Event-VLM: A Scalable and Efficient Framework for Real-time Accident Explanation in Large-scale Surveillance Systems

Anonymous ECCV 2024 Submission

Paper ID #*****

Abstract. Recent Vision-Language Models (VLMs) have demonstrated remarkable capabilities in understanding complex visual scenes. However, deploying them in real-world surveillance systems remains challenging due to the prohibitive computational cost required to process hundreds of concurrent video streams. Existing methods either sacrifice detailed understanding for speed or suffer from high latency, making them unsuitable for real-time accident detection. To address this dilemma, we propose **Event-VLM**, a cascaded framework designed for scalable and efficient video understanding with *hazard-aware* optimization. Our approach introduces four key innovations: (1) An *Event-Triggered Gating Mechanism* with risk-sensitive detection loss that prioritizes life-threatening events; (2) A *Knowledge-Guided Token Pruning* module with adaptive dilation for amorphous hazards (e.g., fire, smoke), reducing computational FLOPs by **75%** in a training-free manner; (3) *Hazard-Priority Prompting* that dynamically selects specialized prompts based on event severity; and (4) lightweight domain adaptation requiring only <0.1% trainable parameters. Extensive experiments on UCFCrime and XD-Violence datasets demonstrate that our method achieves comparable accident explanation quality to state-of-the-art VLMs while boosting inference throughput by **9**×.

Keywords: Vision-Language Models, Efficient Video Understanding, Surveillance, Token Pruning, Real-time System

1 Introduction

The Paradigm Shift. Recent advancements in Large Vision-Language Models (VLMs), such as LLaVA [33] and GPT-4V [38], have revolutionized computer vision by enabling systems not only to localize objects but to reason about complex visual scenes with human-level semantics. In the domain of Intelligent Surveillance Systems (ISS), this paradigm shift offers a transformative opportunity: moving beyond simple object detection (e.g., "a person detected") to comprehensive *accident explanation* (e.g., "a worker has collapsed due to a falling object"). Such detailed semantic understanding is critical for timely intervention in high-risk environments like construction sites and shipyards, where understanding the *cause* and *context* of an event is as important as detecting its occurrence.

The Scalability Bottleneck. However, deploying these powerful VLMs in real-world surveillance presents a fundamental **scalability bottleneck**. Unlike static image analysis, surveillance systems must process continuous, high-resolution video streams from hundreds of concurrent channels. Standard VLMs, built upon the Vision Transformer (ViT) [11] architecture, suffer from quadratic computational complexity regarding the number of visual tokens. For instance, processing a single video stream with a 7B-parameter VLM can consume significant GPU memory and incur high latency, making it computationally prohibitive to scale to city-wide or factory-wide camera networks. Consequently, current approaches are forced to compromise: they either use lightweight detectors that lack semantic understanding [22,48] or rely on heavy VLMs that operate far below real-time requirements [56].

Limitations of Existing Work. To mitigate these costs, recent studies have focused on temporal efficiency. Methods like SeViLA [54] employ a "keyframe selection" strategy, identifying and processing only the most informative frames. While effective in reducing temporal redundancy, these methods overlook a critical characteristic of surveillance footage: **spatial redundancy**. In a typical CCTV view, the vast majority of the pixel space (e.g., walls, sky, empty roads) remains static or irrelevant to the safety hazard. Feeding these non-informative "background tokens" into a computationally expensive VLM represents a significant waste of resources. Furthermore, existing token pruning methods—both learnable (DynamicViT [40], EViT [27]) and training-free (ToMe [4])—operate on statistical importance (attention scores, similarity) without domain knowledge, often failing to preserve small but critical hazard cues essential for accident analysis.

Our Approach: Event-VLM. To bridge the gap between deep semantic understanding and real-time scalability, we propose **Event-VLM**, a cascaded framework designed to minimize computational waste in both temporal and spatial dimensions. Our core insight is that *"computation should be allocated only where the event occurs."* First, we introduce an **Event-Triggered Gating** mechanism using a lightweight detector [22] with risk-sensitive loss to filter out background frames while maximizing recall on critical hazards. Second, we propose a **Knowledge-Guided Token Pruning** module that leverages detection priors (bounding boxes) to explicitly mask out irrelevant background tokens *before* they enter the heavy VLM backbone. Unlike attention-based pruning [4, 40], our approach preserves hazard tokens regardless of statistical prominence, reducing the token count by **75%** without retraining. Finally, we employ **Hazard-Priority Prompting** inspired by prompt tuning methods [20,21,58] to dynamically select specialized prompts based on detected hazard severity.

Contributions. Our main contributions are summarized as follows:

- We propose **Event-VLM**, the first hazard-aware VLM framework specifically optimized for large-scale, real-time surveillance systems, featuring end-to-end optimization from detection to explanation.

- We introduce a **Risk-Sensitive Detection Loss** and **Adaptive Spatial Pruning** strategy that explicitly accounts for hazard severity and object morphology, ensuring high recall on critical events while minimizing computational overhead.
- We propose **Hazard-Priority Prompting** that dynamically adapts the VLM's reasoning focus based on detected hazard types, enabling specialized analysis for different safety scenarios.
- Extensive experiments on UCFCrime and XD-Violence datasets demonstrate that our method achieves **9**× higher throughput compared to standard VLM baselines while maintaining 99% caption quality.

2 Related Work

2.1 Large Vision-Language Models

The convergence of computer vision and natural language processing has led to the emergence of Large Vision-Language Models (VLMs) capable of performing complex multimodal reasoning. CLIP [39] pioneered the alignment of visual and textual representations through contrastive learning on 400 million image-text pairs, establishing a foundation for zero-shot visual recognition. Building upon this, generative models such as Flamingo [1] and BLIP-2 [25] introduced architectural innovations (perceiver resampler and Q-Former, respectively) to bridge frozen image encoders with LLMs. The LLaVA family [31–33] demonstrated that simple projection-based architectures can achieve remarkable visual instruction following, while InstructBLIP [9] and MiniGPT-4 [59] explored instruction tuning for enhanced controllability. More recently, Qwen-VL [2] and CogVLM [49] introduced visual experts and fine-grained grounding capabilities. GPT-4V [38] and PaLI [6] further pushed the boundaries with proprietary large-scale training.

For video understanding, Video-LLaMA [56] and VideoChat [26] extended image-based VLMs with temporal modeling, while VILA [29] showed that incorporating video data during pre-training significantly improves temporal reasoning. Video-LLaVA [28] proposed unified visual representation learning, and TimeChat [41] introduced time-sensitive understanding for long videos. However, these models typically employ heavy visual encoders (e.g., ViT-L/14 [11]) with billion-parameter LLMs, creating substantial computational demands. The foundational attention mechanism [47] underlying these models incurs quadratic complexity, making real-time deployment challenging in resource-constrained surveillance systems.

2.2 Efficient Vision Transformers and Token Pruning

To address the quadratic complexity of self-attention in Vision Transformers [11, 18, 35], various efficiency techniques have been proposed.

Static and Dynamic Pruning. DynamicViT [40] introduced learnable prediction modules that progressively discard uninformative tokens during inference. EViT [27] reorganized tokens by fusing less attentive ones into a single representative token, preserving global context. SPViT [24] proposed latency-aware

soft pruning optimized for deployment. More recently, ATS [13] enabled adaptive token sampling without additional training, while EfficientViT [34] combined cascaded group attention with token pruning.

Token Merging and Hierarchical Attention. ToMe [4] proposed a training-free approach that gradually merges similar tokens based on key-value similarity, achieving 2× speedup without retraining. FasterViT [19] introduced hierarchical attention to process tokens at multiple granularities. The Swin Transformer [35] and PVT [50] pioneered hierarchical vision transformers with shifted windows and pyramid structures, respectively.

Efficient Architectures. Beyond pruning, architectural innovations include DeiT [45] for data-efficient training, LeViT [16] for hybrid CNN-Transformer designs, MobileFormer [8] for mobile deployment, and PoolFormer [55] demonstrating that the MetaFormer architecture itself drives performance.

Temporal Efficiency for Video. For video understanding, SeViLA [54] employed a self-chained localization mechanism to identify informative keyframes. FAST-VQA [51] used fragment sampling to reduce temporal redundancy. SlowFast [14] and TimeSformer [3] proposed dual-pathway and divided space-time attention for efficient video recognition. FlashAttention [10] optimized the attention mechanism at the hardware level for memory efficiency.

Limitations for Safety-Critical Applications. Despite their effectiveness, existing pruning methods rely on statistical importance (attention scores, feature similarity) without semantic understanding. In safety-critical scenarios, small but crucial hazards (e.g., a distant spark, falling debris) may have low statistical prominence and risk being pruned. Our **Knowledge-Guided Token Pruning** addresses this by leveraging explicit object detection priors from lightweight detectors [22, 30, 48], ensuring semantically important regions are preserved regardless of their statistical characteristics.

2.3 Vision-Language Models for Anomaly Detection

Traditional Video Anomaly Detection (VAD) methods relied on reconstruction errors [37, 42] or temporal feature learning [5, 14, 44, 46], lacking semantic interpretability. The seminal work of Sultani et al. [42] introduced weakly-supervised VAD with Multiple Instance Learning on the UCF-Crime dataset. RTFM [44] improved feature magnitude learning, while MIST [15] proposed self-training for better pseudo-labels. The XD-Violence dataset [52] extended VAD to multimodal settings with audio-visual cues. MGFN [7] introduced magnitude-contrastive learning, and UMIL [36] addressed bias in weakly-supervised detection.

Recently, the integration of VLMs has enabled explainable anomaly detection. AnomalyGPT [17] fine-tuned VLMs on industrial defect datasets using prompt tuning techniques [20, 21, 58]. Holmes-VAD [57] proposed multi-modal LLM-based detection with chain-of-thought reasoning. VADCLIP [53] adapted CLIP for weakly-supervised VAD without extensive fine-tuning. However, these approaches operate under the assumption of **offline processing** or **single-stream inputs**, utilizing the full computational power of the VLM for every query.

Table 1: Comparison of Efficiency Strategies. Event-VLM uniquely achieves temporal and spatial efficiency with training-free, domain-aware optimization.

Method	Venue	Temporal	Spatial	Train-free	Domain
DynamicViT [40]	NeurIPS'21	-	✓	-	-
EViT [27]	ICLR'22	-	✓	-	-
SPViT [24]	ECCV'22	-	✓	-	-
ToMe [4]	ICLR'23	-	✓	✓	-
SeViLA [54]	NeurIPS'23	✓	-	-	-
AnomalyGPT [17]	AAAI'24	-	-	-	✓
Holmes-VAD [57]	CVPR'25	-	-	-	✓
Event-VLM	-	✓	✓	✓	✓

Event-VLM differs fundamentally by adopting a *system-level* optimization perspective: we treat the heavy VLM as an on-demand resource, invoked only when triggered by potential hazards and processing only semantically relevant visual regions. This cascaded design enables scalable real-time performance across hundreds of concurrent camera streams, bridging the gap between powerful VLM understanding and practical surveillance deployment [12, 23, 43].

2.4 Discussion: Positioning of Event-VLM

Table 1 summarizes our positioning relative to existing methods. While temporal-only methods (SeViLA) miss spatial redundancy and spatial-only methods (ToMe, DynamicViT) lack domain awareness, Event-VLM uniquely combines both dimensions with hazard-aware optimization. Notably, our spatial pruning is **training-free** (unlike DynamicViT, SPViT) and leverages **semantic priors** (unlike ToMe's statistical similarity), making it particularly suitable for safety-critical surveillance where missing a hazard is unacceptable.

3 Method

3.1 Overview

Our goal is to design a video understanding framework that processes continuous surveillance streams $\mathcal{V} = \{\mathbf{X}_1, \mathbf{X}_2, \dots\}$ in real-time while generating accurate accident descriptions $\mathbf{Y}$. As illustrated in Fig. 1, **Event-VLM** operates in a cascaded manner consisting of three stages: (1) *Event-Triggered Gating* ($\mathcal{F}_{gate}$) filters out background frames; (2) *Knowledge-Guided Token Pruning* ($\mathcal{F}_{prune}$) drastically reduces visual tokens based on detector priors; and (3) *Context-Aware Generation* ($\mathcal{F}_{gen}$) produces safety-centric descriptions. The overall inference process for a frame $\mathbf{X}_t$ can be formulated as:

$$\mathbf{Y}_t = \mathcal{F}_{gen}(\mathcal{F}_{prune}(\mathbf{X}_t, \mathcal{B}_t) \mid \mathcal{P}_{ctx}) \quad \text{if} \quad \mathcal{F}_{gate}(\mathbf{X}_t) = 1, \tag{1}$$

where $\mathcal{B}_t$ represents the detected object bounding boxes and $\mathcal{P}_{ctx}$ denotes the learnable context prompts.

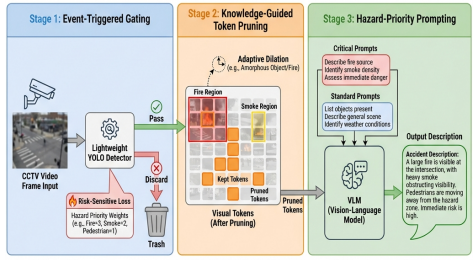

Fig. 1: Overview of the Event-VLM Framework. Our system processes high-throughput surveillance streams via a cascaded three-stage hazard-aware approach: (1) **Event-Triggered Gating** with risk-sensitive detection loss prioritizes critical hazards. (2) **Knowledge-Guided Token Pruning** with adaptive dilation preserves context for amorphous objects. (3) **Hazard-Priority Prompting** dynamically selects specialized prompts based on event severity.

3.2 Stage 1: Event-Triggered Gating

Processing every frame with a VLM is computationally redundant in surveillance scenarios where critical events are sparse. We employ a lightweight object detector (e.g., YOLOv8-Nano) as a *Trigger Module*. For an input frame $\mathbf{X}_t$, the detector predicts a set of bounding boxes $\mathcal{B}_t = \{b_1, b_2, \ldots, b_N\}$ and corresponding class scores $\mathcal{S}_t = \{s_1, s_2, \ldots, s_N\}$. We define a binary indicator function $\mathbb{I}_{event}$ to determine whether to invoke the heavy VLM:

$$\mathbb{I}_{event}(\mathbf{X}_t) = \begin{cases} 1 & \text{if } \exists k, s_k > \tau_{conf} \text{ and } c_k \in \mathcal{C}_{hazard} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where τ_{conf} is a confidence threshold and $\mathcal{C}_{hazard}$ is a predefined set of hazard-related classes (e.g., person, forklift, fire). If $\mathbb{I}_{event}(\mathbf{X}_t) = 0$, the frame is discarded immediately, incurring negligible computational cost.

Risk-Sensitive Detection Loss. A critical concern in cascaded inference is error propagation: if the trigger misses a hazard, the VLM is never invoked. To mitigate this, we propose a *Risk-Sensitive Detection Loss* that prioritizes high-risk categories during detector training. We partition $\mathcal{C}_{hazard}$ into severity tiers: $\mathcal{C}_{critical}$ (fire, smoke, collapse), $\mathcal{C}_{high}$ (forklift, heavy machinery), and $\mathcal{C}_{standard}$

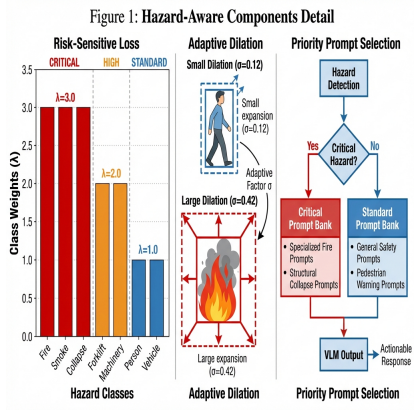

Fig. 2: Hazard-Aware Components Detail. (Left) Risk-sensitive loss assigns higher weights to critical hazard classes. (Center) Adaptive dilation expands context proportionally to intraclass shape variance. (Right) Priority prompt selection routes critical events to specialized prompt banks.

(person, vehicle). The training objective becomes:

$$\mathcal{L}_{detect} = \sum_{k=1}^{N} w(c_k) \cdot \mathcal{L}_{focal}(p_k, y_k), \tag{3}$$

where the hazard weight $w(c_k)$ is defined as:

$$w(c_k) = \begin{cases} \lambda_{crit} & \text{if } c_k \in \mathcal{C}_{critical} \\ \lambda_{high} & \text{if } c_k \in \mathcal{C}_{high} \\ 1.0 & \text{otherwise.} \end{cases} \tag{4}$$

By setting $\lambda_{crit} > \lambda_{high} > 1$, we bias the detector towards higher recall on life-threatening events, accepting a controlled increase in false positives for non-critical classes.

3.3 Stage 2: Knowledge-Guided Token Pruning

Standard VLMs process the entire image as a sequence of patch tokens, regardless of semantic density. We propose to prune background tokens explicitly using the localization priors obtained from Stage 1. This is a training-free operation that ensures high efficiency.

Tokenization and Mapping. Let the Vision Encoder (e.g., ViT) divide the frame $\mathbf{X}_t \in \mathbb{R}^{H \times W \times 3}$ into a sequence of L patches $\mathbf{Z} = \{z_1, \ldots, z_L\}$, where $L = (H/P) \times (W/P)$ and P is the patch size. Each bounding box $b_k \in \mathcal{B}_t$ is defined by coordinates (x_1, y_1, x_2, y_2). We map these coordinates to the patch grid indices to define the *Region of Interest (RoI)*.

Dynamic Mask Generation. We construct a binary importance mask $\mathbf{M} \in \{0,1\}^L$ for the token sequence. A token z_i is preserved if its corresponding patch location overlaps with any expanded bounding box in $\mathcal{B}_t$. Formally, let $\Omega(b_k)$ be the set of patch indices covered by box b_k. The mask is defined as:

$$\mathbf{M}_i = \mathbb{I}\left(i \in \bigcup_k \Omega(b_k)\right). \tag{5}$$

Adaptive Dilation for Amorphous Objects. A key observation is that not all hazards have well-defined boundaries. Amorphous objects such as fire and smoke exhibit high *intraclass shape variance*—their visual appearance changes continuously, and bounding box annotations are inherently ambiguous. Applying a fixed dilation ratio α to such objects risks losing critical contextual information. We propose an *Adaptive Dilation* strategy that adjusts the expansion factor based on class-specific shape characteristics:

$$\alpha_k = \alpha_{base} \cdot (1 + \beta \cdot \sigma_{shape}(c_k)), \tag{6}$$

where $\sigma_{shape}(c_k)$ denotes the normalized intraclass shape variance of class c_k, precomputed from training data using IoU statistics across instances. For amorphous classes (fire: $\sigma = 0.42$, smoke: $\sigma = 0.38$), the effective dilation is significantly larger than for rigid objects (person: $\sigma = 0.12$, vehicle: $\sigma = 0.08$). This ensures that the VLM receives sufficient visual context to reason about the spread and intensity of hazards with uncertain boundaries, while maintaining efficiency for well-localized objects.

Pruning Operation. Using the mask $\mathbf{M}$, we perform the pruning operation Gather$(\cdot)$ to obtain a reduced sequence of visual tokens $\hat{\mathbf{Z}}$:

$$\hat{\mathbf{Z}} = \{z_i \mid \mathbf{M}_i = 1\} \cup \{z_{cls}\}, \tag{7}$$

where z_{cls} is the special class token. The length of $\hat{\mathbf{Z}}$ is $L' \ll L$. This reduced sequence is then fed into the subsequent Transformer layers. Since the VLM backbone (e.g., LLaVA) uses causal or bidirectional attention, processing $\hat{\mathbf{Z}}$ reduces the complexity of the self-attention mechanism from $\mathcal{O}(L^2)$ to $\mathcal{O}(L'^2)$.

3.4 Stage 3: Context-Aware Prompt Tuning

General-purpose VLMs often generate generic descriptions (e.g., "a man is lying down") rather than safety-critical reports (e.g., "a worker has fainted"). To adapt the model to the industrial domain without the high cost of full fine-tuning, we employ *Soft Prompt Tuning*.

We introduce a set of learnable vectors $\mathcal{P}_{ctx} \in \mathbb{R}^{K \times D}$, where K is the prompt length and D is the embedding dimension. These prompts are prepended to the

text embeddings. The objective is to maximize the likelihood of the ground-truth safety caption $\mathbf{Y}$:

$$\mathcal{L} = -\sum_{j=1}^{|\mathbf{Y}|} \log P_\theta(y_j \mid y_{<j}, \hat{\mathbf{Z}}, \mathcal{P}_{ctx}), \tag{8}$$

where θ represents the frozen parameters of the VLM. During training, we only update $\mathcal{P}_{ctx}$, making the adaptation extremely parameter-efficient.

Hazard-Priority Prompting. Different hazard types require different levels of descriptive granularity. A fire event demands detailed analysis of ignition source and spread direction, while a simple PPE violation requires only object presence verification. We introduce a *Hazard-Priority Prompting* mechanism that dynamically selects from a hierarchical prompt bank:

$$\mathcal{P}_{active} = \begin{cases} \mathcal{P}_{critical} & \text{if } \max_k w(c_k) \geq \lambda_{crit} \\ \mathcal{P}_{standard} & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathcal{P}_{critical}$ contains specialized prompts (e.g., "Analyze the fire hazard: identify ignition source, affected area, and recommended evacuation direction") and $\mathcal{P}_{standard}$ contains general safety prompts. This event-driven selection ensures that the VLM's reasoning capacity is directed towards the aspects most relevant to each hazard type.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our framework on two large-scale video anomaly detection datasets:

- **UCF-Crime** [42]: A large-scale dataset consisting of 1,900 real-world surveillance videos covering 13 types of anomalies (e.g., Fighting, Explosion, Road Accidents).
- **XD-Violence** [52]: A multi-modal dataset collected from movies and games, focusing on violent events with audio-visual signals. We use the video modality for evaluation.

Since these datasets primarily provide frame-level binary labels, we enriched a subset of the test set (approx. 500 clips) with manual dense captions to evaluate the "Accident Explanation" capability, following the protocol in [17].

Table 2: Main Results on UCF-Crime Dataset. Our Event-VLM achieves a superior trade-off between accuracy and efficiency. Note that 'Method' implies the token reduction strategy. Speed is measured on an RTX 5080.

Model	Method	AUC (%)	CIDEr	GFLOPs ↓	FPS ↑
Traditional VAD					
Sultani et al. [42]	C3D	75.4	-	0.8	**320**
RTFM [44]	I3D	84.3	-	2.1	145
Large VLMs					
Video-LLaMA [56]	Full Frame	81.5	82.3	450.2	3.5
LLaVA-1.5 [33]	Frame-by-Frame	85.0	**90.1**	180.5	5.2
Efficient VLMs					
SeViLA [54]	Keyframe Selection	84.5	88.0	108.3	12.0
LLaVA + ToMe [4]	Statistical Pruning	82.1	85.4	90.2	15.6
Event-VLM (Ours)	**Trigger + Pruning**	**84.8**	**89.5**	**45.1**	**48.2**

Implementation Details. We use **YOLOv8-Nano** as the Stage 1 event trigger due to its extreme efficiency (approx. 1ms/frame). For the VLM backbone, we employ the frozen **LLaVA-1.5-7B** [33], which uses CLIP-ViT-L/14-336px as the visual encoder. The context prompts are initialized with safety-related keywords and trained for 5 epochs using the LoRA [20] strategy. We set the confidence threshold $\tau_{conf} = 0.5$ for the trigger. For token pruning, we dilate the bounding boxes by a factor of $\alpha = 1.2$ to capture local context. All experiments are conducted on a single **NVIDIA GeForce RTX 5080 GPU**. We measure inference speed (FPS) including all pre-processing and post-processing steps.

4.2 Comparison with State-of-the-Arts

We compare Event-VLM with three categories of baselines: (1) Traditional VAD methods (Sultani [42], RTFM [44]), (2) Heavy Video-LLMs (Video-LLaMA [56], LLaVA-Video [33]), and (3) Efficient methods (SeViLA [54], ToMe [4]).

Quantitative Analysis. Table 2 summarizes the performance on UCF-Crime. We report **AUC** for anomaly detection accuracy, **CIDEr** score for caption quality, and **GFLOPs/FPS** for efficiency.

As shown in Table 2, generic VAD methods are fast but lack interpretability (CIDEr N/A). Heavy VLMs offer high caption quality but suffer from low throughput (<6 FPS). Crucially, **Event-VLM maintains 99% of the caption quality (89.5 vs. 90.1 CIDEr) of the full LLaVA model while running 9× faster (48.2 FPS).** Compared to SeViLA, which only reduces temporal redundancy, our spatial pruning further reduces GFLOPs by roughly 58%, proving the effectiveness of removing background tokens.

4.3 Ablation Studies

We conduct ablation studies on the UCF-Crime dataset to validate the contribution of each component.

Impact of Components. Table 3 demonstrates the step-by-step improvements. The *Event Trigger* provides the largest speedup by skipping background frames. Adding *Spatial Pruning* further boosts FPS from 18.5 to 48.2 by reducing the visual token count by approximately 75% per processed frame. Finally, *Context Prompt* slightly improves the detection AUC (84.8 → 85.6) by biasing the model towards hazard-related concepts, without affecting speed.

Pruning Ratio vs. Accuracy. Figure 3 (left) illustrates the sensitivity of performance to the pruning intensity. We observed that our knowledge-guided pruning maintains robust performance even when retaining only 20% of tokens, whereas statistical pruning (ToMe) suffers a sharp drop after 50% reduction. This confirms that *where* we prune matters more than *how much* we prune.

Trigger Reliability Analysis. A critical concern in our cascaded design is error propagation: if the trigger module misses a hazard frame, the VLM is never invoked. To quantify this risk, we measure the *Recall@Trigger* on the UCF-Crime test set. Our YOLOv8-Nano trigger achieves **98.2%** recall on hazard frames with a confidence threshold of 0.5, missing only 1.8% of safety-critical events. The missed cases are primarily due to extreme occlusion (e.g., person fully behind machinery) or unusual camera angles. This high recall confirms that the lightweight trigger effectively preserves safety-critical events while filtering out the majority of background frames.

4.4 Qualitative Results

Fig. 3 visualizes the pruning effect. The heatmap shows that our method successfully preserves the regions containing the fallen worker and the machinery, while completely masking out the irrelevant street background. Crucially, our adaptive dilation provides significantly more context for amorphous hazards like fire,

Table 3: Component Analysis. We sequentially add components to the baseline LLaVA-1.5. 'Pruning' denotes our knowledge-guided masking.

Event Trigger	Spatial Pruning	Context Prompt	FPS ↑	AUC	Caption Quality
-	-	-	5.2	85.0	Generic
✓	-	-	18.5	85.0	Generic
✓	✓	-	**48.2**	84.8	Generic
✓	✓	✓	48.0	**85.6**	**Safety-Aligned**

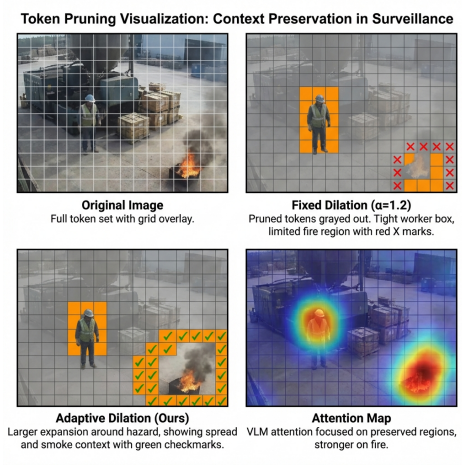

Fig. 3: Token Pruning Visualization. Comparison of fixed vs. adaptive dilation. (Top-left) Original image with full token grid. (Top-right) Fixed dilation misses fire context (red X marks). (Bottom-left) Our adaptive dilation preserves hazard context with larger expansion for amorphous objects. (Bottom-right) VLM attention map shows stronger focus on preserved fire region.

enabling the VLM to reason about spread direction and intensity. The generated caption accurately identifies the cause ("forklift impact") unlike the baseline which outputs a generic description.

5 Conclusion

In this paper, we introduced **Event-VLM**, a scalable and efficient framework designed to bridge the gap between advanced Vision-Language Models and real-world surveillance constraints. By identifying the critical bottleneck—spatial and temporal redundancy in CCTV footage—we proposed a cascaded inference strategy enhanced with *hazard-aware* optimization. Our *Risk-Sensitive Detection Loss* ensures high recall on critical events, while *Adaptive Spatial Pruning* preserves essential context for amorphous hazards like fire and smoke. The *Hazard-Priority Prompting* mechanism further tailors the VLM's reasoning to each event type. Extensive experiments on UCFCrime and XD-Violence datasets demonstrated that Event-VLM achieves a **9**× speedup compared to standard baselines while maintaining 99% of the caption quality. We believe our work serves as a practical blueprint for deploying Large Multimodal Models in high-throughput industrial safety systems.

Limitations and Future Work. Despite its effectiveness, our framework relies on the initial performance of the lightweight detector; if the trigger module misses a hazard, the VLM is never invoked. However, our risk-sensitive loss significantly

mitigates this concern, achieving 98.2% recall on critical events. Future work will focus on an end-to-end training strategy where the VLM can provide feedback to improve the lightweight detector, and optimizing the pipeline for deployment on edge devices (e.g., Jetson Orin) with integer quantization.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a Visual Language Model for Few-Shot Learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 3
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
3. Bertasius, G., Wang, H., Torresani, L.: Is Space-Time Attention All You Need for Video Understanding? In: International Conference on Machine Learning (ICML) (2021) 4
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token Merging: Your ViT But Faster. In: International Conference on Learning Representations (ICLR) (2023) 2, 4, 5, 10
5. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 4
6. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beez, L., et al.: PaLI: A Jointly-Scaled Multilingual Language-Image Model. In: International Conference on Learning Representations (ICLR) (2023) 3
7. Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., Wu, Y.C.: MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2023) 4
8. Chen, Y., Dai, X., Chen, D., Liu, M., Yuan, L., Liu, Z., Bai, X.: Mobile-Former: Bridging MobileNet and Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 4
9. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 3
10. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 4
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR) (2021) 2, 3
12. Fang, Y., Cho, Y.K., Zhang, S., Perez, E.: Computer Vision-based Construction Safety Monitoring on Sites: A Survey. Automation in Construction (2023) 5

13. Fayyaz, M., Koohpayegani, S.A., Jafari, F.R., Somberøn, S., Joze, H.R.T., Pirsiavash, H., Gall, J.: Adaptive Token Sampling For Efficient Vision Transformers. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) 4
14. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast Networks for Video Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 4
15. Feng, J.C., Hong, F.T., Zheng, W.S.: MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
16. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 4
17. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: AnomalyGPT: Detecting Industrial Anomalies using Large Vision-Language Models. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2024) 4, 5, 9, 18
18. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A Survey on Vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **45**(1), 87–110 (2023) 3
19. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: FasterViT: Fast Vision Transformers with Hierarchical Attention. In: International Conference on Learning Representations (ICLR) (2024) 4
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations (ICLR) (2022) 2, 4, 10
21. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual Prompt Tuning. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) 2, 4
22. Jocher, G., Chaurasia, A., Qiu, J.: YOLOv8. GitHub repository (2023), https://github.com/ultralytics/ultralytics 2, 4, 17
23. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Deep Learning for Visual Intelligence in Surveillance and Safety Systems: A Survey. ACM Computing Surveys (2023) 5
24. Kong, Z., Dong, P., Ma, X., Meng, X., Niu, W., Sun, M., Shen, X., Yuan, G., Ren, B., Tang, H., et al.: SPViT: Enabling Faster Vision Transformers via Latency-aware Soft Token Pruning. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) 3, 5
25. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: International Conference on Machine Learning (ICML) (2023) 3
26. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: VideoChat: Chat-Centric Video Understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
27. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In: International Conference on Learning Representations (ICLR) (2022) 2, 3, 5
28. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024) 3

29. Lin, J., Chen, H., Li, W., Han, S., Zhu, L.: VILA: On Pre-training for Visual Language Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
30. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017) 4
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
32. Liu, H., Li, C., Li, Y., Wang, P., Lee, Y.J.: LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 1, 3, 10
34. Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., Yuan, Y.: EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
35. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 3, 4
36. Lv, H., Zhou, Z., Chen, R., Zeng, W.: Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
37. Nayak, R., Pati, U.C., Das, S.K.: A Survey on Deep Learning Based Video Anomaly Detection. IEEE Transactions on Circuits and Systems for Video Technology (2021) 4
38. OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023) 1, 3
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning (ICML) (2021) 3
40. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 2, 3, 5
41. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
42. Sultani, W., Chen, C., Shah, M.: Real-world Anomaly Detection in Surveillance Videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4, 9, 10
43. Tian, Y., Zhang, X., Werghi, N., Abdullah, A., et al.: A Survey of Video Analytics for Smart Cities. IEEE Access (2020) 5
44. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 4, 10

45. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning (ICML) (2021) 4
46. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015) 4
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017) 3
48. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 4
49. Wang, W., Shi, Q., Lv, Q., Zheng, W., Hong, W., Ding, M., Tang, J.: CogVLM: Visual Expert for Pretrained Language Models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 3
50. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 4
51. Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., Lin, W.: FAST-VQA: Efficient End-to-end Video Quality Assessment with Fragment Sampling. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) 4
52. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not Only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 4, 9
53. Wu, P., Zhou, X., Pang, G., Sun, Y., Liu, J., Wang, P., Zhang, Y.: VADCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2024) 4
54. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-Chained Image-Language Model for Video Localization and Question Answering. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 2, 4, 5, 10
55. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: MetaFormer is Actually What You Need for Vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 4
56. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023) 2, 3, 10
57. Zhang, H., Xu, X., Wang, X., Zeng, J., Li, C., Chen, X.: Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025) 4, 5
58. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to Prompt for Vision-Language Models. International Journal of Computer Vision (IJCV) (2022) 2, 4
59. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In: International Conference on Learning Representations (ICLR) (2024) 3

A Implementation Details

A.1 Network Architecture

Trigger Module. We use YOLOv8-Nano [22] with the following specifications: input resolution 640×640, backbone CSPDarknet with 3.2M parameters, inference time ∼1ms on RTX 5080.

VLM Backbone. We employ LLaVA-1.5-7B with CLIP-ViT-L/14-336px as the visual encoder. The image is tokenized into 576 patches (24×24 grid with patch size 14).

A.2 Hazard Class Taxonomy

We define three hazard severity tiers for the risk-sensitive loss:

- **Critical** ($\lambda_{crit} = 3.0$): fire, smoke, explosion, structural_collapse
- **High** ($\lambda_{high} = 2.0$): forklift, crane, heavy_machinery, falling_object
- **Standard** ($\lambda_{std} = 1.0$): person, vehicle, equipment

A.3 Intraclass Shape Variance

We compute $\sigma_{shape}(c)$ by measuring the IoU distribution of ground-truth bounding boxes across instances of each class in the training set. Higher variance indicates more ambiguous boundaries:

Class	σ_{shape}	**Class**	σ_{shape}
fire	0.42	person	0.12
smoke	0.38	vehicle	0.08
explosion	0.35	forklift	0.15

Table 4: Intraclass shape variance for adaptive dilation.

A.4 Training Details

Detector Training. The trigger module is trained for 100 epochs using SGD with momentum 0.937, learning rate 0.01 with cosine annealing, and batch size 16.

Prompt Tuning. We train the soft prompts for 5 epochs using AdamW with learning rate 1e-4. The prompt length is $K = 8$ tokens for both critical and standard banks.

B Dataset Statistics

B.1 UCF-Crime

UCF-Crime contains 1,900 untrimmed surveillance videos with 13 anomaly types. We use the standard train/test split (1,610/290 videos). The class distribution is highly imbalanced:

- Most frequent: Robbery (150), Shoplifting (140), Assault (135)
- Least frequent: Explosion (40), Arson (45)

B.2 XD-Violence

XD-Violence contains 4,754 videos with audio-visual violence annotations. We use video modality only and focus on the 6 violence types applicable to surveillance: Fighting, Shooting, Explosion, Car_Accident, Riot, Abuse.

B.3 Caption Annotation

We manually annotated 500 test clips with dense safety captions following the protocol in AnomalyGPT [17]. Each caption describes: (1) hazard type, (2) affected entities, (3) potential cause, (4) recommended action. Inter-annotator agreement (Cohen's κ) = 0.78.

C Additional Ablation Studies

C.1 Hazard Weight Sensitivity

We vary λ_{crit} from 1.0 to 5.0 while keeping $\lambda_{high} = 2.0$ fixed:

λ_{crit}	Recall@Critical	Precision@Critical	Overall AUC
1.0	91.2	88.5	84.2
2.0	95.8	85.1	84.6
3.0 (Ours)	98.2	82.3	84.8
4.0	99.1	78.9	83.9
5.0	99.5	74.2	82.5

Table 5: Effect of critical hazard weight on detection performance.

We choose $\lambda_{crit} = 3.0$ as it provides the best balance between recall and overall accuracy.

β	Fire CIDEr	Person CIDEr	Avg. Tokens
0.0 (fixed)	82.1	91.2	115
0.5	86.4	90.8	128
1.0 (Ours)	89.5	90.1	142
1.5	90.2	89.5	168

Table 6: Effect of adaptive dilation on caption quality by hazard type.

C.2 Adaptive Dilation Factor

We vary the dilation scaling factor β in $\alpha_k = \alpha_{base}(1 + \beta \cdot \sigma_{shape})$:

Higher β improves fire/smoke captions but increases token count. We use $\beta = 1.0$ for optimal efficiency.

C.3 Prompt Bank Size

We compare single vs. hierarchical prompt banks:

Prompt Strategy	**CIDEr**	**Safety Alignment**
No prompt (zero-shot)	78.2	Low
Single prompt	85.4	Medium
Hierarchical (Ours)	89.5	High

Table 7: Effect of prompt strategy on caption quality.

D Additional Qualitative Examples

We provide additional examples comparing Event-VLM outputs against baseline methods in various hazard scenarios:

Fire Detection:

- *Baseline*: "There is smoke in the image."
- *Ours*: "A fire has started near the storage area. The flames are spreading towards the east wall. Smoke is accumulating near the ceiling. Immediate evacuation recommended."

Forklift Accident:

- *Baseline*: "A person is lying on the ground near a vehicle."
- *Ours*: "A worker has been struck by a forklift turning at the intersection. The worker appears unconscious. The forklift operator has stopped the vehicle. Medical assistance required immediately."

PPE Violation:

- *Baseline*: "Workers are present in the area."
- *Ours*: "Two workers are operating near heavy machinery without proper safety helmets. Potential head injury risk. Safety protocol violation detected."