Shot-Conditioned Vision-Language Adaptation for Effective Harmful Content Detection from Online Short Videos

Shuai Xu[1,2,*], **Zao Qiu**[1], **Xuelin Zhu**[3] and **Yicong Li**[1]
[1]Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]State Key Laboratory of Internet Architecture (Tsinghua University), Beijing, China
[3]The Hong Kong Polytechnic University, Hong Kong, China
{xushuai7,qiuzao,liyicong}@nuaa.edu.cn, zhuxuelin23@gmail.com

Abstract

Short video harmful content detection aims to automatically identify diverse anomalies from user-generated media. This task presents unique challenges due to frequent editing cuts and highly variable anomaly densities, limiting the effectiveness of traditional surveillance-based approaches. Moreover, existing Vision-Language Model-based approaches typically rely on rigid instance selection mechanisms that fail to adapt to the unpredictable duration of anomalies in such unconstrained videos. To address these issues, we propose SVLA, a **S**hot-conditioned **V**ision-**L**anguage **A**daptation framework, for effectively detecting harmful contents from online short videos. Our approach introduces a novel π-adaptive strategy to dynamically estimate shot-level anomaly density, replacing rigid selection with calibrated supervision. Furthermore, we employ a shot-conditioned temporal encoder to respect video hierarchy and adopt a dual-path contextual adapter to resolve semantic ambiguity. To benchmark this task, we construct a new dataset (SVA) covering more genuine online short videos that involve seven anomaly categories. Experiments on the SVA dataset demonstrate that SVLA can achieve the state-of-the-art performance and outperform its competitors across diverse scenarios. Codes and datasets are available at: https://github.com/xushuai7/IJCAI-SVLA.

1 Introduction

In recent years, detecting harmful content (referred to as anomaly in some literature) from online video platforms has gained widespread attention. Weakly Supervised Video Anomaly Detection (WSVAD) only requires video-level annotations to achieve detection capabilities and it is expected to become a general solution to this task. Most studies in this field follow a standardized pipeline: first, videos are formulated as "bags" composed of temporal snippets in the multi-instance learning (MIL) framework [Sultani *et al.*, 2018] ; then, the 3D-CNN backbone network is used to extract visual representations [Wu *et al.*, 2020; Tian *et al.*, 2021]; in the end, anomalous bags are distinguished from normal ones through ranking-based constraints. Although such uni-modal paradigm has matured, it does not fully utilize the rich semantics from cross-modal data, especially when textual information is available. Over the past few years, Vision-Language Models (VLMs) (such as CLIP [Radford *et al.*, 2021]) have made significant progress in this field, achieving considerable anomaly detection effectiveness through semantic matching [Wu *et al.*, 2024b; Dev *et al.*, 2024]. Given the breakthrough performance of the large-scale pre-trained models, directly aligning visual features with anomaly-related textual descriptions has become an emerging research direction and has been widely applied in downstream visual tasks [Xu *et al.*, 2025].

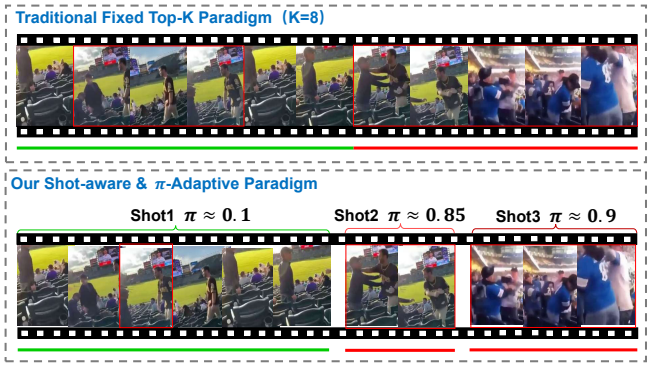

Figure 1: Comparison of instance selection paradigms. Top: The traditional fixed Top-K strategy ignores structural boundaries and density variations. Bottom: Our proposed shot-aware & π-adaptive paradigm dynamically calibrates the selection scope based on shot-level density π, enabling precise detection in both sparse and dense scenarios.

However, existing studies have certain limitations when dealing with the detection of harmful content from online short videos. The methods heavily rely on the fixed priors of surveillance videos, i.e., the assumption of temporal continuity and the assumption of anomaly sparsity. Unlike surveillance videos, user-generated short videos in online platforms like TikTok are usually edited and processed, characterized by rapid shot transitions rather than maintaining temporal co-

* Corresponding author.

herence. If traditional temporal modeling methods are directly applied to such data, the features of different shots will be mixed, resulting in the leakage of temporal information. On the other hand, the "*sparsity assumption*" of surveillance videos is no longer applicable to short videos. The anomaly density of short videos varies greatly, which can be either brief unexpected events or long-lasting abnormal scenes. Therefore, the standard MIL paradigm based on the fixed Top-K strategy may perform poorly: in sparse anomaly scenarios, these methods are prone to incorporating noise information; while in dense anomaly scenarios, they will miss complete abnormal events. Moreover, the unconstrained nature of user-generated videos also brings semantic ambiguity problems, making the static alignment mechanism difficult to meet the requirements.

To address the aforementioned issues, in this paper, we propose SVLA, a novel **S**hot-conditioned **V**ision-**L**anguage **A**daptation framework, for effectively detecting harmful content from online short videos. A key design of SVLA is the shot-conditioned density-adaptive mechanism, which is composed of two modules: the Shot-Conditioned Temporal Encoder (SCTE) and the π-adaptive pooling strategy. SCTE module can clearly define the shot boundaries and hierarchically aggregate features to prevent the leakage of temporal information, while the π-adaptive pooling strategy estimates the anomaly density at the shot level, where the obtained anomaly density π can dynamically adjust the selection range of instances. Figure 1 illustrates the advantage of our instance selection paradigms. If a shot is abnormally dense, we need to expand the range to capture the complete event; while if it is abnormally sparse, we contract it to suppress noise. We can see that our shot-aware and π-adaptive strategy can effectively address the limitations of existing fixed Top-K ranking methods. Moreover, unlike previous studies that only use static alignment, we introduce the **D**ual-path **C**ontextual **A**dapter (DCA), which facilitates fine-grained cross-modal alignment via bilateral interaction: a visual-side adapter is adopted to respect dynamic textual prefixes, and a text-side adapter is used to modulate text prototypes based on global video contexts.

The main contributions of this paper are summarized as follows:

- We propose a novel framework SVLA for effective harmful content detection from online short videos. Unlike surveillance-based approaches, SVLA is specifically tailored to capture the distinct structural hierarchy and variable anomaly density of short video content via a shot-conditioned density-adaptive mechanism.
- The reasoning ability of pre-trained vision-language models is properly leveraged for fine-grained cross-modal alignment. We introduce a dual-path contextual adapter consisting of a visual-side adapter and a text-side adapter to effectively mitigate semantic ambiguity in unconstrained videos.
- For benchmarking the short video anomaly detection task, a new dataset SVA with more short videos in its true sense involving various categories is constructed and partially released (for anonymous review in this stage). Extensive experiments based on SVA dataset validate the effectiveness and superior performance of our approach against the state-of-the-art methods. Besides, it demonstrates competitive robustness on two public surveillance datasets.

2 Related Work

2.1 Weakly Supervised Video Anomaly Detection

Recently, extensive researchers have formulated the WSVAD task under the Multiple Instance Learning (MIL) framework. Sultani et al. [Sultani *et al.*, 2018] firstly propose the deep MIL ranking loss, which considers a video as a bag to distinguish between abnormal and normal ones. Then several follow-up works make efforts to learn more discriminative representations. For example, Tian et al. [Tian *et al.*, 2021] and Chen et al. [Chen *et al.*, 2023] use feature magnitude learning to improve separability. Meanwhile, Zhong et al. [Zhong *et al.*, 2019] and Wu et al. [Wu and Liu, 2021] propose GCN-based methods to model the relationships between snippets, while Cho et al. [Cho *et al.*, 2022] introduce context-motion relation modules. Recognizing the importance of global context, Li et al. [Li *et al.*, 2022] present a Transformer-based architecture, and Huang et al. [Huang *et al.*, 2022] also adopt similar mechanisms to aggregate long-range features. Beyond feature enhancement, some researchers also pay attentions to label noise. For instance, Feng et al. [Feng *et al.*, 2021] and Shi et al. [Shi *et al.*, 2024] propose two-stage self-training frameworks, while other studies [Lv *et al.*, 2021; Zhou *et al.*, 2023; Zhang *et al.*, 2023] present uncertainty-aware schemes.

All the above studies are based on the uni-modal paradigm. However, this paradigm relies on an assumption of temporal continuity. On the contrary, montage-based short videos are characterized by frequent shot cuts, which leads to temporal information leakage when global smoothing is applied.

2.2 Vision-Language Models for VAD

Vision-Language Models (VLMs) such as CLIP have also received significant attention in the field of anomaly detection. Joo et al. [Joo *et al.*, 2023] are the first to use the frozen CLIP encoder as a feature extractor, and subsequent studies have focused on exploring cross-modal correlations. For instance, Wu et al. [Wu *et al.*, 2024b] propose the VadCLIP model to align visual features. Wang et al. [Wang *et al.*, 2025] study the context-aware prompt generator to incorporate local visual context information, and Dev et al. [Dev *et al.*, 2024] explore the reparameterized prompt learning strategy for cross-modal semantic alignment. Apart from studies at the network architecture level, some researchers have also turned their attention to prompt construction. Specifically, Yang et al. [Yang *et al.*, 2023] and Chen et al. [Chen *et al.*, 2024] introduce normality guidance and normal context prompts, while Pu et al. [Pu *et al.*, 2024] utilize external knowledge bases to construct richer semantic information.

Nevertheless, most of these studies are difficult to adapt to online short videos, as they rely solely on the static text-visual alignment while ignoring the dynamic variations of anomaly density. Therefore, applying such rigid selection mechanism

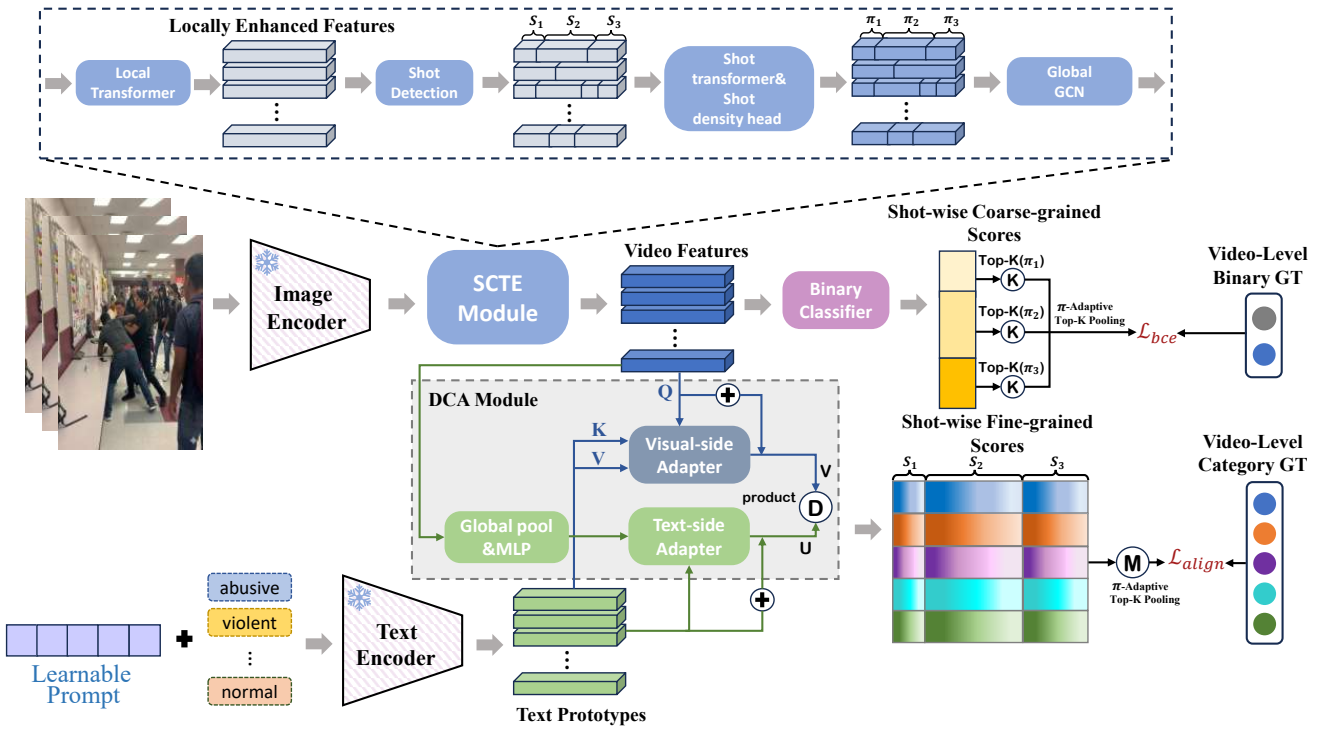

Figure 2: An overview of the proposed SVLA framework.

to short video anomaly detection task will inevitably leads to suboptimal performance.

3 The Approach

Figure 2 displays an overview of the proposed Shot-conditioned Vision-Language Adaptation (SVLA) framework. In the following, we first introduce the preliminaries and then elaborate on the key modules of SVLA.

3.1 Preliminaries

Problem Formulation

Given a set of N training videos $\mathcal{V} = \{V_i\}_{i=1}^{N}$, if all frames of a video do not contain abnormal events (e.g., pornography, violence, blood, etc), this video is defined as normal with the label $y_i = 0$; Otherwise, if there is at least one frame contains abnormal events, this video is labeled as abnormal with the label $y_i = 1$. The goal of short video anomaly detection task is to train a detection model that is able to predict frame-level anomaly scores for unseen videos while only video-level labels are provided.

CLIP-based Embeddings for Image and Text Encoders

The cross-modal reasoning ability of CLIP has revolutionized many tasks in recent years. In this work, we use the frozen visual encoder of CLIP as the backbone to extract snippet-level visual features (as an input of the SCTE module), and utilize the text encoder of CLIP to take full advantage of the semantic association between visual contents and textual concepts. Following the way adopted by [Zhou *et al.*, 2022b], we concatenate learnable context vectors with the category token embeddings and feed them into the frozen text encoder to generate the text prototypes, which are further encoded by the DCA module.

3.2 Shot-Conditioned Temporal Encoder

Standard encoders tend to treat videos as flat sequences. However, we argue that such a paradigm overlooks the hierarchical structure in short videos. In this study, we made an effort to explicitly model the transition from fine-grained frame dynamics to coarse-grained shot semantics. Meanwhile, we also paid specific attentions to preventing information leakage across scene cuts and capturing global dependencies [Wu and Liu, 2021; Zhou *et al.*, 2023]. Taking these factors into account, we devise the shot-conditioned temporal encoder (SCTE).

Local Temporal Modeling and Shot Detection

To begin with, the local transformer is deployed to capture short-range dependencies. It is worth noting that standard global attention inevitably leads to leakage; consequently, we make use of a restricted window mask [Liu *et al.*, 2021]. Following this line, we structuralize the video by simply identifying boundaries where adjacent frame similarity falls below τ_{shot}, thus naturally partitioning it into M shots. In this branch, we opt to aggregate features to represent these units. The representation h_j for the j-th shot is presented as follows:

$$h_j = \phi_{shot}\left(\frac{1}{e_j - s_j}\sum_{t=s_j}^{e_j-1} x_t\right) \tag{1}$$

where a learnable projection ϕ_{shot} is applied to map the aggregated features into the embedding space.

Shot-Conditioned Density Estimation

In the density branch, we make an effort to capture variable anomaly densities via a lightweight head within the SCTE module. Concretely, we process the shot representations via a shot transformer to model inter-shot dependencies. Then, the anomaly density $\pi_j \in (\pi_{min}, 1)$ for the j-th shot is derived as follows:

$$\pi_j = \text{clamp}(\sigma(\phi_\pi(h_j)), \pi_{min}, 1-\epsilon) \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid activation function. Crucially, we simply introduce a learnable density floor π_{min}, thus naturally preventing the supervision signal from collapsing. With the estimated π_j in hand, we use it as a dynamic prior to guide the subsequent adaptive pooling.

Structure-Aware Graph Convolution

Having obtained density priors, we turn our attention to modeling global contexts via a dual-stream GCN [Zhong *et al.*, 2019; Wu *et al.*, 2020]. For the structure-aware stream, we simply reinforce edges using a binary shot mask $\mathbb{M}^{shot}$, which naturally restricts message passing across cuts. In parallel, we also made an effort to model local proximity via a distance-based matrix $\mathbf{A}^{dist}$. Ultimately, we concatenate these outputs to yield the refined features $\hat{\mathbf{F}}$.

Coarse-grained Binary Scoring

In order to capture generic anomaly patterns directly from the visual modality, we construct a lightweight binary branch leveraging the structurally refined features $\hat{\mathbf{F}}$. Concretely, this branch is implemented as a residual MLP cascaded with a sigmoid classifier. Thus, the frame-level anomaly probabilities $\mathbf{A}_{bin} \in \mathbb{R}^{T\times 1}$ are obtained. Such an implementation allows the model to provide the coarse-grained supervision signal $\mathcal{L}_{bce}$. With the scores $\mathbf{A}_{bin}$ in hands, we facilitate visual context extraction in subsequent modules.

3.3 Dual-path Contextual Adapter

To bridge the semantic gap between visual and textual features, we propose the Dual-path Contextual Adapter (DCA). In this module, we employ a bidirectional interaction mechanism, namely, the visual branch leverages fine-grained textual semantics while the text branch modulates prototype features under the guidance of global video context. This design allows us to achieve co-optimization, which naturally ensures robust modal alignment.

Visual-side Adapter mechanism

We inject fine-grained semantics using a cross-modal attention module. Concretely, we simply use $\hat{\mathbf{F}}$ as queries, while the concatenation of dynamic prefixes [Ghadiya *et al.*, 2024; Zhou *et al.*, 2022a] and prototypes serves as keys and values. Then we obtain the attention output $\mathbf{H}_{att}$. We derive a modulation gate $\mathbf{G}$ via context-aware similarity to filter irrelevant associations. Then, we update the features $\mathbf{F}_{cfa}$ as follows:

$$\mathbf{F}_{cfa} = \text{LN}\left(\mathbf{W}_{fuse}\left(\hat{\mathbf{F}} + \beta \cdot (\text{MLP}(\mathbf{H}_{att}) \odot \mathbf{G})\right)\right) \tag{3}$$

where $\odot$ represents element-wise multiplication. We use β to control the injection intensity.

Text-side Adapter mechanism

Since standard text prompts inevitably fail to cover context-dependent variations, we propose a global modulation mechanism. Concretely, the static text prototypes are first projected into the visual feature space to obtain $\mathbf{P}_{base}$. Then we aggregate the visual features $\hat{\mathbf{F}}$ into a global context vector $\mathbf{g}$ via mean pooling, and generate affine scale and shift parameters. Finally, the context-aware text prototypes $\mathbf{P}_{cfa}$ are modulated, which is presented as follows:

$$\mathbf{P}_{cfa} = \text{LN}\left(\mathbf{P}_{base} \odot (1 + \tanh(\boldsymbol{\gamma}(\mathbf{g}))) + \boldsymbol{\beta}(\mathbf{g})\right) \tag{4}$$

In this formulation, we apply the $\tanh$ function to regulate the scale parameter, and use LN for LayerNorm. With the dynamic text prototypes in hand, the embedding space aligns with the specific visual characteristics of the input video.

Dual-path Similarity Scoring

With the aim of balancing the stability of static priors and the specificity of adapted contexts, we propose to fuse the original and refined representations. The final visual features $\mathbf{V}$ and text prototypes $\mathbf{U}$ are obtained via two learnable gates λ_v, λ_t:

$$\mathbf{V} = (1-\lambda_v)\hat{\mathbf{F}} + \lambda_v \mathbf{F}cfa, \quad \mathbf{U} = (1-\lambda_t)\mathbf{P}base + \lambda_t \mathbf{P}_{cfa} \tag{5}$$

where both gates are restricted in the range $(0, 1)$. Then the frame-level anomaly scores $\mathbf{S}$ are computed by the temperature-scaled cosine similarity between $\mathbf{V}$ and $\mathbf{U}$. Finally, these scores are used as the multi-class logits for the alignment loss.

3.4 π-Adaptive Top-K Pooling and Training

In standard MIL approaches, a fixed Top-K [Sultani *et al.*, 2018; Tian *et al.*, 2021] selection strategy is typically employed. However, this static design inevitably falls short for short video anomaly detection as anomaly density varies drastically, ranging from fleeting accidents to continuous fighting scenes. To remedy this, we propose a π-adaptive Top-K pooling strategy, where the supervision signal is dynamically calibrated based on the estimated shot density. Then the model is updated via a joint optimization objective.

Density-Aware Instance Selection

Due to the varying density of the video, using a fixed set of hyperparameters is not applicable to diverse video scenarios. Toward this issue, we propose a dynamic selection strategy. Specifically, the method takes the predicted abnormal density π_j and the shot duration L_j as inputs, and calculates the number of instances k_j to be selected, which is defined as follows:

$$k_j = \text{Round}\left(\frac{L_j}{\delta} \cdot (1 + \eta(\pi_j - \pi_0))\right) \tag{6}$$

In the above formula, we set δ as a reference divisor, and η as the scaling factor. A higher density π_j and a longer shot duration L_j mean that we tend to select more instances in dense shots and vice versa.

Density-weighted Aggregation

For global anomalies, we aggregate frame-level predictions. Specifically, we introduce a density-weighted mechanism. In practice, we weigh the contribution of each shot based on its estimated density π_j, and the final score is presented as:

$$\mathcal{S}_{video} = \sum_{j=1}^{M} \frac{\pi_j}{\sum_{m=1}^{M} \pi_m} \cdot \mathcal{S}_{shot}^{(j)} \tag{7}$$

where $\mathcal{S}_{shot}^{(j)}$ refers to the average score of the top-k_j instances. A higher density π_j means that anomalous shots are prioritized. On the contrary, noise accumulation is suppressed in sparse ones.

Joint Optimization Objective

For the coarse-grained binary branch, we utilize the binary cross entropy mechanism to measure the deviation between video-level predictions and the ground-truth. Then we use the binary cross entropy between predicted probabilities $\mathbf{A}_{bin}$ and label y to compute the classification loss $\mathcal{L}_{bce}$. The objective is formulated as follows,

$$\mathcal{L}_{bce} = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \log(\mathcal{S}_{bin}^{(i)}) + (1 - y_i) \log(1 - \mathcal{S}_{bin}^{(i)}) \Big] \tag{8}$$

For the fine-grained semantic branch, we utilize a MIL-Align mechanism which is similar to vanilla MIL. We hope the video and its paired textual label emit the highest similarity score among others. To achieve this, the alignment loss $\mathcal{L}_{align}$ is firstly computed as follows,

$$\mathcal{L}_{align} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \left(\frac{\exp(\mathcal{S}_c^{(i)})}{\sum_k \exp(\mathcal{S}_k^{(i)})} \right) \tag{9}$$

In addition to classification loss $\mathcal{L}_{bce}$ and alignment loss $\mathcal{L}_{align}$, we also introduce a contrastive loss to slightly push the normal prompt away. Here we first calculate the cosine similarity between normal class embedding $\mathbf{u}_{normal}$ and other abnormal class embeddings $\mathbf{u}_c$, and then compute the regularization loss $\mathcal{L}_{reg}$ as follows,

$$\mathcal{L}_{reg} = \frac{1}{C-1} \sum_{c=2}^{C} |\cos(\mathbf{u}_{normal}, \mathbf{u}_c)| \tag{10}$$

Based on the above constraints, the final total objective is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{bce} + \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{reg} \tag{11}$$

where λ_1 and λ_2 refer to the temperature hyper-parameters for scaling.

3.5 Inference

In the inference phase, SVLA leverages its dual-branch architecture to generate anomaly scores. We compute frame-level probabilities from both the binary classification branch and the semantic alignment branch (derived as $1 - p_{normal}$) [Wu *et al.*, 2022; Wu *et al.*, 2024b], then we select the optimal branch based on validation performance to obtain the results.

4 Experiments

4.1 Datasets and Metrics

Datasets

A new dataset **SVA** that broadly reflects user-generated content characteristics is introduced to address the limitations of static surveillance benchmarks. Collected from online short video platforms like TikTok and YouTube Shorts, SVA contains 6,687 short videos with an average duration of 7.7 seconds. This dataset covers seven distinct anomaly categories alongside normal content, featuring highly dynamic backgrounds, frequent shot cuts, and variable anomaly densities. Besides, **UCF-Crime** [Sultani *et al.*, 2018] and **XD-Violence** [Wu *et al.*, 2020] are also utilized to verify the performance on long-range surveillance scenarios. Table 1 summarizes the statistics of the involved datasets.

Datasets	Source	# Videos	Avg. Len	# Anomaly Types
UCF-Crime [Sultani *et al.*, 2018]	Surveillance	1,900	∼4.0 min	13
XD-Violence [Wu *et al.*, 2020]	Movies/CCTV	4,754	∼2.7 min	6
SVA (Ours)	Online short videos	6,687	∼7.7 sec	7

Table 1: Statistics of the datasets.

Evaluation Metrics

Following previous works [Sultani *et al.*, 2018; Wu *et al.*, 2020], we report the standard Area Under the Curve (AUC) metic values of the frame-level ROC for UCF-Crime dataset. For XD-Violence dataset, we report the Average Precision (AP) metric values. For our proposed SVA dataset, we adopt both AUC and AP metrics to comprehensively evaluate the detection performance.

4.2 Implementation Details

We adopt the frozen CLIP (ViT-B/16) [Radford *et al.*, 2021] as the backbone, projecting visual features to $D = 512$. Inside SCTE, we simply set the shot similarity threshold τ_{shot} as 0.90. We also restrict the minimum shot length to 12 frames. For the π-adaptive strategy, the density floor π_{floor} and reference density π_0 are set as 0.05 and 0.15. For Top-K pooling, we set the base k as $T/16$, where T is the number of total frames. For model training, we use AdamW as the optimizer with batch size of 64. The base learning rate and total epochs are set as 2×10^{-5} and 10, respectively. On the shot density head, we apply a larger learning rate multiplier of 5.0. For the Dual-path Contextual Adapter, the fusion factor β defaults to 0.8 with a bottleneck dimension of 256. All experiments are implemented on a single machine (vGPU-32GB).

4.3 Comparisons with the SOTA Approaches

Table 2 summarizes the experimental results on the proposed SVA dataset and the other two benchmark datasets. On the newly constructed SVA dataset, the proposed approach SVLA achieves an AUC of 95.91% and an AP of 96.52%, outperforming other mainstream methods including the state-of-the-art approaches. The performance advantage is attributed to the precise adaptation of SVLA to the unique attributes of user-generated short videos. Unlike surveillance videos with continuous temporal streams, short videos

Methods	Sources	SVA Dataset		UCF-Crime	XD-Violence
		AUC (%)	AP (%)	AUC (%)	AP (%)
Deep-MIL [Sultani *et al.*, 2018]	CVPR'18	80.06	82.68	75.41	75.18
HL-Net [Wu *et al.*, 2020]	ECCV'20	90.24	91.46	82.44	80.00
RTFM [Tian *et al.*, 2021]	ICCV'21	85.64	87.92	84.30	78.27
MSL [Li *et al.*, 2022]	AAAI'22	90.68	91.84	85.62	78.28
CU-Net [Cho *et al.*, 2022]	CVPR'22	91.04	92.59	86.22	81.31
MGFN [Chen *et al.*, 2023]	AAAI'23	92.47	93.10	86.98	79.19
CU-VAD [Zhang *et al.*, 2023]	CVPR'23	91.84	93.27	86.22	78.74
CLIP-TSA [Joo *et al.*, 2023]	ICIP'23	93.77	94.71	87.58	82.17
OVVAD [Wu *et al.*, 2024a]	CVPR'24	92.89	93.46	86.40	66.53
OE-CTST [Majhi *et al.*, 2024]	WACV'24	92.46	93.65	86.99	81.78
SAA [Fan *et al.*, 2024]	TCSVT'24	92.31	93.87	86.19	84.24
ARMS [Shi *et al.*, 2024]	TMM'24	92.23	93.89	85.79	83.11
TPWNG [Yang *et al.*, 2024]	CVPR'24	93.36	94.27	87.79	83.68
VadCLIP [Wu *et al.*, 2024b]	AAAI'24	91.47	93.41	**88.02**	84.51
DE-Net [Zhang *et al.*, 2025]	TNNLS'25	92.34	93.58	87.86	**86.13**
IFS-VAD [Zhong *et al.*, 2025]	TCSVT'25	93.15	94.23	86.57	83.14
Anomize [Wu *et al.*, 2024a]	CVPR'25	93.46	94.74	87.05	84.24
MEL-VLP [Huang *et al.*, 2025]	TMM'25	94.33	95.21	87.45	83.44
SVLA (Ours)		**95.91**	**96.52**	87.15	85.44

Table 2: Experimental results on the selected datasets, where the **bold** value in each column represents the best performance.

in SVA dataset have an average duration of only 7.7 seconds with numerous editing and switching operations. Traditional methods often smooth the features across edits, thereby causing information leakage problems. In contrast, the shot-condition temporal encoder (SCTE) proposed in this paper can explicitly model this hierarchical structure and avoid feature confusion between unrelated shots. The abnormal density of short videos is usually high, which makes traditional methods that based on the "sparsity assumption" less effective. Besides, the π-adaptive strategy can dynamically expand the instance selection range based on the density estimation results, thereby achieving complete capture of abnormal events in scenarios where rigid selection methods fail.

Beyond the SVA benchmark for short video anomaly detection, SVLA also demonstrates robust generalization on public datasets: the AUC score on UCF-Crime reaching 87.15%, and the AP score on XD-Violence reaching 85.44%. It is worth noting that although SVLA outperforms classic baselines like RTFM, it has not yet fully surpassed the state-of-the-art models specifically tailored for these benchmarks. For UCF-Crime, the dataset predominantly consists of continuous surveillance footage with a fixed perspective and minimal editing. In this scenario, the strict partitioning mechanism in our SCTE module faces a trade-off. Its design rationale is to prevent "cross-segment leakage", which may inadvertently incur over-segmentation in continuous streams. SCTE might fracture the originally continuous long-term context, which is crucial for identifying slowly evolving anomalies (such as a theft incident). In contrast, the traditional global Transformer, which is not constrained by this structure, shall possess a slight advantage in processing continuous streams. For XD-Violence, although it contains some editing, it is generally recognized as long-duration videos (with an average duration of several minutes), indicating that its anomaly density is significantly lower than that of SVA dataset. The proposed π-adaptive strategy is specifically tailored for the high density variability of short videos. In extremely sparse long videos, our shot-level density estimation may be less sensitive than methods specialized in background suppression.

	SCTE	π-Adaptive	DCA_{vis}	DCA_{txt}	AUC (%)	AP (%)
1	✗	✗	✗	✗	91.47	93.41
2	✗	✗	✗	✓	92.88	94.21
3	✗	✗	✓	✗	93.21	94.38
4	✗	✗	✓	✓	94.09	95.03
5	✓	✗	✓	✓	95.12	95.77
6	✓	✓	✓	✓	**95.91**	**96.52**

Table 3: Ablation studies on the SVA dataset. We progressively integrate the Shot-Conditioned Temporal Encoder (SCTE), the π-Adaptive Strategy, and the two components of the Dual-path Contextual Adapter (DCA) to validate the effectiveness of each module.

4.4 Ablation Studies

We conduct extensive ablation studies on the proposed SVA dataset to dissect the contribution of each core component in SVLA. We adopt a re-implemented VadCLIP [Wu *et al.*, 2024b] as our baseline , which yields an AUC of 91.47% and AP of 93.41% using standard temporal modeling and fixed Top-K selection.

Effectiveness of Dual-path Contextual Adapter

We start by checking the cross-modal alignment. Table 3 reveals that the Text-side and Visual-side Adapters, acting on

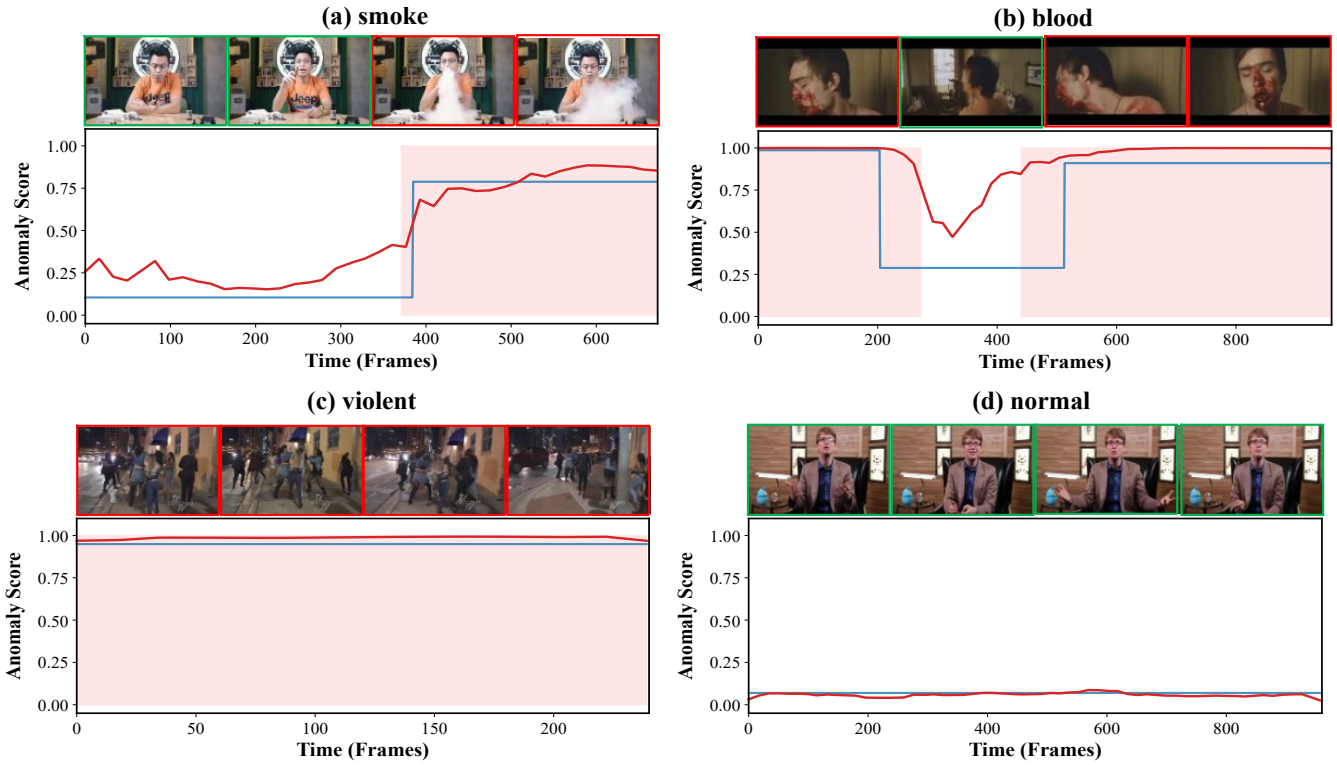

Figure 3: Case studies of SVLA across different short videos selected from SVA dataset. The red curve represents the SVLA prediction score, the blue curve indicates the shot density π, and the pink area shows the anomaly duration, respectively.

their own, lift the AUC to 92.88% and 93.21% respectively. Once working together, the full DCA module further pushes the metric to 94.09%—marking a solid 2.62% jump over the baseline. This results underscore the merit of bilateral adaptation in the DCA module. By tuning text prototypes and creating dynamic visual prefixes, we manage to clear up semantic ambiguity in tricky short-video scenarios where single-modal approaches often fail.

Effectiveness of Shot-Conditioned Temporal Encoder
Building on top of DCA, we further integrate the SCTE module to explicitly enforce the video structure. This addition spikes the performance to 95.12% regarding AUC score. It is clear that strictly walling off shot boundaries can prevent features from bleeding into unrelated scenes, meaning that SCTE module works effectively for chaotic user-generated short videos riddled with sharp cuts.

Effectiveness of π-Adaptive Strategy
Ultimately, integrating this π-adaptive supervision into our framework yields a peak performance of 95.91% w.r.t. AUC metric and 96.52% w.r.t. AP metric. By substituting the rigid Top-K with density-aware calibration, SVLA learns to align with anomaly events of varying durations. This strategy effectively cuts down false negatives in dense anomaly shots while suppressing noise in sparse ones.

4.5 Case Studies

We illustrate typical cases of SVLA detection results in Figure 3, covering three representative anomaly categories (*smoke*, *blood*, *violent*) and one *normal* sample selected from the SVA dataset. We plot the SVLA prediction score (red) and the shot density π (blue) against the ground truth area (pink). As observed in the *smoke* and *violent* cases (a and c), SVLA delivers rapid responsiveness, with scores rising sharply in synchronization with the emergence of shots containing anomaly contents. Notably, in the *blood* case (b), the predicted shot density π dynamically adapts to the video structure, showing a high correlation with the disjointed anomaly segments. Furthermore, in the *normal* case (d), both the anomaly scores and π remain consistently low despite visual complexity, verifying the robustness of SVLA against false alarms. The results demonstrate that our shot-conditioned strategy can effectively adapt to various anomaly scenarios and variable video durations while maintaining high robustness for videos without anomaly contents.

5 Conclusion

In this paper, we propose a novel framework based on shot-conditioned vision-language adaptation for harmful content detection from online short videos. To efficiently handle variable anomaly densities and frequent editing cuts, we devise a π-adaptive strategy to dynamically calibrate supervision signals, and then design mechanisms to improve the alignment between visual hierarchy and textual semantics. Extensive experimental results on multiple datasets validate the effectiveness and robustness of the proposed approach for various detection scenarios.

Acknowledgments

This work is supported by the Natural Science Foundation of China under Grant No. 62302213 and No. 62502204.

References

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

[Chen *et al.*, 2024] Junxi Chen, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18319–18329, 2024.

[Cho *et al.*, 2022] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12137–12146, 2022.

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

[Fan *et al.*, 2024] Yidan Fan, Yongxin Yu, Wenhuan Lu, and Yahong Han. Weakly-supervised video anomaly detection with snippet anomalous attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5480–5492, 2024.

[Feng *et al.*, 2021] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021.

[Ghadiya *et al.*, 2024] Ayush Ghadiya, Purbayan Kar, Vishal Chudasama, and Pankaj Wasnik. Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1965–1974, 2024.

[Huang *et al.*, 2022] Chao Huang, Yabo Liu, Jie Wen, Yong Xu, Qiang Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Transactions on Cybernetics*, 54(5):3197–3210, 2022.

[Huang *et al.*, 2025] Chao Huang, Weiliang Huang, Qiuping Jiang, Wei Wang, Jie Wen, and Bob Zhang. Multimodal evidential learning for open-world weakly-supervised video anomaly detection. *IEEE Transactions on Multimedia*, 2025.

[Joo *et al.*, 2023] Hyukhee Joo, Kibaek Ko, K Yamazaki, and N Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE, 2023.

[Li *et al.*, 2022] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1395–1403, 2022.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Lv *et al.*, 2021] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2021.

[Majhi *et al.*, 2024] Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Oe-ctst: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 8574–8583, 2024.

[Pu *et al.*, 2024] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Multimedia*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Shi *et al.*, 2024] Haoyue Shi, Le Wang, Sanping Zhou, Gang Hua, and Wei Tang. Abnormal ratios guided multi-phase self-training for weakly-supervised video anomaly detection. *IEEE Transactions on Multimedia*, 26:5575–5586, 2024.

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

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

[Wang *et al.*, 2025] Benfeng Wang, Chao Huang, Jie Wen, Wei Wang, Yabo Liu, and Yong Xu. Federated weakly supervised video anomaly detection with multimodal prompt. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. AAAI-25.

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

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

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

[Wu *et al.*, 2024a] Peng Wu, 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*, pages 18297–18307, 2024.

[Wu *et al.*, 2024b] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6074–6082, 2024.

[Xu *et al.*, 2025] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20370–20382, 2025.

[Yang *et al.*, 2023] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18899–18909, 2023.

[Yang *et al.*, 2024] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18899–18908, June 2024.

[Zhang *et al.*, 2023] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280, 2023.

[Zhang *et al.*, 2025] Chen Zhang, Guorong Li, Yuankai Qi, Hanhua Ye, Laiyun Qing, Ming-Hsuan Yang, and Qingming Huang. Dynamic erasing network with adaptive temporal modeling for weakly supervised video anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[Zhong *et al.*, 2019] Jia-Xing Zhong, 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*, pages 1237–1246, 2019.

[Zhong *et al.*, 2025] Yuanhong Zhong, Ruyue Zhu, Ge Yan, Ping Gan, Xuerui Shen, and Dong Zhu. Inter-clip feature similarity based weakly supervised video anomaly detection via multi-scale temporal mlp. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1961–1970, 2025.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[Zhou *et al.*, 2023] Hang Zhou, 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*, volume 37, pages 3769–3777, 2023.