Video Anomaly Detection Using Multi-level Text Prompts and Context-Aware Frame Sampling for Long Duration Videos

Anonymous WACV Algorithms Track submission

Paper ID *****

Abstract

Weakly supervised video anomaly detection (WSVAD) is an essential yet challenging problem, particularly for long-duration surveillance videos. Recent approaches have leveraged contrastive language-image pretraining (CLIP) to enhance frame-level feature representations, yielding significant improvements. However, most existing methods overlook the challenges associated with the duration of the videos. These methods employ uniform or stepwise frame sampling to manage video length, which risks discarding crucial fine-grained temporal information and leads to information bottlenecks. In this paper, we propose MLP-VAD, a novel framework that addresses these limitations through text-guided frame selection using a frozen Image-Text Matching CLIP model. By leveraging class-level semantic descriptions, our approach adaptively filters frames relevant to downstream anomaly detection tasks. We also introduce multi-level prompt-guided feature learning integrating entity-level and scene-level prompts to mimic human cognitive processes more closely in anomaly detection task. Extensive experiments on the UCF-Crime and XD-Violence benchmarks demonstrate that MLPVAD surpasses current state-of-the-art methods, achieving an AUC of 88.84% and an AP of 85.37%, respectively.

1. Introduction

Weakly supervised video anomaly detection (WSVAD) has recently gained significant attention in the fields of video understanding and computer vision. The primary objective of WSVAD is to utilize video-level annotations to detect abnormal activities within video frames, with applications spanning surveillance systems [2], traffic monitoring [1], and beyond. Given the scarcity of frame-level annotations, video-level supervision has become prevalent in WSVAD research.

Existing approaches typically employ backbone networks such as I3D/C3D [5] or transformer-based models like ViT [10] to extract frame-level representations. These features are processed with temporal modeling within a Multiple Instance Learning (MIL) framework [11] to predict normality or abnormality scores. Although effective on several datasets, these approaches mainly rely on single-domain video data, which limits their generalizability in complex scenarios where distinctions between normal and abnormal events are very subtle [7, 13]. The vision-language models have emerged as a promising direction for video anomaly detection [4, 6, 26]. By leveraging the CLIP model [19], these methods obtain rich, generalized visual representations aligned with textual descriptions via contrastive learning on large-scale image-text datasets. Incorporating CLIP as a feature extractor enhances contextual understanding of challenging scenes and significantly improves performance compared to unimodal approaches. Despite these advances, several challenges remain:

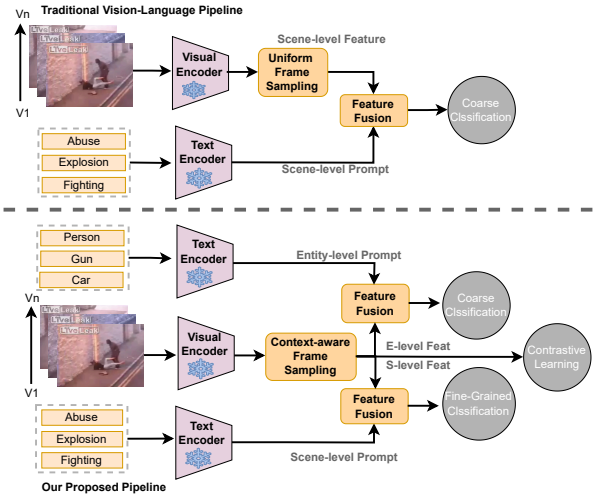

Figure 1. Traditional multimodal VAD vs. our proposed MLP-VAD paradigm.

First, most existing methods are designed for short to medium-length videos. Processing long-duration videos

(e.g., over 30 minutes at 30 fps, containing upwards of 60,000 frames) necessitates frame sampling strategies, commonly uniform or stepwise, to reduce input size to manageable lengths (e.g., 128–256 frames). Note that the coarse sampling induces temporal noise, disregards important fine-grained frame information, and risks excluding critical frames, thereby impairing overall performance.

Second, current multimodal WSVAD methods predominantly utilize coarse class-level textual prompts [26] attended over aggregated video-level features, overlooking fine-grained entity-level information and frame-wise interactions. This coarse granularity limits their ability to capture subtle anomaly cues in the frames.

Third, effective discrimination between normal and abnormal scenes with subtle visual differences remains underexplored, yet is essential for realistic applications.

To address these challenges, we propose a novel paradigm for VAD that emulates human cognitive processes. Our model, **MLPVAD**, illustrated in Figure 1, incorporates the following key components:

- **Context-Aware Frame Sampling (CAFS):** To efficiently process long videos, CAFS selects a compact subset of informative frames by leveraging the CLIP-based pre-trained model *BLIP2* [15]. Through frame-wise image-text matching, CAFS filters redundant frames while retaining those most representative of anomalies. Here, text prompts are textual summaries comprising class-specific entities and actions.
- **Multi-level Text Prompt Learning:** To accurately understand a scene, it is necessary to focus on each object in the frame as well as their interactions over time. Entity-level feature learning emphasizes class entities in frames (e.g., *Fight: person, fist, kick, face*; *RoadAccident: car, truck, collision, fire*). This enables fine-grained frame-level attention through entity prompts, concentrating on object relations within the local context. Scene-guided learning further addresses entity interactions over time by applying attention over the summarized video features and scene prompts. This hierarchical approach improves robustness in complex, long-duration scenarios.
- **Contrastive Representation Learning:** To improve discrimination between subtly different classes, a contrastive learning branch leverages multiple positive and negative samples, encouraging anomaly representations to align closely with queries while separating normal representations, thereby complementing classification objectives.

By closely modeling human visual reasoning and exploiting advanced vision-language representations, **MLPVAD** effectively addresses the challenges above, demonstrating superior performance on large-scale benchmarks.

The main contributions of this work are:

1. A novel VAD framework that emulates human cognitive processes, featuring a context-aware frame sampling (CAFS) module optimized for long-duration videos by leveraging large language models (LLMs).
2. A *Multi-level Text Prompt Guided Network* enhancing contextual understanding at both entity and scene levels, supported by contrastive learning for discriminative coarse- and fine-grained classification.
3. Empirical validation of the **MLPVAD** model on two large-scale benchmark datasets, achieving state-of-the-art performance with an AUC of 88.84% and an AP of 85.37% on UCF-Crime and XD-Violence, respectively.

2. Related Works

2.1. Weakly Supervised Video Anomaly Detection:

Leveraging video-level labels, weakly supervised methods have made significant improvements over the past few years in the video anomaly detection task. **WSVAD** was first introduced by Sultani et al. [20] using an MIL approach for surveillance video and outperformed previous unsupervised models. Since then, numerous advancements have been made for WSVAD, improving the temporal network and better feature learning with transformer and CLIP-based pretrained models. To better capture the temporal dynamics across video frames, Tian et al. [21] used a self-attention network. Zhong et al. reformulated the WSVAD problem as a supervised learning task using a graph convolutional network (GCN) method, generating noisy labels from an off-the-shelf video [30]. Feng et al. attempted to remove noisy labels with a multiple instance pseudo-label generator that produces more reliable pseudo-labels for fine-tuning a task-specific feature encoder [11]. More recently, the CLIP model has also attracted great attention in the VAD community [26]. Based on the visual features of the CLIP model, Lv et al [17] proposed a new MIL framework: Unbiased MIL (UMIL), to learn unbiased anomaly features that improve WSVAD performance. Furthermore, Wu et al. utilize a learnable text prompt that leverages the frozen CLIP model, performing both coarse-grained and fine-grained classification at the video level and class level [26]. Most of these weakly supervised methods perform classification using video labels and do not focus on the entity that plays a crucial role in the anomaly activities. This performance is suboptimal and does not reflect real-world understanding of challenging anomaly activities.

2.2. Large Vision-Language Models:

Multimodal training has gained popularity due to its effectiveness in extracting rich, context-aware features from multimodal data. The application of multimodal training began with the most straightforward tasks, such as image classification, and expanded to more challenging tasks, including video understanding and VAD. One of the first works to include visual and textual information for the VAD task was

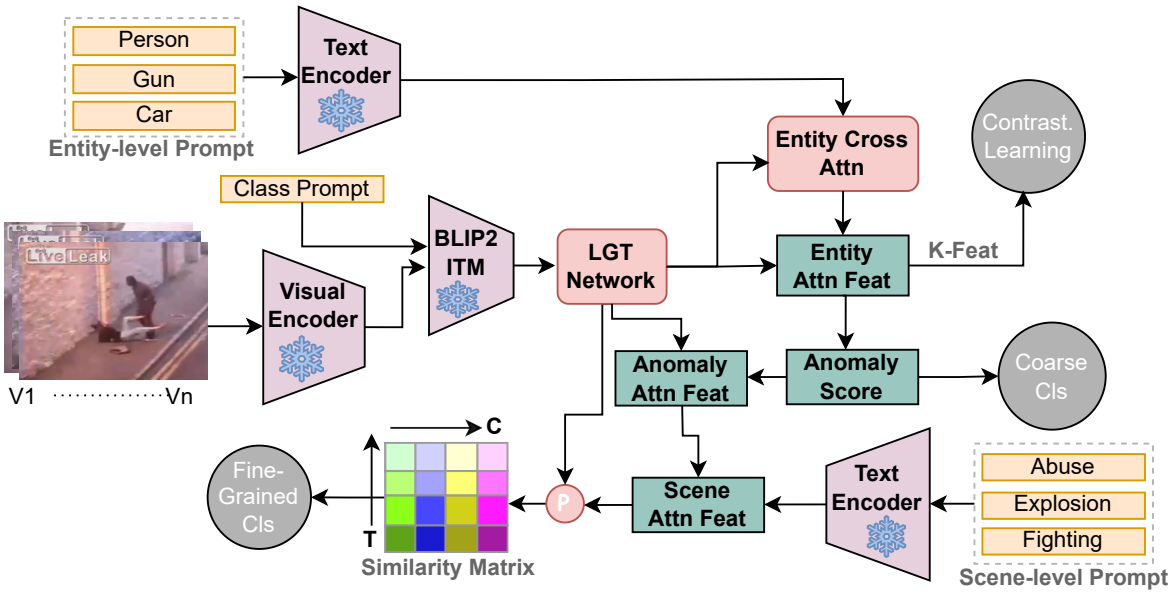

Figure 2. Proposed MLPVAD Architecture with each major component and objective functions.

introduced by Chen et al. [6]. In this work, the authors generated text captions from video frames using the Swin-BERT pretrained model and later performed classification on multimodal features. Later on, the CLIP model demonstrated significant progress in capturing text-vision relations and has been effective for several downstream tasks, such as classification [28], visual question-answering (VQA) [15], semantic segmentation [16], and others. The CLIP model has recently been effectively extended from static image understanding to the spatiotemporal domain of video, enabling joint modeling of visual and textual information over time. VideoCLIP [27] is proposed to align video and textual representations by contrasting temporally overlapping video-text pairs with hard negatives mining. Azad et al. [3] employed hierarchical context learning using multi-scale text prompts for better scene understanding. MA-LMM [12] also leverages the CLIP model for the video-understanding task and incorporates the memory-bank technique to retain long-term temporal information. Next, Zanella et al. [29] and Wu et al. [26] use learnable textual features from CLIP to enhance the representation of the overall features, followed by an MIL-based video anomaly classifier. Existing approaches rarely leverage CLIP models to sample a large number of video frames based on the CLIP text features. Instead, most prior work incorporates language information only at a coarse level for representation learning, primarily focusing on improving classification performance. There is a pressing need to investigate the power of the CLIP model and fully leverage its capabilities in language representation, as well as its effectiveness in the VAD pipeline.

3. Methodology

In this section, we first present the problem definition of the WSVAD task. Next, we briefly introduce the overall architecture of the proposed MLPVAD method. Next, we subsequently elaborate on each of the modules of the MLPVAD method and its role in solving different challenges discussed in earlier sections.

3.1. Problem Definition and Overall Architecture

The goal of this work is to perform the video anomaly detection task in a weakly supervised manner using video-level labels. For example, given a video v, if all the frames of the particular video are normal, the video is annotated as a normal video ($y = 0$). On the other hand, if even a few frames from video v contain abnormal frames, the complete video is annotated as an abnormal video ($y = 1$). The task is to perform binary classification and predict the $\hat{y}$ for a given video v, with annotation $y = 0/1$. The recent successful approaches for the WSVAD task mostly used the MIL approach on frames to predict abnormality. In the MIL approach, previous works use CNN methods such as I3D [5] and C3D [22] to extract consecutive features, which are grouped to create video snippets for the binary classification task. This research focuses equally on the visual and textual information and leverages the power of Large-Language Models (LLMs) to accompany the visual pipeline. For feature extraction, we utilize the encoder of CLIP as the backbone due to its significant performance over the past few years across a wide range of downstream tasks such as VQA [18], VAD [26], Video Understanding

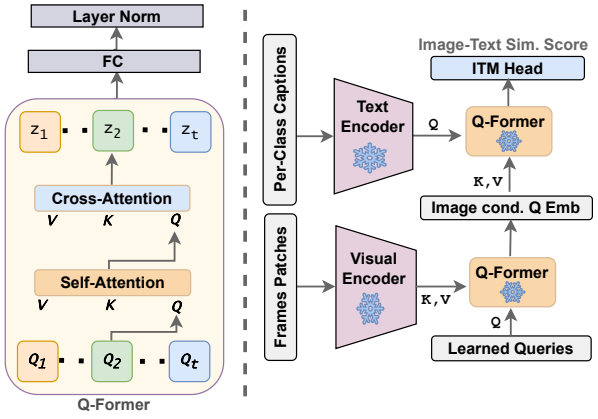

Figure 3. Context-Aware frame sampling (CAFS) with Q-Former module.

[12]. We have three objective functions (See Figure 2): 1) Coarse Classification on entity-attended visual features, 2) Randomized Contrastive Learning on top-k/bottom-k features, and 3) Fine-grained classification from the similarity matrix generated from scene-attended visual features.

3.2. Context Aware Frame Sampling (CAFS)

The current video understanding and VAD methods are prone to information loss and sometimes noisy due to the selection of irrelevant frames through uniform or stepwise frame selection. For the first time, we propose a CAFS method as illustrated in Figure 3, to select frames related to the downstream task. For a video v, we pass it to the CLIP visual encoder (ViT16/ViT14) and extract the visual feature $X \in B \times N \times d_v$. Here, B is the batch size, N is the number of frames (CLS embedding), and d_v is the visual feature dimension. The text-prompt is derived by randomly generating dense captions from several videos per class using VideoLLaMA2 [8]. The robustness of modern LLM-based video models for the caption generation task across different domains is demonstrated in several recent benchmark works [8, 23, 24]. Our experiment also found that the captions generated from VideoLLaMA2 are satisfactorily accurate to guide the frame selection with several strong actions/nouns. We present a few captions generated by VideoLLaMA2 in Figure 4 to prove the caption confidence. We randomly pick 30 videos per class from train set, create sub-samples of the videos, and run caption generations on them. The highest-occurring entities/nouns form a final text summarization prompt and represent the overall actions/activities per class. We perform sub-sampling on videos and majority voting on generated entities to make the final prompt stable and not affected by the model's hallucination. The final prompt is passed to the BLIP2-pretrained BERT text-encoder [9] and generates text embedding $T \in B \times max\_token \times d_t$ for the further frame matching task.

$$\mathbf{Q}_{\text{self}} = \text{softmax}\left(\frac{\mathbf{Q}_l\mathbf{W}_Q\left(\mathbf{Q}_l\mathbf{W}_K\right)^\top}{\sqrt{d_k}}\right)\left(\mathbf{Q}_l\mathbf{W}_V\right) \quad (1)$$

$$\mathbf{Q}'_l = \text{LayerNorm}\left(\mathbf{Q}_l + \mathbf{Q}_{\text{self}}\right)$$

$$\mathbf{Q}_{\text{cross}} = \text{softmax}\left(\frac{\mathbf{Q}'_l\mathbf{W}'_Q\left(\mathbf{V}\mathbf{W}'_K\right)^\top}{\sqrt{d_k}}\right)\left(\mathbf{V}\mathbf{W}'_V\right) \quad (2)$$

$$\mathbf{Q}_{l+1} = \text{LayerNorm}\left(\mathbf{Q}'_l + \mathbf{Q}_{\text{cross}}\right)$$

The CAFS module builds upon the idea of Q-Former as illustrated in Figure 3(Left). The work process of Q-Former starts with M learnable query tokens $\mathbf{Q}_0 \in \mathbb{R}^{M\times D_q}$, where D_q is the query token dimension. Next, the self-attention query to query is defined in Eq. 1 and the cross-attention query-to-key tokens are defined in Eq. 2. We repeat the work process of Eq. 1 to 2 for L layers. Figure 3 shows how the CAFS is designed using two stages of the Q-Former block. In the first stage, we use pre-trained learned queries as the Q_0, whereas key (K) and values (V) come from the Visual Encoder output X. The production of the first stage is Visual attention query embeddings (VA_{emb}). This VA_{emb} is later used as the key (k) and value (v), whereas the Per-class text prompt embedding (T) is used as the Query embeddings. The output of the second-stage Q-former is derived from the last-layer cross-attention (A_{cross}), which represents the relevance between the class prompt and each frame in the particular video. Finally, the cross-attention output (A_{cross}) is passed into the ITM Head as provided in Eq. 3.

$$\mathbf{ITM}_{\text{score}} = Sort(Sigmoid(FC(\mathbf{A_{cross}}))) \in \mathbb{R}^{T\times 2} \quad (3)$$

From the set of T temporally sorted frames, the algorithm selects the Top-K frames with the highest matching scores for downstream anomaly detection. To enhance computational efficiency, the indices of these Top-K frames are cached after the first training epoch, thereby avoiding redundant recomputation in subsequent epochs. This module is task-agnostic and can be seamlessly integrated as an off-the-shelf component within a wide range of video understanding tasks.

3.3. Local-Global Temporal (LGT) Network

Preserving temporal information is crucial for the VAD task. There are several ways to model temporal information, such as adding positional embedding, Multi-scale Temporal Network with different strides of 1-D convolution [6], and Graph Convolutional Network (GCN) for learning the long-term and short-term temporal relationships [30]. Due to the recent effectiveness of the GCN for temporal modeling [26, 30], we leverage a similar implementation from a recent work on WSVAD [26]. The LGT Network in Figure 2 utilizes a self-attention module masked within a local

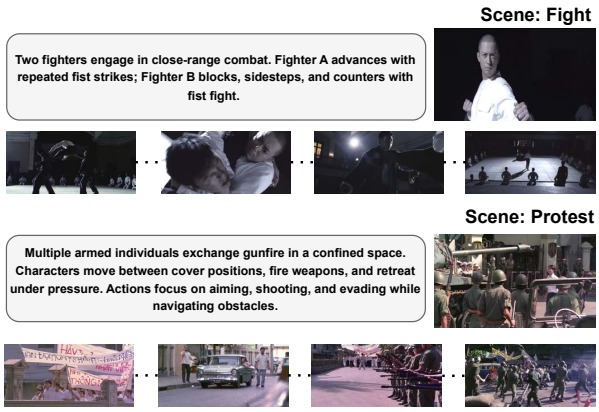

Figure 4. Generated caption using VideoLLaMA2 for different scene types.

window for **local temporal dependencies**. Next, we split the Top-k frame-level features into several uniform-length segments over the temporal dimension to achieve the local self-attention. The goal is to learn the temporal context from nearby features and establish local-context relationships across the video frames, denoted as $\mathbf{X}_l$. Windowed Attention Mask only attends to a fixed window or block, reducing quadratic complexity. To capture the **global temporal dependency**, we utilize the GCN module following the idea introduced by [26]. Wu et al. address the long-term temporal relationship based on feature similarity and relative distance. The following formula summarizes the concept of global temporal learning through the similarity and distance matrices.

$$\begin{aligned} \text{sim}_{\text{adj}} &= \text{GELU}\left(\text{Softmax}\left(H_{\text{sim}}(\mathbf{X}_l)\right)\right) \\ \text{dist}_{\text{adj}} &= \text{GELU}\left(\text{Softmax}\left(H_{\text{dist}}(\mathbf{X}_l)\right)\right) \\ \mathbf{X}_g &= \text{sim}_{\text{adj}} \parallel \text{dist}_{\text{adj}} \end{aligned} \tag{4}$$

The sim_{adj} is the cosine similarity function. The $dist_{adj}$ is the proximity adjacency matrix defined as $-|i-j|/\sigma$, where i and j are frame indices and σ is a hyperparameter to control the influence of the distance relation.

3.4. Multi-Level Prompt Network

To improve VAD performance, we propose a novel paradigm that mimics the human cognitive process through hierarchical focus and multi-level text prompt attention. As shown in Figure 2, the pipeline consists of three branches overall. The top and bottom branches deal with multi-level text prompt learning. On the other hand, the middle branch provides the temporal visual information to accompany the text information.

Entity-level Prompt Attention focuses on important entities that might play a role in a particular class of anomaly. For example: *"Arrest"* anomaly class focuses on entities/nouns such as: ["police", "handcuffs", "officer", "person", "uniform", "vehicle", "badge"]. The number of entities per class is denoted as N_c. We utilize a pre-trained CLIP text encoder (ViT-B/16) to generate entity embeddings (T_e). Next, we perform entity cross-attention with temporal visual features. For cross attention, we use visual feature as Query ($\mathbf{Q} = \mathbf{X}_g$), and per-class entity prompts as key and value ($\mathbf{K} = \mathbf{V} = \mathbf{T}_e^c$). Eq. 5 outlines the cross-attention formula with a learnable class-entity weight W_c to focus more on important entities.

$$\mathbf{A_o} = \textbf{Norm}\left(\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \cdot \mathbf{W}_c\right) \tag{5}$$

Next, we normalize the weighted attention by dividing it by the total attention and add a very small bias (ϵ) to avoid invalid operations and vanishing gradient descent. Finally, we compute the attention output by multiplying the weighted attention by the value matrix (V). The **entity attn. features** becomes $\mathbf{A_o} \in \mathbb{R}^{B\times K\times d_v}$.

Scene-level Prompt Attention Focuses on overall activities in the video over time using a summarized video feature and a scene text prompt. Next, an anomaly-attended video feature is created by the dot product between the anomaly scores and each frame feature in the video, with the following formula:

$$\mathbf{A} = \left(\text{FC}\left(\mathbf{A_o}\right)^\top \cdot \mathbf{X}_g\right) \in \mathbb{R}^{B\times 1\times d_v} \tag{6}$$

Next, we introduce learnable scene prompts to guide the scene-level feature learning [26, 31]. The initial scene prompts are anomaly class labels, such as "Fighting," "Explosion," and "Abuse," etc. Next, we tokenize the class label and place the CLIP tokens between the learnable prefix and the postfix tokens. The complete scene prompt structure follows: $[\texttt{CLS}]$ $\underbrace{[\texttt{P}_1],..,[\texttt{P}_m]}_{\text{Prompt Prefix Tokens}}$ $[\texttt{T}]$ $\underbrace{[\texttt{P}_{m+1}],..,[\texttt{P}_n]}_{\text{Prompt Postfix Tokens}}$ $[\texttt{EOS}]$. To obtain the scene embedding (T_s), we utilize the same pre-trained CLIP text encoder (ViT-B/16) that we employed during entity embedding. Next, we aggregate the scene embedding with anomaly attn. feature, and feed to a feed-forward-network (FFN), with operations in Eq. 7.

$$\mathbf{S} = \left(\text{FFN}\left(A + T\_s\right) + T\_s\right) \tag{7}$$

Finally, we perform matrix multiplication on the normalized frame feature and textual scene attention to derive a similarity map between each frame and anomaly classes. The similarity matrix is calculated in Eq. 8.

$$\mathbf{Sim_{mat}} = \left(\text{Norm}\left(X_g\right)\text{Norm}\left(S\right)^\top\right) \in \mathbb{R}^{B\times K\times C} \tag{8}$$

3.5. Objective Functions

In our MLPVAD pipeline, we use three objective functions to optimize the model: (1) Coarse Classification, (2) Con-

trastive Learning, and (3) Fine-grained Classification. The functions are labeled with the circle shape in Figure 2.

Coarse Classification Determines whether a video is normal/abnormal. To perform MIL, we select top-K confidence frames from the anomaly video and bottom-K confidence frames from the normal video. Next, we perform mean aggregation and derive a single confidence value as the video-level confidence and feed it to the binary cross-entropy loss function. Eq. 9 outlines the loss calculation formula.

$$\mathcal{L}_{bce} = (\text{Sigmoid}\,(FC(A_o))) \in \mathbb{R}^{B\times 1} \tag{9}$$

Contrastive Learning For challenging datasets, the difference between normal and abnormal cases can be very subtle. We require efficient feature learning techniques to classify subtle cases better. We propose using randomized contrastive learning with multiple positive and negative keys to achieve robust separation in feature space.

$$\mathcal{L}_{\text{cons}} = Mean\left(-log\frac{\sum_{i=1}^{P} exp(sim(Q,K_i^+)/\tau)}{\sum_{j=1}^{N} exp(sim(Q,K_j^-)/\tau)}\right) \tag{10}$$

The P and N are the number of positive and negative keys. The Query and Keys come from top-k/bottom-k confidence frames. Using contrastive learning, we can cluster all abnormal frames together and separate normal frames from the abnormal frames in the feature space.

Fine-Grained Classification The imbalanced dataset makes accurate and unbiased VAD detection challenging. Also, actual anomaly frames may span a very short amount of time in the video. For multi-class anomaly classification, some videos might fall under more than one anomaly class. To address these challenges, we introduce fine-grained classification based on a scene-guided frame-class alignment matrix. Similar to the coarse classification, we rely on the MIL approach, considering only the top-K candidates. The steps to calculate fine-grained loss are:

$$\begin{aligned} \mathbf{p}_i &= \text{Softmax}\left(\frac{1}{k_i}\sum_{j=1}^{k_i}\text{TopK}(\mathbf{L}_i)\right) \in \mathbb{R}^C \\ \mathcal{L}_{\text{ce}} &= \text{Mean}\left(-\sum_{c=1}^{C} y_i^{(c)}\log p_i^{(c)}\right) \end{aligned} \tag{11}$$

The final objective function for the MLPVAD model is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{bce} + \mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{ce}} \tag{12}$$

4. Experiments

Datasets are (1) UCF-Crime and (2) XD-Violence benchmarks. Both of the datasets have video-level labels (weakly labeled) for the training set and frame-level labels for the testing set. The UCF-Crime CCTV surveillance videos cover 13 anomaly event categories, whereas XD-Violence covers six anomaly event categories from movie clips.

Methods	UCF (AUC)	XD (AP)
Sultani et al. (2018)	84.06	74.78
Wu et al. (2020)	84.40	78.66
RTFM (2021)	85.22	78.27
AVVD (2022)	82.38	77.50
TEVAD (2023)	84.85	79.52
MGFN (2023)	86.67	80.11
VadCLIP (2024)	87.65	83.88
TPWNG (2024)	87.79	83.68
MLPVAD	**88.84**	**85.37**

Table 1. Coarse-grained comparison using AUC and AP on UCF-Crime and XD-Violence dataset.

Evaluation Metrics Following earlier benchmarks, Average Precision (AP) and Area Under the Curve (AUC) are used as classification metrics [6, 26]. Particularly for coarse-grained classification, we calculate frame-level AUC and frame-level AP for the dataset. For fine-grained classification, we calculate the mean Average Precision (mAP) under different intersection over union (IoU) thresholds ranging from 0.1 to 0.5.

Implementation is adopted from LAVIS [14] and VadCLIP [26] github reporsitories. For feature extraction, we utilize the pre-trained CLIP (ViT-B/16 or ViT-B/14). Our implementation seamlessly supports both versions of the ViT encoder. The BLIP-2 Q-Former pretrained model is also adopted from LAVIS. The VadCLIP implementation serves as inspiration for the LGT Network implementation. In the CAFS module, the value of K is 256. The batch size is set to 64, and the text and visual-feature dimensions are 768 and 512, respectively. For contrastive learning, we select k = 8 out of K (256) frames. For model training, a single NVIDIA RTX A6000 GPU is used with the PyTorch framework. For both datasets, the learning rate and # of epochs were set to 2×10^{-4} and 20, respectively.

4.1. Comparison with State-of-the-art methods

In this section, we compare the proposed MLPVAD model with several of the latest Weakly-supervised models. To perform a fair comparison, we use the same visual and textual encoders (ViT-B/16) and the same feature dimension for all SOTA methods. We also maintain the batch size and the number of epochs constant throughout the evaluation. Most existing SOTA methods downsize the video input to a fixed k=256 frames for training. So, we also select the Top $k = 256$ frames from the CAFS modules to match the performance comparison with other SOTA methods. Moving forward, we use K=256 for all experiments, except in the

mAP@IOU %	0.1	0.2	0.3	0.4	0.5
Method ↓	Dataset: UCF-Crime				
Sultani et al.	5.72	4.41	2.69	0.02	0.01
AVVD	10.27	7.01	6.25	3.42	3.29
VadCLIP	11.72	7.83	6.40	4.53	2.93
MLPVAD	**13.21**	**9.74**	**7.78**	**6.20**	**4.11**
Method ↓	Dataset: XD-Violence				
Sultani et al.	22.72	15.57	9.98	6.20	3.78
AVVD	30.51	25.75	20.18	14.83	9.79
VadCLIP	37.03	30.84	23.38	17.90	14.31
MLPVAD	**38.55**	**32.63**	**25.82**	**19.39**	**15.60**

Table 2. Fine-grained comparisons on XD-Violence and UCF-Crime.

CAFS	Vis-P	Scn.-P	Ent.-P	C-Lrn	AP(%)
Baseline	✓	✓			84.53
✓	✓	✓			86.25
✓	✓		✓		85.33
	✓	✓	✓		86.56
✓	✓	✓	✓		88.20
✓	✓	✓	✓	✓	**88.84**

Table 3. Ablation study on UCF-Crime dataset for different MLPVAD modules.

ablation study, where different values of K are examined.

Coarse-grained WSVAD Performance Table 1 presents the coarse-grained performance comparison with several SOTA methods. The coarse classification based on visual data performs worse than the MLPVAD method [21, 25, 26]. Earlier works, such as [21, 25], showed decent AUC gains of 85.22%, and 84.57%, respectively, on UCF-Crime. Whereas our MLPVAD method achieved an AUC of 88.84%, outperforming its closest competitors, such as VadCLIP (87.65%) and TPWNG (87.79%).

MLPVAD also outperforms the state of the art for the XD-Violence dataset. Most previous works, such as [6], [21], and [7], achieve AP close to 80%. VadCLIP and TPWNG achieved a noticeable improvement by utilizing a more effective temporal network and feature learning, resulting in an AP of 83.88%. Our efficient frame selection technique and entity-guided approach outperformed existing SOTA methods with a margin of $+\Delta 1.5\%$ and gained an AP of 85.37%.

Fine-grained WSVAD Performance. In Table 2, the fine-grained classification is evaluated using mAP at different IOU thresholds. We compare the fine-grained MLPVAD performance with other SOTA work that introduced fine-grained analysis. Fine-grained classification was first proposed by AVVD, later fine-tuned by Sultani et al. and VadCLIP with CLIP features. Our work modifies the fine-grained pipeline, integrating scene attention and task-centric anomaly attention features guided by entity prompts. Compared to AVVD and VadCLIP, MLPVAD achieves a performance improvement of 8% and 1.5% on the UCF-Crime dataset (IOU = 0.1). MLPVAD also showed a decent performance gain of 13.21% mAP (IOU=0.1) in challenging the XD-Violence dataset.

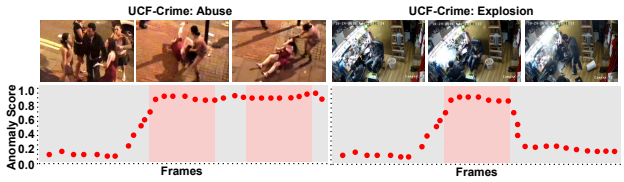

Figure 5. Frame-level anomaly detection using the MLPVAD model for the UCF-Crime dataset.

4.2. Ablation Study

In this section, we try to answer several questions regarding the effectiveness of the proposed modules in the MLPVAD pipeline. The ablation study begins with the baseline components in the pipeline (See Table 3) for the UCF-Crime dataset. We start with textual and visual data in the baseline, as our MLPVAD focuses more on utilizing multimodal features for improved performance. The baseline model achieves an AP of 84.53%. Next, we incorporate the proposed CAFS module for task-specific frame selection, and we observed a significant gain of $+\Delta 1.75\%$. It proves that careful frame selection is very crucial for downstream tasks. Next, we turn off scene prompt learning and use only the entity-visual prompt with the CAFS module. We found that entity learning alone outperforms scene learning, achieving an AP of 85.33%. In the next two experiments, we use visual, entity, and scene prompts with/without the CAFS module. The pipeline with CAFS achieved a $+\Delta 1.64\%$ improvement compared to the one without it, demonstrating the effectiveness of the CAFS module. Finally, we introduce randomized contrastive learning in the pipeline and optimize the model for better feature separation. Using multiple positive and negative keys for contrastive learning, MLPVAD achieved an AP of **88.84%** on the benchmark dataset.

Effectiveness based on video duration The duration of the video is a very important factor to consider for any video-related downstream task. In Table 4, we present a performance comparison between MLPVAD and other SOTA methods, dividing the test videos into short ($< 15mins$) and long ($>= 15mins$) categories. We pick *15 minutes* as the threshold for a balanced evaluation. The UCF-Crime dataset contains 1,795 short and 112 long videos. It is ev-

Methods	Short Videos ($< 15mins$)	Long Videos ($>= 15mins$)
TEVAD	83.76	77.35
MGFN	89.47	84.75
VadCLIP	90.22	85.30
TPWNG	90.08	84.90
MLPVAD	**91.48**	**87.48**

Table 4. Duration-based performance comparison using **AUC** on UCF-Crime dataset.

Frame Select.	UCF AUC(%)	XD AP(%)
Stepwise	85.78	82.10
Mean-Pool	87.30	83.88
CAFS (K=128)	88.25	84.68
CAFS (K=256)	88.84	85.37
CAFS (K=320)	**89.06**	**85.60**

Table 5. Effectiveness of different frame selection methods and performance analysis for CAFS with different values of K.

ident from Table 4 that methods perform better for short videos than long videos. The frame selection and temporal modeling are more effective when the duration is short. The MLPVAD method still outperforms other methods by at least a minimum margin of $+\Delta 1.3\%$ for the UCF-Crime dataset. For long videos, our MLPVAD demonstrates its superiority through efficient frame selection and a multi-level prompt pipeline. It achieves 87.48% AUC on UCF-Crime and beats the closest competitors, such as VadCLIP and TPWNG, by a large margin of $\Delta 2\%$.

Ablation on different frame selection Table 5 illustrates the MLPVAD evaluation results when trained using different frame selection methods. The performance drops by 3% when we use traditional interval-based frame selection. The mean-pool method shows a decent performance gain of $\sim$ 2% for both experimental datasets. The technique still suffers when the video length increases (See Table 4), as more noise is introduced in frames during the mean-pooling operations. We show the CAFS effectiveness for three different values of K. CAFS performs better than most of the SOTA methods, even with half the amount of temporal frames. With K=256, the CAFS stands out as the superior method, achieving $\sim 3\%$ gain in each dataset. Further increasing K to 320 improves performance by $+\Delta 0.3\%$; however, it also requires additional GPU memory and computational power.

Qualitative Results Figure 5 and Figure 6 illustrate the frame-level detection result for a few videos with the anomaly score. The red dots in the Figure denote the anomaly score for a particular frame. The anomaly scores

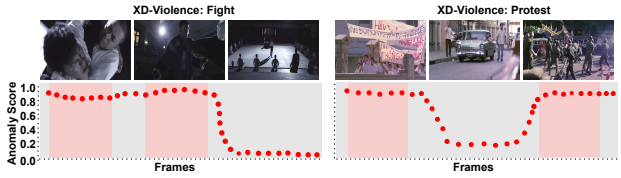

Figure 6. Frame-level anomaly detection using the MLPVAD model for the XD-Violence dataset.

exhibit a sharp increase during the presence of anomalous events, followed by a rapid decline once these events conclude. In contrast, normal events consistently yield lower anomaly scores over time. This behavior indicates that our approach is highly sensitive to abnormal activities, effectively identifying them promptly while preserving low anomaly predictions during normal scenarios.

5. Conclusion and Future Work

In this work, we propose a new paradigm, named MLPVAD, for the WSVAD task, focusing on long-duration videos. We propose an elegant technique for selecting a limited number of task-specific frames from a larger pool. We utilize a frozen Image-Text Matching CLIP Network to perform text-guided frame selection. Next, we propose a multi-level prompt learning architecture that follows human-like cognitive processes. We leverage the power of the latest LLMs and CLIP models to guide the learning at the entity and scene levels. Finally, we introduce MIL learning with three objective functions: coarse and fine-grained classification, and contrastive learning. We empirically verify the effectiveness of MLPVAD modules through SOTA comparison and extensive ablations on WSVAD benchmarks. The ablation study also verifies the superiority of MLPVAD for long-duration VAD. In the future, we plan to improve the VAD performance for cases where the abnormality is very subtle and exhibits high interclass similarity. We also plan to extend the performance analysis for different LLM models for caption generation and verify the LLMs' hallucination for challenging and subtle scenes.

References

[1] Armstrong Aboah. A vision-based system for traffic anomaly detection using deep learning and decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4212, 2021. 1

[2] Jungmo Ahn, JaeYeon Park, Sung Sik Lee, Kyu-Hyuk Lee, Heesung Do, and JeongGil Ko. Safefac: Video-based smart safety monitoring for preventing industrial work accidents. *Expert Systems with Applications*, 215:119397, 2023. 1

[3] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. In *Proceedings of the Computer Vi-*

sion and Pattern Recognition Conference, pages 8545–8556, 2025. 3

[4] Debojyoti Biswas and Jelena Tesic. Mmvad: A vision-language model for cross-domain video anomaly detection with contrastive learning and scale-adaptive frame segmentation. *Expert Systems with Applications*, page 127857, 2025. 1

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 3

[6] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2023. 1, 3, 4, 6, 7

[7] 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*, pages 387–395, 2023. 1, 7

[8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 4

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 4

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[11] 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. 1, 2

[12] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514, 2024. 3, 4

[13] Noussaiba Jaafar and Zied Lachiri. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, 211:118523, 2023. 1

[14] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada, 2023. Association for Computational Linguistics. 6

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 3

[16] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023. 3

[17] Hui Lv, 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*, pages 8022–8031, 2023. 2

[18] Övgü Özdemir and Erdem Akagündüz. Enhancing visual question answering through question-driven image captions as prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1562–1571, 2024. 3

[19] 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. 1

[20] 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. 2

[21] 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. 2, 7

[22] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[23] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4

[24] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 4

[25] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020. 7

[26] 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*, pages 6074–6082, 2024. 1, 2, 3, 4, 5, 6, 7

[27] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[28] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 3

[29] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into clip latent space for video anomaly recognition. *Computer Vision and Image Understanding*, 249:104163, 2024. 3

[30] 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. 2, 4

[31] 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. 5