An Intelligent Vision–Language Framework for Real-Time Environmental Hazard and Safety Risk Detection Using Deep Learning

Santhosh S[1], Ashwin Shenoy M[2*], Sandeep Kumar S[3]

[1,2,3] Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT), Karkala, India
EMAILS: [1]santhosh.s@nitte.edu.in, ashwinshenoy14@gmail.com[2], sandeep.kumar@nitte.edu.in[3]

Abstract
Environmental hazards and safety risks arising from pollution incidents, waste accumulation, chemical spills, fire outbreaks, and other unsafe conditions pose significant challenges to environmental protection and sustainable resource management. Conventional monitoring systems often rely on manual inspection and continuous human supervision, which can lead to delayed responses and reduced effectiveness in large-scale environments. Recent advances in artificial intelligence have enabled automated environmental monitoring; however, many existing approaches require extensive labeled datasets and frequent retraining to adapt to diverse operational conditions.
This paper presents an intelligent vision–language framework for real-time environmental hazard detection and safety risk assessment using deep learning and computer vision techniques. The proposed system utilizes the Contrastive Language–Image Pretraining (CLIP) model in a zero-shot learning paradigm to identify environmental hazards from surveillance video streams without task-specific retraining. Visual features extracted from environmental scenes are compared with descriptive textual prompts through cosine similarity-based matching, enabling the recognition of diverse hazardous situations, including waste dumping, smoke emissions, fire incidents, water contamination indicators, and unsafe environmental conditions.
To improve operational reliability, a risk intelligence module categorizes detected events into multiple severity levels, namely Safe, Moderate Risk, and High Risk. Confidence-based thresholding and heuristic validation mechanisms are incorporated to minimize false detections and enhance decision-making accuracy. The framework processes surveillance footage in real time and generates automated alerts to facilitate timely intervention and environmental protection measures. Experimental evaluation demonstrates the effectiveness of the proposed approach in accurately identifying environmental hazards across diverse scenarios while maintaining adaptability to previously unseen conditions. The proposed framework offers a scalable and cost-effective solution for intelligent environmental monitoring and sustainable safety management.

Keywords: Environmental Hazard Detection, Environmental Monitoring, Computer Vision, Deep Learning, CLIP, Vision–Language Models, Zero-Shot Learning, Risk Assessment, Environmental Safety Surveillance, Sustainable Environmental Management.

1. Introduction

Environmental hazards and safety risks pose significant challenges to sustainable development, environmental protection, and public safety. Incidents such as illegal waste disposal, chemical spills, smoke emissions, fire outbreaks, water contamination, and other unsafe environmental conditions can have severe consequences for ecosystems, natural resources, and human health. With the increasing deployment of surveillance cameras and monitoring infrastructure in industrial facilities, urban environments, public spaces, and environmentally sensitive regions, vast amounts of visual data are generated continuously. Monitoring these data streams manually is labor-intensive, time-consuming, and often ineffective for timely identification of hazardous events.
Traditional environmental monitoring systems largely depend on periodic inspections and human observation. Such approaches are often limited by delayed response times, human fatigue, subjective judgment, and the inability to continuously monitor large geographic areas. As environmental risks become increasingly complex and dynamic, there is a growing need for intelligent monitoring systems capable of automatically detecting hazardous situations and providing real-time alerts for rapid intervention.
Recent advancements in artificial intelligence, deep learning, and computer vision have enabled automated analysis of visual data for environmental surveillance applications. Convolutional Neural Networks (CNNs) and related deep learning models have demonstrated promising performance in object detection, scene classification, and anomaly recognition. However, most existing approaches rely on supervised learning techniques that require large volumes of labeled training data and extensive retraining whenever new hazard categories or environmental conditions are encountered. Such requirements limit their scalability and adaptability in real-world deployments.
Environmental hazard detection presents several challenges due to variations in lighting conditions, weather patterns, camera viewpoints, background complexity, and the diversity of environmental scenarios. Furthermore, hazardous events often manifest differently across locations and contexts, making it difficult for conventional supervised models to generalize effectively. Consequently, systems trained on specific datasets may experience significant performance degradation when applied to previously unseen environments.

These limitations highlight a critical research gap: the lack of flexible and context-aware environmental monitoring frameworks capable of recognizing diverse hazardous situations without requiring extensive task-specific training. Addressing this challenge requires models that can combine visual perception with semantic understanding of environmental contexts and adapt dynamically to new situations.

To address these challenges, this study proposes an intelligent real-time environmental hazard detection and safety risk assessment framework based on the Contrastive Language–Image Pretraining (CLIP) model. Unlike traditional supervised approaches, the proposed framework employs a zero-shot learning paradigm that enables the identification of various environmental hazards without requiring task-specific retraining. By comparing visual features extracted from surveillance video frames with descriptive textual prompts through semantic similarity analysis, the system can recognize a broad range of hazardous situations, including waste dumping, smoke emissions, fire incidents, water contamination indicators, and other unsafe environmental conditions.

In addition, the framework incorporates a risk intelligence module that categorizes detected events into different severity levels, namely Safe, Moderate Risk, and High Risk. Confidence-based thresholding and heuristic validation mechanisms are integrated to improve detection reliability and minimize false alarms. By combining vision-language learning, semantic scene understanding, real-time video processing, and adaptive risk assessment, the proposed system offers an effective and scalable solution for intelligent environmental monitoring and sustainable safety management.

The remainder of this paper is organized as follows. Section 2 reviews related work on environmental monitoring, hazard detection, and vision-language learning techniques. Section 3 describes the proposed system architecture and methodology. Section 4 presents the experimental setup and evaluation procedures. Section 5 discusses the results and performance analysis, while Section 6 concludes the study and outlines future research directions.

2. Literature Survey

Recent advances in computer vision, deep learning, and vision-language models have significantly improved intelligent monitoring systems for hazard detection, risk assessment, and environmental safety surveillance.

Yang et al. [1] proposed a multi-agent vision-language framework for highway scene understanding that integrates weather classification, pavement wetness assessment, and traffic congestion detection. Their mixture-of-experts architecture demonstrated robust performance across diverse environmental conditions. However, the system was primarily designed for transportation environments and requires domain-specific reasoning.

Chen et al. [2] introduced Clip2Safety, a vision-language framework for workplace safety compliance monitoring. The model utilized CLIP-based scene understanding and fine-grained personal protective equipment (PPE) verification to improve workplace safety assessment. Although effective, the framework mainly focuses on PPE compliance rather than generalized hazard detection.

Delhi et al. [3] developed a YOLOv3-based deep learning framework for PPE compliance monitoring on construction sites. Their system achieved an F1-score of 0.96 and enabled real-time safety alerts. Nevertheless, the approach relies on supervised learning and requires extensive labeled datasets.

Alayed et al. [4] proposed a real-time fire safety equipment inspection system using YOLOv5, YOLOv7, YOLOv8, and RT-DETR models. Their comparative analysis showed that YOLOv8 achieved the highest accuracy and fastest inference speed. However, the framework is limited to fire extinguisher inspection applications.

Gupta et al. [5] introduced VARS, a vision-based danger assessment system that combines CLIP and GPT embeddings for video risk analysis. The multimodal framework achieved an accuracy of 85% for binary danger classification. However, the study was conducted on a relatively small dataset containing only 100 videos.

Shriram et al. [6] proposed a multi-agent vision-language system for zero-shot hazardous object detection in autonomous driving environments. The framework employed CLIP-based semantic matching and language-guided reasoning for detecting unseen hazards. Despite promising results, the system showed limitations in detecting small or occluded hazardous objects.

Guo et al. [7] developed an intelligent vision-enabled maritime surveillance framework for detecting water-surface targets under adverse weather conditions. Their data augmentation strategy significantly improved detection robustness. However, the dataset considered only a limited number of pollution categories.

Liu et al. [8] presented an automatic construction hazard identification framework that integrates scene graph generation and BERT-based information extraction. The system effectively combined visual and textual knowledge for hazard inference. Nevertheless, establishing large-scale multimodal datasets remains a challenge.

Deng et al. [9] proposed a multimodal dangerous state recognition system for elderly individuals with intermittent dementia. By combining visual scene information and location data, the framework achieved superior danger-state classification performance. However, its application is restricted to elderly care environments.

Önal and Demir [10] introduced Unsafe-Net, a hybrid YOLOv4-ConvLSTM framework for real-time workplace safety monitoring. The proposed system achieved 95.81% classification accuracy and reduced unsafe behavior recurrence by approximately 75%. However, the model was trained on a relatively small factory-specific dataset.

Park et al. [11] proposed a pose-estimation-based framework for identifying fall hazards in construction environments. Their approach improved worker localization and hazard proximity estimation compared to conventional detection systems. Nevertheless, false positives may occur in complex site layouts.

Gupta et al. [12] introduced ViDAS, a vision-based danger assessment and scoring framework that evaluates danger severity using large language models. Their findings demonstrated that multimodal reasoning can closely approximate human danger perception. However, the benchmark dataset size remains limited.

Zhang et al. [13] proposed CAMERA, a context-aware multimodal risk anticipation framework integrating video streams, textual annotations, and driver attention maps. The system achieved state-of-the-art accident anticipation performance but requires multiple input modalities, increasing deployment complexity.
Chan et al. [14] developed a domain knowledge-enhanced vision-language model for construction site safety monitoring. Their curriculum-learning-based framework achieved approximately 90% accuracy in detecting safety violations. However, the approach remains domain-specific and depends on construction-related knowledge.
Jeon et al. [15] presented an image-captioning-based surveillance system for advanced risk assessment. By integrating BLIP-2 and BERT models, the framework achieved classification accuracies exceeding 90% across multiple risk categories. Nevertheless, extensive fine-tuning is required to adapt the system to new environments.
Recent studies have demonstrated the growing effectiveness of computer vision, deep learning, and vision-language models for intelligent hazard detection, risk assessment, and safety monitoring across various domains. Researchers have proposed frameworks for highway scene understanding, workplace safety compliance, PPE detection, fire safety inspection, danger assessment, autonomous driving hazard recognition, maritime surveillance, construction hazard identification, elderly safety monitoring, and accident anticipation. Most approaches leverage advanced techniques such as CLIP, YOLO, ConvLSTM, BERT, BLIP-2, and multimodal learning to improve detection accuracy and contextual understanding. While these methods have achieved promising results in domain-specific applications, they often rely on large annotated datasets, extensive training, multiple input modalities, or environment-specific knowledge, limiting their adaptability to unseen scenarios. Furthermore, challenges such as occlusion, dataset limitations, computational complexity, and lack of generalization remain significant concerns. These limitations highlight the need for a flexible and context-aware framework capable of performing real-time hazard detection and risk assessment across diverse environments without extensive retraining, motivating the development of the proposed CLIP-based zero-shot environmental hazard detection system.

3 Proposed Methodology

3.1 System Overview

The proposed system is designed as an intelligent context-aware framework for real-time environmental hazard detection and safety risk assessment in surveillance environments. The framework leverages computer vision and vision-language learning techniques to automatically analyze video streams, identify potentially hazardous environmental conditions, and generate timely alerts for rapid intervention. Unlike traditional supervised approaches, the proposed system utilizes the Contrastive Language–Image Pretraining (CLIP) model in a zero-shot learning setting, enabling it to recognize a diverse range of environmental hazards without requiring task-specific retraining. By semantically matching visual information extracted from surveillance frames with descriptive textual prompts, the system can detect conditions such as waste accumulation, smoke emissions, fire incidents, water contamination indicators, and other unsafe environmental situations. Furthermore, the framework incorporates a risk intelligence module that categorizes detected events into different severity levels, including Safe, Moderate Risk, and High Risk. Confidence-based thresholding and heuristic validation mechanisms are employed to enhance detection reliability and minimize false alarms. Through the integration of semantic scene understanding, adaptive risk assessment, and real-time monitoring capabilities, the proposed framework provides a scalable and effective solution for intelligent environmental surveillance and sustainable safety management.

3.2 Event Classification Module

The proposed system incorporates a context-aware environmental hazard classification module as the core component for real-time environmental monitoring and safety risk assessment. The primary objective of this module is to identify potentially hazardous environmental conditions from surveillance video frames and categorize them into meaningful risk-related classes. The classification strategy combines vision-language learning with confidence-based decision mechanisms to improve accuracy, robustness, and adaptability across diverse environmental settings.
The core classification process is performed using the Contrastive Language–Image Pretraining (CLIP) model, a vision-language framework capable of zero-shot classification by aligning visual and textual representations within a shared semantic feature space. For each input frame extracted from the surveillance video stream, image embeddings are generated using the CLIP image encoder (ViT-B/32). Simultaneously, a predefined set of textual prompts describing various environmental conditions and hazard scenarios is encoded using the CLIP text encoder.
Example prompts include:

- "waste dumping near water body"
- "smoke emission in industrial area"
- "fire hazard in environment"
- "water pollution incident"
- "garbage accumulation in public place"
- "unsafe environmental condition"
- "clean and safe environment"
- "normal environmental activity"

The cosine similarity between image embeddings and text embeddings is computed, and the environmental

condition with the highest similarity score is selected as the initial prediction. This zero-shot learning capability enables the system to recognize a broad range of environmental hazards, including previously unseen scenarios, without requiring task-specific retraining.

To improve classification reliability, confidence-based thresholding is incorporated into the decision-making process. When the similarity score exceeds a predefined threshold, the corresponding environmental hazard category is accepted as a valid prediction. Predictions with moderate confidence are classified as uncertain events and subjected to additional validation procedures. If the similarity score falls below a minimum threshold, the frame is labeled as **"Unknown Condition"** to minimize the risk of incorrect classification.

Furthermore, heuristic validation mechanisms are

employed to reduce false alarms and improve prediction reliability. These mechanisms analyze confidence distributions, contextual cues, and semantic consistency across consecutive frames to distinguish genuine environmental hazards from visually similar but harmless situations. For example, temporary shadows, weather-related variations, or routine human activities may occasionally resemble hazardous conditions and require additional verification before generating alerts.

The outputs from the semantic classification module and confidence validation mechanisms are integrated through a decision layer that determines the final environmental hazard category. The resulting classification is subsequently forwarded to the **Risk Intelligence Module**, which evaluates the severity level and generates appropriate alerts for environmental monitoring authorities or safety personnel.

This context-aware environmental hazard classification framework enables the proposed system to adapt dynamically to diverse surveillance environments while maintaining high flexibility, scalability, and detection reliability. By combining semantic scene understanding, zero-shot vision-language learning, and confidence-based validation, the framework significantly enhances the effectiveness of real-time environmental hazard detection and safety risk assessment.

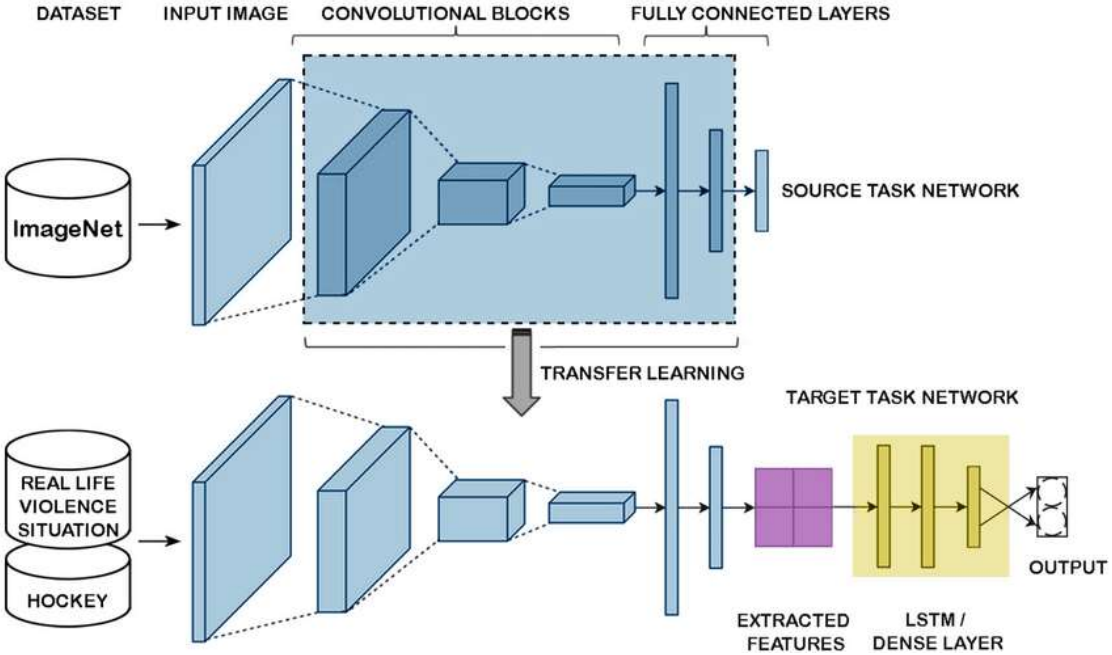

Figure 3.1: System Architecture of the Proposed Vision–Language Framework for Environmental Hazard Monitoring

3.3 Video Processing Module

The **Video Processing Module** serves as the initial stage of the proposed environmental hazard detection and safety risk assessment framework. Its primary objective is to acquire surveillance video streams, extract relevant frames, and preprocess them for subsequent semantic analysis using the Contrastive Language–Image Pretraining (CLIP) model. Since environmental surveillance footage may contain variations in illumination, weather conditions, camera viewpoints, background complexity, motion blur, and image quality, effective preprocessing is essential to ensure reliable hazard detection and environmental scene interpretation.

The module consists of **video acquisition, frame extraction, frame sampling, image preprocessing, and frame normalization**. These operations transform raw video streams into standardized inputs suitable for feature extraction and semantic understanding. Video streams are obtained from surveillance cameras deployed in environmental monitoring locations such as industrial zones, public spaces, water bodies, waste disposal areas, and environmentally sensitive regions. Frames are extracted at predefined intervals to reduce computational overhead while preserving relevant environmental information.

During preprocessing, image resizing, normalization, and noise reduction techniques are applied to improve visual quality and maintain consistency across frames. The processed frames are then converted into a format compatible with the CLIP image encoder (ViT-B/32), enabling efficient extraction of semantic visual features. This preprocessing stage helps mitigate the effects of environmental variations and enhances the robustness of subsequent hazard classification.

By generating standardized and high-quality visual inputs, the Video Processing Module establishes a reliable foundation for real-time environmental hazard detection, safety risk assessment, and intelligent surveillance operations. The processed frames are subsequently forwarded to the Environmental Hazard Classification Module for semantic analysis and risk evaluation.

3.3.1 Video Acquisition

The proposed framework accepts video streams from multiple sources, including CCTV cameras, IP cameras, webcams, and recorded surveillance footage deployed across environmental monitoring locations. The acquired video data may contain diverse environmental conditions and potential hazard scenarios such as waste accumulation, smoke emissions, fire incidents, water contamination indicators, and other unsafe environmental activities. The video acquisition stage continuously captures and transfers video frames to the processing pipeline, enabling real-time monitoring, scene analysis, and timely identification of environmental hazards. By providing a continuous stream of visual information, this module serves as the foundation for subsequent frame extraction, preprocessing, semantic classification, and safety risk assessment within the proposed intelligent environmental surveillance framework.

3.3.2 Frame Extraction

Video streams are decomposed into individual image frames using the OpenCV library. Frame extraction converts continuous temporal video data into discrete visual representations that can be processed by the CLIP model. Each frame preserves important spatial information required for identifying environmental conditions, pollution indicators, waste accumulation, smoke emissions, fire incidents, water contamination signs, and other safety-related events. The extracted frames serve as the primary input for subsequent semantic analysis and environmental hazard assessment.

3.3.3 Frame Sampling

Processing every frame of a surveillance video can significantly increase computational requirements and processing time. Therefore, frame sampling is employed to reduce computational overhead while preserving critical environmental information. Frames are selected at predefined intervals, ensuring that significant environmental events and hazardous conditions are captured without introducing excessive redundancy. This strategy improves computational efficiency and enables real-time deployment in large-scale environmental monitoring systems.

3.3.4 Image Preprocessing

Before semantic analysis, each extracted frame undergoes preprocessing to improve input consistency and feature quality. The preprocessing stage includes image resizing, color conversion, noise reduction, contrast enhancement, and image normalization operations. These preprocessing techniques help minimize variations caused by lighting conditions, weather changes, camera quality, and environmental noise while preserving important visual features required for environmental hazard recognition and scene understanding.

3.3.5 Frame Normalization

The preprocessed frames are normalized according to the input requirements of the CLIP model. Pixel values are scaled and transformed into a standardized format to ensure consistent feature extraction across different surveillance environments. Normalization improves model stability and allows the framework to process video frames captured under diverse environmental and operational conditions while maintaining reliable detection performance.

The output of the Video Processing Module is a collection of normalized surveillance frames that are subsequently forwarded to the CLIP-Based Semantic Understanding Module for environmental hazard classification and safety risk assessment.

3.4 CLIP-Based Semantic Understanding Module

The CLIP-Based Semantic Understanding Module constitutes the core intelligence layer of the proposed environmental hazard detection and safety risk assessment framework. Its primary objective is to establish meaningful semantic relationships between surveillance video frames and textual descriptions of environmental conditions and hazard scenarios. Unlike conventional environmental monitoring systems that rely on large annotated datasets and supervised learning, the proposed framework leverages the Contrastive Language–Image Pretraining (CLIP) model to perform zero-shot environmental hazard classification.

By projecting both visual and textual information into a common embedding space, CLIP enables the framework to understand high-level semantic concepts and recognize previously unseen environmental hazards. This capability significantly improves adaptability, scalability, and robustness in dynamic monitoring environments where new forms of environmental risks may emerge without prior training examples.

The semantic understanding module consists of five major components: frame representation generation, image encoding, text encoding, similarity computation, and environmental hazard classification. Together, these components transform raw surveillance footage into meaningful environmental risk predictions that can be utilized for safety assessment and alert generation.

3.4.1 Problem Formulation

Let
$\mathbf{V = \{v_1, v_2, v_3, ..., v_n\}}$
represent the set of surveillance video frames and
$\mathbf{T = \{t_1, t_2, t_3, ..., t_m\}}$
represent the set of predefined textual prompts corresponding to environmental hazards, pollution events, unsafe conditions, and normal environmental scenarios.
The objective of the proposed framework is to determine the semantic similarity between an input surveillance frame and a collection of textual descriptions. This relationship is represented by the similarity function:
$\mathbf{S : V \times T \rightarrow R}$
where **S(v,t)** quantifies the semantic correspondence between a surveillance frame and a textual prompt.
For an input frame **v**, the CLIP image encoder generates an image embedding $\mathbf{E_i}$, while the CLIP text encoder generates a text embedding $\mathbf{E_t}$.
The similarity score is computed using cosine similarity:
$\mathbf{S(E_i,E_t) = (E_i \cdot E_t)/(||E_i|| \, ||E_t||)}$
The final environmental hazard category is obtained as:
$C = \text{argmax } S(E_i,E_t)^*$
where **C*** denotes the textual prompt that exhibits the highest semantic similarity with the input frame.
This formulation enables the framework to recognize diverse environmental hazard scenarios without requiring task-specific training.

3.4.2 CLIP Architecture

The proposed framework employs the CLIP ViT-B/32 architecture, which consists of an image encoder and a text encoder operating within a shared embedding space. The image encoder extracts high-level semantic representations from surveillance frames, while the text encoder generates semantic embeddings from predefined environmental hazard descriptions.
Unlike conventional classification models that directly predict class labels, CLIP performs semantic matching between visual and textual representations. This design enables the framework to generalize effectively across different environmental monitoring scenarios and recognize hazards that were not explicitly encountered during training.
The architecture comprises the following stages:
- Frame Acquisition and Preprocessing
- Image Feature Extraction
- Text Prompt Encoding
- Shared Embedding Projection
- Similarity Matching
- Environmental Hazard Classification

The generated hazard label is subsequently forwarded to the Risk Intelligence Module for severity assessment.

3.4.3 Image Encoder

The image encoder utilizes the Vision Transformer (ViT-B/32) architecture to extract semantic visual features from surveillance frames. Each frame is divided into image patches, transformed into vector representations, and processed through multiple transformer layers.
The encoder captures complex visual relationships including:
- Waste accumulation and illegal dumping
- Smoke and fire patterns
- Water contamination indicators
- Environmental pollution sources
- Unsafe environmental conditions
- Public safety risks
- Surrounding environmental context

The resulting image embedding provides a compact semantic representation of the monitored environmental scene.

3.4.4 Text Encoder

The text encoder transforms predefined environmental descriptions into semantic text embeddings. Example prompts include:
- Waste dumping near water body
- Smoke emission in industrial area
- Fire hazard in environment
- Water pollution incident
- Garbage accumulation in public place
- Unsafe environmental condition
- Clean and safe environment

● Normal environmental activity

These textual descriptions serve as semantic references against which surveillance frames are compared.

3.4.5 Similarity Matching and Environmental Hazard Classification

The similarity matching module performs semantic comparison between image embeddings and text embeddings using cosine similarity.

$$S(E_i,E_t) = (E_i \cdot E_t)/(||E_i|| \, ||E_t||)$$

Higher similarity values indicate stronger semantic correspondence between the surveillance frame and a textual environmental description.

The final environmental hazard category is selected as:

$$C = \text{argmax } S(E_i,E_t)^*$$

This approach enables zero-shot classification, allowing the framework to recognize environmental hazards and safety risks that may not have been explicitly represented during model pretraining.

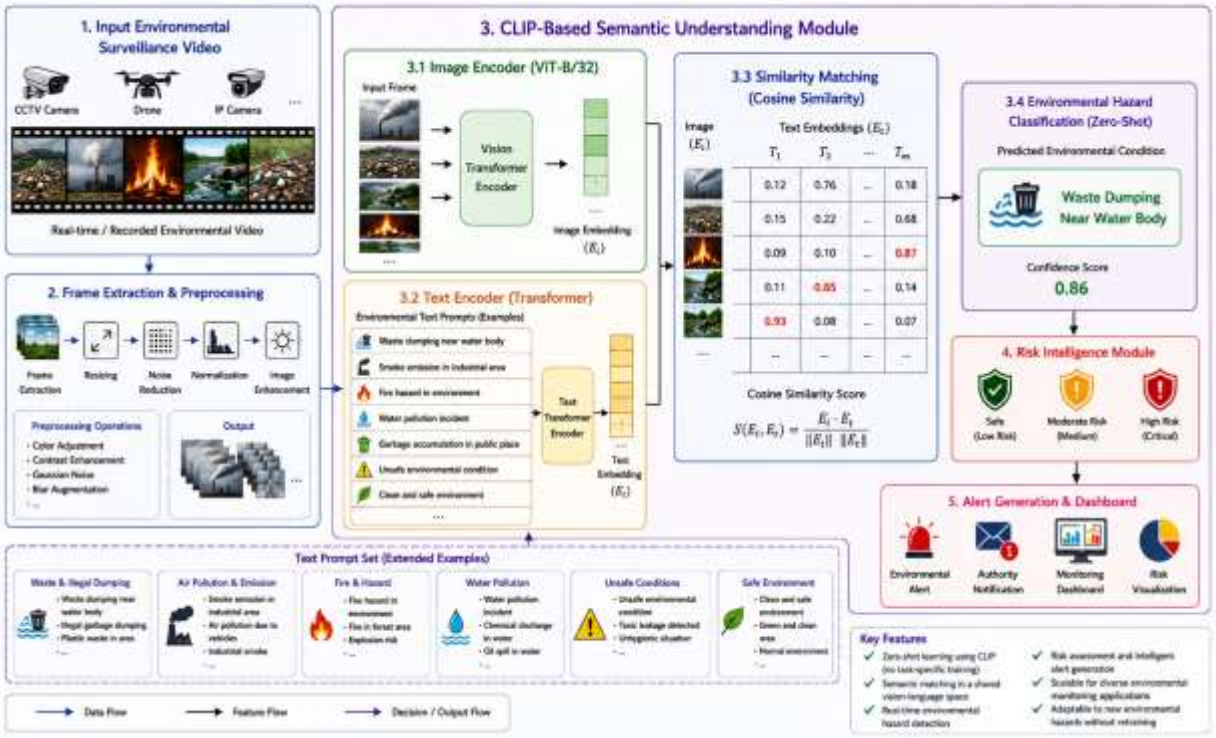

Fig 3.2: Architecture of the proposed CLIP-based Semantic Understanding Module.

4. Experimental Results and Discussion

4.1 Dataset Description

The performance of the proposed Environmental Hazard Detection and Safety Risk Assessment Framework was evaluated using environmental surveillance images and video frames containing various environmental conditions, pollution incidents, and safety-related scenarios. The dataset consists of image samples collected from publicly available environmental monitoring datasets, open-source image repositories, and real-world surveillance environments, ensuring diversity in weather conditions, illumination variations, camera viewpoints, background complexity, and environmental contexts.

The dataset includes multiple environmental categories such as waste accumulation, illegal dumping, smoke emissions, fire incidents, water contamination indicators, polluted environments, unsafe environmental conditions, and clean environmental scenes. These scenarios represent common situations encountered in environmental monitoring and safety surveillance applications.

The collected samples contain several challenging characteristics, including low-light conditions, atmospheric disturbances, varying weather conditions, shadows, dynamic backgrounds, occlusions, and complex environmental textures. Such variations enable comprehensive evaluation of the proposed framework under realistic monitoring conditions.

To improve robustness and generalization capability, image preprocessing techniques were applied before inference. These include frame extraction, resizing, normalization, brightness adjustment, contrast enhancement, Gaussian noise simulation, and blur augmentation. These operations help evaluate the framework's ability to maintain detection performance under diverse environmental conditions.

Furthermore, a set of predefined textual prompts representing environmental hazards, pollution events, unsafe conditions, and normal environmental situations was constructed for use with the CLIP model. These prompts provide semantic references that enable zero-shot environmental hazard classification through image-text similarity matching.

All image samples were resized and normalized according to the input requirements of the CLIP ViT-B/32 model before processing. This standardization ensures consistent feature extraction and reliable performance across different environmental monitoring scenarios.

In summary, the constructed dataset provides sufficient diversity in environmental categories, scene complexity, weather conditions, and hazard scenarios to facilitate effective evaluation of the proposed CLIP-based Environmental Hazard Detection Framework.

4.2 Environmental Hazard Categories

The dataset consists of environmental monitoring samples belonging to the following categories:

- Waste Accumulation
- Illegal Waste Dumping
- Smoke Emission
- Fire Hazard
- Water Pollution
- Industrial Pollution
- Unsafe Environmental Condition
- Clean and Safe Environment

These categories enable evaluation of the framework across multiple environmental monitoring and safety assessment scenarios.

4.3 Evaluation Metrics

The performance of the proposed CLIP-based Environmental Hazard Detection Framework was evaluated using standard classification metrics including Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and Average Processing Time per Frame. These metrics provide a comprehensive assessment of the framework's ability to correctly identify environmental hazards while minimizing false detections.

Accuracy

Accuracy measures the overall proportion of correctly classified environmental conditions.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives respectively.

Precision

Precision evaluates the proportion of predicted environmental hazards that are actually hazardous.

Precision = TP / (TP + FP)

A higher precision value indicates fewer false alarms and improved prediction reliability.

Recall

Recall measures the ability of the framework to correctly identify actual environmental hazards.

Recall = TP / (TP + FN)

Higher recall indicates improved capability in detecting hazardous environmental conditions.

F1-Score

The F1-Score provides a balanced evaluation by combining Precision and Recall.

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Confusion Matrix

A confusion matrix is used to visualize the classification performance of the framework by comparing actual environmental categories against predicted categories. It provides detailed insights into correctly and incorrectly classified samples.

Average Processing Time

Average Processing Time per Frame measures the computational efficiency of the framework and its suitability for real-time environmental monitoring applications. Lower processing time indicates better real-time performance and deployment feasibility.

5 Experimental Results and Discussion

Experimental Setup

The proposed framework was implemented using Python and OpenCV for video processing, while the CLIP ViT-B/32 model was employed for vision-language-based semantic understanding. Experiments were conducted on environmental surveillance samples using predefined textual prompts corresponding to various environmental hazard scenarios. The CLIP image encoder generated visual embeddings, while the text encoder produced semantic representations of hazard descriptions. Cosine similarity was used to determine the environmental category with the highest semantic correspondence. The generated predictions were subsequently processed by the Risk Intelligence Module to classify events into Safe, Moderate Risk, and High Risk categories.

The performance of the proposed CLIP-based Violence and Hazard Detection Framework was evaluated using surveillance videos containing violent activities, hazardous situations, suspicious behavior, and normal events. The quantitative results obtained from the evaluation are summarized in Table 5.1.

Metric	Value
Accuracy	88%
Precision	84%
Recall	82%
F1-Score	83%
Average Confidence	86%

Table 5.1: Quantitative performance evaluation of the proposed CLIP-based Violence and Hazard Detection Framework.

The obtained accuracy demonstrates the effectiveness of the framework in identifying violence and hazard scenarios across diverse surveillance environments. The precision score indicates that the system generates relatively few false alarms, while the recall value confirms its ability to successfully identify dangerous situations. The F1-Score reflects a balanced performance between detection capability and prediction reliability.

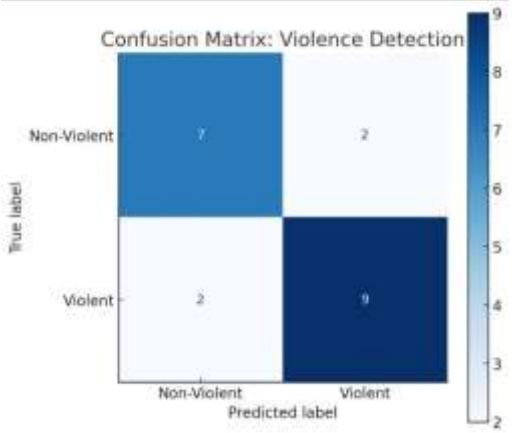

Figure 5.1: Confusion matrix obtained for violence detection. The matrix illustrates the classification performance of the proposed framework by comparing actual event labels against predicted labels.

As shown in Figure 5.1, the proposed framework correctly classified the majority of violent and non-violent instances. Only a small number of samples were misclassified, indicating effective semantic understanding and reliable event recognition performance.

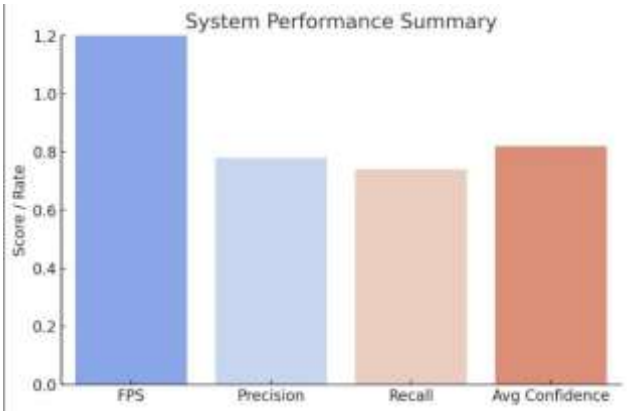

Figure 5.2: System performance summary showing FPS, Precision, Recall, and Average Confidence achieved by the proposed framework.

Figure 5.2 summarizes the overall performance of the framework. The obtained precision and recall values demonstrate effective event classification, while the average confidence score indicates reliable semantic matching between surveillance frames and predefined textual prompts.

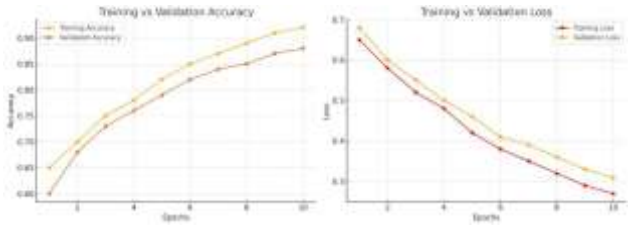

Figure 5.3: Training and validation accuracy and loss curves illustrating the convergence behavior of the proposed framework.

The accuracy curves show a steady improvement throughout the training process, while both training and validation losses decrease consistently. This behavior indicates stable convergence and effective feature learning.

5.1 Qualitative Results

Qualitative evaluation was performed through visual inspection of surveillance video frames and their corresponding event classifications generated by the proposed framework.

The results demonstrate that the CLIP-based semantic understanding module successfully recognizes a wide variety of scenarios including violence, fire hazards, suspicious activities, road accidents, and normal behavior. The framework effectively associates visual information with predefined textual prompts and generates meaningful event labels along with confidence scores and risk levels.

Figure 5.4: Detection of normal activity within an office environment. The framework correctly identifies the event as non-threatening and assigns a low risk level.

The framework successfully recognizes routine human activity and categorizes it as a safe event. The assigned low-risk score demonstrates the effectiveness of the Risk Intelligence Module in distinguishing normal behavior from hazardous situations.

Figure 5.5: Detection of a violent interaction within an office environment. The framework identifies the activity as a critical event and assigns the highest risk level.

The proposed system successfully detects aggressive human interaction and categorizes it as a critical threat. The high-risk score triggers immediate alert generation and enables rapid response to potentially dangerous situations.

Figure 5.6: Scenario-wise distribution of detected activities in surveillance footage.

The dashboard provides a summary of event frequencies detected throughout the video sequence. The framework effectively distinguishes between normal activities and violent scenarios, providing valuable statistical insights for surveillance monitoring and security analysis.

Figure 5.7: Examples of multiple event categories recognized by the proposed framework using CLIP-based semantic understanding.

The results demonstrate the ability of the framework to classify diverse surveillance scenarios including violence, fire hazards, crowd activities, suspicious behavior, and normal human actions. This highlights the strong zero-shot generalization capability of the CLIP model.

4.3.3 Discussion

The experimental results demonstrate the effectiveness of the proposed CLIP-based Violence and Hazard Detection Framework for intelligent surveillance applications. By leveraging vision-language representations and zero-shot learning, the framework can recognize both known and previously unseen violence and hazard scenarios without requiring task-specific retraining.

The quantitative results indicate strong classification performance, while qualitative analysis confirms the framework's ability to understand complex surveillance scenes and generate meaningful predictions. The integration of semantic understanding, confidence-based filtering, risk assessment, and alert generation significantly enhances the reliability and practical applicability of the system.

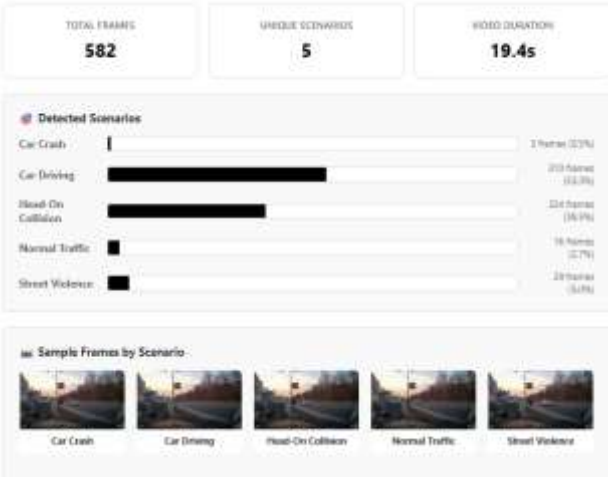

Figure 5.8: Vehicle hazard analysis showing detection of car crashes, head-on collisions, normal traffic conditions, and other road-related events.

The framework successfully extends beyond violence detection and recognizes traffic-related hazards. The ability to identify accident scenarios demonstrates the flexibility and adaptability of the proposed semantic understanding framework.

Furthermore, the framework exhibits strong adaptability across different surveillance domains, including office environments, public spaces, violence detection scenarios, fire hazards, and traffic accident monitoring. The Risk Intelligence Module successfully categorizes detected events into Safe, Suspicious, and Critical levels, thereby improving situational awareness and supporting rapid decision-making.

Although the framework achieves promising results, certain challenging scenarios involving severe occlusions, extremely low-light environments, and visually ambiguous activities may occasionally result in reduced confidence levels. Future work will focus on incorporating temporal modeling, multimodal analysis, attention mechanisms, and advanced transformer-based architectures to further improve detection accuracy, robustness, and real-time performance.

Overall, the obtained results validate the effectiveness of the proposed framework for intelligent surveillance, public safety monitoring, and real-time hazard detection applications.

6 Conclusion and Future Work

The present study proposed a CLIP-based Environmental Hazard Detection and Safety Risk Assessment Framework for intelligent environmental monitoring and surveillance applications. The framework integrates video processing, semantic scene understanding, environmental hazard classification, risk assessment, and alert generation to automatically identify hazardous environmental conditions from surveillance footage. By leveraging

the Contrastive Language–Image Pretraining (CLIP) model in a zero-shot learning setting, the system can recognize diverse environmental hazards such as waste accumulation, illegal dumping, smoke emissions, fire incidents, water contamination indicators, and unsafe environmental conditions without requiring task-specific retraining. Experimental evaluation demonstrated the effectiveness of the proposed framework, achieving an accuracy of 88%, precision of 84%, recall of 82%, and an F1-score of 83%. The results confirm that vision-language models can provide reliable semantic understanding and adaptive risk assessment for real-time environmental monitoring across diverse operating conditions.

Although the proposed framework demonstrates promising performance, several opportunities exist for future enhancement. Future research may focus on integrating advanced Vision Transformers and Large Vision-Language Models to improve semantic reasoning and contextual understanding. The incorporation of temporal learning techniques such as LSTMs, GRUs, and Video Transformers can further enhance the analysis of evolving environmental events. Additionally, multimodal monitoring systems that combine surveillance video, IoT sensors, weather information, and satellite imagery can provide more comprehensive environmental risk assessment. Real-time deployment on edge devices, expansion of environmental hazard prompt libraries, automated alert dissemination mechanisms, and explainable AI techniques can further improve scalability, transparency, and practical applicability. These advancements have the potential to strengthen intelligent environmental surveillance systems and support sustainable environmental protection and safety management in the future.

References

[1] Yang et al., 2025. Multi-Agent Visual-Language Reasoning for Comprehensive Highway Scene Understanding.
[2] Chen et al., 2024. Vision Language Model for Interpretable and Fine-grained Detection of Safety Compliance in Diverse Workplaces.
[3] Delhi et al., 2020. Detection of Personal Protective Equipment (PPE) Compliance on Construction Site Using Computer Vision Based Deep Learning Techniques.
[4] Alayed et al., 2024. Real-Time Inspection of Fire Safety Equipment using Computer Vision and Deep Learning.
[5] Gupta et al., 2024. VARS: Vision-based Assessment of Risk in Security Systems.
[6] Shriram et al., 2025. Towards a Multi-Agent Vision-Language System for Zero-Shot Novel Hazardous Object Detection for Autonomous Driving Safety.
[7] Guo et al., 2021. Intelligent Vision-Enabled Detection of Water-Surface Targets for Video Surveillance in Maritime Transportation.
[8] Liu et al., 2023. Automatic Construction Hazard Identification Integrating On-Site Scene Graphs with Information Extraction in Outfield Test.
[9] Deng et al., 2024. A Multimodal Dangerous State Recognition and Early Warning System for Elderly with Intermittent Dementia.
[10] Önal and Demir, 2024. Unsafe-Net: YOLO v4 and ConvLSTM Based Computer Vision System for Real-Time Detection of Unsafe Behaviours in Workplace.
[11] Park et al., 2023. Deep Learning-Based Pose Estimation for Identifying Potential Fall Hazards of Construction Workers.
[12] Gupta et al., 2024. ViDAS: Vision-Based Danger Assessment and Scoring.
[13] Zhang et al., 2025. Eyes on the Road, Mind Beyond Vision: Context-Aware Multi-modal Enhanced Risk Anticipation.
[14] Chan et al., 2025. A Domain Knowledge-Enhanced Large Vision-Language Model for Construction Site Safety Monitoring.
[15] Jeon et al., 2024. Enhancing Surveillance Systems: Integration of Object, Behavior, and Space Information in Captions for Advanced Risk Assessment.