucination rate. This configuration reports a hazard only when the AND-logic activation conditions defined in Section 3.1 are satisfied. The result shows that explicit evidence requirements reduce unsupported hazard claims to a very low level on the present test set.

The end-to-end report score is lower than the hazard-verification-gate score (precision 0.939 vs. 0.987; F1 0.917 vs. 0.941; hallucination 6.1% vs. 1.0%). This does not mean that RAG-2 changes the confirmed hazard label. Configuration 8 is scored at the verification gate, whereas configuration 9 is scored on hazard claims that appear in the final regulation-conditioned report. In a small number of cases, downstream report generation introduces unsupported hazard-associated wording or associations after regulations are retrieved; these additional claims are counted as false positives or hallucinations in the end-to-end evaluation. RAG-2 reliability itself is evaluated separately using regulation precision@3 and jurisdiction error in Section 5.3.The F1 difference between configuration 8 and the end-to-end report output is $\Delta F1 = -0.024$ (95% CI [−0.052, 0.005]); because the interval includes zero, we do not interpret this as evidence of an F1 improvement. The paired image-level correctness comparison likewise does not indicate a difference (McNemar $p = 0.218$ after Holm–Bonferroni adjustment). We therefore do not claim that adding the compliance stage improves hazard-detection accuracy. Its operational value is the addition of regulation-linked output after hazard confirmation, and its retrieval reliability is assessed separately.

Figure 9 presents descriptive per-category recall for the proposed end-to-end pipeline across the ten analytical hazard categories, with test-set support shown beside every category. The support-weighted average of the displayed category recalls is 89.6% after rounding, consistent with the aggregate recall of 0.896 reported in Table 5. Four categories - fatigue-related (n = 10), noise (n = 6), collapse (n = 5), and fire/explosion (n = 1) - are designated low-support; their values are shown for completeness and are not used for category-specific inference. Electrocution has 49 test instances and is therefore not treated as low support. Because Figure 9 reports recall only, conclusions about false-positive reduction are based on the precision and hallucination analyses rather than on this figure.

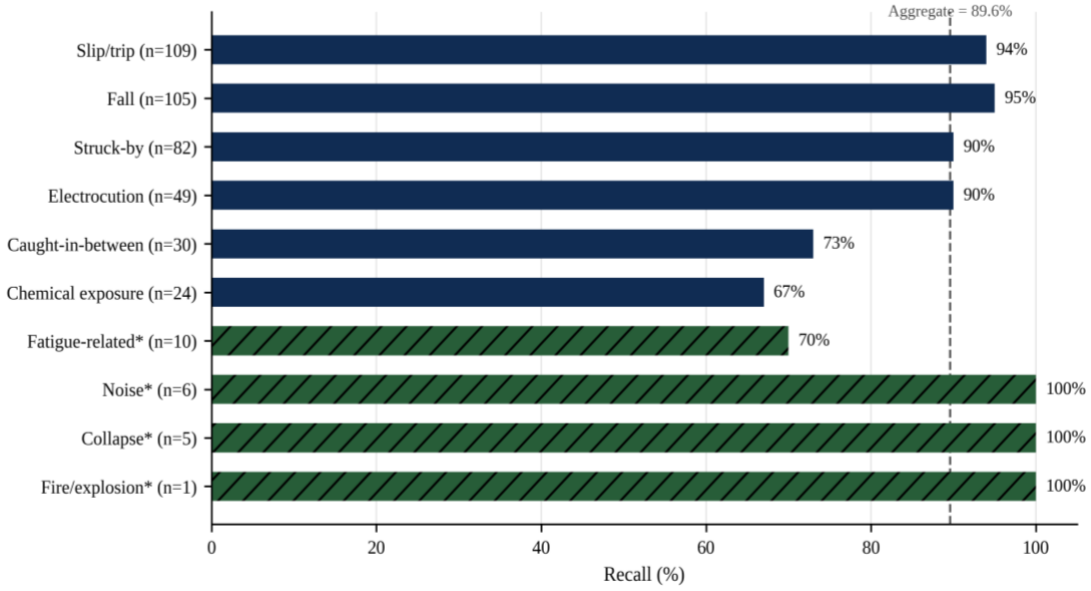

Figure 9: Per-category recall of the end-to-end proposed pipeline across the ten hazard categories. Bars show the category-level recall percentages and test-set support (n); the dashed line marks the aggregate recall of 89.6% from Table 5. Asterisks identify low-support categories (≤10 test instances), which are shown descriptively only.

Table 7: Sensitivity of the end-to-end report pipeline to retrieval depth k.

k	Coverage	Precision	Recall	F1	Hallucination	Latency (s)
1	68.5%	0.902	0.842	0.871	8.9%	2.8
3	83.7%	0.928	0.881	0.904	6.8%	3.2
5	89.8%	0.939	0.896	0.917	6.1%	3.6
10	94.6%	0.932	0.901	0.916	6.5%	4.4
20	97.1%	0.914	0.903	0.908	8.0%	5.8

Table 8: Qualitative comparison of VLM-only vs. proposed model outputs on representative test images.

SCENE DESCRIPTION	VLM-ONLY OUTPUT (GPT-4V)	PROPOSED MODEL OUTPUT	REVIEWER COMMENT
	Active hazards include an operating excavator, a worker inside the excavation, falling soil, and unprotected trench edges.	Fall/trench collapse hazard: worker in trench + unstable walls + no protective system. Struck-by risk from the excavator was also confirmed.	Proposed model provides condition-constrained and structured hazard explanation; VLM output is generic.
	Active hazards include working at height without protection, unstable surface, exposed rebar, and carrying materials.	Fall-from-height hazard: worker at edge + unstable rebar surface + no guardrails. Impalement risk was also identified.	VLM shows hallucinated or unsupported hazards; the proposed model reports only visually grounded risks.
	Active hazards include manual handling risks, obstructed vision, and load instability.	No visually confirmed hazard: activity appears controlled with proper PPE and no clear unsafe conditions.	VLM introduces hallucinated hazards; the proposed model correctly reports none.
	Active hazards include distraction and reduced awareness in a construction environment.	Fall-from-scaffold hazard linked to visible elevated structure; no unsafe worker behaviour observed.	VLM introduces unsupported behavioural hazards; the proposed model remains visually grounded.
	Active hazards include dust exposure, debris, noise, hose risks, and uneven ground.	Dust exposure from visible dust cloud + trip hazard due to uneven ground and debris.	VLM over-detects hazards (hallucination); the proposed model remains visually grounded.

Sensitivity to retrieval depth. Retrieval depth was varied over $k \in \{1, 3, 5, 10, 20\}$ with all other settings held fixed (Table 7). Coverage increases monotonically with k, while end-to-end report F1 plateaus near k = 5; at larger k, weaker visual matches are added and context bleeding increases. The proportion of residual false positives attributed to metadata influence rises from 33% at k = 5 to 52% at k = 20. Latency and token consumption also increase with k, so k = 5 was adopted as the operating point.

Sensitivity to encoder choice. Because both retrieval stages depend on pretrained encoders, encoder choice was tested rather than assumed. Replacing the image encoder with ViT-L/14 changes retrieval coverage from 89.8% to 92.4% and end-to-end report F1 from 0.917 to 0.923. Replacing the regulation encoder with all-MiniLM-L6-v2 changes regulation precision@3 from 91.4% to 88.7%. Within the evaluated alternatives, the framework is therefore not strongly sensitive to encoder choice, and smaller encoders may be useful when local deployment or lower computational cost is prioritized.

The recall pattern is also informative at the aggregate level. The unconstrained baseline achieves recall 0.658 while hallucinating 55.0% of its reported hazard claims. The end-to-end pipeline raises report-level recall to 0.896 while reducing the claim-level hallucination rate to 6.1%. At the verification gate, RAG-1 plus taxonomy reaches recall 0.899 and a 1.0% hallucination rate. These results show that the principal reliability gain occurs before the regulatory stage, while the downstream compliance/report stage adds regulatory context at a modest end-to-end claim-generation cost.

5.2 Qualitative analysis

The exploratory practitioner observations were consistent with the quantitative reliability findings and provided additional insight into report interpretability. Reports produced by the proposed framework were described as grounded and actionable because identified hazards were accompanied by the specific combination of worker behaviour, physical site conditions, and organizational factors that satisfied the activation conditions.

The practitioners noted that this structure made the basis of the system's judgement easier to inspect and prioritize. In contrast, VLM-only outputs were often described as generic and, in several cases, implausible because some reported hazards were not observable in the scene. These observations are used to support claims about interpretability and practical readability, not to establish quantitative gains in decision quality or adoption.

Overall, the qualitative observations complement rather than duplicate the quantitative evaluation. Table 8 presents representative examples of three recurring patterns: (1) the proposed framework provides structured condition-based explanations that are absent from the VLM baseline; (2) it can withhold a hazard label when sufficient visual evidence is not present; and (3) the unconstrained VLM can introduce unsupported hazards that cannot be verified from the image.

5.3 Deployment characteristics: Latency and cost

A framework intended for site monitoring should report its operational burden. Latency was measured over the 260 test images, including API and network round-trip time. Computational cost is reported primarily through input/output token consumption because commercial API prices vary by provider and over time.

Table 9 reports per-stage latency and token requirements, and the text below provides a pricing-independent formula that can be updated using the provider rates applicable at deployment.

Table 9: Per-stage inference latency and token consumption of the proposed pipeline.

Pipeline stage	Mean latency (s)	p95 latency (s)	Input tokens per image	Output tokens per image
CLIP encoding of the query image	0.07	0.12	n/a	n/a
FAISS retrieval (k = 5)	0.01	0.02	n/a	n/a
Condition-constrained reasoning (VLM)	2.46	4.21	2,460	420
Regulation retrieval (RAG-2, top 3)	0.05	0.09	n/a	n/a
Report generation (LLM)	1.03	1.88	1,280	610
End-to-end	**3.62**	**6.03**	**3,740**	**1,030**
Baseline (VLM only), for comparison	2.12	3.74	1,080	310

End-to-end processing requires an average of 3.62 s per image and consumes approximately 3,740 input tokens and 1,030 output tokens. The condition-constrained VLM reasoning stage accounts for approximately 68.0% of total latency and 60.4% of total token consumption, making it the dominant computational component. To convert token use into monetary API cost without fixing the paper to a particular provider price, let $P_{in,V}$ and $P_{out,V}$ denote the VLM input/output prices in USD per million tokens, and $P_{in,L}$ and $P_{out,L}$ the report-generation model prices. The estimated per-image model-call cost is $C_{image} = 0.00246P_{in,V} + 0.00042P_{out,V} + 0.00128P_{in, L} + 0.00061P_{out,L}$. This formula excludes local CLIP/FAISS computation and can be updated directly when provider prices change.

At this throughput, the framework is better suited to periodic inspections, event-triggered image analysis, or near-real-time single-image assessment than to continuous frame-by-frame video processing. Because most latency and token consumption arise from the VLM reasoning and report-generation stages rather than retrieval itself, deployment cost is governed primarily by the selected language-model backbone. The backbone results suggest that smaller or locally deployed open-weight models may provide a practical alternative where data locality or reduced operating cost is prioritized and a modest reduction in predictive performance is acceptable.

Reliability of regulatory retrieval. Because a compliance report depends not only on hazard confirmation but also on the regulatory information attached to each confirmed hazard, RAG-2 was evaluated separately. For each confirmed hazard, the three highest-ranked retrieved provisions were independently rated by two construction-safety professionals as relevant, partially relevant, or irrelevant to the identified hazard and its jurisdictional context. For precision@3, relevant and partially relevant provisions are counted as acceptable retrievals and irrelevant provisions as retrieval errors; precision@3 is the proportion of acceptable items among all top three provision candidates evaluated. The jurisdiction mis-assignment rate is computed over the same retrieved-provision set and counts candidates assigned to the wrong jurisdiction. Regulation precision@3 was 91.4%; 8.6% of retrieved provisions were classified as irrelevant, and the jurisdiction mis-assignment rate was 1.7%. These results show that gating sharply limits premature regulatory retrieval but does not guarantee perfect provision relevance, so final professional verification remains appropriate before a retrieved provision is treated as a definitive compliance determination. The most frequent retrieval failure involved a provision that was broadly related to the correct safety domain but did not address the specific observable hazard configuration. This occurred most frequently for electrocution hazards, where visually similar electrical conditions can correspond to different regulatory requirements depending on equipment state, worker proximity, and the type of protective control present. These findings indicate that gated retrieval substantially restricts irrelevant regulatory content, although a final verification step remains appropriate before regulatory information is presented as a definitive compliance judgement.

5.4 Robustness under degraded imaging conditions

Construction site imagery is frequently affected by illumination, blur, occlusion, and other environmental limitations. The robustness of the proposed pipeline was therefore examined separately under three degraded imaging conditions. Test images were tagged for low illumination, defined as mean luminance below the 15th percentile of the test-set distribution; motion blur or defocus, defined using a variance-of-Laplacian value below 100; and partial occlusion of the hazard-relevant image region, which was identified manually. These tags were assigned independently of model outputs, and an image could belong to more than one degradation category. Performance for each subset is reported in Table 10.

Table 10: Performance of the proposed pipeline under degraded imaging conditions.

Subset	Images	Precision	Recall	F1	Hallucination rate
Unaffected	154	0.961	0.935	0.948	3.0%
Low illumination	44	0.928	0.846	0.885	5.2%
Motion blur or defocus	32	0.907	0.772	0.834	6.8%
Partial occlusion	51	0.919	0.806	0.859	5.9%

Performance decreases under all three degradation conditions, although the largest reduction occurs for motion blur or defocus, where F1 falls from 0.948 for unaffected images to 0.834. Recall is affected more strongly than precision, decreasing to 0.772, while precision remains above 0.90. A similar, although less pronounced, pattern

occurs under partial occlusion and low illumination. This behaviour is consistent with the qualitative residual-error analysis, which identifies image quality as a recurring source of uncertainty. The reduction in recall is also consistent with the framework's condition-constrained decision mechanism. Hazard activation requires sufficient positive support for the hazard-specific observable evidence conditions; when blur, darkness, or occlusion makes one or more required conditions difficult to establish, the system is more likely to withhold a hazard decision than to infer the missing evidence. The motion-blur/defocus subset illustrates both effects: recall falls from 0.935 for unaffected images to 0.772 (-0.163), while the hallucination rate rises from 3.0% to 6.8% (+3.8 percentage points). The larger absolute change is therefore in recall, so degraded imagery primarily increases missed detections, but it also materially increases unsupported hazard claims. From a deployment perspective, this result supports the inclusion of an explicit image-quality screening stage before hazard reasoning. Images that fail minimum quality requirements for sharpness, illumination, visibility, or occlusion should be marked as insufficient for reliable automated assessment and referred for human review or image recapture rather than being interpreted as evidence that no hazard is present. Such a mechanism would preserve the conservative evidence boundary of the framework while reducing the risk of false reassurance under poor imaging conditions.

6. DISCUSSION

The quantitative results presented in Section 5 demonstrate that the proposed framework achieves strong hazard detection performance while maintaining a low hallucination rate of 6.1%, significantly below the baseline. These outcomes can be attributed to three key mechanisms embedded in the framework's design. The proposed framework reduces false detections by changing how hazards are inferred using evidence grounding and structured taxonomy constraints rather than free-form generative interpretation as detailed in Section 3.

First, the framework provides cognitive grounding by anchoring hazard reasoning in similar real-world visual cases. Instead of analysing an image in isolation, the model retrieves visually similar construction scenarios with known hazard contexts. These retrieved cases act as prior experience, allowing the model to reason in a way that is closer to how human safety professionals rely on past observations and near-miss experience. This grounding discourages speculative hazard generation and encourages evidence-based reasoning.

Second, the framework enforces constraint-based reasoning through structured hazard taxonomies and explicit activation conditions. Hazards are not inferred simply because certain objects are present, but only when the required conditions defined in the taxonomy are satisfied. For example, a fall hazard is reported only when specific contributing factors such as unprotected edges, worker proximity, and missing protective measures are visually and contextually supported. These constraints limit the range of possible hazards and prevent the model from reporting unsupported risks.

Finally, the framework reduces hallucination by limiting free-form generation. Conventional vision–language models often generate fluent but unconstrained textual descriptions, which can introduce imagined hazards. In contrast, the proposed approach guides generation through structured hazard categories, retrieved evidence, and staged reasoning steps. By restricting what the model is allowed to claim and requiring justification for each hazard, the system prioritizes reliability over linguistic creativity. Together, cognitive grounding, constraint-based reasoning, and reduced free-form generation fundamentally alter the model's reasoning behaviour rather than guessing based on learned correlations, the system must justify hazards through visual precedent and condition-based logic.

6.1 Practical implications

These results can be placed alongside recent VLM-based construction safety systems, with the caveat that direct numerical comparison across studies is unreliable: hazard definitions, category granularity, dataset composition and hallucination criteria differ substantially, and no shared benchmark yet exists for this task. The comparison is therefore made at the level of mechanism. Systems that improve reliability through prompt design act on how the model expresses its judgement but leave the evidential basis of that judgement unchanged. Systems that retrieve regulatory or ontological knowledge strengthen the compliance step while still interpreting the query image without visual precedent. The results reported here indicate that the perceptual step is where the largest reliability gain is available: image retrieval alone raises F1 from 0.535 to 0.923 before any structured constraint is applied, and the taxonomy contributes a further gain in precision on top of it. The absence of a common benchmark is itself a limitation of this research area, and the explicit reporting of the taxonomy, configuration-level prompt templates,

and evaluation protocol in this paper is intended as a step toward one. Beyond detection accuracy, the framework offers potential value across three operational domains. For image-based site monitoring, it can process photographs from mobile devices, surveillance systems, and inspections to flag hazards with condition-based explanations. For compliance support, the regulation-aware retrieval module links each confirmed hazard to candidate safety provisions, which may streamline documentation while retaining the need for professional verification. For safety training, retrieval of visually similar hazard scenarios and structured explanatory reports can provide a case-based learning resource for hazard-recognition exercises.

6.2 Failure cases and residual hallucination

Despite the substantial reduction in unsupported hazard claims achieved by the proposed framework, a residual end-to-end hallucination rate of 6.1% remains. This is a claim-level rate computed over reported hazards as defined in Section 4.4; the manuscript does not convert it into an image-level prevalence without the corresponding per-image prediction counts. Inspection of residual errors reveals three recurring failure modes: limited visual precedent for rare retrieval categories, context bleeding when only part of the condition structure is visually supported, and degraded image quality. Two of these mechanisms can be quantified from the reported evaluation: at the selected k = 5 operating point, 33% of residual false positives were attributed to metadata influence/context bleeding, while retrieval failure occurred for 43 of 421 ground-truth hazard instances (10.2%). These percentages use different denominators and are not additive; no exhaustive percentage partition for image-quality failures is reported in the available evaluation, so no additional percentage is assigned here.

First, errors are concentrated in categories with sparse retrieval support, particularly fire/explosion (0.96% of retrieval instances) and electrocution (0.78% of retrieval instances). The electrocution denominator difference is substantial: electrocution accounts for 13 of 1,674 retrieval instances (0.78%) but 49 of 421 held-out test hazard instances (11.6%). In these cases, RAG-1 may retrieve adjacent hazard categories rather than true visual precedents, weakening the grounding available to condition-constrained reasoning. This is interpreted as a retrieval-support limitation rather than as a stable category-level performance estimate and motivates targeted expansion of the knowledge base.

Second, some residual false positives occur when only part of the three-class condition structure is clearly visible and retrieved metadata influences interpretation of an unobserved factor, indicating that structured constraints are not fully immune to context bleeding. Third, motion blur, partial occlusion, low illumination, and low contrast recur among residual errors, suggesting that preprocessing, image-quality screening, or confidence-based referral to human review could reduce this failure mode. These failure modes also motivate the image-quality screening and human-review recommendation in Section 5.4.

Retrieval failure is the principal quantified driver of these residual errors. A retrieval failure is defined as a query for which none of the k retrieved cases shares the ground-truth Level 1 hazard label of the input image. Stratifying performance by retrieval outcome exposes the mechanism directly: on instances for which retrieval succeeded, the pipeline achieves an F1-score of 0.937 and a hallucination rate of 3.8%, whereas for instances in which retrieval failed the F1-score decreases to 0.744 and the hallucination rate increases to 26.3%. Residual hallucination is therefore governed principally by knowledge-base coverage rather than by the reasoning stage, identifying targeted dataset expansion, rather than additional architectural complexity, as the highest-value next step.

7. LIMITATIONS

Despite the promising results reported in Section 5, the proposed framework has several limitations that should be acknowledged and addressed in future work.

First, the retrieval knowledge base, while covering ten hazard categories, remains limited in scale and environmental diversity. Construction sites vary significantly in layout, equipment types, worker activities, and lighting conditions. Expanding the dataset with annotated images from a broader range of project types and geographic contexts would improve generalization and strengthen retrieval coverage for underrepresented categories such as fire/explosion and electrocution, which were identified in Section 6.2 as primary sources of residual false positives.

Second, the framework currently operates on static images, which limits its ability to detect hazards that are fundamentally temporal in nature. Dynamic behaviours such as unsafe equipment operation sequences, worker

fatigue indicators, or evolving site conditions cannot be fully characterized from a single frame. Future work should incorporate video-based analysis, wearable sensor data, or building information modelling inputs to extend the system's coverage to these behavioural and time-dependent hazard types.

Third, although the annotation process now follows a documented guideline with double annotation and reported agreement (Section 4.2), it still relies on human judgement, which introduces residual inconsistency for subtle or context-dependent hazards. Agreement was measured on a subsample rather than on the whole corpus, and the annotator pool remains small. Larger annotator pools, adjudication of the full corpus, and cross-institutional annotation would strengthen the reliability of both the hazard labels and the hallucination judgements.

Fourth, the practitioner evaluation reported in Section 5.2 involved two practitioners reviewing a sample of outputs. It is sufficient to establish that the structured reports are interpretable and considered useful, but not to support quantitative claims about adoption, inspection speed, or decision quality. A controlled user study with a larger and more diverse practitioner sample, conducted on live inspection tasks, remains necessary future work.

Finally, while the AND-logic taxonomy proposed in this study is grounded in established accident causation theory and demonstrates strong empirical performance, it represents a rule-based approximation rather than a formal causal model. The framework verifies predefined conditions rather than discovering causal structures from data. A natural and important extension is the incorporation of formal causal modelling including causal graphs, intervention analysis, and counterfactual reasoning trained on large-scale incident and near-miss databases. This would allow the system to learn hazard activation conditions rather than only verify them, moving toward a deeper and more generalizable understanding of construction accident causation.

8. CONCLUSION

This study shows that image-to-image visual grounding is a major contributor to hallucination reduction in VLM-based construction-safety analysis. The proposed multimodal framework grounds hazard inference in visually similar real-world precedents and a structured three-level hazard taxonomy before initiating regulatory retrieval. At the end-to-end report level, the complete pipeline achieves precision 0.939, recall 0.896, F1-score 0.917, and a 6.1% hallucination rate, representing an 88.9% relative reduction from the VLM-only baseline. At the hazard-verification gate, RAG-1 plus taxonomy provides the strongest detection performance (precision 0.987, F1 0.941, hallucination 1.0%), while RAG-1 alone reaches F1 0.923, confirming that visual precedent retrieval provides the largest single performance gain. The study makes four replicable contributions: runtime image-to-image retrieval of annotated visual precedent; an explicit three-level condition-based hazard representation; a grounding-before-compliance architecture that prevents unconfirmed hazards from triggering regulatory retrieval; and component-level validation across ablations, multiple VLM backbones, sensitivity tests, robustness analysis, annotation reliability, and deployment characteristics. Regulatory retrieval remains imperfect (precision@3 = 91.4%), so retrieved provisions should be professionally verified before definitive compliance use. Future work should expand underrepresented hazard categories and site diversity, incorporate temporal/video evidence, and validate the framework across additional regulatory jurisdictions and larger practitioner samples.

REFERENCES

Adil, M., Lee, G., Gonzalez, V. A., & Mei, Q. (2025). Using vision language models for safety hazard identification in construction. arXiv. https://arxiv.org/abs/2504.09083

Adil, M., Ahmed, M., Aqib, M., Gonzalez, V. A., Lee, G., & Mei, Q. (2026). Integration of object detection and small VLMs for construction safety hazard identification. arXiv. https://doi.org/10.48550/arXiv.2604.05210

Alansari, A., & Luqman, H. (2025). Large language models hallucination: A comprehensive survey. arXiv. https://arxiv.org/abs/2510.06265

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv. https://arxiv.org/abs/1606.06565

Ceylan, B. O., Akyuz, E., & Arslan, O. (2021). Systems-Theoretic Accident Model and Processes (STAMP) approach to analyse socio-technical systems of ship allision in narrow waters. Ocean Engineering, 239, 109804. https://doi.org/10.1016/j.oceaneng.2021.109804

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

Chan, C. F., Wong, P. K. Y., Guo, X., Liang, Z., Wu, Y., & Cheng, J. C. P. (2026). Enhancing vision-language model for construction safety inspection via visually grounded reasoning and reinforcement learning. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.6294267

DEKRA. (2024). Safety technology 2024: Examining trends in technology solutions used to reduce serious injuries and fatalities in the workplace. https://www.dekra.com

Duan, R., Deng, H., Tian, M., Deng, Y., & 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., Ding, L., Zhong, B., Love, P. E. D., & Luo, H. (2018). Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. Advanced Engineering Informatics, 37, 139–149. https://doi.org/10.1016/j.aei.2018.05.003

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

Hallowell, M. R., & Gambatese, J. A. (2009). Construction safety risk mitigation. Journal of Construction Engineering and Management, 135(12), 1316–1323. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000107

International Labour Organization. (2016, December 13). Occupational safety and health awareness raising material in the construction sector. https://www.ilo.org/resource/occupational-safety-and-health-awareness-raising-material-construction

Larouzée, J., & Guarnieri, F. (2015). From theory to practice: Itinerary of Reason's Swiss Cheese Model. In Safety and reliability of complex engineered systems (pp. 817–824). CRC Press. https://doi.org/10.1201/b19094-110

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2024). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13872–13882). https://doi.org/10.1109/CVPR52733.2024.01316

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://arxiv.org/abs/2005.11401

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning (Vol. 162, pp. 12888–12900). PMLR. https://proceedings.mlr.press/v162/li22n.html

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 292–305). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.20

Li, Y., Xu, F., Zhang, Z., Mei, X., & Huang, H. (2026a). Construction site fall hazard identification and automated captioning using adapted vision-language models. Automation in Construction, 183, 106790. https://doi.org/10.1016/j.autcon.2026.106790

Li, Y., Xu, F., Mei, X., Zhang, Z., & Jin, Q. (2026b). Vision-grounded safety hazard identification and caption generation for equipment installation construction. Advanced Engineering Informatics, 76, 105008. https://doi.org/10.1016/j.aei.2026.105008

Nakharu, S., & Kumar, P. (2025). Hallucination detection and mitigation in large language models: A comprehensive review. Journal of Information Systems Engineering and Management. https://www.jisem-journal.com

Nath, N. D., & Behzadan, A. H. (2020). Deep convolutional networks for construction object detection under different visual conditions. Frontiers in Built Environment, 6, 97. https://doi.org/10.3389/fbuil.2020.00097

Occupational Safety and Health Administration. (2024). Construction safety photographs and inspection records. U.S. Department of Labor. https://www.osha.gov/

Phinias, R. N. (2023). Benefits and challenges relating to the implementation of health and safety leading indicators in the construction industry: A systematic review. Safety Science, 163, 106131. https://doi.org/10.1016/j.ssci.2023.106131

Qu, X., Chen, Q., Wei, W., Sun, J., & Dong, J. (2024). Alleviating hallucination in large vision-language models with active retrieval augmentation. arXiv. https://arxiv.org/abs/2408.00555

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 8748–8763). PMLR. https://proceedings.mlr.press/v139/radford21a.html

Raman, R., & Mitra, A. (2023). IoT-enhanced workplace safety for real-time monitoring and hazard detection for occupational health. In Proceedings of the International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI 2023). IEEE. https://doi.org/10.1109/ICAIIHI57871.2023.10489803

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4035–4045). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1437

Seo, J., Han, S., Lee, S., & Kim, H. (2015). Computer vision techniques for construction safety and health monitoring. Advanced Engineering Informatics, 29(2), 239–251. https://doi.org/10.1016/j.aei.2015.02.001

Tran, S. V.-T., Yang, J., Hussain, R., Khan, N., Kimito, E. C., Pedro, A., Sotani, M., Lee, U.-K., & Park, C. (2024). Leveraging large language models for enhanced construction safety regulation extraction. Journal of Information Technology in Construction, 29, 1026–1038. https://doi.org/10.36680/j.itcon.2024.045

U.S. Bureau of Labor Statistics. (2024). National census of fatal occupational injuries in 2023 (Report USDL-24-2522). U.S. Department of Labor. https://www.bls.gov/news.release/cfoi.htm

Wang, C., Asadi Shamsabadi, E., Chen, Z., Shen, L., Ahmadian Fard Fini, A., & Dias-da-Costa, D. (2025). Automating construction safety inspections using a multi-modal vision-language RAG framework. arXiv. https://doi.org/10.48550/arXiv.2510.04145

Wang, H., Guo, F., Wang, Q., & Liu, J. (2026). Tandem CCV-VLM: Visual construction safety inspection based on collaborative cross-verification VLM. Advanced Engineering Informatics, 75, 104841. https://doi.org/10.1016/j.aei.2026.104841

Xu, J., Cheung, C., Manu, P., Ejohwomu, O., & Too, J. (2023). Implementing safety leading indicators in construction: Toward a proactive approach to safety management. Safety Science, 157, 105929. https://doi.org/10.1016/j.ssci.2022.105929

Zhang, D., Lei, J., Li, J., Wang, X., Liu, Y., Yang, Z., Li, J., Wang, W., Yang, S., Wu, J., Ye, P., Ouyang, W., & Zhou, D. (2025). Critic-V: VLM critics help catch VLM errors in multimodal reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9050–9061). https://doi.org/10.1109/CVPR52734.2025.00846

Zhang, R., Zhang, H., & Zheng, Z. (2024). VL-Uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv. https://arxiv.org/abs/2411.11919

Zhu, Y., Tao, L., Dong, M., & Xu, C. (2025). Mitigating object hallucinations in large vision-language models via attention calibration. arXiv. https://doi.org/10.48550/arXiv.2502.01969