ASYMMETRICVAD: AN ASYMMETRIC TRI-MODALITY REASONING PARADIGM FOR WEAKLY SUPERVISED VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised video anomaly detection (WSVAD) is a critical task in intelligent surveillance. It aims to detect and localize abnormal events using only video-level labels. However, traditional multimodal WSVAD methods predominantly rely on symmetric fusion paradigms that treat video, audio, and text modalities with equal importance. Such paradigms often overlook representation biases, making the joint features highly susceptible to volatile background audio noise and temporal alignment offsets. To address these challenges, we analyze the role of different modalities and propose AsymmetricVAD, an Asymmetric Tri-modality Reasoning Paradigm with a conditional loop. Specifically, AsymmetricVAD utilizes a Selective Injection and Enhancement (SIE) module as an asymmetric gatekeeper. It selectively injects audio into the video/text feature space only under temporal fluctuation and high cross-modal similarity. Furthermore, a Look-Back mechanism establishes the loop to suppress the representation bias. To justify our method, we build a new dataset TMVAD, which is the first tri-modality dataset. Extensive experiments on the benchmark XD-Violence and TMVAD datasets demonstrate that AsymmetricVAD achieves competitive performance on coarse- and fine-grained WSVAD tasks. We will open-source our code and dataset at https://anonymous.4open.science/r/AsymmetricVAD.

1 INTRODUCTION

Video anomaly detection (VAD) plays a crucial role in intelligent security systems. It aims to automatically detect anomaly events from a large amount of video data. To reduce manual annotation costs, most existing methodsSun et al. (2025); Dev et al. (2024); Zhang et al. (2025) design Weakly Supervised Video Anomaly Detection (WSVAD) models that utilize video-level labels for training. Meanwhile, researchers have explored multimodal data to improve detection accuracy. In addition to video, they utilize audio or text generated by vision-language models (VLMs) since they provide complementary informationHuang et al. (2025); Zanella et al. (2024) during training.

However, dual-modality methods (video-text or video-audio) still lack comprehensive scene understanding. For example, video-text modelsYang et al. (2024); Sun et al. (2024); Yin et al. (2026) may fail to detect sudden, off-screen acoustic anomalies such as shooting, whereas video-audio modelsMeng et al. (2025); Jin et al. (2025) often struggle with high-level semantic categorization without textual guidance. In contrast, tri-modality frameworksSun et al. (2026); Wang et al. (2025b) extract three types of modalities for feature fusion to explore deeper complementary data utilization. They typically utilize frozen encoders and treat the three modalities with equal importance as illustrated in Fig. 1(a, top), which face two major challenges in real surveillance scenarios. First, assuming that three modalities are equally important is often impractical. For example, on a busy street, sounds such as loud laughter or vehicle horns can cause abrupt fluctuations in audio. Treating audio with equal weight can lead to false alarms even when the scene is normal. Second, frozen encoders can also introduce representation biases in featuresLee et al. (2025); Zhang et al. (2024a), which can result in unstable anomaly predictions. These considerations lead us to explore two novel research questions: (1) Can we pinpoint the role of different modalities for training, and (2) can we reduce representation biases during training?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

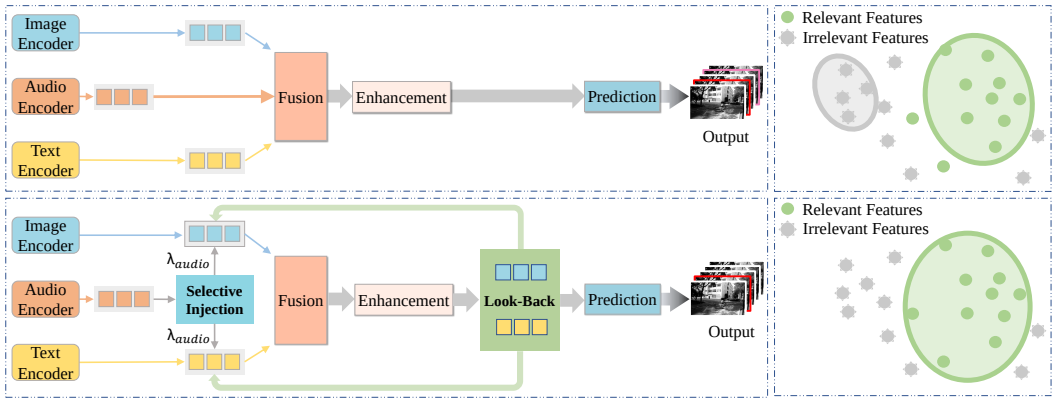

Figure 1: Comparison between the traditional tri-modality WSVAD method and our approach. (a) Unlike traditional methods that fuse modalities symmetrically (top), our framework treats audio as an auxiliary modality via Selective Injection and incorporates a Look-Back mechanism to dynamically recalibrate representations (bottom). (b) Illustration of feature representation biases. Enclosed regions indicate the model's attention. Top: Conventional methods generate a representation bias (grey ellipse) and miss critical features. Bottom: Our method mitigates the representation bias, guiding the model to focus on task-relevant features (green ellipse).

To address these questions, we justify the role of each modality and propose a novel asymmetric tri-modality reasoning paradigm with a conditional loop as illustrated in Fig. 1(a, bottom). First, we identify that audio plays an auxiliary role by statistically analyzing the noise and the corresponding impact of three modalities in XD-violence. Then we further reveal the gap between existing tri-modality paradigms and real-world scenarios. Considering the role of audio, we propose a Selective Injection and Enhancement (SIE) module that utilizes audio as an auxiliary trigger based on the temporal variance of audio and the similarity between video-audio and text-audio features.

Second, we introduce a "Look-Back" mechanism that mitigates the representation bias in complex scenes during multimodal feature alignment. Specifically, it reverts to the initial feature space to restart the reasoning process when the model detects a notable semantic mismatch during training. Thus, it dynamically recalibrates biased feature channels, which effectively prevents error propagation from fused features, as illustrated in Fig. 1(b). Note that our "Look-Back" mechanism does not require a large number of parameters since the processes share the same parameters.

We evaluate our method on XD-ViolenceWu et al. (2020), which is the only open-source VAD dataset that includes audioWang et al. (2025b). To further verify the effectiveness of our method, we build a new dataset called the Tri-Modality Video Anomaly Dataset (TMVAD) and test our method on it. TMVAD features 2,000 untrimmed videos across 22 fine-grained categories. Compared with existing datasets, it is the first to provide synchronized audio and segment-level text descriptions. Experiments demonstrate that our model outperforms existing methods on coarse-grained detection and fine-grained classification.

In summary, our contributions are as follows:

- We justify the role of each modality and propose an asymmetric tri-modality reasoning paradigm. To the best of our knowledge, we are the first to pinpoint the auxiliary role of audio for VAD.
- We design a SIE module for audio feature selection based on the evaluation of audio variance and cross-modal feature alignment.
- We introduce a Look-Back mechanism to optimize the training. It reduces the representation bias during training.
- We construct TMVAD, the first open-source tri-modality VAD dataset. We believe it can facilitate research on VAD.

2 RELATED WORK

2.1 SINGLE-MODALITY WSVAD

Early WSVAD research primarily relied on single-modality video frames Chen et al. (2022); Majhi et al. (2023). Typically formulated via Multiple Instance Learning (MIL) Park et al. (2023), these methods regard videos as bags and clips as instances to identify anomalies under weak labels. To capture temporal dependencies, subsequent works integrated attention mechanisms Fan et al. (2023); Ghadiya et al. (2024); fei Chen et al. (2025) and temporal convolutions Zhang et al. (2024b); Majhi et al. (2024). However, relying on visual data limits performance in ambiguous scenarios, where the visual modality lacks sufficient discriminative power, necessitating supplementary modalities.

2.2 DUAL-MODALITY WSVAD

To address these limitations, recent studies leverage dual-modality data, spanning video-audio and video-text paradigms. Video-audio approaches Ali et al. (2025); Feng et al. (2023) integrated acoustic signals to provide temporal features when visual information is sparse, while video-text methods Wang et al. (2025a); Dai et al. (2026) utilized pre-trained VLMs like CLIP Radford et al. (2021) to enhance semantic category understanding. Despite their progress, both paradigms have limitations: video-audio methods often lack high-level semantic definitions Wu et al. (2023a), whereas video-text methods may miss sudden, unsighted temporal anomalies Li et al. (2025). Consequently, dual-modality configurations remain insufficient for complex, real-world scenes.

2.3 TRI-MODALITY WSVAD

To capture comprehensive anomaly cues, recent research has shifted toward tri-modality WSVAD, integrating video, audio, and text Wang & Chen (2025); Zhou et al. (2026b); Xing et al. (2025). For example, CMHKFWang et al. (2025b) extracted unimodal features via separate visual, textual, and acoustic backbones (e.g., CLIP and VGGishHershey et al. (2016)), and subsequently executed parallel fusion by routing all three modalities through multi-branch cross-attention layers before concatenating them into a unified multi-modal representation. They fundamentally assume that video, audio, and text are equally important. However, this symmetric strategy is highly susceptible to acoustic noise Ding et al. (2025); symmetric fusion propagates background interference into the joint latent space, causing spurious energy fluctuations that trigger false alarms. Furthermore, although many existing methods attempt to mitigate representation biases through attention mechanismsPatrikar et al. (2025); Biswal & Baliarsingh (2026) or alignment constraintsYin et al. (2026); Liu et al. (2025), they predominantly employ open-loop, one-way inference Xu et al. (2025); Zhao et al. (2025); Chu et al. (2026). Thus, they can not detect or correct representation biases during training, which may propagate errors to downstream tasks.

3 METHOD

3.1 PROBLEM DEFINITION

Let V_t and A_t denote the input video and corresponding audio at temporal segment $t \in \{1, 2, \ldots, L\}$, where L is the total number of segments. We first use a pre-trained VLM to generate a descriptive text prompt T_t that captures high-level semantic information of the video segment V_t. We then extract segment-level features from all three modalities using the frozen encoders of ImageBind Girdhar et al. (2023): the visual encoder ϕ_v^{IB}, the text encoder ϕ_t^{IB}, and the audio encoder ϕ_a^{IB}. For each segment t, we obtain:

$$F_v(t) = \phi_v^{IB}(V_t), \quad F_t(t) = \phi_t^{IB}(T_t), \quad F_a(t) = \phi_a^{IB}(A_t). \tag{1}$$

During training, each video is associated with a binary label $y \in \{0, 1\}$, where $y = 1$ indicates the presence of at least one anomaly segment, and $y = 0$ represents a completely normal video. To enhance semantic interpretability and guide feature learning, we further introduce a set of anomaly category label features $\{e_c\}_{c=1}^{C}$ following the work Wu et al. (2023b), where C is the number of anomaly types. Each $e_c \in \mathbb{R}^d$ is obtained by encoding the corresponding category text label via ϕ_t^{IB}, and is used to align segment-level features with semantic anomaly concepts during training. We aim to automatically detect and temporally localize anomaly events in untrimmed videos utilizing only video-level supervision.

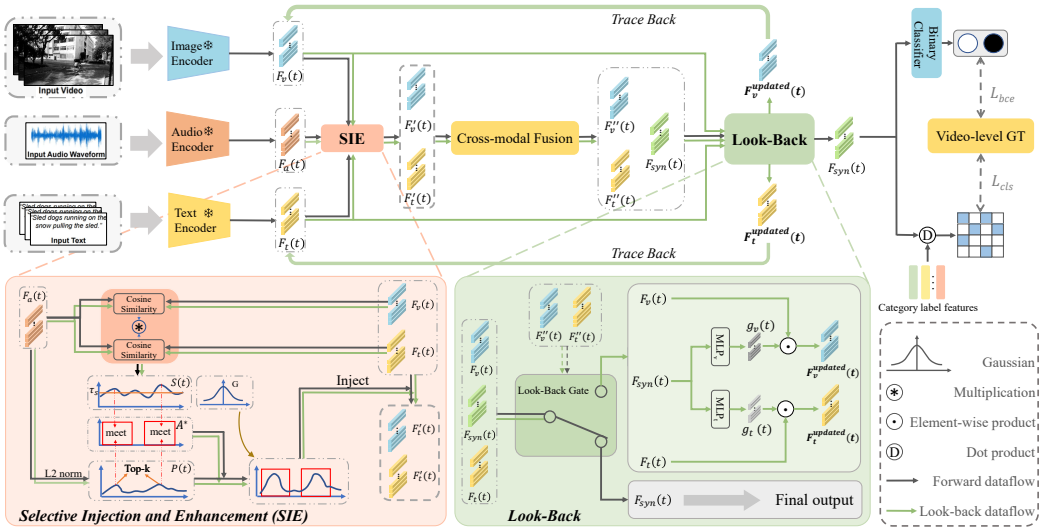

Figure 2: Overall framework of AsymmetricVAD. It consists of two components: (a) SIE selects Top-K salient and semantically aligned audio features to enhance video and text features; and (b) Look-Back detects representation biases by measuring feature alignment discrepancy and then triggers a loop to recalibrate the initial features. The final representations are used for joint video-level anomaly detection and multi-label category classification.

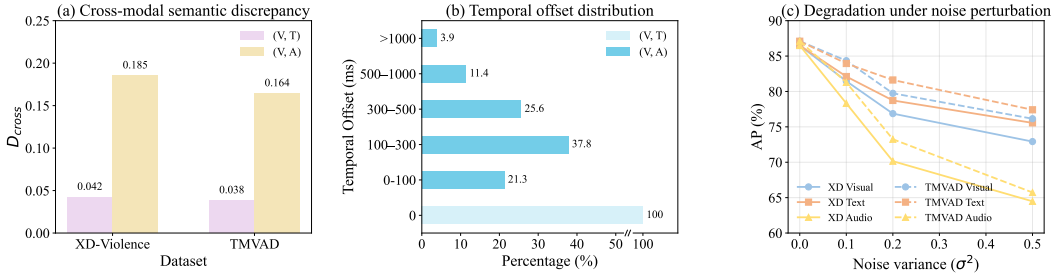

Figure 3: Statistical analysis of modality characteristics. (a) Cross-modal semantic discrepancy D_{cross} of video-text (V, T) and video-audio (V, A) pairs. (b) Distribution of temporal offsets between corresponding video-text (V, T) and video-audio (V, A) events. (c) Downstream anomaly detection performance (AP) under varying levels of visual, textual, and acoustic noise perturbations.

3.2 Modality Role Analysis

To understand the role of each modality, we first conduct a detailed modal analysis on benchmark datasets XD-Violence and TMVAD to examine their features. Specifically, we analyze the semantic and temporal alignment between cross-modal representations. To access the impact of feature misalignment, we further introduce controlled perturbations and evaluate the performance degradation.

We first define a cross-modal metric of stability to measure the semantic feature alignment between modality pairs:

$$D_{\text{cross}}(m_1, m_2) = \frac{1}{L}\sum_{t=1}^{L}\left(1 - \frac{F_{m_1}(t)^\top F_{m_2}(t)}{\|F_{m_1}(t)\|_2 \|F_{m_2}(t)\|_2}\right), \tag{2}$$

where $(m_1, m_2) \in \{(V, T), (V, A)\}$ denote the video-text or video-audio pairs, and $F_{m_1}(t), F_{m_2}(t)$ represent their feature at segment t. As illustrated in Fig. 3(a), we find that the video-audio discrepancy $D_{\text{cross}}(V, A)$ is significantly larger than the video-text discrepancy $D_{\text{cross}}(V, T)$. Furthermore, Fig. 3(b) shows the distribution of temporal misalignment between corresponding video-text and video-audio events from XD-Violence. We find that temporal misalignment is prevalent, with audio events frequently preceding or lagging behind visual counterparts. In contrast, the textual descriptions are generated by a VLM based on video segments, they are intrinsically synchronized with the video timeline and thus possess zero temporal offset. Lastly, Fig. 3(c) demonstrates the impact of

feature misalignment on downstream WSVAD performance (Average Precision, AP). We observe that anomaly detection performance suffers severe degradation under audio perturbations, while the drops under visual and textual perturbations are highly similar and relatively minor. Note that we also justify the impact of temporal alignment offsets in Appendix D.5 due to the page limit. These results reveal a critical insight: video and text have similar impact on training, however, representation biases introduced by audio are highly volatile and detrimental to downstream tasks.

3.3 Selective Injection and Enhancement

To reduce the negative impact of audio, the SIE module identifies audio that can be used for feature enhancement. The audio must satisfy two principles. First, the audio should be salient, which deviates from background sound. Second, the selected audio features must be aligned well with video and text.

To estimate temporal fluctuations of the audio, we first quantify the prominence of the t-th audio segment feature $F_a(t)$ by computing its ℓ_2 norm, yielding the audio energy $E_a(t) = \|F_a(t)\|_2$, which reflects the magnitude of this audio segmentTian et al. (2021). Although we can select the audio feature with the largest fluctuation based on its magnitude, this selection may introduce noise, since audio features can be outliers. Thus, we further define a temporal window w_t around each segment t to mitigate outlier impact. Inside this window, the average magnitude is calculated as:

$$\bar{E}_a(t) = \frac{1}{|w_t|} \sum_{\tau \in w_t} E_a(\tau), \tag{3}$$

where τ indexes segments within w_t.

Then, for each segment t, we compute a bounded score $P(t) \in [0, 1)$ via $\tanh(\cdot)$ based on the magnitude's deviation from the global mean μ_{bg} and standard deviation σ_{bg}:

$$P(t) = \tanh\left(\mathrm{ReLU}\left(\frac{\bar{E}_a(t) - \mu_{bg}}{\sigma_{bg} + \epsilon}\right)\right), \tag{4}$$

where ϵ is a small constant to prevent division by zero, and $\mathrm{ReLU}(\cdot)$ is the Rectified Linear Unit activation function. We describe the details of w_t, μ_{bg} and σ_{bg} in Appendix A.1. A high $P(t)$ score indicates that the audio feature significantly deviates from the background audio.

Then we employ a Top-K ratio which identifies the proportion of temporal steps with the largest fluctuations. This yields a set of salient indices J:

$$J = \mathrm{TopK}\left(\{P(t)\}_{t=1}^{L}\right). \tag{5}$$

After selecting the potential temporal indices J, we evaluate the alignment of video-audio and text-audio to further reduce the impact of audio feature outliers. We compute the joint cross-modal similarity S at each segment t as follows:

$$S(t) = \mathrm{ReLU}\left(\frac{F_v(t)^\top F_a(t)}{\|F_v(t)\|_2 \|F_a(t)\|_2}\right) \cdot \mathrm{ReLU}\left(\frac{F_t(t)^\top F_a(t)}{\|F_t(t)\|_2 \|F_a(t)\|_2}\right), \tag{6}$$

where $\cdot$ denotes multiplication. A high score indicates that the audio is consistent with other two modalities. We retain the temporal indices to form a verified subset $A^* \subset J$ that meets the cross-modal similarity threshold τ_s:

$$A^* = \{j \in J \mid S(j) > \tau_s\}. \tag{7}$$

Thus, audio segments are selected only when they are temporally salient and semantically aligned. Meanwhile, we enhance the adjacent audio magnitude to address the temporal lag between corresponding audio and video events. Specifically, for each $j \in A^*$, we use a Gaussian kernel to generate an enhanced magnitude:

$$E'_a(t) = E_a(t) + \epsilon_{\text{audio}} \sum_{j \in A^*} (\alpha S(j) + (1 - \alpha) P(j)) \exp\left(-\frac{(t - j)^2}{2\sigma^2}\right), \tag{8}$$

where α and ϵ_{audio} are learnable parameters that control the balance between cross-modal similarity and audio variance, and the enhancement strength, respectively. σ is a predefined parameter that controls the spread range. Finally, we inject the enhanced magnitude $E'_a(t)$ into the video and text features:

$$F'_v(t) = (1 + \lambda_{\text{audio}} E'_a(t)) F_v(t), F'_t(t) = (1 + \lambda_{\text{audio}} E'_a(t)) F_t(t), \tag{9}$$

where $\lambda_{\text{audio}} > 0$ is a learnable parameter constrained via $\mathrm{Softplus}(\cdot)$.

3.4 LOOK-BACK MECHANISM

After feature enhancement via the SIE module, we adopt the cross-modal fusion method Wang et al. (2025b) to integrate the enhanced video feature $F'_v(t)$ and text feature $F'_t(t)$, yielding the aligned video feature $F''_v(t)$, text feature $F''_t(t)$, and joint fused representation $F_{\text{syn}}(t)$.

The Look-Back mechanism runs a conditional loop reasoning when it detects representation bias in the fused features. It conditionally resets the text and visual features to correct the representation bias before downstream error propagation.

The representation bias degrades downstream anomaly detection tasks due to cross-modal alignment discrepancy. To decide whether to activate the Look-Back mechanism, we calculate the global cross-modal semantic gap U_{dis} by measuring the average cosine similarity between the aligned video and text features across all segments:

$$U_{\text{dis}} = \frac{1}{L}\sum_{t=1}^{L}\left(1 - \frac{F''_v(t)^\top F''_t(t)}{\|F''_v(t)\|_2 \|F''_t(t)\|_2}\right). \tag{10}$$

We set the gating indicator based on this cross-modal semantic gap:

$$I_{\text{lb}} = I(U_{\text{dis}} > \tau_{\text{unc}}), \tag{11}$$

where τ_{unc} is a predefined threshold for measuring cross-modal alignment discrepancy.

If $I_{\text{lb}} = 0$, the multimodal alignment has low semantic conflict, and the model directly utilizes $F_{\text{syn}}(t)$ as the final joint output. In contrast, if $I_{\text{lb}} = 1$, a severe conflict is flagged, triggering the model to trace back to the raw video feature $F_v(t)$ and text feature $F_t(t)$. Specifically, the model generates gating features for the video and text modalities based on the joint feature $F_{\text{syn}}(t)$:

$$g_v(t) = \sigma(\text{MLP}_v(F_{\text{syn}}(t))), \qquad g_t(t) = \sigma(\text{MLP}_t(F_{\text{syn}}(t))), \tag{12}$$

where $\sigma(\cdot)$ is the sigmoid function and MLP is the multi-layer perceptron network.

Next, the model recalibrates $F_v(t)$ and $F_t(t)$ using element-wise products with $g_v(t)$ and $g_t(t)$, respectively:

$$F_v^{updated}(t) = g_v(t) \odot F_v(t), \qquad F_t^{updated}(t) = g_t(t) \odot F_t(t), \tag{13}$$

where $F_v^{updated}(t)$ and $F_t^{updated}(t)$ are recalibrated video and text feature representations that suppress biased semantic channels.

Finally, $F_v^{updated}(t)$ and $F_t^{updated}(t)$ replace the original input features and are fed back into the SIE module for the next round of processing.

After Look-Back, the final joint representation $F_{\text{syn}}(t)$ is fed into downstream prediction heads. Specifically, a linear classification head projects $F_{\text{syn}}(t)$ to yield the segment-level anomaly confidence score $s_{\text{bin}}(t)$, from which the video-level score is aggregated and supervised by the binary cross-entropy loss L_{bce} against the video-level label $y \in \{0, 1\}$. Simultaneously, fine-grained anomaly category scores $s_{\text{cls}}(t) \in \mathbb{R}^C$ are obtained through dot-product similarities between $F_{\text{syn}}(t)$ and category text embeddings, supervised by the multi-label classification loss L_{cls}. Details of the loss functions are provided in Appendix A.2.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We construct **TMVAD**, a comprehensive tri-modality benchmark designed to cover both fixed and non-fixed camera perspectives. TMVAD comprises 2,000 untrimmed videos (1,987,695 frames in total) spanning 22 distinct anomaly categories, including 1,000 anomaly videos. The normal and abnormal videos are jointly split into training and testing sets at a 4:1 ratio. TMVAD preserves original audio (supplemented with synchronized synthetic audio when absent), incorporates segment-level (16-frame) textual descriptions generated via the VLM Moondream2 Korrapati et al. (2024), and provides fine-grained, frame-level temporal annotations for the test set via multi-annotator cross-verification. More details of TMVAD are described in Appendix B. To demonstrate generalizability and enable fair comparison with existing methods, we also evaluate our model on

Table 1: Coarse-grained comparison on XD-Violence and TMVAD. Unsupervised (Un) and weakly-supervised (Weak) methods are grouped. Best results are **bolded**; second best results are underlined.

	Method	Venue	Modality	XD-Violence (AP %)	TMVAD (AP %)
Un	GODSWang & Cherian (2019)	ICCV	V	61.56	60.71
	FRDTao et al. (2024)	TMM	V	-	74.53
	CLAPAl-Lahham et al. (2024)	CVPR	V	77.65	76.27
Weak	RADSultani et al. (2018)	CVPR	V	73.20	74.91
	RTFMTian et al. (2021)	ICCV	V	77.81	77.25
	RTFMTian et al. (2021)	ICCV	V+A	78.10	78.13
	ST-MSLLi et al. (2022)	AAAI	V	78.28	78.46
	LA-NetPu & Wu (2022)	ICME	V	80.72	-
	CLIP-TSAJoo et al. (2022)	ICIP	V+T	82.19	81.90
	DMUZhou et al. (2023)	AAAI	V	81.66	80.67
	DMUZhou et al. (2023)	AAAI	V+A	81.77	80.92
	VadCLIPWu et al. (2023b)	AAAI	V+T	84.51	84.62
	PEL4VADPu et al. (2023)	TIP	V+T	85.59	85.94
	TPWNGYang et al. (2024)	CVPR	V+T	83.68	83.19
	ITCLiu et al. (2024)	TIP	V+T	85.45	-
	IFS-VADZhong et al. (2025)	TCSVT	V	83.14	82.33
	LEC-VADWang & Chen (2025)	ICML	V+T	86.56	-
	CMHKFWang et al. (2025b)	ACL	V+A+T	86.57	86.95
	LPGBJiang (2025)	ICLR	V+T	<u>86.96</u>	-
	DSANetYin et al. (2026)	AAAI	V+T	86.95	<u>87.68</u>
Weak	**AsymmetricVAD (Ours)**	**ICLR**	**V+A+T**	**87.21**	**88.23**

the widely adopted **XD-Violence** Wu et al. (2020) benchmark, where textual descriptions aligned with each video are generated using Moondream2, following the same procedure as in TMVAD.

Evaluation Metrics. We evaluate our model on two tasks using AP and mean Average Precision (mAP). Specifically, for coarse-grained anomaly detection, we adopt frame-level AP, while for fine-grained anomaly detection, we report mAP under different Intersection over Union (IoU) thresholds.

Implementation details. We train the network using PyTorch on an NVIDIA RTX 4090 GPU. We utilize the AdamW optimizer for 15 epochs with an initial learning rate of 1×10^{-5} and a batch size of 32. The extracted feature dimension D for all modalities is 1024. The hyperparameters are set as follows: $\tau_s = 0.35$ on XD-Violence and 0.3 on TMVAD, $\sigma = 3$, $\tau_{\text{unc}} = 0.4$, and the Top-K ratio $= 17\%$. More details of hyperparameters are provided in Appendix D.6.

4.2 Comparison with State-of-the-Art Methods

We evaluate our framework against state-of-the-art WSVAD methods in terms of coarse-grained anomaly detection and fine-grained anomaly classification and localization.

Coarse-grained WSVAD Results. Table 1 shows the coarse-grained results on both XD-Violence and TMVAD. On XD-Violence, our method achieves improvements of 0.64% and 0.25% over the tri-modality method CMHKF and the second best method LPGB, respectively. Meanwhile, on TM-VAD, we find that AsymmetricVAD obtains AP improvements of 1.28% over CMHKF and 0.55% over DSANet. These gains across benchmarks suggest that audio contributes to the VAD detection, and asymmetric utilization of audio helps mitigate the negative influence of audio noises.

Fine-grained WSVAD Results. Fine-grained WSVAD requires the model to categorize specific anomaly types and localize temporal boundaries under various IoU thresholds. Table 2 displays the fine-grained results on XD-Violence and TMVAD. On XD-Violence, our method achieves an average mAP of 29.30%, which achieves improvements of 0.43% and 2.60% compared to DSANet and CMHKF, respectively. Notably, AsymmetricVAD demonstrates consistent gains across all IoU thresholds. It still achieves 18.47% at a threshold of IoU = 0.5. In addition, our method improves the average mAP by 0.67% over DSANet and 1.22% over CMHKF on TMVAD. It also yields an mAP of 17.85% when IoU = 0.5. These fine-grained results indicate that our framework addresses the semantic shift introduced by representation biases, which contributes to generating more accurate temporal localization of subtle anomaly events.

Table 2: Fine-grained performance comparison on XD-Violence and TMVAD.

Method	Venue	XD-Violence (mAP@IoU %)						TMVAD (mAP@IoU %)					
		0.1	0.2	0.3	0.4	0.5	AVG	0.1	0.2	0.3	0.4	0.5	AVG
RADSultani et al. (2018)	CVPR	20.08	13.72	8.44	5.06	2.81	10.02	22.15	15.38	9.96	6.18	3.47	11.43
3C-NetNarayan et al. (2019)	ICCV	23.77	17.78	11.90	8.28	5.87	13.52	25.63	19.24	13.41	9.36	6.58	14.84
W-TALCPaul et al. (2018)	ECCV	26.27	18.87	13.83	9.50	6.55	15.00	28.11	20.94	15.47	11.08	7.69	16.66
VadCLIPWu et al. (2023b)	AAAI	37.03	30.84	23.38	17.90	14.31	24.69	38.86	32.47	25.41	20.08	15.96	26.56
CMHKFWang et al. (2025b)	ACL	38.31	32.12	25.68	20.93	16.47	26.70	38.92	33.19	26.66	20.57	16.31	27.13
DSANetYin et al. (2026)	AAAI	40.93	34.63	28.21	22.70	17.89	28.87	39.12	33.51	27.01	21.64	17.13	27.68
AsymmetricVAD (Ours)	ICLR	**41.31**	**35.12**	**28.68**	**22.93**	**18.47**	**29.30**	**40.27**	**34.15**	**27.36**	**22.11**	**17.85**	**28.35**

Table 3: Performance comparison of different modality combinations and paradigms.

Config	Modalities	Para	XD-Violence (AP %)	TMVAD (AP %)
1	V+T	Sym	84.68	84.81
2	V+A	Sym	83.24	83.39
3	V+A+T	Sym	85.63	86.68
4 (Ours)	V+A+T	Asym	**87.21**	**88.23**

Table 4: Ablation studies on the effectiveness of SIE and Look-Back (LB).

Index	SIE	LB	XD-Violence (AP %)	TMVAD (AP %)
1 (Full)	✓	✓	**87.21**	**88.23**
2		✓	85.63	86.68
3	✓		85.79	86.62
4			84.31	85.28

4.3 Ablation Study

To evaluate the contributions of each component, we conduct ablation studies across both datasets on the asymmetric tri-modality design, the SIE module, and the Look-Back mechanism. Specifically, Table 3 presents different modality combinations alongside symmetric and asymmetric fusion paradigms. Furthermore, Table 4 evaluates four configurations to investigate the individual and joint effects of the SIE module and Look-Back mechanism.

Effectiveness of the Asymmetric Tri-modality Design. To evaluate our asymmetric tri-modality paradigm, we compare four configurations derived from our architecture: *Config 1* utilizes only video and text features calibrated by the Look-Back without SIE; *Config 2* similarly bypasses SIE to align video and audio; *Config 3* symmetrically integrates all three modalities equipped with Look-Back; and *Config 4* represents our complete asymmetric framework integrating SIE and Look-Back. Table 3 shows the results.

The comparison indicates the effectiveness of the asymmetric tri-modality design. First, the bi-modality baselines *(Config 1 & 2)* yield sub-optimal performance, where the V+T setup outperforms V+A by 1.44% AP on XD-Violence and 1.42% AP on TMVAD. This disparity is primarily due to persistent, unaligned environmental audio noise that degrades the joint representation under symmetric fusion. Second, the symmetric tri-modality baseline (*Config 3*) improves performance over the V+A baseline by 2.39% AP on XD-Violence and 3.29% on TMVAD. Although the text modality provides semantic context to mitigate visual ambiguity, the joint feature space still suffers from unchecked audio noise propagation. In contrast, our asymmetric tri-modality paradigm (*Config 4*) further improves the AP by 1.58% on XD-Violence and 1.55% on TMVAD over the symmetric baseline (*Config 3*). This demonstrates that simply concatenating all three modalities in symmetric fusion introduces audio noise into the joint representation, whereas asymmetric usage of audio preserves the robustness of the video-text representation.

Effectiveness of SIE and Look-Back

As shown in Table 4, *Index 1* denotes our full framework, *Indices 2* and *3* isolate each component by removing SIE and Look-Back respectively, and *Index 4* corresponds to the baseline tri-modality model without SIE and Look-Back. We observe that removing the SIE module leads to an AP drop of 1.58% on XD-Violence and 1.55% on TMVAD, respectively. This is because the SIE module filters out irrelevant audio noise while preserving key temporal information for sudden anomalies like explosions. Meanwhile, removing the Look-Back mechanism causes the AP to decrease by 1.42% on XD-Violence and 1.61% on TMVAD, respectively. This is because it dynamically identifies and suppresses the biased feature channels within the asymmetric representation, which prevents the accumulation of representation biases.

4.4 Qualitative Results

Anomaly Score Visualization. Fig. 4 visualizes the anomaly scores and compares them with CMHKF. In the cluttered *Fighting* scenario (Fig. 4a), CMHKF's anomaly score fluctuates sharply

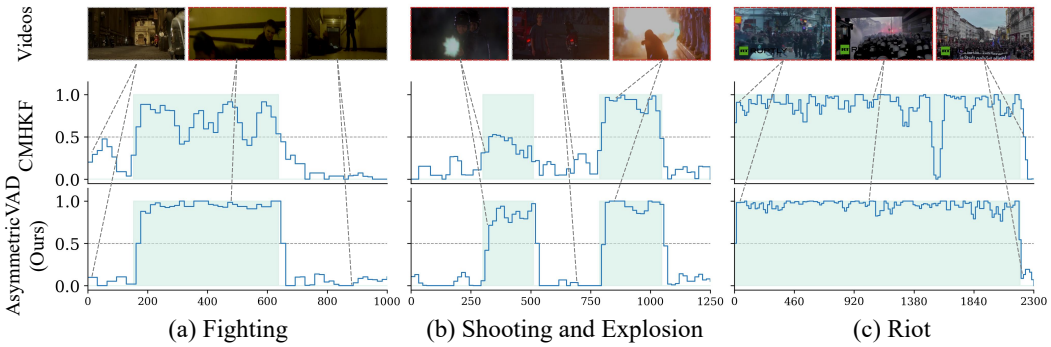

Figure 4: Visualization of anomaly scores on XD-Violence. The visualization displays key video frames and anomaly scores for CMHKF and AsymmetricVAD across three different events: (a) *Fighting*, (b) *Shooting and Explosion*, and (c) *Riot*. Shaded areas denote the Ground-truth.



Figure 5: t-SNE visualizations of feature distributions across five progressive stages of our framework on the XD-Violence dataset: (a) Original features, (b) After SIE, (c) After the first Look-Back loop, (d) After the second Look-Back loop, and (e) After the third Look-Back loop.

between 0.4 and 0.9, whereas our model maintains a confident prediction near 0.9. This robustness is enabled by our Look-Back mechanism, which recalibrates noisy representations and resolves semantic conflicts. For the brief, acoustic-heavy *Shooting and Explosion* scenario (Fig. 4b), CMHKF produces a weak initial peak (0.5), while our method generates two sharp peaks near 1.0. This sensitivity is driven by our SIE module, which utilizes the audio to dynamically enhance video and text features. In the sustained *Riot* scenario (Fig. 4c), CMHKF suffers from pronounced dips, whereas our model maintains a stable score near 1.0 since our asymmetric tri-modality framework filters out transient background noise and refines the joint representation. More results are provided in Appendix D.3.

Feature Distribution Visualization. To analyze the progressive representation refinement, Fig. 5 presents t-SNE visualizations of the latent space across sequential stages on XD-Violence. The original feature space (Fig. 5a) exhibits severe overlap between normal and anomalous samples. After applying the SIE module (Fig. 5b), a subtle separation emerges as certain audio-salient events shift, though ambiguous segments remain mixed in the center. This initial separation is further refined by the Look-Back mechanism. After the first loop (Fig. 5c), drifted boundary samples are partially recalibrated closer to their main distributions. After the second loop (Fig. 5d), the separation between classes becomes more pronounced. Finally, after the third loop (Fig. 5e), the latent space achieves a stable partition with cleaner boundaries and fewer scattered outliers. This progressive transition indicates that our framework steadily resolves the representation bias.

5 Conclusion

In this paper, we present AsymmetricVAD, an asymmetric tri-modality framework that defines the roles of video, text, and audio for WSVAD. Our paradigm achieves robust asymmetric fusion that suppresses audio noise interference using SIE, which pinpoints the most relevant audio features. Additionally, we introduce a Look-Back mechanism to dynamically recalibrate biased features via a conditional loop. Future work will focus on optimizing the efficiency of the Look-Back process and exploring the generalizability of this asymmetric paradigm in open-vocabulary and zero-shot anomaly detection.

AI USE STATEMENT

In this work, we have not used any generative AI tools or AI-assisted technologies, including research ideation, data collection, code implementation, and manuscript writing. All work was conducted entirely by ourselves, and we take full responsibility for the final content of this work.

REFERENCES

Anas Al-Lahham, Muhammad Zaigham Zaheer, Nurbek Tastan, and Karthik Nandakumar. Collaborative learning of anomalies with privacy (clap) for unsupervised video anomaly detection: A new baseline. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12416–12425, 2024. URL https://api.semanticscholar.org/CorpusID:268819481.

Amjid Ali, Zulfiqar Ahmad Khan, Altaf Hussain, Muhammad Munsif, Adnan Hussain, and Sung Wook Baik. Avar-net: A lightweight audio-visual anomaly recognition framework with a benchmark dataset. *ArXiv*, abs/2510.13630, 2025. URL https://api.semanticscholar.org/CorpusID:282102192.

Manas Ranjan Biswal and Santos Kumar Baliarsingh. Integrating vision and language: An improved vad model. *Journal of Applied Science and Technology Trends*, 2026. URL https://api.semanticscholar.org/CorpusID:286321839.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton W. T. Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:254044120.

Chengxi Chu, Nurul Japar, and Chee Kau Lim. Reba: Residual mixture-of-experts and bidirectional video-text alignment for better fine-grained weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8280–8290, 2026.

Zunkai Dai, Kesen Li, Jiajia Liu, Jie Yang, and Yu Qiao. No need for real anomaly: Mllm empowered zero-shot video anomaly detection. *ArXiv*, abs/2602.19248, 2026. URL https://api.semanticscholar.org/CorpusID:285973581.

Prabhu Prasad Dev, Raju Hazari, and Pranesh Das. Reflip-vad: Towards weakly supervised video anomaly detection via vision-language model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. URL https://api.semanticscholar.org/CorpusID:273509763.

Hantao Ding, Shengfeng Lou, Hairong Ye, and Yanbing Chen. Mt-cmvad: A multi-modal transformer framework for cross-modal video anomaly detection. *Applied Sciences*, 2025. URL https://api.semanticscholar.org/CorpusID:279532748.

Yidan Fan, Yongxin Yu, Wenhuan Lu, and Yahong Han. Weakly-supervised video anomaly detection with snippet anomalous attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 34:5480–5492, 2023. URL https://api.semanticscholar.org/CorpusID:263132101.

Peng fei Chen, Shengdong Du, Xiaole Zhao, Jie Hu, Jingjing Li, and Tianrui Li. Dctformer: A dual-branch transformer with cloze tests for video anomaly detection. *IEEE Transactions on Multimedia*, 27:9022–9032, 2025. URL https://api.semanticscholar.org/CorpusID:281561064.

Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10491–10503, 2023. URL https://api.semanticscholar.org/CorpusID:255416023.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng-Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26866–26875, 2024. URL https://api.semanticscholar.org/CorpusID:272722535.

Ayush Ghadiya, Purbayan Kar, Vishal M. Chudasama, and Pankaj Shivdayal Wasnik. Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1965–1974, 2024. URL https://api.semanticscholar.org/CorpusID:272915448.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023. URL https://api.semanticscholar.org/CorpusID:258564264.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. URL https://api.semanticscholar.org/CorpusID:218538102.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. Cnn architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, 2016. URL https://api.semanticscholar.org/CorpusID:8810481.

Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In *International Conference on Machine Learning*, 2025. URL https://api.semanticscholar.org/CorpusID:283567029.

Yalong Jiang. Local patterns generalize better for novel anomalies. In *International Conference on Learning Representations*, 2025. URL https://api.semanticscholar.org/CorpusID:272911563.

Wenping Jin, Li Zhu, and Jing Sun. Aligning first, then fusing: A novel weakly supervised multimodal violence detection method. *Knowl. Based Syst.*, 322:113709, 2025. URL https://api.semanticscholar.org/CorpusID:275471698.

Kevin Hyekang Joo, Khoa T. Vo, Kashu Yamazaki, and Ngan T. H. Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3230–3234, 2022. URL https://api.semanticscholar.org/CorpusID:254563882.

Sumin Kim, Hyemin Jeong, Mingu Kang, Yejin Kim, Yoori Oh, and Joonseok Lee. Triplesumm: Adaptive triple-modality fusion for video summarization. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://api.semanticscholar.org/CorpusID:286222910.

Vikhyat Korrapati et al. Moondream: a tiny vision language model that kicks ass and runs anywhere, 2024. URL https://github.com/m87-labs/moondream.

Junhee Lee, ChaeBeen Bang, MyoungChul Kim, and MyeongAh Cho. Refinevad: Semantic-guided feature recalibration for weakly supervised video anomaly detection. *ArXiv*, abs/2511.13204, 2025. URL https://api.semanticscholar.org/CorpusID:283072549.

Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29203–29212, 2025. URL https://api.semanticscholar.org/CorpusID:277271996.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. URL https://api.semanticscholar.org/CorpusID:256390509.

S. Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:248982052.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2023. URL https://api.semanticscholar.org/CorpusID:263672058.

Tianshan Liu, Kin Man Lam, and Bing-Kun Bao. Injecting text clues for improving anomalous event detection from weakly labeled videos. *IEEE Transactions on Image Processing*, 33:5907–5920, 2024. URL https://api.semanticscholar.org/CorpusID:273372380.

W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yang Liu, Hongjin Wang, Zepu Wang, Xiaoguang Zhu, Jing Liu, Peng Sun, Rui Tang, Jianwei Du, Victor C. M. Leung, and Liang Song. Crcl: Causal representation consistency learning for anomaly detection in surveillance videos. *IEEE Transactions on Image Processing*, 34:2351–2366, 2025. URL https://api.semanticscholar.org/CorpusID:277736219.

Yunze Liu, Chi-Hao Wu, Enmin Zhou, and Junxiao Shen. Omniretriever: Any-to-any audio-video-text retrieval via fusion-as-teacher distillation. *ArXiv*, abs/2605.26641, 2026. URL https://api.semanticscholar.org/CorpusID:288671858.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. *2013 IEEE International Conference on Computer Vision*, pp. 2720–2727, 2013. URL https://api.semanticscholar.org/CorpusID:6070091.

Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1975–1981. IEEE, 2010.

Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection. *ArXiv*, abs/2301.07923, 2023. URL https://api.semanticscholar.org/CorpusID:255999634.

Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Oe-ctst: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8559–8568, 2024. URL https://api.semanticscholar.org/CorpusID:268796415.

Jingke Meng, Huilin Tian, Ge Lin, Jianfang Hu, and Wei-Shi Zheng. Audio-visual collaborative learning for weakly supervised video anomaly detection. *IEEE Transactions on Multimedia*, 2025. URL https://api.semanticscholar.org/CorpusID:275936630.

Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8678–8686, 2019. URL https://api.semanticscholar.org/CorpusID:201313982.

Seong Hyeon Park, H. Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2664–2673, 2023. URL https://api.semanticscholar.org/CorpusID:256653748.

Devashree R. Patrikar, Mayur Rajaram Parate, and Nikhil Dhengre. Dual-stage attention mechanism for robust video anomaly detection and localization. *Signal, Image and Video Processing*, 19, 2025. URL https://api.semanticscholar.org/CorpusID:279331711.

S. Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *European Conference on Computer Vision*, 2018. URL https://api.semanticscholar.org/CorpusID:51866361.

R Gnana Praveen and Jahangir Alam. Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4803–4813, 2024. URL https://api.semanticscholar.org/CorpusID:268537197.

Yujiang Pu and Xiaoyu Wu. Locality-aware attention network with discriminative dynamics learning for weakly supervised anomaly detection. *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022. URL https://api.semanticscholar.org/CorpusID:251492909.

Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 33:4923–4936, 2023. URL https://api.semanticscholar.org/CorpusID:259252177.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ArXiv*, abs/2103.00020, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

Shengyang Sun, Jiashen Hua, Junyi Feng, Dongxu Wei, Baisheng Lai, and Xiaojin Gong. Tdsd: Text-driven scene-decoupled weakly supervised video anomaly detection. *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. URL https://api.semanticscholar.org/CorpusID:273645007.

Shengyang Sun, Jiashen Hua, Junyi Feng, Dongxu Wei, Baisheng Lai, and Xiaojin Gong. Delving into instance modeling for weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35:7655–7666, 2025. URL https://api.semanticscholar.org/CorpusID:276711257.

Shengyang Sun, Jiashen Hua, Junyi Feng, and Xiaojin Gong. Enhancing weakly supervised multimodal video anomaly detection through text guidance. *ArXiv*, abs/2602.10549, 2026. URL https://api.semanticscholar.org/CorpusID:285470238.

Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. WAVE: Learning unified & versatile audio-visual embeddings with multimodal LLM. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://api.semanticscholar.org/CorpusID:281658786.

Chenchen Tao, Chong Wang, Sunqi Lin, Suhang Cai, Di Li, and Jiangbo Qian. Feature reconstruction with disruption for unsupervised video anomaly detection. *IEEE Transactions on Multimedia*, 26:10160–10173, 2024. URL https://api.semanticscholar.org/CorpusID:270075441.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan Verjans, and G. Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4955–4966, 2021. URL https://api.semanticscholar.org/CorpusID:236950792.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019:6558 – 6569, 2019. URL https://api.semanticscholar.org/CorpusID:173990158.

Benfeng Wang, Chao Huang, Jie Wen, Wei Wang, Yabo Liu, and Yong Xu. Federated weakly supervised video anomaly detection with multimodal prompt. In *AAAI Conference on Artificial Intelligence*, 2025a. URL https://api.semanticscholar.org/CorpusID:277758600.

Guohua Wang, Shengping Song, Wuchun He, and Yongsen Zheng. Cmhkf: Cross-modality heterogeneous knowledge fusion for weakly supervised video anomaly detection. In *Annual Meeting of the Association for Computational Linguistics*, 2025b. URL https://api.semanticscholar.org/CorpusID:280018538.

Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8200–8210, 2019. URL https://api.semanticscholar.org/CorpusID:201058600.

Yu Wang and Shiwei Chen. Learning event completeness for weakly supervised video anomaly detection. *ArXiv*, abs/2506.13095, 2025. URL https://api.semanticscholar.org/CorpusID:279402810.

Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision (ECCV)*, 2020.

Peng Wu, Xiaotiao Liu, and Jing Liu. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 25:1674–1685, 2023a. URL https://api.semanticscholar.org/CorpusID:246541120.

Peng Wu, Xu Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, P. Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2023b. URL https://api.semanticscholar.org/CorpusID:261076368.

Peng Wu, Xuerong Zhou, Guan-Song Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. URL https://api.semanticscholar.org/CorpusID:271855764.

Zhenghao Xing, Hao Chen, Binzhu Xie, Jiaqi Xu, Ziyu Guo, Xuemiao Xu, Jianye Hao, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. Echotraffic: Enhancing traffic anomaly understanding with audio-visual insights. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19098–19108, 2025. URL https://api.semanticscholar.org/CorpusID:280096700.

Chen Xu, Ke Xu, Xinghao Jiang, and Tanfeng Sun. Plovad: Prompting vision-language models for open vocabulary video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35:5925–5938, 2025. URL https://api.semanticscholar.org/CorpusID:275453816.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18899–18908, 2024. URL https://api.semanticscholar.org/CorpusID:269137657.

Wenti Yin, Huaxin Zhang, Xiangdong Wang, Yuqing Lu, Yicheng Zhang, Bin Gong, Jialong Zuo, Li Yu, Changxin Gao, and Nong Sang. Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment. In *AAAI Conference on Artificial Intelligence*, 2026. URL https://doi.org/10.1609/aaai.v40i14.38191.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18527–18536, 2024. URL https://api.semanticscholar.org/CorpusID:268819369.

Huaxin Zhang, Xiaohao Xu, Xiangdong Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853, 2024a. URL https://api.semanticscholar.org/CorpusID:274597373.

Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17385–17394, 2024b. URL https://api.semanticscholar.org/CorpusID:271515656.

Sijia Zhang, Maoguo Gong, Yi-Ming Lin, Hao Li, Yuanqing Gao, and Yihong Zhang. Dual distillation fusion for weakly supervised anomaly detection in surveillance videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 35:7512–7525, 2025. URL https://api.semanticscholar.org/CorpusID:276877560.

Shili Zhao, Ran Zhao, Yan Meng, Xiaoyi Gu, Chen Shi, and Daoliang Li. Mstagent-vad: Multi-scale video anomaly detection using time agent mechanism for segments' temporal context mining. *Expert Syst. Appl.*, 276:127154, 2025. URL https://api.semanticscholar.org/CorpusID:276893154.

Yuanhong Zhong, Ruyue Zhu, Ge Yan, Ping Gan, Xuerui Shen, and Dong Zhu. Inter-clip feature similarity based weakly supervised video anomaly detection via multi-scale temporal mlp. *IEEE Transactions on Circuits and Systems for Video Technology*, 35:1961–1970, 2025. URL https://api.semanticscholar.org/CorpusID:273437604.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:256808591.

Mao-Wen Zhou, Erma Rahayu Mohd Faizal Abdullah, Aznul Qalid Md Sabri, and Susanto Rahardja. Learn the interactions: Weakly supervised video anomaly detection with human-object interactions. *PLOS One*, 21, 2026a. URL https://api.semanticscholar.org/CorpusID:292149878.

Maowen Zhou, Erma Rahayu Mohd Faizal Abdullah, Aznul Qalid Md. Sabri, and Susanto Rahardja. Tri-modal video anomaly detection via hierarchical contrastive calibration. *Array*, 30:100884, 2026b. URL https://api.semanticscholar.org/CorpusID:288350065.

APPENDIX CONTENTS

Appendix A. Algorithm Details **17**

A.1 Calculation of w_t, μ_{bg}, and σ_{bg} 17

A.2 Loss Functions 17

Appendix B. TMVAD Dataset **17**

B.1 Dataset Overview and Construction 18

B.2 Comparison with Existing VAD Datasets 18

Appendix C. Tri-Modality Methods in Other Domains **19**

Appendix D. More Experimental Results **19**

D.1 Generalization of the Asymmetric Fusion Paradigm 19

D.2 Comprehensive Evaluation of the Look-Back Mechanism 19

D.3 Additional Qualitative Visualizations 20

D.4 Ablation Study on the Effectiveness of Different VLMs 21

D.5 Robustness to Temporal Misalignment 21

D.6 Hyperparameter Sensitivity Analysis 21

D.7 Computational Efficiency Analysis 23

A Algorithm Details

A.1 Calculation of w_t, μ_{bg}, and σ_{bg}

The temporal window w_t centered at segment t with radius r is defined as:

$$w_t = \{\tau \in \{1, 2, \ldots, L\} \mid \max(1, t-r) \leq \tau \leq \min(L, t+r)\}, \tag{14}$$

where L is the total number of segments, τ denotes the segment index, and the window size is $|w_t| = \min(L, t+r) - \max(1, t-r) + 1$. In our implementation, we set $r = 2$. Given the averaged magnitude sequence $\{\bar{E}_a(t)\}_{t=1}^{L}$, the global mean μ_{bg} and standard deviation σ_{bg} are computed as:

$$\mu_{bg} = \frac{1}{L} \sum_{t=1}^{L} \bar{E}_a(t), \tag{15}$$

$$\sigma_{bg} = \sqrt{\frac{1}{L} \sum_{t=1}^{L} \left(\bar{E}_a(t) - \mu_{bg}\right)^2}. \tag{16}$$

A.2 Loss Functions

For video-level anomaly detection, the segment-level anomaly confidence score is computed via a linear projection head $f_{\text{bin}}(\cdot)$:

$$s_{\text{bin}}(t) = \sigma(f_{\text{bin}}(F_{\text{syn}}(t))), \tag{17}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The video-level anomaly score is obtained by taking the maximum over time, $S_{\text{bin}} = \max_t s_{\text{bin}}(t)$, and is supervised by the binary cross-entropy loss using the video-level label $y \in \{0, 1\}$:

$$L_{\text{bce}} = -y \log(S_{\text{bin}}) - (1-y) \log(1 - S_{\text{bin}}). \tag{18}$$

For fine-grained anomaly category prediction, the segment-level category score vector $s_{\text{cls}}(t) \in \mathbb{R}^C$ is obtained via the dot product between $F_{\text{syn}}(t)$ and the category text embeddings $\{e_c\}_{c=1}^{C}$:

$$s_{\text{cls}}(t) = [\langle F_{\text{syn}}(t), e_1 \rangle, \langle F_{\text{syn}}(t), e_2 \rangle, \ldots, \langle F_{\text{syn}}(t), e_C \rangle], \tag{19}$$

where $s_{\text{cls}}(t)[c]$ denotes the score for the c-th anomaly category at segment t.

The video-level category prediction for the c-th class is aggregated as $S_{\text{cls}}[c] = \max_t \sigma(s_{\text{cls}}(t)[c])$, which is optimized using the multi-label binary cross-entropy loss:

$$L_{\text{cls}} = -\sum_{c=1}^{C} \left[y_c \log(S_{\text{cls}}[c]) + (1-y_c) \log(1 - S_{\text{cls}}[c]) \right], \tag{20}$$

where $y_c \in \{0, 1\}$ indicates whether the c-th anomaly category is present in the video-level category label vector $\mathbf{y} \in \{0, 1\}^C$. The overall training objective is formulated as:

$$L_{\text{total}} = L_{\text{bce}} + \lambda L_{\text{cls}}, \tag{21}$$

where λ is a hyperparameter balancing the two loss terms.

Table 5: Category distribution of the TMVAD dataset.

Group	Classes	No. of Subclasses
Traffic incidents	car accidents, red-light violation	2
Violent & disruptive behavior	fighting, disorderly conduct, quarrelling, robbery, shooting, police assault	6
Public-safety hazards	fire, explosion, natural disaster, animal injury, high-altitude hazard, equipment malfunction	6
Personal emergencies & offenses	falls, drowning, self-harm, child trafficking, fence climbing, chasing, theft, indecent assault	8

B TMVAD Dataset

TMVAD has four notable characteristics: (1) it has the largest number of anomaly categories, (2) it is the first open-source dataset that provides video, audio, and text, (3) fine-grained temporal localization with frame-level boundaries for each anomaly event in the testing set, and (4) diverse real-world scenes with camera motion metadata.

Table 6: Comparison with representative VAD datasets. "Binary" denotes benchmarks organized primarily as normal-versus-anomaly detection rather than explicit anomaly-class recognition.

Dataset	Size	Anomaly Category	Modalities	Camera Position
UCSD Ped2 Mahadevan et al. (2010)	28 videos	Binary	V	fixed
Avenue Lu et al. (2013)	37 videos	Binary	V	fixed
ShanghaiTech Liu et al. (2018)	437 videos	Binary	V	fixed
UCF-Crime Sultani et al. (2018)	1,900 videos	13	V	fixed
XD-Violence Wu et al. (2020)	4,754 videos	6	V + A	non-fixed
TMVAD (ours)	2,000 videos	22	V + A + T	fixed / non-fixed

B.1 Dataset Overview and Construction

TMVAD is a multimodal benchmark collected from open web platforms and real-world scenarios for VAD. It contains 2,000 untrimmed videos, of which 1,000 videos have anomaly events. All videos are saved in MP4 format at 30 frames per second. The number of video frames is 1,987,695, corresponding to approximately 18.41 hours of video footage. Individual video clips have an average duration of 33.13 seconds, ranging from 5.77 seconds to 739.50 seconds. This variation in video length prevents models from assuming a fixed clip length. We partition TMVAD into a training set and a testing set with a 4:1 ratio. The partition is carefully structured to ensure that all 22 anomaly categories are represented in both the training and testing subsets.

Anomaly categories: To cover a wider spectrum of real-world anomaly events, TMVAD defines 22 distinct anomaly categories. These categories are organized into four groups, as summarized in Table 5: *Traffic incidents* (2 classes), *Violent & disruptive behavior* (6 classes), *Public-safety hazards* (6 classes), and *Personal emergencies & offenses* (8 classes). This covers severe non-violent occurrences (such as fainting, falls, and equipment malfunctions) alongside violent behaviors, enabling a more comprehensive and balanced assessment of anomaly detection algorithms.

Tri-modality collection: For the video, we collect part of them from online sources. We also use phones to record videos for more scenarios. For the audio, TMVAD preserves the original audio corresponding to each video. Note that some videos do not have original audio. Thus, we generate the audio and synchronize it with the video. For the text, we introduce textual descriptions for each video segment. Specifically, each video is partitioned into uniform segments of 16 frames. For each video segment, we utilize the VLM Moondream2 Korrapati et al. (2024) to generate textual descriptions of the scenes and actions.

Fine-Grained temporal localization: To facilitate precise evaluation of both anomaly detection and temporal localization, we establish fine-grained, frame-level temporal boundaries for the testing subset. Instead of relying on coarse video-level indicators, we manually annotate the exact start and end frame indices of each anomalous event. This dense temporal annotation is curated through a multi-annotator cross-verification pipeline.

Camera metadata: To evaluate the impact of camera motion on VAD models, each video is annotated with a camera position attribute indicating a fixed or non-fixed camera position. Among the 1600 training videos, there are 829 fixed and 771 non-fixed camera positions, while the 400 testing videos include 233 and 167, respectively.

B.2 Comparison with Existing VAD Datasets

Table 6 compares TMVAD with VAD benchmarks across four indicators of size, anomaly categories, types of modalities, and camera position. Although early datasets such as UCSD Ped2, Avenue, and ShanghaiTech established the foundations of VAD research, they focus on video modality and binary anomaly detection within fixed camera positions. Meanwhile, their size is small, which has a negative impact on the development of models. Thus, larger-scale benchmarks are built, such as UCF-Crime with 1,900 videos and XD-Violence with 4,754 videos. In contrast to these two datasets, TMVAD is the second-largest dataset and provides the most anomaly categories. Furthermore, while most existing benchmarks rely on the visual modality or dual-modality, TMVAD is the only benchmark that jointly provides visual streams, audio, and temporally aligned text descriptions. Meanwhile, unlike fixed scenes or web-based non-fixed settings, TMVAD spans fixed and non-fixed camera positions, which can support systematic evaluation of model robustness under complex scenarios.

Table 7: Comparison of representative tri-modality methods across other domains.

Method	Venue	Application Domain	Paradigm
MulT Tsai et al. (2019)	ACL	Sentiment Analysis	Symmetric
MISA Hazarika et al. (2020)	ACM MM	Sentiment Analysis	Symmetric
RJCMA Praveen & Alam (2024)	CVPRW	Emotion Recognition	Symmetric
EAU Gao et al. (2024)	CVPR	Robust Representation Learning	Symmetric
OmniRetriever Liu et al. (2026)	EMNLP	Cross-Modal Retrieval	Symmetric
WAVE Tang et al. (2026)	ICLR	Cross-Modal Retrieval	Symmetric
TripleSumm Kim et al. (2026)	ICLR	Video Summarization	Symmetric
AsymmetricVAD (Ours)	**ICLR**	Video Anomaly Detection	Asymmetric

Table 8: Integration of our asymmetric mechanism into the existing state-of-the-art method on XD-Violence and TMVAD.

Method	XD-Violence (AP %)	TMVAD (AP %)
CMHKF Wang et al. (2025b) (Original)	86.57	86.95
CMHKF + SIE (Ours) (Asymmetric)	**87.12** (↑ 0.55)	**87.73** (↑ 0.78)

C Tri-Modality Methods in Other Domains

As shown in Table 7), tri-modality has been used in diverse domains. However, they all use symmetric paradigms, which may suffer from problems in some cases. For example, for multimodal sentiment analysis, foundational architectures like MulT Tsai et al. (2019) and MISA Hazarika et al. (2020) employed symmetric all-to-all attention, where noisy audio features inevitably corrupt the primary textual sentiment anchor and lead to semantic drift when cross-modal discrepancies or acoustic ambiguities occur. In multimodal video summarization, symmetric models such as TripleSumm Kim et al. (2026) selected visually ordinary frames due to asynchronous background music. The aforementioned problems are due to the symmetric utilization of tri-modality data during the training. Thus, we believe that our asymmetric paradigm can provide insights to more domains.

D More Experimental Results

D.1 Generalization of the Asymmetric Fusion Paradigm

To further evaluate the portability and generalizability of our proposed asymmetric fusion paradigm, we integrate our asymmetric selective injection mechanism into the existing state-of-the-art multi-modal WSVAD method CMHKF Wang et al. (2025b). Specifically, we replace its default equal-importance tri-modality fusion layers with our asymmetric design, utilizing the audio modality as an auxiliary modulator. As shown in Table 8, integrating our asymmetric mechanism into CMHKF improves performance on both benchmarks, with gains of 0.55% on XD-Violence and 0.78% on TMVAD. This improvement suggests that vulnerability to audio noise is a common limitation of symmetric multi-modal fusion. It also demonstrates that our asymmetric paradigm can serve as a flexible, plug-and-play enhancement that improves the robustness of existing multimodal VAD frameworks.

D.2 Comprehensive Evaluation of the Look-Back Mechanism

We first analyze the percentage of samples that execute the Look-Back loop across sequential iterations in Table 9. Specifically, to evaluate the performance at each loop index i, we manually intercept subsequent recalibration passes and evaluate the corresponding AP, where $i = 0$ denotes the baseline without Look-Back recalibration, and *Ratio* indicates the proportion of video samples that require recalibration after the i-th loop relative to the total sample count. We observe that the Look-Back mechanism progressively reduces the number of biased features: initially ($i = 0$), 28.4% and 32.1% of samples exhibit the representation bias and trigger recalibration, whereas after the third loop ($i = 3$), this ratio rapidly diminishes to 1.1% and 1.4%, yielding cumulative AP gains of 1.42% on XD-Violence and 1.61% on TMVAD over the uncalibrated baseline. Although extending the process to $i = 4$ marginally lowers the ratio to 0.3% and 0.4%, with AP performance plateauing at 87.22% and 88.26%, we select $i = 3$ as our default configuration to strike an optimal trade-off between detection accuracy and computational efficiency.

To further evaluate the generalizability and portability of our proposed Look-Back mechanism, we integrate it into two existing state-of-the-art multimodal frameworks: CMHKF and HCC. CMHKF extracts unimodal features using separate backbones, aligns them with visual semantics, and per-

Table 9: Quantitative analysis of the Look-Back mechanism over sequential loops. Here, i denotes the loop index of the sequential Look-Back process ($i = 0$ corresponds to the baseline without Look-Back), and the Look-Back Ratio indicates the percentage of video samples that still require further recalibration after the i-th loop.

Loop (i)	XD-Violence		TMVAD	
	Ratio (%)	AP (%)	Ratio (%)	AP (%)
$i = 0$	28.4	85.79	32.1	86.62
$i = 1$	12.6	86.48	14.3	87.41
$i = 2$	4.5	86.96	5.1	87.95
$i = 3$	1.1	87.21	1.4	88.23
$i = 4$	0.3	87.22	0.4	88.26

Table 10: Performance evaluation of integrating the Look-Back mechanism into existing frameworks on XD-Violence and TMVAD datasets.

Method	**XD-Violence (AP %)**	**TMVAD (AP %)**
CMHKFWang et al. (2025b) (Original)	86.57	86.95
CMHKFWang et al. (2025b) + Look-Back (Ours)	**87.03** (↑ 0.46)	**87.76** (↑ 0.81)
HCCZhou et al. (2026b) (Original)	86.47	84.81
HCCZhou et al. (2026b) + Look-Back (Ours)	**86.82** (↑ 0.35)	**85.58** (↑ 0.77)

forms parallel fusion through multi-branch cross-attention layers. For CMHKF, we insert Look-Back immediately after the cross-attention stage. Instead of directly outputting the concatenated representation, Look-Back evaluates the semantic alignment gap among the multi-branch outputs. If a semantic conflict is flagged, Look-Back generates gating weights from the fused feature to dynamically recalibrate the initial unimodal backbone features, which are then re-routed through the cross-attention layers for a corrected second pass. A similar conditional feedback loop is integrated into the multi-modal fusion stage of HCC. The quantitative results of these experiments are summarized in Table 10. As observed, incorporating our Look-Back mechanism leads to consistent performance improvements for both baseline frameworks across both benchmarks. Specifically, the AP of CMHKF increases by 0.46% on XD-Violence and 0.81% on TMVAD, while HCC achieves an improvement of 0.35% on XD-Violence and 0.77% on TMVAD. This steady improvement demonstrates that the conditional feedback mechanism is not overfitted to our specific architecture, but rather addresses a fundamental limitation common to multimodal networks. By adaptively checking and correcting misaligned representations, the Look-Back mechanism successfully mitigates the representation bias.

D.3 Additional Qualitative Visualizations

To further validate the temporal localization precision and robustness of our proposed framework, we provide additional qualitative visualizations across both the XD-Violence and our TMVAD benchmarks in Fig. 6. The visual results compare our AsymmetricVAD against the competitive tri-modality baseline CMHKF Wang et al. (2025b).

Abuse Scenario (Fig. 6, top-left): This sequence shows an indoor, low-contrast conflict where audio contains mixed speech and ambient interference. We find that CMHKF suffers from ambiguous temporal boundaries in the orange region, where its anomaly score fluctuates between 0.6 and 0.9. Meanwhile, it produces spurious activations during normal intervals (e.g., around frames 200 to 400). In contrast, our AsymmetricVAD exhibits rising and falling edges strictly aligned with the ground truth, maintaining a confident score near 1.0. This demonstrates that the SIE module effectively filters out irrelevant noise.

Car Accident Scenario (Fig. 6, top-right): Car collisions typically exhibit instantaneous, high-impact audio-video bursts. CMHKF demonstrates a severe response lag, where the predicted score slowly increases from frame 176 and only reaches peak confidence after frame 300. Conversely, our method triggers an immediate and sharp spike to 0.9 at the collision and consistently sustains high anomaly certainty until the sequence ends. This responsiveness confirms the sensitivity of our asymmetric audio injection, which captures salient temporal features.

Theft Scenario (Fig. 6, bottom-left): While CMHKF successfully detects the abnormal periods, it yields ambiguous anomaly scores (the orange region) in the normal window due to visual background motion. In contrast, our AsymmetricVAD suppresses the score close to 0.1 throughout the entire normal period, while preserving confident predictions (near 1.0) during both abnormal stages.

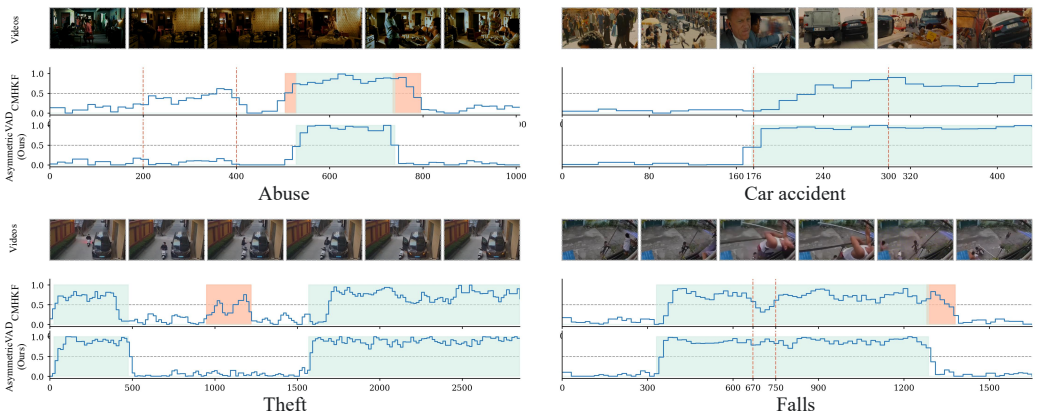

Figure 6: Additional qualitative comparisons of anomaly score predictions between CMHKF and our AsymmetricVAD on XD-Violence (*Abuse*, *Car accident*) and TMVAD (*Theft*, *Falls*).

Table 11: Evaluation of different VLMs on XD-Violence and TMVAD datasets.

VLM	Params	XD-Violence (AP %)	TMVAD (AP %)
BLIP-2 (OPT-2.7B)Li et al. (2023)	3.1 Billion	87.19	<u>88.26</u>
LLaVA-1.5 (7B)Liu et al. (2023)	7.0 Billion	**87.27**	**88.28**
Moondream2Korrapati et al. (2024) (Ours)	1.86 Billion	<u>87.21</u>	88.23

This proves that our Look-Back mechanism effectively identifies cross-modal semantic discrepancies and rectifies biased feature representations before false alarms occur.

Falls Scenario (Fig. 6, bottom-right): Detecting human falls requires tracking subtle physical displacement amid stationary surroundings. In this case, CMHKF exhibits noticeable performance degradation through the incident (around frames 670 to 750) and displays a delayed recovery after the event ends (the orange region). Benefiting from dynamic feature recalibration, our AsymmetricVAD outputs an uninterrupted, stable prediction close to 1.0 across the entire anomaly span and drops to zero once normal conditions resume.

D.4 Ablation Study on the Effectiveness of Different VLMs

To verify the robustness of our framework, we evaluate the performance with different VLMs of BLIP-2, and LLaVA-1.5 for descriptive text prompts of the corresponding video. As presented in Table 11, our method remains stable across these models. It demonstrates that our method is driven by the asymmetric structural designs and the Look-Back mechanism rather than the specific choice of the VLM. Although upgrading to LLaVA-1.5 yields a marginal performance increase of 0.06% on XD-Violence and 0.05% on TMVAD, Moondream2 utilizes 5.14 billion fewer parameters.

D.5 Robustness to Temporal Misalignment

To evaluate robustness against temporal misalignment between video and audio, we compare the original synchronized inputs with asynchronous inputs simulated by random 500 $\sim$ 1000 ms audio delays, as shown in Table 12. The results indicate that CMHKF exhibits vulnerability to asynchronous inputs, with its AP decreasing by 3.30% on XD-Violence and by 3.01% on TMVAD, which can be attributed to phase disruption in its temporal filtering. Conversely, the proposed asymmetric framework demonstrates stability under asynchronous conditions, showing minor AP decreases of only 0.66% and 0.81%, respectively. This stability is assisted by the SIE module, which buffers local offsets via Gaussian-kernel expansion, and the Look-Back mechanism, which dynamically recalibrates features to suppress asynchronous audio channels upon detecting semantic instability.

D.6 Hyperparameter Sensitivity Analysis

Fig. 7 illustrates the sensitivity of anomaly detection performance across different hyperparameters.

As shown in Fig. 7(a), AP exhibits an overall rise-and-fall pattern with cross-modal similarity threshold τ_s. We adopt the peak-performing values of $\tau_s = 0.35$ on XD-Violence and $\tau_s = 0.3$ on TMVAD, a minor variance largely attributable to the different data source characteristics between the two benchmarks.

Table 12: Robustness comparison under artificial video-audio desynchronization with random delays (500 $\sim$ 1000 ms) on XD-Violence and TMVAD.

Temporal State	Method	AP (%)	
		XD-Violence	TMVAD
Original	CMHKFWang et al. (2025b)	86.57	86.95
	AsymmetricVAD (Ours)	**87.21**	**88.23**
Asynchronous	CMHKFWang et al. (2025b)	83.27	83.94
	AsymmetricVAD (Ours)	**86.55**	**87.42**

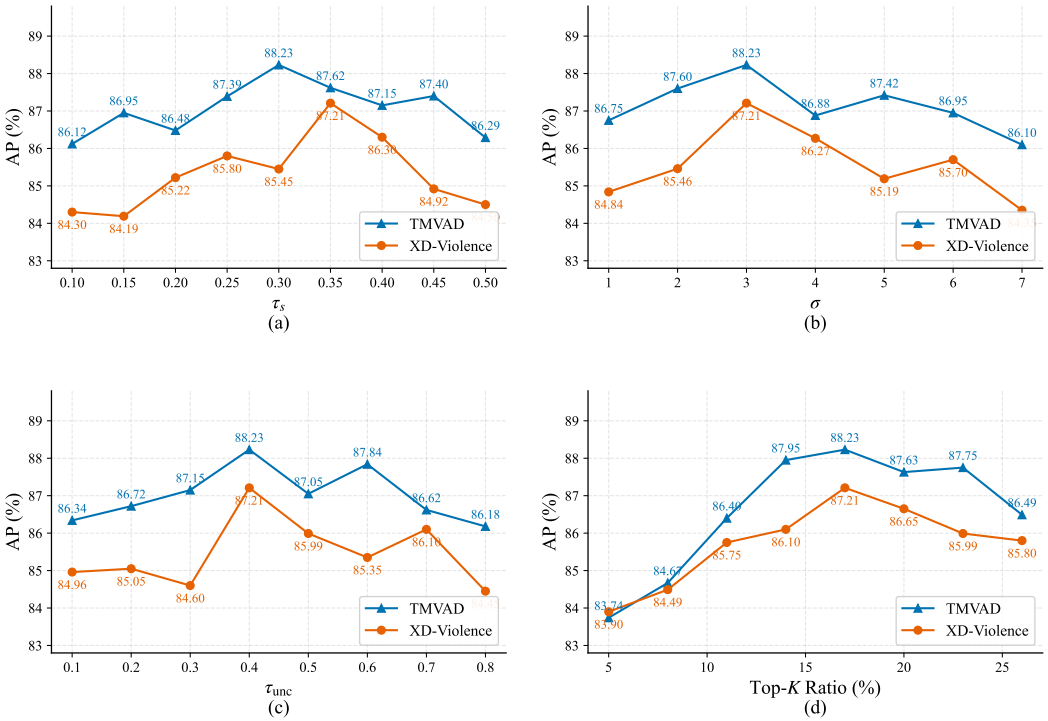

Figure 7: Impact of different hyper-parameters on AP.

Table 13: Computational efficiency comparison of different methods.

Method	Modality	Params (M)↓	FLOPs (G)↓	Inference Latency (ms)↓
VadCLIPWu et al. (2023b)	V+T	**12.64**	5.77	8.1
STPromptWu et al. (2024)	V+T	31.50	44.80	27.4
CMHKFWang et al. (2025b)	V+T +A	12.96	5.99	8.6
VADHOIZhou et al. (2026a)	V	53.84	**2.48**	**7.2**
AsymmetricVAD (Ours)	V+T+A	13.43	14.03	11.4

Regarding the temporal diffusion of acoustic energy, the Gaussian bandwidth in Fig. 7(b) achieves the most favorable balance at our chosen default of $\sigma = 3$ across both benchmarks. Setting $\sigma < 3$ may be insufficient to bridge the potential temporal latency between audio and visual features, whereas a large bandwidth tends to diffuse acoustic energy into distant normal segments, which could introduce undesirable noise.

For the Look-Back discrepancy threshold τ_{unc} (Fig. 7c), balancing gating sensitivity against tolerance to representation biases leads to our default choice of $\tau_{\text{unc}} = 0.4$, around which both datasets attain their maximum AP. A lower threshold (≤ 0.3) may induce unnecessary recalibrations on well-aligned features, while a higher threshold (≥ 0.5) tends to leave multi-modal misalignments insufficiently corrected.

Lastly, as shown in Fig. 7(d), both datasets achieve their best performance around our chosen Top-K ratio of 17%. Setting this ratio too low tends to miss anomalous sound events, while setting it too high may introduce irrelevant background noise into the candidate pool.

D.7 Computational Efficiency Analysis

We analyze computational efficiency and average runtime in Table 13. Compared with CMHKF, AsymmetricVAD introduces an additional 0.47M parameters and 8.04 GFLOPs. This increase is due to the Look-Back mechanism, which requires conditional calibration. However, it maintains a competitive inference latency of 11.4 ms, only 2.8 ms higher than CMHKF, and 4.2 ms higher than VADHOI. This is because the Look-Back loop is triggered only when semantic misalignment occurs. Note that less than 29% and 33% of video samples trigger recalibration on XD-Violence and TMVAD, respectively, as illustrated in Table 9.