Federated Weakly Supervised Video Anomaly Detection with Multimodal Prompt

Benfeng Wang[1], Chao Huang[1*], Jie Wen[2], Wei Wang[1], Yabo Liu[2], Yong Xu[2]

[1]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
wangbf23@mail2.sysu.edu.cn, huangch253@mail.sysu.edu.cn, wenjie@hit.edu.cn, wangwei29@mail.sysu.edu.cn, yaboliu.ug@gmail.com, yongxu@ymail.com

Abstract

Video anomaly detection (VAD) aims at locating the abnormal events in videos. Recently, the Weakly Supervised VAD has made great progress, which only requires video-level annotations when training. In practical applications, different institutions may have different types of abnormal videos. However, the abnormal videos cannot be circulated on the internet due to privacy protection. To train a more generalized anomaly detector that can identify various anomalies, it is reasonable to introduce federated learning into WSVAD. In this paper, we propose Global and Local Context-driven Federated Learning, a new paradigm for privacy protected weakly supervised video anomaly detection. Specifically, we utilize the vision-language association of CLIP to detect whether the video frame is abnormal. Instead of leveraging handcrafted text prompts for CLIP, we propose a text prompt generator. The generated prompt is simultaneously influenced by text and visual. On the one hand, the text provides global context related to anomaly, which improves the model's ability of generalization. On the other hand, the visual provides personalized local context because different clients may have videos with different types of anomalies or scenes. The generated prompt ensures global generalization while processing personalized data from different clients. Extensive experiments show that the proposed method achieves remarkable performance.

Code — https://github.com/wbfwonderful/Fed-WSVAD

Introduction

Video anomaly detection (VAD) is an important task in the field of computer vision (Zhang, Qing, and Miao 2019; Huang et al. 2021, 2022b,c). For example, VAD can be applied to intelligent surveillance systems, which could detect abnormal events to reduce property damage. Besides, VAD can also be used for video content examination systems to better detect the inappropriate content. Weakly supervised VAD (WSVAD) (Sultani, Chen, and Shah 2018; Huang et al. 2022a; Zhang et al. 2022; Wu et al. 2023) is an important branch of VAD. For WSVAD, the model will simultaneously take normal and abnormal videos as inputs. However, there is only video-level annotation provided. Existing methods typically adapt frame-level anomaly score to video-level annotation with multiple instance learning (MIL) (Sultani, Chen, and Shah 2018). WSVAD has great performance while requiring little annotation cost. However, traditional WSVAD only makes use of single modal.

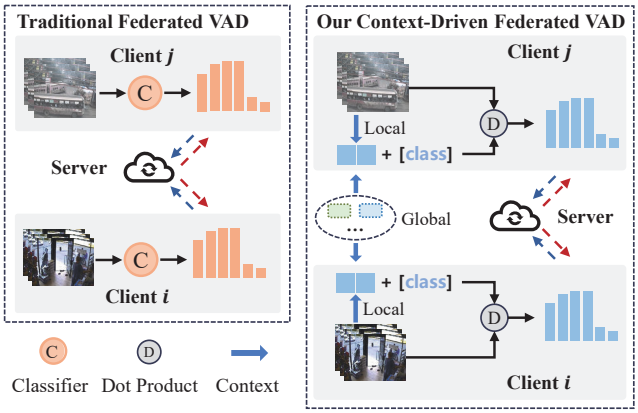


Figure 1: Comparison with different paradigms of federated WSVAD. Traditional federated VAD method (left) simply aggregates the classifier. In contrast, the proposed method (right) focuses on context driven federated VAD.

Recently, large scale vision-language pre-trained models represented by Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) have shown their great performance on various downstream tasks, such as object detection (Du et al. 2022; Zhao et al. 2022; Kim et al. 2024), image segmentation (Xu et al. 2022; Yun et al. 2023; Luo et al. 2024) and video understanding (Wu, Sun, and Ouyang 2023; Jin et al. 2024). The core idea of CLIP is to align image and text in embedding space through contrastive learning. Some researches try to adapt CLIP to WSVAD (Wu et al. 2024a,b,c; Yang, Liu, and Wu 2024). Despite those CLIP-based methods are superior to traditional methods due to utilizing the vison-language association, they are only suitable for centralized training, which means all videos must be send to a central server. In this context, the privacy cannot be guaranteed.

Federated learning (FL) (McMahan et al. 2017) is a distributed machine learning method, which aims to train a uni-

*Corresponding Author
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

fied model among multiple clients without sharing raw data and achieves privacy protection. Recently, some researches have applied FL to computer vision such as image classification (Hsu, Qi, and Brown 2020; Su et al. 2024) and recognition (Liu et al. 2020; Dutto et al. 2024). For VAD, abnormal videos may be distributed in different institutions or owners. Due to sensitivity of video content or privacy protection, it is hard to collect the abnormal videos from different institutions or owners for centralized training. For instance, the transportation department has surveillance videos of traffic accident scenes, but the videos will not be allowed to spread on the internet to protect the perpetrator and victim of the accident. Additionally, the police has videos recorded by law enforcement recorders, which should not be spread due to the violent or terrifying elements of videos. Based on the above situation, it is reasonable to introduce WSVAD into the federation learning environment in order to train a unified anomaly detection model with abnormal videos from different institutions.

To perform privacy protected WSVAD in federated learning environment as well as activate the potential of CLIP in WSVAD, the following three challenges need to be addressed. Firstly, how to adapt CLIP to the task of WSVAD. Secondly, how to utilize CLIP in federated learning environment. Thirdly, how to improve performance of the federated model.

In this paper, we propose a Global and Local Contexts-driven Federated Learning Framework for Privacy Protected Video Anomaly Detection. Figure 1 shows the difference between the proposed method and traditional federated WSVAD. Specifically, **for the first challenge**, we propose a temporal modeling block to capture the temporal dependencies of abnormal events, since the abnormal frames in the video are consecutive and related to each other. After that, texts are introduced to utilize the vision-language association of CLIP. Similar to zero-shot image classification with CLIP, we compute the similarity of text and visual features. Finally, the similarity map represents the multi-category frame-level anomaly confidence. Besides, only video-level annotation is available in WSVAD. The MIL is leveraged to select the most abnormal frames to represent the whole video. **For the second challenge**, CLIP is suitable for federated learning due to its powerful representation ability which can fit various data from different clients (Cui et al. 2024b). However, it is impractical to train the whole CLIP in federated learning due to high computation and transmission costs. Prompt learning such as Context Optimization (CoOp) (Zhou et al. 2022b) adapts CLIP to downstream tasks through additional trainable parameters, which explores the task-related information while maintaining the performance of CLIP. Some researches (Zhao et al. 2023; Yang, Wang, and Wang 2023; Li et al. 2024) extend prompt learning to federated environment, which make full use of the power of CLIP while learning personalized prompt for every client. However, the learned prompts are still limited. Therefore, we propose a generator to dynamically generate unique prompt based on contexts. **For the third challenge**, the proposed prompt generator is driven by both global and local contexts. On the one hand, the global contexts are composed of anomaly categories from all clients, which are global task related. Conditioned by the global context, the prompt generator will maintain generalization as soon as possible when aggregating on the server, which is always simply averaging (FedAvg) (McMahan et al. 2017). In other words, the generated prompt can generalize to the anomaly categories from other clients. On the other hand, the local contexts are the videos with different characteristic among clients. For instance, client i holds the surveillance videos of the street, while client j possess the surveillance videos of grocery store. The local context guides the generated prompt to pay attention on the average distribution of the videos from different clients. That means the model will identify client-specific characteristic moderately while keeping generalization. Last but not least, the combination of global and local contexts strikes a balance between local and global optimum.

In summary, the contributions of this paper are as follows:

- We propose a novel privacy protected federated learning framework for video anomaly detection, which enables multiple clients to train a unified and generalized video anomaly detector with privacy.
- We design a prompt generator driven by both global and local contexts, which achieves global generalization and local personalization.
- We conduct extensive experiments on multiple datasets. In order to simulate real-world scenarios, we re-organize XD-Violence and UCF-Crime datasets. The results demonstrate the superiority of the proposed method.

Related Work

Weakly Supervised Video Anomaly Detection

There is only video-level annotation provided for training WSVAD model. In order to address this limitation, Sultani *et al.* (Sultani, Chen, and Shah 2018) first proposed a multiple instance learning framework for WSVAD, which considers the videos as bags and the snippets of video as instances. The core idea of MIL is to select the snippet with highest anomaly confidence to represent the whole video. Then the rank loss is introduced to pull away the representative snippet in normal and abnormal videos. Based on this, many fellow-up works have made improvements. For example, Huang *et al.* (Huang et al. 2022a) proposed a transformer based temporal feature aggregator to capture semantic similarity and position correlation of snippets from different embedding space. Zhou *et al.* (Zhou, Yu, and Yang 2023) proposed a memory mechanism to store normal and abnormal prototypes to better distinguish hard samples. Lv *et al.* (Lv et al. 2023) proposed an unbiased MIL framework to promote the model to focus on unbiased anomalies. Recently, significant progress has been made in multimodal learning (Xu et al. 2023; Ling et al. 2023; Cui et al. 2023, 2024a). Some researches introduce pre-trained vision-language model like CLIP into WSVAD. Specifically, Wu *et al.* (Wu et al. 2024c) proposed a new novel paradigm named VadCLIP, which adapts CLIP to WSVAD with two prompt mechanisms. Yang *et al.* (Yang, Liu, and Wu 2024) proposed

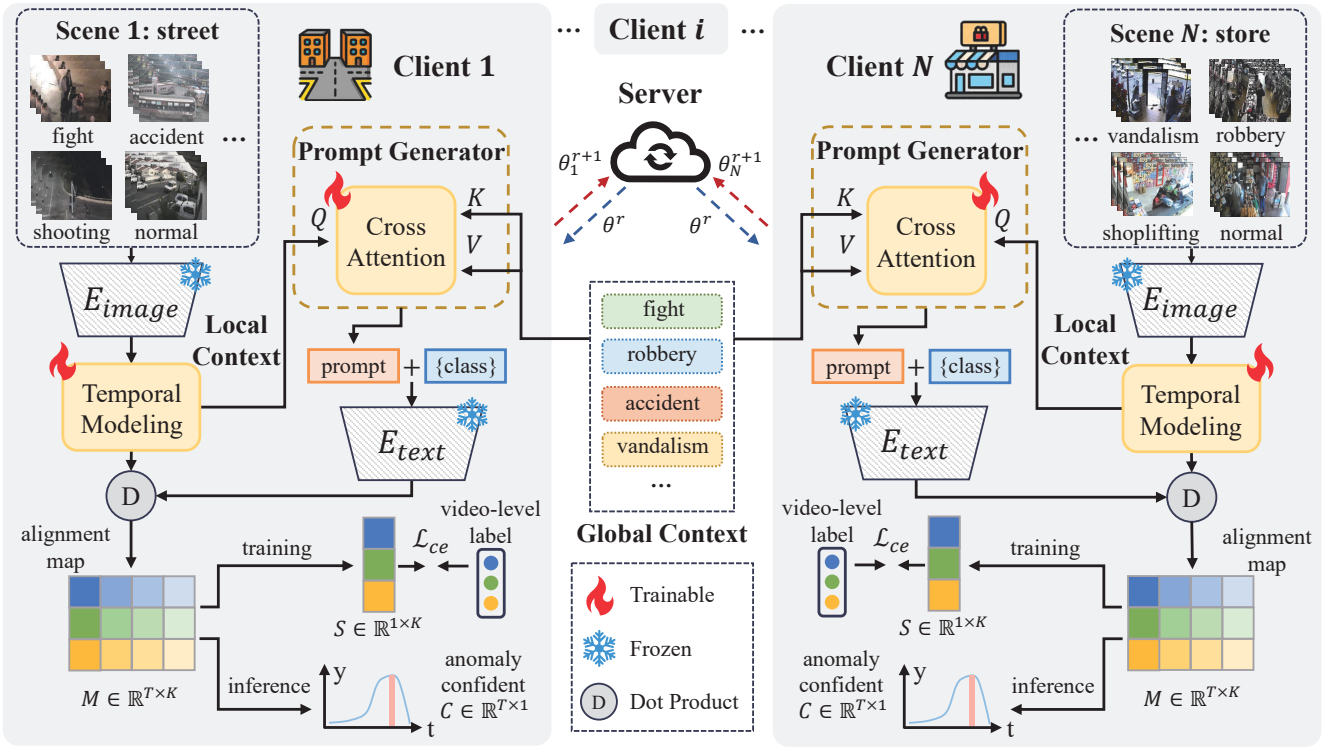


Figure 2: The framework of the proposed method. Client 1 owns the videos recorded on the street while client N owns the videos recorded in the store. Each client is equipped with a frozen CLIP. We propose a generator to dynamically generate prompts driven by local visual contexts and global text contexts for CLIP text encoder.

a framework to transfer the capability of CLIP to generate pseudo-label for WSVAD. Besides, a few works try to apply WSVAD into federated learning. Specifically, Doshi *et al.* (Doshi and Yilmaz 2023) proposed a transformer-based federated learning frame for video anomaly detection and video action recognition. Al-lahham *et al.* (Al-Lahham et al. 2024) proposed a new baseline named CLAP, which contains new test and evaluation scenarios for federated VAD. However, the performance of federated VAD deserves further exploration.

Federated Prompt Learning for CLIP

Prompt tuning effectively adapts CLIP to downstream tasks. For instance, CoOp (Zhou et al. 2022b) replaces the handcrafted prompt with a set of learnable vectors for CLIP text encoder. Furthermore, CoCoOp (Zhou et al. 2022a) dynamically adjusts the learned prompts with tokens generated by the input image. Some works (Guo et al. 2023; Zhao et al. 2023) introduced prompt tuning into federated learning, which not only keeps the capability of CLIP but also save transmission costs. Moreover, pFedPG (Yang, Wang, and Wang 2023) learns a network for generating personalized prompts to tackling the personalized data among clients. FedTPG (Qiu et al. 2024) also trains a prompt generator conditioned by task-related contexts, which enables the generalization capability to unseen classes and datasets. FedAPT (Su et al. 2024) solves the cross domain challenge, which guides CLIP to activate the domain related knowledge by incorporating personalized information into prompts. Fed-DPT (Wei et al. 2023) facilitates domain adaptation with domain-specific prompts coupling visual and textual representations by self-attention. However, in a federated environment, the challenge is to enable each client to adapt or generate prompts that are suited to their personalized data while still contributing to a robust and generalized global model. Some works explore the balance of generalization and personalization. For instance, FedOTP (Li et al. 2024) utilizing unbalanced optimal transport to align global and local prompts while promoting prompts to pay more attention on class related information from image. FedPGP (Cui et al. 2024b) leverages low-rank decomposition to ensure robust generalization as well as introducing additional contrastive loss.

Proposed Method

In this section, we introduce the proposed method in detail, as illustrated in Figure 2. The proposed method adapts CLIP for federated weakly supervised video anomaly detection with global and local contexts driven prompts.

Adapting CLIP for WSVAD

There are only video-level annotations available for WSVAD during training. Given a video v, the video is defined as abnormal if at least one frame contains anomaly and the

corresponding label y is defined as 1. On the contrary, if all frames of v do not contain anomaly, the video will be labeled as normal with corresponding label $y = 0$. The goal of WSVAD is to train an anomaly detector $f(\cdot)$ which predicts the frame-level confidence with video-level annotations. Previous works often use the pre-trained C3D (Tran et al. 2015) or I3D (Carreira and Zisserman 2017) to extract video features. For better feature representation and utilize the vision-language association, we use image encoder of CLIP to extract video features. Specifically, the extracted video features can be represented as $I \in \mathbb{R}^{T\times D}$, where T represents the number of frames and D is the embedding dimension. CLIP is trained with large scale image-text pair and demonstrates great performance in various image processing tasks. However, temporal dependencies will be ignored if videos are only processed with CLIP. For WSVAD, the abnormal event usually lasts for a period of time. It is difficult to detect anomalies with independent frames. In order to eliminate the impact of lacking temporal information for WSVAD, some works (Huang et al. 2022a; Wu et al. 2024c) proposed various temporal modeling blocks. Similarly, we introduce a lightweight temporal modeling block. Specifically, a Transformer Encoder $\varphi(\cdot)$ is added after video features extracted by image encoder of CLIP, which is presented as: $V = \varphi(I)$, where $V \in \mathbb{R}^{T\times D}$ refers to the video features with temporal dependencies. Then, the text labels are encoded by the text encoder of CLIP, which can be represented as $L \in \mathbb{R}^{K\times D}$, where K represents the number of text labels. Subsequently, the alignment map M can be computed as follows:

$$M = V \cdot (L)^{\top}, \tag{1}$$

where $M \in \mathbb{R}^{T\times K}$ refers to the similarity between video frames and text labels. Finally, we adjust the alignment map to adapt the video-level annotations. Specifically, each row of M refers to the similarity of all video frames and the current class. We compute the average of top k values of each row to measure the degree of consistency between the video and the current class, which is represented as $S = \{s_1, s_2, ..., s_K\}$. Then the multi-class prediction is computed as

$$p_i = \frac{exp(s_i/\tau)}{\sum_j exp(s_j/\tau)}, \tag{2}$$

where p_i represents the probability that the video belongs to i_{th} class, and τ represents temperature hyper-parameter. After that, the final loss is computed by cross-entropy. During the inference stage, for the alignment map M, we consider the similarity between current video and normal class. The higher similarity means that the current video is more likely to be a normal video, corresponding to a lower anomaly confidence. So we subtract the similarity from one as the anomaly confidence.

Generating Prompts in Federated WSVAD

CoOp (Zhou et al. 2022b) improves the performance of CLIP on downstream tasks by replacing the handcrafted text prompt with learnable prompt vectors. Similarly, we introduce the text prompt for text encoder to adapt various anomaly classes. However, the learned prompts are still limited in federated learning environment. To better adapt federated WSVAD, we propose a prompt generator to generate text prompts dynamically. Specifically, the original class labels l are first tokenized by the tokenizer of CLIP, which is presented as $tokens = Tokenizer(l)$.

Subsequently, the tokens are concatenated with the generated prompt $\mathrm{P} = \{\mathrm{p}_1, ..., \mathrm{p}_n\}$, and the input of the text encoder of CLIP can be presented as $t = \{\mathrm{p}_1, ..., token, ..., \mathrm{p}_n\}$.

Then let's discuss the process of prompt generation in federated WSVAD. Considering a federated learning environment where N clients collaborated training a model, each client i holds own dataset $\mathcal{D}_i$ and is equipped with a pre-trained CLIP. It should be noted that the trainable parameters of each client consist of the temporal modeling block mentioned in sec 3.2 and the prompt generator, while the CLIP keeps frozen during training. Below we describe the interaction process between clients and server for round r:

- **Step I**: Each client receives and loads the current global parameters θ^r from server.
- **Step II**: For each client i, the generated prompts are concatenated with label tokens and following input into text encoder. Then compute the multi-class prediction P with equation 3. The loss function is defined as

$$\mathcal{L}_i = -\mathbb{E}_{(v,Y)\in\mathcal{D}_i} Y log P, \tag{3}$$

  where $Y \in \mathbb{R}^K$ is a multi-class label for a video. Finally, we update the local parameters with optimizer e.g. SGD.
- **Step III**: After local training with some epochs, all clients send their local parameters θ_i^r to the server for aggregated global parameters θ^{r+1}, which is presented as

$$\theta^{r+1} = \sum_i \frac{|\mathcal{D}_i|}{\sum_j |\mathcal{D}_j|} \theta_i^r. \tag{4}$$

Global and Local Context-driven Prompt

The goal of federated learning is to train a unified model with distributed data. Usually, the distribution of data held by different clients varies. For instance, different institutions may hold different abnormal videos in the task of WSVAD. Thus, in order to achieve local personalization while maintaining model generalization performance, we propose the Global and Local Context-driven Federated Learning for WSVAD. On the one hand, motivated by FedTPG (Qiu et al. 2024), the task-related text can provide context for federated prompt learning. For federated WSVAD, the global class texts essentially contain more general summary information about the task. In other words, despite the fact that there are unseen anomalies for some clients during training, the global contexts can activate the frozen CLIP to gain moderate ability of identifying those unseen anomalies. On the other hand, the local visual information can provide client-specific average context to the prompt, which promotes the model to learn the distribution of local data instead of overly leaning towards a particular anomaly category. The combination of global and local context achieves a balance between personalization and generalization.

Method	Venue	Feature	Random		Event		Scene	
			UCF	XD	UCF	XD	UCF	XD
ZS-CLIP (Radford et al. 2021)	ICML2021	CLIP	71.20	51.34	71.20	51.34	71.20	51.34
Temporal-CLIP (Ju et al. 2022)	ECCV2022	CLIP	83.69	72.07	80.65	69.76	83.89	70.26
FedCoOp (Guo et al. 2023)	TMC2023	CLIP	84.03	71.80	84.72	69.70	**85.36**	73.37
PPVU (Doshi and Yilmaz 2023)	DSC2023	TimeSformer	82.90	-	-	-	-	-
CLAP (Al-Lahham et al. 2024)	CVPR2024	CLIP	**84.99**	68.60	84.75	66.57	84.57	70.54
Ours		CLIP	84.06	**75.32**	**85.07**	**74.23**	84.03	**75.99**

Table 1: AUC (%) on UCF-Crime and AP (%) on XD-Violence of comparisons on three data split strategies: Random split, Event based split, Scene based split.

Specifically, the prompt generator is a cross-attention module. For global contexts, the class texts are first encoded into embeddings as $\mathcal{T} \in \mathbb{R}^{K \times D}$. Then the embeddings $\mathcal{T}$ are transformed into key and value vectors. For local contexts, considering a batch of visual features $V \in \mathbb{R}^{B \times T \times D}$ with batch size B, we compute the average of V in the batch dimension. Similarly, the averaged visual features $\bar{V}$ are transformed into query vectors. Finally, the prompt P is generated with cross attention merging those vectors, which can be presented as follows:

$$\mathrm{P} = CrossAttention(Q_{\bar{V}}, K_{\mathcal{T}}, V_{\mathcal{T}}), \tag{5}$$

where $Q_{\bar{V}} = \bar{V} \times W_Q$, $K_{\mathcal{T}} = \mathcal{T} \times W_K$, $V_{\mathcal{T}} = \mathcal{T} \times W_V$.

Experiment

Experimental Settings

Datasets and evaluation metrics We conduct extensive experiments on the large-scale datasets: UCF-Crime (Sultani, Chen, and Shah 2018) and XD-Violence (Wu et al. 2020). Additionally, we evaluate the generalization to unseen datasets on the test set of ShanghaiTech (Luo, Liu, and Gao 2017). ShanghaiTech is initially designed for semi-supervised VAD. We utilize the re-organized test set by Zhong *et al.* (Zhong et al. 2019).

Referring to CLAP (Al-Lahham et al. 2024), we re-organize the UCF-Crime and XD-Violence datasets. Specifically, there are three strategies to split datasets: random split, event based split and scene based split. **Random split** is a baseline setting that each client holds equal number of videos. **Event based split** means that each client owns videos with different anomalies. For instance, client i possesses the videos of explosion while client j owns the videos of fighting. **Scene based split** is the closest to real application scenarios. In this setting, each client is equipped with videos recorded in specific scenes. For example, client i has videos recorded in stores containing anomalies like robbery and shoplifting, while client j owns videos recorded on the street with car accidents and shooting. Due to limited space, more details about datasets split are provided in the supplementary materials.

As for evaluation metrics, following previous works(Wu et al. 2020, 2024c), we evaluate the performance with the Area Under Curve (AUC) of the frame-level Receiver Operating Characteristics (ROC) for UCF-Crime, ShanghaiTech and UBnormal. For XD-Violence, AUC of the frame-level Precision-Recall Curve (AP) is used.

Mode	Method	UCF	XD
Centralized	Temporal-CLIP (Ju et al. 2022)	83.72	75.73
	FedCoOp (Guo et al. 2023)	84.24	**76.48**
	PPVU (Doshi and Yilmaz 2023)	**86.30**	-
	CLAP (Al-Lahham et al. 2024)	83.85	66.73
	Ours	84.82	71.16
Local	Temporal-CLIP (Ju et al. 2022)	**81.36**	64.66
	FedCoOp (Guo et al. 2023)	80.97	66.19
	CLAP (Al-Lahham et al. 2024)	76.23	61.35
	Ours	81.17	**64.88**
Federated	Temporal-CLIP (Ju et al. 2022)	80.65	69.76
	FedCoOp (Guo et al. 2023)	84.72	69.70
	CLAP (Al-Lahham et al. 2024)	84.75	66.57
	Ours	**85.07**	**74.23**

Table 2: AUC (%) on UCF-Crime and AP (%) on XD-Violence of comparisons on three training settings: Centralized training, Local training and Federated training.

Training settings We compare the proposed method with existing SOTA in three training settings: centralized training, local training and federated training. **Centralized training** requires that all data is collected together for training an anomaly detector without privacy. **Local training** means each client trains own anomaly detector with local data individually, which ensures privacy with possible loss of performance. **Federated training** requires all clients to joint an anomaly detector while ensuring privacy. There is only global test set in this paper, which means each client only possesses local training set.

Implementation details We utilize a pre-trained CLIP (VIT-B/16) to extract the visual and text features. The feature dimension D is 512. The prompt generator is composed of a four-head cross-attention, layer normalization and a linear layer for projection. The embedding dimension for cross-attention is 512. The proposed method is trained on a single NVIDIA RTX 3090 GPU using PyTorch with batch size 128. The learning rate is 1e-5. The global aggregation round and local training epoch is 15 and 10, respectively.

Method	Source	Target	
	XD	UCF	SHTech
ZS-CLIP (Radford et al. 2021)	51.34	71.20	53.69
Temporal-CLIP (Ju et al. 2022)	69.76	**79.81**	48.31
FedCoOp (Guo et al. 2023)	69.70	75.23	45.30
Ours	**74.23**	77.35	**54.62**

Table 3: Training on seen dataset (XD-Violence), the result of generalization to unseen datasets (AUC (%) on XD-Violence and ShanghaiTech).

Method	Source	Target	
	UCF	XD	SHTech
ZS-CLIP (Radford et al. 2021)	71.20	51.34	53.69
Temporal-CLIP (Ju et al. 2022)	83.89	54.80	54.83
Ours	**84.03**	**55.91**	**60.42**

Table 4: Training on seen dataset (UCF-Crime), the result of generalization to unseen datasets (AP (%) on XD-Violence and AUC (%) on ShanghaiTech).

Baselines There are only a few federated WSVAD methods. We compare the proposed method with: (1) **Zero-shot CLIP** (Radford et al. 2021) with the handcrafted prompt template "a photo of {class}"; (2) **Temporal-CLIP** (Ju et al. 2022), a variant of ZS-CLIP. Temporal-CLIP is equipped with a temporal modeling block based on ZS-CLIP; (3) **FedCoOp** (Guo et al. 2023), a FL variant of CoOp. FedCoOp replaces the handcrafted prompt with a set of learnable vectors in federated learning environment, which is also equipped with a temporal modeling block; (4) **PPVU** (Doshi and Yilmaz 2023), a transformer based federated video understanding method; (5) **CLAP** (Al-Lahham et al. 2024), the SOTA method of federated WSVAD. We re-implement CLAP with features extracted from CLIP.

Comparisons on Different Data Splits

In this section, we evaluate the proposed method on different dataset splits. The results are shown in table 1. The proposed method shows better performance than ZS-CLIP (Radford et al. 2021) with handcrafted prompt and FedCoOp (Guo et al. 2023) with learnable prompt vectors, indicating that the proposed prompt generator is more appropriate for federated learning environment with more flexibility. Moreover, the proposed method achieves better performance compared with existing federated WSVAD. Based on transformer, PPVU (Doshi and Yilmaz 2023) simply trains a classifier in federated environment, which is inferior to our method. CLAP (Al-Lahham et al. 2024) proposed a cluster based strategy to generate pseudo labels. For WSVAD, the generated pseudo labels of CLAP are replaced with the video level labels. More importantly, PPRU and CLAP only consider the visual features and ignore generalization and personalization in federated learning environment. In contrast, our method utilizes the association of vison and language as well as try to make a trade-off between global generalization and local personalization with global and local driven-contexts.

Comparisons on Different Training Settings

In this section, we evaluate the proposed method on different training settings. The results are shown in table 2. We re-implemented some existing methods in different training settings for comparison. On the one hand, it is hoped that the performance of federated training is close to centralized training as soon as possible cause centralized training ensures the global generalization of the model to a certain extent. On the other hand, even though the privacy is protected during local training, if all clients only train own model with local dataset, the individual local model usually shows poor performance in global test data. As a compromise solution, federated training ensures global performance by aggregating local model as well as the privacy of local training data. Compared with local training, the proposed method in federated training achieves a gain of 2.89% AUC and 10.44% AP on UCF-Crime and XD-Violence, respectively. Besides, the proposed prompt generator is driven by both global and local contexts, which is specially designed for federated learning environment to balance generalization and personalization. Therefore, the proposed method shows limitation to some extent when centralized training.

Generalization to Unseen Datasets

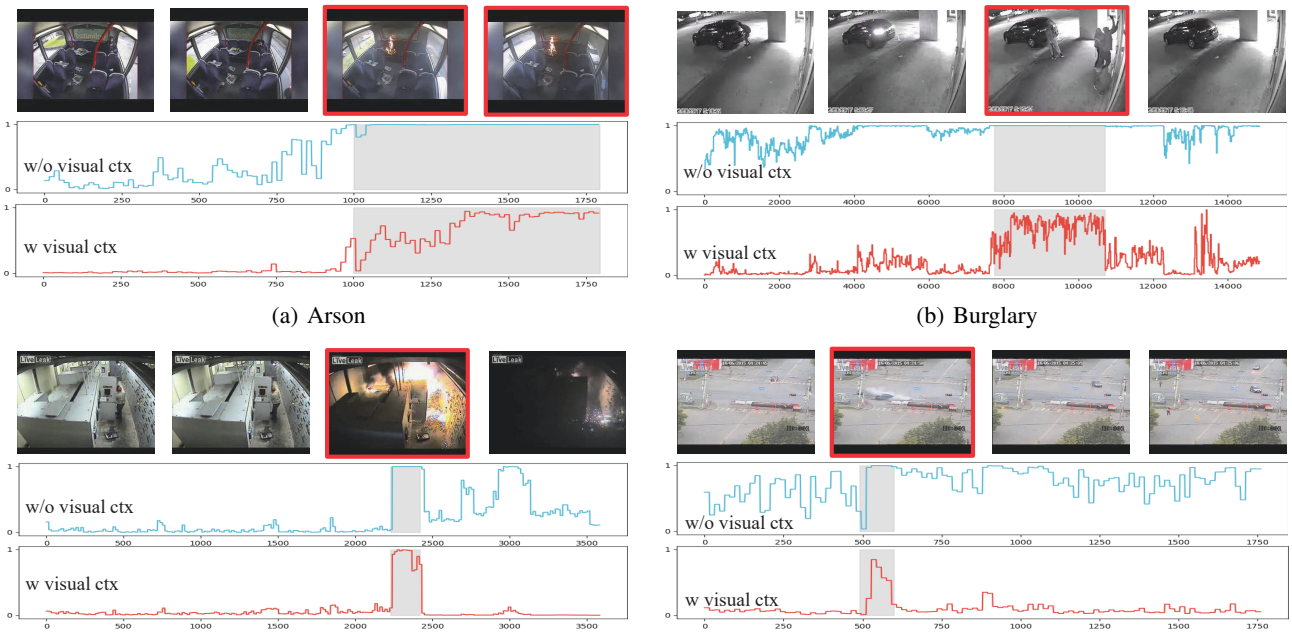
In this section, we evaluate the generalization ability of the proposed method. Specifically, we train the model on UCF-Crime or XD-Violence datasets in federated learning environment, and test the generalization ability on unseen datasets. The results are shown in table 3 and table 4. For unseen dataset ShanghaiTech, the proposed method achieves 54.62% AUC and 60.42% AUC when trained on XD-Violence and UCF-Crime, respectively, which outperforms ZS-CLIP and Temporal-CLIP by 3.83% and 9.32% on average, respectively. The generalization ability can be attributed to two aspects. On the one hand, the pre-trained CLIP provides robust feature representation with abundant semantic information. On the other hand, the proposed prompt generator is driven by global and local contexts. And the global context represents the task-related information, which provides the model with the ability of generalization to unseen datasets.

Ablation Studies

In this section, we focus on evaluating the effectiveness of each component in our method. The results are shown in Table 5. Firstly, we investigate the difference between local text context and global text context. Similar to FedTPG (Qiu et al. 2024), when only providing texts owned by the current client, the average performance decreases by 0.28% AUC and 4.16% AP on UCF-Crime and XD-Violence, respectively. The reason is that global contexts provide full task related information while the local contexts are relatively limited and cannot provide information about the global task. Additionally, it should be noted that the local contexts are identical to global contexts when the datasets are randomly

Local text context	Global text context	Local visual context	UCF Random	UCF Event	UCF Scene	UCF AVG	XD Random	XD Event	XD Scene	XD AVG
✔			83.56	83.87	83.45	83.62	73.59	69.40	71.75	71.58
✔		✔	84.06	84.42	83.82	84.10	75.32	67.72	70.01	71.02
	✔		83.56	84.64	83.89	84.36	73.59	72.39	73.89	73.29
	✔	✔	**84.06**	**85.07**	**84.03**	**84.38**	**75.32**	**74.23**	**75.99**	**75.18**

Table 5: Effectiveness of the modules of the proposed prompt generator.



(a) Arson (b) Burglary

(c) Explosion (d) Car accident

Figure 3: Visualization of anomaly confidence on UCF-Crime of four anomalies: Arson, Burglary, Explosion and Car accident. The blue lines represent the model without local visual context. And the red ones represent the model with local visual context. The gray areas represent ground-truth of anomalies.

split. Secondly, we consider the effectiveness of local visual contexts. When replacing the local visual contexts with a set of learnable vectors, the average performance produces a drop of 1.89% AP. This is because the model cannot fit the local dataset well in absence of the guidance of local visual context.

Qualitative Analyses

We show the qualitative visualization result on UCF-Crime in Figure 3. Specifically, the blue curves represent the situation where the local visual contexts are removed, and the red ones represent the situation with both local visual contexts and global text contexts. The gray areas refer to the ground-truth of the anomalies in the videos. As we can see, in the absence of local visual contexts, the model sometimes predicts relatively high anomaly confidence for the normal frames. In contrast, with the guidance of local visual contexts, the aggregated model shows better performance with a lower false alarm rate, which proves the effectiveness of the local visual contexts again.

Conclusion

In this work, we propose Global and Local Context-driven Federated Learning, a new paradigm for privacy protected video anomaly detection. Specifically, we utilize the vision-language association of CLIP to detect the anomalies in the videos. Additionally, to maintain the powerful capabilities of CLIP in federated learning environment, we replace the handcrafted text prompt with dynamically generated prompt. Moreover, to balance the global generalization and local personalization, we guide the generation of prompts with local visual contexts and global text contexts. Driven by two contexts, the generated prompts are more effective in the federated environment. In the future, we will further explore the privacy protected VAD.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62301621), Shenzhen Science and Technology Program (No. 20231121172359002, No. 202412023000572), and Guangdong Basic and Applied Basic Research Foundation (No. 2023B0303000010).

References

Al-Lahham, A.; Zaheer, M. Z.; Tastan, N.; and Nandakumar, K. 2024. Collaborative Learning of Anomalies with Privacy (CLAP) for Unsupervised Video Anomaly Detection: A New Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12416–12425.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Cui, C.; Ren, Y.; Pu, J.; Li, J.; Pu, X.; Wu, T.; Shi, Y.; and He, L. 2024a. A novel approach for effective multi-view clustering with information-theoretic perspective. *Advances in Neural Information Processing Systems*, 36.

Cui, C.; Ren, Y.; Pu, J.; Pu, X.; and He, L. 2023. Deep multi-view subspace clustering with anchor graph. *arXiv preprint arXiv:2305.06939*.

Cui, T.; Li, H.; Wang, J.; and Shi, Y. 2024b. Harmonizing Generalization and Personalization in Federated Prompt Learning. *arXiv preprint arXiv:2405.09771*.

Doshi, K.; and Yilmaz, Y. 2023. Privacy-preserving video understanding via transformer-based federated learning. In *2023 IEEE Conference on Dependable and Secure Computing (DSC)*, 1–8. IEEE.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14084–14093.

Dutto, M.; Berton, G.; Caldarola, D.; Fanì, E.; Trivigno, G.; and Masone, C. 2024. Collaborative Visual Place Recognition through Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 4215–4225.

Guo, T.; Guo, S.; Wang, J.; Tang, X.; and Xu, W. 2023. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2020. Federated visual classification with real-world data distribution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 76–92. Springer.

Huang, C.; Liu, C.; Wen, J.; Wu, L.; Xu, Y.; Jiang, Q.; and Wang, Y. 2022a. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Transactions on Cybernetics*, 54(5): 3197–3210.

Huang, C.; Liu, Y.; Zhang, Z.; Liu, C.; Wen, J.; Xu, Y.; and Wang, Y. 2022b. Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection. In *Proceedings of the 30th ACM international conference on multimedia*, 307–315.

Huang, C.; Wen, J.; Xu, Y.; Jiang, Q.; Yang, J.; Wang, Y.; and Zhang, D. 2022c. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE transactions on neural networks and learning systems*, 34(11): 9389–9403.

Huang, C.; Wu, Z.; Wen, J.; Xu, Y.; Jiang, Q.; and Wang, Y. 2021. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Transactions on Industrial Informatics*, 18(8): 5171–5179.

Jin, P.; Takanobu, R.; Zhang, W.; Cao, X.; and Yuan, L. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13700–13710.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, 105–124. Springer.

Kim, J.; Ku, Y.; Kim, J.; Cha, J.; and Baek, S. 2024. VLM-PL: Advanced Pseudo Labeling Approach for Class Incremental Object Detection via Vision-Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4170–4181.

Li, H.; Huang, W.; Wang, J.; and Shi, Y. 2024. Global and local prompts cooperation via optimal transport for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12151–12161.

Ling, Y.; Chen, J.; Ren, Y.; Pu, X.; Xu, J.; Zhu, X.; and He, L. 2023. Dual label-guided graph refinement for multi-view graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8791–8798.

Liu, Y.; Huang, A.; Luo, Y.; Huang, H.; Liu, Y.; Chen, Y.; Feng, L.; Chen, T.; Yu, H.; and Yang, Q. 2020. Fedvision: An online visual object detection platform powered by federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13172–13179.

Luo, J.; Khandelwal, S.; Sigal, L.; and Li, B. 2024. Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4029–4040.

Luo, W.; Liu, W.; and Gao, S. 2017. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International conference on multimedia and expo (ICME)*, 439–444. IEEE.

Lv, H.; Yue, Z.; Sun, Q.; Luo, B.; Cui, Z.; and Zhang, H. 2023. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8022–8031.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of

deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Qiu, C.; Li, X.; Mummadi, C. K.; Ganesh, M. R.; Li, Z.; Peng, L.; and Lin, W.-Y. 2024. Federated text-driven prompt generation for vision-language models. In *The Twelfth International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Su, S.; Yang, M.; Li, B.; and Xue, X. 2024. Federated adaptive prompt tuning for multi-domain collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15117–15125.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Wei, G.; Wang, F.; Shah, A.; and Chellappa, R. 2023. Dual prompt tuning for domain-aware federated learning. *arXiv preprint arXiv:2310.03103*.

Wu, L.; Huang, C.; Zhao, S.; Li, J.; Zhao, J.; Cui, Z.; Yu, Z.; Xu, Y.; and Zhang, M. 2023. Robust fall detection in video surveillance based on weakly supervised learning. *Neural networks*, 163: 286–297.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. 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*, 322–339. Springer.

Wu, P.; Zhou, X.; Pang, G.; Sun, Y.; Liu, J.; Wang, P.; and Zhang, Y. 2024a. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18297–18307.

Wu, P.; Zhou, X.; Pang, G.; Yang, Z.; Yan, Q.; WANG, P.; and Zhang, Y. 2024b. Weakly Supervised Video Anomaly Detection and Localization with Spatio-Temporal Prompts. In *ACM Multimedia 2024*.

Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; and Zhang, Y. 2024c. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6074–6082.

Wu, W.; Sun, Z.; and Ouyang, W. 2023. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2847–2855.

Xu, J.; Ren, Y.; Shi, X.; Shen, H. T.; and Zhu, X. 2023. UNTIE: Clustering analysis with disentanglement in multi-view information fusion. *Information Fusion*, 100: 101937.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 736–753. Springer.

Yang, F.-E.; Wang, C.-Y.; and Wang, Y.-C. F. 2023. Efficient model personalization in federated learning via client-specific prompt generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19159–19168.

Yang, Z.; Liu, J.; and Wu, P. 2024. Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18899–18908.

Yun, S.; Park, S. H.; Seo, P. H.; and Shin, J. 2023. Ifseg: Image-free semantic segmentation via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2967–2977.

Zhang, D.; Huang, C.; Liu, C.; and Xu, Y. 2022. Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. *IEEE Signal Processing Letters*, 29: 1197–1201.

Zhang, J.; Qing, L.; and Miao, J. 2019. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, 4030–4034. IEEE.

Zhao, H.; Du, W.; Li, F.; Li, P.; and Liu, G. 2023. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhao, S.; Zhang, Z.; Schulter, S.; Zhao, L.; Vijay Kumar, B.; Stathopoulos, A.; Chandraker, M.; and Metaxas, D. N. 2022. Exploiting unlabeled data with vision and language models for object detection. In *European conference on computer vision*, 159–175. Springer.

Zhong, J.-X.; Li, N.; Kong, W.; Liu, S.; Li, T. H.; and Li, G. 2019. 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*, 1237–1246.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3769–3777.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.