Towards Video Anomaly Detection from Event Streams: A Baseline and Benchmark Datasets

Peng Wu[1], Yuting Yan[1], Guansong Pang[2], Yujia Sun[3], Qingsen Yan[1], Peng Wang[1⋆], and Yanning Zhang[1]

[1] School of Computer Science, Northwestern Polytechnical University, China
{xdwupeng, xgdyanyuting, qingsenyan}@gmail.com, {peng.wang, ynzhang}@nwpu.edu.cn
[2] School of Computing and Information Systems, Singapore Management University, Singapore
gspang@smu.edu.sg
[3] School of Artificial Intelligence, Xidian University, China
yjsun@stu.xidian.edu.cn

Abstract. Event-based vision, characterized by low redundancy, focus on dynamic motion, and inherent privacy-preserving properties, naturally fits the demands of video anomaly detection (VAD). However, the absence of dedicated event-stream anomaly detection datasets and effective modeling strategies has significantly hindered progress in this field. In this work, we take the first major step toward establishing event-based VAD as a unified research direction. We first construct multiple simulated event-stream based benchmarks for video anomaly detection, featuring synchronized event and RGB recordings. Leveraging the unique properties of events, we then propose an EVent-centric spatiotemporal Video Anomaly Detection framework, namely **EWAD**, with three key innovations: an event density aware dynamic sampling strategy to select temporally informative segments; a density-modulated temporal modeling approach that captures contextual relations from sparse event streams; and an RGB-to-event knowledge distillation mechanism to enhance event-based representations under weak supervision. Extensive experiments on three benchmarks demonstrate that our EWAD achieves significant improvements over existing approaches, highlighting the potential and effectiveness of event-driven modeling for video anomaly detection. The benchmark datasets and code are available at https://github.com/kanyutingfeng/EWAD.

Keywords: Video anomaly detection · Event-based vision · Vision-language model

1 Introduction

With the increasing demand for video understanding in public safety, video anomaly detection (VAD) has emerged as a critical task for identifying potential

⋆ Corresponding author.

risks and abnormal behaviors [2, 19]. In recent years, driven by advances in deep learning, RGB-based VAD methods have achieved remarkable progress. However, RGB videos inherently suffer from fixed frame rates, high data redundancy, and delayed dynamic perception, limiting their effectiveness in high-density or rapidly changing environments.

In contrast, event cameras, as an emerging type of vision sensor, offer an alternative paradigm for video analysis. Operating asynchronously at the pixel level, they emit events only when brightness changes exceed a threshold, encoding the time, location, and polarity of such changes. This results in high temporal resolution, low latency and inherently sparse outputs [17]. Rather than densely capturing static scenes, event cameras naturally emphasize motion and dynamic changes, attributes highly aligned with the nature of anomalies in videos. These characteristics make event data particularly well-suited for the VAD task, as they enhance temporal sensitivity, reduce redundant background noise, and preserve privacy.

Despite growing interest in event-based vision for low-level tasks such as image restoration, its application to VAD remains largely unexplored. Two key challenges hinder progress in this field: the lack of high-quality event-stream VAD datasets [21], and the limited transferability of existing VAD models originally designed for synchronous RGB videos to the asynchronous and sparse nature of event data. Currently, there has been little research on effective modeling strategies capable of fully leveraging the spatiotemporal representation inherent in event streams. For example, Qian et al. [21] proposed a VAD framework based on the spiking neural network (SNN) using event data, demonstrating preliminary effectiveness. However, this line of research largely overlooks the powerful cross-modal alignment capabilities of recent vision-language models (VLMs) and the complementary information available in paired RGB frames.

In this paper, we propose a unified framework that tackles two key challenges: the scarcity of dedicated benchmarks and the difficulty of modeling event-based spatiotemporal dynamics. We first construct large-scale simulated event-based benchmarks through temporal alignment of event streams with widely used RGB-based VAD datasets (UCF-Crime [24], CCTV-Fight [20], and UBnormal [1]). These benchmarks enable systematic evaluation of event-based anomaly detection across diverse scenes and scenarios. Building upon this foundation, we then develop an **EV**ent-centric **V**ideo **A**nomaly **D**etection baseline named **EWAD** tailored to spatiotemporal event understanding. Unlike conventional cameras that capture RGB frames at fixed intervals, event cameras asynchronously record changes in brightness at the pixel level, producing sparse signals that primarily reflect scene dynamics. Such representations tend to highlight motion-related patterns while suppressing static background information. These characteristics provide complementary cues to frame-based representations and may be beneficial for motion-centric tasks such as anomaly detection. Motivated by this observation, EWAD is designed to better exploit the temporal dynamics and sparse structure of event streams for anomaly detection.

Specifically, we first propose an event-density aware dynamic sampling (EDS) strategy that adaptively selects temporally diverse and discriminative frames, i.e., emphasizing those moments likely to contain anomalous cues, thus effectively leveraging the inherently sparse and bursty nature of event data while ensuring comprehensive temporal coverage critical for efficent anomaly detection. Then, we design an event-modulated distance-decay attention (EDA) mechanism that jointly encodes event density and inter-event intervals, allowing the model to capture long-range temporal dependencies by dynamically adjusting time perception. By aligning attention dynamics with event saliency, this mechanism helps preserve sparsity while enhancing temporal awareness. Notably, this mechanism is lightweight and can be seamlessly integrated into existing temporal modeling modules. Furthermore, to compensate for the limited supervisions provided by hard labels and deficient information inherent in event-only data, we propose a cross-modal knowledge distillation (KD) strategy, wherein high-level priors from pretrained RGB models are transferred to the event-based model. This distillation strategy enables the event model (student) to acquire the implicit knowledge embedded in the teacher model, such as inter-class similarity relationships and the structural organization of the feature space. This distillation mechanism is only applied during training stage. Finally, we extend our method to explore event-based spatial anomaly localization, demonstrating the capability of event streams in delivering fine-grained cues for regional abnormality understanding.

In summary, our contributions are four-fold:

- We introduce large-scale benchmarks that align event streams with RGB datasets to enable systematic evaluation of event-centric video anomaly detection across varied scenes.
- We propose a dynamic sampling strategy and an event-modulated attention mechanism tailored to event streams, improving temporal sensitivity and efficiency.
- We develop a knowledge distillation framework that transfers high-level semantics from RGB-based models to enhance the representational capacity of event-only detectors.
- Extensive experiments on three benchmarks demonstrate that our EWAD achieves state-of-the-art performance among event-based methods and establishes a strong baseline for future research in event-centric video anomaly detection.

2 Related Work

2.1 Deep Learning for Video Anomaly Detection

Deep learning methods have significantly advanced VAD by enabling data-driven deep neural network (DNN) pipelines. However, their success often hinges on the availability of large-scale, finely annotated datasets, which are expensive and

time-consuming to obtain. To address this limitation, weakly supervised learning has emerged as a promising alternative. It reduces the annotation burden while maintaining competitive detection performance, and has become a vibrant research direction in recent years. Most weakly supervised methods adopt a multiple instance learning (MIL) framework [11], where only video-level labels are provided, and the goal is to infer segment- or frame-level anomaly scores. A seminal work by Sultani et al. [24] introduced MIL to VAD, laying the foundation for subsequent weakly supervised approaches. He et al. [9] tackled the imbalance between normal and abnormal videos via adversarial training and hard-negative mining to enhance detection with limited anomalies. Su et al. [23] proposed a semantic-driven consistency scheme to align object localization and features, enabling interpretable anomaly detection.

With the rapid development of VLMs, especially the CLIP model [22], cross-modal semantic understanding has been significantly enhanced. CLIP projects both images and textual descriptions into a shared embedding space, enabling efficient visual-textual alignment and demonstrating impressive performance across a wide range of vision tasks [35]. In the context of weakly supervised VAD, a number of recent methods, such as VadCLIP [38], OVVAD [36], STPrompts [37], and TPWNG [39], leverage CLIP's vision-language alignment by incorporating textual prompts, which embed high-level prior knowledge into the anomaly detection process. Building on the powerful reasoning capabilities of VLMS and large language models (LLMs) [27], several inference-based approaches have emerged recently, including LAVAD [41], Holmes-VAU [42], Vera [40], Ex-VAD [12], and SlowfastVAD [5]. These methods eliminate the need for additional training and rely instead on zero-shot or few-shot inference mechanisms. While promising, such approaches often suffer from slower inference speed and have room for improvement in detection performance.

2.2 Event Cameras in Computer Vision

Event cameras are bio-inspired vision sensors that offer several unique advantages, including ultra-high temporal resolution, low latency, high dynamic range. Unlike conventional frame-based cameras, event cameras asynchronously capture per-pixel intensity changes, making them particularly suitable for high-speed motion analysis and low-light conditions. These properties have spurred interest in applying event cameras to a variety of computer vision tasks, including image restoration [30], optical flow estimation [8], and object recognition [3]. Despite their advantages, the application of event cameras to video anomaly detection remains in its early stages. Event streams naturally suppress static background and emphasize dynamic changes, which can be advantageous for detecting anomalous behaviors. For instance, Qian et al. [21] extended the existing UCF-Crime dataset by collecting corresponding event data using an event camera, and proposed a SNN–based VAD framework that demonstrated preliminary effectiveness. Nonetheless, this line of research faces several limitations: the high acquisition cost associated with event cameras, the lack of diverse event-based datasets, the underutilization of the powerful multimodal capabilities of models

Table 1: Summary of event-based VAD datasets.

Dataset	Total videos	Length	Train videos	Test videos	Classes	Scenes	Resolution
UCF-Crime	1900	128 hours	1610	290	14	-	346 × 260
CCTV-Fights	1000	18 hours	750	250	1	-	346 × 260
UBnormal	543	2.2 hours	268	211	23	29	640 × 480

like CLIP, the insufficient exploration of complementary information between RGB and event modalities, and the lack of investigation into the potential of event data's low-redundancy characteristics for spatial localization.

3 Dataset

3.1 Dataset Overview

The quality of datasets fundamentally determines the performance and comparability of VAD models. However, the field currently faces a critical shortage of publicly available event-based VAD datasets, severely hindering research progress. Real event data collection is costly, highly sensitive to environmental conditions, and further complicated by the inherent rarity of anomalous events, making the construction of large-scale, real-world datasets exceedingly difficult.

To address this limitation, we adopt a simulation-based strategy to generate high-quality event data from widely used RGB-based VAD benchmarks, including UCF-Crime [24], CCTV-Fights [20], and UBNormal [1], which cover diverse scenes and anomaly types. Table 1 summarizes detailed statistics of datasets.

To the best of our knowledge, this work introduces the first publicly available large-scale event-based VAD benchmark constructed and diverse real-world scenarios, such as urban streets and indoor surveillance. The datasets cover anomalies including shooting, fall, fight, road accident, etc, with each video comprising millions of event points, highlighting the ultra-high temporal resolution of event cameras. This large-scale, high-fidelity resource provides a solid foundation for training and evaluating weakly supervised event-centric VAD models and is expected to significantly accelerate research in this emerging domain.

3.2 Event Data Generation

We employ the state-of-the-art v2e simulator [10], which models optical flow and pixel-level brightness changes to convert RGB videos into realistic event streams. The use of v2e is motivated by its widespread adoption in the event-vision community, where multiple benchmarks and prior studies rely on simulated event data [4, 14, 16, 18] for large-scale evaluation. This process preserves the spatiotemporal dynamics and anomaly labels of the original videos, ensuring temporal continuity and spatial consistency in the generated data. We upsampled

Fig. 1: Examples of generated event data and RGB counterparts, sourced from UCF-Crime, UBnormal, and CCTV-Fight.

all videos to 100 fps during simulation to leverage event cameras' high temporal resolution. More information on setting the v2e simulator parameters can be found in the supplementary materials. As shown in Figure 1, the simulated event data clearly captures the motion of foreground objects while exhibiting minimal redundant dynamic visual information in the background.

3.3 Adaptive Event Frame Generation

Event data can be represented in various forms. In this work, we adopt event frames by discretizing asynchronous events into frame sequences for compatibility with vision models. Traditional methods often fix the number of events per frame, which may cause sparsity-induced blur or density-induced ghosting. To mitigate this, we propose an adaptive frame generation strategy that adjusts event density, striking a balance between temporal resolution and data sparsity.

The size and timestamp of each bin are determined by following the feature extraction strategy used for RGB frames [24]. To maintain consistency with the original RGB video's frame rate, we sample one frame every 16 frames as the center of a time window, and extend the window by 8 frames before and after the center, thereby dividing the event data into non-overlapping time segments to ensure each event frame corresponds to unique and non-redundant event data.

We set a baseline of one-tenth the average event points per bin (i.e., $mean \times 1/10$). We observe that frames with excessive events require downsampling to avoid ghosting artifacts, while those with insufficient events benefit from upsampling to preserve structural integrity. Ghosting refers to density accumulation artifacts, not optical blur. Given that the event count distribution is often skewed by a small number of high-activity frames, we empirically use the median rather than the mean event count per video as a more robust reference for sparsity control. We define a sparsity coefficient as: $sc = N_c/median$, where N_c represents the number of event points contained within the bin, and a larger sc indicates

a denser frame and thus necessitates fewer events per bin, whereas a smaller sc corresponds to a sparser frame requiring more events. The final number of event points per bin is computed as,

$$EventNum = \frac{1}{10} \times mean + \frac{1}{sc} \times median \tag{1}$$

To further mitigate ghosting in videos characterized by intense global motion, we also enforce an upper bound 10000 on the number of events per bin. This constraint ensures temporal coherence and avoids frame saturation in highly dynamic scenes.

4 Methodology

4.1 Overview

The overall framework of EWAD is illustrated in Figure 2. Our method requires paired event and RGB modalities during the training phase but operates solely on event streams during inference. Following feature extraction, we introduce a dynamic sampling strategy EDS to prioritize temporally informative segments that are more likely to contain abnormal patterns. This allows the model to focus on frames with higher anomaly potential during training. Next, we propose a temporal modeling mechanism EDA to model long-range temporal dependencies within the sparse event sequences by jointly encoding event density and inter-event intervals. Then temporally enhanced features are forwarded to two lightweight prediction heads: a binary classification head for frame-level anomaly confidence scores, and a multi-class alignment head for semantic category prediction. In addition to conventional hard supervision via MIL, we further incorporate a soft supervision paradigm based on cross-modal distillation from an RGB-based teacher model. Specifically, we distill both the binary anomaly scores and multi-class predictions to guide the learning of the event-based student model. This multi-level distillation significantly improves the model's discriminative capacity under weak supervision, enabling more effective anomaly detection without relying on dense annotations. Finally, we leverage both the temporal anomaly scores and the spatial distribution of event activations to generate training-free spatial anomaly heatmaps.

4.2 Event-Density Aware Dynamic Sampling

Event cameras produce asynchronous, high-temporal-resolution data triggered by luminance changes, resulting in inherently sparse and non-uniform information distribution over time. Notably, anomalous activities often cause abrupt intensity variations, leading to localized bursts of event density. In this context, uniform sampling strategies may waste resources on redundant frames from low-activity regions while missing discriminative cues in dynamic segments. To address this, we propose an event-density aware dynamic sampling strategy that

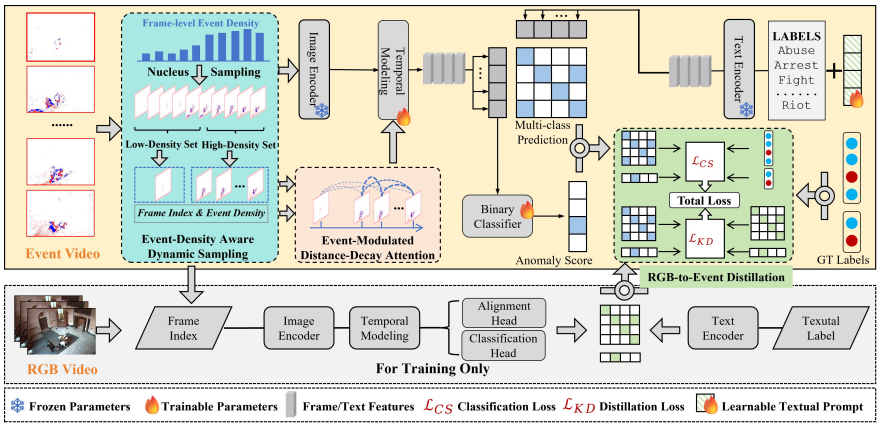

Fig. 2: Overview of our proposed method EWAD.

adaptively selects frames based on temporal event density, enabling the model to focus on regions with high potential anomaly signals. This not only enhances temporal modeling effect but also improves anomaly sensitivity under weak supervision.

Specifically, the event stream is partitioned into discrete frames. Let n_i denote the number of events occurring within the i^{th} frame. The event density d_i for this frame is defined as the proportion of events in the i^{th} frame relative to the total number of events across all frames, and is computed as:

$$d_i = \frac{n_i}{\sum_j n_j} \tag{2}$$

Given the temporal sequence and corresponding event density information, our goal is to prioritize frames from high-density regions, where anomalies are more likely to occur, while retaining frames from low-density regions to preserve semantic context such as background information. This renders traditional top-k sampling suboptimal, as it rigidly excludes low-density frames, potentially discarding informative content. Inspired by sampling strategies in LLMs, we adopt a modified dual-interval nucleus sampling. We first define a density threshold τ_d (e.g., 0.95) and sort all event frames in descending order of density. We then accumulate densities until their sum exceeds τ_d, designating these frames as the high-density set, with the remaining forming the low-density set. Based on a predefined sampling ratio (e.g., 8:2), we severally perform multinomial sampling without replacement within each set to select representative key frames for training. This dynamic sampling strategy EDS effectively emphasizes informative frames from dense event regions while ensuring the inclusion of diverse low-density frames. As a result, the model benefits from enhanced discriminative capability, improved robustness, and better generalization across varying event densities.

4.3 Event-Modulated Distance-Decay Attention

Event streams are inherently non-uniform in time, with varying densities reflecting different motion intensities. Meanwhile, existing studies [45] have shown that temporal perception varies under different event densities. Therefore, incorporating event density into temporal modeling enables the model to dynamically adjust its sense of time, leading to more accurate anomaly detection across diverse motion patterns. To this end, we propose an event-modulated distance-decay attention mechanism, which integrates both temporal distance and event density to dynamically weight temporal dependencies.

Given sampled features with corresponding timestamps $t \in \mathbb{R}^T$ and event densities $d \in \mathbb{R}^T$, we first normalize timestamps per sequence to a fixed scale. For each token pair (i, j), the attention weight is defined as:

$$w_{ij} = \frac{\exp\left(-\lambda \cdot \frac{|\tilde{t}_i - \tilde{t}_j|}{d_j + \epsilon}\right)}{\sum_{k=1}^{T} \exp\left(-\lambda \cdot \frac{|\tilde{t}_i - \tilde{t}_k|}{d_k + \epsilon}\right) + \epsilon} \tag{3}$$

where $\lambda > 0$ controls the decay rate, and ϵ is a small constant added for numerical stability. This formulation prioritizes closer and denser tokens, enhancing long-range temporal modeling while preserving sparsity. In this work, we seamlessly integrate this plug-and-play mechanism into the temporal modeling module of VadCLIP [38].

4.4 RGB-to-Event Knowledge Distillation

Due to the limited supervision provided by hard labels and the inherent learning difficulty associated with event-based data [26], we introduce an RGB-based VAD model as a teacher to guide the training of the event model. Furthermore, considering that foundation models like CLIP are pre-trained extensively on massive RGB-text pairs, their semantic priors are inherently tied to the RGB modality. Direct application of CLIP to sparse and asynchronous event streams often struggles to fully unlock these pre-trained priors due to the significant modality gap. Therefore, employing an RGB-based teacher serves as a crucial bridge, transferring these high-level, RGB-centric semantics to the event-based student. Our cross-modal knowledge distillation strategy comprises two complementary components:

Binary classification logit-level distillation. To align the anomaly confidence scores between the event model and the RGB teacher model, we minimize the mean squared error between their binary classification probabilities. Let $a_i^e \in [0, 1]$ and $a_i^r \in [0, 1]$ denote the predicted anomaly confidence scores of the event model and RGB teacher model, respectively, for the i^{th} video frame. The binary distillation loss is defined as:

$$\mathcal{L}_{\text{bin}} = \frac{1}{T} \sum_{i=1}^{T} (a_i^e - a_i^r)^2 \tag{4}$$

Multi-class logit-level distillation. Inspired by recent advances in multi-class knowledge distillation, particularly those incorporating logit standardization [25], we further align the detailed categorical predictions between the two modalities to enhance the event model's capacity for fine-grained anomaly recognition. Let $Z^e \in \mathbb{R}^{T\times K}$ and $Z^r \in \mathbb{R}^{T\times K}$ represent the logits of the event model and RGB teacher model, respectively, across K categories.

We first normalize the logits, the benefit of logit standardization lies in encouraging the student model to learn the inter-class relationships predicted by the teacher, rather than attempting to precisely match the raw logits. By mitigating such discrepancies, standardization alleviates the difficulty of distillation and enhances the robustness of the student model,

$$\hat{Z} = \frac{Z - \mu}{\sigma + \epsilon} \tag{5}$$

$$\mu = \frac{1}{K}\sum_{k=1}^{K} Z_k, \quad \sigma = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(Z_k - \mu)^2} \tag{6}$$

We then compute the temperature-scaled softmax distributions:

$$p^e_{i,k} = \text{softmax}\left(\frac{\hat{Z}^e_{i,k}}{\tau}\right), \quad p^r_{i,k} = \text{softmax}\left(\frac{\hat{Z}^r_{i,k}}{\tau}\right) \tag{7}$$

where $\tau > 0$ is a temperature parameter. The multi-class distillation loss is calculated as the averaged Kullback–Leibler divergence, weighted by τ^2:

$$\mathcal{L}_{\text{multi}} = \tau^2 \cdot \frac{1}{T}\sum_{i=1}^{T}\sum_{k=1}^{K} p^r_{i,k} \log \frac{p^r_{i,k}}{p^e_{i,k}} \tag{8}$$

The total knowledge distillation loss combines both components:

$$\mathcal{L}_{\text{KD}} = \alpha\mathcal{L}_{\text{bin}} + \beta\mathcal{L}_{\text{multi}} \tag{9}$$

where α and β are weighting coefficients.

It is important to note that the distillation process is applied only during training. At inference time, the event model operates independently without requiring any RGB input, ensuring both efficiency and practical deployment.

4.5 Training and Inference

During training, we optimize a composite objective loss:

$$\mathcal{L} = \mathcal{L}_{\text{CS}} + \mathcal{L}_{\text{KD}} \tag{10}$$

The classification loss $\mathcal{L}_{CS}$ follows VadCLIP [38]:

$$\mathcal{L}_{CS} = \mathcal{L}_{bce} + \mathcal{L}_{nce} + \mathcal{L}_{cts} \tag{11}$$

where $\mathcal{L}_{bce}$, $\mathcal{L}_{nce}$, $\mathcal{L}_{cts}$ are binary classification, alignment and contrastive losses, respectively.

During inference, temporal anomaly detection is performed by jointly leveraging the classification head and alignment head to identify anomalous frames. For each identified frame, we further analyze the corresponding event data to localize potential spatial anomaly regions. Specifically, we apply a pre-defined threshold to the event map to extract candidate regions, followed by image morphological operations to refine these regions and obtain final bounding boxes.

5 Experiments

5.1 Experimental Setup

Evaluation metrics. We evaluate model performance using the frame-level Area Under the Receiver Operating Characteristic Curve (AUC) for anomaly detection. For the spatial anomaly localization task, we report the Temporal Intersection over Union (TIoU) [15] to measure the accuracy of localized anomalous regions over time.

Implementation details. For classification and alignment heads, we follow the setup of VadCLIP. For training, we use the Adam optimizer with a learning rate of 2×10^{-5}. The model is trained for 10 epochs with a batch size of 128. For EDS, the default sampling length is 256. Following previous work [25], we set the loss weighting parameters α and β as 0.1 and 9, respectively, and the temperature τ for knowledge distillation as 2. We extensively ablated all hyperparameters, the detailed results are provided in the supplementary material.

5.2 Main Experimental Results

Temporal anomaly detection results. As shown in Table 2, we compare our method with existing approaches across different feature types and model architectures, including SNN-DNN, SNN-SNN, and ViT-DNN frameworks.

For a strictly fair comparison under the event-centric paradigm, we adapted representative RGB-based methods (i.e., DeepMIL [24], DMU [44], and VadCLIP [38]) to take our extracted event frame features as input, training and evaluating them on our simulated event benchmarks.

Our proposed method EWAD achieves the best performance across all evaluated datasets. Specifically, on the event-camera data of UCF-Crime, EWAD achieves 72.36% AUC, outperforming the previous state-of-the-art (MSF [21], 65.01%) by +7.35%. Furthermore, EWAD generalizes well to our simulated benchmarks (e.g., reaching 76.55% on simulated UCF-Crime), highlighting the practicality of simulation-derived benchmarks for scalable event-based VAD. In addition, even when using the same ViT-based features, EWAD still shows notable improvements over other ViT-DNN methods such as DMU and VadCLIP. These results not only highlight the superior representational power of ViT features but also demonstrate the effectiveness of our event-specific modules. It

Table 2: AUC comparison on different datasets.

Category	Method	Feat-Arch.	AUC (%)		
			UCF	CCTV	UB
Event-Camera	DeepMIL [24]	SNN-DNN	55.56	-	-
	HLNet [34]	SNN-DNN	58.58	-	-
	RTFM [28]	SNN-DNN	52.67	-	-
	TSA [13]	SNN-DNN	51.86	-	-
	ResSNN [6]	SNN-SNN	53.99	-	-
	PLIF [7]	SNN-SNN	54.74	-	-
	SFormer [43]	SNN-SNN	62.78	-	-
	MSF [21]	SNN-SNN	65.01	-	-
	EWAD (Ours)	**ViT-DNN**	**72.36**	**-**	**-**
Simulated	DeepMIL [24]	ViT-DNN	61.11	54.98	56.52
	DMU [44]	ViT-DNN	73.47	57.85	57.46
	VadCLIP [38]	ViT-DNN	73.01	63.08	53.19
	EWAD (Ours)	**ViT-DNN**	**76.55**	**64.17**	**58.30**

is also worth noting that no knowledge distillation is applied on CCTV-Fight and UBnormal, yet EWAD still outperforms previous methods, underscoring the strong generalization ability of our architecture and training strategy.

Spatial anomaly localization results. As shown in Table 3, our EWAD, leveraging a completely training-free approach that relies solely on event modality, achieves a TIoU of 13.28%. While the localization performance is slightly inferior to the latest RGB-based approaches, it remains competitive when compared to earlier RGB-only methods like C3D (7.20%) and NLN (12.20%). This demonstrates that, leveraging the low-redundant spatial representation of event data, our method can effectively deliver solid anomaly localization performance. Note that the performance metrics for classic methods including TSN, C3D, and NLN are cited from Liu et al. [15] We include these earlier approaches to demonstrate that our training-free, event-based baseline can already rival or surpass established, heavily-engineered spatiotemporal RGB architectures (e.g., 3D CNNs). We clarify that the spatial localization results in this work are intended as a feasibility study, aiming to explore whether event data can support spatial anomaly localization. Meanwhile, the relatively lower TIoU compared to RGB-based methods is primarily attributed to limited temporal discriminability of event-only models.

5.3 Ablation Studies

To assess the contribution of each core component, we conduct comprehensive ablation experiments on UCF-Crime dataset. Table 4 reports the frame-level AUC under different model configurations. As we can see, each component brings clear performance gains. Specifically, EDS demonstrates its effectiveness in prioritizing informative event regions and reducing redundancy. Adding EDA further

Table 3: TIoU comparison on UCF-Crime.

Method	Modality	TIoU(%)
TSN [31]	RGB	2.60
C3D [29]	RGB	7.20
NLN [33]	RGB	12.20
Liu et al. [15]	RGB	16.40
VadCLIP [38]	RGB	22.05
STPrompts [37]	RGB	23.09
EWAD(Ours)	Event	13.28

Table 4: Impact of each component on UCF-Crime.

EDS	EDA	B-KD	M-KD	AUC (%)
×	×	×	×	73.01
✓	×	×	×	73.94
✓	✓	×	×	74.17
×	×	✓	✓	74.45
✓	✓	✓	×	74.53
✓	✓	×	✓	74.58
✓	✓	✓	✓	76.55

enhances temporal modeling, yielding a sustained improvement over the EDS-only setting, which validates the importance of jointly encoding event density and temporal intervals. Notably, the RGB-to-event knowledge distillation strategy alone leads to a substantial 1.44% boost over the baseline, highlighting the advantage of transferring high-level semantic priors from the RGB domain. When all components are integrated, the model achieves the highest performance, with a 3.54% absolute gain over the base configuration, confirming the complementary benefits of our design. The last three rows report the performance under different distillation strategies. Here, B-KD refers to binary knowledge distillation, while M-KD denotes multi-class knowledge distillation. Combining both B-KD and M-KD yields additional improvements, suggesting that binary-level and multi-class-level supervision provide complementary guidance to enhance the discriminative capability of the student model. It is worth noting that our framework is designed to establish a strong uni-modal event baseline, enabling a clearer evaluation of the intrinsic potential of event data without relying on RGB inputs.

5.4 Visualization

Visualization of temporal anomaly detection. As shown in Figure 3, the pink-shaded areas indicate the ground-truth anomalous segments, while the blue curves represent the anomaly scores predicted by our EWAD. Notably, certain scene transitions, such as opening or closing credits, share visual patterns with dynamic events, which introduces significant challenges in accurately detecting anomalies. Nevertheless, EWAD consistently generates high confidence scores that align with the ground-truth abnormal regions, while maintaining low responses in normal intervals, demonstrating its discriminative capability. However, compared to RGB-based methods, the detected abnormal segments appear less temporally continuous, and the distinction between normal and abnormal regions is relatively weaker, highlighting the potential for improvement through multimodal (RGB+Event) fusion.

Visualization of spatial anomaly localization. Figure 4 illustrates spatial anomaly localization results on the simulated event data. The figure shows two bounding boxes: the green box represents the ground truth (GT) anomaly region,

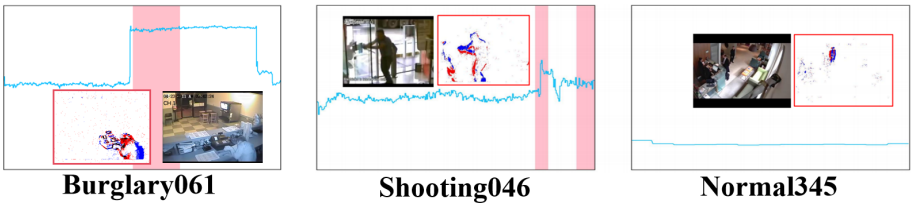

Fig. 3: Qualitative results of temporal anomaly detection.

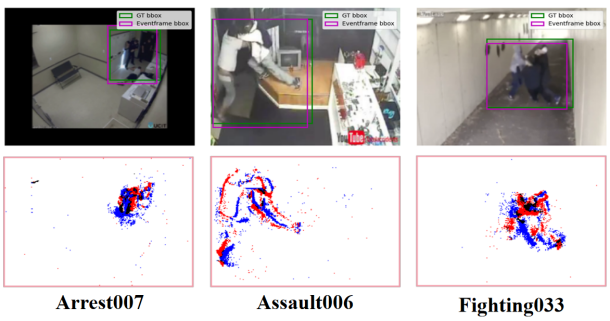

Fig. 4: Visualization of spatial anomaly localization.

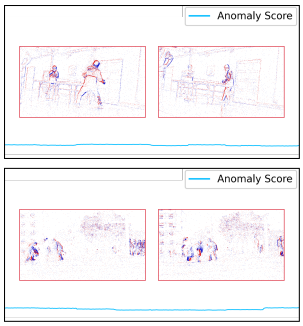

Fig. 5: Anomaly scores of EWAD for examples in EventVOT.

while the pink box indicates our predicted anomaly region obtained from the event frames. As shown, the predicted bounding boxes based on event data align well with the true anomaly locations. This highlights the superiority of the event data for spatial anomaly detection.

5.5 Real Event Data Validation

Event streams in this work are generated using the v2e simulator. As a simulation-based method, the quality of the generated events is influenced by the source RGB videos (e.g., motion blur or exposure conditions), which may introduce artifacts. Despite these limitations, the absence of large-scale event-stream datasets remains a key bottleneck for event-based VAD. Capturing rare anomalous events with real event cameras is extremely challenging in practice, making simulation-based benchmarks a practical step toward enabling systematic study in this direction. Moreover, VAD primarily relies on motion dynamics and temporal irregularities rather than fine-grained visual textures. As a result, moderate simulation artifacts are less detrimental for VAD tasks than for low-level vision problems.

To further assess the real-world applicability of our approach, we conducted validation experiments on 18 randomly selected normal video clips from the real-world EventVOT dataset [32], covering diverse scene categories. Our results show

that 17 out of 18 samples yield consistently low anomaly scores (representative examples in Figure 5), suggesting that EWAD learns motion semantics rather than relying on artifacts introduced by simulation.

6 Conclusion

In this work, we take a foundational step toward event-based video anomaly detection by establishing both a strong baseline EWAD and the first set of large-scale RGB–event benchmarks. EWAD integrates event-density aware dynamic sampling, density-modulated temporal modeling, and cross-modal knowledge transfer into a unified framework, showing that carefully designed mechanisms can effectively bridge the gap between asynchronous event signals and high-level anomaly understanding. Beyond temporal-level detection, our exploration of spatial localization further highlights the potential of event data for fine-grained abnormality analysis. Comprehensive experiments validate the strength and generality of our approach across multiple datasets. Future work will explore constructing real-world event-centric VAD datasets, exploring multimodal fusion with RGB and audio, and developing foundation models tailored to event streams. The paper ends with a conclusion.

References

1. Acsintoae, A., Florescu, A., Georgescu, M.I., Mare, T., Sumedrea, P., Ionescu, R.T., Khan, F.S., Shah, M.: Ubnormal: new benchmark for supervised open-set video anomaly detection. In: CVPR. pp. 20143–20153 (2022) 2, 5
2. Barbosa, R.Z., Oliveira, H.S., Tavares, J.M.R.: A survey on multi-modal and weakly supervised approaches for robust anomaly detection in video data. Information Fusion p. 103388 (2025) 2
3. Chen, N., Xiao, C., Dai, Y., He, S., Li, M., An, W.: Event-based tiny object detection: A benchmark dataset and baseline. arXiv preprint arXiv:2506.23575 (2025) 4
4. Courtois, J., Miramond, B., Pegatoquet, A.: Spiking monocular event based 6d pose estimation for space application. arXiv preprint arXiv:2501.02916 (2025) 5
5. Ding, Z., Zhang, H., Wu, P., Pang, G., Yang, Z., Wang, P., Zhang, Y.: Slowfastvad: video anomaly detection via integrating simple detector and rag-enhanced vision-language model. arXiv preprint arXiv:2504.10320 (2025) 4
6. Fang, W., Yu, Z., Chen, Y., Huang, T., Masquelier, T., Tian, Y.: Deep residual learning in spiking neural networks. Advances in Neural Information Processing Systems **34**, 21056–21069 (2021) 12
7. Fang, W., Yu, Z., Chen, Y., Masquelier, T., Huang, T., Tian, Y.: Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2661–2671 (2021) 12
8. Gehrig, M., Muglikar, M., Scaramuzza, D.: Dense continuous-time optical flow from event cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(7), 4736–4746 (2024) 4
9. He, P., Zhang, F., Li, G., Li, H.: Adversarial and focused training of abnormal videos for weakly-supervised anomaly detection. Pattern Recognition **147**, 110119 (2024) 4
10. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1312–1321 (2021) 5
11. Huang, C., Liu, C., Wen, J., Wu, L., Xu, Y., Jiang, Q., Wang, Y.: Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. IEEE Transactions on Cybernetics **54**(5), 3197–3210 (2022) 4
12. Huang, C., Shi, Y., Wen, J., Wang, W., Xu, Y., Cao, X.: Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In: Forty-second International Conference on Machine Learning (2025) 4
13. Joo, H.K., Vo, K., Yamazaki, K., Le, N.: Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 3230–3234. IEEE (2024) 12
14. Li, J., Li, J., Zhu, L., Xiang, X., Huang, T., Tian, Y.: Asynchronous spatio-temporal memory network for continuous event-based object detection. IEEE Transactions on Image Processing **31**, 2975–2987 (2022) 5
15. Liu, K., Ma, H.: Exploring background-bias for anomaly detection in surveillance videos. In: ACMMM. pp. 1490–1499 (2019) 11, 12, 13
16. Liu, S., Li, J., Zhao, G., Zhang, Y., Jiang, W., Li, M., Ji, X.: Eventflash: Towards efficient mllms for event-based vision. arXiv preprint arXiv:2602.03230 (2026) 5

17. Liu, S., Li, J., Zhao, G., Zhang, Y., Meng, X., Yu, F.R., Ji, X., Li, M.: Eventgpt: Event stream understanding with multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29139–29149 (2025) 2
18. Liu, S., Dragotti, P.L.: Sensing diversity and sparsity models for event generation and video reconstruction from events. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12444–12458 (2023) 5
19. Liu, Y., Yang, D., Wang, Y., Liu, J., Liu, J., Boukerche, A., Sun, P., Song, L.: Generalized video anomaly event detection: systematic taxonomy and comparison of deep models. ACM Computing Surveys **56**(7) (2023) 2
20. Perez, M., Kot, A.C., Rocha, A.: Detection of real-world fights in surveillance videos. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2662–2666. IEEE (2019) 2, 5
21. Qian, Y., Ye, S., Wang, C., Cai, X., Qian, J., Wu, J.: Ucf-crime-dvs: A novel event-based dataset for video anomaly detection with spiking neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6577–6585 (2025) 2, 4, 11, 12
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 4
23. Su, Y., Tan, Y., An, S., Xing, M., Feng, Z.: Semantic-driven dual consistency learning for weakly supervised video anomaly detection. Pattern Recognition **157**, 110898 (2025) 4
24. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018) 2, 4, 5, 6, 11, 12
25. Sun, S., Ren, W., Li, J., Wang, R., Cao, X.: Logit standardization in knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15731–15740 (2024) 10, 11
26. Sun, Y., Dong, W., Wang, S., Wu, P., Feng, M., Li, X., Shi, G.: Distilling hierarchical knowledge from multimodal fusion for unimodal image segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2025) 9
27. Tang, J., Lu, H., Wu, R., Xu, X., Ma, K., Fang, C., Guo, B., Lu, J., Chen, Q., Chen, Y.: Hawk: learning to understand open-world video anomalies. In: NeurIPS. vol. 37, pp. 139751–139785 (2024) 4
28. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4975–4986 (2021) 12
29. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015) 13
30. Wang, B., He, J., Yu, L., Xia, G.S., Yang, W.: Event enhanced high-quality image recovery. In: European Conference on Computer Vision. pp. 155–171. Springer (2020) 4
31. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016) 13

32. Wang, X., Wang, S., Tang, C., Zhu, L., Jiang, B., Tian, Y., Tang, J.: Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19248–19257 (2024) 14
33. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018) 13
34. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: European conference on computer vision. pp. 322–339. Springer (2020) 12
35. Wu, P., Su, W., He, X., Wang, P., Zhang, Y.: Varcmp: Adapting cross-modal pre-training models for video anomaly retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8423–8431 (2025) 4
36. Wu, P., Zhou, X., Pang, G., Sun, Y., Liu, J., Wang, P., Zhang, Y.: Open-vocabulary video anomaly detection. In: CVPR. pp. 18297–18307 (2024) 4
37. Wu, P., Zhou, X., Pang, G., Yang, Z., Yan, Q., Wang, P., Zhang, Y.: Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In: ACMMM (2024) 4, 13
38. Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., Zhang, Y.: Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6074–6082 (2024) 4, 9, 10, 11, 12, 13
39. Yang, Z., Liu, J., Wu, P.: Text prompt with normality guidance for weakly supervised video anomaly detection. In: CVPR. pp. 18899–18908 (2024) 4
40. Ye, M., Liu, W., He, P.: Vera: explainable video anomaly detection via verbalized learning of vision-language models. In: CVPR (2025) 4
41. Zanella, L., Menapace, W., Mancini, M., Wang, Y., Ricci, E.: Harnessing large language models for training-free video anomaly detection. In: CVPR. pp. 18527–18536 (2024) 4
42. Zhang, H., Xu, X., Wang, X., Zuo, J., Huang, X., Gao, C., Zhang, S., Yu, L., Sang, N.: Holmes-vau: Towards long-term video anomaly understanding at any granularity. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13843–13853 (2025) 4
43. Zhou, C., Yu, L., Zhou, Z., Ma, Z., Zhang, H., Zhou, H., Tian, Y.: Spikingformer: Spike-driven residual learning for transformer-based spiking neural network. arXiv preprint arXiv:2304.11954 (2023) 12
44. Zhou, H., Yu, J., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3769–3777 (2023) 11, 12
45. Zhou, S., Buonomano, D.V.: Unified control of temporal and spatial scales of sensorimotor behavior through neuromodulation of short-term synaptic plasticity. Science Advances **10**(18), eadk7257 (2024) 9