Vol. 07, No. 01, pp. 137 –156 (2026)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Integrating Vision and Language: An Improved VAD Model

Manas Ranjan Biswal* and Santos Kumar Baliarsingh

School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar-751024, Odisha, India.
manas.biswalfcs@kiit.ac.in, santos.baliarsinghfcs@kiit.ac.in

**Correspondence: manas.biswalfcs@kiit.ac.in*

Abstract

Automatic anomaly detection in video surveillance is crucial for public and private safety. However, it is challenging because of unclear abnormal events, limited labeled data, and mismatches between different types of data. Traditional video anomaly detection methods mainly focus on spatiotemporal visual features. They often ignore semantic information and interactions between different data types. Additionally, many multimodal approaches use basic fusion methods that do not solve the alignment problems between these types of data. To address these issues, we propose a multimodal framework that includes a Hierarchical Multi-scale Temporal Network (H-MSTN). This network models short-, medium-, and long-term dependencies in visual and textual data. A lightweight cross-modal attention module makes sure the semantics align. Meanwhile, a Multimodal Attention-Based Fusion Transformer (MAFT) refines cross-modal representations in real time. We evaluate this framework using the UCF-Crime and XD-Violence benchmarks. The proposed method achieves 92.42% AUC on UCF-Crime and 88.63% AP on XD-Violence with significantly lower computational cost and faster inference than recent multimodal baselines such as ReFLIP-VAD. These results demonstrate a strong efficiency–accuracy trade-off for real-time deployment while maintaining competitive or improved performance over prior methods such as MVAD and TEVAD.

Keywords: *Video Anomaly Detection (VAD), Vision- Language Models, Multimodal Anomaly Detection, Hierarchical Multi-Scale Temporal Network (H-MSTN), Cross-Modal Attention Module (CMAM), Multimodal Attention-Based Fusion Transformer (MAFT)*

Received: November 20th, 2025 / Revised February 13th, 2026 / Accepted: February 25th, 2026 / Online: March 3rd, 2026

I. INTRODUCTION

VAD is an important part of intelligent video surveillance. It supports applications in public safety, healthcare, and industrial settings. Its goal is to automatically identify unusual events in video streams. This includes violent actions, patient falls, and equipment failures. Despite its applicability, VAD remains a challenging problem due to data imbalance and semantic ambiguity. Traditional supervised methods face challenges because there are many more normal events than unusual ones. This makes the data unbalanced. Also, abnormal activities are often unclear and vary a lot, which makes them hard to detect. Anomalies vary in appearance, context, and duration, which makes them difficult to define and label. Weakly supervised methods have become popular because they only need video-level labels. However, they still struggle to accurately find and describe the exact abnormal events in videos.

The core challenge in video anomaly detection lies in the scarcity and diversity of abnormal events. Existing visual- only methods rely on spatial-temporal representations from models like Temporal Segment Networks (TSN) [1], 3D Convolutional Networks (C3D) [2], or Inflated 3D ConvNet (I3D) [3], which effectively capture motion but often overlook the high-level semantics of events. Abnormal behaviors in real-world surveillance often do not exhibit clear visual cues and may span multiple semantic contexts. It is very hard to capture using low-level spatio-temporal features alone. To overcome this, recent research has turned to multimodal approaches, which integrate visual and textual modalities. Video captioning models trained on large vision-language datasets can create detailed summaries that improve understanding of anomalies. Models like TEVAD [4] combine video and text to create captions for video segments. This method improves both anomaly detection and interpretability. However, most current fusion strategies use shallow or parallel designs. They perform fusion only at the input or decision level, missing out on important interactions between different modalities. These strategies often have difficulty identifying complex, long- duration anomalies in unedited videos. Recent advancements in Video Anomaly Retrieval (VAR) highlight the need for accurate multimodal representations that connect video segments with detailed textual descriptions. This leads to the integration of hierarchical temporal modeling with cross-modal attention, helping to capture varied visual patterns and meaningful text alignment.

doi: 10.38094/jastt71658

To tackle this challenge, we suggest a hybrid framework that combines video evidence with textual descriptions for a better understanding of anomalies. The framework models the timing of visual and textual streams separately through an H-MSTN, which captures motion patterns at different levels. To connect the meanings of both forms, a simple cross-modal attention module is included after the H-MSTN. This component performs mid-level fusion by aligning caption embedding with temporally processed visual features. This enhances cross-modal interaction without introducing excessive model complexity. To aggregate anomaly scores, we apply MIL using top-k pooling followed by binary cross-entropy loss.

Although the individual architectural components (e.g., trans- formers, dilated convolutions, and cross-modal attention) are well established, the novelty of this work lies in their hierarchical integration and task-specific adaptation for weakly supervised multimodal anomaly detection. The proposed design explicitly addresses temporal scale variation, cross-modal misalignment, and noisy correlations that are common in real-world surveillance scenarios.

The proposed method includes the following contributions.

- We introduce a hybrid video-text VAD model that combines temporal motion modelling using an H-MSTN with semantic understanding via video captions.
- We introduce a lightweight mid-level cross-modal attention module to improve semantic alignment between independently processed visual and textual streams.
- We propose a Multimodal Attention-based Fusion Transformer to dynamically align and integrate multi-modal features for enhanced representation and anomaly scoring.
- We validate our method on benchmark datasets and demonstrate that our method outperforms existing caption-aware and visual-only models.

The rest of the article is organized as follows: Section II provides a comprehensive review of related work. Section III presents the proposed methodology. Section IV discusses experiments and results. Section V concludes the paper with future directions.

II. Related works

VAD has been an active area of research due to its critical role in public safety, surveillance, and behavioural analysis. Recent advances in vision–language pretraining have significantly influenced multimodal video understanding tasks. Models such as CLIP and ALIGN learn joint visual–textual representations from large-scale image–text pairs. These representations have been increasingly adopted in VAD to improve semantic understanding beyond purely visual cues. Over the years, several recent VAD frameworks [5], [6], [7], [8] have been developed to detect anomalies. AnomalyCLIP [9] employs CLIP similarity scores between video frames and predefined anomaly prompts, enabling language-driven anomaly detection without explicit temporal modelling. Multi-modal learning surveys [10], [11], [12] highlight best practices such as explicit cross-modal attention, hierarchical fusion, and modality-specific inductive biases to improve robustness and interpretability.

Table I highlights that existing multimodal VAD methods typically incorporate only a subset of vision–language components. TEVAD and MVAD leverage captions but rely on shallow fusion without an explicit temporal hierarchy. VadCLIP focuses on CLIP similarity without modelling long-range temporal dependencies. ReFLIP-VAD introduces cross-modal alignment but lacks structured multi-scale temporal modelling. Recent VLLM-based VAD approaches emphasize reasoning and caption generation but often omit explicit temporal and attention-driven fusion mechanisms. In contrast, the proposed framework uniquely integrates hierarchical multi-scale temporal modelling (H-MSTN), explicit mid-level cross-modal attention (CMAM), and a gated Multimodal Attention-Based Fusion Transformer (MAFT), enabling both fine-grained temporal reasoning and robust semantic alignment. Sultani et al. [13] propose a weakly supervised deep anomaly detection method using an MIL framework. The authors introduce a large- scale dataset of 1900 untrimmed surveillance videos covering 13 types of real-world anomalies. Yang et al. [14] introduced a novel event restoration task, where key frames are used to infer missing intermediate frames. They propose USTN-DSC, a U-shaped Swin Transformer enhanced with dual skip connections. Wu et al. [15] method aims to detect violent video segments using only video-level labels. It also builds audio and visual snippet bags, clusters them into semi-bags, and uses contrastive learning plus self-distillation to improve detection. Chen et al. [5] address the challenge of detecting anomalies in long videos using only weak video-level labels. It introduces a Glance-and-Focus network to integrate spatial and temporal cues. SwinBERT [16] is the first end-to-end transformer model for video captioning that operates directly on raw video frames without relying on separate feature extractors. It uses a Video Swin Transformer to encode spatial-temporal content and a multimodal transformer to generate captions via masked language modelling. Hadsell et al. [17] introduce a contrastive and invariant mapping approach for dimensionality reduction that learns a nonlinear embedding function. Unlike classical methods such as Principal Component Analysis (PCA) or Locally Linear Embedding (LLE), DrLIM does not require a predefined distance metric in input space.

Wu et al. [18] propose a VAR technique, which goes beyond binary anomaly classification to retrieve specific abnormal events from a large corpus using rich textual queries. The paper [19] proposes a Multimodal VAD framework that incorporates audio, visual, and textual (language) modalities to improve anomaly detection performance in intelligent surveillance systems. The paper [20] presents a multi-modal anomaly detection approach that combines audio and visual information to detect abnormal events in surveillance videos. Chen et al. [4] in- troduces a text-enhanced video anomaly detection framework that leverages captions to improve weakly supervised video anomaly detection (WSVAD). Wu et al. [21] proposed a novel weakly supervised video anomaly detection framework that addresses limitations of visual-only detection by incorporating audio-visual collaboration. Lv et al. [22] proposed

a novel framework that integrates Video-based Large Language Models (VLLMs) with traditional VAD to enable threshold-free and explainable anomaly detection in long surveillance videos. Contrastive Predictive Coding (CPC) [23] is a powerful self-supervised approach that learns compact, high-level representations by predicting future latent representations instead of reconstructing raw data. Biswas and Tešić [24] tackle unsupervised domain adaptation for object detection in remote sensing imagery. For this, they have combined support-set guided pseudo-labeling, debiased contrastive learning, and CycleGAN-based image translation. ALPRO [25] is an end- to-end video-language pre-training framework that enhances cross-modal alignment without using explicit object detectors. Jaafar et al. [26] study explores four deep learning–based fusion strategies to detect aggression in surveillance videos by combining audio, video, text, and five meta-features. Roy et al. [27] enhance vision transformer models by embedding morphological operations within transformer blocks to better capture both spectral and spatial structural information in hyperspectral images. The following concluding remarks can be drawn as limitations in existing works.

TABLE I. ARCHITECTURAL COMPARISON OF THE PROPOSED FRAMEWORK WITH RECENT MULTIMODAL AND VISION-LANGUAGE VAD METHODS

Method	Uses Captions	CLIP-based Visual Features	Prompt-based Text	Explicit Hierarchical Temporal Modeling	Cross-modal Attention	Learned Multimodal Fusion Strategy
TEVAD	✓	✗	✗	Limited (Snippet-level)	✗	Late fusion
MVAD	✓	✓	✗	✗	✗	Simple multimodal fusion
VadCLIP	✗	✓	✗	✗	✗	CLIP similarity scoring
ReFLIP-VAD	✓	✓	✗	✗	✓	Cross-modal alignment
VLLM-based VAD	✓	✓	✓	✗	Implicit	Reasoning-based scoring
Ours	✓	✓	✓	✓ (H-MSTN)	✓ (CMAM)	✓ (Gated MAFT)

Implicit indicates functionality embedded within large vision–language models rather than an explicit architectural module.

- Most existing methods rely solely on visual cues, which struggle to disambiguate complex or subtle anomalies.
- Traditional VAD approaches fail to capture high-level semantics, which are used for identifying complex abnormal behaviors.
- Many multimodal systems use simplistic fusion strategies, such as early or late fusion. However, these approaches fail to capture meaningful intermediate interactions across modalities. Consequently, they suffer from semantic misalignment and inadequate representation of long-duration anomalies.
- Without learning cross-modal attention mechanisms, existing systems often fail to align vision and language effectively.

In summary, existing approaches to VAD can be broadly categorized into visual-only methods, caption-guided multi-modal models, and large vision–language reasoning frameworks. While visual-only methods focus primarily on motion cues, they often lack high-level semantic grounding. Caption-aware approaches introduce textual context but typically rely on shallow or generic fusion mechanisms without explicitly addressing sparse anomaly localization or noisy cross-modal correlations. Recent vision–language models emphasize semantic reasoning, yet they do not incorporate structured multi-scale temporal modeling tailored to weak supervision.

In contrast, the proposed framework integrates hierarchical multi-scale temporal modeling with anomaly-aware cross-modal alignment and gated multimodal fusion in a unified design. This structured integration explicitly addresses temporal scale variation, semantic misalignment, and noise robustness under weak supervision, distinguishing our approach from prior multimodal and visual-only methods.

III. METHODOLOGY

The proposed framework aims to enhance VAD by integrating high-level semantic cues from captions with low-level visual motion features. This section provides a detailed description of the proposed model. A Section-III workflow of the proposed multimodal anomaly detection framework is given in Figure 1. Figure 2 illustrates the architecture of the proposed multimodal anomaly detection framework. Untrimmed video segments and prompt templates are used to generate anomaly-focused captions via a prompt-based captioning network. Visual features are extracted using Video-CLIP, and textual features are obtained using Sentence-T5. Both

feature streams are independently processed through H- MSTNs to capture temporal dependencies. A CMAM aligns the modalities at mid-level, and a MAFT performs late fusion. The final anomaly score is computed using top-k pooling under an MIL framework. The overall workflow of the proposed multimodal anomaly detection framework is summarized in Algorithm 1 in the appendix. Unlike generic transformer-based architectures, the proposed framework is specifically structured to handle weak supervision and sparse anomaly signals. Each module (H-MSTN, CMAM, and MAFT) is designed to address a distinct challenge: multi-scale temporal modelling, anomaly-aware cross-modal alignment, and noise-robust multimodal aggregation, respectively.

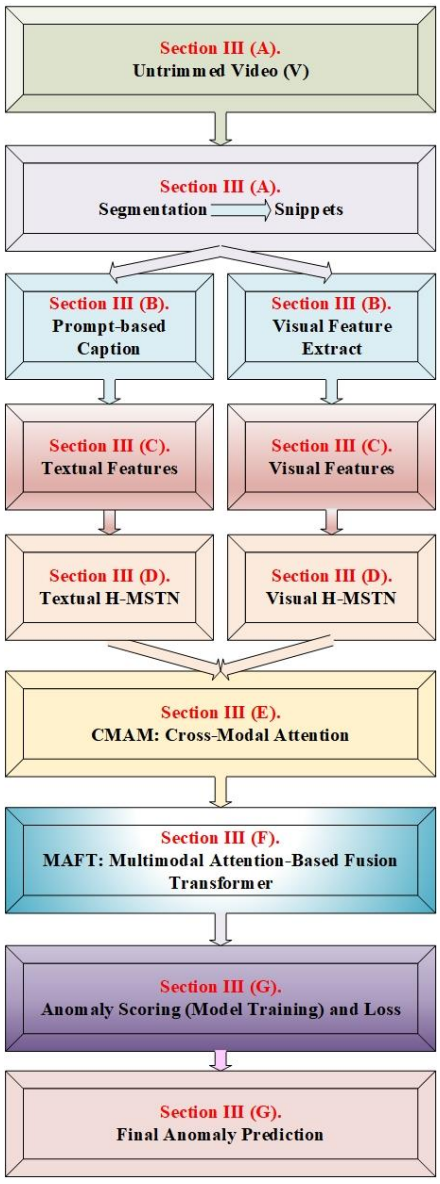

Fig. 1. Overview of the proposed multimodal anomaly detection framework

A. Problem Formulation

Given an untrimmed video $V = \{v_1, v_2, \dots, v_t\}$, where v_t represents a video snippet at time t. A weak video-level label $Y \in \{0,1\}$ indicating whether the video contains anomalies. The goal is to detect anomalous segments without snippet-level supervision. Each snippet is passed through a deep encoder to extract a feature vector. An anomaly score $s_t \in [0,1]$ is assigned to each snippet v_t using a scoring function g_ϕ. The video-level anomaly score is approximated using MIL by taking $\hat{Y} = \max_t s_t$, where at least one snippet must be abnormal for $Y = 1$. The model is optimized using binary cross-entropy loss. The final loss combines Multiple Instance Learning (MIL) loss, total variation regularization, and sparsity constraints.

B. Video Description Network

In the proposed work, we used a prompt-based generation strategy to produce informative and anomaly-relevant captions for each video snippet. Traditional captioning models describe everything neutrally and broadly without focusing on any specific task. The prompt-based captioning uses carefully crafted textual prompts to guide the model to focus only on relevant or abnormal events. It allows generating a descriptive sentence s_t that not only captures the visual content but also emphasizes semantic cues relevant to anomalies (e.g., actions, interactions, abnormal motions).

In a video snippet v_t to generate meaningful captions, we used SwinBERT [16] prompt-based video captioning model $\mathcal{C}(\cdot \mid \mathcal{P})$, where $\mathcal{P}$ is a predefined prompt template. The input video snippet $v_t \in \mathbb{R}^{F\times H\times W\times C}$ consists of F frames with height H, width W, and channels C. The caption generation process is formally defined as: $s_t = \mathcal{C}(v_t \mid \mathcal{P})$, where the model uses both the visual input and the prompt to conditionally generate the descriptive sentence s_t. For instance, consider an input video snippet v_t that shows a person running inside a shopping mall. A standard captioning model might generate a generic description, "A person is running." In contrast, a prompt-based captioning model supports a task-specific prompt like "Describe any unusual behavior in the video." This produces a context-aware and semantically rich caption: "A man is sprinting through a shopping mall, which appears unusual." This demonstrates how prompt-based generation can enhance the relevance and informativeness of captions.

C. Feature Extraction

The proposed framework extracts snippet-level features from both visual and textual modalities. In this work, we use Video-CLIP for visual features and Sentence-T5 for text. This provides better alignment between the two modalities and captures richer context.

Visual Feature Extraction: A vision-language pretraining network, Video-CLIP supports joint learning of spatial-temporal features. Given an input video V, we segment it into T fixed-length non-overlapping snippets $\{v_1, v_2, \dots, v_t\}$. Each contains 16 consecutive frames. Each snippet is passed through the Video-CLIP visual encoder to obtain a dense video embedding using Equation 1.

$$x_t^v = \text{VideoCLIP}(v_t) \in \mathbb{R}^{d_v} \quad (1)$$

Where d_v is the dimensionality of the visual embedding. The resulting sequence of embeddings forms the visual stream, which is denoted by Equation 2.

$$X_v = \{x_1^v, x_2^v, \dots, x_t^v\} \quad (2)$$

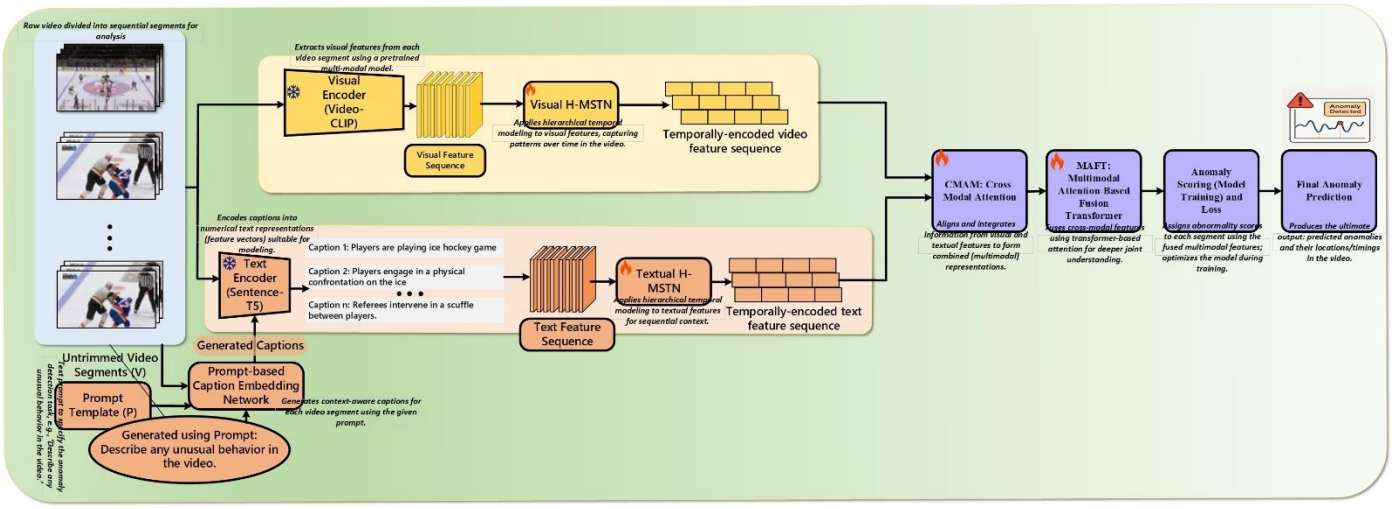

Fig. 2. Proposed multimodal anomaly detection framework.

Textual Feature Extraction (Sentence Embedding): A transformer-based sentence embedding model, Sentence-T5 convert encoding rich contextual information into fixed-dimensional vectors. It provides greater fluency and contextualization. For each video snippet v_t, we generate a descriptive sentence using a pretrained video captioning model. A sample caption might be: "A person is running across a crowded street." Each generated caption s_t is passed through the Sentence-T5 encoder to obtain its sentence-level embedding by using Equation 3.

$$x_t^l = \text{SentenceT5}(s_t) \in \mathbb{R}^{d_l} \quad (3)$$

Where d_l is the dimensionality of the text embedding. The textual stream is aggregated by using Equation 4

$$X_l = \{x_1^l, x_2^l, \dots, x_t^l\} \quad (4)$$

The effectiveness of these feature extractor selections is empirically validated through the ablation studies presented in Section IV-D.

D. Hierarchical Multi-scale Temporal Network (H-MSTN)

The proposed framework employs an H-MSTN to model temporal dependencies at multiple resolutions, enabling effective capture of both short-term and long-term anomaly patterns. Given an untrimmed video divided into t non-overlapping snippets, we extract modality-specific temporal sequences as:

- Visual stream: $X_v = \{x_1^v, x_2^v, \dots, x_t^v\}$, where $x_t^v \in \mathbb{R}^{d_v}$
- Textual stream: $X_l = \{x_1^l, x_2^l, \dots, x_t^l\}$, where $x_t^l \in \mathbb{R}^{d_l}$

H-MSTN consists of L hierarchical temporal levels, each designed to capture temporal patterns at different scales using a Pyramid Dilated Convolution (PDC) block. At level $\ell \in \{1, \dots, L\}$, given the input feature map $F^{(\ell-1)} \in \mathbb{R}^{T \times d}$ (with $F^{(0)} = X_v$ or X_l), parallel 1D convolutions with dilation rates $\delta \in \{1,2,4\}$ are applied to extract multi-scale temporal features. For reproducibility, we fix $L = 3$ to balance temporal modeling capacity and computational complexity. Using fewer levels (e.g., L = 2) limits the ability to capture both short- and long-range dependencies, while deeper hierarchies increase computational cost and risk overfitting under weak supervision. Empirically, three levels provided stable training and sufficient multi-scale temporal representation. Each level contains two Transformer layers with four attention heads, a hidden size of 256, and a feed-forward dimension of 512. Visual and textual features are linearly projected to a shared 256-dimensional embedding space before temporal modelling.

Each dilated convolution branch is computed using Equation 5.

$$F_\delta^{(\ell)} = \text{ReLU}\left(\text{Conv1D}_\delta\left(F^{(\ell-1)}\right)\right) \quad (5)$$

These outputs are concatenated by using Equation 6 to form a multi-scale representation.

$$F'^{(\ell)} = \text{Concat}\left(F_1^{(\ell)}, F_2^{(\ell)}, F_4^{(\ell)}\right) \in \mathbb{R}^{T \times 3d} \quad (6)$$

A linear projection followed by a residual connection is applied by using Equation 7.

$$F^{(\ell)} = W_\ell F'^{(\ell)} + F^{(\ell-1)} \quad (7)$$

Additionally, to enhance global temporal reasoning, we optionally apply a non-local block or self-attention mechanism (Equation 8).

$$F^{(\ell)} = F^{(\ell)} + \text{SelfAttn}\left(F^{(\ell)}\right) \quad (8)$$

This hierarchical design enables efficient modelling of local temporal transitions while preserving long-range dependencies. The same H-MSTN architecture is applied independently to visual and textual streams, which are represented by Equations 9 and 10, respectively.

$$\tilde{X}_v = \text{H-MSTN}(X_v) \in \mathbb{R}^{T \times \tilde{d}} \quad (9)$$

$$\tilde{X}_l = \text{H-MSTN}(X_l) \in \mathbb{R}^{T \times \tilde{d}} \quad (10)$$

While H-MSTN effectively captures intra-modality temporal dependencies, it processes each modality independently. To explicitly model cross-modal interactions and align visual cues with textual semantics, we introduce a Cross-Modal Attention Module (CMAM), described in the following section.

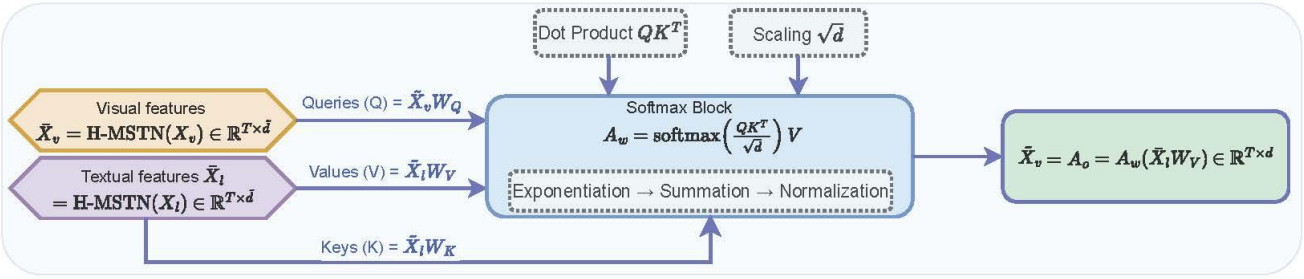

Fig. 3. Working mechanism of CMAM

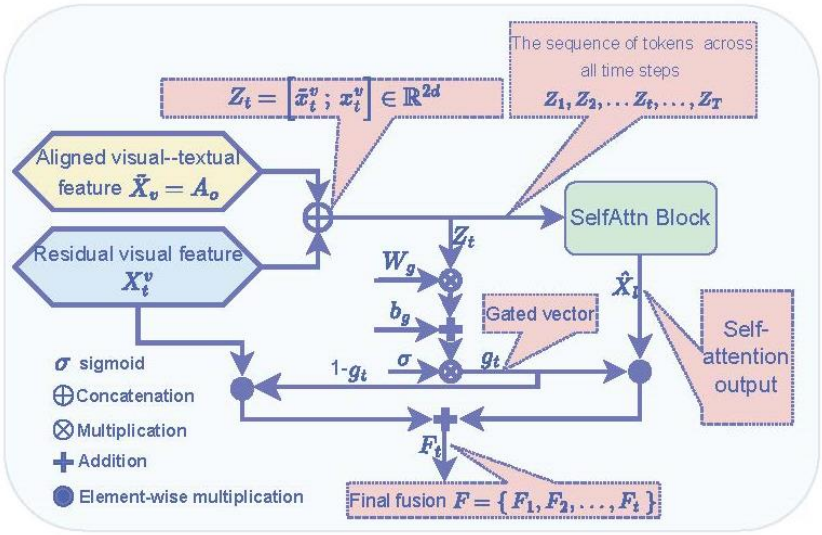

Fig. 4. Working Mechanism of Multimodal Attention-Based Fusion Transformer

E. Cross-Modal Attention

A lightweight CMAM (Figure 3) is introduced after the independent H-MSTNs to align visual and textual embedding. This module serves as a mid-level fusion mechanism that enhances interaction between modalities while keeping the architecture efficient. We calculate how strongly each visual snippet (query) should attend to each textual snippet (key). Let $\tilde{X}_v \in \mathbb{R}^{T\times d}$ and $\tilde{X}_l \in \mathbb{R}^{T\times d}$ denote the outputs of visual and language H-MSTNs, respectively, where t is the number of snippets and d is the feature dimensionality. The CMAM first computes the attention weight matrix $A_w \in \mathbb{R}^{T\times T}$ using scaled dot-product attention given by Equation 11. The final cross-modal aligned representation is obtained by applying the attention weights to the value matrix in Equation 12.

$$A_w = \text{softmax}\left(\frac{(\tilde{X}_v W_Q)(\tilde{X}_l W_K)^\top}{\sqrt{d}}\right) \in \mathbb{R}^{T\times T} \quad (11)$$

$$A_o = A_w(\tilde{X}_l W_V) \in \mathbb{R}^{T\times d} \quad (12)$$

where W_Q, W_K, and W_V are learnable projection matrices that transform the inputs into query, key, and value representations. The aligned output is denoted as $\tilde{X}_v = A_o$. The aligned feature at time t is $\tilde{X}_t^v = (A_o)_t$ (t-th row of the attention output matrix). These projections allow the model to learn how to attend from one modality to the other. Here, $QK^\top$ represents the similarity matrix between visual queries and textual keys. The division by $\sqrt{d}$ acts as a scaling factor to prevent overly large dot products. The softmax function is applied row-wise to ensure that the attention weights in each row sum to 1. In CMAM, the visual stream attends to the textual stream. The queries (Q) are derived from the visual representations, whereas the keys (K) and values (V) are derived from the textual representations. This design ensures that textual semantics guide the refinement of visual features. This is done by performing a weighted sum of the value vectors by using Equation 13.

$$\tilde{x}_t^v = \sum_{j=1}^{T} A_{t,j}\, v_j \quad (13)$$

The resulting $\tilde{X}_t^v$ is a weighted average of the textual value vectors for each visual snippet. $A_{t,j}$ denote how much the t-th visual snippet attends to the j-th language token. v_j denote the value vector of the j-th textual snippet.

In the context of weakly supervised anomaly detection, semantic misalignment arises when caption-level textual descriptions do not precisely correspond to temporally localized anomalous segments. Since anomalies are often sparse and subtle, global cross-modal similarity may introduce noisy or misleading alignments. Standard attention mechanisms perform similarity matching but do not explicitly account for sparse anomaly signals or unreliable cross-modal correlations.

The proposed CMAM refines temporally processed visual features using caption-guided attention at an intermediate stage, enabling anomaly-focused alignment. Subsequently, MAFT applies gated multimodal aggregation to suppress inconsistent cross-modal responses, thereby improving robustness to noisy or weak semantic cues. Although CMAM enables rich cross-modal feature exchange, it does not provide an integrated representation across modalities. In particular, naïve concatenation or attention alone can lead to redundancy and noisy feature fusion. Therefore, we propose the Multimodal Attention-Based Fusion Transformer (MAFT), which builds upon CMAM outputs and introduces gated fusion with transformer layers to produce a unified multimodal sequence representation.

F. Multimodal Attention-Based Fusion Transformer

The CMAM enables one modality (vision) to selectively attend to another (text). It operates directionally and models only pairwise interactions. The MAFT is a late-fusion architecture designed to integrate pre-extracted visual and textual features into a joint embedding space using self-attention mechanisms. The complete process of this module is given in Figure 4. It is a deep learning architecture that leverages the power of both Transformers and attention mechanisms to effectively combine information from multiple modalities. It uses the Transformer's ability to handle sequential data and the attention mechanism's capacity to weigh the importance of different inputs to create a richer, more robust representation of the data. MAFT allows every token to attend globally across modalities. This enhances robustness to noise and improves anomaly detection by leveraging the combined context of both streams. Thus, CMAM aligns modalities, while MAFT fuses them into a unified, context-aware representation. This modality-agnostic fusion leads to improved performance in tasks like classification, prediction, and more. In our implementation, MAFT consists of 2 Transformer layers, each using 4 attention heads and a hidden size of 256. Before entering MAFT, both the CMAM-aligned features and the residual visual features are projected to 256-dimensional vectors to ensure consistent scaling during multimodal self-attention and gated fusion. For each time step t, concatenate the aligned token with its residual visual token using Equation 14.

$$Z_t = [\,\tilde{x}_t^v\,;\, x_t^v\,] \in \mathbb{R}^{2d} \tag{14}$$

MAFT applies a Transformer-based self-attention operation over the full multimodal sequence $Z = \{Z_1, \dots, Z_T\}$ to enable global reasoning across modalities given by Equation 15.

$$\hat{X} = \mathrm{SelfAttn}(Z) \tag{15}$$

where $\hat{X} = \{\hat{X}_1, \dots, \hat{X}_T\}$ represents the multimodal context-enhanced outputs. Equation 15 allows MAFT to dynamically weigh the contribution of both the aligned and original features.

We used a gated fusion mechanism that adaptively balances the contribution of aligned visual-textual features. The final fused representation $F_t \in \mathbb{R}^d$ for each snippet (time step) $t \in \{1, \dots, T\}$ is computed using Equation 16 and 17.

$$g_t = \sigma\big(W_g Z_t + b_g\big), \quad g_t \in [0,1]^d \tag{16}$$

$$F_t = g_t \odot \hat{X}_t + (1 - g_t) \odot x_t^v \tag{17}$$

In Equations 16 and 17 $W_g \in \mathbb{R}^{d\times 2d}$, $b_g \in \mathbb{R}^d$ are learnable parameters. σ is the element-wise sigmoid function that outputs the gate vector $g_t \in [0,1]^d$ and $\odot$ denotes element-wise multiplication. This formulation enables the model to dynamically control the influence of visual and textual cues per time step.

G. Anomaly Scoring (Model Training) and Loss

Final anomaly scores for all snippets are aggregated using top-k pooling, which selects the top-k scores (highest anomalies) from the sequence. To compute the final anomaly score for a video, the model first assigns an anomaly score $\hat{y}_t \in [0,1]$ to each of the T snippets using a trained prediction head. The vector $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_T\}$ contains all snippet scores. Since only video-level labels are available, the model aggregates the most anomalous snippets to estimate the video-level score. A top-k pooling strategy is adopted, where k is a predefined value (e.g., 8 or 16). This technique selects the top-k highest-scoring snippets as $\hat{y}_{\text{top-}k} = \mathrm{TopK}(\hat{y}, k) \subset \hat{y} = \{\hat{y}_{\text{top-1}}, \hat{y}_{\text{top-2}}, \dots, \hat{y}_{\text{top-}k}\}$. The video-level anomaly score is calculated by averaging the scores of these top-k snippets by using Equation 18.

$$\hat{Y} = \frac{1}{k}\sum_{i=1}^{k} \hat{y}^i_{\text{top-}k} \tag{18}$$

This score is then compared against the ground-truth label $y \in \{0,1\}$ using binary cross-entropy loss. The video-level label $y \in \{0,1\}$, the Binary Cross-Entropy loss is computed by using Equation 19.

$$\mathcal{L}_{MIL} = -y\log\big(\hat{Y}\big) - (1 - y)\log\big(1 - \hat{Y}\big) \tag{19}$$

Now, we apply Total Variation (TV) Regularization to encourage smoothness across temporal anomaly scores. A total variation loss is calculated by using Equation 20 that penalizes abrupt changes between consecutive scores. In this equation $\hat{y}_t$ the predicted anomaly score at time step t.

$$\mathcal{L}_{\text{TV}} = \sum_{t=1}^{T-1} |\hat{y}_t - \hat{y}_{t+1}| \tag{20}$$

The sparsity constraint (Equation 21) incorporates the prior knowledge that anomalies are rare in surveillance videos. It prevents the model from overfitting and encourages it to highlight only the most critical anomalous segments. This is achieved by penalizing overly high or dense anomaly scores across all snippets. This encourages the model to assign high scores only to truly abnormal events, thereby making anomaly detection more effective.

$$\mathcal{L}_{\text{sparse}} = \sum_{t=1}^{T} \hat{y}_t \tag{21}$$

The final loss is a weighted sum of the MIL loss, TV regularization, and sparsity constraint as given in Equation 22. In this equation, λ_{TV} and λ_{sparse} are hyperparameters that control the contribution of the regularization terms. These weights are tuned based on validation performance to achieve a balance between anomaly focus and temporal smoothness.

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{MIL}} + \lambda_{TV} \cdot \mathcal{L}_{TV} + \lambda_{\text{sparse}} \cdot \mathcal{L}_{\text{sparse}} \tag{22}$$

For instance, if a video of 5 snippets yields predicted anomaly scores $\hat{y}$ is [0.12,0.85,0.40,0.83,0.80]. From this, for $k = 3$ top 3 scores are $[0.85,0.83,0.80]$. The aggregated anomaly score is $\hat{y}_t = \frac{1}{3}(0.85 + 0.83 + 0.80) = 0.8267$. If the ground-truth label is anomalous ($y = 1$), the MIL loss is computed as $\mathcal{L}_{\text{MIL}} = -y \cdot \log(\hat{y}_t) = -\log(0.8267) \approx 0.190$. This method stabilizes learning by mitigating the effect of noisy max values and encouraging the model to focus on multiple high-confidence anomalous snippets. The TV loss is calculated by summing the absolute differences between consecutive predicted scores: $\mathcal{L}_{TV} = |0.12 - 0.85| + |0.85 - 0.40| + |0.40 - 0.83| + |0.83 - 0.80| = 0.73 + 0.45 + 0.43 + 0.03 = 1.64$. The sparsity loss is the sum of all anomaly scores: $\mathcal{L}_{\text{sparse}} = 0.12 + 0.85 + 0.40 + 0.83 + 0.80 = 3.00$. Finally, the total loss $\mathcal{L}_{\text{final}}$ combines the MIL loss with the weighted TV and sparsity regularization using hyperparameters $\lambda_{TV} = 0.1$ and $\lambda_{\text{sparse}} = 0.01$: $\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{MIL}} + \lambda_{TV} \cdot \mathcal{L}_{TV} + \lambda_{\text{sparse}} \cdot \mathcal{L}_{\text{sparse}} = 0.190 + 0.1 \cdot 1.64 + 0.01 \cdot 3.00 = 0.190 + 0.164 + 0.03 = 0.384$.

IV. Experiment and Results

This section presents the details of the datasets, experimental setup, and a detailed analysis of the results obtained across various benchmarks and evaluation metrics.

A. Datasets, Evaluation Metrics, and Implementation Details

Datasets description: We evaluate our method on two large-scale and realistic benchmark datasets commonly used for weakly supervised VAD. These datasets are UCF-Crime [13] and XD-Violence [28]. A comparative overview of both datasets is provided in Table II.

Evaluation Metrics: The article follows standard metrics that match previous effective methods to evaluate the proposed model. For the UCF-Crime dataset, we calculate the frame-level Area Under the Curve (μ_{AUC}) and the AUC for anomalous videos only (AUCno-n). These metrics measure the model's ability to tell the difference between normal and unusual frames and its sensitivity to anomalies. For the XD-Violence dataset, we calculate the frame-level Average Precision (AP) to understand the precision-recall trade-off in temporal predictions. We also compute the mean Average Precision (μ_{AP}) at different Intersection over Union (IoU) thresholds. This helps us evaluate the timing of the predicted and actual anomaly segments. These metrics assess both coarse and detailed anomaly detection performance. They also allow fair comparisons with vision-language and multimodal frameworks in various video surveillance situations. All experiments follow the official train–test splits of UCF-Crime and XD-Violence, with 10% of the training data held out for validation and hyperparameter tuning. Model selection is based on the best validation AUC (UCF-Crime) or AP (XD-Violence), and the corresponding checkpoint is used for final evaluation without early stopping. A fixed random seed (42) is used across Python, NumPy, and PyTorch to ensure deterministic behavior, with negligible performance variance across runs. VideoCLIP and Sentence-T5 encoders are kept frozen, and only the proposed temporal modeling and fusion modules are trained to improve stability and generalization.

TABLE II. Summary of UCF-Crime and XD-Violence Dataset

Dataset Name	Modalities	Dataset Statistics	Labels
XD-Violence	Visual (RGB), Audio, Text (optional captions) [Visual + Audio, Visual + Text, Visual + Audio + Text (Multimodal Fusion)]	• Total Videos: 4,754 video clips • Total Duration: Over 217 hours • Anomalous Videos: 1,903 • Normal Videos: 2,851 • Scenes: Streets, malls, subways, fights, protests, riots • Sources: YouTube and online platforms	Video-level binary (violent/non-violent)
UCF-Crime	Visual (RGB); Event-camera and frame-level in extensions	• Total Videos: ~1,900 untrimmed videos • Total Duration: ~128 hours • Anomalous: 810 (train), 140 (test) • Normal: 1,610 (train), 150 (test) • Anomaly Categories: 13 (e.g., abuse, arson, robbery) • Scenes: Sparse anomalies in long surveillance sequences	Video-level binary (normal/anomalous), some frame-level

Implementation Details: For the implementation of our proposed framework, we used the Python language and PyTorch framework and ecosystem equipped with an NVIDIA Tesla V100 GPU with 32GB memory. Training the proposed model took about 1.5 to 2 hours on UCF-Crime and 2.5 to 3 hours on XD-Violence, depending on batch size and epoch count. Inference runs at nearly real-time speed, processing each video snippet in under 15 ms. Peak GPU memory usage was about 18 to 20 GB during training. Besides performance accuracy, we also consider practical deployment. The reported inference time of 15 ms per snippet supports real-time surveillance applications. The measured FLOPs and GPU memory usage

stay within the range of standard high-performance GPUs. Freezing the pretrained encoders further cuts down on computational overhead. In resource-constrained environments, caption generation can happen offline or centrally, allowing for efficient online inference.

To make sure the proposed framework is stable and strong, we ran all experiments five times with different random seeds. The results in the main comparison tables show the mean ± standard deviation. The proposed method achieves 92.42 ± 0.18% AUC on UCF-Crime and 88.63 ± 0.21% AP on XD-Violence, demonstrating consistent performance with low variance. Additionally, a paired t-test confirms that the improvements over baseline methods are statistically significant ($p<0.05$). The implementation setup used Adam optimizer with a batch size of B = 64 for the UCF-Crime dataset and B = 128 for the XD-Violence dataset. For the UCF-Crime dataset, the learning rate and total epoch is set as η=4"e" ^(-4) and 60. Similarly, for the XD-Violence dataset, the learning rate and total epoch is set as η=5"e" ^(-4) and 80. We used a weight decay of 1e-3 to prevent overfitting.

TABLE III. FRAME-LEVEL AUC PERFORMANCE COMPARISON FOR THE UCF-CRIME DATASET

Supervision	Source	Method	Features	AUC (%)	Ano-AUC (%)
Un	ICMLC 2003	OCSVM [34]		63.20	51.06
Un	CVPR 2016	Hasan et al. [29]	-	50.20	39.43
Un	ICCV 2019	BODS [35]		68.3	-
Un	ICCV 2019	GODS [35]		70.5	-
Un	Pattern Recog 2020	FSCN [30]	STFF	70.6	-
Weak	IEEE TCS II	Deep-MIL [36]	C3D	75.42	54.25
Weak	CVPR 2018	Deep-MIL [13]	CLIP/C3D	84.14/75.41	63.29/54.25
Weak	ECCV 2020	HL-Net [28]	CLIP/I3D	84.57/82.44	62.21
Weak	IEEE TIP	CRFD [31]	CLIP	84.72	62.60
Weak	ICCV 2021	RTFM [7]	I3D	86.65	63.86
Weak	IEEE TIP	WASAL [37]	TSN	85.38	67.38
Weak	Measurement 2023	AWAD [38]	C3D	81.48	
Weak	TMM 2022	AVVD [15]	CLIP/I3D	82.45/82.44	60.27/-
Weak	CVPR 2023	UB-MIL [32]	X-CLIP	87.58	68.91
Weak	CVPR 2023	TEVAD [4]	SwinBERT/SimCSE	84.90	-
Weak	AAAI 2024	Vad-CLIP [39]	CLIP	88.02	70.23
Weak	IEEE TCSVT	ReFLIP-VAD [33]	CLIP	88.57	72.35
Weak	IEEE TCSVT	ReFLIP-VAD [33]	FLIP	89.14	74.72
	This Work	**Proposed (Ours)**	**CLIP+Sentence-T5**	**92.42**	**74.74**

For the UCF-Crime dataset, the loss weight adjustment factors are set as follows: the total variation regularization coefficient λ_{TV} is set to 0.1, and the sparsity constraint coefficient λ_{sparse} is set to 0.01. Similarly, for the XD-Violence dataset, the loss weight adjustment factors are set as follows: the total variation regularization coefficient λ_{TV} is set to 0.05, and the sparsity constraint coefficient λ_{sparse} is set to 0.005. The regularization weights λ_{TV} and λ_{sparse} were selected through validation-based tuning on each dataset. We evaluated a small range of values and chose the setting that provided the best trade-off between temporal smoothness and anomaly sparsity. Empirically, we observed that the performance was stable within a reasonable range around the selected values, and the reported settings yielded consistent results across both datasets. All hyperparameter tuning, including batch size selection, was conducted using validation splits derived from the training data. The test sets of UCF-Crime and XD-Violence were used exclusively for final performance evaluation to avoid any risk of data leakage. The experiment uses VideoCLIP as the visual

feature extractor and Sentence-T5 as the textual feature extractor for encoding snippet-level video embeddings and caption representations, respectively. The visual sample rate is set to 16 frames per snippet for the UCF-Crime dataset and 20 frames per snippet for the XD-Violence dataset. The sliding window size for visual feature extraction is fixed at 16 frames. The experiment employed Top-k Pooling under an MIL framework, with k set to either 8 or 16. The value of Top-k pooling was selected through validation-based tuning for each dataset. We evaluated a small range of k values and chose k=8 for UCF-Crime and k=12 for XD-Violence, which provided the best trade-off between sensitivity and robustness. All tuning was performed on validation splits derived from the training data, and test sets were used exclusively for final evaluation. For feature integration, the fusion module combines a CMAM with an MAFT. All attention-based modules (CMAM, H-MSTN, and MAFT) operate in a 256-dimensional embedding space. Raw visual features from VideoCLIP with dimension d_v and textual features from Sentence-T5 with dimension d_l are linearly projected to this shared 256-dimensional space before temporal modeling and fusion.

We set the embedding dimension to d = 256 to balance representational capacity and computational efficiency. Smaller dimensions may limit cross-modal expressiveness, while larger ones significantly increase attention cost and memory usage without consistent performance gains. H-MSTN uses $L = 3$ hierarchical levels with 2 layers and 4 attention heads per level, while MAFT employs 2 layers with 4 attention heads each. The visual and textual feature extractors (VideoCLIP and Sentence-T5) are kept frozen during training, and only the proposed temporal modeling and fusion modules are optimized. Freezing the encoders significantly reduces training cost and memory usage, and helps prevent overfitting under weak supervision. This design also improves generalization by leveraging robust pre-trained representations.

B. Comparative analysis of the quantitative results

Table III shows a comprehensive comparison of frame-level AUC performance across various video anomaly detection (VAD) methods on the UCF-Crime dataset. The table categorizes these methods based on the type of supervision used—Unsupervised (Un) and Weakly Supervised (Weak). The performance comparison is based on two key metrics: AUC (%), which measures how well the method can distinguish anomalous frames from normal ones, and Ano-AUC (%), an additional metric focusing on anomaly-specific detection performance. Unsupervised VAD methods, which do not use labeled anomaly data, generally report lower performance. The AUC ranges from 50.20% (Hasan et al. [29]) to 70.6% (FSCN). Weakly supervised approaches that rely only on video-level labels achieve much stronger results. Methods like RTFM (86.65%), CRFD (84.72%), UB-MIL (87.58%), and ReFLIP-VAD (89.14%) show consistent improvements. The latest version of ReFLIP-VAD reached 92.43% by using vision and language features. The performance difference between the proposed method (92.42%) and ReFLIP-VAD (92.43%) on UCF-Crime is marginal and falls within statistical variation. Therefore, the results indicate comparable performance rather than a substantial numerical margin over this recent baseline. The proposed method integrates CLIP and Sentence-T5 for multimodal representation, achieving 92.42% AUC and 74.74% Ano-AUC. This demonstrates the value of combining visual and textual modalities for better anomaly detection under weak supervision.

Table IV compares computational complexity and runtime efficiency. The proposed method achieves lower FLOPs (89.3G), reduced inference time (15 ms per snippet), and lower GPU memory usage (19 GB) compared to recent multimodal approaches. The improvement is mainly attributed to freezing pretrained encoders and employing lightweight hierarchical temporal and fusion modules, making the framework more suitable for real-time deployment. Although H-MSTN incorporates hierarchical temporal modeling with self-attention, it operates on compact snippet-level representations rather than full frame sequences, which keeps the computational overhead manageable; as shown in Table IV, the proposed framework maintains computational complexity within a comparable range of recent multimodal methods, demonstrating that the hierarchical design does not introduce prohibitive cost. Although the proposed model achieves accuracy comparable to the recent ReFLIP-VAD baseline (92.42% vs. 92.43% AUC on UCF-Crime), Table IV demonstrates a clear efficiency advantage. The proposed framework requires substantially lower computational cost (89.3 GFLOPs vs. 112.4 GFLOPs) and faster inference (15 ms vs. 27.3 ms per snippet) with reduced GPU memory (19 GB vs. 26.1 GB). This indicates that the proposed hierarchical multimodal design improves deployment efficiency without sacrificing detection accuracy. Therefore, the primary benefit over ReFLIP-VAD lies in computational efficiency and real-time suitability rather than raw accuracy gains.

TABLE IV. COMPUTATIONAL EFFICIENCY COMPARISON

Method	FLOPs (GFLOPs)	Inference Time (ms/snippet)	GPU Memory (GB)
ReFLIP-VAD	112.4	27.3	26.1
MVAD	104.8	24.6	24.3
Ours	**89.3**	**15.0**	**19.0**

Table V presents a comparison of frame-level Average Precision (μ_{AP}) across weakly supervised VAD methods on UCF-Crime at IoU thresholds from 0.1 to 0.5. The final column shows the mean μ_{AP}. Early methods like Deep-MIL had limited performance at 3.24%. HL-Net improved this to 6.05%. More recent methods, such as Vad-CLIP (6.68%), ReFLIP-VAD (9.62%), and MVAD (12.46%), demonstrate significant gains through vision and language as well as multimodal modeling. The proposed method achieves the highest average μ_{AP} of 13.92%, outperforming MVAD at all IoU levels. It shows

particularly strong improvements at higher thresholds with 10.55% at IoU 0.4 and 7.72% at IoU 0.5. This improvement comes from combining CLIP-based visual features with Sentence-T5 textual embeddings.

TABLE V. FRAME-LEVEL AVERAGE PRECISION PERFORMANCE COMPARISON $\mu_{AP}@IoU$ (%)FOR THE UCF-CRIME DATASET

Method	Source	0.1	0.2	0.3	0.4	0.5	AVG
Deep-MIL [13]	CVPR 2018	5.73	4.41	2.69	1.93	1.44	3.24
HL-Net [28]	ECCV 2020	10.27	7.01	6.25	3.42	3.29	6.05
Vad-CLIP [39]	AAAI 2024	11.72	7.83	6.40	4.53	2.93	6.68
MVAD [19]	IEEE TIM	21.13	15.12	10.96	7.80	7.23	12.46
ReFLIP-VAD [33]	IEEE TCSVT	14.23	10.34	9.32	7.54	6.81	9.62
Ours (Proposed)	**This Work**	**21.32**	**16.42**	**13.60**	**10.55**	**7.72**	**13.92**

TABLE VI. FRAME-LEVEL AVERAGE PRECISION PERFORMANCE COMPARISON $\mu_{AP}@IoU$ (%)FOR THE XD-VIOLENCE DATASET

Supervision	Method	Source	Features	AP (%)
Un	SVM baseline	-	-	50.78
Un	OCSVM [40]	NeurIPS 1999	-	27.25
Un	ConvAE [29]	CVPR 2016	-	30.77
Un	S3R [41]	ECCV 2022	I3D	53.65
Weak	CRFD [31]	IEEE TIP 2011	I3D	75.90
Weak	Deep-MIL [13]	CVPR 2018	CLIP/I3D	75.18/75.07
Weak	Ju et al. [43]	ECVA 2022	I3D	76.57
Weak	HL-net [28]	ECCV 2020	CLIP/I3D	80.07/78.64
Weak	AnomalyCLIP [42]	CVIU 2024	ViT-B/16	78.55
Weak	RTFM [7]	ECCV 2021	I3D	77.81
Weak	MSL [44]	CVPR 2022	I3D	78.45
Weak	MS-BSAD [45]	ICIP 2021	I3D	78.92
Weak	TPWNG [46]	CVPR 2024	CLIP	83.68
Weak	VadCLIP [39]	AAAI 2024	CLIP	84.13
Weak	CLIP-TSA [47]	ICIP 2023	CLIP	82.17
Weak	MVAD [19]	IEEE TIM	CLIP	86.32
Weak	ReFLIP-VAD [33]	IEEE TCSVT	CLIP/FLIP	85.81/86.29
	Ours (Proposed)	**This Work**	**CLIP+Sentence-T5**	**88.63**

TABLE VII. FRAME-LEVEL AVERAGE PRECISION PERFORMANCE COMPARISON $\mu_{AP}@IoU$ (%) FOR THE XD-VIOLENCE

Method	Source	0.1	0.2	0.3	0.4	0.5	AVG
Deep-MIL [13]	CVPR 2018	20.08	13.72	8.44	5.06	2.81	10.02
Wu-TALC [48]	ECCV 2018	26.27	18.87	13.83	9.50	6.55	15.00
3C-Net [49]	ICCV 2019	23.77	17.78	11.90	8.28	5.87	13.52
AVVD [15]	TMM 2022	35.35	28.02	20.94	15.01	10.33	21.93
VadCLIP [39]	AAAI 2024	37.03	30.84	23.18	17.09	14.31	24.70
MVAD [19]	IEEE TIM	43.72	33.53	27.34	22.63	18.59	29.16
ReFLIP-VAD [33]	IEEE TCSVT	39.24	33.45	27.71	20.86	17.22	27.36
Ours (Proposed)	**This Work**	**44.10**	**34.25**	**28.90**	**23.45**	**18.05**	**29.55**

This integration allows for richer multimodal representations and more accurate temporal localization under weak supervision.

TABLE VIII. BATCH SIZE PERFORMANCE ON BOTH DATASETS

Dataset	Batch Size	AUC (%)	AP (%)
UCF-Crime	16	86.22	31.19
UCF-Crime	32	87.32	33.76
UCF-Crime	64	92.42	35.22
UCF-Crime	96	87.42	34.53
UCF-Crime	128	88.89	34.95
UCF-Crime	192	89.96	34.95
UCF-Crime	256	89.72	33.57
XD-Violence	16	91.65	85.10
XD-Violence	32	92.11	85.73
XD-Violence	64	92.76	85.81
XD-Violence	96	94.82	86.12
XD-Violence	128	97.36	88.63
XD-Violence	192	95.15	86.24
XD-Violence	256	95.55	86.32

Table VI presents a performance comparison of various VAD methods on the XD-Violence dataset using Average Precision (AP%) as the evaluation metric. The methods are grouped based on supervision type—unsupervised and weakly supervised. Unsupervised approaches such as OCSVM and ConvAE show limited effectiveness, achieving APs of 27.25% and 30.77%, respectively. The more recent S3R improves performance to 53.65% using self-supervised learning with I3D features. Weakly supervised methods demonstrate significantly better results. Early models like Deep-MIL and CRFD achieve around 75%. Other advanced techniques such as HL-Net, AnomalyCLIP, and VadCLIP show continued progress, reaching up to 84.13%. MVAD and ReFLIP-VAD approach the top with APs of 86.32% and 86.29%, respectively. The proposed method in this work surpasses all prior models by achieving an AP of 88.63% through the integration of CLIP-based visual features with Sentence-T5 textual embeddings. It shows the

strength of multimodal learning in enhancing anomaly detection under weak supervision.

Table VII presents a frame-level Average Precision (μ_{AP}@IoU) comparison across various weakly supervised video anomaly detection methods on the XD-Violence dataset. Early methods like Deep-MIL and Wu-TALC exhibit limited performance, with average APs of 10.02% and 15.00%, respectively. Other methods, such as 3C-Net and AVVD, achieve 13.52% and 21.93% average AP, respectively. VadCLIP uses vision-language alignment with CLIP to improve the results to 24.70%. MVAD achieves the highest performance with an average AP of 29.16%. ReFLIP-VAD achieves an average AP of 27.36%. Our proposed method outperforms the strongest existing baselines. It achieve 44.10%, 34.25%, and 18.05% AP at IoU thresholds 0.1, 0.2, and 0.5, respectively. The result gives an average AP of 29.55% and performs better than the methods listed in Table VII.

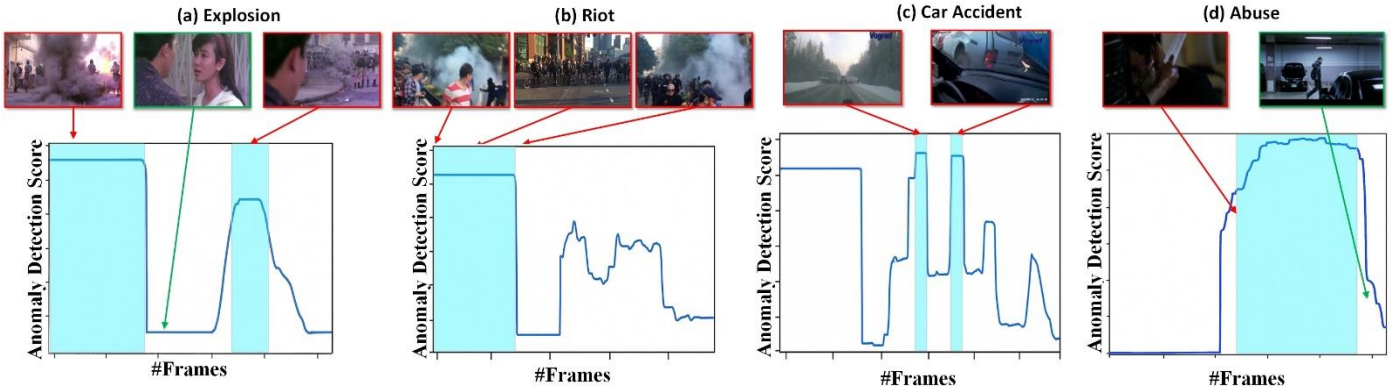

Fig. 5. Visualization Results of ground-truth and anomaly detection score of the proposed method on the UCF-Crime dataset

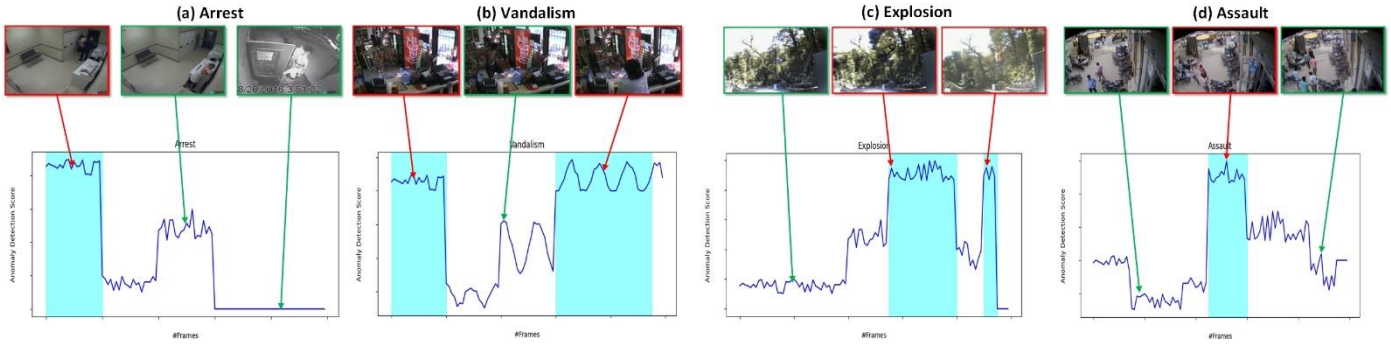

Fig. 6. Visualization Results of ground-truth and anomaly detection score of the proposed method on XD-Violence datasets

TABLE IX. PERFORMANCE COMPARISON (BEFORE AND AFTER) OF PROMPT-BASED CAPTIONING WITH H-MSTN AND CMAM

Variation	UCF AUC (%)	UCF AP (%)	UCF Ano-AUC (%)	XD AUC (%)	XD AP (%)	XD Ano-AUC (%)
Before	90.96	32.72	72.62	95.25	86.47	87.57
After	91.45	35.41	75.82	96.75	87.21	89.38

TABLE X. PERFORMANCE COMPARISON (BEFORE AND AFTER) OF PROPOSED FUSION MODULES

Variation	UCF AUC (%)	UCF AP (%)	UCF Ano-AUC (%)	XD AUC (%)	XD AP (%)	XD Ano-AUC (%)
Baseline (H-MSTN)	90.30	33.20	73.10	95.20	86.00	87.40
CMAM (without MAFT)	91.45	35.41	75.82	96.75	87.21	89.38
MAFT (without CMAM)	91.98	34.90	74.90	96.90	87.80	89.70
H-MSTN + CMAM + MAFT (Proposed)	**92.42**	**35.22**	**74.74**	**97.36**	**88.63**	**90.21**

Table VIII compares the performance of different batch sizes on the UCF-Crime and XD-Violence datasets using AUC and AP metrics. For UCF-Crime, the best performance is observed at batch size 64, achieving 92.42% AUC and 35.22% AP. Smaller or larger batch sizes result in slightly lower performance. It indicates that 64 offers an optimal balance for this dataset. For XD-Violence, the performance consistently improves with batch size, peaking at batch size 128 with 97.36% AUC and 88.63% AP. This suggests that a moderately large batch size is most effective for complex multimodal data. Larger

batch sizes beyond 128 do not significantly improve results. It indicates a saturation point in training effectiveness.

TABLE XI. TEXT ENCODER ABLATION STUDY ON UCF-CRIME AND XD-VIOLENCE DATASETS

Text Encoder	UCF-Crime AUC (%)	XD-Violence AP (%)
BERT-base	90.85	86.90
SimCSE	91.34	87.42
Sentence-T5 (Ours)	92.42	88.63

Table IX presents a performance comparison before and after applying prompt-based captioning. On the UCF-Crime dataset, the AUC improves from 90.96% to 91.45%, AP increases from 32.72% to 35.41%, and anomaly-specific AUC (Ano-AUC) rises from 72.62% to 75.82%. On the XD-Violence dataset, the AUC increases from 95.25% to 96.75%, AP improves from 86.47% to 87.21%, and Ano-AUC goes up from 87.57% to 89.38%. These results show that incorporating prompt-based captioning with the H-MSTN and CMAM modules leads to consistent improvements across all evaluation metrics and both datasets.

C. Qualitative Analysis

Figure 5 and Figure 6 provide a detailed visualization of the performance of a proposed VAD method on the UCF-Crime and XD-Violence datasets. In each subplot, the x-axis labeled "#Frames" (ranging from 0 to 1,500) tracks the chronological progression through video frames. The y-axis labeled "Anomaly Detection Score" (ranging from 0 to 1) reflects the model's confidence in identifying anomalies for each frame. A detection score near 1 signals a high likelihood of an anomaly, and a score near 0 indicates normalcy. Additionally, each figure displays representative video frames for various event types such as explosion, riot, car accident, abuse, arrest, vandalism, and assault. In these plots, the cyan shaded regions mark the ground-truth abnormal intervals, while the overlaid line graphs represent the anomaly scores predicted by the model. Arrows connect key frames to their corresponding positions on the timeline. It illustrates how different events align with the predicted anomalies. The visualizations demonstrate that the model's scores rise sharply during true abnormal periods and remain low during normal segments. It demonstrates its ability to temporally localize different types of real-world anomalies and violent incidents effectively.

Despite strong overall performance, the proposed framework encounters limitations in certain challenging scenarios. Class-wise analysis (Figure 10) reveals higher false positive rates (FPR) for visually ambiguous events such as Arson and elevated false negative rates (FNR) for categories such as Arrest. These patterns indicate that events with subtle motion cues or visually similar background dynamics remain difficult to distinguish.

In particular, abrupt camera motion or crowded scenes may lead to false positives, as rapid movements resemble anomalous behavior. Conversely, subtle or partially occluded abnormal activities may result in missed detections due to weak temporal contrast. These observations highlight opportunities for improving fine-grained temporal modeling and enhancing robustness to visually ambiguous scenarios. Moreover, certain anomaly types are inherently harder due to overlapping visual and semantic patterns. Arson and Assault may share motion or illumination characteristics with normal scenes, increasing false positives, while subtle interactions such as Arrest can lack strong temporal cues, leading to higher false negatives under weak supervision. We further analyzed representative failure cases. False positives are often triggered by sudden illumination changes, camera motion, or crowd activity that visually resemble anomalous patterns (e.g., bright reflections misclassified as Arson). False negatives typically occur when abnormal events are subtle, partially occluded, or lack strong motion contrast, such as restrained interactions in Arrest scenarios. Compared to visual-only baselines, the proposed multimodal framework shows greater robustness in semantically complex scenarios. Baseline methods often misclassify ambiguous motion patterns due to reliance on low-level temporal cues, whereas the integration of caption-based semantics and cross-modal attention improves alignment between visual content and event context, reducing such errors.

Backbone Dependency and Generalization: The current framework relies on fixed pre-trained encoders (Video-CLIP for visual features and Sentence-T5 for textual representations). As a result, the overall performance is partly dependent on the representational quality and domain alignment of these specific backbones. Although encoder freezing improves efficiency and stability under weak supervision, it may limit adaptability to alternative feature spaces. A systematic evaluation of different visual and language backbones (e.g., I3D, ViT variants, BERT-based encoders, or domain-adapted vision–language models) is an important direction for future work to assess backbone sensitivity and generalization across surveillance domains.

D. Ablation Studies

Ablation Studies are conducted to evaluate the individual contributions of different components in the proposed model. This helps validate the effectiveness of our proposed model. Table X reports the ablation of CMAM and MAFT. Starting from the baseline, adding CMAM alone improves performance across all three metrics on both datasets by enhancing mid-level cross-modal alignment between visual and textual features. Incorporating MAFT alone further boosts the results. The combined CMAM+MAFT model achieves the highest performance. On UCF-Crime, the full model increases AUC from 91.45% (with CMAM only) to 92.42%, while maintaining comparable AP and Ano-AUC values. For XD-Violence, all metrics consistently improve, with AUC rising from 96.75% to 97.36%, AP increasing from 87.21% to 88.63%, and Ano-AUC improving from 89.38% to 90.21%. The results confirm that CMAM and MAFT are complementary, and their combination performs best. Although the improvement of MAFT over CMAM on UCF-Crime is relatively modest (+0.97% AUC), the gain is obtained under identical training settings and is consistent across both datasets. Since results are reported as the mean of five runs with low variance, the improvement reflects stable complementary contributions from semantic alignment and gated multimodal fusion rather than random fluctuation.

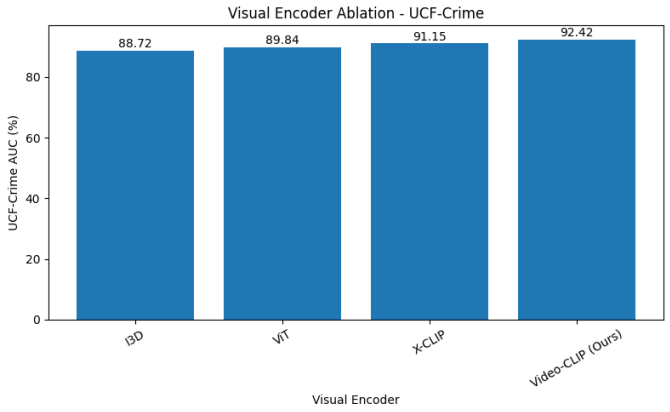

Fig. 7. Comparison of visual feature extractors on UCF-Crime

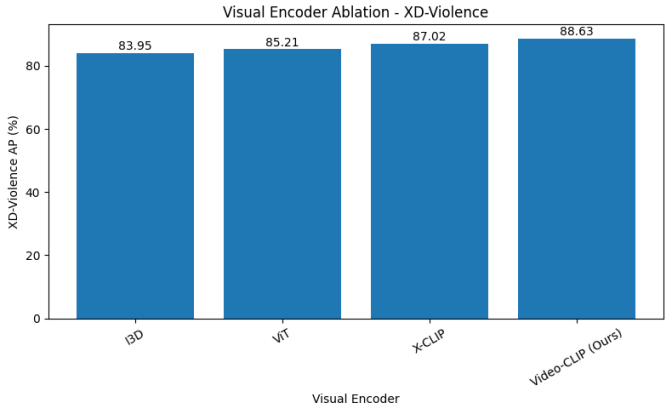

Fig. 8. Comparison of visual feature extractors on XD-Violence

Table XI presents a comparison of different textual encoders under identical training settings. Sentence-T5 achieves the highest performance, obtaining 92.42% AUC on UCF-Crime and 88.63% AP on XD-Violence, outperforming SimCSE by approximately 1.1% and BERT-base by around 1.6%–1.7% across metrics. The consistent improvement indicates that richer sentence-level semantic representations improve cross-modal alignment and anomaly-focused reasoning. These results validate the selection of Sentence-T5 as the most effective textual encoder in the proposed framework.

Table XII compares different temporal encoders under identical experimental settings. H-MSTN achieves the best performance with 92.42% AUC on UCF-Crime and 88.63% AP on XD-Violence, outperforming the pure Transformer by approximately 1.3%–1.7% and LSTM/1D-CNN by larger margins. The improvement confirms that hierarchical multi-scale temporal modelling captures both short-term dynamics and long-range dependencies more effectively than single-scale architectures. These results validate the necessity of H-MSTN in the proposed framework.

TABLE XII. TEMPORAL ENCODER ABLATION STUDY

Temporal Encoder	UCF-Crime AUC (%)	XD-Violence AP (%)
1D-CNN	88.96	84.72
LSTM	89.84	85.63
Transformer	91.15	86.90
H-MSTN (Ours)	92.42	88.63

Figures 7 and 8 compare different visual feature extractors under identical experimental settings. Video-CLIP achieves the best performance with 92.42% AUC on UCF-Crime and 88.63% AP on XD-Violence, outperforming X-CLIP by approximately 1–1.5% and I3D/ViT by larger margins. The results indicate that vision–language pretraining provides stronger semantic representations for anomaly detection. These findings validate the choice of Video-CLIP in the proposed framework.

Table XIII shows the individual and combined effects of prompt-based captioning and visual features on anomaly detection performance for UCF-Crime and XD-Violence datasets. Using Prompt Caption only gives 84.21% AUC and 32.24% AP on UCF-Crime, and 85.72% AUC and 83.57% AP on XD-Violence. The Vision only setting performs slightly better on UCF-Crime with 85.76% AUC and 31.21% AP, while yielding 83.65% AUC and 85.10% AP on XD-Violence. When both modalities are combined (Prompt + Vision), the model achieves the highest performance: 92.42% AUC and 35.22% AP on UCF-Crime, and 97.36% AUC and 88.63% AP on XD-Violence. These results demonstrate that integrating prompt-based captions with visual features leads to significant improvements over using either modality alone.

Table XIV compares model performance across different temporal scales on the UCF-Crime and XD-Violence datasets. When using a single temporal scale, the mid-scale performs best with 87.5% AUC on UCF-Crime and 83.2% AP on XD-Violence, while fine and long scales achieve slightly lower scores. Combining two scales consistently improves results. The Fine + Mid combination reaching 89.7% AUC and 85.9% AP, and Mid + Long and Fine + Long combinations also outperformed single scales. The best performance occurs when all three temporal scales—fine, mid, and long are integrated. This achieve 92.42% AUC on UCF-Crime and 88.63% AP on XD-Violence. This shows that multi-scale temporal fusion effectively captures complex temporal patterns.

TABLE XIII. ABLATION STUDY ON UCF-CRIME AND XD-VIOLENCE DATASETS

Method	UCF AUC (%)	UCF AP (%)	XD AUC (%)	XD AP (%)
Prompt Caption only	84.21	32.24	85.72	83.57
Vision only	85.76	31.21	83.65	85.10
Prompt + Vision (Ours)	**92.42**	**35.22**	**97.36**	**88.63**

TABLE XIV. ABLATION STUDY ON UCF-CRIME AND XD-VIOLENCE DATASETS

Temporal Scale	UCF-Crime AUC (%)	XD-Violence AP (%)
Fine only	86.2	81.6
Mid only	87.5	83.2
Long only	85.3	82.4
Fine + Mid	89.7	85.9
Mid + Long	88.8	85.5
Fine + Long	88.1	85.0
Fine + Mid + Long (All)	**92.42**	**88.63**

Table XV presents a sensitivity analysis of the attention dimension d within the CMAM and its impact on anomaly detection performance across two benchmark datasets: UCF-Crime and XD-Violence. As the attention dimension increases from 64 to 256, performance improves significantly. The UCF-Crime AUC rises from about 88.30% to 92.42%, while the XD-Violence AP goes up from roughly 83.75% to 88.63%. This indicates that a moderate increase in attention capacity improves cross-modal interaction learning. However, when the dimension reaches 512, there is a slight drop in performance. The UCF-Crime AUC is around 91.80%, and the XD-Violence AP is about 87.50%. This suggests potential overfitting or repeated features. The peak performance at d=256 shows it as the best balance between representational power and computational cost. The best experimental result achieves an AUC of 92.42% on the UCF-Crime dataset and an AP of 88.63% on the XD-Violence dataset. This trend aligns with prior observations in multi-modal learning frameworks, where attention dimensionality critically influences alignment quality and detection accuracy.

TABLE XV. EFFECT OF ATTENTION DIMENSION ON PERFORMANCE

Attention Dimension	UCF-Crime AUC (%)	XD-Violence AP (%)
64	~88.30	~83.75
128	~90.20	~86.40
256	**92.42**	**88.63**
512	~91.80	~87.50

Table XVI compares the performance of different fusion strategies used in the proposed VAD framework: early fusion, late fusion, and MAFT. The early fusion strategy, where visual and textual features are combined at the input level, achieving an AUC of 86.20% on the UCF-Crime dataset and an AP of 82.60% on XD-Violence. The late fusion strategy, which processes the modalities independently before combining them at the decision level, shows a slight improvement with an AUC of 87.85% and AP of 85.00%. However, the MAFT strategy proposed in this work outperforms both by a significant margin, achieving 92.42% AUC on UCF-Crime and 88.63% AP on XD-Violence. These results highlight the effectiveness of MAFT in dynamically aligning and integrating visual and textual features.

TABLE XVI. PERFORMANCE COMPARISON OF FUSION STRATEGIES

Fusion Strategy	UCF-Crime AUC (%)	XD-Violence AP (%)
Early fusion	~86.20	~82.60
Late fusion	~87.85	~85.00
MAFT (our work)	**92.42**	**88.63**

Table XVII compares Top-k pooling with attention-based and bilinear pooling strategies. The results show that Top-k pooling achieves superior and more stable performance, consistent with the sensitivity analysis in Figure 9 and the sparse anomaly assumption under MIL.

TABLE XVII. COMPARISON OF DIFFERENT POOLING STRATEGIES UNDER THE MIL FRAMEWORK

Pooling Strategy	UCF-Crime AUC (%)	XD-Violence AP (%)
Attention-based Pooling	91.78	87.21
Bilinear Pooling	90.96	86.14
Top-k Pooling (Ours)	92.42	88.5

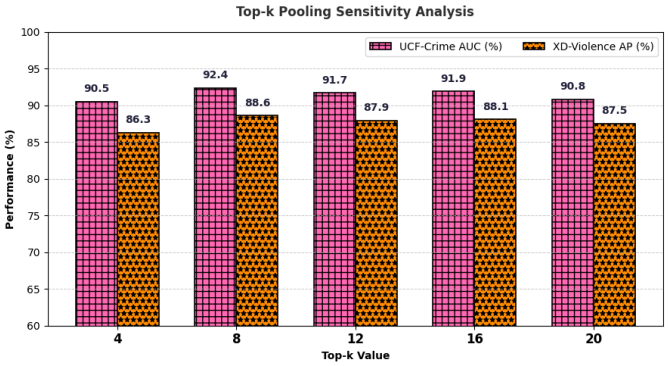

Fig. 9. Sensitivity analysis of Top-k pooling on UCF-Crime and XD-Violence dataset

Figure 9 presents a sensitivity analysis of the Top-k pooling parameter on the UCF-Crime and XD-Violence datasets. It illustrates how different Top-k values affect anomaly detection performance. The x-axis shows Top-k values ranging from 4 to 20, while the y-axis indicates performance in percentage, with AUC (%) for UCF-Crime and AP (%) for XD-Violence. The results demonstrate that performance varies with the choice of k, peaking at a Top-k value of 8 for UCF-Crime (92.4% AUC) and 12 for XD-Violence (88.5% AP). Lower (k=4) or higher (k=20) values tend to reduce performance. It indicates that an appropriate Top-k value is crucial for balancing the inclusion of relevant anomalous frames without introducing excessive noise. This analysis confirms the importance of tuning the Top-k parameter to achieve optimal detection accuracy across different datasets.

Table XVIII compares the performance of standard captioning and prompt-based captioning (the proposed method) in the context of VAD. The standard captioning method generates generic captions for video snippets without any specific task guidance. It achieves an AUC of approximately 88.60% on the UCF-Crime dataset and an AP of around 85.90% on the XD-Violence dataset. In contrast, the prompt-based captioning method, which generates captions based on task-specific prompts, significantly outperforms the standard approach. By focusing on anomaly-relevant behaviors, it achieves a 92.42% AUC on UCF-Crime and 88.63% AP on XD-Violence. These results demonstrate that prompt-based captioning enhances anomaly detection by providing more informative and contextually rich captions.

TABLE XVIII. PERFORMANCE COMPARISON OF CAPTIONING METHODS

Captioning Method	UCF-Crime AUC (%)	XD-Violence AP (%)
Standard Captioning	~88.60	~85.90
Prompt-based (Proposed)	92.42	88.63

Figure 10 presents a class-wise comparison of False Positive Rate (FPR) and False Negative Rate (FNR) percentages for five categories: Abuse, Arrest, Arson, Assault, and Burglary. The orange dashed line with triangle markers represents FPR, while the blue dotted line with star markers indicates FNR. Among the classes, Arson exhibits the highest FPR at 14.7%, labeled as "Max FPR", while Arrest shows the highest FNR at 15.2%,

marked as “Max FNR". The relatively higher FPR for Arson (14.7%) is mainly due to visually similar patterns such as bright illumination, smoke-like textures, or fire-colored regions that may also appear in normal scenes. Lighting variations, reflections, or environmental effects (e.g., dust or fog) can partially resemble fire cues, leading to false alarms under weak supervision. The lowest FPR occurs in the Arrest class (8.9%), whereas the lowest FNR is observed in the Abuse class (9.3%). The graph highlights varying trends between FPR and FNR across different classes. This visual comparison aids in identifying which categories are more prone to false alarms (FPR) or missed detections (FNR), offering valuable insights for improving classification accuracy.

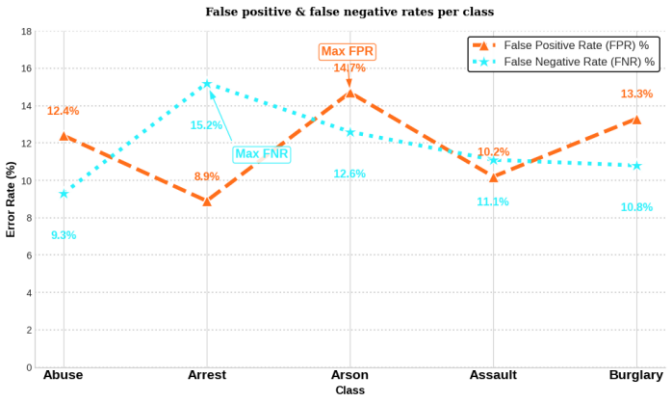

Fig. 10. Class-wise Comparison of False Positive and False Negative Rates (\%)

Figure 11 shows the cross-modal attention heatmap created by the CMAM in the proposed VAD framework. It visualizes the attention between visual snippets and caption tokens. Each cell reflects its interaction strength. Darker areas indicate weaker correlations, and brighter areas show stronger visual-text connections. These results demonstrate how CMAM captures meaningful multimodal relationships and improves anomaly detection through vision-language integration. Beyond aggregate heatmap visualization, the learned cross-modal attention also provides word-level interpretability. We observe that caption tokens associated with abnormal actions or unsafe behaviors (e.g., “fighting,” “running aggressively,” “explosion,” “assault”) consistently receive higher attention weights when computing anomaly scores.

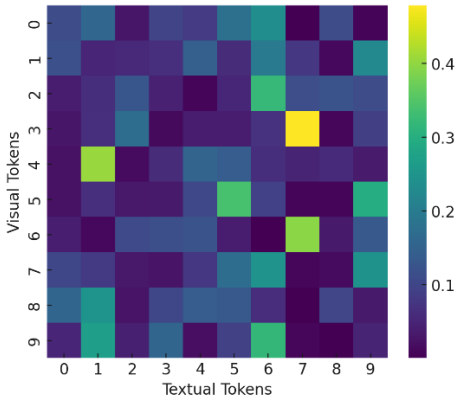

Fig. 11. Attention Heatmap

In contrast, context-neutral or background-related words (e.g., “street,” “people,” “building”) contribute less to the final prediction. This indicates that the CMAM selectively emphasizes semantically meaningful words in the captions that are most relevant for anomaly detection, thereby improving interpretability without requiring additional supervision.

Figure 12 illustrates the detailed class-wise Anomaly Area Under Curve (Ano-AUC) performance of four top methods — UB-MIL [32], DMU [50], ReFLIP-VAD [33], and the proposed method (Ours) on the UCF-Crime dataset. Each bar represents the Ano-AUC (anomaly area under the curve, in percent) achieved by each method across different event categories such as Abuse, Arrest, Arson, Assault, Burglary, and others. The proposed method (green bars) consistently outperforms the other three across all 13 crime categories and the overall average. For instance, in the ‘RoadAccident’ class, the proposed method achieves the highest Ano-AUC, approaching 90%, compared to ~84% for ReFLIP-VAD, ~80% for DMU, and ~74% for UB-MIL. Similarly, in the ‘Fighting’ class, the proposed method scores approximately 86%, which is better than ReFLIP-VAD (~82%), DMU (~78%), and UB-MIL (~70%). For most classes, including Abuse, Arrest, Arson, and Shoplifting, the proposed method maintains 4–10% higher Ano-AUC than ReFLIP-VAD, which is the next best. The ‘Explosion’ and ‘Shooting’ categories also show strong performance (exceeding 85%). On the Average metric, the proposed method reaches approximately 83%, while ReFLIP-VAD, DMU, and UB-MIL score around 78%, 74%, and 69%, respectively. This consistent superiority highlights the robustness and effectiveness of the proposed approach in detecting anomalies across diverse types of crimes.

Figure 13 shows the class-wise average precision (AP) comparison of four leading methods: CLIP-TSA [47], DMU [50], ReFLIP-VAD [33], and the proposed method tested on the XD-Violence dataset. Each group of bars represents a specific event class, including Abuse, Car Accident, Explosion, Fighting, Riot, Shooting, and an overall average. The results indicate that the proposed method achieves the highest AP in all seven classes. It reaches about 88% AP in “Explosion" and 87% in “Shooting," exceeding ReFLIP-VAD, DMU, and CLIP-TSA by 3 to 10%. In “Fighting," it nearly hits 90%, the highest among all models. Similar gains appear in “Abuse," “Car Accident," “Riot," and “Normal." Overall, the method maintains an average AP above 85%, outperforming ReFLIP-VAD (around 82%), DMU (around 78%), and CLIP-TSA (around 74%). These results clearly indicate the superior effectiveness and generalization ability of the proposed model in violence event detection scenarios.

E. Discussion of gains

The performance gains of the proposed framework arise from the complementary contributions of temporal modeling, text guidance, and multimodal fusion rather than from a single component. Prompt-based captioning improves anomaly relevance in the textual modality, yielding consistent gains of 2–3% (Table IX). Hierarchical multi-scale temporal modeling (H-MSTN) enhances the capture of both short and long-duration anomalies, contributing stable improvements across datasets (Table XIV).

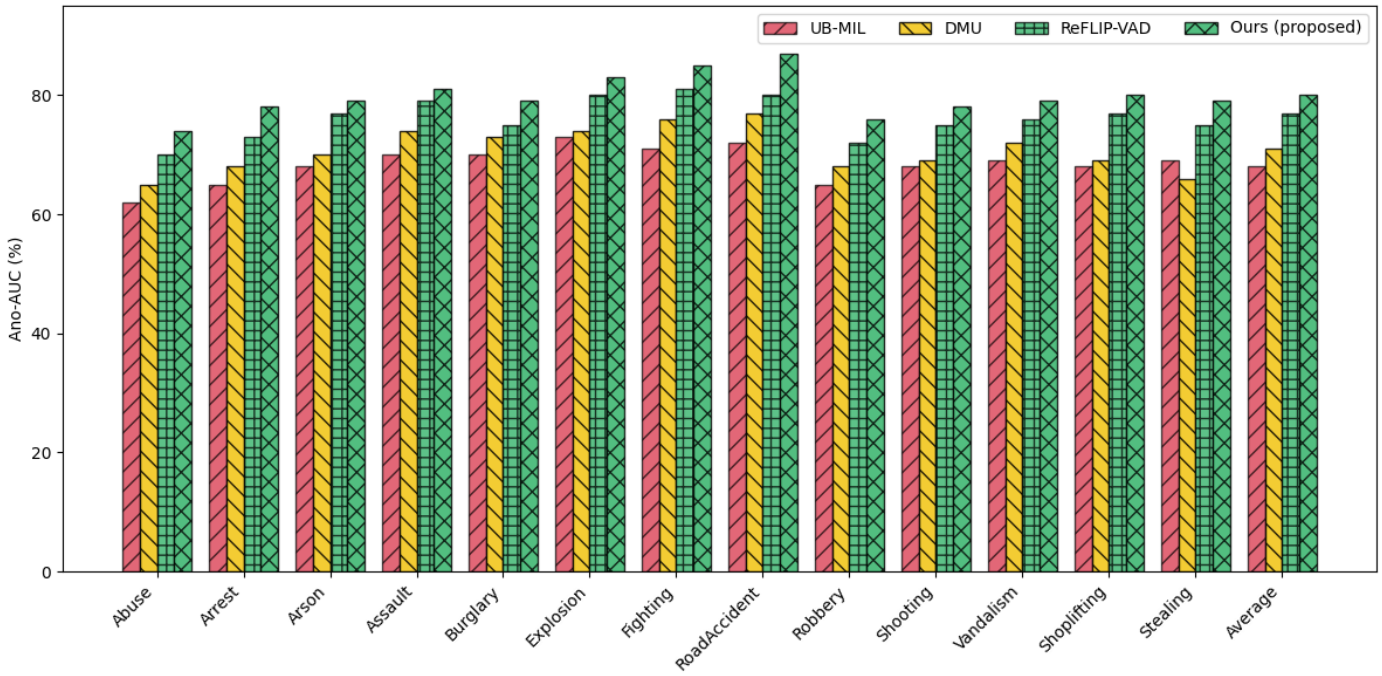

Fig. 12. Detailed Class-wise Anomaly Ano-AUC Performance of Top Methods on UCF-Crime Dataset

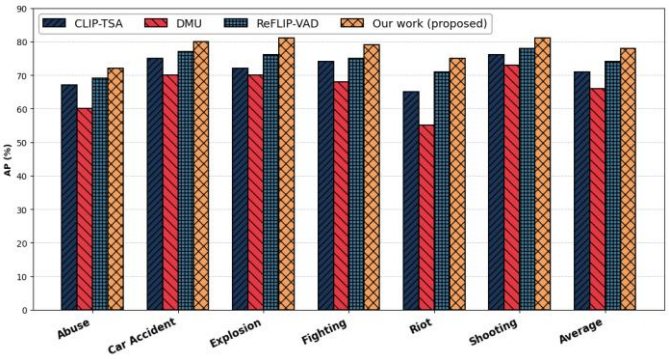

Fig. 13. Detailed Class-wise Average Precision (AP) of Top Methods on XD-Violence Dataset

The largest and most consistent gains stem from the proposed MAFT fusion mechanism, which enables global cross-modal reasoning beyond early or late fusion and accounts for the majority of the final performance improvement (Table X). Overall, the ablation results demonstrate that the improvements are not due to a single factor but emerge from the synergy of (i) anomaly-focused captions, (ii) mid-level cross-modal alignment (CMAM), and (iii) late-fusion with MAFT. This design choice explains why our model achieves higher robustness across both short, highly visual anomalies and long, semantically subtle anomalies compared to existing multimodal baselines.

F. Deployment Considerations and Practical Feasibility

The proposed framework allows for near real-time deployment. It processes video snippets one after the other, without needing future frames. It achieves inference times of under 15 ms per snippet on a standard GPU. Its modular, sliding-window design makes it easy to integrate into online surveillance systems. Feature extraction and captioning can also be handled in edge-cloud settings if needed. Robustness to noisy or low-resolution inputs is ensured through CLIP-based semantic representations, hierarchical temporal modelling, and multimodal fusion, which collectively reduce sensitivity to visual degradation and stabilize anomaly predictions.

G. Limitations, Bias, and Societal Implications

The proposed framework relies on the quality of generated captions, and inaccurate descriptions may affect semantic alignment. Its use of SwinBERT, Video-CLIP, and transformer-based fusion increases computational cost, which can limit deployment on resource-constrained systems. The model may also face domain shift when applied beyond surveillance scenarios. Class-wise analysis reveals higher FNR/FPR for visually ambiguous events such as Arrest and Arson, highlighting potential failure cases.

Beyond technical limitations, VAD systems raise important societal and ethical considerations, particularly in surveillance and public-safety contexts. Bias may arise from imbalanced datasets, subjective definitions of abnormality, or class-dependent error distributions, as reflected in the higher false negative rates for categories such as Arrest and the elevated false positive rates for Arson (Figure 5). In addition, reliance on captioning and large pre-trained vision–language models may inherit semantic or contextual biases present in the training data. Large-scale deployment also introduces privacy concerns and risks of misuse if anomaly predictions are interpreted without contextual awareness. We therefore emphasize the importance of domain-specific calibration, human-in-the-loop decision-making, and transparent threshold selection to ensure responsible and fair real-world deployment.

V. Conclusion and Future Work

This work presents a unified multimodal video anomaly detection framework that integrates hierarchical temporal modeling (H-MSTN) with cross-modal alignment (CMAM) and multimodal fusion (MAFT) to jointly exploit visual and textual information in untrimmed surveillance videos. Prompt-based captioning further enhances the semantic relevance of textual representations. Extensive experiments on the UCF-Crime and XD-Violence benchmarks demonstrate that the proposed approach surpasses prior visual-only and caption-guided methods in frame-level anomaly detection performance. Compared with the strongest recent baseline ReFLIP-VAD, the proposed framework achieves statistically comparable detection accuracy while requiring substantially lower computational complexity and faster inference. These results establish an improved efficiency–accuracy trade-off, highlighting the suitability of the framework for practical real-time multimodal anomaly detection under weak supervision.

Future work may involve integrating additional modalities such as audio or sensor data to enhance contextual understanding. We also plan to explore cross-dataset evaluation to further assess the generalization capability of the proposed framework under domain shift. Additionally, advancing the framework with stronger video–language pretraining and leveraging large language models can further improve semantic reasoning and generalization. We plan to explore adaptive prompt learning techniques that automatically generate context-aware prompts. Lastly, extending the framework to real-time anomaly detection settings and broader application domains like healthcare, industrial monitoring, and traffic analysis are promising avenues for future exploration.

References

[1] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

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

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

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

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

[8] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. 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, 2021.

[9] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.

[10] Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34, 2025.

[11] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[12] Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299–5318, 2024.

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

[14] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14592–14601, 2023.

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

[16] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end trans- formers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022.

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[18] Peng Wu, Jing Liu, Xiangteng He, Yuxin Peng, Peng Wang, and Yanning Zhang. Toward video anomaly retrieval from video anomaly detection: New benchmarks and model. *IEEE Transactions on Image Processing*, 33:2213–2225, 2024.

[19] Dicong Wang, Qilong Wang, Qinghua Hu, and Kaijun Wu. Multimodal vad: Visual anomaly detection in intelligent monitoring system via audio-vision-language. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[20] Ata-Ur Rehman, Hafiz Sami Ullah, Haroon Farooq, Muhammad Salman Khan, Tayyeb Mahmood, and Hafiz Owais Ahmed Khan. Multi-modal anomaly detection by using audio and visual cues. *IEEE Access*, 9:30587–30603, 2021.

[21] Peng Wu, Wanshun Su, Guansong Pang, Yujia Sun, Qingsen Yan, Peng Wang, and Yanning Zhang. Avadclip: Audio-visual collaboration for robust video anomaly detection. *arXiv preprint arXiv:2504.04495*, 2025.

[22] Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024.

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learn- ing with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[24] Debojyoti Biswas and Jelena Tešić. Unsupervised domain adaptation with debiased contrastive learning and support-set guided pseudo- l a beling for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[25] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022.

[26] Noussaiba Jaafar and Zied Lachiri. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, 211:118523, 2023.

[27] Swalpa Kumar Roy, Ankur Deria, Chiranjibi Shah, Juan M Haut, Qian Du, and Antonio Plaza. Spectral–spatial morphological attention transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[28] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pages 322–339. Springer, 2020.

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

[30] Peng Wu, Jing Liu, Mingming Li, Yujia Sun, and Fang Shen. Fast sparse coding networks for anomaly detection in videos. *Pattern Recognition*, 107:107515, 2020.

[31] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021.

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

[33] Prabhu Prasad Dev, Raju Hazari, and Pranesh Das. Reflip-vad: Towards weakly supervised video anomaly detection via vision-language model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[34] Kun-Lun Li, Hou-Kuan Huang, Sheng-Feng Tian, and Wei Xu. Improving one-class svm for anomaly detection. In *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, volume 5, pages 3077–3081. IEEE, 2003.

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

[36] Yang Liu, Jing Liu, Jieyu Lin, Mengyang Zhao, and Liang Song. Appearance-motion united auto-encoder framework for video anomaly detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(5):2498–2502, 2022.

[37] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021.

[38] Lin Wang, Xiangjun Wang, Feng Liu, Mingyang Li, Xin Hao, and Nianfu Zhao. Attention-guided mil weakly supervised visual anomaly detection. *Measurement*, 209:112500, 2023.

[39] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. 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, 2024.

[40] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.

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

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

[43] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022.

[44] Shuo Li, Fang Liu, and Licheng Jiao. 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, 2022.

[45] Yang Zhen, Yuanfang Guo, Jinjie Wei, Xiuguo Bao, and Di Huang. Multi-scale background suppression anomaly detection in surveillance videos. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1114–1118. IEEE, 2021.

[46] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18899–18908, 2024.

[47] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. 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, 2023.

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

[49] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 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, 2019.

[50] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.

APPENDIX

APPENDIX A: PROMPT TEMPLATES AND SELECTION STRATEGY

Prompt-based captioning is used to generate anomaly-focused textual descriptions for each video snippet. We design a small set of task-specific prompt templates that explicitly guide the captioning model to focus on unusual or abnormal behaviors rather than generic scene descriptions.

The following prompt templates are used consistently across all experiments:

"Describe any unusual or abnormal behavior occurring in the video."

"What suspicious or unexpected activity is happening in this scene?"

"Explain if there is any violent, dangerous, or anomalous action in the video."

"Describe the event in the video, focusing on abnormal or unsafe behavior."

During training and inference, a single prompt is randomly selected from this fixed prompt pool for each video snippet. This simple randomization improves robustness to prompt phrasing while keeping the prompt space controlled and reproducible. No prompt tuning or dataset-specific prompt engineering is performed. All prompts are manually designed before experimentation and remain fixed throughout training and evaluation. This ensures that performance gains arise from multimodal learning and temporal modeling rather than prompt overfitting.

APPENDIX B: PROPOSED MULTIMODAL VIDEO ANOMALY DETECTION FRAMEWORK

Algorithm 1: Proposed Multimodal Video Anomaly Detection Framework

Input: Untrimmed video $V = \{v_1, v_2, \ldots, v_T\}$
Output: Video-level anomaly score $\hat{Y}$

Step 1: Caption Generation for *each snippet* v_t **do**
  Generate caption $s_t = \mathcal{C}(v_t \mid \mathcal{P})$ using prompt-based SwinBERT

Step 2: Feature Extraction for *each snippet* v_t **do**
  Visual feature: $x_t^v = \text{VideoCLIP}(v_t)$
  Textual feature: $x_t^l = \text{SentenceT5}(s_t)$

Form visual stream $X_v = \{x_t^v\}$ and textual stream $X_l = \{x_t^l\}$

Step 3: Hierarchical Multi-Scale Temporal Modeling $\bar{X}_v = \text{H-MSTN}(X_v)$

$\bar{X}_l = \text{H-MSTN}(X_l)$

Step 4: Cross-Modal Alignment (CMAM) Compute cross-modal attention

$A = \text{softmax}\left(\frac{Q_v K_l^\top}{\sqrt{d}}\right) V_l$

Aligned visual features: $\tilde{X}_v = A(\bar{X}_v, \bar{X}_l)$

Step 5: Multimodal Fusion using MAFT
Concatenate tokens: $Z_t = [\tilde{x}_t^v; x_t^v]$

Apply self-attention: $\hat{X} = \text{SelfAttn}(Z)$

Gated fusion: $F_t = g_t \odot \hat{X}_t + (1 - g_t) \odot x_t^v$

Step 6: Snippet-level Scoring for *each* t **do**
  Compute anomaly score $\hat{y}_t = g_\phi(F_t)$

Step 7: MIL-Based Top-k Aggregation Select top-k scores: $\text{TopK}(\hat{y}, k)$

Video-level score: $\hat{Y} = \frac{1}{k} \sum_{i=1}^{k} \hat{y}_{\text{top-}i}$

Step 8: Loss Computation MIL loss: $\mathcal{L}_{MIL}$

Temporal smoothness loss: $\mathcal{L}_{TV}$

Sparsity constraint: $\mathcal{L}_{sparse}$

Final loss:
$\mathcal{L}_{final} = \mathcal{L}_{MIL} + \lambda_{TV}\mathcal{L}_{TV} + \lambda_{sparse}\mathcal{L}_{sparse}$

return $\hat{Y}$

APPENDIX C: PROMPT TEMPLATE SPECIFICATION

In this work, anomaly-focused captions are generated using a task-specific prompt template applied to each video snippet before textual encoding. The exact prompt template used in all experiments is:

Prompt Template:

"Describe any unusual, suspicious, or abnormal behavior occurring in this video snippet."

To further evaluate robustness during validation, we also experimented with a refined variant:

"Identify and describe any abnormal activities, violent actions, or suspicious events visible in the scene."

After validation-based tuning, the first prompt template was selected for all final experiments on both UCF-Crime and XD-Violence datasets to ensure consistency and fair comparison.

The prompt text is concatenated with the input video snippet and passed to the SwinBERT captioning model for caption generation. Decoding is performed using greedy decoding with a maximum caption length of 30 tokens. The same prompt configuration and decoding strategy are uniformly applied across all snippets and datasets to ensure reproducibility.