

# VADER: Towards Causal Video Anomaly Understanding with Relation-Aware Large Language Models

Ying Cheng<sup>1</sup>, Yu-Ho Lin<sup>1</sup>, Min-Hung Chen<sup>2</sup>, Fu-En Yang<sup>2</sup>, Shang-Hong Lai<sup>1\*</sup>

<sup>1</sup>National Tsing Hua University <sup>2</sup>NVIDIA

## Abstract

*Video anomaly understanding (VAU) aims to provide detailed interpretation and semantic comprehension of anomalous events within videos, addressing limitations of traditional methods that focus solely on detecting and localizing anomalies. However, existing approaches often neglect the deeper causal relationships and interactions between objects, which are critical for understanding anomalous behaviors. In this paper, we propose VADER, an LLM-driven framework for Video Anomaly unDERstanding, which integrates keyframe object Relation features with visual cues to enhance anomaly comprehension from video. Specifically, VADER first applies an Anomaly Scorer to assign per-frame anomaly scores, followed by a Context-Aware Sampling (CAES) strategy to capture the causal context of each anomalous event. A Relation Feature Extractor and a COntrastive Relation Encoder (CORE) jointly model dynamic object interactions, producing compact relational representations for downstream reasoning. These visual and relational cues are integrated with LLMs to generate detailed, causally grounded descriptions and support robust anomaly-related question answering. Experiments on multiple real-world VAU benchmarks demonstrate that VADER achieves strong results across anomaly description, explanation, and causal reasoning tasks, advancing the frontier of explainable video anomaly analysis. Project page is available at <https://vader-vau.github.io/>.*

## 1. Introduction

Video anomaly detection (VAD) plays a critical role in surveillance, traffic monitoring, and public safety. Conventional methods primarily focus on temporal localization and classification [13, 28, 29, 40, 46, 53, 54, 71], but often lack semantic interpretability, limiting their utility in decision-making. To address this, recent research has shifted from video anomaly detection to video anomaly understanding

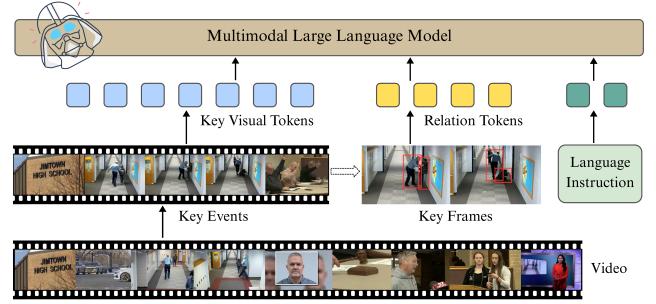


Figure 1. **Concept of VADER.** VADER enables detailed anomaly understanding by extracting key visual and relational cues from selected key frames.

(VAU) [34, 47, 50, 57, 67], emphasizing detailed descriptions, causal reasoning, and context-aware explanations.

Recent advances in Large Language Models (LLMs) [3, 16, 21, 30, 31, 36, 61] have enabled natural language understanding of video content, paving the way for explainable VAU. Several works have leveraged LLMs for open-world VAU, including causation-focused pipeline [11], interaction-aware framework [47], and temporal segment selection [67]. However, these approaches often neglect deeper causal relationships and dynamic object interactions, which are critical for understanding unusual behaviors. See Supp. Material Sec. A.

In this work, we propose **VADER**, an LLM-driven framework for Video Anomaly unDERstanding, integrating keyframe-level object Relation features and visual cues to enhance anomaly comprehension in video, as illustrated in Figure 1. VADER first applies an Anomaly Scorer to assign per-frame anomaly scores, followed by a Context-Aware Sampling (CAES) strategy that adaptively selects keyframes to capture the full causal context of each event. For each sampled keyframe, a Relation Feature Extractor provides rich object and relational representations, which are further aggregated by a COntrastive Relation Encoder

\*Corresponding author

(CORE) into compact relational tokens that encode salient and temporally resolved interaction dynamics. By integrating both visual and relational cues into the LLM, VADER enables detailed, causally consistent narrative generation and robust anomaly-related question answering.

We extensively evaluate VADER on three recent and challenging VAU benchmarks, covering tasks such as anomaly description, question answering, and causal reasoning. Our method outperforms strong baselines and recent state-of-the-art systems across multiple metrics. Ablation analyses further highlight the impact of each major component.

In summary, our main contributions are as follows:

- We introduce VADER, an integrated framework that combines context-aware event sampling with dynamic relational modeling to support causally grounded video anomaly understanding with LLMs.
- We propose a Context-Aware Sampling strategy (CAES) based on anomaly scores and temporal gradients, enabling story-driven keyframe selection.
- We develop a weakly supervised CONorstic Relation Encoder (CORE) that produces dynamic tokens for evolving object interactions, enhancing both interpretability and reasoning depth.
- We validate VADER on multiple challenging benchmarks, achieving state-of-the-art or highly competitive performance in video anomaly description, explanation, and reasoning tasks.

## 2. Related Work

### 2.1. Video Anomaly Detection and Understanding

Video Anomaly Detection (VAD) aims to identify events that deviate from typical patterns in videos. Early approaches [23, 37, 42, 63] relied on handcrafted features and traditional algorithms to capture motion or appearance anomalies. With the advent of deep learning, CNNs and RNNs have been widely adopted to model spatiotemporal dynamics [10, 27, 48, 71] and enhance detection performance. Recent methods [14, 35, 38, 72, 72, 73] often utilize reconstruction-based or weakly supervised approaches to address these tasks, aiming to learn normal patterns and detect deviations as anomalies. To facilitate progress in this field, several datasets [2, 7, 27, 33, 46, 52, 59] have been introduced, covering diverse video content and annotation granularity. Beyond VAD, Video Anomaly Understanding (VAU) involves deeper analysis, such as describing anomaly contents or investigating their contexts. With recent advancements in text generation and multimodal learning, VAU research has seen significant progress. Several benchmarks such as [11, 47, 62, 67] have been developed. Building on these resources, VAU emerged as a new task, and traditional VAD was further extended with MLLM-based approaches; we review these in the following section.

### 2.2. Video Anomaly Analysis with MLLMs

Recent advances in Multimodal Large Language Models (MLLMs) [3, 21, 26, 74] have greatly expanded the scope of video anomaly analysis. By jointly modeling visual content and natural language, MLLMs enable capabilities such as open-vocabulary anomaly detection and reasoning [5, 56]. Video-capable MLLMs [16, 20, 30, 31, 36] further enhance these abilities by reasoning over video data. Related works on VAD include LAVAD [65] and SUVAD [12] for training-free anomaly detection; AnomalyRuler [57] with rule-based reasoning; VAD-LLaMA [34] and VERA [60] as an explainable detector. Works on VAU include Holmes-VAU [67] with an anomaly-focused sampling; CUVA [11] with causation-focused reasoning; HAWK [47] with interaction-aware modeling; and Assist-PDA [58] for real-time analysis. Despite these advances, existing approaches still overlook the deeper causal relationships and interactions between objects, which are critical for understanding anomalous behaviors.

### 2.3. Modeling Relationships with Scene Graphs

Understanding object relationships is vital for high-level video analysis. Many anomalies arise from unusual interactions or collective behaviors, such as a person running against a crowd, which cannot be fully captured by appearance features alone. Scene graphs provide a structured representation of these interactions by modeling objects as nodes and relationships as edges, enabling models to move beyond low-level features toward relational understanding. EGTR [17] extracts relational information from DETR [6] for efficient graph generation, while DecoAD [8] integrates scene and action features via knowledge graphs to better detect human-centric anomalies. Lohner et al. [32] further showed that adding encoded scene graph features improves detection performance, highlighting the value of explicit relational modeling. Building on these advances, we leverage [17] to extract object-level relations and integrate them into our framework, enabling more interpretable and comprehensive anomaly understanding.

## 3. Method

In this section, we present VADER, a novel framework for deep, causal understanding of anomalous events in complex videos. Unlike traditional anomaly detection approaches, VADER is designed to interpret and explain both the occurrence and underlying causes of anomalous events. Our framework consists of two main components: (1) keyframe detection and selection (Sec 3.1) for narrative-driven context sampling, and (2) relation extraction and integration (Sec 3.2) for fine-grained object interaction modeling. An overview of our pipeline is shown in Figure 2.

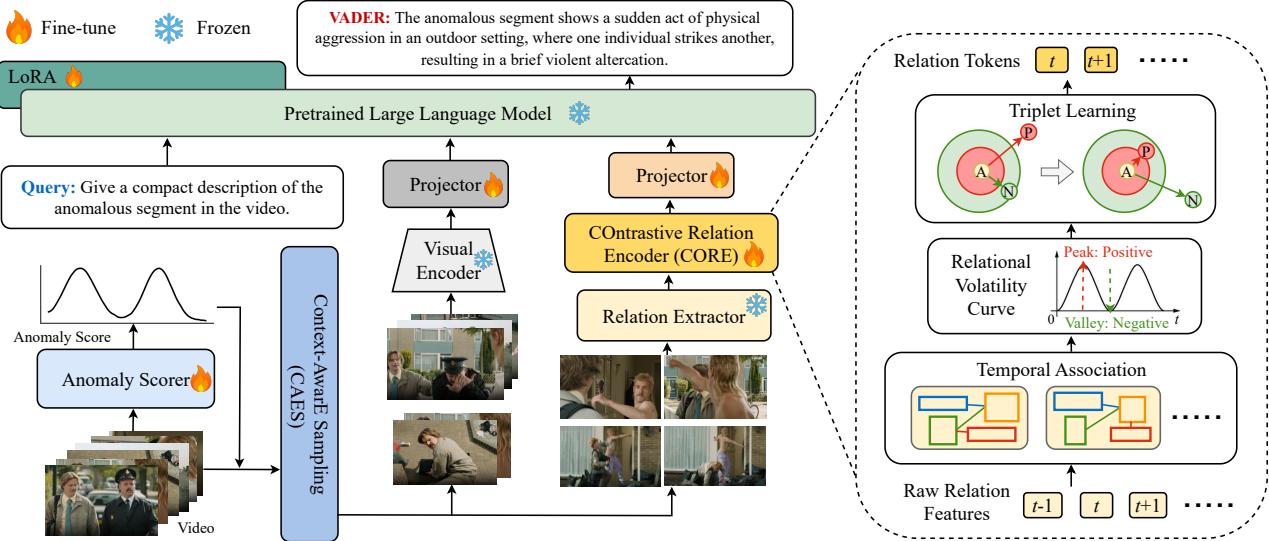


Figure 2. **Overview of VADER framework.** Given an input video, the Anomaly Scorer and Context-AwarE Sampling (CAES) identify keyframes for narrative-driven anomaly analysis. Visual and relational features are extracted and encoded, with dynamic relational patterns distilled by the COntrastive Relation Encoder (CORE). All cues are fused by a pretrained LLM for comprehensive video anomaly understanding. The right panel illustrates the relational branch, including temporal association, volatility mining, and contrastive token learning.

### 3.1. Keyframe Detection and Selection

To focus computational resources on the most informative video content, VADER first assigns an anomaly score to each frame using an Anomaly Scorer and then applies a keyframe selection strategy to filter and refine the candidate frames for downstream analysis.

#### 3.1.1. Anomaly Scoring

To build a robust and widely applicable system, we leverage an Anomaly Scorer inspired by the CLIP-based framework [64]. Rather than training a specialized model for each target dataset [72, 73], our scorer is pre-trained once on a large, aggregated dataset encompassing diverse video scenarios.

Following [64], we compute a normality prototype  $\mathbf{m}$  by averaging features extracted by the CLIP [41] image encoder  $E_I$  on all normal training video frames  $\mathcal{I}_{\text{norm}}$ :

$$\mathbf{m} = \frac{1}{|\mathcal{I}_{\text{norm}}|} \sum_{\mathbf{x} \in \mathcal{I}_{\text{norm}}} E_I(\mathbf{x}), \quad (1)$$

Both visual and text features are then re-centered by subtracting  $\mathbf{m}$ , aligning the origin with normal behavior, where distances indicate abnormality and directions encode semantics. The re-centered frame features are then projected onto class-specific semantic directions  $d_c$  and passed through a softmax to produce a conditional class distribution  $p_{c|A}(I_i)$ , representing the probability of each anomaly class  $c$  given that an anomaly occurs.

The per-frame anomaly probability  $p_A(I_i)$ , estimated by a temporal module that captures short- and long-term de-

pendencies across frame sequences, is combined with the conditional class distribution  $p_{c|A}(I_i)$  to form a joint probability. The frame-level anomaly score is computed as:

$$S_i = \max_c (p_A(I_i) \cdot p_{c|A}(I_i)), \quad (2)$$

The resulting anomaly scores  $S$  are used to guide keyframe selection, while the predicted anomaly classes provide semantic cues for downstream reasoning by LLMs.

#### 3.1.2. Context-AwarE Sampling (CAES)

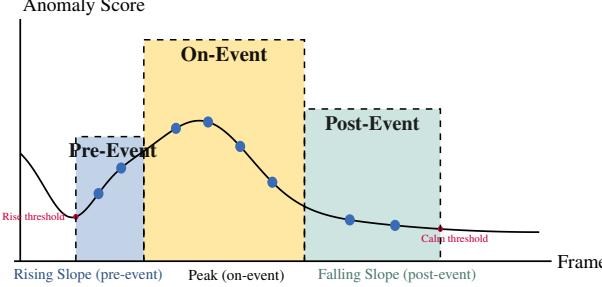
To capture the narrative structure of anomalous events, we introduce CAES, a Context-AwarE keyframe Sampling strategy that preserves both causal context and event diversity. Formally,

$$K = \text{CAES}(S), \quad (3)$$

where  $K$  is the set of selected keyframes and  $S$  is the anomaly score sequence.

As illustrated in Figure 3, CAES first detects all anomalous intervals using a percentile-based adaptive threshold for each video, then automatically expands each event into pre-event and post-event context segments by applying rise and calm thresholds to the slope of anomaly scores. Keyframes are then uniformly sampled from pre-event, on-event, and post-event segments to cover the causal lead-up, climax, and aftermath of each anomaly, ensuring a comprehensive yet compact representation.

If multiple anomalies are detected or the total exceeds the frame budget (e.g., 64), frames with the highest anomaly scores are prioritized. Any remaining slots are filled by



**Figure 3. Illustration of CAES keyframe selection strategy.** The anomaly score curve is segmented into pre-event (blue), on-event (yellow), and post-event (green) intervals. Blue dots are sampled keyframes, and red dots indicate rise and calm thresholds.

sampling from background segments to ensure temporal coherence. This approach ensures a comprehensive yet compact representation for downstream narrative reasoning.

### 3.2. Relation Extraction and Integration

While CAES identifies what and when to analyze, the core of our VADER framework lies in its ability to understand how the relationships between objects dynamically evolve. Moving beyond static scene description, VADER first extracts rich object and relational features from each selected frame, then automatically discovers temporal patterns of interaction using only video-level labels. Finally, it distills these patterns into compact dynamic representations and integrates them into the main LLM for deep, causally-grounded reasoning about complex events.

#### 3.2.1. Relational Feature Extraction

For each keyframe selected by CAES, we utilize a pre-trained DETR-based scene graph generation model [17] as a feature extractor to extract appearance embeddings for detected objects and a relational tensor. This tensor contains the intermediate features computed for each object pair before the relationship classification head, serving as a comprehensive snapshot of the scene’s relational dynamics.

To maintain consistent object identities across frames, we perform object association using a combination of IoU overlap and cosine similarity of appearance embeddings. To capture temporal changes in these interactions, we compute a relational volatility curve for each video by tracking objects and, for each pair of adjacent frames, taking the maximum  $L_2$  norm of the difference between relational features for all co-tracked object pairs:

$$\text{Volatility}(t) = \max_{(i,j)} \|\mathbf{r}_{ij}(t) - \mathbf{r}_{ij}(t-1)\|_2, \quad (4)$$

This sequence of volatility scores, together with the associated before/after relational vectors, is cached for efficient downstream processing.

#### 3.2.2. COntrative Relation Encoder (CORE)

Based on the relational volatility curves, we automatically mine and encode salient patterns of object interaction change in a weakly-supervised manner. To ensure robust peak detection, each volatility curve is first smoothed with a Gaussian filter. For abnormal videos, relational-change pairs, i.e.,  $[\mathbf{r}_{ij}(t-1); \mathbf{r}_{ij}(t)]$ , corresponding to the top  $k\%$  of these smoothed peaks are designated as positive samples, while pairs from volatility valleys and from normal videos serve as negatives.

To distill these transitions into compact representations, we train a lightweight COntrative Relation Encoder (CORE) that maps each relational-change pair to compact relation tokens. To train CORE, we employ a triplet margin loss, formulated as

$$\mathcal{L}_{\text{triplet}} = \max (0, d(f_a, f_p) - d(f_a, f_n) + \alpha), \quad (5)$$

where  $f_a$ ,  $f_p$ , and  $f_n$  denote the anchor, positive, and negative samples encoded by CORE,  $d(\cdot, \cdot)$  is the L2 distance, and  $\alpha$  is the margin. We further adopt semi-hard negative mining [43] to improve discriminative power.

### 3.3. LLM Integration and Fine-tuning

For each video, keyframes selected by CAES are encoded into visual tokens and passed through the Relation Feature Extractor to obtain raw relational tensors, from which relation tokens are generated by CORE. Task instructions are converted into text tokens. All tokens are concatenated and jointly fed into the Multimodal LLM, enabling the model to reason over both visual content and the evolving object relationships that underlie complex events.

During training, we fine-tune only the multimodal projectors and LoRA [15] adapters, keeping the backbone frozen. This setup allows VADER to align both static and dynamic video cues with causal, semantically rich natural language output. As a result, the model can generate detailed, context-aware anomaly descriptions and perform robust anomaly-related reasoning.

**Implementation Details.** Including hyperparameters, threshold values, and training configurations, are provided in the Suppl. Material Sec. G.

## 4. Experiments

In this section, we present the experimental results of applying the proposed VADER model to three latest comprehensive benchmark datasets commonly used for evaluating VAU methods. We also compare the experimental results with the SOTA methods and discuss some ablation studies on the proposed method.

### 4.1. Benchmark Datasets and Evaluation Metrics

To evaluate the validity of VADER, we use three latest comprehensive VAU benchmarks: HIVAU-70k [67], HAWK

| Benchmark | BLEU          | ROUGE | CIDEr | METEOR | MoverScore     | GPT Score   | MMEval |
|-----------|---------------|-------|-------|--------|----------------|-------------|--------|
|           | Lexical-level |       |       |        | Semantic-level | Judge-based |        |
| HIVAU-70k | ✓             | ✓     | ✓     | ✓      |                |             |        |
| HAWK      | ✓             |       |       |        |                | ✓           |        |
| CUVA      | ✓             | ✓     |       |        | ✓              |             | ✓      |

Table 1. Evaluation metrics for each benchmark. Metric types are grouped into lexical-level (blue), semantic-level (yellow), and judge-based (green).

[47], and CUVA[11].

**HIVAU-70k** [67] is a large-scale dataset comprising videos from UCF-Crime [55] and XD-Violence [46], providing hierarchical annotations at clip, event, and video levels for both perception and reasoning tasks.

**HAWK** [47] focuses on open-world video anomaly understanding, collecting videos from seven diverse datasets [2, 7, 27, 33, 46, 52, 59], and supports both event description and question answering tasks.

**CUVA** [11] targets causation understanding in real-world anomaly videos, offering detailed annotations that explain what happened, why, and how for each event.

We adopt the evaluation metrics following the official protocols of each benchmark. These can be categorized into three groups: **(1) Lexical-level metrics**, including BLEU [39], ROUGE [25], METEOR [4], and CIDEr [49], measure n-gram overlap, precision, recall, and synonym matching with reference texts. **(2) Semantic-level metrics**, such as BLEURT [44] and MoverScore [70], use pretrained language models to capture meaning and contextual similarity beyond exact word matches. **(3) Judge-based evaluative metrics**, including GPT score [1] and MMEval [11], directly use LLMs or VLMs as evaluators to judge plausibility, informativeness, and consistency of the generated explanations with video content, simulating human judgment in open-ended settings.

All metrics are reported such that higher scores indicate better performance. Table 1 summarizes the metrics used for each benchmark.

## 4.2. Experimental Results

We present quantitative evaluations of VADER across three challenging benchmarks, highlighting its robust and superior performance.

**Performance on CUVA.** Table 2 demonstrates VADER’s competitive performance across anomaly description and causation tasks. VADER achieves the highest MMEval score for ”Causes” task with an improvement of 7.38 points and maintains a comparable score on the other tasks. While VADER’s BLEURT scores are slightly lower than some baselines, this is due to BLEURT’s focus on surface-level wording similarity. In contrast, VADER excels on human-aligned metrics such as UniEval and

| Method                          | Metric     | Description  | Causes       | Effect       |
|---------------------------------|------------|--------------|--------------|--------------|
| mPLUG-owl [61]                  | BLEU       | 0.55         | 0.65         | 0.47         |
|                                 | ROUGE      | 12.58        | 13.54        | 8.83         |
|                                 | BLEURT     | 40.66        | 43.28        | 37.95        |
|                                 | MoverScore | 51.97        | 52.71        | 50.06        |
|                                 | UniEval    | 67.46        | 62.29        | <u>59.07</u> |
|                                 | MMEval     | 73.42        | 17.15        | 44.31        |
| Video-LLAMA [66]                | BLEU       | 0.60         | 0.53         | 0.35         |
|                                 | ROUGE      | 13.15        | 12.36        | 8.02         |
|                                 | BLEURT     | 40.55        | 43.02        | 39.68        |
|                                 | MoverScore | 51.32        | 51.25        | 49.48        |
|                                 | UniEval    | 52.28        | 47.29        | 43.03        |
|                                 | MMEval     | 65.65        | 16.24        | 32.84        |
| PandaGPT [45]                   | BLEU       | 0.66         | 0.51         | 0.30         |
|                                 | ROUGE      | 13.33        | 14.09        | 8.79         |
|                                 | BLEURT     | 38.23        | 43.95        | <u>39.95</u> |
|                                 | MoverScore | 51.73        | 51.54        | 49.62        |
|                                 | UniEval    | 57.05        | 54.88        | 50.84        |
|                                 | MMEval     | 74.19        | 22.47        | <b>69.45</b> |
| Otter [20]                      | BLEU       | 1.07         | 1.09         | <b>1.11</b>  |
|                                 | ROUGE      | <b>15.19</b> | <u>15.87</u> | <u>11.40</u> |
|                                 | BLEURT     | 29.92        | 32.52        | 28.94        |
|                                 | MoverScore | <b>53.54</b> | <b>54.25</b> | <u>51.91</u> |
|                                 | UniEval    | 45.14        | 49.05        | 47.51        |
|                                 | MMEval     | 76.30        | 3.53         | 39.21        |
| Video-ChatGPT [36]              | BLEU       | 0.30         | 0.29         | 0.41         |
|                                 | ROUGE      | 9.75         | 9.08         | 8.23         |
|                                 | BLEURT     | <u>46.83</u> | <b>49.52</b> | 37.24        |
|                                 | MoverScore | 50.73        | 50.70        | 49.83        |
|                                 | UniEval    | <u>70.82</u> | <u>70.77</u> | 54.35        |
|                                 | MMEval     | 78.55        | 44.57        | 46.08        |
| CUVA* [11]                      | BLEU       | 0.55         | 0.51         | 0.38         |
|                                 | ROUGE      | <u>14.35</u> | 9.08         | 8.23         |
|                                 | BLEURT     | <b>47.10</b> | <u>48.13</u> | <b>48.28</b> |
|                                 | MoverScore | 52.25        | 52.28        | 49.95        |
|                                 | UniEval    | 68.18        | 63.41        | 51.87        |
|                                 | MMEval     | <b>79.65</b> | <u>58.92</u> | 50.64        |
| <b>VADER<sup>†</sup> (Ours)</b> | BLEU       | <b>1.09</b>  | <b>1.19</b>  | <b>1.11</b>  |
|                                 | ROUGE      | 12.85        | <b>16.84</b> | <b>17.38</b> |
|                                 | BLEURT     | 32.34        | 37.48        | 37.60        |
|                                 | MoverScore | <u>52.65</u> | <u>52.78</u> | <b>53.21</b> |
|                                 | UniEval    | <b>85.00</b> | <b>79.37</b> | <b>82.06</b> |
|                                 | MMEval     | 78.89        | <b>66.30</b> | <u>63.26</u> |

Table 2. **Evaluation on the CUVA benchmark.** VADER shows competitive results across anomaly description and causation. **Bold** indicate the best performance, and *underlined* indicate the second best. Models marked with <sup>†</sup> are finetuned on CUVA; those marked with \* employ prompt or adapter tuning; all others are evaluated without domain-specific finetuning.

MMEval, indicating its strength in generating coherent and accurate causal explanations beyond lexical overlap.

**Performance on HIVAU-70k.** As shown in Table 3, VADER achieves state-of-the-art results across all metrics and evaluation levels (clip, event, and video). The per-

| Method                       | BLEU         |              |              | CIDEr        |              |              | METEOR       |              |              | ROUGE        |              |              |
|------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                              | C            | E            | V            | C            | E            | V            | C            | E            | V            | C            | E            | V            |
| Video-ChatGPT [36]           | 0.152        | 0.068        | 0.066        | 0.033        | 0.011        | 0.013        | 0.102        | 0.069        | 0.044        | 0.153        | 0.048        | 0.079        |
| Video-LLAMA [66]             | 0.151        | 0.079        | 0.104        | 0.024        | 0.014        | 0.017        | 0.112        | 0.076        | 0.057        | 0.156        | 0.067        | 0.090        |
| Video-LLAVA [24]             | 0.164        | 0.046        | 0.055        | 0.032        | 0.009        | 0.013        | 0.097        | 0.022        | 0.014        | 0.132        | 0.023        | 0.045        |
| LLAVA-Next-Video [69]        | 0.435        | 0.091        | 0.120        | 0.102        | 0.015        | 0.031        | 0.117        | 0.085        | 0.096        | 0.198        | 0.080        | 0.106        |
| QwenVL2 [51]                 | 0.312        | 0.082        | 0.155        | 0.044        | 0.020        | 0.044        | 0.133        | 0.092        | 0.112        | 0.163        | 0.081        | 0.137        |
| InternVL2 [9]                | 0.331        | 0.101        | 0.145        | 0.052        | 0.022        | 0.035        | 0.141        | 0.095        | 0.101        | 0.182        | 0.102        | 0.122        |
| NVILA [20]                   | 0.610        | 0.340        | 0.283        | 0.261        | 0.154        | 0.098        | 0.157        | 0.096        | 0.072        | 0.273        | 0.218        | 0.198        |
| Holmes-VAU <sup>‡</sup> [67] | 0.913        | 0.804        | 0.566        | 0.467        | 1.519        | <u>1.437</u> | 0.190        | 0.165        | 0.121        | 0.329        | 0.370        | 0.355        |
| VADER <sup>‡</sup> (Ours)    | <u>1.035</u> | <u>1.163</u> | <u>1.068</u> | <u>0.612</u> | <u>1.650</u> | 1.403        | <u>0.215</u> | <u>0.183</u> | <u>0.137</u> | <u>0.376</u> | <u>0.444</u> | <u>0.408</u> |
| VADER <sup>†</sup> (Ours)    | <b>1.266</b> | <b>1.246</b> | <b>1.268</b> | <b>1.040</b> | <b>1.763</b> | <b>1.812</b> | <b>0.247</b> | <b>0.216</b> | <b>0.164</b> | <b>0.429</b> | <b>0.463</b> | <b>0.446</b> |

Table 3. **Evaluation on the HIVAU-70k benchmark.** The dataset supports hierarchical evaluation at the clip (C), event (E), and video (V) levels. VADER achieves consistently strong results across metrics and levels. Models marked with <sup>†</sup> are finetuned on HIVAU-70k; those marked with <sup>‡</sup> are finetuned and use the same backbone size as baseline [67]; all others are evaluated without domain-specific finetuning.

| Method                    | Anomaly Video Description Generation |              |              |              |              |              |              | Anomaly Video Question-Answering |              |              |              |              |              |              |
|---------------------------|--------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|----------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                           | Text-Level                           |              |              |              | GPT-Guided   |              |              | Text-Level                       |              |              |              | GPT-Guided   |              |              |
|                           | BLEU-1                               | BLEU-2       | BLEU-3       | BLEU-4       | Reason.      | Detail       | Consist.     | BLEU-1                           | BLEU-2       | BLEU-3       | BLEU-4       | Reason.      | Detail       | Consist.     |
| Video-ChatGPT [36]        | 0.107                                | 0.046        | 0.017        | 0.008        | 0.084        | 0.108        | 0.055        | 0.177                            | 0.096        | 0.058        | 0.038        | 0.508        | 0.430        | 0.421        |
| VideoChat [22]            | 0.053                                | 0.023        | 0.008        | 0.003        | 0.107        | 0.205        | 0.054        | 0.261                            | 0.133        | 0.074        | 0.043        | 0.699        | 0.631        | 0.598        |
| Video-LLAMA [66]          | 0.062                                | 0.025        | 0.009        | 0.004        | 0.120        | 0.217        | 0.066        | 0.156                            | 0.081        | 0.045        | 0.027        | 0.586        | 0.485        | 0.497        |
| LLAMA-Adapter [68]        | 0.132                                | 0.052        | 0.018        | 0.008        | 0.060        | 0.091        | 0.038        | 0.199                            | 0.109        | 0.067        | 0.043        | 0.646        | 0.559        | 0.549        |
| Video-LLAVA [24]          | 0.071                                | 0.030        | 0.012        | 0.005        | 0.077        | 0.115        | 0.038        | 0.094                            | 0.054        | 0.034        | 0.023        | 0.393        | 0.274        | 0.316        |
| NVILA [20]                | 0.132                                | 0.058        | 0.016        | 0.003        | 0.355        | <b>0.527</b> | <u>0.299</u> | 0.003                            | 0.001        | 0.001        | 0.001        | 0.428        | 0.304        | 0.362        |
| HAWK <sup>†</sup> [47]    | 0.270                                | 0.139        | 0.074        | 0.043        | 0.283        | 0.320        | 0.218        | 0.319                            | 0.179        | 0.112        | 0.073        | <b>0.840</b> | 0.794        | 0.753        |
| VADER <sup>†</sup> (Ours) | <b>0.324</b>                         | <b>0.196</b> | <b>0.127</b> | <b>0.071</b> | <b>0.428</b> | <u>0.442</u> | <b>0.357</b> | <b>0.484</b>                     | <b>0.311</b> | <b>0.210</b> | <b>0.150</b> | <b>0.828</b> | <b>0.825</b> | <b>0.794</b> |

Table 4. **Evaluation on the HAWK benchmark.** VADER achieves leading results in both anomaly description generation and question-answering tasks across text-level and GPT-guided metrics. **Bold** indicate the best performance, and underlined indicate the second best. Models marked with <sup>†</sup> are finetuned on HAWK; all others are evaluated without domain-specific finetuning.

formance gains are particularly significant at the event and video levels, highlighting VADER’s ability to reason over complex, temporally extended anomalous events and generate coherent, causally-aware narratives.

**Performance on HAWK.** Results presented in Table 4 illustrate VADER’s strong performance in anomaly description and question-answering tasks. For description generation, VADER surpasses previous best models by 0.045 in BLEU-1 and 0.026 in BLEU-4. For the question-answering task, VADER demonstrates even greater improvements, outperforming prior methods by 0.164 in BLEU-1 and 0.077 in BLEU-4. Additionally, it achieves leading GPT-guided scores, notably in Detail (description task) and Consistency (QA task), underscoring its ability to generate detailed, plausible, and contextually consistent narratives.

These results underscore VADER’s versatility and high performance across diverse evaluation scenarios.

### 4.3. Ablation Study

We conduct ablation studies on the HAWK [47] description task to systematically assess the contribution of each com-

ponent in VADER.

**Impact of Key Components.** As shown in Table 5, removing the relation reasoning branch, CORE, leads to a consistent drop in both text-level and GPT-guided metrics, confirming the importance of modeling object interactions. Disabling the proposed context-aware sampling strategy, CAES, also degrades performance across all metrics, further validating its effectiveness. Without fine-tuning, performance drops significantly, though the GPT-guided detail score is anomalously high, likely because the model generates lengthy but less focused responses due to the lack of domain adaptation.

**Effect of Sampling Strategy.** Table 6 compares our CAES with alternative keyframe selection methods. Uniform sampling results in lower BLEU and ROUGE scores, as it may overlook key anomalous frames. Top-K focuses on frames with the highest anomaly scores, often over-concentrating on anomalous segments and neglecting essential context. The ATS method [67] adaptively samples frames based on anomaly scores using a density-aware approach; however, in our setting, it shows only marginal

| Method          | Text-Level   |              | GPT-Guided    |              |              |
|-----------------|--------------|--------------|---------------|--------------|--------------|
|                 | BLEU         | ROUGE        | Reasonability | Detail       | Consistency  |
| <b>VADER</b>    | <b>0.718</b> | <b>0.283</b> | <b>0.428</b>  | 0.442        | <b>0.357</b> |
| w/o CORE        | 0.668        | 0.274        | 0.419         | 0.477        | 0.343        |
| w/o CAES        | 0.594        | 0.244        | 0.387         | 0.435        | 0.331        |
| w/o Fine-Tuning | 0.209        | 0.143        | 0.355         | <b>0.527</b> | 0.299        |

Table 5. **Ablation study on key components of VADER.** Removing either CORE, CAES, or fine-tuning leads to performance drops in both text-level and GPT-guided metrics, underscoring the importance of each module to the overall performance.

| Context Sampling Method | Text-Level   |              | GPT-Guided    |              |              |
|-------------------------|--------------|--------------|---------------|--------------|--------------|
|                         | BLEU         | ROUGE        | Reasonability | Detail       | Consistency  |
| Fixed window            | 0.647        | 0.273        | 0.419         | 0.477        | 0.343        |
| Exponential interval    | 0.665        | <b>0.278</b> | 0.392         | 0.421        | 0.342        |
| <b>Dynamic window</b>   | <b>0.668</b> | 0.274        | <b>0.419</b>  | <b>0.477</b> | <b>0.343</b> |

Table 7. **Ablation study on context sampling strategies in CAES.** Dynamic window sampling achieves the best overall performance across both text-level and GPT-guided metrics.

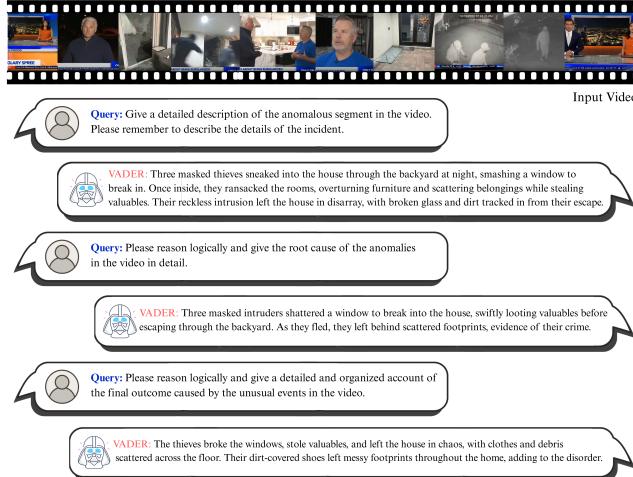


Figure 4. **Qualitative results of VADER’s causal reasoning capabilities.** Given a video depicting a nighttime break-in, VADER generates detailed, context-aware answers for a sequence of causal reasoning queries, capturing both the key actions and underlying cause-effect chains within the event.

gains over Top-K on text-level metrics, likely due to limited context diversity in complex scenarios. Our CAES achieves the best trade-off across both text-level and GPT-guided metrics, effectively balancing anomaly saliency and context diversity to produce coherent event narratives.

**Effect of Context Sampling Strategies.** Table 7 compares different approaches for selecting pre/post-event con-

| Method             | Text-Level   |              | GPT-Guided    |              |              |
|--------------------|--------------|--------------|---------------|--------------|--------------|
|                    | BLEU         | ROUGE        | Reasonability | Detail       | Consistency  |
| Uniform            | 0.594        | 0.244        | 0.387         | 0.435        | 0.331        |
| Top-K              | 0.663        | 0.273        | 0.381         | 0.411        | 0.327        |
| ATS [67]           | 0.641        | 0.273        | 0.383         | 0.423        | 0.335        |
| <b>CAES (Ours)</b> | <b>0.668</b> | <b>0.274</b> | <b>0.419</b>  | <b>0.477</b> | <b>0.343</b> |

Table 6. **Comparison of keyframe selection strategies.** Our Context-Aware Sampling (CAES) outperforms uniform, Top-K, and ATS approaches across text-level and GPT-guided metrics.

| Method                | Text-Level   |              | GPT-Guided    |              |              |
|-----------------------|--------------|--------------|---------------|--------------|--------------|
|                       | BLEU         | ROUGE        | Reasonability | Detail       | Consistency  |
| Relational visual cue | 0.670        | 0.276        | 0.385         | 0.430        | 0.326        |
| Scene graph text      | 0.686        | 0.277        | 0.391         | 0.434        | 0.335        |
| <b>Relation token</b> | <b>0.718</b> | <b>0.283</b> | <b>0.428</b>  | <b>0.442</b> | <b>0.357</b> |

Table 8. **Comparison of relation representations.** Using relation tokens as relation representations outperforms both relational visual cue and scene graph text approaches across all text-level and GPT-guided metrics.

text frames in CAES. Fixed Window samples consecutive frames immediately before the anomaly, which may lead to high redundancy and limited temporal scope. Exponential Interval samples at increasing intervals to provide multi-scale context, but may still miss critical transitions. Our Dynamic Window adaptively determines context boundaries based on anomaly score gradients, enabling event-specific, informative context coverage. The results show that Dynamic Window achieves strong performance across metrics, highlighting the value of adaptive, narrative-aware context selection.

**Relation Representation.** Table 8 compares different strategies for encoding relational information. Using only relational visual cues encodes the detected objects’ static visual features and locations, but fails to capture interactions or evolving relationships between objects. Scene graph text descriptions leverage the relationship labels predicted by the pretrained EGTR [17] on Visual Genome [18], but these generic labels are not tailored for the nuanced, context-specific interactions common in anomaly scenarios. In contrast, our relation token, generated by CORE, provides continuous and fine-grained temporal cues, yielding consistently better results across all metrics.

#### 4.4. Qualitative Results

Figure 4 illustrates a qualitative example that highlights VADER’s strengths in causal reasoning. Given a video of a nighttime break-in, VADER offers a detailed description, root cause, and outcome summary of the anomalous event.

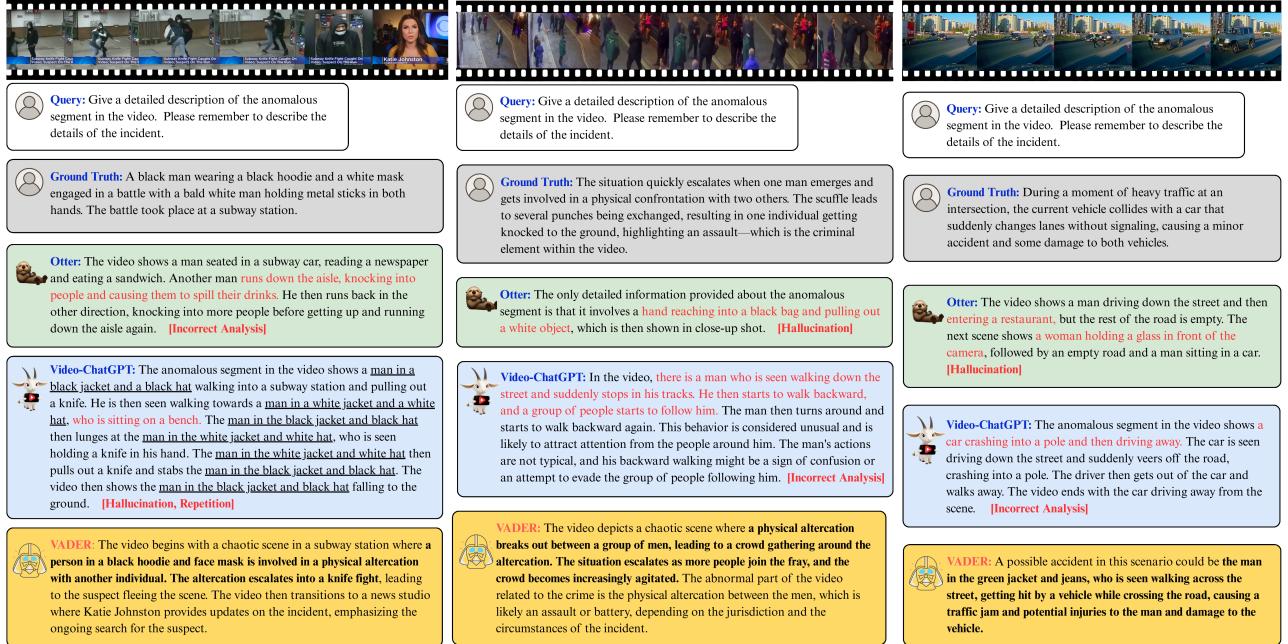


Figure 5. Three examples of the task of describing anonymous videos are depicted here. The descriptions generated by Otter [20] and Video-ChatGPT [36] contain hallucination or incorrect analysis. In contrast, VADER produces concise, contextually grounded, and causally coherent descriptions that accurately reflect the events and their underlying dynamics across various challenging cases.

Its responses capture not only key visual cues such as the broken window and scattered belongings, but also the underlying object-level interactions, such as the intruders’ actions in the environment, demonstrating the benefit of our keyframe selection and relational modeling.

Figure 5 further compares VADER with existing MLLMs [20, 36] on diverse real-world anomalous videos. Across various scenarios, physical altercations, road accidents, and complex crowd behaviors, VADER produces more accurate, context-aware, and semantically coherent descriptions than recent baselines. These results illustrate VADER’s ability to mitigate hallucinations, capture causal structure, and provide comprehensive explanations, thereby advancing explainable video anomaly understanding.

We provide additional qualitative results and failure cases in the Supp. Material Sec. F and C.

## 5. Limitation

While VADER advances video anomaly understanding, several limitations remain:

**Dependency on Upstream Models.** VADER relies on upstream modules such as the Anomaly Scorer and Relation Feature Extractor, where errors in detection or association can propagate and affect the final reasoning results.

**Inherent Bias towards High-Motion Events.** The reliance on relational volatility as the core signal introduces a bias toward anomalies with strong motion, making the framework less sensitive to subtle or low-motion anomalies.

**Limitation to Object-Centric View.** VADER’s current design is object-centric, which limits its ability to capture scene-level or group-level anomalies such as environmental changes or collective crowd behaviors.

Detailed potential mitigation strategies are discussed in Supp. Material Sec. D.

## 6. Conclusion

In this work, we presented VADER, a novel framework for video anomaly understanding that leverages Context-Aware Sampling (CAES) and dynamic relational modeling to deliver detailed, causally grounded interpretations of anomalous events. Extensive experiments on multiple challenging benchmarks demonstrate that VADER achieves state-of-the-art or highly competitive performance across anomaly description, causal explanation, and question answering tasks. Systematic ablation studies further highlight the critical roles of adaptive context sampling and relational dynamics encoding.

**Acknowledgement** This work is supported by the NVIDIA Taiwan AI Research & Development Center (TRDC).

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5
- [2] Andra Acsintoa, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ub-normal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20143–20153, 2022. 2, 5
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2
- [4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5
- [5] Yunkang Cao, Xiaohao Xu, Yuqi Cheng, Chen Sun, Zongwei Du, Liang Gao, and Weiming Shen. Personalizing vision-language models with hybrid prompts for zero-shot anomaly detection. *IEEE Transactions on Cybernetics*, 2025. 2
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2
- [7] Antoni B Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):909–926, 2008. 2, 5
- [8] Chenglizhao Chen, Xinyu Liu, Mengke Song, Luming Li, Shaojiang Yuan, Xu Yu, and Shanchen Pang. Unveiling context-related anomalies: Knowledge graph empowered decoupling of scene and action for human-related video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2
- [9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 6
- [10] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International symposium on neural networks*, pages 189–196. Springer, 2017. 2
- [11] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18793–18803, 2024. 1, 2, 5, 12, 13
- [12] Shibo Gao, Peipei Yang, and Linlin Huang. Suvad: Semantic understanding based video anomaly detection using mllm. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2
- [13] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12742–12752, 2021. 1
- [14] Mahmudul Hasan, Jonghyun Choi, jan Neumann, Amit K Roy-Chowdhury, and Larry Davis. Learning temporal regularity in video sequences. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2016. 2
- [15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4, 15
- [16] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 1, 2
- [17] Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. Egtr: Extracting graph from transformer for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24229–24238, 2024. 2, 4, 7
- [18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 7
- [19] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 15
- [20] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2, 5, 6, 8, 15
- [21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

- [22] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 6
- [23] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. 2
- [24] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning unified visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6
- [25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2
- [27] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection—a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 2, 5
- [28] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12147–12156, 2023. 1
- [29] Zhan Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13588–13597, 2021. 1
- [30] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 1, 2
- [31] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4122–4134, 2025. 1, 2, 14
- [32] Aaron Lohner, Francesco Compagno, Jonathan Francis, and Alessandro Oltramari. Enhancing vision-language models with scene graphs for traffic accident understanding. In *2024 IEEE International Automated Vehicle Validation Conference (IAVVC)*, pages 1–7. IEEE, 2024. 2
- [33] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 2, 5
- [34] Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024. 1, 2
- [35] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8022–8031, 2023. 2
- [36] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 1, 2, 5, 6, 8
- [37] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE conference on computer vision and pattern recognition*, pages 935–942. IEEE, 2009. 2
- [38] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 2021. 2
- [39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5
- [40] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020. 1
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [42] Venkatesh Saligrama and Zhu Chen. Video anomaly detection based on local statistical aggregates. In *2012 IEEE Conference on computer vision and pattern recognition*, pages 2112–2119. IEEE, 2012. 2
- [43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4, 15
- [44] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020. 5
- [45] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 5
- [46] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 1, 2, 5
- [47] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. In *Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 5, 6, 14
- [48] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

- [49] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5
- [50] Hao Wang, Jiayou Qin, Ashish Bastola, Xiwen Chen, John Suchanek, Zihao Gong, and Abolfazl Razi. Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation. *arXiv preprint arXiv:2403.12415*, 2024. 1
- [51] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6
- [52] Shu Wang and Zhenjiang Miao. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*, pages 1220–1223. IEEE, 2010. 2, 5
- [53] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE transactions on neural networks and learning systems*, 33(6):2301–2312, 2021. 1
- [54] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer, 2022. 1
- [55] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, ZhaoYang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pages 322–339. Springer, 2020. 5
- [56] Ruiyao Xu and Kaize Ding. Large language models for anomaly and out-of-distribution detection: A survey. *arXiv preprint arXiv:2409.01980*, 2024. 2
- [57] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2
- [58] Zhiwei Yang, Chen Gao, Jing Liu, Peng Wu, Guansong Pang, and Mike Zheng Shou. Assistpda: An online video surveillance assistant for video anomaly prediction, detection, and analysis. *arXiv preprint arXiv:2503.21904*, 2025. 2
- [59] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 444–459, 2022. 2, 5
- [60] Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8679–8688, 2025. 2
- [61] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 5
- [62] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges, 2023. 2
- [63] Andrei Zaharescu and Richard Wildes. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In *European Conference on Computer Vision*, pages 563–576. Springer, 2010. 2
- [64] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into clip latent space for video anomaly recognition. *Computer Vision and Image Understanding*, 249:104163, 2024. 3
- [65] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024. 2
- [66] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 5, 6
- [67] Huixin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13843–13853, 2025. 1, 2, 4, 5, 6, 7, 12, 13
- [68] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Ao-jun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 6
- [69] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 6
- [70] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*, 2019. 5
- [71] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017. 1, 2
- [72] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3769–3777, 2023. 2, 3
- [73] Yixuan Zhou, Yi Qu, Xing Xu, Fumin Shen, Jingkuan Song, and Heng Tao Shen. Batchnorm-based weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 3
- [74] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

## Supplementary Material

|  |           |
|--|-----------|
| <b>A Illustrative Example of Motivation</b>                  | <b>12</b> |
| <b>B Ablation Study on Sample Mining Hyperparameters</b>     | <b>12</b> |
| <b>C Failure Cases</b>                                       | <b>12</b> |
| <b>D Potential Mitigations for Limitations</b>               | <b>13</b> |
| D.1. Mitigating Dependency on Upstream Modules               | 13        |
| D.2 Reducing Bias Toward High-Motion Events .                | 13        |
| D.3 Moving Beyond Object-Centric Reasoning .                 | 13        |
| <b>E Computational Efficiency</b>                            | <b>14</b> |
| <b>F. Intermediate Visualization of Module Contributions</b> | <b>14</b> |
| <b>G Implementation Details</b>                              | <b>15</b> |

### A. Illustrative Example of Motivation

To better illustrate the motivation behind our work, Figure 6 presents a real-world anomalous event where a dog suddenly attacks a boy walking on the roadside. This example demonstrates the challenges in understanding not only what happens in the scene, but also how it occurs, which requires modeling both causal relationships and dynamic object interactions.

In this case, different methods generate diverse outputs when asked to describe the anomalous segment:

**CUVA** [11] focuses on basic visual elements but produces a description that does not align with the actual event. While certain textual metrics such as ROUGE are relatively high, the description lacks factual accuracy and fails to capture the core abnormal interaction.

**Holmes-VAU** [67] includes some relevant context, such as the presence of danger and a dog, but its narrative remains ambiguous. The description does not clearly reflect the cause-effect chain or the temporal progression of the event, limiting interpretability.

**VADER (Ours)** explicitly models the interactions and causal dynamics between objects, resulting in a coherent description that clearly states who was involved and how the anomaly unfolded. This produces outputs that align well with human judgments, as shown by higher semantic-level and human-evaluation metrics (e.g., BLEURT, UniEval, and mmEval).

The quantitative metrics and qualitative feedback below each output further highlight the importance of deeper reasoning. While lexical overlap metrics alone may not

fully capture factual correctness, human-aligned evaluations reflect the clarity and interpretability of the descriptions. VADER achieves the highest ratings by providing a precise and logically structured explanation of the anomalous event.

### B. Ablation Study on Sample Mining Hyperparameters

We perform an ablation study on the Gaussian smoothing parameter ( $\sigma$ ) and the peak selection threshold (top- $k$  percentile) in our weakly-supervised sample mining pipeline. As shown in Table 9, we report AUC scores on the test set for different combinations of  $\sigma$  and top- $k$ .

Our results reveal two main findings. First, with  $\sigma = 2.0$ , a top- $k$  percentile of 5% yields the best AUC (72.12). Increasing the threshold to 7% slightly reduces performance, likely due to the inclusion of less informative (noisy) samples. Conversely, a stricter threshold of 3% also degrades performance, suggesting that overly selective sampling yields too few positives for effective training. Second, across all thresholds,  $\sigma = 2.0$  consistently outperforms both smaller ( $\sigma = 1.0$ ) and larger ( $\sigma = 3.0$ ) values. We attribute this to improved noise filtering at  $\sigma = 2.0$  without excessive smoothing that would obscure salient events.

In summary,  $\sigma = 2.0$  and top- $k = 5\%$  strike the best balance between sample quality and quantity, and are used in all subsequent experiments.

| Smooth Sigma ( $\sigma$ ) | Top-k Percentile (%) |              |       |
|---------------------------|----------------------|--------------|-------|
|                           | 3.0%                 | 5.0%         | 7.0%  |
| 1.0                       | 69.53                | 70.18        | 68.91 |
| 2.0                       | 70.85                | <b>72.12</b> | 71.64 |
| 3.0                       | 68.17                | 69.25        | 68.55 |

Table 9. **Ablation study on sample mining hyperparameters.** Test set AUC (%) for different combinations of Gaussian smoothing parameter ( $\sigma$ ) and top- $k$  percentile thresholds used in peak selection.  $\sigma = 2.0$  and top- $k = 5\%$  achieve the best performance, highlighting the importance of balancing sample quality and quantity in our weakly-supervised mining strategy.

### C. Failure Cases

Figure 7 presents three examples highlighting VADER’s tendency to favor visually dynamic events. While VADER accurately captures prominent high-motion activities, it often overlooks the underlying causes of anomalies or neglects subtle contextual cues, such as unattended objects or gradual environmental changes. For instance, in the middle example, VADER correctly identifies the car collision but misses the key causal factor, the vehicle ignoring the traf-

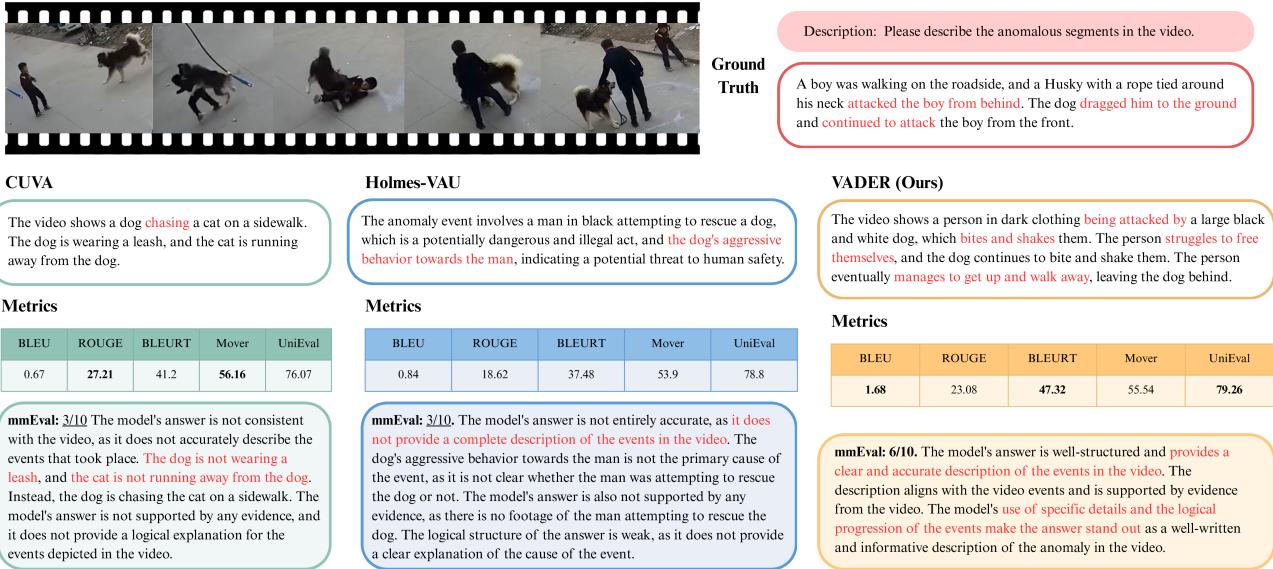


Figure 6. **Illustrative example showing the need for causal and relational modeling.** Given the same video of a dog attacking a boy, CUVA [11] and Holmes-VAU [67] produce incorrect or incomplete descriptions, while VADER captures key interactions and event progression, resulting in accurate and coherent descriptions with higher human-aligned evaluation scores.

fic signal, resulting in an incomplete explanation of why the anomaly occurred.

Figure 8 presents three examples highlighting VADER’s limitation to object-centric reasoning. While VADER accurately captures localized actions and pairwise object interactions, it often fails to represent the broader scene context or collective behaviors. For instance, in the left example, VADER correctly describes individual pedestrians entering and exiting the subway but misses the group of people gathered to watch a street performance, resulting in an incomplete understanding of the true anomaly.

## D. Potential Mitigations for Limitations

To address the three primary limitations discussed in Sec. 5, we outline potential mitigation strategies. These strategies aim to enhance VADER’s robustness, reduce bias, and broaden its capacity to handle diverse anomaly scenarios.

### D.1. Mitigating Dependency on Upstream Modules

VADER relies on the performance of upstream modules, which may cause errors to propagate and impact the final reasoning results. To reduce this dependency and improve the reliability of the pipeline, we consider the following approaches to strengthen upstream components and improve their alignment with downstream tasks:

**End-to-End Joint Training.** Transform the current pipeline into a jointly trainable framework, allowing gradients from the LLM to update the upstream modules so they can better align with the final reasoning objectives.

**Module Robustness Enhancement.** Before end-to-end integration, each module can be strengthened individually by training on more diverse data and employing more advanced tracking algorithms, thereby reducing upstream errors and improving overall robustness

### D.2. Reducing Bias Toward High-Motion Events

VADER’s reliance on relational volatility can lead to a bias toward anomalies involving strong motion while overlooking subtle anomalies. To address this imbalance, additional cues can be incorporated to complement volatility:

**Object State Modeling.** Track and interpret object states, such as transitions from “carried” to “stationary” without an owner, to capture static anomalies like abandoned objects.

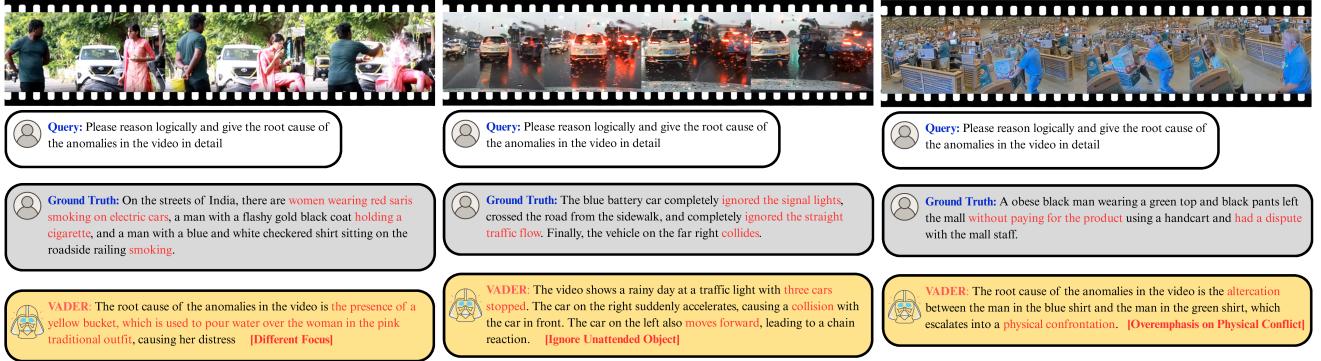
**Global Scene Context.** Model typical activity flow within the scene and identify deviations, enabling the detection of subtle anomalies such as loitering or suspicious inactivity.

### D.3. Moving Beyond Object-Centric Reasoning

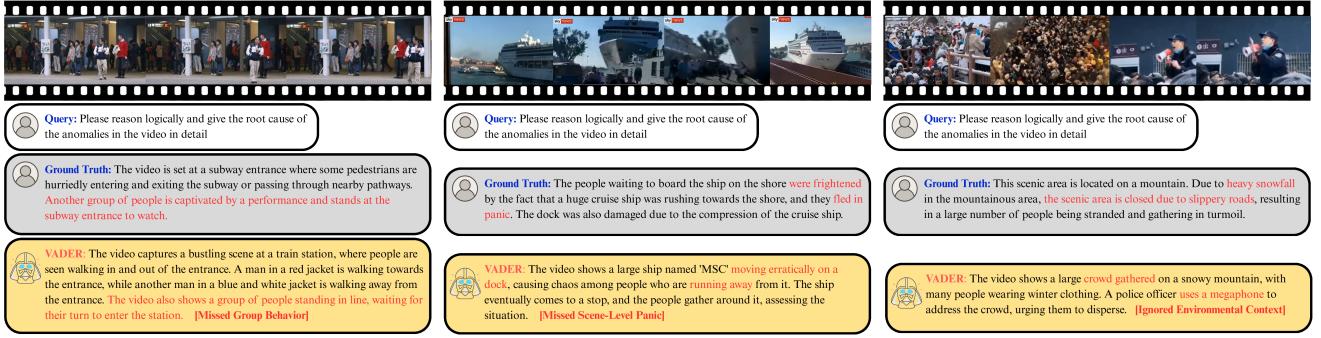
The current design of VADER focuses on pairwise object interactions, limiting its ability to capture scene-level or group-level anomalies. To broaden the scope of anomaly understanding, the following extensions can be considered:

**Global Scene Modeling.** Add a global feature stream that directly encodes entire frames to detect anomalies not tied to specific objects, such as lighting changes or smoke.

**Group-Level Reasoning.** Identify and model groups of



**Figure 7. Examples illustrating VADER’s high-motion bias.** VADER tends to focus on visually prominent or dynamic actions, overlooking subtle or context-dependent cues. In the left example, it emphasizes pouring water while ignoring the public smoking. In the middle example, it identifies the car collision but misses the underlying cause, which is the vehicle running the traffic signal. In the right example, it overemphasizes the physical confrontation while neglecting the initial theft that triggered the anomaly.



**Figure 8. Examples illustrating VADER’s limitation to object-centric reasoning.** VADER focuses on localized pairwise interactions, overlooking group behaviors or scene-level environmental factors. In the left example, it describes individual pedestrian movements but misses the crowd gathered to watch a performance. In the middle example, it identifies the ship’s movement but overlooks the collective panic it causes. In the right example, it focuses on the police officer and the crowd but ignores the environmental cause behind the stranded people.

objects to capture collective behaviors, allowing detection of emergent events like crowd panic or mass movements.

## E. Computational Efficiency

We provide a comparison of inference speed between VADER and NVILA [31] on the HAWK benchmark in Table 10. The table reports both the total inference time (in minutes) and frames per second (fps). Although VADER introduces additional modules for relational reasoning and context-aware sampling, it still operates at a practical speed for real-world applications.

## F. Intermediate Visualization of Module Contributions

Figure 9 illustrates how each module incrementally improves the generated descriptions. The example shows a

| Method     | Total Time (min) | FPS   |
|------------|------------------|-------|
| NVILA [31] | 49.85            | 33.63 |
| VADER      | 87.25            | 19.22 |

Table 10. Comparison of inference time and fps between NVILA [31] and VADER on the HAWK [47] benchmark.

bustling scene at a train station. The Base Model generates a generic narrative without focusing on key details. CAES helps the model attend to important frames, capturing activities like walking and standing in line. CORE further models object interactions, enabling it to distinguish directional actions and explain relationships between people and their surroundings. This step-by-step progression demonstrates how each module contributes to improving both descriptive accuracy and interpretability.



Base Model: The video captures a bustling scene at a train station platform. The platform is sheltered by a roof supported by white pillars, and the ground is paved with light-colored tiles. A group of **people is gathered** near the entrance to the station

+CAES: The video features a bustling scene at a train station, where people are seen **walking and standing in line**, waiting for their train. **The crowd is diverse**, with individuals dressed in various outfits, such as a man in a red jacket and another in a white and blue jacket.

+CAES+CORE (VADER): The video captures a bustling scene at a train station, where people are seen walking in and out of the entrance. A man in a red jacket is **walking towards the entrance**, while another man in a blue and white jacket is **walking away from the entrance**. The video also shows a **group of people standing in line**, waiting for their turn to enter the station.

Figure 9. **Intermediate visualization** showing how CAES and CORE improve descriptions. CAES helps focus on key frames, while CORE models object interactions for more interpretable outputs.

## G. Implementation Details

Anomaly intervals are detected using an adaptive threshold with the 97th percentile computed per video. For each interval, pre- and post-event contexts are determined by the 95th and 85th percentiles of the score slope over a 5-frame window, with a maximum context window of 30 frames. We uniformly sample 4 frames from each context, 8 from the event, and fill to 64 frames with background frames.

For relational analysis, object association combines cosine similarity of appearance embeddings (weight 0.8) and IoU, with matching solved by the Hungarian algorithm [19] and a maximum track age of 15 frames. Relational volatility at each timestep is measured as the maximum L2 distance between all co-tracked relation pairs in adjacent frames, followed by Gaussian smoothing with a standard deviation of 2.0. The top 5% of volatility peaks are used as positives.

The Relational Dynamic Encoder is a two-layer MLP trained with triplet margin loss [43] with margin 0.5 and semi-hard negative mining [43] with pool size 30. The encoder training uses Adam (learning rate  $1 \times 10^{-4}$ ) with StepLR for 50 epochs.

For LLM fine-tuning, we adopt NVILA [20] as the backbone, updating only the projector and LoRA [15] adapters while freezing all other parameters. The learning rate is set to  $2 \times 10^{-5}$  with a cosine schedule and a warm-up ratio of 0.03.