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.

Learning from Noisy Supervision: A Denoising–Debiasing Framework for Weakly Supervised Video Anomaly Detection

Yaxin Zhao[1]*, Yang Wang[1]*, Wenya Guo[1], Sihan Xu[1], Xiangrui Cai[1]†, Xi Lin[2], Ying Zhang[1], Xiaojie Yuan[1]
[1]VCIP, TMCC, DISSec, College of Computer Science, Nankai University
[2]College of Computer Science, Shanghai Jiao Tong University
yaznz@mail.nankai.edu.cn, wangyang@dbis.nankai.edu.cn, caixr@nankai.edu.cn

Abstract

Weakly supervised video anomaly detection (WS-VAD) aims to localize frame-level anomalies using only video-level labels. This task is typically formulated within a multiple instance learning (MIL) paradigm, where each video is treated as a bag of snippets, achieving robust performance without requiring additional information. However, existing methods often struggle with noisy supervision signals. Normal snippets within abnormal bags are frequently misclassified as anomalies due to inaccurate anomaly scores. These misclassified instances act as noisy samples, introducing false supervision that hinders the learning of true anomaly patterns. In this work, we introduce D^{2}MIL, a Denoising–Debiasing framework within the Multiple Instance Learning paradigm designed to suppress noise and improve anomaly discrimination. Our approach integrates two key components: (1) Denoising Module: We introduce a dynamic drop rate to adaptively filter out suspected noisy samples during training, based on the observation that noisy samples incur higher training losses. (2) Debiasing Module: We leverage a vision-language model to re-evaluate the discarded samples. This recovers potentially valuable abnormal instances that were mistakenly removed, as they are similar to noisy samples but difficult for the model to recognize. D^{2}MIL is a general purpose denoising strategy that can be integrated into any MIL-based method. Our extensive experiments on the three benchmark datasets (ShanghaiTech, UCF-Crime, and MSAD) demonstrate that D^{2}MIL is compatible with diverse MIL frameworks and consistently enhances their performance.

1. Introduction

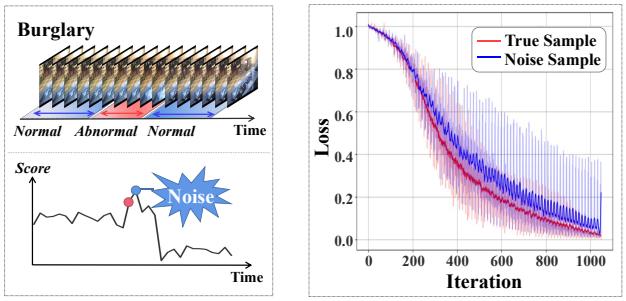
Video Anomaly Detection (VAD) is a critical task in computer vision, aimed at identifying abnormal events that deviate from typical behavioral patterns in video sequences [11, 13], with widespread applications in public safety surveillance [14]. To enhanced generalization across diverse scenes and reduced annotation costs, researchers turn to weakly supervised settings.

Weakly Supervised Video Anomaly Detection (WS-VAD) relies solely on video-level binary labels (normal/abnormal) for training, has emerged as a widely studied research direction [6, 24, 28]. Among various WS-VAD approaches, the Multiple Instance Learning (MIL) framework has emerged as the dominant paradigm due to its effectiveness [19, 22, 25, 34]. In MIL, each video is treated as a "bag" and its snippets are regarded as "instance". Normal videos are treated as negative bags in which all instances are normal, while abnormal videos are regarded as positive bags containing at least one abnormal instance. Typically, models identify the top-k high-confidence instances from positive bags as pseudo-anomalies, using them to guide instance-level representation learning. [3, 7, 9, 17]

Despite the strong adaptability of the MIL framework to WS-VAD, it remains highly susceptible to label noise during training. The noise primarily arises from the inaccurate selection of pseudo-anomalous instances within abnormal videos. Specifically, abnormal videos typically contain far more normal snippets than truly anomalous ones. However, due to unreliable anomaly score estimation in the early training stages, some normal instances are mistakenly assigned high anomaly scores and selected as part of the top-k pseudo-anomalous set. These incorrectly selected instances, referred to as noisy samples, introduce false supervision signals that disrupt the model's ability to learn genuine anomaly patterns.

Our preliminary experiments confirm this noise issue. As illustrated in Figure 1a, in the Burglary example from UCF-Crime, the MIL model assigns the highest anomaly score to a normal instance, while the true abnormal instance is ranked second. This highlights a selection bias that misleads the training process. Furthermore, Figure 1b shows that such noise persists throughout training: misselected normal instances (noise) consistently exhibit high

[1]*Equal contribution. † Corresponding author.



(a) (b)

Figure 1. **(a)** An abnormal video example from "Burglary" category, with abnormal (red) and normal (blue) intervals, and the MIL model's anomaly score curve below. Highlighted points indicate the highest-scoring (blue) and second-highest-scoring (red) instances. **(b)** Loss trend of selected top-k samples during training on UCF-Crime dataset, with the red curve for true abnormal samples (low, stable decline) and the blue curve for noise samples (high, volatile). The shaded area represents raw loss distribution, and the dark line shows the smoothed moving average.

and unstable losses, in contrast to the gradually converging loss curve of true abnormal instances. Motivated by this observation, we propose a **Dynamic Threshold-based Denoising Module** that leverages the high-loss characteristics of noisy samples to adaptively filter out suspected noise, thereby mitigating such misleading effects.

In addition, we observe that some truly abnormal instances also exhibit high training losses, making them indistinguishable from noisy samples (as shown in the red-blue shaded overlap in Figure 1b). These high-loss abnormal instances are typically hard examples that the model struggles to recognize. Simply discarding all high-loss instances risks removing such hard samples, leading the model to overfit easy instances and impairing its generalization ability. To address this, we introduce a **VLM-based Debiasing Module**. Leveraging the powerful visual understanding and reasoning capabilities of vision-language models (VLMs), we progressively recall potential hard samples from the filtered high-loss candidates under the guidance of a curriculum learning strategy [1].

In summary, we propose a novel framework, **Denoising and Debiasing within Multiple Instance Learning (D^{2}MIL)**, which integrates denoising and debiasing modules to enable robust and generalizable WS-VAD. The main contributions of this paper are as follows:

(1) We are the first to explicitly distinguish noisy samples from hard abnormal samples within MIL framework. Our empirical analysis reveals that high-loss samples contain both harmful noise and valuable hard instances.

(2) We propose the D^{2}MIL framework with two complementary modules: a Denoising Module that adaptively suppresses noisy snippets via loss, and a Debiasing Module that uses visual language reasoning to recall informative hard samples mistakenly filtered out.

(3) D^{2}MIL can be integrated into existing MIL-based methods, leading to steady performance improvements across multiple baselines on the ShanghaiTech, UCF-Crime, and MSAD datasets.

2. Related Work

Weakly Supervised Video Anomaly Detection detects and localizes abnormal events using only video-level binary labels, offering a cost-effective alternative to fully supervised methods [10]. A dominant paradigm is MIL, which treats videos as bags of snippets for instance-level anomaly inference, and was first introduced for deep anomaly detection by Sultani et al.[19], establishing a foundational benchmark. RTFM [20] introduces a feature magnitude learning function to alleviate bias from dominant negative instances. TEVAD [4] uses captions to improve VAD by combining semantic textual guidance with visual features to improve discriminability. MGFN [5] proposed a magnitude-contrastive loss and integrated a glance-and-focus strategy to better distinguish anomalies from normal patterns. UR-DMU [34] employs dual memory units to learn discriminative representations of normal and abnormal data. Despite these advances, most MIL-based methods overlook the detrimental effects of noisy supervision during training.

Several recent studies attempt to alleviate noise and bias in MIL-based video anomaly detection. UMIL [15] mitigates bias in MIL-based anomaly detection by separating confident and ambiguous segments to achieve invariant feature learning. Park et al. [16] propose Normality-Guided MIL, which models normal patterns with learned prototypes to refine unreliable MIL predictions. ProDisc-VAD [36] takes an initial step toward addressing this issue by introducing a progressive discrepancy learning framework to mitigate label ambiguity. However, it fails to capture the distinctive loss patterns of noisy samples and does not account for hard abnormal samples. In contrast, D^{2}MIL explicitly addresses both noise and bias through a two-stage collaborative process, leading to more reliable anomaly localization and improved generalization.

Furthermore, there are also studies based on external information. Pu et al.[17] enriches contextual feature representations by leveraging semantic prompts. Chen et al. [3] introduces class-aware textual prompts to provide semantic guidance during anomaly instance selection. Similarly, LEC-VAD [22] incorporates external class-descriptive information to model the completeness of anomalous events. While prompt-enhanced methods show promise, they rely on class-specific prior knowledge. By contrast, our approach requires no additional annotations. It leverages a frozen VLM as a general semantic prior to assist debiasing without introducing extra supervision.

Large Multimodal Models for Video Anomaly Detection. In recent years, Large Language Models (LLMs) and VLMs have been explored in VAD tasks to enhance semantic understanding and interpretability. For example, Holmes-VAD [29] innovates by constructing the large-scale VAD-Instruct 50K benchmark with single-frame annotations and fine-tuned multi-modal LLM to achieve unbiased VAD. However, methods based on fine-tuning LLMs incur excessive computational overhead. VadCLIP [26] adapts CLIP for WS-VAD by aligning snippet-level visual features with class-aware textual prompts. VERA [27], in contrast, achieves interpretable anomaly detection via verbalized learning using off-the-shelf VLMs without requiring any fine-tuning. Inspired by these methods, we leverage a frozen VLM as an auxiliary module to re-evaluate discarded high-loss candidates to recall potential hard abnormal samples. Importantly, our method does not require any additional fine-tuning procedures.

3. Method

In this section, we present our proposed framework, **Denoising and Debiasing in MIL**, as illustrated in Figure 2. It comprises two modules: a dynamic threshold-based denoising module and an VLM-based debiasing module. The denoising module dynamically identifies and filters noisy samples during the training of MIL models by applying a loss-based thresholding mechanism. On top of this, the debiasing module leverages the visual reasoning capabilities of VLM to identify and recover hard samples that may have been mistakenly removed during the denoising process. In the following subsections, we first formalize the problem setting and then elaborate on the design of each module.

3.1. Problem Formulation

We follow the task setting of WS-VAD [28, 32], where each training video is annotated only with a binary video-level label $Y \in \{0, 1\}$, indicating whether the video contains anomalous events. Frame-level annotations are not available. The goal is to perform fine-grained frame-level anomaly localization under this weak supervision setting.

To address this, we adopt the widely-used MIL paradigm [3, 23]. In this framework, each video v is treated as a *bag* and divided into m non-overlapping snippets $\{s_1, s_2, \ldots, s_m\}$, each regarded as an *instance*. We denote normal videos as v^n with bag label $Y^n = 0$, implying that all instances are normal. Conversely, abnormal videos are denoted as v^a with bag label $Y^a = 1$, indicating that at least one instance in the bag is anomalous.

The objective is to learn an anomaly scoring function $f(v)$ that assigns a instance level anomaly score to each video. In typical MIL methods, the top-k highest-scoring instances are selected from both normal and abnormal bags. A contrastive loss is then applied to minimize the anomaly scores of instances from normal videos, while maximizing those from abnormal ones, enabling the model to focus on discriminative anomalous patterns and achieve frame-level anomaly localization.

3.2. Stage1: Dynamic Threshold-based Denoising Module

We take the first MIL method [19] as a representative example. In each training batch, we sample b abnormal videos (B_a) and b normal videos (B_n), where the number of both types is equal. Then, the instances with the highest anomaly scores from the normal and abnormal video bags are selected to form training pairs. Later works [34] extend this paradigm by selecting the top-k scoring instances instead of just one. In this work, we use the $k = 1$ case as an illustrative example, while the underlying principle remains consistent. The model is trained with a contrastive loss to encourage higher anomaly scores for abnormal instances. The loss function is shown in Eq.(1):

$$L(B_a, B_n) = \max(0, 1 - \max_{i \in B_a} f(v_i^a) + \max_{i \in B_n} f(v_i^n)), \quad (1)$$

where $f(v_i^a)$ and $f(v_i^n)$ represent the anomaly scores of all instances in the i-th bag of the anomaly batch B_a and the normal batch B_n, respectively. And max denotes selecting the instance s_i^a and s_i^n with the highest anomaly scores for the bags v_i^a and v_i^n, respectively. The loss encourages the highest-scoring instance from an abnormal video to have a significantly higher score than that from a normal video, enforcing discriminative learning between anomalous and normal patterns.

Specifically, for each normal and abnormal bag (v_i^n, v_i^a) in the batch, we calculate the contrast loss according to Eq.(1) and obtain the loss sequence $Q = \{l_1, l_2, \ldots, l_b\}$, where each element l_i corresponds to the loss between the i-th normal and anomaly bags. We then sort the loss values in ascending order to obtain Q_s:

$$\begin{aligned} Q_s = \{l_{(1)}, l_{(2)}, \ldots, l_{(b)}\}, \\ \text{where } l_{(1)} \leq l_{(2)} \leq \cdots \leq l_{(b)} \end{aligned} \quad (2)$$

where $l_{(i)}$ represents the i-th smallest loss in the batch. Next, we design a dynamic dropout rate to filter out noisy samples during training. Specifically, based on the sorted loss sequence Q_s, we compute the low-loss subset Q_{keep} and the high-loss subset Q_{drop} according to the current dropout rate. The calculation process is shown in Eq.(3),

$$Q_{\text{keep}} = \mathcal{Q}_s\big[: (1 - \tau(T))b\big], \quad (3)$$

where $\tau(T) \in [0, 1]$ denotes the dynamic drop rate at iteration T of the current epoch, which increases gradually

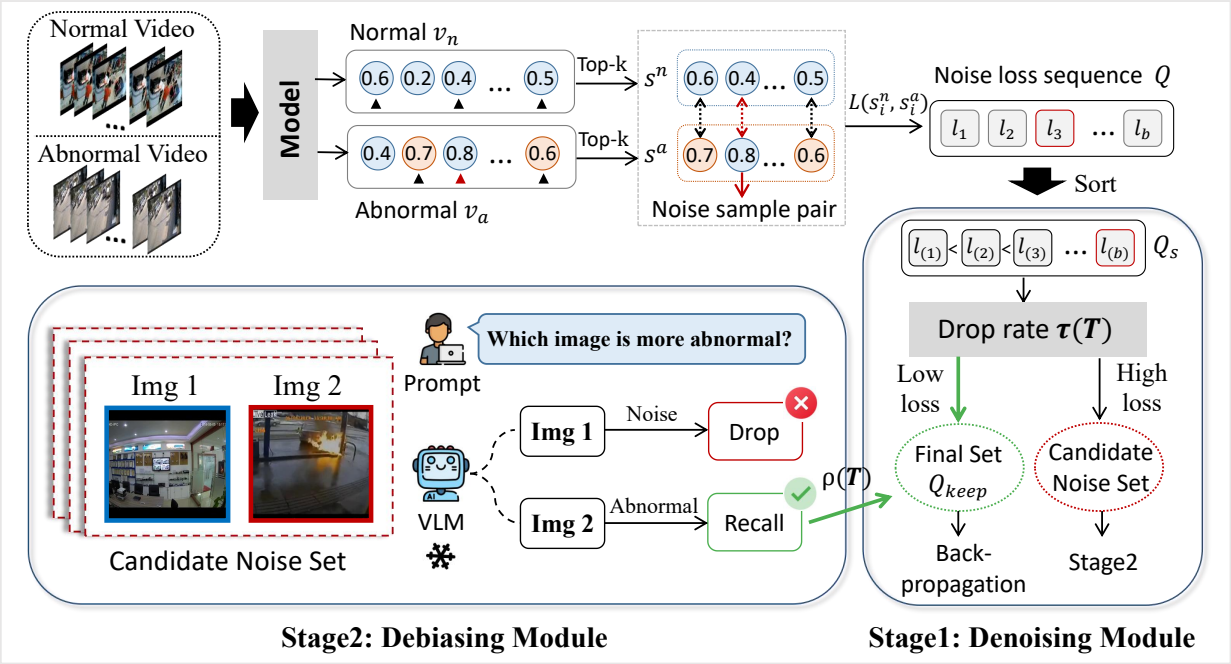


Figure 2. Our proposed D^{2}MIL framework consists of two key modules: a Denoising Module, which adaptively filters out high-loss instances based on a dynamic drop rate, forming a candidate noise set; and a Debiasing Module, which leverages a frozen vision-language model to recover mistakenly discarded abnormal instances. The final loss is computed using a combination of the retained low-loss instances and the recalled hard samples.

over iterations within each epoch. We adopt a simple yet effective linear scheduling strategy for the drop rate, which avoids introducing additional hyperparameters. This design improves controllability and ease of use, allowing the optimal noise filtering rate to be efficiently determined across different datasets with minimal tuning effort. The calculation formula is shown in Eq.(4),

$$\tau(T) = \tau_{\max} \cdot \frac{T}{T_{\text{total}} - 1}, \tag{4}$$

where $\tau_{\max}$ is the maximum drop rate, and T_{total} is the total number of training iterations within one epoch. This strategy enables the model to leverage most samples to learn general patterns at the beginning of each epoch, then progressively reduces the interference from noisy samples, thus facilitating more robust anomaly pattern learning.

The whole algorithm is explained in Alg. 1. The high-loss sample set Q_{drop} discarded after noise filtering is saved and passed to the subsequent debiasing module for further processing.

3.3. Stage2: VLM-based Debiasing Module

While the denoising module effectively removes high-risk noisy samples during training, we observe that some true abnormal instances may be mistakenly filtered out due to their similarity to noise. These high-loss instances are often hard samples that the model struggles to recognize, typically reflecting ambiguous anomaly patterns that have not yet been learned. Over-pruning such samples can cause the model to focus excessively on easier patterns, thereby limiting its generalization ability in real-world scenarios.

To address this issue, we introduce a VLM-based Debiasing Module that leverages zero-shot visual reasoning capabilities to selectively recover potentially informative hard

Algorithm 1 Dynamic Threshold-based Denoising

Require: batches B_n, B_a; total steps T_{total}; warm-up T_{warm}; max drop rate $\tau_{\max}$

for $T = 1$ **to** T_{total} **do**
  Sample b normal and abnormal video bags $B_n = \{v_1^n, \ldots, v_b^n\}$, $B_a = \{v_1^a, \ldots, v_b^a\}$
  Select top-scoring instances s_i^n, s_i^a
  Compute MIL loss with Eq.(1)
  Sort all l_i in ascending order to obtain Q_{sorted}
  Set drop rate $\tau(T)$
  Keep lowest-loss instance pairs $Q_{\text{keep}} = \lfloor (1 - \tau(T)) \cdot b \rfloor$
  Store discard high loss noise sets Q_{drop} for debiasing
end for

samples from the set of discarded candidate noise set Q_{drop} . This module is applied after denoising, aiming to preserve high-value anomaly instances without significantly increasing the risk of introducing noise.

Specifically, for each high-loss $l'_i \in Q_{\text{drop}}$ discarded by the denoising module, we retrieve its corresponding instance pair (s'^n_i, s'^a_i), along with the associated video identifier and segment index. The normal instance s'^n_i, sampled from a normal video bag, serves as a true negative. In contrast, the abnormal instance s'^a_i, previously identified as suspected noise, is the main subject of re-evaluation.

Following prior work [2, 21] that demonstrates the effectiveness of single-frame representations, we specifically leverage the middle frame of each snippet, as it is more likely to capture the semantic center of local actions. Given each normal-abnormal pair, we extract the middle frame of the snippet as a representative keyframe $(\text{img}^n_i, \text{img}^a_i)$ and feed them into the vision-language model Qwen-VL-Max [31]. This model exhibits state-of-the-art cross-modal reasoning capabilities, enabling it to compare the abnormality level between two images in a zero-shot manner.

We design a conservative prompt to guide the model's decision process:

> *"You are a surveillance anomaly analyst. Given two keyframe images (Image 1, Image 2), determine which image clearly shows an abnormal event. Return 0 for Image 1, 1 for Image 2, and 2 if uncertain. Only answer 0 or 1 when the abnormality is obvious and you are highly confident. If unsure, return 2."*

This prompt encourages the model to assess anomaly likelihood conservatively. If the abnormal instance is judged as more anomalous, it is regarded as a potentially misfiltered true anomaly and added to the recalled set. Otherwise, if the normal instance appears more anomalous or the model expresses uncertainty, the candidate is treated as noise and permanently discarded. Given the complexity of real-world video data, we adopt a conservative strategy to minimize false positives, preferring to miss difficult samples rather than reintroduce noise into the model

After sample filtering, we adopt a curriculum learning strategy [1, 8, 30]. This strategy gradually incorporates more hard samples from the recalled set Q_{recall}, aiming to avoid model instability caused by the abrupt reintroduction of hard samples. Similar to the dropout rate $\tau(T)$ used in the denoising module, we employ a linear scheduling strategy to progressively increase the recall ratio$\rho(T)$ of these hard samples, following the same formulation as Eq.(4). This schedule encourages the model to focus on learning simple patterns at the early stage of each epoch while gradually increasing exposure to harder samples, enabling the model to capture richer anomaly patterns. Finally, the recalled hard sample set Q_{recall} is combined with the low-loss set Q_{keep} from the denoising module to compute the final training loss and update the model via backpropagation.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We conduct experiments on three widely-used WS-VAD evaluation datasets:

- **ShanghaiTech** [12] is a medium-scale dataset comprising 437 surveillance videos recorded from 13 fixed-camera street scenes. This dataset was originally designed for unsupervised anomaly detection, where only normal samples are used for training. Following the setup in Zhong et al. [33], we adapt it to a weakly supervised scenario to evaluate our method.
- **UCF-Crime** [19] is a large-scale, real-world dataset containing 1,900 untrimmed videos, recorded in diverse indoor and outdoor environments. The dataset includes 13 anomalous event categories. For training, video-level labels are provided for 1,610 videos, while frame-level annotations are available for 290 test videos, enabling precise frame-level evaluation.
- **MSAD** [35] is a recently introduced high-resolution dataset, consisting of 720 videos collected from 14 distinct scenarios. These scenarios cover varying indoor/outdoor contexts, lighting conditions, and weather variations. It contains 35 human-related and 20 non-human-related anomaly types. We utilize video-level labels for 360 normal and 120 abnormal videos, and conduct evaluations on a test set consisting of 240 abnormal and 120 normal videos.

Evaluation Metrics. Following previous works [20, 34, 35], we used the Area Under the Curve (AUC) of the frame level ROC (Receiver Operating Characteristic) as the main evaluation metric for three datasets.

4.2. Implementation Details

We follow the original hyperparameter settings and model configurations of each baseline to ensure fair evaluation. Specifically, we adopt the settings from [18] for Sultani et al., [35] for RTFM and MGFN, [4] for TEVAD, and [34] for UR-DMU. For the Sultani et al. baseline, we adjust the dropout probability ($drop_p$) to ensure stable training under our setup, while keeping all other configurations unchanged.

For D^{2}MIL , we introduce two additional parameters: the maximum noise drop rate τ_{max} and the maximum hard sample recall rate ρ_{max}. Based on the parameter analysis experiments, we determine the optimal parameter settings for each baseline. For the Sultani et al. method, which selects the top-1 feature, the best performance is achieved with $\tau_{\text{max}} = 0.5$ and $\rho_{\text{max}} = 0.1$ as shown in Table 7. For

Method	Feature	AUC(%)
MIST (CVPR 2021)	I3D	94.83
RTFM (ICCV 2021)	I3D	97.03
MSL (AAAI 2022)	I3D	95.45
TEVAD (CVPR 2023)	I3D + Text	97.80
MGFN (AAAI 2023)	I3D	93.97
UR-DMU (AAAI 2023)	I3D	96.15
ProDisc-VAD (arxiv 2025)	ViT	97.98
RTFM + **D^{2}MIL**	I3D	**97.93** (+0.90)
TEVAD + **D^{2}MIL**	I3D + Text	**98.13** (+0.33)
MGFN + **D^{2}MIL**	I3D	**94.13** (+0.16)
UR-DMU + **D^{2}MIL**	I3D	**97.44** (+1.29)

Table 1. AUC results on the ShanghaiTech dataset. Gray rows indicate baseline methods enhanced by D^{2}MIL . Improvements over original methods are shown in parentheses.

Method	Feature	AUC(%)
Sultani et al. (CVPR 2018)	C3D	75.41
Sultani et al. (CVPR 2018)	I3D	76.21
Sultani et al. (2021)	I3D	83.34
MIST (CVPR 2021)	I3D	82.30
RTFM (ICCV 2021)	I3D	83.14
MSL (AAAI 2022)	I3D	85.30
TEVAD (CVPR 2023)	I3D + Text	84.54
MGFN (AAAI 2023)	I3D	83.30
UR-DMU (AAAI 2023)	I3D	86.19
UMIL (CVPR 2023)	I3D	86.75
Holmes-VAD (CVPR 2024)	Vit	84.61
VERA (CVPR 2025)	VLM	86.55
ProDisc-VAD (arxiv 2025)	ViT	87.12
Sultani et al. + **D^{2}MIL**	I3D	**85.24** (+1.90)
RTFM + **D^{2}MIL**	I3D	**84.30** (+1.16)
TEVAD + **D^{2}MIL**	I3D + Text	**84.76** (+0.22)
MGFN + **D^{2}MIL**	I3D	**83.70** (+0.40)
UR-DMU + **D^{2}MIL**	I3D	**87.80** (+1.61)

Table 2. AUC results on the UCF-Crime dataset. Notation follows Table 1. No Instruction fine-tuning is used for HolmesVAD.

other baselines that adopt top-k feature selection, the optimal performance is typically obtained with τ_{max} ranging from 0.1 to 0.2 and ρ_{max} around 0.1.

4.3. Comparisons with SOTA Methods

We integrate the D^{2}MIL framework into several representative MIL-based baselines to validate the effectiveness of our proposed approach.[4, 5, 19, 20, 34] These include: Sultani et al., the earliest MIL method; RTFM and MGFN, which extend the top-1 instance selection to top-k; TEVAD, the first to incorporate text information into visual anomaly

Method	Feature	AUC(%)
RTFM (ICCV 2021)	I3D	86.65
TEVAD (CVPR 2023)	I3D + Text	86.82
MGFN (AAAI 2023)	I3D	84.96
UR-DMU (AAAI 2023)	I3D	85.02
RTFM + **D^{2}MIL**	I3D	**88.52** (+1.87)
TEVAD + **D^{2}MIL**	I3D + Text	**87.94** (+1.12)
MGFN + **D^{2}MIL**	I3D	**85.32** (+0.36)
UR-DMU + **D^{2}MIL**	I3D	**86.68** (+1.66)

Table 3. AUC results on the MSAD dataset. Notation follows Table 1.

detection; and UR-DMU, a state-of-the-art method under the current MIL framework on the UCF-Crime dataset. To ensure a fair and focused comparison, we adhere to the standard WS-VAD setting by excluding methods that leverage additional information (e.g., anomaly category descriptions) and limiting LLM-based methods to results obtained without instruction fine-tuning. This evaluation protocol more accurately reflects real-world scenarios where supervision and resources are constrained.

Results on ShanghaiTech. The frame-level AUC results on the ShanghaiTech dataset are presented in Table 1. Our proposed D^{2}MIL method consistently enhances performance across all MIL-based baselines. Specifically, integrating D^{2}MIL brings about AUC improvements for each baseline: 0.90% over RTFM, 0.33% over TEVAD, and 0.16% over MGFN, 1.29% over UR-DMU. Even though most baselines already attain strong performance on this dataset, our method still maintains stable enhancements. Meanwhile, TEVAD reaches a score of 98.13% and outperforms ProDisc-VAD[36], a method designed to address ambiguous labels.

Results on UCF-Crime. The frame-level AUC results on the UCF-Crime dataset are presented in Table 2. Across representative MIL-based baselines, our proposed D^{2}MIL method delivers consistent performance boosts. Specifically, it yields a 1.90% improvement over the Sultani et al. method (reproduced using I3D two-stream features) [18], 1.16% over RTFM, 0.22% over TEVAD, 0.40% over MGFN, and 1.61% over UR-DMU. Notably, integrating D^{2}MIL into UR-DMU yields an AUC of 87.80%, establishing a new state-of-the-art among MIL-based methods. It also outperforms the LLM-based Holmes-VAD and the VLM-based VERA, further demonstrating the effectiveness of our approach on large-scale datasets and its ability to substantially enhance MIL model.

Results on MSAD. As a recently released dataset with diverse scenarios, MSAD is notably more complex to evaluate, with limited existing results. As shown in Table 3, our

method still achieves stable gains. Specifically, we observe improvements of 1.87% over RTFM, 1.12% over TEVAD, 0.36% over MGFN, and 1.66% over UR-DMU. Notably, the RTFM + D^{2}MIL combination attains a state-of-the-art performance of 88.52% on this dataset, underscoring the efficacy of our method across diverse scenarios.

4.4. Ablation Studies

Baseline	**Raw**(%)	**Denoise-only**(%)	**D^{2}MIL** (%)
RTFM	97.03	97.85	**97.93**
TEVAD	97.80	98.07	**98.13**
MGFN	93.97	94.11	**94.13**
UR-DMU	96.15	97.35	**97.44**

Table 4. Ablation studies on the ShanghaiTech dataset. We compare the performance of each baseline method under three settings: the original baseline, the baseline with the denoising module, and with the full D^{2}MIL framework.

Baseline	**Raw**(%)	**Denoise-only**(%)	**D^{2}MIL** (%)
Sultani et al.	83.34	84.77	**85.24**
RTFM	83.14	83.96	**84.30**
TEVAD	84.54	85.72	**85.83**
MGFN	83.44	83.60	**83.70**
UR-DMU	86.19	87.53	**87.90**

Table 5. Ablation studies on the UCF-Crime dataset.

Baseline	**Raw**(%)	**Denoise-only**(%)	**D^{2}MIL** (%)
RTFM	86.65	88.03	**88.52**
TEVAD	86.82	87.28	**87.94**
MGFN	84.96	85.13	**85.32**
UR-DMU	85.02	86.40	**86.68**

Table 6. Ablation studies on the MSAD dataset.

Effectiveness of proposed modules. To comprehensively evaluate the contribution of each component, we conduct ablation studies on five representative MIL-based VAD baselines across three benchmark datasets: UCF-Crime, ShanghaiTech, and MSAD. We compare three configurations: the original baseline (Raw), the baseline integrated with the denoising module (Denoise-only), and the full D^{2}MIL framework. On the ShanghaiTech dataset (Table 4), integrating the denoising module consistently improves performance across all baselines, particularly on RTFM (+0.82%) and UR-DMU (+1.20%), demonstrating its good performance in complex crowd-scene environments. The complete D^{2}MIL pipeline yields an additional 0.09% improvement, indicating that the debiasing module remains beneficial even under near-saturated performance conditions. On the UCF-Crime dataset (Table 5), a similar trend can be observed. The denoising module enhances performance by 0.16%–1.34% across baselines, confirming its effectiveness in mitigating noisy supervision within the MIL framework. The introduction of the debiasing module further contributes an additional 0.11%–0.47% improvement, highlighting its complementary role. For the MSAD dataset (Table 6), which contains multi-scene aerial videos, our framework exhibits the most significant relative performance gains. Incorporating the denoising module improves the results by an average of 1.1%, while the debiasing stage provides an additional 0.4%–0.7% boost, underscoring its capability to enhance generalization under challenging cross-scene and multi-view conditions.

Drop $\tau_{\max}$	**Recall** $\rho_{\max}$ **= 0.1**		**Recall** $\rho_{\max}$ **= 0.2**	
	AUC	AP	AUC	AP
0.1	84.82	30.92	84.76	30.79
0.2	84.92	31.16	84.58	31.22
0.3	84.73	31.48	85.15	31.31
0.4	85.08	31.55	84.94	31.50
0.5	**85.24**	**31.60**	85.07	31.52
0.6	84.78	31.43	84.79	31.30
0.7	84.89	31.39	84.93	31.44
0.8	84.79	31.47	84.82	31.29
0.9	83.94	31.28	83.83	31.36

Table 7. Experimental results of the Sultani et al. + D^{2}MIL method with varying drop rate $\tau_{\max}$ (0.1–0.9) and recall rate $\rho_{\max}$ (0.1, 0.2) parameters on the UCF-Crime dataset, evaluated by AUC and AP metrics.

Parameter analysis. To investigate the optimal values for the drop rate $\tau_{\max}$ and recall rate $\rho_{\max}$, we tested 18 parameter combinations to determine the best-performing configuration. As shown in Table 7, we varied the drop rate from 0.1 to 0.9 and evaluated two recall rate settings (0.1 and 0.2), with performance quantified using both AUC and the frame-level Average Precision (AP) metrics. The highest AUC (85.24%) and AP (31.60%) are achieved when $\tau_{\max} = 0.5$ and $\rho_{\max} = 0.1$, confirming this parameter combination as optimal.

4.5. Qualitative Analysis

In Figure 3, we present qualitative comparisons between RTFM (green curve) and RTFM + D^{2}MIL (blue curve) on eight representative test videos from the UCF-Crime and MSAD datasets. Each plot depicts the temporal evolution of the predicted anomaly scores, with the ground-truth anomaly intervals shaded in light red. Representative key frames are shown above each plot, where abnormal frames

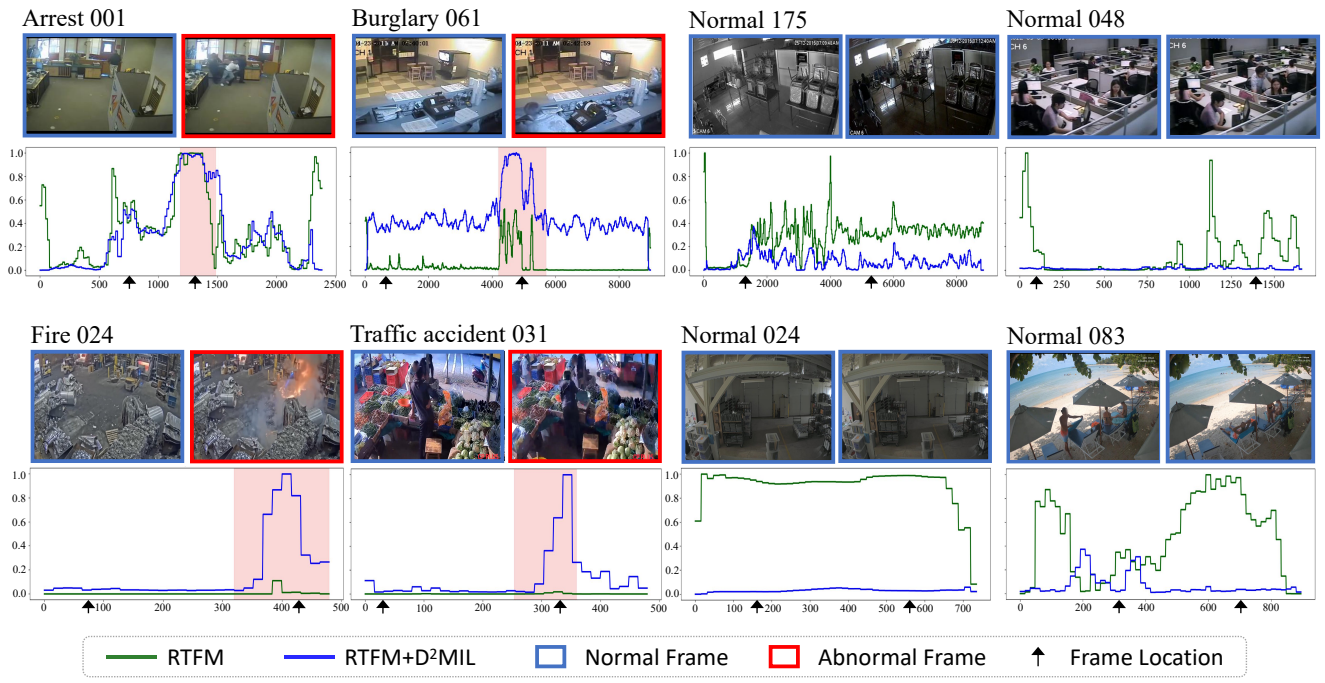


Figure 3. Qualitative comparison results of RTFM and RTFM + D^{2}MIL on UCF-Crime and MSAD datasets, where the horizontal axis represents frames and the vertical axis denotes anomaly scores.

are highlighted in red boxes and normal frames in blue boxes; the corresponding frame locations are marked by arrows along the time axis.

The four examples on the left are abnormal videos. Compared to the baseline, our D^{2}MIL -enhanced model produces sharper and more concentrated anomaly peaks within annotated abnormal intervals, while suppressing spurious responses in normal regions. For example, in Burglary 061 and Fire 024, D^{2}MIL accurately localizes true anomaly segments with high scores, closely matching ground-truth boundaries. This demonstrates improved temporal precision in anomaly localization.

The four examples on the right are normal videos. The original RTFM baseline often produces unstable or false-positive scores in normal scenes, misinterpreting illumination or motion variations as anomalies (e.g., Normal 083). In contrast, RTFM + D^{2}MIL maintains consistently low anomaly scores across the sequence, reducing false alarms and yielding smoother predictions, which reflects improved robustness and generalization. Overall, the qualitative results confirm that our proposed D^{2}MIL framework enhances anomaly localization accuracy on truly abnormal sequences while substantially mitigating false detections in normal scenarios, achieving more stable and trustworthy frame-level anomaly detection.

5. Conclusion

In this paper, we revisit the MIL paradigm and highlight two critical yet underexplored challenges: the contamination of noisy samples and their confusion with genuinely hard abnormal instances. To address these, we introduce D^{2}MIL , a unified framework that combines a dynamic threshold-based denoising module with a VLM-based debiasing module. This design enables the model to suppress unreliable instances while adaptively retaining informative samples, leading to more robust and discriminative learning under weak supervision. D^{2}MIL can be seamlessly integrated into existing MIL-based baselines, delivering consistent performance improvements across three benchmark datasets: ShanghaiTech, UCF-Crime, and MSAD. Despite these promising results, one aspects leave room for improvement. The debiasing performance is somewhat influenced by the reasoning capacity of the underlying VLM.

6. Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 62572260, 62572257, 72342017, 62572311), National Natural Science Foundation of Tianjin, China (No. 25JCQNJC01380), and the Fundamental Research Funds for the Central Universities, Nankai University (No. 63251116)

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2, 5

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 5

[3] 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

[4] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5559, 2023. 2, 5, 6

[5] 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*, pages 387–395, 2023. 2, 6

[6] 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*, pages 12137–12146, 2023. 1

[7] JiaChang Feng, FaTing Hong, and WeiShi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, pages 14009–14018, 2021. 1

[8] Lu Jiang, Zheng Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018. 5

[9] Shuo Li, Fang Liu, and LiCheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI*, pages 1395–1403, 2022. 1

[10] Jing Liu, Yang Liu, Jieyu Lin, Jielin Li, Liang Cao, Peng Sun, Bo Hu, Liang Song, Azzedine Boukerche, and Victor CM Leung. Networking systems for video anomaly detection: A tutorial and survey. *ACM Computing Surveys*, 57 (10):1–37, 2025. 2

[11] Wen Liu, WeiXin Luo, DongZe Lian, and ShengHua Gao. Future frame prediction for anomaly detection - a new baseline. In *CVPR*, pages 6536–6545, 2018. 1

[12] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 5

[13] CeWu Lu, JianPing Shi, and JiaYa Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, pages 2720–2727, 2013. 1

[14] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *TIP*, 2021. 1

[15] 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. 2

[16] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2664–2673, 2023. 2

[17] 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*, 33:4923–4936, 2023. 1, 2

[18] seominseok. Real world anomaly detection in surveillance videos : Pytorch re-implementation. https://github.com/seominseok0429/Real-world-Anomaly-Detection-in-Surveillance-Videos-pytorch, 2021. Accessed: 2025-3-19. 5, 6

[19] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 1, 2, 3, 5, 6

[20] Yu Tian, GuangSong Pang, YuanHong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *ICCV*, pages 4955–4966, 2021. 2, 5, 6

[21] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. In *ECCV*, 2016. 5

[22] Yu Wang and Shiwei Chen. Learning event completeness for weakly supervised video anomaly detection. *ArXiv*, abs/2506.13095, 2025. 1, 2

[23] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3512–3527, 2021. 3

[24] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021. 1

[25] Peng Wu, Jing Liu, YuJia Shi, FangTao Shao, ZhapYang Wu, and ZhiWei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, pages 322–339, 2020. 1

[26] 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. 3

[27] Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. *ArXiv*, abs/2412.01095, 2024. 3

[28] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly

supervised learning with normalcy suppression for anomalous event detection. In *ECCV*, pages 358–376, 2020. 1, 3

[29] 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*, abs/2406.12235, 2024. 3

[30] Yang Zhang, Philip David, and Boqing Gong. Curriculum learning for domain adaptation in semantic segmentation of urban scenes. In *ICCV*, 2017. 5

[31] Yujia Zhang, Xin Huang, Yijia Wang, and et al. Qwen-vl: Scaling vision-language models with decoupled visual encoders. In *arXiv preprint arXiv:2403.17820*, 2024. 5

[32] JiaXing Zhong, NanNan Li, WeiJie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, pages 1237–1246, 2019. 3

[33] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1237–1246, 2019. 5

[34] 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, 5, 6

[35] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems*, 37:89943–89977, 2024. 5

[36] Tao Zhu, Qi Yu, Xinru Dong, Shiyu Li, Yue Liu, Jinlong Jiang, and Lei Shu. Prodisc-vad: An efficient system for weakly-supervised anomaly detection in video surveillance applications, 2025. 2, 6