WHERE TO VERIFY, HOW TO CORRECT: SELECTIVE SEMANTIC VERIFICATION FOR WEAKLY SUPERVISED VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly Supervised Video Anomaly Detection (WSVAD) aims to localize frame-level anomalies using only video-level supervision, yet visually dynamic normal activities can still receive high anomaly scores and cause false alarms. Language-guided verification provides complementary semantic evidence for reassessing such elevated responses. However, applying semantic verification to every snippet is costly for long videos and does not necessarily improve anomaly localization. We propose Semantic-Temporal Reprojection (STR), a training-free framework for selective semantic verification of frozen WSVAD detectors. STR retains dense baseline scoring and applies semantic verification to selected elevated responses. Adaptive temporal anchor mining allocates queries across distinct temporal contexts. Structured semantic verification produces a bounded semantic score, while a video-adaptive perception veto limits corrections at extreme baseline responses. Local Gaussian reprojection constrains the remaining corrections to nearby temporal locations. The resulting one-sided correction operates on candidates exposed by the frozen detector and retains or attenuates existing anomaly scores. Experiments on UCF-Crime, XD-Violence, and MSAD show consistent improvements across multiple frozen WSVAD baselines, achieving 90.80% AUC, 89.08% AP, and 88.40% AUC, respectively. Under matched query budgets, adaptive mining outperforms uniform, random, and score-based Top-K selection. Compared with dense verification, STR uses far fewer semantic queries and achieves higher detection performance with lower verification-stage runtime. Code is available at `https://anonymous.4open.science/r/STR`.

1 INTRODUCTION

Weakly Supervised Video Anomaly Detection (WSVAD) aims to localize frame-level anomalous events using only video-level supervision. Most methods formulate each video as a bag of temporal snippets and learn snippet-level anomaly scores through Multiple Instance Learning (MIL) (Sultani et al., 2018). Although recent methods improve feature discrimination and temporal modeling (Tian et al., 2021; Chen et al., 2023b; Lv et al., 2023), video-level supervision still provides limited constraints on individual normal snippets within abnormal videos. Consequently, visually dynamic normal activities such as dancing, crowd motion, and rapid traffic may receive high anomaly scores even when no anomalous event occurs. Such false alarms are difficult to suppress using visual anomaly scores alone.

Language-derived semantic evidence provides a complementary cue for resolving such visually ambiguous responses. CLIP-based WSVAD methods enrich visual representations with textual priors (Wu et al., 2024b; Yang et al., 2024), while recent foundation-model approaches use video descriptions and language reasoning to identify or explain anomalous events (Zanella et al., 2024; Ye et al., 2025; Shao et al., 2025). Long-video anomaly understanding has also begun to use adaptive sampling to concentrate multimodal reasoning on anomaly-relevant content (Zhang et al., 2025). These studies demonstrate that language-derived semantics and selective processing can complement conventional visual anomaly detection. Existing methods mainly enhance visual-language representations or obtain anomaly predictions directly from foundation models.

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.

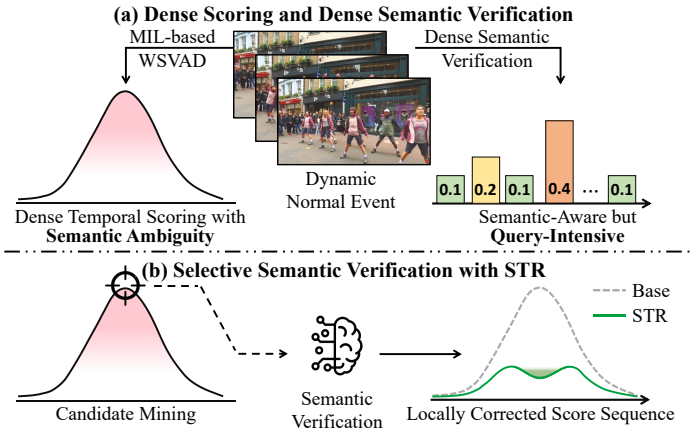


Figure 1: (a) MIL-based WSVAD can yield false alarms on dynamic normal activities, whereas dense semantic verification is query-intensive. (b) Our STR selectively verifies high-score candidates and locally refines baseline scores.

However, dense semantic verification is not necessarily more effective for refining a frozen WSVAD score sequence. High-score responses with ambiguous visual evidence are natural targets for semantic verification because local event context can help distinguish true anomalies from visually dynamic normal activities. Given the dense anomaly scores already produced by the frozen detector, applying the verifier to every snippet substantially increases computation and requires additional semantic judgments throughout the video. A key design challenge is to determine where semantic verification should be applied and how each judgment should modify the corresponding local score sequence. Figure 1 illustrates this joint problem of selective query allocation and constrained local score correction.

These considerations motivate selective semantic verification for frozen WSVAD detectors. The detector retains its dense temporal anomaly-score sequence, while semantic verification is applied to a sparse set of locally elevated candidates. Candidate selection determines where semantic evidence is consulted, and score correction determines how that evidence modifies the temporal score sequence. The correction operates on candidates exposed by the frozen detector and is constrained to retain or attenuate existing scores. This design preserves dense visual coverage, focuses semantic analysis on selected candidates, and keeps the underlying detector unchanged across frozen baselines.

To address this problem, we propose Semantic-Temporal Reprojection (STR), a training-free framework that couples selective semantic verification with bounded local score correction. Adaptive temporal anchor mining combines feature changes with local score structure to distribute semantic queries across distinct temporal contexts. For each selected anchor, structured semantic verification produces a bounded score from its multi-frame visual context. A video-adaptive perception veto limits corrections at extreme baseline responses, while Gaussian reprojection propagates the remaining correction strengths to nearby temporal snippets. Experiments on UCF-Crime, XD-Violence, and MSAD show consistent gains across frozen baselines, reaching 90.80% AUC, 89.08% AP, and 88.40% AUC, respectively. Under matched query budgets, adaptive mining outperforms uniform, random, and score-based Top-K selection on UCF-Crime and XD-Violence. Compared with dense verification, STR achieves higher detection metrics with substantially fewer semantic queries and lower verification-stage runtime.

The main contributions are summarized as follows:

- We introduce STR, a training-free framework for selective semantic verification of frozen WSVAD detectors. STR jointly controls semantic query allocation and local score correction without retraining the underlying detector.
- We develop adaptive temporal anchor mining and a perception-guarded correction mechanism that combines structured semantic verification, a video-adaptive veto, and local Gaussian reprojection. The resulting correction remains bounded and temporally local.

- Experiments across three benchmarks, multiple frozen detectors, different visual features, and multiple vision-language models demonstrate consistent gains across the evaluated settings. Matched-budget comparisons further show the benefit of adaptive query allocation, while a semantic-correspondence control shows that context-score alignment contributes to the observed gains.

2 Related Work

2.1 Weakly Supervised Video Anomaly Detection

Weakly supervised video anomaly detection has progressively improved anomaly discrimination through stronger visual and semantic representations, yet video-level supervision still leaves visually ambiguous normal activities difficult to separate from true anomalies. Sultani et al. (2018) formulated WSVAD as a Multiple Instance Learning problem that learns snippet-level anomaly scores from video-level labels. Subsequent methods improve temporal feature discrimination and reduce supervision bias. RTFM (Tian et al., 2021) and MGFN (Chen et al., 2023b) exploit feature magnitude to distinguish anomalous snippets, while UMIL (Lv et al., 2023) mitigates contextual bias induced by abnormal training videos. More recent approaches introduce language priors to strengthen semantic modeling. VadCLIP (Wu et al., 2024b) transfers fine-grained vision-language associations to WSVAD, and TPWNG (Yang et al., 2024) uses normality-guided prompts to improve pseudo-label generation. STPrompt (Wu et al., 2024a), PE-MIL (Chen et al., 2024), and TDSD (Sun et al., 2024) further incorporate textual prompts or captions into feature learning. LEC-VAD (Wang & Chen, 2025) combines vision-language semantics with anomaly-aware temporal regularization to improve event completeness. Recent methods also improve training-time supervision by exploring diverse and uncertain snippets (Acharya et al., 2026) or re-evaluating discarded samples with a vision-language model (Zhao et al., 2026). This line of work primarily strengthens anomaly discrimination through representation learning, semantic modeling, or training-time supervision refinement.

2.2 Semantic Reasoning and Efficient Verification

Foundation models have extended video anomaly analysis from visual scoring to language-guided detection and explanation. LAVAD (Zanella et al., 2024) converts frame descriptions into temporally aggregated anomaly scores without task-specific training. HiProbe-VAD (Cai et al., 2025) uses intermediate multimodal representations for anomaly scoring and explanation, while VERA (Ye et al., 2025) generates segment-level scores and language-based rationales through verbalized learning. CUVA (Du et al., 2024) further broadens anomaly understanding from temporal localization to descriptions of what occurred, why it happened, and how severe its effects were. These methods show that foundation models can support anomaly scoring, temporal interpretation, and human-readable explanations. Recent methods also improve long-video reasoning through selective temporal processing. Holmes-VAU (Zhang et al., 2025) uses an anomaly-focused sampler to concentrate multimodal analysis on anomaly-relevant frames. EventVAD (Shao et al., 2025) detects event boundaries to reduce the visual content processed by multimodal models and preserve event-level context. Lin et al. (2025) further employ chained test-time reasoning for temporal anomaly detection, spatial localization, and anomaly explanation. These studies demonstrate the value of selective processing and semantic reasoning for video anomaly analysis. STR focuses on frozen WSVAD detectors that already provide dense temporal anomaly scores. It selectively allocates semantic verification to elevated responses and constrains the resulting score correction to local temporal neighborhoods.

3 Method

The proposed Semantic-Temporal Reprojection (STR) framework augments a frozen WSVAD detector with selective semantic verification and perception-guarded local score correction. As illustrated in Figure 2, the framework first selects a small set of temporal candidates from the baseline score sequence, verifies each candidate using its local visual context, and propagates the resulting correction strength to nearby snippets. The following subsections introduce the problem formulation (Section 3.1), adaptive anchor mining (Section 3.2), structured semantic verification (Section 3.3), and local Gaussian score correction (Section 3.4).

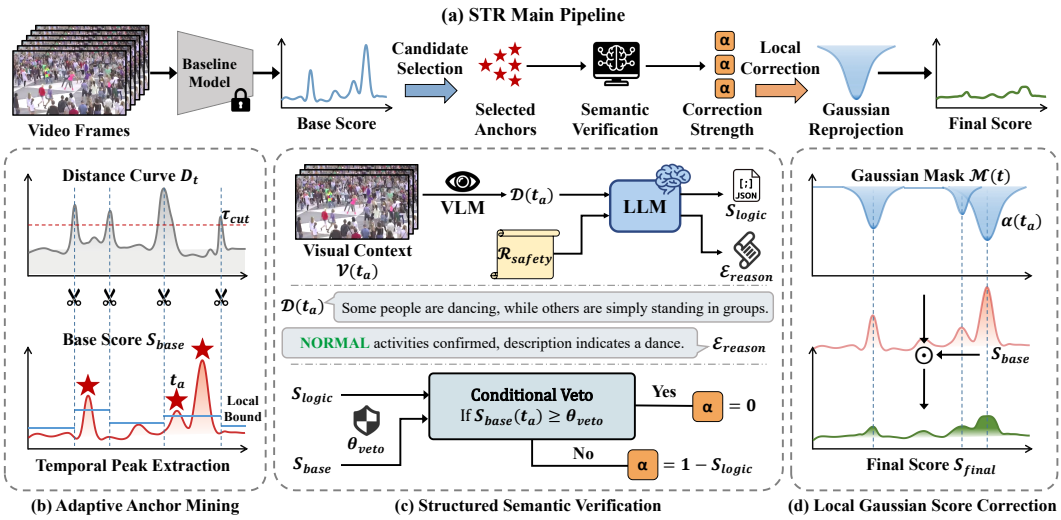


Figure 2: Overview of STR. (a) The complete inference pipeline. (b) Adaptive anchor mining selects local score peaks within feature-based temporal segments. (c) Semantic verification produces $\mathcal{S}_{logic}$ and a human-readable rationale, while the veto limits correction of top-scoring responses. (d) Gaussian reprojection propagates local corrections to obtain $\mathcal{S}_{final}$.

3.1 Problem Formulation

Given an untrimmed video, let $X \in \mathbb{R}^{T \times d}$ denote its temporal feature sequence, where T is the number of feature snippets and d is the feature dimension. A frozen WSVAD baseline maps X to a temporal anomaly-score sequence at the same snippet resolution:

$$\mathcal{S}_{base} = \mathcal{F}_{baseline}(X), \qquad \mathcal{S}_{base} \in [0,1]^T, \tag{1}$$

where $\mathcal{F}_{baseline}(\cdot)$ denotes the scoring function of the underlying detector. STR operates directly on X and $\mathcal{S}_{base}$ without updating the parameters or architecture of $\mathcal{F}_{baseline}$. It selects a sparse anchor set $\mathcal{A} = \{t_a^{(i)}\}_{i=1}^{K}$, obtains structured semantic evidence for these anchors, and produces a corrected score sequence $\mathcal{S}_{final}$. Because the anchor set is derived from $\mathcal{S}_{base}$, STR operates on elevated candidates exposed by the frozen detector. Its correction is one-sided and retains or reduces baseline scores, as defined in Section 3.4. This decoupled formulation enables the same verification and correction procedure to be applied to different frozen baselines and visual features.

3.2 Adaptive Temporal Anchor Mining

Selective semantic verification restricts vision-language inference to a small set of candidate temporal locations. STR mines these candidates from the temporal structure of the frozen feature and score sequences. The selected anchors identify high-score locations for semantic verification. The subsequent two-stage semantic analysis determines whether these locations contain semantically benign or anomalous evidence.

We first use adjacent-feature changes to divide the sequence into local temporal segments. For adjacent feature vectors $X_t, X_{t+1} \in \mathbb{R}^d$, their cosine distance is

$$D_t = 1 - \frac{\langle X_t, X_{t+1} \rangle}{\|X_t\|_2 \|X_{t+1}\|_2}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\| \cdot \|_2$ denotes the L_2 norm. We identify unusually large feature changes using the video-adaptive empirical threshold

$$\tau_{cut} = \mu_D + 3\sigma_D, \tag{3}$$

where μ_D and σ_D are the mean and standard deviation of $\{D_t\}_{t=1}^{T-1}$. Locations satisfying $D_t > \tau_{cut}$ divide the video into local segments $\{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$. The distance threshold is used only to construct local temporal contexts. It does not directly determine whether a segment is anomalous.

For each temporal segment $\mathcal{C}$, STR retains locally elevated baseline scores as candidate locations:

$$\mathcal{P} = \{t \in \mathcal{C} \mid \mathcal{S}_{base}(t) > \mu_{\mathcal{C}} + \beta\sigma_{\mathcal{C}}\}, \tag{4}$$

where $\mu_{\mathcal{C}}$ and $\sigma_{\mathcal{C}}$ denote the mean and standard deviation of the baseline anomaly scores in $\mathcal{C}$. For the current segment, the candidate set $\mathcal{P}$ is divided into temporally contiguous sub-segments $\{\mathcal{P}_1, \ldots, \mathcal{P}_m\}$. For each resulting sub-segment, STR selects the location with the highest baseline response:

$$t_a^{(i)} = \arg\max_{t \in \mathcal{P}_i} \mathcal{S}_{base}(t). \tag{5}$$

Across all temporal segments, the selected locations are collected and indexed as the anchor set $\mathcal{A} = \{t_a^{(i)}\}_{i=1}^{K}$. The frozen detector still processes the complete sequence, while semantic verification is reduced from T dense queries to K anchor queries, where $K \ll T$ in practice.

3.3 Structured Semantic Verification

Semantic verification determines whether the high baseline response at an anchor is supported by visual evidence of an anomalous event. For each $t_a \in \mathcal{A}$, STR extracts a multi-frame local visual context $\mathcal{V}(t_a)$. A vision-language model $\mathcal{F}_{VLM}$ first converts this context into a textual description:

$$\mathcal{D}(t_a) = \mathcal{F}_{VLM}\big(\mathcal{V}(t_a)\big). \tag{6}$$

Conditioned on the fixed rule prompt $\mathcal{R}_{safety}$, the LLM maps the visual description to a structured semantic output:

$$\mathcal{Y}(t_a) = \mathcal{F}_{LLM}\left(\mathcal{D}(t_a) \mid \mathcal{R}_{safety}\right). \tag{7}$$

The structured output $\mathcal{Y}(t_a)$ contains a bounded semantic score $\mathcal{S}_{logic}(t_a) \in [0, 1]$ and a textual rationale $\mathcal{E}_{reason}(t_a)$. A larger semantic score indicates stronger evidence of an anomaly. The rationale provides a human-readable explanation and is not used in numeric score correction. An invalid, missing, or out-of-range structured output is assigned zero correction strength and therefore does not modify the baseline score sequence. The rule prompt assigns low scores to explicitly benign contexts, intermediate scores to uncertain descriptions, and high scores to confirmed violence, crime, fire, explosion, or accidents. For example, dancing, standing, and normal traffic are treated as benign only when the visual context contains no evidence of aggression, collision, fire, weapons, emergency, or distress. The complete prompt and output format are provided in the supplement.

The perception veto excludes the highest-scoring baseline responses from semantic attenuation. We derive a video-adaptive threshold directly from the empirical distribution of baseline scores:

$$\theta_{veto} = Q_\rho\big(\mathcal{S}_{base}\big), \tag{8}$$

where $Q_\rho(\cdot)$ denotes the empirical quantile at level ρ for each input video. The resulting threshold adapts the protected score range to the corresponding video. For each anchor with a valid semantic output, the semantic score determines a correction strength $\alpha(t_a)$:

$$\alpha(t_a) = \begin{cases} 0, & \mathcal{S}_{base}(t_a) \geq \theta_{veto}, \\ 1 - \mathcal{S}_{logic}(t_a), & \mathcal{S}_{base}(t_a) < \theta_{veto}. \end{cases} \tag{9}$$

Anchors above the threshold retain their baseline responses, while the remaining anchors receive attenuation according to their semantic scores. This asymmetric rule bounds how strongly semantic evidence can modify the highest-ranked baseline responses.

3.4 Local Gaussian Score Correction

Semantic verification produces correction strengths only at the selected anchors. STR reprojects these sparse anchor-level corrections onto the dense score sequence by propagating each one to a local temporal neighborhood around its anchor using Gaussian weighting.

For an anchor t_a, the local Gaussian attenuation is determined by the correction strength $\alpha(t_a)$ and the standard deviation ω, measured in snippets. The corrections from all anchors are combined into

$$\mathcal{M}(t) = \min_{t_a \in \mathcal{A}} \left\{ 1 - \alpha(t_a) \exp\left(-\frac{(t - t_a)^2}{2\omega^2} \right) \right\}. \tag{10}$$

The minimum operator selects the strongest single local correction at each temporal position without accumulating multiplicative attenuation from overlapping anchors. When $\mathcal{A}$ is empty, we set $\mathcal{M}(t) = 1$ for all t. The final corrected score sequence is

$$\mathcal{S}_{final} = \mathcal{S}_{base} \odot \mathcal{M}, \tag{11}$$

where $\odot$ denotes element-wise multiplication. The standard deviation ω controls the local temporal extent of each correction and does not define the duration of an anomalous event. Since $\mathcal{M}(t) \in [0, 1]$, STR can only retain or reduce each normalized baseline anomaly score.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate STR on three WSVAD benchmarks covering diverse anomaly categories and recording scenarios. **UCF-Crime** (Sultani et al., 2018) is a large-scale, untrimmed dataset comprising 1,610 training videos and 290 testing videos, totaling 128 hours. Collected from real-world surveillance systems, it covers 13 distinct anomaly categories spanning both indoor and outdoor settings. **XD-Violence** (Wu et al., 2020) is the largest dataset focused on violent events, containing 3,954 training videos and 800 testing videos from movies, games, surveillance cameras, and dashcams, covering 6 violence categories. **MSAD** (Zhu et al., 2024) contains 720 videos across 14 scenarios and 55 fine-grained human-related and environmental anomaly types.

Evaluation Metrics. Following standard WSVAD protocols (Wang & Zhao, 2026; Zhu et al., 2026), we evaluate frame-level anomaly detection. For UCF-Crime and MSAD, we report the area under the receiver-operating-characteristic curve (**AUC**). For XD-Violence, whose normal and anomalous frames are strongly imbalanced, we report average precision (**AP**). The ablation analysis additionally reports frame-level recall at a fixed false-positive rate of 5%.

4.2 Implementation Details

Visual features follow the CLIP ViT-B/16 or I3D configuration of each frozen baseline. We use Qwen3-VL-8B (Bai et al., 2025) for visual description and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) for structured semantic verification. Each anchor uses a 32-frame window, with 16 frames sampled as VLM input. The LLM adopts greedy decoding and returns a bounded score with a textual rationale. The same prompts, semantic score intervals, and decoding strategy are used across all datasets and frozen baselines and remain unchanged in the cross-VLM comparison. Semantic outputs are fixed for detection comparisons and regenerated without caches for runtime measurement. We use a single global configuration with $\beta = 0.5$, $\rho = 99.5\%$, and $\omega = 2$ snippets across all datasets and frozen baselines. All experiments run on a single NVIDIA RTX 4090 GPU with approximately 20 GB peak memory usage.

4.3 Comparison with State-of-the-Art

We compare STR with recent WSVAD and vision-language anomaly detection methods on UCF-Crime, XD-Violence, and MSAD. The main UCF-Crime and XD-Violence results apply STR to frozen DeSC scores (Zheng et al., 2026), while the MSAD result uses frozen RTFM scores (Tian et al., 2021). On XD-Violence, STR raises the frozen DeSC result from 87.18% to 89.08% AP, an absolute gain of 1.90 points. On UCF-Crime, STR raises DeSC from 89.37% to 90.80% AUC, an absolute gain of 1.43 points. These results show that selective semantic verification remains effective when applied to a strong CLIP-based score sequence without retraining the underlying detector. On MSAD, applying STR to frozen RTFM raises AUC from 86.65% to 88.40%, extending the improvement to an I3D-based detector and a benchmark covering both human-related and environmental anomaly types. Together, these results show that the observed gains are not restricted to a single visual feature family.

4.4 Cross-Baseline and Cross-Model Transfer

Transfer across frozen baselines. Table 3 reports positive gains in all eight evaluated baseline–dataset settings. The improvements hold for both I3D and CLIP features and for baselines with

Table 1: Comparison with state-of-the-art methods on XD-Violence and UCF-Crime. Best and second-best results are shown in **bold** and underlined. $*$ indicates additional dense physical modalities, including pose, depth, panoptic masks, and optical flow beyond standard visual features.

Method	Venue	Feature	XD (AP %)	UCF (AUC %)
Sultani et al. (2018)	CVPR'18	I3D	75.68	77.92
MSL (Li et al., 2022)	AAAI'22	I3D	78.28	85.30
DMU (Zhou et al., 2023)	AAAI'23	I3D	81.66	86.97
Sultani et al. (2018)	CVPR'18	CLIP	75.18	84.14
CLIP-TSA (Joo et al., 2023)	ICIP'23	CLIP	82.17	87.58
TDSD (Sun et al., 2024)	MM'24	CLIP	84.69	-
TPWNG (Yang et al., 2024)	CVPR'24	CLIP	83.68	87.79
VadCLIP (Wu et al., 2024b)	AAAI'24	CLIP	84.51	88.02
CMHKF (Wang et al., 2025)	ACL'25	CLIP	86.57	88.24
EventVAD (Shao et al., 2025)	MM'25	CLIP	64.04	82.03
VERA (Ye et al., 2025)	CVPR'25	CLIP	70.54	86.55
π-VAD (Majhi et al., 2025)	CVPR'25	CLIP*	85.37	<u>90.33</u>
Holmes-VAU (Zhang et al., 2025)	CVPR'25	CLIP	<u>87.68</u>	88.96
DeSC (Zheng et al., 2026)	ICME'26	CLIP	87.18	89.37
DSANet (Yin et al., 2026)	AAAI'26	CLIP	86.95	89.44
TLMA (Xu et al., 2026)	CVPR'26	MAE	86.78	89.47
STR (Ours)	-	**CLIP**	**89.08**	**90.80**

Table 2: Comparison with state-of-the-art methods on the MSAD dataset. Best results are in **bold**.

Method	Venue	Feature	MSAD (AUC %)
MIST (Feng et al., 2021)	CVPR'21	I3D	82.65
RTFM (Tian et al., 2021)	ICCV'21	I3D	86.65
MGFN (Chen et al., 2023b)	AAAI'23	I3D	84.96
DMU (Zhou et al., 2023)	AAAI'23	I3D	85.02
TEVAD (Chen et al., 2023a)	CVPR'23	I3D	86.82
RTFM + STR (Ours)	-	**I3D**	**88.40**

substantially different initial performance. This consistency supports the transferability of STR across the evaluated feature spaces and detector architectures. The gains on the strongest DeSC configurations show that STR is not limited to relatively weak baselines alone.

Transfer across vision-language models. We further replace Qwen3-VL-8B (Bai et al., 2025) with LLaVA-NeXT-Video-7B (Zhang et al., 2024) and CogVLM2 (Hong et al., 2024). The UCF-Crime AUC varies by at most 0.30 across the three vision-language models, with similarly small variation on XD-Violence. The small variation across these backbones supports cross-model transfer. Full results are provided in the supplement.

4.5 Query Allocation and Verification Cost

Matched-budget anchor selection. Table 4 provides a controlled comparison of query allocation by assigning exactly the same per-video query count to all competing sparse strategies. Adaptive mining outperforms uniform, random, and score-based Top-K selection on both datasets. Its advantage over Top-K shows that combining feature-based temporal segmentation with local score peaks provides more effective query placement than ranking baseline scores alone.

Dense verification reference. Table 5 complements the matched-budget comparison with dense verification. Dense pointwise verification obtains 88.74% AUC on UCF-Crime and 87.16% AP on XD-Violence. Applying the same Gaussian correction densely gives 88.93% AUC on UCF-Crime and 87.57% AP on XD-Violence, while STR reaches 90.80% AUC and 89.08% AP. Sparse pointwise correction obtains 90.35% AUC and 88.63% AP, showing that both query placement and local reprojection contribute to the final performance. Verification-stage runtime excludes feature extraction, frozen-baseline inference, and model loading, and is averaged over three uncached synchronized runs with batch size 1, including video decoding and frame loading. STR reduces the

Table 3: Cross-baseline transfer on frozen WSVAD detectors. STR is applied without baseline retraining, and absolute gains are shown in parentheses.

Dataset	Baseline	Feature	Base	+ STR (Δ)
XD-Violence (AP %)	DMU (Zhou et al., 2023)	I3D	81.66	**82.89** (↑ 1.23)
	VadCLIP (Wu et al., 2024b)	CLIP	84.51	**86.22** (↑ 1.71)
	DeSC (Zheng et al., 2026)	CLIP	87.18	**89.08** (↑ 1.90)
UCF-Crime (AUC %)	DMU (Zhou et al., 2023)	I3D	86.97	**87.77** (↑ 0.80)
	VadCLIP (Wu et al., 2024b)	CLIP	88.02	**88.97** (↑ 0.95)
	DeSC (Zheng et al., 2026)	CLIP	89.37	**90.80** (↑ 1.43)
MSAD (AUC %)	MGFN (Chen et al., 2023b)	I3D	84.96	**87.23** (↑ 2.27)
	RTFM (Tian et al., 2021)	I3D	86.65	**88.40** (↑ 1.75)

Table 4: Matched-budget anchor selection with equal per-video query counts. Random sampling reports the mean over five fixed seeds.

Selection strategy	UCF (AUC %)	XD (AP %)
Uniform sampling	88.96	87.62
Random sampling	89.12	86.51
Top-K by score	89.53	88.39
Adaptive mining	**90.80**	**89.08**

Table 5: Detection performance and cost under dense and selective semantic verification.

Data	Strategy	AUC / AP (%) ↑	Queries per vid. ↓	Verification sec. / vid. ↓	Speedup ↑
UCF	Dense	88.74	239.20	709.9	1.0×
	STR	**90.80**	**1.46**	**13.8**	**51.4**×
XD	Dense	87.16	182.13	913.1	1.0×
	STR	**89.08**	**1.94**	**8.3**	**110.3**×

average number of semantic queries from 239.20 to 1.46 per video on UCF-Crime and from 182.13 to 1.94 on XD-Violence, corresponding to 51.4× and 110.3× verification-stage speedups.

4.6 Ablation Studies

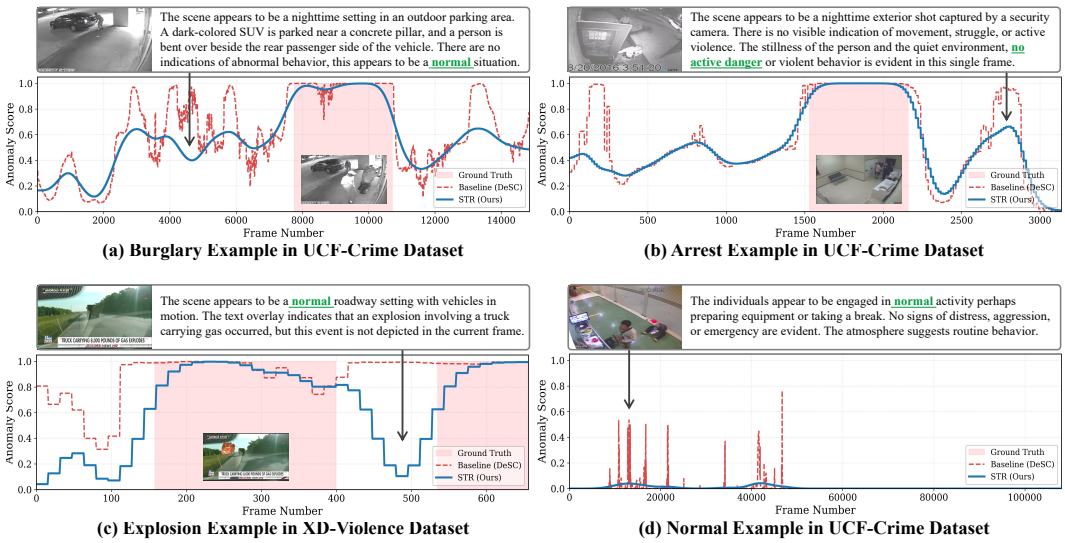
Effectiveness of components. Table 6 shows progressive gains from semantic verification, the perception veto, and Gaussian correction, reaching 90.80% AUC on UCF-Crime and 89.08% AP on XD-Violence. We further permute the semantic scores across the same anchors while preserving their score distribution, veto, and Gaussian correction. Performance decreases to 90.21% AUC and 88.67% AP, indicating that context-score correspondence contributes to the observed gains.

Recall at a fixed operating point. At an FPR of 5%, STR raises frame-level recall from 37.33% to 43.13% on UCF-Crime and from 68.03% to 74.95% on XD-Violence. The recall gains are consistent with the overall AUC and AP improvements.

Aggregation of overlapping corrections. Min aggregation obtains 90.80% AUC on UCF-Crime and 89.08% AP on XD-Violence, compared with 90.34% AUC and 88.60% AP for mean aggregation and 89.33% AUC and 86.26% AP for max aggregation. The min operator retains the strongest local correction without accumulating attenuation from overlapping anchors.

Veto threshold and Gaussian standard deviation. Performance remains stable across veto quantiles from 98.0% to 99.8% and over moderate Gaussian widths. The shared 99.5% veto quantile and Gaussian width of 2 snippets both lie within these stable regions. Restricting the veto to only maximum-score locations at the 100% quantile reduces performance on both datasets.

Robustness of anchor construction. Using a threshold coefficient of 0.5, UCF-Crime reaches 90.80% AUC, compared with 90.46% and 90.10% for coefficients of 0 and 1.0. The 95th-percentile,



(a) Burglary Example in UCF-Crime Dataset **(b) Arrest Example in UCF-Crime Dataset**

(c) Explosion Example in XD-Violence Dataset **(d) Normal Example in UCF-Crime Dataset**

Figure 3: Qualitative results comparing frozen baseline and STR scores. Pink regions denote ground-truth anomalies, and text bubbles show semantic outputs at selected anchors. STR suppresses locally normal responses while retaining dominant anomaly responses.

Table 6: Component ablation with adaptive anchor mining fixed.

Semantic Verification	Perception Veto	Gaussian Correction	UCF (AUC %)	XD (AP %)
✗	✗	✗	89.37	87.18
✓	✗	✗	89.93	87.94
✓	✓	✗	90.35	88.63
✓	✓	✓	**90.80**	**89.08**

97.5th-percentile, and median-plus-MAD boundaries obtain 89.19%, 89.32%, and 89.64% AUC, respectively, supporting robustness to alternative anchor-construction rules.

4.7 Qualitative Analysis

Figure 3 illustrates local semantic correction and response preservation. Figures 3(a) and 3(b) show elevated baseline responses on locally benign contexts, which STR reduces after semantic verification. A residual post-event response remains in Figure 3(b), showing that local correction does not remove every inherited false alarm. In Figure 3(c), the verified context contains no visible explosion in the normal interval between two anomaly events, and STR lowers this interval while preserving both anomaly responses. Figure 3(d) further shows isolated spike suppression in a long normal video. Across the anomalous cases, the dominant responses remain after correction, with only mild boundary attenuation from Gaussian propagation.

5 Conclusion

STR provides a training-free framework for selective semantic verification of frozen WSVAD detectors by combining adaptive query allocation with bounded local score correction. Experiments across three benchmarks, multiple frozen detectors, visual features, and vision-language models demonstrate consistent improvements without retraining the underlying detector. Matched-budget and dense-verification comparisons support the roles of selective query allocation and bounded local correction, while the semantic-correspondence control further shows the value of context-aligned semantic evidence. The current framework operates on candidates surfaced by the frozen detector, and extending semantic verification to a broader candidate space remains a future direction.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and manuscript polishing. All AI-assisted text was reviewed and revised by the authors, who verified the technical claims, citations, and experimental descriptions against the original sources and experimental records. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Implementation details and shared hyperparameters are provided in Section 4.2. The appendix provides the exact prompts, inference pseudocode, structured-output audit, complete cross-model results, robustness analyses, and additional qualitative examples. Anonymous code is linked in the abstract.

REFERENCES

Anusha Acharya, Hitesh Sapkota, Qi Yu, and Xumin Liu. The road less seen: Segment exploration for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14147–14156, June 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

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 33rd ACM International Conference on Multimedia*, pp. 592–601, 2025.

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*, pp. 18319–18329, 2024.

Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5549–5559, 2023a.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 387–395, 2023b.

Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18793–18803, 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*, pp. 14009–14018, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogvlm2: Visual language models for image and video understanding, 2024. URL https://arxiv.org/abs/2408.16500.

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

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

Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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*, pp. 8022–8031, 2023.

Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and Francois Bremond. Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24265–24274, 2025.

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

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

Shengyang Sun, Jiashen Hua, Junyi Feng, Dongxu Wei, Baisheng Lai, and Xiaojin Gong. Tdsd: Text-driven scene-decoupled weakly supervised video anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5055–5064, 2024.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4975–4986, 2021.

Guohua Wang, Shengping Song, Wuchun He, and Yongsen Zheng. Cmhkf: Cross-modality heterogeneous knowledge fusion for weakly supervised video anomaly detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31594–31607, 2025.

Yu Wang and Shiwei Chen. Learning event completeness for weakly supervised video anomaly detection. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62505–62517. PMLR, 13–19 Jul 2025.

Yu Wang and Shengjie Zhao. Weakly supervised video anomaly detection with anomaly-connected components and intention reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28545–28556, June 2026.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pp. 322–339. Springer, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9301–9310, 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*, volume 38, pp. 6074–6082, 2024b.

Rong Xu, Runqi Wang, Yingjun Zhang, Tao Tao, Xiaomeng Li, and Liping Jing. Tlma: Mitigating the impact of weakly labeled information for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35597–35606, June 2026.

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

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

Wenti Yin, Huaxin Zhang, Xiang Wang, Yuqing Lu, Yicheng Zhang, Bingquan Gong, Jialong Zuo, Li Yu, Changxin Gao, and Nong Sang. Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12027–12035, 2026.

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*, pp. 18527–18536, 2024.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13843–13853, 2025.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Yaxin Zhao, Yang Wang, Wenya Guo, Sihan Xu, Xiangrui Cai, Xi Lin, Ying Zhang, and Xiaojie Yuan. Learning from noisy supervision: A denoising-debiasing framework for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21326–21335, June 2026.

Hantao Zheng, Ning Han, Yawen Zeng, and Hao Chen. Decoupled sensitivity-consistency learning for weakly supervised video anomaly detection, 2026. URL `https://arxiv.org/abs/2603.19780`.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3769–3777, 2023.

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

Yiyan Zhu, Menghao Zhang, Haifeng Sun, Pengfei Ren, Xianao Chu, Chenye Xu, Hong Tan, Jinghan Wang, Qi Qi, and Jingyu Wang. Alert-clip: Abnormality-aware latent-enhanced representation tuning of clip for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35545–35554, June 2026.

A SUPPLEMENTARY OVERVIEW

This supplementary material expands the reproducibility and empirical analysis of Semantic-Temporal Reprojection (STR). Section B provides the exact prompts, the complete semantic-verification chain, and a structured-output audit. Section C specifies the inference procedure in algorithmic form. Section D reports transfer results across visual-description models, analyzes sensitivity to the veto threshold and Gaussian width, evaluates semantic-score correspondence, and compares alternative temporal-boundary rules. Section E presents additional qualitative evidence on UCF-Crime and XD-Violence.

B PROMPT DESIGN AND COMPLETE VERIFICATION CHAIN

The semantic branch uses a visual-description model followed by a structured semantic verifier. The first stage summarizes observable activity from a sampled multi-frame context. The second stage maps this description to a bounded semantic score and a short rationale under a fixed rule rubric. The same prompts and score intervals are used unchanged across UCF-Crime, XD-Violence, and MSAD to maintain a consistent semantic-verification protocol.

B.1 THE VISUAL DESCRIPTION STAGE: QWEN3-VL PROMPT

Qwen3-VL-8B (Bai et al., 2025) receives the 16 frames sampled from the 32-frame source context specified in the main paper. This stage uses no system prompt. A user-level instruction directs the model to describe visible interactions and to distinguish active anomalous events from visually dynamic but ordinary activities.

- **Input:** The input consists of the sampled multi-frame visual context and the explicit user prompt.



User Prompt:

Describe the scene in detail. Focus on whether there is any crime, violence, fire, explosion, or accident. If people are just running, dancing, or gathering normally, state it clearly.

The generated description is denoted by $\mathcal{D}(t_a)$ for the selected anchor t_a. The baseline anomaly score is not provided to the visual-description model.

B.2 THE STRUCTURED VERIFICATION STAGE: LLAMA-3.1 PROMPT

Llama-3.1-8B-Instruct (Grattafiori et al., 2024) follows a fixed system prompt that defines the semantic score bands and the required output structure. The generated visual description is then supplied through a separate user prompt. The baseline score is not included in either prompt and enters the pipeline only through anchor selection and the subsequent perception veto.

- **System Prompt:** Defines the semantic rubric and the expected JSON output format, restricting the verifier to a bounded numeric response and a short explanation.

System Prompt:

You are a video anomaly detection verifier.

Logic:

1. If the description confirms NORMAL activities such as running for exercise, dancing, fixing a car, shopping, playing, gathering, walking, or busy street traffic, output a LOW score between 0.0 and 0.2.

2. If the description is AMBIGUOUS or UNKNOWN, output a MODERATE score between 0.4 and 0.6.

3. If the description confirms a REAL ANOMALY such as fighting, shooting, a riot, abuse, a car accident, or an explosion, output a HIGH score between 0.8 and 1.0.

Format:

Return ONLY a JSON object formatted as
{ "refined_score": float,
"reason": "short explanation" }.

- **User Prompt:** Supplies the visual description to the semantic verifier. The returned *refined_score* is parsed as $\mathcal{S}_{logic}(t_a)$, while *reason* provides the human-readable rationale $\mathcal{E}_{reason}(t_a)$.

User Prompt:
Visual Description: "{caption}"
Judge: *LLM begins generation here*

The verifier uses greedy decoding. An invalid, missing, or out-of-range output is assigned zero correction strength and therefore does not modify the baseline score sequence.

B.3 Complete STR Verification Example

A representative context from XD-Violence (Wu et al., 2020) illustrates the complete semantic branch and its connection to local score correction. The example is taken from video *v=m8EkFsaGPzU__#1_label_B4-0-0.mp4* around frame 3488. The baseline score is used for anchor selection and veto activation, but is not supplied to either semantic model.

- **Step 1: Multi-frame visual description.** The sampled context contains several people interacting with objects in a cluttered outdoor area.

VLM Output $\mathcal{D}(t_a)$:

The scene depicts several individuals in an outdoor area that appears disorganized and cluttered with debris, including plastic bags, containers, and scattered objects. One person is handling something near a wooden structure, while another person is seated nearby and examining an object. No visible fire, explosion, accident, or active violent behavior appears in the sampled context. The overall scene shows people interacting with objects in a cluttered area.

- **Step 2: Structured semantic verification.** The verifier maps the description to the LOW interval and returns both the numeric field and its rationale.

LLM Output $\big(\mathcal{S}_{logic}(t_a), \mathcal{E}_{reason}(t_a)\big)$:

{
"refined_score": 0.0,
"reason": "The description does not indicate violent or anomalous behavior; it shows people interacting with objects in a cluttered area."
}

- **Step 3: Perception-guarded local reprojection.** The parsed output gives $\mathcal{S}_{logic}(t_a) = 0$. When the selected anchor lies below the video-adaptive veto threshold, the correction strength is $\alpha(t_a) = 1 - \mathcal{S}_{logic}(t_a) = 1$. The resulting Gaussian mask propagates this correction to nearby snippets with rapidly decaying weights. When the anchor meets or exceeds the veto threshold, $\alpha(t_a) = 0$, so no correction is generated from that anchor.

B.4 Structured Output Reliability

The output audit covers every selected anchor used in the main UCF-Crime and XD-Violence experiments. The 423 UCF-Crime outputs and 1,553 XD-Violence outputs correspond to average query counts of 1.46 and 1.94 per video, respectively. Table 7 reports parsing, field-completeness, and range checks.

Table 7: Reliability audit over all selected anchors.

Audit item	UCF-Crime	XD-Violence
Total outputs	423	1,553
Valid JSON	423 (100%)	1,553 (100%)
Score field present	423 (100%)	1,553 (100%)
Reason field present	423 (100%)	1,553 (100%)
Out-of-range scores	0 (0%)	0 (0%)

Every cached output satisfies the required interface, and no out-of-range score is observed. The audit confirms that the structured numeric signal is available over the complete selected-anchor sets rather than only on a sampled subset. This audit concerns the fixed outputs used for detection comparisons; runtime measurements regenerate the semantic outputs without caches, as described in the main paper.

C Algorithm Pseudocode

Algorithm 1 provides an algorithmic summary of the complete STR pipeline. The frozen detector first produces the dense temporal score sequence. Adaptive mining then selects a sparse anchor set according to the feature segmentation and local score rules defined in Eqs. (2)–(5) of the main paper. For each anchor, the semantic branch produces a structured score, the perception veto determines whether a correction is generated from that anchor, and local Gaussian reprojection updates the surrounding score sequence. No model parameter is changed during this procedure. All reported experiments use the same values $\beta = 0.5$, $\rho = 99.5\%$, and $\omega = 2$ snippets across all three datasets.

When no anchor is selected, $\mathcal{A}$ is empty and $\mathcal{M}(t) = 1$ at every temporal location. STR then returns the original frozen-detector score sequence. The semantic-query complexity is reduced from T calls under dense verification to K calls under adaptive mining, where $K \ll T$ in the reported experiments. No additional global smoothing is applied after local Gaussian reprojection.

Algorithm 1 Inference procedure of Semantic-Temporal Reprojection

Require: Raw video V, feature sequence $X \in \mathbb{R}^{T\times d}$, frozen detector $\mathcal{F}_{\text{baseline}}$
Visual-description model $\mathcal{F}_{\text{VLM}}$, semantic verifier $\mathcal{F}_{\text{LLM}}$, rule prompt $\mathcal{R}_{\text{safety}}$
Hyperparameters β, ρ, and ω
Ensure: Corrected score sequence $\mathcal{S}_{\text{final}}$

Dense scoring and adaptive anchor mining
$\mathcal{S}_{\text{base}} \leftarrow \mathcal{F}_{\text{baseline}}(X)$
$\mathcal{A} \leftarrow \text{AnchorMining}(X, \mathcal{S}_{\text{base}}, \beta)$
$\theta_{\text{veto}} \leftarrow Q_\rho(\mathcal{S}_{\text{base}})$
$\mathcal{M}(t) \leftarrow 1, \quad t = 1, \ldots, T$

Semantic verification and guarded correction
for each anchor $t_a \in \mathcal{A}$ **do**
  $\mathcal{V}(t_a) \leftarrow \text{SampleFrames}(V, t_a)$
  $\mathcal{D}(t_a) \leftarrow \mathcal{F}_{\text{VLM}}(\mathcal{V}(t_a))$
  $\mathcal{Y}(t_a) \leftarrow \mathcal{F}_{\text{LLM}}\big(\mathcal{D}(t_a) \mid \mathcal{R}_{\text{safety}}\big)$
  if $\mathcal{Y}(t_a)$ is invalid, incomplete, or out of range **then**
    $\alpha(t_a) \leftarrow 0$
    continue
  end if
  $S_{\text{logic}}(t_a) \leftarrow \mathcal{Y}(t_a)[\texttt{refined_score}]$
  if $S_{\text{base}}(t_a) \geq \theta_{\text{veto}}$ **then**
    $\alpha(t_a) \leftarrow 0$
  else
    $\alpha(t_a) \leftarrow 1 - S_{\text{logic}}(t_a)$
  end if
  Local Gaussian reprojection
  for $t = 1, \ldots, T$ **do**
    $G_{t_a}(t) \leftarrow 1 - \alpha(t_a) \exp\left(-\dfrac{(t-t_a)^2}{2\omega^2}\right)$
    $\mathcal{M}(t) \leftarrow \min\{\mathcal{M}(t), G_{t_a}(t)\}$
  end for
end for

Final score reprojection
$\mathcal{S}_{\text{final}} \leftarrow \mathcal{S}_{\text{base}} \odot \mathcal{M}$
return $\mathcal{S}_{\text{final}}$

D Additional Quantitative Results

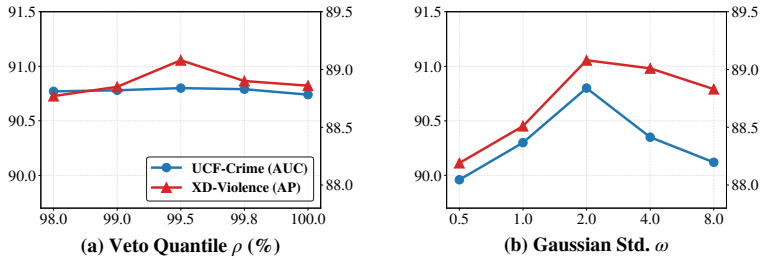
The following experiments expand four findings summarized in the main paper. The first examines transfer across visual-description backbones. The second analyzes sensitivity to the veto threshold and Gaussian width. The third tests whether context-aligned semantic scores contribute beyond generic local attenuation. The fourth examines robustness to alternative temporal-boundary rules.

D.1 Transfer across Visual-Description Models

Table 8 compares three visual-description models while keeping Llama-3.1, the prompt, the selected anchors, and all score-correction settings unchanged. This comparison isolates the effect of the visual-description backbone by keeping the structured verification and correction stages fixed.

Table 8: Transfer across visual-description models.

Visual-description model	UCF (AUC %)	XD (AP %)
Qwen3-VL-8B	**90.80**	**89.08**
LLaVA-NeXT-Video-7B	90.61	88.83
CogVLM2	90.52	88.78



(a) Veto Quantile ρ (%) **(b) Gaussian Std. ω**

Figure 4: Sensitivity to the veto quantile ρ and Gaussian standard deviation ω.

Qwen3-VL-8B reaches 90.80% AUC on UCF-Crime and 89.08% AP on XD-Violence. Replacing it with LLaVA-NeXT-Video-7B (Zhang et al., 2024) changes the two results to 90.61% and 88.83%, while CogVLM2 (Hong et al., 2024) obtains 90.52% and 88.78%. The maximum variation is therefore 0.28 AUC on UCF-Crime and 0.30 AP on XD-Violence. The small variation across these backbones supports cross-model transfer.

D.2 Sensitivity to Veto Threshold and Gaussian Width

Figure 4 visualizes the sensitivity summarized in Section 4.6 of the main paper. Performance remains stable for veto quantiles from 98.0% to 99.8% and over moderate Gaussian widths, with the shared setting $\rho = 99.5\%$ and $\omega = 2$ snippets lying within these stable regions. Restricting the veto to the 100% quantile reduces performance on both datasets.

D.3 Semantic Correspondence Control

To test whether the gains can be explained by anchor mining and generic local attenuation alone, we permute the verified semantic scores among selected anchors using a fixed permutation while retaining the anchor set, the semantic-score values, the perception veto, and the Gaussian correction pipeline. This control preserves the available semantic-score distribution but breaks its correspondence to the local visual contexts.

As shown in Table 9, shuffled semantic correspondence reduces UCF-Crime AUC from 90.80% to 90.21% and XD-Violence AP from 89.08% to 88.67%. The decrease on both datasets indicates that context-aligned semantic evidence provides additional gains beyond anchor selection and generic local score reduction.

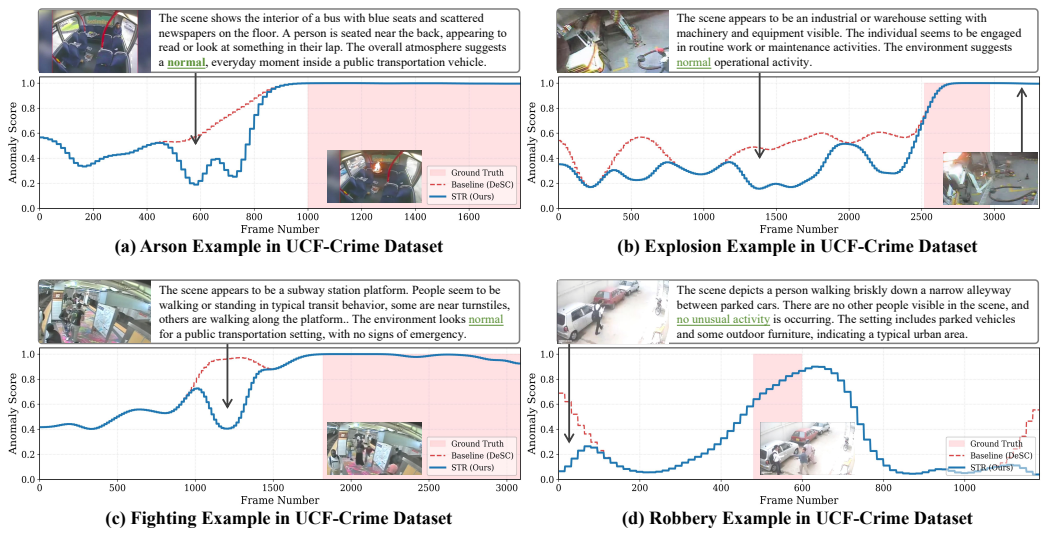
Table 9: Semantic correspondence control with shuffled semantic scores.

Setting	UCF (AUC %)	XD (AP %)
Shuffled semantic scores	90.21	88.67
Full STR	**90.80**	**89.08**

D.4 Robustness of Temporal-Boundary Construction

The feature-distance threshold determines only where the temporal sequence is divided into local contexts. It does not assign anomaly labels or directly modify $\mathcal{S}_{base}$. Table 10 compares the adopted $\mu_D + 3\sigma_D$ boundary with two percentile thresholds and a median-plus-MAD alternative.

All four rules yield UCF-Crime AUC values between 89.19% and 90.80%, showing that anchor construction remains functional under rank-based and robust-statistics alternatives. The adopted $\mu_D+3\sigma_D$ rule is used throughout the reported experiments, while the alternatives serve as robustness checks of the anchor construction. Because these alternatives do not rely on a Gaussian interpretation of D_t, the comparison also shows that STR does not require the three-sigma boundary to represent a probabilistic normality model.



(a) Arson Example in UCF-Crime Dataset **(b) Explosion Example in UCF-Crime Dataset**

(c) Fighting Example in UCF-Crime Dataset **(d) Robbery Example in UCF-Crime Dataset**

Figure 5: Additional qualitative results on UCF-Crime. Red dashed curves denote frozen baseline scores, blue solid curves denote STR scores, shaded regions indicate annotated anomaly intervals, and text boxes show visual descriptions at selected anchors.

Table 10: Alternative temporal-boundary rules on UCF-Crime.

Boundary rule	UCF (AUC %)	Recall@FPR=5% (%)
$\mu_D + 3\sigma_D$	**90.80**	**43.13**
95th percentile	89.19	41.12
97.5th percentile	89.32	39.27
Median + MAD	89.64	39.33

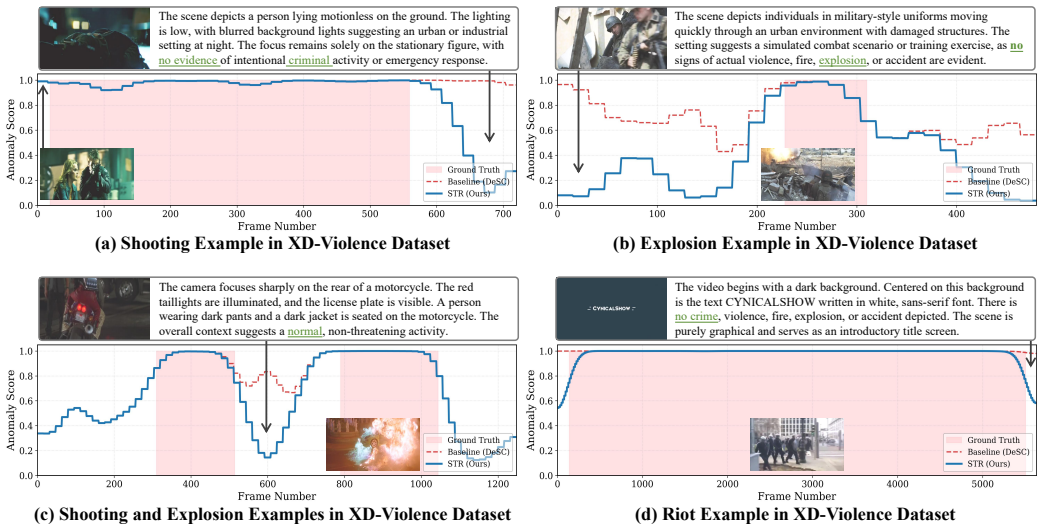
E Additional Qualitative Analysis

The qualitative figures compare frozen DeSC scores (Zheng et al., 2026) with STR scores on UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020). The text boxes reproduce the original visual descriptions generated at selected anchors. Together, the examples illustrate three behaviors: correction of visually ambiguous normal responses, separation of nearby normal and anomalous intervals, and preservation of dominant anomaly responses.

E.1 Correction of Visually Ambiguous Normal Responses

Figure 5(a) shows routine activity inside a bus, where a seated person and scattered objects produce a strong frozen baseline response despite the absence of an active anomaly. Figure 5(b) contains industrial machinery and maintenance activity, while Figure 5(c) shows ordinary movement on a subway platform. In both cases, motion, crowd structure, or unusual background content coincides with elevated responses from the frozen detector. The generated descriptions identify routine activity and STR lowers the corresponding responses around the selected anchors.

Figure 5(d) shows a short normal interval before the annotated robbery event, where a person walks between parked vehicles. The frozen detector produces an elevated response during this motion, whereas semantic verification identifies no crime, accident, or emergency. The corrected curve follows the original temporal structure while reducing the local response amplitude around the verified anchor. The XD-Violence examples in Figures 6(b) and 6(d) show the same localized correction pattern for a simulated combat scene and a graphical title screen.



(a) Shooting Example in XD-Violence Dataset **(b) Explosion Example in XD-Violence Dataset**

(c) Shooting and Explosion Examples in XD-Violence Dataset **(d) Riot Example in XD-Violence Dataset**

Figure 6: Additional qualitative results on XD-Violence. Red dashed curves denote frozen baseline scores, blue solid curves denote STR scores, shaded regions indicate annotated anomaly intervals, and text boxes show visual descriptions at selected anchors.

E.2 Temporal Separation Across Event Boundaries

Figure 6(a) shows a high response that persists after the active event has ended. The selected post-event context contains no ongoing violent action, and STR reduces the residual response without globally smoothing the complete sequence. Figure 6(c) contains two annotated anomaly periods with a normal temporal interval separating them. The frozen detector remains elevated throughout this gap, while the local semantic correction lowers the intermediate response and preserves the dominant responses in both surrounding anomaly intervals.

A related pattern appears near the late portion of Figure 5(b). The example retains a residual response outside the corrected neighborhood, showing that Gaussian reprojection remains centered on the verified anchors. This behavior distinguishes local correction from an independent sequence-level smoothing operation.

E.3 Preservation of Dominant Anomaly Responses

Across the anomalous examples, the primary high-score responses inside the annotated intervals remain after correction. This behavior is consistent with the perception veto, which assigns zero correction strength to anchors above the video-adaptive threshold. It is also consistent with the Recall@FPR analysis in the main paper, where STR increases recall at the same 5% false-positive rate. The qualitative curves therefore show that the framework mainly reassesses ambiguous elevated responses instead of uniformly suppressing the complete anomaly-score sequence.

E.4 Scope of Candidate-Based Verification

STR allocates semantic verification to elevated responses selected from a frozen score sequence. Potential anomalies with uniformly low baseline scores may remain outside the anchor set, and false-positive regions without a nearby selected anchor may receive no correction. Residual responses such as the late region in Figure 5(b) illustrate this baseline-dependent and local behavior. These cases reflect the candidate-based scope of STR and motivate broader candidate discovery without reverting to dense semantic verification.