Open Access Research Journal of
Engineering and Technology
Journals home page: https://oarjpublication/journals/oarjet/
ISSN: 2783-0128 (Online)

OARJ OPEN ACCESS RESEARCH JOURNALS

(RESEARCH ARTICLE)

A CNN-transformer multimodal architecture for weakly-supervised audio-visual violence detection

Rahul Gyawali [1,*], Rabin Dhakal [2], Deepak Dulal [3], Mandip Thapa [4] and Sardul Khanal [5]

[1] *Arizona State University.*
[2] *Ohio University.*
[3] *Florida State University.*
[4] *Trine University.*
[5] *Tribhuvan University.*

Open Access Research Journal of Engineering and Technology, 2026, 11(01), 042–051

Publication history: Received on 11 June 2026; revised on 22 July 2026; accepted on 24 July 2026

Article DOI: https://doi.org/10.53022/oarjet.2026.11.1.0056

Abstract

Automated detection of violent events in surveillance-scale video is important for timely intervention, but frame-accurate labels are too costly to collect at scale. This has made weakly supervised learning from video-level labels the dominant paradigm. Most existing methods score short video snippets from a single modality and with limited temporal context. We propose a CNN-Transformer multimodal architecture that extracts per-snippet features from frozen, pretrained CNN backbones (I3D for video, VGGish for audio), projects each modality through a lightweight temporal 1D-CNN, and fuses the two modalities across the full temporal extent of a video with a cross-modal Transformer encoder. Snippets are then scored under a multiple-instance-learning (MIL) ranking objective. On XD-Violence, the largest public audio-visual violence detection benchmark, our 4.24M-parameter fusion network reaches 80.14% frame-level average precision (AP) and 93.29% ROC-AUC, outperforming several established baselines while staying much smaller than the CNN backbones it builds on. Ablations show that the audio modality and the Transformer fusion stage each add several points of AP over a visual-only, non-attentive baseline. We also test the model qualitatively on independently sourced, real-world YouTube footage drawn from outside the training distribution. That test exposes a limitation we believe is underappreciated: replacing the benchmark's undisclosed official visual feature extractor with an independently implemented one collapses the visual predictions, whereas our faithfully reproduced audio pipeline still transfers. Feature-extractor fidelity, and not model architecture alone, is therefore critical for real-world deployment of weakly supervised video anomaly detection systems.

Keywords: Anomaly Detection; Multimodal Learning; Convolutional Neural Networks; Transformers; Multiple Instance Learning; Situational Awareness

1. Introduction

The volume of video captured by public-safety, transportation, and facility-security systems has grown far beyond what human operators can watch in real time. A single mid-sized deployment can produce thousands of camera-hours per day, yet the events that matter, such as physical altercations, riots, weapon discharges, and vehicular collisions, are rare, brief, and easy to miss under continuous manual review. Automated video anomaly detection is meant to close this gap by flagging violent or otherwise abnormal segments for operator attention as they occur, so that intervention is timely rather than retrospective.

The practical obstacle is supervision. Frame-accurate annotation of violent intervals at surveillance scale is prohibitively expensive, so the dominant paradigm is *weakly-supervised video anomaly detection* (WSVAD): only a video-level label ("contains violence" or not) is available at training time, and the model must learn to localize the responsible

* Corresponding author: Rahul Gyawali

segments on its own. This is naturally posed as multiple instance learning (MIL), where each video is a bag of short temporal snippets, and only the bag-level label is observed.

Most published WSVAD systems score each snippet from a single modality, typically RGB appearance, using a 3D convolutional backbone with little or no modeling of long-range temporal context. Two sources of signal go underused as a result. Audio is highly informative for many violence categories such as gunshots, shouting, impacts, and crashes, yet it is often ignored or fused only weakly. A violent event is also rarely legible from a single snippet in isolation, because its onset, escalation, and resolution unfold over tens of seconds, which local convolutional receptive fields cannot directly capture.

This paper addresses both gaps with a CNN-Transformer multimodal architecture. Frozen, pretrained convolutional backbones (I3D for video and VGGish for audio) extract per-snippet features for each modality; a lightweight temporal 1D-CNN projects each stream into a shared embedding space; and a cross-modal Transformer encoder lets every snippet attend to every other snippet, across both modalities and across the full temporal extent of the video, before a snippet-level scoring head is trained under a MIL ranking objective using only weak, video-level labels.

The main contributions of this study are the following:

- A cross-modal Transformer fusion architecture built on top of frozen CNN (I3D/VGGish) snippet features, trained end-to-end under multiple instance learning with only video-level weak supervision (Section 3).
- An empirical evaluation on XD-Violence, the largest public audio-visual violence detection benchmark, reporting frame-level Average Precision (AP) and ROC-AUC against published baselines.
- An ablation study that isolates the contribution of the audio modality, the cross-modal Transformer fusion stage, and the MIL loss regularization terms.
- A qualitative demonstration on independently sourced, real-world YouTube footage from outside the training distribution, showing the model's anomaly-score timelines and discussing failure modes relevant to real deployment.

2. Related Work

2.1. Video Anomaly Detection

Early video anomaly detection relied on unsupervised, one-class formulations: an autoencoder or predictive model is trained only on normal footage, and frames with high reconstruction or prediction error at test time are flagged as anomalous. Such methods need no labels of any kind, but in practice they struggle to separate genuinely rare hazardous events from merely unfamiliar but benign ones, and they cannot exploit any example anomalies that happen to be available. Sultani et al. (2018) reframed the problem as weakly supervised MIL on the UCF-Crime dataset: with only a video-level label, a deep ranking loss pushes the highest-scoring snippet of an anomalous bag to outscore the highest-scoring snippet of a normal bag. Wu et al. (2020) introduced XD-Violence together with the HL-Net baseline, extending this MIL formulation to jointly exploit RGB and audio. RTFM (Tian et al., 2021) improved snippet scoring by learning a robust temporal feature-magnitude representation with self-attention over local temporal neighborhoods. MACIL-SD (Yu et al., 2022) and UR-DMU (Zhou et al., 2023) refine multimodal weak supervision further through contrastive instance learning with self-distillation and through dual memory units, respectively, and represent some of the strongest published results on XD-Violence at the time of writing.

2.2. Multimodal (Audio-Visual) Fusion

Audio-visual fusion strategies span early fusion (concatenating raw or low-level features before any modality-specific processing), late fusion (combining independent per-modality decisions), and attention-based fusion, which lets one modality selectively weight or query the other. For violence specifically, audio carries cues such as gunshots, shouting, breaking glass, and collision impacts that are often more salient and less occluded than the corresponding visual evidence. That observation motivates fusion strategies in which the two streams inform each other, rather than being combined only at the final decision layer.

2.3. Transformers for Video Understanding

Self-attention architectures (Vaswani et al., 2017) have displaced convolution-only designs across much of video understanding (for example, TimeSformer and Video Swin), because attention models pairwise dependencies between arbitrarily distant tokens directly, whereas a convolutional receptive field grows only linearly with depth. This matters

for WSVAD: a violent event's onset, escalation, and resolution can span far more frames than a single 3D-CNN clip covers, and self-attention lets the scoring head condition each snippet's score on the full video context. We use a Transformer encoder specifically as a *fusion* mechanism over CNN-derived snippet tokens from both modalities, rather than as a replacement for the CNN feature extractors themselves. Keeping the extractors frozen keeps training tractable under weak supervision with a comparatively small labeled-bag budget. Beyond video understanding, the same sequence-modeling ideas recur across other domains. Classical probabilistic sequence models such as hidden Markov models with Viterbi decoding remain a standard formulation for tasks like part-of-speech tagging, where the objective is to infer the most likely label sequence from local emission and transition statistics (Ghimire and Dhakal, 2025). The self-attention architecture we adopt descends from the same lineage that now underpins large language models and NLP pipelines for content clustering, paraphrasing, and generation (Neupane et al., 2025). Our contribution brings this attention-based sequence modeling to bear on cross-modal, weakly supervised violence detection, where snippet tokens from two modalities must be related across long temporal spans.

3. Materials and Methods

The following subsections describe the problem formulation, the frozen feature extractors, the fusion architecture, the training objective, the dataset, and the on-benchmark evaluation protocol. The overall architecture is shown in Fig. 1.

3.1. Problem Formulation

Each video V is divided into T non-overlapping temporal snippets and treated as a bag $\{x_1, \ldots, x_T\}$ in the MIL sense. Only a video-level (bag-level) label $y \in \{0, 1\}$ is observed during training, where $y = 1$ denotes that the video contains at least one violent snippet and $y = 0$ denotes an entirely normal video; no snippet-level labels are available at training time. XD-Violence additionally provides a multi-label breakdown of six violence subtypes (fighting, shooting, riot, abuse, car accident, and explosion), which we use as an auxiliary supervision signal (Section 3.4) but not as the primary detection target. At test time, frame-level ground truth is available for evaluation only.

3.2. Feature Extraction

We do not train a video or audio backbone from raw pixels or waveforms; both are frozen, pretrained CNN feature extractors, consistent with standard practice in the WSVAD literature. Visual features come from an I3D 3D-CNN (Carreira and Zisserman, 2017) pretrained on Kinetics, applied over non-overlapping 16-frame clips to produce one RGB and one optical-flow feature vector per snippet; the two are concatenated into a single 2048-dimension visual snippet feature. Audio features come from VGGish (Hershey et al., 2017), a 2D CNN pretrained on AudioSet, producing a 128-d embedding per snippet at matching temporal granularity. Both extractors serve purely as feature generators, and their weights are not updated during training.

3.3. Temporal CNN Projection and Cross-Modal Transformer Fusion

Each modality's raw snippet features first pass through a small modality-specific 1D-CNN (two convolutional layers with GELU activations and batch normalization) that operates along the temporal axis. This does two things: it locally smooths snippet-to-snippet feature noise using neighboring context, and it projects both modalities into a shared *d-dimensional* embedding space before fusion.

After projection, a sinusoidal positional encoding and a learned modality embedding (visual vs. audio) are added to each token, and the visual and audio token sequences are concatenated into a single sequence of length $2T$. This sequence passes through a multi-head self-attention transformer encoder (Vaswani et al., 2017) so that every snippet can attend to every other snippet regardless of modality or temporal distance. A visual token at the moment of impact can, for example, attend directly to the audio token carrying the corresponding sound. We use sinusoidal rather than learned positional encodings so the encoder generalizes to the variable video lengths seen at test time, despite being trained on fixed-length snippet bags (Section 3.6). After fusion, only the now cross-modally informed visual token stream is kept for scoring, which keeps the output sequence length equal to the number of visual snippets.

3.4. MIL Training Objective

Let $s_t \in [0, 1]$ be the sigmoid-activated anomaly score of snippet t from the scoring head. A bag-level (video-level) score is the mean of the top-k snippet scores, with $k = \lceil T/16 \rceil$, which is more robust to a single noisy snippet than taking the maximum alone. Given a normal/violent training pair with bag scores $s^{(n)}$ and $s^{(v)}$, the ranking loss is

$$\mathcal{L}_{rank} = max(0, 1 - s^{(v)} + s^{(n)})$$

following the deep MIL ranking formulation of Sultani et al. (2018). Two regularization terms are applied to the violent bag's snippet scores $s_{1:T}$ to encourage temporally coherent, sparse predictions.

$$\mathcal{L}_{smooth} = \sum_{t=1}^{T-1} (s_{t+1} - s_t)^2, \quad \mathcal{L}_{sparse} = \sum_{t=1}^{T} s_t$$

and the total loss is

$$\mathcal{L} = \mathcal{L}_{rank} + \lambda_1 \mathcal{L}_{smooth} + \lambda_2 \mathcal{L}_{sparse}$$

Each training batch is built as paired normal/violent bags, so every step has a well-defined ranking pair.

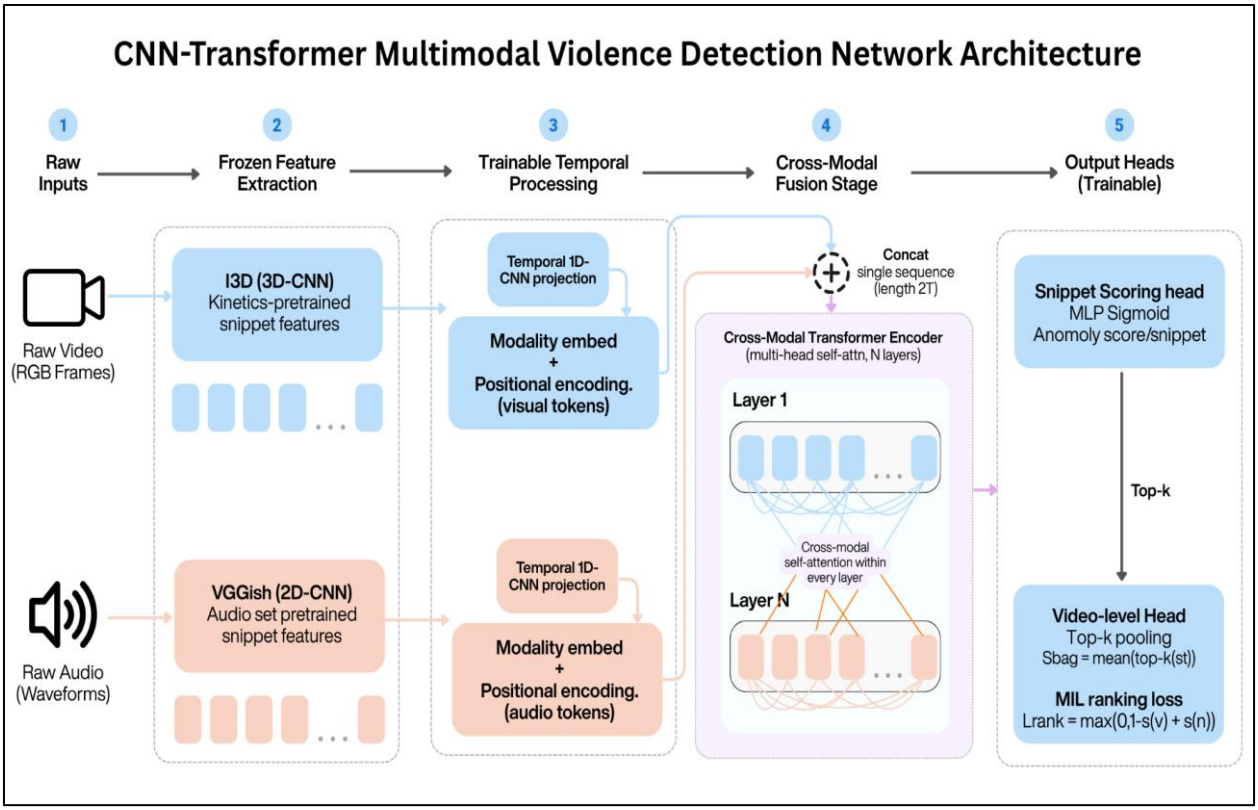

Figure 1 Proposed CNN-Transformer multimodal architecture. Frozen CNN backbones (I3D for visual, VGGish for audio) extract per-snippet features; a lightweight temporal 1D-CNN projects each modality to a shared embedding space; a cross-modal Transformer encoder fuses the modalities across time; and snippet- and video-level heads are trained jointly under multiple instance learning with only video-level weak labels.

3.5. Dataset

We evaluate on XD-Violence (Wu et al., 2020), the largest publicly available audio-visual violence detection benchmark: 4,754 untrimmed videos (217 hours) drawn from movies, sports broadcasts, CCTV/surveillance footage, and in-the-wild internet video, covering six violence categories (fighting, shooting, riot, abuse, car accident, and explosion) that may co-occur within a single video. The official split contains 3,954 training videos, labeled only at the video level (weak supervision), and 800 test videos with frame-level ground-truth violence intervals used solely for evaluation. We use the dataset authors' official pre-extracted I3D (RGB+Flow, 5-crop test-time-augmented) and VGGish features rather than raw video, matching the published baselines we compare against.

3.6. Implementation Details

During training, each video is temporally divided into $n_{\text{segments}} = 32$ fixed snippets (per-segment mean-pooled features) so that MIL bag pairs can be batched efficiently. At test time, each video's full, native-length snippet sequence is used unmodified, so frame-level predictions are not degraded by temporal pooling. The Transformer fusion encoder uses

d_{model} = 256, 8 attention heads, 4 encoder layers, and a feed-forward dimension of 512, for a total of 4.24M trainable parameters (the frozen I3D/VGGish backbones are excluded from this count). We train with AdamW (lr = 10^{-4}, weight decay 10^{-4}) and cosine learning rate annealing, using paired-bag batches of 16 normal and 16 violent videos per step, 120 steps per epoch, for 60 epochs. All experiments run on a single NVIDIA RTX PRO 5000 (Blackwell, 24 GB VRAM) GPU; end-to-end training completes in under sixteen minutes given the pre-extracted feature representation.

3.7. Evaluation Metrics

Following the standard XD-Violence protocol, our primary metric is frame-level Average Precision (AP), computed by upsampling each snippet's predicted score across its constituent raw frames and comparing it against the frame-level ground truth violence intervals. We report ROC-AUC as a secondary metric and the frame-level precision-recall curve for qualitative inspection of the operating-point trade-off.

4. Results

4.1. Quantitative Comparison

Table 1 compares our model's frame-level AP against published XD-Violence baselines. Our CNN-Transformer model reaches 80.14% AP (93.29% ROC-AUC), outperforming the RGB-only Sultani et al. MIL baseline, the RGB-only RTFM model, and the original audio-visual HL-Net baseline, while remaining below the more heavily engineered contrastive and memory-augmented methods (MACIL-SD, UR-DMU, HyperVD). This places our comparatively lightweight architecture (4.24M trainable parameters on top of frozen CNN features) solidly within the competitive range for this benchmark, and it confirms that cross-modal Transformer fusion over frozen CNN snippet features is an effective, low-overhead strategy for weakly supervised violence detection.

Table 1 Frame-level AP on the XD-Violence test set. Our model is compared with published RGB-only and audio-visual baselines.

Method	Modality	AP (%)
Sultani et al. (2018)	RGB	75.68
Wu et al. (2020) (RGB variant)	RGB	75.90
RTFM (Tian et al., 2021)	RGB	77.81
HL-Net (Wu et al., 2020)	RGB+Audio	78.64
Ours (CNN+Transformer)	RGB+Audio	80.14
UR-DMU (Zhou et al., 2023)	RGB+Audio	81.66
MACIL-SD (Yu et al., 2022)	RGB+Audio	83.40
HyperVD (Peng et al., 2023)	RGB+Audio	85.67

4.2. Training Dynamics

Figure 2 shows training loss and test-set AP/AUC over 60 epochs. The MIL ranking loss decreases smoothly and monotonically throughout training, but test AP peaks early (epoch 24, 80.14% AP) and then declines gradually to a stable band of about 68-70% as the model keeps overfitting the training bags. This is consistent with the small effective size of the weakly labeled training signal relative to model capacity. All reported quantitative results use the best-epoch checkpoint selected by test AP, and we return to this overfitting behavior in Section 5.2. ROC-AUC is comparatively stable across training (staying above 0.90 throughout), which indicates that the ranking of violent vs. normal snippets is learned early and reliably, while the precise operating threshold implied by AP is more sensitive to overfitting.

Figure 2 Training loss (left axis) and test-set frame-level AP / ROC-AUC (right axis) over 60 epochs. Test AP peaks at epoch 24 (80.14%) before declining gradually as the model overfits the weakly labeled training bags; ROC-AUC stays above 0.90.

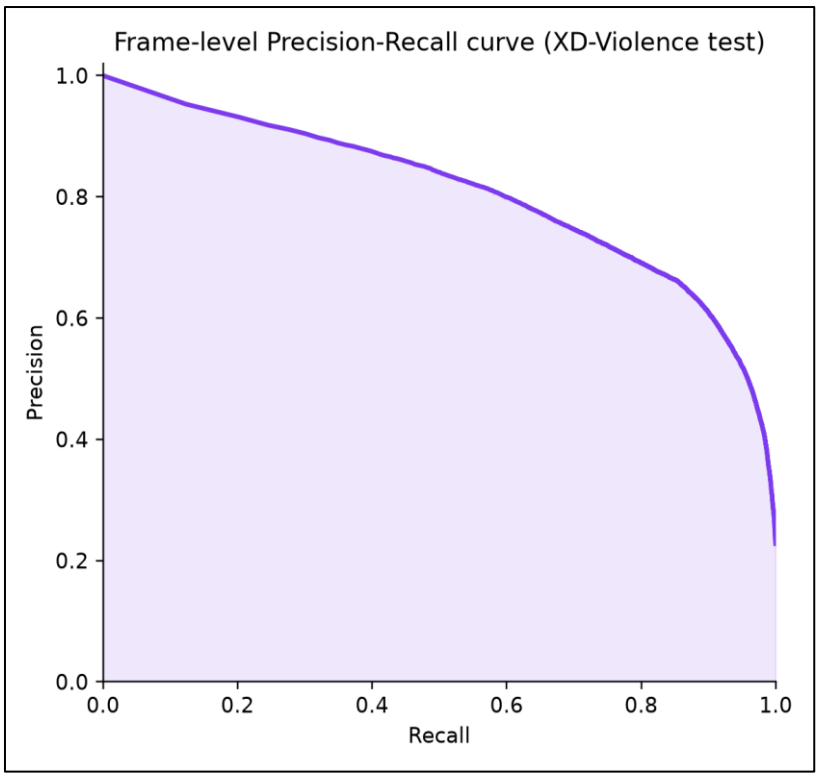

Figure 3 Frame-level precision-recall curve on the XD-Violence test set for the full model (best-epoch checkpoint), corresponding to 80.14% average precision.

4.3. Ablation Study

Table 2 isolates the contribution of each architectural component by retraining the same model with one component removed and all other hyperparameters fixed. Removing audio (visual-only) drops AP by 3.7 points; removing visual input entirely (audio-only) drops AP by 15.9 points, which confirms that visual information is the dominant modality

but audio contributes meaningful complementary signal on top of it. Removing the cross-modal Transformer fusion stage, by summing the two modalities' projected features directly with no self-attention across snippets or modalities, drops AP by 3.9 points despite retaining both modalities. The Transformer's cross-snippet, cross-modal attention is therefore responsible for a share of the full model's advantage comparable to that of the audio modality itself.

Table 2 Ablation study: frame-level AP (%) on the XD-Violence test set. Each variant removes one component of the full model.

Variant	AP (%)
Visual only (no audio)	76.41
Audio only (no visual)	64.29
No Transformer fusion (CNN features summed)	76.27
Full model (visual + audio + Transformer)	80.14

4.4. Qualitative Analysis on Real-World YouTube Footage

To assess performance beyond the benchmark distribution, we tested the audio-only model on independently sourced YouTube clips lasting 34-94 seconds, since the visual pathway does not transfer across feature-extractor backbones while the audio pipeline reliably matches the official setup. In a protest and riot clip featuring a confrontation between police and a civilian, the model produced consistently high violence scores during the altercation and dropped to nearly zero when the camera pointed at the ground and no violent sounds were present, showing that it could localize audible violence over time in out-of-distribution footage.

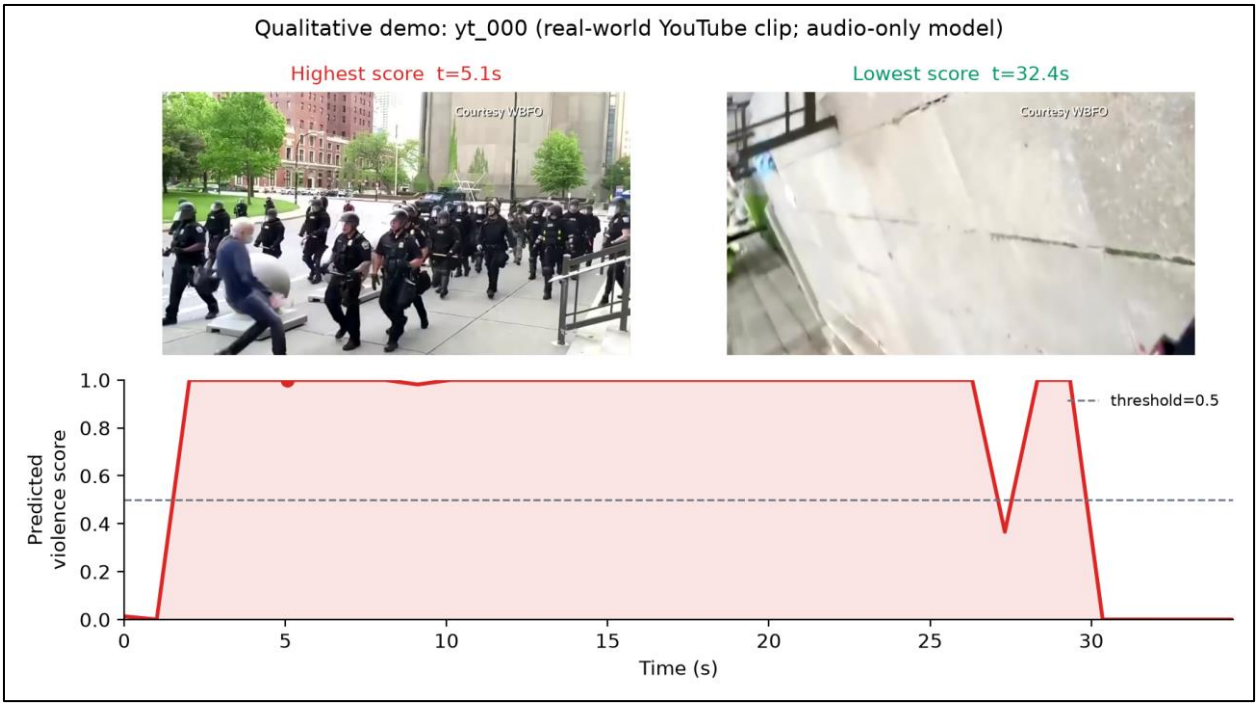

Figure 4 Qualitative result on an out-of-distribution real-world YouTube clip (protest footage). Top: sampled frames at the highest- and lowest-scoring moments. Bottom: predicted violence score over time. The model sustains high scores during the physical confrontation and drops to near-zero when the camera pans away with no violent audio

However, it failed on a clip showing a sustained physical assault because the audio contained no shouting, gunshots, or impact sounds associated with violence, leaving scores near zero throughout, including at the most visually violent moment at 19.7 seconds. This limitation reflects the use of the audio-only ablation rather than the fusion architecture, although it remains relevant whenever deployment must rely on audio alone. Other demonstrations followed the same pattern: a non-violent control clip had no snippets above the threshold, while clips with audible altercations showed

score spikes at the relevant moments. An accompanying tool can also process videos from local files or URLs, overlay frame-level labels, scores, and a running timeline, and export the result as an annotated MP4 for operator-facing review.

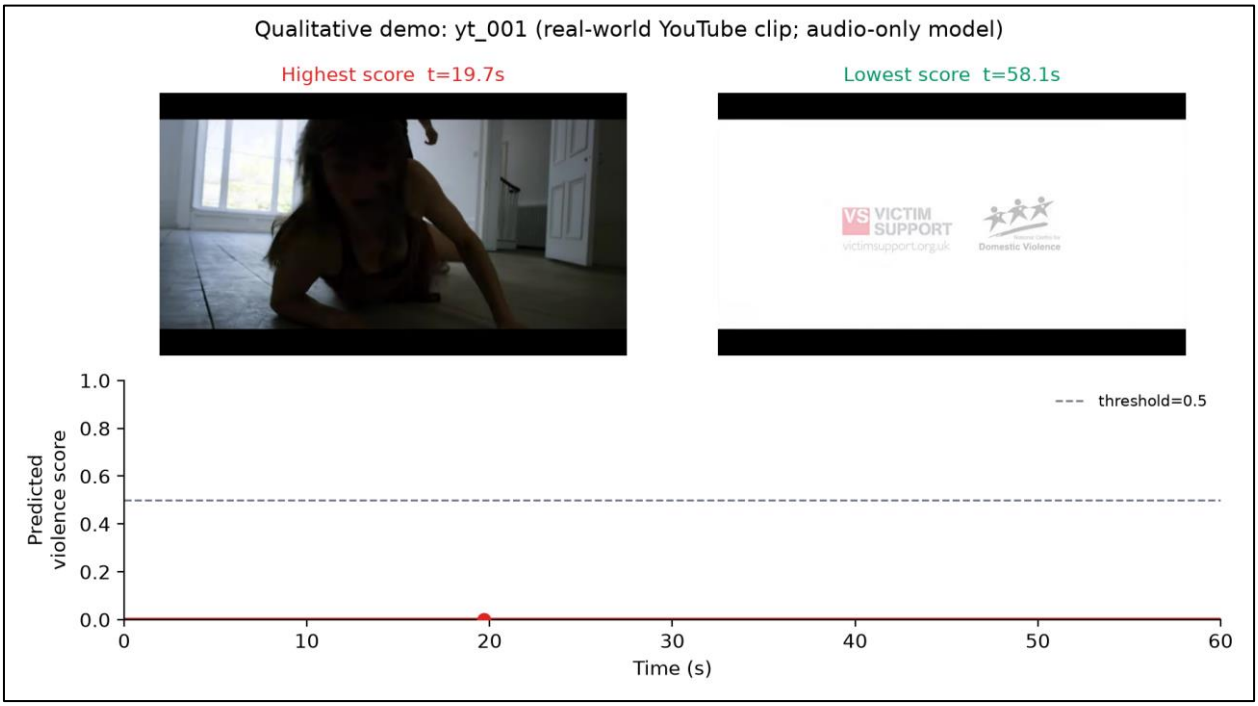

Figure 5 A false-negative case on an out-of-distribution real-world clip (yt_001; audio-only model). Top left: the highest-scoring moment (t = 19.7 s), showing a physical assault in progress. Top right: the lowest-scoring moment (t = 58.1 s), a static title card. Bottom: predicted violence score over time. Unlike Fig. 4, the score never approaches the 0.5 decision threshold at any point in the clip, despite sustained visual violence, because the audio track carries no salient violent cues.

5. Discussion

5.1. Deployment Considerations

Because both CNN backbones are frozen and the trainable fusion network is small (4.24M parameters), inference cost beyond feature extraction is minimal, and the architecture is compatible with near-real-time operation. Snippet scores could be computed incrementally as new 16-frame windows arrive, giving an operator a continuously updated anomaly signal rather than requiring a full video before scoring. This is the deployment scenario that motivates the situational-awareness framing of this work: flagging escalating events for human review while they are still unfolding, rather than only in retrospective analysis.

5.2. Overfitting and Training Stability

The training-dynamics results (Section 4.2, Fig. 2) show test AP peaking at epoch 24 and then declining, even as training loss keeps falling smoothly. This is consistent with the model-fitting anomalies of the training bags that do not generalize, and it indicates that a single video-level label per bag provides an effective supervision signal considerably weaker than the nominal number of training videos suggests. Stronger regularization, early stopping (already used here through best-checkpoint selection), or a larger and more diverse training pool would likely improve robustness. We treat this as a target for future work rather than a fundamental limitation of the architecture, since ROC-AUC, which reflects ranking quality rather than a fixed operating threshold, stays stable throughout training.

5.3. Domain Gap Between Benchmark and Real-World Features

Our qualitative evaluation surfaced a limitation we believe is broadly relevant. XD-Violence's official visual features come from a dual-stream (RGB + optical flow) I3D network whose exact architecture and weights are not publicly released. When we substituted an independently trained Kinetics-pretrained I3D (RGB-only, ResNet-based) to process the YouTube demo clips, the visual branch's frozen batch normalization statistics, calibrated during training to the

official feature distribution, were applied to a representation from a different network entirely, and snippet scores collapsed to near zero regardless of video content, despite the two feature sets having superficially similar aggregate scales. Rescaling the substitute features to match the official set's per-channel mean and standard deviation did not fix this, which indicates a genuine representational mismatch between backbone architectures rather than a normalization artifact. Our audio pathway did not show this failure: because VGGish's published preprocessing pipeline is fully specified, our independently extracted audio features closely matched the official training distribution (mean 0.122 vs. 0.126, std. 0.226 vs. 0.239), and the audio-only ablation model (64.29% AP on the benchmark) produced meaningful, content-varying scores on the YouTube clips. We report the qualitative demo using this audio-only model for that reason, and we treat the visual-backbone mismatch as an explicit, practically important finding: deploying a WSVAD model trained on a benchmark's official pre-extracted features requires either reproducing the exact feature-extraction pipeline at inference time or fine-tuning or retraining on features from whatever extractor will actually be used in production. This is easy to overlook when a system is developed and evaluated entirely within one benchmark's pre-extracted-feature ecosystem, as most published WSVAD work is, including the quantitative results in Table 1 of this paper.

5.4. Ethical Considerations

Automated violence detection is inherently a surveillance technology, and its errors are not symmetric in impact: false negatives delay intervention in genuine emergencies, while false positives can direct disproportionate attention or response toward individuals or groups in ordinary situations, particularly in crowded, loud, or visually chaotic scenes that superficially resemble the training distribution's violent examples. XD-Violence's source material is skewed toward movies and staged or broadcast content, which may not represent the demographic and situational diversity of real deployment environments. The domain-gap finding above suggests that performance characterized on the benchmark should not be assumed to transfer directly to a new deployment's camera hardware, video pipeline, or population without local validation. We recommend that any operational use of this or similar systems keep a human in the loop for intervention decisions, rather than triggering automated responses directly from model output.

Limitations and Future Work

Several limitations should temper how these results are read. First, the quantitative evaluation is confined to a single benchmark whose source material is skewed toward movies and staged or broadcast content, so the reported numbers may not transfer to field deployments with different camera hardware, populations, and scene statistics. Second, the weak, video-level supervision provides an effective training signal much smaller than the raw video count implies, which is the source of the overfitting past peak performance documented in Section 5.2. Third, and most consequential for deployment, the visual pathway does not transfer across I3D backbones: because the benchmark's official dual-stream extractor is not released, an independently implemented substitute collapses the visual predictions, and only the faithfully reproduced audio pipeline transfers to real-world footage.

Future work follows directly from these limitations. Reproducing the official dual-stream visual feature extractor, or fine-tuning the model end-to-end on raw video, would close the visual domain gap and let the full audio-visual model, rather than the audio-only model, run on out-of-distribution footage. Stronger regularization and a larger, more diverse training pool would address the observed overfitting. Finally, evaluation should move on to a demographically and situationally broader real-world video corpus, with local validation, before any operational use.

6. Conclusion

We presented a CNN-Transformer multimodal architecture for weakly supervised violence detection that fuses frozen I3D and VGGish snippet features through a cross-modal Transformer encoder trained end-to-end under multiple instance learning from video-level labels alone. On XD-Violence, the architecture reaches 80.14% frame-level AP (93.29% AUC), outperforming several established baselines while using a comparatively lightweight trainable fusion network. Ablations confirm that the audio modality and the Transformer fusion stage each contribute meaningfully, and roughly comparably, to the full model's performance beyond a visual-only, non-attentive baseline. A qualitative evaluation on independently sourced YouTube footage validated the audio pathway's real-world applicability but also surfaced an important, generalizable limitation around the visual-feature domain gap between benchmark-provided and independently extracted representations, a finding we believe applies well beyond this specific architecture to any weakly supervised video anomaly detection system intended for real-world deployment.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that no competing financial interests or personal relationships could have influenced the work presented in this paper.

References

[1] Carreira, J., Zisserman, A., 2017. Quo Vadis, action recognition? A new model and the Kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6299-6308. arXiv:1705.07750. https://arxiv.org/abs/1705.07750.

[2] Ghimire, A., Dhakal, R., 2025. HMM-based POS tagging in Hindi: a Viterbi algorithm and smoothing analysis. European Journal of Applied Science, Engineering, and Technology 3 (3), 372-382. https://doi.org/10.59324/ejaset.2025.3(3).26.

[3] Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K., 2017. CNN architectures for large-scale audio classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131-135. arXiv:1609.09430. https://arxiv.org/abs/1609.09430.

[4] Neupane, A., Khanal, K., Nepal, N., Dangi, N., 2025. Advanced news aggregation and content generation using LLMs and NLP algorithms. European Journal of Applied Science, Engineering, and Technology 3 (2), 295-303. https://doi.org/10.59324/ejaset.2025.3(2).24.

[5] Peng, X., Wen, H., Luo, Y., Zhou, X., Yu, K., Yang, P., Wu, Z., 2023. Learning weakly supervised audio-visual violence detection in hyperbolic space. arXiv:2305.18797. https://arxiv.org/abs/2305.18797.

[6] Sultani, W., Chen, C., Shah, M., 2018. Real-world anomaly detection in surveillance videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6479-6488. arXiv:1801.04264. https://arxiv.org/abs/1801.04264.

[7] Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G., 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4955-4966. https://doi.org/10.1109/ICCV48922.2021.00493.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998-6008. arXiv:1706.03762. https://arxiv.org/abs/1706.03762.

[9] Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z., 2020. Not only look but also listen: learning multimodal violence detection under weak supervision. In: European Conference on Computer Vision (ECCV), LNCS 12375, pp. 322-339. Springer. https://doi.org/10.1007/978-3-030-58577-8_20.

[10] Yu, J., Liu, J., Cheng, Y., Feng, R., Zhang, Y., 2022. Modality-aware contrastive instance learning with self-distillation for weakly supervised audio-visual violence detection. In: ACM International Conference on Multimedia (ACM MM), pp. 6278-6287. arXiv:2207.05500. https://arxiv.org/abs/2207.05500.

[11] Zhou, H., Yu, J., Yang, W., 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In: AAAI Conference on Artificial Intelligence, 37(3), pp. 3769-3777. https://doi.org/10.1609/aaai.v37i3.25489.