VADTree: Explainable Training-Free Video Anomaly Detection via Hierarchical Granularity-Aware Tree

Wenlong Li[1] **Yifei Xu**[1,4] **Yuan Rao**[1] **Zhenhua Wang**[2] **Shuiguang Deng**[3]
[1]School of Software, Xi'an Jiaotong University
[2]China Railway Xi'an Group
[3]College of Computer Science and Technology, Zhejiang University
[4] Xi'an Jiaotong University Suzhou Institute
wenlongli@stu.xjtu.edu.cn belonxu_1@xjtu.edu.cn

Abstract

Video anomaly detection (VAD) focuses on identifying anomalies in videos. Supervised methods demand substantial in-domain training data and fail to deliver clear explanations for anomalies. In contrast, training-free methods leverage the knowledge reserves and language interactivity of large pre-trained models to detect anomalies. However, the current fixed-length temporal window sampling approaches struggle to accurately capture anomalies with varying temporal spans. Therefore, we propose **VADTree** that utilizes a Hierarchical Granularity-aware Tree (HGTree) structure for flexible sampling in VAD. VADTree leverages the knowledge embedded in a pre-trained Generic Event Boundary Detection (GEBD) model to characterize potential anomaly event boundaries. Specifically, VADTree decomposes the video into generic event nodes based on boundary confidence, and performs adaptive coarse-fine hierarchical structuring and redundancy removal to construct the HGTree. Then, the multi-dimensional priors are injected into the visual language models (VLMs) to enhance the node-wise anomaly perception, and anomaly reasoning for generic event nodes is achieved via large language models (LLMs). Finally, an inter-cluster node correlation method is used to integrate the multi-granularity anomaly scores. Extensive experiments on three challenging datasets demonstrate that VADTree achieves state-of-the-art performance in training-free settings while drastically reducing the number of sampled video segments. The code will be available at https://github.com/wenlongli10/VADTree.

1 Introduction

Video Anomaly Detection (VAD) aims at temporally locating unexpected and unusual events in videos, thereby facilitating widespread applications including autonomous driving [49, 23] and industrial manufacturing [28]. Most traditional VAD approaches primarily locate anomalous frames by learning the normal or abnormal patterns from training samples with either fully-supervised [1, 39], weakly-supervised [34, 2, 48, 46, 25] or unsupervised learning [51, 62, 22].

While the aforementioned methods perform competitively on experimental VAD benchmarks, their inherent drawbacks limit the capabilities of interpretability, generalization, and interaction in real-world applications. The rapid development of pre-trained Large Language Models (LLMs) and Visual Language Models (VLMs) facilitates the combination of visual comprehending and language interaction, which are particularly well-suited for explainable VAD in real-world surveillance scenarios. Recent research on explainable VAD generates semantic segments of long-term videos with temporal window strategy and equips VLMs with auxiliary guidance to make interpretable anomaly scoring

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

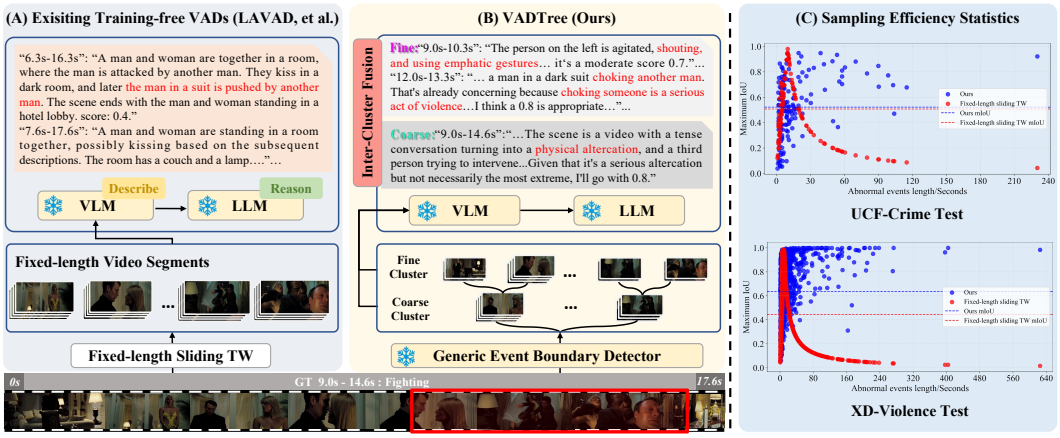


Figure 1: Comparison of our methods with popular paradigms. As illustrated in (A), prevailing training-free VAD methods relying on fixed-length sliding temporal window sampling inherently fail to adapt to dynamic anomaly durations. (B) demonstrates our VADTree is based on pre-trained knowledge of Generic Event Boundary Detection to achieve adaptive coarse-fine hierarchical representation of videos, and support multi-granularity anomaly understanding and score fusion. (C) displays the maximum IoU between all sampled video segments and ground-truth abnormal events across two VAD datasets. The sampling results of 10 seconds long fixed-length sliding temporal window (TW) [50, 55] can only achieve higher IoU with abnormal events that are close in length to itself (mIoU $= 0.51$ on UCF-Crime and mIoU $= 0.44$ on XD-Violence). Our granularity-aware tree demonstrates strong flexibility, and achieves higher IoU for events from 3 seconds to 630 seconds, which is the basis for subsequent understanding and localization of anomalies (mIoU $= 0.52$ on UCF-Crime and mIoU $= 0.64$ on XD-Violence).

[24, 55, 50]. As the pioneering training-free VAD, LAVAD [55] exploits an off-the-shelf VLM to caption each video frame, and enables LLMs to aggregate and score scene semantic dynamics over time in each temporal window. Inspired by Verbalized Machine Learning (VML), VERA [50] leverages video-level annotation data and verbalized learning to optimize a set of guidance questions to drive the frozen VLMs to make abnormal judgments on semantics within the temporal window, and yields frame-level anomaly scores in a coarse-to-fine manner without parameter modifications.

Towards explainable VAD in more practical real-world scenarios, there remains a significant gap to comprehend and reason about anomalies with different durations. A key challenge lies in accurately localizing diverse anomalies under a training-free setting. The video segments sampled by fixed temporal windows are straightforward to implement but remain far from the ground-truth abnormal event boundaries [24, 55, 50]. More critically, this strategy risks abrupt semantic discontinuities or the conflation of irrelevant semantics, which exacerbates the noise in abnormal semantics and amplifies hallucinations of VLMs. Although HolmesVAU [60] introduces an anomaly-focused temporal sampler to handle the anomalies of varying durations, this approach trained on domain-specific videos is prone to underperform in practical videos recorded in changed domains. Another fundamental limitation is the inability to comprehensively understand multi-granular anomalies. While existing explainable VADs excel at detecting transient anomalies, such as traffic accidents or explosions, they often fail to model more complex events like burglaries and arrests. These complex events require extended contextual reasoning. Previous studies [55, 50] have attempted to integrate semantics from sliding temporal windows. However, the fixed-length windows inherently conflicts with the dynamic characteristics of event durations in real-world scenarios, struggling to address frame redundancy and inevitable noise.

To address these challenges, we propose VADTree , a training-free VAD framework that realizes multi-granularity anomaly reasoning via hierarchical event-aware video understanding. Unlike existing temporal window-based approaches, VADTree adaptively organizes video content into a hierarchical granularity-aware tree structure by leveraging pre-trained generic event boundary detectors. This tree structure naturally aligns with the temporal dynamics of real-world events, allowing for adaptive sampling of video segments that match anomaly durations. We address semantic noise in anomaly scoring by introducing intra-cluster node refinement that aggregates contextually relevant nodes to

refine initial predictions. To resolve the conflicts between coarse-grained and fine-grained cues, we develop inter-cluster node correlation to dynamically integrate anomaly evidence across temporal granularities, enhancing detection robustness through score consistent aggregation. We evaluate VADTree on three benchmark datasets: UCF-Crime [34], XD-Violence [43], and MSAD [67]. Our empirical results demonstrate that VADTree outperforms unsupervised, one-class, and training-free VAD methods. This work makes the following contributions:

- We propose **VADTree**, a training-free generic event-centric video anomaly detection framework that flexibly leverages pre-trained GEBD knowledge to localize anomalous events in temporal positions. VADTree overcomes the inefficiency and roughness of dense sampling while providing a multi-granularity perception and reasoning capability for training-free VAD.
- We propose a hierarchical granularity-aware tree that utilizes a coarse-fine representation of anomalous videos based on potential generic event boundaries. Additionally, we design an event-centric anomaly scoring and refining approach to derive generic event anomaly scores from tree nodes, which integrates multidimensional prior information and multi-granularity scores to enhance VAD performance and reasoning ability.
- VADTree achieves SOTA performance among training-free, unsupervised, and one-class methods on both UCF-Crime and XD-Violence datasets, and even surpasses some weakly supervised methods on MSAD dataset.

2 Related Work

2.1 Video Anomaly Detection

Traditional VAD approaches primarily employ deep neural networks (DNNs) through three dominant learning paradigms. Fully-supervised methods [1, 39] utilize frame-level annotations to learn the distinction between normal and abnormal frames, but they entail a prohibitive cost of acquiring large-scale labeled datasets. Weakly-supervised approaches [34, 2, 48, 57, 25] address this limitation by training discriminative models using video-level labels from both normal and abnormal samples, learning to identify anomalous patterns without precise temporal annotations. Unsupervised learning approaches [21, 51, 62, 40, 22, 38] bypass annotation requirements entirely by solving frame reconstruction or prediction tasks to construct distinct representation spaces for normal and anomalous video content. The traditional methods still lack interactivity and rely heavily on the availability of training data.

Recent advances [60, 59, 55, 47, 24, 50, 11] have successfully leveraged VLMs to generate interpretable textual descriptions of detected anomalies. Current approaches primarily follow two paradigms: (1) Methods that rely on frozen models first split videos via sliding temporal windows, then analyze potential anomalies through multiple pre-trained models [55, 50, 11, 9]. (2) The instruction fine-tuning based methods utilize DNN-based VAD models to filter out potential abnormal frames, which are then fed into the VLMs along with prompts for further abnormal description and judgment [60, 59]. However, the sliding temporal window employed in current training-free methods suffers from inflexibility and sampling redundancy, making it challenging to accurately capture anomalous events with varying content lengths. Instruction fine-tuning based methods require additional data and computational resources to identify potential anomalous video segments. In our work, we explore an adaptive temporal sampling approach for potential anomalous events under training-free conditions.

2.2 Event-based Video Understanding

An event is an inherent semantic unit of videos, serving as a critical foundation for scene context understanding. Recent advances in video understanding have extensively explored event-centric representations to achieve compact and effective modeling [41, 13, 6, 15]. Specifically, HEMLLM [6] designs an adaptive sequence segmentation mechanism to partition long videos into coherent event segments. Similarly, LLMEPET [15] employs pseudo-events to guide precise moment prediction within event boundaries. TRACE [13] introduces a causal event modeling framework to deconstruct videos into event sequences, where the current event is predicted based on previous event information and textual instructions. Meanwhile, VideoTree [41] constructs a query-adaptive hierarchical representation grounded in the inherent event and scene structure of videos. As an event-aware

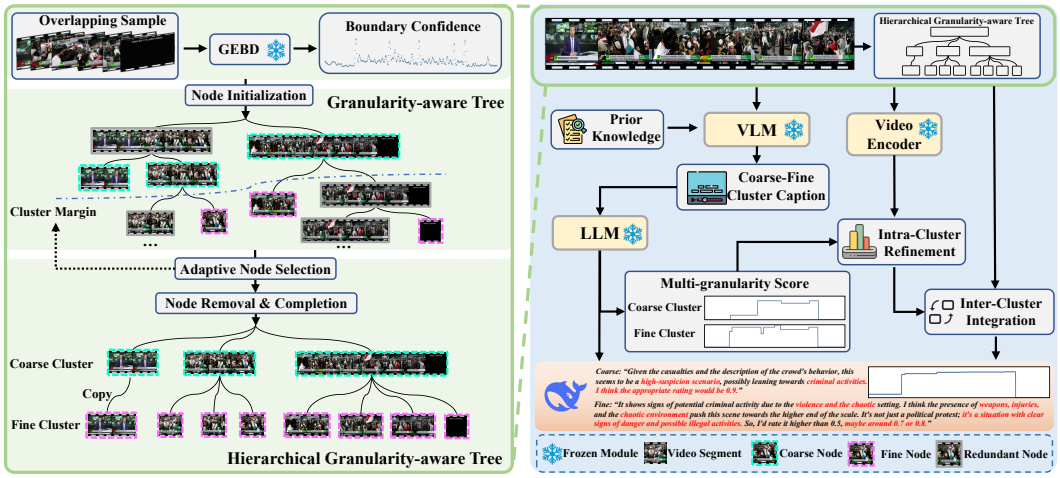


Figure 2: The architecture of our proposed VADTree. The left side shows the construction of a hierarchical granularity-aware tree, which provides flexible multi-granularity characterization for the understanding and location of abnormal events. Then, as shown on the right, the description, reasoning, and refinement are implemented in a node-wise manner, and finally abnormal score integration is completed based on the topological relationship of the HGTree.

training-free VAD framework, EventVAD [30] integrates dynamic spatiotemporal graph modeling and VLMs to detect anomaly events. However, the generalization of the event-aware method it constructs has not been verified, and its robustness in identifying complex boundaries is limited. Differently, our work focuses on training-free VAD that addresses anomalous semantic understanding within multi-granularity generic event-structured video representations.

3 Methodology

Given an input video sequence $V = \{I_t\}_{t=1}^{T}$ with T frames, our training-free approach aims to directly locate and detect the anomalous events within V without any parameter updates or fine-tuning on external datasets. The overall pipeline of our VADTree is illustrated in Figure 2, which is composed of a hierarchical granularity-aware tree, generic event-centric anomaly scoring and refining, and inter-cluster node correlation. Firstly, we utilize the GEBD pre-trained model and depth-first traversal to construct a granularity-aware tree. Further, we use k-means clustering to stratify and simplify the tree, resulting in a hierarchical granularity-aware tree with coarse and fine clusters (Section 3.1). Then, the generic event-centric anomaly scoring module produces initial anomaly score based on a video content description and intrinsic prior knowledge, and ensures contextual relevance and reduces scoring inaccuracy by aggregating score from semantically similar segments in the intra-cluster. For the fusion of inter-cluster anomaly scores, we design a cohesion-driven correlation mechanism to ensure semantic integrity and complementarity across different hierarchical structures (Section 3.2).

3.1 Hierarchical Granularity-aware Tree

To address the inherent limitations of uniform sampling for arbitrary-length anomalies, we construct a hierarchical granularity-aware tree by leveraging pre-trained GEBD knowledge [64, 31], which adaptively accommodates events with diverse temporal scales through dynamic multi-granularity decomposition. It mainly includes three operations: segmentation confidence sequence generation, generic event node initialization, and adaptive node stratification.

Segmentation Confidence Sequence As conventional GEBD models are limited to processing short video clips (duration of l_{raw} frames), we extend their capability to long videos through an overlapping sliding window strategy. Inspired by [68, 33], we first partition the input video V into K overlapping temporal segments $\{V_{local}^{(k)}\}_{k=1}^{K}$. Each segment is independently processed by the

pre-trained GEBD model to generate a local boundary confidence sequence $C_{local}^{(k)}$ of length l_{raw}:

$$C_{local}^{(k)} = [(t_1, c_1), (t_2, c_2), \ldots, (t_{l_{raw}}, c_{l_{raw}})] = f_{\text{GEBD}}\left(V_{local}^{(k)}\right), \tag{1}$$

where $t_i \in \mathbb{Z}^+$ represents the global frame index in V, and $c_i \in [0, 1]$ is the confidence score at position t_i. To mitigate edge effects from windowing, we retain only the central $l_{raw}/2$ frames from each $C_{local}^{(k)}$. These partial sequences are concatenated into a unified global confidence sequence C that preserves both positional information and confidence scores:

$$\begin{aligned} C &= \text{Concat}_{k=1}^{K}\left(C_{local}^{(k)}\left[\lfloor \tfrac{1}{4} l_{raw} \rfloor : \lfloor \tfrac{3}{4} l_{raw} \rfloor\right]\right), \\ \hat{C} &= \{(t, \hat{c}) \mid \text{LocalMax}(C, t)\}, \end{aligned} \tag{2}$$

where $\text{Concat}(\cdot)$ operator aligns confidence scores by their global indices t, and $\lfloor \cdot \rfloor$ ensures the central frames is an integer. This operation essentially discards the fractional part of the division result. $\text{LocalMax}(\cdot)$ identifies peak positions at index t if the confidence score $C(t)$ satisfies:

$$C(t) \geq C(t \pm 1). \tag{3}$$

Generic Event Node Initialization VADTree constructs a granularity-aware binary tree $\mathcal{T} = \{\mathcal{N}_i\}_{i=0}^{M}$ where each node $\mathcal{N}_i = (\hat{c}_l^{(i)}, \hat{c}_r^{(i)}, V_{l:r}^{(i)})$ represents a video segment $V_{l:r}^{(i)}$ with associated confidence scores $\hat{c}_l^{(i)}$ and $\hat{c}_r^{(i)}$ for its left and right temporal boundaries. The tree structure is built via the TreeInit algorithm (Appendix A.1), which recursively splits segments at the most confident event boundaries $\hat{c}_{\text{max}} \in \hat{C}$. Initialization starts with the root node $\mathcal{N}_0 = (1, 1, V_{1:T}^{(0)})$, where 1 means that the confidence of the factual boundary for the beginning and end frames. The algorithm performs depth-first partitioning until either exhausting all candidate boundaries or encountering confidence values below the threshold γ_{min}. The resulting tree $\mathcal{T}$ inherently encodes temporal granularity through its hierarchical organization, where internal nodes represent segmentation decisions and leaf nodes correspond to atomic events.

Adaptive Node Stratification After initializing the granularity-aware tree $\mathcal{T}$, we stratify it hierarchically to enable multi-granular representation of videos. Given the continuous confidence scores and the inherent content uncertainty in videos, $\mathcal{T}$ allows for decomposition into an arbitrary number of hierarchical clusters. Considering the marginal performance gains diminishing with excessive layers and the resultant increase in inference overhead, we adopt the classic two cluster granularity semantic perception strategy to partition $\mathcal{T}$ into coarse cluster parent-wise nodes and fine-cluster child-wise nodes [45]. The former captures those clear event boundaries, while the latter captures localized motion patterns over shorter temporal intervals. To account for varying boundary clarity across scenarios and filming conditions, we dynamically determine these clusters via K-means clustering, thereby evolving $\mathcal{T}$ into a hierarchical granularity-aware tree $\mathcal{T}^{'} = \{\mathcal{S}_{coarse}^{'}, \mathcal{S}_{fine}^{'}\}$.

$$\begin{aligned} &(\hat{C}_{coarse}, \hat{C}_{fine}) = \text{K-Means}(\hat{C}, 2), \\ &\mathcal{S}_{coarse} = \left\{\mathcal{N}_i \,\middle|\, \min(\hat{c}_l^i, \hat{c}_r^i) \geq \min(\hat{C}_{coarse})\right\}, \quad \mathcal{S}_{fine} = \left\{\mathcal{N}_i \,\middle|\, \min(\hat{c}_l^i, \hat{c}_r^i) \leq \max(\hat{C}_{fine})\right\}, \\ &\mathcal{S}_{coarse}^{'} = \text{RemoveDup}(\mathcal{S}_{coarse}), \quad \mathcal{S}_{fine}^{'} = \text{Complete}(\text{RemoveDup}(\mathcal{S}_{fine})), \end{aligned} \tag{4}$$

where $\text{RemoveDup}(\cdot)$ and $\text{Complete}(\cdot)$ denote the redundancy elimination and node completion operators, respectively. These two clusters $\mathcal{S}_{coarse}^{'}$ and $\mathcal{S}_{fine}^{'}$ are constructed per Eq. 4, by applying K-means clustering and the comparison operation to the confidence scores $\hat{c}$.

To achieve maximal granularity, $\text{RemoveDup}(\cdot)$ is applied to retain only the finest-grained nodes along each tree path while pruning redundant ancestor nodes. Besides, as some leaf nodes in $\mathcal{S}_{coarse}^{'}$ cannot be further split and therefore lack corresponding fine-grained segments, $\text{Complete}(\cdot)$ function replicates these critical nodes to ensure comprehensive coverage. Crucially, the nodes in both $\mathcal{S}_{coarse}^{'}$ and $\mathcal{S}_{fine}^{'}$ can guarantee complete video representation. The details of $\text{RemoveDup}(\cdot)$ and $\text{Complete}(\cdot)$ along with the proof of comprehensive coverage are provided in A.2.

3.2 Generic Event-centric Anomaly Scoring and Refining

Prior-infused Node Scoring When humans recognize behavior, well-learned societal scripts inherently trigger cognitive associations [29]. Building on this foundation, we consider these observed patterns as priors for anomaly analysis, which can be systematically categorized along three dimensions: event scene b_{scene}, specific characters/objects b_{obj}, and actions/behaviors b_{act}. Particularly, we explicitly exclude two categories of ill-posed semantic cues for VLMs: (1) *micro-expressions* (e.g., distracted gaze indicating theft intention) that demand prohibitively high image resolution, and (2) *audio-dependent semantic triggers* (e.g., loud sounds suggesting explosions) that are unavailable in visual-only surveillance systems. According to the above findings, the LLM processes both generation instructions P_b and constraint instructions P_c to derive multidimensional priors as Eq. 5:

$$B = (b_{scene}, b_{obj}, b_{act}) = f_{\text{LLM}_{gen}}(P_b \circ P_c). \tag{5}$$

These priors are then injected into VLMs to facilitate human-like reasoning during video content description. For the HGTree $\mathcal{T}^{'} = \{\mathcal{S}^{'}_{coarse}, \mathcal{S}^{'}_{fine}\}$, let V_u^g represent the sampled frames at node u in $\mathcal{S}^{'}_g$. The VLM generates content captions via Eq. 6:

$$d_u^g = f_{\text{VLM}}(V_u^g, B \circ P_d). \tag{6}$$

Following LAVAD [55], we instruct an LLM to quantify anomaly likelihood through discrete scoring $a \in \{0, 0.1, ..., 1\}$, with 0 and 1 encoding normal and anomalous extremes respectively. The score derivation from prompt P_s follows:

$$a_u^g = f_{\text{LLM}}(d_u^g, P_s). \tag{7}$$

Intra-cluster Node Refinement The score derived from Eq. 7 only examines a partial interval in the entire video without considering long-term context, which is prone to local false positive anomalies caused by mutations. To alleviate this issue, we refine the initial anomaly score by taking into account the context of intra-cluster event nodes. Obviously, within the same cluster, nodes with high semantic similarity should logically exhibit converging anomaly scores. To quantify semantic similarity between different nodes, we compute cosine similarity $\text{sim}(\cdot, \cdot)$ based on their feature representations extracted from a pre-trained vision encoder [12]. For the V_u^g of node u in $\mathcal{S}^{'}_g$, let $\kappa_u = [\kappa_u^{(1)}, \ldots, \kappa_u^{(K)}]$ index the top-K most similar nodes. As shown in Eq. 8, the refined anomaly score $\hat{a}_u$ is computed as an ensemble of initial scores of top-K nodes relevant to V_u^g.

$$\hat{a}_u^g = \underbrace{\sum_{i=1}^{K} a_{\kappa_u^{(i)}}}_{\text{Initial scores}} \cdot \underbrace{\frac{\exp(\text{sim}(u, \kappa_u^{(i)})/\tau)}{\sum_j^K \exp(\text{sim}(u, \kappa_u^{(j)})/\tau)}}_{\text{Softmax weights}} \tag{8}$$

Inter-cluster Node Correlation Previous VAD studies [62, 60] have demonstrated the significance of multi-scale learning, given the varying temporal durations of anomalies and the influence of contextual lengths on anomaly determination. This observation aligns with our experimental findings that abnormal event cues exhibit cluster-specific variations. Therefore, we elucidate a cohesion-driven fusion mechanism for multi-granularity decision fusion based on coarse and fine cluster nodes in the hierarchical granularity-aware tree $\mathcal{T}^{'}$. To integrate multi-granularity anomaly cues while suppressing hierarchical inconsistencies, this mechanism dynamically weights the contributions of parent and child nodes through intra-cluster cohesion metrics.

Specifically, parent nodes and child nodes come from the coarse and fine clusters of the event tree, respectively. For each parent node $\mathcal{N}_i$ containing m child nodes $\{\mathcal{N}_{i1}, ..., \mathcal{N}_{im}\}$, we compute the intra-cluster cohesion w_i as the variance of their denoised anomaly scores via Eq. 9:

$$w_i = \frac{1}{m}\sum_{j=1}^{m}\left(\hat{a}_{n_{ij}} - \mu_i\right)^2, \quad \text{where } \mu_i = \frac{1}{m}\sum_{j=1}^{m}\hat{a}_{n_{ij}}. \tag{9}$$

Subsequently, we conduct normalization within the coarse clustering process to obtain the $\hat{w}_i$. A lower $\hat{w}_i$ signifies strong semantic consistency among child nodes, implying that the parent node should dominate the fusion process. Conversely, higher $\hat{w}_i$ implies conflicting child node evidence,

Table 1: Results on UCF-Crime dataset show that VADTree substantially outperforms all Training-free, one-class, and unsupervised methods, even surpassing some weakly-supervised approaches.

Method	Supervision	AUC (%)
Non-Explainable VAD Methods		
Sultani *et al.* [34]	Weakly Supervised	75.41
Sultani *et al.* [34]	Weakly Supervised	77.92
IBL [61]	Weakly Supervised	78.66
GCL [53]	Weakly Supervised	79.84
GCN [65]	Weakly Supervised	82.12
MIST [10]	Weakly Supervised	82.30
Wu *et al.* [43]	Weakly Supervised	82.44
CLAWS [54]	Weakly Supervised	83.03
RTFM [37]	Weakly Supervised	83.31
RTFM [37]	Weakly Supervised	84.03
Wu & Liu *et al.* [44]	Weakly Supervised	84.89
MSL [19]	Weakly Supervised	85.30
MSL [19]	Weakly Supervised	85.62
S3R [42]	Weakly Supervised	85.99
MGFN [4]	Weakly Supervised	86.67
MGFN [4]	Weakly Supervised	86.98
SSRL [17]	Weakly Supervised	87.43
CLIP-TSA [16]	Weakly Supervised	87.58
GS-MoE [7]	Weakly Supervised	91.58
UR-DMU [66]	Weakly Supervised	86.97
UMIL [25]	Weakly Supervised	86.75
π-VAD [26]	Weakly Supervised	90.33
SVM [34]	One Class	50.00
SSV [32]	One Class	58.50
BODS [40]	One Class	68.26
GODS [40]	One Class	70.46
GCL [53]	Unsupervised	74.20
Tur [38]	Unsupervised	66.85
DyAnNet [36]	Unsupervised	79.76
Explainable VAD Methods		
VADor [24]	Fine-tuning	88.13
Holmes-VAD [59]	Fine-tuning	89.51
Holmes-VAU [60]	Fine-tuning	88.96
VERA [50]	Verbalized Learning	86.55
Blip2 [18]	Training-free	46.42
ZS CLIP [27]	Training-free	53.16
ZS ImageBind (Image) [12]	Training-free	53.65
ZS ImageBind (Video) [12]	Training-free	55.78
LLaVA-1.5 [20]	Training-free	72.84
Video-Llama2 [58]	Training-free	74.42
LAVAD [55]	Training-free	80.28
SUVAD [11]	Training-free	83.90
MCANet [9]	Training-free	82.47
EventVAD [30]	Training-free	82.03
VADTree(Ours)	Training-free	**84.74**

Table 2: Results on XD-Violence dataset demonstrate that VADTree achieves significantly superior performance over current Training-free approaches in terms of AUC ROC, while also outperforming all one-class and unsupervised methods. The best results among training-free methods are highlighted in bold. * denotes the method that incorporates an additional audio modality. VADTree* employs Kimi-Audio-7B-Instruct to extract audio captions and enables the LLM to perform anomaly reasoning based on the multimodal text information.

Method	Supervision	AP (%)	AUC (%)
Non-Explainable VAD Methods			
Wu *et al.* [43]	Weakly Supervised	73.20	-
Wu *et al.** [43]	Weakly Supervised	78.64	-
MSL [19]	Weakly Supervised	75.53	-
Wu and Liu [44]	Weakly Supervised	75.90	-
RTFM [37]	Weakly Supervised	77.81	-
RTFM* [37]	Weakly Supervised	78.54	-
MSL [19]	Weakly Supervised	78.28	
MSL [19]	Weakly Supervised	78.58	-
S3R [42]	Weakly Supervised	80.26	-
MGFN [4]	Weakly Supervised	79.19	-
MGFN [4]	Weakly Supervised	80.11	-
CLIP-TSA [16]	Weakly Supervised	82.19	-
GS-MoE [7]	Weakly Supervised	82.89	94.52
MACIL-SD* [52]	Weakly Supervised	83.40	-
UR-DMU* [26]	Weakly Supervised	81.77	-
π-VAD* [26]	Weakly Supervised	85.37	-
Hasan *et al.* [14]	One Class	-	50.32
Lu *et al.* [22]	One Class	-	53.56
BODS [40]	One Class	-	57.32
GODS [40]	One Class	-	61.56
RareAnom [35]	Unsupervised	-	68.33
Explainable VAD Methods			
Holmes-VAD [59]	Fine-tuning	90.67	-
Holmes-VAU [60]	Fine-tuning	87.68	-
VERA [50]	Verbalized Learning	70.54	88.26
Blip2 [18]	Training-free	10.89	29.43
ZS CLIP [27]	Training-free	17.83	38.21
ZS ImageBind (Image) [12]	Training-free	27.25	58.81
ZS ImageBind (Video) [12]	Training-free	25.36	55.06
LLaVA-1.5 [20]	Training-free	50.26	79.62
Video-Llama2 [58]	Training-free	53.57	80.21
LAVAD [55]	Training-free	62.01	85.36
SUVAD [11]	Training-free	**70.10**	-
MCANet* [9]	Training-free	69.72	87.43
EventVAD [30]	Training-free	64.04	87.51
VADTree (Ours)	Training-free	67.82	**90.44**
VADTree* (Ours)	Training-free	68.85	**90.55**

and the parent node may have missed some instantaneous cue, necessitating greater reliance on the child node's fine semantics.

By adjusting the initial fusion weight of the 0.5 through the control coefficient $\beta \in [-1, 1]$, the final frame-wise anomaly score $\bar{a}$ for each segment is determined based on the anomaly scores of final fine cluster nodes $\bar{a}_{n_{ij}}$:

$$\bar{a}_{n_{ij}} = \frac{1}{2}(1 - \beta\hat{w}_i)\hat{a}_{n_i} + \frac{1}{2}(1 + \beta\hat{w}_i)\hat{a}_{n_{ij}}. \tag{10}$$

4 Experiments

We validate the performance of VADTree on three datasets against state-of-the-art VAD methods trained with different types of supervision, as well as other training-free baselines. To verify the necessity of each core module, we conduct systematic ablation studies to demonstrate the rationality and effectiveness of VADTree's proposed components. In the following, we first describe the experimental setup in terms of datasets and performance metrics. We then present and discuss the results in Section 4.1, followed by the ablation studies in Section 4.2, and conclude with qualitative experiments in Section 4.3. For more experimental analysis and qualitative results, please refer to the Appendix C.

Table 3: Results on MSAD dataset demonstrate that training-free approach VADTree delivers competitive performance against existing state-of-the-art weakly-supervised methods.

Method	Supervision	AUC (%)	AUC_a (%)	AP (%)	AP_a (%)
RTFM [37]	Weakly Supervised	86.65	-	-	-
MGFN [4]	Weakly Supervised	84.96	-	-	-
TEVAD [3]	Weakly Supervised	86.82	-	-	-
UR-DMU [26]	Weakly Supervised	85.78	67.95	67.35	75.30
GS-MoE [7]	Weakly Supervised	87.72	69.54	68.26	76.68
π-VAD [26]	Weakly Supervised	88.68	**71.25**	71.26	**77.86**
VADTree (Ours)	Training-free	**89.32**	67.85	**71.41**	75.49

Table 4: Results of VADTree on UCF-Crime dataset with different HGTree construction configuration.

γ_{min}	Cluster Tool	Clusters	AUC (%)
0.3	-	Fine	80.89
0.4	-	Fine	82.81
0.5	-	Fine	80.85
0.3	K-Means	Coarse + Fine	83.74
0.4	K-Means	Coarse + Fine	84.74
0.5	K-Means	Coarse + Fine	82.40
0.4	K-Medoids	Coarse + Fine	**85.24**

Table 5: Effect of different components on UCF-Crime dataset.

Module	AUC (%)
HGTree Fine Cluster	71.57
+ Prior-infused Node Scoring	75.67
+ Intra-cluster Node Refinement	83.05
+ Inter-cluster Node Correlation	**84.74**

Table 6: Comparison of performance of VADTree under different VLM and LLM on UCF-Crime dataset.

VLM	LLM	AUC (%)
LLaVA-NeXT-Video-7B	DeepSeek-R1-Distill-Qwen-14B	**84.74**
InternVL2_5-8B	DeepSeek-R1-Distill-Qwen-14B	83.74
LLaVA-NeXT-Video-7B	t5gemma-9B-2B	84.00
InternVL2_5-8B	t5gemma-9B-2B	83.56

Datasets We evaluate our method using three commonly used VAD datasets featuring real-world surveillance scenarios, i.e., UCF-Crime [34], XD-Violence [43], and MSAD [67]. **UCF-Crime** is a large-scale dataset comprising 1900 long untrimmed real-world surveillance videos with 13 types of anomalies. The training set consists of 800 normal and 810 anomalous videos, while the test set includes 150 normal and 140 anomalous videos. **XD-Violence** is another large-scale dataset for violence detection, comprising 4754 untrimmed videos with audio signals and weak labels that are collected from both movies and YouTube. XD-Violence captures 6 categories of anomalies and it is divided into a training set of 3954 videos and a test set of 800 videos. We also evaluate VADTree on **MSAD** dataset, which provides a greater diversity of real-world scenarios than existing benchmarks.

Performance Metrics We measure the VAD performance using the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC) as it is agnostic to thresholding for the detection task. For XD-Violence dataset, we also report the average precision (AP), which refers to the area under the frame-level precision-recall curve, following the established evaluation protocol in [43].

Implementation Details We use EfficientGEBD [64] as the model f_{GEBD} for generic event boundary knowledge acquisition, and the overlapping sampling window length l_{raw} follows the 10s window of Kinetics-GEBD [31]. The video description model f_{VLM} and the anomaly reasoning model f_{LLM} use LLaVA-Video-7B-Qwen2 [63] and DeepSeek-R1-Distill-Qwen-14B [8] respectively. In all experiments, the VLM input is configured to a maximum of 64 frames, with the LLM having the thinking mode turned on by default. Although the "Think" mode of DeepSeek-R1-Distill-Qwen-14B incurs additional inference overhead, we still intentionally retain it because it generates valuable intermediate reasoning steps that significantly enhance anomaly interpretation. The video encoder f_{VE} is provided by ImageBind [12].

4.1 Comparison with State of the Art

Most videos in **UCF-Crime** dataset have low resolution, with mild semantic changes between events within the same video and few shot transitions. Table 1 demonstrates that VADTree achieves a substantial superiority over all training-free methods. In particular, its performance exceeds LAVAD [55] by 4.5% and surpasses EventsVAD [30] by 2.7%.

The videos in **XD-Violence** dataset are primarily sourced from films and TV shows. Consequently, the content is more deliberately composed and contains significantly more frequent shot transitions. Table 2 reveals that VADTree achieves an AUC ROC 5.1% higher than LAVAD [55]. And it shows a 2.9% gain over EventsVAD [30], thereby establishing a new state-of-the-art. Furthermore, VADTree exhibits superior anomaly detection performance compared to all single-class and unsupervised methods on UCF-Crime and XD-Violence datasets.

Surprisingly, VADTree achieves state-of-the-art performance on **MSAD** dataset (as shown in Table 3) and attaining the highest scores in overall metrics. This demonstrates that our training-free approach outperforms even weakly-supervised methods that rely on extensive training data. While π-VAD shows a slight advantage on anomaly-specific metrics AUC_a and AP_a, likely due to its supervised fine-tuning on anomalous segments, VADTree's competitive performance without any dataset-specific training highlights its superior generalization capability.

4.2 Ablation Study

In this section, we present the ablation study conducted on UCF-Crime dataset. We first ablate the effectiveness of each proposed component of VADTree, and then elaborate on the effect of the HGTree on the final detection accuracy under different construction parameters. Finally, we discuss the impact of different pre-trained model combinations on the performance of VADTree .

Effect of Each Proposed Component We ablate different modules of our proposed method VADTree to prove the effectiveness of the four proposed components, including HGTree fine cluster, prior-infused node scoring, intra-cluster node refinement and inter-cluster node correlation. As shown in Table 5, we first use HGTree fine cluster to express the entire video and build a baseline. When we input the prior knowledge for the prompt of VLM, the AUC ROC of the method is further improved, indicating that the understanding and accurate description of anomalies can benefit from the prior information of anomalies. If we further refine the initial anomaly scores of each node within intra-cluster, the AUC ROC will be significantly improved; this is because the module can effectively suppress the inference noise and hallucination of VLM and LLM, and introduce references for independent reasoning of each segment. Finally, inter-cluster node correlation further increased the AUC ROC to 84.7%, indicating that the HGTree based parent-child node structure information can effectively guide multi-granularity scores fusion.

Effect of Different HGTree Construction Configurations The empirical analysis systematically examines how HGTree's configuration parameters govern VAD performance. As illustrated in Table 4, varying the γ_{min} directly modulates the granularity of segmented videos; lower thresholds($\gamma_{min} = 0.3$) induce noisy event boundaries due to over-segmentation, whereas higher thresholds($\gamma_{min} = 0.5$) restrict hierarchical results diversity. Our framework achieves optimal balance at $\gamma_{min} = 0.4$, which consistently delivers peak performance across benchmarks. Crucially, ablation studies reveal that using only leaf nodes reduces HGTree to a single cluster structure, yielding a 1.9% lower AUC ROC compared to the hierarchical two clusters architecture. This indicates the importance of hierarchical granularity-aware representation and decision correlation. Furthermore, replacing K-Means with K-Medoids for clustering produces an additional performance gain, demonstrating the advantage of using more robust centroid selection when dealing with potential outliers in generic event boundary.

Effect of Different VLM and LLM Configurations To comprehensively evaluate the generalizability, we conduct tests with alternative model architectures. Specifically, we select InternVL2_5-8B [5] as an additional VLM with distinct input specifications of 32 frames at 448×448 resolution, contrasting with LLaVA-NeXT-Video-7B's 64 frames, 336×336 inputs. This means that InternVL2_5-8B focuses more on perceiving spatial details rather than the temporal continuity of actions. Furthermore, we incorporate t5gemma-9B-2B [56] as additional LLM variant featuring a unique 9B-encoder and 2B-decoder configuration. This differs from DeepSeek-R1-Distill-Qwen-14B's autoregressive architecture. As evidenced by the results in Table 6, the performance fluctuation of VADTree is acceptable under different model combinations, which demonstrates the strong generalization ability of our framework.

4.3 Qualitative Analysis

Figure 3 presents qualitative results of VADTree using sample videos from UCF-Crime and XD-Violence. Benefiting from its flexible granularity-aware tree video representation, VADTree accurately segments the boundaries between anomalous and normal events. Additionally, due to the scoring stability brought by inter-cluster node correlation, the anomaly scores generated by our method are overall smoother. In addition, we observe that the same video can obtain divergent anomaly scores across its coarse and fine cluster representations. This divergence stems primarily from the lack of a

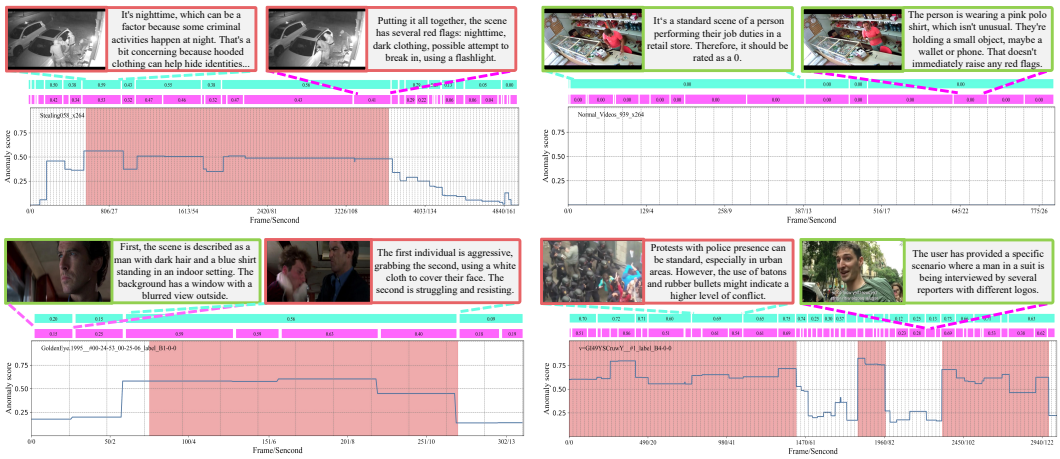


Figure 3: Qualitative results from VADTree on four test videos: two from UCF-Crime (top row) and two from XD-Violence (bottom row). The hierarchical video segment representations and corresponding anomaly scores are visualized alongside their key language explanations, with cyan and rose rectangles denoting coarse cluster and fine cluster nodes respectively. Each video's final anomaly scores (blue solid line) are computed by inter-cluster node correlation. Ground-truth anomalies are highlighted by red regions.

uniform standard for the independent inference of nodes by VLMs and LLMs, which inevitably leads to scoring fluctuations. Our inter-cluster correlation can reduce the impact of such instability.

5 Conclusion

This paper presents a novel training-free framework VADTree for adaptive multi-granularity VAD. By constructing a hierarchical granularity-aware tree to achieve node-wise anomaly understanding and score refinement based on tree structure information, our method overcomes the limitation of poor flexibility in anomaly detection based on fixed-length sliding temporal window sampling in existing methods. The elimination of domain-specific training requirements and explicit explainability through pre-trained model reasoning make our framework particularly suitable for real-world surveillance applications.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China 62572387 and U22B2036, and Jiangsu Agricultural Science and Technology Innovation Fund (CX(24)3132), and Natural Science Basic Research Program of Shaanxi (Program No.2024JC-YBMS-498), and Shaanxi Provincial Key Research and Development Program - Key Project - Qinchuangyuan Original Innovation Window "Four Chains" Integration Project (2024PT-ZCK-93).

References

[1] Shuai Bai, Zhiqun He, Yu Lei, Wei Wu, Chengkai Zhu, Ming Sun, and Junjie Yan. Traffic anomaly detection via perspective map based on spatial-temporal information matrix. In *CVPRW*, 2019. 1, 3

[2] J. Chen, L. Li, L. Su, Z. Zha, and Q. Huang. Prompt-enhanced mulieee transactions on image processingle instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 3

[3] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 8

[4] 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*, 2023. 7, 8

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 9

[6] Dingxin Cheng, Mingda Li, Jingyu Liu, Yongxin Guo, Bin Jiang, Qingbin Liu, Xi Chen, and Bo Zhao. Enhancing long video understanding via hierarchical event-based memory. *arXiv preprint arXiv:2409.06299*, 2024. 3

[7] Giacomo D'Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, François Bremond, and Egor Bondarev. Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection. *arXiv preprint arXiv:2508.06318*, 2025. 7, 8

[8] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 8, 24

[9] Prabhu Prasad Dev, Raju Hazari, and Pranesh Das. Mcanet: Multimodal caption aware training-free video anomaly detection via large language model. In *International Conference on Pattern Recognition*, pages 362–379. Springer, 2025. 3, 7

[10] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Mulieee transactions on image processingle instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 7

[11] Shibo Gao, Peipei Yang, and Linlin Huang. Suvad: Semantic understanding based video anomaly detection using mllm. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3, 7

[12] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 6, 7, 8

[13] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024. 3

[14] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 7

[15] Yiyang Jiang, Wengyu Zhang, Xulu Zhang, Xiao-Yong Wei, Chang Wen Chen, and Qing Li. Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7249–7258, 2024. 3

[16] H. Joo, K. Vo, K. Yamazaki, and N. Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *Proceedings of the IEEE International Conference on Image Processing*, 2023. 7

[17] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *Proceedings of the European Conference on Computer Vision*, 2022. 7

[18] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 7

[19] S. Li, F. Liu, and L. Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 7

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv*, 2023. 7

[21] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3

[22] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 1, 3, 7

[23] Hao Lu, Jiaqi Tang, Xinli Xu, Xu Cao, Yunpeng Zhang, Guoqing Wang, Dalong Du, Hao Chen, and Yingcong Chen. Scaling multi-camera 3d object detection through weak-to-strong eliciting. *arXiv*, 2024. 1

[24] H. Lv and Q. Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024. 2, 3, 7

[25] 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, 3, 7

[26] 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. 7, 8

[27] 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*, 2021. 7

[28] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[29] Roger C Schank and Robert P Abelson. Scripts, plans, goals, and understanding hillsdale, 1977. 6

[30] 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. *arXiv preprint arXiv:2504.13092*, 2025. 4, 7, 8

[31] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8075–8084, 2021. 4, 8, 28

[32] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *ICPR*, 2018. 7

[33] Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221, 2024. 4

[34] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3, 7, 8

[35] Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim. Rareanom: A benchmark video dataset for rare type anomalies. *Pattern Recognition*, 140:109567, 2023. 7

[36] Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In *WACV*, 2023. 7

[37] 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 International Conference on Computer Vision*, 2021. 7, 8

[38] A. Tur, N. Dall'Asen, C. Beyan, and E. Ricci. Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. In *International Conference on Image Analysis and Processing*, 2023. 3, 7

[39] Gaoang Wang, Xinyu Yuan, Aotian Zheng, Hung-Min Hsu, and Jenq-Neng Hwang. Anomaly candidate identification and starting time estimation of vehicles from traffic videos. In *CVPRW*, 2019. 1, 3

[40] J. Wang and A. Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3, 7

[41] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*, 2024. 3

[42] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *Proceedings of the European Conference on Computer Vision*, 2022. 7

[43] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision*, 2020. 3, 7, 8

[44] 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

[45] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv:2407.15841*, 2024. 5

[46] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *ICCV*, 2023. 1

[47] Y. Yang, K. Lee, B. Dariush, Y. Cao, and S. Lo. Follow the rules: reasoning for video anomaly detection with large language models. *arXiv preprint arXiv:2407.10299*, 2024. 3

[48] Z. Yang, J. Liu, and P. 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*, 2024. 1, 3

[49] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. Dota: unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):444–459, 2022. 1

[50] M. Ye, W. Liu, and P. He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. *arXiv preprint arXiv:2412.01095*, 2024. 2, 3, 7, 29, 30

[51] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *ACM International Conference on Multimedia*, 2019. 1, 3

[52] 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

[53] M. Zaheer, A. Mahmood, M. Khan, M. Segu, F. Yu, and S. Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7

[54] 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 *Proceedings of the European Conference on Computer Vision*, 2020. 7

[55] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 6, 7, 8, 29, 30

[56] Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation. *arXiv preprint arXiv:2504.06225*, 2025. 9

[57] C. Zhang, G. Li, Y. Qi, S. Wang, L. Qing, Q. Huang, and M. 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*, 2023. 3

[58] H. Zhang, X. Li, and L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Empirical Methods in Natural Language Processing*, 2023. 7

[59] H. Zhang, X. Xu, X. Wang, J. Zuo, C. Han, X. Huang, C. Gao, Y. Wang, and N. Sang. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*, 2024. 3, 7

[60] H. Zhang, X. Xu, X. Wang, J. Zuo, X. Huang, C. Gao, S. Zhang, L. Yu, and N. Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. *arXiv preprint arXiv:2412.06171*, 2024. 2, 3, 6, 7

[61] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *ICIP*, 2019. 7

[62] M. Zhang, J. Wang, Q. Qi, H. Sun, Z. Zhuang, P. Ren, R. Ma, and J. Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 3, 6

[63] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. 8

[64] Ziwei Zheng, Zechuan Zhang, Yulin Wang, Shiji Song, Gao Huang, and Le Yang. Rethinking the architecture design for efficient generic event boundary detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1215–1224, 2024. 4, 8

[65] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 7

[66] 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, pages 3769–3777, 2023. 7

[67] L. Zhu, L. Wang, A. Raj, T. Gedeon, and C. Chen. Advancing video anomaly detection: A concise review and a new dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 3, 8

[68] Wentao Zhu, Yufang Huang, Xiufeng Xie, Wenxian Liu, Jincan Deng, Debing Zhang, Zhangyang Wang, and Ji Liu. Autoshot: A short video dataset and state-of-the-art shot boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2238–2247, 2023. 4

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the contributions of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitation Section is provided in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The proof is provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We discuss the details required to reproduce the experiments in this paper in the Experiments and Appendix sections.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code in the appendix to reproduce our results, and link our code release in the anonymous github in section abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We use standard datasets and splits, we provide hyperparameters in experimental details along with ablations in experiment sections and appendix to understand the contribution of each component in our algorithm.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Considering the limitation of computing resources, we repeated the main experiments and reported the mean and standard deviation in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided details about compute resources used in supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have read the ethics guidelines and confirm that we do not use human subjects, use existing datasets, explicitly discuss social impacts.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential social impact of this paper is discussed in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No new datasets or generative models are released, and public datasets are used, so no protection measures are required.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: UCF-Crime and XD-Violence are cited appropriately.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes] ,

    Justification: We provide code and instructions.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This article does not involve subjects and related approval requirements.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our core methodology integrates large language models (LLMs) as a non-standard component for semantic anomaly reasoning in video analysis.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

VADTree: Explainable Training-Free Video Anomaly Detection via Hierarchical Granularity-Aware Tree

Technical Appendices

Table of Contents

The appendix begins by detailing the algorithmic process underlying the construction of the HGTree and establishing the proof of its representational completeness. Furthermore, it presents additional experimental details, including extensive ablation studies and comparative analyses. Finally, the appendix examines the limitations of this work and its potential societal impact.

A Hierarchical Granularity-aware Tree

A.1 TreeInit: Granularity-Aware Binary Tree Construction

Algorithm 1 TreeInit: Granularity-Aware Binary Tree Construction Algorithm (Section 3.1)

Require:
Video $V_{1:T}$ with T frames,
Confidence scores $\hat{C} = \{(\tau_i, \hat{c}_i)\}_{i=1}^{N}$,
Confidence threshold γ_{min}
Ensure: Binary tree $\mathcal{T} = \{([s_j, e_j], [\hat{c}_s^j, \hat{c}_e^j])\}_{j=1}^{M}$

$\mathcal{T} \leftarrow \emptyset, \mathcal{U} \leftarrow \emptyset$ ▷ Result set & consumed split points
Push root node $\mathcal{D} \leftarrow [[1, T]]$ ▷ DFS stack initialization
while $\mathcal{D} \neq \emptyset$ **do**
  $[l, r] \leftarrow \mathcal{D}.\texttt{pop}()$
  $\hat{c}_l \leftarrow \mathbb{I}(l = 1) \cdot 1 + \mathbb{I}(l > 1) \cdot \hat{c}_l$ ▷ Left boundary confidence
  $\hat{c}_r \leftarrow \mathbb{I}(r = T) \cdot 1 + \mathbb{I}(r < T) \cdot \hat{c}_r$ ▷ Right boundary confidence
  $\mathcal{T}.\texttt{add}([l, r], [\hat{c}_l, \hat{c}_r])$
  Find split $\tau^* \leftarrow \underset{\tau \in (\Psi \setminus \mathcal{U}) \cap (l, r)}{\arg\max} \hat{c}_\tau$ ▷ Select the highest remaining confidence point
  if $\hat{c}_{\tau^*} \geq \gamma_{\text{min}}$ **then**
    $\mathcal{U}.\texttt{add}(\tau^*)$
    $\mathcal{D}.\texttt{push}([\tau^*, r])$ ▷ Right child
    $\mathcal{D}.\texttt{push}([l, \tau^*])$ ▷ Left child
  end if
end while
return $\text{Sort}(\mathcal{T}, l_j \uparrow)$ ▷ Sort by start time

A.2 Proof of Coverage Completeness in Hierarchical Coarse-Fine Clustering

Theorem 1 (Coverage Completeness)**.** *Based on the method described in Section 3.1, we get* $\mathcal{T}' = (\mathcal{S}'_{coarse}, \mathcal{S}'_{fine})$*, where* $|\mathcal{S}'_{coarse}| = M'_c$ *and* $|\mathcal{S}'_{fine}| = M'_f$ *. Then:*

The original video sequence $V_{1:T}$ ***can be exactly reconstructed through temporal concatenation of segments from either the coarse cluster*** $\mathcal{S}'_{coarse}$ ***or the fine cluster*** $\mathcal{S}'_{fine}$***:***

$$\bigcup_{\mathcal{N}_i \in \mathcal{S}'_{coarse}} [l_i, r_i] = [1, T], \quad \bigcup_{\mathcal{N}_i \in \mathcal{S}'_{fine}} [l_i, r_i] = [1, T]. \tag{11}$$

Notations:

- $\mathcal{N}_i = ([l_i, r_i], [\hat{c}_l^{(i)}, \hat{c}_r^{(i)}])$: A tree node represents a generalized event video segment, with boundary frames and their confidences as $[l_i, ri]$ and $[\hat{c}_l^{(i)}, \hat{c}_r^{(i)}]$ respectively.
- $\prec$: Parent-child relation in $\mathcal{T}$ ($\mathcal{N}_j \prec \mathcal{N}_i \iff \mathcal{N}_i$ is a child of $\mathcal{N}_j$)
- $\mathcal{T}_{leaf} \triangleq \{\mathcal{N}_i \in \mathcal{T} \mid \nexists\, \mathcal{N}_j \prec \mathcal{N}_i\}$: Leaf node set of $\mathcal{T}$

Proof. **Part 1: Initial Coverage Guarantee** The root node $\mathcal{N}_0 = ([1, T], [1, 1]) \in \mathcal{T}$ spans the full video by definition. Through iterative splitting in Algorithm 1, each parent node $\mathcal{N}_p = ([l_p, r_p], [\hat{c}_l^{(p)}, \hat{c}_r^{(p)}])$ is partitioned into non-overlapping child nodes:

$$\mathcal{N}_c^L = ([l_p, \tau^*], [\hat{c}_l^{(c)}, \hat{c}_{\tau^*}^{(c)}]), \quad \mathcal{N}_c^R = ([\tau^*, r_p], [\hat{c}_{\tau^*}^{(c)}, \hat{c}_r^{(c)}]) \tag{12}$$

where $\tau^* \in (l_p, r_p)$. This implies:

$$\bigcup_{\mathcal{N}_i \in \mathcal{T}_{leaf}} [l_i, r_i] = [1, T] \tag{13}$$

Part 2: Coarse Cluster Guarantee The RemoveDup operator filters nodes through:

$$\mathcal{S}'_{coarse} = \{\mathcal{N}_i \in \mathcal{S}_{coarse} \mid \nexists\, \mathcal{N}_i \prec \mathcal{N}_j\} \tag{14}$$

As the above operation exclusively targets non-leaf nodes in $\mathcal{S}_{coarse}$ and leaves leaf nodes unchanged. Therefore, the current leaf nodes satisfies the expression completeness of the original video shown in Eq. 13, and then satisfies the first item of Eq. 11: $\bigcup_{\mathcal{N}_i \in \mathcal{S}'_{coarse}} [l_i, r_i] = [1, T]$.

Part 3: Fine Cluster Guarantee

The Complete operator ensures coverage via two mechanisms:

1. Boundary alignment: For edge cases:

$$\begin{aligned} &\text{if } \min_{\mathcal{N}_i \in \mathcal{S}'_{fine}} l_i > 1 : \text{insert } \mathcal{N}_1 \text{ from } \mathcal{S}'_{coarse} \\ &\text{if } \max_{\mathcal{N}_i \in \mathcal{S}'_{fine}} r_i < T : \text{append } \mathcal{N}_{M'_c} \text{ from } \mathcal{S}'_{coarse} \end{aligned} \tag{15}$$

2. Bridge the gap between nodes: For any adjacent nodes $\mathcal{N}_i = ([l_i, r_i], [\hat{c}_l^{(i)}, \hat{c}_r^{(i)}])$ and $\mathcal{N}_{i+1} = ([l_{i+1}, r_{i+1}], [\hat{c}_l^{(i+1)}, \hat{c}_r^{(i+1)}])$ in $\mathcal{S}'_{fine}$ with $r_i < l_{i+1}$:

$$\exists\, \{\mathcal{N}_c\} \subset \mathcal{S}'_{coarse} \text{ s.t. } \bigcup_c [l_c, r_c] = [r_i, l_{i+1}] \tag{16}$$

Through Eq. 15 and Eq. 16, the second term of Eq. 11 is satisfied: $\bigcup_{\mathcal{N}_i \in \mathcal{S}'_{fine}} [l_i, r_i] = [1, T]$.

Conclusion: Both coarse and fine cluster maintain complete temporal coverage through Section 3.1 process. □

B Generic Event-centric Anomaly Scoring and Refining

B.1 Prior-infused Node Scoring

This section mainly supplements the prompt details used by VLM and LLM. First, by employing $P_b \circ P_c$ (as demonstrated in Section B.1.1), we input the prompt into the LLM [8] [1] to derive prior knowledge that excludes ill-posed semantic cues. The prior knowledge B is shown in Table 7 and Table 8 respectively.

The model configuration details of the VLM for describing video content and the LLM for scoring anomalies are consistent with their open-source repositories [2] [3].

B.1.1 Multidimensional Prior Knowledge Generation Prompt

UCF-Crime *"To help video anomaly detection agent review the occurrence of abnormal events, it is now necessary to pre-analyze possible anomalies to establish a prior knowledge base that matches abnormal events. The video taken has no sound, and may have a long distance or a blurry picture. There may be Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, RoadAccidents, Robbery, Shooting, Shoplifting, Stealing and Vandalism 13 types of events. Please carefully analyze these scenes. Then point out the characteristics of each abnormal event from the following three perspectives: the scene environment, characters or specific objects, actions or behaviors that occurred."*

[1] https://chat.deepseek.com/

[2] https://huggingface.co/lmms-lab/LLaVA-NeXT-Video-7B

[3] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B

XD-Violence *"To help video anomaly detection agent review the occurrence of abnormal events, it is now necessary to pre-analyze possible anomalies to establish a prior knowledge base that matches abnormal events. The video taken has no sound, and may have a long distance or a blurry picture. There may be Abuse, Explosion, Fighting, Car Accident, Shooting and Riot 6 types of events. Please carefully analyze these scenes. Then point out the characteristics of each abnormal event from the following three perspectives: the scene environment, characters or specific objects, actions or behaviors that occurred."*

B.1.2 Multidimensional Prior Knowledge

The prior knowledge B generated for the UCF-Crime and XD-Violence datasets are shown in Table 7 and Table 8 respectively.

UCF-Crime

Table 7: Multidimensional Prior Knowledge of UCF-Crime Dataset.

Abnormal Event Type	Scene Environment Features	Character/Object Features	Action/Behavior Features
Abuse	Secluded spaces (indoors/corners), non-public areas (private locations)	Two parties in physical conflict (perpetrator/victim), dragging tools (ropes/clubs)	Shoving/dragging, repeated hitting, restraining movement (pinning down)
Arrest	Public areas (streets/squares), zones with police vehicles or officers	Uniformed police, handcuffs, batons or firearms	Forced restraint, frisking, escorting to vehicles, lying on the ground
Arson	Areas with flammable materials (warehouses/vehicles), abnormal smoke/flames	Individuals holding flammable containers (gasoline bottles), ignition tools (lighters)	Throwing incendiary objects, fleeing quickly, repeatedly checking the fire
Assault	Narrow passages, crowded areas with sudden dispersion (subway stations/bar entrances)	Armed individuals (knives/blunt weapons), victims struggling on the ground	Sudden lunging, weapon swinging, victims adopting defensive postures
Burglary	Damaged doors/windows, unlit buildings at night, surveillance blind spots (back alleys)	Masked/dark-clothed individuals, lock-picking tools (pliers), backpacks (for loot)	Peering through windows, picking locks, rummaging through items
Explosion	Smoke spreading, flying debris, crowds fleeing outward from a central point	Suspicious packages/vehicles, post-explosion wreckage (metal fragments)	Throwing motions, sudden flash of flames, crowds crouching/running
Fighting	Public spaces (restaurants/stadiums) with concentrated physical conflicts, overturned furniture	Multiple people entangled, bleeding faces, torn clothing	Punching/kicking, hair-pulling, siege

Road Accidents	Collision points (intersections/curves), skid marks, scattered debris, traffic congestion	Deformed vehicles, deployed airbags, paramedics (uniforms/stretchers)	Sudden braking, vehicle rollovers, pedestrians being hit
Robbery	Streets/ATM areas, fast-moving vehicles (motorcycles/cars)	Threats with guns/knives, motorcycle helmets (face concealment), stolen items (bags)	Snatching and fleeing, threatening gestures, vehicles abruptly stopping/accelerating
Shooting	Crowds suddenly ducking/fleeing, vehicles braking abruptly, bullet holes in windows	Gun-wielding individuals, gunshot victims falling, spent shell casings	Aiming firearms, continuous firing, seeking cover
Shoplifting	Loitering near shelves, surveillance blind spots (corners), suspicious concealment (coats)	Frequently observing staff, hiding items (in bags/under clothing)	Concealing items in clothing, glancing around nervously, quickly leaving shelves
Stealing	Crowded areas (subways/markets), sudden disappearance of target items (wallets/-phones)	Close proximity to victims, distractions (e.g., bumping), rapid transfer of stolen goods	Pickpocketing (hands reaching into pockets), passing loot to accomplices
Vandalism	Graffiti-covered walls, shattered glass, toppled public facilities (trash cans/fences)	Spray paint cans, hammers/stones, targets (cameras/glass)	Smashing motions, spraying walls, kicking facilities

Table Notes:
1. Scene environment features capture spatial anomalies (e.g., secluded corners) and physical damage patterns
2. Character/object features focus on suspicious entities and high-risk items
3. Action/behavior features characterize motion dynamics critical for low-quality video analysis

Recognition Tips:
1. Blurry footage: Track group behavior changes (crowd fleeing patterns)
2. Long-distance: Monitor environmental dynamics (smoke/glass shattering)
3. Silent videos: Analyze action intensity (repeated hitting motions)

XD-Violence

Table 8: Multidimensional Prior Knowledge of XD-Violence Dataset.

Abnormal Event Type	**Scene Environment Features**	**Character/Object Features**	**Action/Behavior Features**
Abuse	Secluded or private settings (alleyways, dimly lit rooms), lack of bystanders	Dominant/submissive individuals with indistinct blunt objects (belts, sticks)	Sudden aggressive movements (hitting/-grabbing), victim recoiling/fleeing, prolonged physical contact

Explosion	Sudden bright flash with smoke/fire, structural damage (collapsed walls)	Chaotically moving people, objects near blast source (vehicles, trash bins)	Rapid light/smoke expansion, crowd scattering, lingering smoke/flames
Fighting	Public areas with bystanders (streets, bars), overturned furniture	Aggressive postures, broken objects nearby (bottles, chairs)	Repetitive punching/kicking, high-intensity movements, bystander reactions
Car Accident	Roads with sudden traffic stops, skid marks/debris (broken glass)	Collided vehicles (distorted shapes), injured individuals	Rapid deceleration/impact, post-crash behaviors (crowd gathering, emergency services)
Shooting	Public spaces with visible escape routes (malls, streets)	Shooter's stance (arm extended), collapsing victims	Crowd panic (ducking/sprinting), post-shooting patterns (bodies, law enforcement)
Riot	Large crowds with fires/smoke, broken infrastructure (smashed windows)	Protesters with makeshift weapons, riot police/armored vehicles	Group violence (throwing objects), chaotic dispersal patterns (wave-like movements)

Low-quality Analysis Guidelines:

- Focus on movement patterns (sudden scattering/aggressive postures)
- Detect environmental changes (smoke/debris) over fine details
- Analyze group dynamics (crowd panic/clustering)
- Prioritize spatiotemporal anomalies (light flashes/vehicle stops)

B.1.3 Video Content Description Prompt

If directly driving the VLM to describe the video, its prompt P_d is as follows:

"Please describe the content in the video in detail."

If prior knowledge B is incorporated, the form of $B \circ P_d$ is adopted.

B.1.4 Abnormal Scoring Prompt

The prompt P_s used in this section is divided into two parts: Scoring Criteria and Output Format Control.

Scoring Criteria *"How would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious or potentially criminal activities?"*

Output Format Control *"Please provide the response in the form of a Python list and respond with only one number in the provided list below [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] without any textual explanation. It should begin with '[' and end with ']'."*

C More Results

C.1 More Experimental Details

Based on the experimental details described in Section 4, the $\gamma_{min} = 0.4$ and K-Means clustering algorithm are used to generate the HGTree for inference. In the inter-cluster node refinement process,

we implemented a top-K control for the final weighted neighborhood node numbers. Additionally, this process also includes the temperature parameter τ of softmax. In the Inter-cluster Node Correlation, the hyperparameter β affects the weight of coarse and fine clusters in the final anomaly score.

C.2 Effect of Different GEBD Methods

As shown in Section 3.1, we suppress the negative impact of low-quality generalized event boundaries on VAD performance in several ways. In addition, we select different GEBD models trained on Kinetics-GEBD dataset [59] to demonstrate the stability of the above strategies. Table 9 reveals a direct correlation between their VAD performance on UCF-Crime and their original Kinetics-GEBD [31] dataset results. This confirmation highlights two key findings: (1) The quality of GEBD models remains an influential factor, as improved GEBD implementations consistently yield better performance; (2) Our architectural innovations demonstrate robust adaptability to boundary quality variations.

From a domain shift perspective, Kinetics-GEBD's open-world diversity provides transferable representations superior to those of constrained anomaly datasets (UCF-Crime, XD-Violence, and MSAD). This GEBD-based pre-training aligns with established transfer learning paradigms, boosting cross-domain detection robustness.

Table 9: Results of VADTree based on different GEBD models on UCF-Crime dataset.

GEBD Method	Kinetics-GEBD-Val F1(%)	Kinetics-GEBD-Test F1(%)	UCF-Crime AUC(%)
SceneDetect[4]	-	-	80.00
BasicGEBD-ResNet50	73.70	76.80	82.85
EfficientGEBD-ResNet18-L4	78.20	-	84.70
EfficientGEBD-ResNet50-L4 (Ours)	**78.64**	**78.70**	**84.74**

C.3 Effect of Intra-cluster Node Refinement Configuration

The neighborhood size parameter K in Eq. 8 governs the trade-off between localized feature precision and noise suppression. Table 10 demonstrates substantial AUC gains from neighborhood node refinement: 9.88% for VADTree-Fine and 7.14% for VADTree-Coarse when expanding K from 0 to 10. Both clusters exhibit maximal improvements within this critical initialization range. Performance stabilizes between K=10 and K=15, with the fine and coarse clusters maintaining AUC of 83.03–83.05% and 82.55–82.81%, respectively. A gradual degradation when K exceeds 15 indicates that the optimal balance between contextual integration and noise suppression has been reached within this range.

Table 10: Influence of top-**K** weighted neighborhood nodes on AUC (%).

K	0	5	10	15	20	25
VADTree-Coarse	75.67	81.84	82.81	82.55	81.96	81.73
VADTree-Fine	73.17	79.77	83.05	83.03	82.65	82.43

The temperature coefficient τ in Eq. 8 regulates the entropy characteristics of Softmax-derived distributions while maintaining ordinal relationships between elements. Our empirical analysis (Table 11) reveals that as τ approaches zero ($\tau = 0.001$), the distribution collapses into a degenerate form concentrated solely on the maximal element, equivalent to a non-weighted selection. Progressively increasing the parameter to moderate values produces an AUC plateau of 83.05% for the VADTree-Fine exhibiting minimal variance. Notably, excessive temperature values ($\tau = 100$) induce uniform distributions, degrading performance to 82.43% AUC for VADTree-Fine. This analysis indicates that the optimal parameter range is $\tau \in [0.01, 1]$, where an optimal balance is achieved between distribution sharpness and model stability. For the experiments of VADTree-Coarse, we can get similar conclusions.

[4] https://github.com/Breakthrough/PySceneDetect

Table 11: Influence of softmax temperature τ on AUC (%).

τ	0.001	0.01	0.1	1	10	100
VADTree-Coarse	78.72	80.68	82.81	82.42	82.21	82.20
VADTree-Fine	77.83	80.72	83.05	83.05	83.02	83.02

Table 13: Ablation study of VADTree components on the UCF-Crime dataset. The upper and lower panels present experiments using HGTree and 10 seconds fixed-length sliding temporal window (TW) sampling respectively.

HGTree Fine Cluster	Prior-infused f_{VLM}	Refinement	Correlation	AUC (%)
✓	✗	✓	✓	83.08
✓	✓	✗	✓	77.97
✓	✓	✓	✗	83.05
✓	✓	✓	✓	**84.74**
✗	✗	✗	✗	72.93
✗	✓	✗	✗	75.21
✗	✗	✓	✗	80.62
✗	✓	✓	✗	82.81

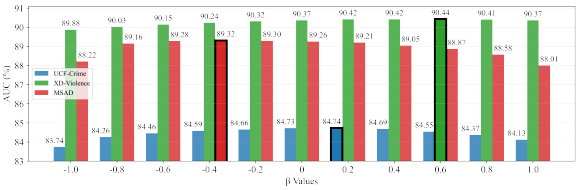
C.4 Effect of Inter-cluster Node Correlation Configuration

The β coefficient regulates parent-child node interplay in our cohesion-driven correlation (Eq. 10). Notably, when $\beta = 0$, the correlation operation degenerates to a simple average of anomaly scores from parent and child nodes. As quantified in Figure 4, the optimal control coefficient $\beta = 0.4$ delivers peak AUC performance at 84.74% for UCF-Crime dataset, indicating an effective equilibrium between parent node contextual integration and child nodes semantic specificity. Additionally, limited AUC fluctuation demonstrates the hierarchy's inherent noise suppression capability. This validates our weighted design as an effective strategy for multi-granularity fusion.

Table 12: Performance comparison demonstrating the efficacy of inter-cluster correlation. The integration of hierarchical clusters in VADTree yields the highest AUC.

Datasets	UCF-Crime	XD-Violence	MSAD
VADTree-Fine	83.05	90.04	86.71
VADTree-Coarse	82.81	89.36	87.01
VADTree	**84.74**	**90.44**	**89.32**

Figure 4: Influence of inter-cluster-correlation control coefficient β on AUC.



C.5 More Ablation Experiments

Our additional ablation analysis examines the contribution of each component in VADTree , namely the HGTree fine cluster, prior-infused node scoring, intra-cluster node refinement, and inter-cluster node correlation, to assess their individual impact on performance. We also evaluate the effectiveness of our components on the 10s fixed-length sliding temporal window (TW) sampling method. Table 13 shows the results of all ablated variants of VADTree. The experiment shows that each component has a significant impact on our final results. At the same time, these components are still effective for methods using fixed-length sliding temporal window sampling.

C.6 Comparison of VADTree and Different Video Sampling Methods

In this experiment, we conducted a comparative analysis of VADTree against mainstream video sampling approaches, focusing on final anomaly detection performance and computational efficiency. The fixed-length sliding temporal window (TW) method, employed by LAVAD and VERA [55, 50], serves as our primary comparison method. Additionally, we propose three metrics to evaluate video sampling efficiency: (1) Number of Segments (NoS), defined as the total number of video segments

sampled from the test dataset; (2) Mean Intersection over Union (mIoU), computed by first identifying the maximum temporal IoU between each anomalous event and all sampled segments within a video, then averaging these maximum values across all events; (3) Mean Intersection Frames (mIF), similar to mIoU, replaces the IoU metric with the number of intersecting frames between sampled video segments and ground truth abnormal segments.

Table 14: Results of VADTree variants with different video sampling methods on the UCF-Crime Dataset. 16f represents a stride of 16 frames. **NoS** indicates the number of generated video segments. **mIoU** and **mIF** are used to measure the quality of video sampling.

Method	TW Length	Stride	NoS↓	mIoU↑	mIF↑	AUC (%)
sliding TW	5s	5s	7558	0.41	122	82.06
sliding TW	10s	10s	3852	0.40	191	82.81
sliding TW	20s	20s	1994	0.33	265	81.33
sliding TW	10s	16f [55, 50]	69634	0.51	210	82.87
VADTree-Coarse	-	-	2248	0.37	369	82.81
VADTree-Fine	-	-	6365	0.40	233	83.05
VADTree	-	-	8613	0.47	343	**84.74**
VADTree + Redundant	-	-	12440	0.52	456	-

As demonstrated in Table 14, non-overlapping implementations of the TW strategy exhibit poor alignment with anomalous events. While dense overlapping sampling with short strides (16 frames) [55, 50] marginally improves AUC ROC it produces 8× more segments than VADTree, incurring significant computational costs without commensurate performance benefits. Our proposed VADTree achieves superior anomaly detection performance while maintaining comparable computational efficiency to non-overlapping TW baselines, demonstrating effective balance between precision and resource utilization.

C.7 Computational Analysis

According to the performance report by LAVAD, its VLM Caption module integrates the results of five BLIP-2 models. The parameter counts used by LAVAD are 3.6 times that of our method VADTree (as shown in Table 15) . We display the total inference time (GPU hours) of LAVAD and VADTree on two NVIDIA GeForce RTX 3090 GPUs in Table 16. The time consumption of the VLM Caption, LLM Summary, and LLM Scoring parts of LAVAD is estimated.

A closer examination of Table 16 reveals the following key observations: (1) VADTree-Coarse requires less than 30% of LAVAD's GPU hours (16.5 vs. 55.9) while achieving a 2.53% higher AUC (82.81 vs. 80.28) on UCF-Crime. This confirms that our method achieves a better trade-off between computational efficiency and detection accuracy compared to LAVAD. (2) Our VADTree framework is highly flexible, with the core HGTree construction process being computationally efficient. Both the VLM and LLM components are modular, allowing for adjustments based on computational constraints. Importantly, high-cost inference models are not essential for VADTree's effectiveness. (3) The default VADTree's inference time is primarily influenced by the reasoning phase of DeepSeek-R1-Distill-Qwen-14B-think in the LLM scoring module. Replacing it with faster variants (e.g., DeepSeek-R1-Distill-Qwen-14B-no-think or t5gemma-9B-2B) significantly reduces inference time. We intentionally preserve the "Think" process because it generates valuable intermediate reasoning steps that significantly enhance anomaly interpretation. Even without this phase, our variants outperform LAVAD in AUC performance while maintaining lower inference times in all cases.

Table 15: Component-level parameter analysis of VADTree and LAVAD.

Methods	HGTree Construction	Video/Text Encoding	VLM Caption	LLM Summary	LLM Scoring	Total
LAVAD	-	ImageBind_Huge-1.2B	OPT-6.7B × 2 + FLAN-T5XL-3B × 2 + FLAN-T5XXL-33B	Llama-2-13B-chat	Llama-2-13B-chat	79.6B
VADTree	ResNet50-25.6M	ImageBind_Huge-1.2B	LLaVA-NeXT-Video-7B	-	DeepSeek-R1-Distill-Qwen-14B	22.2B

Table 16: Component-level inference time consumption analysis of VADTree and LAVAD on UCF-Crime dataset. The bold font indicates that VADTree's total GPU hours is lower than that of LAVAD.

Methods	HGTree Construction	Video/Text Encoding	VLM Caption	LLM Summary	LLM Scoring	Total (GPU hours)	AUC(%)
LAVAD	-	5.1h	20h	7.7h × 2	7.7h × 2	55.9	80.28
VADTree-Coarse	0.3h	0.2h	5.2h × 2	-	2.8h × 2	**16.5**	82.81
VADTree-Fine	0.3h	0.4h	14.8h × 2	-	7.9h × 2	**46.1**	83.05
VADTree	0.3h	0.6h	20.0h × 2	-	10.7h × 2	62.3	84.74
VADTree-Coarse	0.3h	0.2h	5.2h × 2	-	0.6h × 2 (no Think)	**12.1**	82.83
VADTree-Fine	0.3h	0.4h	14.8h × 2	-	1.2h × 2 (no Think)	**32.7**	82.72
VADTree	0.3h	0.6h	20.0h × 2	-	1.8h × 2 (no Think)	**44.5**	84.65
VADTree-Coarse	0.3h	0.2h	5.2h × 2	-	0.1h × 2 (t5gemma-9B-2B)	**11.1**	82.21
VADTree-Fine	0.3h	0.4h	14.8h × 2	-	0.2h × 2 (t5gemma-9B-2B)	**30.0**	82.19
VADTree	0.3h	0.6h	20.0h × 2	-	0.3h × 2 (t5gemma-9B-2B)	**41.5**	84.00

C.8 Stability Analysis

We conduct error analysis of VADTreeon UCF-Crime dataset and report their mean and variance. The randomness of the experimental results mainly comes from the randomness of the generated content during VLM and LLM inference. As shown in Table 17, the δ across all configurations is statistically insignificant compared to the performance gaps between different methods (Table 1 and Table 2). This confirms that the observed performance is robust against experimental randomness.

Table 17: Stability analysis of VADTree with different HGTree configurations.

Method	Exp-1	Exp-2	Exp-3	Mean Results	δ
VADTree-Coarse	82.81	82.75	82.92	82.83	0.17
VADTree-Fine	83.05	82.86	83.05	82.99	0.19
VADTree	84.74	84.49	84.73	84.65	0.25

C.9 Additional Case Studies and Qualitative Results

For complex anomalies such as arrests and burglaries, the fixed temporal sampling used in LAVAD and frame-level caption aggregation can lead to missed long-range semantic dependencies, resulting in inaccurate or incomplete interpretations of abnormal events. To show the superiority of our VADTree, we conduct a qualitative analysis using two video samples from the UCF-Crime dataset. The Figure 5 displays some key inference information. LAVAD's frame-level semantic aggregation often leads to hallucinations and struggles to accurately identify long-range abnormal events. In contrast, VADTree excels at detecting sub-events (e.g., "chased by another person," "attempted break-in") while also synthesizing long-range contextual clues (e.g., "putting it all together"). The above qualitative results will be added to our revision.

In the example shown in Figure 6, the overall score of fine clusters fluctuates greatly ([0.7, 0.9, 0.9, 0.9, 0.6] in ② and [0.9, 0.6, 0.9, 1.0] in ③), while the anomaly score of coarse cluster ② node is low. After refinement, the above situation is improved, but the correct anomaly score is suppressed (③: 0.8→0.64). In addition, the anomaly reasoning of each node is independent and lacks mutual reference and dimensions; this may lead to unstable anomaly scores in the same video segment. The final inter-cluster correlation resolves these shortcomings well. This example indicates that VADTree can significantly enhance the robustness of training-free VAD inference.

D Limitations

Like existing training-free VAD methods, VADTree's performance relies heavily on the visual perception capabilities of VLMs. Most VLMs mainly focus on more complex semantic understanding and reasoning, and there are still significant limitations in accurately obtaining various small shallow abnormal semantics (such as the flame of a lighter after an explosion). This constraint may hinder accurate anomaly detection. If essential visual characteristics are not captured during the encoding stage, it becomes unlikely for VADTreeto effectively carry out abnormal reasoning or perform

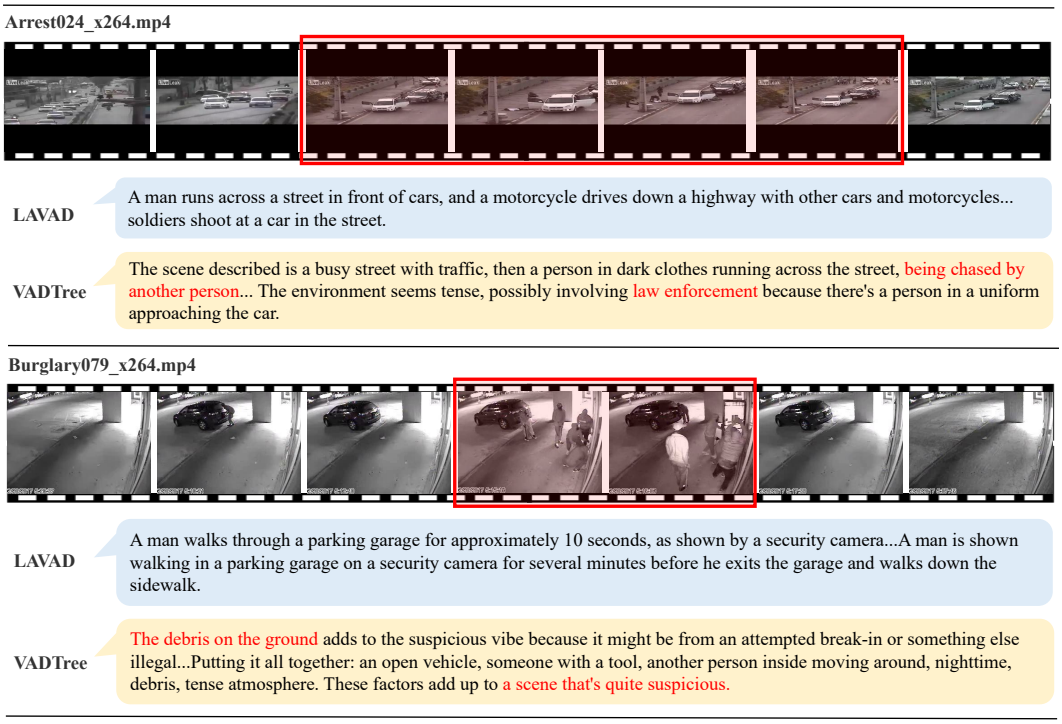


Figure 5: Case studies of complex anomalies: arrests and burglaries. VADTree excels in generating accurate explanations by modeling hierarchical events and long-range dependencies, whereas LAVAD produces incomplete descriptions due to frame-level limitations.

temporal inter-cluster corrections. Therefore, a primary challenge for VLM based VAD is to guarantee that both visual and temporal features are adequately captured.

E Broader Societal Impacts

Our training-free paradigm enables efficient video anomaly detection with minimal computational costs, yet its deployment in safety-critical scenarios (e.g., public surveillance) requires careful consideration of privacy-preserving mechanisms. While avoiding explicit biometric data processing, the prior knowledge base could theoretically retain sensitive environmental patterns. We advocate transparency audits to mitigate potential privacy risks in real-world implementations.

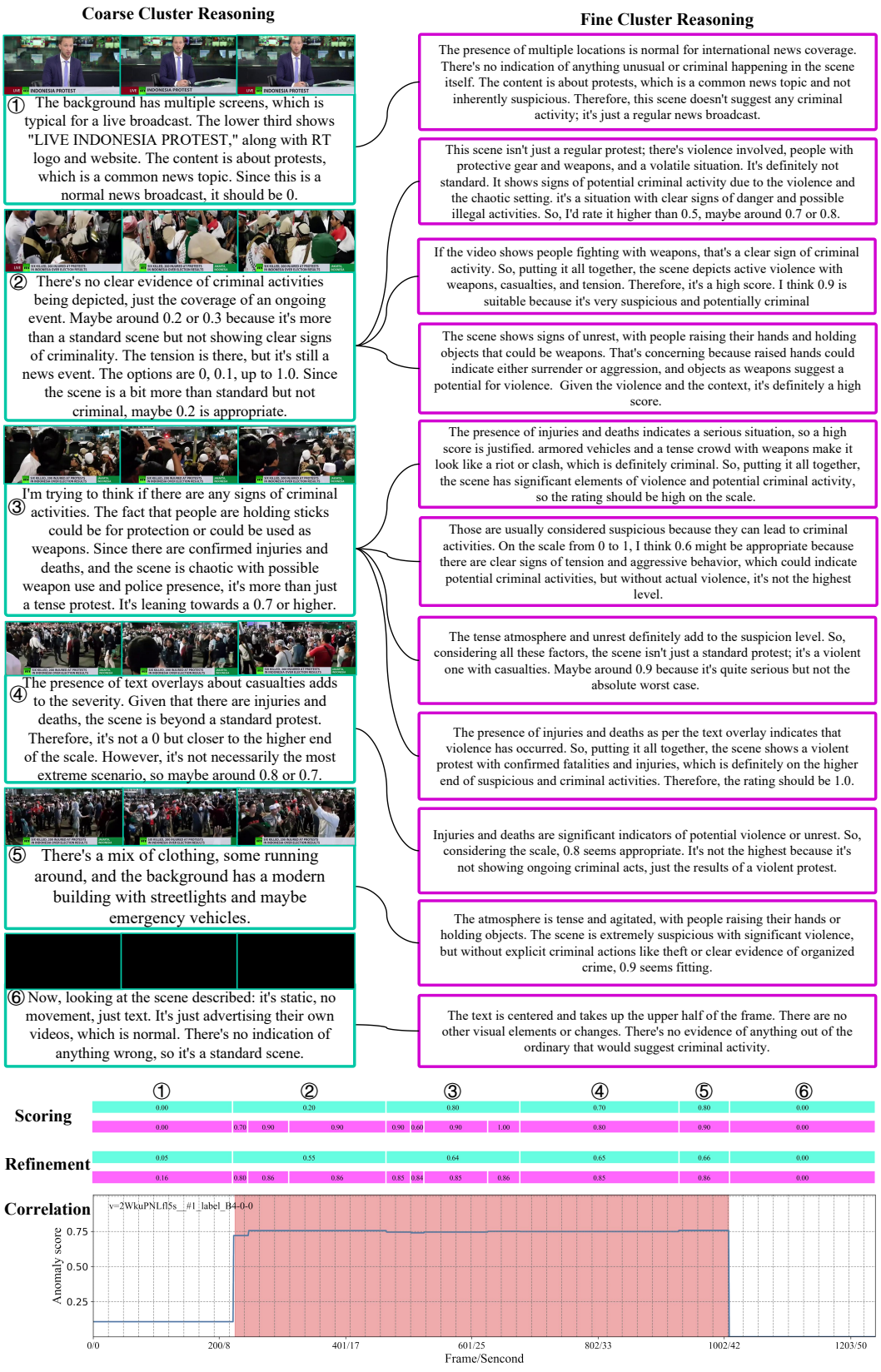


Figure 6: Qualitative results of VADTree on a test video, showcasing anomaly explanation and scoring based on HGTree representation. The scoring and text explanation include initial anomaly scores (Scoring), refined scores (Refinement), and final anomaly scores (Correlation) Based on HGTree for video representation, the different granularity reasoning results of coarse and fine clusters on anomalies can complement each other.