Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Improving Temporal Localization in Vision-Language Video Anomaly Detection

Siddharth Shah[1], Narendrasinh Chauhan[2]

[1]Research Scholar, The Charutar Vidya Mandal (CVM) University, Anand, Gujarat, India.
[2]Department of Information Technology, A. D. Patel Institute of Technology, The Charutar Vidya Mandal (CVM) University, Anand, Gujarat, India.

E-mail: [1]siddharth7763@icloud.com, [2]narendrasinh.chauhan@cvmu.edu.in

Orcid ID: [1]0000-0003-4607-4761, [2]0000-0002-2165-9694

Abstract

The recent developments in the vision-language-based model framework for weakly supervised video anomaly detection (WSVAD) have significantly enhanced anomaly detection performance. The dual-branch framework consisting of two branches for performing binary classification and aligning textual descriptors with visual snippets, has been found efficient concerning anomaly detection. Nonetheless, a significant problem of temporal localization still persists. The existing solutions use a fixed value of top-k snippets without consideration for either short or long anomalies. Moreover, prediction inconsistency in terms of temporal localization with the presence of spikes in normal periods and gaps in anomalies is another issue that arises. The current solution is based on the VadCLIP framework and only modifies some specific aspects of it. First, Confidence-Adaptive MIL (CA-MIL) computes a per-video threshold from the score distribution, selecting fewer snippets when confidence lowers and more when an anomalous event has a larger time frame. Second, a temporal smoothness term penalizes abrupt score transitions between adjacent snippets. Third, two parallel scoring heads, one point-wise MLP, and one local-context convolution are fused through learned gating that accounts for disagreement. Lastly, at test time, Score-level Temporal Context Aggregation (STCA) smooths the final predictions using local averaging and global statistics. Cross-modal attention provides a small additional boost to AUC. In UCF-Crime, the average mAP between the 0.1–0.5 IoU thresholds increases from 6.68 to 9.37 (+40.3%), with mAP@0.5. XD-Violence sees an average increase in mAP from 24.70 to 31.63 (+28.1%). Detection performance is preserved (UCF-Crime AUC decreases by 0.10% from 88.02 to 87.92; XD-Violence AP increases by 0.22% from 84.51 to 84.73).

Keywords: Video Anomaly Detection, Weakly Supervised Learning, Vision-Language Models, Temporal Localization.

1. Introduction

Automated surveillance generates vast amounts of video data, and it is not feasible to manually review every video. Video anomaly detection (VAD) is used to flag unusual events such as fights, accidents, thefts, or fires so that human operators can focus attention where it is most needed. Anomalies are rare by definition, and the space of possible anomalies is

DOI: https://doi.org/10.36548/jiip.2026.3.002
Received: 02.04.2026, received in revised form: 03.05.2026, accepted: 24.05.2026, published: 05.06.2026

essentially unbounded. It requires a prohibitive amount of annotation effort to collect labelled examples for all normal and abnormal cases.

Weakly supervised VAD overcomes this bottleneck as it requires only video-level labels instead of frame- level labels, meaning, each video is labelled. Frame-level scores must be inferred at test time. Sultani et al. introduced this in 2018 along with the UCF-Crime benchmark dataset [1], and the field has grown rapidly since then. Most methods rely on Multiple Instance Learning (MIL), treating each video as a bag of snippet-level instances and aggregating their scores to match the video label.

Early MIL approaches used C3D or I3D features and simple classifiers. Graph convolutional networks [2] and attention mechanisms [3] later improved temporal modeling. Methods based on the transformer architecture [4,5] also helped in further raising the AUC values. During all these advancements, the emphasis has been on differentiating between normal and anomalous snippets; little attention was given to temporal localization, which was measured through mAP under various IoU criteria. Previous research has highlighted that MIL-based methods often suffer from temporally inconsistent predictions due to the lack of temporal supervision in those approaches [3,9]. Tian et al. [3] found that snippet scores are highly variant within continuous anomalies. This inconsistency leads to a decline in localization metrics (mAP) while having little impact on detection metrics (AUC).

CLIP uses features of a pre-trained visual encoder trained on millions of images along with texts. The encoder can capture rich semantic correlations, which could be beneficial for surveillance scenarios. VadCLIP [7] demonstrated how to use both the visual and textual aspects of CLIP by training one branch for the classification of snippets and using another branch for the alignment of visual and textual anomaly classes. However, two main issues have been observed with respect to the performance of the VadCLIP algorithm's predictions: 1.) The traditional MIL chooses the same number of snippets independently of the span covered by the anomaly within the video, leading to the learning of irrelevant information or missing out on critical snippets altogether 2.) Due to abnormal peaks in normal snippets and dips between anomalies, false positives may occur.

Based on the above insights, several modifications to the existing approach in VadCLIP [7], focusing on temporal localization, are proposed. These are primarily methodological improvements, including:

Confidence-Adaptive MIL (CA-MIL): Instead of a static selection process, a dynamic threshold is employed, which depends on the scores predicted for each video. Videos with high score predictions choose fewer snippets, whereas videos with low score predictions select more snippets.

1. Temporal smoothness regularization: A simple penalty in $|s_{t+1} - s_t|$ makes it more likely that adjacent snippets will have similar scores. This would avoid jitter, but not hamper the occurrence of natural shifts from normal to abnormal and vice versa.

2. Attention-based dual scoring fusion with uncertainty-aware gating: Two independent scoring heads are used, each point-wise and local temporal approaches, respectively. The fusion method incorporates the differences between these scoring methods and combines them based on uncertainty.

3. Score-level Temporal Context Aggregation (STCA): This is a straightforward post-inference-time step that includes neighborhood-based averaging and the use of statistics from the entire video.

An additional cross-modal attention block is used to improve consistency between vision and language. Analysis of this model shows that this helps coarse-grained detection rather than fine-grained localization. Attention mechanisms across all anomaly types show remarkable similarities, implying that this technique is able to detect anomalies irrespective of the class type, given minimal supervision.

Extensive experiments have been conducted for the UCF-Crime and XD-Violence datasets. The gain in mean average precision (mAP) is seen to be 40.3% and 28.1%, respectively, whereas the detection accuracy is largely maintained (UCF-Crime AUC: -0.10%; XD-Violence AP: +0.22%). It is observed from ablation studies that CA-MIL leads to the greatest gain, followed by temporal smoothness. From these observations, we can infer that localization gains in vision-language VAD are achievable.

2. Related Work

Weakly supervised video anomaly detection using the Multiple Instance Learning approach was suggested by Sultani et al. [1], which allows learning an algorithm to separate anomalous instances from normal ones with video-level labels. At the same time, their work laid the foundation for the UCF-Crime dataset as well as the fixed top k snippet score aggregation method. Wu et al. [8] further explored this technique in a multimodal environment, considering audio information alongside visual features using the XD-Violence dataset.

Graph convolutional network-based models capturing feature similarity and temporal consistency were developed by Zhong et al. [2]. Tian et al. [3] designed a robust temporal feature magnitude learning model that utilizes the fact that anomalous snippets have higher feature magnitudes than normal snippets. Self-training models that allow refining pseudo-labels iteratively and improving anomaly detection were developed by Feng et al. [9]. Recently, Chen et al. [10] designed a magnitude-contrastive learning algorithm with a glance-and-focus module, whereas Zhou et al. [5] proposes the dual-memory unit with an uncertainty regulation technique. Inspired by the glance-and-focus mechanism of [10] and feature magnitudes of [3], Shah et al. [11] proposed their work relying on contrastive-transformer hybrid modeling.

In another dimension of features for VAD, the Contrastive Language-Image Pre-training Model (CLIP) [6] has demonstrated a remarkable capability to learn visual representations that are aligned with textual descriptions. These CLIP features are pre-trained on hundreds of millions of image-text pairs to provide learned visual-semantic correspondences. The success of prompt-based adaptation techniques, presented by Context Optimization (CoOp) [12], has further expanded the applicability of CLIP to specialized domains.

Early work in video anomaly detection CLIP as a feature extractor. Joo et al. [13] introduced CLIP-TSA, which merges CLIP visual features with temporal self-attention to identify dependencies among frames. Lv et al. [14] proposed an unbiased multiple instance learning (UMIL) method to mitigate selection bias in training with CLIP features. These methods only used the visual encoder and did not take advantage of CLIP's ability to understand language.

VadCLIP [7] introduced a dual-branch architecture that jointly leverages visual and textual representations. One branch performs binary classification for coarse-grained detection, while the other aligns visual features with textual class descriptions for fine-grained categorization. This architecture achieved state-of-the-art performance on both the UCF-Crime and XD-Violence benchmarks specifically in the detection and localization of anomalies. Wu et al. [15] extended the framework to open-vocabulary settings, enabling the detection of new categories of anomalies. Yang et al. [16] proposed normality-guided text prompts to improve alignment between textual descriptions and normal frames. More recently VadCLIP++ [17] incorporates dynamic text prompts to capture temporal changes through frame difference features, while WSVAD-CLIP [18] combines axial transformers with graph attention for improved temporal modelling.

Temporal smoothness has also been applied for action recognition and localization [19]. For anomaly detection tasks, previous works have conducted temporal modelling at the feature level instead of making predictions consistent over time. Huang et al. [20] introduced self-guided temporal discriminative transformers to learn long-range dependencies. However, explicit regularization that enforces temporally smooth and consistent anomaly predictions on consecutive frames has not been explored in vision-language VAD. Therefore, a temporal smoothness branch is introduced to penalize discontinuous anomaly scores.

3. Proposed Work

This section describes the proposed modifications to the VadCLIP framework [7] to improve temporal localization. The underlying architecture—including the frozen CLIP encoders, the Local-Global Temporal adapter, and the dual-branch design—is retained from VadCLIP without structural changes. The contributions of this work lie in four additions to the training objectives and inference procedure: (1) Confidence-Adaptive Multiple Instance Learning (CA-MIL) for dynamic snippet selection, (2) temporal smoothness regularization for prediction consistency, (3) Dual-Head Fusion for robust anomaly scoring, and (4) Score-level Temporal Context Aggregation (STCA) for test-time refinement. A cross-modal attention module is also added for vision-language feature alignment. Figure 1 illustrates the overall framework, where the base pipeline is inherited from VadCLIP [7] and the proposed additions are indicated.

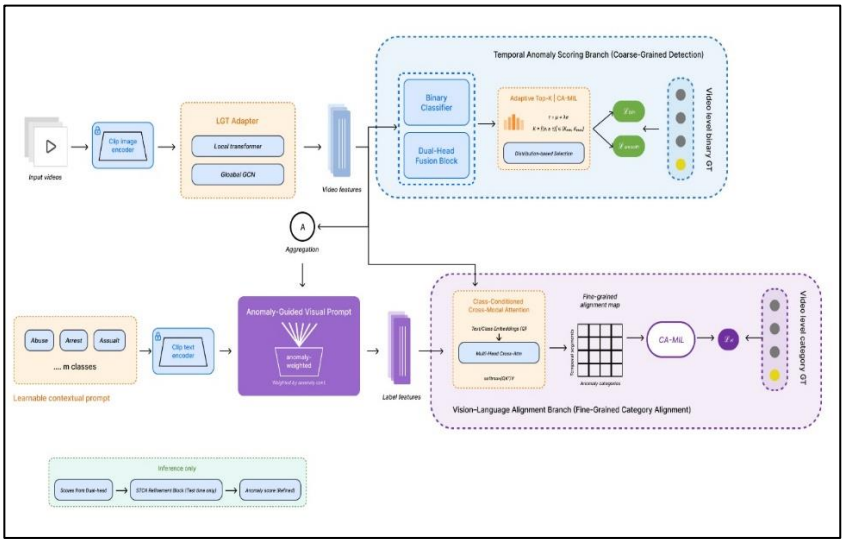

Figure 1. Overall Framework of the Proposed System

Figure 1 illustrates the Temporal-Adaptive Vision–Language Architecture, built upon the VadCLIP [7] with modifications introduced in this work: CA-MIL, Dual-Head Fusion, temporal smoothness regularization, cross-modal attention, and STCA.

3.1 Problem Formulation

Given an untrimmed video $\mathcal{V}$, the goal is to produce snippet-level anomaly scores using only video-level supervision during training. The video is divided into T non-overlapping snippets, and visual features $X = \{x_t\}_{t=1}^{T}$ where $x_t \in \mathbb{R}^d$ are extracted using a pre-trained CLIP visual encoder. During training, only video-level labels $y \in \{0,1\}$ are available, where $y = 1$ indicates the presence of at least one anomalous event. The objective is to learn a scoring function $f: X \rightarrow s$ that produces snippet-level anomaly scores $s = \{s_t\}_{t=1}^{T}$ where $s_t \in [0,1]$.

For vision-language alignment, a set of C class labels $\mathcal{Y} = \{y_0, y_1, \ldots, y_{C-1}\}$ is defined, where y_0 represents "normal" and $y_1, \ldots, y_{C-1}$ represent different anomaly categories. Text embeddings $E = \{e_c\}_{c=0}^{C-1}$ are obtained from the CLIP text encoder using learnable prompt templates.

3.2 Base Architecture

The proposed enhancements are built upon the VadCLIP framework [7], which employs a dual-branch architecture for coarse-grained classification and fine-grained vision-language alignment. The base architecture consists of three stages:

Feature Extraction: Visual features are extracted using a frozen CLIP ViT-B/16 encoder, producing $x_t \in \mathbb{R}^{512}$ for each snippet. Text features $e_c \in \mathbb{R}^{512}$ are obtained by encoding class-specific prompts through the CLIP text encoder with learnable context tokens, following the approach in [7, 12].

Temporal Adaptation: Following VadCLIP, the visual features are processed through a Local-Global Temporal (LGT) adapter that combines local temporal modeling via windowed transformer layers with global structure learning via graph convolution:

$$H = \mathrm{GCN}\big(\mathrm{LocalTransformer}(X + P)\big) \tag{1}$$

where P denotes learnable positional embeddings and $H \in \mathbb{R}^{T \times d}$ represents temporally-adapted features. The GCN component models both feature similarity and temporal distance relationships as described in [7, 2].

Dual-Branch Scoring: The C-branch performs binary classification while the A-branch computes vision-language alignment scores. The proposed enhancements primarily target the training objectives and scoring mechanisms within this dual-branch structure.

The three stages above are reproduced from VadCLIP and are summarized here for completeness; the reader is referred to [7] for full details. The remainder of this section describes the components introduced in the present work.

3.3 Confidence-Adaptive Multiple Instance Learning (CA-MIL)

Standard MIL approaches for VAD employ a fixed top-k selection strategy to aggregate snippet scores into video-level predictions [1, 8]. However, anomalous events vary significantly

in duration—a shooting may last only a few seconds while a burglary may span several minutes. Using a fixed k leads to either including irrelevant snippets (over-selection) or missing critical anomalous segments (under-selection).

CA-MIL addresses this limitation by adaptively determining the number of snippets to aggregate based on the confidence distribution of anomaly scores within each video.

3.3.1 Adaptive Threshold Computation

For a video with L valid snippets and predicted scores $\mathbf{s} = \{s_t\}_{t=1}^{L}$, an adaptive threshold τ is computed based on the score statistics:

$$\tau = \mu_s + \lambda_{std} \cdot \sigma_s \tag{2}$$

where $\mu_s = \frac{1}{L}\sum_{t=1}^{L} s_t$ is the mean score, $\sigma_s = \sqrt{\frac{1}{L}\sum_{t=1}^{L}(s_t - \mu_s)^2}$ is the standard deviation, and λ_{std} is a hyperparameter controlling selectivity (set to 0.5 in experiments). The threshold is clamped to $[\tau_{min}, \tau_{max}]$ to prevent extreme values.

The choice of mean + standard deviation is motivated by four considerations specific to weakly-supervised VAD. First, the formulation is parameter-free: no auxiliary network or learned scalar is introduced, which matters under weak supervision where additional parameters tend to overfit to video-level labels rather than recover frame-level structure. Second, it is per-video adaptive: the threshold automatically follows the score distribution of each input, without conditioning on a video-level feature that would itself need to be aggregated. Third, it admits a clear statistical interpretation—a snippet is selected when its score lies at least λ_{std} standard deviations above the video's own mean, i.e. when it is a positive outlier relative to its local context. Fourth, the formulation is architecture-agnostic and can be dropped into any MIL pipeline at zero parameter cost. Two alternatives were considered and rejected during development: a global confidence threshold ($\tau = c$, fixed across videos), which does not adapt to per-video score scales and consistently under-selects on hard videos; and a percentile rule ($\tau = Q_p(\mathbf{s})$), which is more brittle to score outliers and produced unstable training in early experiments.

Under weak supervision, the MIL loss provides no direct signal on the optimal selectivity: a learned λ_{std} tends to drift downward (selecting more snippets reduces the loss on anomaly videos in the short term) and erases per-video adaptivity. Fixing λ_{std} at a moderate positive value preserves the intended positive outlier semantics. The specific value 0.5 was selected by grid search over $\{0.0, 0.25, 0.5, 0.75, 1.0\}$ on a held-out validation split, and the method is empirically insensitive to nearby values—Avg mAP on UCF-Crime varied within ± 0.15 over $\lambda_{std} \in [0.25, 0.75]$.

The threshold $\tau = \mu_s + 0.5\,\sigma_s$ has a clean probabilistic reading: under a Gaussian approximation to the per-video score distribution, it marks the upper $\sim 31\%$ tail. Combined with the cap $k_{max} \approx 0.15L$ from §3.3.2, this yields effective selection of the top $\sim 15\%$ of snippets on well-separated anomaly videos, with the floor $k_{min} \approx 0.05L$ taking over on flat distributions. Smaller values ($\lambda_{std} \in \{0.0, 0.25\}$) admit marginal snippets just above the mean and blunt the anomaly–normal contrast; larger values ($\lambda_{std} \geq 1.0$) raise τ past the secondary mode in bimodal anomaly videos and miss the boundaries of short anomalies, hurting mAP at

strict IoU. $\lambda_{std} = 0.5$ is the smallest grid value that keeps τ strictly above the per-video mean while remaining within the k_{max} cap on typical videos.

3.3.2 Bounded Adaptive Selection

The adaptive k is determined as the number of snippets exceeding the threshold:

$$k_{adaptive} = |\{t: s_t \geq \tau\}| \tag{3}$$

To ensure training stability, this value is bounded relative to a base selection ratio r:

$$k = \text{clamp}(k_{adaptive}, \lfloor \alpha_{min} \cdot r \cdot L \rfloor, \lfloor \alpha_{max} \cdot r \cdot L \rfloor) \tag{4}$$

where $\alpha_{min} = 0.5$ and $\alpha_{max} = 1.5$ define the allowable range, and $r = 0.1$ following standard practice.

α_{min} and α_{max} bounds determine how far the adaptive count k may deviate from the base count $r \cdot L = 0.1L$. The lower bound $\alpha_{min} = 0.5$ ($k_{min} \approx 0.05L$) prevents pathological under-selection when the adaptive threshold collapses on flat-scored videos: keeping at least $\sim 5\%$ of snippets stabilizes the top-k mean $\hat{y}_{bin}$ across batches. The upper bound $\alpha_{max} = 1.5$ ($k_{max} \approx 0.15L$) caps selection when an unusually large fraction of snippets exceeds τ, accommodating the longest realistic anomalies on UCF-Crime without diluting the aggregate with normal context. Figure 2 confirms that, after training, the empirical distribution of k stays interior to $[0.05L, 0.15L]$ for most test videos, so neither bound is routinely binding. The setting $(0.5,1.5)$ was the most stable in a coarse grid over $(\alpha_{min}, \alpha_{max}) \in \{(0.5,1.5), (0.7,1.3), (0.3,1.7), (0.5,2.0)\}$: narrower ranges erased CA-MIL's per-video adaptivity; wider ranges increased run-to-run variance without improving mean Avg mAP.

3.3.3 MIL Loss Formulation

For binary anomaly detection, the video-level prediction is obtained by averaging the top-k scores:

$$\hat{y}_{bin} = \frac{1}{k} \sum_{t \in \mathcal{T}_k} s_t \tag{5}$$

where $\mathcal{T}_k$ denotes the indices of the top-k scoring snippets. The binary MIL loss follows standard formulation:

$$\mathcal{L}_{bin} = -y\log(\hat{y}_{bin}) - (1-y)\log(1-\hat{y}_{bin}) \tag{6}$$

For multi-class vision-language alignment, class logits $\mathbf{l}_t \in \mathbb{R}^C$ are similarly aggregated:

$$\hat{l} = \frac{1}{k} \sum_{t \in \mathcal{T}_k} l_t \tag{7}$$

with cross-entropy loss over the C classes:

$$\mathcal{L}_{vl} = -\sum_{c=0}^{C-1} y_c \log\left(\frac{\exp(\hat{l}_c)}{\sum_{j=0}^{C-1} \exp(\hat{l}_j)}\right) \tag{8}$$

The key property of CA-MIL is per-video adaptivity driven by the shape of the score distribution. When the predicted scores are strongly bimodal (high σ_s, indicating a clear separation between an anomaly cluster and a normal cluster), the threshold $\mu_s + \lambda_{std}\sigma_s$ rises and CA-MIL selects only the few snippets in the high-score mode. When scores are diffuse (low σ_s, indicating an ambiguous distribution), the threshold collapses near the mean and CA-MIL admits more snippets, with the bounds k_{min} and k_{max} preventing extremes in either direction. Anomaly duration m enters only indirectly through its effect on the score distribution, not as a direct input to the selection rule. This adaptivity addresses both over-selection and under-selection problems inherent in fixed-k approaches.

The adaptive threshold $\tau = \mu_s + \lambda_{std}\sigma_s$ could be miscalibrated on normal videos, where all snippets are expected to score low. CA-MIL guards against this through three coupled mechanisms.

1. Weak-label discipline: For a normal video, $y = 0$, so the binary MIL loss reduces to $\mathcal{L}_{bin} = -\log(1 - \hat{y}_{bin})$, which directly penalizes any high-confidence top-k aggregate. Even if CA-MIL selects the most anomaly-like snippets in a normal video, training pushes those very scores toward zero. The selection mechanism therefore cannot, by construction, learn to inflate scores on normal videos without incurring loss.

2. Threshold clamping: The raw threshold is clipped to $[\tau_{min}, \tau_{max}] = [0.2, 0.8]$ before counting snippets. This prevents pathological behavior in two regimes: on confident normal videos where σ_s is very small (threshold would collapse to μ_s and select almost every snippet), and on uniformly noisy videos where the threshold would otherwise drift outside the valid score range.

3. Bounded k: The number of selected snippets is constrained to $k \in [\lfloor \alpha_{min} \cdot r \cdot L \rfloor, \min(\lfloor \alpha_{max} \cdot r \cdot L \rfloor, \lfloor 0.2L \rfloor)]$, i.e. at most roughly 15–20% of snippets for $\alpha_{max} = 1.5$, $r = 0.1$. The upper bound prevents over-selection on a normal video whose threshold happens to fall below the bulk of its scores. Empirically, the impact is visible in Figure 3: on the UCF-Crime test set, normal frames concentrate at a mean predicted score of 0.251 while anomalous frames concentrate at 0.843, indicating that the trained selection mechanism does not produce systematically inflated scores on normal content.

3.3.4 Theoretical Analysis of Adaptive MIL

The gradient of the binary MIL loss reaches only the k selected snippets:

$$\frac{\partial \mathcal{L}_{bin}}{\partial s_t} = \frac{1}{k} \frac{\partial \mathcal{L}_{bin}}{\partial \hat{y}_{bin}} \mathbb{1}[t \in \mathcal{T}_k] \tag{9}$$

Under fixed k, two pathologies arise depending on the true anomaly extent m. When $m > k$, only the highest-scoring snippets receive gradients while $m - k$ genuine anomaly snippets are ignored (under-coverage). When $m < k$, the selection admits $k - m$ normal snippets, diluting the gradient with conflicting signal. Adaptive $k = |\{t: s_t \geq \tau(s)\}|$ tracks per-video anomaly mass, providing extent-aware gradient coverage that removes both mismatches.

This adaptive selection also synergizes with the smoothness regularizer $\mathcal{L}_{smooth}$. Under fixed k, on short anomalies $\mathcal{T}_k$ extends beyond the true interval, causing smoothness to

incorrectly pull boundary snippets upward. Under adaptive k, $\mathcal{T}_k$ stays interior, so smoothness compresses scores within the anomaly while boundaries fall steeply. Table 6 confirms this compounding: CA-MIL plus smoothness yields mAP@0.5 improvement from 2.93 to 5.96 (+103%), exceeding either component alone.

Frame-level AUC depends only on rank order and is insensitive to extent mismatch; mAP at strict IoU instead penalizes temporal support deviation. This predicts the empirical pattern in Table 6: CA-MIL alone delivers the largest single mAP increment (6.68 → 9.05, +2.37 absolute) at the cost of a modest AUC drop (88.02 → 87.21); the remaining components (CCMA, smoothness, dual-head, STCA) recover most of that AUC, so the full pipeline ends at only −0.10% AUC for +40.3% Avg mAP, with the largest gain at mAP@0.5 (+103%).

3.4 Temporal Smoothness Regularization

Anomaly scores often exhibit temporal inconsistencies, such as isolated spikes in normal regions or unexpected dips in anomalous scores during anomalous events. These unusual peaks and dips in scores arise because the models optimize for video-level classification without explicit encouragement for temporal coherence.

A regularization term penalizes abrupt changes between consecutive predictions:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{T-1}\sum_{t=1}^{T-1} | s_{t+1} - s_t | \tag{10}$$

This loss encourages temporally consistent predictions while still allowing genuine transitions between normal and anomalous segments. The regularization weight λ_{smooth} is set to 0.1 to balance smoothness against discriminative capacity.

3.5 Dual Head Fusion

Single classification heads can produce overconfident predictions on ambiguous snippets. A dual-head fusion mechanism combines two scoring heads with an inductive bias through uncertainty-aware gating.

3.5.1 Complementary Scoring Heads

Two parallel classification heads with different architectures are employed:

$$s_t^{(1)} = \sigma\big(\text{MLP}(h_t)\big) \tag{11}$$

$$s_t^{(2)} \& = \sigma(\text{Conv1D}(H)_t \tag{12}$$

where MLP is a two-layer perceptron that captures point-wise patterns, and Conv1D is a 1D convolution (kernel size 3) capturing local temporal patterns. Both produce scores $s_t^{(1)}, s_t^{(2)} \in [0,1]$.

3.5.2 Uncertainty-Aware Gating

The disagreement between heads serves as an uncertainty indicator:

$$u_t = \left| s_t^{(1)} - s_t^{(2)} \right| \tag{13}$$

High disagreement indicates ambiguous regions. The fusion weights are computed through a learnable gating network:

$$g_t = \sigma(\text{MLP}_g([s_t^{(1)}; s_t^{(2)}; u_t; h_t])) \tag{14}$$

The final fused score combines both predictions:

$$s_t = g_t \cdot s_t^{(1)} + (1 - g_t) \cdot s_t^{(2)} \tag{15}$$

This mechanism allows the model to adaptively weight the contribution of each head based on their relative confidence for each snippet.

3.5.3 Training of the Gating Network

The gating network MLP_g receives no direct snippet-level supervision; it is trained jointly end-to-end with the rest of the model through the primary weakly-supervised MIL loss. Gradients reach the gate parameters from $\mathcal{L}_{\text{bin}}$ and $\mathcal{L}_{\text{vl}}$ via the fused score $s_t = g_t \cdot s_t^{(1)} + (1 - g_t) \cdot s_t^{(2)}$: any snippet that contributes to the top-k aggregate $\hat{y}_{\text{bin}}$ propagates a gradient back through both heads and through g_t, so the gate is implicitly supervised to upweight whichever head reduces the loss at that snippet.

Two auxiliary regularizers further shape gate behavior beyond what the MIL signal alone enforces. The first is a confidence-weighted agreement penalty,

$$\mathcal{L}_{\text{agree}} = \frac{1}{T}\sum_{t=1}^{T} max\left(max(s_t^{(1)}, 1 - s_t^{(1)}), max(s_t^{(2)}, 1 - s_t^{(2)})\right) \cdot u_t, \tag{16}$$

which penalizes disagreement precisely where at least one head is already confident (score close to 0 or 1), giving the gate a clean signal in unambiguous regions and concentrating the gating decision on genuinely uncertain snippets. The second is a diversity margin on inter-head cosine similarity,

$$\mathcal{L}_{\text{div}} = max\big(0, \cos\big(s^{(1)}, s^{(2)}\big) - 0.9\big), \tag{17}$$

which activates only when the two heads become near-identical and prevents collapse to a degenerate solution in which the gate is uninformative.

Both regularizers are added to the overall training objective, with weights $\lambda_{\text{agree}} = 0.1$ and $\lambda_{\text{div}} = 0.05$. The gating network is a lightweight 3-layer MLP adding roughly 2K parameters and is optimized with the same Adam settings as the rest of the model.

3.6 Score-level Temporal Context Aggregation (STCA)

STCA is a lightweight inference-time module that refines predictions by incorporating temporal context at the score level, requiring no additional training.

3.6.1 Context Encoding

For each snippet t with initial score s_t, local and global context are computed:

Local context captures the immediate temporal neighborhood:

$$c_t^{\text{local}} = \frac{1}{2w+1}\sum_{i=-w}^{w} s_{t+i} \tag{18}$$

where w is the window half-size (set to 1, corresponding to a 3-snippet local kernel).

Global context captures video-level statistics:

$$c_t^{\text{global}} = [\mu_s; \sigma_s; s_{max}; \text{rank}(s_t)] \tag{19}$$

where $\text{rank}(s_t)$ is the normalized rank of s_t among all scores.

3.6.2 Score Refinement

The refined score combines original prediction with contextual information:

$$\hat{s}_t = \alpha \cdot s_t + \beta \cdot c_t^{\text{local}} + \gamma \cdot f_{\text{global}}(c_t^{\text{global}}) \tag{20}$$

where α, β, γ are weights satisfying $\alpha + \beta + \gamma = 1$, and f_{global} maps global context to a scalar adjustment via a sigmoid over the per-snippet z-score $(s_t - \mu_s)/\sigma_s$. In practice, $\alpha = 0.8$, $\beta = 0.15$, $\gamma = 0.05$ are used.

The three weights encode an explicit prior on how much each information source should influence the final score, and were not selected arbitrarily. The dominant weight on the raw score ($\alpha = 0.8$) ensures STCA acts as a refinement rather than a re-prediction: the discriminative signal learned during training is preserved, and STCA only nudges scores when the temporal evidence is strong. The moderate weight on local context ($\beta = 0.15$) provides short-range smoothing that suppresses isolated single-snippet spikes and fills brief gaps inside anomaly intervals, without erasing genuine sharp transitions at event boundaries. The small weight on global context ($\gamma = 0.05$) introduces a gentle, video-level calibration: snippets that are atypical relative to the rest of the video are slightly elevated, snippets near the video mean are slightly suppressed. The constraint $\alpha + \beta + \gamma = 1$ keeps $\hat{s}_t$ within the valid score range $[0,1]$, removing the need for a post-hoc rescaling step.

The specific values were selected by a coarse grid search on a held-out validation split, with $\alpha \in \{0.6, 0.7, 0.8, 0.9\}$ and the remaining mass split between β and γ at a 3: 1 ratio (chosen because local context is the more informative signal at the snippet scale). The setting $(0.8, 0.15, 0.05)$ produced the best Avg mAP on UCF-Crime; the method is largely insensitive within $\alpha \in [0.7, 0.85]$ (Avg mAP changing by less than 0.1). Aggressive smoothing ($\alpha < 0.6$) noticeably hurts performance at strict IoU thresholds because it blurs boundaries; almost no smoothing ($\alpha > 0.9$) reduces STCA's contribution to noise. A learnable-weight variant was also implemented, with (α, β, γ) produced by a softmax over learned logits initialized to $(0.6, 0.3, 0.1)$. In our setting it performed comparably to the fixed scheme but with higher run-to-run variance, so the fixed configuration was retained for reporting. STCA operates purely at the score level, making it computationally efficient and applicable as a plug-in module during inference.

3.7 Cross-Modal Attention for Vision-Language Alignment

Cross-modal attention mechanism is employed to aggregate visual features guided by text embeddings. CCMA leverages the vision-language alignment within VadCLIP's dual branch framework.

3.7.1 Attention Mechanism

Given visual features $H \in \mathbb{R}^{T\times d}$ and text embeddings $E \in \mathbb{R}^{C\times d}$, attention weights are computed through cross-attention:

$$A = \text{softmax}\left(\frac{EW_Q(HW_K)^{\top}}{\sqrt{d_h}}\right) \tag{21}$$

where $W_Q, W_K \in \mathbb{R}^{d\times d_h}$ are learnable projections. This produces attention maps $A \in \mathbb{R}^{C\times T}$ where $A_{c,t}$ represents the relevance of snippet t for class c.

The attended visual features for each class are computed as:

$$h_c = \sum_{t=1}^{T} A_{c,t} \cdot h_t \tag{22}$$

Vision-language similarity is then computed between aggregated features and text embeddings for classification. Quantitative analysis confirms a strong correlation between the underlying mechanism and the resulting anomaly scores, with an AUPRC that exceeds 0.90 when comparing attention weights to predicted values. However, class-specific behaviors reveal a limitation. The mean cosine similarity of 0.885 suggests that attention patterns in various types of anomalies remain remarkably similar. Consequently, the cross-modal attention component contributes to coarse-grained detection with an AUC improvement of 0.07%, while the gains in temporal localization are primarily CA-MIL, temporal smoothness, and STCA.

3.8 Training and Inference

The complete training objective combines all loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bin}} + \mathcal{L}_{\text{vl}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{agree}}\mathcal{L}_{\text{agree}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} \tag{23}$$

where $\lambda_{\text{smooth}} = 0.1$ balances the smoothness regularization, and $\lambda_{\text{agree}} = 0.1$, $\lambda_{\text{div}} = 0.05$ weight are the gate-shaping regularizers.

The model is trained using the AdamW optimizer with a specified learning rate 10^{-4} and weight decay 10^{-5}. The learning rate is decayed by a factor of 0.5 at epochs 5 and 10. Training uses a batch size of 8 with gradient clipping at norm 1.0 and proceeds for 15 epochs. The best model is selected based on validation AUC. During inference, visual features are extracted and processed through the LGT adapter, cross-modal attention, and dual-head fusion to obtain initial scores. STCA then refines these scores using temporal context. The final snippet-level scores are used for both frame-level AUC evaluation and temporal localization with mAP metrics.

4. Results and Discussion

This section presents an experimental evaluation of the proposed training enhancements. The evaluation covers both detection performance (AUC, AP) and temporal localization accuracy (mAP) at multiple Intersection over Union (IoU) thresholds, following the temporal action localization protocol [21], with ablation studies examining the contribution of each component.

4.1 Datasets

The proposed methodology was evaluated on two widely-used benchmarks for weakly supervised video anomaly detection, summarized in Table 1.

Table 1. Summary of Evaluation Datasets

Property	UCF-Crime [1]	XD-Violence [8]
Total Videos	1,900	4,754
Total Duration	128 hours	217 hours
Anomaly Classes	13	6
Training Videos	1,610	3,954
Test Videos	290	800
Training Labels	Video-level	Video-level
Test Labels	Frame-level	Frame-level

4.2 Implementation Details

The proposed enhancements were implemented on top of the VadCLIP codebase [7]. Visual features were extracted using a frozen CLIP ViT-B/16 encoder, producing 512-dimensional features for each video snippet. Text embeddings for anomaly categories were obtained through the CLIP text encoder with learnable prompt tokens (context length 20). Training was performed using the AdamW optimizer with an initial learning rate 10^{-4} and weight decay 10^{-5}. The learning rate was reduced by a factor of 0.5 at epochs 5 and 10. Models were trained for 15 epochs with a batch size of 8 on a single NVIDIA RTX 3090 GPU. Gradient clipping with a maximum norm of 1.0 was applied for training stability.

For CA-MIL, the threshold parameter λ_{std} was set to 0.5, with base selection ratio $r = 0.1$ and bounds $\alpha_{min} = 0.5$, $\alpha_{max} = 1.5$. The temporal smoothness weight λ_{smooth} was set to 0.1. For STCA, the local window half-size w was set to 1 (3-snippet kernel), with refinement weights $\alpha = 0.8$, $\beta = 0.15$, $\gamma = 0.05$. These hyperparameters were selected based on validation performance and remained fixed across both datasets.

4.3 Quantitative Results

4.3.1 Coarse-Grained Detection Results

As shown in Table 2, on UCF-Crime the proposed approach achieved an 87.92% AUC, comparable to the VadCLIP baseline (88.02%). The difference in AUC (-0.10%) is expected since the proposed improvements were designed to enhance temporal localization rather than detection itself. Table 3 provides the results obtained for XD-Violence, where the proposed architecture achieved 84.73% AP performance, outperforming VadCLIP (84.51%, +0.22%) by a small margin. Such differences are usual for this benchmark and correspond to its inherent run-to-run variance. However, they also illustrate the tradeoff between the two approaches: the proposed modules guide scores based on their localization quality, not frame-level separability. Temporal smoothness regularization and STCA score refinement attempt to cluster the neighboring scores, reducing individual peaks rewarded by the AUC criterion, whereas CA-MIL scoring changes top-k selection for videos, slightly shifting the optimization signal for the binary classifier. Ablation study results provided in Table 6 confirm this intuition: CA-MIL drops AUC to 87.21%, and then CCMA, smoothness regularization, dual-head fusion, and STCA bring some improvement, leading to the final value of 87.92%. Meanwhile, all of these features increase Avg mAP from 6.68 to 9.37. On the XD-Violence benchmark, 84.73% AP

accuracy was achieved, beating VadCLIP (84.51%, +0.22%). These results indicate that the proposed training modifications maintained competitive detection performance while substantially improving localization.

Table 2. Coarse-Grained Detection Results on UCF-Crime

Method	Venue	AUC (%)
Sultani et al. [1]	CVPR 2018	84.14
Wu et al. [8]	ECCV 2020	84.57
RTFM [3]	ICCV 2021	85.66
AVVD [21]	ECCV 2022	82.45
CLIP-TSA [13]	ICIP 2023	87.58
DMU [5]	AAAI 2023	86.75
HACSPT [11]	SIVP 2025	87.05
VadCLIP [7]	AAAI 2024	88.02
Proposed	-	87.92

Table 3. Coarse-Grained Detection Results on XD-Violence

Method	Venue	AP (%)
Sultani et al. [1]	CVPR 2018	75.18
Wu et al. [8]	ECCV 2020	80.00
RTFM [3]	ICCV 2021	78.27
AVVD [21]	ECCV 2022	78.10
CLIP-TSA [13]	ICIP 2023	82.17
DMU [5]	AAAI 2023	82.41
VadCLIP [7]	AAAI 2024	84.51
Proposed	-	84.73

4.3.2 Fine-Grained Detection Results

The results presented in both Table 4 and Table 5 show a significant increase in localization performance on both datasets. On UCF-Crime, the average mAP improves from 6.68 to 9.37, an absolute improvement of 40.3% over VadCLIP. Scores shown a notable increase on XD-Violence, where the average mAP is 31.63%, an improvement of 28.1%. This improvement is consistent across all IoU thresholds and demonstrates the model's robust ability to handle both coarse localization and precise boundary detection. This robustness is most evident at the stricter thresholds; where the mAP@0.5 scores have a notable increase of 103.4% on UCF-Crime and 40.0% on XD-Violence. Absolute mAP on UCF-Crime is universally low across the weakly-supervised literature because the metric and the supervision regime are fundamentally mismatched. Training receives only video-level labels, so no frame-level signal is available to align predictions with anomaly boundaries during optimization; the mAP protocol, in contrast, averages precision over IoU thresholds 0.1–0.5 and penalizes boundary error severely. Within this regime the baselines on the same protocol (Table 4) reach 3.24 (Sultani), 6.05 (AVVD), and 6.68 (VadCLIP); 9.37 is the highest reported value among comparable WSVAD methods and exceeds VadCLIP by a wider margin ($+2.69$ absolute) than the gap between any two prior methods in the table. The disproportionately large $+103\%$ gain at the strictest threshold mAP@0.5 ($2.93 \rightarrow 5.96$) further indicates that the proposed pipeline pushes localization noticeably closer to the ground-truth boundaries.

Table 4. Fine-Grained Temporal Localization Results on UCF-Crime (mAP %)

Method	@0.1	@0.2	@0.3	@0.4	@0.5	Avg
Random	0.21	0.14	0.04	0.02	0.01	0.08
Sultani [1]	5.73	4.41	2.69	1.93	1.44	3.24
AVVD [21]	10.27	7.01	6.25	3.42	3.29	6.05

VadCLIP [7]	11.72	7.83	6.40	4.53	2.93	6.68
Proposed	14.00	11.64	8.94	6.32	5.96	9.37
Improvement	+19.5%	+48.7%	+39.7%	+39.5%	+103%	+40.3%

Table 5. Fine-Grained Temporal Localization Results on XD-Violence (mAP %)

Method	**@0.1**	**@0.2**	**@0.3**	**@0.4**	**@0.5**	**Avg**
Random	1.82	0.92	0.48	0.23	0.09	0.71
Sultani [1]	22.72	15.57	9.98	6.20	3.78	11.65
AVVD [21]	30.51	25.75	20.18	14.83	9.79	20.21
VadCLIP [7]	37.03	30.84	23.38	17.90	14.31	24.70
Proposed	45.17	37.62	30.76	24.55	20.04	31.63
Improvement	+22.0%	+22.0%	+31.6%	+37.2%	+40.0%	+28.1%

4.4 Ablation Study

To understand the contribution of each proposed component, ablation experiments were conducted by incrementally adding components to the VadCLIP baseline.

Table 6. Component Ablation Study on UCF-Crime

CA-MIL	**CCMA**	**Smooth**	**Dual-Head**	**STCA**	**AUC (%)**	**Avg mAP**
✓	×	×	×	×	87.21	9.05
✓	✓	×	×	×	87.28	9.27
✓	✓	✓	×	×	87.36	9.29
✓	✓	✓	✓	×	87.96	9.29
✓	✓	✓	✓	✓	87.92	9.37

4.4.1 Impact of CA-MIL

Replacing the fixed top-k selection with confidence-adaptive selection significantly increased Avg mAP from 6.68 to 9.05 (+2.37). The AUC dropped to 87.21%, a trade-off that is expected: adapting k to the score distribution favors tighter temporal boundaries at the cost of slightly less stable coarse-grained classification.

4.4.2 Impact of Cross-Modal Attention

Cross-modal attention on top of CA-MIL brought a modest mAP gain (+0.22, from 9.05 to 9.27) together with a small recovery of AUC (+0.07%, from 87.21 to 87.28). Analysis of the learned attention maps showed that all anomaly classes develop nearly identical patterns, with an average pairwise cosine similarity of 0.885. In practice, CCMA appears to sharpen the boundary between anomalous and normal content, in general, rather than learning category-specific cues, which explains why its localization benefit remains limited.

4.4.3 Impact of Temporal Smoothness

Temporal smoothness adds a penalty and has a positive effect. It increases the mAP from 9.27 to 9.29 and further increases AUC to 87.36%. This penalty helps to make the output more stable by reducing random fluctuations in normal areas and bridging short gaps during actual anomaly periods. It focuses on the variation between consecutive score differences, and the penalty effectively removes noisy single-snippet errors without affecting the transitions at real anomaly boundaries.

4.4.4 Impact of Dual-Head Fusion

Dual-head mechanisms led to an increase in AUC by +0.60%, reaching 87.96% while maintaining the average mAP at 9.29. The point-wise MLP and the local-context convolution head pick up different feature signals, which are then combined using uncertainty-aware gating. The significant increase in AUC indicates that this fusion strategy effectively addresses ambiguities in snippets that previously obstructed a single classifier.

4.4.5 Role of STCA

Score-level Temporal Context Aggregation is applied at the inference stage. STCA improves Avg mAP from 9.29 to 9.37 $(+0.08)$ while slightly reducing AUC from 87.96% to 87.92% (-0.04%). The small AUC drop is consistent with the trade-off argued throughout this section: STCA's local averaging dampens the isolated single-frame score spikes that AUC's ranking metric rewards, in exchange for a tighter alignment between predicted and ground-truth anomaly intervals. STCA adds negligible computational overhead and requires no additional learnable parameters, operating directly on the predicted score sequence at inference time.

4.4.6 Statistical Significance

Results from Tables 4-6 are obtained through a single training run using a fixed seed provided along with the codebase. It should be admitted that performing multi-seed experiments with variance reporting could increase the validity of our experimental study. Given the limited computing resources, it was impossible to perform multi-seed testing in the course of this project. Nevertheless, the level of performance increase (+2.69 mAP, +3.03 mAP@0.5) is relatively large compared to usual one-digit percentage increases in the field of VAD. Multi-seed validation is our next step.

4.5 Analysis

4.5.1 Behavior of Confidence-Adaptive Selection

Figure 2 shows the effectiveness of CA-MIL by examining the distribution of selection values (k) across the 290 UCF-Crime test videos. Most MIL techniques apply a static $k = \lfloor T/16 \rfloor + 1$, which results in a constant $k = 23$ regardless of video and anomaly context. In contrast, CA-MIL produces a right-skewed distribution where it maintains a comparable mean $\bar{k} = 24.4$ and exhibits significantly higher variance.

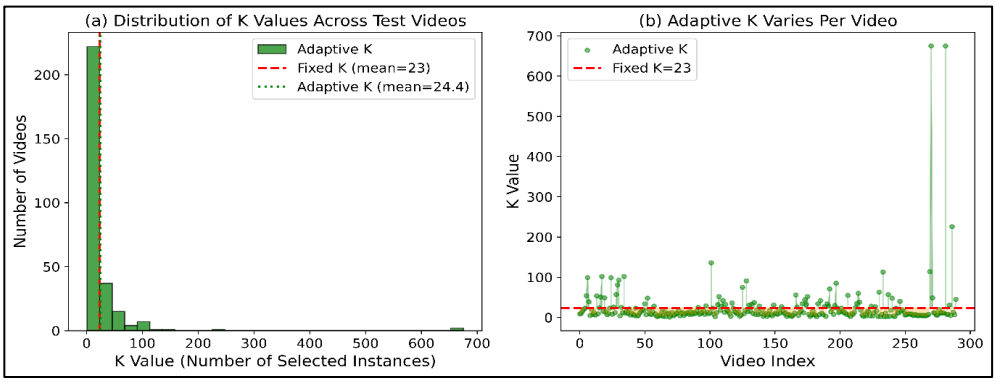

Figure 2. CA-MIL Frames Selection on UCF-Crime Test Videos. (a) Distribution of k Values (b) Per-Video k

4.5.2 Score Distribution Analysis

Figure 3 presents the effectiveness of the scoring function in the UCF-Crime test set by examining the distribution of predicted anomaly scores in normal and anomalous frames. The results demonstrate a clear separation between normal and abnormal frames where normal frames are concentrated around a mean $\mu = 0.251$, whereas anomalous frames shift towards $\mu = 0.843$. This substantial gap of approximately 0.6 highlights the model's robustness.

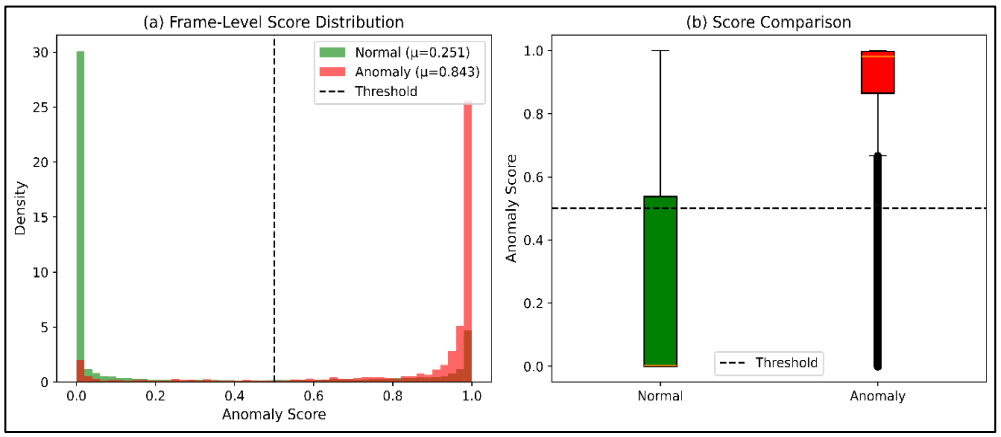

Figure 3. Distribution of Predicted Anomaly Scores on UCF-Crime. (a) Density Plot (b) Box Plot

4.5.3 Cross-Modal Attention Pattern Analysis

As illustrated in Figure 4, the cosine similarity matrix for pairs of attention maps from different classes has been depicted. As expected, all 14 classes of anomalies share extremely high levels of cosine similarity (on average 0.885) with very low diversity (0.115). Such a result indicates that, while performing well in the detection of "abnormality", it is incapable of detecting the characteristic elements of each class.

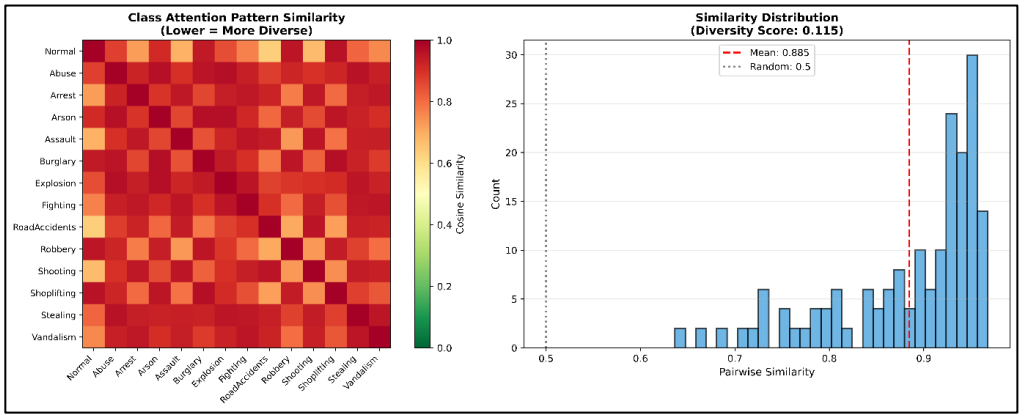

Figure 4. Analysis of Cross-Modal Attention Diversity across Anomaly Classes. (a) Pairwise Cosine Similarity matrix (b) Distribution of Similarity Value

4.5.4 Qualitative Evaluation

Figure 5 shows the anomaly scores for four test videos from the UCF-Crime dataset. In each figure, the combined anomaly scores (in blue) are plotted against the ground-truth annotations (in gray bars). In all instances, the model predicts high confidence scores (0.87–1.00) for anomalies, whereas zero scores are generated for normal frames.

To quantify the temporal smoothness visible in Figure 5, the mean absolute first difference of predicted scores, $\Delta_s = \frac{1}{L-1}\sum_{t=1}^{L-1} |s_{t+1} - s_t|$, was computed on every UCF-Crime test video. The proposed pipeline yields a mean $\Delta_s = 0.0063 \pm 0.0069$ across the 290 test videos, indicating that consecutive snippet scores remain close to each other within both normal and anomalous regions. The same qualitative pattern is visible in Figure 5: the trajectories transition cleanly into and out of the annotated intervals rather than oscillating around the decision boundary, which is the primary mechanism behind the observed gains at strict IoU thresholds.

4.6 Computational Analysis

The inference-time computational overhead of the proposed pipeline is evaluated on a single video of length $T = 256$ snippets with $D = 512$-dimensional visual features using an NVIDIA RTX 4060 GPU. The complete proposed model requires 169.86 M parameters, executing 41.42 G FLOPs with an end-to-end latency of 36.57 ms per video. Within this architecture, the STCA module introduces minimal overhead; as an inference-only component consisting of a single 1-D convolution and three element-wise terms, it adds zero training cost and represents the most computationally efficient component relative to its performance gains.

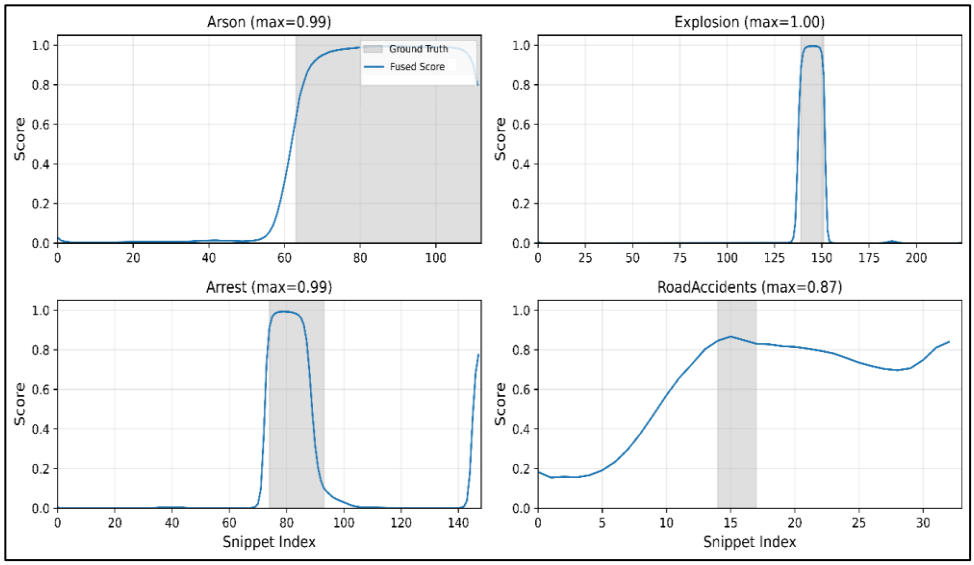

Figure 5. Qualitative Results on UCF-Crime Test Videos Showing Anomaly Score Trajectories for the Arson, Explosion, Arrest, And Road Accidents Categories

5. Limitations

There are several limitations that need discussion. 1) Weakly supervised fine-grained anomaly category detection is still underexplored, since CCMA attention maps are largely similar across different anomaly categories. 2) Extremely short anomalies (1 or 2 snippets) are hard to localize because of limited temporal context within each snippet.

6. Conclusion

This paper addressed temporal localization in vision-language based video anomaly detection by introducing four targeted training and inference enhancements on the VadCLIP

backbone: Confidence-Adaptive Multiple Instance Learning (CA-MIL), which adaptively selects snippets based on the per-video score distribution; temporal smoothness regularization, which discourages abrupt score transitions; dual-head score fusion, which combines point-wise and temporal scoring perspectives through uncertainty-aware gating; and Score-level Temporal Context Aggregation (STCA), an inference-time refinement requiring no additional parameters. On UCF-Crime, average mAP rose from 6.68 to 9.37 (+40.3%), with mAP@0.5 improving from 2.93 to 5.96 (+103%); on XD-Violence, average mAP increased from 24.70 to 31.63 (+28.1%). Detection performance was preserved (UCF-Crime AUC −0.10%, XD-Violence AP +0.22%), indicating that localization and detection benefit from different optimization pressures. Future work will explore richer multimodal fusion strategies, weakly supervised techniques for tighter temporal boundaries, and real-time deployment scenarios where localization precision is critical.

References

[1] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-World Anomaly Detection in Surveillance Videos." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 6479-6488.

[2] Zhong, Jia-Xing, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. "Graph Convolutional Label Noise Cleaner: Train a Plug-And-Play Action Classifier for Anomaly Detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 1237-1246.

[3] Tian, Yu, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. "Weakly-Supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning." In Proceedings of the IEEE/CVF international conference on computer vision, 2021, 4975-4986.

[4] Li, Shuo, Fang Liu, and Licheng Jiao. "Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection." In Proceedings of the AAAI conference on artificial intelligence, vol. 36, no. 2, 2022, 1395-1403.

[5] Zhou, Hang, Junqing Yu, and Wei Yang. "Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 3, 2023, 3769-3777.

[6] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning Transferable Visual Models from Natural Language Supervision." In International conference on machine learning, PmLR, 2021, 8748-8763.

[7] Wu, Peng, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. "Vadclip: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection." In Proceedings of the AAAI conference on artificial intelligence, vol. 38, no. 6, 2024, 6074-6082.

[8] Wu, Peng, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. "Not Only Look, But Also Listen: Learning Multimodal Violence Detection Under

Weak Supervision." In European conference on computer vision, Cham: Springer International Publishing, 2020, 322-339.

[9] Feng, Jia-Chang, Fa-Ting Hong, and Wei-Shi Zheng. "Mist: Multiple Instance Self-Training Framework for Video Anomaly Detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, 14009-14018.

[10] Chen, Yingxian, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. "Mgfn: Magnitude-Contrastive Glance-And-Focus Network for Weakly-Supervised Video Anomaly Detection." In Proceedings of the AAAI conference on artificial intelligence, vol. 37, no. 1, 2023, 387-395.

[11] Shah, Siddharth, and Dr Narendrasinh Chauhan. "HACSPT: A Hybrid Adaptive Contrastive Self-Paced Transformer for Video Anomaly Detection: S. Shah, Dr. N. Chauhan." Signal, Image and Video Processing 19, no. 14 (2025): 1187.

[12] Zhou, Kaiyang, Jingkang Yang, Chen Change Loy, and Ziwei Liu. "Learning to Prompt for Vision-Language Models." International journal of computer vision 130, no. 9 (2022): 2337-2348.

[13] Joo, Hyekang Kevin, Khoa Vo, Kashu Yamazaki, and Ngan Le. "Clip-TSA: Clip-Assisted Temporal Self-Attention for Weakly-Supervised Video Anomaly Detection." In 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, 3230-3234.

[14] Lv, Hui, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. "Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, 8022-8031.

[15] Wu, Peng, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. "Open-Vocabulary Video Anomaly Detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 18297-18307.

[16] Yang, Zhiwei, Jing Liu, and Peng Wu. "Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, 18899-18908.

[17] Liu, Long, Jianjun Li, Guang Li, Yunfeng Zhai, and Ming Zhang. "VadCLIP++: Dynamic Vision-Language Model for Weakly Supervised Video Anomaly Detection." Digital Signal Processing (2025): 105560.

[18] Li, Min, Jing Sang, Yuanyao Lu, and Lina Du. "WSVAD-CLIP: Temporally Aware and Prompt Learning with CLIP for Weakly Supervised Video Anomaly Detection." Journal of Imaging 11, no. 10 (2025): 354.

[19] Lan, Tian, Yang Wang, and Greg Mori. "Discriminative Figure-Centric Models for Joint Action Localization and Recognition." In 2011 International conference on computer vision, IEEE, 2011, 2003-2010.

[20] Huang, Chao, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. "Weakly Supervised Video Anomaly Detection via Self-Guided Temporal Discriminative Transformer." IEEE Transactions on Cybernetics 54, no. 5 (2022): 3197-3210.

[21] Wu, Peng, Xiaotao Liu, and Jing Liu. "Weakly Supervised Audio-Visual Violence Detection." IEEE Transactions on Multimedia 25 (2022): 1674-1685.