This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

The Road Less Seen: Segment Exploration for Weakly Supervised Video Anomaly Detection

Anusha Acharya[1] Hitesh Sapkota[2*] Qi Yu[1,2*] Xumin Liu[1]
[1]Rochester Institute of Technology [2]Amazon
aa6301@rit.edu sapkoh@amazon.com qi.yu@rit.edu xmlics@rit.edu

Abstract

*Weakly supervised learning (WSL) provides a cost-effective learning paradigm for video anomaly detection (VAD) from data with video-level annotation instead of requiring costly fine-grained segment-level annotation. Although contemporary methods have shown promising results on challenging real-world surveillance videos, most of them are evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC). We reveal that a high AUROC could result in a very low recall for meaningful False Positive Rate (FPR) thresholds. Thus, these models suffer from limited practical values, especially in high-stake domains (*e.g. *public safety and medical diagnosis), where missing the true anomalies incur high cost. This surprising phenomenon is rooted in the interplay of weak supervision and the highly imbalanced distribution between normal and anomalous video segments. To tackle this key challenge in VAD systems, we propose a novel dual exploration strategy that combines temporal clustering with uncertainty-based segment exploration. Temporal clustering selects diverse segments based on both semantic and temporal similarity, while uncertainty-based sampling targets low-scoring segments with high model uncertainty. The main aim of exploration is to ensure that the model learns from a wide range of patterns, both diverse and ambiguous, resulting in more informed and robust decision-making, and reduction in false negatives. Meanwhile, we adopt two practical metrics to replace the commonly used AUROC score for a more effective measure for evaluation. Experiments conducted in challenging real-world videos demonstrate that our exploration strategy improves VAD performance compared to the baselines on these metrics, which justifies its improved practical value in real-world settings.*

1. Introduction

Video anomaly detection aims to identify unusual events within video footage. One of its primary applications is in intelligent surveillance camera systems [17] in assisting public safety officers by automatically detecting abnormal activities. In these security-critical contexts, accurate anomaly detection is essential, as missed detections can lead to severe consequences such as delayed emergency responses, property damage, or even threats to human life. Weakly supervised learning provides a practical and scalable solution for learning VAD systems as it requires only video-level annotation, thereby avoiding the labor-intensive and costly process of annotating each video segment. Multiple instance learning (MIL), a WSL method, treats each video as a bag of segments labeled either positive (containing at least one abnormal segment) or negative (containing only normal segments) [6]. Features are extracted from video segments using pretrained models such as C3D [19], I3D [3], or ViT [7], which are then used to train anomaly detectors that output anomaly scores. These detectors are usually trained to maximize the score margin between segments of the positive and negative bags, respectively.

State-of-the-art weakly supervised video anomaly detection (WSVAD) methods have achieve high AUROC scores [1, 4, 8, 10, 11, 27, 32]. As a threshold-free metric, the AUROC has been commonly used in the literature for performance evaluation. However, our detailed analysis reveals that the metric is an overly optimistic and doesn't precisely reflect on the model's utility in practical settings. Specifically, AUROC does not account for the significant difference in the cost of missing an abnormal event or false negatives (which is more costly to miss) compared to a normal one or false positives [12, 16]. For example, the UR-DMU model [32] attains a competitive AUROC score of 86.97 on the UCF-Crime dataset [17], but achieves the lowest recall compared to the other two models, including VadCLIP[24] and RTFM [18], as shown in Figure 1 (b).

A fundamental reason for the low recall is the lack of segment-level supervision. To take advantage of (weak) video-level labels, existing works usually consider top-1 or top-k segments with the highest anomaly scores as abnormal segments during training [29]. However, a fixed value of k is rigid and can lead to either overestimation or under-

*Work not related to the position at Amazon.

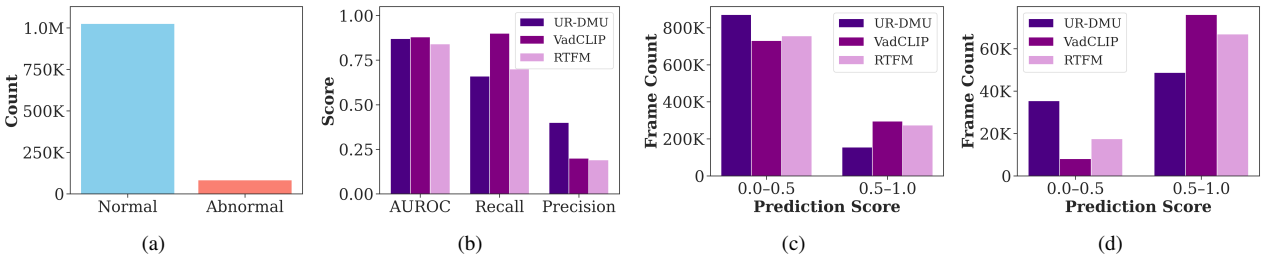


Figure 1. (a) Imbalanced distribution of normal and abnormal segments in UCF-Crime dataset (b) AUROC, recall, and precision (at threshold 0.5) of three representative VAD models (c) & (d) Prediction scores of the models for normal and abnormal segments, respectively.

estimation of the true number of abnormal events in different videos. Moreover, segments with high anomaly scores often tend to cluster around a single event within a narrow temporal window, restricting the top-k selection to a single event and thus failing to capture multiple distinct abnormal events. Finally, this strategy tends to favor segments with prominent motion changes and makes the model biased toward simple contexts while overlooking more subtle or complex anomalies [10]. Techniques, such as distributionally robust optimization (DRO) [15] and inclusion of temporally distant segments [14], have attempted to alleviate the problem. However, segment selection still relies on the highest anomaly scores, which are typically biased toward simpler contexts and miss out on complex patterns of anomalies, resulting in high false negatives.

Unbiased MIL mitigates context bias by splitting video segments into confident and ambiguous sets, using unsupervised clustering to detect anomalies in the ambiguous set [10]. However, it assumes that all anomalies share similar semantics, limiting its ability to capture diverse patterns. It is also worth noting that making predictions biased towards abnormal segments does not address the problem either. For example, in Figure 1 (b), VadCLIP achieves a very high recall at 0.5 but it introduces more than 290K false positive, which almost double the size of UR-DMU, as shown Figure 1 (c). Figure 1 (d) further confirms that VadCLIP tends to predict higher anomaly scores. Due to the highly imbalanced distribution of normal and abnormal segments, it misclassifies a large number of normal segments, leading to a very high false positive rate that makes it less useful.

To address the challenges as outline above, we propose a dual diversity-ambiguity exploration strategy for WSVAD. It is designed to specifically target the discovery of all relevant abnormal segments, rather than focusing solely on those with the highest anomaly scores. It performs two complementary forms of exploration: (1) temporal clustering-based selection (TCBS), which selects semantically and temporally diverse segments to preserve local continuity and capture distinct abnormal events, and (2) uncertainty-based selection (UBS), which targets ambiguous segments with low anomaly scores but high model uncertainty. For uncertainty-guided exploration, we further leverage the memory unit that represents diverse abnormal events. Among the segments with high prediction uncertainty, we prioritize those that closely resemble the memory features for further exploration. Moreover, to mitigate the model's limited knowledge of truly novel or unseen anomalies, we leverage the extensive pretrained semantic knowledge of a vision-language model (VLM) and fuse its prediction scores with our dual-exploration strategy to improve overall detection performance. Our main contributions are summarized as follows:

- We identify a critical weakness in using AUROC for performance evaluation of video anomaly detection by showing its overly optimistic assessment on the model's utility in practical settings.
- We propose a dual exploration framework that leverages the temporal-semantic similarity of segments and model uncertainty under weak supervision to perform systematic exploration, enabling the discovery of diverse and ambiguous abnormal patterns.
- We recommend two practical evaluation metrics to replace the commonly used AUROC score for a more effective measure for evaluation.

Experiments conducted in challenging real-world videos demonstrate better dual exploration performance compared to competitive baselines on these metrics, which justifies its improved practical value in real-world settings.

2. Related Work

Sultani et al. first used deep MIL framework to train VAD models, and also introduced a large-scale dataset for the training purpose [17]. The model is trained to maximize the separation between the highest-scoring segment of normal and abnormal videos. This reliance on a single abnormal segment in the entire video makes it vulnerable to outliers [14]. Robust Temporal Feature Magnitude Learning (RTFM) considers top-k abnormal segments for learning [18]. A few other recent methods also adopt the top-k approach for segment selection with improved performance [4, 24, 32]. However, they still suffer from some key limitations. First, using a fixed k is too rigid, as the actual number of abnormal segments in a video is typically unknown, and

different for different videos. Second, these methods often tend to select temporally adjacent high-scoring segments, potentially overlooking anomalies that are temporally dispersed. Finally, they tend to be confident and biased towards easy-to-classify segments, which might miss more complex and subtle anomalies.

To overcome these limitations, a Distributionally Robust Optimization (DRO) framework has been developed that adapts k according to the content of each video [15]. However, this approach still does not ensure the coverage of all distinct abnormal events. Dynamic non-parametric clustering has also been leveraged that encourages temporally adjacent segments to lie within the same cluster [14]. Then, the segment with the highest anomaly score is selected from each cluster. However, this model still favors simple contextual anomalies and may miss complex or subtle patterns during segment selection and model training. In contrast, Unbiased MIL divides video segments into confident and ambiguous sets based on prediction variance, and clusters the ambiguous set into presumably normal and abnormal segments [10]. Segments similar to high-scoring confident segments are labeled abnormal and considered for training. However, it assumes semantic similarity among anomalies. Diverse or distinct abnormal events may be missed, limiting exploration of the full anomaly spectrum, thereby compromising detection performance.

In parallel, increasing attention has been paid to the integration of textual features of CLIP-based models and the use of learnable prompts to enhance semantic diversity [4, 8, 24]. For instance, Zhong et al. [24] proposed VadCLIP, which combines visual and textual modalities by leveraging the frozen CLIP backbone for improved anomaly localization and classification. Although CLIP-based models and memory-augmented architectures [32] offer rich and discriminative feature representations, these methods are still limited by top-k selection. Beyond CLIP, recent vision-language models such as InternVL [5, 20, 33], LLaVA [9], and VideoLLaMA [30] have demonstrated strong performance on general video understanding tasks. While strategies such as learning reasoning modules, guided prompts [26], or instruction tuning [31] can improve their performance on VAD task, these approaches incur additional costs for fine-tuning. Instead, we leverage the prediction confidence of these VLMs in a training-free manner to enhance detection performance.

3. Diversity-Ambiguity Dual Exploration for WSVAD

3.1. Preliminaries

In a MIL setting, each abnormal video is treated as a positive bag $\mathcal{B}_{\text{pos}}$, and has at least one abnormal segment, whereas each normal video is treated as a negative bag $\mathcal{B}_{\text{neg}}$, containing only normal segments. During training, the model is supervised by video-level labels, but is trained to to assign an anomaly score at the segment level. The model $f(.)$ is encouraged to assign high anomaly scores to abnormal segments (close to 1) and low scores (close to 0) to normal ones. Then the MIL loss is defined as:

$$\mathcal{L}_{MIL} = \text{BCE}(y, \hat{y}) \tag{1}$$

where $y = [y_{\text{pos}} = 1,\ y_{\text{neg}} = 0]$ is the ground truth bag level. For most existing methods, the positive bag label is determined by either selecting the maximum scoring segment or by averaging the scores of top-k segments in the positive bag. Following the MIL assumption, let $\mathcal{K} \subset \mathcal{B}_{\text{pos}}$ be the set of selected segments from the positive bag such that:

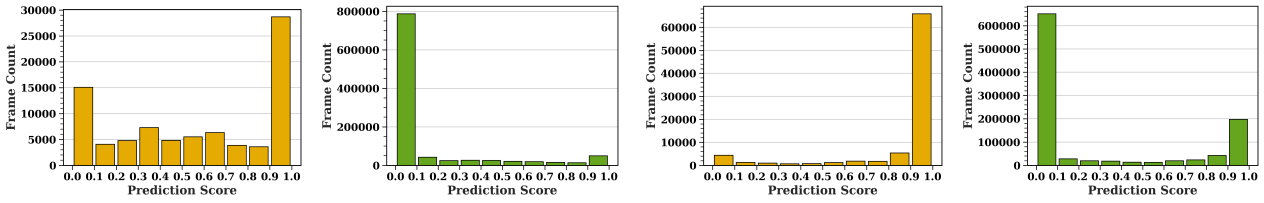
$$\hat{y} = \left[\hat{y}_{\text{pos}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} f(\mathbf{x}^{+}_{[k]}),\ \hat{y}_{\text{neg}} = \max_{j \in \mathcal{B}_{\text{neg}}} f(\mathbf{x}^{-}_{j}) \right] \tag{2}$$

where $\mathbf{x}_i$ denotes the feature of the i-th video segment with superscript $^+$ or $^-$ indicating that the segment belongs to the positive and negative bags, respectively. In a negative bag, since all video segments are normal, minimizing the maximum anomaly score encourages the model to assign low scores across the entire video.

3.2. Why is AUROC not a good metric for Evaluating Video Anomaly Detection?

In cases of severe data imbalance where negatives vastly outnumber positives, even a very low false positive rate (FPR) may still generate a large number of false alarms. To maintain precision at a usable level, the classification threshold must be raised, which in turn compromises recall. We refer to this as a **low practical recall phenomenon**. Meanwhile, since AUROC measures the probability that a model ranks a randomly chosen positive instance above a negative one, it can still appear high if the model correctly identify many negatives (which are more frequent) at a good rate given the class imbalance setting. Thus, AUROC gives a false sense of effectiveness in such scenarios, masking the severe recall loss required to maintain precision.

This limitation becomes apparent in the preliminary analysis of the WSVAD models. For instance, UR-DMU fails to detect about 15,000 abnormal segments even at a low threshold, yet maintains a high AUROC score of 0.8697 because it correctly classifies a large portion of normal segments [see Figures 2a and 2b] . Conversely, models like VadCLIP that leverages large pre-trained vision–language encoders such as CLIP achieve better recall at the expense of excessive false alarms. In the Figure 2d, VadCLIP misclassifies roughly 2×10^5 normal segments in the 0.9–1.0 prediction score bin, which inflates AUROC, as the false positive rate remains small relative to millions of normal



(a) URDMU Abnormal (b) URDMU Normal (c) VadCLIP Abnormal (d) VadCLIP Normal

Figure 2. Prediction score distributions from two WSVAD methods on UCF-Crime dataset, highlighting AUROC limitations. (a, b) URDMU misses 15k abnormal frames (low recall) but classifies nearly all normal frames correctly. (c, d) VadCLIP detects more abnormal frames, improving recall, but misclassifies 200k+ normal frames at a 0.9 threshold. Severe imbalance makes AUROC deceptively high.

segments. However, in a real-world surveillance setting, this is not a reliable solution, as a security officer needs to review tens and thousands of false alarms. Ultimately, these models exhibit a low practical recall at effective operating thresholds (see the results at Table 1).

The cost-sensitivity limitation of AUROC has been alleviated via weighted optimization methods based on AUC [16] or partial AUC (pAUC) [12] have been formulated. However, they are not suitable for WSVAD. Considering the significantly higher cost of false negatives in this context, we argue that recall evaluated at a meaningful FPR threshold, along with Average Precision, provides a more practical and effective evaluation metric.

3.3. Overview of the Dual Exploration Framework

The conventional approach in WSVAD is to obtain segment-level features from pretrained models. Although recent works increasingly prefer the CLIP model, as pointed out in the observation section, these methods tend to produce significant false alarms. Following the common practices and memory augmented baselines [32], we extract features from the I3D model pretrained on the Kinetics dataset [3], and pass them to the Global-Local Temporal Module and Memory Units (see Figure 3). The classifier then generates segment-level anomaly scores, which are used to compute MIL Loss. Most previous approaches then select only the top-k segments with the highest anomaly scores, restricting the model's ability to consider all segments in a positive bag, potentially overlooking other abnormal segments, and limiting its understanding of the full range of abnormalities present. To mitigate this bias, a broader exploration of segments within positive bags is essential. The model can make much more informed decisions if we allow it to observe a wider range of segments during training. However, adopting standard exploration strategies such as reinforcement learning is not possible due to the lack of fine-grained segment-level supervision. Therefore, we propose a soft exploration strategy that combines the selection of temporally diverse and ambiguous segments. The key idea is to enable the model to consider all potentially abnormal segments and make decisions based on a comprehensive view of the video.

3.4. Temporal Clustering based Selection (TCBS)

We adopt a clustering-based approach to encourage semantic diversity among selected segments. Video segments are clustered based on both feature and temporal similarity, ensuring that each cluster contains visually similar segments that occur closer in time. This encourages each cluster to represent a single scene while separating segments from different scenes into distinct clusters. During selection, we allow at most one segment to be chosen from any given cluster, promoting diversity across scenes without overrepresenting any single one.

Let $\mathbf{x}_i$ denote the feature representation of the ith segment. We define a similarity function $S(\mathbf{x}_i, \mathbf{x}_j)$ as the cosine similarity between segments. We perform temporal clustering as outlined in the Algorithm in the Supplementary Material. Specifically, the first segment initializes the first cluster. For each subsequent segment, if its similarity to the representative feature of the previous cluster ($\mathbf{x}_m^c$) exceeds the predefined threshold τ_s, it is added to that cluster; otherwise, a new cluster is created.

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} \tag{3}$$

We set the threshold τ_s based on the cosine similarity between adjacent segments within each video. For videos with drastic motion or scene changes, like explosions, the similarity between consecutive segments tends to be lower compared to more subtle anomaly scenarios, such as abuse or shoplifting. To make the threshold adaptive, we compute τ_s for each video individually by selecting the q-quantile (e.g., $q = 0.03$) of its similarity scores. The hyperparameter q controls the number of clusters formed. Lowering q leads to fewer clusters, since segments must exhibit greater dissimilarity to initiate a new cluster. The distribution of the similarity threshold τ_s across both datasets is presented in the Supplementary Material. As q represents the quantile of the similarity distribution, increasing q naturally leads to a higher similarity threshold. The UCF-Crime

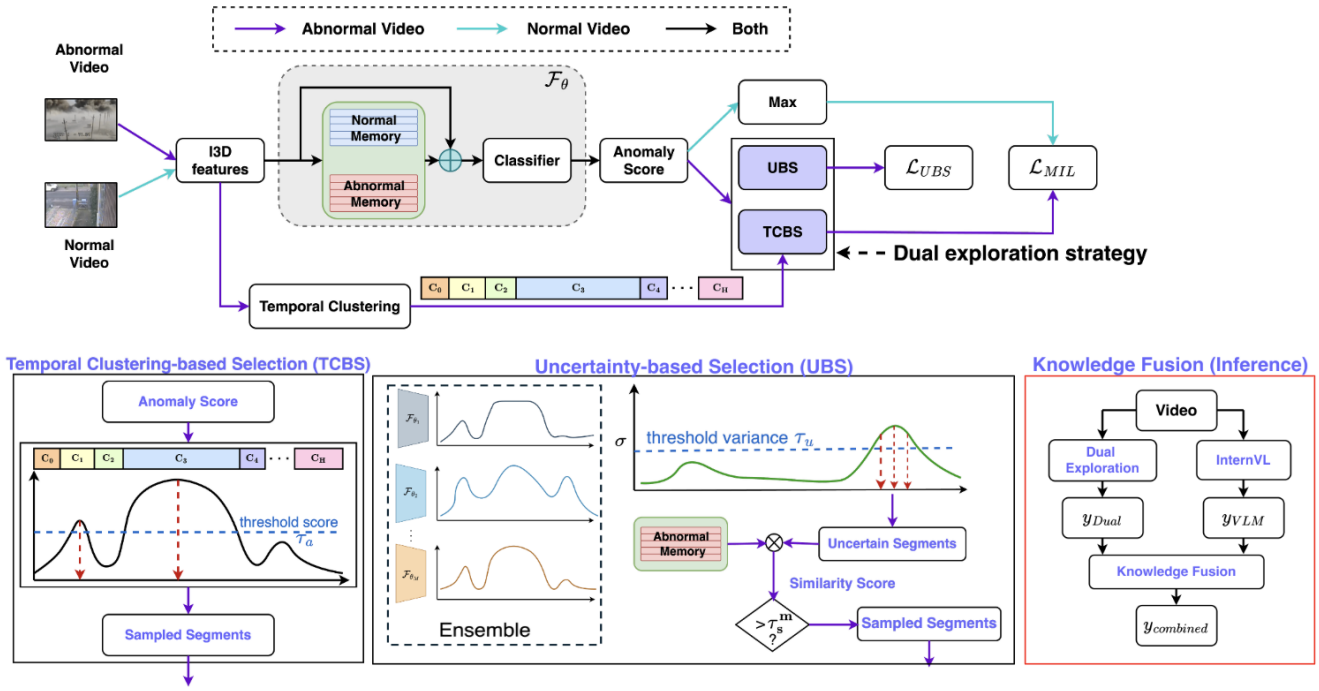


Figure 3. Overview of the proposed MIL training framework with a dual exploration strategy. After the model generates segment-level anomaly scores, we apply segment exploration and selection for computing the MIL loss. For normal videos, the segment with the highest anomaly score is selected to compute the MIL loss. For abnormal videos, segments are selected using a dual exploration strategy: (1) Temporal clustering-based selection, which chooses top-scoring segments from distinct temporal clusters to ensure diversity among selected events; and (2) Uncertainty-based selection, which targets segments with high predictive uncertainty and strong similarity to the abnormal memory, encouraging exploration of ambiguous yet informative patterns. During inference, the anomaly score from the dual exploration model is fused with the score generated by the VLM (InternVL) to obtain the final prediction score.

dataset has generally higher similarity thresholds because it consists surveillance videos with visually similar segments, whereas the XD-Violence dataset contains movie clips with varied camera angles and cuts. Despite these differences, our adaptive thresholding approach works effectively across both datasets.

To ensure selection from diverse abnormal events, we choose the segment with the highest anomaly score from each cluster. Let $f(\mathbf{x}_i^+)$ be the anomaly score of the ith segment. A segment is included in the set $\mathcal{K}$ only if its anomaly score exceeds a predefined threshold τ_a.

$$\mathcal{K} = \bigcup_{i=1}^{H} \{k : k = \arg\max_{j \in C_i} p_j, \quad f(\mathbf{x}_k^+) \geq \tau_a\} \qquad (4)$$

where H is the total number of clusters, k represents the segment with the highest anomaly score within cluster C_i, and τ_a is a dynamic threshold, which is chosen as a percentile of the anomaly scores computed over the segments of a given video.

TCBS ensures that segments representing abnormal events at different times are selected, rather than consecutively ranked segments that may belong to a single event.

3.5. Uncertainty-Based Selection

We observe segments in abnormal video with low anomaly score that may not be captured by temporal clustering alone, as it relies on certain anomaly score thresholds for detection (Figure in Supplementary Material). However, we find that these segments often exhibit high prediction uncertainty. Uncertainty-based exploration encourages the model to focus on such segments during the early training stages, allowing it to develop a robust understanding of abnormal patterns by prioritizing regions with high predictive uncertainty. Such exploration prevents the model from overfitting on highly confident segments, ensuring it becomes well-informed about its ambiguities.

Segment-level uncertainty u_i is obtained by calculating the standard deviations of the anomaly score obtained from an ensemble of M models $f(.)$ trained with different initialization. To differentiate abnormal segments from normal ones among the uncertain cases, we select the segments that closely resemble abnormal memory. We then define a set of candidate segments $\mathcal{U}$ that satisfy both criteria: (i) high uncertainty ($u_i > \tau_u$) and (ii) high semantic similarity to abnormal memory ($\frac{1}{|AM|}\sum_j S(\mathbf{x_i^+}, \mathbf{x_j^{AM}}) > \tau_s^m$),

Table 1. Results on UCF-Crime dataset. * indicates our own implementation

Method	Feature	AP (%)	Recall@FPR=1%	Recall@FPR=2%	Recall@FPR=3%
MIL [17]	I3D	25.03*	-	-	-
RTFM [18]	I3D	29.46*	-	-	-
BNSVP [14]	I3D	30.68*	-	-	-
UR-DMU [32]	I3D	35.48*	0.170	0.170	0.212
UR-DMU (various k):					
$k = 1$	I3D	33.94*	-	-	-
$k = 5$	I3D	33.54*	-	-	-
PEL4VAD [13]	I3D+Text	33.99	-	-	-
VadCLIP [24]	CLIP	33.55	0.109	0.155	0.217
InternVL3-14B [33]	Training Free	29.50	**0.301**	**0.301**	**0.418**
Ours (Dual Exploration + InternVL)	I3D + InternVL	**38.33**	<u>0.173</u>	<u>0.263</u>	<u>0.336</u>

where j indicates the memory slot, and $|AM|$ is the size of abnormal memory (refer to Algorithm in Supplementary Material). For segments in $\mathcal{U}$, we compute the BCE loss. Since uncertainty-based selection is applied only to abnormal videos, the target score for these segments is set to 1.

$$\mathcal{L}_{UNC} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \text{BCE}(f(\mathbf{x}_k^+), 1), \tag{5}$$

Uncertainty is calculated at each epoch and updated during the training. As the training starts with random initialization, the initial uncertainty estimates are unreliable. It is reasonable to use uncertainty estimates only after the model has undergone a few training steps. On the other hand, we aim to avoid overemphasizing uncertain segments, as they might overfit to noisy or borderline samples. Excessive focus on such segments can mislead the model and degrade it's generalization performance.

A good approach is to gradually reduce the influence of uncertainty as training progresses like implementing a cosine annealing scheduler. While we do not explicitly implement such a scheduler, we find that by carefully setting the hyperparameters τ_s^m (memory similarity threshold) and τ_u (uncertainty threshold), we can naturally obtain a similar scheduling effect. In the early training phase, the memory is randomly initialized, and the cosine similarity between memory entries and uncertain input segments is generally below 0.7. After a few training steps, as the model begins to learn abnormal patterns, the similarity of potential abnormal segments with abnormal memory starts to increase. Furthermore, by selecting uncertain but highly similar segments across all ensemble models and including them in training, the overall model uncertainty decreases over time.

Finally, we integrate MIL loss ($\mathcal{L}_{\text{MIL}}$) and uncertainty based selection loss ($\mathcal{L}_{\text{UNC}}$) along with the baseline model loss ($\mathcal{L}_{\text{URDMU}}$):

$$\mathcal{L}_{\text{Dual}} = \mathcal{L}_{\text{MIL}} + \gamma \mathcal{L}_{\text{UNC}} + \mathcal{L}_{\text{URDMU}} \tag{6}$$

3.6. Knowledge Fusion with VLM

In our approach, we primarily rely on abnormal memory for the exploration of uncertain samples, which is biased towards the model's prior knowledge of abnormality. As a result, truly novel or unseen anomalies that differ significantly from the memory unit may be overlooked. To address this limitation, we propose to leverage the generalization ability of powerful VLMs by fusing their prediction scores with the dual exploration model. We leverage the InternVL3 model [33] as the VLM.

Let y_{Dual} denote the prediction from the Dual Exploration Model trained with the loss $\mathcal{L}_{Dual}$, and y_{VLM} denote the prediction from the Vision-Language Model (VLM) pretrained on a large-scale dataset. y_{VLM} represents the VLM's confidence computed as the average of output across multiple runs using training-free inference with a structured prompt (see Supplementary Material). The combined prediction y_{combined} is obtained as a weighted sum of the two predictions, where the weight $\lambda \in [0, 1]$ is tuned for optimal performance.

$$y_{\text{combined}} = \lambda\, y_{\text{Dual}} + (1 - \lambda)\, y_{\text{VLM}} \tag{7}$$

4. Experiments

Datasets. We evaluate our model on two publicly available WSVAD benchmark datasets. The **UCF-Crime** dataset [17] consists of unedited surveillance footage covering 13 types of abnormal events. It contains 1,900 videos, in which the training split has 1,610 videos with video-level annotations, and the test set has 290 videos with segment-level annotations. The **XD-Violence** dataset [23] includes 4,754

Table 2. Results on XD-Violence dataset. * indicates our own implementation

Method	Feature	AP (%)	Recall@FPR=4%	Recall@FPR=5%	Recall@FPR=6%
MIL [17]	C3D	75.68	-	-	-
RTFM [18]	I3D	77.81	-	-	-
CRFD [22]	I3D	75.90	-	-	-
MSL [21]	V-Swim	78.28	-	-	-
UR-DMU [32]	I3D	79.14*	0.590	0.638	0.682
TSA [8]	CLIP	82.17	-	-	-
MACIL-SD [28]	I3D+audio	83.40	-	-	-
TPWNG [25]	CLIP	83.68	-	-	-
VadCLIP [24]	CLIP	84.50	0.635	0.672	0.701
InternVL3-14B [33]	Training Free	69.85	0.637	0.637	**0.755**
Ours (Dual Exploration + InternVL)	I3D + InternVL	**84.58**	**0.657**	**0.715**	<u>0.752</u>

Table 3. Ablation of different components on the UCF-Crime dataset.

Baseline	TCBS	UBS	VLM	AP	Recall@FPR=1%	Recall@FPR=2%	Recall@FPR=3%
✓				35.48	0.170	0.170	0.212
✓	✓			34.25	**0.182**	0.182	0.207
✓		✓		34.58	0.146	0.192	0.243
✓	✓	✓		36.42	0.167	0.232	0.256
✓	✓	✓	✓	**38.33**	0.173	**0.263**	**0.336**

untrimmed videos with only video-level labels, consisting of diverse content captured from movies and YouTube.

Evaluation metrics. The standard evaluation across all prior works uses AUROC for UCF-Crime and Average Precision (AP) for XD-Violence [17, 23]. However, as discussed earlier, AUROC tends to become overly optimistic and is not a reliable metric for safety-critical scenarios. Therefore, we evaluate both datasets using the **Area Under the Precision-Recall Curve**, and **Recall at a False Positive Rate (Recall@FPR)**. Recall@FPR $= \alpha$ measures the proportion of true anomalies correctly detected when the false positive rate is constrained to a fixed threshold α. For instance, Recall@FPR $= 1\%$ quantifies the recall achieved while keeping the false alarm rate below 1%. Unlike AUROC, which averages performance across all thresholds, including those corresponding to unrealistically high false positive rates, Recall@FPR provides a targeted measure of detection capability under practically feasible false alarm levels, making it more suitable for real-world deployment. We also compare the misclassification cost for the baseline URDMU and our exploration method by assigning a cost W of missing abnormal segments [2].

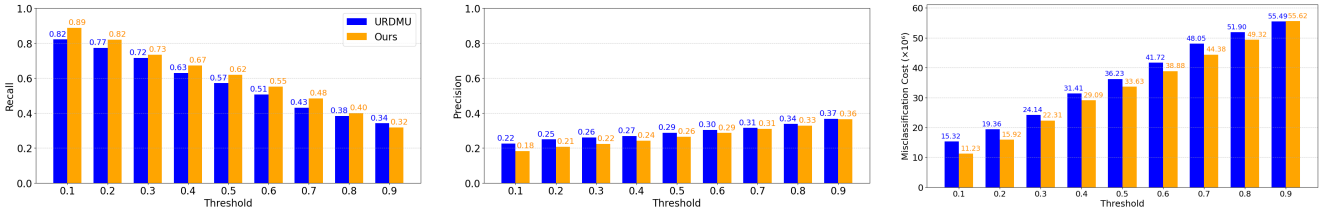
$$\texttt{MCC} = W \times \texttt{FN} + \texttt{FP} \tag{8}$$

where FN and FP are the count of false negatives and false positives, respectively, and W is the penalty assigned to false negatives.

Details on implementation, hyperparameters, and experimental setup are included in the Supplementary Material.

4.1. Comparison Results

We report AP and Recall@FPR for UCF-Crime and XD-Violence in Tables 1 and 2. Since most UCF-Crime baselines report AUROC, we re-computed their AP scores using available code and models. We also present the Recall@FPR of the best-performing open-source WSVAD models in single-modality and multi-modality. Our method, combining dual exploration and VLM fusion, achieves the highest AP (38.33%) among both weakly supervised models and training-free VLMs. This improvement stems from temporal clustering and uncertainty-based exploration, in contrast to baselines that rely only on top-k highest scoring segments. At 5% FPR on the UCF-Crime dataset model produces over 50,000 false alarms ($\approx$ 50 extra videos), making such thresholds impractical. We therefore also report recall at lower FPRs, where our method outperforms other weakly supervised approaches. The training-free InternVL model shows stronger recall at very low FPRs, making it valuable for fusion despite its lower AP score. Its higher recall at low FPR indicates that we can rely on its confident predictions, but the model confuses many less obvious anomalies with normal frames, leading to poor overall ranking and increased false positives at lower thresholds. Similarly, for XD-Violence, our method consistently



(a) Recall (b) Precision (c) Misclassification Cost

Figure 4. Comparison of recall, precision, and misclassification cost with the baseline (UR-DMU) across different thresholds on the UCF-Crime dataset.

Table 4. AUROC per event type for Dual Exploration and VLM on UCF-Crime.

Method	**AUROC (Abnormal)**	**Abuse**	**Arrest**	**Arson**	**Assault**	**Burglary**	**Explosion**
Dual Exploration	0.708	**0.81**	0.64	**0.60**	0.89	**0.77**	0.47
InternVL	0.700	0.61	**0.70**	0.58	**0.92**	0.64	**0.77**
Method	**Fighting**	**Accidents**	**Robbery**	**Shooting**	**Shoplifting**	**Stealing**	**Vandalism**
Dual Exploration	**0.79**	0.68	**0.83**	**0.77**	**0.66**	**0.85**	**0.90**
InternVL	0.76	**0.82**	0.81	0.74	0.52	0.84	0.76

achieves higher recall than baselines, even at FPRs of 4% and above.

4.2. Ablation Study

Effectiveness of Dual Exploration: Table 3 shows the effect of individual exploration components. Temporal clustering-based exploration achieves the highest recall at 1% FPR because it selects segments from different clusters rather than consecutive frames, ensuring coverage of diverse parts of the video. By limiting selection to one segment per cluster, it also reduces the likelihood of multiple false alarms from the same region. In contrast, uncertainty-based exploration alone yields lower recall at very low FPR, as it may select noisy normal segments that are semantically similar to anomalies. However, as FPR increases to 2–3% within the practical operating range, it captures segments that temporal clustering (TCBS) misses, specifically, segments with lower anomaly scores but high model uncertainty. By combining both strategies, our method achieves better practical recall, as shown by improvements at 2% and 3% FPR, and better AP despite a few noisy selections. Figure 4 demonstrates improved recall of dual exploration across all prediction thresholds, except at 0.9 due to TCBS, with minimal loss in precision and a lower misclassification cost (W = 1000) compared to the baseline model.

Effectiveness of VLM Fusion: To understand the effectiveness of VLM fusion, we report AUROC separately for abnormal videos and for different types of abnormal events using dual exploration prediction and VLM prediction. In Table 4, we observe that VLM shows stronger detection capability for certain event types, such as Explosion, Arrest, and Road Accidents, but performs poorly on human-centric abnormal events that occur within short temporal windows (e.g., Abuse and Shooting) as well as events that require temporal reasoning (e.g., Shoplifting). This limitation may arise from the lack of reasoning ability in VLMs trained on generalized domains. In contrast, our dual exploration method achieves comparatively better performance on these challenging categories. Therefore, combining dual exploration with VLM enables more effective detection by leveraging knowledge sharing between generalized anomalies and domain-specific anomalies.

5. Conclusion and Future Work

In this work, we emphasize the need for cost-sensitive evaluation metrics for WSVAD, as commonly used metrics like AUROC can either mask poor recall or obscure significant false alarm rates, leading to a low practical recall that limits real-world applicability. To improve recall, we introduce a soft exploration strategy that informs the model about potentially diverse and ambiguous abnormal events during the training process, rather than being restricted to top-k segments. Additionally, we introduce a fusion approach that leverages knowledge sharing between vision-language models (VLMs) trained on generalized domains and exploration models trained on specialized domains to mitigate the exploration model's bias toward its learned abnormal features during uncertainty exploration. In the future, we plan to extend this knowledge sharing to the training phase, enabling both VLMs and weakly supervised models to improve their detection capabilities.

References

[1] 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*, pages 10275–10285, 2025. 1

[2] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995. 7

[3] 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*, pages 6299–6308, 2017. 1, 4

[4] 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*, pages 18319–18329, 2024. 1, 2, 3

[5] 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*, pages 24185–24198, 2024. 3

[6] 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. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[8] 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)*, pages 3230–3234. IEEE, 2023. 1, 3, 7

[9] 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. 3

[10] 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*, pages 8022–8031, 2023. 1, 2, 3

[11] 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*, pages 24265–24274, 2025. 1

[12] Andreas Maurer and Massimiliano Pontil. Estimating weighted areas under the roc curve. *Advances in Neural Information Processing Systems*, 33:7733–7742, 2020. 1, 4

[13] 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. 6

[14] 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)*, pages 3212–3221, 2022. 2, 3, 6

[15] Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally robust optimization for deep kernel multiple instance learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 2188–2196. PMLR, 2021. 2, 3

[16] Huiyang Shao, Qianqian Xu, Zhiyong Yang, Peisong Wen, Gao Peifeng, and Qingming Huang. Weighted roc curve in cost space: Extending auc to cost-sensitive learning. *Advances in Neural Information Processing Systems*, 36: 17357–17368, 2023. 1, 4

[17] 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, 2018. 1, 2, 6, 7

[18] 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*, pages 4975–4986, 2021. 1, 2, 6, 7

[19] 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*, pages 4489–4497, 2015. 1

[20] 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. 3

[21] Dong-Lai Wei, Chen-Geng Liu, Yang Liu, Jing Liu, Xiao-Guang Zhu, and Xin-Hua Zeng. Look, listen and pay more attention: Fusing multi-modal information for video violence detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1980–1984. IEEE, 2022. 7

[22] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021. 7

[23] 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*, pages 322–339. Springer, 2020. 6, 7

[24] 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*, pages 6074–6082, 2024. 1, 2, 3, 6, 7

[25] 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*, pages 18899–18908, 2024. 7

[26] 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*, pages 8679–8688, 2025. 3

[27] 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. *arXiv preprint arXiv:2511.10334*, 2025. 1

[28] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6278–6287, 2022. 7

[29] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate probability estimates. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002. 1

[30] 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. 3

[31] 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. 3

[32] 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*, pages 3769–3777, 2023. 1, 2, 3, 4, 6, 7

[33] 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. 3, 6, 7