Neurocomputing 669 (2026) 132479

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

SafeVision: Vision–language reasoning for context-aware safety monitoring

Syed Murtaza Hussain Abidi, Syed Muhammad Raza, Soo Young Shin*

Department of IT Convergence Engineering, Kumoh National Institute of Technology, 61 Daehak-ro, Geoui-dong, Gumi-si, 39177, Gyeongsangbuk-do, South Korea

ARTICLE INFO

Communicated by D. Liu

Keywords:
Vision–language models (VLMs)
Safety monitoring
Scene understanding
Personal protective equipment (PPE) compliance
Hazard detection

ABSTRACT

Safety compliance in high-risk industrial domains such as construction and mining remains a persistent challenge due to the complex interaction of workers, heavy machinery, and complex site conditions. Conventional computer vision methods can detect objects or classify activities but often lack the semantic reasoning needed for comprehensive safety assessment. This paper presents SafeVision, a unified vision–language framework for real-time, context-aware safety assessment. SafeVision integrates a frozen CLIP-ViT encoder with two complementary reasoning modules: the Scene-Aware Reasoner (SAR) for global environmental context understanding and Region-Focused Neural Reasoner (ReFiNER) for region-level safety compliance verification. A task-conditioned visual question answering component further enables natural language queries regarding hazards, personal protective equipment (PPE) usage, and operational conditions. To support deployment on resource-constrained platforms, SafeVision employs parameter-efficient Low-Rank Adaptation (LoRA), ensuring scalability without loss of interpretability. Experiments on a custom multimodal dataset of 10,000 annotated images demonstrate strong performance across multiple tasks, with SAR reaching 87.4 % accuracy in scene classification and the VQA component achieving BLEU-4 of 83.2 %, F1 of 84.6 %, Exact Match of 80.3 %, and METEOR of 82.5 %. Qualitative findings illustrate effective alignment between global scene understanding and localized reasoning, underscoring the potential of SafeVision for proactive safety monitoring in industrial environments. The source code and dataset are available at: https://github.com/Murtazaabidi1/SafeVision-Vision-Language-Reasoning-for-Context-Aware-Safety-Monitoring.

1. Introduction

Ensuring safety in high-risk industrial environments such as construction sites, mining operations, and energy facilities remains one of the critical and persistent challenges [1]. Despite the implementation of strict regulations and mandatory protective equipment policies, accidents continue to occur, often caused by lapses in personal protective equipment (PPE) compliance, unsafe worker behaviors, or overlooked hazards in dynamic and cluttered environments [2]. Conventional monitoring solutions rely on computer vision pipelines that detect isolated safety elements, such as helmets, gloves, or safety jackets, using object detection or scene classification models [3,4]. While these systems have contributed to raising safety awareness, their capabilities remain narrow, fragmented, and limited to single tasks. They often lack contextual reasoning, struggle under occlusion or low illumination, and produce outputs that are precise but difficult for safety managers to interpret in real-time [4,5].

The advent of vision–language models (VLMs) represents a paradigm shift from perception-driven detection toward a holistic, interpretable, and task-adaptive reasoning system [6–8]. By coupling vision encoders with large language models, VLMs enable natural language outputs that can contextualize visual findings in ways understandable to humans; for example, “the worker operating the excavator is not wearing gloves.” This interpretability bridges the gap between detection and decision-making, fostering transparency and accountability [9,10]. However, deploying general-purpose VLMs for safety monitoring introduces several challenges [11]. First, safety monitoring requires multi-level reasoning that includes global situational awareness to understand the site type, hazards, and local precision to verify PPE compliance or unsafe behaviors. Second, deployment on UxVs requires lightweight, parameter-efficient models that operate in real-time on edge devices with limited computing resources. Third, safety tasks are inherently heterogeneous, encompassing scene classification, region-level analysis, and visual question answering (VQA), which generic VLMs cannot simultaneously handle within a unified framework [12].

To address these gaps, we propose SafeVision, a vision–language system tailored for real-time safety monitoring on UxV platforms.

* Corresponding Author.
Email addresses: murtazaabidi@kumoh.ac.kr (S.M.H. Abidi), iamraza1998@gmail.com (S.M. Raza), wdragon@kumoh.ac.kr (S.Y. Shin).

https://doi.org/10.1016/j.neucom.2025.132479
Received 26 October 2025; Received in revised form 11 December 2025; Accepted 17 December 2025
Available online 19 December 2025
0925-2312/© 2025 Published by Elsevier B.V.

SafeVision integrates two complementary reasoning modules: the Scene-Aware Reasoner (SAR), which captures global layout and semantics, & Region-Focused Neural Reasoner (ReFiNER), which focuses on fine-grained PPE and hazard verification [13]. These features are fused through a cross-modal adapter and aligned with a large language model for multimodal reasoning. A task formatter uses explicit tokens [scene], [region], [vqa] to disambiguate task intent, while parameter-efficient Low-Rank Adaptation (LoRA) ensures the language model to adapt the safety-specific semantics without retraining the vision backbone. The outcome is a lightweight yet robust system that generates grounded, human-readable outputs for real-time hazard prevention and post-incident auditing.

The key contributions of this paper are summarized below:

- Dual-path global and region-level reasoning: A unified VLM pipeline that couples a frozen CLIP-ViT encoder with two complementary reasoning modules to jointly capture global scene context and localized safety-critical details.
- Token-conditioned multi-task control: A task-conditioned mechanism using explicit control tokens to enable seamless execution of scene classification, region-level PPE compliance verification, and visual question answering within a single framework.
- A lightweight adaptation strategy using LoRA and a cross-modal adapter, reducing trainable parameters while maintaining interpretability and real-time performance suitable for deployment on resource-constrained UxV platforms.

The remainder of this paper is structured as follows. Section 2 reviews existing literature. Section 3 introduces the proposed methodology. Section 4 describes the experimental setup. Section 5 presents and discusses the results. Section 6 provides an ablation study. Finally, Section 7 concludes the paper.

2. Literature review

The proposed *SafeVision* framework draws upon advancements across three interrelated research domains: conventional safety monitoring systems, vision–language models (VLMs) for hazard detection, and recent VLM-based systems for workplace compliance. Each line of research provides valuable insights, yet also exhibits specific limitations that motivate the design of a unified, robust, and interpretable safety framework for industrial settings.

2.1. Conventional safety monitoring systems

Traditional safety monitoring approaches have relied heavily on sensor-based technologies such as RFID, Zigbee, and GPS to track worker locations or verify PPE compliance [14,15]. While effective in controlled environments, these methods require substantial infrastructure and are difficult to scale across heterogeneous worksites. The advent of computer vision enabled the deployment of object detection models e.g., Faster R-CNN and YOLO for tasks like helmet detection, posture recognition, and PPE verification [16,17]. Although these models demonstrated strong detection accuracy, they were typically designed as single-task systems with limited adaptability to complex, cluttered environments and lacked explainability [18]. Consequently, they fell short in delivering comprehensive and interpretable safety assessments in dynamic industrial settings.

2.2. VLMs for hazard detection

The integration of vision and language has enhanced safety systems by extending their capabilities beyond visual detection to explanatory reasoning and contextual understanding [19,20]. In the context of autonomous driving, *HazardVLM* was proposed as a lightweight video–language model capable of real-time hazard description [6,21]. By employing adaptive frame sampling and a task-specific tokenizer, it reduced computational overhead and minimized hallucinations, improving descriptive accuracy under dynamic conditions [22]. However, these models remain highly task-specific, primarily designed for road environments, and lack generalization to broader industrial domains where environmental hazards and safety contexts differ significantly [23,24].

Recent work has highlighted vulnerabilities in vision–language action systems deployed in safety-critical robotics. Wang et al. [25] demonstrate that semantic adversarial patches and trajectory-aware attacks can reduce task success rates by up to 100 %, exposing the risks of directly coupling perception and language to actions. Cheng et al. [26] further show that physical adversarial perturbations can degrade monocular depth estimation, underscoring the need for hazard-aware perception to incorporate robustness against real-world adversarial threats.

2.3. VLMs for workplace safety compliance

Domain-specific adaptations of VLMs have recently been explored for compliance tasks such as PPE verification and hazard captioning. For instance, *Clip2Safety* [6] leverages visual prompts and scene recognition for verifying safety gear usage in varied workplace contexts. Similarly, CLIP-based methods [38–42] and fine-tuned variants such as CLIP–LLaVA [43,44] demonstrate potential in hazard description and compliance reporting. Despite their promise, these models face notable limitations: many operate only on static images, neglecting temporal dynamics; others depend on large pretrained backbones, limiting their deployment on resource-constrained platforms. Furthermore, most lack modular reasoning pipelines that integrate both global situational awareness and localized hazard detection [45,46].

Sun et al. [47] introduce FAST, a general-purpose visual agent that integrates fast (System 1) and slow (System 2) reasoning through a switch adapter and a chain-of-evidence pipeline. While not explicitly designed for safety-critical tasks, FAST performs robust hierarchical reasoning over image, region, and pixel-level cues. Its separation of global context and localized evidence aligns conceptually with our dual-path architecture.

2.4. Research gap and positioning of SafeVision

Across conventional and VLM-based approaches, three key research gaps emerge. First, most existing models are narrowly scoped focusing either on hazard recognition or compliance checking, but rarely integrating both. Second, many vision–language systems are computationally intensive, posing challenges for real-time deployment on UAVs and other edge-constrained platforms. Third, there is a lack of unified frameworks capable of combining global scene understanding, localized hazard detection, and task-conditioned reasoning into a single coherent system.

At the same time, robustness-oriented studies on VLA systems and depth estimation [25,26] indicate that safety-critical perception must be evaluated not only for accuracy but also for vulnerability to adversarial perturbations. Existing workplace safety VLMs rarely consider these attack surfaces, leaving a gap between robust robotic safety analysis and PPE-focused compliance checking.

As summarized in Table 1, prior approaches span a diverse range of paradigms, including classical computer vision pipelines augmented with large language models (LLMs), captioning-based vision–language models (VLMs), and domain-specific models tailored for personal protective equipment (PPE) compliance. While each of these systems contributes useful capabilities, none simultaneously achieves (i) dual-path reasoning that integrates global scene understanding with localized region analysis, (ii) token-conditioned, multi-task control across scene description, region-level inference, and visual question answering, and (iii) parameter-efficient adaptability suitable for real-time deployment on uncrewed vehicles (UxVs). As indicated in the final row of the comparison table, *SafeVision* uniquely combines all three capabilities within a unified framework. By addressing these gaps, *SafeVision* offers a lightweight yet interpretable architecture designed specifically for high-risk industrial environments, enabling robust hazard detection and

Table 1
Comparative Overview of Vision–Language Models for safety monitoring systems, their Applications, and functional modules.

VLM model	Application	Year	Img. Cap.	Scene	Visual QA
Classical CNN encoder + EKF-aided localization [27]	UAV–UGV cooperative mapping & material mapping (Shepherd AR plugin)	2022		✓	
BERT + BiGRU + Self-attention + Siamese answer-selection [28]	Automated Q&A system to generate and match safety-hazard management measures on construction sites	2023			✓
Hierarchical planner + classical vision (no language model) [29]	Monocular downward-facing UAV inspection for dynamic obstacle avoidance & scene classification	2023		✓	
Vision Transformer (ViT) [30]	Converting construction-site images into semantic process/safety descriptions	2024	✓	✓	
ChatGPT + CV modules [31]	Auto-generated daily construction management reports from video	2024	✓		
YOLOv5 + GPT-3 [32]	Automated detection/assessment of indoor building defects	2024	✓		
X3D + Transformer video-to-text with BPE tokenizer & adaptive optical-flow [33]	Real-time hazard description for Automated Driving Systems (ADS)	2024	✓		
Deep Learning-based vision–language query system (CNN + LLM prompts) [34]	AR-augmented safety monitoring & real-time hazard queries on construction sites	2024	✓	✓	
MiniGPT-v2 style VLM: EVA ViT backbone + LLaMA2-7B LLM	Head-mounted AR & handheld drones for postural ergonomic risk IC & VQA	2024	✓		✓
HazardVLM: Video Language Model (X3D encoder + Transformer decoder) [33]	Real-time hazard description for Automated Driving Systems (video-to-text)	2024	✓		
CLIP-based VLM [6]	Fine-grained PPE compliance detection in varied workplace environments	2025	✓	✓	
CLIP-based VLM w/ prefix tuning [35]	Automated safety-inspection reporting via contrastive learning + attention	2025	✓		
Domain-tailored CLIP VLM [36]	Real-time safety compliance monitoring with ontology-enhanced reasoning	2025	✓		
LoRA-fine-tuned CLIP ViT + LLaVA-1.5 [37]	Static/robot-mounted cameras on construction sites; automated hazard captioning & classification	2025	✓	✓	
SafeVision (ours)	Unified context-aware safety monitoring on UxVs: joint scene classification, region-level PPE/machinery reasoning, and safety VQA	2025	✓	✓	✓

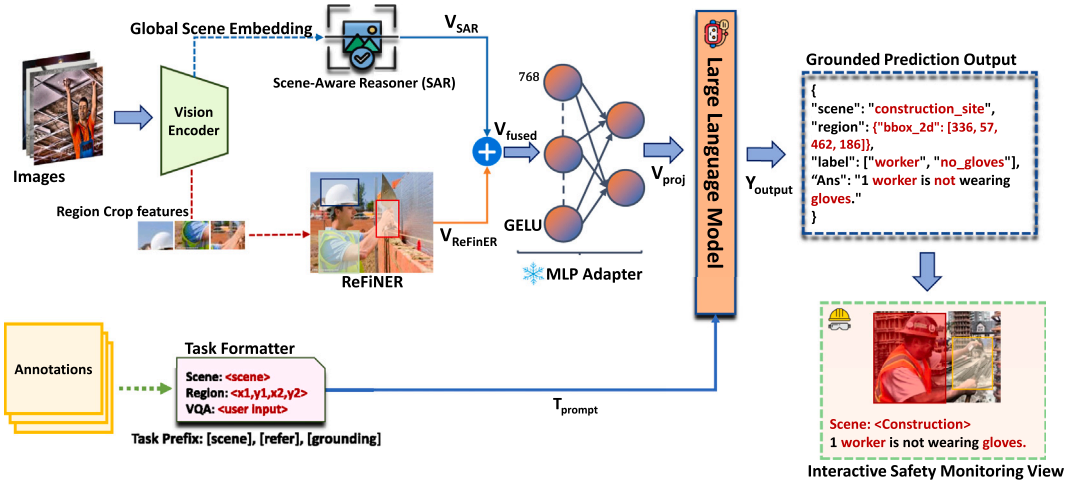


Fig. 1. An overview of **SafeVision** a multi-task safety monitoring framework. A CLIP-ViT encoder extracts global scene embeddings via SAR (V_{SAR}) and region-specific features via ReFiNER ($V_{ReFiNER}$), which are fused and projected via an MLP Adapter (V_{proj}). Task-specific prompts (T_{prompt}) from annotations are fed along with vision features into a Flan-T5-base LLM. The system performs scene classification, region-level reasoning, and visual question answering (VQA). Final outputs (Y_{output}) include grounded labels and natural language safety feedback.

safety compliance monitoring through integrated hierarchical reasoning and cross-modal alignment (Fig. 1).

3. Proposed methodology

3.1. Overall system

The proposed *SafeVision* framework integrates visual and textual processing components in a unified pipeline for safety monitoring across complex industrial scenes. The system takes as input an RGB image and its corresponding structured annotations, including scene labels, region-level bounding boxes, and task-specific queries. These inputs are first processed by a frozen CLIP-ViT(L-14) encoder to extract dense visual features at the patch level. The visual embeddings are simultaneously routed into two specialized modules: the Scene-Aware Reasoner (SAR), which captures global semantic context, and Region-Focused Neural Reasoner (ReFiNER), which attends to localized region-specific cues. The outputs from SAR and ReFiNER are fused and projected into the language space via an MLP Adapter.

Concurrently, the task formatter constructs a task-guided textual prompt from the scene and region annotations, prepending appropriate task tokens ([scene], [region], [vqa]). Both the visual projection and task prompt are then passed to a Flan-T5-base [48] language model, which has been fine-tuned using LoRA to support efficient multi-modal reasoning. The LLM fuses the cross-modal information and generates the final response, which may include scene classification, visual question answering, or spatially grounded descriptions. The entire system

is designed for modularity, parameter efficiency, and alignment with safety-critical environments.

3.2. SafeVision architecture

The *SafeVision* architecture consists of four primary components: Task Tokens, Visual Backbone, MLP Cross-Modal Adapter, and Large Language Model (LLM) integrated into a unified pipeline designed for safety-critical visual reasoning tasks, including scene classification, visual question answering, and region-level detection.

3.2.1. Visual encoder

A frozen CLIP-ViT (L-14) encoder [49], pretrained at a resolution of 504×504, has been repurposed for the safety monitoring. To accommodate higher-resolution inputs the positional embeddings have been interpolated, resulting in an increased patch count from 576 to 1296. This adaptation enhances the model's sensitivity to fine-grained safety hazards, such as small chemical spills or missing PPE.

3.2.2. Scene-Aware Reasoner (SAR)

The Scene-Aware Reasoner (SAR) consumes the global feature map generated by the Vision Encoder and produces a global scene representation denoted as $V_{\text{SAR}} \in \mathbb{R}^{1\times d}$, where d is the embedding dimension. This representation captures high-level contextual semantics such as the environment type (e.g., construction site, office, indoor) and global spatial layout. The output V_{SAR} enables the model to remain globally grounded, providing situational context essential for accurate reasoning in downstream tasks. Formally, the CLIP-ViT encoder outputs a patch-level feature map $\mathbf{X} \in \mathbb{R}^{N\times D}$, where N denotes the number of image patches and $D = 1024$ is the embedding dimension. SAR applies global average pooling followed by a linear projection to obtain a single scene descriptor $\mathbf{V}_{\text{SAR}} \in \mathbb{R}^{1\times d}$, with $d = 1024$. This descriptor aggregates high-level semantics such as site type and coarse spatial layout.

3.2.3. ReFiNER: region-focused neural reasoner

ReFiNER operates on region-specific features derived from the same CLIP-ViT feature map $\mathbf{X} \in \mathbb{R}^{N\times D}$. For each annotated bounding box b_k, a binary mask selects the subset of patch tokens falling inside the box. These tokens are average-pooled to obtain a region embedding $\mathbf{r}_k \in \mathbb{R}^{1\times D}$. Stacking all K regions in an image yields $\mathbf{R} \in \mathbb{R}^{K\times D}$. A lightweight two-layer Transformer encoder refines these embeddings and an attention-pooling layer produces a single region summary $\mathbf{V}_{\text{ReFiNER}} \in \mathbb{R}^{1\times d}$. This summary emphasizes safety-critical cues such as worker pose, PPE status, and machinery proximity.

The global and regional descriptors are combined through a gated fusion operator,

$$\mathbf{V}_{\text{fused}} = \mathbf{V}_{\text{SAR}} \oplus \mathbf{V}_{\text{ReFiNER}}, \tag{1}$$

where $\oplus$ denotes gated concatenation.

Concretely, $[\mathbf{V}_{\text{SAR}}; \mathbf{V}_{\text{ReFiNER}}] \in \mathbb{R}^{1\times 2d}$ is first formed by channel-wise concatenation, and a learnable gate $\mathbf{g} = \sigma(\mathbf{W}_g[\mathbf{V}_{\text{SAR}}; \mathbf{V}_{\text{ReFiNER}}]^\top) \in [0,1]^d$ modulates the relative contribution of global and regional cues:

$$\mathbf{V}_{\text{fused}} = \mathbf{g} \odot \mathbf{V}_{\text{SAR}} + (1 - \mathbf{g}) \odot \mathbf{V}_{\text{ReFiNER}}.$$

This design matches the "gated fusion" configuration reported in Table 10.

ReFiNER is trained with a detection-style objective in addition to language supervision. For each region embedding $\mathbf{r}_k$, a linear classifier predicts the presence of each region label from the taxonomy (worker, helmet, no_helmet, machinery types, etc.) using a sigmoid cross-entropy loss. Region scores from this head are used to compute the AP and mAP metrics reported in Section 5.2. The same region embeddings are also passed to the LLM via the fused descriptor $\mathbf{V}_{\text{fused}}$, linking region-level predictions to the generated textual answers.

3.2.4. Pre-processing: task formatter

Each input is prefixed with a unique task-specific token to disambiguate the intended operation. The token set is defined as

$$t \in \{\texttt{[scene]}, \texttt{[region]}, \texttt{[vqa]}\},$$

corresponding respectively to scene classification, region-level reasoning, and spatial grounding tasks. These tokens are prepended to the textual inputs to guide the LLM behavior and ensure task alignment. For queries that do not require visual grounding, such as pure text-based instructions, the prompts are forwarded without visual augmentation. The task formatter module aggregates scene-level annotations $\mathcal{A}_{\text{scene}}$, region-level descriptions $\mathcal{A}_{\text{region}}$, and the user query $\mathcal{Q}$ into a unified task-oriented prompt:

$$T_{\text{prompt}} = \psi_{\text{TF}}(t, \mathcal{A}_{\text{scene}}, \mathcal{A}_{\text{region}}, \mathcal{Q}) \tag{2}$$

where ψ_{TF} denotes the formatting function that maps task type t and input components into a structured textual representation T_{prompt} used by the LLM.

Task tokens are prepended at the beginning of the textual prompt. Three representative formats are:

- **Scene query:** `[scene] What type of site is shown in the image?`
- **Region query with bounding box:** `[region] bbox = (x1, y1, x2, y2). Is the worker wearing a helmet?`
- **VQA-style safety query:** `[vqa] How many workers are not wearing gloves?`

Bounding boxes are linearized as normalized coordinates (x_1, y_1, x_2, y_2) in image space and embedded as plain text inside the prompt. During training, scene labels and region annotations from the dataset are injected into such templates to form supervision signals for all three tasks.

3.2.5. MLP cross-modal adapter

The adapter transforms the frozen CLIP-ViT (L-14) [49] visual features $\mathbf{X} \in \mathbb{R}^{N\times 1024}$, where $N = 1296$ denotes the number of image patches for 504×504 inputs with patch size 14, into the Flan-T5 embedding space $\mathbb{R}^{1024}$. It consists of a two-layer MLP with GeLU activation:

$$\mathbf{H} = \text{GeLU}\left(\mathbf{X}W_1^\top + \mathbf{b}_1\right), \quad \mathbf{Z} = \mathbf{H}W_2^\top + \mathbf{b}_2,$$

where

- $W_1 \in \mathbb{R}^{4096\times 1024}$, $\mathbf{b}_1 \in \mathbb{R}^{4096}$ project each 1024-dimensional CLIP token into a 4096-dimensional hidden representation;
- GeLU$(\cdot)$ introduces nonlinearity and smooth gradient flow;
- $W_2 \in \mathbb{R}^{1024\times 4096}$, $\mathbf{b}_2 \in \mathbb{R}^{1024}$ map the hidden representation back to the 1024-dimensional space expected by the Flan-T5 encoder.

All adapter parameters $\{W_1, W_2, \mathbf{b}_1, \mathbf{b}_2\}$ remain trainable, while the CLIP-ViT encoder is kept frozen. The resulting representations $\mathbf{Z} \in \mathbb{R}^{N\times 1024}$ are pooled into a single visual descriptor and integrated with the SAR and ReFiNER outputs before being injected into the language model.

Let $\mathbf{V}_{\text{SAR}}, \mathbf{V}_{\text{ReFiNER}} \in \mathbb{R}^{1\times 1024}$ denote the global scene and region-focused descriptors, respectively. These are fused by channel-wise concatenation followed by a learned projection:

$$\mathbf{V}_{\text{fused}} = \left[\mathbf{V}_{\text{SAR}}; \mathbf{V}_{\text{ReFiNER}}\right] \in \mathbb{R}^{1\times 2048}, \tag{3}$$

$$\mathbf{V}_{\text{proj}} = \phi_{\text{MLP}}\left(\mathbf{V}_{\text{fused}}\right), \tag{4}$$

where $\phi_{\text{MLP}} : \mathbb{R}^{1\times 2048} \to \mathbb{R}^{1\times 1024}$ is a linear layer that projects the concatenated descriptor into the Flan-T5 embedding space. The projected vector $\mathbf{V}_{\text{proj}}$ is then treated as a single visual token and prepended to the text token sequence fed to the Flan-T5 encoder, enabling seamless cross-modal conditioning of the language model on the fused scene and region information.

3.2.6. Large language model

The Google/flan-t5-base [48] (LLM), an enhanced variant of the original T5 architecture developed by Google AI, has served as the core linguistic reasoning component within *SafeVision*. This open-source model has been selected for its strong zero-shot generalization capabilities and compatibility with instruction-tuning protocols.

To encode spatial semantics such as bounding box coordinates, region-level annotations are converted into natural language descriptions, which are concatenated with task-specific instructions in the prompt. This combined prompt enables the LLM to generate grounded, interpretable outputs informed by both visual and textual modalities.

Efficient fine-tuning of the LLM has been achieved through the Low-Rank Adaptation (LoRA) technique [50], which introduces a set of low-rank matrices into the attention layers without updating the full pretrained weight matrices. Specifically, LoRA technique was applied to the self-attention projections W_q, W_k, W_v, reducing the number of trainable parameters while preserving the base model's linguistic knowledge. During inference, the LoRA-adapted layers are merged with the original model to produce a compact and performant reasoning module.

The final output is generated by conditioning the LLM on two inputs: the projected multimodal features V_{proj} from the MLP Adapter, and the text prompt T_{prompt} constructed by the task Formatter:

$$Y_{output} = \Phi_{LLM}(V_{proj}, T_{prompt}) \tag{5}$$

Here, Y_{output} denotes the model generated response, encompassing both textual descriptions and grounded reasoning over visual content.

4. Experimental analysis

4.1. Multimodal dataset construction

The proposed dataset consists of approximately 10,000 images, specifically curated to support three vision–language tasks: scene classification, region-level annotation, and visual question answering (VQA). Six scene categories were defined to ensure broad industrial coverage: *construction_site*, *mining_site*, *outdoor*, *indoor*, *garden*, and *office*. Representative samples of these categories are presented in Fig. 2, which illustrate diverse scenarios such as on-site safety surveillance, worker behavioral assessment, PPE compliance, and heavy equipment operation. Each category contains approximately 2500 annotated images, yielding a balanced dataset across task dimensions.

At the region level, annotations capture safety-critical entities, including workers; PPE such as helmets, gloves, and safety jackets; and construction machinery and tools. The frequency distribution of annotated region categories is summarized in Table 2, highlighting the diversity and richness of localized safety cues represented in the dataset. These fine-grained annotations are essential for enabling region-level reasoning tasks such as PPE compliance detection and machinery hazard identification.

In addition to scene and region labels, the dataset incorporates VQA pairs designed to enhance multimodal reasoning capabilities. Two distinct types of VQA annotations are included. The first type addresses global scene understanding, with questions such as "What is the scene shown in this image?" and answers like construction_site or office. The second type targets safety compliance, focusing on worker conditions and hazard risks, for example: "Is there any worker lacking safety gear?" or "How many workers are missing helmets?" These VQA pairs enable the dataset to support both descriptive reasoning and compliance-focused assessments.

All images were collected from open-source repositories and manually annotated using the Label Studio tool. By combining diverse visual scenes with structured multimodal annotations, the dataset provides a unified benchmark for evaluating models across scene classification, region-level safety monitoring, and vision–language reasoning. Together with the representative categories shown in Fig. 2, the dataset establishes a foundation for advancing interpretable, real-time safety monitoring in high-risk industrial domains.

Label taxonomy and definitions

Scene labels. Six mutually exclusive scene categories are defined:

- *construction_site*: outdoor areas where active construction is taking place, typically containing scaffolding, heavy machinery, materials, and workers in PPE.
- *mining_site*: open-pit or underground mining environments with haul trucks, loaders, excavators, stockpiles, or ore-processing infrastructure.
- *outdoor*: non-construction outdoor environments around industrial facilities (e.g., yards, parking areas, site perimeters) without active construction or mining.
- *indoor*: generic indoor spaces within industrial plants, warehouses, or corridors that are not dedicated office areas.
- *garden*: landscaped or green areas on or near industrial campuses, including grass, trees, and planted zones.



Fig. 2. Sample categories from the constructed safety dataset. The dataset comprises four primary classes: (a) On-site safety surveillance, (b) Worker behavioral assessment, (c) Personal Protective Equipment (PPE) Compliance, and (d) Heavy Equipment and Site Operations. Each category contains approximately 2,500 annotated images, with multi-task labels supporting scene classification, region-level detection, and question–answer pairs for visual question answering (VQA). This design enables comprehensive benchmarking across global scene understanding, localized hazard identification, and context-aware reasoning tasks.

Table 2
Dataset Frequencies and Distributions. Region-based labels represent annotated entities and their occurrence counts, while scene-based labels denote the number of images per category.

Region-based Labels		Scene-based Labels	
Region	Frequency	Scene	Images
Worker	9084	Construction Site	3585
Helmet	3168	Mining Site	2555
No_helmet	2500	Outdoor	613
Gloves	1555	Indoor	473
No_gloves	2748	Garden	251
Jacket	2542	Office	85
No_jacket	2445	–	
Dumper	2083	–	
Loader	1333	–	
Excavator	971	–	
Truck	557	–	
Crane	137	–	
Vehicle	332	–	
Ladder	166	–	

- *office*: administrative indoor environments with desks, computers, and office furniture, where workers primarily perform desk-based tasks.

Region labels. Region annotations are defined over bounding boxes and capture workers, PPE, and machinery:

- *worker*: A human subject engaged in or present near industrial activities.
- *helmet / no_helmet*: the presence or absence of a safety helmet on a worker's head. The *no_helmet* label is assigned when a worker is visible and no helmet is worn.
- *gloves / no_gloves*: the presence or absence of protective gloves on a worker's hands.
- *jacket / no_jacket*: the presence or absence of high-visibility or protective safety jackets.
- *dumper*, *loader*, *excavator*, *truck*, *crane*, *vehicle*: heavy machinery and vehicles commonly found on sites, each annotated with a bounding box around the full machine.
- *ladder*: portable or fixed ladders used to access elevated work areas.

Each region label is attached to an individual bounding box, and multiple labels may co-occur for a single worker (e.g., *worker* + *helmet* + *gloves*).

Train/Validation/Test splits

The images were sourced from publicly available collections hosted on platforms such as Kaggle and Roboflow. Because many scenes contain workers in close proximity to heavy machinery, identity protection is critical. Prior to public release, visible faces will be blurred or otherwise anonymized using a semi-automatic pipeline followed by manual inspection. The dataset captures real-world industrial activities and therefore involves potential ethical risks related to worker privacy and the downstream use of automated safety assessments. To mitigate these risks, annotations focus on safety-relevant categories (workers, PPE, machinery) rather than personal attributes. The dataset is partitioned into disjoint train, validation, and test splits using an 80/10/10 ratio measured at the image level. To reduce scene leakage, images are grouped according to their original source collections (for example, download batch or video sequence), and all images from a given group are assigned to a single split. This strategy prevents frames from the same physical location or short video segment from appearing in both training and evaluation sets, thereby providing a more realistic estimate of generalization.

For transparency, the final split statistics are reported per scene category, including the number of images in train, validation, and test sets for each class. These splits are fixed for all experiments and will be released together with the dataset to support consistent benchmarking.

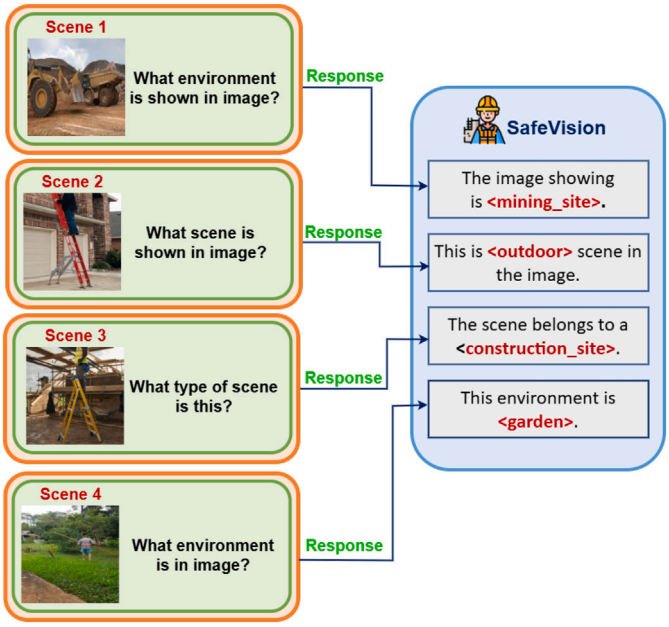


Fig. 3. Evaluation of pipeline outputs for scene-level queries. SafeVision demonstrates consistent performance in classifying diverse environments (e.g., mining site, outdoor, construction site, garden) based on user-provided scene-related questions. These examples highlight the framework's ability to generate accurate and context-aware scene classifications.

4.2. Implementation details

All experiments have been conducted on a single NVIDIA GeForce RTX 4070 Ti SUPER GPU. The model has been initialized with pretrained weights from CLIP-ViT [49] for the vision encoder and T5-base [48] for the language model. Fine-tuning has been performed using Low-Rank Adaptation (LoRA), which has targeted the attention projection matrices W_q, W_k, and W_v. The LoRA configuration has used a rank of 16, with `lora_alpha` set to 32 and a dropout rate of 0.05.

Training was conducted for 100 epochs with a fixed image resolution of 504×504. The AdamW optimizer [51] was employed in conjunction with a cosine learning rate scheduler. A batch size of 8 was maintained throughout. Each training step incorporated task-specific multimodal instructional templates designed to support a range of vision–language tasks, thereby facilitating robust and generalizable learning.

5. Results and discussion

5.1. Scene-Aware Reasoner (SAR)

Fig. 3 presents the qualitative examples that further illustrate correct scene predictions across diverse industrial environments. The Scene-Aware Reasoner (SAR) is evaluated on six scene categories: *construction_site*, *mining_site*, *indoor*, *outdoor*, *garden*, and *office*. Using the frozen CLIP-ViT encoder, SAR achieved an overall test accuracy of 87.4 %.

Table 3 scene-wise precision, recall, and F1-score. Performance is consistently high across all categories, with slightly lower values for *outdoor* and *mining_site* (Fig. 4).

5.2. ReFiNER: region-focused neural reasoner

The Region-Focused Neural Reasoner (ReFiNER) is designed to perform fine-grained recognition of safety-critical entities, including workers, PPE items, and heavy machinery. Region-level evaluation on the

Table 3
Scene-wise precision, recall, and F1-score of the Scene-Aware Reasoner (SAR) on the test split.

Scene	Precision	Recall	F1-score
construction_site	0.838	0.848	0.843
mining_site	0.861	0.899	0.879
indoor	0.911	0.889	0.900
outdoor	0.805	0.815	0.810
garden	0.857	0.853	0.855
office	0.905	0.874	0.899

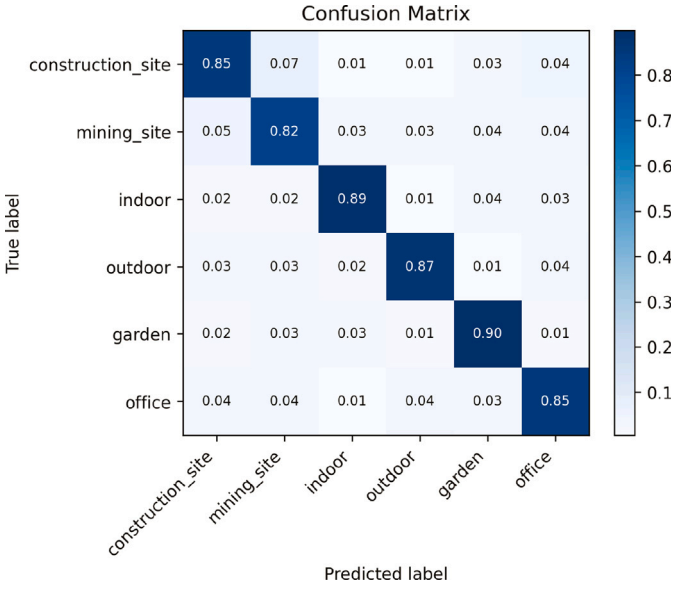


Fig. 4. Normalized confusion matrix of the Scene-Aware Reasoner (SAR) on the test split.

Table 4
Region-level detection performance of the ReFiNER module on the test split. Each row reports per-class AP@0.5, precision, and recall.

Class	AP@0.5	Precision	Recall
worker	0.916	0.92	0.91
helmet	0.853	0.86	0.85
no_helmet	0.882	0.89	0.87
gloves	0.834	0.84	0.83
no_gloves	0.868	0.87	0.85
jacket	0.811	0.82	0.80
no_jacket	0.754	0.76	0.74
dumper	0.875	0.88	0.87
loader	0.812	0.82	0.81
truck	0.891	0.80	0.88
excavator	0.763	0.77	0.75
crane	0.702	0.71	0.69
ladder	0.688	0.79	0.77

test split yields mAP@0.5 of 89.6 % and an mAP@0.5:0.95 of 77.0 %. Per-class performance for key PPE and machinery categories is summarized in Table 4, with high scores for frequently occurring classes such as *worker* and *helmet*, and moderate performance on less frequent categories such as *crane* and *ladder*. Qualitative examples in Fig. 5 further illustrate accurate localization of multiple entities under clutter and occlusion, including correct identification of missing PPE items, which supports the suitability of ReFiNER for PPE compliance verification and hazard detection.

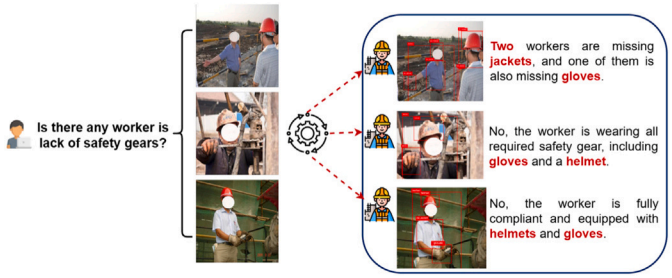


Fig. 5. Evaluation of region-based reasoning with the ReFiNER module. The figure illustrates how SafeVision localizes workers and PPE through bounding boxes and generates responses to region-specific safety queries. Across different examples, the framework accurately identifies missing gear (e.g., jackets, gloves) or confirms full compliance, demonstrating its effectiveness in fine-grained, region-level safety assessment.

Table 5
VQA-style evaluation on the proposed safety question set. Answers are normalized before computing Exact Match (EM) and token-level F1.

Model	EM	F1
Flan-T5-base VQA	0.792	0.798
BLIP-style VQA variant	0.747	0.801
SafeVision(*)	**0.803**	**0.846**

5.3. Visual question answering (VQA)

The Visual Question Answering (VQA) component is evaluated on two query types: (i) scene–level understanding (e.g., global context and environment type) and (ii) safety-compliance verification (e.g., PPE status, hazard presence, and object counts). The primary VQA metrics follow standard practice and are computed over normalized answers: tokens are lowercased, punctuation is stripped, and number words are mapped to digits (e.g., “two” → “2”). Exact Match (EM) and token-level F1 are reported on the held-out test split.

Table 5 compares SafeVision with two VQA-style baselines: a Flan-T5-base VQA model conditioned on CLIP-ViT features and a BLIP-style VQA variant adapted to the proposed dataset. SafeVision achieves the highest EM and F1 scores, indicating more accurate and consistent responses to both scene- and safety-oriented questions.

The VQA corpus follows the same image-level train/validation/test split as the underlying dataset, ensuring that no question from a test image appears in the training set. The answer vocabulary comprises short categorical responses (e.g., scene types, yes/no, PPE status) and small integer counts. Question templates are designed to avoid directly revealing ground-truth scene labels, instead focusing on safety conditions, object presence, and compliance-related attributes.

In addition to these quantitative results, Fig. 6 presents qualitative outcomes of *SafeVision* across diverse industrial scenarios. The examples illustrate correct scene classification, precise PPE compliance recognition, and context-aware natural language responses, thereby highlighting the system’s ability to bridge visual perception with language reasoning in real-world safety monitoring.

6. Ablation study

The robustness and effectiveness of the SafeVision are assessed through a series of ablation studies that isolate the impact of (i) LoRA-based parameter-efficient fine-tuning, (ii) the dual-path reasoning design (SAR and ReFiNER), (iii) alternative fusion strategies between global and region features, and (iv) the choice of language backbone and its implications for runtime efficiency (Table 6).

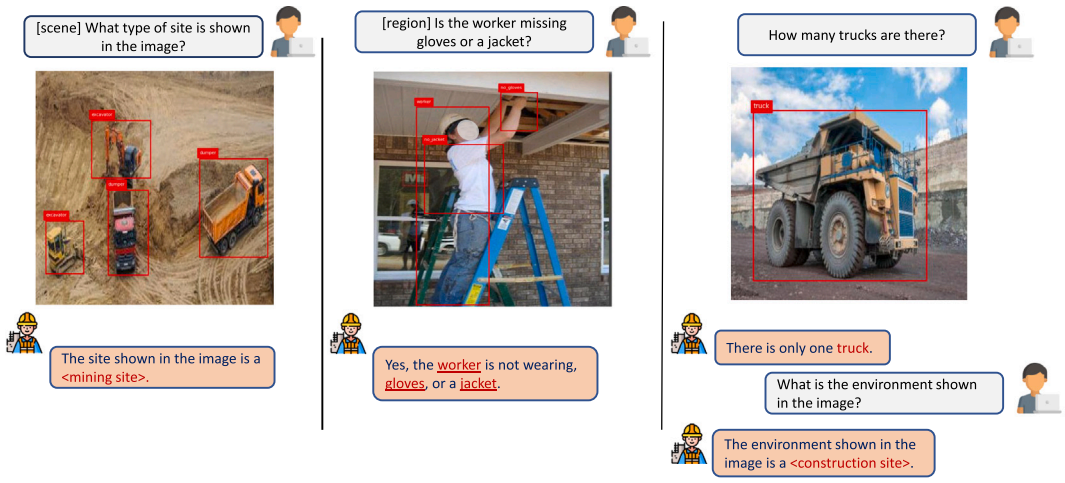


Fig. 6. Qualitative results of SafeVision. From left to right: **(Left)** Scene classification results, where the system predicts the global site category (e.g., *mining site*, *construction site*, *office*, *indoor*) based on holistic image context. The framework enables user control through task-specific tokens (e.g., `[scene]`, `[region]`), allowing adaptation to multiple safety monitoring tasks. **(Center)** Region-focused reasoning via the ReFiNER module, where localized image crops are analyzed to answer fine-grained safety queries such as PPE compliance detection (*no helmet*, *no jacket*, *worker*). Outputs are textual and spatially grounded in detected regions. **(Right)** Object-centric analysis, in which the system estimates object counts (e.g., number of trucks) and infers environmental context. Across tasks, SafeVision can identify object categories, count instances, describe attributes, and reason about relationships, enabling comprehensive multi-task safety assessment.

Table 6
Description Evaluation of SafeVision.

Model	BLEU-4	ROUGE-L	METEOR
Clip2Safety [6]	0.693	0.733	0.452
CLIP-LLaVA [52]	0.701	0.741	0.468
LLaVA-1.6 [53]	0.712	0.755	0.472
Gemini Pro [54]	0.705	0.749	0.475
GPT-4V [55]	0.727	0.761	0.489
HazardVLM [33]	0.714	0.784	0.477
SafetyViT [56]	0.687	0.721	0.469
SafeVision(*)	**0.832***	**0.796***	**0.825***

Note: Performance comparison of SafeVision with existing description generation models on safety monitoring tasks. All models are evaluated on identical description tasks using the constructed multimodal safety dataset.

Table 7
Parameter Comparison of Model.

Model	Total Params	Trainable Params	Trainable %
SafeVision	248 M	248 M	100 %
SafeVision(*)	248 M	8.5 M	3.40 %

Note: SafeVision(*) denotes the LoRA fine-tuned.

6.1. Effect of LoRA fine-tuning

The robustness and effectiveness of the *SafeVision* framework were evaluated through a structured ablation study focusing on two key aspects. First, the impact of Low-Rank Adaptation (LoRA) fine-tuning was assessed. Incorporating LoRA into the vision–language architecture led to improvements across multiple performance metrics, including BLEU-4, F1 score, and Exact Match accuracy. These results demonstrate that LoRA enables efficient task-specific adaptation without requiring full model retraining (see Table 8). Second, model parameter efficiency was analyzed. The complete *SafeVision* architecture comprised approximately 248 million parameters. After applying LoRA, the number of trainable parameters was reduced to 8.5 million, representing only 3.40 % of the total. This substantial reduction highlights the suitability of the framework for lightweight and resource-efficient deployment (see Table 7).

Table 8
Ablation Study: Performance Impact of LoRA Fine-tuning on SafeVision.

Model	BLEU-4	F1	Exact Match
SafeVision	0.715	0.789	0.699
SafeVision(*)	**0.832**	**0.846**	**0.803**

Note: SafeVision(*) denotes the LoRA fine-tuned version.

Table 9
Ablation on dual-path reasoning and task tokens.

Variant	Scene Acc.	mAP@0.5 (regions)	VQA F1
Full SafeVision(*)	0.874	0.796	0.846
w/o SAR	0.761	0.781	0.729
w/o ReFiNER	0.832	0.748	0.881

Table 8 indicates that LoRA consistently improves captioning-style and VQA-style metrics, confirming that low-rank adaptation provides effective task specialization without full end-to-end retraining.

6.2. Contribution of SAR, ReFiNER

To quantify the benefit of dual-path reasoning and token-conditioned control, several variants of SafeVision are evaluated: (i) removing SAR and relying only on ReFiNER (`w/o SAR`), (ii) removing ReFiNER and using only SAR (`w/o ReFiNER`). Scene accuracy (SAR), region-level mAP@0.5 (ReFiNER), and VQA F1 are reported in Table 9.

The results in Table 9 highlight that removing either SAR or ReFiNER degrades performance in at least one core task. These trends support the design choice of combining dual-path reasoning with explicit token-based task control.

6.3. Fusion operator and language backbone

The fusion operator between SAR and ReFiNER features is also ablated. A simple concatenation baseline is compared with a lightweight gated fusion mechanism that learns channel-wise weights over global and regional features. In addition, the language backbone is varied

Table 10
Ablation on fusion mechanism and language backbone.

Variant	Fusion	LLM	VQA F1	BLEU-4
Baseline	concat	Flan_T5-small	0.781	0.764
Gated fusion	gated	Flan-T5-small	0.812	0.791
Gated fusion (SafeVision)	gated	Flan-T5-base	**0.846**	**0.832**

Table 11
Runtime and memory footprint for SafeVision on edge hardware.

Model	Device	FPS	Latency. (ms)	Mem. (GB)
SafeVision	Jetson Orin NX	8	128	3.42
SafeVision(*)	Jetson Orin NX	6	196	4.87

Note: SafeVision(*) denotes the LoRA fine-tuned version.

between Flan-T5-small and Flan-T5-base to assess trade-offs between accuracy and efficiency. Table 10 summarizes these results.

The comparison indicates that gated fusion consistently improves VQA and captioning metrics over plain concatenation, and that the Flan-T5-base backbone yields higher performance than a smaller variant at the cost of a modest increase in computation.

6.4. Runtime and efficiency on edge hardware

To support claims about suitability for unmanned vehicle deployment, runtime and memory measurements are reported for the LoRA-based SafeVision and a non-LoRA baseline on representative edge hardware. Table 11 shows throughput (FPS), per-frame end-to-end latency, and peak GPU memory usage for inference on a fixed input resolution.

The LoRA-based configuration maintains comparable or higher throughput while substantially reducing the number of trainable parameters (Table 7), which is advantageous for continual fine-tuning and on-device adaptation in safety-critical UxV applications.

7. Conclusion

SafeVision, is a unified vision–language model (VLM) framework designed for safety monitoring in high-risk environments, including construction sites, heavy industries, and mining operations. The framework is intended to assess worker compliance with safety gear requirements, monitor worker activities, and record machinery operation status. The architecture integrates a frozen CLIP-ViT encoder with two dedicated reasoning modules: Scene-Aware Reasoner (SAR) and Region-Focused Neural Reasoner (ReFiNER) to extract both global and localized semantic representations. These visual features are fused and projected into the language embedding space of a Flan-T5-based Large Language Model (LLM) through a trainable MLP adapter. A parameter-efficient Low-Rank Adaptation (LoRA) strategy is employed, enabling fine-tuning with only 3.40 % of the total model parameters while preserving the generalization capabilities of the pretrained LLM. The system is developed and evaluated on a custom annotated dataset with multi-task supervision, encompassing scene classification, visual question answering, and region-level grounding. Experimental results demonstrate strong performance in both zero-shot and fine-tuned configurations, generating grounded and context-aware outputs. These findings position SafeVision as a robust and efficient solution for vision–language reasoning in real-world safety-critical applications.

CRediT authorship contribution statement

Syed Murtaza Hussain Abidi: Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Syed Muhammad Raza:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology. **Soo Young Shin:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships that may be considered potential competing interests:

Soo Young Shin reports that support was provided by Kumoh National Institute of Technology. Soo Young Shin reports a relationship with Kumoh National Institute of Technology that includes employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2024–00437190, 50 %). This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2022–00156394, 50 %).

Data availability

No data was used for the research described in the article.

References

[1] L. Albert, C. Routh, Designing impactful construction safety training interventions, Safety 7 (2) (2021) 42.
[2] S. Rasouli, Y. Alipouri, S. Chamanzad, Smart Personal Protective Equipment (PPE) for construction safety: a literature review, Safety Sci. 170 (2024) 106368.
[3] N.D. Nath, A.H. Behzadan, S.G. Paal, Deep learning for site safety: real-time detection of personal protective equipment, Autom. Constr. 112 (2020) 103085.
[4] A.M. Vukicevic, M. Petrovic, P. Milosevic, A. Peulic, K. Jovanovic, A. Novakovic, A systematic review of computer vision-based personal protective equipment compliance in industry practice: advancements, challenges and future directions, Artif. Intell. Rev. 57 (12) (2024) 319.
[5] M.T. Hosain, A. Zaman, M.R. Abir, S. Akter, S. Mursalin, S.S. Khan, Synchronizing object detection: applications, advancements and existing challenges, IEEE Access 12 (2024) 54129–54167.
[6] Z. Chen, H. Chen, M. Imani, R. Chen, F. Imani, Vision language model for interpretable and fine-grained detection of safety compliance in diverse workplaces, Expert Syst. Appl. 265 (2025) 125769.
[7] M. Ye, X. Rong, W. Huang, B. Du, N. Yu, D. Tao, A survey of safety on large vision-language models: attacks, defenses and evaluations, arXiv preprint arXiv:2502.14881, 2025.
[8] Y. Lee, K. Kim, K. Park, I. Jung, S. Jang, S. Lee, Y.-J. Lee, S.J. Hwang, Holisafe: Holistic safety benchmarking and modeling with safety meta token for vision-language model, arXiv preprint arXiv:2506.04704, 2025.
[9] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 46 (8) (2024) 5625–5644.
[10] F. Bordes, R.Y. Pang, A. Ajay, A.C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, et al., An introduction to vision-language modeling, arXiv preprint arXiv:2405.17247, 2024.
[11] X. Wu, R. Xian, T. Guan, J. Liang, S. Chakraborty, F. Liu, B.M. Sadler, D. Manocha, A. Bedi, On the safety concerns of deploying LLMS/VLMS in robotics: highlighting the risks and vulnerabilities, in: First Vision and Language for Autonomous Driving and Robotics Workshop, 2024.
[12] Z. Liu, Y. Nie, Y. Tan, X. Yue, Q. Cui, C. Wang, X. Zhu, B. Zheng, Safety alignment for vision language models, arXiv preprint arXiv:2405.13581, 2024.
[13] C. Cai, T. Liu, J. Gao, W. Liu, K. Wu, R. Wang, Y. Wang, S.C. Liew, From semantics, scene to instance-awareness: Distilling foundation model for open-vocabulary situation recognition, arXiv preprint arXiv:2507.14686, 2025.
[14] E. Svertoka, S. Saafi, A. Rusu-Casandra, R. Burget, I. Marghescu, J. Hosek, A. Ometov, Wearables for industrial work safety: a survey, Sensors 21 (11) (2021) 3844.
[15] A. Rashidi, G.L. Woon, M. Dasandara, M. Bazghaleh, P. Pasbakhsh, Smart personal protective equipment for intelligent construction safety monitoring, Smart Sustain. Built Environ. 14 (3) (2025) 835–858.
[16] Y. Zhang, S. Huang, J. Qin, X. Li, Z. Zhang, Q. Fan, Q. Tan, Detection of helmet use among construction workers via helmet-head region matching and state tracking, Autom. Constr. 171 (2025) 105987.
[17] J. Alnahas, Faster-PPENet: advancing logistic intelligence for PPE recognition at construction sites, IEEE Access 13 (2025).
[18] Q. Chen, D. Long, S. Wang, Q. Chen, B. Yuan, Real-time detection of personal protective equipment violations for construction workers using semisupervised learning and video clips, J. Constr. Eng. Manag. 151 (3) (2025) 04024213.

[19] Y. Wang, H. Luo, W. Fang, An integrated approach for automatic safety inspection in construction: domain knowledge with multimodal large language model, Adv. Eng. Inform. 65 (2025) 103246.
[20] A.K. Rajpoot, G. Agrawal, Synergistic fusion: vision-language models in advancing autonomous driving and intelligent transportation systems, in: Building Embodied AI Systems: the Agents, the Architecture Principles, Challenges, and Application Domains, Springer, 2025, pp. 205–221.
[21] X. Han, S. Chen, Z. Fu, Z. Feng, L. Fan, D. An, C. Wang, L. Guo, W. Meng, X. Zhang, et al., Multimodal fusion and vision-language models: a survey for robot vision, arXiv preprint arXiv:2504.02477, 2025.
[22] A. Baiocchi, I. Spinelli, A. Nicolosi, S. Scardapane, Adaptive token selection for scalable point cloud transformers, Neural Netw. (2025) 107477.
[23] M. Adil, G. Lee, V.A. Gonzalez, Q. Mei, Using vision language models for safety hazard identification in construction, arXiv preprint arXiv:2504.09083, 2025.
[24] S. Shriram, S. Perisetla, A. Keskar, H. Krishnaswamy, T.E.W. Bossen, A. Møgelmose, R. Greer, Towards a multi-agent vision-language system for zero-shot novel hazardous object detection for autonomous driving safety, arXiv preprint arXiv:2504.13399, 2025.
[25] T. Wang, C. Han, J. Liang, W. Yang, D. Liu, L.X. Zhang, Q. Wang, J. Luo, R. Tang, Exploring the adversarial vulnerabilities of vision-language-action models in Robotics, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 6948–6958.
[26] Z. Cheng, J. Liang, H. Choi, G. Tao, Z. Cao, D. Liu, X. Zhang, Physical attack on monocular depth estimation with optimal adversarial patches, in: European Conference on Computer Vision, Springer, 2022, pp. 514–532.
[27] O. Elmakis, T. Shaked, A. Degani, Vision-based UAV-UGV collaboration for autonomous construction site preparation, IEEE Access 10 (2022) 51209–51220.
[28] D. Tian, M. Li, Q. Ren, X. Zhang, S. Han, Y. Shen, Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining, Autom. Constr. 145 (2023) 104670.
[29] Z. Xu, B. Chen, X. Zhan, Y. Xiu, C. Suzuki, K. Shimada, A vision-based autonomous UAV inspection framework for unknown tunnel construction sites with dynamic obstacles, IEEE Robot. Autom. Lett. 8 (8) (2023) 4983–4990.
[30] B. Yang, B. Zhang, Y. Han, B. Liu, J. Hu, Y. Jin, Vision transformer-based visual language understanding of the construction process, Alex. Eng. J. 99 (2024) 242–256.
[31] B. Xiao, Y. Wang, Y. Zhang, C. Chen, A. Darko, Automated daily report generation from construction videos using Chatgpt and computer vision, Autom. Constr. 168 (2024) 105874.
[32] Y. Wen, K. Chen, Autonomous detection and assessment of indoor building defects using multimodal learning and GPT, in: Construction Research Congress 2024, 2024, pp. 1001–1009.
[33] D. Xiao, M. Dianati, P. Jennings, R. Woodman, HazardVLM: a video language model for real-time hazard description in automated driving systems, IEEE Trans. Intell. Veh. 10 (2024).
[34] H. Chen, L. Hou, S. Wu, G. Zhang, Y. Zou, S. Moon, M. Bhuiyan, Augmented reality, deep learning and vision-language query system for construction worker safety, Autom. Constr. 157 (2024) 105158.
[35] W.-L. Tsai, P.-L. Le, W.-F. Ho, N.-W. Chi, J.J. Lin, S. Tang, S.-H. Hsieh, Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention, Autom. Constr. 169 (2025) 105863.
[36] C.-F. Chan, P.K.-Y. Wong, X. Guo, J.C.P. Cheng, J.P.-C. Chan, P.-H. Leung, X. Tao, Context-aware vision-language model agent enriched with domain-specific ontology for construction site safety monitoring, Autom. Constr. 177 (2025) 106305.
[37] T. Kim, S. Kim, W.-C. Chern, S. Park, D. Kim, H. Kim, Optimizing large vision-language models for context-aware construction safety assessment, Autom. Constr. 180 (2025) 106510.
[38] R. Pei, J. Liu, W. Li, B. Shao, S. Xu, P. Dai, J. Lu, Y. Yan, Clipping: distilling clip-based models with a student base for video-language retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18983–18992.
[39] Y. Li, F. Xu, Z. Zhang, X. Mei, H. Huang, Construction Site Fall Hazard Identification and Automated Captioning Using Adapted Vision-Language Models, 2025. Available at SSRN 5340544.
[40] Y.-T. Chen, T.-C. Hu, Automated detection of construction opening safety risks using computer vision, Int. J. Saf. Secur. Eng. 15 (5) (2025).
[41] C. Che, Q. Lin, X. Zhao, J. Huang, L. Yu, Enhancing multimodal understanding with clip-based image-to-text transformation, in: Proceedings of the 2023 6th International Conference on Big Data Technologies, 2023, pp. 414–418.
[42] R. Liu, J. Huang, G. Li, J. Feng, X. Wu, T.H. Li, Revisiting temporal modeling for clip-based image-to-video knowledge transferring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6555–6564.
[43] Y. Zhai, S. Tong, X. Li, M. Cai, Q. Qu, Y.J. Lee, Y. Ma, Investigating the catastrophic forgetting in multimodal large language model fine-tuning, in: Conference on Parsimony and Learning, PMLR, 2024, pp. 202–227.
[44] Y. Zhai, S. Tong, X. Li, M. Cai, Q. Qu, Y.J. Lee, Y. Ma, Investigating the catastrophic forgetting in multimodal large language models, arXiv preprint arXiv:2309.10313, 2023.
[45] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, S. Liu, Spatialrgpt: grounded Spatial reasoning in vision-language models, Adv. Neural Inf. Process. Syst. 37 (2024) 135062–135093.
[46] X. Cao, T. Zhou, Y. Ma, W. Ye, C. Cui, K. Tang, Z. Cao, K. Liang, Z. Wang, J.M. Rehg, et al., MAPLM: a real-world large-scale vision-language benchmark for map and traffic scene understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21819–21830.
[47] G. Sun, M. Jin, Z. Wang, C.-L. Wang, S. Ma, Q. Wang, T. Geng, Y.N. Wu, Y. Zhang, D. Liu, Visual agents as fast and slow thinkers, arXiv preprint arXiv:2408.08862, 2024.
[48] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, J. Mach. Learn. Res. 25 (70) (2024) 1–53.
[49] H. Xue, Y. Sun, B. Liu, J. Fu, R. Song, H. Li, J. Luo, CLIP-ViP: adapting pre-trained image-text model to video-language alignment, in: The Eleventh International Conference on Learning Representations, 2023.
[50] S. Hayou, N. Ghosh, B. Yu, Lora+: Efficient low rank adaptation of large models, arXiv preprint arXiv:2402.12354, 2024.
[51] M. Reyad, A.M. Sarhan, M. Arafa, A modified Adam algorithm for deep neural network optimization, Neural Comput. Appl. 35 (23) (2023) 17095–17112.
[52] T. Kim, S. Kim, W.-C. Chern, S. Park, D. Kim, H. Kim, Optimizing large vision-language models for context-aware construction safety assessment, Autom. Constr. 180 (2025) 106510.
[53] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, Y.J. Lee, LLaVANeXT: improved reasoning, OCR, and world knowledge (2024).
[54] Team, Gemini and Anil, Rohan and Borgeaud, Sebastian and Alayrac, Jean-Baptiste and Yu, Jiahui and Soricut, Radu and Schalkwyk, Johan and Dai, Andrew M and Hauth, Anja and Millican, Katie and others, Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805, 2023.
[55] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., GPT-4 technical report, arXiv preprint arXiv:2303.08774, 2023.
[56] Q. Chen, X. Yin, Tailored vision-language framework for automated hazard identification and report generation in construction sites, Adv. Eng. Inform. 66 (2025) 103478.

Author biography



Syed Murtaza Hussain Abidi is pursuing a Master's degree in IT Convergence Engineering at the Wireless and Emerging Network Systems Laboratory (WENS Lab), Kumoh National Institute of Technology (KIT), South Korea. Currently, he is a graduate fellow and researcher at KIT. His research interests primarily focus on Computer Vision, Multi-modal Data Fusion, Vision-Language Models, 2D/3D Vision, and lightweight deep learning models, with practical applications in ROS (Robot Operating System).



Syed Muhammad Raza completed his graduate (M.S.) degree majoring in IT Convergence Engineering in 2023 at WENS Lab, Kumoh National Institute of Technology (KIT), South Korea. Currently, he is working as an AI Software Engineer. His research interests mainly include image processing, lightweight deep learning models, and computer vision applications.



Soo Young Shin (M'07–SM'17) received his Ph.D. in electrical engineering and computer science from Seoul National University in 2006. He served at Samsung Electronics (2007–2010) and joined Kumoh National Institute of Technology, where he is an Associate Professor. He was a postdoctoral researcher at the University of Washington (2007) and a visiting scholar at the University of British Columbia (2017). His research interests include future wireless communications, signal processing, unmanned mobility, quantum neural network, Internet of Things, AI/Deep learning applications, etc.