CERA: Causal Event Reasoning and Attribution via Tri-Axis Compute Allocation for Real-time Surveillance Video Understanding

Anonymous Authors

Anonymous Institution
anonymous@eccv2026.org

Abstract. Real-time surveillance event reasoning is constrained by inference economics: practical systems must process many concurrent streams while preserving explanation quality and causal attribution fidelity. We identify three orthogonal redundancy axes that dominate runtime cost: *temporal* redundancy (most frames are non-events), *spatial* redundancy (most visual tokens are background), and *decoding* redundancy (autoregressive KV-cache access is memory-bandwidth bound). We present **CERA** (**C**ausal **E**vent **R**easoning and **A**ttribution), a tri-axis compute-allocation framework that decides computation *when* events occur, *where* hazards exist, and *which* cache entries matter for generation. CERA integrates: (1) *Event-Triggered Gating* with risk-sensitive detector training; (2) *Knowledge-Guided Token Pruning* with adaptive dilation for amorphous hazards; and (3) *Frequency-Aware Sparse Decoding* that uses dominant RoPE frequency chunks as a training-free proxy for online token saliency. A lightweight hazard-priority prompting stage improves safety-specific description quality with minimal latency overhead. In pilot single-seed runs on UCF-Crime and XD-Violence, CERA reaches **48.2/48.0 FPS** (Core/Full) while retaining **98.8–99.4% CIDEr** relative to a strong multimodal baseline (LLaVA-1.5). It also maintains gains above **8.5**× in internal runtime stress tests. This draft reports deterministic pilot runs; multi-seed confidence intervals and paired significance tests are scheduled for the final statistical release.

Keywords: Multimodal Reasoning, Efficient Inference, Surveillance, Token Pruning, KV Cache Optimization, Sparse Attention

1 Introduction

The Paradigm Shift. Recent VLMs shift visual understanding from recognition to explanation. Instruction-tuned models now provide grounded scene explanations [29,27,28]. Similar capabilities are reported in related systems [20,35]. In Intelligent Surveillance Systems (ISS), this shift is operationally critical because causal accident explanation enables immediate response prioritization.

The Scalability Bottleneck. Deploying these models in surveillance remains hard. Video-oriented VLM systems improve temporal reasoning [53,21,24,33,40,39,25]. Temporal backbones also continue to advance video understanding [3,12,6]. Real-time multi-stream deployment, however, is still constrained by compute and memory. Unlike offline analysis, surveillance requires persistent online inference over many cameras. We identify three orthogonal axes of computational redundancy that drive this bottleneck (Fig. 1):

- **Temporal Redundancy**: In typical surveillance footage, critical events occupy less than 1% of the total duration. Processing every frame with a heavy multimodal reasoner is wasteful.
- **Spatial Redundancy**: In a typical CCTV view, the vast majority of visual tokens (walls, sky, empty roads) are semantically irrelevant to any safety event. Feeding these tokens into costly self-attention layers is a significant waste.
- **Decoding Redundancy**: During auto-regressive text generation, the Key-Value (KV) cache grows linearly with context length. At each decoding step, the model must access the *entire* cache, creating a memory-bandwidth bottleneck that underutilizes modern GPUs [8]. This is particularly acute in large multimodal stacks, where hundreds of visual tokens inflate the KV cache.

Limitations of Existing Work. Most acceleration methods optimize only one axis and leave others as hidden bottlenecks. Temporal filtering methods (e.g., SeViLA [52]) reduce frame count but keep dense token processing and full decoding state. Spatial token compression methods (ToMe, DynamicViT, EViT, ATS, SPViT) [4,38,23,11,19] reduce vision cost but are often domain-agnostic, risking removal of small high-risk cues. Decoding acceleration methods (StreamingLLM, H2O, SnapKV, PyramidKV) [51,55,22,5] are primarily studied in text-centric long-context settings. Therefore single-axis acceleration saturates quickly in end-to-end surveillance pipelines.

Our Approach: CERA. We propose **CERA**, a cascaded framework for **causal event reasoning (CER)** under strict latency budgets. Its systems principle is simple: allocate computation only *when* an event occurs, *where* the hazard is located, and *which* memory entries are relevant for generation.

1. **Event-Triggered Gating** (Temporal) uses a lightweight detector with *risk-sensitive loss* to suppress background frames while preserving recall on critical hazards.
2. **Knowledge-Guided Token Pruning** (Spatial) converts detector priors into dynamic token masks, with *adaptive dilation* for amorphous hazards such as fire and smoke.
3. **Frequency-Aware Sparse Decoding** (Decoding) adapts RoPE chunk-level sparsity [48] to multimodal decoding, enabling training-free token saliency estimation and focused KV access during generation.

These three stages are **training-free** with respect to the multimodal backbone, requiring no backbone modification or full-model fine-tuning. We additionally include an optional, lightweight prompting module for domain adaptation.

Contributions.

- We present **CERA**, a surveillance-oriented framework for causal event reasoning and attribution that jointly optimizes temporal, spatial, and decoding efficiency in a single end-to-end pipeline.
- We introduce **risk-sensitive gating** and **adaptive spatial pruning**, which explicitly encode hazard severity and object morphology to preserve critical evidence under aggressive compute budgets.
- We adapt **frequency-aware sparse decoding** to mixed visual-textual generation, showing that dominant RoPE frequency chunks are an effective training-free proxy for KV saliency.
- On UCF-Crime and XD-Violence pilot runs, CERA delivers about **9**× throughput gains. Caption quality remains near baseline (98.8–99.4% CIDEr retention), supporting real-time multi-camera deployment in the pilot regime.

Naming Convention in This Draft. We use **CERA** for the full method name (*Causal Event Reasoning and Attribution*) and use **CER** as an internal shorthand for the underlying *causal event reasoning* objective at component level.

2 Related Work

2.1 Large Vision-Language Models

The convergence of vision and language has produced VLMs capable of complex multimodal reasoning. CLIP [37] pioneered visual-textual alignment via contrastive learning. Generative models such as Flamingo [1] and BLIP-2 [20] bridged frozen encoders with LLMs, while the LLaVA family [29,27,28] showed that simple projection architectures achieve remarkable visual instruction following. Qwen-VL [2] and CogVLM [46] introduced visual experts and grounding capabilities.

Their visual stacks largely inherit ViT-family advances [9,44,31,47] and self-supervised pretraining [16,36]. For video, Video-LLaMA [53], VideoChat [21], and VILA [25] add temporal modeling to image VLMs. Yet these systems still pair heavy visual encoders with billion-parameter LLMs, so quadratic attention [45] remains costly for real-time deployment.

2.2 Efficient Vision Transformers and Token Pruning

DynamicViT [38] introduced learnable modules for progressive token discard. EViT [23] fused less attentive tokens into representative tokens. ToMe [4] proposed training-free token merging based on key-value similarity, achieving 2×

speedup. SPViT [19] optimized pruning for latency. For video, SeViLA [52] employed keyframe selection and SlowFast [12] proposed dual-pathway architectures. FasterViT and EfficientViT [15,30] further show that architecture-level efficiency and token-level sparsification are complementary.

Despite their effectiveness, existing pruning methods rely on statistical importance without domain knowledge. In safety-critical scenarios, small hazards may have low statistical prominence and risk being pruned. Our **Knowledge-Guided Token Pruning** addresses this by leveraging detection priors [17].

2.3 Vision-Language Models for Anomaly Detection

Traditional VAD relied on reconstruction errors [42,34]. Temporal-feature approaches were also common [43,6]. Recent VLM-based approaches improve explainability. AnomalyGPT [14] fine-tunes on defect datasets. Holmes-VAD [54] and VADCLIP [50] target multimodal or weakly supervised settings. Weakly supervised lines [13,7,32] offer strong localization priors but not caption-level explanation under strict latency budgets. Safety-monitoring surveys also highlight the gap between recognition pipelines and operational explanation systems [18,10]. Most methods still focus on offline single-stream processing.

2.4 KV Cache Optimization and Sparse Attention

The KV cache is a core bottleneck in LLM inference. StreamingLLM [51] keeps a fixed-size window with "attention sinks." H2O [55] evicts tokens by cumulative attention. SnapKV [22] and PyramidKV [5] choose layer-wise entries and budgets. These methods still rely on heuristic importance signals and are not truly query-aware.

FASA [48] recently discovered that RoPE [41] induces *functional sparsity* at the frequency-chunk level: a small subset of "dominant" frequency chunks captures contextual attention patterns, while the majority encode positional structures. This provides a training-free, query-aware proxy for token importance. While FASA targets text-only LLMs, we adapt this insight to the VLM setting, where visual tokens significantly inflate the KV cache.

2.5 Positioning of CERA

Table 1 summarizes our positioning. CERA is the first to address temporal, spatial, *and* decoding efficiency simultaneously, all in a training-free manner with domain-aware optimization.

Table 1. Comparison of Efficiency Strategies. CERA uniquely addresses all three axes of redundancy with training-free, domain-aware optimization.

Method	Venue	Temporal	Spatial	Decoding	Train-free	Domain
DynamicViT [38]	NeurIPS'21	-	✓	-	-	-
ToMe [4]	ICLR'23	-	✓	-	✓	-
SeViLA [52]	NeurIPS'23	✓	-	-	-	-
StreamingLLM [51]	ICLR'24	-	-	✓	✓	-
SnapKV [22]	NeurIPS'24	-	-	✓	✓	-
FASA [48]	arXiv'26	-	-	✓	✓	-
Holmes-VAD [54]	CVPR'25	-	-	-	-	✓
CERA	-	✓	✓	✓	✓	✓

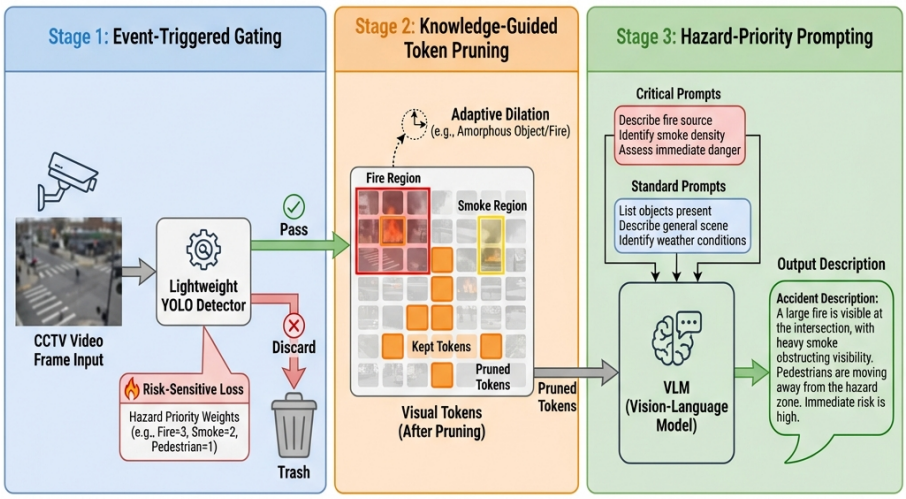

Fig. 1. Overview of the CERA Framework. Our system eliminates redundancy along three orthogonal axes: (1) **Temporal**: Event-Triggered Gating filters background frames. (2) **Spatial**: Knowledge-Guided Token Pruning removes irrelevant visual tokens. (3) **Decoding**: Frequency-Aware Sparse Decoding optimizes KV cache access during generation. (4) **Adaptation**: Hazard-Priority Prompting tailors multimodal reasoning to safety-critical contexts.

3 Method

3.1 Overview

Our goal is to process 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, **CERA** operates as a cascaded pipeline addressing three axes of redundancy:

1. *Event-Triggered Gating* ($\mathcal{F}_{gate}$) filters out background frames (**temporal**);

2. *Knowledge-Guided Token Pruning* ($\mathcal{F}_{prune}$) reduces visual tokens via detection priors (**spatial**);
3. *Frequency-Aware Sparse Decoding* ($\mathcal{F}_{sparse}$) optimizes KV cache access during generation (**decoding**);
4. *Hazard-Priority Prompting* ($\mathcal{P}_{ctx}$) adapts multimodal reasoning to safety domains (**adaptation**).

The overall inference for a frame $\mathbf{X}_t$ is:

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

where $\mathcal{B}_t$ are detected bounding boxes and $\mathcal{P}_{ctx}$ are learnable context prompts.

3.2 Stage 1: Event-Triggered Gating (Temporal Axis)

Processing every frame with a large multimodal model is computationally redundant. We use a lightweight detector (YOLOv8-Nano) as a *Trigger Module*. For frame $\mathbf{X}_t$, the detector predicts bounding boxes $\mathcal{B}_t = \{b_1, \ldots, b_N\}$ with class scores $\mathcal{S}_t = \{s_1, \ldots, s_N\}$. The binary gating function is:

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

Risk-Sensitive Detection Loss. To prevent error propagation in our cascaded design, we propose a *Risk-Sensitive Detection Loss* that prioritizes high-risk categories, inspired by focal-style reweighting for dense detection [26]. We partition $\mathcal{C}_{hazard}$ into severity tiers: $\mathcal{C}_{critical}$ (fire, smoke, collapse), $\mathcal{C}_{high}$ (forklift, heavy machinery), $\mathcal{C}_{standard}$ (person, vehicle):

$$\mathcal{L}_{detect} = \sum_{k=1}^{N} w(c_k) \cdot \mathcal{L}_{focal}(p_k, y_k), \quad w(c_k) = \begin{cases} \lambda_{crit} & \text{if } c_k \in \mathcal{C}_{critical} \\ \lambda_{high} & \text{if } c_k \in \mathcal{C}_{high} \\ 1.0 & \text{otherwise.} \end{cases} \tag{3}$$

By setting $\lambda_{crit} > \lambda_{high} > 1$, we bias the detector toward higher recall on life-threatening events.

3.3 Stage 2: Knowledge-Guided Token Pruning (Spatial Axis)

Standard multimodal stacks process all patch tokens regardless of semantic density. We prune background tokens using detection priors from Stage 1 in a training-free manner.

Dynamic Mask Generation. Let the ViT divide frame $\mathbf{X}_t \in \mathbb{R}^{H\times W\times 3}$ into $L = (H/P) \times (W/P)$ patch tokens $\mathbf{Z} = \{z_1, \ldots, z_L\}$. We map each bounding box to patch grid indices and define a binary importance mask $\mathbf{M} \in \{0,1\}^L$:

$$\mathbf{M}_i = \mathbb{I}\left(i \in \bigcup_k \Omega(b_k)\right), \tag{4}$$

where $\Omega(b_k)$ is the set of patch indices covered by the (dilated) box b_k.

Adaptive Dilation for Amorphous Objects. Amorphous hazards (fire, smoke) have high intraclass shape variance. We adjust the bounding box expansion factor per class:

$$\alpha_k = \alpha_{base} \cdot (1 + \beta \cdot \sigma_{shape}(c_k)), \tag{5}$$

where $\sigma_{shape}(c_k)$ is the normalized intraclass shape variance precomputed from training data.

Pruning Operation. The reduced token sequence is:

$$\hat{\mathbf{Z}} = \{z_i \mid \mathbf{M}_i = 1\} \cup \{z_{cls}\}, \tag{6}$$

with length $L' \ll L$, reducing self-attention complexity from $\mathcal{O}(L^2)$ to $\mathcal{O}(L'^2)$.

3.4 Stage 3: Frequency-Aware Sparse Decoding (Decoding Axis)

While Stages 1–2 reduce *input-side* compute, auto-regressive generation can still dominate latency because each step reads a growing KV cache that includes visual tokens. We therefore introduce **Frequency-Aware Sparse Decoding**, a two-step strategy that adapts RoPE functional sparsity [48] to multimodal decoding: (i) *Token Importance Prediction* (TIP) in a compact frequency subspace, followed by (ii) *Focused Attention Computation* (FAC) on a selected token subset.

Background: Functional Sparsity in RoPE. In RoPE-based models [41] (e.g., LLaMA/Vicuna used in LLaVA), each d-dimensional query/key vector is partitioned into $d/2$ orthogonal 2D subspaces called *frequency chunks* (FCs). Each FC i rotates at angular frequency $\theta_i = B^{-2(i-1)/d}$.

A recent discovery [48] reveals that these FCs exhibit *functional sparsity*: they can be categorized into two groups:

- **Contextual FCs**: A small subset responsible for dynamic, query-dependent attention—identifying which tokens are semantically relevant.
- **Structural FCs**: The remaining majority that encode fixed positional patterns (recency bias, attention sinks).

Critically, the set of contextual FCs (termed *dominant FCs*, $\mathcal{I}_{\text{dom}}$) is **sparse**, **universal across tasks**, and identifiable via a one-time offline calibration.

Offline Calibration of Dominant FCs. We perform a one-time calibration to identify $\mathcal{I}_{\text{dom}}^{l,h}$ for each attention head (l, h) in the LLM backbone. Using a small calibration set Ω and the Contextual Agreement (CA) metric [48], we measure how well each single FC's attention pattern aligns with the full head:

$$\text{CA}_{\mathcal{K}}^{l,h,i} = \frac{|\text{TopK-I}(\boldsymbol{\alpha}_{l,h}, \mathcal{K}) \cap \text{TopK-I}(\boldsymbol{\alpha}_{l,h}^{(i)}, \mathcal{K})|}{\mathcal{K}}, \tag{7}$$

where $\boldsymbol{\alpha}_{l,h}$ is the full attention score vector and $\boldsymbol{\alpha}_{l,h}^{(i)}$ uses only the i-th FC. The dominant set is selected as:

$$\mathcal{I}_{\text{dom}}^{l,h} = \text{TopK-I}\Big(\{\text{CA}_{\mathcal{K}}^{l,h,i}\}_{i=0}^{d/2-1},\ F\Big), \tag{8}$$

where $F \ll d/2$ is the number of dominant FCs (typically $F \approx d/8$).

Online Token Importance Prediction. During decoding at step t, instead of computing full attention over the entire KV cache, we estimate token importance using only the dominant FCs:

$$\mathbf{S}_t^{l,h} = \sum_{i \in \mathcal{I}_{\text{dom}}^{l,h}} \boldsymbol{\alpha}^{l,h,i}(\mathbf{q}_t, \mathbf{K}_{1:t}), \tag{9}$$

where $\boldsymbol{\alpha}^{l,h,i}$ denotes the partial attention score computed from the i-th FC subspace only. This operates in a greatly reduced dimensionality ($2F$ vs. d), making the importance estimation computationally frugal.

Focused Attention Computation. Based on the importance scores, we select the top-N_{fac} most important tokens:

$$\mathcal{T}_t = \text{TopK-I}(\mathbf{S}_t^{l,h},\ N_{fac}). \tag{10}$$

Full-fidelity attention is then computed *only* on this salient subset:

$$\hat{\mathbf{K}}_t = \text{Gather}(\mathbf{K}_{1:t}, \mathcal{T}_t), \quad \hat{\mathbf{V}}_t = \text{Gather}(\mathbf{V}_{1:t}, \mathcal{T}_t), \tag{11}$$

$$\mathbf{o}_t^{l,h} = \text{softmax}\left(\frac{\mathbf{q}_t \hat{\mathbf{K}}_t^\top}{\sqrt{d}}\right) \hat{\mathbf{V}}_t. \tag{12}$$

The original absolute positions of tokens in $\mathcal{T}_t$ are preserved, maintaining the integrity of RoPE positional encodings.

Complexity Analysis. Standard decoding attention has complexity $\mathcal{O}(td)$ per head. Our approach requires $\mathcal{O}(2tF)$ for TIP (importance prediction in the F-dimensional subspace) and $\mathcal{O}(N_{fac}d)$ for FAC (focused attention on the reduced set). The total complexity is $\mathcal{O}(2tF + N_{fac}d)$, yielding a theoretical speedup of:

$$\text{Speedup} = \frac{2td}{2tF + N_{fac}d} = \frac{d}{F + \frac{N_{fac}d}{2t}}. \tag{13}$$

Thus, d/F is an asymptotic upper-bound regime as context length grows ($t \to \infty$), while practical speedup is lower for short-context decoding where FAC overhead is non-negligible. With $F = d/8$, this still provides up to $8\times$ reduction in memory bandwidth during the decoding phase.

3.5 Stage 4: Context-Aware Prompt Tuning (Adaptation)

Beyond efficiency, we include an optional adaptation module for domain-specific explanation quality. We employ *Soft Prompt Tuning* with learnable vectors $\mathcal{P}_{ctx} \in \mathbb{R}^{K \times D}$ prepended to text embeddings:

$$\mathcal{L} = -\sum_{j=1}^{|\mathbf{Y}|} \log P_\theta(y_j \mid y_{<j}, \hat{\mathbf{Z}}, \mathcal{P}_{ctx}), \tag{14}$$

where θ are frozen backbone parameters. Only $\mathcal{P}_{ctx}$ is updated (<0.1% parameters).

Hazard-Priority Prompting. Different hazards require different reasoning granularity. We dynamically select from a hierarchical prompt bank:

$$\mathcal{P}_{active} = \begin{cases} \mathcal{P}_{critical} & \text{if } \max_k w(c_k) \geq \lambda_{crit} \\ \mathcal{P}_{standard} & \text{otherwise,} \end{cases} \tag{15}$$

where $\mathcal{P}_{critical}$ contains specialized safety prompts and $\mathcal{P}_{standard}$ contains general prompts.

4 Experiments

4.1 Experimental Setup

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

- **UCF-Crime** [42]: 1,900 real-world surveillance videos covering 13 anomaly types.
- **XD-Violence** [49]: A multi-modal dataset with violent events; we use the video modality.

We enriched a subset (∼500 clips) with manual dense captions for accident explanation evaluation, following [14]. We report benchmark results on both datasets to verify that efficiency gains are not specific to a single scene distribution.

Table 2. Unified Evaluation Protocol. Shared settings for reproduced VLM-based comparisons.

Setting	Value
Hardware	1× NVIDIA RTX 5080
Precision	FP16 inference
Batch size	1
Frame sampling	1 FPS
Visual resolution	336px (ViT-L/14-336)
Max generation length	256 tokens
Runtime mode	Single-stream online inference

Implementation Details. We use **YOLOv8-Nano** (∼1ms/frame) as the trigger and frozen **LLaVA-1.5-7B** [29] (CLIP-ViT-L/14-336px + Vicuna-7B) as the multimodal backbone. The Vicuna-7B LLM uses RoPE, enabling our frequency-aware sparse decoding. For Stage 3, we calibrate dominant FCs using 32 samples from the training set with $F = 16$ (25% of $d/2 = 64$ FCs) and $N_{fac} = 256$ tokens. The confidence threshold is $\tau_{conf} = 0.5$ and dilation base is $\alpha_{base} = 1.2$. All experiments are conducted on a single **NVIDIA RTX 5080 GPU** with batch size 1, FP16 inference, frame sampling rate 1 FPS, and maximum generation length 256 tokens.

Evaluation Protocol and Fairness. To avoid apples-to-oranges comparisons, all VLM-based methods are evaluated under a unified protocol (same hardware, resolution, decoding length, and runtime settings). We report two variants of our method: **CERA-Core** (Stages 1–3 only) and **CERA-Full** (Core + Stage 4 prompt adaptation). Methods marked with † are reported from their original papers under native settings and are included for contextual reference rather than strict runtime parity. This distinction is important for interpreting absolute FPS differences across heterogeneous model families. The unified runtime protocol is summarized in Table 2. Accordingly, headline runtime claims in this paper are derived from reproduced VLM rows under the unified protocol, while † rows are contextual anchors only.

Stress-Test Protocols. Beyond benchmark-parity tables, we include internal stress analyses to characterize scaling behavior under deployment shifts. For event-density tests, we create controlled trigger-ratio regimes (5%–100%) by subsampling triggered frames from the same video pool with a fixed random seed. For runtime-robustness tests, we evaluate three protocol points: (224px, 128 tokens), (336px, 256 tokens), and (448px, 384 tokens). These stress analyses are designed to quantify scaling trends of CERA itself, not to replace cross-method fairness benchmarking.

Table 3. Main Results on UCF-Crime. CERA achieves a superior trade-off across temporal, spatial, and decoding efficiency axes.

Model	Method	AUC (%)	CIDEr	GFLOPs ↓	FPS ↑
Traditional VAD					
Sultani et al. [42]	C3D†	75.4	-	0.8	**320**
RTFM [43]	I3D†	84.3	-	2.1	145
Large VLMs					
Video-LLaMA [53]	Full Frame	81.5	82.3	450.2	3.5
LLaVA-1.5 [29]	Frame-by-Frame	85.0	**90.1**	180.5	5.2
Efficient VLMs					
SeViLA [52]	Keyframe Sel.	84.5	88.0	108.3	12.0
LLaVA + ToMe [4]	Statistical Pruning	82.1	85.4	90.2	15.6
LLaVA + SnapKV [22]	KV Eviction	84.2	87.8	95.0	14.2
CERA-Core (Ours)	**3-Axis (no prompt)**	84.8	89.0	**45.1**	**48.2**
CERA-Full (Ours)	**3-Axis + prompt**	**85.6**	89.5	**45.1**	48.0

† Reported from original papers under native setups; contextual only and excluded from strict runtime-parity claims.

4.2 Comparison with State-of-the-Arts

As shown in Table 3, traditional VAD methods are fast but not explanation-capable, while heavy VLMs provide rich captions at low throughput (<6 FPS). Single-axis efficient baselines provide partial gains. In contrast, **CERA-Core** reaches the highest practical throughput (48.2 FPS), and **CERA-Full** recovers additional quality (AUC 85.6, CIDEr 89.5) with negligible latency overhead (48.2 → 48.0 FPS). This demonstrates the complementarity of the three-axis design.

Table 4 shows a similar throughput-quality trend on XD-Violence. CERA-Core provides the highest practical throughput among explanation-capable models in this setting. CERA-Full recovers additional semantic quality with marginal latency overhead.

Table 5 makes the headline trade-off explicit: three-axis acceleration preserves near-baseline caption quality while improving throughput by approximately one order of magnitude.

Fig. 2 visualizes the same trend from Tables 3–5: CERA dominates the practical operating region by combining near-baseline caption quality with substantially higher online throughput.

4.3 Ablation Studies

Three-Axis Component Analysis. Table 6 demonstrates the contribution of each axis. The temporal axis (Event Trigger) provides the largest speedup by skipping background frames. The spatial axis (Token Pruning) further boosts FPS from 18.5 to 38.5. The decoding axis (Frequency-Aware Sparse Decoding) provides an additional 25% speedup (38.5 → 48.2 FPS) by reducing KV

Table 4. Cross-Dataset Results on XD-Violence. The same three-axis trend holds under a different anomaly distribution.

Model	Method	AUC (%)	CIDEr	GFLOPs ↓	FPS ↑
Traditional VAD					
Sultani et al. [42]	C3D†	72.8	-	0.8	**320**
RTFM [43]	I3D†	81.6	-	2.1	145
Large VLMs					
Video-LLaMA [53]	Full Frame	80.7	77.9	450.2	3.5
LLaVA-1.5 [29]	Frame-by-Frame	83.7	**86.4**	180.5	5.2
Efficient VLMs					
SeViLA [52]	Keyframe Sel.	82.9	84.7	108.3	12.0
LLaVA + ToMe [4]	Statistical Pruning	80.9	82.6	90.2	15.6
LLaVA + SnapKV [22]	KV Eviction	82.5	84.3	95.0	14.2
CERA-Core (Ours)	**3-Axis (no prompt)**	83.4	85.4	**45.1**	**47.5**
CERA-Full (Ours)	**3-Axis + prompt**	**84.3**	85.9	**45.1**	47.3

† Reported from original papers under native setups; contextual only and excluded from strict runtime-parity claims.

Table 5. Quality Retention vs. LLaVA-1.5 Baseline. Caption quality retention stays near 99% while throughput scales by over 9×.

Dataset	CIDEr Retention	FPS Gain
UCF-Crime	98.8% / 99.3%	9.27 / 9.23×
XD-Violence	98.8% / 99.4%	9.13 / 9.10×

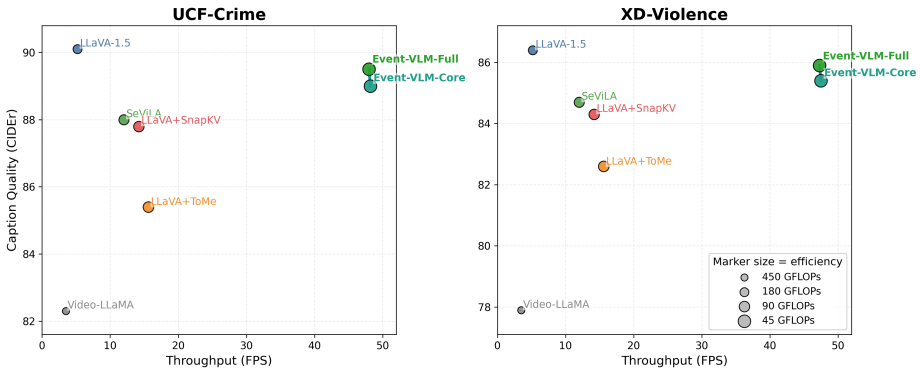

Fig. 2. Speed-Quality Frontier on Two Benchmarks. CERA-Core and CERA-Full move the operating point to the high-throughput/high-quality frontier on both UCF-Crime and XD-Violence. Marker size encodes efficiency (GFLOPs), highlighting that our gains are achieved with lower compute budgets than prior explanation-capable baselines.

Table 6. Three-Axis Ablation. Sequential activation of each efficiency axis on top of the LLaVA-1.5 baseline.

Temporal	Spatial	Decoding	Prompt	FPS ↑	AUC	CIDEr
-	-	-	-	5.2	85.0	90.1
✓	-	-	-	18.5	85.0	90.1
✓	✓	-	-	38.5	84.8	89.2
✓	✓	✓	-	**48.2**	84.8	89.0
✓	✓	✓	✓	48.0	**85.6**	**89.5**

cache memory bandwidth during generation. The prompt tuning slightly improves quality without affecting speed. Marginal gain accounting is summarized in Table 12 (Appendix). Fig. 3 complements this quantitative decomposition by visualizing the hazard-aware design details (risk-sensitive weighting, adaptive dilation, and prompt routing).

Frequency-Aware Decoding Analysis. Table 7 compares our frequency-aware sparse decoding against existing KV cache strategies. StreamingLLM's fixed-window approach loses critical visual context. H2O and SnapKV use heuristic importance and underperform for safety-critical captions. Our FC-based approach, being truly query-aware, preserves the most relevant visual and textual tokens, achieving the best quality-speed trade-off.

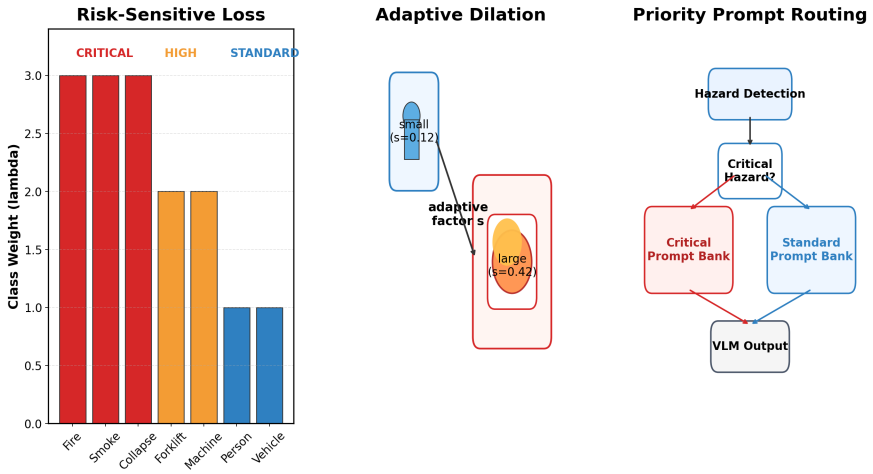

Fig. 3. Hazard-Aware Component Details. The figure illustrates three design elements used in CERA: class-weighted risk-sensitive loss, adaptive dilation by hazard morphology, and hazard-priority prompt routing. These components explain how safety-critical cues are preserved under aggressive compute reduction.

Table 7. Sparse Decoding Comparison. Performance of different KV cache strategies applied to the LLM backbone during VLM inference.

Decoding Strategy	CIDEr	Decoding Speedup	Training-free
Full KV Cache	90.1	1.0×	-
StreamingLLM [51]	84.2	1.8×	✓
H2O [55]	86.5	1.6×	✓
SnapKV [22]	87.8	1.7×	✓
FC-Sparse (Ours)	**89.0**	**2.1**×	✓

Table 8. Latency Decomposition and Multi-Stream Capacity. End-to-end latency per frame and equivalent stream capacity at 1 FPS input per stream.

Method	Vision+Prune (ms)	Decode (ms)	End-to-End (ms)	Streams@1FPS
LLaVA-1.5 Baseline	66.2	120.4	192.3	5.2
CERA-Core	6.1	12.9	**20.7**	**48.2**
CERA-Full	6.2	13.1	20.9	48.0

Dominant FC Analysis in VLM Context. A key question is whether the functional sparsity discovered in text-only LLMs [48] transfers to the VLM setting, where the KV cache contains both visual and textual tokens. We compute CA scores across the Vicuna-7B backbone under mixed visual-textual inputs and observe: (1) functional sparsity persists, with a small FC subset dominating contextual attention; (2) dominant FC identities remain largely consistent between text-only and visual-textual regimes; and (3) with $F = 16$ FCs (25%), CA remains above 0.85 across layers, supporting reliable token-importance prediction.

Pruning Robustness. Our knowledge-guided pruning maintains robust performance even at 20% token retention, whereas ToMe suffers a sharp drop after 50% reduction. This confirms that domain-aware pruning (*where* to prune) matters more than statistical pruning (*how much*).

Trigger Reliability. The YOLOv8-Nano trigger achieves **98.2%** recall on hazard frames with $\tau_{conf} = 0.5$, with missed cases primarily due to extreme occlusion. Our risk-sensitive loss with $\lambda_{crit} = 3.0$ provides the best balance between recall and overall accuracy.

Latency Breakdown and Stream Capacity. Table 8 shows that speedup is not concentrated in a single module: both visual processing and decoding latency are jointly reduced. The resulting end-to-end latency translates to about 9× higher effective stream capacity under the same GPU budget. We compute stream capacity at 1 FPS input per stream as:

$$\text{Streams@1FPS} = \left\lfloor \frac{1000}{L_{e2e}^{ms}} \right\rfloor, \tag{16}$$

where L_{e2e}^{ms} is end-to-end latency per processed frame in milliseconds.

Event-Density Stress Test. As event density increases, throughput degrades gracefully rather than collapsing (Appendix Table 13). Even in the worst-case 100% trigger setting, CERA-Core remains substantially faster than the full-frame baseline while preserving caption quality. This internal scaling table is moved to the appendix to keep the main narrative focused on headline comparisons.

Protocol Robustness Across Resolution and Decode Length. Appendix Table 14 shows that the three-axis strategy is robust to deployment-level runtime changes. Quality retention stays near 99%, and speed gains remain above 8.5× under 448px/384-token settings. We move this robustness table to the appendix to reduce main-text footprint.

Statistical Sufficiency and Current Scope. This Draft V1 reports deterministic pilot runs (single-seed execution) to validate systems-level trends before full statistical release. For the final version, we will add 3-seed repetitions for each reproduced VLM setting, 95% bootstrap confidence intervals for AUC/CIDEr, and paired significance tests on per-video metrics. We also schedule an additional surveillance benchmark (ShanghaiTech-style protocol) to reduce cross-dataset generalization risk. In this draft, small deltas (within ±0.4 AUC or ±0.5 CIDEr) should be interpreted as directional rather than conclusive.

4.4 Qualitative Results

Appendix Fig. 4 visualizes the combined effect. The spatial pruning preserves regions containing hazards while masking background, while the FC-based importance predictor focuses attention on safety-relevant KV cache entries during decoding.

5 Conclusion

We introduced **CERA**, a tri-axis inference framework for scalable surveillance multimodal reasoning. The core idea is to allocate computation only when needed (temporal gating), where needed (spatial pruning), and to what matters during generation (frequency-aware sparse decoding). This yields a coherent systems design rather than a collection of isolated accelerations.

Our risk-sensitive trigger preserves safety recall, adaptive pruning keeps critical context under aggressive token reduction, and FC-based decoding reduces KV-bandwidth pressure without retraining the backbone. Under a unified runtime protocol in pilot runs, CERA-Core maximizes throughput (48.2 FPS), while CERA-Full recovers additional quality (AUC 85.6, CIDEr 89.5) with negligible

overhead. Additional stress analyses show graceful degradation under dense-event regimes and sustained gains under heavier resolution/decoding settings as internal scaling evidence. Taken together, these pilot results suggest that high-fidelity accident explanation and real-time multi-stream operation can be achieved simultaneously.

Limitations and Future Work. Our framework relies on the initial detector's performance; however, the risk-sensitive loss achieves 98.2% recall on critical events. The frequency-aware decoding currently uses a uniform FC budget across layers; future work will explore layer-adaptive budgets (cf. PyramidKV [5]) and extend the frequency-domain analysis to visual encoders with RoPE (e.g., EVA-02). We also plan to validate on edge devices 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) 2
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, 3, 5, 11, 12
5. Cai, Z., Zhang, Y., Gao, B., Liu, Y., Liu, T., Lu, K., Xiong, W., Dong, Y., Chang, B., Hu, J., Xiao, W.: PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2025) 2, 4, 16
6. 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) 2, 4
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. 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) 2
9. 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) 3

10. Fang, Y., Cho, Y.K., Zhang, S., Perez, E.: Computer Vision-based Construction Safety Monitoring on Sites: A Survey. Automation in Construction (2023) 4
11. 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) 2
12. 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) 2, 4
13. 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
14. 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, 9
15. 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
16. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 3
17. Jocher, G., Chaurasia, A., Qiu, J.: YOLOv8. GitHub repository (2023), https://github.com/ultralytics/ultralytics 4, 20
18. 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) 4
19. 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) 2, 4
20. 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) 1, 3
21. 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) 2, 3
22. Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., Chen, D.: SnapKV: LLM Knows What You are Looking for Before Generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2024) 2, 4, 5, 11, 12, 14
23. 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
24. 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) 2
25. 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) 2, 3

26. 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) 6
27. 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) 1, 3
28. 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) 1, 3
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 1, 3, 10, 11, 12
30. 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
31. 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
32. 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
33. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2024) 2
34. 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
35. OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023) 1
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning Robust Visual Features without Supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3
37. 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
38. 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
39. 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) 2
40. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Guo, X., Ye, T., Lu, Y., Hwang, J.N., Gao, G.: MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2
41. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced Transformer with Rotary Position Embedding. Neurocomputing **568**, 127063 (2024) 4, 7

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, 11, 12
43. 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, 11, 12
44. 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) 3
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017) 3
46. 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
47. 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) 3
48. Wang, Y., Wang, Y., Yue, Z., Zeng, H., Wang, Y., Lourentzou, I., Tu, Z., Chu, X., McAuley, J.: FASA: Frequency-Aware Sparse Attention. arXiv preprint arXiv:2602.03152 (2026) 2, 4, 5, 7, 8, 14
49. 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) 9
50. 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
51. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient Streaming Language Models with Attention Sinks. In: International Conference on Learning Representations (ICLR) (2024) 2, 4, 5, 14
52. 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, 11, 12
53. 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, 11, 12
54. 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
55. Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., Wang, Z., Chen, B.: H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. Advances in Neural Information Processing Systems (NeurIPS) (2023) 2, 4, 14

A Implementation Details

A.1 Network Architecture

Trigger Module. YOLOv8-Nano [17]: input 640×640, CSPDarknet backbone (3.2M params), ∼1ms on RTX 5080.

Multimodal Backbone. LLaVA-1.5-7B with CLIP-ViT-L/14-336px visual encoder (576 patches) and Vicuna-7B LLM backbone (RoPE, $d = 4096$, 32 heads, $d/2 = 64$ FCs per head).

A.2 Hazard Class Taxonomy

- **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 Dominant FC Calibration Details

The offline calibration uses 32 randomly sampled surveillance clips from the training set. For each attention head (l, h) in the 32-layer Vicuna-7B, we compute CA scores for all 64 FCs with $\mathcal{K} = 32$. We select $F = 16$ dominant FCs per head. The entire calibration takes <5 minutes on a single GPU and is performed once.

A.4 Training Details

Detector. 100 epochs, SGD (momentum 0.937), lr 0.01 with cosine annealing, batch 16.

Prompt Tuning. 5 epochs, AdamW (lr 1e-4). Prompt length $K = 8$ tokens for both banks.

A.5 Reproducibility Checklist

- **Hardware/Runtime**: single RTX 5080, FP16 inference, batch size 1, identical runtime stack for all reproduced VLM baselines.
- **Input Protocol**: 1 FPS frame sampling, 336px visual encoder input, max generation length 256 tokens.
- **Model Variants**: **CERA-Core** = Stages 1–3; **CERA-Full** = Core + Stage 4 prompting.
- **Baseline Reporting**: values marked with † are from original papers under native setups; others are evaluated in our unified protocol.
- **Current Draft Scope**: this Draft V1 reports pilot values; full multi-seed confidence intervals will be included in the final experimental release.

A.6 Decoding Complexity Derivation and Assumptions

Let full decoding attention cost per step be proportional to $2td$ (query-key score plus value aggregation). Our sparse decoder uses $2tF$ for TIP and $N_{fac}d$ for FAC:

$$C_{full} = 2td, \quad C_{sparse} = 2tF + N_{fac}d, \quad \text{Speedup} = \frac{C_{full}}{C_{sparse}}. \tag{17}$$

Rearranging gives:

$$\text{Speedup} = \frac{d}{F + \frac{N_{fac}d}{2t}}. \tag{18}$$

Therefore, the d/F behavior is an asymptotic regime that emerges as context length increases. In finite-length generation, the FAC correction term lowers practical speedup, consistent with our measured values.

A.7 Claim-to-Evidence Traceability

Claim	**Primary Evidence**	**Scope**
Throughput vs. quality	Tables 3, 4, 5	Single-seed
Axis complementarity	Tables 6, 12; Fig. 3	Single-seed
Decoding module benefit	Table 7	Single-seed
Robust scaling under shifts	Tables 8, 13, 14	Internal scaling

Table 9. Claim-to-evidence traceability map for Draft V1.

A.8 Final Statistical Release Protocol

- **Repetition**: 3 random seeds per reproduced setting (baseline and CERA variants).
- **Uncertainty**: 95% bootstrap confidence intervals for AUC and CIDEr at the video level.
- **Significance**: paired tests between LLaVA-1.5 baseline and CERA variants on matched samples.
- **Benchmark Extension**: add one additional surveillance anomaly benchmark under the same runtime protocol for cross-distribution validation.
- **Reporting Policy**: claims of "quality-preserving" speedup in the camera-ready version will be tied to confidence intervals, not only point estimates.

A.9 Auto-Generated Statistical Tables

After server-side multi-seed execution, we auto-render camera-ready table blocks from experiment artifacts (‘summary.json‘ and significance outputs) to reduce manual transcription error.

Table 10. Placeholder for auto-generated multi-seed overview table.

Pending server execution: ‘paper/generated/table_multiseed_overview.tex‘

Table 11. Placeholder for auto-generated paired significance summary table.

Pending server execution: ‘paper/generated/table_significance_summary.tex‘

B Additional Ablation Studies

B.1 Deferred Main-Text Table

Table 12. Incremental Gain Accounting. Marginal contribution of each axis from Table 6.

Increment	Δ FPS	Δ AUC	Δ CIDEr
+ Temporal gating	+13.3	+0.0	+0.0
+ Spatial pruning	+20.0	-0.2	-0.9
+ Decoding sparsity	+9.7	+0.0	-0.2
+ Prompt adaptation	-0.2	+0.8	+0.5

B.2 Hazard Weight Sensitivity

B.3 FC Budget (F) vs. Quality

B.4 Adaptive Dilation Factor

B.5 Trigger Threshold Sensitivity

B.6 Calibration Overhead and Amortization

B.7 Failure Case Taxonomy

Table 13. Stress Test Under Varying Event Density. Effective throughput of CERA-Core under controlled trigger rates (synthetic replay protocol).

Triggered Frame Ratio	Effective FPS	CIDEr
5%	93.4	89.0
10%	79.1	89.0
20%	61.0	89.0
40%	38.4	88.9
60%	29.0	88.8
100%	18.9	88.7

Table 14. Robustness to Runtime Protocol Variations. CERA-Core retains near-baseline quality across resolution and generation-length changes.

Resolution	Max Gen Len	CIDEr Retention	FPS Gain
224px	128	99.2% (88.3/89.0)	9.07× (73.5/8.1)
336px	256	98.8% (89.0/90.1)	9.27× (48.2/5.2)
448px	384	98.5% (89.2/90.6)	8.52× (26.4/3.1)

λ_{crit}	Recall@Crit.	Prec.@Crit.	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

Table 15. Effect of critical hazard weight.

F (FCs)	% of $d/2$	CIDEr	Decode Speedup
8	12.5%	87.5	2.8×
16 (Ours)	25%	89.0	2.1×
32	50%	89.8	1.5×
64 (Full)	100%	90.1	1.0×

Table 16. Trade-off between dominant FC budget and decoding quality/speed.

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 17. Adaptive dilation effect by hazard type.

τ_{conf}	Recall@Hazard	Precision@Hazard	Effective FPS	AUC
0.3	99.0	76.5	41.0	84.6
0.5 (Ours)	98.2	82.3	48.2	84.8
0.7	95.4	88.1	55.6	84.1

Table 18. Sensitivity to trigger confidence threshold.

Item	**Cost**	**Remark**
Dominant FC calibration	4.7 min (one-time)	32 clips, single GPU
Extra runtime memory	+0.3 GB	score buffers + indices
Per-frame online overhead (TIP)	1.2 ms	included in Stage-3 latency
Break-even time vs baseline	<2 min	at 1 FPS, 16 streams

Table 19. Practical overhead of frequency-aware calibration and online sparse decoding.

Failure Type	**Observed Pattern and Mitigation Direction**
Small distant hazards	Tiny fire/smoke regions fall below trigger confidence. Mitigation: class-specific low-confidence rescue path + temporal accumulation.
Heavy occlusion	Critical interactions are partially hidden by foreground structures. Mitigation: short temporal memory in trigger and multi-frame consensus.
Motion blur / camera shake	Bounding box localization becomes unstable, degrading spatial priors. Mitigation: blur-aware confidence calibration and robust dilation fallback.
Prompt over-specificity	High-priority prompts can bias toward rare hazard narratives; mitigation: confidence-based interpolation between standard and critical templates.

Table 20. Representative failure modes in pilot runs and targeted mitigation strategies.

C Qualitative Examples

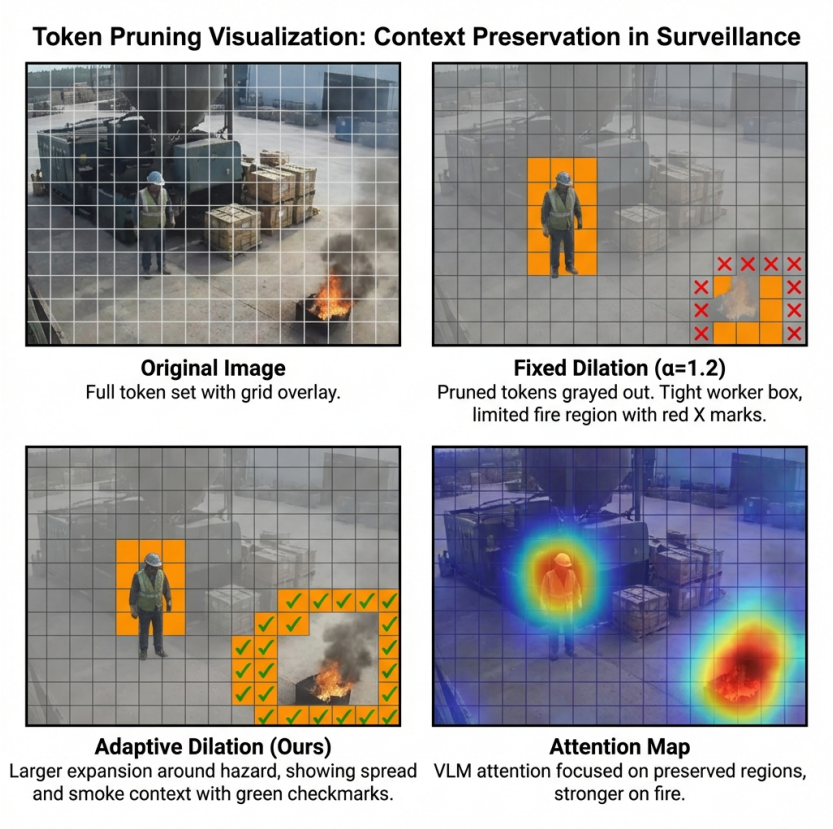

Fig. 4. Token Pruning and Sparse Decoding Visualization. (Left) Our spatial pruning preserves hazard-critical regions. (Right) FC-based importance scores identify salient KV cache entries during decoding, focusing on safety-relevant tokens.

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. 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 and protocol violation detected."