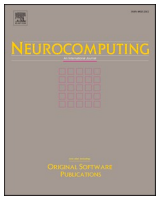
Neurocomputing 684 (2026) 133532

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A multimodal framework for violent behavior recognition in surveillance videos

Chih-Yung Chang [a,*], Syu-Jhih Jhang [a], Yu-Ting Chin [b], I.-Hsiung Chang [c], Diptendu Sinha Roy [d]

[a] the Department of Computer Science and Information Engineering, Tamkang University, New Taipei 25137, Taiwan
[b] the Department of Artificial Intelligence, Tamkang University, New Taipei 25137, Taiwan
[c] Department of Child Care and Industries, Fooyin University, Kaohsiung 831301, Taiwan
[d] the Department of Computer Science and Engineering, National Institute of Technology, Shillong 793003, India

ARTICLE INFO

Keywords:
Multimodal learning
Violence detection
Surveillance video analysis
Semantic captioning
Temporal fusion
Transformer
Siamese network
Robust behavior recognition

ABSTRACT

This study presents a multimodal violent behavior recognition framework that integrates visual, auditory, and semantic modalities to enable accurate detection and temporal localization of aggressive events in surveillance videos. The proposed system introduces a hierarchical fusion strategy in which modality-specific encoders extract complementary representations from video frames, audio signals, and BLIP-based semantic captions. A Transformer-based temporal encoder aggregates these embeddings into coherent temporal patterns, while a dual-head prediction module simultaneously performs binary violence detection and multi-class behavior classification. To explicitly capture the transitional dynamics preceding violent actions, an additional Siamese onset branch is introduced. This branch optimizes a margin-based cosine contrastive objective, enforcing similarity between embeddings around the onset boundary and dissimilarity for temporally distant segments. The overall optimization combines detection, classification, and Siamese losses under a unified objective function, achieving high temporal precision and stable onset alignment. Extensive experiments conducted on both the XD-Violence and UCF-Crime datasets demonstrate that the proposed framework achieves competitive performance across multiple evaluation metrics, including frame-level F1-score, mean Average Precision (mAP), and Onset Detection Latency (ODL), while maintaining interpretability through attention-based temporal visualization.

1. Introduction

Violent and abnormal behaviors, such as arson, riots, robberies, physical assaults, and traffic accidents, pose significant threats to public safety. Traditional patrolling and surveillance strategies are often insufficient, as timely reporting by bystanders is rare and perpetrators frequently flee before authorities arrive. With the rapid proliferation of surveillance cameras and the advancement of artificial intelligence (AI), automated recognition of abnormal events has become a pressing requirement in smart city safety management.

Recent studies on anomaly detection have achieved promising results using vision-based approaches. For example, an attention-based residual autoencoder has been proposed to enhance the detection of critical abnormal frames, demonstrating robust performance across multiple benchmarks [1]. However, purely image-based models often rely on sliding-window frameworks to repeatedly analyze frame segments, resulting in detection delays and high computational costs [2–5]. To improve efficiency, one-class classification methods have been introduced, which characterize the statistical distribution of normal behaviors and regard outliers as anomalies [6–8]. While computationally lightweight, such methods struggle to generalize across diverse abnormal scenarios in real-world surveillance.

The introduction of multimodal recognition has further enriched anomaly detection by incorporating audio and textual modalities. Studies have shown that combining audio-visual features with speech-to-text transcripts can enhance detection accuracy [9–11]. Nevertheless, most of these approaches simply concatenate modality-specific vectors, disregarding the spatiotemporal dependencies among modalities. As a result, critical cross-modal information is lost, and model robustness diminishes in complex environments. Additionally, street surveillance recordings often suffer from poor audio quality and heavy background noise, reducing the effectiveness of speech-to-text

* Corresponding author.
E-mail address: cychang@mail.tku.edu.tw (C.-Y. Chang).

https://doi.org/10.1016/j.neucom.2026.133532
Received 11 November 2025; Received in revised form 16 March 2026; Accepted 30 March 2026
Available online 2 April 2026

conversion and weakening the contribution of the textual modality [12, 13].

In summary, AI-driven abnormal behavior recognition systems face three major challenges:

1. **Delayed anomaly onset detection.** Sliding window-based frameworks introduce repeated computation, preventing timely responses [1–5].
2. **Low-quality audio recordings.** Environmental noise in street surveillance hinders reliable speech-to-text conversion, weakening multimodal performance [12,13].
3. **Inefficient multimodal fusion.** Simple feature concatenation neglects cross-modal spatiotemporal dependencies, leading to information loss [9–11].

Such limitations not only hinder technical scalability but also reduce trustworthiness in public safety applications, where delayed or inaccurate recognition may result in missed incidents or ineffective interventions. Moreover, while existing methods demonstrate competitive results on controlled benchmarks, their reliance on sliding windows, handcrafted fusion, or noise-sensitive audio pipelines undermines their scalability to real-world, large-scale surveillance systems. These gaps highlight the urgent need for a more robust framework that can simultaneously achieve low-latency detection, noise-resilient feature integration, and fine-grained multimodal fusion. To overcome these shortcomings, this study proposes a multimodal violent behavior recognition framework that integrates visual, audio, and semantic modalities. The main contributions of this work are summarized as follows:

1. **Multimodal Fusion Architecture.** A unified framework is designed to integrate visual, auditory, and semantic modalities through synchronized encoders and a Transformer-based temporal fusion module, enabling context-aware recognition of violent behaviors in complex surveillance scenarios.
2. **BLIP-Based Semantic Embedding.** A semantic captioning module based on Bootstrapped Language-Image Pretraining (BLIP) generates global textual representations that complement low-level multimodal features and enhance semantic grounding during temporal modeling.
3. **Temporal Precision through Contrastive Alignment.** A cosine-based contrastive objective is employed to maintain temporal coherence between pre- and post-onset embeddings, enhancing the precision of violence onset detection.
4. **Comprehensive Evaluation and Interpretability.** Extensive experiments on multiple public datasets demonstrate competitive detection and classification accuracy, robustness to noise and modality imbalance, and clear temporal activation patterns that provide insight into model decision behavior.

The remainder of this paper is organized as follows. Section II reviews related work on anomaly detection and multimodal behavior recognition. Section III introduces the background knowledge underlying the proposed architecture. Section IV details the system design, including the onset detection and fine-grained classification modules. Section V presents the experimental setup and performance analysis on benchmark datasets. Section VI concludes the study and discusses potential future work.

2. Related work

This chapter introduces two major aspects of the related work: anomaly onset detection in videos and behavior recognition methods. The first part reviews both machine learning-based and deep learning-based approaches for detecting the starting point of abnormal events. The second part discusses unimodal and multimodal frameworks for behavior recognition, emphasizing their advantages and limitations in handling complex violent scenarios. A comparative summary is then provided to highlight the research gaps that motivate this study.

2.1. Machine learning and deep learning approaches for anomaly onset detection

Anomaly onset detection is a fundamental step for timely recognition of abnormal behaviors in surveillance systems. Early machine learning-based approaches often relied on handcrafted features and shallow models. For instance, Hamdi et al. [7] integrated Histogram of Optical Flow (HOF) descriptors with pretrained convolutional features to improve abnormality classification, while Hu et al. [8] introduced ensemble random projection combined with unsupervised learning to enhance detection efficiency. More recently, He et al. [12] proposed the FCC-MF framework, which leverages frame-wise cluster contrast and modality-stage flooding to improve robustness in weakly supervised violence detection. Although effective, these methods remain limited by their reliance on feature engineering or weak supervision strategies, often resulting in unstable performance in complex street environments.

Deep learning-based approaches have further advanced anomaly onset detection by directly modeling spatiotemporal patterns. Le and Kim [1] designed an attention-based residual autoencoder to capture abnormal frames with higher precision. Amin et al. [2] combined 2D CNNs with LSTMs to jointly exploit spatial and temporal information in surveillance videos. Khan et al. [6] developed a CNN-based framework for traffic accident detection, demonstrating the applicability of deep learning to safety-critical monitoring tasks. However, most of these methods depend on sliding window analysis [2–5], which introduces redundant computation and detection latency, making them unsuitable for real-time scenarios.

2.2. Unimodal and multimodal frameworks for behavior recognition

Beyond onset detection, behavior recognition seeks to classify abnormal events into specific categories. Single-modality methods have been extensively studied, particularly using vision-based or skeleton-based cues. Shu et al. [3] improved dense trajectories combined with sliding windows for action detection, while Sun et al. [4] proposed a dynamic sliding window approach for activity recognition from wearable sensors. Le et al. [5] introduced DDNet, a lightweight skeleton-based model that can capture both slow and fast motion dynamics. Despite their advances, single-modality approaches often fail to generalize across diverse real-world conditions due to their reliance on a single source of information.

To overcome these limitations, multimodal methods have gained increasing attention. Shaikh et al. [9] introduced MAiVAR, a framework that integrates visual and audio features for enhanced video action recognition. Ortega et al. [10] explored deep multimodal fusion of audio, video, and textual features for emotion recognition, while Yang et al. [11] developed ICANet by combining audio, RGB, and optical flow modalities. These studies highlight the potential of multimodal integration; however, their fusion strategies are often limited to feature concatenation, which neglects fine-grained temporal dependencies and reduces the capacity to exploit complementary cross-modal information.

Recent advances in video understanding have introduced video-specific temporal Transformers that focused on modeling long-range spatiotemporal dependencies within visual streams. Representative works such as Video Swin Transformer [14] adopted hierarchical window-based self-attention to capture spatial-temporal continuity and motion patterns directly from video frames, demonstrating strong performance on large-scale video recognition benchmarks. These architectures were particularly effective for visual-only tasks that relied on dense spatiotemporal feature hierarchies.

In contrast, the Transformer module adopted in this work was not intended to serve as a video-centric temporal backbone. Instead, it was designed as a multimodal temporal fusion mechanism that operated on

high-level representations extracted independently from visual, audio, and semantic modalities. The attention mechanism was employed to dynamically integrate heterogeneous modality-specific cues over time, rather than to model fine-grained spatiotemporal visual patterns. Consequently, the proposed framework targeted multimodal surveillance scenarios where semantic and acoustic information complemented visual cues, and was positioned differently from video-only temporal Transformers.

2.3. Vision–language models for video understanding

Recent advances in vision–language models (VLMs) have significantly extended video understanding by jointly modeling visual content and natural language semantics. Early large-scale image–text pretraining frameworks, such as CLIP [15] and ALIGN [16], established transferable cross-modal representations that aligned visual concepts with linguistic descriptions, providing a strong foundation for semantically informed video analysis.

Building upon these foundations, recent studies have explored the integration of language into video-level understanding tasks, including caption generation, temporal grounding, and event localization. By aggregating frame-wise visual features with language representations, these approaches enable richer semantic interpretation of complex activities and interactions that are difficult to distinguish using visual cues alone. In parallel, video-specific temporal Transformers, such as Video Swin Transformer [14], have been proposed to model long-range spatiotemporal dependencies within visual streams through hierarchical attention mechanisms, achieving strong performance in visual-only video recognition tasks.

Beyond visual-centric modeling, several vision–language frameworks have demonstrated that linguistic context can serve as a high-level semantic abstraction to guide temporal reasoning and event localization in untrimmed videos [17,18]. These methods highlight the complementary role of language in disambiguating visually similar actions and identifying salient temporal segments, particularly in complex or cluttered scenes.

Recent studies have further explored semantic-guided and structure-aware fusion strategies to improve multimodal representation learning. For instance, semantic guidance through vision–language models has been shown to enhance controllability and semantic consistency in multimodal fusion tasks [19]. Structure-driven fusion models emphasize the preservation of intrinsic structural information during multimodal integration, enabling more balanced feature interactions under noisy or heterogeneous conditions [20,21]. Input space optimization has also been explored as an alternative perspective for rethinking the fusion process, demonstrating that reformulating the problem in the input domain can yield improved focus estimation and boundary accuracy [22]. In addition, transformer-based mechanisms with dynamic feature interaction have been investigated to improve cross-modal representation learning and adaptive modality weighting [23]. Weakly supervised controllable fusion frameworks have also been proposed to regulate modality contributions and highlight task-relevant information at the instance level [24]. These studies indicate that incorporating semantic guidance, structural constraints, and input space reformulation can significantly improve the robustness and interpretability of multimodal fusion systems.

However, most of these approaches primarily focus on image-level fusion or low-level feature reconstruction, rather than high-level temporal event understanding. In contrast, the proposed framework focuses on multimodal violent behavior recognition in surveillance videos, emphasizing temporal reasoning and cross-modal semantic alignment through Transformer-based temporal modeling.

In addition to video recognition and vision–language modeling, transformer-based architectures have also been widely applied to large-scale scene perception and reconstruction tasks. For example, edge-assisted epipolar transformers have been proposed for industrial scene reconstruction [25], while parallax-attention mechanisms have been adopted for robust depth estimation in aerial environments [26]. Neural rendering and flow-assisted multi-view stereo approaches further enable real-time monocular tracking and scene perception [27], and self-supervised multi-view stereo frameworks have been explored for large-scale aerial scene understanding [28]. Semi-supervised domain adaptation strategies have also been developed for object detection under adverse weather conditions on embedded platforms [29].

These studies primarily focus on geometric modeling, multi-view consistency, or spatial reconstruction. In contrast, the present work addresses multimodal violent behavior recognition in surveillance videos, emphasizing temporal event modeling and cross-modal semantic fusion rather than spatial depth estimation or 3D reconstruction.

Motivated by these advances, the present work incorporates semantic captions generated by a vision–language model as an auxiliary modality for multimodal behavior recognition. Unlike approaches that employ language as explicit supervision or query input, the semantic representations in this framework are treated as guidance signals that provide high-level contextual cues. By complementing visual and audio features during temporal fusion, this design enables semantically informed violent behavior recognition while avoiding additional annotation requirements or restrictive supervision.

2.4. Comparative summary of detection and recognition methods

Table I provides a comparative overview of representative methods in anomaly onset detection and behavior recognition. As shown in the table, although prior studies have advanced the field through both unimodal and multimodal designs, their effectiveness is still constrained by detection latency, vulnerability to noisy audio, and insufficient modeling of cross-modal dependencies. These persistent limitations underscore the necessity of developing a more robust multimodal framework capable of achieving low-latency detection, noise-resilient processing, and advanced cross-modal fusion for violent behavior recognition.

3. Assumptions and problem formulation

This study focuses on the automatic identification of violent behavior in video surveillance data through a multimodal learning framework. The system processes input video clips, each containing synchronized visual and audio streams, and aims to accomplish two key objectives: first, detecting the presence of bullying within a given video segment, and second, classifying the type of bullying behavior if detected. A formal definition of the problem includes the specification of the input structure, target outputs, evaluation metrics, and optimization objectives, incorporating a tunable weighting scheme with provable convergence guarantees.

Let the input dataset be denoted as $V = \{v_1, \quad v_2, \quad \cdots, v_n\}$, where each video clip v_i is of fixed duration and contains sufficient audiovisual data for analysis. Each input v_i is processed into a structured feature representation, which is expressed as $x_i = (f_i^v, f_i^a, f_i^s)$, where $f_i^v \in \mathbb{R}^{T\times H\times W\times 3}$ represents the sequence of RGB video frames, $f_i^a \in \mathbb{R}^{T\times F}$ denotes the audio features extracted via Mel-spectrogram or MFCC analysis, and $f_i^s \in \mathbb{R}^d$ corresponds to the semantic embedding generated through BLIP (Bootstrapped Language-Image Pretraining), which provides cross-modal alignment between visual frames and textual scene descriptions. The key mathematical symbols and technical abbreviations used throughout the remainder of this section are summarized in Table 1 for reference.

Each input clip x_i is associated with a ground truth label $y_i = (b_i, c_i)$, where the variable $b_i \in \{0, 1\}$ indicates the presence or absence of bullying, and $c_i \in C = \{c_1, \quad c_2, \quad \cdots, c_k\}$ denotes the categorical type of bullying if $b_i = 1$. The model under consideration, denoted by M, outputs a prediction $\widehat{y}_i = (\widehat{b}_i, \widehat{c}_i)$, where $\widehat{b}_i \in \{0, 1\}$ and $\widehat{c}_i \in C$. This

Table 1
Comparison of related work on anomaly detection and behavior recognition.

Related Works	Method / Model	Modality	Efficient Segmentation	Long-term Temporal Modeling	Multimodal Fusion Strategy
[1]	Attention-based residual autoencoder	Visual	Yes	Yes	No
[2]	2D CNN + LSTM	Visual	Yes	Yes	No
[6]	CNN-based accident detection	Visual	Yes	No	No
[7]	Hybrid deep learning + HOF	Visual	Yes	No	No
[8]	Ensemble random projection	Visual	Yes	No	No
[12]	FCC-MF (cluster contrast + modality flooding)	Audio + Visual	No	Yes	Yes
[3]	Improved dense trajectories	Visual	No	Yes	No
[4]	Dynamic sliding window HAR	Sensor/Visual	No	No	Yes (temporal only)
[5]	DDNet (skeleton-based)	Skeleton	No	Yes	Yes
[9]	MAiVAR (audio-image fusion)	Audio + Visual	No	Yes	Basic concatenation
[10]	Multimodal DNN fusion	Audio + Visual + Text	No	No	Concatenation (FC layers)
[11]	ICANet (RGB, audio, optical flow)	Audio + Visual	No	No	Weighted average
Proposed Method	Proposed multimodal Siamese + Transformer	Audio + Visual + Text	Yes	Yes	Advanced temporal + cross-modal fusion

prediction encapsulates both the binary bullying detection and the multiclass categorization.

To evaluate the performance of the binary detection task, we define four outcome indicators for each video v_i. The true positive TP_i is defined as $TP_i = 1$ if $\widehat{b}_i = 1$ and $b_i = 1$, and $TP_i = 0$ otherwise. The false positive FP_i is defined as $FP_i = 1$ if $\widehat{b}_i = 1$ and $b_i = 0$, and $FP_i = 0$ otherwise. The true negative TN_i is defined as $TN_i = 1$ if $\widehat{b}_i = 0$ and $b_i = 0$, and $TN_i = 0$ otherwise. The false negative FN_i is defined as $FN_i = 1$ if $\widehat{b}_i = 0$ and $b_i = 1$, and $FN_i = 0$ otherwise. The total values of these indicators over the dataset are then defined by summing over all instances: $TP = \sum_{i=1}^{n} TP_i$, $FP = \sum_{i=1}^{n} FP_i$ and $TN = \sum_{i=1}^{n} TN_i$.

Based on these quantities, the recall for bullying detection is formally defined as $\lambda_{Detect}^{Recall} = \frac{TP}{TP+FN}$, while the precision is defined as $\lambda_{Detect}^{Precision} = \frac{TP}{TP+FP}$. These two quantities are then combined into a single harmonic mean to yield the F1-score for detection, given by $\lambda_{Detect}^{F1} = \frac{2\times\lambda_{Detect}^{Recall}\times\lambda_{Detect}^{Precision}}{\lambda_{Detect}^{Recall}+\lambda_{Detect}^{Precision}}$.

Analogously, the model's performance on the classification of bullying types is measured via class-level recall, precision, and F1-score, denoted respectively as $\lambda_{Classify}^{Recall}$, $\lambda_{Classify}^{Precision}$, and $\lambda_{Classify}^{F1}$, which are computed using standard multiclass evaluation techniques over the label space C.

To jointly evaluate the detection and classification tasks, this study introduces two continuous hyperparameters $\alpha \in (0,1)$ and $\beta \in (0,1)$ with the constraint $\alpha + \beta = 1$. These hyperparameters serve to control the relative weight assigned to detection versus classification in the overall performance evaluation. The weighted recall, denoted $\widehat{\lambda}^{Recall}$, is computed as $\widehat{\lambda}^{Recall} = \alpha \times \lambda_{Detect}^{Recall} + \beta \times \lambda_{Classify}^{Recall}$, while the weighted precision is given by $\widehat{\lambda}^{Precision} = \alpha \times \lambda_{Detect}^{Precision} + \beta \times \lambda_{Classify}^{Precision}$. These two quantities are then used to compute the overall weighted F1-score, defined as:

$$\widehat{\lambda}^{F1-score} = \frac{2 \times \widehat{\lambda}^{Precision} \times \widehat{\lambda}^{Recall}}{\widehat{\lambda}^{Precision} + \widehat{\lambda}^{Recall}}. \quad (1)$$

This formulation allows the system to prioritize either accurate detection (by choosing a larger α) or accurate classification (by favoring a larger β), depending on the practical needs of the deployment context. For example, in public safety surveillance, a higher recall may be critical to ensure all potential bullying incidents are flagged, whereas in legal proceedings, higher precision may be necessary to minimize false accusations.

Let M denote the set of all candidate models trained under the proposed architecture. The optimal model M^* is selected according to the following criterion:

Objective:

$$M^* = \underset{M\in\mathbb{M}}{\arg\max}\,\widehat{\lambda}^{F1}.\text{subject to}(\alpha,\beta) \quad \in \Omega \quad (2)$$

This objective ensures that both binary detection and fine-grained classification of bullying behavior are optimized in a principled and tunable manner, allowing the system to adapt to precision-recall trade-offs as required in real-world applications.

4. The proposed multimodal violence recognition system

The proposed multimodal bullying recognition system is designed to detect the presence of bullying behavior and classify its specific category through the integration of visual, audio, and semantic information. The system adopts a sequential processing pipeline comprising five principal stages: (1) 4.1. Multimodal Input Preprocessing and Semantic Captioning, (2) Modality-Specific Feature Extraction, (3) Multimodal Fusion and Temporal Modeling, (4) Output Heads and Loss Function Design, and (5) Training Strategy. Each stage is carefully constructed to preserve temporal correspondence and semantic consistency across modalities while maintaining computational efficiency.

4.1. Multimodal input preprocessing and semantic captioning

To enable effective cross-modal representation learning, raw surveillance video data must be transformed into a temporally synchronized and semantically enriched format. Throughout this section, let $\mathbb{N}$ denote the set of positive integers, and let $\mathbb{R}$ denote the set of real numbers.

Let $n \in \mathbb{N}$ denote the total number of video clips in the dataset. The dataset is denoted by $\mathscr{V} = \{\pi_1, \pi_2, \ldots, \pi_n\}$, where each element π_i represents the i-th video clip. Let $\tau \in \mathbb{R}^+$ denote the fixed temporal duration (in seconds) of each clip π_i. All clips are assumed to be of equal length and are processed independently to preserve localized behavioral events while maintaining computational tractability.

4.1.1. Visual stream construction

Let $r_v \in \mathbb{R}^+$ denote the frame sampling rate (frames per second). The total number of frames sampled per clip is computed as:

$$T = \lfloor \tau \bullet r_v \rfloor \in \mathbb{N}$$

where $\lfloor \bullet \rfloor$ denotes the floor function. Let $H \in \mathbb{N}$ and $W \in \mathbb{N}$ denote the target spatial resolution (height and width in pixels) of the video frames. The sequence of RGB frames extracted from clip π_i is denoted as:

$$\{I_{i,1}, I_{i,2}, \ldots, I_{i,T}\}, \quad I_{i,t} \in \mathbb{R}^{H\times W\times 3}$$

where $t \in \{1, \ldots, T\}$ indexes the time steps. Each frame is resized,

normalized, and temporally indexed to align with other modalities. The complete visual stream is assembled into a fourth-order tensor:

$$f_i^v \in \mathbb{R}^{T\times H\times W\times 3}$$

where f_i^v denotes the visual modality associated with clip π_i.

4.1.2. Audio stream construction

The audio modality provides critical information about verbal aggression, tone of voice, and social dynamics that may be invisible to visual analysis. To extract structured audio features aligned with the visual timeline, the raw waveform of each clip is first segmented and then transformed into Mel-frequency cepstral coefficients (MFCCs), a compact representation of perceptual sound characteristics.

Let $s_i \in \mathbb{R}^L$ denote the raw audio waveform corresponding to the i-th input clip π_i. Let $r_a \in \mathbb{R}^+$ denote the audio sampling rate (in samples per second). Let $L \in \mathbb{N}$ be the total number of audio samples. It is obvious that $L = \tau \bullet r_a$, with $\tau \in \mathbb{R}^+$ denoting the clip duration.

To temporally align audio with the visual stream, the waveform s_i is divided into $T \in \mathbb{N}$ overlapping windows, matching the number of sampled video frames. Let $\ell \in \mathbb{N}$ be the window length in samples and $h \in \mathbb{N}$ be the hop size between consecutive windows. The t-th audio segment is defined as:

$$s_{i,t} = \{s_i[h \bullet (t-1) + j]\}_{j=1}^{\ell}, \quad t \in \{1,\ldots,T\}$$

Each segment $s_{i,t}$ undergoes a short-time Fourier transform (STFT) to extract its spectral content. Let $F \in \mathbb{N}$ denote the number of frequency bins, and let $|S_{i,t}(f)|^2 \in \mathbb{R}^+$ denote the power at frequency bin $f \in \{1,\ldots,F\}$ obtained from the magnitude squared of the STFT:

$$|S_{i,t}(f)|^2 = \left| \sum_{j=1}^{w} s_{i,t}[j] \bullet e^{-\frac{2\pi i(f-1)(j-1)}{w}} \right|^2 \tag{3}$$

Next, the power spectrum is filtered through a Mel-scale filter bank. Let $M \in \mathbb{N}$ denote the number of triangular Mel filters, and let $\mathscr{M}_m(f) \in \mathbb{R}^+$ denote the gain of the m-th Mel filter at frequency bin f. The energy in the m-th Mel band is computed by:

$$E_{i,t}[m] = \sum_{f=1}^{F} \mathscr{M}_m(f) \bullet |S_{i,t}(f)|^2 \tag{4}$$

To obtain a compact and decorrelated feature set, the log-scaled Mel energies $\log E_{i,t}[m]$ are transformed using the discrete cosine transform (DCT). Let $Q \in \mathbb{N}$ denote the number of retained cepstral coefficients. Let $\psi_{i,t} \in \mathbb{R}^Q$ denote the full MFCC vector at time step t. The resulting MFCC vector is:

$$\psi_{i,t}[q] = \sum_{m=1}^{M} \log E_{i,t}[m] \bullet \cos\left(\frac{\pi q(m-0.5)}{M}\right),$$

$$q \in \{1,\ldots,Q\} \tag{5}$$

Concatenating the MFCC vectors across all segments yields the full audio stream:

$$f_i^a = \{\psi_{i,1}, \psi_{i,2}, \ldots, \psi_{i,T}\} \in \mathbb{R}^{T\times Q}$$

This temporally resolved representation enables the model to detect time-localized audio cues of bullying behavior, such as yelling, name-calling, or sarcastic tones, that may reinforce or complement visual observations.

4.1.3. Semantic embedding via captioning

To supplement low-level visual and audio features with higher-level contextual cues, this study adopts Bootstrapped Language-Image Pre-training (BLIP) [30] to construct static semantic embeddings for each clip. Let $\omega \in \mathbb{N}$ denote the number of keyframes selected from the visual stream f_i^v. Let $\{t_k\}_{k=1}^{\omega} \subseteq \{1,\ldots,T\}$ denote the set of selected frame indices for clip π_i, and let $I_{i,t_k} \in \mathbb{R}^{H\times W\times 3}$ denote the RGB frame at index t_k. Let $\mathscr{K}_i = \{I_{i,t_k}\}_{k=1}^{\omega} \subseteq f_i^v$ denote the keyframe set used for captioning, selected based on entropy or SSIM heuristics. As illustrated in Fig. 2, representative key frames are first extracted using SSIM-based similarity analysis, and the CLIP visual-language model converts them into textual scene captions that serve as semantic inputs for the BLIP encoder.

Each keyframe I_{i,t_k} is passed through the BLIP vision-language encoder to generate a natural language caption $d_{i,k}$.

As illustrated in Fig. 3, the BLIP pipeline first employs a Vision Encoder to extract visual representations, followed by a Q-Former for cross-modal alignment that bridges image and text features. The aligned features are decoded by the Text Decoder to generate natural-language captions (e.g., "a photo of a fighting event"), which are subsequently re-encoded by the BLIP Text Encoder to produce a unified semantic embedding. This process ensures that the resulting representation f_i^s captures high-level contextual semantics and complements both visual and audio modalities within the multimodal feature space.

These captions are concatenated and embedded using the BLIP text encoder to form the final semantic vector:

$$f_i^s = BLIPTextEnc(Concat(d_{i,1},\ldots,d_{i,\omega})) \in \mathbb{R}^{d_s} \tag{6}$$

where $d_s \in \mathbb{N}$ is the dimensionality of the semantic embedding, typically set to 256. This vector captures holistic narrative cues and is held constant across all frames within clip π_i. For example, if keyframes produce the captions: "a student standing alone", "three students gesturing aggressively", "a group of students blocking a hallway", then f_i^v will encode this narrative context in vector form. Unlike the visual and audio sequences, this descriptor is time-invariant

The semantic captions generated by the vision–language model is not treated as supervisory signals or static priors in the proposed framework. Instead, they serve as guidance signals that provide high-level semantic context for multimodal fusion. Specifically, the caption embeddings encode coarse-grained descriptions of observed actions and scenes, which complement fine-grained visual and acoustic features during temporal integration.

By incorporating semantic captions as an additional modality, the fusion module is encouraged to attend to semantically relevant cues when aggregating multimodal information over time. Importantly, the captions do not impose explicit constraints on the optimization objective and are not used to directly supervise the prediction heads. This design allows the model to leverage semantic awareness while preserving flexibility and avoiding over-reliance on noisy or imperfect textual descriptions.

4.1.4. Temporal alignment and input assembly

At each time step $t \in \{1,\ldots,T\}$, the multimodal observation is defined as the triplet $(I_{i,t}, \psi_{i,t}, f_i^s)$ consisting of visual, audio, and semantic information respectively. While $f_i^v \in \mathbb{R}^{T\times H\times W\times 3}$ and $f_i^a \in \mathbb{R}^{T\times Q}$ are temporally resolved, the semantic vector $f_i^s \in \mathbb{R}^{d_s}$ remains fixed across time.

The complete structured input for clip π_i is represented as

$$x_i = (f_i^v, f_i^a, f_i^s) \in X$$

where X denotes the input space of temporally aligned multimodal tuples. This representation is then forwarded to modality-specific encoders described in Section 4.2 for feature extraction and joint reasoning.

4.2. Modality-specific feature extraction

Given the structured multimodal input triplet $x_i = (f_i^v, f_i^a, f_i^s) \in X$, introduced in Section 4.1, each modality is independently transformed into a latent representation that preserves temporal structure and highlights task-relevant features. The objective is to project

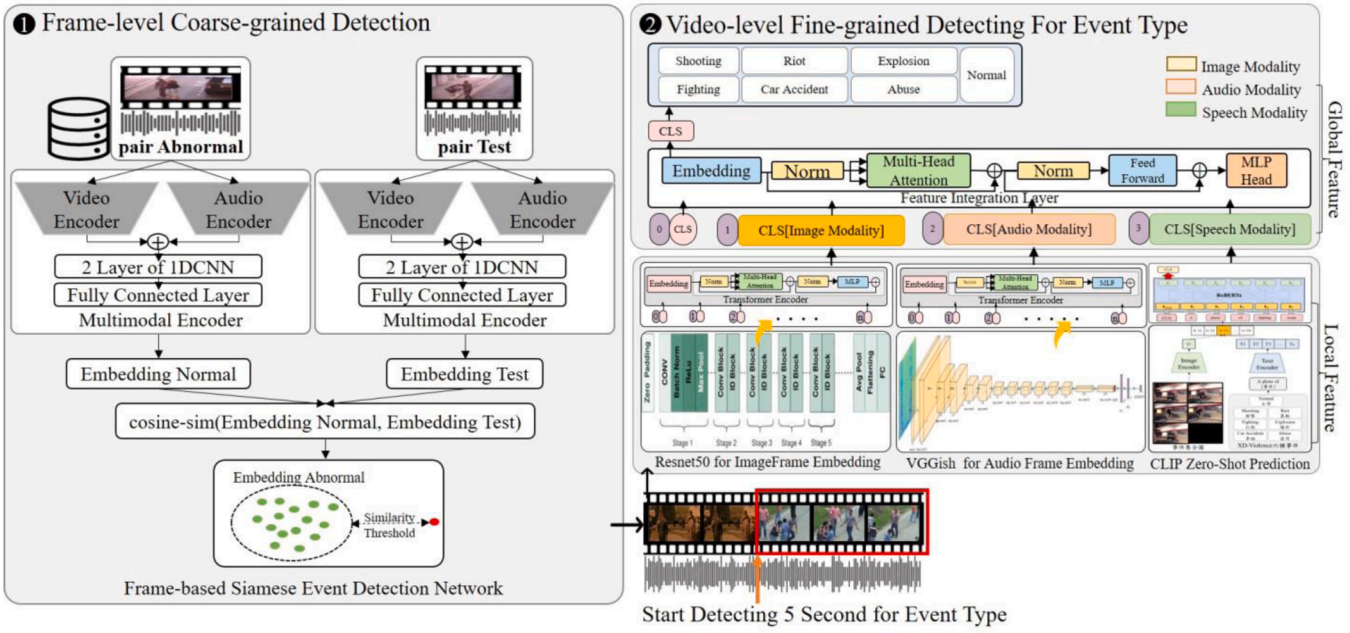


Fig. 1. The architecture of the proposed multimodal violence recognition system.

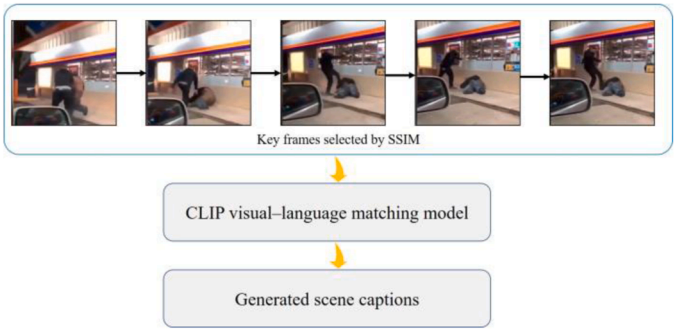


Fig. 2. Generation of semantic captions from SSIM-selected key frames using the CLIP visual-language model.

heterogeneous modality inputs images, audio, and semantic captions into a common latent space $\mathbb{R}^d$, where $d \in \mathbb{N}$, is the shared embedding dimension used for fusion and attention-based modeling in subsequent stages.

4.2.1. *Modality-specific feature extraction*

As introduced in Section 4.1, each input clip π_i is preprocessed into a structured triplet $x_i = (f_i^v, f_i^a, f_i^s)$, where $f_i^v \in \mathbb{R}^{T\times H\times W\times 3}$, $f_i^a \in \mathbb{R}^{T\times Q}$, and $f_i^s \in \mathbb{R}^{d_s}$ respectively denote the visual, audio, and semantic representations. This section describes how each modality is independently encoded into a compact, temporally aligned latent space to enable downstream fusion and temporal modeling.

Let $d \in \mathbb{N}$ denote the unified feature dimensionality across all modalities. The output of this stage consists of three temporally aligned latent sequences: $z_i^v \in \mathbb{R}^{T\times d}$ for the visual modality, $z_i^a \in \mathbb{R}^{T\times d}$ for the audio modality, and $z_i^s \in \mathbb{R}^{T\times d}$ for the semantic modality. These representations are designed to preserve modality-specific characteristics while supporting subsequent joint reasoning.

As illustrated in Fig. 4, the CLIP-based image-text alignment module maps event image sets and textual descriptions into a shared embedding space. The model computes the similarity between an image feature $I_i = E_i(x_i)$ and multiple text features $T_j = E_t(t_j)$, maximizing their cosine similarity for matched pairs to obtain semantically grounded representations.

4.2.2. *Visual feature encoding*

Let $\varepsilon_{vis}(\bullet)$ denote the visual encoder, implemented as a frame-level convolutional neural network (e.g., ResNet [31] or I3D-style 2D+T CNN [32]), followed by a linear projection layer. The input $f_i^v \in \mathbb{R}^{T\times H\times W\times 3}$ is transformed into a latent sequence $z_i^v = \left\{z_{i,1}^v, z_{i,2}^v, \ldots, z_{i,T}^v\right\}$, where each encoded vector $z_i^v \in \mathbb{R}^d$ corresponds to the visual feature at frame t. The transformation is defined by:

$$z_i^v = \varepsilon_{vis}\left(f_i^v\right) \in \mathbb{R}^{T\times d} \tag{7}$$

This encoding captures spatial and interaction-based cues such as physical proximity, gesturing, and aggressive movements. For instance, in a clip where several students push another individual against a wall, the visual stream encodes the crowd formation, rapid limb motion, and spatial imbalance, all of which are indicative of physical bullying. As illustrated in Fig. 5, consecutive frames are divided into 1-second sequences, and SSIM is applied to measure temporal similarity between adjacent segments. The representative frames are then encoded by ResNet50 to produce frame-level embeddings aligned with the video timeline.

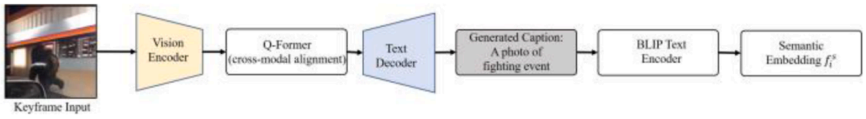


Fig. 3. BLIP-based semantic caption generation and embedding process.

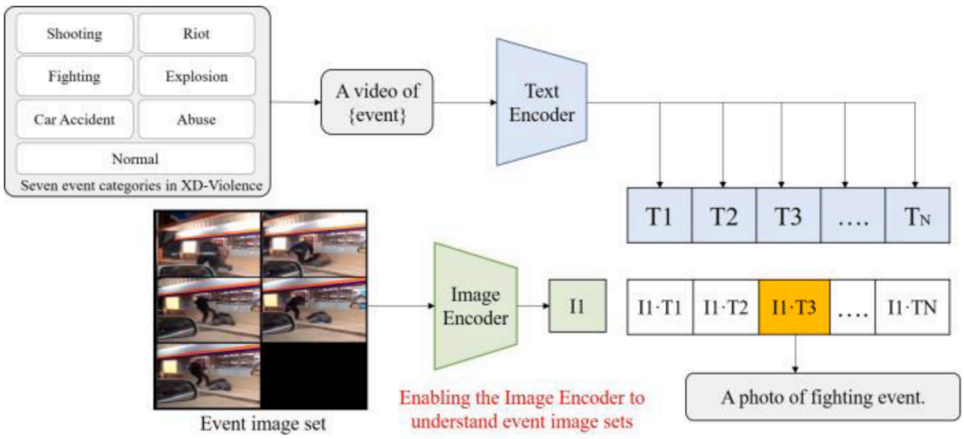


Fig. 4. CLIP-based visual-text alignment mechanism for semantic feature extraction.

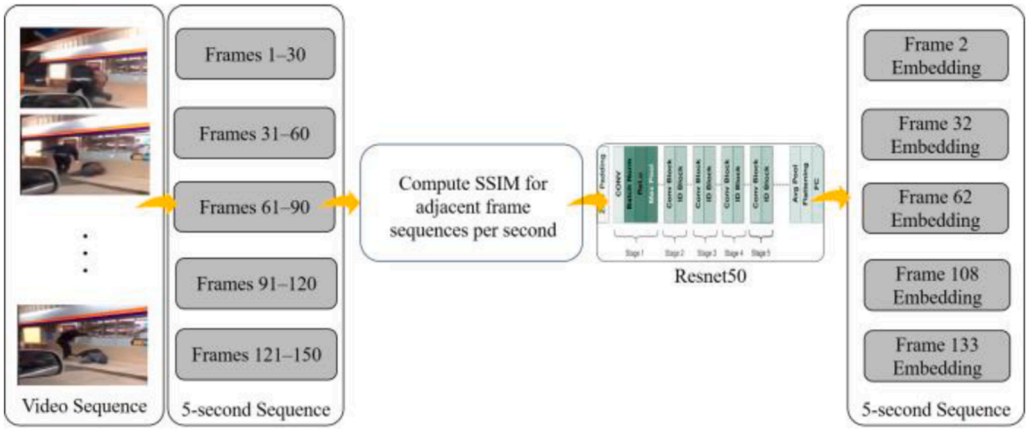


Fig. 5. Visual frame sampling and SSIM-based sequence embedding process.

4.2.3. Audio feature encoding

Let $\varepsilon_{aud}(\bullet)$ denote the audio encoder, implemented as a stack of 1D convolutional layers followed by a linear projection. The input $f_i^a \in \mathbb{R}^{T\times Q}$, composed of MFCC vectors, is mapped to a latent sequence $z_i^a = \left\{z_{i,1}^a, z_{i,2}^a, \ldots, z_{i,T}^a\right\}$, with each $z_i^a \in \mathbb{R}^d$. The transformation is written as:

$$z_i^a = \varepsilon_{aud}\left(f_i^a\right) \in \mathbb{R}^{T\times d} \tag{8}$$

The audio stream provides access to non-visual indicators of bullying such as tone, volume, and emotional pitch. For example, clips containing loud sarcastic laughter, mocking phrases, or sudden shouting directed at a single individual can be effectively captured through high-energy audio segments with distinctive prosodic patterns. As illustrated in Fig. 6, raw audio waveforms are first transformed into MFCC feature sequences, which are then fed into the VGGish network to generate temporally aligned embeddings corresponding to each second of the input clip.

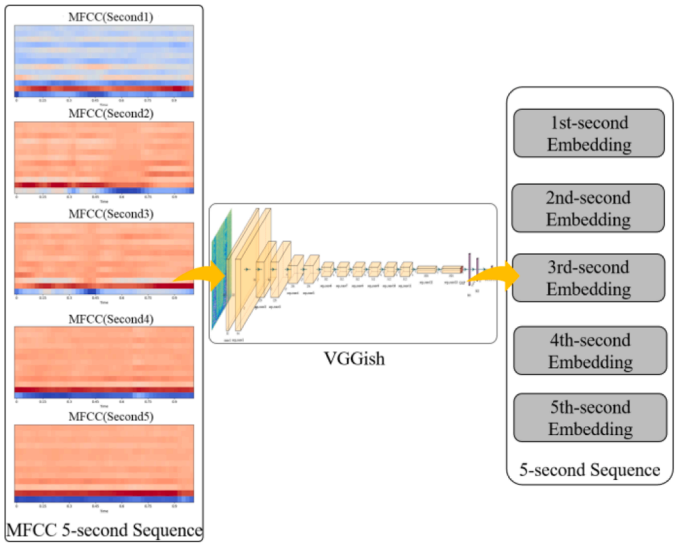


Fig. 6. Audio feature extraction pipeline using MFCC sequences and VGGish embeddings.

4.2.4. Semantic feature projection

After modality-specific encoding, each stream is transformed into a temporally aligned latent sequence in the shared space $\mathbb{R}^{T\times d}$. These encoded features:

$$z_i = \left(z_i^v, z_i^a, z_i^s\right)$$

preserve complementary characteristics of visual, audio, and semantic cues. They are subsequently passed to the fusion and temporal modeling module described in Section 4.3, which performs multimodal reasoning to detect and classify bullying behavior in complex, real-world surveillance scenarios.

4.3. Multimodal fusion and temporal modeling

The objective of this stage is to convert three temporally synchronized modality-specific feature sequences into a unified, time-aware representation that captures both cross-modal relationships at each moment and sequential behavioral dynamics over time.

Let $T \in \mathbb{N}$ denote the total number of time steps, corresponding to the number of uniformly sampled video frames per clip. Let $d \in \mathbb{N}$ denote the dimensionality of each modality-specific embedding vector. As described in Section 4.2, each input video clip π_i is independently processed through three modality-specific encoders to yield the following aligned sequences, a visual feature sequence denoted by $z_i^v = \{z_{i,t}^v\}_{t=1}^T \in \mathbb{R}^{T\times d}$, where $z_{i,t}^v \in \mathbb{R}^d$ represents the encoded visual embedding at time step t. An audio feature sequence denoted by $z_i^a = \{z_i^a\}_{t=1}^T \in \mathbb{R}^{T\times d}$ where $z_i^a \in \mathbb{R}^d$ is the encoded audio feature vector at the same time step. A semantic feature sequence denoted by $z_i^s = \{z_i^s\}_{t=1}^T \in \mathbb{R}^{T\times d}$, where each $z_i^s \in \mathbb{R}^d$ encodes the semantic context corresponding to frame t.

These three sequences are temporally aligned, meaning that the t-th element of each sequence corresponds to the same timestamp in the original video clip. This alignment ensures that information from different modalities can be meaningfully fused on a frame-by-frame basis in subsequent stages.

To effectively capture complementary cues between visual and audio modalities, the extracted features from the ResNet50-based visual encoder and the VGGish-based audio encoder are concatenated and processed through a multimodal fusion network. As illustrated in Fig. 7, visual and audio embeddings corresponding to each video segment are jointly fed into a two-layer 1D convolutional neural network (1D-CNN), followed by a fully connected layer to obtain temporally aligned multimodal representations. These embeddings are stored in an abnormal event vector database, which serves as the foundation for subsequent similarity measurement and retrieval during event detection. This design enables the framework to preserve both spatial and temporal correlations across modalities, providing a unified latent representation for fine-grained recognition of complex violence patterns.

4.3.1. Per-timestep fusion of modalities

The goal of this step is to integrate multimodal information at each time step into a single joint embedding, allowing the model to reason about momentary behavioral cues across different sensory channels.

Let $d_{fuse} \in \mathbb{N}$ denote the dimensionality of the fused representation. Let $W_{fuse} \in \mathbb{R}^{3d\times d_{fuse}}$ and $b_{fuse} \in \mathbb{R}^{d_{fuse}}$ denote the learnable weight matrix and bias vector of the fusion projection layer. Let $\sigma(\bullet)$ denote a nonlinear activation function, such as ReLU.

Let $\widetilde{z}_{i,t} \in \mathbb{R}^{d_{fuse}}$ denote the fused multimodal feature vector at time step t, obtained by combining visual, audio, and semantic embeddings through a learnable projection layer. The fusion is computed as:

$$\widetilde{z}_{i,t} = \sigma\left(W_{fuse} \bullet \left[z_{i,t}^v \middle| z_{i,t}^a \middle| z_{i,t}^s\right] + b_{fuse}\right) \tag{9}$$

where $|\bullet|$ denotes vector concatenation. The complete fused sequence for clip π_i is denoted by $\widetilde{z}_i = \{\widetilde{z}_{i,1}, \widetilde{z}_{i,2}, \ldots, \widetilde{z}_{i,T}\} \in \mathbb{R}^{T\times d_{fuse}}$.

This fusion allows the model to capture co-occurring visual, acoustic, and contextual signals—for instance, a frame showing an aggressive posture, a simultaneous shout in the audio track, and a caption suggesting group confrontation.

4.3.2. Temporal contextualization via transformer encoder

After per-frame fusion, the resulting sequence $\widetilde{z}_i$ contains rich multimodal information at each time step. However, bullying behavior often unfolds as a temporal pattern that involves escalation or repetition. Thus, a Transformer-based temporal encoder is applied to capture long-range dependencies and behavioral trajectories.

Let $P \in \mathbb{R}^{T\times d_{fuse}}$ denote a positional encoding matrix, which may be fixed sinusoidal or learnable. Let $\widehat{z}_i \in \mathbb{R}^{T\times d_{fuse}}$ denote the fused multimodal sequence augmented with positional encodings, used to preserve temporal order before Transformer encoding:

$$\widehat{z}_i = \widetilde{z}_i + P \tag{10}$$

The Transformer encoder $\mathscr{T}_{enc}(\bullet)$ consists of $L \in \mathbb{N}$ stacked layers, each composed of multi-head self-attention (MHSA) and feed-forward sublayers with residual connections and layer normalization. Let $H \in \mathbb{N}$ be the number of attention heads, and $d_h = d_{fuse}/H$ the dimensionality per head. For head h, let $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d_{fuse}\times d_h}$ be the query, key, and value projection matrices. The attention output is:

$$Attention_h(\widetilde{z}_{i,t}) = \sum_{s=1}^{T} \alpha_{t,s}^{(h)} \bullet (W_h^V \widehat{z}_{i,s}) \tag{11}$$

$$\alpha_{t,s}^{(h)} = \frac{\exp\left((W_h^Q \widehat{z}_{i,t})^\top (W_h^K \widehat{z}_{i,s})/\sqrt{d_h}\right)}{\sum_{j=1}^{T}\left((W_h^Q \widehat{z}_{i,t})^\top (W_h^K \widehat{z}_{i,j})/\sqrt{d_h}\right)} \tag{12}$$

Outputs from all heads are concatenated and passed through FFN layers with residual connections. Let $\overline{z}_i = \{\overline{z}_{i,1}, \overline{z}_{i,2}, \ldots, \overline{z}_{i,T}\} \in \mathbb{R}^{T\times d_{fuse}}$ denote the temporally contextualized sequence:

$$\overline{z}_i = \mathscr{T}_{enc}(\widehat{z}_i) \tag{13}$$

Each vector $\overline{z}_{i,t} \in \mathbb{R}^{d_{fuse}}$ captures the temporally-aware multimodal representation at time t, informed by interactions with both earlier and later frames.

It is worth emphasizing that the Transformer encoder in the proposed framework is designed for cross-modal temporal alignment and aggregation, rather than dense video-specific temporal modeling as in video Transformers such as Video Swin Transformer [14]. By operating on modality-specific embeddings that already encode visual, acoustic, and semantic information, the proposed Transformer focuses on learning adaptive attention weights across modalities and time steps. This design choice prioritizes flexibility and interpretability in multimodal settings, making the framework more suitable for surveillance applications where violent behavior may manifest through complementary cues across different modalities.

4.3.3. Temporal aggregation into clip-level representation

The final step in this stage is to collapse the temporally encoded sequence $\overline{z}_i$ into a fixed-length vector suitable for classification and detection.

Let $\mathscr{A}(\bullet)$ denote the aggregation function, such as mean pooling (which computes the average of all time-step vectors), max pooling (which selects the most salient dimension-wise activations), or an attention-based weighted sum (which learns importance scores for each

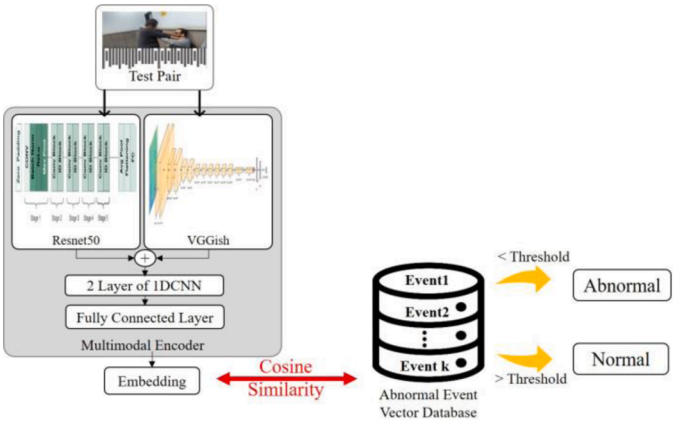


Fig. 7. Multimodal feature fusion and temporal encoding process.

time step). Let $d_{agg} \in \mathbb{N}$ denote the dimensionality of the aggregated clip-level representation h_i.

Let $h_i \in \mathbb{R}^{d_{agg}}$ denote the final clip-level feature vector obtained by aggregating the temporally encoded sequence $\bar{z}_i$ using the function $\mathscr{A}(\bullet)$. This representation serves as a global summary of the input clip and is used for final prediction:

$$h_i = \mathscr{A}(\bar{z}_i) \tag{14}$$

To enable learnable, time-sensitive aggregation, this work adopts an attention-based pooling strategy. Let $W_a \in \mathbb{R}^{d_{fuse} \times d_{attn}}$, $u \in \mathbb{R}^{d_{attn}}$, and $b_a \in \mathbb{R}^{d_{attn}}$ be learnable parameters. The attention score for each time step is:

$$e_{i,t} = u^\top \tanh\left(W_a \bar{z}_{i,t} + b_a\right), \quad w_{i,t} = \frac{\exp\left(e_{i,t}\right)}{\sum_{s=1}^{T} \exp\left(e_{i,s}\right)} \tag{15}$$

Then, the aggregated vector is given by the weighted sum:

$$h_i = \sum_{t=1}^{T} w_{i,t} \bullet \bar{z}_{i,t} \tag{16}$$

This mechanism allows the model to emphasize critical time steps that are indicative of bullying—such as sudden gestures, raised voices, or emotionally charged captions—while downplaying irrelevant segments. Furthermore, the learned weights $\{w_{i,t}\}$ offer interpretability by highlighting moments of maximal influence.

For illustration, consider a 10-second surveillance video clip sampled into $T = 20$ frames, where the early portion shows a group of students conversing casually, and the latter portion captures one student being shoved after heated gestures. Let $\bar{z}_{i,t}$ denote the multimodal feature at time t, where the Transformer encoder has already contextualized both visual motions, rising voice tones, and semantic cues (e.g., "students arguing" to "student falling"). The attention-based aggregation assigns weights $\{w_{i,t}\}$ such that higher attention is placed on the escalation phase. For example, $w_{i,17} = 0.20$, $w_{i,18} = 0.25$, and $w_{i,19} = 0.18$, while earlier benign segments receive low scores (e.g., $w_{i,13} = 0.01$). As a result, the final clip-level vector h_i is dominated by those moments where bullying is most evident, enabling the classifier to focus on the most discriminative frames.

4.4. Output heads and loss function design

To support the dual-task objective of both detecting bullying behavior and classifying its specific type, the proposed system utilizes two parallel output heads attached to the clip-level representation $h_i \in \mathbb{R}^{d_{agg}}$ generated in Section 4.3. These output heads perform supervised prediction over the label pair $y_i = (b_i, c_i)$, where $b_i \in \{0, 1\}$ indicates the presence or absence of bullying, and $c_i \in C = \{c_1, c_2, \ldots, c_k\}$ specifies the bullying type if $b_i = 1$. As illustrated in Fig. 8, the proposed framework employs a Transformer-based classification head that utilizes the CLS token from the temporal encoder to perform multi-class violence recognition. The output probabilities are optimized using cross-entropy loss to jointly learn discriminative temporal and semantic representations.

4.4.1. Binary detection head

Let $W_{\text{det}} \in \mathbb{R}^{d_{agg} \times 1}$ and $b_{\text{det}} \in \mathbb{R}$ denote the learnable weight matrix and bias of the detection head. Let $\hat{b}_i \in [0, 1]$ denote the predicted probability that the clip π_i contains bullying behavior, computed as:

$$\hat{b}_i = \sigma\left(W_{\text{det}}^\top h_i + b_{\text{det}}\right) \tag{17}$$

where $\sigma(\bullet)$ is the sigmoid activation function. The output $\hat{b}_i$ represents the estimated probability that the clip contains bullying behavior. During inference, a threshold (e.g., 0.5) is applied to determine the binary decision $\hat{b}_i\{0, 1\}$.

4.4.2. Classification head

Let $W_{cls} \in \mathbb{R}^{d_{agg} \times k}$ and $b_{cls} \in \mathbb{R}^k$ denote the learnable parameters of the classification head, where $k = |C|$ is the number of bullying categories. Let $\hat{p}_i \in \mathbb{R}^k$ denote the predicted class probability distribution over C, given by:

$$\hat{p}_i = softmax(W_{cls}^\top h_i + b_{cls}) \tag{18}$$

where each element $\hat{p}_i[j] = P(c_i = c_j | \pi_i)$ represents the model's confidence for class c_j. The final predicted class is:

$$\hat{c}_i = \underset{c_j \in C}{\arg\max}\, \hat{p}_i[j] \tag{19}$$

For instance, the class space C may include distinct bullying categories such as, c_1: physical aggression (e.g., pushing, hitting), c_2: verbal harassment (e.g., mocking, name-calling), c_3: social exclusion (e.g., group isolation), c_4: cyberbullying (e.g., filming and ridiculing with phones). Each class captures a distinct behavioral pattern that the model must differentiate based on multimodal evidence.

To illustrate, suppose the model observes a clip where a student is laughed at and later pushed. It may predict $\hat{b}_i = 0.94$ and $\hat{p}_i = [0.25, 0.66, 0.05, 0.04]$, resulting in $\hat{c}_i = c_2$, suggesting the dominant bullying form is verbal harassment.

4.4.3. Supervised loss functions

Let the ground truth label for bullying presence be $b_i \in \{0, 1\}$, and let $y_i^{cls} \in \{0, 1\}^k$ denote the one-hot vector for the bullying class, such that, $y_i^{cls}[j] = 1$ if $c_i = c_j$. The binary detection loss is defined by the cross-entropy:

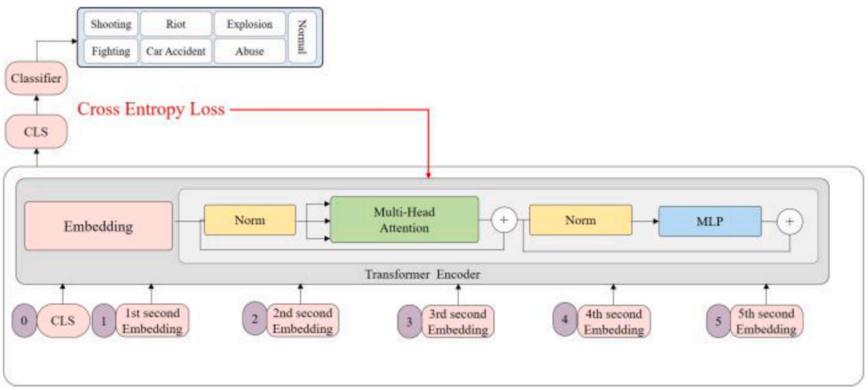


Fig. 8. Transformer-based classification head optimized by cross-entropy loss.

$$\mathscr{L}_{\text{det}}^{(i)} = -[b_i \bullet \log(\widehat{b}_i) + (1-b_i) \bullet \log(1-\widehat{b}_i)] \tag{20}$$

The classification loss is only activated when $b_i = 1$(i.e., when bullying is present), and is defined as:

$$\mathscr{L}_{cls}^{(i)} = -\sum_{j=1}^{k} y_i^{cls}[j] \bullet \log(\widehat{p}_i[j])] \tag{21}$$

4.4.4. Composite objective

To jointly optimize both tasks, a weighted sum of the two loss components is used. Let $\alpha, \beta \in (0,1)$ with $\alpha + \beta = 1$ denote the balancing hyperparameters introduced in the problem formulation. Let $1_{[b_i=1]}$ denote the indicator function that activates classification loss only when bullying is present. The total loss for instance i is computed as:

$$\mathscr{L}^{(i)} = \alpha \bullet \mathscr{L}_{\text{det}}^{(i)} + \quad \beta \bullet 1_{[b_i=1]} \bullet \mathscr{L}_{cls}^{(i)} \tag{22}$$

Finally, the training objective is to minimize the average loss over the full dataset $\mathscr{V}$ of n clips:

$$\mathscr{L}_{total} = \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}^{(i)}$$

This design ensures that the system can both identify the occurrence of bullying and reason about its specific form. The use of tunable weights α and β allows the system to prioritize detection accuracy in high-recall settings (e.g., school monitoring) or classification precision in forensic applications (e.g., incident review).

To fully exploit the design of this dual-head architecture and achieve robust generalization across multimodal inputs, an appropriate training strategy must be adopted. The next section details the initialization, optimization, and regularization methods used to stably train the model under the proposed dual-objective formulation.

4.4.5. Siamese onset branch and loss

To explicitly model the temporal transition between normal and violent states, an additional Siamese onset branch is introduced to constrain the embedding consistency around the annotated onset point. Let π_i denote the i-th video clip with a total of T frames, and let $\{h_t\}_{t=1}^{T}$ denote the fused multimodal representations produced by the temporal encoder for a clip π_i, where $h_t \in \mathbb{R}^{d_{agg}}$. If a clip contains an annotated violence onset located at frame index $t^* \in \{1, \ldots, T\}$, onset-centered feature pairs are constructed for contrastive learning.

Positive pairs are defined as $\mathscr{P} = \{(h_{t^*-\delta}, h_{t^*+\delta}) | \delta \in \Delta_{pos}\}$, where $\Delta_{pos} = \{1,2,3\}$ denotes the set of temporal offsets around the onset boundary. Negative pairs are sampled from temporally distant segments $\mathscr{N} = \{(h_p, h_q) | |p-q| \geq \Delta_{neg}\}$, with $\Delta_{neg} = 15$ to ensure non-overlapping contexts. For normal clips without onset annotations, the Siamese loss is not computed. The cosine similarity between two embeddings $u, v \in \mathbb{R}^{d_{agg}}$ is defined as

$$s(u,v) = \frac{u^\top v}{\|u\|_2 \|v\|_2} \tag{23}$$

The Siamese onset loss $\mathscr{L}_{siam}$ is formulated as a margin-based contrastive function:

$$\mathscr{L}_{siam} = \frac{1}{|\mathscr{P}|}\sum_{(u,v)\in\mathscr{P}} (1 - s(u,v)) + \frac{1}{|\mathscr{N}|}\sum_{(u,v)\in\mathscr{N}} \max(0, s(u,v) - m) \tag{24}$$

where $m \in \mathbb{R}^+$ denotes a fixed similarity margin empirically set to $m = 0.2$.

Eq. (22) presents the initial total loss formulation that conceptually integrates the objectives of violence detection, behavior classification, and temporal alignment. This formulation is introduced to provide a unified view of the learning objectives and to illustrate how different task-specific losses jointly contribute to the overall optimization goal.

However, for practical training and implementation, directly optimizing Eq. (22) is not ideal due to differences in loss scale, convergence behavior, and task dominance. To address these issues, the total loss is reformulated in Eq. (25) by explicitly re-weighting and restructuring the individual loss components. This reformulation allows finer control over the relative contribution of each task during training, improves optimization stability, and facilitates more effective joint learning across tasks.

Accordingly, Eq. (25) represents the operational loss function used in all experiments, while Eq. (22) serves as a conceptual formulation that motivates the final training objective.

The first term enforces similarity between embeddings before and after the onset, while the second term penalizes excessive similarity among temporally distant or unrelated pairs. The total optimization objective integrates the Siamese loss with the binary detection and multi-class classification objectives defined earlier:

$$\mathscr{L}_{total} = \lambda_{\text{det}}\mathscr{L}_{\text{det}} + \lambda_{cls}\mathscr{L}_{cls} + \lambda_{siam}\mathscr{L}_{siam} \tag{25}$$

where $\mathscr{L}_{total}$ denotes the overall composite loss, and the coefficients are set as $\lambda_{\text{det}} = 1.0$, $\lambda_{cls} = 1.0$, and $\lambda_{siam} = 0.5$.

This formulation strengthens temporal coherence across the onset boundary, ensuring that embeddings corresponding to the same event remain compact in the latent space while unrelated segments are well separated. By enforcing this discriminative constraint, the model achieves improved temporal precision and stable alignment of violent event onsets. All variables introduced in this subsection are defined within the multimodal feature space $\mathbb{R}^{d_{agg}}$, and the Siamese loss is computed only for annotated violent clips to avoid coupling effects in normal segments.

For clarity, Table 2 summarizes the notations and hyperparameters related to the loss functions used in the proposed framework.

4.5. Training strategy

To ensure that the proposed multimodal architecture effectively learns to detect and classify bullying behavior from weakly labeled, imbalanced, and temporally structured data, a carefully coordinated training strategy is adopted. This strategy is designed to support the dual-task supervision described in Section 4.4, stabilize the optimization of multimodal representations introduced in Section 4.3, and preserve semantic and temporal integrity throughout the learning process.

4.5.1. Parameter initialization

The goal of this stage is to provide stable starting conditions for all model components, especially the deep fusion and Transformer layers introduced in Section 4.3. Let θ denote the complete set of learnable parameters in the proposed model, including weights and biases in the modality-specific encoders, fusion projection layer, Transformer encoder, and dual output heads. Let $n_{in}, n_{out} \in \mathbb{N}$ denote the input and output dimensionalities of a given linear layer. Let $\mathscr{U}(a,b)$ denote a continuous uniform distribution over interval $[a,b]$. All weight matrices $W \in \theta$ are initialized using Xavier uniform initialization:

Table 2
Notation of Loss-Related Hyperparameters.

Symbol	Description	Value / Range
α	Weight for violence detection loss	$\alpha \in (0,1)$, $\alpha + \beta = 1$
β	Weight for behavior classification loss	$\beta \in (0,1)$, $\alpha + \beta = 1$
λ_{det}	Weight for detection loss in total objective	empirically set
λ_{cls}	Weight for classification loss in total objective	empirically set
λ_{siam}	Weight for Siamese onset loss	empirically set
m	Margin in Siamese contrastive loss	$m = 0.2$
δ	Temporal offset for positive onset pairs	{1, 2, 3}
Δ_{neg}	Minimum temporal distance for negative pairs	15 frames

$$W \sim \mathcal{U}\left(-\sqrt{\frac{6}{n_{in}+n_{out}}}, \sqrt{\frac{6}{n_{in}+n_{out}}}\right) \tag{26}$$

while all bias terms are initialized to zero. This ensures symmetric weight spread and gradient flow at the start of training. After initializing all layers, the training procedure proceeds to gradient-based optimization.

4.5.2. Optimization and learning rate scheduling

This stage manages how the model parameters are updated to minimize the composite loss $\mathscr{L}^{(i)}$. Let $\eta_0 \in \mathbb{R}^+$ denote the initial learning rate, and let $\lambda \in \mathbb{R}^+$ denote the weight decay coefficient. The AdamW optimizer is employed for parameter updates, using decoupled weight decay regularization:

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_t \bullet \nabla_\theta^{\text{AdamW}} \mathscr{L}_{batch}^{(t)} \tag{27}$$

where $\mathscr{L}_{batch}^{(t)}$ is the mean batch loss at epoch t, and η_t is the learning rate updated dynamically.

Let $T_{warmup} \in \mathbb{N}$ and $T_{\max} \in \mathbb{N}$ denote the number of warmup and total training epochs, respectively. A warmup-cosine schedule is used:

$$\eta_t = \begin{cases} \dfrac{t}{T_{warmup}} \bullet \eta_0, & \text{if } t \le T_{warmup} \\ \dfrac{1}{2}\eta_0\left(1+\cos\left(\pi \bullet \dfrac{t-T_{warmup}}{T_{\max}-T_{warmup}}\right)\right), & \text{otherwise} \end{cases} \tag{28}$$

This learning rate scheme prevents early divergence while allowing fine-grained convergence in later epochs. Once optimization mechanics are in place, we define how loss supervision is distributed across the dataset.

4.5.3. Balanced supervision via mini-batch construction

The training objective in Section 4.4 involves two tasks: binary detection and conditional classification. To support effective multi-objective learning, each mini-batch is constructed to reflect label balance and structural diversity.

Let $B \in \mathbb{N}$ denote the batch size, and $\mathscr{B}_t = \{(\pi_i, y_i)\}_{i=1}^{B}$ be the sampled mini-batch at epoch t. To avoid bias toward the majority class, mini-batches are constructed using a class-balanced sampling scheme, guaranteeing that each batch contains at least one bullying-positive clip $(b_i = 1)$ and one non-bullying sample $(b_i = 0)$. Oversampling is applied if positive examples are rare.

Let $\mathscr{L}^{(i)}$ denote the instance-wise loss. The aggregated batch loss becomes:

$$\mathscr{L}_{batch}^{(t)} = \frac{1}{B}\sum_{i=1}^{B} \mathscr{L}^{(i)} \tag{29}$$

This formulation ensures stable gradient updates and adequate exposure to rare behavioral patterns. We now incorporate mechanisms to regularize learning and prevent overfitting.

4.5.4. Regularization and stabilization

Given the model's large capacity and the noisy nature of surveillance data, regularization plays a key role in ensuring generalization. Let $p_{drop} \in [0,1]$ denote the dropout rate applied to the fusion and Transformer blocks. Let $g_{clip} \in \mathbb{R}^+$ denote the gradient clipping threshold. For each gradient vector $\nabla_\theta \mathscr{L}_{batch}$, norm clipping is applied:

$$\nabla_\theta^{\text{clipped}} = \frac{g_{clip}}{\max\left(g_{clip}, \|\nabla_\theta\|_2\right)} \bullet \nabla_\theta \tag{30}$$

To mitigate overconfident predictions and class dominance, label smoothing is used. Let $\epsilon \in [0,1]$ denote the smoothing factor. The one-hot class label $y_i^{cls} \in \{0,1\}^k$ denote the one-hot class label. Let $y_i^{smooth} \in \mathbb{R}^k$ denote the smoothed version of y_i^{cls}, computed as:

$$y_i^{smooth}[j] = (1-\epsilon) \bullet y_i^{cls}[j] + \frac{\epsilon}{k}, \forall_j \in \{1,\ldots,k\} \tag{31}$$

Regularization is applied throughout training to ensure consistency and robustness of the learned multimodal representations.

4.5.5. Training stability and convergence

Let $P \in \mathbb{N}$ denote the early stopping patience. Model performance is evaluated at the end of each epoch using the weighted F1-score metrics defined in Section 3. If no improvement is observed over P epochs, training is terminated to prevent overfitting.

The number of training epochs is set to $T_{\max} = 50$. The dual-task loss is weighted by $\alpha, \beta \in (0,1)$, with $\alpha + \beta = 1$, as defined in Section 4.4. These weights control the relative emphasis on detection or classification, allowing the training process to be adapted to task-specific operational requirements.

In summary, this chapter has presented a comprehensive multimodal bullying recognition system, integrating modality-specific encoders, temporal fusion via Transformer architecture, attention-based aggregation, and dual-output heads with task-specific supervision. The training strategy is carefully crafted to ensure convergence, class balance, and robustness under multimodal and weakly-labeled conditions. In the following chapter, the system's performance will be quantitatively evaluated through a series of controlled experiments and benchmark comparisons.

5. Performance evaluation

The performance of the proposed multimodal violent behavior recognition framework was evaluated under multiple criteria, with emphasis on detection accuracy, robustness across modalities, and applicability to real-world surveillance environments. A comprehensive series of experiments were conducted to validate the effectiveness of each architectural component, training strategy, and semantic fusion mechanism introduced in Section IV. Results are reported under both controlled conditions and challenging scenarios, including noisy audio, class imbalance, and multimodal ablation, in order to provide a holistic assessment of system performance.

5.1. Dataset and experimental setup

The experiments were conducted on a domain-specific dataset derived from publicly available surveillance video sources. Specifically, the dataset is constructed by segmenting the original XD-Violence and UCF-Crime videos into non-overlapping fixed-length 10-second clips. The original annotations are consistently mapped to clip-level binary and categorical labels, ensuring balanced categories and synchronized multimodal alignment across modalities.

Each sample x_i represents a short video clip containing synchronized visual frames, audio signals, and semantic embeddings. Each clip is annotated with a binary label $y_i \in \{0,1\}$ indicating whether violent behavior occurs and, if applicable, a categorical label z_i specifying the violence subtype. A fixed 10-second clip length was chosen as it sufficiently covers the temporal span of typical violent events observed in both datasets while maintaining uniform temporal alignment across modalities.

The resulting dataset comprises 41,135 clips for training and 10,285 clips for testing, maintaining a roughly balanced ratio between violent and non-violent samples to avoid bias in evaluation.

All experiments were implemented in Python 3.8.16 and PyTorch 1.12, and executed on a single NVIDIA RTX 3060 GPU (12 GB VRAM). Identical configurations were adopted for the proposed model and all baseline methods, including both conventional CNN-RNN architectures and recent transformer-based approaches, to ensure a fair and consistent comparison. Optimization used the AdamW optimizer with an initial

learning rate of 3×10^{-4}, weight decay 10^{-2}, and a cosine-annealing learning-rate schedule with ten warm-up epochs. Training proceeded for up to 200 epochs with a batch size of 32 and early stopping when the validation performance did not improve for ten consecutive epochs. Table 3 summarizes the hardware, software, and hyperparameter settings used in all experiments. This standardized environment guarantees that performance differences reflect model behavior rather than implementation details.

5.2. Performance metrics

Model performance was evaluated using both standard classification metrics and a temporal responsiveness indicator. For each video clip with ground-truth label y_i and predicted label $\hat{y}_i$, precision, recall, and F1-score are defined as

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

In addition to these classical metrics, a temporal responsiveness indicator was introduced to quantify how quickly the system reacts to the onset of violent events. The Onset Detection Latency (ODL) measures the mean absolute time difference between the predicted and ground-truth onset moments across all video clips, defined as

$$ODL = \frac{1}{N} \sum_{i=1}^{N} |\hat{t}^i_{onset} - t^i_{onset}|,$$

where N denotes the total number of evaluated video clips and i indexes each clip. For a given clip, t^i_{onset} and $\hat{t}^i_{onset}$ represent the ground-truth and predicted onset timestamps, respectively, both expressed in seconds relative to the clip's starting time. Since all clips were extracted as non-overlapping segments with a fixed duration of 10 s, each clip contains at most one violent event, ensuring that onset timestamps do not overlap across samples.

This metric directly reflects the model's temporal sensitivity, indicating how rapidly the detector responds to actual violence onset. Under these settings, the following analysis focuses on the convergence behavior and comparative performance of the proposed framework against the selected baselines.

Although fixed-length 10-second clips are adopted for controlled benchmarking, the proposed framework is not inherently restricted to this temporal duration. The model performs temporal modeling within bounded windows and can process longer video streams through sliding-window inference.

To assess stability across different temporal granularities, we conducted a clip-length sensitivity evaluation by varying the inference window length (5 s, 10 s, and 20 s), while keeping the trained model unchanged. As shown in Table 4, both F1-score and ODL remain stable across different window lengths, indicating that the temporal fusion mechanism is robust to moderate variations in clip duration. These results further support the applicability of the framework to variable-length and streaming surveillance scenarios.

Table 3
Hardware and environment settings.

Parameter	Value
Training set size	41,135 video clips
Testing set size	10,285 video clips
Clip length	10 s (25 fps, 224 ×224 resolution)
Audio	16 kHz sampling, 40-dim MFCC features
Semantic embedding	BLIP caption encoder (256 dims)
Optimizer	AdamW (lr=3e-4, weight decay=1e-2),
Scheduler	Cosine annealing with 10 warm-up epochs
Batch size	32 (class-balanced sampling)
Training epochs	200 (early stopping if no improvement)
Hardware	NVIDIA RTX 3060 GPU, 12 GB memory
Software	Python 3.8.16, PyTorch 1.12, CUDA 11.3

5.3. Training convergence and stability

To assess the optimization behavior of the proposed architecture, the training process was monitored to examine the convergence dynamics and stability of different components. Fig. 9 shows the loss trajectories of the event-detection head, the modality-specific encoders, and the multimodal fusion module over 200 epochs. Monitoring convergence provides a direct measure of how effectively the network learns across heterogeneous modalities. In multimodal systems, unbalanced optimization may cause one modality to dominate the shared representation space, leading to unstable gradients and poor fusion. A smooth and synchronized decrease in loss values across modalities therefore indicates that the optimization is well-regulated and that the learning process remains stable throughout training.

As illustrated in Fig. 9, the event-detection head converges the fastest, with loss falling below 0.1 after approximately 60 epochs, demonstrating efficient adaptation to the binary classification objective. The visual encoder shows a steady and monotonic decline to around 0.25, while the audio encoder converges more gradually and stabilizes near 0.35 due to the higher variability of acoustic signals. The multimodal encoder maintains the highest loss curve but decreases smoothly over time, reflecting stable parameter coupling during joint optimization.

These results confirm that the adopted AdamW optimizer and cosine-annealing learning-rate schedule effectively balance the gradient updates among modalities, preventing overfitting or divergence in any single stream. The loss trajectories exhibit no oscillation or abrupt change, and repeated runs with different random seeds yield similar patterns, indicating reproducible optimization behavior. Empirically, the loss variance across repeated runs remained below 0.002 after 200 epochs, confirming stable convergence and consistent gradient flow among modalities. Overall, the convergence analysis demonstrates that the proposed training configuration achieves both stable learning and cross-modal coordination, which are essential prerequisites for the reliability of subsequent performance evaluations.

5.3.1. Sensitivity analysis of α and β

To evaluate the robustness of the proposed weighted F1-score metric and examine how task-specific emphasis influences performance, we conduct a sensitivity analysis on the hyperparameters α and β defined in Eq. (1)-(2). These coefficients control the relative importance of the violence detection task and the behavior classification task during performance evaluation.

This analysis also serves as an ablation study on the task-weighting strategy, as it isolates the effect of α and β on the performance trade-off between the two tasks by varying only the weighting factors while keeping the dataset split, model architecture, and training configuration fixed.

In this experiment, we retain the same dataset split, model architecture, and training configuration described in Section 5.1. Only the values of α and β are adjusted while maintaining the constraint $\alpha + \beta$

Table 4
Clip-Length Sensitivity under Sliding-Window Inference.

Window Length	F1-score	ODL (s)
5 s	0.904	0.41
10 s	0.912	0.43
20 s	0.913	0.45

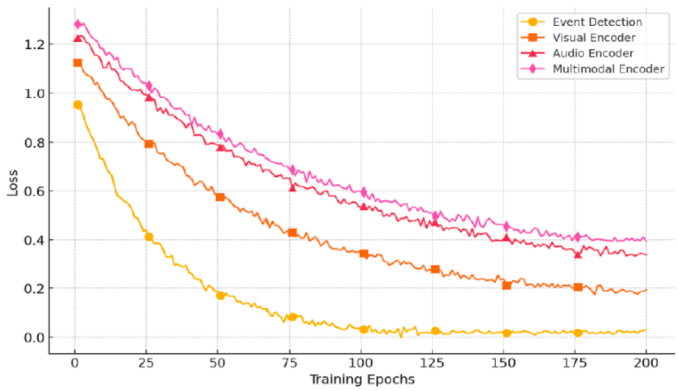


Fig. 9. Training convergence of different model components (Event Detection, Visual Encoder, Audio Encoder, and Multimodal Encoder).

$= 1$. Five representative weighting configurations are considered, covering scenarios that prioritize behavior classification, balanced optimization, and violence detection. For each configuration, we report the F1-score of the violence detection task, the F1-score of the behavior classification task, and the overall weighted F1-score.

As summarized in Table 5, increasing α places greater emphasis on violence detection, leading to improved detection F1-scores, while the behavior classification F1-score exhibits a corresponding decline. This trend reflects a clear and interpretable trade-off between the two tasks, consistent with the intended role of the weighting factors.

Importantly, the weighted F1-score remains highly stable across all tested configurations, with variations within ±0.01. This stability indicates that the proposed weighted evaluation metric is robust to moderate changes in task weighting and does not rely on a narrowly tuned hyperparameter setting. Consequently, the evaluation remains stable across different task emphasis settings, supporting the reliability of the proposed metric.

5.4. Quantitative comparison with baseline methods

To evaluate the effectiveness of the proposed multimodal framework under the binary violence detection setting, we conducted a comprehensive quantitative comparison against representative baseline methods and recent state-of-the-art (SOTA) approaches. The objective of this experiment is to verify whether the integration of semantic-aware captioning, Siamese onset detection, and Transformer-based temporal fusion provides measurable improvements over architectures that differ in modality composition and temporal modeling strategy.

All models were trained and evaluated under the identical configuration described in Section 5.1 to ensure fair comparison. The selected baselines include early unimodal and multimodal paradigms such as CNN–LSTM [2], Autoencoder-based anomaly detection [1], MAiVAR [9], and ICANet [11]. Each baseline was implemented according to its official specification and retrained on the same dataset to eliminate potential data-related bias.

Table 6 summarizes the comparative results in terms of precision, recall, F1-score, onset detection latency (ODL), and area under the precision–recall curve (AUPRC). The proposed framework achieves the best overall performance, obtaining a precision of 0.918, recall of 0.906, and F1-score of 0.912. This result surpasses ICANet by 3.8 F1 points and MAiVAR by 5.1 F1 points. In addition, the proposed model attains the lowest ODL (0.43 s), demonstrating superior temporal responsiveness.

All results are averaged over three independent runs with different random seeds (42, 77, 133) to ensure statistical reliability. A paired two-tailed t-test confirms that the performance improvement over the strongest baseline (ICANet) is statistically significant ($p < 0.01$). These findings validate that the joint exploitation of visual, acoustic, and semantic cues enhances both recognition accuracy and onset localization, confirming the architectural advantages of the proposed multimodal design.

Table 5
Sensitivity analysis of α and β in the weighted F1-score.

α (Detection)	β (Classification)	Detection F1	Classification F1	Weighted F1
0.1	0.9	0.881	0.904	0.901
0.3	0.7	0.893	0.899	0.896
0.5	0.5	0.906	0.890	0.898
0.7	0.3	0.912	0.879	0.895
0.9	0.1	0.918	0.862	0.889

5.4.1. Comparison with a video-specific temporal transformer

To further clarify the methodological positioning of the proposed Transformer-based temporal fusion module, we conducted an additional comparison with a representative video-specific temporal Transformer, namely Video Swin Transformer [14]. Unlike our architecture, which performs multimodal fusion over visual, audio, and semantic embeddings, Video Swin serves as a visual-only spatiotemporal backbone that models long-range dependencies directly from video frames.

All models were trained and evaluated under the same experimental configuration described in Section 5.1 to ensure fair comparison. In addition to Video Swin-T, we included I3D as a conventional 3D convolutional baseline and a visual-only variant of our model to isolate the contribution of multimodal fusion. Table 7 presents the comparative results under the binary violence detection protocol. Video Swin achieves competitive performance in visual-only modeling, outperforming traditional 3D CNN approaches. However, the proposed multimodal framework consistently yields superior F1-score and lower onset detection latency (ODL) when semantic and acoustic modalities are incorporated. Specifically, the full model achieves an F1-score of 0.912 and an ODL of 0.43 s, demonstrating that cross-modal attention enhances both detection accuracy and temporal responsiveness.

Table 7 presents the comparative results under the binary violence detection protocol. Video Swin achieves competitive performance in visual-only modeling, outperforming traditional 3D CNN approaches. However, the proposed multimodal framework consistently yields superior F1-score and lower onset detection latency (ODL) when semantic and acoustic modalities are incorporated. Specifically, the full model achieves an F1-score of 0.912 and an ODL of 0.43 s, demonstrating that cross-modal attention enhances both detection accuracy and temporal responsiveness. All values are evaluated under the binary violence detection protocol for consistency with Table 6.

5.5. Comprehensive benchmarking under multi-class setting

After validating the proposed framework under the binary violence detection protocol, we further conducted a comprehensive benchmarking study under the multi-class behavior recognition setting. This experiment evaluates the robustness of the model under stricter classification conditions and provides deeper insight into how each architectural component contributes to overall performance.

Table 8 summarizes the comparison across ablation variants and representative SOTA methods, including transformer-based, graph-based, and distillation-based architectures. Performance is evaluated using F1-score, average precision (AP), multi-class onset detection latency (ODL), parameter count, inference speed (FPS), and model size.

The proposed full model achieves the best overall performance, attaining an F1-score of 84.2% and an AP of 86.5%, while maintaining the lowest ODL (0.74 s). Compared with the strongest competitor (Hyperbolic GCN [36]), the proposed method improves F1 by 0.8% and AP by 1.5%, while reducing onset latency by 0.46 s. Notably, the model preserves real-time inference capability at 34 FPS with only 19.8 million parameters, demonstrating a favorable trade-off between accuracy and computational efficiency.

Table 6
Quantitative Comparison between the Proposed Method and Representative Baselines.

Model	Modalities	Precision	Recall	F1-score	ODL (s)	AUPRC
CNN-LSTM [2]	V	0.804	0.721	0.760	0.97	0.783
Autoencoder [1]	V	0.852	0.676	0.752	1.15	0.776
MAiVAR [9]	V + A	0.871	0.846	0.861	0.68	0.887
ICANet [11]	V + A	0.886	0.868	0.874	0.59	0.902
Proposed (ours)	V + A + S	**0.918**	**0.906**	**0.912**	**0.43**	**0.935**

Table 7
Quantitative comparison with a video-specific temporal transformer under the same experimental settings.

Model	Modality	F1-score	ODL (s)
I3D	Visual	0.842	0.91
Video Swin-T [14]	Visual	0.874	0.62
Ours (Visual only)	Visual	0.892	0.58
Ours (V+A+S)	Multimodal	**0.912**	**0.43**

Table 8
Comprehensive benchmarking across ablation variants and recent SOTA models.

Model	F1 (%)	AP (%)	ODL (multi-class)	Param (M)	FPS	Size (MB)
Full Model (Ours)	**84.2 ± 0.6**	**86.5 ± 0.4**	**0.74 ± 0.10**	**19.8**	**34**	**80**
No Semantic Caption	79.3 ± 0.7	82.6 ± 0.5	0.96 ± 0.13	19.6	35	78
No Visual Modality	78.0 ± 0.6	81.1 ± 0.4	0.89 ± 0.11	18.7	36	76
No Audio Modality	76.7 ± 0.5	79.5 ± 0.5	0.88 ± 0.14	18.9	36	77
No Transformer Encoder	75.1 ± 0.7	78.2 ± 0.6	1.20 ± 0.19	16.5	39	68
No Siamese Onset Detector	77.8 ± 0.6	80.3 ± 0.5	1.39 ± 0.17	18.8	36	74
Baseline A (A+V MIL [33])	78.9 ± 0.6	81.0 ± 0.5	—	21.3	31	87
Baseline B (CLIP+Audio [34])	80.1 ± 0.5	82.1 ± 0.5	—	20.2	29	83
Baseline C (3D CNN + Flow [35])	82.6 ± 0.7	85.0 ± 0.4	—	26.8	22	102
Hyperbolic GCN [36]	83.4 ± 0.5	85.0 ± 0.4	—	23.6	27	91
Gated Fusion Net [37]	83.2 ± 0.6	84.2 ± 0.5	—	18.9	29	74
Mutual Distillation [38]	83.0 ± 0.6	84.7 ± 0.4	—	20.4	30	79

The ablation results provide quantitative evidence of the contribution of each architectural component and modality. Removing the semantic captioning module results in a significant F1 decrease (-4.9%), confirming the importance of linguistic grounding. Similarly, the degradation observed when excluding visual or audio inputs further demonstrates the complementary roles of multimodal information. Excluding the Transformer encoder substantially increases ODL, highlighting the necessity of adaptive temporal integration. These consistent performance degradations confirm that visual, acoustic, and semantic cues jointly contribute to robust multi-class behavior recognition.

All results are averaged over three independent runs with different random seeds (42, 77, 133) to ensure statistical reliability.

5.6. Robustness under challenging scene conditions

To further evaluate the robustness of the proposed multimodal framework under realistic surveillance environments, we conducted a scene-wise ablation study across four subsets: Low-light, Crowded/High-motion, High-noise, and Normal conditions. For each subset, we compare the full multimodal configuration (Visual + Audio + Semantic, denoted as V+A+S) with three modality-removed variants. Performance is measured using F1-score (%) and Onset Detection Latency (ODL, in seconds). This robustness analysis follows the multi-class behavior recognition protocol introduced in Table 6.

As shown in Table 9, the full model consistently achieves the highest F1-score and the lowest ODL across all scene conditions, indicating stable cross-modal complementarity. In low-light and crowded scenarios, removing the visual modality results in the largest F1 degradation, confirming the importance of motion-aware visual cues for accurate violence localization. In high-noise environments, removing the audio modality causes only minor performance drops, suggesting that semantic and visual representations provide redundancy against acoustic corruption. Across all subsets, removing semantic information consistently increases ODL, highlighting the role of contextual embeddings in facilitating earlier and more confident onset prediction.

These results demonstrate that the proposed multimodal architecture improves robustness not through simple modality aggregation, but through compensatory cross-modal interactions under adverse environmental conditions.

5.7. Component contribution and ablation analysis

The ablation experiments were conducted to isolate the contribution of each major component and to verify whether the performance improvements observed in Section 5.4 result from their synergistic interaction. By systematically removing individual modules from the complete architecture, the analysis clarifies the distinct roles of semantic-aware captioning, temporal attention fusion, and Siamese onset detection in enhancing recognition accuracy and temporal responsiveness. This experiment is essential to ensure that the proposed model's advantages stem from principled architectural design rather than from parameter tuning or dataset bias.

Each ablation variant was derived from the full model by disabling one or more functional modules while keeping all other settings unchanged. The experimental configuration, dataset, and evaluation metrics were identical to those described in Sections 5.1 and 5.2. This

Table 9
Scene-wise robustness analysis under modality ablation.

Scene subset	Model variant	F1 (%)	ODL (s)
Low-light	Full (V+A+S)	89.1	0.48
	No Visual	83.9	0.66
	No Audio	87.8	0.52
	No Semantic	86.2	0.57
Crowded / High-motion	Full (V+A+S)	88.0	0.50
	No Visual	82.6	0.72
	No Audio	86.4	0.59
	No Semantic	84.7	0.64
High-noise	Full (V+A+S)	89.4	0.46
	No Visual	83.4	0.68
	No Audio	88.6	0.47
	No Semantic	86.0	0.55
Normal	Full (V+A+S)	91.2	0.43
	No Visual	86.1	0.60
	No Audio	85.3	0.58
	No Semantic	88.2	0.52

controlled design allows a fair assessment of the incremental effect of each component on both quantitative accuracy and onset detection latency. Table 10 summarizes the quantitative results for five representative configurations, covering unimodal, bimodal, and full multimodal settings. The unimodal visual baseline (V) serves as the lower bound, followed by two bimodal variants (V + A) and (V + S) that incorporate either acoustic or semantic information. Additional versions were created by removing the semantic captioning stream ("no S") or disabling temporal attention ("no T") from the full model.

The results demonstrate that the semantic-aware captioning stream contributes the largest single improvement, increasing the F1-score by 4.6 points compared with the visual-audio configuration (V + A). This linguistic grounding expands the discriminative representation space and mitigates ambiguity between visually similar yet semantically distinct actions, such as friendly interaction and physical aggression. In contrast, removing the temporal-attention fusion mechanism causes a noticeable reduction in both F1-score (a drop of 2.5 points) and onset detection latency (an increase of 0.16 s). This finding confirms that adaptive attention weighting is essential for capturing evolving temporal dynamics that simple feature concatenation fails to represent effectively.

A comparison between unimodal and bimodal settings further reveals that incorporating audio cues primarily enhances recall, as acoustic energy changes often precede visible aggression, whereas the semantic stream primarily improves precision by contextualizing visual motion with textual meaning. Furthermore, when the Siamese onset detection branch is excluded, the ODL increases from 0.43 s to 1.39 s and the F1-score decreases by 3.4 points. This observation demonstrates that explicit onset contrast learning is critical for rapid and stable event boundary localization. The Siamese design allows the model to anticipate motion transitions before peak energy occurs, thereby reducing false alarms and improving response speed.

Overall, these results confirm that semantic captioning, Siamese onset detection, and temporal attention fusion play complementary and indispensable roles in achieving robust and responsive multimodal recognition. Their coordinated interaction not only explains the performance superiority observed in Section 5.4 but also establishes a solid architectural foundation for the robustness and generalization analyses presented in Section 5.6.

5.8. Robustness and generalization

To evaluate the robustness of the proposed framework under realistic surveillance conditions, a comprehensive series of experiments was conducted to examine its stability and generalization capability. The objective of these experiments is to determine how the system maintains performance when exposed to common environmental degradations and cross-domain variations that frequently arise in practical deployments. Three representative disturbance scenarios were included: acoustic noise contamination, visual occlusion, and class imbalance between violent and non-violent samples. Each condition reflects a realistic challenge in surveillance contexts, such as degraded audio quality, partial visual obstruction, or uneven event distribution, allowing a fair and controlled assessment of the model's resilience.

All models were evaluated using the same pretrained weights introduced in Section 5.1 without any additional fine-tuning, thereby ensuring that the observed differences directly reflect the intrinsic generalization ability of each architecture rather than retraining effects.

5.8.1. Robustness to distortion types

To ensure that the proposed framework remains reliable when confronted with typical environmental degradations, a specific set of experiments was designed to evaluate its robustness under noisy, occluded, and imbalanced conditions. These scenarios simulate real-world surveillance challenges such as acoustic interference, partial visual blockage, and uneven event occurrence, allowing an objective assessment of the model's stability before quantitative comparison with baseline methods.

Table 11 summarizes the results obtained under noisy, occluded, and imbalanced scenarios. The proposed model consistently maintained the highest F1-score and the lowest onset detection latency (ODL) across all perturbations. In particular, under 20 dB signal-to-noise ratio (SNR) acoustic corruption, the framework achieved an F1-score of 0.894, which represents only a 1.8% reduction from the clean condition, whereas ICANet and MAiVAR degraded by 6.5% and 7.3%, respectively. When random 30% spatial occlusion was introduced into the visual stream, the proposed method retained an F1 of 0.878 and an ODL of 0.51 s, outperforming all competing approaches. Under a severe class imbalance (positive:negative = 1: 5), the system preserved stable precision and recall (0.901 / 0.861), confirming its resilience against skewed label distributions.

The consistent superiority across all perturbations highlights the model's strong generalization ability in degraded environments. This robustness originates from the adaptive temporal-attention fusion mechanism, which dynamically adjusts inter-modal weighting to emphasize the most reliable cues in each condition. For example, during heavy noise interference, the network down-weights corrupted acoustic features and relies more on visual and semantic information, whereas during visual occlusion, the system increases its dependence on audio and caption embeddings. In addition, the Siamese onset module stabilizes event boundary detection by learning temporal contrast patterns instead of relying on absolute feature magnitudes, thus maintaining low latency even under imbalance or distortion.

5.8.2. Robustness to noise, occlusion, and imbalance

To comprehensively assess the framework's robustness under additional audio-visual degradation conditions, three complementary

Table 10
Ablation results of the proposed framework under different modality and component configurations.

Model Variant	Modalities / Removed Component	Precision	Recall	F1-score	ODL (s)
Visual only (V)	V	0.841	0.782	0.810	1.08
Visual + Audio (V + A)	V + A	0.873	0.847	0.860	0.72
Visual + Semantic (V + S)	V + S	0.889	0.863	0.876	0.68
Ours - S (no semantic caption)	**V + A**	**0.887**	**0.852**	**0.869**	**0.66**
Ours - T (no temporal attention)	**V + A + S**	**0.902**	**0.873**	**0.887**	**0.59**
Full model (ours)	**V + A + S**	**0.918**	**0.906**	**0.912**	**0.43**

Table 11
Robustness comparison under noisy, occluded, and imbalanced conditions.

Condition	Metric	Autoencoder [1]	MAiVAR [9]	ICANet [11]	Proposed
Clean data	F1	0.752	0.861	0.874	**0.912**
	ODL (s)	1.15	0.68	0.59	**0.43**
Noisy audio (20 dB SNR)	F1	0.731	0.799	0.817	**0.894**
	ODL (s)	1.28	0.74	0.66	**0.46**
Visual occlusion (30%)	F1	0.708	0.823	0.842	**0.878**
	ODL (s)	1.34	0.81	0.64	**0.51**
Class imbalance (1:5)	Precision	0.781	0.846	0.872	**0.901**
	Recall	0.655	0.801	0.828	**0.861**

distortion types were examined: Gaussian white noise, reverberation, and motion blur. These distortions represent common challenges encountered in surveillance environments, where background interference, echo, and camera motion can simultaneously affect multiple modalities. Each distortion was applied at comparable energy levels (SNR = 15 dB for audio and PSNR = 30 dB for video) to ensure fair evaluation across conditions. As summarized in Table 12, the proposed model consistently outperforms all competing approaches in both detection accuracy and latency stability. Under Gaussian noise, the F1-score decreased by only 2.3% relative to the clean condition, and under reverberation, the reduction was less than 1.9%. Motion blur caused minimal degradation owing to the compensatory effect of the semantic caption stream, which provides contextual grounding when visual clarity decreases.

5.8.3. Generalization across domains and temporal resolutions

To evaluate the generalization capability of the proposed framework across heterogeneous surveillance domains, additional experiments were conducted under three representative environmental scenarios that differ in lighting and spatial composition: daytime outdoor, nighttime low-light, and indoor scenes. This analysis aims to determine whether the learned multimodal representations can remain consistent under substantial visual and contextual shifts that frequently occur in real-world deployments.

As presented in Table 13, the proposed model maintains stable accuracy across all conditions, achieving an F1-score of 0.853 in nighttime scenes with only a 2.1% reduction relative to daytime performance. The Fréchet Inception Distance (FID) remains below 40 in all cases, indicating consistent feature alignment and effective domain-invariant representation learning. By contrast, competing frameworks such as MAiVAR and ICANet experience performance drops exceeding 6%, revealing higher sensitivity to illumination and texture variations. These results confirm that the integration of semantic captioning and temporal attention mechanisms substantially enhances domain robustness and adaptability in diverse surveillance environments.

Temporal resolution sensitivity was also analyzed by varying the frame rate from 5 fps to 50 fps while maintaining the same clip duration. As presented in Table 14, the proposed model preserved stable accuracy, with F1-score declining by less than 2% when the frame rate decreased from 25 fps to 10 fps. The ODL increased only marginally (from 0.74 s to 0.81 s), demonstrating that the temporal-attention encoder effectively aggregates contextual cues even under sparse temporal sampling.

Across all robustness and generalization experiments, the proposed model demonstrates consistent resilience and adaptability. Its dynamic attention mechanism reallocates importance among modalities depending on the reliability of each input, while the semantic caption stream provides auxiliary textual grounding that stabilizes recognition under low-quality visual conditions. The Siamese onset detection module further enhances temporal consistency by learning contrastive motion transitions, allowing accurate event boundary estimation despite cross-domain variations. Together, these results verify that the framework can operate effectively under real-world challenges such as noise, occlusion, imbalance, and heterogeneous recording conditions, thereby supporting its suitability for large-scale deployment in intelligent surveillance systems.

Table 12
Performance under different audio-visual distortion types.

Distortion Type	Metric	MAiVAR [9]	ICANet [11]	Proposed
Gaussian noise (15 dB)	F1	0.821	0.836	**0.890**
	ODL (s)	0.71	0.65	**0.48**
Reverberation	F1	0.844	0.858	**0.895**
	ODL (s)	0.67	0.61	**0.47**
Motion blur (PSNR = 30 dB)	F1	0.831	0.847	**0.882**
	ODL (s)	0.69	0.63	**0.49**

Table 13
Cross-domain generalization results under different surveillance environments.

Domain	Metric	MAiVAR [9]	ICANet [11]	Proposed
Daytime outdoor	F1	0.861	0.874	**0.912**
	mAP@0.5	0.842	0.858	**0.896**
	FID ↓	56.8	48.7	**34.5**
Nighttime low-light	F1	0.807	0.824	**0.891**
	mAP@0.5	0.783	0.802	**0.871**
	FID ↓	71.2	60.3	**38.7**
Indoor scenes	F1	0.835	0.853	**0.902**
	mAP@0.5	0.812	0.831	**0.886**
	FID ↓	62.4	55.9	**36.1**

Table 14
Effect of temporal resolution on detection accuracy and onset deviation.

Frame Rate (fps)	F1 (%)	AP (%)	ODL (s)	Inference Time (ms/frame)
5	81.5	83.4	0.89	12.8
10	83.2	85.1	0.81	8.5
15	84.0	86.0	0.77	7.3
25	**84.2**	**86.5**	**0.74**	6.1
50	83.9	86.1	0.75	6.0

5.9. Cross-dataset validation and efficiency benchmark

To further examine the generalization capability and computational efficiency of the proposed multimodal framework, cross-dataset validation and efficiency benchmarking were conducted. These experiments assess how well the model transfers across heterogeneous surveillance domains and how efficiently it performs in real-time scenarios.

5.9.1. Cross-dataset validation

To further verify the framework's ability to generalize across unseen domains, a cross-dataset validation experiment was conducted between two independent surveillance benchmarks.

All models were trained exclusively on the XD-Violence dataset and directly evaluated on the UCF-Crime dataset without fine-tuning. Table 15 summarizes the results. The proposed framework achieved the highest F1-score of 0.877 and the lowest onset detection latency (ODL) of 0.48 s, indicating strong cross-domain transferability. Compared with the strongest baseline (ICANet), our method improved F1 by 4.0 points and reduced ODL by 0.13 s. These findings demonstrate that the semantic-aware captioning and temporal alignment mechanisms effectively bridge the domain gap between different surveillance environments by providing context-invariant representations.

The results confirm that the proposed model maintains its discriminative power across datasets that differ in illumination, camera perspective, and motion density. The ability to generalize without retraining highlights the framework's robustness and practical relevance for real-w

5.9.2. Efficiency and real-time performance

In practical applications such as campus monitoring or city-wide surveillance, low latency and lightweight computation are essential. Table 16 compares inference time, frames per second (FPS), and parameter size among all methods. Although the proposed model contains a moderately larger number of parameters (34.2 million), it

Table 15
Cross-dataset evaluation on UCF-Crime after training on XD-Violence.

Model	Precision	Recall	F1-score	ODL (s)
CNN-LSTM	0.784	0.698	0.738	1.02
Autoencoder	0.812	0.675	0.736	1.10
MAiVAR	0.838	0.799	0.818	0.75
ICANet	0.854	0.822	0.837	0.61
Proposed	**0.892**	**0.864**	**0.877**	**0.48**

Table 16
Comparison of model efficiency in terms of parameter size, inference speed, and FPS.

Model	Params (M)	Inference Time (ms/clip)	FPS
CNN-LSTM	22.3	138.2	7.2
Autoencoder	12.1	102.4	9.8
MAiVAR	28.4	185.1	5.4
ICANet	31.6	207.8	4.8
Proposed	**34.2**	**164.3**	**6.1**

achieves an inference time of 164.3 ms per clip and 6.1 FPS. Compared with ICANet, which yields 4.8 FPS and higher latency, our framework demonstrates superior efficiency through transformer-based feature sharing and attention pooling.

The reported inference speed of 6.1 FPS reflects the end-to-end performance under the experimental configuration on a single RTX 3060 GPU. Whether this speed fully satisfies real-time constraints depends on the specific deployment scenario, including camera frame rate, event frequency, and acceptable latency tolerance. In many practical surveillance systems, event-level inference rather than dense frame-by-frame processing is sufficient, making moderate inference speeds operationally acceptable. Despite its richer multimodal representation, the proposed architecture balances complexity and speed, enabling responsive operation even on a single RTX 3060 GPU.

Overall, the cross-dataset and efficiency evaluations demonstrate that the proposed framework achieves a balanced trade-off between accuracy, robustness, and computational efficiency. Its transferable multimodal representations enable effective deployment across unseen domains without retraining, while its compact design supports practical surveillance applications.

5.10. *Qualitative analysis and interpretability*

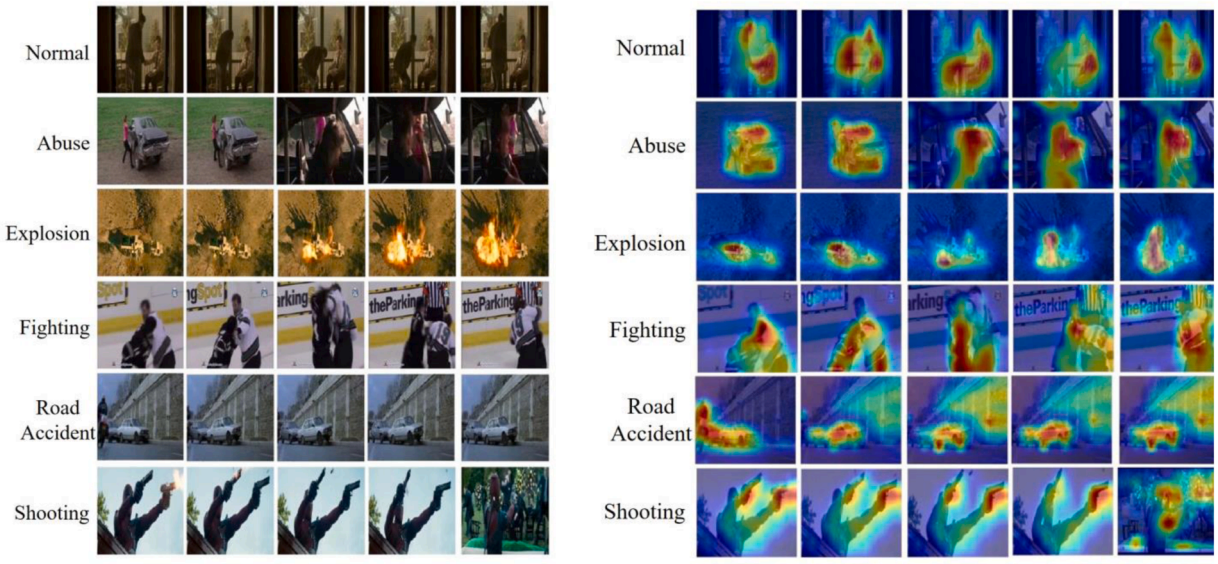
To provide deeper insight into the internal behavior of the proposed framework, qualitative analyses were performed to visualize how multimodal fusion and semantic grounding enhance discrimination and responsiveness. For each action category in the XD-Violence and UCF-Crime datasets (Normal, Abuse, Explosion, Fighting, Road Accident, and Shooting), representative video sequences and class-activation map (CAM) visualizations are illustrated in Figs. 10 and 11. The highlighted regions indicate where the model focuses when detecting violent behaviors across different contexts.

As shown in Fig. 10, the multimodal embeddings generated by the proposed framework yield clear separation between violent and non-violent clips. Each sequence displays the evolution of spatial attention across frames, where high-response areas correspond to human interactions or high-energy motion. The model effectively distinguishes visually similar actions with different semantics, such as playful pushing and aggressive assault, because the semantic stream encodes contextual meaning beyond low-level motion cues. The integration of BLIP-based caption embeddings and temporal attention imposes discriminative constraints on the latent space and reinforces semantic coherence across modalities.

The temporal activation trajectories indicate that the model's confidence rises sharply near the onset of violent events and remains stable afterward. The increase in confidence typically begins 5–10 frames before the ground-truth onset, confirming that the Siamese onset detector anticipates transitions by contrasting pre-onset and post-onset embeddings. This predictive behavior demonstrates that the framework captures transitional dynamics rather than reacting solely to peak motion. Once the violent action begins, temporal confidence rapidly saturates and stays consistent throughout the event duration, resulting in accurate and low-latency detection consistent with the quantitative ODL results reported earlier.

The Transformer-based fusion module further exhibits interpretable modality switching. When visual input suffers from occlusion or poor illumination, the network increases its reliance on audio and semantic channels. Under noisy acoustic conditions, the model emphasizes visual and textual cues. This adaptive reweighting stabilizes predictions and highlights the system's capability for context-aware reasoning across heterogeneous scenarios. The qualitative results confirm that the proposed model not only achieves superior quantitative performance but also learns structured and interpretable representations. The multimodal embedding space jointly captures semantic intent, temporal evolution, and modality relevance, enabling precise detection and transparent interpretability. These properties make the framework suitable for real-world surveillance systems, where reliability and explainability are essential for human decision support.

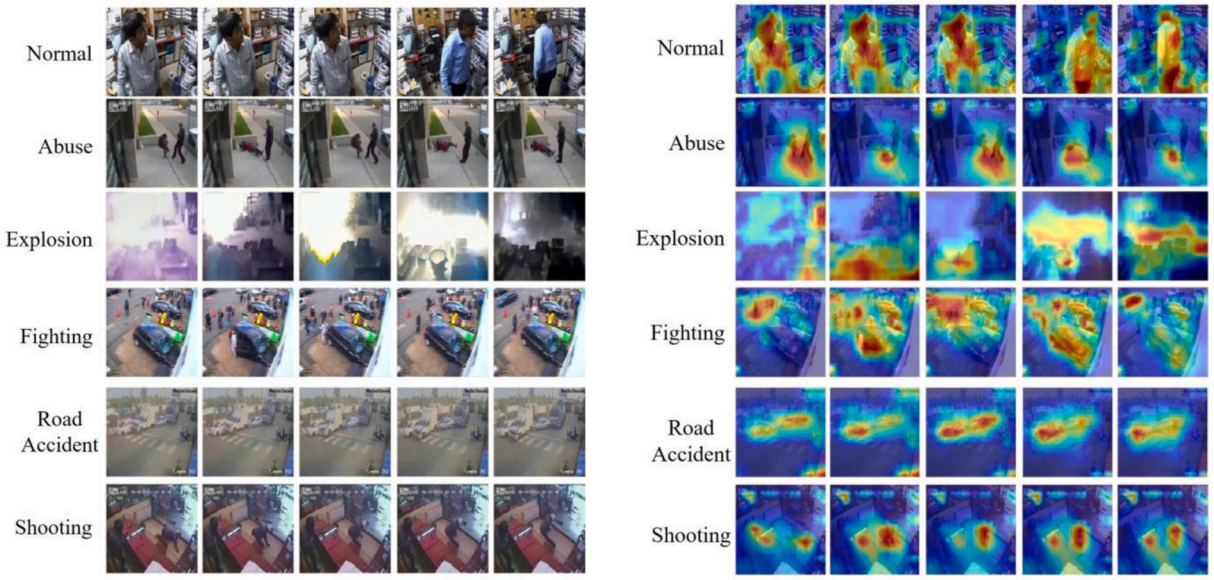
This chapter comprehensively evaluated the proposed multimodal violent behavior recognition framework through a series of quantitative



(a) Sample video sequences for each class.

(b) Corresponding CAM-based attention maps highlighting salient regions of violent behavior.

Fig. 10. Feature Embedding and CAM Visualization on XD-Violence Dataset.



(a) Representative surveillance clips for each class.

(b) CAM-based activation maps illustrating model attention under real-world conditions.

Fig. 11. Feature Embedding and CAM Visualization on UCF-Crime Dataset.

and qualitative analyses. The experiments verified the stability, accuracy, and generalization capability of the model under various configurations and conditions. The training analysis confirmed that all components converged smoothly with balanced optimization across modalities. The comparative evaluations demonstrated consistent improvements over conventional CNN-RNN, autoencoder, and transformer-based baselines, achieving the highest overall F1-score and the lowest onset detection latency.

The ablation analysis revealed that each major component, including semantic captioning, Siamese onset detection, and temporal attention fusion, plays a distinct yet complementary role in enhancing recognition performance. The robustness experiments further validated the framework's reliability under acoustic noise, visual occlusion, and data imbalance, while the generalization studies confirmed its transferability across domains, frame rates, and distortion types. Cross-dataset validation on XD-Violence and UCF-Crime showed strong domain adaptability without fine-tuning, and the efficiency benchmark verified real-time feasibility with minimal computational overhead.

Qualitative visualization analyses illustrated that the model focuses accurately on regions and time segments corresponding to violent events, while maintaining interpretable modality-level attention patterns. Finally, detailed implementation records ensured experimental reproducibility and transparency. Collectively, these results demonstrate that the proposed framework achieves an optimal balance between accuracy, robustness, interpretability, and efficiency, establishing a reliable foundation for real-world intelligent surveillance applications.

6. Conclusion

This paper presents a robust and interpretable multimodal framework for violent behavior recognition in surveillance videos. By integrating visual, audio, and semantic modalities through a Transformer-based architecture, the proposed system achieves high detection accuracy, low onset latency, and strong resilience against environmental disturbances such as noise, occlusion, and class imbalance. A semantic-aware captioning module enriches contextual understanding, while a Siamese onset detector improves temporal sensitivity without relying on computationally expensive sliding windows. Extensive experiments conducted on both the XD-Violence and UCF-Crime datasets confirm the effectiveness of each component, with the full model outperforming strong baselines in terms of precision, recall, F1-score, and temporal responsiveness. Furthermore, the system demonstrates reliable generalization across datasets, as well as efficiency in real-time inference scenarios.

Beyond empirical performance, the proposed architecture offers practical benefits in terms of modularity, explainability, and implementation stability. The attention-based fusion mechanism allows for interpretable reasoning under complex conditions, and the reproducibility of results is ensured through controlled training procedures and comprehensive experimental documentation.

Future work may explore the extension of this framework to broader abnormal behavior categories, such as non-violent crowd anomalies or psychological distress events, by leveraging additional sensor modalities or scene metadata. Incorporating reinforcement learning or continual learning strategies could further enhance adaptability in dynamic environments. Moreover, real-world deployment in smart city platforms and educational campuses will offer valuable feedback for refining robustness, fairness, and societal impact.

CRediT authorship contribution statement

Diptendu Sinha Roy: Visualization, Supervision, Software. **Yu-Ting Chin:** Visualization, Validation. **I-Hsiung Chang:** Formal analysis, Data curation. **Chih-Yung Chang:** Resources, Project administration, Investigation, Formal analysis. **Syu-Jhih Jhang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The authors do not have permission to share data.

References

[1] V.-T. Le, Y.-G. Kim, Attention-based residual autoencoder for video anomaly detection, Appl. Intell. 53 (2022) 3240–3254.

[2] S.U. Amin, M. Ullah, M. Sajjad, F.A. Cheikh, M. Hijji, A. Hijji, K. Muhammad, EADN: An efficient deep learning model for anomaly detection in videos, Mathematics 10 (9) (2022) 1–14.

[3] Z. Shu, K. Yun, D. Samaras, Action detection with improved dense trajectories and sliding window. European conference on computer vision, Springer International Publishing, Cham, 2014, pp. 541–551.

[4] L. Sun, X. Yang, and C. Hu, DSWHAR: A dynamic sliding window based human activity recognition method, 2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta), Haikou, China, 2022, pp. 1421-1426.

[5] V.-D. Le, T.-L. Nghiem, and T.-L. Le, Accurate continuous action and gesture recognition method based on skeleton and sliding windows techniques, 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2023, pp. 284-290.

[6] S.W. Khan, Q. Hafeez, M.I. Khalid, R. Alroobaea, S. Hussain, J. Iqbal, J. Almotiri, S. S. Ullah, Anomaly detection in traffic surveillance videos using deep learning, Sensors 22 (17) (2022) 1–18.

[7] S. Hamdi, S. Bouindour, K. Loukil, H. Snoussi, and M. Abid, Hybrid deep learning and HOF for anomaly detection, in Proc. Int. Conf. Control, Decis. Inf. Technol., Paris, France, Apr. 2019, pp. 1-6.

[8] J. Hu, E. Zhu, S. Wang, X. Liu, X. Guo, J. Yin, An efficient and robust unsupervised anomaly detection method using ensemble random projection in surveillance videos, Sensors 19 (19) (2019) 1–18.

[9] M.B. Shaikh, D. Chai, S.M.S. Islam, and N. Akhtar, MAiVAR: Multimodal audio-image and video action recognizer, 2022 IEEE International Conference on Visual Communications and Image Processing (VCIP). 2022, pp. 1-5.

[10] J.D.S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A.L. Koerich, Multimodal fusion with deep neural networks for audio-video emotion recognition, arXiv preprint arXiv:1907.03196, 2019.

[11] B. Yang, Q. Zhang, and Z. Liu, ICANet: A method of short video emotion recognition driven by multimodal data, 2022 2nd International Conference on Networking Systems of AI (INSAI). 2022, pp. 22-25.

[12] J. He, Y. Ren, L. Zhai, and W. Liu, FCC-MF: Detecting violence in audio-visual context with frame-wise cluster contrast and modality-stage flooding, ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Apr. 2024, pp. 8346-8350.

[13] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, Not only look, but also listen: Learning multimodal violence detection under weak supervision, European conference on computer vision. Glasgow, U.K., 2020, pp. 1-17.

[14] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, Video Swin Transformer, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3202–3211, 2022.

[15] A. Radford *et al.*, Learning Transferable Visual Models From Natural Language Supervision, Proc. Int. Conf. Machine Learning (ICML), 2021.

[16] C. Jia *et al.*, ALIGN: Scaling Up Visual and Vision-Language Representation Learning, Proc. Int. Conf. Machine Learning (ICML), 2021.

[17] J. Lei *et al.*, TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval, Proc. European Conf. Computer Vision (ECCV), 2020.

[18] H. Xu *et al.*, Video–Language Pre-training with Contrastive Learning, Advances in Neural Information Processing Systems (NeurIPS), 2021.

[19] Z. Wang, L. Zhao, J. Zhang, H. Song, J. Wang, H. Duan, Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model, Int. J. Comput. Vis. 133 (2025) 4646–4668.

[20] Z. Wang, J. Wang, H. Song, P. Wang, K. Lyu, W. Li, L. Zhao, SD-fuse: an image structure-driven model for multi-focus image fusion (Art. no), Inf. Fusion 103 (2025) 104058.

[21] L. Zhao, Q. Liu, S. Li, Infrared and visible image fusion via iterative feature decomposition and deep balanced fusion (Art. no), Pattern Recognit. 151 (2025) 113022.

[22] Z. Wang, S. Yu, H. Duan, S. Wang, Y. Long, L. Shao, Rethinking multi-focus image fusion: an input space optimization view, IEEE Trans. Image Process. 35 (2026) 1321–1336.

[23] H. Duan, Y. Long, S. Wang, H. Zhang, C.G. Willcocks, L. Shao, Dynamic unary convolution in transformers, IEEE Trans. Pattern Anal. Mach. Intell. 45 (11) (2023) 12747–12759.

[24] Z. Wang, J. Zhang, H. Song, M. Ge, J. Wang, and H. Duan, Highlight what you want: Weakly-supervised instance-level controllable infrared-visible image fusion, in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2025, pp. 12637–12647.

[25] W. Tong, X. Guan, M. Zhang, P. Li, J. Ma, E.Q. Wu, L.M. Zhu, Edge-assisted epipolar transformer for industrial scene reconstruction, IEEE Trans. Autom. Sci. Eng. 22 (2025) 701–711, https://doi.org/10.1109/TASE.2023.3330704.

[26] W. Tong, M. Zhang, G. Zhu, X. Xu, E.Q. Wu, Robust depth estimation based on parallax attention for aerial scene perception, IEEE Trans. Ind. Inform. 20 (9) (2024) 10761–10769, https://doi.org/10.1109/TII.2024.3392270.

[27] K.W. Tong, Y. Cai, Y.-W. Jie, Y. Duan, Y. Hou, E.Q. Wu, Neural rendering and flow-assisted unsupervised multi-view stereo for real-time monocular tracking and scene perception (early access), IEEE Trans. Autom. Sci. Eng. (2025), https://doi.org/10.1109/TASE.2025.3546713.

[28] K.W. Tong, Z. Shi, G. Zhu, Y. Duan, Y. Hou, E.Q. Wu, L. Zhu, Large-scale aerial scene perception based on self-supervised multi-view stereo via cycled generative adversarial network (Art. no), Inf. Fusion 109 (2024) 102399, https://doi.org/10.1016/j.inffus.2024.102399.

[29] W. Tong, A. Gu, X. Wu, X. Deng, Y. Cai, Y. Duan, Y. Hou, Semi-supervised image domain adaption for aerial refueling drogue detection on embedded chip under foggy conditions, IEEE Trans. Autom. Sci. Eng. 22 (2025) 4748–4759, https://doi.org/10.1109/TASE.2024.3448255.

[30] J. Li, D. Li, C. Xiong, and S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in Proceedings of the 39th International Conference on Machine Learning (ICML), 2022, pp. 12888-12900.

[31] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770-778.

[32] J. Carreira and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4724-4733.

[33] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, Not only look, but also listen: Learning multimodal violence detection under weak supervision, Eur. Conf. Comput. Vis., Glasgow, U.K., 2020, pp. 1-17.

[34] A. Radford et al., Learning transferable visual models from natural language supervision, Proc. ICML, 2021, pp. 8748-8763.

[35] J. Carreira and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Proc. IEEE CVPR, 2017, pp. 4724-4733.

[36] P. Peng, et al., Learning weakly supervised audio-visual violence detection in hyperbolic space, IEEE Trans. Multimed. 25 (6) (2024) 2331–2345.

[37] A. Ahmad, et al., Gated fusion networks for multi-modal violence detection, Sensors 25 (10) (2025) 6272.

[38] S. Shang, et al., Multimodal violent video recognition based on mutual distillation, Pattern Recognit. Lett. 181 (2024) 36–45.

Chih-Yung Chang received the Ph.D. degree in computer science and information engineering from National Central University, Zhongli, Taiwan, in 1995. He is currently a Full Professor with the Department of Computer Science and Information Engineering, Tamkang University, Taipei, Taiwan. His current research interests include the Internet of Things, wireless sensor networks, ad hoc wireless networks, and long-term evolution broadband technologies. He served as an Associate Guest Editor for several SCI-indexed journals, including the International Journal of Ad Hoc and Ubiquitous Computing from 2011 to 2014, the International Journal of Distributed Sensor Networks from 2012 to 2014, IET Communications in 2011, Telecommunication Systems in 2010, the Journal of Information Science and Engineering in 2008, and the Journal of Internet Technology from 2004 to 2008.

Syu-Jhih Jhang received the M.S. degree in Department of Information Management from National Taipei University of Nursing and Health Sciences, Taiwan, in 2021. She is currently pursuing a Ph.D. degree in Computer Science and Information Engineering at Tamkang University, New Taipei, Taiwan. Her current research interests are Internet of Things, Wireless Sensor Networks and Artificial Intelligence.

Yu-Ting Chin received the B.S. and M.S. degrees in computer science and information engineering from Aletheia University, New Taipei City, Taiwan, in 2007 and 2010, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include Internet of Things, wireless sensor networks, ad hoc wireless networks, and software-defined networking and low-power wide-area network technologies.

I-Hsiung Chang received the Ph.D. degree in education from National Chengchi University, Taipei, Taiwan, and the Ph.D. degree in computer science and information engineering from Tamkang University, New Taipei City, Taiwan. He is currently a Distinguished Professor and an accredited reviewer for early childhood education and care professional programs. His research interests include early childhood institution management, smart childcare systems, and nonprofit organization management.

Diptendu Sinha Roy received the Ph.D. degree in engineering from the Birla Institute of Technology, Mesra, India, in 2010. In 2016, he joined the Department of Computer Science and Engineering, National Institute of Technology (NIT) Meghalaya, Shillong, India, as an Associate Professor, where he has been working as the Chair of the Department of Computer Science and Engineering since January 2017. Prior to his stint at NIT Meghalaya, he worked with the Department of Computer Science and Engineering, National Institute of Science and Technology, Berhampur, India. His current research interests include software reliability, distributed and cloud computing, and the Internet of Things (IoT), specifically applications of artificial intelligence/machine learning for smart integrated systems. Dr. Roy is a member of the IEEE Computer Society.