THE DECODER IS THE TEMPORAL MODEL: VERDICT-FIRST REASONING FOR ONLINE VIDEO ANOMALY UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurately localizing anomalous moments in videos typically relies on additional classification heads or temporal detection modules, which are effective for fast and precise anomaly localization. However, in scenarios that also require semantic interpretability, existing approaches often introduce separate detection components, making it difficult to fully exploit the inherent video understanding and language reasoning capabilities of vision-language models (VLMs). We observe that the causal context modeling and fixed vocabulary projection of a language model can directly provide a continuous online anomaly score. Based on this insight, we propose Vigil, a VLM-native framework for online video anomaly detection that directly turns a vision-language model into a window-level anomaly detector while preserving its capability to analyze anomalous events. To accommodate streaming-video scenarios, Vigil optimizes the prefill stage of LLM inference and jointly encodes historical frames and current-window frames into a unified model context, without introducing an additional temporal aggregator or recurrent cross-window hidden state. During supervised fine-tuning and inference, the model predicts `Normal` or `Anomaly` at the first generation position, and the corresponding token logits are normalized to produce a continuous anomaly score, thereby eliminating the need for a separate detection head. Under a predefined 20-second context setting, Vigil achieves low-latency real-time inference while outperforming the strongest offline baselines on both UCF-Crime and XD-Violence. These results suggest a simple and reproducible pathway toward video anomaly detection systems that retain both efficient online detection and semantic understanding, particularly for edge deployment scenarios.

1 INTRODUCTION

Video anomaly understanding asks a model to detect abnormal behaviour in a video and to explain it, and it underpins applications ranging from scene surveillance and violence monitoring to autonomous driving. Early work framed the task as anomaly detection alone: a temporal model assigns an anomaly score to every frame or segment, and the system output is a curve rather than an account of the event (Tian et al., 2021; Zhou et al., 2023; Wu et al., 2024). Such a system can tell an operator that something is wrong, but not what is happening.

The progress of multimodal large language models (MLLMs) has moved video anomaly analysis from detection alone toward semantic understanding. Beyond deciding whether an anomaly is present, video anomaly understanding (VAU) requires a model to identify the content of the anomalous event and to describe it in language (Zhang et al., 2025; Ye et al., 2025). Because VAU asks the model to attend to the decisive moment of an anomaly, many methods treat understanding as an add-on to detection: a dedicated temporal model first selects the segments that appear anomalous, and only those segments are passed to an MLLM for interpretation (Zhang et al., 2025; Zanella et al., 2024). More recent systems place this pipeline under an online constraint and emit decisions as the stream arrives (Feng et al., 2026; Yang et al., 2026; Liu et al., 2026), which is what a deployed system needs. The design still carries three costs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

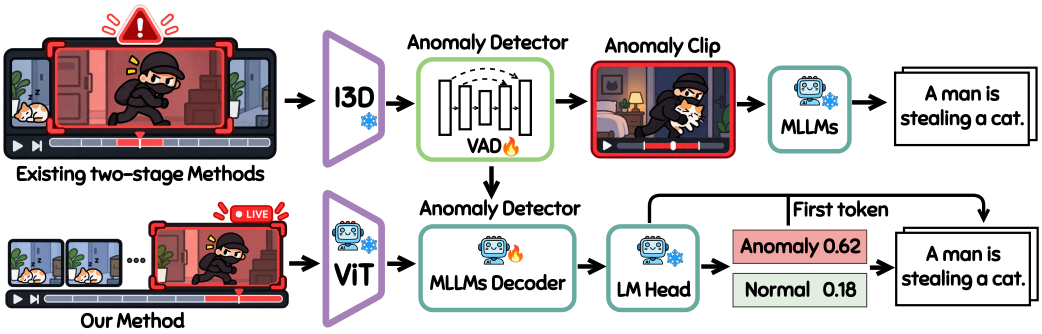

Figure 1: Motivation of Vigil. Top: existing two-stage methods train a separate detector on frozen features, then pass the detected anomaly clip to an MLLM for a description. Bottom: Vigil runs on the live stream and uses the fine-tuned MLLM decoder as the detector.

Redundant understanding is expensive. Most windows in a real stream are normal, yet running semantic interpretation on every one of them dominates the inference cost of an MLLM-based system. *An external detector limits generalization and adds latency.* It overfits the anomaly definition of its training data, and because detection and understanding run in series, every alarm needs a second call to the MLLM before any explanation appears. *Detection and explanation live in different spaces.* The score comes from the detector's own parameters and the explanation from a decoder called afterwards, so the sentence shown to an operator is not produced by the state that raised the alarm. The MLLM's long-context attention is also spent on interpretation rather than on the decision.

This paper takes two positions. First, **training remains necessary**. An online system must learn which behaviours count as anomalous against the current history, and that definition differs across deployments. Second, **an autoregressive model already exposes a decision interface, and anomaly detection should be aligned with it rather than attached beside it**. Renormalizing the `Normal` and `Anomaly` logits at the first answer position gives a continuous anomaly score from a single forward pass. Read this way, an untuned Qwen3-VL-2B (Bai et al., 2025a) already performs well on the annotated XD test windows of HIVAU-70k (Zhang et al., 2025) (Table 1). Most methods instead generate reasoning or descriptions before the verdict (Zhu et al., 2025b; Huang et al., 2026), and the same table shows that this costs more than decoding time: text placed before the decision lowers performance even when it describes the current window, and much more when it comes from another window, although the correct video is still in the input.

Table 1: Reading the verdict token under different test-time inputs.

Input at test time	AP ↑	AUC ↑
Video only	**95.38**	**97.10**
Video + ground-truth description	87.91	89.24
Ground-truth description only	87.55	89.32
Video + shuffled description	31.50	35.23
Vigil, Video only	**99.22**	**99.60**

We therefore propose **Vigil** (Fig. 1),, a VLM-native framework for online video anomaly understanding. Its design follows from the two positions above: the causal attention of the language decoder takes over the role of the temporal aggregator, consuming the historical frames and the current decision window as one ordered context, and detection is aligned with the native autoregressive decision interface, so that detection and semantic generation share exactly the same hidden state and the same output space. The score is read from the pretrained vocabulary projection, whose weights are not updated during fine-tuning, and a description, when one is wanted, continues the decoding from the very state that produced the score rather than from a second pass of a second model. Supervised fine-tuning uses a judgment-first output protocol, so that the position read at inference is the position that is trained, and description supervision is attached only to windows whose annotated clip is strictly aligned with them. For streaming deployment, successive windows share most of their context, and Vigil reuses the overlapping visual key-value states with positional correction instead of re-encoding them, while remaining causal. Routine detection decodes a single token; a description is generated only for the windows that warrant one, so the cost of understanding is paid on anomalies rather than on the stream. A judgment therefore costs one decoded token over a partially reused context, which keeps per-window inference well inside the duration of the window it scores and lets the system run in real time on commodity hardware.

Our contributions are as:

- We formulate online video anomaly detection natively inside a VLM, aligning detection with the autoregressive decision interface of the model, with no additional temporal aggregator and no recurrent cross-window state.
- We introduce a judgment-first scheme with separated judgment and description objectives, which trains the exact position used for detection while keeping on-demand semantic analysis in the same model.
- We develop a streaming inference procedure that achieves real-time inference on edge devices while outperforming all offline and online methods.

2 Related Works

2.1 Video Anomaly Detection

Video Anomaly Detection (VAD) aims to localize abnormal events in untrimmed video, and the weakly supervised setting, where only video-level labels are available during training, has become the dominant formulation (Sultani et al., 2018). Methods are generally divided by the features their temporal model consumes: frozen action features (Tian et al., 2021; Li et al., 2022; Wu et al., 2022; Zhou et al., 2023) and vision-language features (Joo et al., 2023; Wu et al., 2024; Yang et al., 2024; Huang et al., 2025; Lee et al., 2026). In the first group, RTFM (Tian et al., 2021) separates segments by temporal feature magnitude, while UR-DMU (Zhou et al., 2023) maintains separate normal and abnormal memories under an uncertainty term. In the second, CLIP-TSA (Joo et al., 2023) applies temporal self-attention to CLIP features; VadCLIP (Wu et al., 2024) adapts CLIP with learned prompts and aligns segments with class text; and π-VAD (Majhi et al., 2025) injects pose, depth and panoptic cues at training time while keeping an RGB-only detector at inference. Most of these methods consume an entire video before deciding, and REWARD (Karim et al., 2024) reports one architecture under both offline and online protocols. Across the family the output is a score curve emitted by a task-specific head in its own output space, which tells an operator that something is wrong but not what is happening.

2.2 Video Anomaly Understanding

Video Anomaly Understanding (VAU) goes a step further by requiring the model to identify the anomalous event and describe it in language, a shift made possible by multimodal large language models. Existing methods are generally divided into training-free pipelines (Zanella et al., 2024; Shao et al., 2025; Ye et al., 2025) and trained ones (Zhang et al., 2025). LAVAD (Zanella et al., 2024) captions frames and derives a score from LLM-cleaned text; EventVAD (Shao et al., 2025) segments the video into events and reasons over those units rather than fixed windows; VERA (Ye et al., 2025) keeps the vision-language model frozen and optimizes the guiding questions it is asked; while Holmes-VAU (Zhang et al., 2025) builds hierarchical annotations at clip, event and video granularity and trains an MLLM on them. More recently, to lift the constraint that a whole video be available before inference, online systems have been introduced: MoniTor (Feng et al., 2026) drives an LLM with instructions over a stream, PANDA (Yang et al., 2026) plans a detection strategy and reflects when its judgment is uncertain, ESOM (Liu et al., 2026) targets streaming efficiency under anomaly definitions that change at test time. In all of them detection and explanation remain in separate spaces, since the sentence shown to an operator is written afterwards by a decoder that did not produce the score.

3 Methodology

Vigil formulates online video anomaly understanding as judgment-first prediction within a single vision–language model (VLM). As shown in Fig. 2, model jointly processes historical observations and the current decision window, produces a continuous anomaly score from its first answer position, and supports subsequent language generation for semantic analysis. Temporal context is integrated by the pretrained model, while anomaly scoring uses its existing vocabulary projection. Supervised fine-tuning aligns these capabilities with the anomaly definitions and temporal granularity of the task.

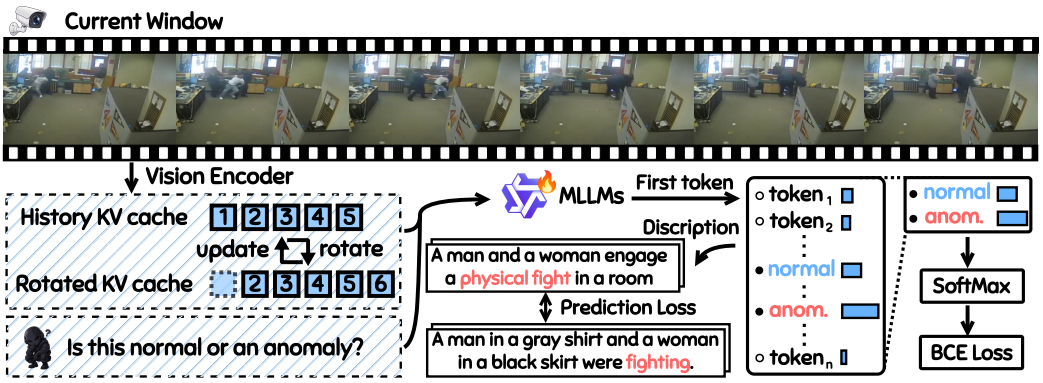

Figure 2: Overview of Vigil. The current window is encoded and appended to the rotated KV cache of the history, so only new frames are encoded. At the first answer position, the `normal` and `anomaly` logits are renormalized into the anomaly score and trained with a BCE loss. For anomalous windows, decoding continues to generate a description, trained with a prediction loss.

3.1 Problem Definition

Given an untrimmed video stream of T frames, $\mathcal{V} = (I_1, I_2, \ldots, I_T)$, where I_i denotes the i-th frame, we study the problem of online video anomaly understanding. At decision time t, we define the current decision window as $\mathcal{C}_t = (I_{t-L_c+1}, \ldots, I_t)$ and its preceding causal history as $\mathcal{H}_t = (I_{t-L_c-L_h+1}, \ldots, I_{t-L_c})$, where L_c and L_h denote the numbers of frames in the current window and the historical context, respectively. The model receives their temporally ordered concatenation

$$\mathcal{X}_t = [\mathcal{H}_t; \mathcal{C}_t], \tag{1}$$

and cannot access any future frame I_i with $i > t$ when making the prediction at time t. Each prediction therefore depends exclusively on video observations available at or before the current decision time.

Training Data. We use HIVAU-70k (Zhang et al., 2025) for supervised fine-tuning. Built upon the training videos of UCF-Crime and XD-Violence, HIVAU-70k augments conventional video-level labels with event boundaries, anomaly judgments, and hierarchical textual descriptions. We denote the event annotations of the n-th video by

$$\mathcal{E}^{(n)} = \{([b_j^{(n)}, e_j^{(n)}], u_j^{(n)}, \mathcal{K}_j^{(n)})\}_{j=1}^{M_n}, \quad \mathcal{K}_j^{(n)} = \{([b_{jk}^{(n)}, e_{jk}^{(n)}], d_{jk}^{(n)})\}_{k=1}^{K_j}, \tag{2}$$

where $u_j^{(n)} \in \{0, 1\}$ indicates whether the event is anomalous, and $\mathcal{K}_j^{(n)}$ contains temporally localized clip descriptions within that event. Events judged as anomalous induce the anomalous frame set $\mathcal{A}^{(n)}$. To construct window-level SFT samples, we partition each video into non-overlapping current windows of L_c frames and define their anomaly coverage ratios and training labels as

$$r_t^{(n)} = \frac{1}{L_c} \sum_{i=t-L_c+1}^{t} \mathbb{I}(i \in \mathcal{A}^{(n)}), \qquad y_t^{(n)} = \mathbb{I}(r_t^{(n)} \geq \rho). \tag{3}$$

We set $\rho = 0.5$ so that each positive SFT sample contains sufficient observable anomaly content. This threshold is used only to construct window-level training labels; evaluation follows the official frame-level annotations of UCF-Crime and XD-Violence. The temporally supervised SFT data can thus be written as $\mathcal{D}_{\text{SFT}} = \{(\mathcal{X}_t^{(n)}, q, y_t^{(n)}, \tilde{d}_t^{(n)})\}$, where $\tilde{d}_t^{(n)}$ is used as a description target only when a clip description is reliably aligned with the current window.

3.2 Anomaly Scoring

Representative two-stage MLLM-based VAU methods (Zhang et al., 2025; Lv & Sun, 2024) typically employ a temporal anomaly detector g_ϕ to aggregate video features and a task-specific classifier c_ψ to determine whether the current segment is anomalous. Only segments whose anomaly scores

exceed a threshold are subsequently passed to an MLLM m_ω for semantic analysis. This pipeline can be abstracted as

$$\mathbf{z}_t = g_\phi(\mathcal{X}_t), \qquad a_t = c_\psi(\mathbf{z}_t), \qquad \hat{d}_t = \begin{cases} m_\omega(\mathcal{X}_t, q), & a_t \geq \eta, \\ \varnothing, & a_t < \eta, \end{cases} \tag{4}$$

where η is the detection threshold. Vigil unifies temporal modeling, anomaly judgment, and language analysis within a single MLLM. Historical frames and current-window frames are temporally ordered in $\mathcal{X}_t$ and directly aggregated by self-attention in the language decoder, without an additional temporal aggregator or cross-window hidden state. Unlike existing approaches such as VAU-R1 (Zhu et al., 2025b) and Vad-R1 (Huang et al., 2026), which explicitly introduce chain-of-thought reasoning (Wei et al., 2022) for anomaly analysis, Vigil avoids constructing a complete reasoning trajectory before anomaly prediction. We argue that, when the language decoder itself serves as the temporal modeling module, additional reasoning context that is not specifically optimized for anomaly detection may introduce irrelevant semantic information and distract the model from the temporal evidence critical to anomaly judgment. Therefore, Vigil performs anomaly prediction directly from the temporally aggregated decoder representation, while retaining the language modeling capability for subsequent semantic analysis. Let $\mathbf{h}_t$ denote the decoder hidden state at the first answer position. We have

$$\mathbf{h}_t = F_\theta(\mathcal{X}_t, q), \qquad \boldsymbol{\ell}_t = W_{\text{vocab}}\mathbf{h}_t, \tag{5}$$

where W_{vocab} is the vocabulary projection matrix inherited from the pretrained VLM. Let v_N denote the `Normal` token and v_A the first token of `Anomaly`. The continuous anomaly score of the current window is obtained by binary renormalization:

$$s_t = \frac{\exp(\ell_{t,v_A})}{\exp(\ell_{t,v_N}) + \exp(\ell_{t,v_A})} = \sigma(\ell_{t,v_A} - \ell_{t,v_N}), \qquad \hat{y}_t = \mathbb{I}(s_t \geq \eta). \tag{6}$$

For the tokenizer used in this work, v_N and v_A correspond to `Normal` and `Anomaly`, respectively. During supervised fine-tuning, W_{vocab} is kept frozen, while LoRA (Hu et al., 2022) adapts $\mathbf{h}_t$ for anomaly judgment, eliminating the need for an additional detection head. If semantic analysis is required, decoding simply continues after the anomaly decision to generate $\hat{d}_t$.

3.3 Judgment-First Supervised Fine-Tuning

The output protocol places the judgment before any semantic description. Suppressing the video index for readability, the target text is

$$a_t = \begin{cases} \texttt{Normal}, & y_t = 0, \\ \texttt{Anomaly}, & y_t = 1,\ m_t = 0, \\ \texttt{Anomaly} \parallel \texttt{newline} \parallel \tilde{d}_t, & y_t = 1,\ m_t = 1, \end{cases} \tag{7}$$

where $\parallel$ denotes text concatenation. This ordering directly trains the first answer position used for detection, while retaining description supervision on aligned positive windows.

Separated judgment and description losses. Let $(a_{t,1}, \ldots, a_{t,K_t})$ be the tokenized ta rget, including a terminal token where it is supervised. We separate its supervised positions into a judgment segment $\mathcal{J}_t$ and a description segment $\mathcal{D}_t$. The judgment segment contains all tokens of the judgment string, together with the separating newline when a description is present. The description segment contains the description and its terminal token. Judgment-only samples supervise the judgment string without supervising the following terminal token.

Writing

$$p_{t,k} = p_\theta(a_{t,k} \mid \mathcal{X}_t, q, a_{t,<k}), \tag{8}$$

the segment-wise language losses are

$$\mathcal{L}_{\text{jud},t} = -\frac{1}{|\mathcal{J}_t|} \sum_{k \in \mathcal{J}_t} \log p_{t,k},$$
$$\mathcal{L}_{\text{desc},t} = \begin{cases} -\dfrac{1}{|\mathcal{D}_t|} \displaystyle\sum_{k \in \mathcal{D}_t} \log p_{t,k}, & m_t = 1, \\ 0, & m_t = 0. \end{cases} \tag{9}$$

Normalizing the segments separately prevents a long description from reducing the relative contribution of judgment supervision. In addition, we explicitly optimize the binary score used at inference:

$$\mathcal{L}_{\mathrm{bin},t} = -y_t \log s_t - (1 - y_t) \log(1 - s_t). \tag{10}$$

The complete objective is

$$\mathcal{L} = \mathbb{E}_{t \sim \mathcal{P}_{\mathrm{train}}} \left[\mathcal{L}_{\mathrm{jud},t} + \lambda_{\mathrm{bin}} \mathcal{L}_{\mathrm{bin},t} + \lambda_{\mathrm{desc}} m_t \mathcal{L}_{\mathrm{desc},t} \right], \tag{11}$$

with $\lambda_{\mathrm{bin}} = 0.25$ and $\lambda_{\mathrm{desc}} = 1$. Here $\mathcal{P}_{\mathrm{train}}$ denotes the training sampling distribution, in which positive windows are repeated twice. The language loss teaches the output protocol, and the binary loss directly supervises the judgment-logit margin used for scoring.

3.4 Streaming Inference with Approximate Cache Reuse

Successive decision windows share most of their historical observations. Repeating a full prefill would re-encode these frames and recompute their decoder KV states at every decision. Vigil reduces this repeated computation by maintaining a streaming cache. The main configuration uses a 4-second current window and its preceding 16 seconds of history, sampled uniformly at 4 frames per second. A complete context therefore contains 80 sampled frames, with 16 newly arriving frames per decision.

Incremental prefill. The first decision uses a full prefill, while subsequent decisions reuse the overlapping visual KV cache and only encode newly arrived frames. The retained cache is updated with positional correction, excluding the task suffix and generated responses to prevent semantic outputs from leaking into later windows. After 20 reuse steps, a full-context refresh is performed, and the cache is reset between videos. This cache reuse introduces cross-window KV states but no additional learned temporal module, while preserving online causality.

Scoring and semantic output. Routine detection directly uses the first-position judgment logits without generating descriptions. Semantic descriptions are produced only when needed by continuing autoregressive decoding after the anomaly judgment. Thus, anomaly scoring and semantic analysis share the same model, while description generation remains optional.

4 Experiments

4.1 Experiments Settings

Datasets and Baselines. We evaluate Vigil on four video anomaly benchmarks. UCF-Crime Sultani et al. (2018) and XD-Violence Wu et al. (2020) serve as the primary detection benchmarks, where models are trained with the temporally localized annotations and descriptions provided by HIVAU-70k Zhang et al. (2025) and evaluated against the official frame-level annotations. MSAD Zhu et al. (2024) is used to evaluate temporal grounding under its official weakly supervised split, while Vad-Reasoning Huang et al. (2026) evaluates video-level anomaly judgment and reasoning. All datasets are converted into causal windows with a 4-second current window, and validation videos are held out from the training sets when available. For UCF-Crime and XD-Violence, we compare Vigil with representative offline methods, including LAVAD Zanella et al. (2024), Event-VAD Shao et al. (2025), RTFM Tian et al. (2021), VERA Ye et al. (2025), UR-DMU Zhou et al. (2023), REWARD Karim et al. (2024), CLIP-TSA Joo et al. (2023), TPWNG Yang et al. (2024), VadCLIP Wu et al. (2024), Ex-VAD Huang et al. (2025), RefineVAD Lee et al. (2026), Holmes-VAU Zhang et al. (2025), and π-VAD Majhi et al. (2025), as well as online methods MoniTor Feng et al. (2026), PANDA Yang et al. (2026), ESOM Liu et al. (2026), and the online configuration of REWARD Karim et al. (2024).

Evaluation Protocol We follow the standard evaluation protocols of each benchmark. On UCF-Crime, we report frame-level ROC-AUC, while on XD-Violence we report frame-level average precision (AP). For MSAD, we report frame-level AUC for anomaly detection and mIoU together with R@$0.3, 0.5, 0.7$ for temporal grounding. On Vad-Reasoning, we evaluate video-level anomaly judgment using Accuracy, F1, Precision, and Recall, where the video-level anomaly score is obtained by aggregating window-level predictions.

Table 2: Comparison with representative methods on UCF-Crime and XD-Violence. We report frame-level ROC-AUC on UCF-Crime (UCF) and frame-level average precision on XD-Violence (XD), both in percent (↑). *Type*: T = training, GF = gradient-free, TF = training-free.

Method	Backbone	Sup.	Type	UCF AUC ↑	XD AP ↑
Offline					
LAVAD Zanella et al. (2024)	BLIP2+Llama-2-13B	N/A	TF	80.28	62.01
EventVAD Shao et al. (2025)	CLIP	N/A	TF	82.03	64.04
RTFM Tian et al. (2021)	I3D	Video	T	84.30	77.81
VERA Ye et al. (2025)	InternVL2-2B	Video	GF	86.55	70.54
UR-DMU Zhou et al. (2023)	I3D	Video	T	86.97	81.66
REWARD Karim et al. (2024)	Uniformer-32	Video	T	87.48	80.30
CLIP-TSA Joo et al. (2023)	CLIP	Video	T	87.58	82.19
TPWNG Yang et al. (2024)	CLIP	Video	T	87.79	83.68
VadCLIP Wu et al. (2024)	CLIP	Video	T	88.02	84.51
Ex-VAD Huang et al. (2025)	CLIP	Video	T	88.29	86.52
RefineVAD Lee et al. (2026)	CLIP	Video	T	88.92	88.66
π-VAD Majhi et al. (2025)	I3D	Video	T	90.33	85.37
Holmes-VAU Zhang et al. (2025)	InternVL2-2B	Full	T	88.96	87.68
Online					
MoniTor Feng et al. (2026)	BLIP2+GLM	N/A	TF	82.57	55.01
PANDA Yang et al. (2026)	Qwen2.5VL-7B+Gemini	N/A	TF	82.57	63.57
ESOM Liu et al. (2026)	Qwen3-VL-8B	N/A	TF	86.18	71.68
REWARD Karim et al. (2024)	Uniformer-32	Video	T	86.94	77.71
Vigil (Ours)	Qwen3-VL-2B	Full	T	**90.42**	**91.00**

Implementation Details All models build on Qwen3-VL-2B-Instruct Bai et al. (2025a), with videos decoded at 4 fps. LoRA (Hu et al., 2022) ($r = 16$, $\alpha = 32$, dropout 0.05) adapts the decoder attention and feed-forward projections; the visual encoder and vocabulary projection remain frozen. We optimize with AdamW (Loshchilov & Hutter, 2019) at a learning rate of 10^{-4} without weight decay, under a cosine schedule (Loshchilov & Hutter, 2017) with 3% linear warm-up, and clip gradients at norm 1.0. Training runs in bfloat16 with a maximum sequence length of 8,192 tokens. All experiment are conducted on a single NVIDIA A100. Positive windows are repeated twice, and the objective uses $\lambda_{\text{bin}} = 0.25$ and $\lambda_{\text{desc}} = 1$.

4.2 Main Results

Comparison with the state of the art. Table 2 reports frame-level ROC-AUC on UCF-Crime and frame-level AP on XD-Violence. Vigil reaches 90.42 AUC and 91.00 AP, the best value in both columns, under the online protocol that scores each frame from past context alone. Among online methods the margin is wide: the strongest prior results both belong to REWARD, at 86.94 AUC and 77.71 AP, and Vigil adds 3.48 and 13.29 over them. The training-free systems that pair a captioner with a large language model trail further, with MoniTor and PANDA both stopping at 82.57 AUC and ESOM, the best of them, reaching 86.18 AUC and 71.68 AP.

Vigil also passes the offline methods that see the whole video before deciding. π-VAD holds the best offline AUC at 90.33, and Vigil is 0.09 above it without using a single future frame; on XD-Violence it adds 2.34 AP over RefineVAD at 88.66. REWARD is the only method reported in both settings, and moving it online costs 0.54 AUC and 2.59 AP, while the online numbers here sit above every offline entry. The backbone accounts for none of this: Vigil runs on Qwen3-VL-2B against Qwen3-VL-8B (ESOM) and Qwen2.5VL-7B with Gemini (PANDA), and VERA, the closest match in capacity at 2B, reaches 86.55 AUC and 70.54 AP. The gain comes from neither a larger model nor access to future frames.

Temporal grounding. Table 3 reports temporal localization results on the anomalous test videos from the official MSAD split (Zhu et al., 2024). Vigil retains windows whose margin exceeds a threshold τ, merges adjacent windows, and selects the segment with the largest cumulative margin as the final prediction. With a 2B backbone, Vigil achieves an mIoU of 63.75, substantially outperforming VAU-R1 at 35.77 and the 7B base model at 45.90. Fixing the threshold to 0.5 results in only a 0.77-point drop in mIoU, indicating that the performance gain is not sensitive to validation-based threshold tuning. It is worth noting that this evaluation includes only anomalous videos and there-

Table 3: Temporal groundig on MSAD (Zhu et al., 2024) (official WS split, 120 anomalous test videos). Rows above the line are quoted from VAU-R1 (Zhu et al., 2025b).

Method	Vis.	mIoU ↑	R@0.3	R@0.5	R@0.7
Qwen2.5-VL-3B Bai et al. (2025b)	yes	21.27	30.00	10.83	4.17
Qwen2.5-VL-3B + SFT	yes	30.65	47.50	30.00	9.17
Qwen2.5-VL-3B + RFT (VAU-R1 Zhu et al. (2025b))	yes	35.77	53.33	34.17	15.83
Qwen2.5-VL-7B Bai et al. (2025b)	yes	45.90	70.83	45.83	21.67
Qwen3-VL-2B + **Vigil**, threshold on val	yes	**63.75**	84.17	67.50	**51.67**
Qwen3-VL-2B + **Vigil**, threshold fixed at 0.5	yes	62.98	**85.00**	**68.33**	49.17

Table 4: Anomaly description quality. B@1 and B@4: BLEU (Papineni et al., 2002); MET.: METEOR (Banerjee & Lavie, 2005); R-L: ROUGE-L (Lin, 2004); CIDEr (Vedantam et al., 2015).

Method	B@1	B@4	CIDEr ↑	MET.	R-L
InternVL3-2B Zhu et al. (2025a)	0.2655	0.0518	0.5663	0.2426	0.2312
Qwen2-VL-2B Wang et al. (2024)	0.2792	0.0480	0.4752	0.2348	0.2260
Qwen2.5-VL-3B Bai et al. (2025b)	0.3093	0.0501	0.4349	0.2209	0.2202
Qwen3-VL-2B Bai et al. (2025a)	0.2206	0.0373	0.5225	0.2442	0.2126
Qwen3-VL-4B Bai et al. (2025a)	0.2106	0.0365	0.5042	0.2625	0.1988
Qwen3-VL-8B Bai et al. (2025a)	0.2270	0.0389	0.4931	0.2516	0.2044
Holmes-VAU-2B Zhang et al. (2025)	0.3561	0.1064	0.8832	0.3228	0.2830
Vigil	**0.4588**	**0.2000**	**1.5574**	**0.3959**	**0.3719**

fore does not capture false positives on normal videos. On the 120 normal videos, Vigil produces 12 false alarms, while achieving a frame-level AUC of 94.91 over the full set of 240 videos.

Anomaly description. Table 4 evaluates the descriptions generated by continuing decoding from the same model state that produces the anomaly score. Vigil emits a description only on windows it judges anomalous (Eq. 7), so normal windows carry no description to score and the evaluation is restricted to anomalous windows. Vigil achieves a CIDEr score of 1.5574 and a B@4 score of 0.2000, substantially outperforming Holmes-VAU-2B, which obtains 0.8832 and 0.1064, respectively, and is trained with the same annotation style. It should be noted, however, that Holmes-VAU-2B operates in an offline inference setting, whereas Vigil performs online inference. Despite using the same dataset and a similar training paradigm, Vigil demonstrates stronger anomaly understanding. We attribute this improvement to two factors. First, reconstructing the training data into an online sequential format enables the model to better capture temporal dependencies and causal relationships across adjacent observations. Second, the model appears to benefit from a form of decision-conditioned belief: once it identifies a window as anomalous, the subsequent generation is conditioned toward explaining the evidence underlying that anomaly decision.

Online cost. Table 5 is best read through the compute column, the processing time spent per second of video. MoniTor and PANDA both need more than one second to process one second of stream, so their queues never drain: the delays in the table hold only for the first window, and the lag keeps growing afterwards. Vigil uses less than 5% of real time, and its delay is set almost entirely by the window length rather than by computation. A shorter window therefore lowers the delay at a small cost in accuracy, a trade-off an agentic pipeline cannot make, because its delay never falls below its own processing time. REWARD, a lightweight specialized detector, costs about as much as Vigil at a 2 s window, so against it the difference is accuracy rather than cost.

Domain transfer. Table 6 presents the performance under joint training and cross-domain transfer settings. The first is what joint training costs: 89.91 against 90.42 on UCF-Crime and 90.54 against 91.00 on XD-Violence, both inside the training-noise floor of 1.39 AUC and 0.65 AP that three seeds of the same recipe establish. One model serves both benchmarks at no measurable cost. The second is what a decision boundary carries across domains. Transfer holds the level of the strongest training-free online system in Table 2, 86.18 AUC and 71.68 AP, with a 2B backbone against its 8B, so the boundary learned in one domain is worth as much as the open-world prior of a model four times the size. Training on the target domain then adds 4.43 AUC and 18.01 AP on top of that, and that gap is what

Table 6: Domain transfer.

	UCF-Crime	XD-Violence
Training set	AUC	AP
UCF	90.42	72.49
XD	85.99	91.00
Joint	89.91	90.54

Table 5: Online decision cost and detection delay. *Decision period* is how often a score is emitted, *processing* is the per-decision inference cost, *delay* is the lag from the start of a decision period to a usable alarm (period + processing), and *compute* is processing per second of video.

Method	Decision period (s)	Processing (ms)	Delay (s)	Compute (ms / s)	UCF AUC	XD AP
MoniTor Feng et al. (2026)	0.6	5900	6.5†	9833	82.57	55.01
PANDA Yang et al. (2026)	5.0	6100	11.1†	1220	82.57	63.57
REWARD (online) Karim et al. (2024)	6.4	500	6.9	78	86.94	77.71
Vigil, 1 s window	1.0	120	1.1	120	89.71	90.79
Vigil, 2 s window	2.0	155	2.2	77	89.49	90.48
Vigil, 4 s window	4.0	184	**4.2**	**46**	**90.42**	**91.00**

a light per-deployment adaptation buys. Training remains necessary, and it is necessary per domain rather than once.

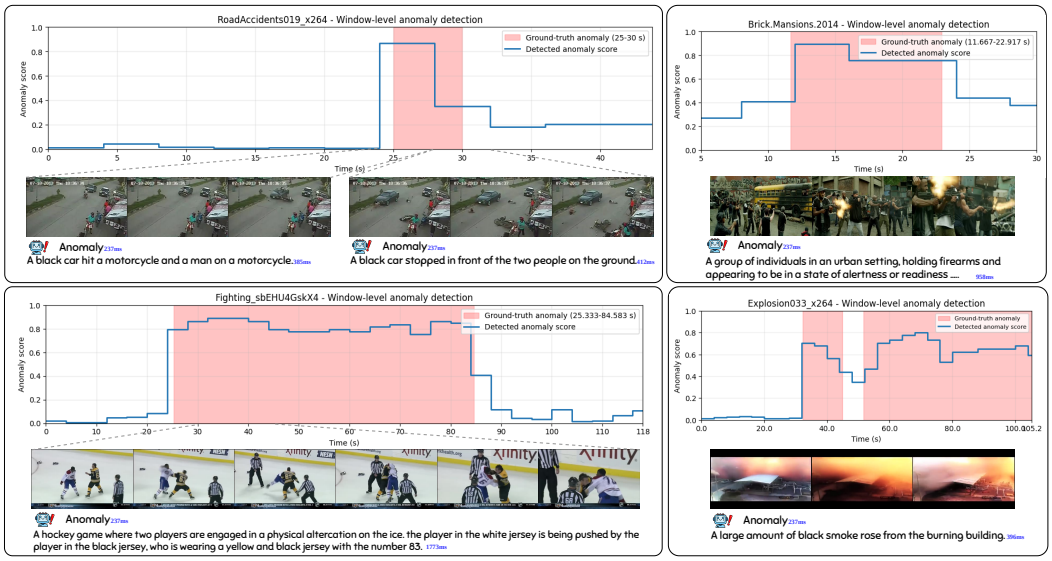

Figure 3: Case studies of online anomaly detection and explanation. The predicted window-level anomaly scores closely track the ground-truth anomalous intervals (red) across diverse scenarios.

4.3 Visualization

Fig. 3 shows traffic accidents, shootings, fights and explosions from UCF-Crime and XD-Violence. In each video the score rises inside the annotated interval, and the descriptions name what happens in the frames. The latencies in the figure are measured on a single NVIDIA RTX 4060 Ti. The verdict takes the same short time in every window because it is one decoded token, while a description takes longer as it gets longer and is generated only for windows the model flags. Appendix shows cases where Vigil fails.

5 Conclusion

We presented Vigil, which reads an online anomaly score from the `Normal` and `Anomaly` logits at the first answer position of a vision-language model, with the decoder's causal attention taking the place of a separate temporal model. With a 2B backbone it outperforms both online and offline methods on UCF-Crime and XD-Violence, and runs well within real time. We also found that text placed before the verdict hurts detection, even when the text is a correct description of the window. The failure cases show that Vigil still reacts more to large visual changes than to small interactions such as a barely visible weapon.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing, and to interpret results: parts of our code were modified with AI assistance to make the implementation more general-purpose, and LLM-based assistants helped us analyze experimental outcomes, for example by aggregating metrics across runs and drafting explanations of ablation results. We have not used generative AI tools to generate synthetic data sets, to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, or to clean and reformat datasets; all experiments were designed by the authors. Formulating or proving mathematical claims, writing proofs, and qualitative or thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools to edit the paper for readability and to create and edit the code that orchestrates our experiments, i.e., scripts that configure, launch, and log runs. We have reviewed all AI-assisted work, and all results reported in this paper are taken from actual experimental runs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Separately from the research and writing assistance above, generative AI is also a component of the proposed method: at inference time, a vision-language model generates textual analyses of anomalous samples, and these analyses are outputs of the method and are reported as such.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report, 2025b.

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*, pp. 65–72, 2005.

Yue Feng, Yingshi Liu, Jingrou Zhang, Jie Qin, et al. Monitor: Exploiting large language models with instruction for online video anomaly detection. *Advances in Neural Information Processing Systems*, 38:35851–35871, 2026.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In *Forty-second International Conference on Machine Learning*, 2025.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *Advances in neural information processing systems*, 38:118486–118518, 2026.

Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3230–3234. IEEE, 2023.

Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6834–6842. IEEE, 2024.

Junhee Lee, ChaeBeen Bang, MyoungChul Kim, and MyeongAh Cho. Refinevad: Semantic-guided feature recalibration for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5899–5907, 2026.

Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1395–1403, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Zihao Liu, Xiaoyu Wu, Wenna Li, Jianqin Wu, and Linlin Yang. Esom: Efficiently understanding streaming video anomalies with open-world dynamic definitions. *arXiv preprint arXiv:2604.07772*, 2026.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models, 2024. URL https://arxiv.org/abs/2401.05702.

Snehashis Majhi, Giacomo D'amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and Francois Bremond. Just dance with π! a poly-modal inductor for weakly-supervised video anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24265–24274. IEEE, 2025.

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*, pp. 311–318, 2002.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2586–2595, 2025.

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 (CVPR)*, June 2018.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4955–4966. IEEE, 2021.

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 (CVPR)*, June 2015.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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*, pp. 729–745. Springer, 2022.

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*, pp. 322–339. Springer, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, et al. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6074–6082, 2024.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18899–18908. IEEE, 2024.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. *Advances in Neural Information Processing Systems*, 38:83182–83211, 2026.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8679–8688. IEEE, 2025.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18527–18536. IEEE, 2024.

Huaxin 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853. IEEE, 2025.

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*, volume 37, pp. 3769–3777, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems*, 37:89943–89977, 2024.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025b.

A Appendix

A.1 Adapter capacity.

Table 7 shows that performance is largely insensitive to adaptation capacity. Increasing the number of trainable parameters, even up to full fine-tuning, provides no clear advantage over a lightweight rank-8 adapter, while performance also tends to peak early during optimization. These results suggest that training does not primarily supply additional capacity, but instead learns a task-specific decision boundary over the events already represented by the backbone. The 2B model appears to possess the visual understanding and causal reasoning required for the task, and a low-rank update is sufficient to align these existing capabilities with anomaly detection.

Table 7: LoRA rank.

	UCF-Crime	XD-Violence
Rank	AUC	AP
r=8	90.02	90.19
r=16	90.42	91.00
r=32	90.63	90.10
r=64	89.69	90.68
Full	90.06	90.25

A.2 Effect of window length on online detection.

Table 8 compares window lengths at a fixed operating point of at most one false-alarm window per hour. For each window length, we select the most sensitive threshold that meets this budget on the 150 fully normal test videos, and match detected events one-to-one against the official annotations at $\mathrm{tIoU} \geq 0.1$. The 4 s window gives the best trade-off, achieving the highest event and first-event recall with precision comparable to the best setting. Very short windows suffer from low precision, as they tend to fragment a single anomaly into multiple detections and fire on unannotated segments of anomalous videos. The 2 s window is more conservative, trading recall for precision. This non-monotonic trend is partly due to threshold calibration: the strict budget permits only a few false alarms on the normal videos, so the selected threshold is sensitive to a handful of high-scoring normal clips. In addition, a fixed per-hour budget is proportionally looser for longer windows, since they produce fewer windows per hour. Across all settings, first-event recall exceeds event recall, indicating that the model reliably catches the onset of anomalous activity, which matters most for online alerting. We therefore use the 4 s window by default.

Table 8: Online event detection on UCF-Crime under a budget of one false-alarm window per hour.

Window	Event P	Event R	First-event R
1 s	22.2	31.4	40.7
2 s	47.3	22.4	26.4
4 s	45.7	**48.1**	**49.3**

A.3 VIDEO-LEVEL SUPERVISION.

Table 9 removes the temporal annotation and retains only the video-level label, matching the standard weakly supervised setting (Sultani et al., 2018) while preserving the online constraint. For an anomalous video, we randomly sample a bag $\mathcal{B}$ of B windows at each step and rank them by the anomaly margin $m(w) = \ell_{\texttt{An}}(w) - \ell_{\texttt{Normal}}(w)$. Only the top-$k$ windows (Tian et al., 2021) are used to form the bag-level score,

$$\bar{m}_{\mathcal{B}} = \frac{1}{k} \sum_{i \in \text{top-}\,k(\mathcal{B})} m(w_i),$$

and the positive video label is imposed through

$$\mathcal{L}_{\text{MIL}} = \lambda_{\text{bin}}\, w_+ \,\text{softplus}(-\bar{m}_{\mathcal{B}}),$$

which is equivalent to applying a binary cross-entropy loss with target 1 to the aggregated top-k margin. For a normal video, the video-level label implies that every sampled window is normal, so we retain window-level supervision:

$$\mathcal{L}_{\text{norm}} = \frac{1}{n} \sum_{j=1}^{n} \left[\text{CE}(w_j, \texttt{Normal}) + \lambda_{\text{bin}} \,\text{softplus}(m(w_j)) \right].$$

The overall objective is $\mathcal{L} = \mathcal{L}_{\text{MIL}} + \mathcal{L}_{\text{norm}}$. We disable the description-level CE branch in this setting, since attaching a semantic description to a specific clip would reveal its temporal location and therefore introduce supervision beyond the video-level label.

Despite this substantially weaker supervision, the resulting model remains competitive with prior weakly supervised methods on both UCF-Crime and XD-Violence, while operating strictly online and using the same general-purpose backbone. The comparison with frame-level supervision further isolates the value of temporal labels: when the backbone, window construction, and read-out position are held fixed, explicitly identifying where the anomaly occurs yields a clear improvement. This suggests that the performance gap is driven primarily by the availability of temporal supervision rather than by a change in architecture, and supports the role of our window construction in exploiting such localized labels.

Table 9: Comparision of differebr supervision.

Method	Backbone	Labels	UCF AUC ↑	XD AP ↑
offline, video-level labels				
RTFM Tian et al. (2021)	I3D	video	84.30	77.81
UR-DMU Zhou et al. (2023)	I3D	video	86.97	81.66
REWARD Karim et al. (2024)	Uniformer	video	87.48	80.30
CLIP-TSA Joo et al. (2023)	ViT	video	87.58	82.19
TPWNG Yang et al. (2024)	ViT	video	87.79	83.68
VadCLIP Wu et al. (2024)	ViT	video	88.02	84.51
online, video-level labels				
REWARD Karim et al. (2024)	Uniformer	video	86.94	77.71
Vigil + Top-k MIL	Qwen3-VL-2B	video	87.17	82.12
online, frame-level labels				
Vigil	Qwen3-VL-2B	frame	90.42	91.00

A.4 PROMPT SENSITIVITY.

Table 10 evaluates the robustness of our model to the wording of the query prompt. Besides the original question, we consider two prompt variants: (1) prepending a role prefix, "You are monitoring a surveillance camera. Is this normal or an anomaly?", and (2) replacing the original question entirely with the five guiding questions proposed by VERA (Ye et al., 2025).

UCF-Crime.

1. *Are there any people in the video who are not in their typical positions or engaging in activities that are not consistent with their usual behavior?*
2. *Are there any vehicles in the video that are not in their typical positions or being used in a way that is not consistent with their usual function?*
3. *Are there any objects in the video that are not in their typical positions or being used in a way that is not consistent with their usual function?*
4. *Is there any visible damage or unusual movement in the video that indicates an anomaly?*
5. *Are there any unusual sounds or noises in the video that suggest an anomaly?*

XD-Violence.

1. *Are there any frames where the video content shows an object or person performing an unusual action that does not fit the usual pattern of the video content?*
2. *Are there any frames where the video content shows an object or person behaving in a way that is clearly abnormal or unexpected?*
3. *Are there any frames where the video content shows a clear indication of an anomaly in the behavior of objects or people, such as an object moving in an unexpected way or a person performing an unusual action?*
4. *Are there any frames where the video content shows a player or object being out of place or not fitting the usual pattern of the video content?*
5. *Are there any frames where the video content shows a disruption in the normal flow of the video, such as a sudden change in the camera angle or a loss of video quality?*

Readout mass denotes the probability mass assigned to the answer tokens used for scoring. It remains at 1.000 under all prompts, indicating that the readout is stable and that the model never drifts outside the expected answer space. Performance changes only marginally across prompts on both datasets, with the original question performing best. Neither the role prefix nor the more elaborate VERA questions provides any gain. These questions were originally designed to guide training-free VLM reasoning, but they appear redundant once the model has been adapted to the task. Overall, our method does not rely on careful prompt engineering, and its performance is largely insensitive to the exact phrasing of the query.

Table 10: Sensitivity to the prompt engineering.

Prompt	Readout mass	UCF AUC	XD AP
Original question	1.000	90.42	91.00
+ role prefix	1.000	89.88	90.85
VERA (5 questions)	1.000	89.61	90.84

A.5 Effect of the binary judgment loss.

Table 11 ablates the binary judgment (BCE) loss, keeping all other settings identical. $\Delta\mu$ denotes the gap between the mean margins of anomalous and normal windows. Recall and F1 are measured at the fixed decision point of 0.5, and the last column reports the F1 lost by using 0.5 instead of the best threshold found by sweeping, i.e., the cost a deployment pays for not tuning a threshold. Adding the BCE loss slightly enlarges the margin gap on both datasets, but its main effect lies in calibration rather than separation. With BCE, the fixed threshold of 0.5 lies close to the optimal operating point, substantially reducing the F1 lost and yielding clearly higher recall at 0.5, especially on UCF-Crime. Without BCE, the scores remain discriminative, and a swept threshold achieves comparable best-case F1, but the default threshold is poorly placed and a practitioner must tune it on labeled data. Since such labels are rarely available at deployment time, the BCE loss makes the model usable out of the box with a fixed, interpretable threshold.

Table 11: Effect of the binary judgment loss on window scoring at a fixed decision point.

	$\Delta\mu \uparrow$	Recall@0.5 $\uparrow$	F1@0.5 $\uparrow$	F1 lost at 0.5 $\downarrow$
UCF-Crime				
with BCE	**4.85**	**38.9**	**42.87**	**4.01**
without BCE	4.63	23.7	38.89	9.62
XD-Violence				
with BCE	**8.04**	**78.6**	**82.62**	**0.25**
without BCE	7.61	72.5	81.14	1.37

A.6 Video-level judgment on Vad-Reasoning.

We further evaluate video-level anomaly judgment on the Vad-Reasoning test set (Huang et al., 2026), which contains 436 videos, 202 of them anomalous (Table 12). Window scores are aggregated into a video score by taking the maximum. We report results both at the fixed decision point of 0.5 and under a threshold selected by two-fold cross-fitting within the test set. With a much smaller backbone, Vigil outperforms Vad-R1 in accuracy, precision and F1. Notably, this holds already at the fixed threshold of 0.5, which requires no tuning on labeled data. Cross-fitting the threshold brings only a marginal further gain, mainly trading a little precision for recall. This confirms that the binary judgment loss places the default threshold close to the optimal operating point (Table 11). In contrast, the base model degenerates to predicting every video as anomalous at 0.5, so we report it only under the fitted threshold. Even then, it lags clearly behind Vigil, showing that the improvement stems from training rather than from threshold selection.

Table 12: Video-level anomaly judgment on Vad-Reasoning.

Method	Backbone	Acc $\uparrow$	Prec $\uparrow$	Rec $\uparrow$	F1 $\uparrow$
Vad-R1 Huang et al. (2026)	Qwen2.5-VL-7B	87.50	88.20	—	86.20
Qwen3-VL-2B (base), τ cross-fit	Qwen3-VL-2B	82.11	80.69	80.69	80.69
Vigil, $\tau = 0.5$	Qwen3-VL-2B	90.37	**94.94**	83.66	88.95
Vigil, τ cross-fit	Qwen3-VL-2B	**90.83**	91.33	**88.61**	**89.95**

A.7 ADDITIONAL QUALITATIVE RESULTS.

Fig. 4 and Fig. 5 show two videos that Vigil handles well, a surveillance clip from UCF-Crime and a film clip from XD-Violence. In Robbery102 the score stays below 0.27 until the robbery starts at 36 seconds. It then jumps to 0.85 and falls a little in each window as the attacker struggles with the woman and runs off. The descriptions follow the same order, from reaching the car to fighting to running away, but they do not say who attacks whom, and the alarm ends one window after the annotation does. Ip.Man.2008 shows a martial artist fighting several soldiers in an alley, edited into many short shots. The score rises in the first window of the fight, stays above 0.8 while it lasts, and falls close to zero as soon as it ends. The descriptions pick up the uniforms, the fight and a soldier being restrained, although one of them mistakes a soldier for a woman. A smaller rise at 44 seconds, when the man picks up a child, stays well below the scores of the fight.

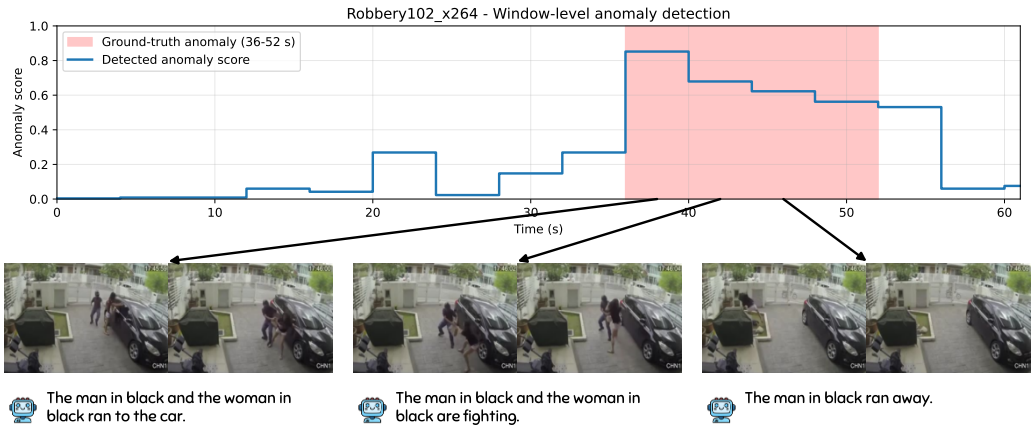

Figure 4: Robbery102 from UCF-Crime. Blue: window-level anomaly scores. Red: ground-truth interval. Below each frame group: the description generated for that window.

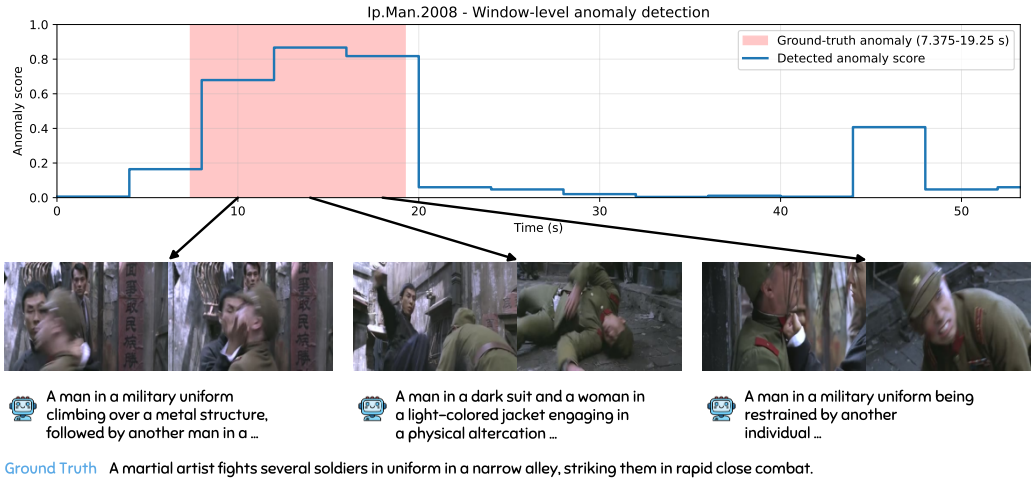

Figure 5: Ip.Man.2008 from XD-Violence. Blue: window-level anomaly scores. Red: ground-truth interval. Below each frame group: the description generated for that window.

A.8 FAILURE CASES.

Fig. 6 and Fig. 7 show two typical failures of Vigil. In Burglary079, the annotated anomaly spans 258 to 357 seconds, and Vigil scores this interval correctly. It also gives high scores to two segments outside it. Around 95 seconds a man walks into the empty garage and opens the trunk of a car, and the score jumps from near zero to about 0.85 and stays high for roughly 100 seconds. Around 450 seconds two men run out of the house, and the score rises to the same level. The camera is fixed in both cases, so the change comes from the frame content: a person appears in a static scene, or people move quickly. Vigil tends to treat such abrupt changes as evidence of an anomaly. Both segments are plausibly part of the burglary as preparation and escape, yet the annotation does not cover them, so they count as false alarms.

Robbery106 fails in a different way. Vigil places its highest score on the annotated window from 16 to 20 seconds, but the generated descriptions do not match the event. According to the ground truth, two men on a motorcycle stop next to a parked car and one of them threatens a man with a weapon. Vigil misses both the weapon and the motorcycle. It describes the victim as being pushed to the ground, which never happens in the frames, and it calls the same victim "the man in white" in one window and "the man in black" in the next. The weapon occupies only a few pixels in the surveillance footage, and a 2B backbone that sees a handful of frames per window has difficulty perceiving it. Without the weapon, the confrontation looks like ordinary physical contact between two people. Taken together, these cases suggest that Vigil responds more readily to large visual changes than to the specific interactions that define an anomaly.

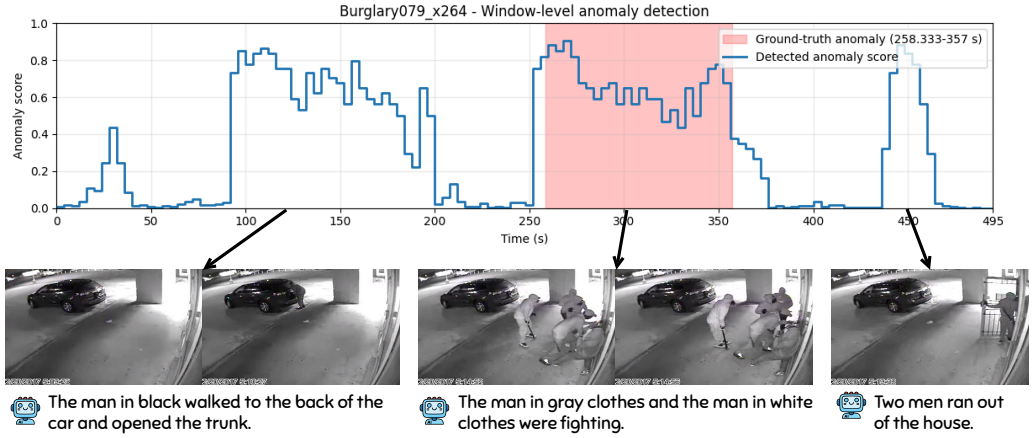

Figure 6: Failure case on Burglary079 from UCF-Crime.

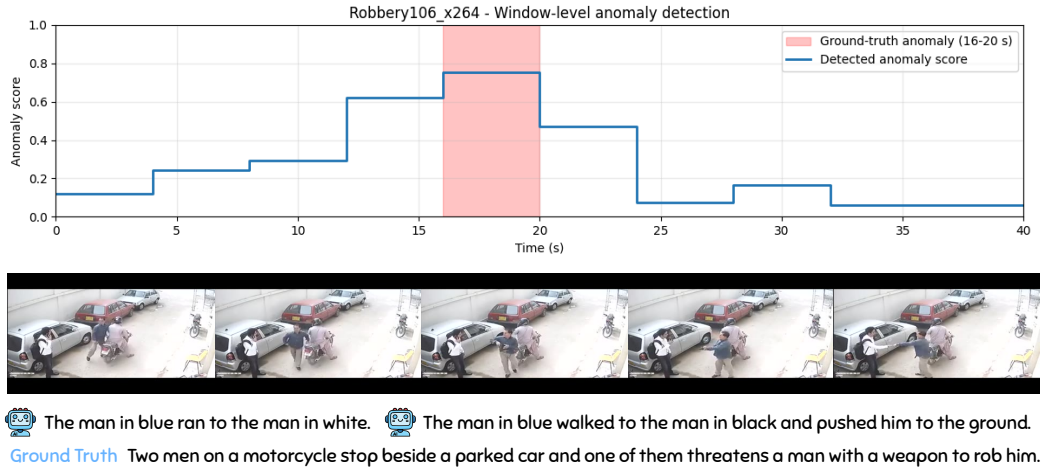

Figure 7: Failure case on Robbery106 from UCF-Crime.

A.9 JUDGMENT ORDER.

Table 13 studies where the verdict token is placed in the target. Under *ja*, the model emits the verdict first and the description after it. Under *aj*, it generates the description first and the verdict last. In both settings, normal windows also carry clean clip descriptions, so the presence of a description does not reveal the label and the comparison isolates the effect of ordering. *ja* outperforms *aj* on both AP and AUC. When the verdict comes last, it is condition ed on a self-generated description, and any errors in that description propagate to the final judgment. Predicting the verdict first lets the model commit to a decision directly from the visual evidence, while the description that follows serves as an explanation. Placing the verdict first also allows a window to be scored in a single forward pass without generating the description, which is important for online detection. We therefore adopt the *ja* order by default.

Table 13: Effect of the judgment order on XD-Violence.

Order	AP ↑	AUC ↑
ja	**91.05**	**96.95**
aj	88.82	96.04

B FUTURE WORK

In this paper, training data comes from HIVAU-70k, so the model inherits the anomaly definitions of UCF-Crime and XD-Violence and loses accuracy across domains. We plan to build window-level data for specific deployments ourselves. The existing descriptions also lack reasoning chains that explain why a window is anomalous given its history, and writing one for every short window of a long video is costly. Reinforcement fine-tuning needs only window labels for its reward, but a think-then-answer policy decodes a long trace where Vigil decodes a single token. We plan to reason only on windows with uncertain scores and to penalize length in the reward.