SEEING TWICE: HIERARCHICAL RE-OBSERVATION FOR TRAINING-FREE VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free video anomaly detection relies on large pretrained models to detect abnormal events without task-specific training. However, most existing methods inspect model-generated descriptions only once, making them vulnerable to implicit and ambiguous anomaly cues. Specifically, context-dependent anomaly cues may remain hidden at a single temporal granularity and become discernible only through multi-granularity observation, while semantically ambiguous cues may persist when the initial visual description is incomplete or unreliable. In this paper, we propose ***Seeing Twice***, a hierarchical re-observation framework that uncovers implicit cues at the observation level and resolves semantic ambiguity at the reasoning level. Our method constructs a hierarchical event representation with intervals at multiple temporal granularities. At the observation level, to recover implicit cues, the framework selectively revisits intervals insufficiently observed at a single granularity, recovering context-dependent evidence by jointly examining them with their hierarchical companions. At the reasoning level, to resolve semantic ambiguity, the framework selects between the initial and recovered descriptions according to their agreement with the visual evidence. It then reconciles the corresponding anomaly scores through hierarchical score propagation to produce coherent frame-level predictions. *Seeing Twice* achieves 87.12% ROC-AUC on UCF-Crime and 75.93% AP on XD-Violence, surpassing existing methods by a clear margin. These results demonstrate the effectiveness of the proposed framework for training-free video anomaly detection. Moreover, our method provides interpretable, human-readable cues for anomalous event analysis.

1 INTRODUCTION

Video anomaly detection (VAD) aims to detect abnormal events in long, untrimmed videos. Weakly supervised methods reduce annotation costs by learning from video-level labels rather than frame-level annotations (Sultani et al., 2018, Tian et al., 2021, Wu et al., 2020). However, they still depend on labeled training videos and task-specific optimization. Training-free VAD instead exploits pretrained models without optimizing them for the target VAD task. These methods follow a common pipeline: a visual model describes video intervals, and a language model converts these descriptions into anomaly scores (Zanella et al., 2024, Li et al., 2025c). The descriptions thus link visual observation to the language model's assessment of abnormality. They also provide human-readable evidence that can be inspected alongside the predictions. Detection quality therefore depends on the visual evidence captured in these descriptions and how it is interpreted.

Recent training-free approaches improve this pipeline by structuring videos into events, asking context-aware visual questions, and refining predictions with temporal context (Shao et al., 2025, Li et al., 2025c, Ahn et al., 2026, Zanella et al., 2024). Yet a single-pass description can still omit a decisive action or support multiple plausible interpretations. Figure 1 illustrates both failure modes with real video examples. In (a), the initial description mentions handling an item but omits the action of moving it from a display into a bag. Without this action detail, the anomaly cue remains implicit in the description. In (b), the initial description suggests that someone is being restrained, whereas the visual evidence better supports the interpretation of a person waiting for an elevator. Together, these examples highlight two related challenges: recovering implicit anomaly cues and resolving semantic ambiguity in video descriptions.

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.

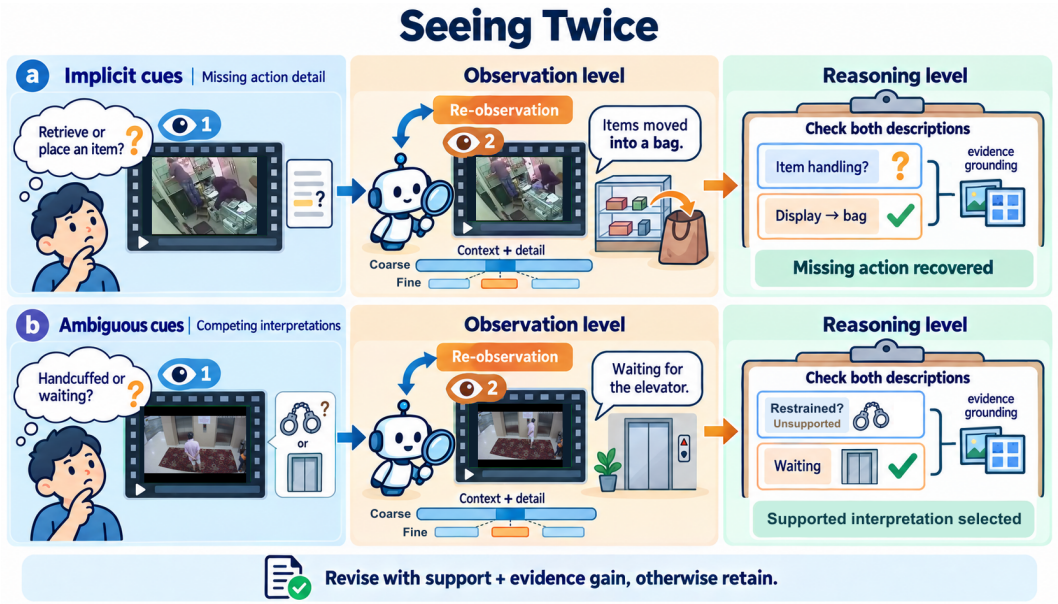

Figure 1: **Why see twice?** (a) Re-observation recovers an omitted action: moving items from a display into a bag. (b) Visual evidence distinguishes waiting for an elevator from an unsupported restraint interpretation. The two levels connect evidence recovery with description selection. Descriptions are condensed for illustration.

Implicit anomaly cues. An anomaly may depend jointly on a brief action and the context in which it occurs. A coarse interval provides broader event context but may obscure a brief, decisive action. A fine interval reveals local detail but may lack the context needed to interpret it. Consequently, an anomaly cue can remain implicit at either granularity even when the relevant frames are present. The challenge is therefore to connect local action evidence with the broader event in which it occurs.

Semantically ambiguous cues. Even when an action is described, its abnormality may remain uncertain. Video multimodal models can misinterpret actions and temporal relationships (Li et al., 2025a), so a more detailed description is not necessarily a more reliable one. An omitted action may conceal an anomaly, whereas an unsupported claim may make a normal event appear abnormal. The challenge is therefore not merely to generate another description, but to identify the interpretation best supported by visual evidence from the interval and its surrounding event.

In this paper, we propose SEEING TWICE, a hierarchical re-observation framework that addresses these two challenges at complementary levels. At the observation level, the framework identifies gaps in the initial descriptions and selectively revisits the corresponding intervals. It also re-observes their parent or child companions to obtain complementary temporal views. Coarse views provide broader event context, while fine views recover local action detail. At the reasoning level, the framework compares initial and recovered descriptions using visual evidence agreement and local claim support, admitting a revision only when it is better grounded in the video. The resulting anomaly scores are then reconciled across the event hierarchy and temporally smoothed to produce coherent frame-level predictions.

Our contributions are threefold:

- We identify implicit cues and semantic ambiguity as complementary challenges in description-based training-free VAD.
- We introduce SEEING TWICE, which combines hierarchical re-observation with evidence-grounded disambiguation to recover missing cues and resolve competing interpretations without task-specific training.
- Seeing Twice achieves 87.12% ROC-AUC on UCF-Crime and 75.93% AP on XD-Violence, demonstrating the effectiveness of the proposed two-level re-observation framework.

2 Related Work

Learning-based video anomaly detection. Deep VAD spans normality modeling, weak supervision, and semantic reasoning (Zhu et al., 2024). Early approaches reconstruct normal appearance and motion or predict future frames (Hasan et al., 2016, Liu et al., 2018), while Tang et al. (2020) combine prediction and reconstruction. Memory-based models constrain reconstruction with normal patterns (Gong et al., 2019, Park et al., 2020). Object-centric methods learn appearance and motion representations through auto-encoding or self-supervised proxy tasks (Ionescu et al., 2019, Georgescu et al., 2021), whereas cooperative learning exploits anomaly rarity in unlabeled videos (Zaheer et al., 2022). Weakly supervised methods learn from video-level labels (Sultani et al., 2018), improving feature discrimination (Tian et al., 2021), instance selection (Lv et al., 2023), and pseudo-label completeness and uncertainty (Zhang et al., 2023). Vision–language adaptation and learned prompts introduce semantic knowledge (Wu et al., 2024b, Chen et al., 2024), with open-vocabulary methods extending detection to unseen anomaly categories (Wu et al., 2024a, Li et al., 2025b). Even with frozen backbones, prompt or head learning retains task-specific optimization (Ye et al., 2025, Cai et al., 2025, 2026), motivating inference without a VAD training stage.

Training-free semantic reasoning. Human-readable anomaly evidence predates large language models, with visual concepts supporting joint detection and recounting of abnormal events (Hinami et al., 2017). Training-free VAD now uses frozen models for description-based inference. LAVAD combines video captions, language-model reasoning, and score refinement (Zanella et al., 2024). Detection based on semantic understanding and multimodal caption integration further develop this route (Gao et al., 2025, Dev et al., 2025). AnyAnomaly uses context-aware visual questioning to detect text-specified events, while DR-VAD makes anomaly criteria explicit through definition-guided reasoning (Ahn et al., 2026, Pei et al., 2026). These decisions depend on which visual details enter the descriptions and how faithfully they represent the event, linking reliable anomaly reasoning to complete and accurate visual observations.

Temporal context and evidence reliability. Caption cleaning improves the alignment between descriptions and visual content (Zanella et al., 2024). Event-aware segmentation and coarse–fine temporal organization preserve action context at different durations (Shao et al., 2025, Li et al., 2025c). PANDA adds tool-assisted reflection and memory, while ReCI retrieves contextual references to support anomaly judgments (Yang et al., 2025, Mo et al., 2026). These advances motivate treating evidence acquisition and interpretation together. The proposed Seeing Twice method addresses missing or ambiguous evidence across temporal granularities through selective hierarchical re-observation and evidence-grounded disambiguation. Recovered descriptions influence predictions only when their visual support justifies a revision.

3 Method

3.1 Problem Formulation and Overview

Given an untrimmed video $V = \{v_t\}_{t=1}^{T}$, we predict frame-level anomaly scores $\mathbf{y} = \{y_t\}_{t=1}^{T}$, where $y_t \in [0, 1]$ increases with abnormality. Description-based training-free VAD first describes video content with a frozen visual model, then scores its abnormality with a frozen language model. An incomplete description can therefore hide an anomaly or leave its meaning ambiguous.

Seeing Twice makes two decisions (Figure 2): which intervals need a second observation, and whether the recovered description should replace the initial one. The *observation level* selects intervals with evidence gaps and obtains complementary coarse or fine views. The *reasoning level* checks the recovered descriptions against visual evidence, scores admitted revisions, and reconciles them with the surrounding event to produce frame-level predictions. All pretrained models remain frozen, with no task-specific fitting of weights, prompts, or prediction heads.

3.2 Initial Observation and Event Representation

We use a frozen event-boundary model (Zheng et al., 2024) to detect content changes, then recursively divide the video into coarse and fine intervals. These intervals form the hierarchical event

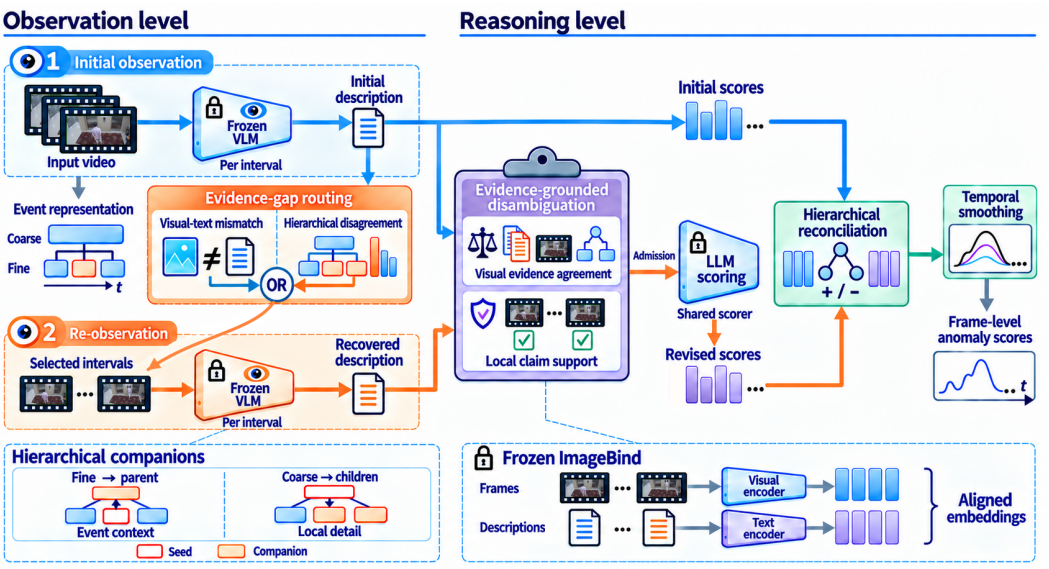

Figure 2: **Overview of SEEING TWICE.** Left: evidence-gap routing selects intervals and hierarchical companions for re-observation. Right: evidence-grounded disambiguation admits supported descriptions before score reconciliation and temporal smoothing. Red outlines indicate seed intervals, shaded intervals denote hierarchical companions, and lock icons represent frozen models.

representation $\mathcal{G} = (\mathcal{C} \cup \mathcal{F}, \mathcal{E})$, with coarse intervals $\mathcal{C}$, fine intervals $\mathcal{F}$, and temporal relations $\mathcal{E}$. Interval i spans $I_i = [b_i, e_i)$. A fine interval's parent is the coarse interval with greatest temporal overlap:

$$p(i) = \arg\max_{c \in \mathcal{C}} |I_i \cap I_c|. \tag{1}$$

The fine intervals assigned to a coarse interval are its children. Intervals sharing a parent are siblings, denoted by $\mathcal{S}(i) = \{j \in \mathcal{F} : p(j) = p(i),\ j \neq i\}$. Ties follow the stored order of coarse intervals.

For the initial observation, a frozen video-language model describes each interval independently as $x_i^{(1)}$. A frozen language model assigns the anomaly score $s_i^{(1)} = f_{\text{score}}(x_i^{(1)}) \in [0, 1]$. We also encode sampled frames with ImageBind (Girdhar et al., 2023), and their normalized mean gives the unit visual representation $\mathbf{v}_i$. The corresponding text encoder maps a description x to a unit vector $\mathbf{t}(x)$ in the same space.

3.3 OBSERVATION LEVEL: SELECTIVE HIERARCHICAL RE-OBSERVATION

Evidence-gap routing. We select intervals for re-observation using two signs of incomplete evidence. First, visual–text mismatch measures how poorly the initial description agrees with the observed frames. We define visual–text agreement as $c_i(x) = (1 + \mathbf{v}_i^\top \mathbf{t}(x))/2$ and mismatch as $d_i = 1 - c_i(x_i^{(1)})$. Second, hierarchical disagreement measures score differences between visually related intervals that may reflect missing context or detail. Let $\mathcal{R}(i)$ contain the available parent, children, and siblings of interval i. The hierarchical disagreement is

$$h_i = \max_{j \in \mathcal{R}(i)} [\mathbf{v}_i^\top \mathbf{v}_j]_+ |s_i^{(1)} - s_j^{(1)}|, \tag{2}$$

where $[z]_+ = \max(z, 0)$ and an empty relation set gives $h_i = 0$. Similarity weighting suppresses disagreement between unrelated content. The routing priority $u_i = \max(d_i, h_i)$ implements the OR decision in Figure 2: either signal can identify an interval for re-observation. Among intervals with $u_i > \tau_r$, we select at most B with the highest priorities as the seed set $\mathcal{B}$. No re-observation occurs when no interval exceeds the threshold.

Complementary hierarchical views. A selected interval, or *seed*, is paired with hierarchical companions that supply a complementary temporal view. A fine seed recruits its parent for broader

event context. A coarse seed recruits up to K children to expose details that its longer interval may obscure. Children are ranked by $q(j \mid i) = u_j + \frac{1}{2}(1 - \mathbf{v}_i^\top \mathbf{v}_j)$, which combines their evidence gaps and visual difference from the parent. The complete set of intervals for re-observation is

$$\mathcal{H} = \mathcal{B} \cup \bigcup_{i \in \mathcal{B}} \mathrm{Comp}_K(i, q). \tag{3}$$

Here Comp_K implements the parent/child rule, with one expansion step and duplicate intervals removed. The frozen *re-observer* produces a recovered description for each selected interval:

$$\bar{x}_i = f_{\mathrm{obs}}\big(V[I_i], \pi_{\mathrm{describe}}\big), \qquad i \in \mathcal{H}. \tag{4}$$

The prompt requests visible actors, actions, objects, and temporal changes without enumerating anomaly categories.

3.4 Reasoning Level: Evidence-grounded Disambiguation

The initial and recovered descriptions may give different interpretations of the same event, leaving its abnormality semantically ambiguous. We resolve this ambiguity through two complementary checks in Figure 2: visual evidence agreement compares the descriptions against the target interval and its event context, while local claim support checks their specific actions and objects. Both checks use the shared visual–text space of frozen ImageBind, and only admitted descriptions are passed to the language model for rescoring.

Visual evidence agreement. Let $\mathbf{v}_i^H$ be the normalized aggregation of parent/child features, weighted by nonnegative similarity to $\mathbf{v}_i$, with $\mathbf{v}_i$ as fallback. A description's evidence agreement combines target and hierarchical support:

$$E_i(x) = \tfrac{1}{2} c_i(x) + \tfrac{1}{4}\big(1 + (\mathbf{v}_i^H)^\top \mathbf{t}(x)\big). \tag{5}$$

Scene-level agreement can nevertheless conceal an unsupported action, such as placing an object into a bag. We therefore extract visible actor, action, object, or state claims $\mathcal{L}(x)$ with the frozen language model. We divide the sampled frames into consecutive local windows $\mathcal{W}_i$, each represented by a unit visual feature $\mathbf{v}_{iw}$. Local claim support is

$$Q_i(x) = \min_{\ell \in \mathcal{L}(x)} \max_{w \in \mathcal{W}_i} \tfrac{1}{2}\big(1 + \mathbf{v}_{iw}^\top \mathbf{t}(\ell)\big). \tag{6}$$

Each claim must find support in at least one window, allowing brief actions without requiring them to persist throughout the interval. Malformed or empty claim sets receive $Q_i(x) = -\infty$ and cannot support a revision.

Admission and score revision. Let $\mathcal{X}_i$ contain the initial and recovered descriptions. All candidates are evaluated on the same visual evidence. We select the description with greatest evidence agreement among those meeting the local-support threshold:

$$x_i^\star = \arg \max_{\substack{x \in \mathcal{X}_i \\ Q_i(x) \geq \tau_q}} E_i(x). \tag{7}$$

The *admission indicator* accepts a changed description only if it improves evidence agreement over the initial description by more than δ:

$$a_i = \mathbf{1}[i \in \mathcal{H}]\, \mathbf{1}[x_i^\star \neq x_i^{(1)}]\, \mathbf{1}[E_i(x_i^\star) - E_i(x_i^{(1)}) > \delta]. \tag{8}$$

Ties favor the initial description. For unqueried intervals or empty eligible sets, we set $x_i^\star = x_i^{(1)}$ and $a_i = 0$. The retained description is $x_i^{(2)} = x_i^\star$ if admitted, and $x_i^{(1)}$ otherwise. Only admitted revisions are rescored:

$$s_i^{(2)} = s_i^{(1)} + a_i\big(f_{\mathrm{score}}(x_i^{(2)}) - s_i^{(1)}\big). \tag{9}$$

The resulting *initial branch* $\{s_i^{(1)}\}$ and *recovered branch* $\{s_i^{(2)}\}$ differ only at admitted revisions. Selection depends on visual support, so the recovered score may increase or decrease.

3.5 Hierarchical Reconciliation and Frame-level Prediction

The reasoning level also relates local scores to the surrounding event through hierarchical score propagation. Parent and similarity-weighted sibling messages contextualize the initial branch, while admitted scores are preserved in the recovered branch during this step. Both branches then undergo visual-neighborhood aggregation and parent adaptation, producing contextualized scores $\bar{\mathbf{s}}^{(b)}$ for $b \in \{1, 2\}$. Let P_f and P_c project fine and coarse interval scores to frames. Each branch's frame-level prediction uses the same coarse weight ω:

$$y_{t,b} = (1 - \omega) P_f(\bar{\mathbf{s}}^{(b)})_t + \omega P_c(\bar{\mathbf{s}}^{(b)})_t. \tag{10}$$

We combine the two predictions according to where description revisions are admitted. Let $A_f(t)$ and $A_c(t)$ denote the fractions of fine and coarse intervals covering frame t with admitted revisions. Their weighted coverage, $\rho_t = (1 - \omega) A_f(t) + \omega A_c(t)$, controls the correction:

$$y_t^{\text{hier}} = y_{t,1} + \rho_t (y_{t,2} - y_{t,1}). \tag{11}$$

Frames without an admitted revision retain the initial branch prediction. Elsewhere, the signed correction allows recovered evidence to raise or lower the score. Finally, shared multi-scale temporal smoothing reduces discontinuities at interval boundaries, yielding the final frame-level prediction $\mathbf{y}$.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate on the official test splits of UCF-Crime (290 videos) and XD-Violence (800 videos) (Sultani et al., 2018, Wu et al., 2020). UCF-Crime covers 13 anomaly categories in surveillance footage, while XD-Violence contains six categories from movies and online sources. We use visual input only and report frame-level ROC-AUC on UCF-Crime and AP on XD-Violence, both as percentages.

Implementation details. LLaVA-Video-7B generates initial descriptions, and CHAI-SFT-8B performs re-observation (Zhang et al., 2025, Lin et al., 2026). DeepSeek-R1-Distill-Qwen-14B assigns anomaly scores, and ImageBind provides the shared visual–text features (Guo et al., 2025, Girdhar et al., 2023). Each interval supplies at most 64 frames. All models remain frozen, with identical inference settings on both datasets.

4.2 Comparison with State-of-the-Art Methods

Table 1 shows that Seeing Twice achieves 87.12% ROC-AUC on UCF-Crime and 75.93% AP on XD-Violence, outperforming all listed training-free methods. The gains over the strongest listed training-free comparators are 2.23 points over PANDA (Yang et al., 2025) on UCF-Crime and 3.70 over ReCI (Mo et al., 2026) on XD-Violence. Compared with LAVAD (Zanella et al., 2024), the gains reach 6.84 ROC-AUC and 13.92 AP points, respectively. These gains span surveillance and movie or online-video scenes under shared inference settings.

The table separates training-free inference from methods with task-specific learning. On UCF-Crime, Seeing Twice also exceeds weakly supervised RTFM (Tian et al., 2021), UR-DMU (Zhou et al., 2023), and MGFN (Chen et al., 2023) without task-specific training. It also exceeds prompt-learning VERA (Ye et al., 2025) on both benchmarks, while requiring no learned prompts or prediction heads.

4.3 Ablation Studies

We assess the contributions of evidence-gap routing, hierarchical companions, and evidence-grounded disambiguation, then test re-observation gains across different frozen observers.

Evidence-gap routing. Table 2A holds seed counts, coarse/fine allocations, companion rules, and downstream processing fixed. Combining visual–text mismatch and hierarchical disagreement improves on random selection by 2.45 UCF ROC-AUC and 5.35 XD AP points. The two components also outperform either alone, supporting complementary interval selection. They jointly address mismatches within an interval and inconsistencies across temporal views.

Table 1: Comparison on UCF-Crime and XD-Violence. Bold marks Seeing Twice.

Mode	Methods	Params	UCF	XD
			ROC-AUC (%)	AP (%)
Weakly Supervised	MIL (Sultani et al., 2018)	–	75.41	–
	RTFM (Tian et al., 2021)	–	84.30	77.81
	CoMo (Cho et al., 2023)	–	86.07	81.31
	UMIL (Lv et al., 2023)	–	86.75	–
	UR-DMU (Zhou et al., 2023)	–	86.97	81.66
	MGFN (Chen et al., 2023)	–	86.98	80.11
	VadCLIP (Wu et al., 2024b)	–	88.02	84.51
Frozen Models with Task-Specific Learning	*Prompt learning*			
	VERA (Ye et al., 2025)	8B	86.55	70.54
	Head learning			
	HiProbe-VAD (Cai et al., 2025)	8B	86.72	82.15
	SteerVAD (Cai et al., 2026)	8B	87.15	83.02
Training-Free	*Direct frozen-model baselines*			
	ZS CLIP (Radford et al., 2021)	0.15B	53.16	17.83
	LLaVA-1.5 (Liu et al., 2024)	13B	72.84	50.26
	Training-free VAD			
	LAVAD (Zanella et al., 2024)	6.7B	80.28	62.01
	SUVAD (Gao et al., 2025)	9B	83.90	70.10
	MCANet (Dev et al., 2025)	–	82.47	69.72*
	EventVAD (Shao et al., 2025)	7B	82.03	64.04
	VADTree (Li et al., 2025c)	7B	84.74	67.82
	PANDA (Yang et al., 2025)	7B	84.89	70.16
	AnyAnomaly (Ahn et al., 2026)	8B	80.70	–
	ReCI (Mo et al., 2026)	8B	81.51	72.23
	Seeing Twice (Ours)	8B	**87.12**	**75.93**

∗: additional audio on XD. Params denotes the size of the largest individual VLM, excluding separate scoring LLMs and feature encoders. –: not reported.

Table 2: Component ablations. UCF: ROC-AUC (%), XD: AP (%). Full uses both companion types.

Configuration	UCF	XD
Full Seeing Twice	**87.12**	**75.93**
A. Evidence-gap routing		
Random selection	84.67	70.58
Visual–text mismatch only	86.41	72.12
Hierarchical disagreement only	85.89	73.96
B. Hierarchical companions		
Seeds only, no companions	86.23	74.67
Seeds + parent companions	86.55	75.24
Seeds + child companions	86.78	75.01
C. Evidence-grounded disambiguation		
Always accept recovered	85.07	72.04
Visual evidence agreement only	86.94	73.92
Local claim support only	86.12	74.35

Table 3: Re-observation across model configurations. Metrics as in Table 2.

Observer(s)	UCF	XD
Single observation (one model)		
LLaVA	84.86	68.02
CHAI	84.65	69.20
Seeing Twice, same model		
LLaVA → LLaVA	86.87	75.11
CHAI → CHAI	87.09	75.86
Seeing Twice, mixed models		
CHAI → LLaVA	86.92	75.32
LLaVA → CHAI	**87.12**	**75.93**

Arrows denote initial observer → re-observer. Compare each pair with its initial model's single observation. LLaVA: LLaVA-Video-7B. CHAI: CHAI-SFT-8B. Bold: default.

Hierarchical companions. Table 2B compares companion types with fixed routing seeds, observers, and downstream processing. Both parent-only and child-only expansion improve on seeds alone, with larger parent gains on XD and larger child gains on UCF. Parent companions place a local action within its surrounding event, while child companions expose brief details that a coarse description can overlook. Combining both types performs best, exceeding seeds alone by 0.89 ROC-AUC and 1.26 AP points. These gains support complementary temporal views for recovering implicit cues, rather than relying on re-observation at the seed's granularity alone.

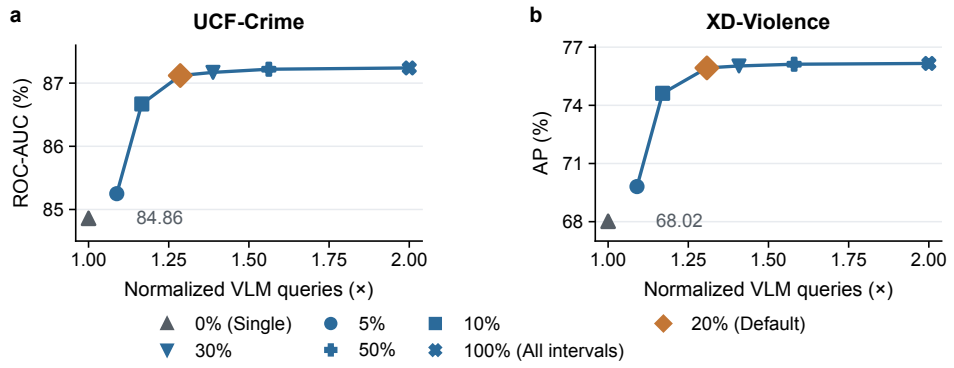

Figure 3: Detection performance versus VLM query count on (a) UCF-Crime and (b) XD-Violence. The horizontal axis counts initial and re-observation queries relative to single observation (1×), with all-interval re-observation at 2×. Legend percentages denote seed-routing budgets. Orange diamonds mark the default 20% budget.

Table 4: Inference costs on UCF-Crime, including all stages. Mean per-video cost is Total/290. Random selection matches the query count at the default 20% budget.

Method	Budget (%)	VLM queries	GPU cost (h)		ROC-AUC (%)
			Total	Per video	
Single observation	0	8,613	62.3	0.215	84.86
Random second look	20	11,076	79.8	0.275	85.04
Seeing Twice	20	11,076	79.1	0.273	**87.12**
All-interval re-observation	100	17,226	131.5	0.453	87.24

Evidence-grounded disambiguation. Table 2C compares selection rules for the same initial and recovered descriptions. Combining visual evidence agreement with local claim support exceeds unconditional acceptance by 2.05 UCF ROC-AUC and 3.89 XD AP points. Since the candidates are unchanged, these gains reflect evidence selection rather than additional observation. The combined rule also outperforms either criterion alone, showing the value of checking overall visual agreement and support for specific actions and objects. Scene-level agreement alone can miss an unsupported action claim, while local claim support alone does not establish which description better matches the event context.

Re-observation across model configurations. Seeing Twice improves on single observation across all tested same-model and mixed-model configurations (Table 3), using the same initial observer and a fixed scorer. With LLaVA for both observations, it gains 2.01 UCF ROC-AUC and 7.09 XD AP points, compared with 2.44 and 6.66 points for CHAI. Both mixed-model configurations also improve on their respective single-observation baselines. Thus, Seeing Twice benefits both model choices without requiring different models for the two observations.

4.4 Computational Efficiency

Relative to single observation, Seeing Twice delivers substantial detection gains with limited additional GPU time per video. On UCF, mean per-video cost rises from 0.215 to 0.273 GPU hours (Table 4), adding 3.48 GPU minutes for a 2.26-point ROC-AUC gain. On XD, AP rises from 68.02% to 75.93%, a 7.91-point gain, with 3.45 additional GPU minutes per video on average. At the same query count, Seeing Twice also exceeds random second look by 2.08 UCF ROC-AUC points without increasing GPU cost. This comparison shows that selecting informative intervals matters beyond simply allocating more visual queries. Figure 3 further shows that the default 20% budget approaches all-interval performance on both datasets. All-interval re-observation doubles VLM queries but adds only 0.12 UCF ROC-AUC and 0.23 XD AP points over the default, indicating diminishing returns. Selective re-observation therefore captures most of the detection benefit with a few additional GPU minutes per video.

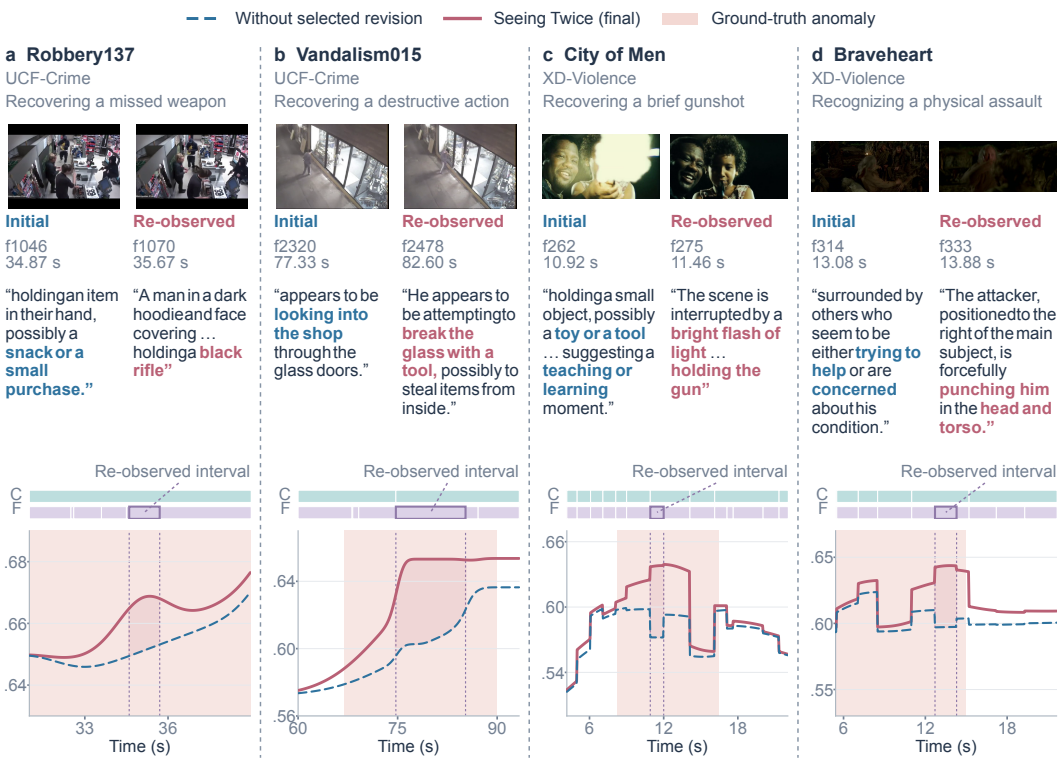

Figure 4: **Qualitative results on UCF-Crime (a–b) and XD-Violence (c–d).** Each column pairs initial and recovered descriptions with sampled frames and frame-level anomaly scores. C/F ribbons denote coarse/fine intervals, and purple guides mark re-observed intervals. Blue dashed curves show scores without the selected revision, while rose solid curves show the final anomaly scores. Shading indicates ground-truth anomalies.

4.5 Qualitative Analysis

Figure 4 shows how Seeing Twice addresses implicit anomaly cues and semantic ambiguity when incomplete descriptions make anomalous events appear ordinary. Re-observation identifies a rifle rather than an everyday item in *Robbery137* (a), and glass-breaking with a tool rather than looking inside in *Vandalism015* (b). In *City of Men* (c), a gun and a brief flash replace the interpretation of a toy or learning activity. In *Braveheart* (d), punches to the head and torso replace the interpretation of assistance. Across these cases, recovering missing objects and actions provides evidence for distinguishing anomalies from plausible normal interpretations. Visual evidence agreement and local claim support then determine whether the recovered description is admitted for rescoring. Relative to predictions without the selected revision, the final curves show sustained score increases in (a–b) and localized increases in (c–d) around anomalous events. Together, these cases show how recovering implicit cues and resolving semantic ambiguity improve frame-level predictions, while the descriptions explain the evidence behind score revisions.

5 Conclusion

We present Seeing Twice, a hierarchical re-observation framework for training-free video anomaly detection that addresses implicit anomaly cues and semantic ambiguity. At the observation level, selective re-observation revisits intervals and their hierarchical companions to recover missing event context and local action details across temporal granularities. At the reasoning level, visual evidence agreement and local claim support guide the selection of initial or recovered descriptions. Hierarchical score propagation then relates their anomaly scores to the surrounding event, producing coherent frame-level predictions. Experimental results on UCF-Crime and XD-Violence demonstrate that Seeing Twice achieves substantial performance gains over existing training-free methods.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and polishing, including improving grammar, clarity, wording, and manuscript organization. The authors reviewed and revised all AI-assisted text and take full responsibility for the content, claims, results, and conclusions presented in the paper. Generative AI was not used to generate experimental results or make scientific conclusions.

REFERENCES

Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. AnyAnomaly: Zero-shot customizable video anomaly detection with LVLM. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3026–3035, 2026.

Zhaolin Cai, Fan Li, Ziwei Zheng, and Yanjun Qin. HiProbe-VAD: Video anomaly detection via hidden states probing in tuning-free multimodal LLMs. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 592–601, 2025.

Zhaolin Cai, Fan Li, Huiyu Duan, Lijun He, and Guangtao Zhai. Steering and rectifying latent representation manifolds in frozen multi-modal LLMs for video anomaly detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Junxi Chen, Liang Li, Li Su, Zheng jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18319–18329, 2024.

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 (AAAI)*, pp. 387–395, 2023.

MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context–motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12137–12146, 2023.

Prabhu Prasad Dev, Raju Hazari, and Pranesh Das. MCANet: Multimodal caption aware training-free video anomaly detection via large language model. In *Pattern Recognition: International Conference on Pattern Recognition (ICPR 2024)*, volume 15332 of *Lecture Notes in Computer Science*, pp. 362–379. Springer, 2025. Published online 5 December 2024.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. MIST: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14009–14018, 2021.

Shibo Gao, Peipei Yang, and Linlin Huang. SUVAD: Semantic understanding based video anomaly detection using MLLM. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

Mariana-Iuliana Georgescu, Antonio Bărbălău, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12742–12752, 2021.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1705–1714, 2019.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 733–742, 2016.

Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3619–3627, 2017.

Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7842–7851, 2019.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13723–13733, 2025a.

Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29203–29212, 2025b.

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 (AAAI)*, pp. 1395–1403, 2022.

Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. VADTree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 148372–148404, 2025c.

Zhiqiu Lin, Siyuan Cen, Chancharik Mitra, Isaac Li, Yuhan Huang, Yu Tong Tiffany Ling, Hewei Wang, Irene Pi, Shihang Zhu, Yili Han, Yilun Du, and Deva Ramanan. Building a precise video language with human–ai oversight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11334–11345, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection – a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6536–6545, 2018.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8022–8031, 2023.

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 ACM International Conference on Multimedia Retrieval (ICMR)*, pp. 930–939, 2026.

Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14372–14381, 2020.

Yanting Pei, Minhao Hao, Yaping Huang, and Qi Zou. DR-VAD: Definition-guided reasoning for training-free video anomaly detection. *Neurocomputing*, 674:132836, 2026.

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. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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 ACM International Conference on Multimedia (ACM MM)*, 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)*, pp. 6479–6488, 2018.

Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020.

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 (ICCV)*, pp. 4975–4986, 2021.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 322–339, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18297–18307, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6074–6082, 2024b.

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 (NeurIPS)*, volume 38, pp. 92745–92774, 2025.

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

M. Zaigham Zaheer, Arif Mahmood, M. Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14744–14754, 2022.

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 (CVPR)*, pp. 18527–18536, 2024.

Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16271–16280, 2023.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025.

Ziwei Zheng, Zechuan Zhang, Yulin Wang, Shiji Song, Gao Huang, and Le Yang. Rethinking the architecture design for efficient generic event boundary detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 1215–1224, 2024.

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 (AAAI)*, pp. 3769–3777, 2023.

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 (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 89943–89977, 2024.

A PROTOCOL AND REPRODUCIBILITY

This appendix details the observation and reasoning levels in Section 3, followed by the controls and additional results supporting Section 4.

Section A specifies interval construction, model prompts, routing, disambiguation, and frame-level prediction. Section B examines component contributions, observer choices, scoring variability, and computational cost. Section C documents comparison protocols and additional qualitative cases.

Table S1: Reading guide from the main claims to their supplementary support.

Main-paper question	Method detail	Evidence or check
When is a second observation needed?	Mismatch/disagreement gate, seed budget, one-hop companions	Query-count bounds and the budget sweep
Which description is retained?	Visual evidence agreement, local claim support, admission margin	Decision cases and disambiguation controls
How do descriptions affect frames?	Signed score revision, branch-specific propagation, admission coverage	Range and fallback properties
Does the method require a different observer?	Initial-model-matched comparisons	Same-model and mixed-model gains
What does selective inference cost?	Unique visual queries and complete GPU totals	Both-dataset budget and cost accounting

The observation level changes which intervals receive another visual description. The reasoning level changes which description supplies the score and how local scores relate to the surrounding event. These operations remain distinct throughout the appendix. A *seed* is an interval selected by the router, a *companion* supplies a complementary temporal view, and a *queried interval* belongs to their deduplicated union. An *admitted revision* is a recovered description that passes the reasoning-level decision, not merely an interval that receives a second observation.

A.1 EVENT CONSTRUCTION AND PARENT ASSIGNMENT

EfficientGEBD uses the frozen ResNet-50 x2x3x4 checkpoint. Frames are uniformly sampled at at most 10 Hz, resized to 224×224, and processed in 100-sample windows with stride 50. Only the central 50 score predictions are retained. Event construction reconstructs this score grid and keeps local peaks with confidence at least 0.4. It recursively splits the full interval at unused peaks in descending confidence, subject to minimum length 1. Two-cluster K-means on legal peak scores (random seed 0, ten initializations) sets the coarse threshold to the minimum score in the high cluster. During depth-first traversal, an interval is coarse only when the lower of its endpoint confidences meets this threshold. Other intervals are fine. Containment-redundant intervals are removed and fine coverage is completed. If fewer than two distinct peak scores exist, both coarse and fine sets use the leaf intervals.

All stages use the parent map in Equation 1, with stable input order resolving ties. The same hierarchy supplies initial descriptions, hierarchical companions, and the relations used for score propagation.

Temporal identities and shared evidence. Each interval carries a granularity, interval endpoints, and a stable identifier. Coarse and fine intervals can overlap while providing different temporal views. Their scores remain separate until frame projection. Repeated references to the same interval do not trigger repeated description requests.

Frame sampling and feature aggregation serve different purposes. Sampling selects the visual input for a description, whereas normalized feature aggregation provides the common evidence space for routing and disambiguation. The same interval representation participates in both stages. The router does not obtain an extra description merely to estimate its priority. Parent and child features are available from the initial event representation before the queried set is constructed.

A.2 VISUAL OBSERVATION PROMPTS

The two observation stages of SEEING TWICE use distinct prompts with the same objective: describing visible evidence without judging abnormality. Neither prompt enumerates dataset-specific anomaly categories. Both datasets use the same prompt pair.

The initial prompt requests a factual description of the interval. The re-observation prompt emphasizes brief actions, object interactions, and temporal changes that may be missed during a single observation. The re-observer receives the sampled frames but not the initial description, making the second observation a fresh visual-description step rather than a text-only revision.

Initial observation prompt. For every coarse and fine interval, the initial visual observer receives the sampled frames of that interval together with the following prompt:

> You are given a short video interval.
>
> Describe the visible content of the video in detail. Focus on the scene, visible actors, salient objects, actions, interactions, and temporal changes.
>
> Describe only information that is supported by the visual evidence. If a detail is uncertain, express the uncertainty rather than guessing.
>
> Do not infer hidden intentions, causes, or events that are not visually observable. Do not classify the video into any anomaly category and do not assign an anomaly score.
>
> Return a concise factual description of the video interval.

The output is the initial description $x_i^{(1)}$. The language-model interface in Section A.3 then assigns its anomaly score $s_i^{(1)} = f_{\text{score}}(x_i^{(1)})$.

Re-observation prompt. For each interval selected by evidence-gap routing or introduced through hierarchical companion expansion, the re-observer receives the sampled frames of that interval together with the following prompt:

> You are given a short video interval for a second visual observation.
>
> Carefully re-observe the video and provide a precise description of the visible event. Pay particular attention to visual details that may be easy to miss in a single observation.
>
> Focus on:
> (1) visible actors and salient objects,
> (2) concrete actions and interactions,
> (3) how objects are held, moved, exchanged, or used,
> (4) short-duration or subtle actions,
> (5) changes in motion, state, or activity over time, and
> (6) visible consequences of an action.
>
> Describe only information that is supported by the visual evidence. If a detail is uncertain, express the uncertainty rather than guessing.
>
> Do not infer hidden intentions, causes, or events that are not visually observable. Do not classify the video into any anomaly category and do not assign an anomaly score.
>
> Return a concise factual description of the video interval.

The output is the recovered description $\bar{x}_i$. Evidence-grounded disambiguation subsequently compares it with $x_i^{(1)}$ using the shared visual evidence and admission rule in Section A.4.

A.3 LANGUAGE-MODEL PROMPTS

The frozen DeepSeek-R1-Distill-Qwen-14B language model serves two prompt interfaces. Anomaly scoring maps an interval description to an anomaly score. Local claim extraction decomposes a candidate description into visually checkable statements without judging abnormality. ImageBind features and the admission rule, rather than a language-model verdict, determine which description is retained.

Both datasets use the same prompts and decoding settings. Neither interface receives a dataset-specific list of anomaly categories.

Anomaly scoring prompt. The anomaly-scoring interface receives only the visual description of an interval. Initial descriptions are scored once during the first observation, and only recovered descriptions admitted by evidence-grounded disambiguation are subsequently rescored. The same prompt is used in both cases:

> You are given a description of a short video interval.
>
> Assess the degree of abnormality of the event described in the text. Base your judgment only on the information explicitly provided in the description.
>
> Consider the described actions, interactions, object use, temporal behavior, visible consequences, and surrounding event context.
>
> Ordinary and benign everyday activities should receive low anomaly scores. Clearly unusual, dangerous, violent, destructive, illegal, or otherwise strongly abnormal activities should receive high anomaly scores. Use intermediate scores when the described evidence is ambiguous, weak, or only partially indicative of abnormality.
>
> Do not treat uncertainty or incomplete information as evidence of abnormality. Do not introduce people, objects, actions, intentions, causes, or outcomes that are not stated in the description. Do not assume an anomaly category that is unsupported by the described evidence.
>
> Return a single anomaly score in the range $[0, 1]$, where 0 indicates clearly normal activity and 1 indicates clearly abnormal activity.
>
> Return the result in the following format:
>
> `Score: <number between 0 and 1>`

This interface implements f_{score} in Equation 9. Unqueried intervals and rejected revisions retain their initial scores without another scoring call.

Local claim-extraction prompt. Scene-level agreement can conceal an unsupported action or object claim. For each candidate description, the second interface extracts concrete statements for comparison with local visual windows:

> You are given a description of a short video interval.
>
> Extract the concrete visual claims contained in the description. Each claim must correspond to a visually checkable fact stated in the text.
>
> A valid claim should describe one of the following:
> - a visible actor or person,
> - a visible object,
> - a concrete action,
> - an interaction between actors or objects, or
> - a visible state or state change.
>
> Follow these rules:
>
> 1. Extract only information that is explicitly stated in the description.
> 2. Keep each claim short, concrete, and self-contained.
> 3. Separate different actions, objects, interactions, or visible states into different claims when they can be checked independently.
> 4. Preserve uncertainty when the original description is uncertain. Do not convert phrases such as "possibly," "appears to," or "may be" into certain statements.
> 5. Do not infer intentions, motivations, causes, future outcomes, moral judgments, or anomaly labels.
> 6. Do not add facts that are absent from the description.
> 7. Return no more than eight claims.
>
> Return only the extracted claims in the following format:
>
> ```
> 1. <claim>
> 2. <claim>
> ...
> ```

The output is the claim set $\mathcal{L}(x)$ in Equation 6. Each claim is embedded separately and compared with consecutive visual windows. Section A.4 specifies how these comparisons determine local support, handle malformed outputs, and control description admission before rescoring.

Decoding settings. Claim extraction uses greedy decoding, with the prompt limiting the output to eight claims. Anomaly scoring uses temperature 0.6, top-p 0.9, and top-k 50. These settings remain fixed across datasets and between initial scoring and rescoring.

A.4 Routing, Companions, and Disambiguation

Evidence-gap routing. The initial observation provides descriptions, scores, and normalized ImageBind visual/text embeddings. Each interval's visual embedding is the normalized average of its sampled frame embeddings. Section 3.3 defines visual–text mismatch d_i and hierarchical disagreement h_i, combined as $u_i = \max(d_i, h_i)$. With $N = |\mathcal{C}| + |\mathcal{F}|$, the router selects up to $B = \lceil b_{\text{obs}} N \rceil$ intervals above τ_r, with earlier start, shorter duration, and stored ID resolving ties. A fine seed adds its parent. A coarse seed adds up to K children ranked by $q(j \mid i)$. The union in Equation 3 uses the original seed set, removes duplicate intervals, and stops after one expansion.

Thresholding and seed budget. Let $\mathcal{E}_r = \{i : u_i > \tau_r\}$ contain the intervals that pass the evidence-gap threshold. The threshold determines eligibility, whereas the budget limits how many eligible intervals become seeds:

$$|\mathcal{B}| = \min(B, |\mathcal{E}_r|). \tag{12}$$

Intervals exactly at the threshold are excluded. If fewer than B intervals qualify, all eligible intervals become seeds and the remaining budget is unused. The seed budget is therefore an upper bound, not a required fraction of every video.

The maximum in $u_i = \max(d_i, h_i)$ allows either signal to establish eligibility. Visual–text mismatch can identify an interval even when related intervals have similar scores. Hierarchical disagreement can identify inconsistent interpretations despite a locally plausible description. The single-signal controls in Table 2A examine their respective contributions.

Companion expansion and query bound. Let n_f and n_c count fine and coarse seeds, so $|\mathcal{B}| = n_f + n_c$. A fine seed adds at most one parent, and a coarse seed adds at most K children. Before deduplication, the number of proposed intervals is at most $2n_f + (K+1)n_c$. Hence,

$$|\mathcal{B}| \le |\mathcal{H}| \le \min\{N, 2n_f + (K+1)n_c\}. \tag{13}$$

For $K = 2$, the upper bound becomes $\min(N, 3|\mathcal{B}|)$. Missing companions and duplicate requests reduce the count. A 20% seed budget thus limits expansion without requiring the final queried-interval ratio to equal 20%.

Visual evidence construction. Each interval supplies at most 64 uniformly spaced frames. For $\mathbf{v}_i^H$, let $\mathcal{D}(i)$ contain its parent and children, and set $w_{ij} = [\mathbf{v}_i^\top \mathbf{v}_j]_+$. Normalize $\sum_{j \in \mathcal{D}(i)} w_{ij} \mathbf{v}_j$ to unit length. If its norm is below 10^{-8}, use $\mathbf{v}_i$. For local claim support, partition sampled timestamps into up to four nonempty consecutive windows and normalize each window's mean embedding. Long descriptions are split into sentence/clause chunks within the text encoder's token limit. Normalize their mean embedding. All candidates share the same preprocessing and raw unit embeddings rather than scaled retrieval logits. The implementation uses text embeddings and local-window features in addition to cached vision similarities.

Local claim support. The claim-extraction interface in Section A.3 supplies up to eight visible claims per description. Parsing and text preprocessing are identical across candidates. Malformed outputs and empty claim sets receive $Q_i = -\infty$ and are ineligible for admission. Long claims split at clause boundaries to fit the text encoder. Equation 6 evaluates all resulting claim segments and takes the minimum support.

Both descriptions use the same sampled frames, local windows, and parent/child features during comparison. ImageBind features measure support for the extracted claims without generating another visual description.

Admission and fallback. For each re-observed interval, set $\mathcal{X}_i = \{x_i^{(1)}, \bar{x}_i\}$ as in Section 3.4. Among descriptions satisfying $Q_i \geq \tau_q$, select the one with greatest visual evidence agreement E_i. Admit a recovered description only when its agreement exceeds that of the initial description by more than δ. Ties favor the initial description. If no description is eligible, retain $x_i^{(1)}$ and set $a_i = 0$. Unqueried or rejected intervals retain their initial scores without another scoring call. Only admitted descriptions receive new anomaly scores, which can be higher or lower than the initial scores.

Agreement and support have complementary roles. Writing $c_i^H(x) = (1 + (\mathbf{v}_i^H)^\top \mathbf{t}(x))/2$, the agreement score is $E_i(x) = [c_i(x) + c_i^H(x)]/2$. It gives equal weight to the target interval and its hierarchical context. Because the features have unit norm, both terms lie in $[0, 1]$. When no usable companion evidence exists, the fallback $\mathbf{v}_i^H = \mathbf{v}_i$ reduces E_i to target-interval agreement, without changing the scale of the score.

Local claim support addresses a different question. For each claim ℓ, define $r_{i\ell} = \max_w(1 + \mathbf{v}_{iw}^\top \mathbf{t}(\ell))/2$. Then $Q_i(x) = \min_\ell r_{i\ell}$. The inner maximum allows a brief action to receive support from one local window. The outer minimum requires every extracted claim to have such support. Replacing the minimum by a mean would allow several well-supported scene statements to hide one poorly supported decisive action.

An equivalent two-description rule. With two candidates and $\delta \geq 0$, admission can be implemented without a general candidate-search routine. A recovered description replaces the initial one exactly when it differs from the initial text, satisfies $Q_i(\bar{x}_i) \geq \tau_q$, and has $E_i(\bar{x}_i) > E_i(x_i^{(1)}) + \delta$. If these conditions hold, it also wins the eligible-candidate maximization. If any condition fails, no changed description can satisfy Equation 8. This equivalence follows directly from the two-element candidate set and the preference for the initial description in a tie.

Table S2: Complete admission cases for a queried interval. The same rule applies on both datasets.

Recovered description	Agreement relative to initial	Decision
Malformed or empty claims	Any	Retain initial, no rescoring
Local support below τ_q	Any	Retain initial, no rescoring
Eligible, unchanged text	Any	Retain initial, no rescoring
Eligible, changed text	Improvement $\leq \delta$	Retain initial, no rescoring
Eligible, changed text	Improvement $> \delta$	Admit and rescore recovered

The initial description's agreement remains the reference in Equation 8, even when its local support falls below the threshold. Admission compares visual support, not candidate anomaly scores. A supported revision can therefore remove an incorrect suspicious interpretation or recover a missing anomaly cue, allowing the revised score to decrease or increase.

A.5 Shared Configuration

Table S3: Shared inference settings, applied identically to both datasets.

Parameter	Setting
Seed fraction b_{obs} / route threshold τ_r	$0.20/0.50$
Companion budget K	2
Admission margin δ / local support τ_q	$0.02/0.50$
Local windows per interval	up to 4
Tree (α, β) / branch strengths (η_1, η_2)	$(0.05, 0.5)/(1, 0)$
Visual neighbors (k, w, τ)	$(10, 10, 0.1)$
Parent adaptation a / coarse weight ω	$0.2/0.2$
Temporal scales Σ / final smoothing σ_0 (frames)	$\{10, 25, 50, 100, 200\}/25$
Background reference	current video

All models remain frozen, and both datasets use the prompts in Sections A.2 and A.3. No dataset-specific anomaly-category list enters observation, routing, admission, or scoring.

A.6 Hierarchical Reconciliation and Frame-level Prediction

This section specifies the contextualized scores $\bar{\mathbf{s}}^{(b)}$ and frame projections used in Section 3.5. Hierarchical score propagation, visual-neighborhood aggregation, and parent adaptation precede frame projection, admission-weighted branch fusion, and temporal smoothing. Let $[\cdot]_0^1$ denote clipping to $[0, 1]$.

Hierarchical score propagation. The initial and recovered branches use the same event hierarchy, with branch-specific propagation strengths defined below. For granularity $g \in \{c, f\}$, let $\mathcal{V}_g(V)$ contain the corresponding intervals in the current video. The reference mean is computed separately for each branch $b \in \{1, 2\}$,

$$\mu_g^{(b)} = \frac{1}{|\mathcal{V}_g(V)|} \sum_{r \in \mathcal{V}_g(V)} s_r^{(b)}. \tag{14}$$

These branch-specific means provide a video-local background reference within each granularity and remain fixed during propagation. A fine interval receives event context from its parent and support from visually similar siblings,

$$m_{i,b}^{\text{par}} = s_{p(i)}^{(b)} - \mu_c^{(b)}, \tag{15}$$

$$m_{i,b}^{\text{sib}} = \begin{cases} \max\limits_{j \in \mathcal{S}(i)} \big(s_j^{(b)} - \mu_f^{(b)}\big)[\text{sim}(\mathbf{v}_i, \mathbf{v}_j)]_0^1, & \mathcal{S}(i) \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Parent messages supply event context. Sibling messages retain the largest similarity-weighted deviation from the fine-granularity reference mean. A negative message results only when all weighted deviations are negative.

For a fine interval, the two messages update its score as

$$s_{i,b}^{\text{tree}} = (1 - \eta_b) s_i^{(b)} + \eta_b \big[s_i^{(b)} + \alpha m_{i,b}^{\text{par}} + \beta m_{i,b}^{\text{sib}}\big]_0^1. \tag{17}$$

Coarse intervals remain anchors in this tree step. With $(\eta_1, \eta_2) = (1, 0)$, tree messages update the initial branch and preserve recovered scores. Both branches then undergo the same visual-neighborhood aggregation and parent adaptation,

$$\bar{\mathbf{s}}^{(b)} = \mathcal{R}\big(\mathbf{s}^{\text{tree},(b)}, \{\mathbf{v}_i\}, \mathcal{G}\big), \tag{18}$$

where $\mathcal{R}$ comprises the operators specified below.

Only admitted revisions are scored again. The residual in Equation 9 is identical for both datasets:

$$\Delta_i^{\text{adm}} = \begin{cases} f_{\text{score}}(x_i^{(2)}) - s_i^{(1)}, & a_i = 1, \\ 0, & a_i = 0, \end{cases} \qquad s_i^{(2)} = s_i^{(1)} + \Delta_i^{\text{adm}}. \tag{19}$$

Positive and negative residuals are retained. Rejected or unchanged descriptions have zero residual.

For frame t, let $\mathcal{J}_g(t) = \{i \in \mathcal{V}_g(V) : t \in I_i\}$ for $g \in \{f, c\}$. The admission coverage used in Equation 11 is

$$A_g(t) = \frac{\sum_{i \in \mathcal{J}_g(t)} a_i}{\max\{1, |\mathcal{J}_g(t)|\}}, \qquad \rho_t = (1 - \omega) A_f(t) + \omega A_c(t). \tag{20}$$

The mask uses interval membership without sigmoid transformation, so empty or unmodified coverage gives zero weight. The shared coarse–fine weight ω combines both granularities without an additional fusion hyperparameter. Since $\rho_t \in [0, 1]$, fusion is a convex combination of the two contextualized frame scores. It falls back to the initial branch when $\rho_t = 0$ and uses the recovered branch when $\rho_t = 1$. This weight measures admission coverage, not calibrated correctness. It restricts direct cross-branch corrections to frames with admitted local revisions, even when propagation changes scores elsewhere.

Within each video and granularity, neighbors are selected in stored interval order. For zero-based position i in a stream of length n, the candidate positions satisfy

$$\max(0, i - \lfloor w/2 \rfloor) \leq j < \min(n, i + \lfloor w/2 \rfloor).$$

The set $\mathcal{N}_{k,w}(i)$ contains the up to k candidates with highest visual similarity. The target interval is not excluded. With the shared $k = w = 10$, all available window entries are retained, including the target, so every valid interval has a nonempty neighborhood. With $\|$ denoting concatenation, the visual-consistency operator computes

$$\nu_i = \max\left(1 - \mathrm{Std}\left(s_{i,b}^{\mathrm{tree}} \,\|\, [s_{j,b}^{\mathrm{tree}}]_{j\in\mathcal{N}_{k,w}(i)}\right), \epsilon\right), \tag{21}$$

$$\ell_{ij} = \frac{\mathrm{sim}(\mathbf{v}_i, \mathbf{v}_j)}{\tau} - \frac{|s_{i,b}^{\mathrm{tree}} - s_{j,b}^{\mathrm{tree}}|}{\nu_i}, \qquad z_{i,b} = \sum_{j\in\mathcal{N}_{k,w}(i)} \mathrm{softmax}_j(\ell_{ij}) s_{j,b}^{\mathrm{tree}}. \tag{22}$$

Here, Std uses the population standard deviation, and $\epsilon = 2^{-52} \approx 2.22 \times 10^{-16}$ is the float64 numerical floor. The concatenated vector includes the target score explicitly, in addition to its occurrence in the neighborhood. The parent-adaptation weight combines parent agreement, neighborhood consistency, and update magnitude. Let $\tilde{z}_{p(i),b}$ denote the parent's score from the refined coarse stream. Both M_i and MAD_i use $j \in \mathcal{N}_{k,w}(i)$, with $M_i = \mathrm{median}_j\, z_{j,b}$, $\mathrm{MAD}_i = \mathrm{median}_j\, |z_{j,b} - M_i| + 0.01$, and $m_i = |z_{i,b} - s_{i,b}^{\mathrm{tree}}|$:

$$r_i^H = \left[1 - \sqrt{|z_{i,b} - \tilde{z}_{p(i),b}|}\right]_0^1, \qquad \kappa_i = \left[1 - \frac{|z_{i,b} - M_i|}{\mathrm{MAD}_i}\right]_0^1,$$
$$r_i = \left[r_i^H (0.5 + 0.2\kappa_i)\left(0.5 + 0.08\left[1 - \frac{m_i}{0.5}\right]_0^1\right)\right]_0^1. \tag{23}$$

For adaptation strength a, let $\beta_{\min} = 0.5a$, $\beta_{\max} = \min(1, 1.5a)$, and $\beta_i = \beta_{\max} - (\beta_{\max} - \beta_{\min}) r_i$. The stabilized fine-interval score is $z'_{i,b} = (1 - \beta_i) z_{i,b} + \beta_i \tilde{z}_{p(i),b}$. Coarse intervals use their refined coarse scores directly. Table S3 reports the shared neighborhood, temperature, and adaptation settings.

Frame projection and temporal smoothing. Observation, projection, and admission masks use the same half-open intervals. For integer frame indices, $[b_i, e_i)$ writes frames $b_i, \ldots, e_i - 1$. Each projection first initializes a frame sequence to zero, then writes interval scores in stored order, with later writes replacing overlaps. It applies the fixed transform $\mathrm{sigmoid}(z) = 1/(1 + \exp(-z))$ to the resulting sequence, so uncovered frames receive 0.5. This is a score transformation, not probability calibration. Admission masks use interval membership directly, without the sigmoid transform. After branch fusion in Equation 11, multi-scale temporal smoothing produces

$$\mathbf{y} = G_{\sigma_0} * \left[\max_{\sigma\in\Sigma}(G_\sigma * \mathbf{y}^{\mathrm{hier}})\right], \tag{24}$$

where G_σ is a unit-sum Gaussian kernel, $*$ denotes temporal convolution, and the maximum is frame-wise. Temporal kernels are truncated at 3σ, use centered zero padding without edge renormalization, and reduce to identity for sequences shorter than three frames. These choices and the scales in Table S3 are shared across datasets and controls.

A.7 Range, Fallback, and Locality of the Prediction Operators

This section derives the range, fallback, and locality properties of the prediction operators. The analysis establishes how admitted revisions change scores and how contextual processing preserves their bounds.

Description revision is signed and selective. Before contextual processing, $a_i = 0$ gives $s_i^{(2)} = s_i^{(1)}$ exactly. If $a_i = 1$, the recovered score is the frozen scorer's output for the admitted description, rather than an average of the two description scores. Since both scores lie in $[0, 1]$, the residual satisfies

$$-s_i^{(1)} \le \Delta_i^{\mathrm{adm}} \le 1 - s_i^{(1)}. \tag{25}$$

The residual can be positive or negative. Equality between unchanged interval scores applies before contextual processing, where the branches use different tree-propagation strengths as specified in Section A.6.

Contextual scores remain bounded. The clipped term in Equation 17 lies in $[0, 1]$. For $0 \leq \eta_b \leq 1$, its mixture with the original score also lies in this interval. Visual-neighborhood aggregation uses nonnegative softmax weights that sum to one. Parent adaptation is another convex combination because $0 \leq \beta_i \leq 1$. Thus, each stage preserves the interval-score range. Coarse scores are refined before they supply the parent values for fine-interval adaptation, avoiding an implicit circular update.

The reference means and tree messages are computed from branch inputs, not from partially updated scores encountered during traversal. Likewise, neighborhood weights use the completed tree scores for the branch. This ordering makes the mathematical result independent of the order in which interval updates are evaluated. Stored order remains relevant for tie resolution and overlapping frame writes, where the protocol explicitly specifies it.

Admission coverage bounds frame corrections. For nonempty coverage, $A_g(t)$ is an average of binary admission indicators. Empty coverage gives zero. Hence $A_g(t) \in [0, 1]$ and $\rho_t \in [0, 1]$. Equation 11 can be written as

$$y_t^{\text{hier}} = (1 - \rho_t)y_{t,1} + \rho_t y_{t,2}, \qquad |y_t^{\text{hier}} - y_{t,1}| = \rho_t |y_{t,2} - y_{t,1}|. \tag{26}$$

The fused score lies between the branch predictions. The correction follows the sign of their difference and scales linearly with admission coverage for a fixed branch difference. When $\rho_t = 0$, the initial branch is recovered exactly at that frame. When $\rho_t = 1$, the recovered branch supplies the frame score.

Temporal smoothing follows local fusion. The exact fallback above applies before temporal smoothing. Subsequent convolution mixes nearby frame scores, allowing a local revision to influence neighboring predictions. Gaussian kernels have nonnegative, unit-sum weights, and zero padding adds no value outside $[0, 1]$. Each smoothed sequence therefore remains bounded. Taking the frame-wise maximum across scales and applying the final Gaussian also preserves that range.

With the shared scales, the largest first-stage radius is $3 \max \Sigma = 600$ frames and the final radius is $3\sigma_0 = 75$ frames. Temporal smoothing alone can therefore spread an input difference by at most 675 frames in either direction. This is a bound on that operator's support, not on the full method's temporal dependence. Parent context and video-local reference means already connect more distant intervals before smoothing.

Video-local inference. The reference mean in Equation 14, the hierarchy, and the visual neighbors all use the current video. For fixed model outputs and fixed stochastic draws, adding another video to the evaluation collection does not change these inputs. Dataset-level query ratios aggregate completed per-video counts only for reporting. They do not feed back into routing, admission, or scoring. This separates inference within a video from the collection-level computation of evaluation metrics.

B Additional Experiments

All performance metrics are percentages. Entries for the default full configuration use its five-run mean, with sample standard deviations reported separately in Section B.4. Controlled comparisons keep models, prompts, decoding settings, and cached inputs fixed except for the factor under evaluation. Observation and claim extraction use greedy decoding. Anomaly scoring uses temperature 0.6, top-p 0.9, and top-k 50.

B.1 Frozen Observer Protocol

Table 3 reports four initial-observer and re-observer combinations using frozen LLaVA-Video-7B and CHAI-SFT-8B. Every row keeps DeepSeek-R1-Distill-Qwen-14B fixed for anomaly scoring. The single-observation baseline uses the corresponding initial observer and retains hierarchical reconciliation, but performs no re-observation. Each paired comparison evaluates re-observation and evidence-grounded disambiguation with the same initial observer and downstream score processing.

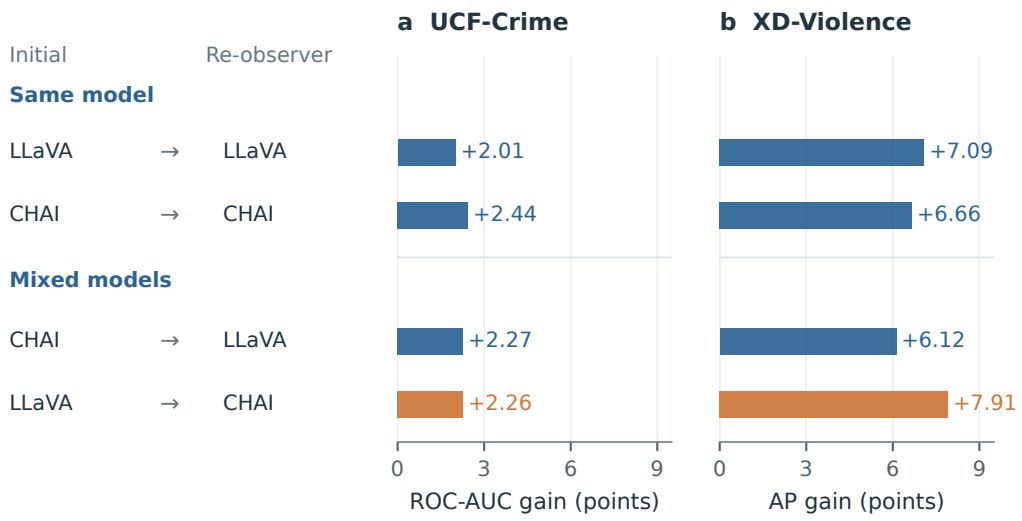

Figure S1: Seeing Twice gains over single observation, computed from Table 3. Each row names the initial observer and re-observer. Bars show percentage-point gains over the corresponding initial model alone, with the default configuration in orange. The LLaVA baselines are 84.86 UCF ROC-AUC and 68.02 XD AP, and the CHAI baselines are 84.65 and 69.20.

Figure S1 measures gains over single observation with the same initial observer. With LLaVA as the initial observer, same-model Seeing Twice improves UCF ROC-AUC and XD AP by 2.01 and 7.09 points. Using CHAI for re-observation increases these gains to 2.26 and 7.91 points. With CHAI as the initial observer, the gains are 2.44/6.66 points for CHAI re-observation and 2.27/6.12 for LLaVA re-observation. All configurations use the same anomaly scorer, and Seeing Twice improves detection with either initial model.

Reversing the mixed-model order changes both initial descriptions and recovered evidence, giving 87.12/75.93 for LLaVA followed by CHAI and 86.92/75.32 for CHAI followed by LLaVA. The same-model rows show that changing the observer is not required for gains from the full framework.

Control scope. These comparisons assess the complete framework against single observation. Table 2 instead fixes the downstream pipeline to examine routing, companion context, and description selection individually.

B.2 Evidence-gap Routing and Evidence-grounded Disambiguation

Table 2A compares random selection, visual–text mismatch, and hierarchical disagreement with the full routing strategy. Routing policies use matched seed counts and coarse/fine allocations per video, with the same companion-expansion and downstream prediction rules. Matching seed budgets controls the number of routing seeds, not total visual queries or pipeline latency.

Hierarchical companion controls. Table 2B compares three restricted companion configurations with full Seeing Twice at the top of the table. The seeds-only configuration queries $\mathcal{B}$ without companion expansion and achieves 86.23/74.67 UCF ROC-AUC/XD AP. Parent-only expansion adds the parents of fine seeds and achieves 86.55/75.24. Child-only expansion adds the selected children of coarse seeds and achieves 86.78/75.01. Full Seeing Twice applies both expansion rules and reaches 87.12/75.93. All configurations keep the routing seeds, observers, prompts, disambiguation, and score processing fixed. These comparisons vary the source of hierarchical context, rather than enforcing equal query counts.

Evidence-grounded disambiguation controls. Table 2C compares unconditional acceptance, visual evidence agreement, and local claim support with full evidence-grounded disambiguation. All variants use the same initial and recovered descriptions and the same downstream score processing.

The full method combines visual evidence agreement and local claim support through the admission rule in Equation 8. Each row changes how the two descriptions are selected, without adding another observation or candidate description.



Figure S2: Ablation comparisons organized by three questions: (A) which intervals to re-observe, (B) which hierarchical context to include, and (C) which recovered descriptions to use. Points show the ROC-AUC/AP values in Table 2. Orange diamonds repeat full Seeing Twice in each column, while blue circles show the corresponding controls. Routing comparisons match seed counts. Companion comparisons fix the seeds, not total query counts. Disambiguation comparisons reuse the same candidate descriptions.

Complementary evidence across the two levels. Figure S2 separates the three experimental questions. On XD, combined routing exceeds mismatch alone by 3.81 AP points and disagreement alone by 1.97 points. Parent-and-child expansion improves on parent-only and child-only expansion by 0.69 and 0.92 points. Combining agreement and claim support exceeds the respective single-criterion controls by 2.01 and 1.58 points. Full Seeing Twice also outperforms every control on UCF.

Unconditional acceptance trails the full method by 2.05 UCF ROC-AUC points and 3.89 XD AP points. Because this comparison reuses the same descriptions, it isolates evidence-grounded selection from additional visual generation. Companion comparisons instead assess complementary temporal views for fixed routing seeds.

B.3 Hierarchical Reconciliation and Temporal Smoothing

Each comparison removes one contextual operator while retaining the remaining coefficients and the video-local reference. Removing sibling messages or temporal smoothing produces the largest performance decreases in Table S4. Parent messages, visual-neighborhood aggregation, and parent adaptation also contribute, complementing re-observation and evidence-grounded disambiguation.

Table S4: Leave-one-out analysis of hierarchical reconciliation and temporal smoothing. All remaining operators retain the shared configuration on both datasets.

Configuration	UCF ROC-AUC	XD AP
Full Seeing Twice	**87.12**	**75.93**
Without parent messages ($\alpha = 0$)	86.98	75.72
Without sibling messages ($\beta = 0$)	86.61	74.36
Without visual-neighborhood aggregation	86.82	74.81
Without parent adaptation ($a = 0$)	87.00	75.56
Without temporal smoothing	86.69	74.33

B.4 Repeated Runs

The full configuration is evaluated with scoring seeds $\{0, 1, 2, 3, 4\}$. Table S5 reports the individual metrics, arithmetic means, and sample standard deviations with denominator $n - 1$.

Table S5: Five runs of the full configuration. All metrics are percentages.

Seed	UCF ROC-AUC	XD AP
0	87.03	75.68
1	87.18	76.02
2	87.08	75.81
3	87.24	76.19
4	87.07	75.95
Mean ± SD	87.12±0.09	75.93±0.20

UCF PR-AUC is 45.62±0.18. Run-to-run variability describes stochastic scoring and is distinct from the point differences between ablation configurations.

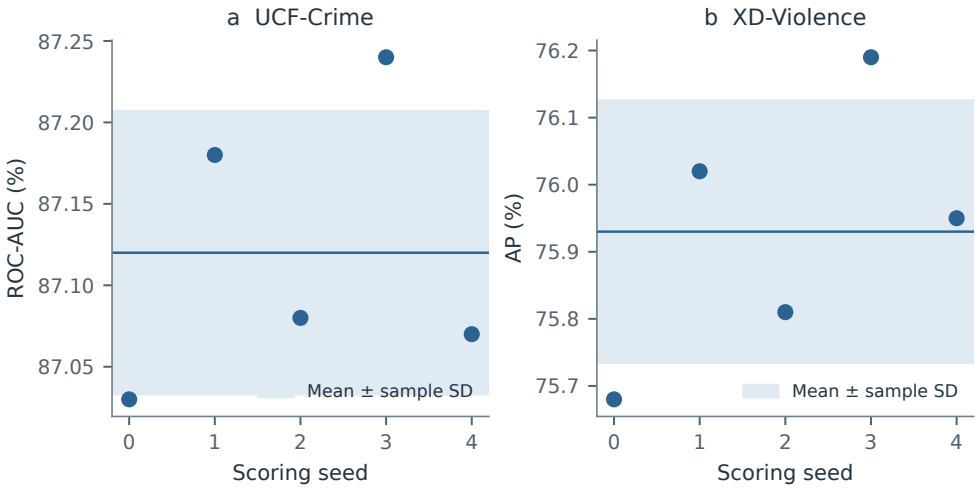

Figure S3: Repeated scoring with the full configuration. Points show all five runs from Table S5. Horizontal lines mark arithmetic means, and shaded bands show one sample standard deviation across scoring seeds. UCF-Crime uses ROC-AUC and XD-Violence uses AP.

Figure S3 shows UCF ROC-AUC ranging from 87.03 to 87.24 and XD AP ranging from 75.68 to 76.19. Every run exceeds single observation by at least 2.17 ROC-AUC points and 7.66 AP points. The improvement therefore persists across all five scoring runs. These ranges describe the full configuration, not paired significance tests for individual ablations.

B.5 Computational Efficiency Protocol

At the default 20% seed-routing budget, companion expansion results in re-observation of 2,463 out of 8,613 UCF-Crime intervals (28.6%) and 7,392 out of 24,000 XD-Violence intervals (30.8%). These actual queried-interval ratios differ from the seed budget because the queried set also includes hierarchical companions.

Section 4.4 varies $b_{\text{obs}} \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$ with the threshold, companion rule, and downstream processing fixed. For each video, $B = \lceil b_{\text{obs}} N \rceil$ limits seeds and $M = |\mathcal{H}|$ counts distinct queried intervals. Equation 13 bounds their expansion. Dataset-level ratios use summed counts, $\sum_V M_V / \sum_V N_V$. The 100% endpoint bypasses selection and queries all intervals, so it is identical to all-interval re-observation.

All budgets retain hierarchical reconciliation and temporal smoothing. At 0%, only initial LLaVA descriptions are used, yielding 84.86% UCF ROC-AUC and 68.02% XD AP. Positive budgets add re-observation and evidence-grounded disambiguation, with downstream processing fixed. The efficiency random control matches unique queried intervals per video, unlike the seed-matched control in Table 2A. Two-description disambiguation requires no additional VLM generation, giving $N + M$ visual requests and $2N$ for all-interval re-observation. GPU costs also include LLM claim extraction and scoring.

Costs use two NVIDIA GeForce RTX 4090 GPUs. Single-observation cost covers event construction, feature encoding, initial VLM descriptions, LLM scoring, hierarchical reconciliation, and temporal smoothing, totaling 62.3 GPU hours on UCF and 173.6 on XD. Re-observation totals additionally include recovered descriptions, claim extraction and encoding, disambiguation, and rescoring. Initial scoring is counted once. Mean per-video cost divides total GPU hours by 290 UCF or 800 XD videos. Figure 3 normalizes VLM requests as $(N + M)/N$, from 1 for single observation to 2 for all-interval re-observation. The random row's 20% label identifies the reference budget whose actual query count it matches. This query-matched control is available only for UCF.

Mean additional GPU time per video is the difference in dataset-level totals divided by the number of videos. On UCF-Crime, $(79.1 - 62.3)/290 \times 60 \approx 3.48$ GPU minutes. On XD-Violence, $(219.6 - 173.6)/800 \times 60 = 3.45$ GPU minutes. These dataset averages do not imply a constant increment for every video or quantify timing variability. They measure accumulated GPU compute time, not wall-clock latency.

Complete budget records. Table S6 connects the routing budget, distinct intervals queried after companion expansion, and complete inference cost. VLM queries include the initial description of every interval.

Table S6: Budget, unique queried intervals, and complete cost on both datasets. UCF reports ROC-AUC and XD reports AP, both in percent. Bold identifies the default 20% budget. All budgets retain hierarchical reconciliation. The 0% rows use single observation, and the 100% rows query all intervals.

Dataset	Budget (%)	Re-observed intervals	VLM queries	GPU h	Metric
UCF	0	0	8,613	62.3	84.86
	5	758	9,371	67.4	85.25
	10	1,430	10,043	71.4	86.67
	20	**2,463**	**11,076**	**79.1**	**87.12**
	30	3,342	11,955	83.0	87.17
	50	4,841	13,454	92.0	87.22
	100	8,613	17,226	131.5	87.24
XD	0	0	24,000	173.6	68.02
	5	2,160	26,160	188.0	69.81
	10	4,080	28,080	199.6	74.62
	20	**7,392**	**31,392**	**219.6**	**75.93**
	30	9,792	33,792	234.1	76.03
	50	13,920	37,920	259.0	76.12
	100	24,000	48,000	319.9	76.16

Realized query coverage. Figure S4 reports the actual queried-interval ratios for every selective budget. At a 10% seed budget, the ratios are 16.6% on UCF and 17.0% on XD. At the default 20% budget, they reach 28.6% and 30.8%, respectively. Companion expansion therefore provides additional temporal views while leaving most intervals unqueried a second time. Increasing the budget from 20% to 50% nearly doubles coverage, to 56.2% and 58.0%, but adds only 0.10 ROC-AUC and 0.19 AP points. The default configuration captures most of the observed detection gain before this additional expansion.

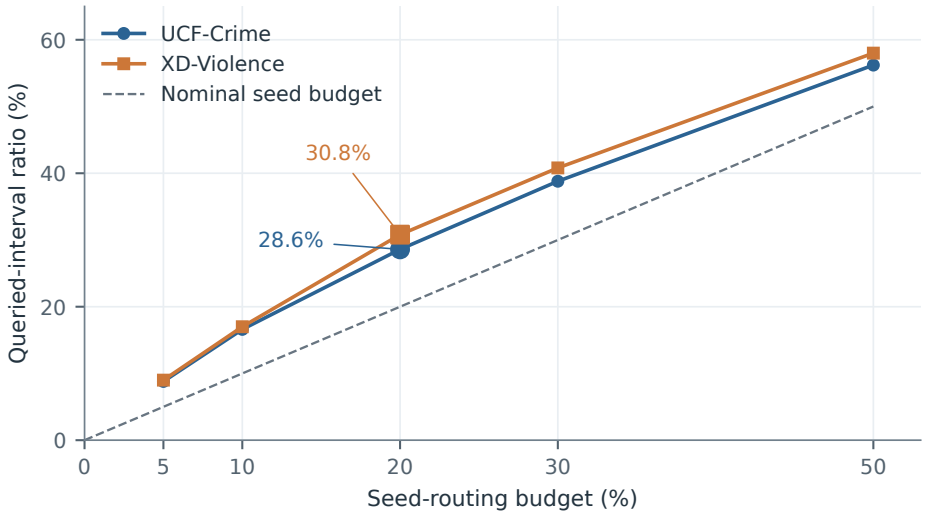

Figure S4: Actual re-observation coverage after companion expansion and deduplication. Both datasets include all five selective seed budgets from Table S6. The dashed diagonal is the nominal seed budget, not a measured query count. Labels identify the default 20% configuration. Coverage counts distinct queried intervals, including companions.

Marginal computation and detection gains. Figure S5 expresses the same sweep in mean additional GPU minutes per video and performance gains over single observation. The default configuration adds an average of 3.48 minutes on UCF and 3.45 minutes on XD. Increasing the UCF budget from 20% to exhaustive re-observation adds a further 52.4 GPU hours, averaging 10.84 minutes per video, for 0.12 ROC-AUC points. On XD, the corresponding increment is 100.3 GPU hours, averaging 7.52 minutes per video, for 0.23 AP points. Thus, the dominant gains occur before exhaustive re-observation, while the marginal cost continues to grow.

The query-matched random control uses the same 11,076 UCF visual requests but reaches 85.04 ROC-AUC at 79.8 GPU hours. Seeing Twice reaches 87.12 at 79.1 GPU hours, a 2.08-point advantage without more visual queries. GPU totals can differ because generated text and scoring work also contribute to cost.

All-interval re-observation doubles visual queries, not necessarily GPU time. Its GPU ratio is $131.5/62.3 = 2.11$ on UCF and $319.9/173.6 = 1.84$ on XD. Feature extraction is not repeated for each description, whereas language-model work contributes to total cost. Query normalization gives the main figure a common endpoint without altering measured GPU totals.

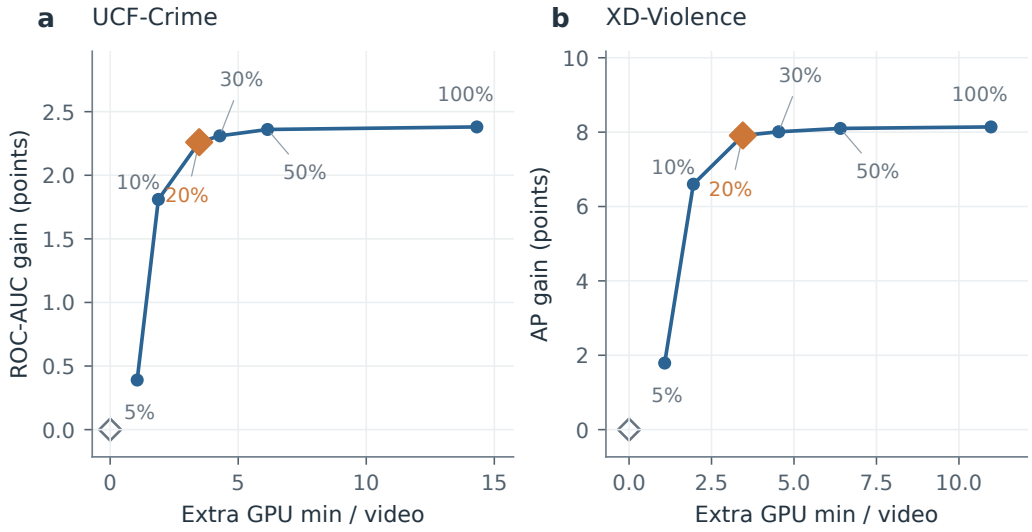

Figure S5: Performance gain versus mean additional GPU time per video, calculated from Table S6. Open diamonds denote single observation, and point labels give seed-routing budgets. Orange diamonds identify the default 20% budget, and 100% denotes all-interval re-observation. GPU minutes measure accumulated compute time rather than elapsed latency. Both panels include the complete budget sweep.

C Comparison Protocols and Additional Qualitative Analysis

Additional weakly supervised comparisons. Table S7 retains further published baselines, while the main table emphasizes training-free inference. These methods use video-level supervision and are not controlled variants of Seeing Twice.

Table S7: Additional weakly supervised VAD methods under their reported protocols (%).

Method	UCF ROC-AUC	XD AP
MIST (Feng et al., 2021)	82.30	–
MSL (Li et al., 2022)	85.30	78.28

Published comparison protocols. Tables 1 and S7 cover representative formally published VAD methods, drawing on original papers and the comparisons in Li et al. (2025c), Cai et al. (2026). MIL uses C3D, whereas MIST, RTFM, MSL, and UR-DMU use I3D, with RGB-only UR-DMU results. CoMo uses RGB I3D. Its 86.07/81.31 pair is reported in the authors' main-paper and supplementary evaluations. UMIL and VadCLIP use CLIP. MGFN uses I3D on UCF and VideoSwin on XD. VADTree uses its visual-only results.

Frozen backbones do not by themselves imply training-free VAD. VERA optimizes verbal prompts from labeled training examples, while HiProbe-VAD trains a lightweight scorer with InternVL2.5. SteerVAD trains an InternVL3-based controller and scorer on a 1% training subset. These methods are separated from inference without task-specific learning. Table 1 is a published-protocol comparison, not a common-backbone reimplementation. Its rankings refer to the listed methods and supervision groups.

Direct frozen-model baselines and description-based VAD. Zero-shot CLIP and LLaVA-1.5 are direct frozen-model baselines, with UCF ROC-AUC/XD AP of 53.16/17.83 and 72.84/50.26. Their model citations identify the pretrained architectures, while these VAD scores come from Tables 1–2 of Zanella et al. (2024). SUVAD uses semantic descriptions for detection, while MCANet combines visual and audio descriptions. Their respective 83.90/70.10 and 82.47/69.72 results follow Tables

1–2 of Li et al. (2025c). The asterisk on MCANet's XD result identifies its additional audio input. The comparison includes eight training-free VAD competitors, two direct frozen-model baselines, and Seeing Twice.

Context-aware and agentic training-free methods. PANDA (NeurIPS 2025) combines scene-aware planning, tool-assisted reflection, and memory (Yang et al., 2025). We report its default offline configuration, 84.89 UCF ROC-AUC and 70.16 XD AP, from its Table 1, and its online configuration is a separate operating mode. AnyAnomaly (WACV 2026) reports 80.70 UCF ROC-AUC in its Table 5 (Ahn et al., 2026). It queries the model with anomaly-category text and aggregates scores across these queries, following its published zero-shot protocol. XD AP is unreported and remains blank. ReCI (ICMR 2026) uses InternVL2-8B with semantic context, an online historical reference pool, and confidence-aware refinement. We use its full-benchmark 81.51/72.23 values, not its curated low-light subset results.

C.1 Additional Qualitative Analysis

Four additional videos illustrate implicit anomaly cues and semantic ambiguity. Each case pairs original frames with stored initial and recovered description excerpts. Frame indices are zero-based, timestamps are relative to the dataset clip, and ellipses mark omitted text.

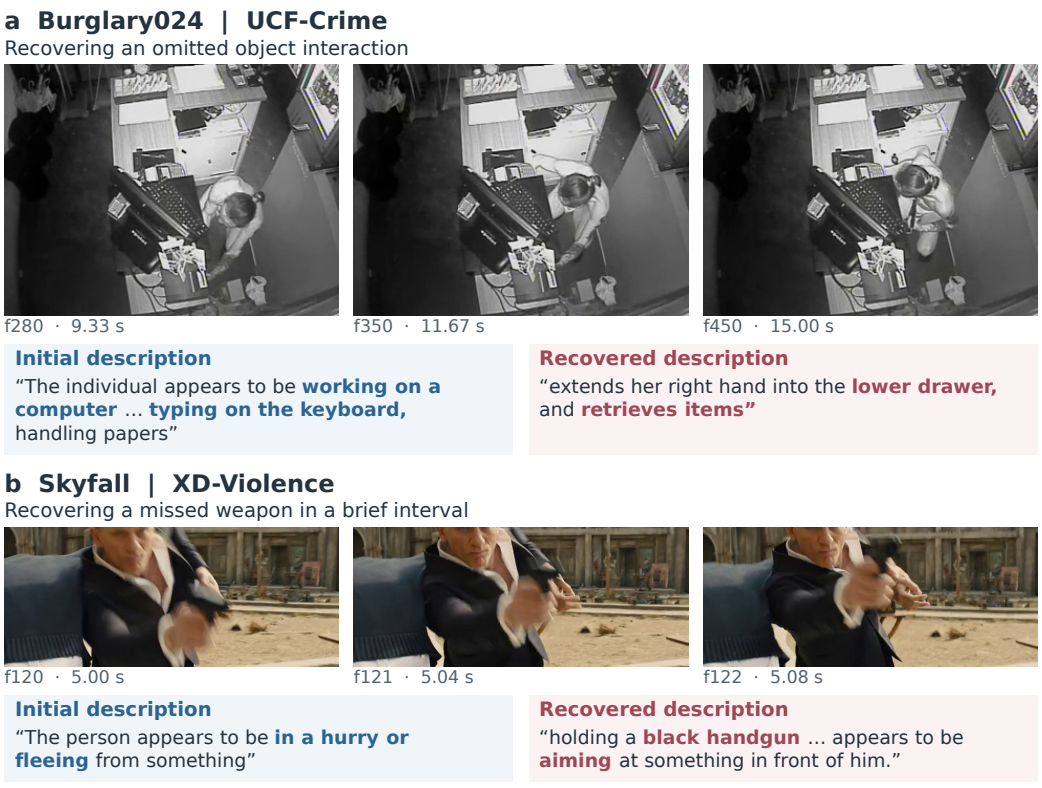

Figure S6: Recovery of implicit anomaly cues. (a) Burglary024, frames 270–488: a drawer interaction replaces the interpretation of routine computer work. (b) Skyfall, frames 120–122: a handgun becomes explicit in an initially movement-focused description. Colored bold phrases highlight the evidence in the verbatim excerpts.

Recovering an omitted interaction. In Figure S6a, the initial description substitutes typing and handling papers for the person's interaction with the drawer. The recovered description instead identifies reaching and retrieving. The frames show the person bending below the counter and bringing a hand back towards its surface. Recognizing the surrounding workplace is therefore insufficient to capture the anomaly-relevant action.

Recovering a brief object cue. In Figure S6b, the initial description captures movement but omits the handgun visible as the person's arm extends. The recovered description identifies the object and aiming action. Together, the cases show two forms of missing evidence: an interaction obscured by a broad scene description and an object omitted in a brief interval. Re-observation makes the relevant action or object explicit in both cases.

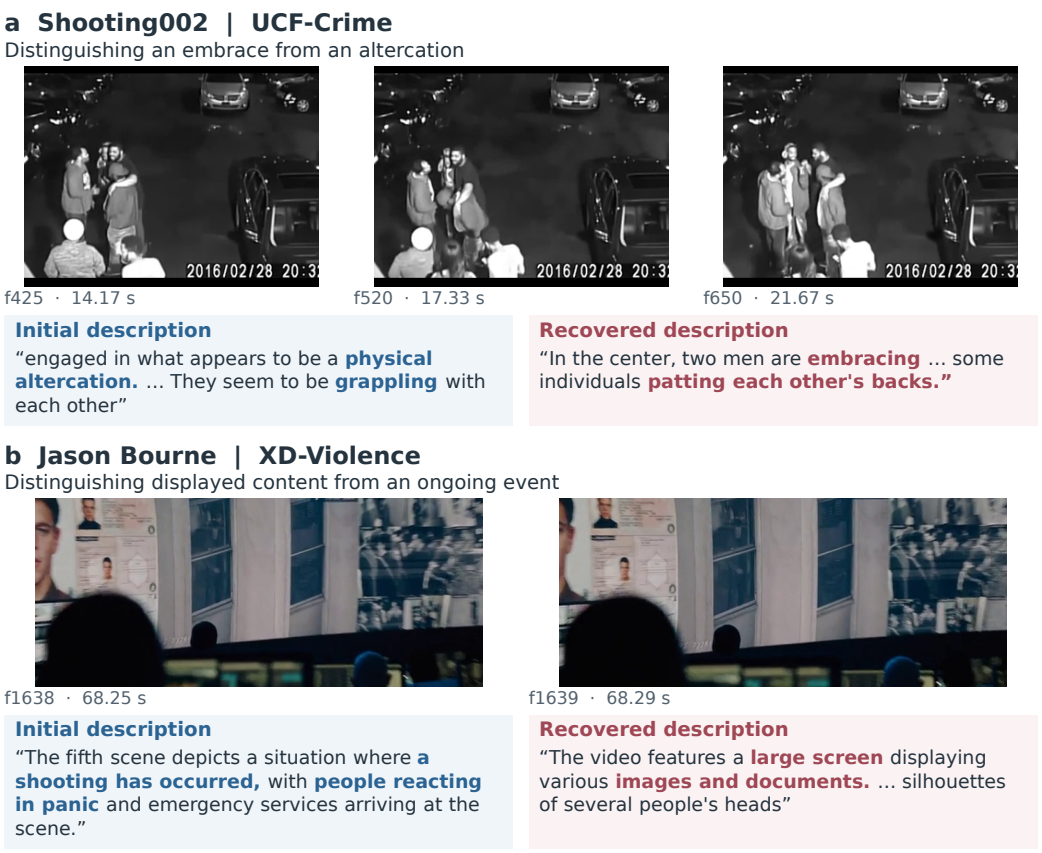

Figure S7: Resolution of semantic ambiguity. (a) Shooting002, frames 414–696: embracing replaces an altercation interpretation in an interval preceding the annotated shooting. (b) Jason Bourne, frames 1638–1639: a display of images and documents replaces an unsupported shooting interpretation. Key phrases are highlighted in the verbatim excerpts.

Distinguishing contact from conflict. Figure S7a concerns the meaning of an interaction rather than a missing object. Both descriptions capture close physical contact, but the initial description frames it as a fight. The recovered description identifies embracing and back-patting, consistent with an arm resting around another person's shoulders across the sampled frames. The later shooting does not make this preceding interaction anomalous, so the video's category cannot determine the meaning of every interval.

Distinguishing a depicted event from the observed scene. In Figure S7b, the initial description introduces a shooting and emergency response. The recovered description identifies the large screen, its images and documents, and the viewers in front of it. These relationships distinguish a monitoring scene from the ongoing incident asserted by the initial text. The surrounding visual context is essential because recognizing content on a screen does not establish that the depicted event occurs in the observed scene.

These cases complement evidence recovery by showing that a useful revision need not make an event sound more anomalous. Re-observation can reveal a missing action or weapon, but it can also replace unsupported anomaly claims with a visually grounded interpretation.