ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Memory Module Guided for Federated Weakly-Supervised Abnormal event detection in videos

Yuhan Zhu[1,*]

[1] School of Computer Science and Technology (National Exemplary Software School), 400065, Chongqing, China

SUMMARY: *Abnormal event detection in videos (VAD) serves as a key component in applications like intelligent surveillance and video content review, particularly in the context of growing concerns over video sequences privacy. Traditional anomaly detection methods typically require centralized training data, which poses privacy risks when dealing with sensitive data, such as surveillance videos or traffic accident footage. In practical scenarios, video sequences are often distributed across multiple institutions or devices, and due to privacy protection regulations, these videos cannot be shared. Therefore, how to efficiently per- form abnormal event detection in videos while ensuring privacy still constitutes a formidable challenge. To alleviate this issue, we propose a novel framework Federated Weakly Supervised Abnormal event detection in videos (Fed-WSVAD), which leverages federated learning to solve the problem of data privacy between institutions. Specifically, we combine the massive visual-textual pre-trained model CLIP with federated learning, introducing a global-level and local context-driven dynamic text prompt generator. This generator creates text prompts that enhance global-level generalization while maintaining local personalization dependent on the unique data characteristics of each client, consequently improving the effectiveness of abnormal event detection in videos. To further enhance the generalization ability of the model, we incorporate a memory module that stores and updates feature prototypes for normal and anomalous samples. This allows knowledge sharing and transfer across clients without requiring data sharing. Our model not only reduces data transmission and storage risks due to privacy concerns but also improves anomaly detection performance through the introduction of compactness and separateness losses. We implemented comprehensive experiments on the XD-Violence dataset and UCF-Crime dataset datasets, and the results show that compared to traditional federated learning methods, our Fed-WSVAD approach significantly outperforms in both global-level generalization and local personalization. This method effectively balances privacy protection and performance optimization, enabling the training of efficient abnormal event detection in videos models without sharing sensitive data. This research provides a novel solution for privacy-preserving abnormal event detection in videos and demonstrates the potential and practical application value of federated learning in weakly supervised abnormal event detection in videos. In the future, we will further explore how to optimize this method in more complex scenarios to enhance its feasibility and robustness in practical-scenario applications.*

KEYWORDS: *Abnormal event detection in videos, Federated Learning, Dual Memory, Weakly-Supervised*

*2023214370@stu.cqupt.edu.cn
https://doi.org/10.65102/is20261240
Published: 10 June, 2026

1 Introduction

VAD is an indispensable task in computer vision, with applications in intelligent surveillance, security monitoring, and video content review [1]. The goal of VAD is to detect abnormal events or behaviors in videos, such as accidents, crimes, or unexpected behaviors in surveillance footage. Dependent on supervision intensity, VAD can be classified into semi-supervised, weakly supervised, self-supervised, fully supervised and unsupervised methods: Fully supervised VAD achieves high accuracy but incurs prohibitive annotation costs, unsupervised VAD requires no annotations yet has limited generalization, while weakly supervised VAD (WSVAD), typically formulated under the multiple instance learning (MIL) framework, where only coarse-grained video annotations are required in place of VAD is an indispensable task in computer vision [2]. The goal of VAD is to detect abnormal events or behaviors in videos, such as accidents, crimes, or unexpected behaviors in surveillance footage. Dependent on supervision intensity, VAD can be classified into semi-supervised, weakly supervised, self-supervised, fully supervised and unsupervised methods: Fully supervised VAD achieves high accuracy but incurs prohibitive annotation costs, unsupervised VAD requires no annotations yet has limited generalization, while weakly supervised VAD (WSVAD), typically formulated under the MIL framework, where only coarse-grained video annotations are required in place of fine-grained frame labels, greatly reducing the annotation cost and making these methods more practical for practical-scenario applications [3].

However, despite the progress made in WSVAD, existing methods face formidable drawbacks when dealing with the privacy-sensitive nature of video sequences. In many practical scenarios, such as law enforcement or surveillance, video sequences cannot be shared due to privacy concerns. For example, traffic accident footage and police body cam recordings contain sensitive information about individuals that cannot be disseminated publicly or stored centrally [4]. As a result, there is a pressing need for anomaly detection models that can effectively work in a decentralized manner, protecting privacy while still achieving high performance.

FL has evolved into a promising privacy-preserving solution to alleviate these drawbacks. FL allows multiple institutions or devices to collaboratively train shared models, and this distributed processing method ensures that sensitive data is retained on local devices, consequently achieving privacy data protection. However, there are many difficulties in directly applying FL to abnormal event detection in videos, as it requires generalization of different video sequences from multiple sources [5].

To alleviate these issues, we propose a new algorithm framework called Fed-WSVAD. The proposed solution integrates federated learning methods and weakly supervised abnormal event detection in videos methods, and uses a dynamic text prompt generator dependent on the CLIP model to generate context driven prompts through CLIP's visual language association, helping the model learn global-level anomaly categories and client specific patterns, consequently ensuring that the proposed model can generalize data across different institutions [6].

The main innovative work of this article is as follows: (1) A privacy preserving federated learning framework for weakly supervised abnormal event detection in videos is proposed, which ensures the privacy and security of video sequences while achieving generalized model training across multiple clients. (2) Designed a dynamic prompt generator driven by global-level and local contexts, which improves the model's generalization ability while maintaining client specific data features. (3) We demonstrate the effectiveness of our approach through comprehensive experiments on two widely-used datasets, UCF-Crime dataset and XD-Violence dataset, where our method outperforms existing federated learning-based anomaly detection methods.

In the following sections, we first provide a review of related work in weakly supervised abnormal event detection in videos and federated learning. Then, we present our proposed framework in detail, followed by a comprehensive experimental evaluation and conclusion.

2 Related Work

We review the existing approaches in VAD, focusing on weakly supervised abnormal event detection in videos (WSVAD), federated learning (FL) for abnormal event detection in videos, and related methods that combine federated learning with anomaly detection [7].

2.1 Abnormal event detection in videos

Abnormal event detection in videos aims to identify abnormal events or behaviors in videos. It can be classified dependent on the level of supervision used in the training process. Fully supervised methods require frame-wise annotations, which yield high performance but are costly in terms of data labeling. In contrast, unsupervised methods do not require any annotations and rely on learning patterns from the data itself, but they often struggle with generalization. Weakly supervised methods strike a balance between performance and annotation cost by using clip-level labels in place of frame-wise annotations. Among these, WSVAD typically employ a MIL framework where the task is to predict the anomaly confidence for each video dependent on the clip-level label [8].

Recently, significant progress has been made in WSVAD, where various methods have utilized deep learning-based temporal modeling, attention mechanisms, and self-supervised learning. The above-mentioned methods have improved the detection performance of multiple video anomaly datasets such as UCF crime and XD violence, but there is still significant room for improvement in alleviateing privacy protection and data isolation issues.

2.2 Federated Learning for Abnormal event detection in videos

FL is mainly used for training machine learning models with dispersed data sources and can effectively protect data privacy, making it a promising privacy-preserving solution. For specific application scenarios of abnormal event detection in videos, the FL method enables multiple institutions or devices to obtain a unified and effective model through collaborative training, without high requirements for sharing raw video sequences, which is extremely important for privacy sensitive fields such as law enforcement, healthcare, and surveillance [9].

Fed WSVAD is the first to combine the FL method with visual language cue learning methods, proposing a new method for weakly supervised abnormal event detection in videos. It utilizes CLIP's visual language association and introduces a global-level local context driven cue generator, which ensures improved global-level generalization performance while adapting to client specific data. It also allows for cross institutional collaboration of video detection data without sharing sensitive video sequences, effectively alleviateing privacy protection issues.

CLAP is an unsupervised federated VAD method, which has the advantage of adapting to the heterogeneous characteristics of video client data and utilizing a three-stage process of knowledge separation, accumulation, and feedback to train a detection model for video anomaly data, with high detection robustness. However, when dealing with complex and heterogeneous scenarios, the effectiveness of the CLAP algorithm decreases because it heavily relies on the generated pseudo labels, and the accuracy of the generated pseudo labels cannot meet the corresponding requirements.

Other methods, such as DLPP, have also proposed frame- works for semi-supervised federated abnormal event detection in videos [10]. DLPP alleviates drawbacks like non-IID

data and data imbalance in federated learning by introducing the Gradient Dynamic Update (GDU) and Client Reputation Weighting (CRW) modules. While these methods improve the fairness and stability of federated training, they do not alleviate system heterogeneity and the privacy issues arising from malicious client attacks.

2.3 Memory Mechanisms in Federated Learning

Memory mechanisms have been introduced to federated learning to enable long-term knowledge transfer between clients. Recent works such as FedVAD and Luo et al. have proposed models that integrate memory networks with federated learning to improve generalization. FedVAD, for instance, incorporates federated framework for visual consistency clustering with GPT-driven adaptive cross-client semantic-enhanced distillation, allowing it to alleviate data heterogeneity issues and improve communication efficiency in weakly supervised and unsupervised scenarios [11].

Our work builds on these efforts by introducing a dynamic memory mechanism that combines both normal and abnormal memory banks. This approach facilitates the sharing of long-term anomaly knowledge between clients without violating privacy. We employ K-means clustering to aggregate memory prototypes from different clients, enabling cross-client knowledge transfer under privacy protection.

2.4 Drawbacks and Gaps

Although existing federated VAD learning methods have achieved certain results, there are still some issues that need to be alleviated [12]: (1) Fed WSVAD and other methods have not completely solved the problem of data heterogeneity in video clients, which can lead to high requirements for the proposed algorithm's data distribution differences in video clients. If the data distribution differences are significant, it will directly cause a decline in the effectiveness of the learning model; (2) CLAP and other methods rely heavily on pseudo labels, resulting in poor adaptability to complex environments; (3) The existing methods such as DLPP have not designed a good data transmission and storage mechanism, and their adaptability to the transfer mode of long-term abnormal data between video clients is poor.

For the above analysis problem, the proposed method combines dynamic storage mechanism and weakly supervised video anomaly federated learning detection method, which ensures the security of privacy data while obtaining a more robust detection model, providing a more efficient and adaptive solution for video anomaly joint detection.

3 Methodology

This section proposes a Federated Weakly Supervised Abnormal event detection in videos (WSVAD) algorithm model that integrates federated learning methods with dynamic storage methods to significantly enhance the model's global-level generalization and local personalization capabilities, while protecting client data privacy.

3.1 Overall Framework Overview

Let $\boldsymbol{F}_t$ denote the feature vector of the *t*-th frame in a video. The memory at client *i* is updated dependent on the resemblance between the contemporary feature and stored memory items, as follows [13]:

$$\mathcal{M}_i^{(t+1)} = \mathcal{M}_i^{(t)} + \alpha \cdot \text{Update}\left(\mathcal{M}_i^{(t)}, f_{\text{features}}\left(\boldsymbol{F}_t\right)\right) \tag{1}$$

where α controls the update rate, and $f_{\text{features}}(\boldsymbol{F}_t)$ represents the feature extraction function for the current frame.

In parallel, the dynamic prompt generator creates text prompts by synergy of global-level anomaly categories C_{global} and local visual features $\boldsymbol{F}_i$. The prompt $\boldsymbol{P}_i$ for client *i* is generated as:

$$\boldsymbol{P}_i = g\left(C_{\text{global}}, \boldsymbol{F}_i\right) \tag{2}$$

where $g(\cdot)$ is the function that generates prompts dependent on both global-level and local contexts.

Federated learning enables the model to be trained without sharing raw data. The model parameters θ_i at each client are updated locally, and only these updates are sent to the server. This aggregation procedure is given by:

$$\theta_{\text{global}}^{(r+1)} = \frac{1}{N}\sum_{i=1}^{N}\theta_i^{(r)} \tag{3}$$

where *N* is the clients' number, and *r* represents the current training round. The aggregated model is then used to continue the training in the next round.

The overall process of the proposed method is shown in Figure 1, where different clients have videos from different scenes and collaborate to train the model without sharing raw data.

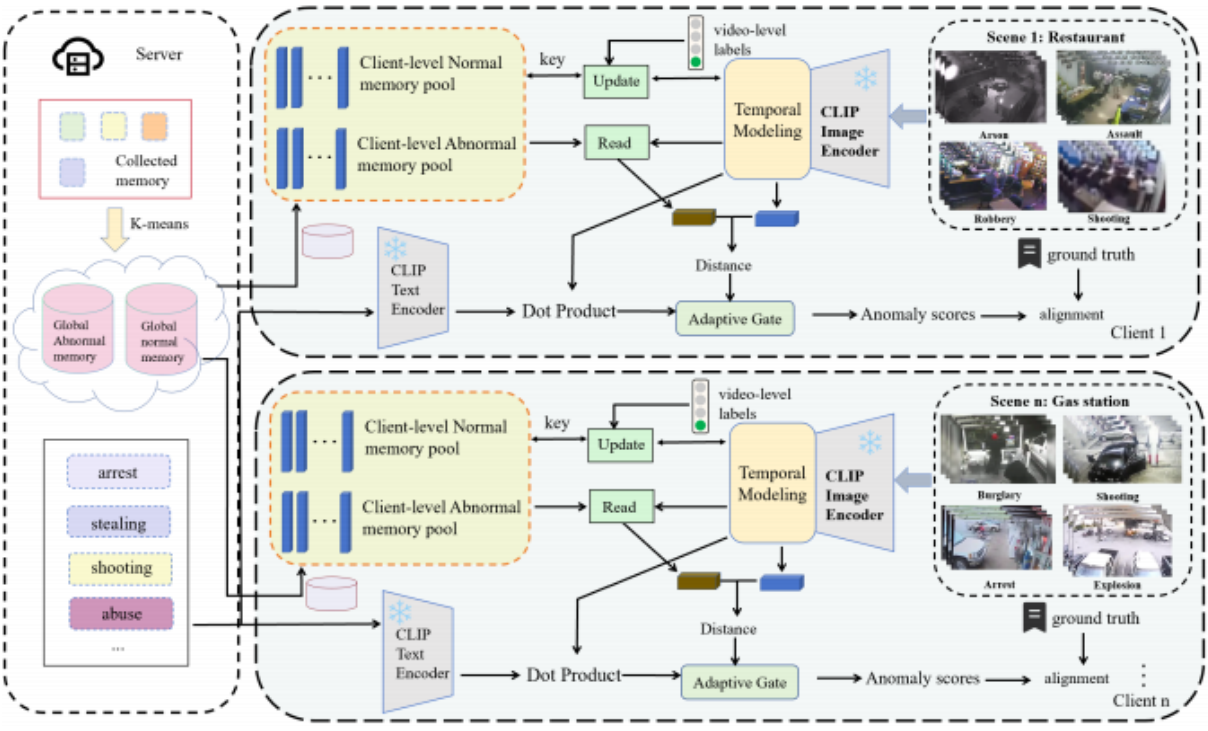

Figure 1: The proposed framework

3.2 Client-Side Memory Update

The memory is up-dated iteratively dependent on the resemblance between the current video frame features and the stored memory items [14]. Specifically, the features extracted from the video frames are passed through a temporal modeling block, which captures dependencies

between frames. A dynamic prompt is generated that combines both global-level anomaly categories and the local context of the video frames, guiding the model to integrate both generalizable knowledge and personalized data. The update process can be formulated as:

$$\mathcal{M}_i^{(t+1)} = \mathcal{M}_i^{(t)} + \alpha \cdot \text{Update}\left(\mathcal{M}_i^{(t)}, f_{\text{features}}\left(X_t\right)\right) \tag{4}$$

where $\mathcal{M}_i^{(t)}$ represents the memory at time t, and α is a learning rate controlling the update speed.

The memory update is driven by two components: $\mathcal{L}_{\text{compact}} = CompactnessLoss$, $\mathcal{L}_{\text{separate}} = SeparatenessLoss$, which ensure that memory items are both compact (similar items are grouped together) and distinct (dissimilar items are kept apart).

3.3 Server-Side Memory Aggregation

On the server side, the memory states of each client are aggregated to form a global-level memory state [15]. The aggregation procedure accounts for the varying sizes of client datasets by weighting the contributions of each client in accordance with the scope of its dataset. Formally, the global-level memory aggregation procedure can be written as:

$$\mathcal{M}_g = \sum_{i=1}^{N} w_i \mathcal{M}_i \tag{5}$$

where w_i represents the weight of client i, typically proportional to the scope of its dataset, and $\mathcal{M}_i$ is the memory of client i.

3.4 Loss Functions

The total loss is the weighted sum of three components: compactness loss, separateness loss, and cross-entropy loss. The total loss function can be expressed as [16]:

$$\mathcal{L} = \lambda_{\text{compact}} \mathcal{L}_{\text{compact}} + \lambda_{\text{separate}} \mathcal{L}_{\text{separate}} + \mathcal{L}_{\text{CE}} \tag{6}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss used for anomaly classification. $\mathcal{L}_{\text{compact}}$ encourages memory items to be compact. $\mathcal{L}_{\text{separate}}$ ensures that memory items are distinct.

Here, λ_{compact} and $\lambda_{\text{separate}}$ are hyperparameters controlling the relative importance of the compactness and separateness losses.

3.5 Federated Training Process

The federated training process proceeds in multiple rounds. In each round, the server distributes the global-level model parameters and memory state to all clients. Each client performs local training on its private dataset and updates both its model and memory. The update and aggregation procedure are formally described as:

$$\theta^{(r+1)} = \frac{1}{N}\sum_{i=1}^{N} \theta_i^{(r)}, \ \mathcal{M}^{(r+1)} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{M}_i^{(r)} \tag{7}$$

where $\theta_i^{(r)}$ and $\mathcal{M}_i^{(r)}$ represent the model parameters and memory state of client i in round r, respectively.

The detailed workflow of federated training and memory interaction is shown in Figure 2.

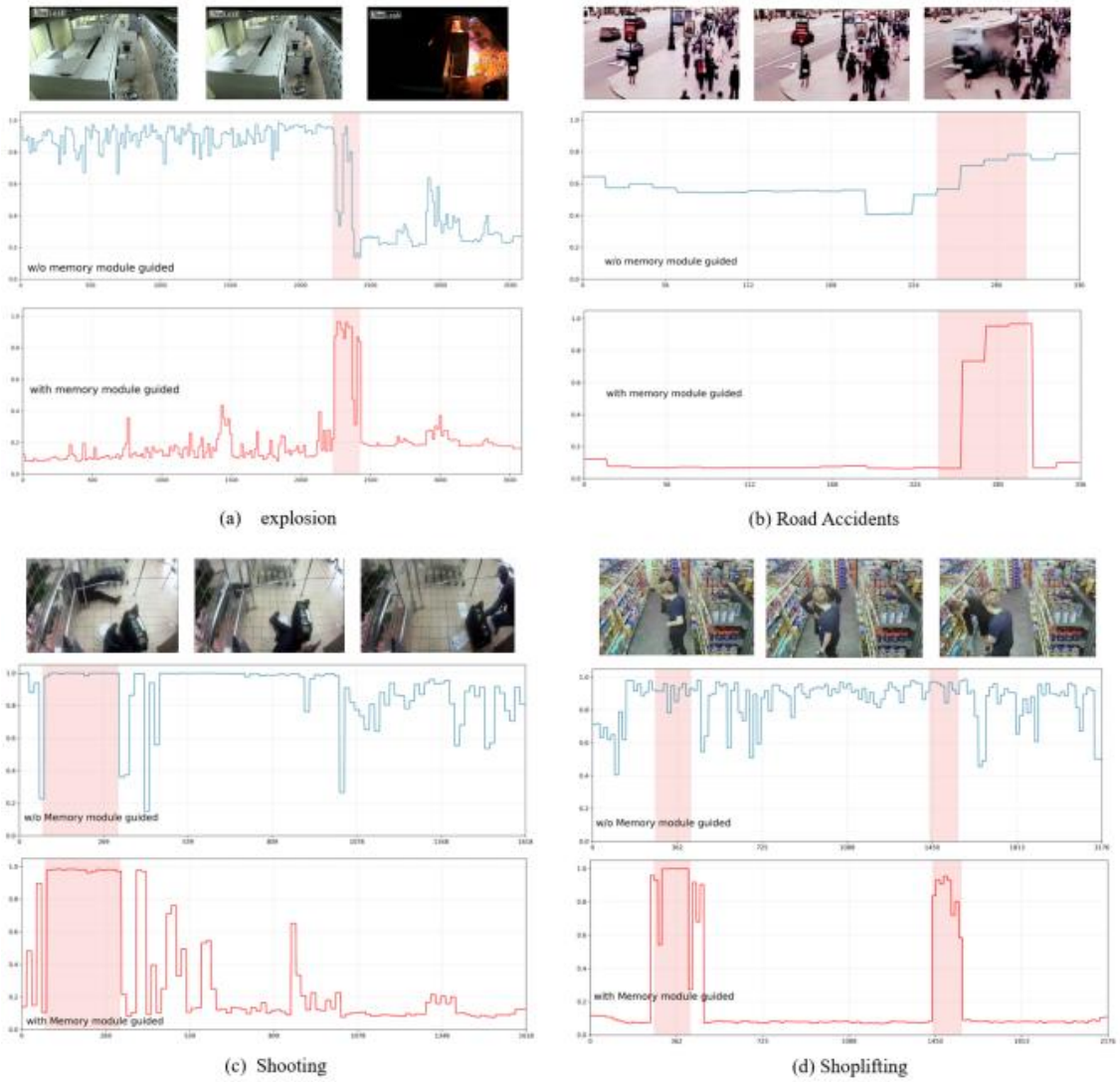

Figure 2: Enter Caption

To verify the effectiveness of the proposed method, comparative experiments were implemented on the UCF crime and XD violence datasets under three data segmentation strategies. The experimental results are shown in Table 1.

Table 1: Comparisons on three data split strategies

Method	**Venue**	**Feature**	**Random**		**Event**		**Scene**	
			UCF	XD	UCF	XD	UCF	XD
ZS-CLIP	ICML2021	CLIP	71.20	51.34	71.20	51.34	71.20	51.34
Temporal-CLIP	ECCV2022	CLIP	83.69	72.07	80.65	69.76	83.89	70.26
FedCoOp	TMC2023	CLIP	84.03	71.80	84.72	69.70	85.36	73.37
PPVU	DSC2023	TimeSformer	82.90	-	-	-	-	-
CLAP	CVPR2024	CLIP	84.99	68.60	84.75	66.57	84.57	70.54
Fed-WSVAD	AAAI2025	CLIP	84.06	75.32	85.07	74.23	84.03	75.99
Ours		CLIP	**84.29**	**79.64**	**85.64**	**74.51**	**84.35**	**80.29**

3.6 Abnormality degree Calculation

In our approach, the abnormality degree for each video frame is calculated by synergy of two key components: the CLIP-based feature alignment score and the distance between the current video frame's features and the memory items for both normal and abnormal categories [17].

Formally, the abnormality degrees anomaly for a given video frame *Ft* is calculated as the weighted fusion of the CLIP alignment score and the memory-based distance score:

$$S_{\text{anomaly}} = \lambda_1 \cdot S_{\text{clip}}(\boldsymbol{F}_i) + \lambda_2 \cdot \left[\text{Dis}(\boldsymbol{F}_i, \mathcal{M}_{\text{normal}}) - \text{Dis}(\boldsymbol{F}_i, \mathcal{M}_{\text{abnormal}})\right] \tag{8}$$

where, $S_{\text{clip}}(\boldsymbol{F}_i)$ represents the resemblance score between the current frame's feature $\boldsymbol{F}_i$ and a reference text prompt generated using CLIP. This score measures the alignment between the frame and global-level anomaly categories. $\text{Dis}(\boldsymbol{F}_i, \mathcal{M}_{\text{normal}})$ and $\text{Dis}(\boldsymbol{F}_i, \mathcal{M}_{\text{abnormal}})$ denote the distances between the frame's feature and the stored memory items for normal and abnormal events, respectively. These distances are computed using a distance function like cosine resemblance or Euclidean distance. λ_1 and λ_2 are the weights for the CLIP alignment score and the memory distance score, respectively, balancing their contributions to the final abnormality degree.

The calculation of $S_{\text{clip}}(\boldsymbol{F}_i)$ is dependent on the cosine resemblance between the video frame's feature and the generated prompt embedding from CLIP:

$$S_{\text{clip}}(\boldsymbol{F}_t) = \frac{\boldsymbol{F}_t \cdot \boldsymbol{P}}{\|\boldsymbol{F}_t\| \|\boldsymbol{P}\|} \tag{9}$$

where $\boldsymbol{P}$ is the generated prompt embedding for the current video, and $\|\cdot\|$ represents the L2 norm.

The distance score between the video frame $\boldsymbol{F}_t$ and the memory items for both normal and abnormal categories is computed as:

$$\text{Dis}(\boldsymbol{F}_t, \mathcal{M}_{\text{normal}}) = \min_{m_i \in \mathcal{M}_{\text{normal}}} \|\boldsymbol{F}_t - m_i\|_2 \tag{10}$$

$$\text{Dis}(\boldsymbol{F}_t, \mathcal{M}_{\text{abnormal}}) = \min_{m_i \in \mathcal{M}_{\text{abnormal}}} \|\boldsymbol{F}_t - m_i\|_2 \tag{11}$$

where $\mathcal{M}_{\text{normal}}$ and $\mathcal{M}_{\text{abnormal}}$ are the memory pools for normal and abnormal events, respectively, and $\|\cdot\|_2$ represents the L2 norm.

Thus, the abnormality degree is a weighted fusion of these two components, where higher scores indicate a higher likelihood of anomaly in the current frame. This combined score allows the model to effectively detect anomalies by leveraging both global-level semantic alignment (via CLIP) and local feature similarities (via memory) [18].

4 Experiments

4.1 Experimental Setup

Experimental datasets: UCF crime dataset and XD violence dataset. Data segmentation strategies: random segmentation, event-based segmentation, and scene-based segmentation, as follows: (1) In the random segmentation strategy, allocate an equal amount of video sequences

to each client. (2) In the event based splitting strategy, assign different types of abnormal video sequences to each client. (3) In scene-based segmentation strategies, client video sequences from specific environments such as street scenes and store monitoring are allocated to simulate practical-scenario applications, ensuring that the model has more contextual diversity during training.

4.2 Comparison with State-of-the-Art Methods

Comparing the proposed method with several state-of-the-art anomaly video detection techniques, the experimental results show that the proposed method consistently outperforms the selected baseline comparison model in different data segmentation strategies. For the selected UCF crime dataset and XD violence dataset, the proposed model algorithm achieves higher AUC and AP scores compared to baseline comparison methods such as ZS-CLIP, Temporary CLIP, FedCoOp, PPVU, and CLAP. The proposed method demonstrates robust performance in both random and context driven distributed video sequences.

4.3 Effect of Different Training Settings

In this experiment, we evaluate the effectiveness of our method under three different training settings: centralized training, local training, and federated training. These settings allow us to assess the effectiveness of federated learning and memory modules in improving the effectiveness of abnormal event detection in videos.

In the centralized training setting, all data from all clients is available to a single model, representing the ideal scenario where no privacy constraints are present. This setup serves as a baseline to compare against the federated and local training settings. In the local training setting, each client trains its own model locally, without exchanging model parameters with others. This setup simulates a scenario where each client has isolated data and is unable to share information with other clients. Finally, in the federated training setting, clients collaborate by exchanging model parameters while keeping their data local, preserving privacy while still benefiting from shared model updates.

Table 2 shows the ablation study data of memory modules on the UCF Crime dataset.

Table 2: Ablation study of memory modules on the UCF-Crime dataset.

Normal Mem.	Abnormal Mem.	Rand.	Event	Scene	Avg.
✓	✓	84.06	85.07	84.03	84.38
		82.55	84.26	83.28	83.36
		81.73	83.55	83.73	83.00
✓	✓	84.29	85.64	84.35	84.76

In accordance with the experimental data shown in Table 2, the model equipped with normal and abnormal memory can obtain the highest average AUC index data on the UCF Crime dataset, which is 84.76%. This is 0.38%, 1.40%, and 1.76% higher than the models equipped with only normal memory, only abnormal memory, and no memory, respectively. This indicates that the dual memory mechanism adopted in this paper effectively enhances the feature representation of video sequences by explicitly modeling normal and abnormal patterns.

In the robustness test of the model, the dual memory mechanism model used in this article achieved an AUC index of 82.47% in the UCF crime dataset and an AP index of 77.47% in the XD violence dataset. In contrast, the effectiveness of local training has significantly decreased, with AUC index dropping to 75.01% and AP index dropping to 64.47%. This indicates that data isolation severely limits the generalization performance of the algorithm model. In a federated

environment, the proposed model method not only outperforms all baseline methods, but also surpasses its own centralized performance, proving that federated collaboration effectively alleviates data heterogeneity and utilizes distributed knowledge to improve detection accuracy, achieving an excellent balance between privacy protection and performance.

Table 3 shows the comparative data of AUC (%) for UCF crimes and AP (%) for XD violence in three training environments.

Table 3: Comparisons on three training settings training.

Mode	Method	UCF	XD
Centralized	Temporal-CLIP	83.72	75.73
	FedCoOp	84.24	76.48
	PPVU	86.30	-
	CLAP	83.85	66.73
	Fed-WSVAD	84.82	71.16
	Ours	82.47	77.47
Local	Temporal-CLIP	81.36	64.66
	FedCoOp	80.97	66.19
	PPVU	76.23	61.35
	CLAP	81.17	64.88
	Fed-WSVAD	75.01	64.47
	Ours		
Federated	Temporal-CLIP	80.65	69.76
	FedCoOp	84.72	69.70
	PPVU	84.75	66.57
	CLAP	85.07	74.23
	Fed-WSVAD	85.64	80.29
	Ours		

In accordance with the experimental data shown in Table 3: (1) In centralized training, the proposed method achieved an AUC index value of 82.47% on the UCF crime dataset and an AP value of 77.47% on the XD violence dataset, which is significantly better than most of the selected baseline comparison methods. (2) In local training, the effectiveness of the proposed method showed a decline, with AUC index dropping to 75.01% and AP index dropping to 64.47%, respectively by 7.46% and 13.00%. This indicates that data isolation seriously damages the model's generalization ability. (3) In the federal environment, the proposed method has an AUC index of 85.64% in the UCF crime dataset and an AP index of 80.29% in the XD violence dataset, both of which are superior to the selected baseline comparison method. The above experimental results indicate that federated collaboration effectively alleviates data heterogeneity, while the proposed dual memory module enhances the feature representation performance of video sequences, achieving a good balance between data privacy protection and detection accuracy.

4.4 Ablation Study

The main purpose of conducting ablation experiments in this section is to evaluate the contribution of each component in the proposed algorithm framework. The ablation experiment results are shown in Table 4. Table 4 shows the ablation research data of the proposed memory module on the UCF crime and XD violence datasets.

Table 4: Ablation study of the proposed memory modules on both XD-Violence dataset and UCF-Crime dataset datasets

without Memory	Client Memory	Server Memory	UCF-Crime dataset				XD-Violence dataset			
			Random	Event	Scene	Average	Random	Event	Scene	Average
✓			84.06	85.07	84.03	84.38	75.32	74.23	75.99	75.18
	✓		84.02	85.50	84.12	84.54	77.07	81.81	76.09	78.32
	✓	✓	84.29	85.64	84.35	84.76	79.64	74.51	80.29	78.14

In accordance with the experimental data in Table 4: (1) On the UCF crime dataset, the average AUC index of the baseline model without any memory is 84.54%; When only client memory is introduced, the average AUC metric of the algorithm model increases to 84.38%, with slight improvements in all sub metrics. The complete model equipped with client and server memory has an average AUC index of 84.76%, which is the best AUC index among the comparison models. It is 0.22% higher than the baseline comparison model without memory and 0.38% higher than the single client memory variant model. (2) On the XD Violence dataset, the average AP score of the baseline comparison model without memory mechanism was 78.32%; When using only client memory, the average AP metric of the model drops to 75.18%; The average AP metric of the complete model with dual memory is 78.14%, which is 3.14% higher than the single client memory variant model.

5 Conclusion

This article proposes a memory module guided federated weakly supervised abnormal event detection in videos method (Fed-WSVAD) to alleviate the two core issues of data privacy protection requirements and cross institutional data silos in abnormal event detection in videos tasks: (1) constructing a federated learning framework for weakly supervised abnormal event detection in videos, integrating multi instance learning with federated averaging strategy, and completing cross client knowledge transfer only through model parameters and memory prototypes, so that sensitive monitoring data is fully retained at local nodes, meeting privacy compliance requirements from the data flow level. (2) Design a global-level local context driven dynamic text prompt generator that adaptively generates prompt features dependent on CLIP visual language alignment ability. While improving the model's cross scene anomaly category generalization, it preserves the personalized distribution characteristics of each client data and adapts to non-independent and identically distributed data conditions in practical-scenario scenarios. (3) Introducing a dual branch memory module to construct feature prototype libraries for normal and abnormal samples, local memory updates are completed on the client side, and global-level memory aggregation is achieved on the server side. By utilizing compactness loss and separability loss to optimize feature representation, the accuracy of anomaly discrimination and model robustness are significantly improved. The experimental results show that the proposed algorithm model has better performance on the UCF Crime and XD Violence standard datasets, achieving a good balance between privacy protection, detection accuracy, and heterogeneous data adaptability. It provides a feasible distributed anomaly detection scheme for privacy sensitive scenarios such as security monitoring, traffic supervision, and public safety.

Future research: (1) Introducing temporal context modeling and anomaly semantic priors to improve the alignment accuracy of fine-grained anomaly prompts in complex scenes, and reduce false detections caused by background interference and category confusion. (2) Design an adaptive weight allocation and incremental update mechanism to improve the training efficiency and convergence speed of the system for long time series and massive clients. (3)

Research is implemented on more stringent heterogeneous scenarios to alleviate practical drawbacks such as differences in camera perspectives, lighting variations, and uneven behavior categories, enhancing the stability of the model in practical-scenario deployment environments. (4) To alleviate the risks of malicious client poisoning attacks, gradient leakage, and model theft, an integrated framework for privacy enhancement and model security is constructed to promote the implementation of our method in high security video surveillance systems.

About the Author

Yuhan Zhu was born in Chongqing, China in 2004. He is currently an undergraduate student at School of Computer Science and Technology (National Exemplary Software School), mainly researching computer vision and video anomaly detection.

References

[1] Acsintoae A, Florescu A, Georgescu M I, et al. Ubnormal: New benchmark for supervised open-set video anomaly detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 20143-20153.

[2] Barbalau A, Ionescu R T, Georgescu M I, et al. SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection[J]. Computer Vision and Image Understanding, 2023, 229: 103656.

[3] Qasim M, Verdu E. Video anomaly detection system using deep convolutional and recurrent models[J]. Results in Engineering, 2023, 18: 101026.

[4] Patwal A, Diwakar M, Tripathi V, et al. An investigation of videos for abnormal behavior detection[J]. Procedia Computer Science, 2023, 218: 2264-2272.

[5] Flaborea A, Collorone L, Di Melendugno G M D A, et al. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 10318-10329.

[6] Ul Amin S, Ullah M, Sajjad M, et al. EADN: An efficient deep learning model for anomaly detection in videos[J]. Mathematics, 2022, 10(9): 1555.

[7] Zanella L, Menapace W, Mancini M, et al. Harnessing large language models for training-free video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 18527-18536.

[8] Ullah W, Ullah A, Hussain T, et al. Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data[J]. Future Generation Computer Systems, 2022, 129: 286-297.

[9] Tholl C, Bickmann P, Wechsler K, et al. Musculoskeletal disorders in video gamers–a systematic review[J]. BMC musculoskeletal disorders, 2022, 23(1): 678.

[10] Zaheer M Z, Mahmood A, Khan M H, et al. Generative cooperative learning for unsupervised video anomaly detection[C]//Proceedings of the IEEE/CVF conference on

computer vision and pattern recognition. 2022: 14744-14754.

[11] Cho M A, Kim T, Kim W J, et al. Unsupervised video anomaly detection via normalizing flows with implicit latent features[J]. Pattern Recognition, 2022, 129: 108703.

[12] Rezaee K, Rezakhani S M, Khosravi M R, et al. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance[J]. Personal and Ubiquitous Computing, 2024, 28(1): 135-151.

[13] Ullah W, Hussain T, Khan Z A, et al. Intelligent dual stream CNN and echo state network for anomaly detection[J]. Knowledge-Based Systems, 2022, 253: 109456.

[14] Pallavicini F, Pepe A, Mantovani F. The effects of playing video games on stress, anxiety, depression, loneliness, and gaming disorder during the early stages of the COVID-19 pandemic: PRISMA systematic review[J]. Cyberpsychology, Behavior, and Social Networking, 2022, 25(6): 334-354.

[15] Cesari M, Heidbreder A, St. Louis E K, et al. Video-polysomnography procedures for diagnosis of rapid eye movement sleep behavior disorder (RBD) and the identification of its prodromal stages: guidelines from the International RBD Study Group[J]. Sleep, 2022, 45(3): zsab257.

[16] Penuelas-Calvo I, Jiang-Lin L K, Girela-Serrano B, et al. Video games for the assessment and treatment of attention-deficit/hyperactivity disorder: a systematic review[J]. European child & adolescent psychiatry, 2022, 31(1): 5-20.

[17] Jiménez-Muñoz L, Peñuelas-Calvo I, Calvo-Rivera P, et al. Video games for the treatment of autism spectrum disorder: A systematic review[J]. Journal of Autism and Developmental Disorders, 2022, 52(1): 169-188.

[18] Kaur A, Noori Hoshyar A, Saikrishna V, et al. Deepfake video detection: challenges and opportunities[J]. Artificial Intelligence Review, 2024, 57(6): 159.