Not Only Detection But Also Explain: An Interpretable Multimodal System for Anomaly Detection

Chih-Yung Chang, *Member, IEEE*, Wen-Dong Jiang, *Graduate Student Member, IEEE*, I-Hsiung Chang, and Diptendu Sinha Roy, *Senior Member, IEEE*

***Abstract*—Anomaly detection is often formulated as a Multiple Instance Learning (MIL) problem, where an anomaly detector learns to generate frame-level anomaly scores under multi-modal supervision. However, most previous studies have two drawbacks: 1) the feature fusion between attributes overlooks the model's interpretability; and 2) the training strategy based on pre-trained models leads to excessively long training times. To address these issues, this paper proposes an Interpretable Multi-modal Anomaly Detection System called IMADS. The proposed IMADS abandons the traditional multi-modal fusion approach and adopts single-modal training with late-stage decision fusion optimized by Bayesian techniques to resolve the interpretability problem caused by traditional feature fusion. Moreover, the proposed IMADS introduces a transfer learning method from frame-level to video-level, which essentially adds a temporal dimension to the 2D-to-3D conversion by randomizing parameters along the time dimension during training. This approach not only improves training efficiency but also ensures the model's interpretability. Experimental results demonstrate that the proposed IMADS significantly outperforms existing methods in terms of performance, proving its innovativeness and effectiveness.**

***Index Terms*—Multiple instance learning (MIL), interpretable machine learning, violence detection, behavior recognition.**

I. Introduction

VIDEO Anomaly Detection (VAD) is a fundamental task in computer vision that aims to identify and localize abnormal behaviors in videos [1], [2], [3]. With the increasing deployment of surveillance cameras in various scenarios, VAD has played an increasingly important role in intelligent video surveillance, replacing the previous labor-intensive and time-consuming manual monitoring. In the early stages, VAD tasks typically relied on a single visual modality to recognize anomalous behaviors in videos [1]. However, solely depending on visual information may overlook the context and environment in which anomalous behaviors occur, leading to false positives or false negatives. Benefiting from the rapid development of computers, some researchers have attempted to utilize multimodal and cross-modal techniques [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], integrating data from different modalities such as images, audio, and text, to achieve more comprehensive and accurate anomalous behavior recognition.

Although adopting multimodal approaches can improve the detection performance of anomaly detection systems in videos, two problems still exist: 1) the interpretability issue caused by modal fusion; and 2) the prolonged training process of multimodal models.

For the first issue, the different modalities of data have their own unique characteristics and representations. For instance, video frames contain rich visual information, while audio includes speech and environmental sound features [15]. When fusing these modalities together, the model needs to learn how to coordinate and balance the contributions from different modalities to make the final anomaly judgment [8], [9]. However, this fusion process may lead to the opacity of the decision-making process, making it difficult for users to understand how the model comprehensively considers information from different modalities and reaches its conclusions.

Moreover, features from different modalities may exhibit redundancy or complementarity. Some anomalous behaviors may be reflected in multiple modalities, while others may only appear in specific modalities. [31] Modal fusion needs to effectively handle this redundancy and complementarity, avoiding over-reliance on a single modality while fully leveraging the discriminative information provided by different modalities. However, finding the right balance and interpreting the contribution of each modality remains a challenging topic.

Furthermore, the strategy of modal fusion also affects interpretability [21], [22], [23]. Common fusion strategies include early fusion (feature-level fusion) and late fusion (decision-level fusion). Early fusion combines features from different modalities at an early stage, forming a joint representation, while late fusion integrates decisions made independently by each

Received 25 September 2025; accepted 22 November 2025. Date of publication 28 January 2026; date of current version 26 March 2026. This work was supported by the National Science and Technology Council of Taiwan, Republic of China, under Grant NSTC 112-2622-E-032-005 and in part by the National Science and Technology Council of Taiwan NSTC Under Grant 113-2221-E-032-033-MY2. *(Corresponding author: Chih-Yung Chang.)*

Chih-Yung Chang and Wen-Dong Jiang are with the Department of Computer Science and Information Engineering, Tamkang University, Taipei 25137, Taiwan (e-mail: cychang@mail.tku.edu.tw; 812414018@o365.tku.edu.tw).

I-Hsiung Chang is with the Department of Child Care and Education, Fooying University, Kaohsiung 831301, Taiwan (e-mail: elite5931.tw@gmail.com).

Diptendu Sinha Roy is with the Department of Computer Science and Engineering, National Institute of Technology, Shillong 793003, India (e-mail: diptendu.sr@nitm.ac.in).

This article has supplementary downloadable material available at https://doi.org/10.1109/TETCI.2026.3654456, provided by the authors.

Recommended for acceptance by Y. Zhang.

Digital Object Identifier 10.1109/TETCI.2026.3654456

Fig. 1. The Interpretability Issue Caused by Modal Fusion.

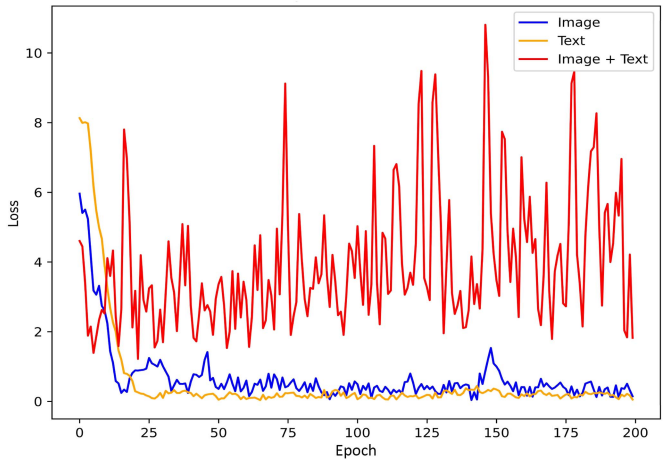


Fig. 2. Convergence Difficulties Caused by Feature Fusion.

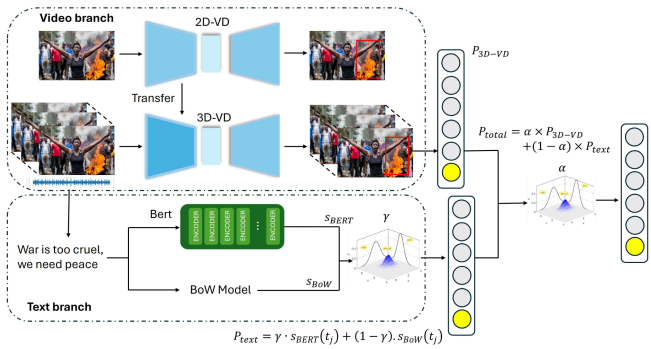


Fig. 3. The process of IMADS.

modality. Different fusion strategies have varying impacts on interpretability. Early fusion may lead to confusion between modalities, making it difficult to analyze the independent contribution of each modality, while late fusion, although preserving the independence of modalities, still faces numerous challenges when interpreting the overall decision.

Fig. 1 shows an example of decision visualization for a multimodal model. As shown in Fig. 1, although the model successfully identifies the anomalous behaviors occurring in these cases, when presented using a heatmap, people find that the violent regions recognized by the model are often difficult to understand. This is because at the decision level, information from various modalities has already been fused into an embedded representation, making it challenging to understand why the model identifies these regions as anomalous behaviors through analysis.

For the Second issue, the anomalous behaviors often involve training on time series data, which can make the model training process extremely lengthy. The characteristics of time series data, such as dynamic changes, long-term dependencies, and periodicity, require the model to capture complex temporal relationships. This necessitates numerous iterations during the training process to ensure that the model learns effective temporal patterns from historical data. This requirement significantly increases training time, especially when dealing with high-dimensional time series data, such as the analysis of video frame continuity. These data are not only voluminous but also have complex temporal correlations among features.

Currently, most studies adopt a fixed training paradigm: first, they perform rapid feature extraction on data from different modalities (e.g., C3D [2], [4], I3D [3], [5], VIT [24]), then fuse the features of the embeddings from different modalities through weighted averaging. Subsequently, they utilize a temporal module (e.g., Transformer [25]) to mine temporal and spatial information, and finally, they employ multi-instance learning for anomalous behavior recognition. Although this approach is relatively efficient in terms of processing speed, it has certain limitations, particularly in terms of gaining a deep understanding of temporal dynamics.

Moreover, data from different modalities have different feature distributions and representation methods [26], [27]. For example, image data are typically high-dimensional, dense pixel matrices, while text data are discrete, sparse word vector sequences. Effectively mapping these heterogeneous data into a common feature space and aligning them at the semantic level is a challenging problem. This leads to difficulties in the rapid convergence of multimodal models during the initial training stages.

Furthermore, the contribution of information from different modalities to the anomaly detection task varies. Some anomalous behaviors may be primarily reflected in visual information, while others may rely more on audio or textual cues. However, during the training process, multimodal models need to automatically learn how to balance and fuse this modal information. This requires the model to continuously adjust the weights and contribution levels of each modality during training to find the optimal fusion strategy. This undoubtedly increases the training difficulty and the required number of iterations for the model.

Fig. 2 illustrates the severe challenges encountered when training multimodal models [28], [29], [30], [32]. As shown in Fig. 2, the training results of single modalities (image and text) indicate that as the number of training rounds increases, the loss function remains relatively stable and gradually converges. However, when the features of image and text are fused together for training, the fluctuations in the loss function significantly increase, and the convergence process becomes exceptionally difficult and unstable.

These notable fluctuations and convergence difficulties reflect the inherent differences in feature distributions and representation methods among different modalities. Fusing these heterogeneous data into a unified semantic space not only requires the

model to overcome a large amount of noise and inconsistencies but also necessitates continuous adjustment and optimization of the weight allocation among modalities during the training process. This process is not only complex and time-consuming but also fraught with uncertainty, severely challenging existing model architectures and learning strategies.

To address the aforementioned issues, this paper proposes a novel, interpretable multimodal anomaly detection system called IMADS. As shown in Fig 3, the system adopts a late fusion strategy for model training, effectively achieving semantic-level alignment of different modalities by separately designing interpretable modules for image and text. Furthermore, the proposed IMADS introduces Bayesian optimization techniques in the modal decision fusion stage, automatically selecting optimal weights to enhance the system's interpretability. The system consists of two main modules: image and text. The image module is divided into a 2D AutoEncoder (2D-VD) and a 3D AutoEncoder (3D-VD), with each layer incorporating deconvolution to enhance the interpretability of the image module. These two modules utilize encoders for feature extraction and perform object detection and classification tasks on the generators. During the training phase, the IMADS first employs the 2D image module for frame-level training, and then uses transfer learning techniques to transfer the encoder weights of 2D-VD to the encoder of 3D-VD, achieving rapid training of the temporal dimension and fine-tuning of 2D parameters. In the text module, the proposed IMADS combines pre-trained models and bag-of-words models for joint decision-making, further improving the decision accuracy and interpretability of the text module through Bayesian optimization. Experiments are conducted on multiple public benchmark datasets and compared with previous works to demonstrate that the proposed IMADS not only improves the accuracy of anomaly detection but also enhances the interpretability of model decisions. The contributions of this study can be summarized in two key aspects:

1) Late fusion strategy with Bayesian optimization optimizes weights and preserves modality independence, improving accuracy and interpretability.
2) Innovative 2D-3D architecture combines spatial and temporal feature extraction, accelerating training and improving accuracy through transfer learning.

The remainder of the paper is organized as follows. Section II discusses and compares previous relevant studies. Section III discusses Assumptions and problems in this paper. Section IV introduce the details the proposed IMADS. Section V provides the experiments and performance evaluation. The conclusion is discussed in Section VI.

II. Related Work

Previous works related to video anomaly detection include single-modal and multi-modal domains. The following are specific related works.

A. Anomaly Detection on Single Modal

In the single-modal works of video anomaly detection, the focus is mainly on processing single images or videos. *Scholkopf et al.* [1] proposed the OCSVM for detecting anomalous behaviors. Their research primarily focused on the recognition of single anomalous images and inputting frame-by-frame data into machine learning models for identifying anomalous behaviors. *Sultani et al.* [2] and *Hasan et al.* [3] were the first to propose using deep multi-instance learning models for anomaly detection. Their research treated video frames as bags, and multiple segments within the frames were considered as instances to enhance the detection performance of the model. *Tran et al.* [4] and *Carreira and Zisserman* [5] used C3D and I3D to extract video features, which were then fed into MIL-based binary classifiers. However, these works focused on single modalities and had oversimplified model designs. *D'Amicantonio et al.* [33] introduced a Gaussian Splatting guided mixture-of-experts framework that specialized experts for different anomaly categories under weak supervision. The temporal splatting loss provided soft temporal guidance from video-level labels. *Zhang et al.* [34] proposed a multi-timescale feature learning approach that extracted short, medium, and long-range temporal tubelets with a video transformer to capture fine motion and contextual cues. The method aggregated multi-timescale representations to adapt pre-trained video features for weakly-supervised anomaly detection in surveillance settings. *Prathibha and Tamizharasan* [35] proposed a lightweight spatio–temporal attention network for weakly supervised video anomaly detection that selected salient spatial regions and temporally informative snippets from binary-labeled videos. They further introduced a dual attention module with few trainable parameters and a weighted loss to sharpen the separation between normal and abnormal segments, enabling deployment on resource-constrained devices. These studies expand the scope of collective or structural anomaly detection—that is, detecting groups or subsets of data points that jointly deviate from normal behavior—from different methodological dimensions. For example, paper [37] proposed dividing data into subsets based on their natural neighbor topological relationships to detect anomalies at multiple granularities. Paper [38] introduced a method that integrates local density with global structural information to enhance sensitivity to group-level anomalies. Paper [39] adopted an adversarial neighbor perception mechanism combined with feature distillation to further strengthen the identification of anomalous patterns across multi-scale neighborhoods. From a methodological perspective, these works help extend the manuscript's current focus on individual anomaly detection toward a broader view of group- or structure-based anomalies, thereby enriching both the related work and the problem formulation sections.

Although multi-instance learning provides advantages in handling time-series data, it may be affected by unclear labels. This makes it difficult for these single-modal works to cope with complex or dynamic real-world anomaly data. Furthermore, the black-box problem inherent in deep learning prevents these models from providing clear decision analysis to users.

B. Anomaly Detection on Multi-Modal

In the multi-modal works of video anomaly detection, the focus is mainly on combining images with audio or images with

text. In the image and audio part, *Zhong et al.* [6] proposed a method based on Graph Convolutional Networks (GCN) to model the feature similarity and temporal consistency between video segments' images and audio. *Tian et al.* [7] used self-attention networks to capture the global temporal context of the joint image and audio embeddings in videos. *Wu et al.* [8] proposed a global and local attention module to capture the temporal dependencies of the joint image and audio embeddings in videos, obtaining more expressive embeddings. *Wu et al.* [9] also proposed a combination method based on Transformer and GCN to distinguish different types of anomalous video frames. *Huang et al.* [17] proposed a transformer-styled temporal feature aggregator and a self-guided discriminative feature encoder to capture the anomalous video feature.

However, due to the often overly sensitive nature of audiomodal data, its effect in the multi-modal training process is not very prominent. In recent years, with significant progress in visual-language pre-training, most research has begun using images combined with text for anomaly detection. For example, *Zhou et al.* [10] proposed classification method based on improved CLIP; *Joo et al.* [11] use the visual features of CLIP to efficiently extract discriminative representations, long-term, and short-term temporal dependencies through temporal self-attention, and nominate clips; *Zhou et al.* [12] applied contrastive learning methods to object detection; *Yu et al.* [13] achieved success in the field of anomalous scene text detection. *Rao et al.* [14] conducted research on dense crowd prediction. *Lv et al.* [15] proposed a multi-instance learning framework called UMIL to learn unbiased anomaly features and improve the performance of the anomalous behavior detection task. *Wu et al.* [16] proposed a simple but powerful baseline that effectively adapts pre-trained image-based visual-language models to leverage their powerful capabilities in general video understanding and applies them to video anomaly detection. *Zhang et al.* [36] proposed a text-guided framework that injected prior knowledge into a frame-prediction pipeline to emphasize crucial relationships, both object–object and object–scene—underlying anomalies in surveillance videos. They encoded explicit priors via a multi-layer GCN embedding network and introduced implicit priors through prompt tuning, which improved generalization, adaptation to new scenarios, and recognition of anomaly types.

However, these multi-modal methods all use pre-trained models for training, which easily leads to the extracted features not being perfectly adaptable to video anomaly detection. Most multi-modal tasks adopt early fusion of features for anomalous behavior detection, resulting in the interpretability of these models being widely questioned.

Table I compares the proposed method with these previous works in terms of whether they are multi-modal and have interpretability, where "O" represents the presence of the attribute and "X" represents the absence of the attribute.

TABLE I
COMPARISON WITH OTHER WORK

Related Works	Model	Multi-Modal	interpretability
[1] (1999)	OCSVM	X	X
[2] [4] (2016)	C3D	X	X
[3] [5] (2018)	I3D	X	X
[33] (2025)	GS-MoE	X	X
[34] (2024)	MTFL	X	X
[35] (2025)	STVAD	X	X
[7] (2021)	RTFM	O	X
[8] (2020)	HD-Net	O	X
[9] (2021)	AVVD	O	X
[17] (2024)	WSTD	O	X
[12] (2022)	CLIP	O	X
[11] (2023)	CLIP-TSA	O	X
[13] (2023)	DMU	O	X
[15] (2023)	UMIL	O	X
[16] (2024)	VadCLIP	O	X
[36] (2024)	CG-VAD	O	X
Our Method	IMADS	O	O

III. ASSUMPTIONS AND PROBLEM FORMULATION

This section introduces the assumptions and problem statement of this study. Given a video V with a duration of t, this paper aims to identify whether or not the anomaly in V. Let $\hat{V}_i \in V$ and $V_i \in V$ denote non-anomaly and anomaly videos, respectively. Let video $V = \{\Phi_1, \Phi_2, \cdots, \Phi_n\}$ be divided into n equal-length segments, each containing non-anomaly, anomaly or a combination of both. Each segment $\Phi_i = \hat{V}_i \cup V_i$might comprise both $\hat{V}_i$ and V_i behaviors.

Let $C = \{c_1, c_2, \cdots, c_m, \hat{c}\}$ denote all possible classes, where c_i denotes the m anomaly classes for $1 \leq i \leq m$ and $\hat{c}$ denotes the non-anomaly class. Let $L = \{l_1, l_2, \cdots, l_m, \hat{l}\}$ be the set of labels, where l_i is the corresponding label of class c_i for $1 \leq i \leq m$ and $\hat{l}$ denote the label of non-anomaly class $\hat{c}$. This study assumes that each V_i falls into a specific anomaly category c_j. Consider an anomaly detection mechanism $\mathcal{M}$ which aims to detect the occurrence of an anomaly event. Let $R^{\mathcal{M}}$ be a segment of video which is detected as the anomaly event by applying mechanism $\mathcal{M}$. That is, R^M can be represented as

$$R^M = \left\{\hat{R}_1^M, R_1^M, \hat{R}_2^M, R_2^M \ldots.., \hat{R}_n^M, R_n^M\right\}.$$

Let δ_i be a Boolean variable that indicates whether or not a given video segment $\tilde{R}_i$ actually contains an anomaly event. That is,

$$\delta_i = \begin{Bmatrix} 1, & \tilde{R}_i \in c_j \\ 0, & \tilde{R}_i \in \hat{c} \end{Bmatrix}.$$

Let θ be the prediction threshold. Let $\delta_i^M(\theta)$ denote whether or not the result of video segment $\tilde{R}_i$ predicted by M contains anomaly event. Let probability P_i denote the prediction score of $\tilde{R}_i \in R^M$. That is,

$$\delta_i^M(\theta) = \begin{cases} 1, & \text{if } P_i \geq \theta \\ 0, & otherwise. \end{cases}$$

Let TP_i, TN_i, FP_i and FN_i represent True Positive, True Negative, False Positive, and False Negative, respectively, for the prediction result of input segment $\tilde{R}_i$ by applying mechanism M. These per-segment metrics evaluate the accuracy for each individual video segment, where TP_i indicates correct anomaly detection, TN_i indicates correct non-anomaly classification, FP_i indicates a false alarm (non-anomaly misclassified as

anomaly), and FN_i indicates a missed anomaly. Their values can be calculated using Exp. (1):

$$\begin{aligned} TP_i &= \delta_i \times \delta_i^M(\theta), \\ FP_i &= (1-\delta_i) \times \delta_i^M(\theta), \\ FN_i &= \delta_i \times \left(1-\delta_i^M(\theta)\right), \\ TN_i &= (1-\delta_i) \times \left(1-\delta_i^M(\theta)\right). \end{aligned} \tag{1}$$

Let TP, TN, FP, and FN represent the cumulative True Positive, True Negative, False Positive, and False Negative, respectively, of the prediction results by applying mechanism $\mathcal{M}$ to all segments $\tilde{R}_i \in R$, for $1 \le i \le n$. These cumulative metrics aggregate the per-segment results to assess overall performance across the entire video. The values of TP, TN, FP, and FN can be calculated by Exp. (2):

$$\begin{aligned} TP &= \sum_{i=1}^{n} TP_i,\ FP = \sum_{i=1}^{n} FP_i \\ TN &= \sum_{i=1}^{n} TN_i\, FN = \sum_{i=1}^{n} FN_i. \end{aligned} \tag{2}$$

Let $\mathcal{P}^M$ and $\mathcal{R}^M$ denote the Precision and Recall of the predictions by applying mechanism *M* to predict a given video V. The values of $\mathcal{P}^M$ and $\mathcal{R}^M$ can be calculated by Exp. (3) and Exp. (4):

$$\mathcal{P}^M = \frac{TP}{TP+FP} \tag{3}$$

$$\mathcal{R}^M = \frac{TP}{TP+FN}. \tag{4}$$

A. Average Precision ($\mathcal{AP}$)

Let θ_j denote the *j*-th prediction threshold, let $\mathcal{AP}^{\mathcal{M}}$ denote the *Average Precision* of mechanism $\mathcal{M}$. The $\mathcal{AP}^{\mathcal{M}}$ can be calculated by Exp. (5),

$$\mathcal{AP}^{\mathcal{M}} = \sum_{j=1}^{n-1} \left(\mathcal{R}(\theta_{j+1})^{\mathcal{M}} - \mathcal{R}(\theta_j)^{\mathcal{M}}\right) \times \mathcal{P}(\theta_{j+1})^{\mathcal{M}}. \tag{5}$$

The first objective of this paper is to develop mechanism M^{best} that satisfies Exp. (6).

First Objective in XD-Violence:

$$M^{best} = arg \max_{M \in \mathcal{M}} \left(\mathcal{AP}^{\mathcal{M}}\right). \tag{6}$$

B. Area Under the Curve (AUC)

Let $\mathcal{A}_M$ denote the *AUC* of mechanism *M*. The value of $\mathcal{A}_M$ can be calculated by Exp. (7),

$$\begin{aligned} \mathcal{A}_M &= \int_0^1 TPR(\theta)\, d\left(FPR(\theta)\right) \\ &= \int_0^1 \frac{TP(\theta)}{TP(\theta)+FN(\theta)} d\left(\frac{FP(\theta)}{FP(\theta)+TN(\theta)}\right). \end{aligned} \tag{7}$$

The second objective of this paper is to develop mechanism M^{best} that satisfies Exp. (8).

TABLE II
2D-VD MODAL CONFIGURATION

Layer Name	Output Size	Configuration
$Conv1$	112×112	$7 \times 7,\ 64,\ Stride2$
		$3 \times 3\ max\ pool, Stride\ 2$
$Conv_2^{2D-VD}$	56×56	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix} \times 2$
$Encoder_1$		$[1 \times 1, 64]$
$Deconv2_x$	56×56	$1 \times 1, 64$
$Decoder_1$		$[1 \times 1, 64]$
$Conv_3^{2D-VD}$	28×28	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 2$
$Encoder_2$		$[1 \times 1, 128]$
$Deconv3_x$	28×28	$1 \times 1, 128$
$Decoder_2$		$[1 \times 1, 128]$
$Conv_4^{2D-VD}$	14×14	$\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 256 \end{bmatrix} \times 2$
$Encoder_3$		$[1 \times 1, 256]$
$Deconv4_x$	14×14	$1 \times 1, 256$
$Decoder_3$		$[1 \times 1, 256]$
$Conv_5^{2D-VD}$	7×7	$\begin{bmatrix} 3 \times 3, & 512 \\ 3 \times 3, & 512 \end{bmatrix} \times 2$
$Encoder_4$		$[1 \times 1, 512]$
$DeConv5_x$	7×7	$1 \times 1, 512$
$Decoder_4$		$[1 \times 1, 512]$
R_{2D-VD}		Fully connected layer
C_{2D-VD}		Fully connected layer
	1×1	Average bool,1000-d fc, softmax
$FLOPs$		1.8×10^9

Objective in UCF-Crime:

$$M^{best} = arg \max_{M \in \mathcal{M}} \left(\mathcal{A}_M\right). \tag{8}$$

This section introduced the assumptions and problem formulation of this paper. The next section will introduce the proposed IMADS.

IV. THE PROPOSED IMADS

This section will introduce the proposed IMADS. Fig. 1 shows the brief process of the designed IMADS. The proposed IMADS consists of two branches: the image branch and the text branch. The image branch includes two modules, 2D-VD and 3D-VD, while the text branch includes two modules, Bert and BoW. The goal of IMADS is to combine the prediction results from the image branch P_{3D-VD} and the prediction results from the text branch P_{text} to find the optimal weight α,for late fusion of the modalities using Bayesian optimization. Herein, P_{3D-VD} represents the prediction results of 3D-VD on image data and P_{text} represents the decision fusion of Bert and BoW, which also uses Bayesian optimization to find the optimal weight γ to complete the text branch decision fusion. The following is a detailed introduction to IMADS:

A. Video Branch

The video branch consists of two modules: the 2D-VD module and the 3D-VD module. Tables II and III show the specific configurations of the 2D-VD module and the 3D-VD module.

TABLE III
3D-VD MODAL CONFIGURATION

Layer Name	Output Size	Configuration
$Conv1$	$112 \times 112 \times 112$	$7 \times 7 \times 7,\ 64,\ Stride3$
		$3 \times 3 \times 3\ max\ pool\ , Stride\ 3$
$Conv_2^{3D-VD}$	$56 \times 56 \times 56$	$\begin{bmatrix} 3 \times 3 \times 3, & 64 \\ 3 \times 3 \times 3, & 64 \end{bmatrix} \times 2$
$Encoder_1$		$[1 \times 1\ , 64]$
$Deconv2_x$	$56 \times 56 \times 56$	$1 \times 1, 64$
$Decoder_1$		$[1 \times 1\ , 64]$
$Conv_3^{3D-VD}$	$28 \times 28 \times 28$	$\begin{bmatrix} 3 \times 3 \times 3, & 128 \\ 3 \times 3 \times 3, & 128 \end{bmatrix} \times 2$
$Encoder_2$		$[1 \times 1 \times 1\ , 128]$
$Deconv3_x$	$28 \times 28 \times 28$	$1 \times 1 \times 1, 128$
$Decoder_2$		$[1 \times 1 \times 1\ , 128]$
$Conv_4^{3D-VD}$	$14 \times 14 \times 14$	$\begin{bmatrix} 3 \times 3 \times 3, & 256 \\ 3 \times 3 \times 3, & 256 \end{bmatrix} \times 2$
$Encoder_3$		$[1 \times 1 \times 1\ , 256]$
$Deconv4_x$	$14 \times 14 \times 14$	$1 \times 1 \times 1\ , 256$
$Decoder_3$		$[1 \times 1 \times 1\ , 256]$
$Conv_5^{3D-VD}$	$7 \times 7 \times 7$	$\begin{bmatrix} 3 \times 3 \times 3, & 512 \\ 3 \times 3 \times 3, & 512 \end{bmatrix} \times 2$
$Encoder_4$		$[1 \times 1 \times 1\ , 512]$
$DeConv5_x$	$7 \times 7 \times 7$	$1 \times 1 \times 1, 512$
$Decoder_4$		$[1 \times 1 \times 1\ , 512]$
$\mathcal{R}$		Fully connected layer
$\mathcal{C}$		Fully connected layer
	$1 \times 1 \times 1$	Average bool, fc, softmax
$FLOPs$		1.8×10^9

The image branch consists of two modules: the 2D-VD module and the 3D-VD module. The 2D-VD module adopts a standard 2D residual neural network architecture, uses an AutoEncoder architecture for anomaly detection due to its specific focus on violence, enhancing system efficiency and speed. where each layer includes convolutional layers, encoders, deconvolutional layers, and decoders for extracting image features and providing interpretability.

Specifically, the first layer uses a 7×7 convolutional kernel with an output channel number of 64 and a stride of 2, followed by a 3×3 max-pooling layer with a stride of 2. The following layers are $Conv_2^{2D-VD}$, $Conv_3^{2D-VD}$, $Conv_4^{2D-VD}$, and $Conv_5^{2D-VD}$, each containing two 3×3 convolutional layers with output channel numbers of 64, 128, 256, and 512, respectively. After each convolutional layer, there is an encoder (using a 1×1 convolutional kernel to adjust the number of channels), a deconvolutional layer (using a 1×1 convolutional kernel), and a decoder (using a 1×1 convolutional kernel).

Finally, the classification is performed through an average pooling layer and two fully connected layer, using the softmax function to output the results. The FLOPs are setting in 1.8×10^9. The regression and classification layers for object detection are denoted by $R_{\text{2D-VD}}$ and $C_{\text{2D-VD}}$ respectively.

The training loss function for $R_{\text{2D-VD}}$, L_{box} is shown in Exp. (9) as follows:

$$L_{box} = 1 - \frac{|B_g \cap B_p|}{|B_g \cup B_p| + \epsilon}, \tag{9}$$

where B_g and B_p are the ground truth and predicted bounding boxes, respectively, and ϵ is a small constant to avoid division by zero. The training loss function for $C_{\text{2D-VD}}$, L_{cls} is defined as the binary cross-entropy loss in Exp (10):

$$L_{cls} = -y\ log\ (\hat{y}) - (1 - y) \log\ (1 - \hat{y})\ , \tag{10}$$

where y is the ground truth label and $\hat{y}$ is the predicted probability.

The 3D-VD module adopts a 3D residual neural network architecture to handle the spatiotemporal features in video data. The architecture of the AutoEncoder designed by 2D-VD is the same, except that the convolution in the time dimension is added. Specifically, the first layer uses a $7 \times 7 \times 7$ convolutional kernel with an output channel number of 64 and a stride of 3, followed by a $3 \times 3 \times 3$ max-pooling layer with a stride of 3. The subsequent layers are $Conv_2^{3D-VD}$, $Conv_3^{3D-VD}$, $Conv_4^{3D-VD}$ and $Conv_5^{3D-VD}$. each containing two $3 \times 3 \times 3$ convolutional layers with output channel numbers of 64, 128, 256, and 512, respectively. After each convolutional layer, there is an encoder (using a $1 \times 1 \times 1$ convolutional kernel to adjust the number of channels), a deconvolutional layer (using a $1 \times 1 \times 1$ convolutional kernel), and a decoder (using a $1 \times 1 \times 1$ convolutional kernel). Finally, classification is performed through an average pooling layer and two fully connected layers, using the softmax function to output the results. The FLOPs are setting in 1.8×10^9.The regression and classification layers for object detection are denoted by $R_{\text{3D-VD}}$ and $C_{\text{3D-VD}}$, respectively.

The training loss function for $R_{\text{3D-VD}}$,$L_{box-smooth}$ incorporates a temporal smoothness term for bounding boxes in Exp. (11):

$$L_{box-smooth} = L_{boxt} + \lambda_{smooth} \sum_{i=t}^{t+1} \| \left(B_{g_i} - B_{p_i}\right) - \left(B_{g_{i+1}} - B_{p_{i+1}}\right) \|_2^2, \tag{11}$$

where B_{g_i} and B_{p_i} ground truth and predicted bounding boxes at time step, respectively. The training loss function for $C_{\text{3D-VD}}$, $L_{cls-smooth}$ incorporates a temporal smoothness term Exp. (12):

$$L_{cls-smooth} = L_{clst} + \lambda_{smooth} \sum_{i=t}^{t+1} \|p_i - p_{i+1}\|_2^2, \tag{12}$$

Where p_i and p_{i+1} are the predicted probabilities at consecutive time steps, and λ_{smooth} is the smoothness coefficient.

To quickly train the 3D-VD module, the proposed IMADS adopts a transfer learning method. First, the 2D-VD model is trained using frame-level annotations. After the 2D-VD training is completed, the parameters from the 2D-VD encoder are transferred to the 3D-VD module, and videos are used to train the 3D-VD model. This method not only allows for rapid training but also reduces the data requirements for the 3D-VD, achieving good results with fewer frame-level annotations. The specific steps are as follows:

First, the 2D-VD is trained. After the training is completed, the parameters are transferred to the 3D-VD for transfer learning. The weights of the l-th layer encoder in the 2D-VD module,

denoted as $\boldsymbol{W}_l^{2D}$, are defined with the shape in Exp. (13):

$$\boldsymbol{W}_l^{2D} \in \mathbb{R}^{C_1 \times C_{l-1} \times 1 \times k \times k}, \tag{13}$$

Herein, $C_1$and C_{l-1} represent the number of channels in the l-th and $l-1$-th layers, respectively, and k denotes the kernel size.

Similarly, the weights of the l -th layer encoder in the 3D-VD module, denoted as $\boldsymbol{W}_l^{3D}$, are defined with the shape in Exp. (14):

$$\boldsymbol{W}_l^{3D} \in \mathbb{R}^{C_1 \times C_{l-1} \times t \times k \times k}. \tag{14}$$

Herein, t denotes the temporal dimension. During the transfer process, the weights from $\boldsymbol{W}_l^{2D}$ are copied to the spatial dimensions of $\boldsymbol{W}_l^{3D}$ as follows in Exp. (15):

$$w_{l,i,j,r,p,q}^{3D} = \begin{cases} w_{l,i,j,0,p,q}^{2D}, & r = 0 \\ 0, & r \neq 0, \end{cases} \tag{15}$$

Where $i \in \{1, 2, \ldots, C_l\}$, $j \in \{1, 2, \ldots, C_{l-1}\}$, $r \in \{1, 2, \ldots, t-1\}$, $p, q \in \{1, 2, \ldots, k\}$.

Next, Xavier initialization is used to initialize the weights of $\boldsymbol{W}_l^{3D}$ along the temporal dimension as shown in Exp. (16):

$$w_{l,i,j,r,p,q}^{3D} \sim \mathcal{N}\left(0, \sqrt{\frac{2}{C_l + C_{l-1}}}\right), \; r \neq 0. \tag{16}$$

In the fine-tuning stage, $\boldsymbol{W}_l^{3D}$ is trained end-to-end with the objective:

$$\min_{W^{3D}} L_{cls-smooth}\left(f_{cls}\left(X; W^{3D}\right), Y\right) + \lambda \|\boldsymbol{W}^{3D} - \boldsymbol{W}^{2D}\|_F^2 + L_{box-smooth}\left(f_{reg}\left(X; W^{3D}\right), B\right),$$

Herein, $L_{cls-smooth}$ and $L_{box-smooth}$ are the smooth classification and regression loss functions respectively. $f(X; W^{3D})$ represents the prediction function using the 3D-VD module for input X, Y denotes the true labels, B denotes the true labels the bounding box regression. $\|.\|_F$ denotes the Frobenius norm, λ is a hyperparameter balancing the parameter transfer loss and the task loss, and L denotes the number of layers in the network. Expand this, the objective is shown in Exp. (17):

$$\begin{aligned} &\min_{W^{3D}} \left(\frac{1}{N} \sum_{n=1}^{N} L_{cls-smooth}\left(f_{cls}\left(x_n; W^{3D}\right), y_n\right) \right. \\ &\left. + \frac{1}{N} \sum_{n=1}^{N} L_{box-smooth}\left(f_{reg}\left(x_n; W^{3D}\right), b_n\right) \right) \\ &+ \lambda \sum_{l=1}^{L} \sum_{i=1}^{C_l} \sum_{j=1}^{C_{l-1}} \sum_{r=0}^{t-1} \sum_{p=1}^{k} \sum_{q=1}^{k} \left(w_{l,i,j,r,p,q}^{3D} - w_{l,i,j,0,p,q}^{2D}\right)^2. \end{aligned} \tag{17}$$

Through this transfer learning method, the 3D-VD module can quickly perform fine-tuning in the spatial dimension and training in the temporal dimension. Additionally, this method allows the 3D-VD module to achieve good model performance with less video annotation data.

B. Text Branch

The text branch primarily includes two modules: the BERT pre-trained module and the Bag-of-Words (BoW) module. The input text comes from the speech-to-text transcription of video data. The output is a combined decision from both modules. Unlike manual settings or expert decisions, the decision weights in the text branch are determined using Bayesian optimization. This approach not only enables automated selection but also provides interpretability, as Bayesian optimization focuses on the posterior probability distribution, considering both prior knowledge and the likelihood of the data. The specific operation of the text branch is as follows:

Let $x = \{x_1, x_2, x_3, \ldots.., x_n\}$ denote the input video, where x_j represents each 1-second video segment, and the corresponding transcribed text is $t = \{t_1, t_2, t_3, \ldots.., t_n\}$, whrere t_j represents the transcribed text for each 1-second video segment. The operation of BERT is denoted as $f_{BERT}(\cdot)$, represent as: $s_{BERT}(t_j) = f_{BERT}(t_j)$, where $s_{BERT}(t_j)$ represents the prediction score for the text t_j of BERT model. The operation of BoW is denoted as $f_{BoW}(\cdot)$, $s_{BoW}(t_j) = f_{BoW}(t_j)$, where $s_{BoW}(t_j)$ represents the prediction score for the text t_j of BoW model.

Bayesian optimization is then performed with the goal of finding the optimal weight γ, where $\gamma_{BERT} = \gamma$ and $\gamma_{BoW} = 1 - \gamma$. The objective is to maximize the posterior probability $p(\gamma|x, y)$, as shown in Exp. (18):

$$\gamma^* = \arg\max_{\gamma} p(\gamma|t, y). \tag{18}$$

According to Bayes' theorem, the posterior probability can be decomposed as shown in Exp. (19):

$$p(\gamma|t, y) = \frac{p(\gamma|t, y) \cdot (\gamma)}{p(t, y)} \propto p(\gamma|t, y) \cdot (\gamma). \tag{19}$$

Where $p(\gamma|t, y)$ is the likelihood term and $p(\gamma)$ is the prior term. For the likelihood term, it can be further decomposed as shown in Exp. (20):

$$p(\gamma|t, y) = \prod_{j=1}^{n} p(\gamma|t_j, y) = \prod_{j=1}^{n} p(t_j|y, \gamma) \cdot (\gamma|y). \tag{20}$$

Assuming the prediction scores of each segment t_j are conditionally independent, as shown in Exp. (21):

$$\begin{aligned} p(\gamma|t, y) &= \prod_{j=1}^{n} p(t_j|y, \gamma) \cdot (\gamma|y) \\ &= \prod_{j=1}^{n} \left[\gamma \cdot s_{BERT}(t_j) + (1 - \gamma) \cdot s_{BoW}(t_j)\right] \cdot (\gamma|y). \end{aligned} \tag{21}$$

For the prior term $p(\gamma)$, a Gaussian distribution that encourages sparsity and small magnitude of the weights can be chosen, as shown in Exp. (22):

$$p(\gamma) = \mathcal{N}(\gamma|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\gamma - \mu)^2}{2\sigma^2}\right). \tag{22}$$

Substituting the likelihood and prior terms into the posterior probability expression and taking the logarithm, as indicated in

Exp. (23):

$$\begin{aligned} logp\left(\gamma|t,y\right) &= \sum_{j=1}^{n} log\left[\gamma \cdot s_{BERT}\left(t_j\right) + \left(1-\gamma\right).s_{BoW}\left(t_j\right)\right] \\ &\quad + \log p\left(\gamma|y\right) + \log p\left(\gamma\right) + const \\ &= \sum_{j=1}^{n} log\left[\gamma \cdot s_{BERT}\left(t_j\right) + \left(1-\gamma\right).s_{BoW}\left(t_j\right)\right] \\ &\quad + \log p\left(\gamma|y\right) - \frac{\left(\gamma-\mu\right)^2}{2\sigma^2} + const. \end{aligned} \tag{23}$$

The optimal weight γ^*can be obtained by maximizing the above objective function in Exp. (24):

$$\begin{aligned} \gamma^* = \arg\max_{\gamma} &\sum_{j=1}^{n} \log\left[\gamma \cdot s_{BERT}\left(t_j\right) + \left(1-\gamma\right) \cdot s_{BoW}\left(t_j\right)\right] \\ &+ \log p\left(\gamma|y\right) - \frac{\left(\gamma-\mu\right)^2}{2\sigma^2}. \end{aligned} \tag{24}$$

This optimization problem can be solved using gradient descent or its variants. Let the objective function be $\mathcal{L}(\gamma)$, then the update rule for gradient descent is given in Exp. (25):

$$\gamma^{(t+1)} = \gamma^{(t)} \ - \eta \cdot \nabla_{\gamma} \mathcal{L}\left(\gamma^{(t)}\right), \tag{25}$$

Where η is the learning rate, and $\nabla_{\gamma}\mathcal{L}(\gamma^{(t)})$ is the gradient of the objective function at $\gamma^{(t)}$.By iteratively updating the weight until convergence, the optimal weight γ^* can be obtained. During this process, the prior term $-\frac{(\gamma-\mu)^2}{2\sigma^2}$ acts as a regularize, encouraging sparsity and small magnitude of the weights, thereby improving model interpretability. By adjusting the hyperparameter σ, the influence of the prior on the posterior can be controlled, balancing model performance and interpretability.

C. Late Fusion

After both the text branch and the video branch have been trained, during the usage phase, the video is input into the IMADS system. The decision output for the video part, $P_{3D-VD}$is determined by the 3D−VD module. The decision for the text branch, $P_{text} = \gamma \cdot s_{BERT}(t_j) + (1-\gamma).s_{BoW}(t_j)$ is composed by the BERT module and the BoW module, and the optimal weight γ is found through Bayesian optimization.

The final comprehensive decision, $P_{total} = \alpha \cdot P_{3D-VD} + (1-\alpha) \cdot P_{text}$,is determined by repeating Exp. (18)–(25) to find the optimal final weight α^*.

V. Performance Evaluation

In this section, relevant experimental performance and analysis are presented.

A. Dataset

The experiments are conducted on two popular video anomaly detection datasets: UCF-Crime and XD-Violence. The XD-Violence dataset is currently the largest publicly available dataset in the field of violence detection. It includes 4,754 videos, totaling 217 hours, covering six types of violent events: verbal abuse, car accidents, explosions, brawls, riots, and shootings. This dataset is randomly divided into a training set with 3,954 videos and a test set with 800 videos. The test set is further categorized into 500 violent videos and 300 non-violent videos. The UCF-Crime dataset contains 1,900 real-world surveillance videos, with 1,610 for training and 290 for testing. It is important to note that the training videos in both the XD-Violence and UCF-Crime datasets only have video-level labels. For the 2D-VD training experiments, 600 videos were selected from the training sets of both XD-Violence and UCF-Crime for frame-by-frame annotation, processing each video at 30 frames per second. In XD-Violence, each video is selected to be 5 seconds long; in UCF-Crime, each video is chosen to be 3 seconds long. For the textual part, for XD-Violence, video-to-text transcription is performed every 1 second. Moreover, since the UCF-Crime dataset only includes image modality without sound and text, the experiments also followed the settings from references [15][16][17], using a CLIP pre-trained model designed for anomaly detection to process each frame of the videos to match corresponding text descriptions, meeting the experimental needs.

B. Implementation Details

The IMADS is trained on a single NVIDIA A100 GPU using PyTorch, with AdamW as the optimizer. For the choice of pre-trained textual models, the module selected is english-abusive-MuRIL, and the bag-of-words model chosen is XG-Boost.

In the settings of hyperparameters, for the 3D-VD transfer learning in Exp. (18), λ is set to 0.7, and for the Bayesian optimization in Exp (25), σ is set to 1.2. The learning rate η for Bayesian optimization is uniformly set to 10^{-4} in Exp. (26). The initial values of γ and α are both set to 0.1.

In coarse-grained tasks, the learning rates for the image branch in 2D-VD are set to 5×10^{-4} for both XD-Violence and UCF-Crime, and the batch size is 256, with 200 training epochs. A sigmoid function is used to implement binary classification for coarse granularity. The learning rates for 3D-VD in XD-Violence and UCF-Crime are set to 3×10^{-4} and 4×10^{-4}, respectively, with 150 training epochs. Similarly, the sigmoid function is used to achieve binary classification for coarse granularity. In the textual branch, for XG-Boost, the learning rate in XD-Violence is set to 4×10^{-4}, with 250 trees and a depth of 100 per tree; in UCF-Crime, the learning rate is set to 5×10^{-4}, with 500 trees and a depth of 250 per tree.

In fine-grained tasks, the learning rates for the image branch in 2D-VD are set to 2×10^{-4} for XD-Violence and 1×10^{-4} for UCF-Crime, with 300 training epochs. The learning rates for 3D-VD in XD-Violence and UCF-Crime are 3×10^{-4} and 4×10^{-4}, respectively, with 200 training epochs. In the textual branch, for XG-Boost, the learning rate in XD-Violence is set to 4×10^{-4}, with 500 trees and a depth of 300 per tree; in UCF-Crime, the learning rate is set to 5×10^{-4}, with 530 trees and a depth of 250 per tree.

C. Comparison With State-of-the-Art Methods

The proposed IMADS can realize coarse-grained and fine-grained violence detection. The experiment presents the performance of the proposed IMADS and compares it with several

TABLE IV
COARSE-GRAINED COMPARISONS ON XD-VIOLENCE (AP%) AND UCF-CRIME (AUC% AND ANO-AUC%)

Modal	Method	AP (%)	AUC (%)	Ano-AUC (%)
Single-Modal	SVM baseline	50.80	50.10	50.00
	OCSVM	28.63	63.20	51.06
	C3D	31.25	51.20	39.43
	I3D	45.23	50.48	50.09
	STVAD	83.33	84.98	71.34
	GS-MoE	82.89	**91.58**	**83.86**
	MTFL	84.57	87.16	67.26
Multi-Modal	CLIP	76.57	84.72	62.60
	I3D	75.18	84.14	63.29
	HD-Net	80.00	84.57	62.21
	RTFM	78.27	85.66	63.86
	AVVD	78.10	82.45	60.27
	DMU	82.41	86.75	68.62
	WSTD	83.56	87.95	68.84
	CLIP-TSA	82.17	87.58	62.45
	VadCLIP	84.51	88.02	70.26
	UMIL	81.68	86.75	68.68
	CG-VAD	84.57	86.70	68.30
	IMADS (Ours)	**86.61**	88.03	73.68

TABLE V
FINE-GRAINED COMPARISONS ON XD-VIOLENCE

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
Random Baseline	1.82	0.92	0.48	0.23	0.09	0.71
SVM	18.64	12.53	11.05	8.26	4.32	10.96
OCSVM	19.85	9.06	8.56	6.55	6.03	10.01
C3D	20.28	13.72	8.44	5.06	2.81	10.06
I3D	22.72	15.57	9.98	6.20	6.78	12.25
GS-MoE	37.44	31.11	26.34	16.85	12.76	24.90
STVAD	36.44	27.34	24.51	14.42	11.56	22.85
MTFL	32.09	26.12	27.46	17.44	15.54	23.53
HD-Net	35.35	28.02	20.94	15.01	10.33	21.93
CLIP	29.68	23.11	20.08	13.58	11.24	19.53
AVVD	30.51	25.75	20.18	14.83	9.79	20.21
RTFM	31.25	26.85	21.94	13.56	12.54	21.23
DMU	32.33	28.88	22.57	14.33	13.68	22.36
WSTD	33.95	26.33	23.22	17.56	14.53	23.12
UMIL	34.44	27.13	22.63	**19.85**	13.24	23.46
CLIP-TSA	34.53	32.88	28.11	13.65	10.01	23.84
VadCLIP	37.03	30.84	23.38	17.90	14.31	24.70
CG-VAD	35.76	29.33	21.56	17.21	14.22	23.61
IMADS (Ours)	**39.88**	**31.46**	**28.55**	19.48	**16.49**	**27.17**

TABLE VI
FINE-GRAINED COMPARISONS ON UCF-CRIME

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
Random Baseline	0.21	0.14	0.04	0.02	0.01	0.08
SVM	3.64	3.59	2.39	1.34	0.98	2.39
OCSVM	4.85	4.44	2.33	1.85	1.56	3.01
C3D	4.33	3.88	2.46	1.66	2.11	2.89
I3D	5.73	4.41	2.69	1.93	1.44	3.24
GS-MoE	5.24	4.87	5.32	3.24	3.01	4.34
STVAD	10.81	6.28	5.48	3.58	2.24	5.67
MTFL	10.26	5.23	5.55	3.72	2.59	5.47
HD-Net	8.47	6.88	5.62	4.91	3.38	5.85
CLIP	10.11	6.56	5.34	4.01	1.77	5.56
AVVD	10.27	7.01	6.25	3.42	3.29	6.05
RTFM	12.59	7.54	6.44	5.42	1.54	6.71
DMU	11.32	7.62	5.97	4.33	2.36	6.32
WSTD	11.95	6.98	5.88	4.26	3.57	6.53
UMIL	11.84	7.85	6.52	3.97	2.84	6.60
CLIP-TSA	12.62	8.13	6.66	4.28	1.91	6.72
VadCLIP	11.72	7.83	6.40	4.53	2.93	6.68
CG-VAD	11.52	6.18	5.26	3.49	1.78	5.65
IMADS (Ours)	**12.66**	**8.24**	**6.69**	**5.49**	**3.83**	**7.38**

state-of-the-art methods on coarse-grained and fine- grained VAD tasks.

Coarse-grained VAD Results: Table IV reports coarse-grained anomaly detection results on XD-Violence (AP) and UCF-Crime (AUC, Ano-AUC). Overall, multimodal methods outperform unimodal ones on most evaluation dimensions. Against this backdrop, our proposed IMADS demonstrates stable and strong overall performance on both benchmarks: it achieves 86.61% AP on XD-Violence, an outstanding result among the compared methods and 88.03% AUC on UCF-Crime. It slightly surpassing strong competitors VadCLIP (88.02%) and WSTD (87.95%), while also attaining the highest Ano-AUC (73.68%) among multimodal methods. Furthermore, relative to VadCLIP/WSTD, the proposed IMADS improves AP on XD-Violence by +2.10/+3.05 percentage points; on UCF-Crime, the AUC gains are +0.01/+0.08, reflecting a small but consistent lead. Notably, although AVVD uses multi-category labels, it reaches only 78.10% AP on XD-Violence and 82.45% AUC on UCF-Crime, substantially below the proposed IMADS, which indicating that increasing label granularity is not sufficient; effective utilization of annotations is crucial.

At the methodological level, prior multimodal works often adopt early fusion at the embedding layer, which can induce modality imbalance and complicate optimization. IMADS employs a modular, decoupled design for image/text branches with late fusion: each modality is fully trained in its own stage, after which Bayesian optimization adaptively searches fusion weights, delivering better convergence stability and cross-dataset generalization while maintaining efficiency. From a broader perspective, we also acknowledge several limitations and boundaries: (i) on UCF-Crime AUC, certain unimodal methods (e.g., GS-MoE: 91.58%) outperform IMADS (88.03%), suggesting that different architectures may excel under specific distributions and metrics; (ii) the AUC margins over VadCLIP/WSTD on UCF-Crime are small, and statistical significance should be corroborated with multiple random seeds and interval estimates; and (iii) the benefits of late fusion in IMADS depend to some extent on pretraining quality and the fusion weight search space—when one modality degrades markedly, overall gains may narrow. Nevertheless, considering the leading AP on XD-Violence, the comparable or slightly better AUC on UCF-Crime, and the best Ano-AUC among multimodal methods, IMADS exhibits effectiveness, balance, and robustness on mainstream coarse-grained evaluations.

Fine-grained VAD Results: Tables V and VI report mAP at multiple IoU thresholds on XD-Violence and UCF-Crime. Fine-grained detection is more demanding than coarse-grained detection because it requires precise localization and category-level discrimination; consequently, mAP consistently drops as the IoU threshold increases. On XD-Violence (Table V), IMADS attains the best AVG mAP (27.17%) among all compared methods: +2.47 pp over the strongest multimodal baseline VadCLIP

TABLE VII
ABLATION STUDY

Video Branch		Text Branch		XD-Violence	UCF-Crime
2D-VD	3D-VD	Bert	BoW	AP (%)	AUC (%)
√				36.64	36.89
	√			48.88	51.25
		√		54.65	65.56
			√	36.56	47.56
√		√		59.64	62.45
√			√	52.88	57.69
	√	√		68.48	70.49
	√		√	61.23	63.89
√	√			71.86	69.97
		√	√	66.23	69.44
√	√	√	√	**86.61**	**88.03**

(24.70%), +3.33 pp over CLIP-TSA (23.84%), and +2.27 pp over the strong unimodal GS-MoE (24.90%). Per-threshold, IMADS leads at IoU = 0.1, 0.3, 0.5, while CLIP-TSA slightly edges out at IoU = 0.2 and UMIL is marginally better at IoU = 0.4, indicating complementary strengths across regimes and IMADS's advantage in overall balance and high-threshold robustness.

On UCF-Crime (Table VI), IMADS achieves the top mAP at every IoU threshold (0.1–0.5) and the highest AVG mAP (7.38%), surpassing CLIP-TSA (6.72%) and RTFM (6.71%) by +0.66 and +0.67 pp, respectively. It is suggesting consistent generalization across datasets. Architecturally, prior multimodal works often rely on early fusion at the embedding level, which can cause modality imbalance and hamper convergence. IMADS employs modular, independent branches with late fusion and Bayesian weight optimization, enabling each modality to fully learn before adaptive decision-level integration. While there remains room to improve at certain thresholds, IMADS demonstrates leading or on-par overall performance with stronger stability, validating the effectiveness of the proposed framework.

D. Ablation Study

Experiments are conducted on the XD-Violence and UCF-Crime dataset with extensive ablation. Table VII presents the ablation study of IMADS on two different datasets. The analysis of the ablation study results leads to the following conclusions: Using only the Video Branch or Text Branch results in relatively low performance, indicating that relying solely on video or text information is insufficient for violence detection. In the Video Branch, the performance of 3D-VD is superior to that of 2D-VD, demonstrating that 3D convolutions can better capture the spatiotemporal information in videos. However, the experimental results also show that without transfer learning, directly training 3D-VD yields worse results compared to using transfer learning, with AP decreasing by 23.98% in the XD-Violence dataset and AUC decreasing by 18.72% in the UCF-Crime dataset. This is because the noise in the video data affects the ability of the 3D-VD module to extract information in multiple dimensions. Transfer learning, on the other hand, by first training on 2D-VD and then transferring to 3D-VD, allows 3D-VD training to focus more on the temporal and spatial dimensions. Similarly, in the Text Branch, the performance of BERT is superior to that of BoW, indicating that BERT can better understand text semantics and extract effective textual features. Furthermore, when the decisions of both branches are weighted, the decision scores improve. The experimental results show that using both the Video Branch and the Text Branch together significantly outperforms using either branch alone, indicating that video and text information are complementary, and their fusion can significantly enhance violence detection performance. Using both 2D-VD and 3D-VD in the Video Branch, or both BERT and BoW in the Text Branch, further improves performance, indicating that there is also complementarity between different model components. When all model components are used together, the best performance is achieved on both datasets, with an AP of 86.61% and an AUC of 88.03% , indicating that this fusion approach can maximize the use of information from different modalities, thereby improving the accuracy of anomaly behavior detection.

E. Bayesian Optimization and Interpretable Analysis

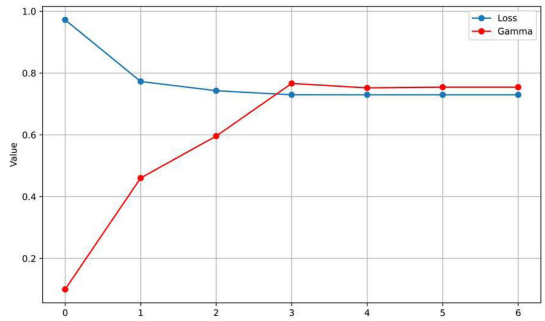
Fig. 4 aims to demonstrate the process and results of finding the optimal weights γ^* and α^* for the text branch and late fusion of the proposed IMADS after Bayesian optimization on two different datasets. The x-axis denotes the number of iterations during Bayesian optimization, while the y-axis represents the values of the weight parameters and the loss. Fig. 4 illustrates how Bayesian optimization determines the optimal weights γ^* and α^* for the text branch and late fusion after multiple iterations and convergence of the loss function.

Specifically, For the text branch, on the XD-Violence dataset, the optimal weight $\gamma^* \approx 0.7543$ indicates that the branch's decisions rely more on $s_{BERT}(t_j)$, whereas on the UCF-Crime dataset, the optimal weight $\gamma^* \approx 0.4821$ suggests that the decisions lean more toward $s_{BoW}(t_j)$.

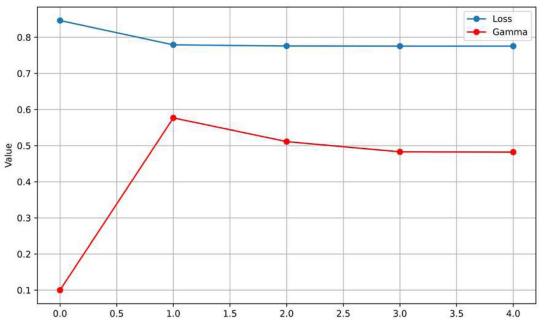
For late fusion, on the XD-Violence dataset, the optimal weight $\alpha^* \approx 0.6567$ shows that the IMADS results are more influenced by the text branch's decisions. Conversely, on the UCF-Crime dataset, the optimal weight $\alpha^* \approx 0.4116$ demonstrates that the results are more focused on the 3D-VD branch. Through this optimization process, IMADS achieves an effective balance between the textual and visual branches, thereby attaining stronger model performance across datasets.

This interesting phenomenon shows that the weight settings differ across different datasets. More precisely, Bayesian optimization can find the best weight distribution points for different data distributions across various datasets. To explain and visualize this phenomenon, the experiments used LIME [18] and CAM-based [20] approaches to perform interpretable and visual analyses on each module in different branches of the IMADS.

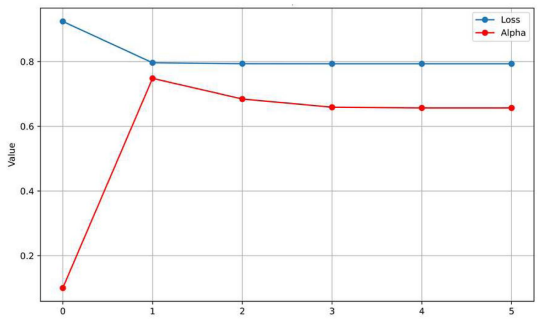
For the text branch, the experiments used LIME [18] to interpret the text content. To ensure interpretability stability and consistency, the experiments followed the approach in [19], setting the kernel function in LIME to 0.25. The interpretable texts were selected from the data in each test set. To maintain interpretative fairness, the original linear regression in LIME was replaced with a tree model to avoid collinearity among attributes.



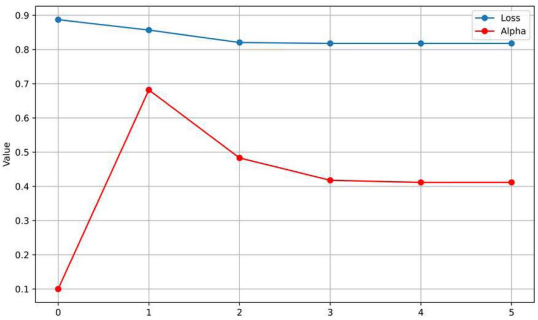
(a) optimal weight γ^*=0.7543061551133637 in XD-Violence



(b) optimal weight γ^*=0.48205059308642356 in UCF-Crime



(c) optimal weight α^*=0.6567825135511175 in XD-Violence



(d) optimal weight α^*=0.41157187316060234in UCF-Crime

Fig. 4. The process and results of obtaining the optimal weights γ^* and α^* for the text branch and late fusion through Bayesian optimization on two datasets.

For the video branch, the experiments used a CAM-based method [20] to visualize the image content. The CAM is a heatmap-based interpretability method that visualizes the model's decision process, helping users understand the decisions the model makes based on the data.

Fig. 5 shows the interpretable visualization of two modules in the text branch using LIME. The Fig. contains four subplots. Subplots (a) and (b) are visualizations of the BoW and BERT modules in XD-Violence, respectively. Subplots (c) and (d) are visualizations of the BoW and BERT modules in UCF-Crime, respectively.

According to subplots (a) and (b), the text data in the XD-Violence dataset is obtained through speech-to-text conversion, with most of the sentences being emotional. As shown in the subplots (a), the interpretable visualizations reveal that the decisions of the BoW model precisely focus on certain adjectives or nouns, such as "garbage" and "mental". This indicates that the training of the BoW model meets the expected standards, enabling it to make relevant decisions and judgments based on emotional content.

Furthermore, in subplot (b), the interpretable results show that the pre-trained language model better integrates contextual and semantic information in the experimental results, but its decision scores are not as precise as those of the BoW model. This suggests that while the pre-trained model excels in understanding the overall semantics and context of the text, it is not as accurate as the BoW model in specific emotional judgments and detail handling.

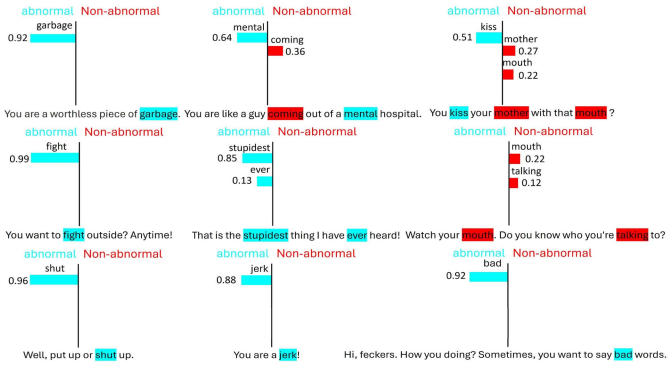
Compared to the experimental results and the selection of γ^* in Fig. 4, it can be seen that the weight distribution in XD-Violence, which prioritizes the pre-trained model and supplements it with the BoW model, is very reasonable.

According to subplots (c) and (d), the text data in the UCF-Crime dataset is obtained through CLIP, with most of it describing scenes and behaviors. The interpretable visualizations reveal that in subplot (c), the decisions of the BoW model precisely focus on words used to describe events and scenes, such as "fighting" and "violent." This indicates that the training of the BoW model on UCF-Crime meets the expected standards, allowing it to make decisions and judgments that align with user expectations.

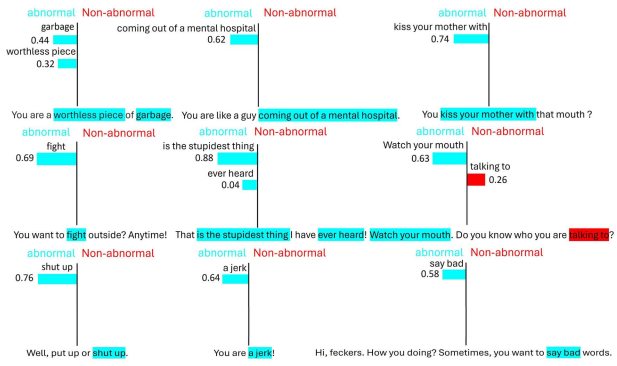
Furthermore, in subplot (d), the interpretable results show that the pre-trained language model integrates context and semantics in the experimental results, but its decision scores are not as clear as those of the BoW model. This suggests that while the pre-trained model excels in understanding overall semantics and contextual connections, it is not as precise as the BoW model in describing specific events and making detailed judgments. Compared to the selection of γ^* in the experimental results of Fig. 4, it can be seen that the weight distribution in UCF-Crime, prioritizing the BoW model and supplementing it with the pre-trained model, is very reasonable.

Fig. 6 shows the visualization results of 3D-VD on two different datasets using CAM heatmaps with two simple examples. As shown in Fig. 6, the judgments of 3D-VD in these two anomaly scenes focus on violent behavior, indicating that the 3D-VD model can effectively capture and emphasize key areas related to violence in the images, thereby improving the model's ability to detect abnormal behavior.

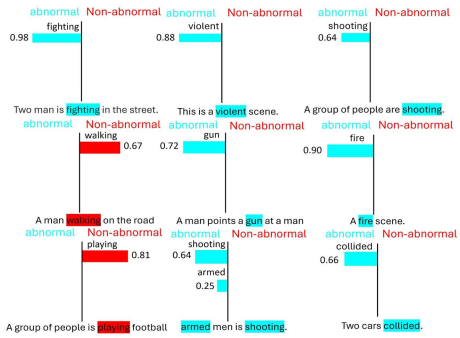
Specifically, in the example from the XD-Violence dataset, the model's focus is on the chaotic scenes in the crowd, while in the example from the UCF-Crime dataset, the model's focus is



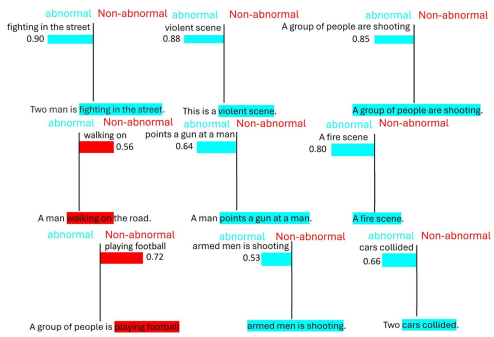
(a) Some examples of interpretable visualizations of the Bow model using a LIME-based approach on the XD-Violence dataset.



(b) Some examples of interpretable visualizations of the Bert model using a LIME-based approach on the XD-Violence dataset.



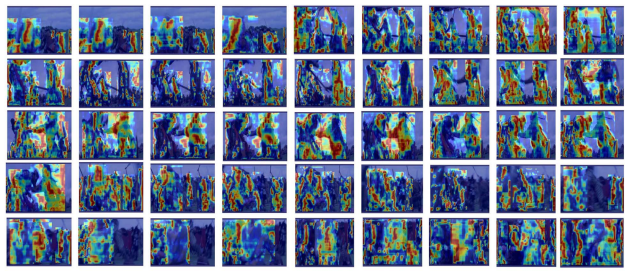
(c) Some examples of interpretable visualizations of the Bow model using a LIME-based approach on the UCF-Crime dataset.



(d) Some examples of interpretable visualizations of the Bert model using a LIME-based approach on the UCF-Crime dataset.

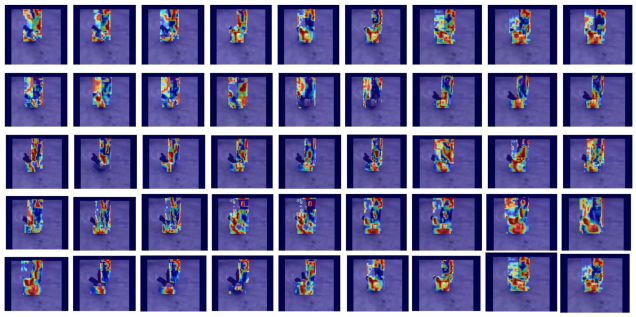
Fig. 5. LIME Interpretable Visualization.

(a) The examples of the XD-Violence dataset.



(b) The interpretable visualizations of the 3D-VD model using a CAM-based method approach on the XD-Violence dataset.

(c) The examples of the XD-Violence dataset.



(d) The interpretable visualizations of the 3D-VD model using a CAM-based method approach on the XD-Violence dataset.

Fig. 6. The CAM-based Interpretable Visualization.

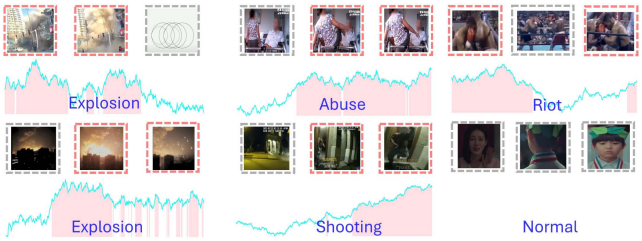


Fig. 7. Qualitative results of coarse-grained in XD-Violence.

on areas where fighting occurs. These results not only validate the stability and reliability of the 3D-VD model in different scenes but also demonstrate its strong generalization ability, maintaining high accuracy in various violent scenarios.

Similarly, combining the textually interpretable cases in Fig. 5, compared to the weight α^*optimized by Bayesian optimization, it is evident that α^* has a greater weight in the video branch due to the complexity of the text content in the XD-Violence dataset. In the UCF-Crime case, because of higher recognizability of the text content, the weight α^* in the video branch is slightly lower than in the text branch.

This also reflects the flexibility and effectiveness of Bayesian optimization in weight allocation.

Experimental results also prove that the Bayesian optimization method, by adaptively adjusting the weights of the video and text branches, enables the model to fully utilize the advantages of their respective features across different datasets and scenarios, thereby enhancing overall detection performance. At the same time, this optimization approach has strong interpretability. Specifically, through the adjustment of weights, users can clearly understand the model's emphasis on video and text features in different scenarios, which not only improves the model's performance but also enhances its transparency and credibility.

F. Qualitative Results

Fig. 7 aims to demonstrate the qualitative results of the proposed IMADS for video anomaly detection on the XD-Violence dataset. The Fig. includes various types of anomaly examples and their corresponding anomaly prediction curves. Each type of anomaly (e.g., explosions, abuse, riots, shootings, and normal) includes several example video frames. In these frames, red boxes indicate the areas where the model predicts anomalies. Below each set of anomaly-type frames, there is a blue anomaly prediction curve. The pink background areas correspond to the actual anomaly periods. When the blue curve rises above the average value, the area is marked with a pink background, indicating that the model detects anomalies at those time points.

As shown in Fig. 7, the experimental results indicate that the proposed IMADS model can accurately predict anomaly events in most cases while maintaining a low false positive rate for normal videos. The low false positive rate of the IMADS is due to its ability to effectively distinguish between normal and abnormal patterns by leveraging visual and textual features. Similarly, the Bayesian-optimized late fusion allows the model to achieve optimal balance between different modalities, reducing the likelihood of misclassifying normal events as anomalies.

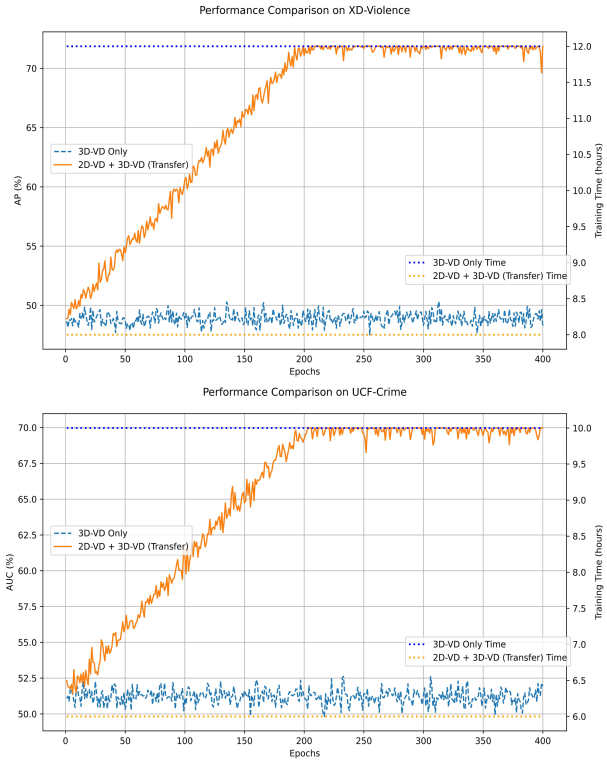


Fig. 8. Transfer learning efficiency test.

G. Transfer Learning Efficiency Test

Fig. 8 shows the comparison of performance and training time between 2D-VD transferred to 3D-VD and standalone 3D-VD on the UCF-Crime and XD-Violence datasets. As shown in Fig. 8, the accuracy of standalone 3D-VD does not significantly improve with an increasing number of training epochs. In contrast, the performance of the 2D-VD transferred to 3D-VD method significantly improves as the number of training epochs increases. Additionally, as indicated by the right Y-axis, the 2D-VD + 3D-VD (Transfer) method significantly reduces the training time. On the XD-Violence dataset, the training time is reduced from 12 hours to 8 hours; on the UCF-Crime dataset, the training time is reduced from 10 hours to 6 hours. This demonstrates that transfer learning not only effectively accelerates the training process but also improves model performance.

VI. Conclusion

This paper introduced IMADS, an interpretable multimodal framework for video anomaly detection that couples text and video branches and resolves cross-modal decisions through Bayesian-optimized late fusion. On the video side, we adopt a transfer strategy that pretrains on frame-level 2D-VD and then fine-tunes only the temporal weights on video-level 3D-VD, yielding long-range temporal sensitivity at modest computational cost. A comprehensive suite of experiments and ablations shows that IMADS delivers competitive accuracy, retains efficient inference, and provides interpretation-friendly evidence, while remaining stable across both coarse- and fine-grained tasks.

At the same time, we acknowledge several technical and practical boundaries. Although IMADS achieves state-of-the-art or near-state-of-the-art performance on key metrics across XD-Violence and UCF-Crime, we observe that, under certain evaluation settings, unimodal architectures such as GS-MoE can surpass multimodal late fusion on UCF-Crime AUC, underscoring that architectural advantages depend on data distribution and metric choice. Moreover, the AUC margins over strong multimodal baselines on UCF-Crime are modest and warrant confirmation via multiple random seeds and interval estimation. The benefits of late fusion are also contingent on the quality of single-modality pretraining and the fusion search space; when one modality deteriorates substantially due to noise, occlusion, or semantic sparsity, overall gains may diminish. While temporal-only fine-tuning improves efficiency and long-range modeling, its spatial adaptability under severe domain shift remains improvable. Finally, our interpretability evaluation relies primarily on intrinsic scores and visual rationales; human-subject studies and task-level assessments are still missing, and uncertainty estimation and calibration for open-set and OOD cases are not yet integrated into a unified treatment.

Looking forward, we will develop IMADS along three intertwined directions. First, at the modality scale, we aim to incorporate audio and trajectory cues, strengthen robustness to misalignment and missing modalities, and leverage open-vocabulary textual priors to reduce annotation cost. Second, at the learning mechanism level, we plan to make fusion adaptive and uncertainty-aware. It is extending Bayesian optimization to per-video or per-segment dynamic weighting, complemented by temperature calibration and confidence estimation for reliable open-world deployment. Third, at the system validation level, we will establish rigorous protocols for cross-domain transfer, test-time adaptation, and continual learning; on the edge, we will pursue distillation, pruning, quantization, and sliding-window inference to optimize latency–energy–accuracy jointly, while closing the loop on interpretability with human-subject studies and task-level impact metrics. In sum, IMADS provides an efficient, robust, and extensible blueprint for interpretable multimodal anomaly detection, and we expect these directions to further strengthen its reliability and usability in complex, open-world, real-time applications.

REFERENCES

[1] B. Scholkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 582–588, 1999, vol. 12.

[2] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6479–6488.

[3] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 733–742.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[6] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1237–1246.

[7] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 4955–4966.

[8] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 322–339.

[9] P. Wu and J. Liu, "Learning causal temporal relation and feature discrimination for anomaly detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3513–3527, 2021.

[10] X. Zhou, R. Girdhar, A. Joulin, P. Krahenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 350–368.

[11] H. K. Joo, K. Vo, K. Yamazaki, and N. Le, "CLIP-TSA: CLIP-assisted temporal self-attention for weakly-supervised video anomaly detection," in *Proc. IEEE Int. Conf. Image Process.*, 2023, pp. 3230–3234.

[12] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, pp. 3769–3777.

[13] W. Yu, Y. Liu, W. Hua, D. Jiang, B. Ren, and X. Bai, "Turning a CLIP model into a scene text detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 6978–6988.

[14] Y. Rao et al., "DenseCLIP: Language-guided dense prediction with context-aware prompting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18061–18070.

[15] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8022–8031.

[16] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, and P. Wang, "VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 6074–6082.

[17] C. Huang et al., "Weakly supervised video anomaly detection via self-guided temporal discriminative transformer," *IEEE Trans. Cybern.*, vol. 54, no. 5, pp. 3197–3210, May 2024.

[18] M. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? : Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 97–101.

[19] Z. Tan, Y. Tian, and J. Li, "GLIME: General, stable and local LIME explanation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp.36250–36277.

[20] C. Qiu, F. Jin, and Y. Zhang, "Empowering CAM-based methods with capability to generate fine-grained and high-faithfulness explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 4587–4595.

[21] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 5, pp. 726–742, Oct. 2021.

[22] H. Li, X. Shi, X. Zhu, S. Wang, and Z. Zhang, "FSNet: Dual interpretable graph convolutional network for Alzheimer's disease analysis," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 15–25, Feb. 2023.

[23] D. Lyu, F. Yang, H. Kwon, W. Dong, L. Yilmaz, and B. Liu, "TDM: Trustworthy decision-making via interpretability enhancement," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 3, pp. 450–461, Jun. 2022.

[24] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[25] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 6000–6010, 2017, vol. 30.

[26] Y. Zhang, S. Yang, W. Huang, C. -D. Wang, and H. Cai, "Unified representation learning for multi-view clustering by between/within view deep majorization," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 615–626, Feb. 2024.

[27] Z. Xiao et al., "Deep contrastive representation learning with self-distillation," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 3–15, Feb. 2024.

[28] F. Ming, W. Gong, L. Wang, and L. Gao, "Balancing convergence and diversity in objective and decision spaces for multimodal multi-objective optimization," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 2, pp. 474–486, Apr. 2023.

[29] S. Zhang, C. Yin, and Z. Yin, "Multimodal sentiment recognition with multi-task learning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 200–209, Feb. 2023.

[30] D. Zhu et al., "MTCAM: A novel weakly-supervised audio-visual saliency prediction model with multi-modal transformer," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 2, pp. 1756–1771, Apr. 2024.
[31] F. Pourpanah, C. P. Lim, A. Etemad, and Q. M. J. Wu, "An ensemble semi-supervised adaptive resonance theory model with explanation capability for pattern classification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 814–827, Feb. 2024.
[32] W. Liu, Q. He, Z. Li, and Y. Li, "Self-learning modeling in possibilistic model checking," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 264–278, Feb. 2024.
[33] G. D'Amicantonio et al., "Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV),*, Honolulu, HI, USA, Aug. 2025.
[34] Y. Zhang, E. Akdag, E. Bondarev, and P. H. N. De With, "MTFL: Multi-timescale feature learning for weakly-supervised anomaly detection in surveillance videos," *in Seventeenth Int. Conf. Mach. Vis. (ICMV 2024),*, vol. 13517, Oct. 2024, Art. no. 135170M.
[35] P. P. G. and P. S. Tamizharasan, "STVAD: A lightweight spatio–temporal attention network for video anomaly detection," *IEEE Trans. Comput. Social Syst.*, vol. 12, no. 6, pp. 5015–5030, Dec. 2025, doi: 10.1109/TCSS.2025.3584160.
[36] M. Zhang, J. Wang, Q. Qi, Z. Zhuang, H. Sun, and J. Liao, "Cognition guided video anomaly detection framework for surveillance services," *IEEE Trans. Serv. Comput.*, vol. 17, no. 5, pp. 2109–2123, Sep./Oct. 2024.
[37] Q. Chen, M. Zhao, Y. Ji, X. Luo, Y. Zhang, and Y. Zhang, "MGOD: Multi-granular outlier detection with clustlier analysis," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2024, pp. 3105–3110.
[38] H. Liu, S. Zhang, Z. Wu, and X. Li, "Outlier detection using local density and global structure," *Pattern Recognit.*, vol. 157, 2025, Art. no. 110947.
[39] Y. Su, E. Su, W. Wang, P. Jing, D. Ma, and F. L. Wang, "Adversarial neighbor perception network with feature distillation for anomaly detection," *Expert Syst. Appl.*, vol. 274, 2025, Art. no. 126911.

Chih-Yung Chang (Member, IEEE) received the Ph.D. degree in computer science and information engineering from National Central University, Taoyuan City, Taiwan, in 1995. He is currently a Distinguished Professor with the Department of Computer Science and Information Engineering and the Department of Artificial Intelligence, Tamkang University, New Taipei, Taiwan. His research interests include artificial intelligence, deep learning and machine learning, the Internet of Things, and wireless sensor networks.

He is currently the Co-Chair of ACM SIGMOBILE, Taiwan. He has been an Associate Guest Editor for several SCIindexed journals, including International Journal of Ad Hoc and Ubiquitous Computing since 2011, Journal of Applied Science and Engineering since 2018, International Journal of Distributed Sensor Networks from 2012 to 2014, IET Communications in 2011, Telecommunication Systems in 2010, Journal of Information Science and Engineering in 2008, and Journal of Internet Technology from 2004 to 2008.

Wen-Dong Jiang (Graduate Student Member, IEEE) received the B.S. degree from the Department of Multimedia and Game Science Development, Lunghwa University of Science and Technology, Taoyuan, Taiwan, in 2021, and the M.S. degree from the Department of Computer Science and Information Engineering, Ming Chuan University, Taoyuan, in 2023. He is currently working toward the Ph.D. degree in computer science and information engineering with Tamkang University, New Taipei City, Taiwan. His research interests include explainable machine learning and its applications in smart cities. He was the recipient of the Best Student Paper Award at WASN 2024 and DLT 2024. He is also a reviewer for ETASR.

I-Hsiung Chang received the Ph.D. degree in education from National Chengchi University, Taiwan, in 2012. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering, Tamkang University, Taiwan. His research interests cover the joint of education and information technologies in AI-IoT, robot, and big data

Diptendu Sinha Roy (Senior Member, IEEE) received the Ph.D. degree in engineering from the Birla Institute of Technology, Ranchi, India, in 2010. He was with the Department of Computer Science and Engineering, National Institute of Science and Technology, Berhampur, India. In 2016, he joined as an Associate Professor with the Department of Computer Science and Engineering, National Institute of Technology (NIT) Meghalaya, Shillong, India, where he has been working as the Chair since 2017. His research interests include software reliability, distributed and cloud computing, and the Internet of Things, specifically applications of artificial intelligence/machine learning for smart integrated systems.

He is a member of the IEEE Computer Society.