A Vision-Language-Driven UxV System for Safety Monitoring

Syed Murtaza Hussain Abidi,
Dept. of IT Convergence Engineering
Kumoh National Institute of Technology
Gumi, 39177, South Korea

Syed Muhammad Raza,
Dept. of IT Convergence Engineering
Kumoh National Institute of Technology
Gumi, 39177, South Korea

Jinsu Park,
Dept. of IT Convergence Engineering
Kumoh National Institute of Technology
Gumi, 39177, South Korea

Soo Young Shin,
Dept. of IT Convergence Engineering
Kumoh National Institute of Technology
Gumi, 39177, South Korea

2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN) | 979-8-3315-2487-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICUFN65838.2025.11169839

***Abstract*—This paper present a Vision-Language Model (VLM)-driven unmanned vehicle (UxV) system designed for real-time safety monitoring in high-risk industrial environments such as construction and oil and gas operations. The proposed system integrates advanced computer vision and natural language processing (NLP) to generate context-aware textual descriptions of safety conditions. It employs unmanned aerial vehicles (UAVs) for wide-area surveillance and unmanned ground vehicles (UGVs) for close-range inspection, thereby enhancing spatial coverage and operational flexibility. To support VLM fine-tuning, a custom dataset comprising 2,000 safety-related images was developed, enabling dynamic hazard detection and assessment of regulatory compliance. Unlike conventional object detection methods, the system leverages zero-shot learning to facilitate rapid adaptation to new environments with minimal retraining. Preliminary evaluations indicate that the system-generated captions closely align with human annotations, demonstrating high interpretability and real-time applicability. The framework exhibits strong scalability, offering a proactive approach to hazard detection and safety compliance monitoring in complex industrial settings.**

***Index Terms*—Vision Language Model, Unmanned Vehicles, Safety.**

I. INTRODUCTION

Workplace safety remains a critical concern in high-risk industries such as manufacturing, construction, oil and gas, electrical work, and chemical processing, where hazardous environments pose persistent threats to workers. Despite the enforcement of strict safety regulations, the use of mandated personal protective equipment (PPE), and the implementation of comprehensive training programs, workplace accidents continue to occur. A key limitation is the absence of real-time, context-aware monitoring systems capable of adapting to dynamic and evolving industrial conditions.

Conventional computer vision-based safety systems have advanced PPE detection and improved compliance monitoring. However, these systems typically rely on static, rule-based object detection and lack the capacity to interpret complex worker behaviors or contextual nuances. This restricts their effectiveness in identifying emerging hazards or assessing safety compliance in novel scenarios.

To overcome these limitations, a Vision-Language Model (VLM)-based system is proposed for real-time safety monitoring in industrial environments. This system combines advanced computer vision and natural language processing (NLP) to produce context-aware textual descriptions of safety conditions, enhancing interpretability and decision-making. By leveraging zero-shot learning, the framework supports rapid adaptation to new work environments with minimal retraining, thereby improving scalability and operational flexibility. Additionally, the integration of unmanned ground vehicles (UGVs) for detailed on-site inspections and unmanned aerial vehicles (UAVs) for broad-area surveillance enables comprehensive and adaptive safety monitoring.

The contributions of this work are summarized as follows:

- Develop a custom safety dataset comprising 2,000 annotated images to support real-time compliance monitoring across diverse industrial settings.
- Design of a VLM-driven unmanned vehicle (UxV) system that surpasses traditional object detection approaches by generating interactive, context-sensitive natural language descriptions of workplace safety conditions.

II. RELATED WORK

A. Safety Monitoring with Vision-Language Models

Recent advancements in vision-language models (VLMs) have significantly expanded the potential of safety monitoring systems across various industrial contexts. Chen et al. [1] introduced a fine-grained VLM tailored for safety compliance detection, demonstrating that interpretable multimodal models can effectively enhance workplace monitoring. Similarly, Yang et al. [2] leveraged vision transformer architectures to improve the understanding of construction workflows, thereby contributing to safety-focused applications.

Several studies have also emphasized the importance of leveraging large-scale pre-training for safety-related tasks.

*Corresponding Author: Soo Young Shin (wdragon@kumoh.ac.kr).

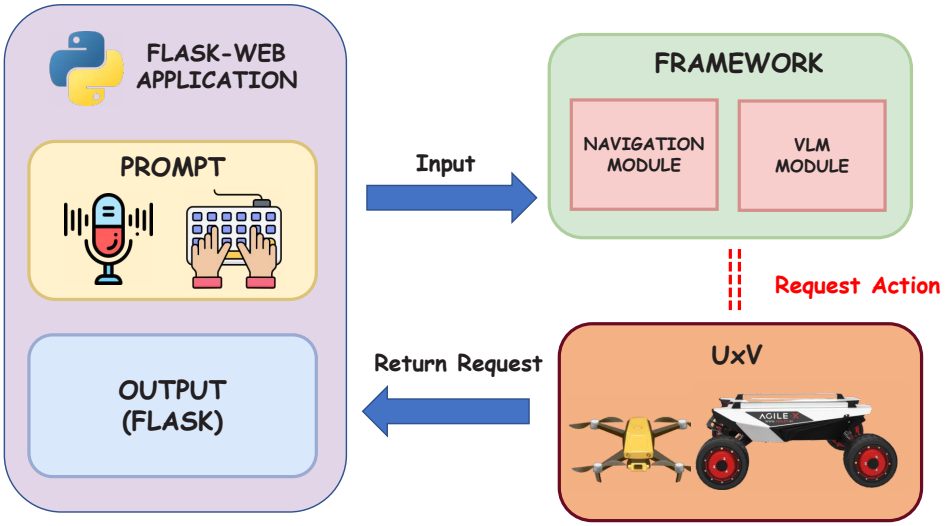

Fig. 1. High-level schematic of the proposed VLM-driven UxV safety monitoring system.

Tsai et al. [3] explored the application of Contrastive Language–Image Pretraining (CLIP) in the context of construction site inspections, while Xiao et al. [4] proposed an automated framework that combines ChatGPT with computer vision for daily site report generation. Wen et al. [5] further extended this line of work by incorporating GPT-based assessments and multimodal learning for autonomous detection of structural defects in indoor environments.

B. Advances in Vision-Language Architectures

The development of VLMs has been catalyzed by recent breakthroughs in multimodal representation learning. Notable contributions include Flamingo [6], which introduced a flexible architecture for handling diverse modalities, and CLIP [7], which achieved state-of-the-art performance in image-text alignment through large-scale semantic pretraining. Most instruction-following VLMs build upon a robust visual backbone, such as the Vision Transformer (ViT) [8], and integrate modality fusion strategies to enable joint reasoning.

Subsequent models, such as BLIP-2 [9], have further demonstrated the capacity of VLMs to generalize across a wide range of vision-language tasks. This capability has also been extended to video understanding applications, as evidenced by Maaz et al. [10]. Collectively, these developments highlight the transformative role of VLMs in enabling automated hazard recognition, real-time compliance verification, and contextualized reporting in safety-critical domains.

III. Proposed Methodology

A. Overall System

As illustrated in Fig.1, the proposed system enables real-time safety monitoring using Unmanned Ground and Aerial Vehicles (UxVs), integrating a fine-tuned Vision-Language Model (VLM) with an interactive web interface. The system automatically generates semantic descriptions of visual scenes to enhance situational awareness and support proactive safety management in industrial environments.

B. System Architecture and User Interaction

The core interface is implemented as a Flask-based web application, facilitating human-system interaction. Users can issue commands via either speech or text, which are processed and transmitted to the UxV subsystem. In response, unmanned aerial vehicles (UAVs) and unmanned ground vehicles (UGVs) capture images and transmit them back to the web application. Each image is displayed alongside a descriptive caption generated by the VLM. This setup supports real-time interaction, enabling operators to iteratively refine commands and request updated visual data dynamically.

C. Vision-Language Model and Fine-Tuning

The captioning component is powered by BLIP-2, a state-of-the-art Vision-Language Model that integrates computer vision with natural language processing to produce contextually rich image descriptions. To adapt the model for industrial safety applications, BLIP-2 is fine-tuned on a custom dataset focused

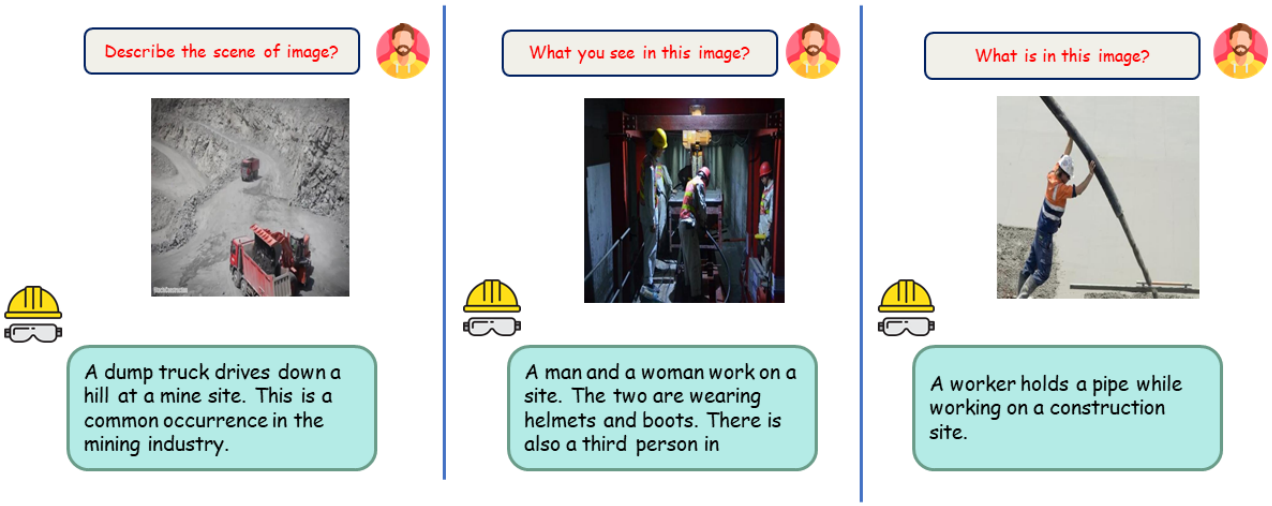

Fig. 2. Qualitative results of proposed system

on hazard detection and compliance verification. Fine-tuning is conducted to enhance the model's ability to identify domain-specific risks and compliance indicators within construction and manufacturing contexts.

D. Custom Dataset and Annotation Process

Fig. 3. Sample scenarios from the custom construction site dataset, categorized into three primary annotation tasks: Worker Compliance (top row), PPE Detection (middle row), and Construction Site Safety (bottom row). These categories support comprehensive safety and compliance monitoring.

The custom dataset comprises images drawn from open-source repositories and organized into three categories:

- **Construction Site Safety:** Scenes featuring heavy machinery and environmental hazards.
- **Worker Compliance:** Images depicting workers on ladders, with or without appropriate safety equipment.
- **PPE Detection:** Instances illustrating the presence or absence of personal protective equipment such as helmets, gloves, and vests.

Manual annotation was performed using the Label Studio platform, assigning each image both a scene label—selected from a controlled vocabulary (*construction site*, *mining site*, *indoor*, *garden*)—and a visual question–answer pair focused on safety assessment. Captions are constrained to 50 characters to satisfy BLIP-2 dual-encoder input requirements, producing concise, domain-specific labels that enhance fine-tuning efficacy and support high-accuracy, real-time safety monitoring.

E. Output Module and Real-Time Feedback

The final output is rendered through the web application interface, which presents each captured image alongside its generated caption. This real-time feedback loop provides immediate safety insights and enables users to issue follow-up queries or request additional image captures. The system thereby supports continuous, adaptive monitoring, offering a scalable solution for safety oversight in complex operational environments.

IV. EXPERIMENTAL ANALYSIS

A. Training Parameter

The BLIP-2 Vision-Language Model (VLM) is employed for image captioning tasks. Training is conducted using a batch size of 2, with data shuffling enabled to ensure stochastic variation across epochs. A total of 100 training epochs are executed. The full model architecture consists of approximately 3.75 billion parameters, of which 5.24 million parameters (approximately 0.14%) are designated as trainable. Optimization is performed using the Adam algorithm with a fixed learning rate of (lr) of 1e-4, while a cosine annealing learning rate scheduler is applied to facilitate smooth convergence. Gradient clipping is employed to stabilize updates and prevent exploding gradients during optimization.

B. Fine-Tunning Parameter

Model fine-tuning is conducted using Low-Rank Adaptation (LoRA) with 8-bit quantization to reduce computational overhead. The LoRA configuration includes a rank r of 16,, a scaling factor lora_alpha of 32, and a dropout rate lora_dropout of 0.05. No modifications are made to the model's bias terms. Adaptation is restricted to the q_proj and k_proj modules, enabling task-specific tuning while preserving the model's general-purpose zero-shot reasoning capabilities. This setup allows the system to efficiently adapt to the domain-specific demands of safety monitoring in industrial environments.

V. RESULT

Figure 2 illustrates qualitative examples generated by the system, where each user prompt elicits a descriptive or safety-related caption in real time. The evaluation dataset comprises 357 images sampled from diverse industrial scenarios, including construction sites, indoor facilities, and equipment-intensive environments.

Results indicate a high degree of alignment between the generated captions and manually annotated ground truth descriptions, suggesting strong interpretability and semantic coherence. The system demonstrates robustness in handling varied safety contexts, such as detecting missing PPE, recognizing hazardous postures, and identifying environmental threats. These findings support the viability of employing a fine-tuned VLM for automated safety assessment and suggest practical applicability in real-world industrial settings.

A. Quantitative Evaluation via Standard Metrics

TABLE I
EVALUATION METRICS FOR ANALYSIS

Metric	Score (%)	Note
BLEU@1	19.66	Unigram precision
BLEU@2	8.80	Bigram precision
BLEU@3	3.89	Trigram precision
METEOR	12.45	Includes synonyms

Table I presents the system's performance using well-established automatic evaluation metrics for image captioning. The BLEU scores (BLEU@1 = 19.66%, BLEU@2 = 8.80%, BLEU@3 = 3.89%) reflect progressively decreasing precision as the n-gram size increases, a common trend in natural language generation tasks when outputs are semantically aligned but lexically diverse. The relatively low BLEU@2 and BLEU@3 values suggest limited exact phrase overlap between model-generated and reference captions, which is typical in real-world descriptive captioning due to lexical variability. In contrast, the METEOR score of 12.45%—which accounts for synonymy and stem matches—indicates stronger semantic alignment between predictions and references, supporting the notion that the model captures essential meaning even with lexical variations.

B. Domain-Specific Keyword Recall Analysis

TABLE II
KEYWORD RECALL ANALYSIS

Keyword	Recall (%)	Found / Total
helmet	33.33	(1 / 3)
worker	50.00	(4 / 8)
safety	100.00	(5 / 5)
dump truck	33.33	(1 / 3)
gloves	33.33	(1 / 3)

To complement the standard metrics, Table II presents keyword-level recall scores for five critical safety-related terms: helmet, worker, safety, dump truck, & gloves. This analysis quantifies how frequently the model reproduces essential terms found in the reference captions. The keyword safety achieved perfect recall (100.00%), indicating consistent recognition and generation of this central concept across all relevant test instances. The term worker also yielded moderate recall (50.00%), showing partial success in referencing human presence in the scene.

However, recall rates for helmet, dump truck, & gloves remain at 33.33%, revealing challenges in the consistent identification and generation of more specific objects or entities. This suggests a need for improved grounding of fine-grained object semantics within the caption generation process, particularly when such elements are contextually important but visually subtle.

These keyword-level findings are critical because they highlight the system's effectiveness in prioritizing safety-relevant content, even when BLEU scores remain low due to stylistic differences between generated and reference captions. The recall analysis directly evaluates the system's ability to include high-priority terminology crucial for automated safety monitoring.

VI. CONCLUSION

This study presents a Vision-Language Model (VLM)-based unmanned vehicle (UxV) system for real-time safety monitoring in high-risk industrial environments. By integrating computer vision and natural language processing, the system facilitates dynamic hazard detection, real-time compliance verification, and interpretable safety assessments. The proposed

framework advances existing safety monitoring techniques by moving beyond static object detection toward interactive, context-aware analysis.

VII. FUTURE WORK

Future work will enhance system capabilities by integrating advanced VLM components, such as scene classification for better environmental understanding, referring expression comprehension for precise object localization, and multi-turn conversational interactions to enable interactive communication with human operators. These enhancements will improve contextual awareness, facilitate proactive safety interventions, and foster intelligent human-machine collaboration. By extending its adaptability and decision-making capabilities, the proposed system aims to set a new benchmark for workplace safety monitoring and regulatory compliance enforcement across various industries.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2025-RS-2022-00156394, 50%) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation)

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2024-00437190, 50%) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

REFERENCES

[1] Z. Chen, H. Chen, M. Imani, R. Chen, and F. Imani, "Vision language model for interpretable and fine-grained detection of safety compliance in diverse workplaces," *Expert Systems with Applications*, vol. 265, p. 125769, 2025.

[2] B. Yang, B. Zhang, Y. Han, B. Liu, J. Hu, and Y. Jin, "Vision transformer-based visual language understanding of the construction process," *Alexandria Engineering Journal*, vol. 99, pp. 242–256, 2024.

[3] W.-L. Tsai, P.-L. Le, W.-F. Ho, N.-W. Chi, J. J. Lin, S. Tang, and S.-H. Hsieh, "Construction safety inspection with contrastive language-image pre-training (clip) image captioning and attention," *Automation in Construction*, vol. 169, p. 105863, 2025.

[4] B. Xiao, Y. Wang, Y. Zhang, C. Chen, and A. Darko, "Automated daily report generation from construction videos using chatgpt and computer vision," *Automation in Construction*, vol. 168, p. 105874, 2024.

[5] Y. Wen and K. Chen, "Autonomous detection and assessment of indoor building defects using multimodal learning and gpt," in *Construction Research Congress 2024*, 2024, pp. 1001–1009.

[6] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[10] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.