



ARTICLE

Two-Stage Pseudo-Label Refinement for Localization-Aware Weakly Supervised Video Anomaly Detection

Meng-Luen Wu^{*} and Yan-Hsin Lin

Department of Computer Science and Information Engineering, Tamkang University, New Taipei, Taiwan

*Corresponding Author: Meng-Luen Wu. Email: 158769@mail.tku.edu.tw

Received: 25 March 2026; Accepted: 09 September 2026; Published: 27 October 2026

ABSTRACT: Weakly supervised video anomaly detection aims to localize abnormal events in untrimmed videos using only video-level labels, making it applicable in scenarios where fine-grained annotations are unavailable. However, weak supervision often introduces noisy and ambiguous supervisory signals, and existing approaches may still face challenges in balancing temporal modeling and semantic discrimination. To investigate this issue, this paper presents a two-stage pseudo-label learning framework that functionally decouples temporal dependency modeling from semantic refinement. In the first stage, segment-level anomaly scores are generated using multi-scale temporal transformers and dual-branch graph convolutional networks to capture both local temporal dynamics and global contextual dependencies. In the second stage, these scores are treated as soft pseudo-label cues and are further refined through a CLIP-guided multimodal prompt learning strategy, where learnable textual prompts facilitate semantic alignment between visual and textual representations. Experiments on the UCF-Crime dataset show that the proposed framework achieves competitive performance in frame-level binary anomaly detection. Notably, more consistent improvements are observed in localization-aware multi-class evaluation (mAP@IoU), suggesting that the decoupled design may be more beneficial for segment-level anomaly localization than for global anomaly ranking. Rather than claiming broad superiority, the results indicate that separating temporal modeling from semantic refinement provides a structured way to examine how pseudo-label cues interact with multimodal alignment under weak supervision. These findings offer empirical insights into the behavior and limitations of pseudo-label refinement in weakly supervised video anomaly detection.

KEYWORDS: Weakly supervised learning; video anomaly detection; pseudo-label learning; temporal modeling; graph convolutional networks; vision-language models; CLIP; prompt learning

1 Introduction

Intelligent surveillance video systems are increasingly deployed to support public safety and operational monitoring in transportation hubs, industrial sites, and urban areas. A key capability is video anomaly detection (VAD), which aims to localize and recognize rare abnormal events, such as violence, theft, accidents, and explosions, from long untrimmed video streams. Effective VAD depends on spatio-temporal representation learning that captures both appearance cues and temporal evolution. Early work commonly adopted supervised video backbones such as C3D [1] and I3D [2], while more recent architectures with space-time attention, such as VideoSwin [3], have shown strong performance for general video understanding. However, these supervised pipelines typically require large amounts of fine-grained annotations, which are expensive to obtain for surveillance data. The rarity, diversity, and temporal ambiguity of anomalies further make dense frame-level or segment-level annotation difficult to scale.

This annotation bottleneck motivates weakly supervised video anomaly detection, where only video-level labels are available and the model must infer segment-level anomaly scores. Multiple benchmarks have been established to support evaluation, including UCSD Ped1/Ped2 [4], Avenue [5], ShanghaiTech Campus [6], UCF-Crime [7], and XD-Violence [8]. In practical deployments, surveillance systems may also face constraints related to latency, model size, edge readiness, and annotation uncertainty [9,10]. Therefore, the challenge is not only to recognize whether a video contains abnormal events, but also to infer temporally meaningful anomaly cues from coarse supervision.

Recent WSVAD research has increasingly moved beyond coarse video-level anomaly recognition toward fine-grained temporal localization. Under weak supervision, this shift makes pseudo-label reliability particularly important, because temporal localization must be inferred from video-level labels rather than directly supervised segment annotations. Noisy or incomplete pseudo labels may allow a model to classify an abnormal video correctly while still failing to localize abnormal intervals. Therefore, this study treats pseudo-label refinement as a localization-related problem and examines how temporal modeling and semantic refinement interact to support segment-level anomaly localization.

Despite recent progress, weakly supervised VAD still faces important challenges when only coarse video-level labels are available. The ambiguity of weak supervision can propagate noise into segment-level predictions, while insufficient temporal modeling may limit the ability to capture contextual dependencies across video segments. At the same time, directly applying semantic alignment to unreliable segment-level predictions may reinforce noisy pseudo-label signals. These observations suggest that pseudo-label refinement should not be treated as a fully reliable supervision mechanism. Instead, it is necessary to examine how temporally derived anomaly cues are generated, transferred, and adjusted during semantic refinement.

This work proposes a two-stage framework to investigate how temporal pseudo-label generation and semantic refinement interact under weak supervision, with an emphasis on localization-aware behavior rather than broad model superiority. In Stage 1, temporal modeling is used to generate structured segment-level anomaly cues. In Stage 2, these cues are treated as soft pseudo-label signals and refined through CLIP-guided semantic alignment. This staged formulation provides a more explicit way to analyze the interaction between temporal cue generation and multimodal refinement.

Most weakly supervised VAD methods are formulated under the Multiple Instance Learning (MIL) paradigm, where each video is treated as a bag of segments and the video label indicates whether any segment contains an anomaly. UCF-Crime has become one of the representative benchmarks for weakly supervised VAD using video-level supervision. MIST [11] iteratively refines pseudo labels via self-training, while RTFM [12] leverages temporal feature magnitudes to strengthen discriminative learning. Li et al. [13] further incorporate transformers for multi-sequence learning under weak supervision, and more recent approaches such as UMIL [14] aim to reduce bias and uncertainty. These works indicate that robust pseudo supervision is particularly important when training data is limited and annotation resources are constrained.

Another key difficulty lies in temporal reasoning. Anomalies can be short-lived (such as explosions) or extended (such as fights), requiring models to capture both short-range cues and long-range dependencies. Conventional temporal models, such as non-local neural networks [15], can encode sequential relations but may struggle to balance patterns across multiple temporal scales. Transformer-based video architectures, including TimeSformer [16] and ViViT [17], offer an effective mechanism to model long-range temporal dynamics, motivating multi-scale temporal modeling for weakly supervised VAD under limited supervisory signals.

In parallel, vision-language models provide a complementary direction by introducing semantic priors through multimodal alignment. CLIP [18], pre-trained on large-scale image-text pairs, aligns visual

representations with textual embeddings and supports knowledge transfer to downstream recognition. Prompt learning methods such as CoOp [19], CoCoOp [20], and VPT [21] further adapt CLIP to specific tasks via learnable prompts. Recent studies have explored CLIP for weakly supervised anomaly detection, including VadCLIP [22] and AnomalyCLIP [23], suggesting that multimodal alignment can improve semantic discrimination for complex video events.

Based on the above observations, this study focuses on weakly supervised video anomaly detection under limited annotation conditions, where only coarse video-level labels are available. Here, limited annotation refers to the absence of dense frame-level or segment-level labels, rather than to a lightweight architectural design. Rather than aiming to establish a universally superior anomaly detection model, we examine how temporally derived segment-level anomaly cues interact with subsequent semantic refinement.

To this end, we adopt a two-stage design that separates temporal dependency modeling from multimodal alignment, enabling a clearer analysis of how pseudo labels are transferred and adjusted across stages. In Stage 1, the temporal modeling module is used to produce segment-level anomaly scores that capture contextual dependencies across time. These scores are not treated as reliable segment annotations, but as noisy intermediate representations that encode temporal saliency for subsequent semantic refinement. In Stage 2, a CLIP-guided multimodal prompt learning mechanism is introduced to refine these cues through semantic alignment, allowing the model to adjust noisy predictions by incorporating vision-language representations. This staged formulation provides a structured way to study the role of pseudo-label refinement, rather than assuming that such refinement always leads to improved performance.

In this design, Stage 1 addresses the temporal modeling limitation of CLIP by generating temporally structured anomaly cues, while Stage 2 uses CLIP-guided prompt learning only to semantically recalibrate these cues rather than to model temporal dependencies from scratch.

The technical contribution of this work is not the introduction of a new transformer, GCN, CLIP encoder, or prompt-learning module. Rather, the contribution lies in the staged pseudo-label refinement formulation, where temporally derived segment scores are preserved as continuous soft cues and then used in CLIP-guided semantic refinement. This formulation makes the transfer of pseudo-label information between temporal modeling and semantic alignment explicit, allowing the interaction between temporal cues and multimodal representations to be analyzed more clearly.

The main contributions of this work are as follows:

- (1) This paper investigates a two-stage framework that functionally decouples temporal pseudo-label generation from semantic refinement, while keeping the two stages sequentially connected through intermediate soft pseudo-label cues. This formulation enables a more explicit analysis of pseudo-label transfer under weak supervision.
- (2) This paper adopts multi-scale temporal modeling with dual-branch graph convolutional networks to generate structured segment-level anomaly cues, which are used as intermediate soft supervision rather than explicit ground-truth annotations.
- (3) This paper examines CLIP-guided prompt-based refinement as a semantic recalibration mechanism for temporally derived pseudo-label cues, rather than as a replacement for temporal modeling.
- (4) Experimental results on UCF-Crime show competitive frame-level anomaly ranking performance and more consistent gains in localization-aware mAP@IoU evaluation. These findings suggest that the staged design is more closely related to localization-oriented segment alignment than to global anomaly ranking under the UCF-Crime setting.

The remainder of this paper is organized as follows. [Section 2](#) reviews related work. [Section 3](#) presents the proposed method. [Section 4](#) reports the experimental results. Finally, [Section 5](#) concludes the paper.

2 Related Work

Recent progress in video anomaly detection (VAD) mainly concerns two practical constraints: limited fine-grained supervision and the need for reliable temporal reasoning in long, untrimmed videos. In real-world surveillance deployments, these challenges are often amplified by limited annotation conditions where dense temporal annotations and extensive manual labeling are not available. This section summarizes representative directions related to our study, including weakly supervised learning with pseudo-label refinement, temporal and graph-based sequence modeling, and vision-language adaptation.

2.1 Weakly Supervised VAD and Pseudo-Label Learning

Weakly supervised VAD is often studied under the multiple instance learning (MIL) setting, where video-level labels are used to train segment-level scoring functions. Sultani et al. [7] introduced UCF-Crime and demonstrated MIL-based training with only video-level supervision. This formulation is particularly useful when fine-grained annotations are unavailable or costly to obtain. Later studies addressed robustness issues under weak supervision, including temporal feature magnitude learning [12], bias mitigation in MIL training [14], and self-paced refinement strategies [24]. Peng et al. [25] further investigated temporal-resolution-aware learning under the MIL formulation, showing that snippet construction and temporal granularity can substantially influence weakly supervised anomaly scoring.

Pseudo-label learning is also used to compensate for missing segment-level annotations by generating auxiliary labels from weak supervision. Two-stage self-training is explored in MIST [11]. Since pseudo labels can be imperfect, subsequent work considers reliability-enhancing mechanisms such as confidence- and completeness-based filtering [26] and attention designs that emphasize anomalous snippets [27]. Text-prompt-based normality guidance has also been investigated to incorporate semantic priors into weakly supervised VAD [28]. These approaches highlight the importance of improving pseudo-label reliability when only coarse supervision is available, particularly in large-scale surveillance environments where annotation resources are limited. Overall, these efforts reflect continuing interest in improving pseudo supervision under weak labels.

2.2 Temporal and Graph-Based Modeling

Temporal modeling is fundamental in VAD because anomalous events can vary in duration and depend heavily on contextual evolution. Prior work has studied temporal regularities in video sequences for anomaly detection [29], and spatio-temporal backbones such as 3D CNNs are widely used to encode appearance and motion cues [1]. For sequential dependency modeling, recurrent models are commonly adopted as baselines, while non-local neural networks extend receptive fields through attention mechanisms [15]. Transformer-based video architectures such as TimeSformer [16] and ViViT [17] further enhance the ability to model long-range temporal relations, which is particularly beneficial under weak supervision when segment-level anomaly patterns must be inferred from coarse labels.

In the weakly supervised setting, Qiu and Hou [30] emphasized that temporal context remains central in WSVAD and that more representative temporal feature learning is required to alleviate redundancy and capture long-range dependencies. Liang et al. [31] further explored spatial-temporal feature fusion to enhance the discriminative power of anomaly representations, highlighting the role of temporal continuity under coarse supervision. In addition, Shao et al. [32] revisited temporal convolution under the MIL framework, demonstrating that explicitly modeled temporal correlations can effectively mitigate the imbalance between rare anomalous instances and dominant normal patterns.

Graph neural networks provide another perspective for representing structured relations. GCNs are introduced for semi-supervised learning by Kipf and Welling [33], and GAT incorporates attention over graph edges for adaptive aggregation [34]. These graph-based methods provide complementary mechanisms for capturing structured dependencies in video sequences, which can improve representation learning when training signals are sparse or weakly supervised. These studies offer general techniques that can be leveraged in video representation learning.

2.3 Vision-Language Adaptation

Vision-language models align visual and textual representations and have been used as general-purpose multimodal feature extractors. CLIP [18] is a representative example trained with contrastive learning on large-scale image-text pairs. Although originally developed for image-text alignment, CLIP has inspired subsequent extensions to video understanding and multimodal event analysis. For example, CLIP-Event [35] incorporates event structural knowledge, including event types and argument roles, to improve fine-grained language-vision alignment. In weakly supervised VAD, VadCLIP leverages CLIP representations for anomaly scoring and uses textual prompts for classification [22], while AnomalyCLIP explores CLIP's latent space for anomaly recognition [23]. Open-vocabulary video anomaly detection further investigates the use of language priors to handle unseen anomaly categories [36]. These multimodal approaches provide additional semantic cues that can complement weak supervision, particularly when annotation resources are limited. Wang et al. [37] further explored relational learning with vision-language models for weakly supervised video anomaly detection, showing that auxiliary semantic supervision and temporal attention can improve robustness in complex scenarios.

Prompt learning is a parameter-efficient strategy to adapt vision-language models to downstream tasks. Representative prompt learning approaches for CLIP-style models include CoOp [19], CoCoOp [20], VPT [21], PromptSRC [38], and MaPL [39], which explore learnable textual or multimodal prompts for parameter-efficient adaptation and improved transferability. Such parameter-efficient adaptation strategies are especially useful in limited annotation settings where retraining large models from scratch is impractical. Li et al. [40] further extended this line of research to weakly supervised video anomaly detection by integrating prompt learning with temporal modeling and cross-modal alignment, demonstrating the effectiveness of combining semantic adaptation and temporal dependency modeling in WSVAD.

More recently, Wang and Chen proposed LEC-VAD [41], a learning event completeness framework for weakly supervised video anomaly detection. LEC-VAD addresses incomplete anomaly localization by modeling both category-aware and category-agnostic vision-language semantics. It further introduces anomaly-aware Gaussian mixture modeling to learn event boundaries and memory-bank-based prototype learning to enrich category-level textual descriptions. This line of work highlights the increasing importance of localization completeness and multimodal semantic modeling in recent WSVAD research.

2.4 Discussion and Motivation

Recent WSVAD studies provide important reference points for this work, particularly in pseudo-label reliability, CLIP-based semantic adaptation, and localization-aware anomaly detection. This trend indicates that the central challenge is no longer only deciding whether a video is abnormal, but also determining when and where abnormal events occur under weak supervision. Zhang et al. [26] investigated the completeness and uncertainty of pseudo labels, showing that pseudo-label quality is critical for weakly supervised anomaly localization. VadCLIP [22] adapts vision-language representations to weakly supervised anomaly detection and demonstrates the usefulness of CLIP-based semantic matching for anomaly scoring. AnomalyCLIP [23] further investigates the latent space of CLIP for video anomaly recognition. Yang et al. [28] introduced

normality-guided textual prompts to improve weakly supervised anomaly detection, while Wang and Chen [41] proposed LEC-VAD, which focuses on event completeness and boundary localization through vision-language semantics.

Compared with these studies, the present work focuses on a different aspect of CLIP-guided WSVAD. Rather than directly using CLIP-based matching as the main anomaly scoring mechanism, explicitly modeling pseudo-label uncertainty, or learning event completeness, this study examines how temporally derived soft pseudo-label cues can be transferred to and refined by a CLIP-guided semantic refinement stage. Therefore, the contribution lies in the staged analysis of temporal pseudo-label generation and semantic refinement, rather than in claiming a universally superior CLIP-based anomaly detector.

In summary, existing approaches to weakly supervised video anomaly detection exhibit complementary strengths but also notable limitations. MIL-based methods provide a practical solution when dense annotations are unavailable; however, they often suffer from noisy and ambiguous supervision, which can degrade the reliability of segment-level predictions. Temporal and graph-based modeling techniques enhance the ability to capture long-range dependencies and structured relations, but they primarily focus on representation learning without explicitly addressing the noise introduced by weak labels. Vision-language approaches incorporate semantic priors through multimodal alignment and improve discriminative capability, yet they may be sensitive to inaccurate or noisy segment-level supervision when directly applied in weakly supervised settings.

These observations suggest that jointly improving pseudo supervision quality while incorporating both temporal reasoning and semantic alignment remains a critical challenge. To address this issue, this work adopts a two-stage pseudo-label learning framework. In the first stage, temporal modeling is used to generate informative yet coarse pseudo labels that capture contextual dependencies. In the second stage, vision-language alignment is leveraged to refine these pseudo labels, enhancing semantic consistency and discrimination. By treating temporal cue generation and semantic refinement as functionally distinct stages, the proposed framework allows pseudo-label cues to be generated first and then semantically re-evaluated, instead of treating pseudo labels as fixed hard supervision.

3 Proposed Method

This section presents the proposed two-stage framework for weakly supervised video anomaly detection. The framework organizes temporal pseudo-label generation and semantic refinement into functionally distinct stages, so that the role of each component can be analyzed more explicitly. In Stage 1, segment-level anomaly scores are generated through multi-scale temporal modeling and graph-based reasoning. These scores are not treated as ground-truth annotations, but as soft pseudo-label cues that reflect temporally salient patterns inferred from video-level supervision. In Stage 2, the pseudo-label cues are further examined through CLIP-guided multimodal prompt learning, where learnable prompts and vision-language alignment provide additional semantic constraints for segment-level prediction.

3.1 Problem Formulation and Notation

Given an untrimmed video V with a video-level weak label $y \in \{0, 1\}$, the objective is to learn segment-level anomaly scores under weak supervision. In the binary setting, the label indicates whether the video contains an anomaly and is used for anomaly scoring. In the semantic refinement stage, category names, including 13 anomaly categories and the normal class, are used to construct CLIP-guided prompts for class-aware segment scoring. Segment-level temporal annotations and anomaly categories are used only for evaluation, including frame-level AUC and localization-aware mAP@IoU.

For clarity, the notations used in this paper are summarized in [Table 1](#).

Table 1: Definition of notations.

Notation	Description
v_m	Visual feature map after multi-scale temporal modeling
v_l	Local temporal features from the transformer encoder
v_p	Visual prompt derived from segment-level visual representation
A_s	Adjacency matrix based on cosine similarity
A_d	Adjacency matrix based on temporal distance
t_p	Textual feature after applying learnable prompts
τ	Temperature scaling factor for softmax
k	Number of top segments used in pseudo-label generation
T	Number of video segments
D	Embedding dimension of visual/textual features
$M_{s,c}$	Similarity score between the s -th segment and the c -th class (anomaly type)
$\hat{y}_v$	Video-level score
y_s	Segment-level ground-truth label used only for evaluation, not for training
$\hat{y}_c$	Video-level predicted probability for class c after temporal aggregation and softmax
p_{a_n}	Embedding for the abnormal class n -th prompt
p_{norm}	Embedding for the normal class prompt
q_s	Segment-level class probability distribution for segment s
a_s	Stage-2 anomaly probability for segment s , computed as $1 - q_s$ [normal]
s_i	Stage-1 anomaly score of the i -th segment
pl	Soft pseudo-label cue derived from Stage-1 anomaly score
$\mathcal{L}_{bce}$	Stage-1 video-level binary cross-entropy loss
$\mathcal{L}_{pl}$	Segment-level pseudo-label consistency loss
$\mathcal{L}_{\text{align}}$	Video-level semantic alignment loss
$\mathcal{L}_{\text{cts}}$	Prompt contrastive loss
$\mathcal{L}_{\text{total}}$	Total Stage-2 training objective

3.2 Framework Overview

This section presents the proposed two-stage framework for weakly supervised video anomaly detection under limited annotation conditions. The framework is designed to examine two issues in prior approaches: insufficient temporal modeling in CLIP-based methods and limited fine-grained supervision in MIL-based models. To address these challenges, the first stage generates pseudo labels by reconstructing temporal dependencies through multi-scale temporal modeling, while the second stage refines anomaly recognition by enhancing semantic discrimination with CLIP-guided multimodal prompt learning. An overview of the architecture is illustrated in [Fig. 1](#).

3.3 CLIP-Based Visual and Text Representations

Each video is divided into fixed-length segments, such as 16-frame clips. Visual features $v \in \mathbb{R}^{T \times D}$ are extracted using the CLIP visual encoder (ViT-B/16), where T denotes the number of segments and D the embedding dimension.

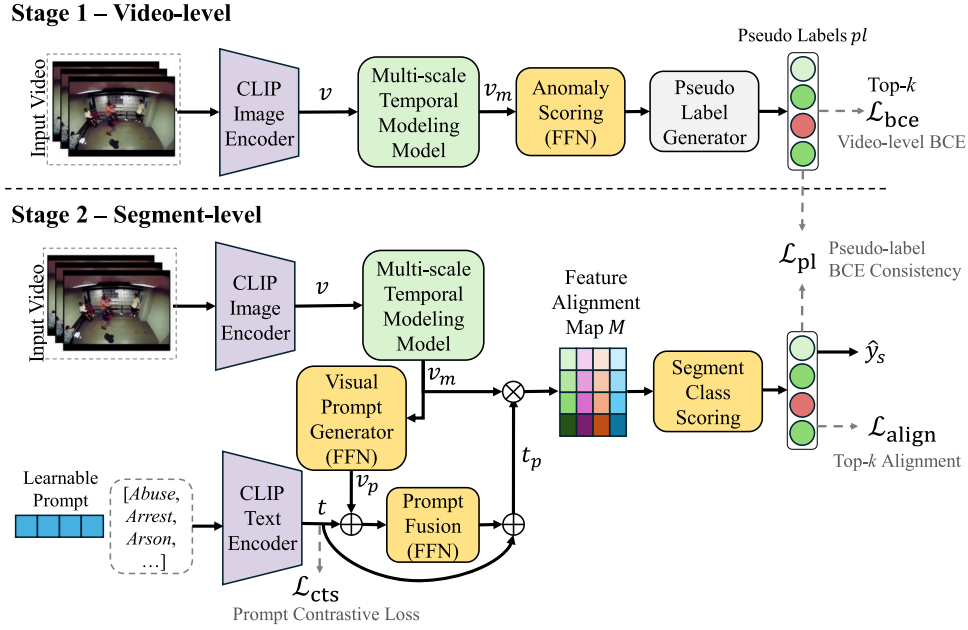


Figure 1: Overview of the proposed two-stage pseudo-label refinement framework. In Stage 1, CLIP visual features are processed by multi-scale temporal modeling to produce segment-level anomaly scores, which are further aggregated by top- k pooling for video-level BCE supervision. The resulting segment-level scores are retained as soft pseudo labels pl , and the segment-level visual representation v_m is passed to Stage 2. In Stage 2, v_m is used to generate visual prompts v_p , while learnable prompts and class names are encoded by the CLIP text encoder to obtain textual features t . These are fused to form prompt-enhanced textual representations t_p , which are matched with segment-level visual features for similarity-based segment scoring. The training objective combines pseudo-label consistency loss, alignment loss, and prompt contrastive loss.

To align textual features more effectively with visual representations, we adopt the CoOp method [19]. Unlike earlier approaches that rely on manually designed prompt sentences, we embed the names of 14 categories (13 anomaly types plus the normal class) into learnable prompt vectors. These category names are used to construct semantic prompts in Stage 2, rather than to indicate that segment-level category labels are available during training. These learnable prompts mitigate the bias introduced by handcrafted templates, allowing for more flexible semantic adaptation across categories, which is particularly valuable in weakly supervised settings where the supervision is sparse and less structured. For each class, a set of M learnable prompt vectors is concatenated with the tokenized class name, and the sequence is encoded by the CLIP text encoder:

$$t = \text{encoder}(L_1, \dots, L_m, \dots, L_M, \text{Tokenizer}(\text{class})) \quad (1)$$

here, t denotes the textual feature extracted by the CLIP text encoder, L_m represents the m -th learnable prompt vector, and Tokenizer converts the class name such as “robbery” or “assault” into token embeddings compatible with CLIP.

3.4 Pseudo-Label Generation via Temporal and Graph Modeling

In environments with limited annotations, the lack of fine-grained annotations is a significant challenge for video anomaly detection. Current methods often struggle to generate reliable segment-level pseudo-labels, particularly when temporal dependencies are not adequately captured [25,30]. To address this, we propose combining multi-scale temporal modeling with graph convolutional networks to capture both

local and global temporal dependencies, improving the accuracy of pseudo-label generation in weakly supervised settings.

3.4.1 Goal and Design Rationale

In the first stage, we focus on pseudo-label generation to overcome the challenge of sparse annotations. By using a multi-scale temporal transformer and dual-branch graph convolutional networks (GCNs), we can model temporal dependencies that are not explicitly captured by CLIP, a model initially trained on static images. The pseudo-label generator in this stage learns to produce segment-level anomaly scores based on temporal patterns in the video, which are then combined with video-level weak labels to strengthen supervision and improve model accuracy.

To address the lack of temporal reasoning in CLIP [30], we propose a multi-scale temporal modeling module that captures both short-term and long-term dependencies. This module consists of two main components: a multi-scale temporal transformer that extracts local temporal features and a graph convolutional network that encodes global contextual relations across video segments. Our design builds on prior temporal modeling strategies [12] by incorporating a multi-scale transformer, inspired by the effectiveness of recent video transformer architectures such as TimeSformer [16] and ViViT [17].

3.4.2 Multi-Scale Temporal Transformer

Anomalous events exhibit diverse temporal characteristics: for example, an explosion typically occurs within a short duration, whereas a fight may extend over a longer period. To address such variability under limited annotation conditions, we propose a multi-scale temporal transformer that applies multiple transformer encoders with distinct attention windows $\{w_1, w_2, w_3\}$. Each encoder is restricted by an attention mask to simulate a specific temporal scope, thereby extracting local temporal features denoted as v_l . These multi-scale features enhance temporal robustness and enable more accurate segment-level anomaly predictions [30].

The overall architecture of the multi-scale temporal modeling module, together with the dual-branch GCN for global reasoning, is illustrated in Fig. 2.

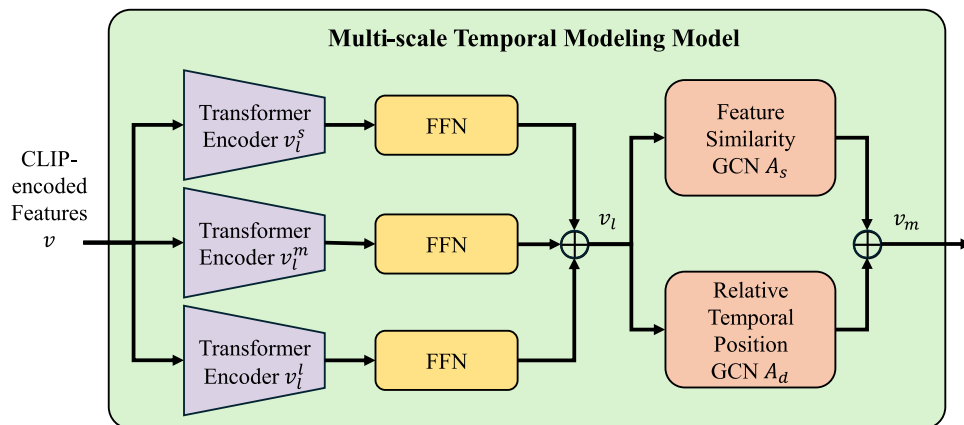


Figure 2: Multi-scale temporal modeling with transformers and dual-branch GCNs for global temporal context learning.

3.4.3 Dual-Branch GCN Global Context

To capture long-range dependencies, we construct two graph convolutional branches based on feature similarity and relative temporal position. We adopt a dual-branch GCN to integrate feature similarity and temporal distance, extending the foundations of graph convolution and temporal graph modeling. Let A_s and A_d be the adjacency matrices derived from cosine similarity and temporal distance, respectively. The feature similarity graph is defined as:

$$A_s = \text{softmax} \left(\frac{v_l v_l^T}{\|v_l\| \cdot \|v_l\|} \right), A_s \in \mathbb{R}^{T \times T} \quad (2)$$

where v_l represents the local features extracted by the multi-scale temporal transformer. The relative temporal position graph is defined as:

$$A_d = \text{softmax} \left(\frac{-|u - o|}{\sigma} \right), A_d \in \mathbb{R}^{T \times T} \quad (3)$$

where u and o denote the temporal indices of video segments, and σ is a hyperparameter controlling the influence range of relative distance. A smaller σ emphasizes nearby temporal segments, whereas a larger σ allows broader temporal aggregation across more distant segments. In this study, σ was fixed to 1 for all reported experiments.

The updated global representation is obtained by combining the outputs of the two branches:

$$v_m = \text{Concat}(\text{GCN}(A_s, v_l), \text{GCN}(A_d, v_l)) \quad (4)$$

Fig. 3 illustrates the construction of the two adjacency graphs. The feature similarity graph A_s links segments with similar embeddings in the feature space, while the relative temporal position graph A_d links segments according to their temporal distance. These complementary graphs are designed to provide feature-based and temporal-distance context for segment-level representation learning.

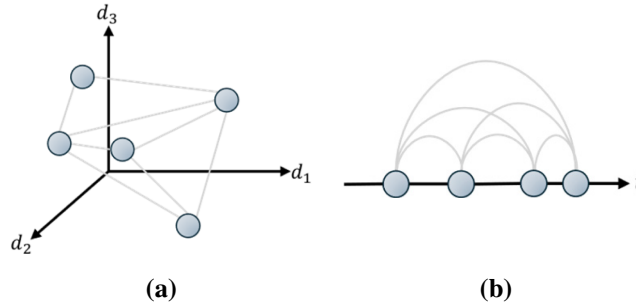


Figure 3: Illustration of graph construction strategies in the temporal modeling module. (a) Feature similarity graph A_s connects segments with similar visual representations in feature space (d is for the dimension); (b) relative temporal position graph A_d connects segments according to their temporal distance along the video timeline (t is for the time).

It should be noted that the graph construction does not guarantee that all edges are beneficial. Feature-similarity edges may connect visually similar but semantically different segments, including unrelated normal and abnormal regions. Temporal-distance edges may also propagate anomaly responses to nearby normal segments, especially when the anomaly boundary is ambiguous. Therefore, the dual-branch graph is used as a structured context modeling mechanism, but graph sparsity, edge reliability, and harmful-edge analysis remain important issues for further investigation.

3.4.4 Segment-Level Pseudo Labels and Training

The combined representation v_m is passed through a feedforward classification head to produce segment-level anomaly scores $\{s_1, s_2, \dots, s_T\}$. A video-level prediction is obtained by averaging the top- k segment scores, where $k = T/16$ and T denotes the number of video segments. The resulting video-level score $\hat{y}_v$ is compared with the video-level label using binary cross-entropy. The binary cross-entropy (BCE) loss is defined as follows:

$$\mathcal{L}_{bce} = -y \log(\hat{y}_v) - (1 - y) \log(1 - \hat{y}_v) \quad (5)$$

Unlike fully supervised settings, the segment-level scores obtained in this stage are not treated as ground-truth annotations. Instead, they are interpreted as soft pseudo-label cues that reflect temporally salient patterns inferred under weak supervision. These scores are continuous values in $[0, 1]$, and no binarization or thresholding is applied before they are passed to the second stage.

This design choice is motivated by the observation that, under weak supervision, segment-level predictions may contain substantial noise due to ambiguous or incomplete video-level labels. Rather than assuming that pseudo labels are reliable, the proposed framework retains them as soft supervisory signals, which are later refined through additional constraints in Stage 2. In particular, the second stage introduces semantic alignment and contrastive objectives that provide complementary guidance beyond the temporal cues captured in Stage 1.

Therefore, the role of Stage 1 is to generate temporally structured intermediate cues, not to produce clean segment annotations. These cues are noisy but informative, and are used as soft supervision for Stage 2, where semantic alignment and contrastive objectives provide additional constraints for refinement.

The algorithm for generating segment-level pseudo-labels is presented in Algorithm 1.

Algorithm 1: Segment-level pseudo label generation (Stage 1)

	Input: Video V with video-level label $y \in \{0, 1\}$
	Output: Anomaly scores $\{s_i\}$ for $i = 1, \dots, T$
1:	Divide V into T segments and extract features v
2:	Obtain local temporal features v_l using multi-scale transformer
3:	$v_m \leftarrow \text{Concat}(\text{GCN}(A_s, v_l), \text{GCN}(A_d, v_l))$
4:	Segment anomaly scores: $s_i = \text{FFN}(v_m[i])$, for $i = 1, \dots, T$
5:	Select top- k scores, where $k = T/16$
6:	Compute video-level prediction: $\hat{y}_v = \text{MEAN}(\text{TOP-K}(\{s_i\}, k))$
7:	Compute BCE loss: $\mathcal{L}_{bce} = -y \log(\hat{y}_v) - (1 - y) \log(1 - \hat{y}_v)$
8:	Update model parameters with $\mathcal{L}_{bce}$
9:	After training, use $\{s_i\}$ as intermediate soft pseudo-label cues for Stage 2

3.5 Prompt-Guided Multimodal Refinement

While many existing methods rely on single-modality representations, they often struggle to align temporal features with semantic information under weak supervision, where segment-level labels are noisy or unavailable. To address this limitation, the proposed framework introduces a prompt-guided multimodal refinement stage that leverages the Stage-1 scores as intermediate soft supervision for semantic recalibration. Instead of directly trusting these pseudo labels, this stage refines them through vision-language alignment, enabling semantic recalibration of noisy segment-level predictions.

3.5.1 Goal and Design Rationale

The second stage introduces a multimodal prompt learning mechanism to semantically recalibrate the temporally structured cues generated in Stage 1. The two-stage design is motivated by the observation that direct end-to-end training under weak supervision can propagate noise from inaccurate pseudo labels into high-level semantic alignment. By decoupling temporal modeling and semantic refinement, the framework first generates structured pseudo labels from temporal context in Stage 1 and then refines them using vision-language alignment in Stage 2, so that CLIP-guided prompts provide semantic recalibration rather than temporal dependency modeling.

In this stage, textual embeddings are refined with learnable prompts, following recent advances in prompt tuning [19–21]. Unlike earlier CLIP-based anomaly detection approaches [22,23], the proposed framework does not rely solely on direct multimodal matching, but instead uses Stage-1 pseudo labels as structured auxiliary supervision for semantic refinement. This design enables prompt-driven multimodal alignment to correct noisy segment-level predictions and improve anomaly discrimination under weak supervision.

3.5.2 Learnable Textual Prompts

To better align visual and textual embeddings, we adopt a learnable prompt mechanism inspired by CoOp [19]. For each class, a set of learnable prompt vectors $P = \{p_1, \dots, p_M\}$ is concatenated with the tokenized class name before being fed into the CLIP text encoder. The resulting textual embedding is refined by a feedforward network combined with a visual prompt:

$$t_p = \text{FFN}(v_p + t) + t \quad (6)$$

where t_p denotes the prompt-enhanced textual feature, t is the textual feature produced by the CLIP text encoder, and v_p is the visual prompt derived from the video representation.

3.5.3 Similarity-Based Matching

Each segment's visual feature v_m is compared to the prompt-enhanced textual embeddings using cosine similarity:

$$M = \frac{v_m \cdot t_p^\top}{\|v_m\| \|t_p\|} \quad (7)$$

This produces a similarity matrix $M \in \mathbb{R}^{T \times (K+1)}$, which serves as a soft prediction of anomaly class per segment.

3.5.4 Loss Functions and Optimization

We employ three loss functions to optimize the Stage-2 refinement.

Pseudo-Label Loss: The pseudo-label loss enforces consistency between Stage-2 predictions and the soft anomaly cues generated in Stage 1. Let $pl \in [0, 1]$ denote the normalized anomaly score of segment s produced in Stage 1, and $\hat{y}_s \in [0, 1]$ denote the corresponding prediction from Stage 2. The pseudo-label loss is defined as:

$$\mathcal{L}_{pl} = -(pl) \log(\hat{y}_s) - (1 - pl) \log(1 - \hat{y}_s) \quad (8)$$

here, pl is directly derived from the Stage-1 output without binarization and is treated as a soft target rather than a definitive label. This formulation allows the model to preserve uncertainty in segment-level predictions instead of forcing hard assignments.

It is important to note that this loss alone does not guarantee correction of noisy pseudo labels. Instead, it provides a temporal consistency constraint that encourages Stage-2 predictions to remain aligned with the coarse anomaly cues identified in Stage 1. To mitigate the effect of noise, the pseudo-label loss is jointly optimized with the alignment loss and the contrastive loss, which introduce global semantic consistency and inter-class separation. These additional objectives allow the model to adjust or counterbalance unreliable pseudo-label signals when they conflict with multimodal semantic evidence.

In this sense, the pseudo-label term acts as a weak supervisory signal within a multi-objective framework, rather than a dominant source of supervision. This design reduces the risk of directly propagating errors from Stage 1 and provides a more flexible mechanism for refinement under weak supervision.

Alignment Loss: Since pseudo labels may contain noise, we introduce an additional alignment loss following [11] to enhance consistency between video features and class embeddings. The alignment loss follows prior work on top- k pooling for anomaly classification [11], while the contrastive formulation is motivated by the vision-language alignment strategy of CLIP [18]. Specifically, for each class c_n , we compute an aggregated similarity score a_n by pooling the top- k segment-class similarity scores across time (or by averaging):

$$a_n = \text{Aggregate} \left(\{M_{s,n}\}_{s=1}^T \right) \quad (9)$$

where $M_{s,n}$ denotes the similarity between the s -th video segment and the n -th anomaly class. An additional score a_{norm} is computed for the normal class in the same manner. $\text{Aggregate}(\cdot)$ denotes a temporal aggregation operator that summarizes segment-level anomaly scores into a video-level signal for weak supervision.

The aggregated scores are normalized through a softmax with temperature τ :

$$\hat{y}_{c_n} = \text{softmax}(a_n/\tau) = \frac{\exp(a_n/\tau)}{\sum_{j=1}^{N+1} \exp(a_j/\tau)} \quad (10)$$

where N is the number of anomaly classes and the additional term corresponds to the normal class. The temperature parameter τ controls the sharpness of the distribution.

Finally, a cross-entropy loss is computed between the predicted distribution and the ground-truth class label:

$$\mathcal{L}_{\text{align}} = - \sum_n y_{c_n} \log(\hat{y}_{c_n}) \quad (11)$$

This alignment mechanism reduces the impact of noisy pseudo labels by introducing a global semantic consistency constraint at the video level.

Contrastive Loss: To enhance the separation between normal and abnormal representations, we introduce a contrastive loss that enforces dissimilarity between the two types of prompts. For each abnormal class c_n , we obtain an embedding p_{a_n} by feeding its learnable prompts into the CLIP text encoder. Similarly, the embedding p_{norm} is obtained by encoding the learnable prompts associated with the normal class. The contrastive loss is then defined as:

$$\mathcal{L}_{\text{cts}} = \sum_m \max \left(0, \frac{p_{a_n} p_{\text{norm}}^\top}{\|p_{a_n}\| \cdot \|p_{\text{norm}}\|} \right) \quad (12)$$

This objective penalizes high cosine similarity between abnormal and normal prompt embeddings, thereby encouraging greater semantic separation across categories.

The final training objective is a weighted sum of the three loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pl}} + \mathcal{L}_{\text{align}} + \lambda \mathcal{L}_{\text{cts}} \quad (13)$$

The weight λ controls the contribution of the contrastive term. In our experiments, λ was fixed to 0.1 for all reported results. This setting was used to keep the prompt contrastive objective as a regularization term, rather than allowing it to dominate the pseudo-label loss and semantic alignment loss.

This design further improves robustness by preventing semantic overlap between normal and anomalous categories, which is critical under weak supervision where class boundaries are often ambiguous.

Together, these objectives enable the model to progressively refine pseudo supervision by combining temporal structure with semantic alignment, rather than relying solely on noisy instance-level predictions.

3.5.5 Training Procedure

Training in Stage 2 is performed by minimizing a weighted sum of the pseudo-label loss, the alignment loss, and the contrastive loss. The temporal modeling module from Stage 1 is reused to extract segment-level features, while Stage 2 focuses on refining these representations through multimodal alignment. The prompt-guided refinement process in Stage 2 in Algorithm 2.

Algorithm 2: Prompt-guided refinement (Stage 2)

Input: Video V , pseudo labels $pl = \{s_i\}$, for $i = 1, \dots, T$,
class set C including normal, learnable prompts $P = \{L_1, \dots, L_M\}$
Output: Refined segment predictions and class scores

```

1:  $\{v_i\} \leftarrow \text{CLIP-ImageEncoder}(V)$ ;
2:  $v_l \leftarrow \text{MultiScaleTransformer}(\{v_i\})$ 
3:  $v_m \leftarrow \text{Concat}(\text{GCN}(A_s, v_l), \text{GCN}(A_d, v_l))$ 
4:  $v_p \leftarrow \text{FFN}(v_m)$  # visual prompt per segment
5: for each class  $c$  in  $C$  do
6:    $t_c \leftarrow \text{encoder}(L_1, \dots, L_M, \text{Tokenizer}(c))$ 
7:    $t_p^c \leftarrow \text{FFN}(v_p + t_c) + t_c$ 
8: end for
9: for  $s = 1, \dots, T$  do
10:   for each class  $c$  in  $C$  do
11:      $M_{\{s,c\}} \leftarrow \text{cosine}(v_m[s], t_p^c)$ 
12:   end for
13:    $q_s \leftarrow \text{softmax}_c(M_{\{s,c\}})$  # class distribution per segment
14:    $a_s \leftarrow 1 - q_s[\text{normal}]$  # anomaly probability per segment
15: end for
16:  $\mathcal{L}_{\text{pl}} \leftarrow \text{mean}_s \text{ BCE}(a_s, pl)$  # segment supervision by pseudo labels
17: for each class  $c$  in  $C$  do

```

(Continued)

Algorithm 2 (continued)

```

18:          $a_c \leftarrow \text{Aggregate}_s(M_{\{s,c\}})$  # top- $k$  pooling or mean over time
19:     end for
20:      $\hat{y}_c \leftarrow \text{softmax}_c(a_c/\tau)$ 
         $\mathcal{L}_{\text{align}} \leftarrow \text{CE}(\hat{y}_c, y_{\text{video\_class}})$  # one-hot including normal
22:      $p_{\text{norm}} \leftarrow \text{encoder}(L_1, \dots, L_M, \text{Tokenizer}(\text{normal}))$ 
23:      $\mathcal{L}_{\text{cts}} \leftarrow \sum_{\{c \in C \setminus \{\text{normal}\}\}} \max(0, \text{cosine}(p_c, p_{\text{norm}}))$ 
24:      $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{pl} + \mathcal{L}_{\text{align}} + \lambda \cdot \mathcal{L}_{\text{cts}}$ 
25:     Update Stage-2 parameters by minimizing  $\mathcal{L}_{\text{total}}$ 

```

3.6 Summary

The proposed framework integrates pseudo-label generation and multimodal refinement into a unified two-stage design. In Stage 1, segment-level anomaly scores are learned under weak supervision by combining multi-scale temporal modeling with a top- k aggregation strategy, producing pseudo labels that capture fine-grained temporal cues. In Stage 2, these pseudo labels guide multimodal prompt-based refinement, where learnable textual prompts and similarity matching enhance semantic discrimination across classes. By organizing these operations into functionally distinct stages, the framework enables temporal cues to be generated before semantic refinement is applied.

4 Experimental Results

To evaluate the proposed two-stage pseudo-label learning framework, we conduct experiments on the UCF-Crime dataset. We report frame-level AUC for binary anomaly ranking and mAP@IoU for localization-aware multi-class evaluation. The former evaluates whether anomalous frames are ranked higher than normal frames, whereas the latter evaluates whether predicted anomaly segments are temporally aligned with ground-truth intervals under category-aware evaluation.

4.1 Dataset Description

We evaluate our method on UCF-Crime [7], a widely used benchmark for weakly supervised video anomaly detection. The dataset consists of 1900 untrimmed surveillance videos captured from various real-world settings, including urban environments, industrial sites, and public spaces. It covers 13 anomaly categories (Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Accident, Robbery, Shooting, Stealing, Vandalism, and Violence) and one normal category (Normal). While the evaluation in this study is conducted on UCF-Crime, this dataset is widely adopted in prior WSVAD studies and provides a representative benchmark for weakly supervised anomaly detection. Following prior works [14,22], we use UCF-Crime because it provides video-level labels and temporal annotations required for evaluating both frame-level anomaly ranking and localization-aware mAP@IoU.

4.2 Evaluation Metrics

We report two types of performance metrics:

- (1) Binary Classification (Normal vs. Anomaly): Performance is measured by the Area Under the ROC Curve (AUC) at the frame level. Given a decision threshold ρ , the true positive rate (TPR) and false positive rate (FPR) are defined as:

$$\text{TPR}(\rho) = \frac{TP(\rho)}{TP(\rho) + FN(\rho)}, \text{FPR}(\rho) = \frac{FP(\rho)}{FP(\rho) + TN(\rho)} \quad (14)$$

where TP , FP , TN , FN denote true positives, false positives, true negatives, and false negatives, respectively. The ROC curve is obtained by plotting $\text{TPR}(\rho)$ against $\text{FPR}(\rho)$ across thresholds, and the AUC is computed as:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x)) dx \quad (15)$$

- (2) **Multi-Class Anomaly Classification:** Following prior works [12,14], we report mean Average Precision (mAP) under different Intersection-over-Union (IoU) thresholds (0.1 to 0.5) between detected and ground-truth anomaly segments. The IoU between a predicted region B_p and a ground-truth region B_{gt} is defined as:

$$\text{IoU}(B_p, B_{gt}) = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \quad (16)$$

For each anomaly class c , the average precision (AP) is computed as the area under the precision–recall curve:

$$\text{AP}(c) = \int_0^1 p_c(r) dr \quad (17)$$

where $p_c(r)$ is the precision at recall r for class c . The mAP across C classes is then:

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^C \text{AP}(c) \quad (18)$$

We note that these metrics capture different aspects of performance but do not exhaustively characterize VAD behavior. False-alarm-oriented metrics, abnormal-only evaluation, and precision-recall-based measures may provide additional insight under class imbalance and should be considered in future evaluations.

Although frame-level AUC is widely used in WSVAD, it mainly evaluates ranking behavior across decision thresholds and may not fully reflect detection quality under severe class imbalance. Precision-recall-based metrics, such as AP or AUPRC, can provide additional insight into the detection of rare anomalous instances. In this study, we report frame-level AUC for consistency with prior UCF-Crime evaluation protocols and use localization-aware mAP@IoU to further examine temporal localization behavior. Nevertheless, coarse-level AP/AUPRC is not included in the current evaluation and is discussed as a limitation.

4.3 Implementation Details

All experiments were implemented using PyTorch 2.1 with CUDA 12.1, and executed on a workstation equipped with an NVIDIA RTX 4080 GPU, an Intel Core i9-13980HX CPU, and 32 GB RAM. For feature extraction, we utilized the pretrained CLIP model (ViT-B/16), which generates 512-dimensional visual and textual features by aligning vision-language embeddings. The model was used as a backbone for both visual and textual encoding, leveraging CLIP's ability to transfer knowledge learned from large-scale image-text pairs to our anomaly detection task.

Each video was processed by dividing it into non-overlapping segments of 16 frames, with a maximum of 256 segments retained per video. This segmentation strategy helps to standardize the input size, ensuring efficient processing while maintaining temporal context. For multi-scale temporal modeling, we employed

three transformer encoders with attention windows of sizes 2, 4, and 8, enabling the model to capture both short-term and long-term dependencies. The varying attention window sizes allow the model to adaptively focus on different temporal scales, facilitating the detection of anomalies occurring at multiple time intervals.

The dual-branch Graph Convolutional Network (GCN) was implemented to construct adjacency matrices for both cosine similarity and temporal distance. These matrices capture global context by linking visually similar and temporally related video segments. The cosine similarity branch represents feature-based relationships, while the temporal distance branch models the chronological ordering of the segments. This dual-branch design enhances the model's understanding of both feature and temporal alignment.

In the prompt learning module, we assigned 20 learnable prompts per class, allowing for flexibility in multimodal alignment. The learnable prompts are integrated into CLIP's textual embeddings to guide the model in classifying anomaly types based on both visual and semantic cues. This approach helps address the issue of semantic ambiguity in video anomaly detection by introducing flexible, adaptive textual prompts that can align with the model's learned features.

The model was trained in two stages, with each stage lasting for 10 epochs. We used the AdamW optimizer with a learning rate of $2e-5$ and a batch size of 32. All feature extractors were frozen during training to preserve the pretrained knowledge from CLIP, ensuring that the model focused on learning from the pseudo-labels and temporal relationships. Several hyperparameters, such as the weight of the contrastive loss term and the temperature scaling factor in the softmax function, were selected based on performance on a held-out validation set. These choices were fine-tuned to balance the contributions of each loss component (pseudo-label loss, alignment loss, and contrastive loss) to the overall model performance.

For top- k aggregation, the highest k segment scores within each video are selected and averaged, where $k = T/16$ and T denotes the number of video segments. For the temporal-distance adjacency matrix, the scaling factor σ was fixed to 1 in all experiments. The temperature parameter τ was set to 0.07, and the contrastive loss weight λ was set to 0.1.

4.4 Performance Comparison

To assess the effectiveness of our method, we compare it with representative weakly supervised video anomaly detection approaches on UCF-Crime, covering both binary classification and multi-class settings.

For binary anomaly detection, Table 2 reports frame-level AUC results. As shown, the proposed method achieves 87.41% AUC, which is competitive with recent weakly supervised baselines. However, compared to VadCLIP, the proposed method does not achieve the highest AUC, indicating that its advantage is not reflected in global anomaly ranking performance alone.

Table 3 provides a different perspective by reporting localization-aware mAP@IoU. Unlike frame-level AUC, which primarily measures ranking quality across the entire video, mAP@IoU [42] emphasizes temporal alignment between predicted anomaly segments and ground-truth intervals. As shown in Table 3, the proposed method consistently achieves higher mAP values than both Sultani et al. [7] and VadCLIP [22] across multiple IoU thresholds. However, the gain over VadCLIP is modest at some IoU thresholds, especially at stricter temporal overlap settings. Therefore, these results should be interpreted as indicative localization-aware evidence rather than as strong statistical evidence of robust superiority. In addition, the current results are based on a single run, and multiple runs with standard deviation are not reported. Therefore, the stability of the observed mAP gains remains to be further verified.

Table 2: Binary anomaly detection (frame-level AUC) on UCF-Crime.

Method	Feature	AUC (%)
Sultani et al. [7]	C3D [1]	75.41
MIST [11]	I3D [2]	82.30
RTFM [12]	I3D [2]	84.30
MSL [13]	VideoSwin [3]	85.62
Zhang et al. [26]	I3D [2]	86.22
Fan et al. [27]	I3D [2]	86.19
UMIL [14]	I3D [2]	86.75
AnomalyCLIP [23]	CLIP [18]	86.36
VadCLIP [22]	CLIP [18]	88.02
Ours	CLIP [18] + PL	87.41

Table 3: Localization-aware multi-class evaluation (mAP@IoU) on UCF-Crime.

Method	mAP@IoU					
	0.1	0.2	0.3	0.4	0.5	AVG
Sultani et al. [7]	5.73	4.41	2.69	1.93	1.44	3.24
VadCLIP [22]	11.72	7.83	6.40	4.53	2.93	6.68
Ours	13.60	9.82	6.42	4.86	3.40	7.62

This contrast suggests that the proposed framework exhibits different strengths under different evaluation criteria. While its performance in global ranking (AUC) remains comparable to existing methods, its improvements are more evident in localization-aware evaluation. This observation indicates that the decoupled design may be more beneficial for segment-level anomaly localization than for coarse video-level anomaly scoring. Accordingly, the AUC comparison should be interpreted as complementary evidence rather than as a claim of superiority in coarse anomaly ranking.

Therefore, Tables 2 and 3 provide complementary but distinct evidence: Table 2 reflects coarse frame-level anomaly ranking, whereas Table 3 evaluates localization-aware temporal alignment. The proposed framework should therefore be interpreted as showing stronger localization-oriented behavior rather than uniformly superior performance across all evaluation aspects.

4.5 Ablation Study

To investigate the effectiveness of the proposed design, we conduct an ablation study focusing on the main design components in the two-stage framework, including pseudo-label learning and multi-scale temporal modeling. The ablation study is intended to isolate the contribution of key design components, rather than to provide an exhaustive evaluation of all modules. The results are summarized in Table 4.

Table 4: Ablation study on binary AUC (%).

Base Model	Pseudo Label Learning	Multi-Scale Model	AUC (%)
✓	✗	✗	84.59
✓	✓	✗	85.28
✓	✗	✓	86.24
✓	✓	✓	87.41

As shown in Table 4, incorporating pseudo-label learning improves the baseline performance from 84.59% to 85.28% in terms of AUC, indicating that pseudo supervision provides additional training signals beyond coarse video-level labels. Similarly, introducing multi-scale temporal modeling improves performance to 86.24%, demonstrating the importance of capturing temporal dependencies for anomaly detection. When both components are combined, the model achieves the best performance of 87.41%, suggesting that the two components provide complementary benefits.

These results further suggest that, under weak supervision, the primary bottleneck may not only be the absence of segment-level annotations, but also the lack of temporally structured representations from which reliable pseudo labels can be derived. In other words, pseudo labels are useful only when the underlying temporal cues are sufficiently informative. This finding supports the staged design of the proposed framework, where temporal reasoning first constructs structured anomaly cues and pseudo-label refinement subsequently uses them as soft supervision.

From a design perspective, these results provide empirical support for the effectiveness of integrating temporal modeling with pseudo-label refinement. In particular, temporal modeling enhances the structural consistency of segment-level predictions, while pseudo-label learning introduces additional supervision that helps guide the learning process under weak labels. Their combination leads to more reliable anomaly scoring compared to using either component alone.

The ablation results in Table 4 provide preliminary evidence that temporal modeling and pseudo-label learning are complementary under the current setting. However, this ablation does not provide complete module-level attribution. In particular, the current experiments do not separately evaluate the feature-similarity GCN branch, the temporal-distance GCN branch, the prompt module, the fusion module, or the individual loss terms. Therefore, the ablation results should be interpreted as evidence for the overall staged design rather than as conclusive evidence for the independent contribution of each component.

An interesting observation from Table 4 is that multi-scale temporal modeling provides a larger individual gain than pseudo-label learning. This suggests that, under the frame-level AUC metric, reconstructing temporal dependencies has a more direct effect on anomaly ranking than introducing pseudo supervision alone. However, pseudo-label learning becomes more beneficial when combined with temporal modeling, increasing the AUC from 86.24% to 87.41%. This indicates that pseudo labels may be more reliable when they are generated from temporally structured representations. Therefore, the results suggest that pseudo-label refinement should not be viewed as an isolated supervision strategy, but as a mechanism whose effectiveness depends on the quality of temporal cues produced in the first stage.

4.6 Qualitative Analysis

In addition to the quantitative comparison with other methods, we conduct a qualitative analysis to examine whether predicted anomaly scores align with annotated anomaly intervals. The results are shown in Fig. 4, where the blue line represents the predicted anomaly scores, and the red areas represent the labeled

anomaly time segments. The x -axis represents the temporal index of the video, and the y -axis represents the anomaly scores. The four plots represent normal, robbery, explosion, and vandalism events. The experimental results show that the proposed method aligns well with the labeled anomaly intervals in several cases. In Fig. 4a, the low scores maintained in normal videos suggest that the model can suppress false responses in this example. In Fig. 4b, for a continuous event like robbery, the model maintains high scores throughout the event duration, and in Fig. 4c, for abrupt events such as explosions, the anomaly score exhibits a rapid response. These examples suggest that the model can capture certain temporal patterns, but they should not be interpreted as comprehensive evidence of robustness.

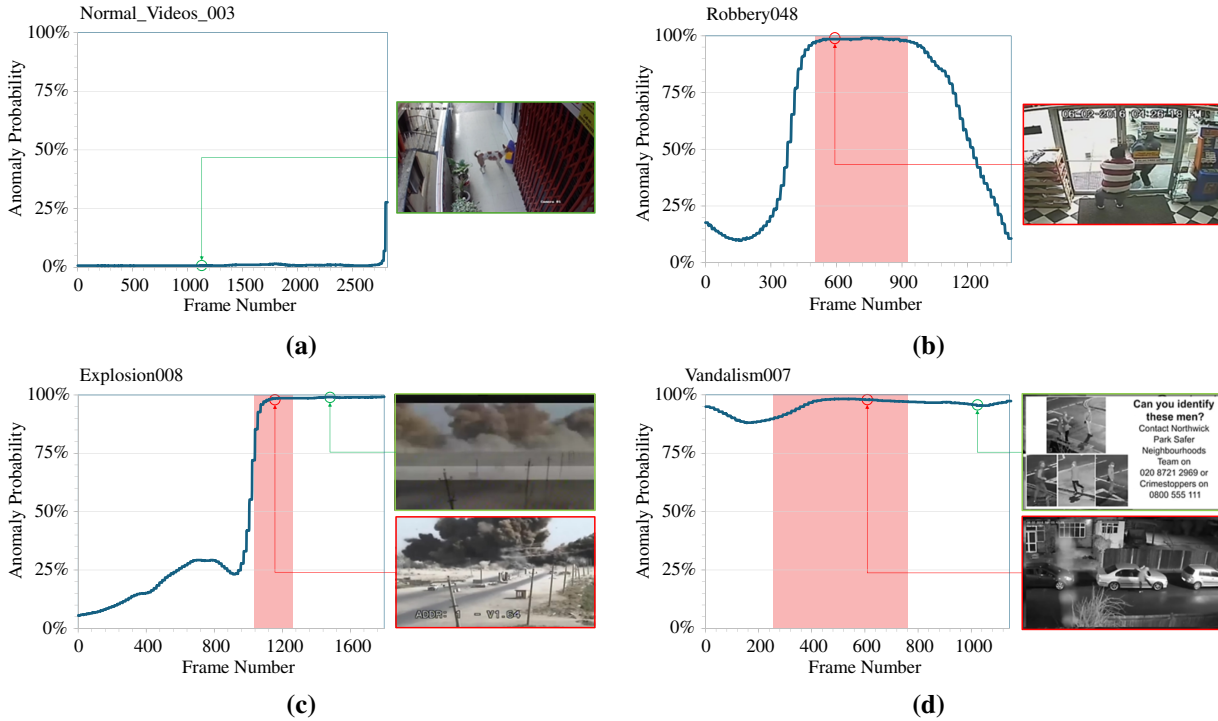


Figure 4: Qualitative examples on UCF-Crime. (a) Normal_videos_003 (Normal); (b) Robbery048 (Robbery); (c) Explosion008 (Explosion); (d) Vandalism007 (Vandalism). The red regions indicate the ground-truth anomaly intervals, and the blue curve denotes the predicted anomaly score over time.

However, not all videos are correctly detected in terms of the anomaly time points. In Fig. 4c, the video contains an explosion event. Although the annotated anomaly interval is detected, the anomaly score remains high after the annotated period. One possible explanation is that post-explosion visual effects, such as persistent smoke, continue to provide strong anomaly cues even after the labeled interval ends. Additionally, a similar issue occurs in Fig. 4d, which represents a vandalism event. Our method detects anomalies throughout the entire video, while the annotated anomaly interval covers only part of the video. We attribute this behavior to temporal persistence in the learned anomaly scoring: once an anomalous pattern is activated, weak supervision and segment-level aggregation can cause the model to propagate high scores to adjacent segments, leading to delayed score decay after the event ends.

In addition, we present a representative failure case. As shown in Fig. 5, the video depicts a robbery event where two individuals are seen robbing a person inside a white vehicle. Despite the presence of a weapon, the anomaly score remains consistently low throughout the video, indicating that the model failed to detect the expected abnormal event. The lack of significant or violent confrontation, and the fact that

part of the incident occurred inside the car window, may have contributed to the missed detection. This suggests that while temporal modeling captures contextual persistence, additional mechanisms for fine-grained semantic reasoning may be required to better distinguish subtle anomalies. This failure underscores the complexity of anomaly detection, where the identification of anomalies is based not only on the actions performed by individuals, but also on contextual factors such as the objects they carry, the environment, and the background of the scene.

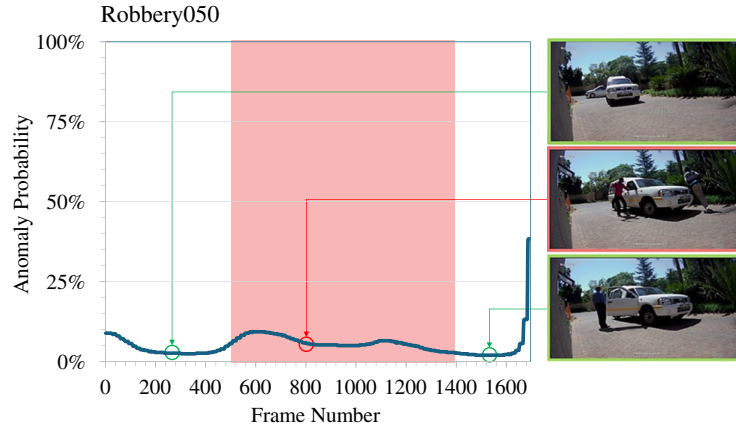


Figure 5: Failure case on UCF-Crime. Example from Robbery050 (Robbery). Despite the annotated anomaly interval, the predicted anomaly score remains low, illustrating a miss in a subtle event scenario.

The qualitative examples are intended to illustrate representative behaviors rather than to provide a comprehensive failure analysis. A broader qualitative evaluation covering more difficult videos, ambiguous temporal boundaries, subtle abnormal events, and hard negative cases would be necessary to fully characterize the limitations of the proposed framework.

4.7 Discussion

The proposed framework was evaluated on UCF-Crime under both binary anomaly detection and multi-class anomaly classification settings. Under the standard frame-level evaluation protocol, the method achieves competitive performance in terms of AUC. However, as discussed in Section 4.4, frame-level AUC mainly reflects global anomaly ranking and does not fully capture the quality of temporal localization.

Compared to recent vision-language-based methods such as VadCLIP, which adapt large-scale multi-modal representations for weakly supervised VAD, our approach places greater emphasis on reconstructing temporal dependencies that may not be fully captured by static pre-trained encoders. The improved localization-aware mAP@IoU results suggest that integrating multi-scale temporal reasoning with prompt-guided alignment can provide complementary benefits for segment-level localization, rather than replacing or uniformly outperforming direct multimodal adaptation. Therefore, in the proposed framework, CLIP is not used to replace temporal modeling. Instead, it provides semantic constraints after temporally structured cues have been generated in the first stage.

The localization-aware mAP@IoU results provide additional insights into the behavior of the proposed framework. The consistent improvement across IoU thresholds suggests that the decoupled design, which separates temporal pseudo-label generation from semantic refinement, may help produce segment-level predictions that better align with ground-truth anomaly intervals. This behavior is consistent with the design

intention of treating Stage-1 outputs as structured temporal cues and refining them through multimodal alignment in Stage 2.

The ablation study further supports the complementary roles of temporal modeling and pseudo-label refinement. Temporal modeling contributes to capturing contextual dependencies across segments, while pseudo-label learning introduces additional supervisory signals under weak labels. Their combination leads to improved consistency in segment-level anomaly scoring. However, it should be noted that the ablation is limited to a subset of design factors and does not exhaustively analyze all components of the framework.

The experimental results suggest that temporal modeling and pseudo-label refinement may contribute differently to weakly supervised VAD. Temporal reasoning appears to have a stronger effect on coarse anomaly ranking, as reflected by the larger AUC gain in the ablation study. In contrast, the more consistent improvement in localization-aware mAP@IoU suggests that pseudo-label refinement may play a more important role in improving temporal alignment and segment-level consistency.

Several empirical limitations remain in the current study. First, the evaluation is limited to UCF-Crime and does not establish cross-dataset generalization. Additional datasets such as XD-Violence would be needed to assess transferability across different anomaly categories, scene distributions, and annotation protocols. Second, the current comparison is representative rather than exhaustive, and it does not include all recent localization-aware or CLIP-based WSVAD methods under a unified implementation protocol. Third, the current results are based on a single run, and multiple runs with standard deviation are not reported. Therefore, the stability of the observed mAP gains remains to be further verified. Fourth, the current study does not include a direct comparison with an end-to-end implementation using the same components, so the results should not be interpreted as conclusive evidence that staged training is superior to all end-to-end alternatives. Fifth, the ablation study does not separately evaluate the feature-similarity GCN branch, the temporal-distance GCN branch, the prompt module, the fusion module, or the individual loss terms. Sixth, the current analysis does not examine graph sparsity, harmful graph edges, or computational cost, including parameter count, FLOPs, inference speed, and memory usage. Therefore, the results should be interpreted as benchmark-specific evidence for analyzing staged pseudo-label refinement, rather than as proof of broad superiority, full component-level attribution, or deployment-ready robustness.

5 Conclusion

This paper presented a two-stage pseudo-label learning framework for weakly supervised video anomaly detection under limited annotation conditions. The proposed design separates the training roles of temporal cue generation and semantic refinement, while still allowing Stage-2 refinement to use Stage-1 outputs as intermediate soft supervision.

In the first stage, multi-scale temporal transformers and dual-branch graph convolutional networks are used to generate structured segment-level anomaly scores. These scores are treated as soft pseudo-label cues rather than definitive segment annotations. In the second stage, the cues are refined through learnable textual prompts and vision-language alignment, providing an additional semantic constraint for segment-level prediction.

Experiments on UCF-Crime show that the proposed framework achieves competitive frame-level AUC and more consistent gains in localization-aware mAP@IoU. These results suggest that the observed benefit of the staged design is more closely associated with localization-oriented segment alignment than with global anomaly ranking under the UCF-Crime setting. Therefore, the main contribution lies in the proposed staged framework and its empirical analysis of pseudo-label refinement behavior under weak supervision, rather than in claiming broad superiority over existing methods.

The current study still has several limitations. The evaluation is restricted to a single benchmark dataset, repeated-run statistics are not reported, and the effects of individual modules and loss terms have not been exhaustively analyzed. Future work should extend the evaluation to additional datasets, incorporate precision-recall-based and false-alarm-oriented metrics, compare against end-to-end alternatives, conduct component-level ablations, analyze graph sparsity and harmful edges, and report computational profiling results.

Acknowledgement: Not applicable.

Funding Statement: The authors thank the National Science and Technology Council for supporting this work in part under Grant NSTC 114-2221-E-032-010-.

Author Contributions: The authors confirm their contribution to the paper as follows: study conception and design: Meng-Luen Wu, Yan-Hsin Lin; data collection: Yan-Hsin Lin; analysis and interpretation of results: Meng-Luen Wu, Yan-Hsin Lin; programming and experiments: Yan-Hsin Lin; draft manuscript preparation: Meng-Luen Wu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available from the UCF-Crime dataset at <https://www.crcv.ucf.edu/projects/real-world/>.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 4489–97. doi:10.1109/ICCV.2015.510.
2. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 4724–33. doi:10.1109/cvpr.2017.502.
3. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video swin transformer. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 3192–201. doi:10.1109/cvpr52688.2022.00320.
4. Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes. *IEEE Trans Pattern Anal Mach Intell*. 2014;36(1):18–32. doi:10.1109/TPAMI.2013.111.
5. Lu C, Shi J, Jia J. Abnormal event detection at 150 FPS in MATLAB. In: Proceedings of the 2013 IEEE International Conference on Computer Vision; 2013 Dec 1–8; Sydney, Australia. p. 2720–7. doi:10.1109/iccv.2013.338.
6. Liu W, Luo W, Lian D, Gao S. Future frame prediction for anomaly detection—a new baseline. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 6536–45. doi:10.1109/CVPR.2018.00684.
7. Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 6479–88. doi:10.1109/CVPR.2018.00678.
8. Wu P, Liu J, Shi Y, Sun Y, Shao F, Wu Z, et al. Not only look, but also listen: learning multimodal violence detection under weak supervision. In: *Computer vision—ECCV 2020*. Cham, Switzerland: Springer; 2020. p. 322–39. doi:10.1007/978-3-030-58577-8_20.
9. Le H, Lu CK, Hsu CC. Video anomaly detection for edge-based IoT systems: a survey of input modalities and real-time applications. *Intell Syst Appl*. 2026;29(3):200635. doi:10.1016/j.iswa.2026.200635.

10. Wang Y, Zhou J, Guan J. A lightweight video anomaly detection model with weak supervision and adaptive instance selection. *Neurocomputing*. 2025;613(6):128698. doi:10.1016/j.neucom.2024.128698.
11. Feng JC, Hong FT, Zheng WS. MIST: multiple instance self-training framework for video anomaly detection. In: *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2021 Jun 20–25; Nashville, TN, USA. p. 14004–13. doi:10.1109/cvpr46437.2021.01379.
12. Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021 Oct 10–17; Montreal, QC, Canada. p. 4955–66. doi:10.1109/iccv48922.2021.00493.
13. Li S, Liu F, Jiao L. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *Proc AAAI Conf Artif Intell*. 2022;36(2):1395–403. doi:10.1609/aaai.v36i2.20028.
14. Lv H, Yue Z, Sun Q, Luo B, Cui Z, Zhang H. Unbiased multiple instance learning for weakly supervised video anomaly detection. In: *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2023 Jun 17–24; Vancouver, BC, Canada. p. 8022–31. doi:10.1109/CVPR52729.2023.00775.
15. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7794–803. doi:10.1109/cvpr.2018.00813.
16. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? In: *Proceedings of the 38th International Conference on Machine Learning*; 2021 Jul 18–24; Online. p. 813–24.
17. Arnab A, Dehghani M, Heigold G, Sun C, Lucic M, Schmid C. ViViT: a video vision transformer. In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021 Oct 10–17; Montreal, QC, Canada. p. 6816–26. doi:10.1109/iccv48922.2021.00676.
18. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*; 2021 Jul 18–24; Online. p. 8748–63.
19. Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. *Int J Comput Vis*. 2022;130(9):2337–48. doi:10.1007/s11263-022-01653-1.
20. Zhou K, Yang J, Loy CC, Liu Z. Conditional prompt learning for vision-language models. In: *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022 Jun 18–24; New Orleans, LA, USA. p. 16795–804. doi:10.1109/cvpr52688.2022.01631.
21. Jia M, Tang L, Chen BC, Cardie C, Belongie S, Hariharan B, et al. Visual prompt tuning. In: *Computer vision—ECCV 2022*. Cham, Switzerland: Springer; 2022. p. 709–27. doi:10.1007/978-3-031-19827-4_41.
22. Wu P, Zhou X, Pang G, Zhou L, Yan Q, Wang P, et al. VadCLIP: adapting vision-language models for weakly supervised video anomaly detection. *Proc AAAI Conf Artif Intell*. 2024;38(6):6074–82. doi:10.1609/aaai.v38i6.28423.
23. Zanella L, Liberatori B, Menapace W, Poiesi F, Wang Y, Ricci E. Delving into CLIP latent space for video anomaly recognition. *Comput Vis Image Underst*. 2024;249:104163. doi:10.1016/j.cviu.2024.104163.
24. Yu G, Wang S, Cai Z, Liu X, Xu C, Wu C. Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement. In: *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022 Jun 18–24; New Orleans, LA, USA. p. 13967–78. doi:10.1109/cvpr52688.2022.01360.
25. Peng S, Cai Y, Yao Z, Tan M. Weakly-supervised video anomaly detection via temporal resolution feature learning. *Appl Intell*. 2023;53(24):30607–25. doi:10.1007/s10489-023-05072-8.
26. Zhang C, Li G, Qi Y, Wang S, Qing L, Huang Q, et al. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In: *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2023 Jun 17–24; Vancouver, BC, Canada. p. 16271–80. doi:10.1109/CVPR52729.2023.01561.
27. Fan Y, Yu Y, Lu W, Han Y. Weakly-supervised video anomaly detection with snippet anomalous attention. *IEEE Trans Circuits Syst Video Technol*. 2024;34(7):5480–92. doi:10.1109/TCSVT.2024.3350084.

28. Yang Z, Liu J, Wu P. Text prompt with normality guidance for weakly supervised video anomaly detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 18899–908. doi:10.1109/CVPR52733.2024.01788.
29. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS. Learning temporal regularity in video sequences. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 733–42. doi:10.1109/CVPR.2016.86.
30. Qiu H, Hou B. Temporal context and representative feature learning for weakly supervised video anomaly detection. *J Inf Intell.* 2025;3(6):481–91. doi:10.1016/j.jiixd.2025.06.001.
31. Liang W, Zhang J, Zhan Y. Weakly supervised video anomaly detection based on spatial-temporal feature fusion enhancement. *Signal Image Video Process.* 2024;18(2):1111–8. doi:10.1007/s11760-023-02828-0.
32. Shao W, Xiao R, Rajapaksha P, Wang M, Crespi N, Luo Z, et al. Video anomaly detection with NTCN-ML: a novel TCN for multi-instance learning. *Pattern Recognit.* 2023;143(5):109765. doi:10.1016/j.patcog.2023.109765.
33. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations (ICLR); 2017 Apr 24–26; Toulon, France.
34. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: Proceedings of the 2018 International Conference on Learning Representations (ICLR); 2018 Apr 30–May 3; Vancouver, BC, Canada.
35. Li M, Xu R, Wang S, Zhou L, Lin X, Zhu C, et al. CLIP-event: connecting text and images with event structures. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 16399–408. doi:10.1109/cvpr52688.2022.01593.
36. Wu P, Zhou X, Pang G, Sun Y, Liu J, Wang P, et al. Open-vocabulary video anomaly detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 18297–307. doi:10.1109/CVPR52733.2024.01732.
37. Wang J, Li G, Liu J, Xu Z, Chen X, Wei J. RelVid: relational learning with vision-language models for weakly video anomaly detection. *Sensors.* 2025;25(7):2037. doi:10.3390/s25072037.
38. Khattak MU, Wasim ST, Naseer M, Khan S, Yang MH, Khan FS. Self-regulating prompts: foundational model adaptation without forgetting. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 15144–54. doi:10.1109/iccv51070.2023.01394.
39. Khattak MU, Rasheed H, Maaz M, Khan S, Khan FS. MaPLe: multi-modal prompt learning. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 19113–22. doi:10.1109/CVPR52729.2023.01832.
40. Li M, Sang J, Lu Y, Du L. WSVAD-CLIP: temporally aware and prompt learning with CLIP for weakly supervised video anomaly detection. *J Imaging.* 2025;11(10):354. doi:10.3390/jimaging11100354.
41. Wang Y, Chen S. Learning event completeness for weakly supervised video anomaly detection. In: Proceedings of the 42nd International Conference on Machine Learning (ICML); 2025 Jul 13–19; Vancouver, BC, Canada. p. 62505–17.
42. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. *Int J Comput Vis.* 2010;88(2):303–38. doi:10.1007/s11263-009-0275-4.