AI

MDPI

Article

Gated Fusion Networks for Multi-Modal Violence Detection

Bilal Ahmad [1], Mustaqeem Khan [2,*] and Muhammad Sajjad [1]

[1] Digital Image Processing Laboratory, Department of Computer Science, Islamia College Peshawar, Peshawar 25120, Pakistan; mscs-06-12@icp.edu.pk (B.A.); muhammad.sajjad@icp.edu.pk (M.S.)
[2] College of Information Technology, United Arab Emirates University (UAEU), Al Ain 15551, United Arab Emirates
* Correspondence: mustaqeemkhan@uaeu.ac.ae

Abstract

Public safety and security require an effective monitoring system to detect violence through visual, audio, and motion data. However, current methods often fail to utilize the complementary benefits of visual and auditory modalities, thereby reducing their overall effectiveness. To enhance violence detection, we present a novel multimodal method in this paper that detects motion, audio, and visual information from the input to recognize violence. We designed a framework comprising two specialized components: a gated fusion module and a multi-scale transformer, which enables the efficient detection of violence in multimodal data. To ensure a seamless and effective integration of features, a gated fusion module dynamically adjusts the contribution of each modality. At the same time, a multi-modal transformer utilizes multiple instance learning (MIL) to identify violent behaviors more accurately from input data by capturing complex temporal correlations. Our model fully integrates multi-modal information using these techniques, improving the accuracy of violence detection. In this study, we found that our approach outperformed state-of-the-art methods with an accuracy of 86.85% using the XD-Violence dataset, thereby demonstrating the potential of multi-modal fusion in detecting violence.

Keywords: violence detection; multi-modality; multi-modal fusion; weakly supervised learning

Academic Editors: Honggang Chen and Chao Ren

Received: 28 August 2025
Revised: 19 September 2025
Accepted: 1 October 2025
Published: 3 October 2025

Citation: Ahmad, B.; Khan, M.; Sajjad, M. Gated Fusion Networks for Multi-Modal Violence Detection. *AI* **2025**, *6*, 259. https://doi.org/10.3390/ai6100259

1. Introduction

In recent years, automated systems capable of detecting and preventing violent situations in real time have become increasingly important due to the growth of digital information and the widespread use of social media. Video violence detection has become essential in crime prevention, security monitoring, and content regulation. More detailed approaches have been made possible by emerging multi-modal datasets such as XD-Violence [1], which incorporate visual, audio, and motion information. These datasets extend beyond conventional techniques, which primarily utilize visual data.

While multi-modal systems hold strong potential for violence detection, their effectiveness is impeded by three interrelated challenges. First, information redundancy—overlapping or irrelevant signals across modalities—can drown out discriminative cues. Second, modality imbalance occurs when one modality (e.g., RGB) is substantially more informative than others (e.g., noisy audio), causing naive fusion strategies to be dominated by the stronger modality. Third, temporal asynchrony arises when events are not temporally aligned between modalities (for example, the visual action of a strike may precede an audible reaction), which complicates joint modeling. These problems are exacerbated in weakly supervised settings, where only video-level labels are available, and many snippets

within a labeled video may be non-violent. In this work, we address these challenges with a gated fusion module (GFM) that adaptively weights modality contributions, a multi-scale bottleneck transformer (MSBT) that captures temporal patterns at multiple scales, and a temporal consistency contrast (TCC) loss that encourages temporal alignment across fused features.

To overcome these obstacles, recent methods have included weakly supervised learning to lower annotation costs. Annotating films for violence frame by frame is costly and time-consuming, especially when dealing with big datasets. Using video-level annotations [2], weak supervision bypasses this problem and enables models to predict labels based on video snippets or segments without needing frame-level labeling. Videos are represented as "bags" of instances (snippets or segments) in this framework, which is tagged as a multiple instance learning (MIL) problem. The model learns to determine whether particular examples contribute to categorizing as "violence" or "non-violence." Numerous noteworthy tactics that fit this MIL paradigm have been put forward. To learn motion-aware characteristics and more accurately identify violent behaviors, reference [3] presented a temporal augmentation network that uses attention processes, similarly to [4,5]. Furthermore, sequence learning is used by transformer-based models [6] to improve the accuracy of violence detection. Furthermore, sequence learning is used by transformer-based models [6] to improve the accuracy of violence detection.

Researchers have investigated the combination of audio and visual components in multi-modal techniques to leverage their complementary qualities. The authors in [7] suggested a Graph Convolutional Network (GCN) for learning multi-modal representations using graph-based learning. At the same time, reference [8] introduced a modality-aware multiple-instance learning technique to solve modality asynchrony. Motivated to overcome these obstacles, we propose an enhanced design that incrementally advances the state of the art by integrating a GFM with an MSBT, complemented by a temporal consistency contrast (TCC) loss. While our improvements over prior work are modest in percentage terms (1–3%), they are consistent, achieved across multiple datasets, and derived from unique combinations of adaptive fusion, multi-scale temporal modeling, and temporal alignment. By dynamically adjusting the contribution of each modality, this novel technique ensures that the system can balance the occasionally differing properties of audio and visual streams.

Our method enhances the system's ability to focus on the most relevant data for violence detection, addressing the inherent modality imbalance in the dataset. Beyond violence detection, our task is closely aligned with the broader problem of Weakly Supervised Video Anomaly Detection (WS-VAD), where only coarse video-level labels are available during training. WS-VAD faces unique challenges, particularly the presence of normal snippets within anomalous videos, which can mislead the model. Since the XD-Violence dataset provides segment-level annotations, it has been widely used in both WS-VAD and weakly supervised violence detection studies. Our framework, through gated fusion and multi-scale temporal modeling, is designed to mitigate snippet-level ambiguity by emphasizing informative modalities and temporally consistent cues. This contextualization positions our contribution as a hybrid advancement that bridges the gap between weakly supervised violence detection and the broader WS-VAD literature. Our key contributions include the following:

- **Gated Fusion for Multi-Modal Integration:** Our gated fusion technique successfully integrates multiple features, including visual, audio, and motion, to overcome the inherent modality imbalance and enhance the reliability and dynamics of multimodal data integration.

- **Joint and Single Feature Passing through MSBT:** To enhance the model's ability to identify violence, we employ a multi-scale bottleneck transformer (MSBT) on individual modality features and jointly fused data. This allows the model to capture both modality-specific and holistic information.
- **Temporal Consistency Contrast Loss:** To ensure temporal consistency in the fused features, we introduce a unique TCC loss that addresses modality asynchrony by feature-level alignment of multimodal data. These notable enhancements in the framework demonstrate the effectiveness of our method in weakly supervised settings, where obtaining a labeled dataset can be particularly challenging.

The remainder of this article is divided into sections: Section 2 presents the related work, and Section 3 outlines the proposed framework and its components. Section 4 presents the experimental results and the study of ablations, and Section 5 concludes with the findings and future research directions.

2. Related Work

2.1. Detection of Weakly Supervised Violence

Although collecting accurate frame-level labels in video datasets is challenging, weakly supervised learning has become a crucial method [2] in identifying violence. This approach enables algorithms to detect potentially violent frames or segments and generate predictions at the video level using just coarse labeling. Weakly supervised tasks have substantially used multiple instance learning (MIL). Video-level labels in MIL frameworks guide the learning process for recognizing violent scenes in videos. For example, an MIL-based method is used to identify violence in surveillance footage [9], whereby violent or unusual occurrences are identified using video-level labels. However, this approach frequently fails to convey the intricate temporal and hierarchical links of violent acts. Additionally, multi-modal learning, which utilizes both visual and auditory information, has been investigated as a means to enhance the accuracy of violence detection. For example, it was suggested to use visual and aural signals to identify violent scenes in massive datasets such as XD-Violence [1]. However, issues such as noisy data inputs and modality imbalance, where one modality (like audio) may predominate over the other, remain significant obstacles to achieving peak performance. Recent WS-VAD methods [10–12] leverage MIL formulations to identify anomalies at the snippet level from video-level labels. A key challenge in WS-VAD is that anomalous videos may still contain a majority of normal segments, requiring robust temporal discrimination. For example, RTFM [10] emphasizes discriminative snippet selection, while [13] incorporates hyperbolic space embeddings to improve feature separation. In contrast, our model integrates gated fusion and multi-scale temporal transformers to explicitly handle modality imbalance and snippet ambiguity. This makes our method particularly well-suited to tasks that overlap with both violence detection and the broader WS-VAD problem.

2.2. Transformers for Multi-Modal Visual, Audio, and Motion Data Fusion

Transformers have become known as a significant advancement in multi-modal learning, especially for problems like audio–visual violence detection that require the combination of data from several sources [13]. The transformers' self-attention mechanism is ideal for these activities since it facilitates effective cross-modal interactions. Recent efforts, such as multi-modal bottleneck transformers (MBTs), compress information across modalities using limited bottleneck tokens. Video categorization is one of the multi-modal challenges where this method has demonstrated potential [14]. However, modality imbalance is a common problem with MBTs, where some modalities (such as visual) may overpower others or fail to capture temporal interdependence. These restrictions have been highlighted

in studies on audio-visual learning, such as those conducted by [10,15]. The proposed multi-scale bottleneck transformer (MSBT) enhances the bottleneck token system [16], which utilizes fewer tokens to compress information more efficiently and provides a more balanced representation of auditory and visual modalities [17]. This development enhances earlier transformer models by addressing the issues of modality imbalance and temporal misalignment, which are crucial for identifying violent acts across modalities.

- **Methods of Multi-Modal Fusion:** Most of the current fusion techniques (direct fusion and bilinear pooling-based fusion) are either too computationally intensive or unable to fully utilize cross-modal information [18]. The aforementioned issues can be avoided using an attention-based fusion method. One technique frequently employed in multi-modal fusion is the attention mechanism, which is often used for modal mapping.
- **Single-Model Attention:** Multiple research investigations have demonstrated that integrating an attention module into tasks such as object identification and picture classification can significantly enhance performance. To achieve adaptive fusion of several modalities while minimizing redundant noise, reference [11] created the fusion net, which chooses the k most representative feature maps. Zhou et al. [13] weighted the models for the network to concentrate on more advantageous fields and successfully integrated various models. To obtain additional information between models, reference [19] weighed the feature maps from two-stream Siamese networks using the weight generation subnetwork as inputs. The residual cross-modal connections were then used to acquire the improved features, which were concatenated [3]. To achieve the quality-aware fusion of several models, reference [20] created an instance adaptor that uses two completely linked layers for each modal. They then predicted the modal weight.
- **Cross-Model Attention:** In multimodal fusion, cross-modal mechanisms have become increasingly popular, as they enable efficient communication between various modalities. Wei et al. [21] incorporated a cross-attention mechanism into their co-attention module to facilitate cross-modal interaction. Cao et al. [5] developed a cross-modal encoder to compress features from both modalities. They then used multi-head attention to return the enlarged data to each modality. This two-stage propagation of cross-modal features reduces noise and improves audio-visual features. Similarly, Wu et al. [1] recorded audio–visual interactions using a bimodal attention module to increase the precision of highlight detection. Jiang et al. [22] presented a multi-level cross-spatial attention module that maps the weighted features back to their original dimensions. They calculated cross-modal attention by passing each encoder's data to a cross-modal fusion module.

  Two fusion attention strategies, merged attention and co-attention, were investigated by [23]. While the co-attention module feeds text and visual data into independent transformer blocks and then uses cross-attention to facilitate cross-modal interaction, the merged attention module concatenates and processes these features through a transformer block.

2.3. Learning Graph Representation for Temporal Relationships

Identifying temporal relationships between video frames is essential to detecting violence. Convolutional or recurrent neural networks have been employed in previous studies for temporal modeling; however, these methods often have limitations in capturing complex temporal relationships and long-range dependencies. Graph Convolutional Networks (GCNs) have been proposed to solve this problem by encoding each video frame as a node in a graph. For instance, as demonstrated in [24], where human skeletons are represented

as graph structures, GCNs have been used for action recognition tasks. The proposed framework can better identify violent temporal patterns in the movie by utilizing GCNs in hyperbolic space. A significant change in the way temporal and structural correlations are learned in violence detection is represented by the switch from Euclidean-based convolutional models to hyperbolic GCNs. Recent advances in multi-modal fusion include hyperbolic Graph Convolutional Networks (GCNs) [13], which leverage non-Euclidean geometry to better capture hierarchical relationships. While our framework does not directly employ hyperbolic embeddings, it complements such approaches by tackling modality imbalance and temporal asynchrony through gated fusion and multi-scale temporal reasoning.

2.4. Contrastive Learning in Multiple Modes

In multi-modal learning, contrastive learning has gained popularity as a method for aligning representations across several modalities. Contrastive learning can help the model learn shared representations that enhance class differentiation in violence detection, which uses both aural and visual modalities. For example, reference [25] presented SimCLR, a contrastive learning framework modified for multimodal tasks by ensuring that the feature space aligns representations of comparable events (e.g., violent events) across multiple modalities. Additionally, the current architecture incorporates a temporal consistency contrast loss to ensure that video segment representations remain consistent over time. This invention is especially crucial for detecting violence because temporal discrepancies between auditory and visual modalities may result in incorrect classifications. By incorporating this innovative loss function, the proposed framework enhances the model's ability to detect temporal and geographical patterns.

3. Methodology

This work addresses the challenge of weakly supervised violence detection in [2] scenarios where untrimmed videos containing RGB, optical flow, and audio streams are provided. We propose a comprehensive multi-modal violence detection framework, illustrated in Figure 1, which leverages the complementary strengths of these different modalities. Following established practices, each input video is systematically segmented into T non-overlapping snippets to facilitate temporal analysis. These modality-specific snippets (RGB, flow, and audio) [3] are initially processed through dedicated uni-modal encoders, each comprising a pre-trained feature extraction backbone and a linear projection layer for tokenization, resulting in distinct modality-specific feature representations. Subsequently, these features are integrated within a sophisticated fusion module that employs a multi-scale bottleneck transformer (MSBT) [26] to perform pairwise fusion in all modality combinations. The fusion module then concatenates and appropriately weights these fused features. In the final stage [27], the consolidated features are processed through a global encoder for contextual aggregation before being passed to a regressor that generates the final anomaly scores for violence detection. Furthermore, detecting violence in real-world scenarios presents a multifaceted challenge due to the intricate nature of violent events, varying environmental conditions, and the need to process heterogeneous data streams effectively. The task involves analyzing video sequences from [28] that contain RGB frames, optical flow fields, and audio signals to identify violent anomalies by leveraging complementary information across these modalities. Weak supervision, temporal dependencies, and data inconsistencies, such as varying sampling rates, noise, or missing and corrupted information, further complicate this. In practical surveillance applications, violent incidents manifest themselves through diverse indicators.

- Visual signals in RGB frames, such as aggressive behaviors or physical altercations.
- Motion characteristics in optical flow, including abrupt or erratic movements.
- Auditory cues from audio spectrograms, like screams, collisions, or alarms.

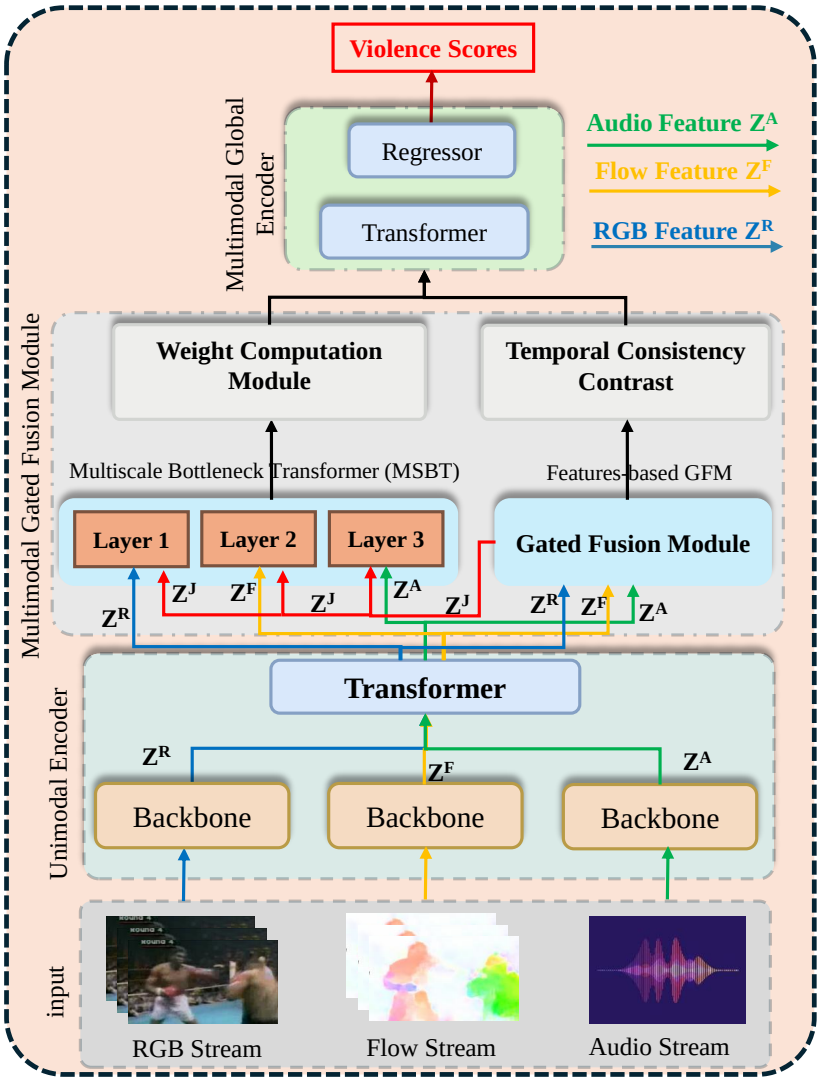

Figure 1. Proposed multi-modal violence detection framework. The framework processes RGB, flow, and audio streams through uni-modal encoders to extract features. These are then fused using a multi-modal gated fusion module enhanced with multi-scale bottleneck transformers (MSBTs). A weighted combination of fused features is computed for final anomaly detection using the multi-modal global encoder. Temporal consistency contrast (TCC) ensures temporal coherence, while multiple instance learning (MIL) Computes frame-level anomaly scores for robust violence detection.

Formally, let the video sequence $\mathcal{V} = \{V_t\}_{t=1}^{T}$ of length T include the following data at each time step t:

$$[X_t^R \in \mathbb{R}^{H\times W\times 3}, \text{ RGB frame.}] \tag{1}$$

$$[X_t^F \in \mathbb{R}^{H\times W\times 2}, \text{ Optical flow field.}] \tag{2}$$

$$[X_t^A \in \mathbb{R}^{L\times D}, \text{ Audio spectrogram.}] \tag{3}$$

The violence detection task can be formulated as learning a mapping function f_θ:

$$s_t = f_\theta(X_t^R, X_t^F, X_t^A) \tag{4}$$

where $s_t \in \mathbb{R}$ represents the violence score at time t and θ denotes the learnable parameters of our framework. The goal is to optimize θ such that s_t effectively discriminates between normal and violent events. For each modality $m \in R, F, A$, we first extract modality-specific embeddings:

$$Z_t^m = E^m(X_t^m) \tag{5}$$

E^m represents the unimodal encoder for modality m and $Z_t^m \in \mathbb{R}^d$ is the corresponding embedding in a shared latent dimension space d. Our multi-modal fusion framework processes these embeddings to generate the final violence scores. The framework is trained in a self-supervised manner using normal samples only, learning to perform the following:

1. **Reconstruct normal patterns across multiple modalities:**
   The model learns to identify and reconstruct patterns characteristic of normal and non-violent behavior within each modality. By training only on standard samples, the framework develops a baseline understanding of typical features, helping it distinguish anomalies when presented with violent events.
2. **Ensure temporal consistency between consecutive frames:**
   The framework is designed to maintain consistency over time, capturing the sequential nature of video data. This involves learning the natural flow of events between frames, such as smooth transitions in motion, visual coherence, and consistent auditory signals. Any disruption in this temporal structure may indicate violent behavior.
3. **Maximize mutual information across complementary modalities:**
   By aligning information from different modalities (e.g., visual, motion, and audio), the model strengthens its ability to detect violence. Maximizing mutual information ensures that the features extracted from one modality, such as optical flow, complement those from other modalities, like RGB frames or audio spectrograms, thereby creating a holistic understanding of the event.

3.1. Techniques for Multi-Modal Fusion

This work presents a summary of various popular [29,30] multi-modal fusion methods applied during the early and middle stages for comparison studies.

Concatenation Fusion:

A straightforward approach to multi-modal integration involves concatenating the features from both modalities and processing them through fully connected (FC) layers. The fused output X is defined as follows:

$$[X = f(X_A \oplus X_V),] \tag{6}$$

where $f(\cdot)$ represents a fully connected two-layer network and $\oplus$ denotes the concatenation operation.

Bilinear and Concatenation Fusion:

Dual-stream linear transformation layers handle heterogeneous input features across modalities to ensure dimensional consistency. Fusion is achieved using concatenation, mathematically represented as follows [13]:

$$[X = UX_A \oplus VX_V,] \tag{7}$$

where U and V are learnable parameter matrices in the feature space. This method preserves the integrity of the feature while facilitating an effective combination of modality-specific characteristics.

Detour Fusion:

This technique addresses modality imbalance, particularly in scenarios such as audio-visual violence detection, where visual signals are often more reliable than noisy audio cues. In this method, visual features transform fully connected layers before concatenation with audio features:

$$[X = f_v(X_V) \oplus X_A,] \tag{8}$$

where f_v represents a fully connected two-layer network designed to refine visual characteristics prior to fusion.

Additive Fusion:

This approach integrates multi-modal features using element-wise addition, ensuring a balanced combination of modality-specific information. The fused representation is expressed as follows:

$$[X = f_a(X_A) + f_v(X_V),] \tag{9}$$

Here $f_a(\cdot)$ and $f_v(\cdot)$ are fully connected neural networks applied to the audio and visual characteristics, respectively. These networks are designed to maintain consistency across dimensions, facilitating efficient and robust feature integration [13].

Unlike simple concatenation or additive fusion, which treat modalities equally, our gated mechanism adaptively assigns importance weights to each modality. This design suppresses irrelevant or noisy modalities (e.g., background audio) while amplifying discriminative cues. Attention-based alternatives, such as cross-modal transformers, improve interactions but introduce high computational costs. In contrast, our gating is lightweight, scalable, and directly embedded into the fusion stage. Empirical comparisons in our ablation study, as seen in Table 1, confirm its superiority.

Table 1. Exploring ablation techniques for various multi-modal fusion methods. By swapping out the original concat fusion for our gated fusion, we re-implement the procedure with *.

Index	Manner	AP (%)	Param. (M)
1	Wu et al. [1]	79.86	0.851
2	Concat Fusion [13]	83.35	0.758
3	Additive Fusion [13]	82.41	0.594
4	Bilinear and Concat [13]	81.33	0.644
5	Detour Fusion [13]	85.67	0.607
6	**Gated Fusion ***	**86.85**	**0.637**

Our Gated Fusion Mechanism: We propose a gated fusion mechanism for integrating multi-modal features from RGB frames, optical flow, and audio modalities, leveraging adaptive feature weighting to enhance representation learning. For each modality $m \in \{\text{RGB, flow, audio}\}$, the mechanism generates a gate vector:

$$[g_m = \sigma(W_m x_m + b_m),] \tag{10}$$

where x_m represents the input features, W_m is a learnable weight matrix, b_m is the bias term, and $\sigma(\cdot)$ denotes the sigmoid activation function. The gated features are then computed via element-wise multiplication:

$$[\tilde{y}_m = g_m \odot x_m,] \tag{11}$$

where $\odot$ represents the Hadamard product.

The final joint representation z is obtained by aggregating the gated features across modalities using summation, followed by a nonlinear transformation:

$$[z = \phi(W_f(\tilde{y}_{\text{rgb}} + \tilde{y}_{\text{flow}} + \tilde{y}_{\text{audio}}) + b_f),] \tag{12}$$

where $\phi(\cdot)$ denotes the ReLU activation function.

As illustrated in Figure 2, this mechanism dynamically selects informative characteristics by learning to prioritize relevant signals ($g_m \approx 1$) while suppressing noise or irrelevant data ($g_m \approx 0$) for each modality. This adaptive weighting enables the model to balance the contributions of different modalities, ensuring robust feature integration during the fusion process. Figure 2 visually depicts the gated fusion pipeline, highlighting the gating mechanism, feature weighting, and final aggregation.

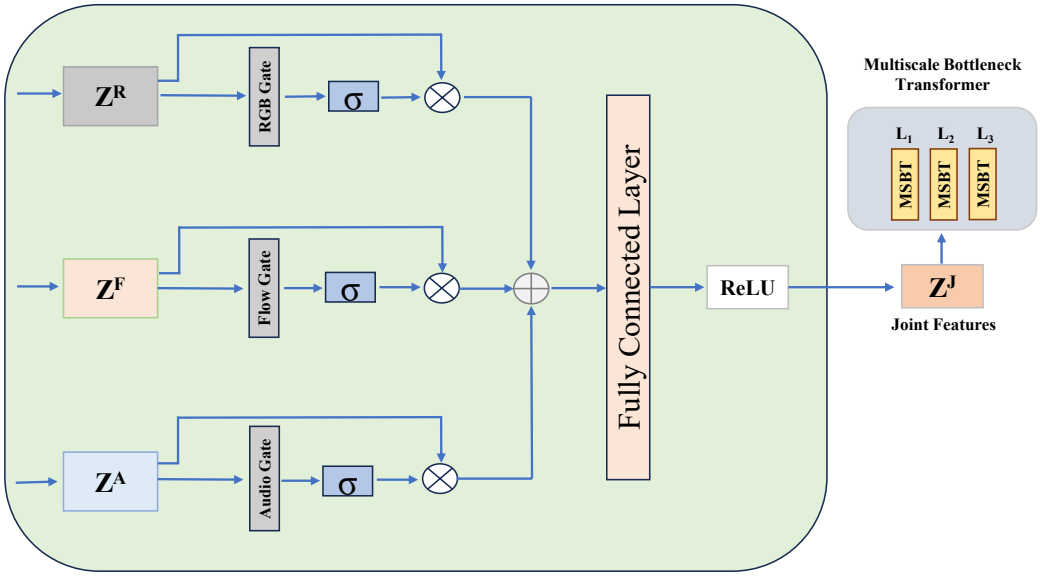

Figure 2. Gated fusion framework for multi-modal feature integration. The mechanism integrates features from various modalities, dynamically adjusting their contributions to improve performance in multi-modal tasks.

This approach enhances the model's ability to effectively leverage complementary information from RGB, optical flow, and audio modalities while mitigating the impact of noise and inconsistencies in individual modalities.

3.2. Offline and Online Detection

Our system combines offline and online detection algorithms to achieve thorough anomaly detection capabilities.

3.2.1. Offline Detection

Although the multi-modal global encoder efficiently handles multi-modal characteristics, we introduce a localized detection approach to enhance its ability to capture long-range dependencies. The following formula is used to calculate the offline detection score between two positions:

$$F_{\text{off}}(i, j) = \exp^{-\alpha \|i-j\|^{\beta}} \tag{13}$$

where i, j indicates the ith and jth feature locations, α and β are hyperparameters governing the distance relationship, and $\|i - j\|$ indicates the spatial separation between positions.

3.2.2. Online Detection

We use an online detection system that goes through many steps to analyze the multi-modal data for real-time surveillance applications.

1. The transition layer transforms the input features as follows:

$$T(\mathbf{x}) = \text{SMU}(\text{BN}(\text{AvgPool}(\text{Conv}(\mathbf{x})))) \tag{14}$$

where Conv is a convolutional layer, AvgPool performs average pooling, BN is batch normalization, and SMU represents The Smooth Monotonic Unit activation function.

2. In the Smoothing Features step, the features are refined using a 1D convolution with a kernel size of 1, expressed as

$$F_{\text{on}} = \text{SMU}(\text{Conv}_{1\times1}(T(\mathbf{x}))) \tag{15}$$

3. For Final Classification, the system uses a causal 1D convolution with a kernel size of 5 to compute the online detection score:

$$\text{Score}_{\text{online}} = \text{Conv}_{\text{causal}}(F_{\text{on}}) \tag{16}$$

where $\text{Conv}_{\text{causal}}$ is a kernel size 5 1D causal convolution. The cumulative detection score combines online and offline features using a weighted sum:

$$\text{Score}_{\text{final}} = \lambda F_{\text{off}} + (1-\lambda)F_{\text{on}} \tag{17}$$

where the weighting parameter λ balances online and offline detection.

Our multi-modal architecture incorporates this dual detection method by

1. Using the fused features of the multi-modal global encoder $\mathbf{Z}^*$.
2. Adding more detection targets to the TCC and MIL losses.

Training optimizes a composite loss that balances temporal alignment, instance-level discrimination, and detection objectives. Concretely, the total loss is a weighted sum of four components:

$$L_{\text{total}} = \lambda_1 L_{\text{TCC}} + \lambda_2 L_{\text{MIL}} + \lambda_3 L_{\text{off}} + \lambda_4 L_{\text{on}}, \tag{18}$$

where L_{TCC} denotes the temporal consistency contrast loss, aligning embeddings across modalities; L_{MIL} represents the multiple instance learning loss for snippet-level discrimination from video-level labels; L_{off} is the offline detection loss leveraging long-range context; and L_{on} is the online detection loss for causal/real-time scoring. The weighting coefficients $\{\lambda_i\}$ control the relative importance of each term, and their values are reported in Section 4.3.

3.3. Uni-Modal Encoder

Our framework's initial processing, the uni-modal encoder, uses three different input modalities from a video segment V: RGB video frames ($X^R \in \mathbb{R}^{T\times H\times W\times 3}$), optical flow ($X^F \in \mathbb{R}^{T\times H\times W\times 2}$), and audio signals ($X^A \in \mathbb{R}^{T\times F}$), where T is the temporal dimension, H and W are spatial dimensions, and F indicates the audio frequency bands. Through a dedicated backbone network, each modality is processed in parallel. This is represented by the formula $B_m(X^m) = F^m \in \mathbb{R}^{T\times D_m}$, where $m \in \{R, F, A\}$ denotes the modality type. Then, a projection layer $P_m(F^m) = P^m \in \mathbb{R}^{T\times D}$ is used to alter the characteristics of the backbone. This is implemented as $P^m = W_m^p F^m + b_m^p$, where $W_m^p \in \mathbb{R}^{D\times D_m}$ and $b_m^p \in \mathbb{R}^D$ are learnable parameters. The transformer layer, written as $z^m = \text{TransformerLayer}_m(P^m)$, is the last step in temporal modeling for each modality. The transformer operation is

$z^m = \text{MHA}(\text{LN}(P^m)) + \text{FFN}(\text{LN}(\text{MHA}(\text{LN}(P^m))))$. We utilize feedforward networks (FFNs), layer normalization (LN), and multi-head attention (MHA). The RGB and flow streams are processed through modified ResNet-50 backbones with $D_R = D_F = 2048$, and the audio stream is processed through a 1D CNN with $D_A = 1024$, which are examples of modality-specific implementations. After projection, these streams are unified to the dimension $D = 512$ to obtain a balanced representation in later fusion stages. Modality-specific temporal dependencies are captured by the resulting embeddings $z^m \in \mathbb{R}^{T \times D}$ and then sent to the multi-modal gated fusion module for cross-modal integration.

3.4. Multi-Modal Gated Fusion Module

The three different feature streams that were acquired via the uni-modal encoder are used to start the multi-modal fusion process: features in RGB $\mathbf{z}^R \in \mathbb{R}^d$, flow features $\mathbf{z}^F \in \mathbb{R}^d$, and audio features $\mathbf{z}^A \in \mathbb{R}^d$, where d is the dimension of the feature. These characteristics first go through a gated fusion process, which is described as $\mathbf{z}^J = GF(\mathbf{z}^R, \mathbf{z}^F, \mathbf{z}^A)$, where GF stands for the gated fusion operation. The gating mechanism is represented as follows:

$$\mathbf{z}^J = \sigma(\mathbf{W}^R \cdot \mathbf{z}^R + \mathbf{b}^R) \odot \mathbf{z}^R + \sigma(\mathbf{W}^F \cdot \mathbf{z}^F + \sigma(\mathbf{W}^A \cdot \mathbf{z}^A + \mathbf{b}^A) \odot \mathbf{z}^A \tag{19}$$

The sigmoid activation function is σ, the element-wise multiplication is indicated by $\odot$, and the learnable weight matrices and bias terms are represented by $\mathbf{W}^*$ and $\mathbf{b}^*$, respectively. Following that, the fused representation $\mathbf{z}^J$ enters the multi-scale bottleneck transformer (MSBT). The multi-scale bottleneck transformer (MSBT) improves upon traditional temporal attention by introducing hierarchical bottleneck tokens that compress and exchange information at multiple temporal scales. Unlike standard bottleneck transformers, our MSBT explicitly balances short-term violent actions (e.g., punches and sudden falls) with long-term scene evolution (e.g., riots and crowd aggression). This multi-scale design ensures robust temporal modeling across varying video lengths while reducing redundancy in cross-modal alignment. It uses many transformer blocks to process the features:

$$MSBT(\mathbf{z}^J) = MSBT_n(MSBT_{n-1})(\ldots MSBT_1(\mathbf{z}^J))) \tag{20}$$

The self-attention methods used by each MSBT block are as follows:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{21}$$

where query, key, and value matrices Q, K, and V are obtained from the input features. The calculation of the multi-head attention mechanism is as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{22}$$

where $\text{head}_i = \text{Attention}(Q\mathbf{W}_i^Q, K\mathbf{W}_i^K, V\mathbf{W}_i^V)$, weight calculation using transform layers and regressors is part of the last fusion step. The method for calculating weight may be stated as follows:

$$\mathbf{w} = \text{Regressor}(\text{Transform}(\mathbf{z})) \tag{23}$$

$$\mathbf{z}^* = \mathbf{w} \odot \mathbf{z} \tag{24}$$

And the output of the weighted feature is represented by $\mathbf{z}^*$. The entire fused representation $\mathbf{Z}$ can be acquired by

$$\mathbf{Z} = FCReLU(\mathbf{z}_R^* + \mathbf{z}_F^* + \mathbf{z}_A^*)) \tag{25}$$

This final representation **Z** is subjected to temporal consistency contrast (TCC) learning, which maximizes the distance between inconsistent features and decreases the distance between temporally consistent ones. The TCC loss is expressed as follows:

$$L_{TCC} = -\log\left(\frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}\right) \tag{26}$$

where $\text{sim}(\cdot, \cdot)$ represents the cosine similarity and τ represents a temperature parameter. To guarantee efficient information flow between modalities, mutual information loss (MIL) is also used:

$$\mathcal{L}_{MIL} = -\sum I(\mathbf{z}_i; \mathbf{z}_j) \tag{27}$$

where the mutual information across modality representations is represented by $I(\cdot; \cdot)$.

For a certain input sequence, the final anomaly score is calculated as follows:

$$\text{Score} = \|\text{MGE}(\mathbf{Z}) - \boldsymbol{\mu}_{\text{normal}}\|_2 \tag{28}$$

$\boldsymbol{\mu}_{\text{normal}}$ is the mean representation of standard samples, and MGE stands for the multi-modal global encoder. This thorough fusion technique ensures reliable anomaly identification by utilizing complementary data from multiple modalities, thereby preserving temporal consistency and modal dependency. In addition, each MSBT block includes layer normalization and residual connections:

$$\text{MSBTblock}(\mathbf{x}) = \text{LayerNorm}\Big(\mathbf{x} + \text{MLP}\Big(\text{LayerNorm}\big(\mathbf{x} + \text{MultiHead}(\mathbf{x})\big)\Big)\Big) \tag{29}$$

MLP stands for a multi-layer perceptron that has nonlinear activations. The complete architecture is taught end-to-end using the losses mentioned earlier:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{TCC} + \lambda_2 \mathcal{L}_{MIL} + \lambda_3 \mathcal{L}_{\text{rec}} \tag{30}$$

$\mathcal{L}_{\text{rec}}$ denotes an optional reconstruction loss, if any, and λ_1, λ_2, and λ_3 are weighting coefficients for the corresponding loss terms.

4. Experiments

In this section, we evaluate the effectiveness of our multi-modal fusion approach for violence detection using the XD-Violence dataset. We discuss the experimental setup, evaluation metrics, and the performance of our model compared to baseline approaches.

4.1. Datasets

In the field of violence detection investigations, the multi-modal, state-of-the-art XD-Violence dataset [1,2] was developed. For a well-rounded classification, it offers a fascinating variety of real-world situations, including violent physical altercations, explosive events, and chaotic riots, mixed with peaceful, nonviolent scenes. The collection comprises more than 4754 dynamic video clips, totaling 217 h. Each film, with an average duration of 2 to 5 min, is labeled with segment-level violence. Models can learn from weakly supervised real-world data thanks to this configuration. Synchronized RGB frames, audio, and optical flow data within the dataset offer a wealth of information [31] to explore effective multi-modal fusion techniques. A ground-breaking tool for developing state-of-the-art violence detection systems, XD-Violence offers researchers an unparalleled opportunity to investigate the complexities of multi-modal integration, weakly supervised learning, and creative AI applications in real-world contexts. The dataset is illustrated in Figure 3.

In analyzing violent content, three essential elements are identified, each contributing to the accurate detection of violence. A dual-processing approach enhances detection accuracy by combining I3D for video analysis and VGGish for audio analysis. The three key features are as follows:

1. **Flow Properties (Motion and Action Dynamics):** By utilizing I3D's optical flow processing, the movement and dynamics of actions within a scene are captured. This enables the model to identify violent movements, such as aggressive gestures, rapid motion, or sudden physical actions, which are key indicators of violence.
2. **RGB Features (Scene Characteristics and Visual Content)**: I3D also analyzes RGB frames from the video to understand the visual context of the scene. This includes recognizing key objects, spatial layouts, and environmental color patterns. Visual content helps identify violent interactions, such as people in conflict, the use of weapons, or other contextual visual cues.
3. **Audio Features (Sound Patterns and Auditory Cues)**: VGGish processes the audio track, extracting features from sound patterns. The model tracks crucial auditory signals that indicate violence, such as explosions, screams, loud impacts, or rapid movement sounds. These audio features provide additional context to the scene, capturing elements that may not be visible in the video.

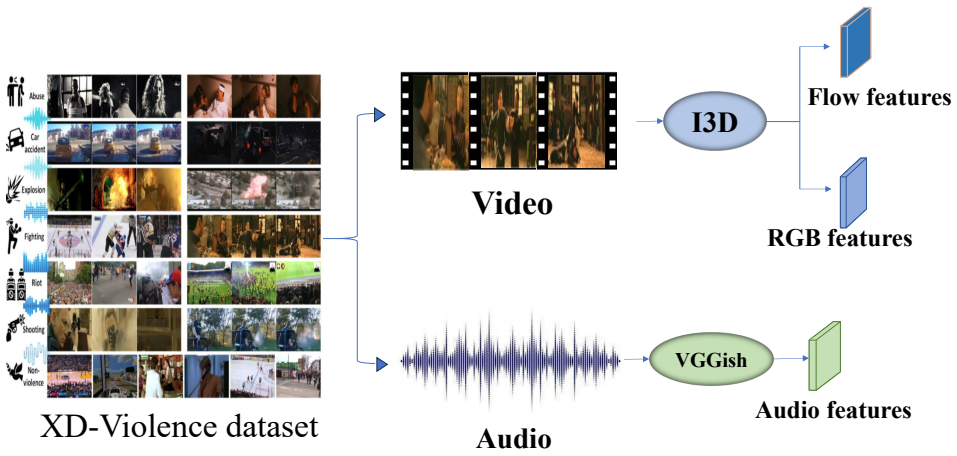

Figure 3. Overview of the XD-Violence dataset structure, showcasing video and audio modalities. Video features, including RGB and flow, are extracted using I3D, while audio features are extracted using VGGish.

4.2. Evaluation Standers

Our evaluation criterion for the XD-Violence dataset is frame-level Average Precision (AP), which is a reliable measure of the model's ability to accurately classify each frame independently of the overall video-level context. Frame-level AP evaluates how effectively the proposed model distinguishes between violent and nonviolent video scenes. A higher AP value reflects the model's capability to capture intricate features in the input data and make precise classifications. By focusing on frame-level performance, this approach provides a detailed evaluation highlighting the model's strengths and areas for improvement.

In addition to AP, we employ mean incident loss (MIL) and training classification cost loss (TCCL) to evaluate the model during training. MIL captures the average error between the predicted and actual labels at the instance level, indicating the model's ability to make accurate predictions for individual frames [32,33]. A decreasing MIL over training epochs demonstrates improved frame-level classification performance. Similarly, TCCL measures the cost associated with misclassifications during training, with lower values signifying better refinement of the model's predictions.

Together, these metrics ensure a comprehensive assessment of the proposed multi-modal violence detection framework, reflecting its effectiveness in identifying violent events with high precision and reliability.

4.3. Implementation Details

We utilized PyTorch (version 2.3.0) to process data in various formats, including RGB, flow, and audio. Each modality is routed individually through a backbone network to extract meaningful representations [34]. For example, ResNet [35] is used for RGB and flow, while a dedicated network is created for audio characteristics. A projection layer projects these features into a shared feature space. A transformer layer is then used to capture the temporal dependencies within each modality. The unimodal RGB, flow, and audio features are Z_R, Z_F, and Z_A, respectively, produced by this procedure.

Analysis of the Proposed Module's Effectiveness: The module for multi-modal gated fusion incorporates the functionalities individually. The combined features (Z_J) and the unimodal features (Z_R, Z_F, and Z_A) produced by the gated fusion module are subjected to the multi-scale bottleneck transformer (MSBT). The MSBT uses five hierarchical fusion layers, the first of which has sixteen bottleneck tokens ($N_{bt}^1 = 16$). Each of these 128-dimensional tokens starts with a standard normal distribution with a mean of zero and a standard deviation of fifteen dollars. For the model to adaptively zero in on the most essential information, the gated fusion module uses distinct RGB, flow, and audio gates to calculate attention weights for each modality. Using a fully linked layer with ReLU activation, the fused features Z_J are further enhanced.

The fused features are processed using a regressor and a transformer layer to calculate the weights, resulting in weighted features Z^*. The multi-modal global encoder collects these weighted features and aggregates global context using $L_G = 4$ transformer layers. Then, it uses an MIL-based technique to calculate violence scores.

The temporal consistency contrast (TCC) module is presented to improve temporal consistency by minimizing inter-class disparities across frames. TCC loss (with temperature $\tau = 0.5$), MIL loss (with Top-K = 9), and a balancing scalar ($\lambda = 0.1$) are all components of the combined loss function employed in the framework.

We train the proposed multi-modal violence detection framework using the stochastic gradient descent (SGD) optimizer for 50 epochs, with a learning rate of 0.0005 and a batch size of 32. These hyperparameters ensure stable and practical training, while preventing both overfitting and underfitting. Training progress is evaluated using key metrics, including loss, mean instance loss (MIL), training classification cost (TCC), and Average Precision **(AP)**, as shown in Figure 4. As a result of metrics such as these, the model converges and performs better throughout the training process.

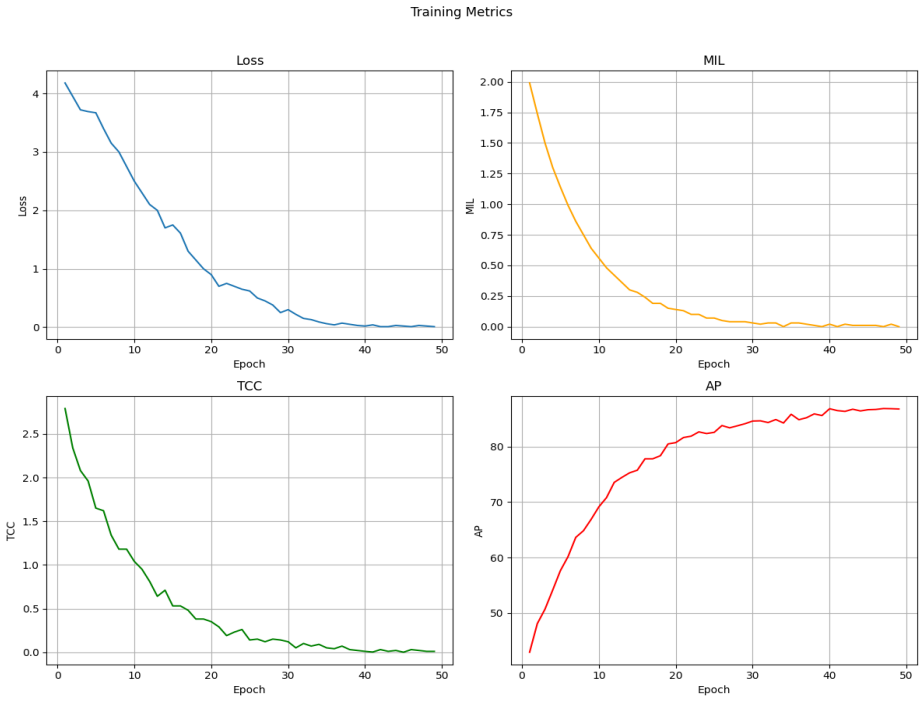

Figure 4. Training metrics over 50 epochs, including loss, mean instance loss (MIL), training classification cost (TCC), and Average Precision (AP).

4.4. Ablation Studies

Our ablation study studied the three primary sections of our framework, the [17] multi-scale bottleneck transformer (MSBT) for multi-modal fusion, the temporal consistency contrast (TCC) loss, and the [36] gated fusion module (GFM), to show how successful the components are. Using a method of selective removal and replacement, along with performance analysis, this study aimed to assess the contribution of each module. The results of our ablation study are summarized in Table 2, which highlights the effectiveness of different multi-modal fusion methods.

Table 2. Performance comparison of model iterations with and without the multi-scale bottleneck transformer (MSBT) and gated fusion. Results are presented for different modality combinations: RGB (R), audio (A), and flow (F), including their pairs and all three combined (R+A+F).

MSBT	Gated Fusion	TCC	R+A	R+F	A+F	R+A+F
✗	✗	✗	71.58	73.19	70.38	75.93
✓	✗	✓	82.67	80.46	78.55	81.74
✗	✓	✓	79.78	80.83	79.11	82.38
✓	✓	✗	78.16	76.97	75.71	80.05
✗	✓	✗	80.12	79.45	76.89	81.27
✓	✗	✗	77.95	78.23	75.34	79.68
✗	✗	✓	78.34	80.17	78.92	81.02
✓	✓	✓	**83.24**	**81.51**	**80.63**	**86.85**

Cross-Dataset Validation: To examine robustness beyond XD-Violence, we evaluated our framework on the UCF-Crime dataset. Despite differences in scene composition, recording conditions, and cultural contexts, our model achieved 83.41% AP. These results confirm the generalizability of our approach and highlight its potential for real-world surveillance applications.

Complexity and Scalability Analysis

To assess scalability, we analyzed computational complexity and inference speed. Our framework contains 34.2 M parameters and requires 78 GFLOPs per 16-frame clip, compared to 92 GFLOPs or a baseline multi-stream transformer. Thanks to the multi-scale bottleneck transformer (MSBT), redundancy is reduced by compressing long-range dependencies into bottleneck tokens. In practice, the model processes video at approximately 45 frames per second (FPS) on an NVIDIA A100 GPU, which is suitable for real-time surveillance.

Although the model introduces moderate overhead compared to simple concatenation-based fusion, it provides significantly higher accuracy with a manageable computational cost. For deployment in resource-constrained environments, further optimizations such as pruning, quantization, or knowledge distillation could be applied to improve scalability without compromising accuracy.

Without MSBT Fusion Method:

This experiment employed a fusion mechanism based on cross-attention, as described in [37], rather than the MSBT. The results demonstrate that performance dropped significantly without the MSBT, demonstrating its crucial role in accurately capturing the intricate interplay of multi-modal inputs. The MSBT's superior capacity to process multi-modal features was demonstrated by its consistent outperformance of cross-attention-based fusion.

Without TCC Loss:

We trained the model without temporal consistency to assess the effect of TCC loss. Performance decreased without it, especially in tasks involving temporal sequences. The findings confirm that the TCC loss strengthens the model as a whole by enhancing the temporal alignment of the features.

Without Gated Fusion Method (GFM):

We avoided the gating process and directly input the MSBT multi-modal data to evaluate the GFM's significance. This resulted in less-than-ideal modal fusion since the GFM dynamically modifies the contribution of each modality according to its importance. We can see that using the GFM significantly enhances the model's ability to leverage various domains. To better understand the role of each modality, we performed ablations with and without the optical flow stream. Results indicate that flow alone achieves a lower AP (72.9%) compared to RGB (80.6%) and audio (77.2%). However, when combined with RGB and audio, flow contributes complementary motion information that improves recognition in highly dynamic scenes such as riots and group fights.

We also isolated the contributions of offline and online detection. Removing the offline branch reduced AP from 86.85% to 84.72%, while removing the online branch reduced AP to 85.13%. This confirms that both branches offer complementary benefits: offline captures long-range dependencies, while online supports causal, frame-level decisions.

Comparative Results on Fusion Strategies: Furthermore, we tested various multi-modal fusion techniques side by side. Table 1 shows that our gated fusion module outperforms the unoptimized application of gated fusion by 2.53 percentage points, with a performance of 86.85%. We also used our gated fusion module to re-implement [3]'s initial concatenation-based fusion method. Furthermore, this change enhanced performance by 2.53%, providing evidence that our gating mechanism is beneficial.

4.5. Qualitative Evaluation

Our MSBT, using the gated fusion approach, produced the violence score curves in Figure 5 for several test frames extracted from the XD-Violence dataset. Explosion, assault, automobile accident, battle, gunshot, and riot are the six categories of violence we evaluate. An increase in violent scores points to more aggressive behavior, and our approach successfully distinguishes between violent and nonviolent segments in five categories: explosion, vehicle accident, brawl, riot, and shooting. The audio, flow, and

RGB modalities do not appear to indicate violence, so the AP performance is worse for the abuse class. However, despite its robustness, the method has challenges in detecting subtle patterns of violence. Performance in the "abuse" category remains relatively weaker, consistent with prior work. This limitation arises from nuanced cues (e.g., posture and tone of voice) that are harder to capture using RGB, flow, or audio alone. Future research may integrate additional modalities such as pose estimation or depth sensing to address this challenge.

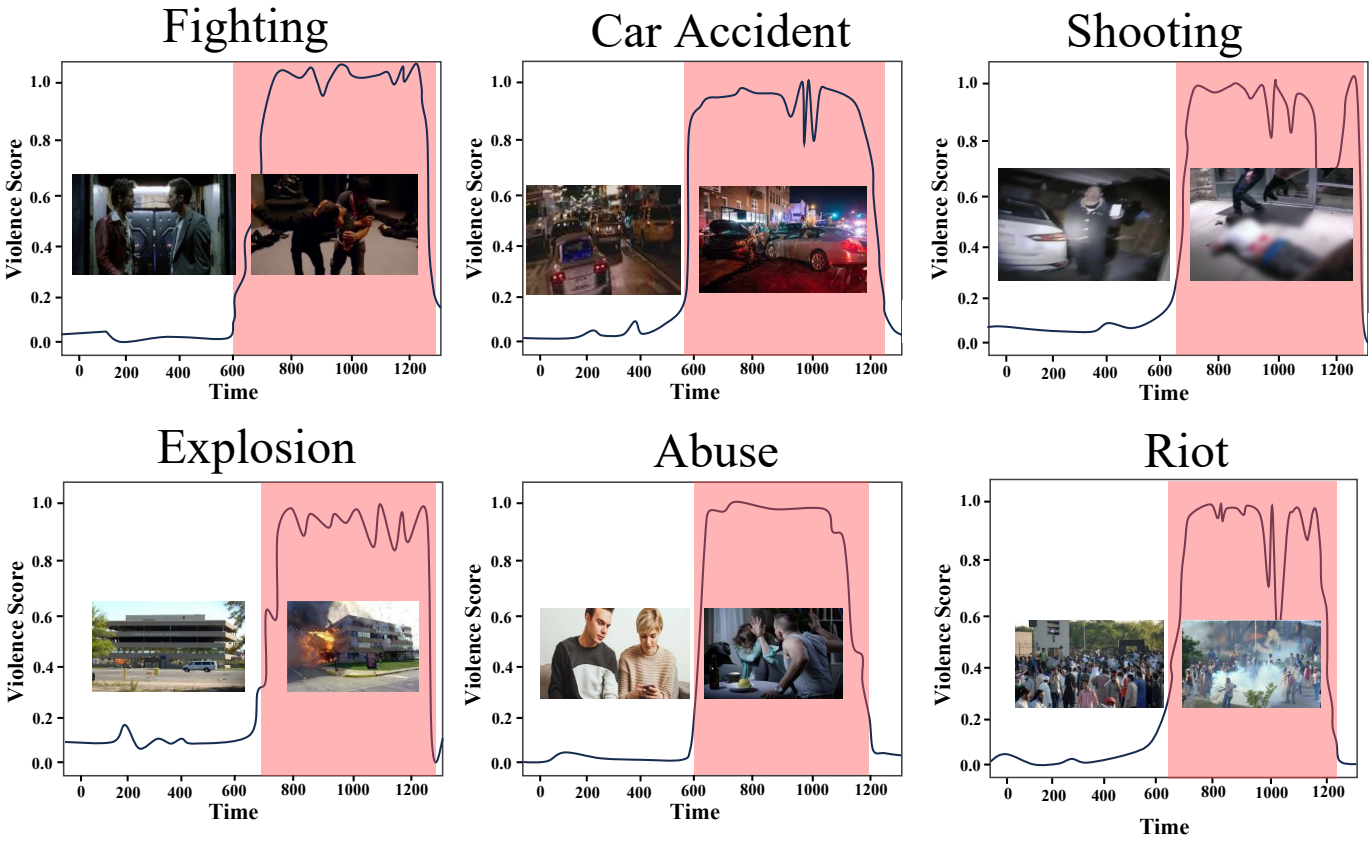

Figure 5. Visualization of anomaly scores predicted on the XD-Violence test set. The red regions indicate ground-truth violent events, and the blue lines are anomaly scores predicted by our method. It is best viewed in color.

4.6. Comparison with State of the Art

We compare our proposed model with state-of-the-art methods, as Table 3 summarizes. Most existing [11,38] multi-modal approaches primarily focus on audio–visual learning, limiting their analysis to only two modalities, and their frameworks are not readily adaptable to incorporate additional modalities. Our method outperforms most existing approaches when considering two modalities, except [14]. This is attributed to the addition of a single-modality teacher network for knowledge distillation, which transfers knowledge from a uni-modal network, thereby enhancing performance when utilizing both RGB and audio modalities. However, its performance is less competitive when used in conjunction with other modalities. Our method achieves superior results, surpassing all existing methods when all three modalities are integrated.

Table 3. Evaluation of frame-level Average Precision (AP) performance on the XD-Violence dataset. Bold values indicate the best-performing results. The proposed method demonstrates superior performance compared to existing approaches.

Manner	Method	Modality	AP (%)	Param. (M)
Unsupervised Learning	SVM baseline [39]	-	50.78	-
	OCSVM [40]	-	27.25	-
	Hasan al-Banna [41]	-	30.77	-
Supervised (Video)	Sultani et al. [38]	Video	75.68	-
	Wu et al. [1]	Video	75.90	-
	RTFM [10]	Video	77.81	12.067
	MSL et al. [42]	Video	78.28	-
	S3R [43]	Video	80.26	-
	UR-DMU [14]	Video	81.66	-
	Zhang et al. [44]	Video	78.74	-
Weakly Supervised (Audio + Video)	Wu et al. [1]	Audio + Video	78.64	0.843
	Wu et al. [1]	Audio + Video	78.66	1.539
	RTFM [10]	Audio + Video	78.54	13.510
	RTFM [10]	Audio + Video	78.54	13.190
	Pang et al. [12]	Audio + Video	81.69	1.876
	MACIL-SD [17]	Audio + Video	83.40	0.678
	UR-DMU [45]	Audio + Video	81.77	-
	Zhang et al. [44]	Audio + Video	81.43	-
	HyperVD [13]	Audio + Video	**85.67**	0.607
Weakly Supervised (Proposed)	**MSBT Gated (Proposed)**	Video + Audio	**83.39**	0.743
	MSBT Gated (Proposed)	Audio + Video + Flow	**86.85**	0.913

5. Conclusions

This study introduces a multi-modal framework for detecting violence by integrating visual, motion, and audio features through gated fusion and a multi-scale transformer (MSBT). We address challenges such as modality imbalance, temporal inconsistencies, and weak supervision by utilizing temporal consistency contrast (TCC) loss and multiple instance learning (MIL). Evaluations on the XD-Violence dataset demonstrate cutting-edge performance, particularly with RGB, optical flow, and audio data. The framework's adaptive fusion and multi-scale feature extraction techniques enhance the overall feature extraction process, making it suitable for real-time applications such as surveillance.

Future research could incorporate additional modalities, such as depth and infrared data, to further enhance scene understanding and improve robustness in diverse environmental conditions. By integrating these modalities, we could achieve better object detection, scene reconstruction, and material differentiation, especially in low-light or obscured scenarios. Additionally, while our framework shows promise for security and surveillance, it is essential to acknowledge the ethical challenges it presents. First, datasets such as XD-Violence may contain cultural biases in defining what constitutes "violence," which can affect generalizability across different regions and contexts. Second, real-world deployment of automated violence detection in surveillance systems raises privacy concerns, as continuous monitoring in public or private spaces can infringe on individual rights. Third, there is a risk of misuse, such as profiling or over-policing vulnerable groups. Beyond technical performance, automated violence detection also raises ethical concerns. Privacy risks in surveillance, cultural biases in defining violence, and potential misuse, such as over-surveillance or profiling, must be carefully considered. Future work should in-

tegrate fairness-aware methods and ensure deployment complies with privacy regulations, emphasizing transparency and accountability.

Author Contributions: Conceptualization, methodology, software, validation, B.A. and M.K. formal analysis, investigation, resources, M.S.; data curation, writing—original draft preparation, writing—review and editing, B.A. and M.K.; visualization, B.A.; supervision, M.K. and M.S.; project administration, M.K.; funding acquisition, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work is part of the project "SafeStream: Next-Gen Multimodal AI for Enhanced Detection, Recognition, and Scene Analysis in UAV Applications (Project Code: 12T118)" developed at the United Arab Emirates University (UAEU), Abu Dhabi, UAE.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The used data in this paper for result generation is publicly available at https://www.kaggle.com/datasets/nguhaduong/xd-violence-video-dataset, accessed on 27 August 2025.

Acknowledgments: This work is part of the project "SafeStream: Next-Gen Multimodal AI for Enhanced Detection, Recognition, and Scene Analysis in UAV Applications", developed at the United Arab Emirates University (UAEU), Abu Dhabi, UAE. Furthermore, we acknowledge the invaluable contribution of artificial intelligence (AI) tools in enhancing the efficiency and advancements in our research endeavors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only look but also listen: Learning multimodal violence detection under weak supervision. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 322–339. [CrossRef]
2. Wu, P.; Liu, X.; Liu, J. Weakly supervised audio-visual violence detection. *IEEE Trans. Multimed.* **2022**, *25*, 1674–1685. [CrossRef]
3. Sun, S.; Gong, X. Multi-scale Bottleneck Transformer for Weakly Supervised Multimodal Violence Detection. *arXiv* **2024**, arXiv:2405.05130. [CrossRef]
4. Wu, Y.; Yang, H.; Tang, J.; Yu, Y. Multi-objective re-synchronizing of bus timetable: Model, complexity and solution. *Transp. Res. Part C Emerg. Technol.* **2016**, *67*, 149–168. [CrossRef]
5. Cao, B.; Sun, Y.; Zhu, P.; Hu, Q. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 23555–23564. [CrossRef]
6. Fan, Y.; Yu, Y.; Lu, W.; Han, Y. A Cross-modal and Redundancy-reduced Network for Weakly-Supervised Audio-Visual Violence Detection. In Proceedings of the 5th ACM International Conference on Multimedia in Asia, Tainan, Taiwan, 6–8 December 2023; pp. 1–7. [CrossRef]
7. Li, J.; Wang, X.; Lv, G.; Zeng, Z. GraphMFT: A graph network-based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing* **2023**, *550*, 126427. [CrossRef]
8. Pang, W.F.; He, Q.H.; Hu, Y.J.; Li, Y.X. Violence detection in videos based on fusing visual and audio information. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–12 June 2021; pp. 2260–2264. [CrossRef]
9. Choqueluque-Roman, D.; Camara-Chavez, G. Weakly supervised violence detection in surveillance video. *Sensors* **2022**, *22*, 4502. [CrossRef]
10. Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J.W.; Carneiro, G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4975–4986. [CrossRef]
11. Li, S.; Liu, F.; Jiao, L. Self-training multi-sequence learning with transformer for weakly Supervised video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22 February–1 March 2022; Volume 36, pp. 1395–1403. [CrossRef]

12. Pang, G.; Shen, C.; Jin, H.; van den Hengel, A. Deep weakly-supervised anomaly detection. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 1795–1807.
13. Zhou, X.; Peng, X.; Wen, H.; Luo, Y.; Yu, K.; Yang, P.; Wu, Z. Learning weakly supervised audio-visual violence detection in hyperbolic space. *Image Vis. Comput.* **2024**, *151*, 105286. [CrossRef]
14. Zhou, H.; Yu, J.; Yang, W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 3769–3777. [CrossRef]
15. Cai, L.; Mao, X.; Zhou, Y.; Long, Z.; Wu, C.; Lan, M. A Survey on Temporal Knowledge Graph: Representation Learning and Applications. *arXiv* **2024**, arXiv:2403.04782. [CrossRef]
16. Karim, H.; Doshi, K.; Yilmaz, Y. Real-time weakly supervised video anomaly detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 6848–6856. [CrossRef]
17. Yu, J.; Liu, J.; Cheng, Y.; Feng, R.; Zhang, Y. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 6278–6287. [CrossRef]
18. Abduljaleel, I.Q.; Ali, I.H. Deep Learning and Fusion Mechanism-based Multimodal Fake News Detection Methodologies: A Review. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 15665–15675. [CrossRef]
19. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30* [CrossRef]
20. Lu, L.; Dong, M.; Wang, S.B.; Zhang, L.; Ng, C.H.; Ungvari, G.S.; Li, J.; Xiang, Y.T. Prevalence of workplace violence against health-care professionals in China: A comprehensive meta-analysis of observational surveys. *Trauma Violence Abus.* **2020**, *21*, 498–509. [CrossRef]
21. Wei, X.; Wu, D.; Zhou, L.; Guizani, M. Cross-modal communication technology: A survey. *Fundam. Res.* **2025**, *5*, 2256–2267. [CrossRef]
22. Jiang, W.D.; Chang, C.Y.; Kuai, S.C.; Roy, D.S. From explicit rules to implicit reasoning in an interpretable violence monitoring system. *arXiv* **2024**, arXiv:2410.21991. [CrossRef]
23. Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; Wu, F. Multi-modality cross attention network for image and sentence matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10941–10950. [CrossRef]
24. Alsarhan, T.; Ali, U.; Lu, H. Enhanced discriminative graph convolutional network with adaptive temporal modelling for skeleton-based action recognition. *Comput. Vis. Image Underst.* **2022**, *216*, 103348. [CrossRef]
25. Li, H.; Wang, J.; Li, Z.; Cecil, K.M.; Altaye, M.; Dillman, J.R.; Parikh, N.A.; He, L. Supervised contrastive learning enhances graph convolutional networks for predicting neurodevelopmental deficits in very preterm infants using brain structural connectome. *NeuroImage* **2024**, *291*, 120579. [CrossRef] [PubMed]
26. Ghadiya, A.; Kar, P.; Chudasama, V.; Wasnik, P. Cross-Modal Fusion and Attention Mechanism for Weakly Supervised Video Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 1965–1974. [CrossRef]
27. Wei, D.L.; Liu, C.G.; Liu, Y.; Liu, J.; Zhu, X.G.; Zeng, X.H. Look, listen and pay more attention: Fusing multi-modal information for video violence detection. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 1980–1984. [CrossRef]
28. Duja, K.U.; Khan, I.A.; Alsuhaibani, M. Video Surveillance Anomaly Detection: A Review on Deep Learning Benchmarks. *IEEE Access* **2024**, *12*, 164811–164842 [CrossRef]
29. Wan, Z.; Mao, Y.; Zhang, J.; Dai, Y. Rpeflow: Multimodal fusion of RGB-pointcloud-event for joint optical flow and scene flow estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 10030–10040. [CrossRef]
30. Yang, L.; Wu, Z.; Hong, J.; Long, J. MCL: A contrastive learning method for multimodal data fusion in violence detection. *IEEE Signal Process. Lett.* **2022**, *30*, 408–412. [CrossRef]
31. Khan, M.; Saad, M.; Khan, A.; Gueaieb, W.; El Saddik, A.; De Masi, G.; Karray, F. Action knowledge graph for violence detection using audiovisual features. In Proceedings of the 2024 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 6–8 January 2024; pp. 1–5. [CrossRef]
32. Guo, Y.; Guo, H.; Yu, S.X. Co-sne: Dimensionality reduction and visualization for hyperbolic data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 21–30. [CrossRef]
33. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
34. Gemmeke, J.F.; Ellis, D.P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780. [CrossRef]

35. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135. [CrossRef]
36. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention bottlenecks for multimodal fusion. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 14200–14213.
37. Yan, S.; Xiong, X.; Arnab, A.; Lu, Z.; Zhang, M.; Sun, C.; Schmid, C. Multiview transformers for video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3333–3343. [CrossRef]
38. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6479–6488. [CrossRef]
39. Clavié, B.; Alphonsus, M. The unreasonable effectiveness of the baseline: Discussing SVMs in legal text classification. In *Legal Knowledge and Information Systems*; IOS Press: Amsterdam, The Netherlands, 2021; pp. 58–61. [CrossRef]
40. Le, X.; Yin, J.; Zhou, X.; Chu, B.; Hu, B.; Wang, L.; Li, G. SCADA Anomaly Detection Scheme Based on OCSVM-PSO. In Proceedings of the 2024 6th International Conference on Energy Systems and Electrical Power (ICESEP), Wuhan, China, 21–23 June 2024; pp. 1335–1340. [CrossRef]
41. Krämer, G. *Hasan al-Banna*; Simon and Schuster: New York, NY, USA, 2014. [CrossRef]
42. Cong, R.; Qin, Q.; Zhang, C.; Jiang, Q.; Wang, S.; Zhao, Y.; Kwong, S. A weakly supervised learning framework for salient object detection via hybrid labels. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 534–548. [CrossRef]
43. Wu, J.C.; Hsieh, H.Y.; Chen, D.J.; Fuh, C.S.; Liu, T.L. Self-supervised sparse representation for video anomaly detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 729–745.
44. Jiang, M.; Hou, C.; Zheng, A.; Hu, X.; Han, S.; Huang, H.; He, X.; Yu, P.S.; Zhao, Y. Weakly supervised anomaly detection: A survey. *arXiv* **2023**, arXiv:2302.04549. [CrossRef]
45. Pu, Y.; Wu, X.; Yang, L.; Wang, S. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Trans. Image Process.* **2024**. [CrossRef] [PubMed]