CMHKF: Cross-Modality Heterogeneous Knowledge Fusion for Weakly Supervised Video Anomaly Detection

Guohua Wang[1*], **Shengping Song**[1*], **Wuchun He**[1], **Yongsen Zheng**[2†]
[1]South China Agricultural University, China
[2]Nanyang Technological University, Singapore
wangguohua@scau.edu.cn, songshengping@stu.scau.edu.cn
hewuchun@stu.scau.edu.cn, yongsen.zheng@ntu.edu.sg

Abstract

Weakly supervised video anomaly detection (WSVAD) presents a challenging task focused on detecting frame-level anomalies using only video-level labels. However, existing methods focus mainly on visual modalities, neglecting rich multi-modality information. This paper proposes a novel framework, Cross-Modality Heterogeneous Knowledge Fusion (CMHKF), that integrates cross-modality knowledge from video, audio, and text to improve anomaly detection and localization. To achieve adaptive cross-modality heterogeneous knowledge learning, we designed two components: Cross-Modality Video-Text Knowledge Alignment (CVKA) and Audio Modality Feature Adaptive Extraction (AFAE). They extract and aggregate features by exploring inter-modality correlations. By leveraging abundant cross-modality knowledge, our approach improves the discrimination between normal and anomalous segments. Extensive experiments on XD-Violence show our method significantly enhances accuracy and robustness in both coarse-grained and fine-grained anomaly detection.

1 Introduction

Weakly supervised video anomaly detection (WSVAD) aims to use video-level labels (normal or abnormal) to evaluate frame-level anomaly scores, thereby reducing manual annotation costs. Most existing WSVAD methods (Feng et al., 2021; Cho et al., 2023; Karim et al., 2024) primarily distinguish between frame-level normal and abnormal events by learning information from a single modality, specifically the video modality. However, relying solely on video modality struggles to localize anomalies in ambiguous scenarios. Complementary modalities, such as audio and text, provide additional contextual cues that help disambiguate complex anomaly patterns. Thus, integrating multi-modality information is essential for WSVAD.

Current WSVAD methods consist of one-stage Multiple Instance Learning (MIL) methods (Karim et al., 2024; Sultani et al., 2018) and pseudo-label self-training two-stage methods (Li et al., 2022; Zhang et al., 2023). One-stage MIL methods use ranking loss to prioritize higher scores in abnormal segments. Nevertheless, they often miss minor anomalies, limiting detection completeness. Two-stage approaches use MIL for pseudo labels first, then train the classifier in the second stage. However, pseudo label accuracy is unreliable, potentially causing misdetections. Historically, most WSVAD methods relied on single-modality data. Recently, studies (Peng et al., 2023; Wu et al., 2024c) have attempted to leverage multi-modality information to enhance performance. However, they often superficially fuse multi-modality data without fully exploiting its potential.

Despite some progress, existing methods still face two major challenges: 1) **Single Modality.** Currently, most of both single-stage and two-stage methods (Lv et al., 2023; Zhang et al., 2023) focus on learning from video data, emphasizing conspicuous visual anomalies. However, this strategy often falls short in visually ambiguous scenarios where the video modality lacks the sufficient discriminative power. For example, in a dust-filled scene, video data might struggle to distinguish between an anomalous event, such as an explosion, and a normal occurrence like strong winds stirring up dust. In these instances, audio and text modalities can provide critical supplementary information. Therefore, how to effectively fuse multi-modality knowledge, including video, audio, and text, plays an important role in these visual ambiguities. This not only enhances detection robustness but also improves system performance in complex or dynamic environments. 2) **Fusion Strategy.** While a few studies (Wu et al., 2020, 2022b) have started to

[*]Both authors contributed equally to this research.
[†]Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31594–31607

harness the power of multi-modality knowledge, they predominantly rely on independently learning feature representations for each modality, achieving multi-modality fusion through straightforward feature concatenation. However, this concatenation fusion strategy only accomplishes a surface-level integration of modalities at the feature level, fundamentally overlooking the potential correlations between modalities. In contrast, the core principle behind adaptive cross-modality knowledge learning is to allow the model to autonomously adjust the interdependencies and contributions of each modality. By dynamically adapting these relationships based on the complexity of specific scenes and anomalous events, this method provides a multi-dimensional representation of anomalies, thus facilitating the learning of high-level features. Therefore, this more nuanced integration of cross-modal knowledge is essential for advancing WSVAD.

In this study, we propose CMHKF to address these challenges. To address the first challenge, we integrate video, audio, and text to jointly learn feature representations. This enhances the differentiation between normal and anomalous segments. To address the second challenge, we propose the Cross-Modality Video-Text Knowledge Alignment (CVKA) and the Audio Modality Feature Adaptive Extraction (AFAE). CVKA leverages CLIP to capture visual-semantic similarity, dynamically aligning and adaptively aggregating relevant text features. AFAE maps the visual-semantic similarity obtained from CVKA to the temporal saliency of audio features. To address temporal misalignment, AFAE uses a Top-k window to bridge the fine-grained feature distribution differences between video and audio. Finally, we propose Multi-Modality Knowledge Adaptive Fusion (MKAF) by extending Joint Cross-Attention Model (Praveen et al., 2022). MKAF effectively captures intra-modality and cross-modality correlations while reducing inter-modality heterogeneity.

Overall, our contributioons are threefold:

- We propose CMHKF to adaptively fuse cross-modality heterogeneous knowledge from video, audio, and text in WSVAD.
- We propose CVKA, AFAE, and MKAF for adaptive modality adjustment and fusion.
- Experiments on XD-Violence show our method outperforms others in coarse-grained and fine-grained WSVAD tasks.

2 Related Work

2.1 Weakly Supervised Video Anomaly Detection

Weakly supervised video anomaly detection (Cao et al., 2023; Majhi et al., 2024; AlMarri et al., 2024) has become a prominent research focus. Most WSVAD methods use MIL(Sultani et al., 2018) to learn features from weakly labeled data for frame-level detection. The pioneering work (Sultani et al., 2018) proposed a deep MIL model with ranking loss. Subsequently, Zhou et al. (Zhou et al., 2023) presented an Uncertainty Regulated Dual Memory Units model to improving the representation learning of normal and anomalous data. Since the aforementioned methods are constrained to single-modality anomaly detection, recent studies have proposed multi-modality approaches (Yang et al., 2024; Peng et al., 2023) that outperform their single-modality counterparts. However, these approaches primarily rely on basic multi-modality feature fusion, overlooking latent interdependencies among modalities. To address these limitations, we propose an adaptive cross-modal knowledge learning approach that integrates video, audio, and text modalities.

2.2 Vision-Text Multi-Modality Models

Vision-text multi-modality has become a key research area for tasks like pre-training (Li et al., 2021; Lei et al., 2021) and vision-text retrieval(Tian et al., 2024; Deng et al., 2023). One influential work is Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021), which uses two encoders to map images and text into a shared space for contrastive learning. Trained on hundreds of millions of image-text pairs, CLIP demonstrates remarkable zero-shot transfer capabilities. Recently, several studies (Wu et al., 2024c; Sun et al., 2024) explored CLIP's application in WSVAD. These works are categorized into two types. The first type (Joo et al., 2023; Sharif et al., 2023) uses CLIP's image encoder as a strong initialization for the video encoder. The second type (Zanella et al., 2024; Wu et al., 2024c) extends CLIP to video-label matching for anomaly detection. However, these studies only superficially exploit CLIP's embedded knowledge. In contrast, our research explores the relationship between visual and semantic features, adaptively fuses text information, and maps CLIP's image-text alignment capabilities to audio, deeply mining its temporal saliency regions.

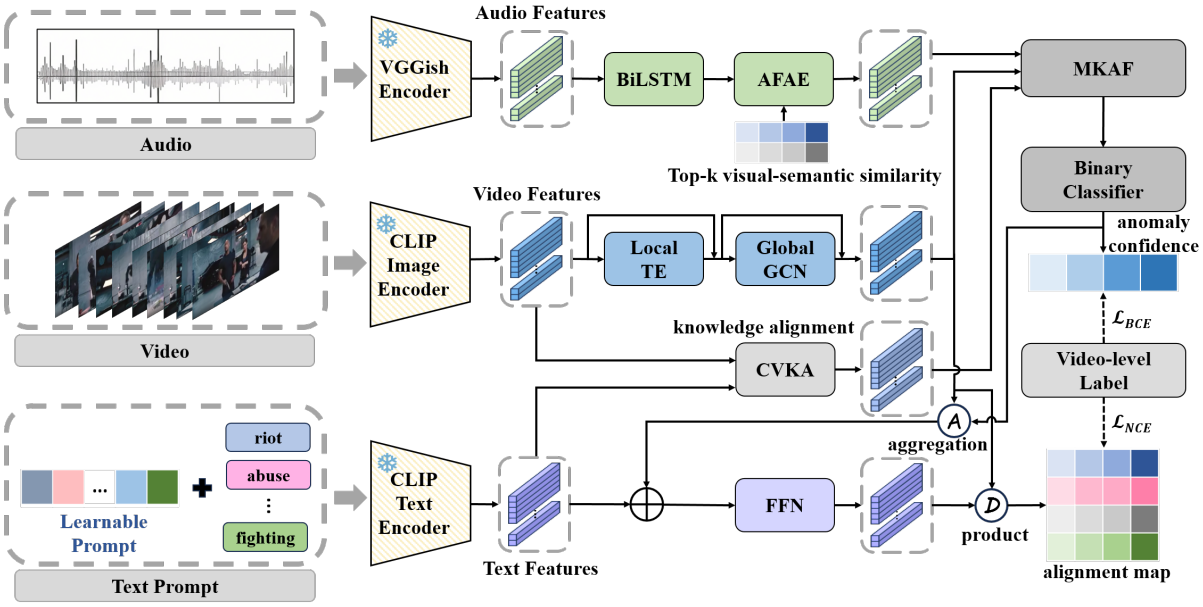


Figure 1: The overall architecture of our proposed CMHKF.

3 CMHKF

This section begins by defining the task of WSVAD. Following this, we describe the overall architecture of CMHKF. Finally, we provide a detailed explanation of each module and its implementation.

3.1 Problem Statement

Given a dataset of M videos and audios, $X = \{x_i\}_{i=1}^{M}$, $Z = \{z_i\}_{i=1}^{M}$, and video-level labels $Y = \{y_i\}_{i=1}^{M}$, where $y_i \in \{0, 1\}$. If $y_i = 0$, the video is normal, meaning all its frames are free of anomalous events. If $y_i = 1$, the video is anomalous, containing at least one anomalous frame. During training, only video-level labels are employed to supervise the model, whereas during testing, the model predicts frame-level anomaly confidence for precise temporal localization.

3.2 Overall Architecture

Figure 1 illustrates the overall workflow of CMHKF. Our method encodes videos, learnable class prompt texts, and audio into feature embeddings using CLIP's image and text encoders, along with the VGGish network (Hershey et al., 2017). This yields video features $F_v \in \mathcal{R}^{N\times D_v}$, text features $F_t \in \mathcal{R}^{C\times D_t}$ and audio features $F_a \in \mathcal{R}^{N\times D_a}$. D_v, D_t, and D_a represent the feature dimensions of video, text, and audio, while N and C represent the number of frames and class prompt texts, respectively. Given CLIP is pretrained on large-scale image-text pairs, it exhibits limitations in effectively modeling temporal dependencies. Inspired by (Wu et al., 2024c), we introduce Local Transformer Encoder (Local TE) and Global Graph Convolutional Network (Global GCN) to obtain temporally dependent video features $\tilde{F}_v \in \mathcal{R}^{N\times D_v}$. We propose the Cross-Modality Video-Text Knowledge Alignment (CVKA) to leverage the correlation between visual and textual modalities in the CLIP feature space. CVKA dynamically aligns and adaptively aggregates text features most semantically similar to the video, yielding text features $\tilde{F}_t \in \mathcal{R}^{N\times D_t}$. Concurrently, we employ Bidirectional Long Short-Term Memory (BiLSTM) (Huang et al., 2015) and Audio Modality Feature Adaptive Extraction (AFAE) to capture temporal saliency in audio features, yielding the audio feature $\tilde{F}_a \in \mathcal{R}^{N\times D_a}$. Subsequently, we fuse $\tilde{F}_v$, $\tilde{F}_t$, and $\tilde{F}_a$ using Multi-Modality Knowledge Adaptive Fusion (MKAF), resulting in the fused feature $F_{fused} \in \mathcal{R}^{N\times D_f}$. Finally, F_{fused} is processed by a Binary Classifier to generate frame-level anomaly confidence, enabling coarse-grained anomaly detection (i.e., classifying frames as either normal or anomalous). Additionally, by weighting the video features with frame-level anomaly confidence and fusing them with all category text features, we obtain the feature $\tilde{F}_{vt}$. Then, by calculating the similarity between $\tilde{F}_v$ and $\tilde{F}_{vt}$, we further achieve fine-grained anomaly classification (i.e., determining whether a video frame is normal or belongs to a specific type of anomaly).

3.3 Multi-Modality Heterogeneous Knowledge

In WSVAD, multi-modality heterogeneous knowledge integrates complementary information from video, audio, and text, each providing distinct insights into anomalous events. Specifically, the video modality offers spatial and temporal context for anomaly detection. Anomalies manifest through dynamic object interactions and scene transitions. However, when visual information is sparse or ambiguous, the audio modality provides a crucial complementary perspective. Unlike video, audio conveys auditory signals correspond to the environment and deliver explicit cues for anomalous events. Furthermore, the text modality enhances semantic understanding by associating category labels with video segments. By providing clear definitions and features for each anomaly type, the text modality ensures that the detection process accurately captures the true nature of anomalous events. Thus, by integrating knowledge from these modalities, the model gains a more comprehensive understanding of anomalous events.

3.4 Cross-Modality Knowledge Fusion for WSVAD

To effectively integrate multi-modality heterogeneous knowledge, we propose two key components: CVKA and AFAE. CVKA modulates text features based on the semantic alignment between video frames and text, thereby enhancing text-to-video relevance. AFAE identifies temporally salient regions in audio, focusing on the most anomalous intervals to extract relevant audio features. These modules refine the fusion of cross-modality knowledge in our CMHKF framework.

Cross-Modality Video-Text Knowledge Alignment. This section discusses the dynamic alignment and adaptive fusion text features based on the visual-textual semantic correlation, which enhances the distinction between normal and anomalous segments. CVKA is illustrated in Figure 2.

To obtain the text features that are most semantically relevant to the video features, we employ text features as queries to precisely capture the correlation between text and video frames. Specifically, we encode the video frames and class label texts using the pretrained CLIP model, generating video frame features $F_v = \{v_n\}_{n=1}^{N}$ and text features $F_t = \{t_c\}_{c=1}^{C}$. We compute the cosine similarity between each text and frame feature to create a visual-semantic similarity matrix $s_{ij} \in \mathcal{R}^{N\times C}$, capturing fine-grained correlations. Mathematically, this process can be expressed as:

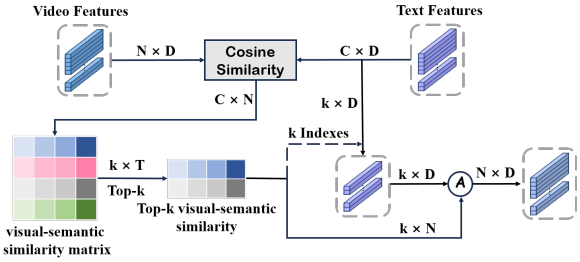


Figure 2: Illustration of Cross-Modality Video-Text Knowledge Alignment (CVKA). CVKA computes the visual-semantic similarity matrix between video features F_v and text features F_t. It then applies the Top-k algorithm to dynamically select k texts and their corresponding visual-semantic similarities. Finally, it adaptively aggregates the selected text features based on their Top-k visual-semantic similarity.

$$s_{ij} = \frac{v_i t_j^\top}{\|v_i\|\|t_j\|} \tag{1}$$

where v_i represents the feature of the i-th frame and t_j represents the text feature of the j-th class.

Next, we aggregate the similarity matrix along the temporal dimension N to obtain the similarity score for each class text, resulting in the corresponding video-level similarity vector $S_j \in \mathcal{R}^C$, which represents the global relationship between feature of the j-th class and the video. Subsequently, we apply the Top-k algorithm to dynamically select the top k category texts most relevant to the video semantics, forming an index set $\mathcal{K}$. Top-k visual-semantic similarity can be obtained based on $\mathcal{K}$. The formulas are as follows:

$$S_j = \sum_{i=1}^{N} s_{ij}, \quad j \in \{1, 2, \ldots, C\} \tag{2}$$

$$\mathcal{K} = \arg\operatorname{top}_k(S) = \{c_1, c_2, \ldots, c_k\} \tag{3}$$

Finally, we perform adaptive aggregation of the selected k class texts based on the Top-k visual-semantic similarity to achieve video-text knowledge alignment:

$$\begin{aligned} F_i &= \frac{1}{k} \sum_{j\in\mathcal{K}} \frac{\exp(s_{ij}/\tau)}{\sum_{i=1}^{N} \exp(s_{ij}/\tau)} t_j \\ \tilde{F}_t &= [F_1; F_2; \ldots; F_N] \in \mathcal{R}^{N\times D_t} \end{aligned} \tag{4}$$

where τ is a temperature parameter that controls the distribution of similarity scores.

Audio Modality Feature Adaptive Extraction. In the CVKA module, we performed semantic alignment between video and text along the temporal dimension. Next, we leverage video-text semantic relevance to capture the temporal saliency of audio features, mitigating audio noise interference. AFAE is shown in Figure 3. Leveraging the robust image-text alignment capabilities within the CLIP feature space, we project the Top-k visual-semantic similarity from CVKA onto the audio's temporal saliency. The projection method follows:

$$S_a^{(n)} = \frac{1}{k} \sum_{j \in \mathcal{K}} \frac{exp(s_{nj}/\tau)}{\sum_{i=1}^{N} exp(s_{ij}/\tau)}, n \in \{1, 2, ..., N\} \tag{5}$$

where $S_a^{(n)}$ represents the temporal saliency of the n-th audio frame. However, Top-k visual-semantic similarity follows the video's temporal sequence. Fine-grained misalignment between video and audio may lead to loss of crucial information if only temporal saliency is used. Moreover, preserving the continuity of anomalous events is essential. To address these challenges, we propose a Top-k window mechanism to mitigate the fine-grained temporal discrepancies between video and audio, ensuring more accurate extraction and retention of key event information. Specifically, we select the top k salient frames from the audio, forming the index set $K = \{K_1, K_2, \ldots, K_k\}$. k is chosen as $\lfloor T/16 \rfloor + 1)$. For each a_j $(j \in K)$, we define a temporal window of size $2w + 1$ around a_j. This window encompasses a_j and its w preceding and succeeding frames, forming the local region $W_j = \{a_j - w, \ldots, a_j, \ldots, a_j + w\}$. We then extend the temporal saliency of the central frame a_j to the other frames within the window, forming a continuous and accurate temporal saliency region. Mathematically, these can be formulated as:

$$\widetilde{S}_a^{(n)} = \begin{cases} S_a^{(j)} + S_a^{(n)}, & \text{if } \exists j \in K \text{ and } n \in R_j \\ S_a^{(n)}, & \text{otherwise} \end{cases} \tag{6}$$

where $R_j \in [\max(1, a_j - w), \min(a_j + w, N)]$. Finally, we aggregate the audio features using $\widetilde{S}_a^{(n)}$ as follows:

$$\tilde{F}_a = \sum_{n=1}^{N} F_a \widetilde{S}_a^{(n)} \tag{7}$$

where $\tilde{F}_a$ represents the audio features extracted by adaptively focusing on temporal saliency regions.

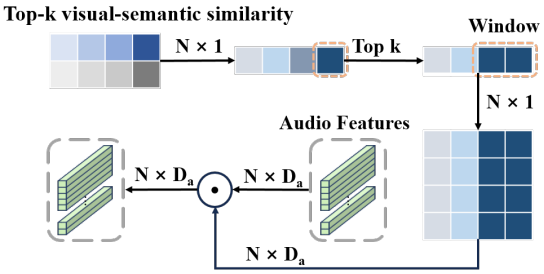


Figure 3: Illustration of the Audio Modality Feature Adaptive Extraction (AFAE). The AFAE is introduced to process audio features. Initially, AFAE projects the Top-k visual-semantic similarity onto the temporal saliency of the audio. This is followed by the application of the Top-k windowing mechanism to mine temporal saliency regions. Finally, the temporal saliency regions are aggregated with the audio features to yield enhanced audio features.

Multi-Modality Knowledge Adaptive Fusion. After obtaining cross-modal knowledge from video, text, and audio, we need to effectively fuse them for WSVAD.

At the outset, we merge the video features $\tilde{F}_v$, text features $\tilde{F}_t$, and audio features $\tilde{F}_a$ to construct the multi-modality joint representation $J \in \mathcal{R}^{N \times D}$. Subsequently, each modality feature $\tilde{F}_v$, $\tilde{F}_t$, and $\tilde{F}_a$ undergoes individual processing with J through a cross-attention mechanism, followed by a feed-forward layer to derive the interacted modality representations F_v^{int}, F_t^{int}, and F_a^{int}. These interacted features are concatenated and then added with J, and collectively passed through a linear projection layer to yield the ultimate fused multi-modality representation $F_{\text{fused}} \in \mathcal{R}^{N \times D_f}$. This holistic methodology effectively amalgamates the distinct modality features to form a cohesive representation that encapsulates the diverse multi-modality information inherent in the dataset.

3.5 Objective Function

For coarse-grained binary classification, we use a Binary Classifier to project F_{fused} into category space, yielding a frame-level anomaly score. Following (Wu et al., 2022b), we average the top k anomaly scores for the video-level prediction p_c. We then compute the binary cross entropy between p_c and ground-truth for classification loss $\mathcal{L}_{BCE}$.

For fine-grained multi-class classification, we introduce the MIL-Align mechanism (Wu et al., 2024c). We compute the alignment map A by evaluating the similarity between video frame features and all category text features. The top k similar-

ity scores for each category are averaged, yielding the vector $M = \{m_1, \ldots, m_C\}$. We then perform multi-class prediction to ensure that the similarity between the video and the correct text surpasses that of the incorrect texts. First, the procedure for computing the multi-class prediction is as follows:

$$p_f^i = \frac{exp\,(m_i/\tau)}{\sum_{j=1}^{C} exp\,(m_j/\tau)} \tag{8}$$

Next, we compute the binary cross entropy between the video prediction p_f^i and ground-truth to obtain the loss $\mathcal{L}_{NCE}$.

In addition, we introduce a contrastive loss $\mathcal{L}_{NA}$ based on normal and abnormal categories to distinguish between normal and abnormal text features. We propose a contrastive loss $\mathcal{L}_{AA}$ based on abnormal categories to differentiate between features of different abnormal classes. The two loss functions are defined as follows:

$$\mathcal{L}_{NA} = \frac{1}{C-1} \sum_{i=2}^{C} \left(1 + \frac{t_n \cdot t_{a_i}}{\|t_n\| \|t_{a_i}\|}\right) \tag{9}$$

$$\mathcal{L}_{AA} = \frac{2}{(C-1)(C-2)} \sum_{i=2}^{C-1} \sum_{j=i+1}^{C} \left| \frac{t_{a_i} \cdot t_{a_j}}{\|t_{a_i}\| \|t_{a_j}\|} \right| \tag{10}$$

where $i = 1$ corresponds to the text feature t_n of the normal category, and t_{a_j} represents the text feature for the j-th anomalous category.

Overall, objective function $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{NCE} + \lambda_1 \mathcal{L}_{NA} + \lambda_2 \mathcal{L}_{AA} \tag{11}$$

4 Experiments and Results

4.1 Dataset and Evaluation Metric

XD-Violence (Wu et al., 2020) is widely used in WSVAD and is the only dataset that includes both visual and auditory modalities. It consists of 4,757 untrimmed videos (totaling 217 hours) from real-world domains, including films, sports, online platforms, and surveillance, featuring six types of violent incidents. The dataset poses challenges due to its rich artistic content, such as perspective shifts and dynamic camera movements. Previous approaches used datasets (Sultani et al., 2018; Liu et al., 2018) as benchmarks, but these unimodal datasets are inadequate for evaluating cross-modal interactions.

Table 1: Coarse-grained comparisons on XD-Violence. Best result is **bolded** and second best result is underlined. * indicates re-implemented by fusing audio and visual features as inputs.

Method	Publication	Modality	AP (%)
Unsupervised learning based methods			
SVM baseline	NIPS'99	Video	50.78
Hasan et al. (Hasan et al., 2016)	CVPR'16	Video	30.77
GODS(Wang and Cherian, 2019)	ICCV'19	Video	61.56
CLAP (Al-Lahham et al., 2024)	CVPR'24	Video	77.65
Weakly supervised learning based methods			
Sultani et al. (Sultani et al., 2018)	CVPR'18	Video	73.20
HL-Net(Wu et al., 2020)	ECCV'20	Video + Audio	78.64
RTFM (Tian et al., 2021)	ICCV'21	Video	77.81
RTFM* (Tian et al., 2021)	ICCV'21	Video + Audio	78.10
Li et al. (Li et al., 2022)	AAAI'22	Video	78.28
S3R (Wu et al., 2022a)	ECCV'22	Video	80.26
MACIL-SD (Yu et al., 2022)	ACMMM'22	Video + Audio	83.40
TEVAD (Chen et al., 2023a)	CVPR'23	Video + Text	79.80
Mgfn (Chen et al., 2023b)	AAAI'23	Video	80.11
Cho et al. (Cho et al., 2023)	CVPR'23	Video	81.30
UR-DMU (Zhou et al., 2023)	AAAI'23	Video	81.66
UR-DMU* (Zhou et al., 2023)	AAAI'23	Video + Audio	81.77
Zhang et al. (Zhang et al., 2023)	CVPR'23	Video + Audio	81.43
REWARD (Karim et al., 2024)	WACV'24	Video	77.71
Wu et al. (Wu et al., 2024a)	CVPR'24	Video + Text	76.03
TPWNG (Yang et al., 2024)	CVPR'24	Video + Text	83.68
VadCLIP (Wu et al., 2024c)	AAAI'24	Video + Text	84.51
Ours(light)	—	Video + Audio + Text	84.65
Ours(full)	—	**Video + Audio + Text**	**86.57**

Table 2: Fine-grained comparisons on XD-Violence. Best result is **bolded** and second best result is underlined.

Method	mAP@IoU (%)					
	0.1	0.2	0.3	0.4	0.5	AVG
Random Baseline	0.37	0.27	0.06	0.03	0.01	0.15
Sultani et al. (Sultani et al., 2018)	20.08	13.72	8.44	5.06	2.81	10.02
3C-Net (Narayan et al., 2019)	23.77	17.78	11.90	8.28	5.87	13.52
W-TALC (Paul et al., 2018)	26.27	18.87	13.83	9.50	6.55	15.00
Wu et al. (Wu et al., 2022b)	35.35	28.02	20.94	15.01	10.33	21.93
VadCLIP (Wu et al., 2024c)	37.03	30.84	23.38	17.90	14.31	24.70
Ours(light)	37.74	29.98	24.09	18.67	13.84	24.86
Ours(full)	**38.31**	**32.12**	**25.68**	**20.93**	**16.47**	**26.70**

Evaluation Metrics. We follow standardized protocols for fair comparisons. For coarse-grained WSVAD, we use frame-level Average Precision (AP) as the evaluation metric (Chen et al., 2023b; Tan et al., 2024). For fine-grained WSVAD, we use mean Average Precision (mAP) under different Intersection over Union (IoU) thresholds, following video action detection protocols (Wu et al., 2023).

4.2 Implementation Details

In our network architecture, the image and text encoders use pre-trained CLIP (ViT-B/16)(Radford et al., 2021), and VGGish for audio features(Hershey et al., 2017). Feature dimension D is 512. Hyperparameters include $k = 2$ (Eq. 3), $w = 2$ (Eq. 6), $\tau = 0.07$ (Eq. 8), $\lambda_1 = 1 \times 10^{-3}$, and $\lambda_2 = 1 \times 10^{-4}$ (Eq. 11). Video sequences are capped at 256 frames during training. The model was trained using PyTorch on an NVIDIA RTX 3090 GPU, with the AdamW optimizer for 10 epochs and learning rate of 1×10^{-5}.

4.3 Comparison with State-of-the-Art Methods

Our method addresses both coarse-grained and fine-grained WSVAD tasks. We present its performance and compare it with state-of-the-art methods on both WSVAD tasks.

Coarse-grained WSVAD Results. We compare our method with state-of-the-art methods and present AP results on XD-Violence (Table 1). Our method achieves an AP of 86.57%, surpassing the best unsupervised method (Al-Lahham et al., 2024) by 8.92% and previous single-modality methods (Zhou et al., 2023) by 4.91%. Furthermore, compared to recent multi-modality weakly supervised methods, our approach outperforms the best-performing visual-audio method (Yu et al., 2022) by 3.17%, and exceeds the best video-text method (Wu et al., 2024c) by 2.06%. To further validate the effectiveness of our CVKA and AFAE, we performed light fusion of audio, video, and text features while retaining the structure of other modules. Without CVKA, we input all text features to demonstrate the performance of simple multi-modality integration, resulting in a lower AP of 84.65%. Results indicate that superior performance arises not only from the richer combination of audio, video, and text modalities but also from CVKA and AFAE facilitating modality-aware interactions for more accurate detection. In conclusion, our method outperforms both identical input modality and other multimodal approaches, validating its effectiveness.

Fine-grained WSVAD Results. Table 2 presents a comparison between CMHKF and other methods on the fine-grained task. As the data indicate, fine-grained WSVAD is evidently more challenging than coarse-grained WSVAD, as it requires considering both multi-class classification accuracy and the continuity of detected segments. From Table 2, it can be observed that our method significantly outperforms previous works on the XD-Violence dataset, achieving an AVG mAP of 26.70%, which is 2.00% higher than the best performance of previous works (24.70%). Similarly, we performed light fusion of the audio, video, and text features, resulting in a lower AVG mAP of 24.86%. This further demonstrates the effectiveness of our method in adaptive cross-modal knowledge learning, optimizing the utilization of multi-modality data for WSVAD tasks.

Table 3: Effectiveness of the CVKA. Best result is **bolded** and second best result is underlined.

Method	k							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
w/o CVKA@AP(%)	83.65	—	—	—	—	—	—	—
w/ CVKA@AP(%)	—	85.26	**86.57**	86.24	85.86	85.72	85.49	85.41
w/o CVKA@AVG mAP(%)	22.49	—	—	—	—	—	—	—
w/ CVKA@AVG mAP(%)	—	25.48	**26.70**	26.13	25.74	25.45	25.37	25.19

Table 4: Effectiveness of the AFAE. Best result is **bolded** and second best result is underlined.

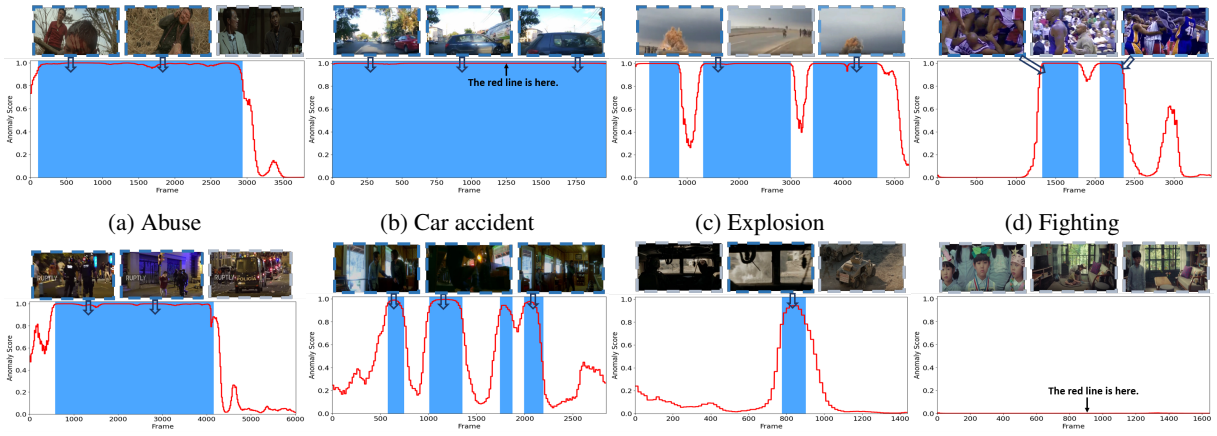
Method	w/o Top-k window	w/ Top-k window	AP (%)	AVG mAP (%)
w/o AFAE	✓		85.06	25.15
w/ AFAE	✓		85.81	26.03
w/ AFAE		✓	**86.57**	**26.70**

4.4 Ablation Study

Next, we conduct a series of ablation studies to investigate the contributions of each component.

Effectiveness of the CVKA. We evaluated CVKA's impact on performance (Table 3). Initially, without CVKA (i.e., $k = 0$), text features were not integrated. The model achieved only 83.65% AP and 22.49% AVG mAP. After integrating CVKA and fusing all text categories ($k = 7$), model achieved 85.41% AP and 25.19% AVG mAP. This shows that semantically clear textual features complement WSVAD. Furthermore, experiments with k values ($k = 0, 1, 2, 3, 4, 5, 6, 7$) represented fusion of the top k semantically relevant text categories. When $k = 2$, the model achieved the highest AP of 86.57% and AVG mAP of 26.70%. This confirms that selectively processing specific text features in CVKA reduces interference from irrelevant texts, aiding the model in capturing the intrinsic characteristics of anomalous events.

Effectiveness of AFAE. To demonstrate AFAE's effectiveness on audio features, we conducted an ablation study (Table 4). Results show that without AFAE, baseline method achieved 85.06% AP and 25.15% mAP. As described in Section 2.1, AFAE uses a Top-k window mechanism to mine audio temporal saliency regions. When AFAE is introduced with only visual-semantic similarity mapping for weighting audio features, excluding the Top-k window mechanism, yields 85.81% AP and 26.03% AVG mAP. Complete AFAE delivers the best performance. This demonstrates the effectiveness of focusing on temporal saliency regions for audio features, generating more compact and beneficial features. Additionally, we tested different window sizes in the Top-k mechanism (Table 5). Window sizes W were set to 1, 3, 5, 7, 9, and 11. Optimal performance was achieved at $W = 5$.



(a) Abuse (b) Car accident (c) Explosion (d) Fighting

(e) Riot (f) Shooting (g) Explosion (h) Normal

Figure 4: The qualitative results of our method on the test videos of the XD-Violence dataset. The red line indicates anomaly score distribution. Blue shaded regions mark anomalous event segments, with anomaly scores on the Y-axis and video frame numbers on the X-axis. Frames with blue borders highlight anomalies, while gray-bordered frames indicate normal segments.

Table 5: The AP and mAP results of CMHKF under different window sizes W.

Method	W					
	1	**3**	**5**	**7**	**9**	**11**
AP(%)	85.97	86.22	**86.57**	86.23	86.11	85.73
AVG mAP(%)	25.67	26.41	**26.70**	26.03	25.87	25.78

Table 6: Effectiveness of the Multi-Modality. Best result is **bolded** and second best result is underlined.

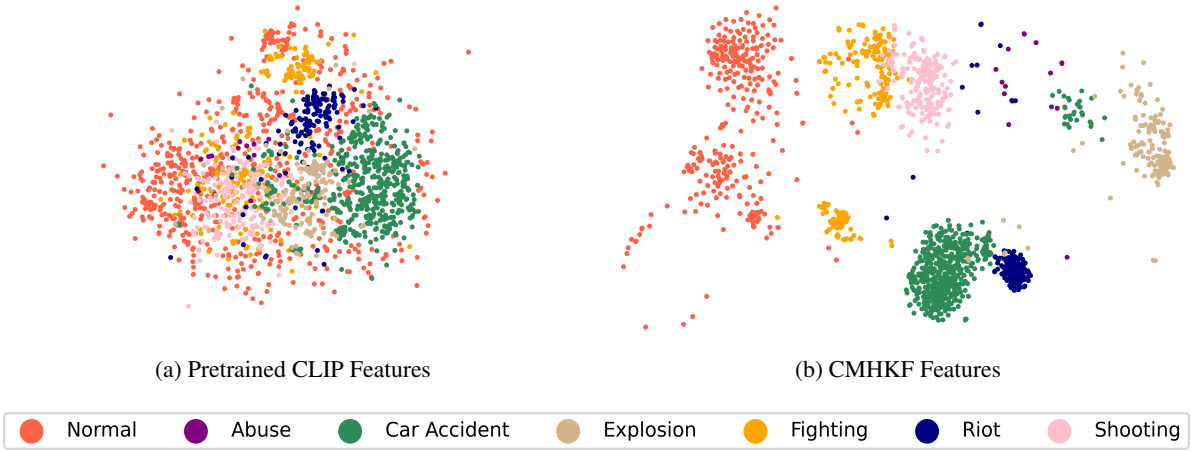
Index	Video	Audio	Text	AP (%)	AVG mAP (%)
1	✓			81.59	21.13
2	✓	✓		83.65	22.49
3	✓		✓	84.81	24.32
4	✓	✓	✓	**86.57**	**26.70**

Effectiveness of the Multi-Modality. Most video anomaly detection studies have traditionally focused on single modalities; only recently have researchers begun exploring dual-modality approaches, such as video–audio or video–text. In contrast, our approach integrates video, audio, and text. We tested four modality combinations: video-only, video-audio, video-text, and video-audio-text, as shown in Table 6. The absence of a modality can disable the corresponding processing module. Our experiments show that multi-modal combinations, particularly video-audio-text, outperform single-modal methods, improving AP and AVG mAP by 4.98% and 5.57%, respectively, compared to video alone, and by 1.76% and 2.38% compared to the best bimodal setup (video-text). These findings validate the efficacy of cross-modal knowledge integration in enhancing video anomaly detection.

Table 7: Evaluation on Abnormal Videos Only. * indicates the method was re-implemented for evaluating performance exclusively on anomalous videos.

Method	AP (%)	Ano-AP (%)	AVG mAP (%)	Ano-AVG mAP (%)
VadCLIP *	84.49	84.89	24.04	17.40
Ours	86.57	86.69	26.70	21.62

Evaluation on Abnormal Videos Only. Previous methods evaluated overall performance using both normal and abnormal videos during testing. However, evaluating performance solely on abnormal videos is a crucial metric for assessing anomaly detection capabilities. Thus, we excluded normal videos from testing (see Table 7). Results show that this exclusion improves coarse-grained AP. We attribute this improvement to the more diverse and random distribution of multi-modal features (e.g., video and audio) in normal videos, which introduces cross-modal inconsistencies and increases uncertainty in the model's decision-making. In contrast, abnormal videos contain more consistent and salient anomaly cues. For example, explosions show visual flashes and loud audio, offering aligned signals. Such consistency facilitates anomaly detection. However, fine-grained AVG mAP declines. We believe that, although the consistency between modalities enables the model to excel in distinguishing normal and anomalous events, different types of anomalous events (e.g., car accidents and explosions) are often correlated and share certain similarities, posing a challenge in accurately identifying distinct anomalous events.



(a) Pretrained CLIP Features

(b) CMHKF Features

Figure 5: The representations learned by (a) the original CLIP-based visual features and (b) the proposed CMHKF method are visualized using t-SNE on the XD-Violence dataset.

4.5 Qualitative Results

To fully evaluate our proposed CMHKF, we also conduct extensive qualitative experiments.

Coarse-grained Qualitative Visualization. Figure 4 shows CMHKF's predicted anomaly scores on the XD-Violence dataset, demonstrating its effectiveness. In Figures 4a, 4e, and 4g, our method accurately localizes anomalous intervals and predicts precise scores for prolonged abnormal regions. Figures 4c, 4d, and 4f further demonstrate its capability to detect and score discontinuous anomalies. Notably, in the darker environment of Figure 4f, the method detects a shooting event, highlighting audio and text contributions. Figure 4b illustrates detection of anomalies spanning entire regions. Moreover, Figure 4h shows the method's capacity to minimize false positives in challenging normal videos. High anomaly sensitivity and low normal scores confirm the method's robustness.

Embedding features. We used t-SNE (Van der Maaten and Hinton, 2008) to visualize the raw CLIP features and the representations learned by our CMHKF model. As shown in Figure 5, the learned representations from our model exhibit a more distinguishable and separable feature distribution than the raw input features extracted by the CLIP model. In the feature space generated by our model, normal and anomalous features are clearly separated, with a distinct boundary between them. Additionally, different types of anomalies are distinguishable, indicating that the CMHKF model excels in both coarse-grained detection of normal versus anomalous events and fine-grained differentiation between specific anomalous event types.

5 Conclusions

In this work, we propose a novel framework named CMHKF, aimed at effectively integrating cross-modality heterogeneous knowledge from video, audio, and text. To further enable adaptive cross-modality knowledge learning, we developed two key modules: CVKA and AFAE. Specifically, CVKA dynamically aligns and aggregates semantically relevant text features. In parallel, AFAE is engineered to mine temporally salient regions of audio features, focusing on capturing key characteristics within the audio. We conducted extensive evaluations on XD-Violence, the only large-scale audio-visual dataset, and the experimental results demonstrate its effectiveness. In the future, we plan to construct a new audio-visual dataset to further validate and benchmark the effectiveness of multi-modality approaches in this domain.

6 Limitations

While the CMHKF framework achieves significant results, there are still some limitations. Firstly, the evaluation of the model is currently limited to the XD-Violence dataset, which is the only available multimodal dataset for audio-visual violence detection. Future work will focus on constructing new multimodal datasets to more comprehensively validate and assess the performance of the method. Secondly, due to the limitations of weakly supervised learning, the text modality relies on learnable class labels. Future research will explore the use of large language models to generate richer textual descriptions for video data to enhance the expressive power of the text modality.

7 Ethics Statement

The data used in this paper are sourced from open-access repositories, and do not pose any privacy concerns. We are confident that our research adheres to the ethical standards set forth by ACL.

References

Anas Al-Lahham, Muhammad Zaigham Zaheer, Nurbek Tastan, and Karthik Nandakumar. 2024. Collaborative learning of anomalies with privacy (clap) for unsupervised video anomaly detection: A new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425.

Salem AlMarri, Muhammad Zaigham Zaheer, and Karthik Nandakumar. 2024. A multi-head approach with shuffled segments for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 132–142.

Congqi Cao, Xin Zhang, Shizhou Zhang, Peng Wang, and Yanning Zhang. 2023. Weakly supervised video anomaly detection based on cross-batch clustering guidance. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2723–2728. IEEE.

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

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. 2023b. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 387–395.

MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. 2023. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12137–12146.

Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. 2023. Prompt switch: Efficient clip adaptation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15648–15658.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE.

Hamza Karim, Keval Doshi, and Yasin Yilmaz. 2024. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6848–6856.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Shuo Li, Fang Liu, and Licheng Jiao. 2022. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1395–1403.

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

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. 2023. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8022–8031.

Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond.

2024. Oe-ctst: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 8574–8583.

Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 2019. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8679–8687.

Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. 2018. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–579.

Xiaogang Peng, Hao Wen, Yikai Luo, Xiao Zhou, Keyang Yu, Ping Yang, and Zizhao Wu. 2023. Learning weakly supervised audio-visual violence detection in hyperbolic space. *arXiv preprint arXiv:2305.18797*.

R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, Patrick Cardinal, et al. 2022. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2486–2495.

Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. 2024. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Md Haidar Sharif, Lei Jiao, and Christian W Omlin. 2023. Cnn-vit supported weakly-supervised video segment level anomaly detection. *Sensors*, 23(18):7734.

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

Shengyang Sun, Jiashen Hua, Junyi Feng, Dongxu Wei, Baisheng Lai, and Xiaojin Gong. 2024. Tdsd: Text-driven scene-decoupled weakly supervised video anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5055–5064.

Weijun Tan, Qi Yao, and Jingfeng Liu. 2024. Overlooked video classification in weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 202–210.

Kaibin Tian, Ruixiang Zhao, Zijie Xin, Bangxiang Lan, and Xirong Li. 2024. Holistic features are almost sufficient for text-to-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17138–17147.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Jue Wang and Anoop Cherian. 2019. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211.

Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. 2022a. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer.

Peng Wu, Xiaotao Liu, and Jing Liu. 2022b. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 25:1674–1685.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. 2024a. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307.

Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. 2024b. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9301–9310.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. 2024c. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6074–6082.

Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. 2023. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, pages 6620–6630.

Zhiwei Yang, Jing Liu, and Peng Wu. 2024. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18899–18908.

Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. 2022. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6278–6287.

Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. 2024. Delving into clip latent space for video anomaly recognition. *Computer Vision and Image Understanding*, 249:104163.

Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. 2023. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280.

Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. 2019. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246.

Hang Zhou, Junqing Yu, and Wei Yang. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3769–3777.

A Appendix

A.1 Added Dataset and Result

In this section, we evaluate our model on the UCF-Crime dataset, performing both coarse-grained and fine-grained WSVAD experiments, despite the lack of audio information, to demonstrate its effectiveness.

UCF-Crime is a widely used, large-scale video surveillance anomaly detection dataset, spanning 128 hours of real-world footage. It comprises 1,900 untrimmed videos across 13 anomaly categories—such as explosions, arrests, and road accidents—providing diverse scenario coverage. The dataset is split into 800 normal and 810 anomalous videos for training, while 140 normal and 150 anomalous videos are used for testing.

Implementation Details. The image and text encoders are based on the pre-trained CLIP model (ViT-B/16) (Radford et al., 2021), with a feature dimension of $D = 512$. We set the following key hyperparameters: the Top-k selection parameter $k = 4$ in Eq. 3; temperature $\tau = 0.07$ in Eq. 8; loss-balancing weights $\lambda_1 = 1 \times 10^{-3}$ and $\lambda_2 = 1 \times 10^{-4}$ in Eq. 11. We trained the model in PyTorch on a single NVIDIA RTX 3090 GPU, employing the AdamW optimizer for 20 epochs with a batch size of 32 and an initial learning rate of 2×10^{-5}.

Coarse-grained WSVAD Results. We compare our method with state-of-the-art methods, presenting the AUC results on the UCF-Crime dataset in Table 8. Given that the UCF-Crime dataset lacks audio data, we omit audio-related modules in our framework and adapt the MKAF module to facilitate the fusion of video and text modalities. Despite operating with an incomplete model, our method achieves competitive performance, attaining an AUC score of 88.24%. This outperforms the best-performing video-text dual-modal method, STPrompt (Wu et al., 2024b), and surpasses the strongest single-modal approach, UR-DMU (Zhou et al., 2023), by 1.27% in AUC. The superior performance of our method underscores the effectiveness of the proposed architecture, particularly the CVKA and MKAF modules. These components demonstrate a robust ability to learn cross-modality heterogeneous knowledge between video and text, establishing our method as highly effective in coarse-grained anomaly detection, even under conditions with incomplete modality integration.

Table 8: Coarse-grained comparisons on UCF-Crime. Best result is **bolded** and second best result is underlined.

Method	Publication	Modality	AUC (%)
Unsupervised learning based methods			
SVM baseline	NIPS'99	Video	50.10
Conv-AE (Hasan et al., 2016)	CVPR'16	Video	50.60
CLAP (Al-Lahham et al., 2024)	CVPR'24	Video	78.02
Weakly supervised learning based methods			
Sultani et al. (Sultani et al., 2018)	CVPR'18	Video	75.41
GCN (Zhong et al., 2019)	CVPR'19	Video	82.12
HL-Net(Wu et al., 2020)	ECCV'20	Video	82.44
MIST (Feng et al., 2021)	CVPR'21	Video	82.30
RTFM (Tian et al., 2021)	ICCV'21	Video	84.30
MSL (Li et al., 2022)	AAAI'22	Video	85.30
S3R (Wu et al., 2022a)	ECCV'22	Video	85.99
Cho et al. (Cho et al., 2023)	CVPR'23	Video	86.10
Zhang et al. (Zhang et al., 2023)	CVPR'23	Video	86.22
UMIL (Lv et al., 2023)	CVPR'23	Video	86.95
UR-DMU (Zhou et al., 2023)	AAAI'23	Video	86.97
Pu et al. (Pu et al., 2024)	TIP'24	Video + Text	86.76
TPWNG (Yang et al., 2024)	CVPR'24	Video + Text	87.79
VadCLIP (Wu et al., 2024c)	AAAI'24	Video + Text	88.02
STPrompt (Wu et al., 2024b)	ACMMM'24	Video + Text	88.08
Ours	—	**Video + Text**	**88.24**

Table 9: Fine-grained comparisons on UCF-Crime. Best result is **bolded** and second best result is underlined.

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
Sultani et al. (Sultani et al., 2018)	5.73	4.41	2.69	1.93	1.44	3.24
Wu et al. (Wu et al., 2022b)	10.27	7.01	6.25	3.42	3.29	6.05
VadCLIP (Wu et al., 2024c)	11.72	7.83	6.40	4.53	2.93	6.68
Ours	**15.49**	**11.99**	**8.95**	**6.40**	**4.88**	**9.54**

Fine-grained WSVAD Results. Table 9 presents a comparison of CMHKF with other methods for the fine-grained task. Despite the absence of audio data and corresponding modules, our method achieves a notable AVG mAP score of 9.54%, surpassing the previously best-performing method (Wu et al., 2024c) by 2.86%. This result highlights the effectiveness of the CVKA module in semantically aligning video and text features. By incorporating text features that provide clearer and more precise definitions of anomalous events, our model effectively captures the intrinsic attributes of various anomaly types, thus enabling more accurate differentiation between distinct anomaly categories.

A.2 Added Ablation Study

Effectiveness of Loss. we conducted an ablation study on the contrastive losses $\mathcal{L}_{\mathrm{NA}}$ and $\mathcal{L}_{\mathrm{AA}}$, as shown in Table 10. The results demonstrate that $\mathcal{L}_{\mathrm{NA}}$ effectively separates normal from anomalous textual features, while $\mathcal{L}_{\mathrm{AA}}$ distinguishes different anomaly types. These losses improve the separation of text embeddings, enabling the model to better identify normal events and anomalies.

Table 10: Effectiveness of the contrastive loss based on anomaly categories $\mathcal{L}_{\mathrm{NA}}$ and $\mathcal{L}_{\mathrm{AA}}$. Best result is **bolded** and second best result is underlined.

Index	$\mathcal{L}_{BCE}$	$\mathcal{L}_{NCE}$	$\mathcal{L}_{NA}$	$\mathcal{L}_{AA}$	AP(%)	AVG mAP(%)
1	✓	✓			86.15	25.98
2	✓	✓	✓		86.41	26.39
3	✓	✓		✓	86.34	26.18
4	✓	✓	✓	✓	**86.57**	**26.70**

Table 11: Effectiveness of Multi-Modality Fusion Methods. Best result is **bolded** and second best result is underlined.

Index	Method	AP (%)	AVG mAP (%)
1	Bilinear & Concat	85.67	25.31
2	Bilinear & Additive	86.19	25.73
3	Concat Fusion	86.34	26.38
4	**MKAF**	**86.57**	**26.70**

Effectiveness of Multi-Modality Fusion Methods. We conducted ablation experiments to compare different multi-modality fusion methods on the XD-Violence dataset, with results presented in Table 11. The fusion methods include Bilinear & Concat, Bilinear & Additive, Concat Fusion, and the Multi-Modality Knowledge Adaptive Fusion (MKAF) method that we propose. Bilinear & Concat processes features through linear layers to ensure consistent dimensions, followed by concatenation. Bilinear & Additive through linear layers to ensure consistent dimensions, and combines modality information through element-wise summation. Concat Fusion concatenates features from the three modalities directly. The results demonstrate that MKAF outperforms other fusion methods, achieving 86.57% AP and 26.70% AVG mAP. This demonstrates the superiority of MKAF for multi-modality fusion.