DISCO: DISTILLATION VIA CONFLICT FOR WEAKLY SUPERVISED VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video anomaly detection (VAD) has recently progressed along two promising directions: weakly supervised learning and adaptation of multimodal vision–language models (VLMs). Weakly supervised VAD is cost-effective because it relies only on video-level labels, but the lack of segment-level annotations makes it difficult to localize and disentangle multiple or diverse abnormal events. In contrast, VLMs pretrained on large-scale multimodal data encode rich, general semantic knowledge, yet they are not explicitly optimized for the domain-specific temporal reasoning required for anomaly detection. In this work, we propose a knowledge-sharing framework that transfers generalized semantics from VLMs to a weakly supervised VAD model via knowledge distillation. Empirically, we observe complementary behavior: zero-shot VLM predictions excel on certain anomaly categories (e.g., explosions), whereas weakly supervised models are stronger on human-centric abnormalities (e.g., abuse). To enable selective distillation, we introduce a utility score that jointly models semantic predictions and motion at the regional level. We aggregate this score across training videos with the same event category and distill only high-conflict events. Extensive experiments on two benchmark datasets show that our distillation via conflict (**DisCo**) consistently improves anomaly detection performance over both weakly supervised and tuning-free VLM baselines, while requiring only a lightweight model at inference time.

1 INTRODUCTION

Video anomaly detection (VAD) identifies events in video footage that deviate from normal patterns and is central to real-world applications such as intelligent surveillance Sultani et al. (2018), industrial monitoring, traffic analysis, and automated content analysis Abdalla et al. (2025); Wu et al. (2020). Fully supervised VAD remains impractical in many scenarios as it requires obtaining frame-level temporal annotations for abnormal events in long, untrimmed videos, which is costly. To alleviate this, weakly supervised video anomaly detection (WSVAD) has emerged as a cost-effective alternative that trains models using only video-level labels. VAD is commonly formulated as a Multiple Instance Learning (MIL) problem Dietterich et al. (1997), which is a practical Weakly Supervised Learning paradigm. Each video is treated as a bag of temporal segments. A positive bag corresponds to an abnormal video: video that contains at least one abnormal segment, whereas a negative bag corresponds to a normal video that contains only normal segments. Given we have only bag-level labels in the training data, the goal is to train a model that outputs an anomaly score to each segment of a given input video. For this, the model is then trained to first identify the abnormal segments from positive bags, called the representative segments, and then to maximize the score difference between these segments and the segments of normal video. In this way, the model learns to localize anomalies without segment-level supervision.

Early work on identifying abnormal segments Zhu & Newsam (2019); Sultani et al. (2018) highlighted the importance of motion-aware representations, since anomalies often manifest as irregular or abrupt temporal dynamics. As a result, most WSVAD frameworks Amicantonio et al. (2025); Zhou et al. (2023); Zhang et al. (2023); Tian et al. (2021); Sapkota & Yu (2022); Sapkota et al. (2021) encode each video segment with spatiotemporal features extracted from pretrained 3D convolutional backbones (e.g., C3D Tran et al. (2015) or I3D Carreira & Zisserman (2017)) or enhance ViT-based features Radford et al. (2021) with temporal adapters that explicitly model motion dy-

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.

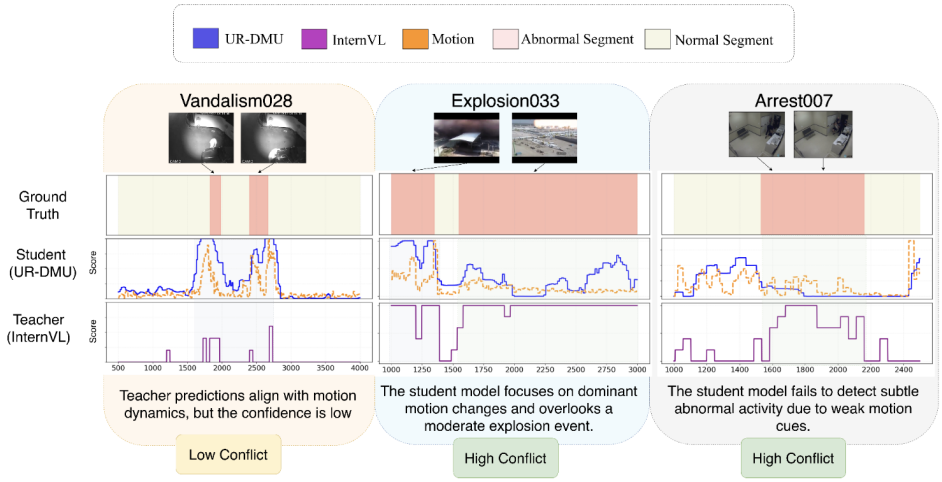


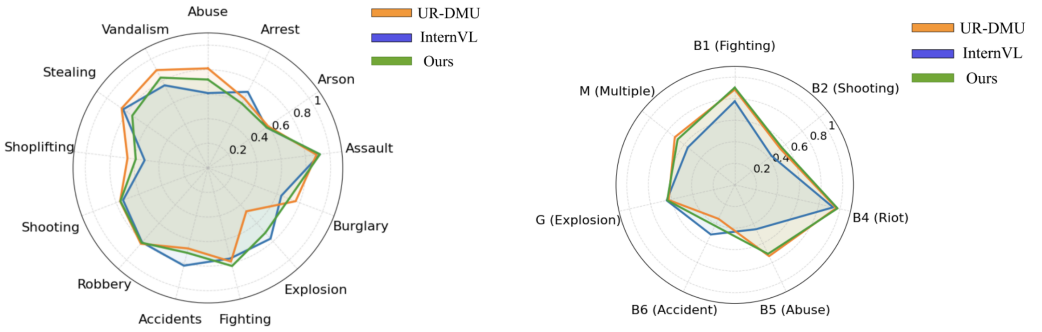
Figure 1: Examples from UCF-Crime showing complementary failure modes: UR-DMU is biased toward high-motion segments and misses subtle anomalies (e.g., Arrest007, Explosion033), whereas InternVL captures semantic events but underperforms in motion-dependent scenarios (e.g., Vandalism028).

namics Wu et al. (2024); Joo et al. (2023). These segment representations are then passed to an MIL-based detector that outputs an anomaly score for each segment. During training, the model typically selects the top-k highest-scoring segments from each positive bag to compute the MIL loss and update parameters Tian et al. (2021). However, the widely used top-k selection strategy can bias training toward segments with drastic motion changes Lv et al. (2023). This bias causes subtle or low-motion anomalies to be overlooked, while high-motion yet semantically normal segments may spuriously trigger false alarms. For instance, in Fig. 1, the baseline model assigns high anomaly scores to segments dominated by strong motion variations, whereas segments with weaker motion receive substantially lower scores.

Several efforts aim to improve upon this limitation of top-k selection, such as enforcing diversity constraints by sampling high-scoring segments from different temporal regions Sapkota & Yu (2022); Amicantonio et al. (2025); Acharya et al. (2026), or adopting uncertainty-aware sampling Lv et al. (2023); Acharya et al. (2026). However, these approaches still depend on prediction scores that are themselves learned primarily from high-motion segments. Consequently, subtle abnormal events that initially receive low scores (e.g., Arrest007 in Fig. 1) are rarely selected during training, and the underlying bias remains largely unresolved.

To address these limitations, it is desirable to introduce an additional supervisory signal to WSVAD models that provides complementary, non-motion-dominated cues. Multimodal vision–language models (VLMs) are a natural candidate: they exhibit strong generalization across a wide range of visual recognition tasks Chen et al. (2024); Li et al. (2024); Zhang et al. (2025) and have recently been explored for VAD Ye et al. (2025); Zhang et al. (2024); Zanella et al. (2024); Cai et al. (2025). Pretrained on large-scale multimodal corpora, VLMs encode rich semantic and contextual knowledge, allowing them to recognize events such as explosions or arrests from scene context even when motion cues are weak (Fig. 1). However, the zero-shot performance of VLMs on VAD remains limited, partly because abnormal events are sparsely represented in their pretraining data Ye et al. (2025); Cai et al. (2026). Moreover, many VLMs Chen et al. (2024); Li et al. (2024) perform video understanding via sparse frame sampling with minimal explicit temporal modeling, which constrains their ability to capture fine-grained motion dynamics that are critical for detecting certain anomalies.

As shown in Fig. 2, a category-wise comparison between the VLM and the WSVAD model on the test sets of both benchmarks reveals clear complementarity. The VLM performs best on anomalies that are recognizable from semantic context (e.g., Explosion), whereas the WSVAD model is stronger on categories that benefit from continuous temporal modeling (e.g., Abuse and Vandalism). Motivated by this observation, we propose a knowledge-sharing framework that leverages the complementary strengths of both models to improve VAD performance.



(a) (b)

Figure 2: Class-wise comparison between InternVL (VLM), UR-DMU (weakly supervised model) and DisCo (ours) on two benchmark datasets: (a) UCF-Crime evaluated using AUROC and (b) XD-Violence evaluated using AP. The models demonstrate complementary strengths across anomaly categories, with performance varying in a class-dependent manner.

A natural way to fuse these complementary strengths is knowledge distillation Hinton et al. (2015). However, naively distilling all teacher predictions can introduce noisy or misleading supervision, since VLMs are unreliable on certain anomaly categories. To address this, we propose a *conflict-aware* knowledge distillation framework for WSVAD with three key components:

- **Utility score estimation** that quantifies utility of VLM predictions for a given temporal region.
- **Utility-based event selection** that aggregates utility scores at the event level and identifies high-conflict categories for distillation.
- **Event-wise knowledge distillation** that selectively transfers semantic supervision to conflict-dominant categories.

Extensive experiments on two benchmark datasets show that our approach consistently outperforms both WSVAD and VLM baselines, while retaining a lightweight single-modality model at inference time.

2 Related Work

Weakly Supervised Learning for VAD: To reduce the high cost of frame-level annotation, Sultani et al. formulated VAD as MIL problem. In this setting, models are trained only on video-level labels, and several works Tian et al. (2021); Zhou et al. (2023); Wu et al. (2024); Joo et al. (2023); Majhi et al. (2025); Li et al. (2022) select the top-k highest-scoring segments from each video to distinguish between abnormal and normal segments. However, the top-k selection strategy provides a biased approximation of abnormal events, as it restricts supervision to a small subset of high-scoring segments that are often dominated by the most salient features Lv et al. (2023) and lack temporal diversity Sapkota & Yu (2022); Acharya et al. (2026). To enhance the robustness of the MIL framework, several prior studies introduce refinements such as adaptive k strategies Sapkota et al. (2021), promoting temporal diversity through non-parametric clustering Sapkota & Yu (2022) or peak-based selection techniques Amicantonio et al. (2025), and exploring segments based on prediction uncertainty Lv et al. (2023); Acharya et al. (2026). Nevertheless, these approaches inherently rely on the anomaly score distribution produced by the same model, whose representations are already shaped by dominant features under weak supervision. Consequently, subtle abnormal events may remain underrepresented during training, necessitating external semantic guidance. Dual-Exploration Acharya et al. (2026) further addresses this limitation by combining the predictions of a multimodal VLM and an MIL model at inference. Although this late-fusion strategy provides complementary semantic information, it requires both models to process every test video, substantially increasing inference cost and limiting its practicality for large-scale real-world surveillance.

Vision–Language Models for VAD: Recently, VLMs have been increasingly adopted for WSVAD because their pretrained representations can provide rich semantic information. VadCLIP Wu et al. (2024), for example, employs frozen CLIP visual and textual encoders, and learnable prompts constructed from predefined anomaly-category names. DSANet Yin et al. (2026) further models diverse normal patterns by selecting low-anomaly-score frames as candidate normal instances and extracting dynamic normal prototypes for feature reconstruction. Nevertheless, both methods remain dependent on anomaly scores learned through top-k selection. More powerful multimodal large language models (MLLMs) Chen et al. (2024); Wang et al. (2025); Li et al. (2024); Zhang et al. (2025) have also been explored for VAD. However, frozen MLLMs may inherit biases from multimodal pretraining and struggle to distinguish anomaly categories that are poorly represented in their latent representations Cai et al. (2026), resulting in suboptimal zero-shot performance. Existing approaches Zhang et al. (2024); Ye et al. (2025); Pu et al. (2024); Cai et al. (2026) often address this limitation through full or parameter-efficient adaptation using additional instruction-tuning or calibration data, introducing additional annotation and computational costs. In contrast, we aim to selectively exploit the category-specific strengths of a frozen MLLM without requiring task-specific fine-tuning.

3 METHOD

We consider a WSVAD setting in which each training video is annotated with a binary anomaly label and an event label, but no segment-level annotations are available. The goal is to learn a model that assigns an anomaly score to each video segment. Following prior WSVAD methods, we formulate this task using MIL and introduce the corresponding notation and objective below.

3.1 PRELIMINARIES

Let $f_\theta(\cdot)$ denote a learnable model trained on a dataset $\mathcal{D} = (V_i, y_i, e_i)_{i=1}^{N}$, where V_i is an untrimmed video, $y_i \in \{0, 1\}$ is its binary video-level label, and e_i denotes its video-level event category. Under the MIL formulation, each video V_i is treated as a bag $\mathcal{B}_i = \{x_{i1}, x_{i2}, \ldots, x_{iL_i}\}$ containing L_i instances (segments). The prediction by the model $f_\theta(\cdot)$ is given by:

$$p_{it} = f_\theta(x_{it}) \tag{1}$$

The video-level prediction $\hat{y}_i$ is obtained by aggregating the top-k segment scores within each bag. The MIL loss is then defined as:

$$\mathcal{L}_{\text{MIL}} = \text{BCE}(y_i, \hat{y}_i), \tag{2}$$

where $\text{BCE}(\cdot, \cdot)$ denotes the binary cross-entropy loss.

Following prior works Wu et al. (2024); Yin et al. (2026); Amicantonio et al. (2025), we use training-video event labels to guide teacher prompts and group videos.

3.2 OVERVIEW OF THE PROPOSED FRAMEWORK

WSVAD models tend to favor regions with large motion changes due to the weakly supervised training paradigm and the top-k selection mechanism Lv et al. (2023); Acharya et al. (2026). As a result, they often fail to detect anomaly events characterized by subtle or moderate motion patterns. To address this limitation, we propose a conflict-aware knowledge sharing framework that selectively transfers semantic knowledge from a multimodal VLM teacher to the VAD model in addition to the weak supervision from the MIL loss.

The overall framework of the proposed method is illustrated in Fig. 3. We adopt a teacher–student architecture, where a pretrained vision–language model, InternVL3-14B Zhu et al. (2025), serves as the semantic teacher, while the motion-centric weakly supervised model UR-DMU Zhou et al. (2023) acts as the student. The student is trained using a combination of the MIL objective and a distillation loss, enabling it to leverage complementary semantic cues from the teacher while preserving its motion-based anomaly detection capability.

3.3 DISTILLATION VIA SEMANTIC-MOTION CONFLICT

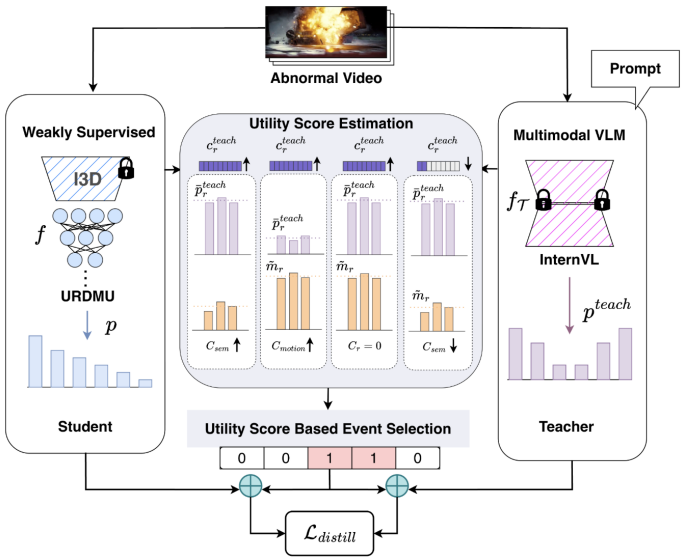


Figure 3: Overview of the proposed **Distillation via Conflict (DisCo)** framework. Abnormal videos and textual prompts are passed to the frozen InternVL teacher to obtain confidence scores p^{teach} in an offline manner. For each temporal region, a utility score C_r is computed by comparing teacher confidence with motion dynamics $\tilde{m}_r$. The utility scores are aggregated at the anomaly-category level, and knowledge distillation is applied only to high-conflict categories using the $\mathcal{L}_{\text{distill}}$ loss.

The student model benefits from knowledge distillation only when the teacher provides complementary information that is not already captured by motion-based supervision. In other words, effective distillation should maximize informative knowledge transfer rather than reinforce redundant signals. As observed in Figures 1 and 4, the student model's anomaly predictions for abnormal segments are highly correlated with motion magnitude. Motion magnitude represents the degree of temporal feature change between adjacent video segments. Let $\mathbf{f}_t \in \mathbb{R}^d$ denote the I3D feature representation of segment t. We quantify motion magnitude as the ℓ_2 norm of the difference between adjacent segment features:

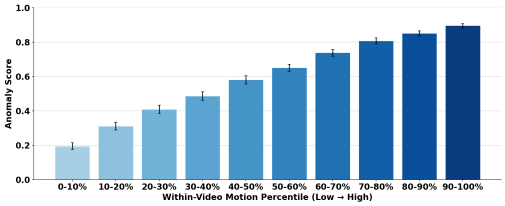


Figure 4: Mean anomaly score of the trained UR-DMU model across within-video percentiles of adjacent I3D feature change on UCF-Crime. Error bars denote 95% confidence intervals.

$$m_t = \|\mathbf{f}_t - \mathbf{f}_{t-1}\|_2 \tag{3}$$

When the semantic prediction of the VLM deviates from motion magnitude, this discrepancy ("conflict") indicates potential complementary information. We therefore quantify the disagreement between semantic and motion signals as the difference between their normalized scores:

$$\Delta_r = \bar{p}_r^{teach} - \tilde{m}_r, \tag{4}$$

where $\bar{p}_r^{teach}$ and $\tilde{m}_r$ denote the semantic and motion scores, respectively, each normalized across all segments in a video and aggregated over a temporal region r comprising a few consecutive segments.

Multimodal VLM Inference Teacher predictions are obtained through training-free inference using InternVL3. Each video is partitioned into non-overlapping temporal segments, from which representative frames are uniformly sampled and provided to the VLM along with textual prompts to generate segment-level anomaly predictions. Each inference produces a binary output (0 or 1), where 1 indicates the presence of an abnormal event. To quantify prediction confidence, multiple

inference runs are performed for each segment. Knowledge from the VLM is most valuable when its predictions conflict with the motion-dominated cues learned by the student under the MIL objective, as such disagreements may reveal anomaly categories that are difficult to capture from motion alone. Moreover, VLM agreement across repeated stochastic generations has been shown to provide a useful signal for output aggregation and uncertainty estimation Wang et al. (2022); Manakul et al. (2023). Thus, a VLM's *utility score* should depend on the disagreement score Δ_r, and the teacher reliability score. Teacher predictions, together with pretrained I3D feature representations, are used to estimate event-wise utility scores, as described in the next section. For student training, the distillation loss is selectively applied to conflict-dominant events and only for abnormal videos. We do not require VLM during inference; only the student model is used to predict anomaly scores.

Teacher Utility for Semantic-dominant Disagreement When semantic evidence exceeds motion evidence ($\Delta_r > 0$), this suggests a potential low-motion anomaly that may be overlooked by the student model. If the teacher predicts this region as abnormal with high confidence and consistency, its supervision is more trustworthy. Accordingly, the magnitude of this disagreement is weighted by the teacher's reliability. This behavior is illustrated in Fig. 1, where the abnormal segments in Explosion033 and Arrest007 exhibit relatively low motion, yet the VLM identifies these segments as abnormal with high confidence, resulting in a higher semantic utility:

$$C_{\text{sem}} = \max(0, \Delta_r)\, p_r^{teach} c_r^{teach} \tag{5}$$

where p_r^{teach} is teacher confidence measured as the mean prediction score, and $c_r^{teach} = 1 - \sigma_r^{teach}$ denotes the teacher's consistency score, estimated as the standard deviation σ_r^{teach}, across multiple inference runs.

Teacher Utility for Motion-dominant Disagreement Conversely, when motion evidence exceeds semantic evidence ($\Delta_r < 0$), this indicates motion-dominant disagreement, where strong motion signals lack semantic support. Such regions may correspond to spurious motion spikes or non-anomalous high-motion events (e.g., scene changes). To suppress the influence of such regions, we define the motion suppression term as:

$$C_{\text{motion}} = \max(0, -\Delta_r)\,(1 - p_r^{teach}) c_r^{teach} \tag{6}$$

where $(1 - p_r^{teach})$ reflects the teacher's confidence that the region is normal. Finally, the overall regional utility score is defined as:

$$\mathcal{C}_r = C_{\text{sem}} + \beta\, C_{\text{motion}} \tag{7}$$

where β regulates the relative importance of motion suppression in the utility estimation.

3.4 Utility-Based Event Selection

For knowledge distillation, our objective is to identify training samples for which the VLM provides the most useful supervision. Let $\mathcal{R} = \{r_1, \ldots, r_N\}$ denote the set of all candidate temporal regions in the training set, where N is the total number of regions. A straightforward strategy is to select the K regions with the highest utility scores:

$$\mathcal{A}_K = \underset{\mathcal{A} \subseteq \mathcal{R}:\, |\mathcal{A}| = K}{\arg\max} \frac{1}{K} \sum_{r \in \mathcal{A}} C_r \tag{8}$$

which maximizes the empirical average utility score among all subsets of size K. However, the observed utility scores are inherently noisy due to variability in the video data and uncertainty in VLM predictions arising from stochastic inference. As a result, regions with the highest observed utility scores may correspond to noisy outliers rather than genuine semantic–motion discrepancies.

To obtain a more robust estimate, we assume that regions sharing similar semantics are drawn from a common event-conditioned distribution. Aggregating utility scores within each semantic group can therefore reduce region-level variability. Motivated by the class-wise complementarity between the VLM teacher and the weakly supervised student illustrated in Fig. 2, we define semantic groups using video-level anomaly events (e.g., *Explosion*, *Arrest*). Let $e(r)$ denote the event label associated with region r, inherited from its video's anomaly category. We model the observed utility score as:

$$C_r = \mu_{e(r)} + \epsilon_r, \tag{9}$$

where $\mu_{e(r)}$ denotes the expected utility score of event $e(r)$, and ϵ_r is a zero-mean random variable capturing region-level and VLM-inference noise.

Let $\mathcal{E}$ denote the set of anomaly events and $\mathcal{R}_e \subseteq \mathcal{R}$ the set of regions belonging to event e. We estimate the event-level utility score as:

$$\widehat{C}_e = \frac{1}{|\mathcal{R}_e|} \sum_{r \in \mathcal{R}_e} C_r. \tag{10}$$

Remark 1 (Stability of event-wise aggregation). Assuming independent zero-mean regional noise with variance σ_e^2, the estimator in Eq. equation 10 satisfies:

$$\mathbb{E}[\widehat{C}_e] = \mu_e, \qquad \mathrm{Var}(\widehat{C}_e) = \frac{\sigma_e^2}{|\mathcal{R}_e|}. \tag{11}$$

Therefore, event-wise aggregation reduces the estimation variance from σ_e^2 for an individual region to $\sigma_e^2/|\mathcal{R}_e|$, yielding a more stable estimate of the underlying utility score.

We then perform selection at the event level rather than the region level. The selected event set is defined as

$$\mathcal{S} = \left\{ e \in \mathcal{E} \mid \widehat{C}_e > \gamma \right\}, \tag{12}$$

where γ is determined adaptively using Otsu thresholding Otsu (1975) over the event-level utility scores. Knowledge distillation is then applied only to videos whose event labels belong to $\mathcal{S}$. Empirically, we find that this event-based selection strategy consistently outperforms direct region-level selection based on the top utility scores, which we discuss in later section.

3.5 Training Objective

Given selected event set $\mathcal{S}$, we perform event-aware knowledge distillation to align the student's segment-level predictions with the teacher's confidence using mean squared error.

$$\mathcal{L}_{\text{distill}} = \mathbf{1}(y_i = 1) \cdot \mathbf{1}(e_i \in \mathcal{S}) \cdot \frac{1}{L_i} \sum_{t=1}^{L_i} \left\| p_{it} - p_{it}^{teach} \right\|_2^2 \tag{13}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function, which equals 1 when $e_i \in \mathcal{S}$ and 0 otherwise for abnormal videos.

The overall training objective combines the MIL loss, additional baseline loss components Zhou et al. (2023), collectively denoted by $\mathcal{L}_{\text{aux}}$, together with the proposed distillation loss:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{MIL}} + \mathcal{L}_{\text{aux}} + \lambda\, \mathcal{L}_{\text{distill}}, \tag{14}$$

where λ is a hyperparameter that controls the trade-off between $\mathcal{L}_{\text{MIL}}$ and $\mathcal{L}_{\text{distill}}$.

4 Experiments

We evaluate our method on two benchmark datasets: UCF-Crime Sultani et al. (2018) and XD-Violence Wu et al. (2020). UCF-Crime has 1,900 untrimmed videos across 13 anomaly categories, with a train-test split of 1,610 and 290 videos, respectively. XD-Violence comprises 4,754 videos across 6 abnormal event types, with a train-test split of 3,954 and 800 videos. In both datasets, the train set has video-level binary and event labels; the videos in the test set have frame-level annotations used only for evaluation.

4.1 Evaluation Metric

Motivated by the XD-Violence evaluation protocol Wu et al. (2020) and recent findings from Dual-Exploration Acharya et al. (2026), we use **Average Precision (AP)** as the primary metric on both datasets. XD-Violence adopts AP because Area Under the Receiver Operating Characteristic Curve (AUROC) may overstate detection performance when abnormal samples constitute only a small proportion of the data. Dual-Exploration further demonstrates this concern on UCF-Crime, showing

Table 1: Performance comparison on UCF-Crime and XD-Violence. R@1, R@2, and R@3 denote recall at FPRs of 1%, 2%, and 3%, respectively. * denotes performance obtained without adapter tuning to ensure a fair comparison.

			UCF-Crime (%)				XD-Violence (%)			
Mode	**Model**	**Encoder**	**AP**	**R@1**	**R@2**	**R@3**	**AP**	**R@1**	**R@2**	**R@3**
Training-free	EventVAD-7B Shao et al. (2025)	ViT	–	–	–	–	64.04	–	–	–
	InternVL3-14B Zhu et al. (2025)	ViT	29.50	**30.1**	30.1	**41.8**	69.85	**63.78**	**63.78**	**63.78**
Weakly Supervised (Multimodal)	PEL4VAD Pu et al. (2024)	I3D+Text	33.99	–	–	–	85.59	–	–	–
	VadCLIP Wu et al. (2024)	ViT	33.55	10.09	15.5	21.7	84.50	41.37	51.9	58.65
	DSANet Yin et al. (2026)	ViT	37.41	–	–	–	82.88*	–	–	–
	Dual-Exploration Acharya et al. (2026)	I3D+ViT	38.33	17.3	26.3	33.6	84.58	33.98	48.54	58.91
Weakly Supervised (RGB-only)	MIL Sultani et al. (2018)	I3D	25.03	–	–	–	75.68	–	–	–
	RTFM Tian et al. (2021)	I3D	29.46	–	–	–	77.81	–	–	–
	BNSVP Sapkota & Yu (2022)	I3D	30.68	–	–	–	–	–	–	–
	UR-DMU Zhou et al. (2023)	I3D	35.48	17.0	17.0	21.2	79.14	26.93	38.67	46.56
	GS-MoE Amicantonio et al. (2025)	I3D	34.08	–	–	–	82.89	–	–	–
	Dual-Exploration Acharya et al. (2026)	I3D	36.42	16.7	23.2	25.6	80.85	30.84	41.12	49.16
	DisCo (Ours)	I3D	**42.29**	22.87	**30.32**	34.57	**85.96**	44.99	55.33	62.22

that a high AUROC does not necessarily translate into high anomaly recall at practically relevant low false-positive rate (FPR) thresholds. This issue is particularly important for highly imbalanced WSVAD benchmarks such as UCF-Crime, where abnormal frames constitute only 7.58% of the test set. We therefore evaluate our method using both AP and recall at low FPR. For reference, AUROC results are also provided in the Appendix.

4.2 Performance Comparison

Table 1 compares DisCo with training-free VLM baselines and state-of-the-art weakly supervised methods. For baselines whose original papers do not report AP, we use the values reported by Dual-Exploration Acharya et al. (2026). The table shows that DisCo improves considerably over UR-DMU with MIL-only training, achieving approximately +6.8 AP gains on both datasets. Moreover, DisCo outperforms Dual-Exploration (which uses RGB-only features for inference), that learns by exploring diverse and ambiguous abnormal segments during training instead of relying on top-k selection. This suggests that segment exploration guided solely by the motion view (dominant temporal feature changes) is insufficient, whereas the complementary semantic guidance provided by the VLM teacher helps identify informative events that are overlooked by motion cues alone. Our method also outperforms the multimodal variant of Dual-Exploration, which ensembles UR-DMU and InternVL predictions for all events at inference only, underscoring the benefit of selective distillation during training. On UCF-Crime, DisCo further outperforms representative multimodal baselines such as VadCLIP and DSANet. For XD-Violence, DSANet reports improved performance by fine-tuning the VLM text encoder (Yin et al. (2026); Table 6), and shows a 4.07% improvement over the variant without text encoder tuning. To ensure a fair comparison with DisCo and other prompt-based methods such as VadCLIP, which do not involve fine-tuning, we report the result of the variant without text encoder tuning. Under this setting, DisCo significantly outperforms these multimodal baselines.

4.3 Ablation Study

Effectiveness of Selective Distillation: Table 2 compares vanilla knowledge distillation (transferring VLM knowledge for all abnormal events), with our conflict-aware strategy, which selectively distills knowledge based on semantic-motion conflict criteria. Vanilla distillation improves performance over the baseline on both datasets, indicating that the VLM provides useful semantic priors for anomaly detection.However, it achieves lower performance than our method because the VLM struggles with anomalies that are not visually evident in static frames but instead depend heavily on motion dynamics. In fact, when categories are selected using alternative criteria that yield different event sets

Table 2: Ablation of selective event-wise knowledge distillation.

Method	**Distilled Events**	**UCF AP (%)**	**XD AP (%)**
Baseline	None	35.48	79.14
Vanilla KD	All	35.94	82.50
Ours	**Conflicting**	**42.29**	**85.96**

than ours for distillation, performance on UCF-Crime drops below that of the baseline (discussed later in Table 6).In contrast, our conflict-based strategy preserves the motion-based representations learned by the weakly supervised model while selectively transferring semantic knowledge from the VLM only for events where it provides complementary information. Specifically, our method selects 6 of the 13 anomaly categories for distillation: Explosion, Arrest, Assault, Fighting, Robbery, and Shooting. This selective distillation improves AP from 35.48 to 42.29 compared with the baseline model. For XD-Violence, we treat videos containing multiple abnormal events as a separate category, resulting in seven anomaly classes. Our conflict-based strategy selects 5 of the 7 classes for distillation while excluding Abuse and Riot.

Effectiveness of Event-Level Conflict Aggregation: Table 3 compares segment-, video-, and event-level granularities for aggregating semantic–motion conflict. Segment-level aggregation follows Eq. 8, distilling the top 10% conflicting segments across 504 videos, which underperforms the baseline (31.77%), as it is sensitive to outliers. Video-level aggregation instead averages conflict over all regions within a video and ranks videos, selecting 139 videos via Otsu thresholding and improving over the baseline (38.36%). Notably, 108 of these videos overlap with those selected by event-level aggregation, which in total selects 338 videos and achieves the best performance (42.29%). To verify that this gain is not simply due to distilling more videos, we construct a controlled variant (Video Top-K) that uses the same video-level aggregation but selects the top 338 videos by utility score, matching the distillation budget of Event-level aggregation. This variant exhibits lower performance, suggesting that aggregating utility score over semantically related videos (events) provides a more reliable criterion for selecting distillation samples, which is consistent with the variance-reduction argument in Remark 1.

Table 3: Effect of utility score aggregation granularity on the performance, reported on UCF-Crime.

Granularity	# Videos	AP (%)	R@1 (%)	R@2 (%)	R@3 (%)
Segment	504	31.77	9.82	15.78	23.62
Video	139	38.36	15.94	25.20	31.91
Video (Top-K)	338	34.48	13.90	21.19	26.69
Event	**338**	**42.29**	**22.87**	**30.32**	**34.57**

Performance without Event-Label Supervision: Table 4 evaluates DisCo under two settings with reduced event-level supervision. In the first setting, we remove the event from the teacher prompt while retaining the event labels for the utility score aggregation. In the second, fully event-agnostic setting, we remove event information from both the teacher prompt and the aggregation procedure. Specifically, for each abnormal video, the VLM generates a short textual description for each segment. We encode each description using the pretrained all-mpnet-base-v2 SentenceTransformer Reimers & Gurevych (2019); Song et al. (2020) and obtain a video-level representation as a abnormality confidence-weighted average of the segment embeddings. The resulting representations are reduced to 32 dimensions and grouped into clusters using agglomerative clustering Ward Jr (1963). Our fully event-agnostic variant improves over the non-distilled baseline by 3.79 AP points and also surpasses the strongest WSVAD baseline (Dual-Exploration), under comparable supervision, i.e., without event-label supervision, while retaining VLM guidance. Additional results and implementation details are provided in the Appendix.

Table 4: Effect of reduced event-label supervision on UCF-Crime.

Method	AP(%)	R@1(%)	R@2(%)	R@3(%)
Baseline	35.48	17.0	17.0	21.2
DisCo	**42.29**	**22.87**	**30.32**	**34.57**
w/o class-guided prompt	39.62	17.63	26.92	33.75
w/o event-label supervision	39.27	16.13	25.45	32.93

5 Conclusion

DisCo selectively distills semantic knowledge from a frozen VLM into a weakly supervised VAD model by identifying anomaly events with high semantic–motion conflict. Event-level aggregation reduces noisy region-wise selection, and experiments on UCF-Crime and XD-Violence show consistent gains over the student backbone, vanilla distillation, and training-free VLM baselines, while retaining an RGB-only student at inference. These results demonstrate that conflict-aware selection is more effective than indiscriminate VLM supervision.

AI USE STATEMENT

In this work, we used generative AI tools to improve writing quality, including grammar correction and general sentence-level editing of text originally written by the authors. Generative AI tools were not used to determine the research direction.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide detailed methodological steps in the main paper, along with additional implementation details, prompts, and related information in the Appendix. We also plan to release the code upon acceptance. The datasets used in our experiments, including both the original videos and the extracted I3D features, are publicly available.

REFERENCES

Moshira Abdalla, Sajid Javed, Muaz Al Radi, Anwaar Ulhaq, and Naoufel Werghi. Video anomaly detection in 10 years: A survey and outlook. *Neural Computing and Applications*, 37(32):26321–26364, 2025.

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, 2026.

Giacomo D' Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Francois Bremond, and Egor Bondarev. Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10275–10285, 2025.

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.

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. *arXiv preprint arXiv:2602.24021*, 2026.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

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*, pp. 12137–12146, 2023.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

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.

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*, pp. 15180–15190, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022.

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*, pp. 26296–26306, 2024.

Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021.

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 François Brémond. 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.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11:285–296, 1975.

Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3212–3221, June 2022.

Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally robust optimization for deep kernel multiple instance learning. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2188–2196. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/sapkota21a.html.

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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. doi: 10.1109/cvpr.2018.00678. URL http://dx.doi.org/10.1109/CVPR.2018.00678.

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.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 322–339. Springer, 2020.

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, 2024.

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.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

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*, pp. 16271–16280, 2023.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*, 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*, volume 37, pp. 3769–3777, 2023.

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

Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.

APPENDIX CONTENTS

A List of Symbols **15**

B Implementation Details **15**

C Additional Results **16**

C.1 AUROC Results 16

C.2 Additional Ablation Study 17

C.2.1 Effectiveness of Semantic–Motion Event Selection 17

C.2.2 Effectiveness of Utility Score Components 17

C.2.3 Effect of Distilling Abnormal and Normal Segments 18

C.2.4 Prompt Sensitivity Analysis 19

C.2.5 Effect of Different Teacher Models 20

C.2.6 Teacher Confidence Estimation 20

C.2.7 Evaluation with Different Student Models 21

C.3 Hyperparameter Sensitivity 21

C.3.1 Effect of hyperparameter β 21

C.3.2 Effect of hyperparameter λ 23

C.3.3 Effect of event selection ranking 23

C.3.4 Effect of number of VLM inference runs 23

C.3.5 Effect of temporal window 24

C.4 Qualitative Results 25

D Theoretical Analysis **25**

D.1 Proof of Remark 1 25

D.2 A Multi-View Interpretation of Semantic Distillation 26

A List of Symbols

Notation and Symbols

$f_{\mathcal{T}}(\cdot)$	Frozen teacher model (VLM)
$f_\theta(\cdot)$	Learnable student model
p^{teach}	Teacher's confidence
p	Student's confidence
$\bar{p}^{teach}$	Normalized teacher's confidence per video
m	Motion magnitude between adjacent segments
$\tilde{m}$	Normalized motion magnitude per video
$\mathbf{f}_t$	I3D feature representation of segment t
Δ_r	Disagreement in semantic representation and motion at region r
C_{sem}	Semantic-dominant utility score
C_{motion}	Motion-dominant utility score
$\mathcal{C}_r$	Total utility score for region r
β	Weight to motion-dominant utility score
$\mathcal{S}$	Set of selected categories
e_i	Event category of video i
γ	Threshold for selecting conflicting events
$\mathcal{C}_e$	Total utility score for event e
$\mathcal{L}_i$	Total number of segments in a video
λ	Weighting parameter in distillation loss
c^{teach}	Teacher's consistency score
σ^{teach}	Standard deviation of teacher predictions across inference runs
AP	Average Precision
$AUROC$	Area Under ROC Curve
FPR	False Positive Rate
$R@1$	Recall@FPR = 1%

B Implementation Details

We use InternVL3-14B Zhu et al. (2025) as the VLM teacher. Each video is divided into 48-frame segments, from which 8 frames are uniformly sampled and provided to the VLM together with an instruction prompt that asks the model to determine whether the segment contains an anomaly. We use the model's default decoding configuration, including `do_sample=True`, `temperature=1`, `top_p=1`, `top_k=50`, etc. To estimate teacher confidence, we perform five independent inference runs for each segment and aggregate the resulting predictions.

For the student model, each video is divided into non-overlapping 16-frame segments, and we follow the default hyperparameter settings of UR-DMU Zhou et al. (2023). For utility-score estimation, each region (r) spans 48 frames. The motion magnitude of a region is computed by averaging the motion scores of the corresponding three consecutive 16-frame segments. We set $\beta = 0.3$ for UCF-Crime and $\beta = 0.8$ for XD-Violence. For the distillation objective, we set $\lambda = 0.9$ for conflicting

events on both datasets. To balance the magnitudes of the MIL and distillation losses, we scale the distillation loss by 4.5 on UCF-Crime and 5.5 on XD-Violence.

Student-model training was performed on a single NVIDIA RTX A6000 GPU each with 48 GB memory, while VLM inference was distributed across four GPUs. Training the weakly supervised learning models takes approximately 30 minutes on the UCF-Crime dataset and 1 hour on the XD-Violence dataset. The only difference from the UR-DMU baseline is that we perform inference every 100 iterations instead of every 10 iterations to reduce inference time during training. For VLM inference, it takes approximately 2 days using 4 GPUs in parallel.

C Additional Results

C.1 AUROC Results

Table 5: Performance Comparison on UCF-Crime dataset

Mode	Model	Encoder	AUROC (%)	AP (%)
Training-free	Zero-Shot CLIP Radford et al. (2021)	ViT	53.13	–
	ZS ImageBind Girdhar et al. (2023)	ViT	53.65	–
	LLAVA-1.5-7B Liu et al. (2024)	ViT	72.84	–
	LAVAD-14B Zanella et al. (2024)	ViT	78.33	–
	EventVAD-7B Shao et al. (2025)	ViT	82.03	–
	InternVL3-14B Zhu et al. (2025)	ViT	80.28	29.50
Weakly Supervised (Multimodal Inference)	TPWNG Yang et al. (2024)	ViT	87.79	-
	PEL4VADPu et al. (2024)	I3D+Text	86.76	33.99
	VadCLIP Wu et al. (2024)	ViT	88.02	33.55
	Ex-VAD Huang et al. (2025)	ViT	88.29	-
	DSANet Yin et al. (2026)	ViT	89.44	37.41
	Dual-Exploration Acharya et al. (2026)	I3D+ViT	87.80	38.33
Weakly Supervised (RGB-only Inference)	MIL Sultani et al. (2018)	I3D	77.42	25.03
	MIST Feng et al. (2021)	I3D	82.30	–
	RTFM Tian et al. (2021)	I3D	84.30	29.46
	WSAL Lv et al. (2021)	I3D	85.38	–
	BNSVP Sapkota & Yu (2022)	I3D	83.39	30.68
	CLAV Cho et al. (2023)	I3D	86.10	–
	MGFN Chen et al. (2023)	V-Swin	86.67	–
	UMIL Lv et al. (2023)	ViT	86.75	–
	UR-DMU Zhou et al. (2023)	I3D	86.97	35.48
	TSA Joo et al. (2023)	ViT	87.58	–
	GS-MoEAmicantonio et al. (2025)	I3D	91.58	34.08
	Dual-Exploration Acharya et al. (2026)	I3D	85.63	36.42
	DisCo (Ours)	I3D	86.44	**42.29**

In Table 5, we report both AUROC and AP on UCF-Crime for reference. The results show that a higher AUROC does not necessarily correspond to a higher AP. For example, VadCLIP achieves a higher AUROC than UR-DMU (88.02% vs. 86.97%), yet its AP is approximately 2 percentage points lower. Following Dual-Exploration, we also report the AP of representative baselines from each modality. Compared with the UR-DMU baseline, our method improves AP from 35.48% to 42.29% while maintaining comparable AUROC performance.

Table 6: Comparison of event-selection criteria on UCF-Crime. All methods select the top six events. OR denotes the overlap ratio between the selected event sets of each baseline criterion and DisCo.

Criterion	AP(%)	R@1(%)	OR	Top-3 Events
Teacher reliability	35.19	13.64	0.333	Explosion, Stealing, Shoplifting
Prediction disagreement	32.12	13.60	0.167	Abuse, Shoplifting, RoadAccidents
Weighted prediction disagreement	32.77	12.12	0.333	Shoplifting, Abuse, RoadAccidents
Semantic–motion disagreement (Ours)	**42.29**	**22.87**	**1.00**	**Explosion, Arrest, Fighting**

C.2 Additional Ablation Study

C.2.1 Effectiveness of Semantic–Motion Event Selection

Table 6 compares our semantic–motion conflict criterion with three alternative event-selection strategies. To determine whether the criteria identify similar events, we measure their overlap with the events selected by DisCo:

$$\mathrm{OR}(A, B) = |A \cap B| \,/\, |A| \tag{15}$$

where A and B denote the event sets selected by an alternative criterion and DisCo, respectively, and $|A| = |B| = 6$. A lower overlap ratio indicates more distinct selections. Teacher reliability measured by c^{teach} selects Explosion and Stealing, for which the teacher is accurate, but also Shoplifting ($\mathrm{Recall@0.5} = 18.57\%$) and Burglary ($\mathrm{Recall@0.5} = 31.75\%$), where it confidently misses many abnormal instances.[1] Consequently, R@1% decreases from 17.00% to 13.64%, showing that teacher reliability alone does not imply useful supervision. Prediction disagreement ($|p^{\mathrm{teach}} - p^{\text{UR-DMU}}|$) ranks events according to the difference between the predictions of the teacher and the MIL-trained UR-DMU model. It selects Abuse and Shoplifting among the top events, where UR-DMU and the teacher exhibit a performance discrepancy, but UR-DMU outperforms the teacher (see Fig. 2). Consequently, distillation based solely on disagreement can transfer teacher errors rather than complementary semantic knowledge, resulting in performance below both the student baseline and vanilla distillation. In contrast, DisCo prioritizes segments for which the teacher provides additional semantic evidence that is not explained solely by dominant motion cues. Moreover, the motion-dominant disagreement term is downweighted by $\beta < 1$, giving greater priority to discovering abnormalities missed by the motion-based student than to suppressing its predictions. Thus, for events such as Abuse and Shoplifting, where strong motion is informative but the teacher yields frequent false-negative predictions, DisCo limits the teacher's suppressive influence. Compared with prediction disagreement, it selects a largely distinct event set ($\mathrm{OR} = 0.167$; one shared event) and achieves the best performance. Additional analyses, including the complete list of events selected by each criterion are present in Table 8.

C.2.2 Effectiveness of Utility Score Components

Table 7 reports the contribution of each component of the proposed utility score, while Table 8 shows the corresponding events selected by each component and by the alternative criteria considered in the main paper.

Using only the semantic component, C_{sem}, improves over vanilla distillation on UCF-Crime by selecting events with relatively high VLM recall, such as Fighting, Assault, and Arrest. However, it performs substantially worse on XD-Violence because it fails to select motion-dominant events, such as Accident, where significant motion changes are often introduced by scene transitions or changes in camera viewpoint rather than the anomaly itself. In contrast, using only the motion

[1] $\mathrm{Recall@0.5}$ denotes class-wise recall computed using 0.5 as the threshold on the teacher prediction (p^{teach}).

Table 7: Ablation of the components used to compute the utility score.

Method	AP (%)	
	UCF-Crime	XD-Violence
C_{sem} only	39.20	80.61
C_{motion} only	29.14	82.68
$\max(0, \Delta_r) + \beta \max(0, -\Delta_r)$	38.75	84.97
$C_{\text{sem}} + \beta C_{\text{motion}}$	**42.29**	**85.96**

Table 8: Event selections produced by alternative selection criteria, individual utility-score components, and the complete proposed utility score on UCF-Crime. The InternVL3-14B column reports event-level recall at a prediction threshold of 0.5. A checkmark indicates that the corresponding event is selected. Events are ordered according to the ranking produced by our utility score.

Event	InternVL3-14B	Alternative Criteria			Utility Score Components			Ours
	Recall@0.5	Teacher Reliability	Prediction Disagreement	Weighted Disagreement	C_{sem}	C_{motion}	$\max(0, \Delta_r) + \beta \max(0, -\Delta_r)$	
Explosion	0.657	✓		✓	✓		✓	✓
Arrest	0.741				✓		✓	✓
Fighting	0.853				✓		✓	✓
Assault	0.847				✓		✓	✓
Robbery	0.642	✓			✓		✓	✓
Shooting	0.643		✓	✓	✓		✓	✓
RoadAccident	0.734	✓	✓	✓	✓	✓	✓	
Burglary	0.317	✓					✓	
Arson	0.281							
Abuse	0.168		✓	✓		✓		
Vandalism	0.571		✓			✓		
Stealing	0.713	✓	✓	✓		✓		
Shoplifting	0.187	✓	✓	✓		✓		
# Selected	–	6	6	6	7	5	8	6

component, C_{motion}, yields the lowest AP on UCF-Crime. As shown in Table 8, this criterion primarily selects events with low VLM recall, including Abuse, Shoplifting, and Vandalism, for which the student model outperforms the teacher. In these cases, the teacher primarily suppresses potentially abnormal events rather than providing complementary semantic information.

Combining semantic and motion cues using $\max(0, \Delta_r) + \beta \max(0, -\Delta_r)$ substantially improves performance by selecting semantic-dominant events regularized by motion suppression, even without explicitly considering teacher reliability. This result suggests that teacher utility can be meaningfully estimated by using motion as a proxy, particularly because MIL models tend to favor dominant feature changes. Nevertheless, this criterion selects eight events on UCF-Crime, including Burglary, for which the teacher is consistently incorrect. This is reflected by its high ranking under the teacher-reliability criterion despite its low recall, thereby introducing lower-utility events that reduce the effectiveness of distillation. Incorporating teacher reliability into the complete utility score further refines the selection to six events by suppressing unreliable categories while retaining informative ones. Consequently, the complete utility score achieves the highest AP on both UCF-Crime and XD-Violence.

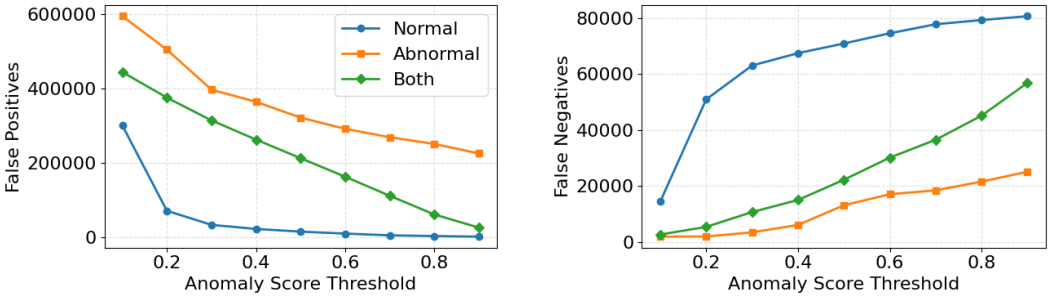
C.2.3 Effect of Distilling Abnormal and Normal Segments

In Table 9, we analyze the effect of distilling segments based on the confidence provided by the VLM. In the first experiment, we distill only potential abnormal segments. Since VLM inference is performed five times per segment, a segment is considered abnormal if it is predicted as 1 in at least three runs, corresponding to a confidence greater than 0.6. In the second experiment, we distill only normal segments with confidence lower than 0.4. As shown in Fig. 5, distilling only abnormal segments leads to more false positives across all thresholds, indicating that the decision boundary becomes biased toward abnormal predictions. Conversely, distilling only normal segments reduces false positives but increases missed anomalies (false negatives). Therefore, we distill both potential

abnormal and normal segments to provide a more balanced signal and improve overall detection performance.

Table 9: Ablation study on distilling different segments.

Method	Confidence Filter	UCF-Crime AP (%)
Baseline	–	35.48
DisCo	Abnormal only ($p^{teach} \geq 0.6$)	36.60
DisCo	Normal only ($p^{teach} \leq 0.4$)	30.36
DisCo	Abnormal + Normal	**42.29**



(a) (b)

Figure 5: Effect of knowledge distillation on detection errors when knowledge is distilled from pseudo-abnormal, normal, or both frame types on the UCF-Crime dataset. Abnormal and normal frames are determined by VLM confidence.

C.2.4 Prompt Sensitivity Analysis

Table 10: Sensitivity analysis of prompt templates on UCF-Crime.

Prompt Template	Selection Strategy	AP (%)
Definition-Guided	All Events	28.51
Event-Specific	All Events	35.94
General	Conflicting Events	38.23
Definition-Guided	Conflicting Events	39.62
Event-Specific	Conflicting Events	**42.29**

Table 10 reports the performance of our framework under different prompt templates used for teacher inference. In the first setting (*General Prompt*), the model is asked to determine whether the input video contains any abnormal event without providing any explicit definition or examples of abnormal activities. In the second setting (*Definition-Guided Prompt*), we provide a short description of abnormal events along with representative categories corresponding to the anomaly classes in the dataset. This additional semantic guidance helps the vision-language model better understand what constitutes abnormal activity. As shown in Table 10, the definition-guided prompt slightly improves detection performance over the general prompt, increasing AP from **38.23%** to **39.62%**. We further evaluate an *Event-Specific Prompt*, in which the teacher is asked to verify whether a particular abnormal event is present in the video (e.g., whether an *explosion* occurs). This setting achieves the best performance.

However, for conflict estimation, we use the *Definition-Guided Prompt* because it allows us to assess the VLM's general understanding of abnormal events without conditioning it on a specific event label. Providing an event-specific prior at this stage may increase false alarms by encouraging the teacher to confirm the queried event even when the visual evidence is weak. Therefore, we

Table 11: Performance Comparison of different teacher models on the UCF-Crime.

Model	Size	AP (%)
InternVL3-14B	14B	39.62
InternVL3-9B	9B	35.75
LLaVA-NeXT-Video	7B	35.89

first identify events with high teacher utility using the definition-guided prompt and then condition the prompt on the corresponding event category to further improve performance. Unless otherwise specified, the different teacher-model, confidence-estimation, student-model generalization, and temporal-window ablations use the *Definition-Guided Prompt*, whereas the utility-component, distillation, hyperparameter, event-ranking, and inference-run ablations use the final *Event-Specific Prompt*. For a fair comparison, we adopt the *General Prompt* and *Definition-Guided Prompt* templates following Acharya et al..

Event-Specific Prompt

Task: You are an anomaly detection assistant. Analyze the given video clip and determine whether it contains the specified abnormal event.
Target Event: {`event_name`}
Definitions:
- Target Event: The video shows the specified abnormal event ({`event_name`}), which involves harmful, dangerous, unlawful, or threatening activity related to this category.
- Non-Target Event: The video does not contain the specified event and instead shows normal daily activities such as walking, talking, driving normally, shopping, exercising, or working.

Evaluation Criteria:
- Carefully examine the actions, objects, and interactions visible in the video clip.
- Consider the context and whether the scene clearly corresponds to the target event.
- The clip may contain only a short portion of the full event.

Output Rule:
- If the video clip contains the target event ({`event_name`}) → Reply with only: 1
- If the video clip does NOT contain the target event → Reply with only: 0
- Do not include explanations, reasoning, or extra text.

C.2.5 Effect of Different Teacher Models

Table 11 compares the effect of different VLMs used as teacher models in our framework. Among the evaluated teacher models, InternVL3-14B achieves the highest AP. Within the InternVL3 family, the 14B model outperforms the 9B variant, suggesting that increased model capacity can improve the quality of the supervisory signal. Fig. 6 further illustrates the event-wise performance comparison between the baseline weakly supervised model (UR-DMU) and two different VLM teachers. The results show that LLaVA-NeXT-Video Li et al. (2024) also performs better than UR-DMU for certain categories, such as *Explosion*, indicating that VLMs possess generalized semantic knowledge that can help identify specific abnormal events. However, its performance remains weaker for most other categories, likely due to the smaller model size and limited video understanding capability. For a fair comparison, we keep the distillation setting consistent across all teacher models. Both results are evaluated using the definition-guided prompt.

C.2.6 Teacher Confidence Estimation

VLMs are prone to hallucinations and overconfident predictions; therefore, reliable confidence estimation is important when using them as teachers for knowledge distillation. Table 12 reports the performance of our method using three training-free confidence metrics derived from the VLM outputs. *Token probability* measures the softmax probability of the generated token. During inference, the language module of the VLM generates a binary response indicating whether a segment is normal or abnormal (i.e., 0 or 1), along with the corresponding logits for the generated token. Since the

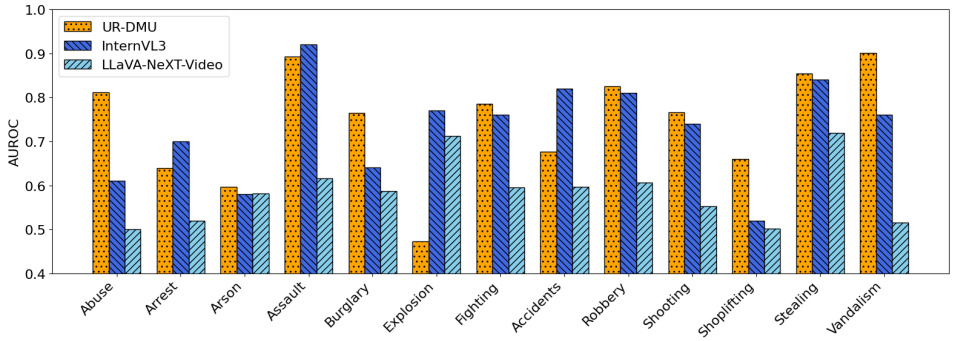


Figure 6: Event-wise comparison between InternVL3-14B, LLaVA-NeXT-Video and UR-DMU.

Table 12: Performance comparison under different teacher confidence estimation on UCF-Crime.

Confidence Metric	AP (%)
Token Probability	36.33
Verbalized Confidence	36.58
Multi Inference	**39.62**

output is restricted to a single-token response, the confidence score is computed as the softmax probability of the generated token. The second metric is *verbalized confidence*, where we slightly modify the prompt to estimate the likelihood that a segment is abnormal. Instead of producing a binary response, the model outputs a score between 0 and 100, which is then normalized to the range [0,1]. For *multi-inference confidence*, we perform multiple inference runs for a single segment to obtain binary predictions while keeping the default temperature setting of InternVL3. The confidence is defined as the average prediction across these runs. The results indicate that multi-inference confidence provides a more reliable confidence estimate than token probability and verbalized confidence.

C.2.7 Evaluation with Different Student Models

Table 13: Compatibility of DisCo with different weakly supervised baselines.

	UCF-Crime			XD-Violence		
Method	**Original**	**+ DisCo**	**Gain**	**Original**	**+ DisCo**	**Gain**
UR-DMU	35.48	**39.62**	+4.14	79.14	**84.77**	+5.63
Dual-Explore	36.42	**39.22**	+2.8	80.85	**85.83**	+4.98

We further evaluate DisCo with Dual-Exploration, a weakly supervised baseline that performs segment exploration to identify additional potentially abnormal regions. For a fair comparison, we use the *Definition-Guided Prompt*, since the event-specific prompt is not applicable to the multimodal inference setting of Dual-Exploration. As shown in Table 13, selective distillation consistently improves Dual-Exploration on both datasets, increasing AP by 2.80 on UCF-Crime and 4.98 on XD-Violence. These results also outperform the test-time ensemble reported in Table 1 of the main paper, which achieves 38.33% AP on UCF-Crime and 84.58% AP on XD-Violence. The gains demonstrate that, even when the student already performs uncertainty-based segment exploration, selective distillation provides complementary semantic knowledge that would otherwise remain overlooked.

C.3 Hyperparameter Sensitivity

C.3.1 Effect of hyperparameter β

Table 14 shows that DisCo is robust to β under fixed Top-k selection: all settings outperform the baseline and vanilla KD, with optimal performance maintained over $\beta = 0$–0.4 on UCF-Crime

Table 14: Effect of the motion-dominance weight β on AP under Top-k and Otsu-based event selection for UCF-Crime and XD-Violence. Dark gray rows denote the robust operating region ($\beta = 0.3$–0.4) across the datasets, while light gray cells indicate uniform performance. Bold values denote the best AP.

β	**UCF-Crime**	**UCF-Crime (Otsu)**		**XD-Violence**	**XD-Violence (Otsu)**	
	AP (%) (Top-k)	AP (%)	#Events	AP (%) (Top-k)	AP(%)	#Events
0.0	**42.29**	39.20	7	84.97	80.61	5
0.1	**42.29**	**42.29**	6	84.97	80.61	5
0.2	**42.29**	**42.29**	6	84.97	80.61	5
0.3	**42.29**	**42.29**	6	84.97	84.97	6
0.4	**42.29**	**42.29**	6	84.97	84.97	6
0.5	40.18	37.83	3	84.97	84.97	6
0.6	40.18	37.83	3	84.97	84.97	6
0.7	37.18	36.42	9	84.97	84.97	6
0.8	37.18	36.42	9	84.97	**85.96**	5
0.9	38.20	36.42	9	84.97	**85.96**	5
1.0	38.20	38.20	6	84.97	**85.96**	5

Table 15: Event selections under different values of the motion-dominance weight β on UCF-Crime. Events are ordered according to the ranking obtained using the optimal value of β in our method.

Event	**Top-k Selection**					**Otsu Selection**				
	β=0	β=0.3	β=0.5	β=0.7	β=1	β=0	β=0.3	β=0.5	β=0.7	β=1
Explosion	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Arrest	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fighting	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Assault	✓	✓	✓	✓		✓	✓		✓	
Robbery	✓	✓	✓			✓	✓		✓	
Shooting	✓	✓				✓	✓		✓	
RoadAccident			✓	✓	✓	✓			✓	✓
Burglary										
Arson										
Abuse					✓				✓	✓
Vandalism										
Stealing										
Shoplifting				✓	✓				✓	✓
# Selected	6	6	6	6	6	7	6	3	9	6

and across the full range on XD-Violence. Under Otsu selection, $\beta = 0.3$–0.4 provides a stable operating region across both datasets, while larger values can reduce performance on UCF-Crime. This indicates that semantic conflict is the primary selection signal, while β regularizes motion-dominant disagreements during adaptive selection.

Table 15 explains this robustness at the event level. On UCF-Crime, the selected event set changes little across β: informative categories such as *Explosion*, *Arrest*, and *Fighting* are consistently retained, while unreliable categories such as *Burglary*, *Arson*, and *Vandalism* are consistently excluded. Only at extreme β values do less reliable categories such as *Abuse* and *Shoplifting* enter the selected set, which reduces AP relative to the optimal setting but never falls below the baseline or alternative selection criteria, since the core high-utility events remain distilled regardless of β. The role of β is more pronounced under Otsu-based adaptive selection, particularly on XD-Violence: when $\beta < 0.3$, performance falls below vanilla KD, underscoring the role of β as a regularizer that suppresses motion-dominant false positives before adaptive thresholding. Beyond $\beta \geq 0.3$, both datasets enter a stable operating region in which performance no longer degrades relative to vanilla KD, confirming that DisCo does not require precise tuning of β to achieve reliable gains.

Table 16: Effect of the weighting coefficient λ on UCF-Crime and XD-Violence.

λ	**UCF-Crime AP (%)**	**XD-Violence AP (%)**
1.0	39.01	**86.11**
0.9	**42.29**	85.96
0.8	39.59	84.69
0.7	40.46	84.00
0.6	40.83	81.63
0.5	37.20	82.26
0.4	37.81	82.56
0.3	36.76	82.44
0.2	38.65	83.10
0.1	39.49	83.56
0.0	35.48	79.14

Table 17: Effect of the number of teacher inference runs on UCF-Crime.

VLM Inference Runs	**AP (%)**
1	39.94
3	41.29
5	**42.29**

While any $\beta \geq 0.3$ is sufficient for robust performance, identifying the dataset-specific *optimal* value can be further guided by the relative motion magnitude of normal and abnormal videos in each dataset. Intuitively, a dataset in which normal videos already exhibit substantial motion requires a stronger motion-suppression weight to avoid mistaking high-motion but non-anomalous segments for genuine conflict, whereas a dataset with comparatively static normal videos permits a lower β without introducing spurious motion-driven false positives. Consistent with this intuition, the median motion magnitude of normal videos is higher on XD-Violence (9.82) than on UCF-Crime (5.71), aligning with the higher optimal β observed for XD-Violence in our sensitivity analysis. This suggests that the ratio between normal and abnormal video motion magnitude within a dataset can serve as a practical, low-cost signal for selecting a near-optimal β without requiring exhaustive tuning.

C.3.2 Effect of hyperparameter λ

Table 16 evaluates sensitivity to the loss weighting coefficient λ. Every setting with $\lambda > 0$ outperforms the corresponding baseline on both datasets. The highest AP is obtained at $\lambda = 0.9$ on UCF-Crime and $\lambda = 1.0$ on XD-Violence. We adopt $\lambda = 0.9$ as a strong shared setting across both datasets while retaining a small contribution from the MIL objective. In cases of abnormal videos where the teacher's maximum segment-level prediction is less than 1, the residual $\mathcal{L}_{\text{MIL}}$ weight ensures that the bag-level classification objective still contributes to training.

C.3.3 Effect of event selection ranking

Fig. 7 further validates our utility-based event ranking. As events are cumulatively added in order of decreasing utility score, AP peaks exactly at the Otsu-selected cutoff (Top-6) and declines thereafter, converging to the vanilla KD performance (35.94%) when all 13 events are included. This confirms that the proposed utility score correctly ranks events by their informativeness for distillation, and that Otsu thresholding reliably identifies the point of maximum benefit without requiring manual tuning of the number of selected events.

C.3.4 Effect of number of VLM inference runs

Table 17 reports the performance for different numbers of teacher inference runs. AP improves as the number of runs increases from 1 to 5, as averaging over more stochastic generations yields a more reliable estimate of teacher confidence and consistency. We do not go beyond 5 runs due to the increased computational overhead.

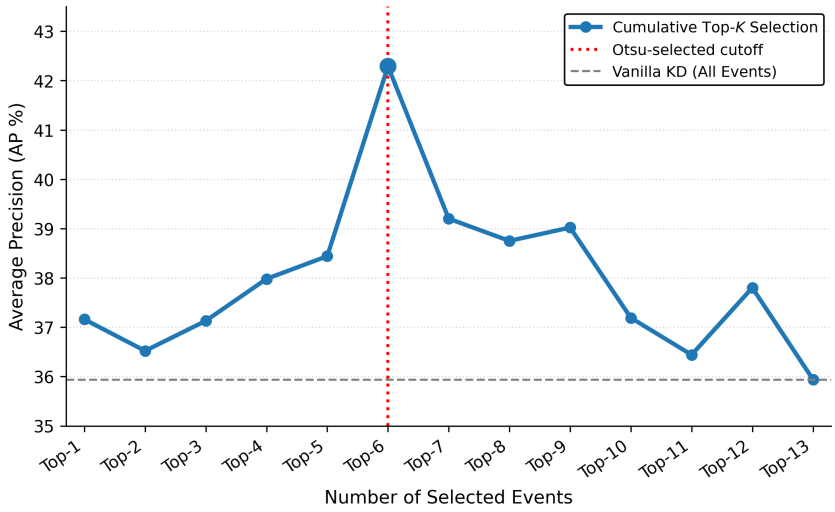


Figure 7: Performance as a function of the number of selected events on UCF-Crime. Events are cumulatively added according to the ranking produced by the proposed utility score. The red dotted line denotes the cutoff automatically selected by Otsu thresholding. Performance reaches its maximum at the Otsu-selected cutoff, indicating that the proposed ranking successfully prioritizes informative events while excluding lower-utility events from distillation.

C.3.5 Effect of Temporal Window

Table 18: Comparison of 16-frame and 48-frame VLM inference windows. The 16-frame setting matches the temporal resolution of the I3D features, while the 48-frame setting is the default used in our method.

Temporal Window	**UCF-Crime**	**XD-Violence**
	AP (%)	**AP (%)**
16	38.31	83.57
48	39.62	84.77

Since the student model's I3D features are extracted over 16-frame segments, matching this temporal resolution, we first evaluate the VLM at the same 16-frame window. Table 18 compares this setting against our default 48-frame window on both datasets. The 16-frame window degrades performance on both UCF-Crime (38.31% vs. 39.62%) and XD-Violence (83.57% vs. 84.77%), as it represents too short a temporal context for the VLM to understand the ongoing event, with consecutive frames often appearing visually similar. In contrast, the 48-frame window, spanning approximately 1.6 seconds at 30 fps, provides sufficient context and improves performance, and is therefore adopted as our default.

Table 19: Performance comparison under different temporal window settings on UCF-Crime.

Temporal Window	**UCF-Crime AP (%)**
48	**39.62**
144	38.03
288	36.05

We further investigate whether increasing the temporal window beyond 48 frames yields additional gains. Table 19 reports the performance of DisCo under temporal windows of 48, 144, and 288 frames, uniformly sampling 8 frames from each window. Performance decreases monotonically as the window grows on UCF-Crime, partly due to information loss caused by the reduced sampling

ratio, which drops from 8/48 (∼16%) to 8/144 (∼5%) and further to 8/288 (∼2.8%). In addition, larger temporal windows increase the likelihood of missing rare abnormal events, which are typically short and localized. Since long-duration anomalies such as shoplifting and burglary are already excluded by the utility score, shorter temporal windows are more suitable for capturing the remaining brief abnormal events.

C.4 Qualitative Results

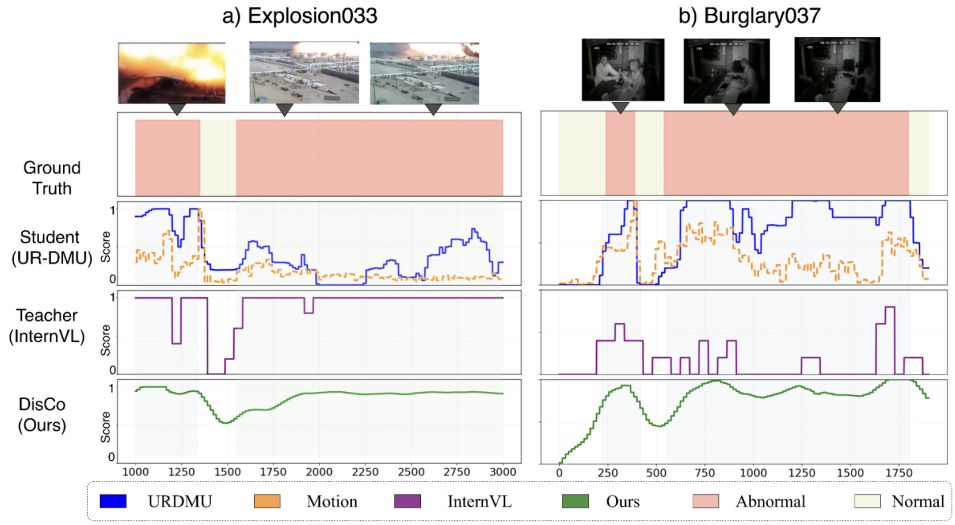


Figure 8: Qualitative comparison of anomaly scores on two UCF-Crime videos: (a) Explosion033 and (b) Burglary037. In Explosion033, the motion-based WSVAD model (UR-DMU) is biased toward dominant motion changes and fails to detect the second explosion event, while the VLM teacher (InternVL) captures both abnormal events through its semantic understanding. In contrast, in Burglary037 the teacher produces unreliable predictions, whereas the student model captures motion patterns associated with the abnormal activity. Our DisCo framework selectively integrates semantic cues from the VLM when beneficial while preserving reliable motion cues from the student.

Fig. 8 presents a qualitative analysis illustrating that our model learns semantic knowledge from the VLM while preserving information captured by the student.

D Theoretical Analysis

D.1 Proof of Remark 1

Proof. By linearity of expectation,

$$\mathbb{E}[\widehat{C}_e \mid e] = \mathbb{E}\left[\frac{1}{N_e}\sum_{r=1}^{N_e} C_r \middle| e\right] = \frac{1}{N_e}\sum_{r=1}^{N_e}\mathbb{E}[C_r \mid e]. \tag{16}$$

Since $C_r = \mu_e + \epsilon_r$ and $\mathbb{E}[\epsilon_r \mid e] = 0$, we have

$$\mathbb{E}[C_r \mid e] = \mu_e. \tag{17}$$

Therefore,

$$\mathbb{E}[\widehat{C}_e \mid e] = \frac{1}{N_e}\sum_{r=1}^{N_e}\mu_e = \mu_e. \tag{18}$$

Assuming that the region-level utility scores are conditionally independent given event e, with $\mathrm{Var}(C_r \mid e) = \sigma_e^2$, we obtain

$$\mathrm{Var}(\widehat{C}_e \mid e) = \mathrm{Var}\left(\frac{1}{N_e}\sum_{r=1}^{N_e} C_r \,\middle|\, e\right) \tag{19}$$

$$= \frac{1}{N_e^2}\sum_{r=1}^{N_e} \mathrm{Var}(C_r \mid e) = \frac{\sigma_e^2}{N_e}. \tag{20}$$

Thus, event-level aggregation reduces the variance of the utility estimate as the number of regions increases. □

D.2 A Multi-View Interpretation of Semantic Distillation

Let X^M denote the student-view representation, X^S the semantic view available to the VLM teacher, and $Y \in \{0, 1\}$ the unobserved segment-level anomaly label. At inference, the student observes only X^M. Under squared loss, define

$$R^\star(X^M) = \inf_f \mathbb{E}\Big[\big(f(X^M) - Y\big)^2\Big], \tag{21}$$

$$R^\star(X^M, X^S) = \inf_g \mathbb{E}\Big[\big(g(X^M, X^S) - Y\big)^2\Big]. \tag{22}$$

The corresponding Bayes-optimal predictors are

$$f_M^\star(X^M) = \mathbb{P}(Y = 1 \mid X^M), \tag{23}$$

$$g^\star(X^M, X^S) = \mathbb{P}(Y = 1 \mid X^M, X^S). \tag{24}$$

Proposition 1 (Complementary semantic information). The two-view Bayes risk satisfies

$$R^\star(X^M, X^S) \le R^\star(X^M), \tag{25}$$

and the corresponding risk gap is

$$\Delta R := R^\star(X^M) - R^\star(X^M, X^S) \tag{26}$$

$$= \mathbb{E}\Big[\big(g^\star(X^M, X^S) - f_M^\star(X^M)\big)^2\Big] \tag{27}$$

$$= \mathbb{E}_{X^M}\left[\mathrm{Var}_{X^S \mid X^M}\left(g^\star(X^M, X^S)\right)\right] \ge 0. \tag{28}$$

Equality holds if and only if $I(Y; X^S \mid X^M) = 0$.

Proof. Because predictors using (X^M, X^S) can ignore X^S, their minimum achievable risk cannot exceed that of predictors restricted to X^M. Moreover, by the tower property,

$$f_M^\star(X^M) = \mathbb{E}\big[g^\star(X^M, X^S) \mid X^M\big]. \tag{29}$$

The stated decomposition then follows from the law of total variance. The gap vanishes exactly when the semantic view provides no additional information about Y conditioned on X^M. □

This result provides an interpretation of semantic distillation: when $\Delta R > 0$, the semantic view contains label-relevant information not fully represented by the student view. An ideal semantic teacher can therefore provide complementary training cues. However, because the student observes only X^M at inference, its minimum achievable risk remains $R^\star(X^M)$.