ANOMALY-GUIDED OBSERVATION AND TWO-LEVEL SEMANTIC FEEDBACK FOR TRAINING-FREE VIDEO ANOMALY ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

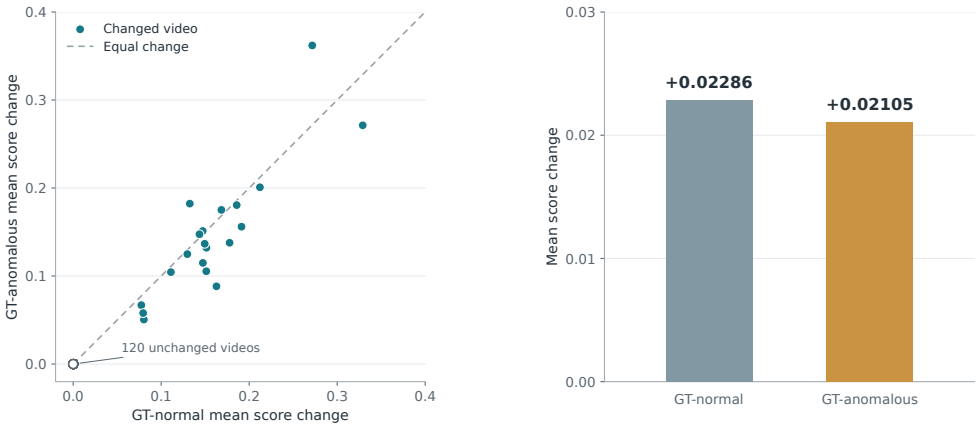
Training-free language-based video anomaly detection (VAD) uses frozen vision–language models (VLMs) and large language models (LLMs) to localize abnormal events without task-specific model-weight optimization. Existing approaches increasingly exploit broader temporal and semantic context, yet generic context does not necessarily foreground abnormal events or specify how video-level semantics should modify local anomaly scores. We introduce a training-free semantic feedback framework that uses VAD-guided active observation for video anomaly understanding (VAU) and returns the resulting anomaly-focused description to refine temporal detection. Starting from an upstream anomaly-score curve, we adapt reasoning-based iterative frame selection: high-score regions seed the observations, and anomaly scores together with VLM relevance assessments guide subsequent temporal exploration. A frozen video-language model generates a single open-ended anomaly description from the selected frames. The description is returned through two complementary feedback levels: video-level feedback determines whether the score distribution should be suppressed, preserved, or sharpened, with selective visual rechecking, while local feedback retrieves event-related windows and promotes scores only after target-local visual verification. All model weights remain frozen, and feedback is applied once without regenerating the description. On the HIVAU video-summary evaluations of UCF-Crime and XD-Violence, VAD-guided observation improves GPT-rated reasonability, detail, and consistency over the matched CLIP-guided variant. For VAD, semantic feedback improves the re-evaluated upstream detector from 84.36% to 86.25% AUC on UCF-Crime and from 68.07% to 72.58% AP on XD-Violence, where the final AUC is 91.74%.

1 INTRODUCTION

Video anomaly detection (VAD) aims to identify and temporally localize abnormal events in long, untrimmed videos (Sultani et al., 2018; Wu et al., 2020). Recent language-based approaches use pretrained vision–language models (VLMs) and large language models (LLMs) to interpret visual observations and estimate anomaly scores without task-specific model-weight updates (Zanella et al., 2024; Lin et al., 2025). In this setting, detection depends on the evidence and context supplied to frozen models at inference time.

Local observations alone can be insufficient for temporally extended or ambiguous events. Training-free VAD methods therefore increasingly incorporate broader temporal and semantic context through caption aggregation, contextual reasoning, and score refinement (Zanella et al., 2024; Lim & Hur, 2026; Mo et al., 2026b). However, broader context is not necessarily *anomaly-focused*. A generic video summary may devote substantial capacity to common scene content while representing the abnormal event only indirectly. Video anomaly understanding (VAU), in contrast, explicitly characterizes what happened and why it is anomalous (Zhang et al., 2025b; Lin et al., 2025). This suggests a complementary role for VAU: beyond serving as an explanation output, its anomaly-focused semantics can provide hypotheses for reconsidering temporal VAD predictions.

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.



(a) (b)

Figure 1: Changes from initial (RAW) to URF-HVAA-refined scores on UCF-Crime. (a) Per-video changes in GT-normal and GT-anomalous mean scores; the dashed line denotes equal changes, and 120 unchanged videos overlap at the origin. (b) Mean changes averaged equally across all 140 test videos containing both GT region types.

Obtaining such semantics from an imperfect detector is itself challenging. Initial anomaly scores help locate suspicious evidence (Lin et al., 2025), but one-shot sampling cannot adapt to evidence discovered during observation. A.I.R. (Zou et al., 2026) iteratively evaluates frame relevance and reallocates subsequent observations for video question answering, while Learning to Watch (Mo et al., 2026a) learns interleaved observation–reasoning policies for VAU. We adapt active observation to training-free VAU: the detector score curve initializes and guides observation, while VLM relevance assessments obtained during acquisition influence subsequent temporal exploration.

Returning the resulting description to VAD raises a second problem. A video-level judgment that an anomaly occurred does not identify which temporal regions warrant score promotion, and textual agreement with a local caption does not establish visible abnormal evidence in that interval. In our UCF-Crime diagnostic, URF-HVAA refinement produces similar video-averaged score increases in ground-truth normal and anomalous regions (Fig. 1). This observation motivates distinguishing *how the video-level score distribution should be treated* from *where local promotion is visually warranted*. Prior work connects detection and understanding through chained anomaly analysis, training-free semantic grounding, or detector–LVLM learning (Lin et al., 2025; Gao et al., 2025; Hashimoto et al., 2026). We study a specific return mechanism: a completed VAU description guides a single refinement pass over an existing temporal anomaly-score sequence, separating video-level modulation from local promotion gated by visual evidence.

Our framework couples VAD-guided active observation with this two-level semantic feedback. The description is generated once, evaluated as a VAU output, and returned to refine detection without regeneration. Experiments show improved GPT-rated description quality and complementary detection gains from the two feedback levels, with the local contribution more apparent in AP than in AUC.

Our main contributions are:

- We introduce a training-free semantic feedback framework that separates video-level score modulation from local refinement supported by visual evidence.
- We develop VAD-guided active observation for anomaly-focused video understanding, allowing newly acquired visual evidence to guide subsequent temporal exploration.
- We show that the generated descriptions are useful both as anomaly explanations and as feedback for temporal detection, evaluating description quality and detection refinement separately.

2 Related Work

Context-aware training-free VAD. LAVAD combines frame captioning, temporal summarization, and score refinement using video–text similarity (Zanella et al., 2024), while EventVAD performs event-aware temporal decomposition before MLLM-based anomaly reasoning (Shao et al., 2025). CoReVAD uses a frozen VLM to generate local decisions and descriptions and then incorporates broader visual–semantic context (Lim & Hur, 2026). ReCI combines video-level semantic context, retrieved reference context, and confidence-guided temporal refinement (Mo et al., 2026b). PANDA introduces agentic planning and knowledge retrieval (Yang et al., 2025), whereas URF-HVAA chains multiple anomaly-oriented priors through a unified zero-shot analysis pipeline (Lin et al., 2025). We use an anomaly-focused interpretation to guide temporal score refinement.

Anomaly-focused video understanding. VAU extends anomaly analysis from temporal scores to semantic descriptions of abnormal events and their context. Holmes-VAU combines hierarchical instruction data with anomaly-focused temporal sampling (Zhang et al., 2025b), while VADER incorporates context-aware sampling and relation-aware representations (Cheng et al., 2026). URF-HVAA passes temporal, textual, and spatial anomaly priors into final understanding (Lin et al., 2025), and the VAGU benchmark and GtS framework address joint anomaly grounding and understanding (Gao et al., 2025). Recent methods additionally make visual observation adaptive. A.I.R. combines query-conditioned initialization, VLM relevance assessment, and localized resampling (Zou et al., 2026); Learning to Watch learns interleaved observation–reasoning trajectories (Mo et al., 2026a); and AgenticVAU coordinates exploration and verification across multiple agents (Duan et al., 2026). We use adaptive observation to construct the description returned to the detector.

VAD–VAU interaction and semantic feedback. Detection and understanding can exchange information at different stages of anomaly analysis. URF-HVAA propagates intermediate anomaly priors across detection and understanding stages (Lin et al., 2025). Glance then Scrutinize is a training-free framework that uses semantic interpretations to guide subsequent temporal grounding (Gao et al., 2025). MuST-VAD instead forms an alternating learning loop in which detector predictions guide LVLM adaptation and updated semantic representations are returned to retrain the detector (Hashimoto et al., 2026). Our method targets a different interaction regime: a completed VAU description is returned to an existing anomaly-score sequence in a single inference-time pass while all model weights remain frozen. The return path separates video-level score modulation from event-conditioned, visually authorized local refinement without treating the generated description as dense temporal supervision.

3 Method

3.1 Problem setup and overview

Given an untrimmed video $V = (v_0, \ldots, v_{T-1})$ of T frames, where v_t denotes frame t, our framework outputs an open-ended anomaly description Y and a refined temporal anomaly-score sequence.

A frozen upstream analyzer provides local captions C, an initial anomaly-score sequence $b^{(0)}$, a suspicious-activity phrase h, and optional spatial overlays. In the main configuration, $b^{(0)}$ is the URF-HVAA refined score sequence. The same scores guide active observation and provide the starting point for semantic refinement. The index i ranges over the available score anchors, with $b_i^{(0)}$ anchored at original frame $16i$. For observation, the scores are smoothed and expanded to frame-level values as described in Appendix B.

Figure 2 summarizes the three components. Score-guided observation constructs visual evidence and a VAU response. The response then supplies a video-level interpretation and local event queries:

$$F = \Pi_{16}\Big(\mathcal{A}(V, b^{(0)}, q)\Big), \qquad Y = \mathcal{G}(\widetilde{V}_F, q, h), \tag{1}$$

$$b^{(v)} = \mathcal{R}_v(b^{(0)}, Y, C, V), \qquad b^{(1)} = \mathcal{R}_\ell(b^{(v)}; b^{(0)}, Y, C, V). \tag{2}$$

Here q denotes the common VAU instruction, $\mathcal{A}$ the iterative observation procedure, and $\mathcal{G}$ the frozen VAU answerer. The projection Π_{16} produces a chronologically ordered sequence F of 16 original-

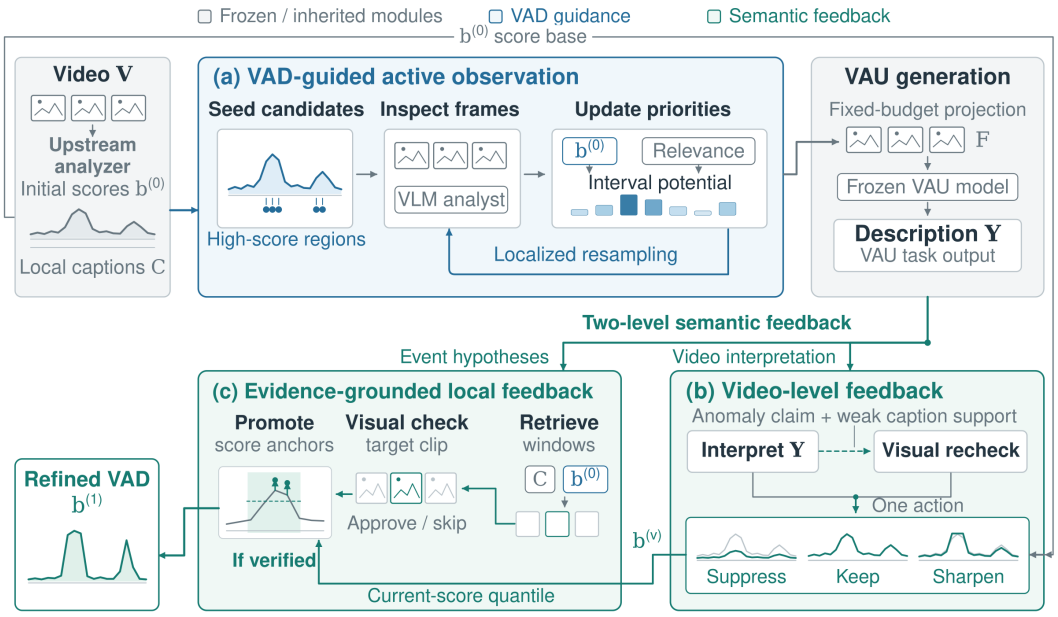


Figure 2: Anomaly-guided observation and two-level semantic feedback. (a) Initial scores $b^{(0)}$ guide seeding and interval priorities within the inherited A.I.R. observation loop; the frames indexed by F yield one description Y. (b) The video interpretation selects one score modulation, with selective visual rechecking of anomaly hypotheses. (c) For anomaly judgments, event hypotheses from Y, captions C, and $b^{(0)}$ retrieve windows; target-clip evidence authorizes promotion, and current-score quantiles select the anchors. The score path is $b^{(0)} \to b^{(v)} \to b^{(1)}$. All visual models use frames derived from V and remain frozen. Curves and tokens are schematic; the single return pass precedes final smoothing and expansion.

frame indices. The visual input $\widetilde{V}_F$ contains the corresponding frames after optional upstream spatial rendering. The two return operators act at different temporal scales: $\mathcal{R}_v$ modifies the video-wide score distribution, whereas $\mathcal{R}_\ell$ retrieves candidate windows using Y, C, and $b^{(0)}$ and promotes selected anchors in the current scores after target-local visual verification. All model weights remain frozen. The description Y is generated once and is not rewritten after feedback.

3.2 Score-guided active evidence acquisition for VAU

The common VAU instruction asks for anomalous activities without naming a video-specific event. We therefore use $b^{(0)}$ as a video-specific prior for locating evidence, replacing query–frame CLIP similarity (Radford et al., 2021) in the A.I.R. observation procedure (Zou et al., 2026).

High-score regions seed the candidate frames. During subsequent observation, the inherited interval-potential mechanism combines anomaly scores with VLM relevance assessments of inspected candidates. Localized density sampling then proposes further frames around retained locations. Newly inspected evidence therefore affects where the procedure looks next. The resulting selection is projected to the answerer's 16-frame input. The interval potential and observation settings are given in Appendix B.

The answerer $\mathcal{G}$ generates Y from these frames, retaining URF-HVAA's suspicious-phrase prompting and gated spatial conditioning (Lin et al., 2025). The same description is evaluated as the VAU output and supplied to both feedback levels. Generation settings are provided in Appendix C.

3.3 VIDEO-LEVEL SEMANTIC FEEDBACK

The first feedback level determines how the video's overall score distribution should be treated. It combines a tentative interpretation from Y with selective visual rechecking when that interpretation is weakly reflected in the local captions.

Tentative interpretation and visual rechecking. A lexical parser assigns Y to one of four coarse states: `normalish`, `unclear`, `mixed`, or `anomalyish`. These states induce tentative SUPPRESS, KEEP, SHARPEN, and SHARPEN actions, respectively. The state `mixed` only denotes simultaneous matches to normality and anomaly expressions and does not resolve their semantic relationship.

We selectively invoke visual rechecking when the VAU description indicates a potential anomaly that is weakly supported by the existing local captions. The frozen verifier observes frames sampled from the full video and from a high-score interval together with the VAU hypothesis. The hypothesis is treated as untrusted text: the verifier is instructed to rely on visible abnormal evidence rather than inferred intent or criminality. Ambiguous or internally inconsistent verification conservatively preserves the input scores, whereas clearly unsupported anomaly hypotheses may trigger suppression; otherwise the tentative action is retained. Exactly one video-level operation is applied. The eligibility threshold and visual sampling settings are provided in Appendix D.

Score modulation. Let a be the final action, Q_τ the τ-quantile operator computed with linear interpolation, and $m \in [0,1]$ the minimum peak threshold. The peak set is

$$\theta_v = \max\{m, \mathrm{Q}_{0.70}(b^{(0)})\}, \qquad H = \{i : b_i^{(0)} \geq \theta_v\}. \tag{3}$$

Let $\alpha, \beta \in [0,1]$ denote the suppression and non-peak attenuation factors, respectively, and let $f \in [0,1]$ denote the peak-score floor. Video-level modulation is

$$b_i^{(v)} = \begin{cases} \alpha b_i^{(0)}, & a = \text{SUPPRESS}, \\ b_i^{(0)}, & a = \text{KEEP}, \\ \max\{b_i^{(0)}, f\}, & a = \text{SHARPEN},\ i \in H, \\ \beta b_i^{(0)}, & a = \text{SHARPEN},\ i \notin H, \end{cases} \tag{4}$$

followed by clipping to $[0,1]$. Suppression reduces the overall score scale. Sharpening attenuates non-peak scores and imposes a floor on scores in H, when this set is nonempty. These operations modulate the existing score sequence without independently localizing an event interval. A single shared parameter configuration is used across all evaluated datasets and compared variants; its values are reported in Appendix D. The transformed scores are not interpreted as calibrated anomaly probabilities.

3.4 EVIDENCE-GROUNDED LOCAL FEEDBACK

Local feedback separates three decisions: event semantics proposes candidate windows, visual verification authorizes promotion, and current-score ranks select the anchors to update.

Event-conditioned retrieval. A separate parser extracts visible event hypotheses from Y. Local promotion is enabled only when this parser judges the video anomalous, independently of the video-level action. TF–IDF matching against C retrieves non-overlapping caption windows for each hypothesis, with $b^{(0)}$ providing additional ranking and filtering cues.

Target-local visual verification. A verifier inspects each candidate window and surrounding frames, using the event hypothesis and matched captions as noisy retrieval cues. Promotion requires the verifier to identify direct abnormal evidence inside the target clip; surrounding context alone cannot authorize it. The verifier authorizes a window, while anchor selection remains score-based.

Local score promotion. Initialize the current score sequence $x = b^{(v)}$. For an approved window w in original-frame coordinates containing at least one score anchor, let $x_w = (x_i : 16i \in w)$. Let $\tau_\ell \in [0,1]$ be the local quantile level and $\lambda \in [0,1]$ the promotion strength:

$$M_w = \{i : 16i \in w,\ x_i \geq \mathrm{Q}_{\tau_\ell}(x_w)\}, \qquad x_i \leftarrow x_i + \lambda(1 - x_i), \quad i \in M_w. \tag{5}$$

Other anchors remain unchanged. Approved windows are processed sequentially, recomputing the quantile from the current scores after preceding updates. After all updates, set $b^{(1)} = x$. Feedback settings are given in Appendix D; final smoothing, frame-level expansion, and evaluation are described in Appendix E.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate video anomaly detection on UCF-Crime, XD-Violence, UBNormal, and MSAD, following the evaluation protocols used by prior VAD methods (Sultani et al., 2018; Wu et al., 2020; Acsintoae et al., 2022; Zhu et al., 2024). For UCF-Crime and UBNormal, the main benchmark metric is frame-level AUC; we additionally report UCF-Crime AP in the feedback analyses. For XD-Violence and MSAD, we report frame-level AUC and average precision (AP); MSAD follows its official micro-AUC/AP evaluation protocol. For VAU, we follow the HIVAU-70K video-summary setting used by URF-HVAA, with 251 UCF-Crime and 800 XD-Violence videos (Zhang et al., 2025b; Lin et al., 2025). We report BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavie, 2005), and ROUGE-L (Lin, 2004). We also report GPT-based scores for reasonability, detail, and consistency following Hawk (Tang et al., 2024). The GPT evaluator compares generated and reference descriptions without access to video pixels.

Models and implementation. All model weights remain frozen throughout inference. The upstream URF-HVAA system provides local captions and the refined temporal anomaly-score curve. Qwen2.5-VL-7B-Instruct (Bai et al., 2025) serves as the candidate analyst during active observation, and VideoLLaMA3-7B (Zhang et al., 2025a) generates the final open-ended VAU description. Qwen2-VL-7B-Instruct (Wang et al., 2024) is used independently for video-level rechecking and target-local visual verification. The final VAU answerer receives 16 selected frames; this is the answer-generation budget rather than the total number of visual observations made by the complete pipeline. Implementation details are given in Appendices B–D; metric definitions and inference costs are reported in Appendix E.

Main configuration and comparisons. The main configuration uses URF-HVAA refined scores for both observation and feedback initialization, and applies video-level followed by local feedback. Tables 1 and 2 compare this configuration with existing methods; ablations examine observation strategies and feedback stages. A single feedback parameter configuration is shared across all evaluated datasets and held fixed across the compared variants.

4.2 VIDEO ANOMALY DETECTION

Table 1 compares the complete framework with zero-shot, training-free methods and includes trained or reference-based methods for broader coverage of UBNormal and MSAD.

Our framework obtains 86.25% AUC on UCF-Crime, 91.74% AUC / 72.58% AP on XD-Violence, and 86.0% AUC / 76.9% AP on MSAD. Relative to the published URF-HVAA results, these correspond to gains of 1.97 AUC points on UCF-Crime, 0.40/4.51 AUC/AP points on XD-Violence, and 0.1/0.5 AUC/AP points on MSAD. On UBNormal, our method obtains 68.67% AUC, compared with 68.98% for URF-HVAA. These comparisons favor our method on the three real-world benchmarks, with a small decrease on synthetic UBNormal. Table 4 separately compares the feedback stages with the upstream scores under the same evaluation setting.

4.3 VIDEO ANOMALY UNDERSTANDING

Table 2 evaluates the generated descriptions before they are used for detector refinement. Compared with the re-evaluated URF-HVAA descriptions, our method improves all three GPT-rated dimensions on both datasets. Lexical metrics also improve, except for UCF-Crime CIDEr, which is unchanged at the reported precision. Table 3 further examines the effects of anomaly guidance and active observation.

Table 1: Comparison with representative video anomaly detection methods. The upper block includes methods using additional training or reference data. The lower block reports selected zero-shot, training-free methods. Legacy comparison values follow the published VAD comparison in URF-HVAA (Lin et al., 2025); MSAD results for RTFM and MGFN follow the official MSAD benchmark (Zhu et al., 2024). Bold denotes the highest reported value within the lower block. All values are percentages; "–" denotes an unreported metric.

Method	Zero-shot	Training-free	UCF-Crime	XD-Violence		UBNormal	MSAD	
			AUC(%)	AUC(%)	AP(%)	AUC(%)	AUC(%)	AP(%)
Additional training or reference data								
RTFM (Tian et al., 2021)	×	×	83.31	–	77.81	60.94	86.7	66.3
MGFN (Chen et al., 2023)	×	×	86.98	–	80.11	–	85.0	63.5
MULDE (Micorek et al., 2024)	×	×	78.50	–	–	72.80	–	–
EGO (Ding et al., 2025)	×	×	81.71	–	65.77	–	87.3	64.4
AnomalyRuler (Yang et al., 2024)	×	✓	–	–	–	71.90	–	–
Zero-shot, training-free								
LAVAD (Zanella et al., 2024)	✓	✓	80.28	85.36	62.01	–	–	–
EventVAD (Shao et al., 2025)	✓	✓	82.03	87.51	64.04	–	–	–
URF-HVAA (Lin et al., 2025)	✓	✓	84.28	91.34	68.07	**68.98**	85.9	76.4
CoReVAD (Lim & Hur, 2026)	✓	✓	82.51	91.44	70.94	–	–	–
ReCI (Mo et al., 2026b)	✓	✓	81.51	91.60	72.23	–	–	–
Ours	✓	✓	**86.25**	**91.74**	**72.58**	68.67	**86.0**	**76.9**

Table 2: Video anomaly understanding on the HIVAU video-summary setting. GPT-R, GPT-D, and GPT-C denote reasonability, detail, and consistency, respectively. Published baseline values follow URF-HVAA (Lin et al., 2025); the final row reports our VAD-guided active observation pipeline. URF-HVAA (re-eval.) reports the released URF-HVAA descriptions under the same evaluator used for our generated descriptions.

Method	UCF-Crime							XD-Violence						
	BLEU	CIDEr	METEOR	ROUGE-L	GPT-R	GPT-D	GPT-C	BLEU	CIDEr	METEOR	ROUGE-L	GPT-R	GPT-D	GPT-C
Supervised Fine-tuning														
Hawk (Tang et al., 2024)	0.379	0.008	0.217	0.187	0.255	0.580	0.214	0.375	0.016	0.176	0.188	0.408	0.586	0.365
HolmesVAU (Zhang et al., 2025b)	0.435	0.021	0.194	0.257	0.448	0.356	0.391	0.376	0.011	0.182	0.253	0.715	0.581	0.673
Zero-shot, training-free														
InternVideo2.5-8B (Wang et al., 2025)	0.159	0.011	0.088	0.103	0.240	0.266	0.205	0.209	0.013	0.119	0.130	0.456	0.447	0.433
VideoChat-Flash-2B (Li et al., 2026)	0.165	0.008	0.108	0.168	0.488	0.283	0.404	0.277	0.026	0.144	0.186	0.690	0.576	0.627
VideoLLaMA3-7B (Zhang et al., 2025a)	0.215	0.014	0.117	0.156	0.463	0.289	0.384	0.290	0.022	0.141	0.169	0.568	0.487	0.499
URF-HVAA (Lin et al., 2025)	0.345	**0.023**	0.175	0.188	0.512	0.428	0.444	0.399	0.029	0.198	0.200	0.721	0.707	0.668
URF-HVAA (re-eval.)	0.345	**0.023**	0.175	0.188	0.506	0.443	0.451	0.404	0.028	0.202	0.200	0.721	0.708	0.670
Ours	**0.386**	**0.023**	**0.183**	**0.196**	**0.524**	**0.475**	**0.466**	**0.465**	**0.047**	**0.211**	**0.212**	**0.731**	**0.730**	**0.682**

Qualitative analysis. Figure 3 illustrates the progression from VAD-guided observation to semantic feedback in a motorcycle-fire example. The displayed local caption describes a motorcycle and a person on the ground, while the final VAU description identifies a fire and people moving away from the scene. The description guides video-level sharpening, followed by visually authorized local promotion. At the illustrated location, the score increases from 0.600 to 0.850 after video-level feedback and to 0.9925 after local feedback.

4.4 Ablation Studies

Anomaly guidance and active observation. Table 3 provides two controlled comparisons. The one-shot baseline is adapted from Holmes-VAU's Anomaly-focused Temporal Sampler (ATS) (Zhang et al., 2025b), which samples frames according to the cumulative anomaly-score distribution. In our comparison, the baseline uses the same VAD prior and 16-frame final-generation conditions as our method, without updating its selection from intermediate visual observations. Within the active-observation pipeline, replacing CLIP guidance with VAD guidance raises GPT-

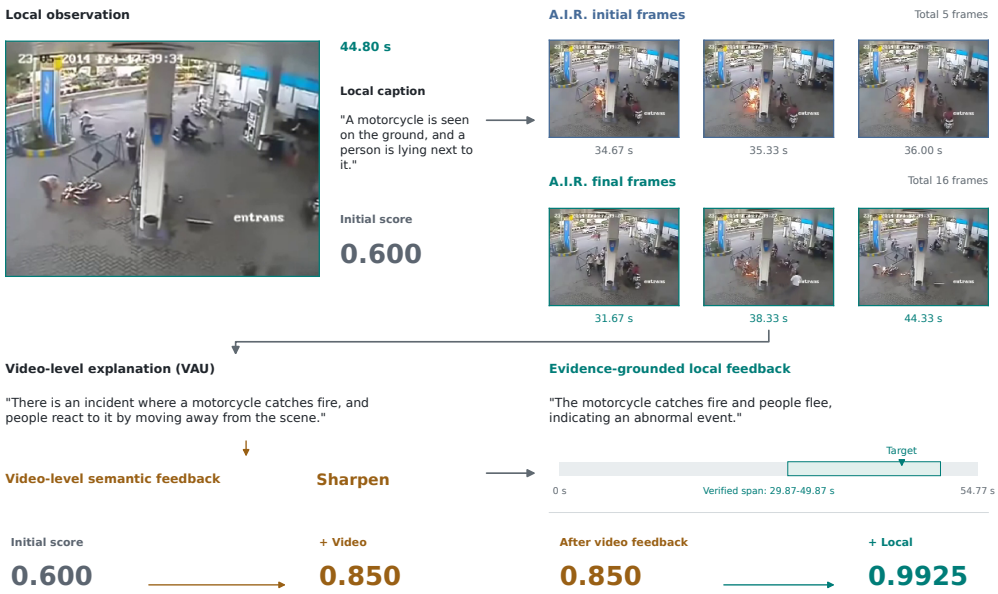


Figure 3: Motorcycle-fire example. VAD-guided observations support the VAU description, which guides video-level sharpening and visually authorized local promotion. The displayed scores track both feedback stages at the illustrated location.

Table 3: VAU observation ablation. All rows share the 16-frame answerer input budget, answerer, prompting, spatial conditioning, and generation settings. ATS and Ours share the VAD prior; AIR–CLIP and Ours share the active-observation pipeline. Active observation additionally uses candidate-analysis calls. GPT-Mean averages GPT-R, GPT-D, and GPT-C. Higher is better.

Method	Guidance	Active	UCF-Crime					XD-Violence				
			BLEU	METEOR	ROUGE-L	CIDEr	GPT-Mean	BLEU	METEOR	ROUGE-L	CIDEr	GPT-Mean
ATS	VAD	×	0.3854	0.1829	**0.1970**	**0.0357**	0.4812	0.4344	0.1977	0.2089	**0.0488**	0.6724
AIR–CLIP	CLIP	✓	0.3541	0.1735	0.1897	0.0230	0.4327	0.4334	0.2035	0.2068	0.0320	0.6897
Ours	VAD	✓	**0.3856**	**0.1833**	0.1959	0.0232	**0.4884**	**0.4647**	**0.2111**	**0.2117**	0.0468	**0.7144**

Table 4: Cumulative ablation of the two semantic-feedback levels. All rows use the URF-refined score base; feedback rows use the VAU description generated by our VAD-guided active observation. Video-level feedback includes selective visual rechecking, and the final row additionally applies evidence-grounded local promotion.

	UCF-Crime		XD-Violence	
Configuration	AUC(%)	AP(%)	AUC(%)	AP(%)
URF-refined base	84.36	36.16	91.32	68.07
+ Video-level feedback	86.23	37.57	91.64	71.37
+ Local feedback (**Ours**)	**86.25**	**38.36**	**91.74**	**72.58**

Mean by 0.0557 on UCF-Crime and 0.0247 on XD-Violence, supporting anomaly-score guidance for the broad VAU instruction.

With the VAD prior and final-generation settings fixed, active observation raises GPT-Mean over ATS by 0.0072 on UCF-Crime and 0.0420 on XD-Violence. Lexical metrics are mixed on UCF-Crime. On XD-Violence, BLEU, METEOR, and ROUGE-L improve, while CIDEr remains slightly higher with ATS.

Table 5: Mean scores and incremental changes under sequential refinement on UCF-Crime. GT-normal and GT-anomaly denote normal and anomalous temporal regions according to the ground-truth annotations. Parentheses report changes from the preceding row.

Score sequence	GT-normal (Δ)	GT-anomaly (Δ)	Difference
RAW	0.2217 (–)	0.3013 (–)	0.0796
+URF-HVAA (Lin et al., 2025)	0.2445 (+0.02286)	0.3223 (+0.02105)	0.0778
+Ours	0.2449 (+0.00036)	0.3366 (+0.01428)	0.0917

Two-level semantic feedback. Table 4 shows that the video-level stage yields the larger incremental gain on all four metrics in this cumulative ablation. Local feedback provides further improvements: 0.79 AP points on UCF-Crime and 1.21 AP points on XD-Violence, alongside AUC gains of 0.02 and 0.10 points, respectively. Its additional contribution is thus more apparent in AP than in AUC after video-level modulation.

Score changes in normal and anomalous regions. Table 5 tracks mean scores from the RAW baseline through URF-HVAA refinement and our semantic feedback. URF-HVAA refinement produces similar mean increases in normal and anomalous regions (+0.02286 and +0.02105, respectively). Our semantic feedback further increases the anomalous-region mean by +0.01428, while leaving the normal-region mean nearly unchanged (+0.00036). Consequently, the mean within-video anomaly-background score difference increases from 0.0778 to 0.0917. This more differentiated score response supports the effectiveness of considering which temporal regions warrant refinement when applying semantic feedback.

Parameter sensitivity. Appendix D evaluates each parameter while holding the others fixed. Semantic feedback improves both AUC and AP on both datasets across all tested values of α, f, and λ, and across multiple settings of β, m, and the quantile thresholds. The shared configuration balances AUC and AP gains across the two datasets while maintaining selective score updates. These results indicate that the improvements are not confined to a single parameter setting.

Description source and refinement base. Appendix A examines the effect of the description source under a fixed feedback configuration. With the URF-refined base and both feedback stages, our descriptions yield higher AP than the released URF-HVAA descriptions on UCF-Crime (38.36 vs. 38.16) and XD-Violence (72.58 vs. 71.49). This comparison supports their use for detector refinement, complementing the independent VAU evaluation.

5 Conclusion

We presented a training-free framework that returns open-ended anomaly understanding to temporal detection. Its two feedback levels separate video-wide score modulation from local promotion supported by target-clip evidence. Controlled comparisons show that VAD-guided active observation improves GPT-rated description quality, while cumulative ablations show complementary gains from the two feedback stages. These results support using a VAU description both as an explanation and as a source of hypotheses for refining an existing anomaly-score sequence.

AI Use Statement

Generative AI tools were used to assist with literature retrieval, critical analysis of the proposed methodology, experimental analysis and interpretation, and the drafting and revision of portions of the manuscript. The motivation, problem formulation, and principal methodological ideas of this work were developed by the authors. AI-assisted analysis of experimental outcomes was also used to support subsequent methodological refinements, including the exploration and adjustment of experimental parameters and implementation choices. All AI-assisted suggestions, analyses, and generated text were reviewed and verified by the authors, who take full responsibility for the methodology, experiments, results, claims, and final content of the paper.

REFERENCES

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UBnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.08644.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

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. URL https://aclanthology.org/W05-0909/.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 387–395, 2023. doi: 10.1609/aaai.v37i1.25112. URL https://ojs.aaai.org/index.php/AAAI/article/view/25112.

Ying Cheng, Yu-Ho Lin, Min-Hung Chen, Fu-En Yang, and Shang-Hong Lai. VADER: Towards causal video anomaly understanding with relation-aware large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. URL https://arxiv.org/abs/2511.07299.

Dexuan Ding, Lei Wang, Liyun Zhu, Tom Gedeon, and Piotr Koniusz. Learnable expansion of graph operators for multi-modal feature fusion. In *International Conference on Learning Representations*, 2025.

Yuxiang Duan, Huining Li, Ao Li, Shuai Feng, Lanju Kong, Ning Liu, Jian Zhang, Xingdong Sheng, and Yuntao Du. AgenticVAU: Multi-agent explore–verify reasoning for video anomaly understanding. *arXiv preprint arXiv:2608.03779*, 2026. URL https://arxiv.org/abs/2608.03779.

Shibo Gao, Peipei Yang, Yangyang Liu, Yi Chen, Han Zhu, Xuyao Zhang, and Linlin Huang. VAGU & GtS: LLM-based benchmark and framework for joint video anomaly grounding and understanding. *arXiv preprint arXiv:2507.21507*, 2025. URL https://arxiv.org/abs/2507.21507.

Satoshi Hashimoto, Hitoshi Nishimura, and Mori Kurokawa. MuST-VAD: Mutual structured learning for video anomaly detection. *arXiv preprint arXiv:2608.06913*, 2026. URL https://arxiv.org/abs/2608.06913.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. VideoChat-Flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b14d7175755b180dc2163e15e3110cb6-Abstract-Conference.html.

Hyeongmuk Lim and Youngbum Hur. CoReVAD: A contextual reasoning framework for training-free video anomaly detection. *arXiv preprint arXiv:2605.23116*, 2026. URL https://arxiv.org/abs/2605.23116.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004. URL https://aclanthology.org/W04-1013/.

Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2511.00962.

Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18868–18877, 2024.

Mengjingcheng Mo, Jiaxu Leng, and Xinbo Gao. Learning to watch: Active video anomaly understanding via interleaved policy optimization. In *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2607.00622.

Mengjingcheng Mo, Jiankang Zheng, Jiaxu Leng, and Xinbo Gao. Retrieval-guided contextual inference for training-free video anomaly detection in low-light scenarios. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 930–939, 2026b. doi: 10.1145/3805622.3810823. URL https://doi.org/10.1145/3805622.3810823.

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. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, Xiaochen Wang, Hao Tang, Yan Wang, and Shuyan Li. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2586–2595, 2025. doi: 10.1145/3746027.3754500.

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*, 2018. URL https://arxiv.org/abs/1801.04264.

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 *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fca83589e85cb061631b7ebc5db5d6bd-Abstract-Conference.html.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4975–4986, 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*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html.

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. URL https://arxiv.org/abs/2409.12191.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2.5: Empowering video MLLMs with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. URL https://arxiv.org/abs/2501.12386.

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 *Computer Vision – ECCV 2020*, pp. 322–339, 2020. doi: 10.1007/978-3-030-58577-8_20. URL https://arxiv.org/abs/2007.04687.

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 *European Conference on Computer Vision*, 2024.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. PANDA: Towards generalist video anomaly detection via agentic AI engineer. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2509.26386.

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*, 2024. URL https://arxiv.org/abs/2404.01014.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL https://arxiv.org/abs/2501.13106.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. URL https://arxiv.org/abs/2412.06171.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2856. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a3c5af1f56fc73eef1ba0f442739f5ca-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.04428.

A ADDITIONAL ANALYSIS OF SEMANTIC FEEDBACK

We additionally examine how semantic feedback behaves under different description sources and refinement bases. We compare the released URF-HVAA description with the description generated by our VAD-guided active observation, while applying one shared feedback configuration across both datasets and all compared variants. We also evaluate both the original RAW score sequence and the URF-refined score sequence as refinement bases. Table 6 reports these variants as additional analysis rather than as separate main methods.

Table 6: Additional analysis of description sources, refinement bases, and feedback stages. "Video" denotes video-level semantic feedback including selective visual rechecking; "Video + local" additionally applies evidence-grounded local promotion. The same shared feedback configuration is used across datasets, description sources, and refinement bases.

Refinement base	Description / feedback	UCF-Crime		XD-Violence	
		AUC	AP	AUC	AP
RAW	URF-HVAA / Video	84.33	35.73	88.78	64.26
	URF-HVAA / Video + local	84.31	35.92	89.23	66.17
	Ours / Video	84.33	36.36	88.95	64.69
	Ours / Video + local	**84.35**	**37.10**	**89.29**	**66.61**
URF-refined	URF-HVAA / Video	86.21	37.61	91.55	70.63
	URF-HVAA / Video + local	86.19	38.16	91.61	71.49
	Ours / Video	86.23	37.57	91.64	71.37
	Ours / Video + local	**86.25**	**38.36**	**91.74**	**72.58**

Description source and refinement base. Across both refinement bases, adding local feedback consistently improves AP for all evaluated description sources and datasets. On XD-Violence, local feedback also improves AUC in every configuration. On UCF-Crime, the AUC effect is small and mixed: descriptions from our VAD-guided active observation yield slight improvements, whereas URF-HVAA descriptions produce decreases of only 0.02 points. With both video-level and local feedback enabled, descriptions from our VAD-guided active observation yield higher AP than URF-HVAA descriptions for both refinement bases on UCF-Crime and XD-Violence. The more consistent gains in AP than in AUC indicate that local promotion benefits precision–recall performance in these comparisons, although the aggregate metrics do not directly measure accuracy within the retrieved windows.

B OBSERVATION AND FIXED-BUDGET CONSTRUCTION

Anomaly prior and initialization. We Gaussian-smooth $b^{(0)}$ with $\sigma = 10$ and expand the scores to frame level at stride 16, without additional min–max normalization. A two-component Gaussian mixture identifies high-score regions from which the initial candidate frames are selected.

Active observation and final selection. We adapt the iterative observation procedure of A.I.R. (Zou et al., 2026), using anomaly scores to guide initialization and interval prioritization. The frozen candidate analyst assigns relevance scores from 1 to 5, which are combined with the anomaly scores to guide subsequent candidate selection and localized resampling. All runs use six observation rounds, with up to 24 candidate indices per round and a nominal internal selection budget of 32 frames.

For the final answerer, we construct a 16-frame input using anomaly and analyst scores with a temporal spacing preference. Any deficit is filled with unused indices from a uniform temporal grid, and the final frames are ordered chronologically. The 16-frame budget applies to answer generation, not to the total visual observations made during selection.

Interval potential. We retain the interval-ranking mechanism of A.I.R. (Zou et al., 2026). For a temporal interval I, its core potential combines relevance, temporal variation, and interval length:

$$P_{\text{core}}(I) = \underbrace{R(I)}_{\text{relevance}} \underbrace{\big(1 + 0.5C(I)\big)}_{\text{complexity}} \underbrace{\big(1 + 0.2 \log L_I\big)}_{\text{length}}, \tag{6}$$

where $R(I)$ aggregates relevance within the interval, $C(I)$ measures its temporal variation, and L_I is the interval length in original-frame coordinates. These factors favor relevant, temporally varying, and longer intervals for further inspection.

Our adaptation replaces the underlying CLIP similarity signal with the smoothed, frame-expanded scores derived from $b^{(0)}$, while retaining the analyst's relevance feedback. Score aggregation and the additional exploration-density adjustment follow the released A.I.R. implementation. Endpoints of high-potential intervals provide candidates for the next analyst evaluation.

C Visual Inputs and Generation Settings

Candidate analyst. Qwen2.5-VL-7B-Instruct assesses candidate frames independently using the inherited relevance rubric, returning a score from 1 to 5 and an explanation. Generation uses a 64-token output limit, temperature 0.01, top-$k = 1$, and top-$p = 0.0001$.

Final VAU generation. VideoLLaMA3-7B receives 16 selected frames in chronological order with their timestamps. We retain URF-HVAA's suspicious-phrase prompting and gated spatial conditioning, using either original frames or the cached spatially annotated source. The final input contains neither analyst responses nor reference descriptions.

The instruction requests anomalous activities and their explanation without timestamps, or a normal-scene description if no anomaly is found. The answerer generates one description per video with a 1,024-token output limit, temperature 0.1, and sampling enabled. Compared observation conditions share the final-generation settings and per-video seed. The candidate and final-generation prompts are included with the LaTeX source.

D Feedback Settings and Parameter Sensitivity

Table 7: Feedback calibration parameters shared across all evaluated datasets.

α	β	f	m	Global quantile	λ	τ_ℓ
0.40	0.80	0.85	0.45	0.70	0.95	0.85

Parameter sensitivity. Table 7 gives the shared configuration, and Tables 8–14 report one-parameter sensitivity analyses. We evaluate one parameter at a time while fixing all remaining parameters to the shared main configuration. The "Main config." row reports the performance obtained with the single parameter setting used across all evaluated datasets. For each sweep, the bold entry marks the value used in this shared configuration.

Overall, the shared configuration lies in a relatively stable performance region for most parameters rather than at an isolated per-metric optimum. The suppression scale α is particularly stable around the selected value, while β, f, and λ mainly alter the balance between AUC and AP. The minimum peak threshold m, global quantile, and local quantile τ_ℓ exhibit clearer transition regions: overly permissive settings can broaden score updates and substantially reduce performance. Several individual sweep values outperform the main configuration on particular metrics, indicating that the shared setting is not selected as the independent optimum of each dataset or metric.

Video-level rechecking. The lexical interpretation distinguishes descriptions containing normality expressions, anomaly expressions, both, or neither. Visual rechecking is eligible when the description contains anomaly expressions but at most 1% of local captions do. Qwen2-VL-7B-Instruct receives a 16-frame video overview and 12 frames from a 300-frame high-score target interval. The

textual input contains the VAU description and up to four suspicious-category cues obtained by fixed rule matching of that description. These cues do not use ground-truth categories. Ambiguous or inconsistent verification preserves the input scores.

Local retrieval and verification. A separate parser extracts event hypotheses and enables local promotion only when its overall judgment indicates an anomaly. Caption windows use width 600 and stride 64 in original-frame coordinates. TF–IDF cosine similarity ranks the windows, with original-base score statistics resolving ties. Up to three non-overlapping candidates are selected per event, then filtered by a similarity threshold of 0.10 and a minimum original-base score of 0.50 at caption anchors. Rejected candidates are not replaced.

The verifier receives the event hypothesis, matched captions, 16 target frames, and eight context frames sampled over the target interval extended by 300 frames on either side. Promotion requires direct abnormal evidence within the target. Approved windows use the sequential, current-score quantile update defined in the main text.

Verifier generation. Both feedback stages use frozen Qwen2-VL-7B-Instruct with greedy generation and output limits of 260 tokens for video-level checks and 240 tokens for local checks. Final score reconstruction and evaluation are described in Appendix E.

Table 8: Sensitivity to α with all other calibration parameters fixed to the shared main configuration. Bold marks the value used in the main configuration. Parentheses report changes from the re-evaluated URF-refined base in percentage points.

α	UCF-Crime		XD-Violence	
	AUC (%)	AP (%)	AUC (%)	AP (%)
Baseline (Lin et al., 2025)	84.356	36.164	91.323	68.071
Main config.	86.246 (+1.890)	38.363 (+2.199)	91.738 (+0.414)	72.584 (+4.514)
0.05	86.551 (+2.195)	38.437 (+2.273)	91.656 (+0.332)	72.557 (+4.486)
0.1	86.572 (+2.216)	38.439 (+2.274)	91.667 (+0.343)	72.559 (+4.489)
0.15	86.587 (+2.231)	38.439 (+2.275)	91.680 (+0.357)	72.563 (+4.492)
0.2	86.571 (+2.215)	38.435 (+2.271)	91.693 (+0.370)	72.567 (+4.496)
0.3	86.461 (+2.105)	38.411 (+2.247)	91.718 (+0.395)	72.576 (+4.505)
0.4	**86.246 (+1.890)**	**38.363 (+2.199)**	**91.738 (+0.414)**	**72.584 (+4.514)**
0.425	86.184 (+1.828)	38.348 (+2.184)	91.742 (+0.418)	72.586 (+4.516)
0.6	85.674 (+1.318)	38.212 (+2.048)	91.762 (+0.438)	72.597 (+4.527)
0.8	84.975 (+0.618)	37.957 (+1.793)	91.767 (+0.443)	72.601 (+4.530)
0.95	84.367 (+0.010)	37.649 (+1.485)	91.761 (+0.438)	72.597 (+4.526)

Table 9: Sensitivity to β with all other calibration parameters fixed to the shared main configuration. Bold marks the value used in the main configuration. Parentheses report changes from the re-evaluated URF-refined base in percentage points.

β	UCF-Crime		XD-Violence	
	AUC (%)	AP (%)	AUC (%)	AP (%)
Baseline (Lin et al., 2025)	84.356	36.164	91.323	68.071
Main config.	86.246 (+1.890)	38.363 (+2.199)	91.738 (+0.414)	72.584 (+4.514)
0.1	82.938 (-1.418)	39.858 (+3.694)	90.113 (-1.211)	72.556 (+4.485)
0.25	84.323 (-0.033)	39.973 (+3.809)	90.880 (-0.443)	73.199 (+5.128)
0.4	85.251 (+0.895)	39.863 (+3.699)	91.354 (+0.030)	73.514 (+5.444)
0.55	85.801 (+1.445)	39.513 (+3.349)	91.619 (+0.296)	73.492 (+5.421)
0.7	86.121 (+1.765)	38.900 (+2.736)	91.731 (+0.408)	73.095 (+5.024)
0.8	**86.246 (+1.890)**	**38.363 (+2.199)**	**91.738 (+0.414)**	**72.584 (+4.514)**
0.9	86.324 (+1.968)	37.785 (+1.620)	91.699 (+0.376)	71.853 (+3.782)
0.95	86.347 (+1.991)	37.467 (+1.303)	91.666 (+0.342)	71.388 (+3.318)

Table 10: Sensitivity to f with all other calibration parameters fixed to the shared main configuration. Bold marks the value used in the main configuration. Parentheses report changes from the re-evaluated URF-refined base in percentage points.

f	UCF-Crime		XD-Violence	
	AUC (%)	AP (%)	AUC (%)	AP (%)
Baseline (Lin et al., 2025)	84.356	36.164	91.323	68.071
Main config.	86.246 (+1.890)	38.363 (+2.199)	91.738 (+0.414)	72.584 (+4.514)
0.5	86.197 (+1.840)	39.276 (+3.111)	91.522 (+0.199)	70.878 (+2.807)
0.55	86.219 (+1.863)	39.319 (+3.155)	91.529 (+0.206)	70.882 (+2.811)
0.65	86.267 (+1.910)	39.417 (+3.252)	91.575 (+0.252)	71.036 (+2.965)
0.75	86.275 (+1.919)	39.367 (+3.202)	91.671 (+0.348)	71.521 (+3.451)
0.85	**86.246 (+1.890)**	**38.363 (+2.199)**	**91.738 (+0.414)**	**72.584 (+4.514)**
0.95	86.177 (+1.820)	36.940 (+0.776)	91.765 (+0.442)	72.704 (+4.634)
1	86.134 (+1.778)	36.284 (+0.120)	91.755 (+0.432)	72.680 (+4.610)

Table 11: Sensitivity to m with all other calibration parameters fixed to the shared main configuration. Bold marks the value used in the main configuration. Parentheses report changes from the re-evaluated URF-refined base in percentage points.

m	UCF-Crime		XD-Violence	
	AUC (%)	AP (%)	AUC (%)	AP (%)
Baseline (Lin et al., 2025)	84.356	36.164	91.323	68.071
Main config.	86.246 (+1.890)	38.363 (+2.199)	91.738 (+0.414)	72.584 (+4.514)
0.05	82.655 (-1.701)	26.925 (-9.240)	89.623 (-1.700)	68.973 (+0.902)
0.1	82.655 (-1.701)	26.925 (-9.240)	89.623 (-1.700)	68.973 (+0.902)
0.15	82.787 (-1.569)	27.005 (-9.159)	89.727 (-1.596)	69.090 (+1.020)
0.2	82.787 (-1.569)	27.005 (-9.159)	89.727 (-1.596)	69.090 (+1.020)
0.25	86.211 (+1.855)	37.045 (+0.881)	91.428 (+0.105)	72.120 (+4.049)
0.35	86.341 (+1.984)	38.711 (+2.547)	91.662 (+0.339)	72.462 (+4.391)
0.45	**86.246 (+1.890)**	**38.363 (+2.199)**	**91.738 (+0.414)**	**72.584 (+4.514)**
0.55	86.092 (+1.735)	38.410 (+2.246)	91.703 (+0.380)	72.639 (+4.568)
0.6	86.092 (+1.735)	38.410 (+2.246)	91.703 (+0.380)	72.639 (+4.568)

Table 12: Sensitivity to the global quantile with all other calibration parameters fixed to the shared main configuration. Bold marks the value used in the main configuration. Parentheses report changes from the re-evaluated URF-refined base in percentage points.

Global quantile	UCF-Crime		XD-Violence	
	AUC (%)	AP (%)	AUC (%)	AP (%)
Baseline (Lin et al., 2025)	84.356	36.164	91.323	68.071
Main config.	86.246 (+1.890)	38.363 (+2.199)	91.738 (+0.414)	72.584 (+4.514)
0.1	86.117 (+1.761)	34.583 (-1.581)	91.432 (+0.108)	70.496 (+2.425)
0.25	86.117 (+1.761)	34.583 (-1.581)	91.553 (+0.229)	71.716 (+3.645)
0.4	86.138 (+1.782)	35.209 (-0.956)	91.733 (+0.410)	72.589 (+4.518)
0.55	86.172 (+1.816)	35.875 (-0.289)	91.788 (+0.465)	72.952 (+4.881)
0.7	**86.246 (+1.890)**	**38.363 (+2.199)**	**91.738 (+0.414)**	**72.584 (+4.514)**
0.85	86.317 (+1.961)	39.755 (+3.591)	91.487 (+0.163)	70.661 (+2.591)
0.95	86.386 (+2.030)	40.335 (+4.171)	91.484 (+0.161)	70.718 (+2.647)

Table 13: Sensitivity to λ with all other calibration parameters fixed to the shared main configuration. Bold marks the value used in the main configuration. Parentheses report changes from the re-evaluated URF-refined base in percentage points.

λ	UCF-Crime		XD-Violence	
	AUC (%)	AP (%)	AUC (%)	AP (%)
Baseline (Lin et al., 2025)	84.356	36.164	91.323	68.071
Main config.	86.246 (+1.890)	38.363 (+2.199)	91.738 (+0.414)	72.584 (+4.514)
0.05	86.234 (+1.877)	37.654 (+1.489)	91.655 (+0.332)	71.460 (+3.389)
0.1	86.241 (+1.885)	37.738 (+1.573)	91.667 (+0.344)	71.553 (+3.482)
0.2	86.251 (+1.895)	37.862 (+1.698)	91.688 (+0.365)	71.734 (+3.663)
0.3	86.257 (+1.901)	37.966 (+1.802)	91.704 (+0.381)	71.893 (+3.823)
0.45	86.261 (+1.905)	38.113 (+1.949)	91.721 (+0.398)	72.114 (+4.043)
0.65	86.259 (+1.902)	38.225 (+2.061)	91.734 (+0.410)	72.358 (+4.287)
0.8	86.253 (+1.897)	38.298 (+2.133)	91.738 (+0.415)	72.488 (+4.417)
0.95	**86.246 (+1.890)**	**38.363 (+2.199)**	**91.738 (+0.414)**	**72.584 (+4.514)**

Table 14: Sensitivity to τ_ℓ with all other calibration parameters fixed to the shared main configuration. Bold marks the value used in the main configuration. Parentheses report changes from the re-evaluated URF-refined base in percentage points.

τ_ℓ	UCF-Crime		XD-Violence	
	AUC (%)	AP (%)	AUC (%)	AP (%)
Baseline (Lin et al., 2025)	84.356	36.164	91.323	68.071
Main config.	86.246 (+1.890)	38.363 (+2.199)	91.738 (+0.414)	72.584 (+4.514)
0.1	85.889 (+1.533)	34.835 (-1.330)	89.747 (-1.576)	61.320 (-6.751)
0.25	85.884 (+1.527)	33.707 (-2.457)	90.214 (-1.109)	63.830 (-4.241)
0.4	86.018 (+1.662)	34.179 (-1.985)	90.962 (-0.361)	67.271 (-0.800)
0.5	85.984 (+1.628)	34.666 (-1.499)	91.239 (-0.084)	68.766 (+0.695)
0.65	86.141 (+1.784)	35.987 (-0.177)	91.789 (+0.466)	72.337 (+4.266)
0.75	86.265 (+1.909)	37.829 (+1.664)	91.750 (+0.427)	72.274 (+4.203)
0.85	**86.246 (+1.890)**	**38.363 (+2.199)**	**91.738 (+0.414)**	**72.584 (+4.514)**
0.9	86.227 (+1.871)	37.840 (+1.676)	91.740 (+0.417)	72.473 (+4.403)

E METRIC DEFINITIONS AND INFERENCE COST

Lexical evaluation. We retain the metric labels and column order of the supplied URF-HVAA evaluation. In that implementation, the reported BLEU value is

$$\mathrm{BLEU}_{\mathrm{reported}} = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{4}\mathrm{SentenceBLEU}_k(\hat{y}_j, y_j). \tag{7}$$

Here n is the number of evaluated videos, and $\hat{y}_j$ and y_j denote the generated and reference descriptions for video j, respectively. The reported value sums sentence BLEU-1 through BLEU-4, rather than reporting BLEU-4 alone. NLTK smoothing method 4 and whitespace tokenization are used, with BLEU-3 weights $(0.33, 0.33, 0.33, 0)$. METEOR uses NLTK `single_meteor_score`; ROUGE is ROUGE-L F-measure with stemming; CIDEr uses `pycocoevalcap`. Each dataset table uses its own CIDEr document-frequency corpus. Published reference values and re-evaluated released descriptions remain in separate blocks because they need not share identical evaluator outputs.

GPT evaluation. GPT evaluation uses `gpt-4.1-2025-04-14` at temperature zero. Reasonability, detail, and consistency are assessed separately from the generated and reference descriptions, without video input, and averaged over each dataset.

Detection evaluation. After video and local updates, each score-anchor sequence is Gaussian-smoothed with $\sigma = 10$, repeated at stride 16, truncated or zero-padded to the evaluation length, and concatenated with the corresponding frame labels. AUC and AP are computed from the pooled frame sequence with `roc_auc_score` and `average_precision_score`. AP is not a trapezoidal precision–recall integral. Neither per-video min–max normalization nor averaging of video-wise AUC is applied.

VAU inference cost. The analyst makes 18.22 image judgments per video on average, and the final answerer receives 16 frames. Recorded selection time averages 9.34 seconds per video. Final generation takes 4.51 seconds within a 6.23-second final-stage wall time. These cache-based RTX 3090-class measurements exclude upstream caption, tag, and spatial-annotation generation, model loading, and feedback; they do not represent end-to-end latency.

Incremental VAD feedback cost. With scores, local captions, and the required descriptions or tags already available, our feedback pipeline is $6.22\times$ faster than full caption rescoring. On the full UCF-Crime test set, Ours takes 743.12 seconds (0.000668 seconds per frame), compared with 4622.26 seconds for full caption rescoring, using two RTX 4090 GPUs. Ours starts from URF-refined scores, while the control starts from RAW. Timing includes preparation, decoding, fresh inference, score updates, and I/O, but excludes upstream generation, model loading, warm-up, final smoothing and expansion, and evaluation. These single-run measurements use an earlier feedback calibration and demonstrate an incremental processing advantage with precomputed inputs, rather than an end-to-end VAD speedup.