Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.073850

ARTICLE

Enhancing Anomaly Detection with Causal Reasoning and Semantic Guidance

Weishan Gao[1,2], Ye Wang[1,2], Xiaoyin Wang[1,2] and Xiaochuan Jing[1,2,*]

[1]China Aerospace Academy of Systems Science and Engineering, Beijing, 100048, China
[2]Aerospace Hongka Intelligent Technology (Beijing) Co., Ltd., Beijing, 100048, China

*Corresponding Author: Xiaochuan Jing. Email: w654672829@163.com

Received: 27 September 2025; Accepted: 05 November 2025; Published: 12 January 2026

ABSTRACT: In the field of intelligent surveillance, weakly supervised video anomaly detection (WSVAD) has garnered widespread attention as a key technology that identifies anomalous events using only video-level labels. Although multiple instance learning (MIL) has dominated the WSVAD for a long time, its reliance solely on video-level labels without semantic grounding hinders a fine-grained understanding of visually similar yet semantically distinct events. In addition, insufficient temporal modeling obscures causal relationships between events, making anomaly decisions reactive rather than reasoning-based. To overcome the limitations above, this paper proposes an adaptive knowledge-based guidance method that integrates external structured knowledge. The approach combines hierarchical category information with learnable prompt vectors. It then constructs continuously updated contextual references within the feature space, enabling fine-grained meaning-based guidance over video content. Building on this, the work introduces an event relation analysis module. This module explicitly models temporal dependencies and causal correlations between video snippets. It constructs an evolving logic chain of anomalous events, revealing the process by which isolated anomalous snippets develop into a complete event. Experiments on multiple benchmark datasets show that the proposed method achieves highly competitive performance, achieving an AUC of 88.19% on UCF-Crime and an AP of 86.49% on XD-Violence. More importantly, the method provides temporal and causal explanations derived from event relationships alongside its detection results. This capability significantly advances WSVAD from a simple binary classification to a new level of interpretable behavior analysis.

KEYWORDS: Video anomaly detection (VAD); computer vision; deep learning; explainable AI (XAI); video understanding

1 Introduction

Video anomaly detection (VAD) is a critical technology in domains like public safety and industrial automation, providing a scalable alternative to the inherent limitations of human-led surveillance [1]. Within this field, the weakly supervised video anomaly detection (WSVAD) has become particularly influential. By training on video-level labels, WSVAD alleviates the intensive labor of temporal annotation and promotes better generalization across varied scenarios. However, despite these benefits, WSVAD frameworks struggle with a significant limitation: a lack of semantic specificity and predictive insight regarding anomalous events [2]. Contemporary models can flag the occurrence of an anomaly but often fail to classify its specific nature or offer actionable diagnostic information. In an industrial context, for example, a system might misinterpret harmless steam as smoke from equipment failure, or mistake routine welding sparks for a fire hazard, leading to costly operational disruptions.

This conceptual confusion stems from foundational constraints in WSVAD methodologies. Firstly, coarse, video-level binary labels provide a weak supervisory signal. This signal only confirms an anomaly's presence and is devoid of information about its specific nature or cause. The resulting lack of meaning-based grounding is the primary source of ambiguity. Secondly, dominant multiple instance learning (MIL) paradigms [3] are effective for leveraging weak labels. However, they typically aggregate snippet-level predictions through oversimplified temporal pooling strategies [4]. These strategies yield weak temporal modeling. They fail to capture the evolving progression and causal logic of events, treating videos as unordered sets of snippets rather than coherent narratives.

To overcome these challenges, research has progressed along two primary avenues. The first seeks to refine the MIL framework with more sophisticated temporal modeling [5–7] or saliency-based feature extraction [8]. While these techniques improve anomaly localization, they remain constrained by the semantically weak video-level labels and do not address the fundamental need for semantic discrimination. The second direction leverages large-scale, pre-trained multimodal models, such as VadCLIP [9], to align visual data with language representations. Although promising, this approach often creates superficial cross-modal links without deep, structured knowledge, and its decision-making process remains opaque. More critically, neither direction adequately addresses the weak temporal modeling inherent in standard MIL at a causal level, leading to a persistent inability to differentiate fine-grained anomalies, a lack of interpretability, and a purely reactive framework with no predictive capability for proactive intervention.

To directly address the intertwined problems of semantic ambiguity and weak temporal modelling, recent state-of-the-art approaches present specific characteristics that leave room for advancement. In the domain of semantic integration, vision-language models such as AnomalyCLIP [10] leverage valuable semantic priors, yet their dependence on static, pre-defined textual prompts can limit the capture of fine-grained, dynamic semantics required to distinguish complex anomaly categories. Concurrently, in temporal modelling, methods like PE-MIL [11] enhance multi-scale feature extraction but typically do not model the causal logic and evolutionary pathways of events, maintaining a focus on post-hoc detection.

This work proposes a novel framework to advance beyond these characteristics. Its strength lies in the synergistic interaction of its core contributions. The framework first introduces an adaptive knowledge-based guidance mechanism. This component constructs hierarchical "concept clouds" from Wikidata [12] and encodes them via BLIP [13]. The goal is to achieve fine-grained conceptual discrimination. This mechanism provides the precise meaning-based concepts necessary for the event relation analysis (ERA) module. The ERA module can then construct meaningful, causal-temporal chains of events. In turn, the ERA module provides predictive capability and causal interpretability. It does so by explicitly modeling temporal dependencies and causal correlations. Critically, the contextual and causal understanding from the ERA guides the conceptual alignment. This focuses attention on the most spatiotemporally relevant visual evidence, creating a closed-loop reasoning system. A temporal context fusion (TCF) module supports this synergy by enriching long-range dependency modeling. Ultimately, this enables a transformative shift from merely detecting anomalies to interpreting and forecasting event evolution.

The remainder of this paper proceeds as follows: Section 2 provides an overview of related work, analyzing the current state of weakly supervised video anomaly detection and prompt learning. Section 3 introduces the overall architecture of the proposed framework, detailing the temporal context fusion, prompt learning using external knowledge, and the event relation analysis modules. Section 4 presents the extensive experimental results and corresponding analyses to validate the effectiveness of the method. Finally, Section 5 contains the conclusion and discussion, outlining the key findings of this research and discussing potential directions for future studies.

2 Related Works

2.1 Weakly Supervised Video Anomaly Detection

VAD research has evolved along several distinct paradigms. A significant body of work focuses on unsupervised learning, which typically operates by learning a model of normal behavior from training data containing only normal events. Anomalies are then identified as deviations from this learned normality. Within this paradigm, one prominent approach is reconstruction-based modeling. For instance, the Inter-fused Autoencoder proposed by Aslam and Kolekar [14] utilizes a combination of CNN and LSTM layers to learn spatio-temporal patterns, flagging anomalies based on high reconstruction errors. An alternative and equally effective direction is prediction-based modeling. TransGANomaly [15] exemplifies this by using a Transformer-based generative adversarial network (GAN) to predict future video frames; a failure to accurately predict a frame indicates an anomalous event. While these methods are powerful, their effectiveness is centered on the assumption that anomalous events are patterns that deviate significantly from a well-defined normality. This assumption can present challenges in complex environments where the variety of normal events is vast and ever-changing.

In contrast, the key challenge of WSVAD is to achieve precise temporal localization of anomalous segments and effective prediction of risk evolution, using only video-level labels. Sultani et al. [16] pioneered a ranking loss–based framework that built the groundwork for later developments. Based on this foundation, Zhong et al. [17] introduced an in-packet consistency loss to enhance training efficiency. However, these early methods typically relied on selecting the highest-scoring segments within a video and treated them as independent instances. Such an approach disregarded the temporal continuity of behavior, resulting in the loss of critical contextual information.

To address this shortcoming, the focus of subsequent research shifted towards recovering temporal dependencies through contextual modeling. For example, Zhu and Newsam [18] applied an attention-based mechanism to capture inter-segment relations. This partially restored temporal structure. However, this approach lacked scene-level structural priors and demonstrated limited performance when handling complex behavioral patterns. To further improve temporal understanding, Cho et al. [3] employed Graph Convolutional Networks (GCNs) to explicitly model hierarchical dependencies within visual representations. These strategies progressively enhanced the modeling of evolving behaviors by optimizing contextual associations. Nevertheless, they all shared a critical and fundamental limitation: they are fundamentally reactive.

This “reactive” nature constitutes a significant gap in current WSVAD methods. Even as temporal modeling becomes increasingly sophisticated, existing frameworks still only detect an event as it occurs or after it has occurred. They cannot anticipate the evolution of a situation. This reactive nature persists even in recent methods with advanced temporal modeling. For instance, PE-MIL [11] employs pyramidal encoding to capture multi-scale temporal contexts; this enhances anomaly localization. However, its temporal reasoning remains implicit and correlational, confined within the MIL framework. The method lacks an explicit mechanism to model the causal logic and evolutionary pathways of events. Such a mechanism is essential for predictive risk assessment. This status quo necessitates a shift in research focus. To transition from passive detection to active risk assessment, the proposed ERA module is designed precisely to fill this gap. By learning these dynamic evolutionary patterns of events, the model transitions from simply identifying anomalies to proactively predicting their potential progression. Furthermore, to provide the rich contextual representation needed for such analysis, the TCF module is designed to capture both short-term and long-range dependencies more effectively than prior temporal modeling approaches.

Beyond the core algorithmic challenges of semantic grounding and causal reasoning, recent research has also explored other practical dimensions of VAD. One critical direction is improving system-level efficiency for real-time deployment. For example, some work proposes an edge-assisted framework. This framework utilizes a lightweight network on edge devices for initial anomaly screening. It transmits only suspicious frames to the cloud for in-depth recognition, saving bandwidth and computational resources [19]. Another important challenge is enabling systems to adapt to new, previously unseen anomaly types without requiring complete retraining. To this end, class-incremental learning networks have been developed, allowing models to learn new anomaly classes on the fly while retaining knowledge of existing ones [20]. While these directions address important system-level and lifelong learning challenges, this work remains focused on the fundamental problem of enhancing the core detection and reasoning capabilities of WSVAD models for known anomaly categories.

2.2 Applications of Prompt Learning in Video Understanding

Prompt learning has gained attention for enhancing conceptual awareness in video understanding. By using knowledge-based references in cross-modal learning, prompts guide model attention toward meaningful regions during inference. This improves alignment and recognition across modalities. In the context of anomaly detection, several studies have begun exploring the integration of prompt learning into WSVAD. Wang et al. [21] proposed using fixed natural language templates to guide action recognition, enabling the model to focus on action-relevant features. However, this manually defined prompting strategy lacked flexibility and struggled to generalize to the diverse and complex anomaly types encountered in WSVAD. To resolve this issue, Ju et al. [22] introduced learnable prompt vectors that can adaptively align with video features. This approach improves the detection of specific anomaly categories. However, these prompt-based methods remained limited to predefined class labels, lacking the capacity to represent visual phenomena.

Most existing prompt-based approaches rely on coarse-grained alignment strategies that are ill-suited for the fine-grained demands of anomaly detection. However, studies that attempt to incorporate external knowledge [5,11] tend to align concepts at the video level. However, they still fail to localize prompts to the relevant anomalous regions accurately. Yang et al. [23] applied implicit graph-based alignment through joint prediction, modestly improving feature-prompt correlation. However, this approach often underperformed in weak anomaly signals, misaligning prompts with irrelevant background regions and leading to a decline in detection accuracy. The limitations of implicit and coarse-grained alignment are also evident in state-of-the-art vision-language models adapted for VAD. Methods like AnomalyCLIP [10] leverage the powerful pre-trained knowledge of models like CLIP. However, they typically rely on static, hand-crafted textual prompts. These prompts are insufficient for capturing the fine-grained, hierarchical meaning of complex anomalous events. As a result, this often leads to superficial cross-modal alignment and limited conceptual discriminability.

To tackle these problems fundamentally, this study introduces a fine-grained visual-text alignment method for noise-sensitive WSVAD. A conceptual prompting module facilitates the precise integration of external knowledge. Specifically, the knowledge-based anchors module moves beyond static prompts. It constructs adaptive “concept clouds” from structured knowledge bases and leverages learnable prompt vectors. This integration achieves fine-grained, context-aware visual-text alignment. It provides the precise conceptual grounding needed to guide model attention. In parallel, an event relation analysis module models the temporal evolution of events. Together, these components significantly enhance the model’s capacity for subtle anomaly detection and proactive risk assessment.

3 Proposed Methodology

3.1 Overview

The proposed method restructures the conventional end-to-end anomaly detection process into three distinct yet interconnected stages, each designed to address a specific aspect of weakly supervised video anomaly detection. As illustrated in Fig. 1, these stages operate collaboratively and are jointly optimized through backpropagation, ensuring seamless information flow and unified performance enhancement.

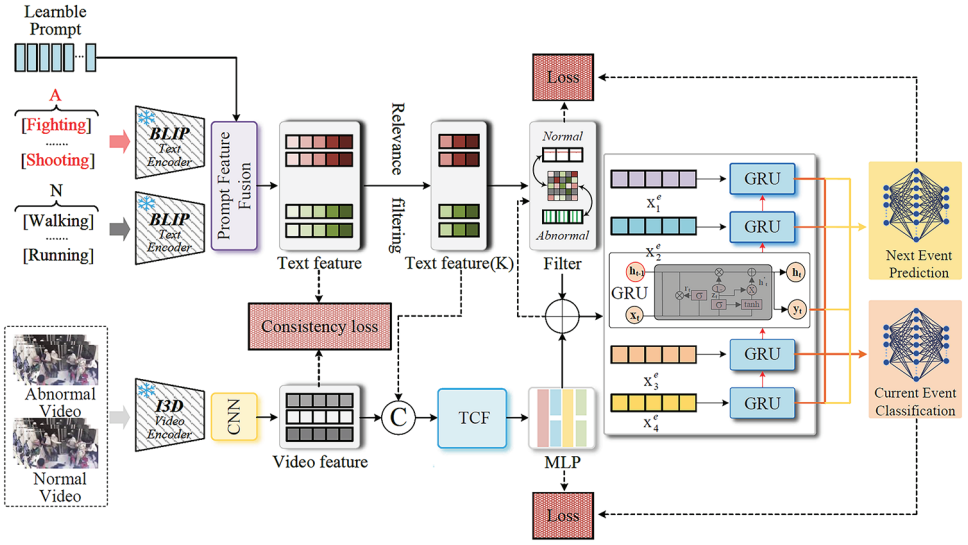

Figure 1: The overall architecture of the proposed framework. The model processes inputs through two parallel streams: a video stream for extracting visual features and a prompt stream for generating semantic anchors. These streams are aligned and fused, with the resulting features fed into the final event relation analysis module to model event evolution for classification and prediction

The first stage, the basic detection module. It determines if an anomaly exists and identifies its temporal location. This stage uses two submodules: the temporal context fusion component (Section 3.2), which captures short- and long-range dependencies, while the causal temporal predictor (Section 3.3), which out-puts segment-level anomaly scores. These scores serve as attention cues for subsequent stages.

The second stage is the semantic enhancement module. It performs a critical transformation of the visual features by integrating them with external conceptual knowledge. It leverages the anomaly scores from the first stage to distinguish between foreground (potentially anomalous) and background regions. Concept-based anchors, derived from knowledge graphs, are aligned with the most relevant foreground segments using an auxiliary alignment loss (Section 3.4). This process refines the raw visual representations into contextually enriched features that are more suitable for downstream tasks.

The third stage, the event relation analysis module, elevates the framework's capability from passive analysis to active pre-warning. This module models the evolving and logical sequence of events, transcending judgments of isolated moments. Building on the semantically enriched features from the second stage, it utilizes a gated recurrent unit (GRU) network to encode the temporal progression (Section 3.5). Finally, it employs two parallel prediction heads to forecast the current and subsequent event probabilities. This enables the system to anticipate future risks based on learned causal patterns.

The entire method is trained end-to-end with a multi-task objective. The primary MIL-based loss from the first stage drives accurate anomaly detection. In contrast, the conceptual alignment loss from the second

stage encourages feature representations to become more interpretable and discriminative. Joint training enables effective gradient sharing across all components. This allows the model to achieve high detection performance and meaningful semantic understanding simultaneously.

3.2 Temporal Context Fusion

For the fundamental anomaly detection phase, the TCF module is designed to capture intricate temporal connections adeptly. Current methodologies either fail to effectively simulate long-range interdependence by emphasizing local interactions [8] or inadequately represent local correlations by depending on global self-attention [24]. The overall architecture of the TCF module, which addresses these challenges, is illustrated in Fig. 2. The TCF module concurrently gathers information from extensive dependencies and neighbourhood connections, integrating them adaptively via an adaptive gating mechanism. The module focuses on creating an affinity matrix that represents the inherent relationships among pieces. An input feature sequence $Z \in R^{N \times D_{in}}$ is initially transformed into queries (Q), keys (K), and values (V) by three distinct linear projection functions: $g_q(\cdot)$, $g_k(\cdot)$ and $g_v(\cdot)$. The initial inter-segment similarity S_{raw} is calculated using the dot product of Q and K, with the outcome serving as the output of the linear projection functions $g_q(\cdot)$, $g_k(\cdot)$ and $g_v(\cdot)$, defined in Eq. (1).

$$S_{raw} = g_q(Z) \cdot g_k(Z)^T \tag{1}$$

The explicit integration of positional information into the similarity computation is achieved through a temporal relationality embedding (TRE), $E_{temporal}$. The components of this embedding matrix ($E_{temporal}(idx_1, idx_2)$) are created dynamically according to the absolute position indices of the segments (idx_1, idx_2) and are modulated by learnable scaling factors (κ and δ). These are generated dynamically and regulated by a learnable scale factor, κ, and a bias, δ, as given in Eq. (2).

$$E_{temporal}(idx_1, idx_2) = \exp(-|\kappa(idx_1 - idx_2)^2 + \delta|) \tag{2}$$

The temporal relevance embedding is incorporated into the original similarity to create the location-aware affinity matrix, $S_{pos} = S_{raw} + E_{temporal}$. The Gaussian kernel embedding inherently accentuates relationships among adjacent segments, allowing the model to adjust to varying spatial representations and disparate video durations. The wide-area dependency flow aims to capture contextual relationships on a global scale. The location-aware affinity matrix, S_{pos}, undergoes a conventional scaled dot-product attention calculation to derive a wide-area dependence map, $W_{global} = \text{softmax}\left(\frac{\tilde{S}_{pos}}{\sqrt{d_{key}}}\right)$. This map is subsequently employed to evaluate and aggregate the value representation V, producing the wide-area contextual features $C_{global} = W_{global} \cdot g_v(Z)$. A learnable adaptive balancing gate, $g_\alpha \in [0, 1]$, is utilized to combine the contextual information derived from the two streams. This adaptive balancing gate enables the model to adjust the contribution weights of global and local information in response to the inputs, as shown in Eq. (3).

$$C_{merged} = g_\alpha \odot C_{global} + (1 - g_\alpha) \odot C_{local} \tag{3}$$

In this context, $\odot$ signifies element-wise multiplication, and the g_α function can be inferred from input features via a compact network. The integrated feature, C_{merged}, is subjected to particular normalization and a linear transformation, $g_h(\cdot)$, before being amalgamated with the original input, Z, by residual concatenation. Ultimately, layer normalization (LN) is employed to provide the final, context-augmented output, C_{out}, which can be expressed as:

$$C_{out} = \text{LN}(Z + g_h(\mathcal{N}_{P2L}(C_{merged}))) \tag{4}$$

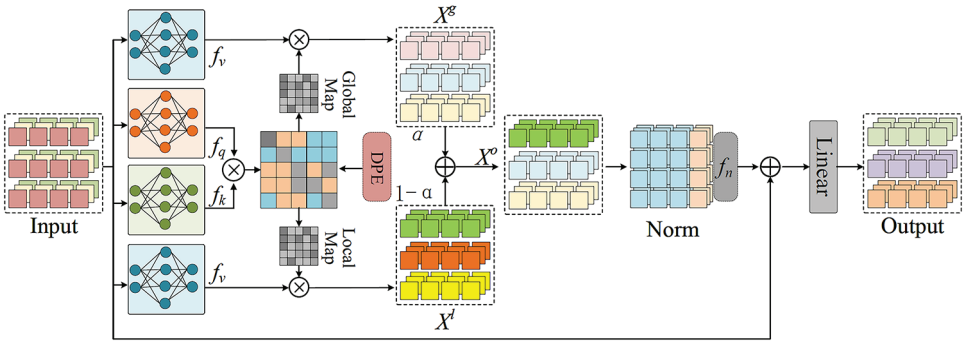

Figure 2: Architecture of the temporal context fusion module

3.3 Causal Temporal Predictor and Classifier

After the enhancement of the TCF module, to further refine the anomaly discrimination-oriented features and make predictions, the causal temporal predictor and classifier is introduced. The module consists of two sequential one-dimensional convolutional layers. Each layer incorporates a GELU activation function and a dropout layer, which enhances the model's nonlinear representation capability while mitigating overfitting. The procedure is depicted as follows:

$$\mathrm{H}_{\text{inter}} = \text{Dropout}(\text{GELU}(\text{Conv1D}_1(\mathrm{C}_{out}))) \tag{5}$$

$$\mathrm{H}_{final} = \text{Dropout}(\text{GELU}(\text{Conv1D}_2(\mathrm{H}_{\text{inter}})))$$

Utilizing the finely tuned discriminative features H_{final}, a temporal causal convolution head $h_{score}(\cdot)$ is employed to forecast the likelihood of anomalies in each video segment. This causal convolution ensures that only present and past data are utilized when forecasting the abnormal condition of the current clip, which is crucial for online or real-time detection contexts. The ultimate sequence of clip-level anomaly probabilities, $\mathrm{P}_{seg} \in [0,1]^N$, is derived by the sigmoid activation function, $\sigma(\cdot)$:$\mathrm{P}_{seg} = \sigma(h_{score}(\mathrm{H}_{final}))$, where $\mathrm{P}_{seg,t}$ represents the anomaly probability of the t-th clip in the video.

The MIL paradigm is employed to formulate the principal anomaly detection loss function, $\mathcal{L}_{\text{an}}$ [24]. A video-level anomaly prediction score p_{vid} is produced by aggregating the segment-level anomaly probability sequences P_{seg}. For a positive packet, the anomaly prediction score for the video is determined as the average of the k_{pos} probability values corresponding to the highest segment-level anomaly probabilities, where k_{pos} is defined as $\lfloor N_{seg}/16 + 1 \rfloor$. Generally, the highest clip-level anomaly probability is chosen as the anomaly prediction score for the video, specifically $k_{neg} = 1$, as demonstrated in Fig. 3. For a small batch comprising B video samples, each with video-level accurate labels $y_{vid,b}$ from the set, and associated video-level prediction scores $p_{vid,b}$, the MIL-based binary cross-entropy loss $\mathcal{L}_{\text{an}}$ is defined as follows:

$$\mathcal{L}_{\text{an}} = -\frac{1}{B}\sum_{b=1}^{B}\left[y_{vid,b}\log(p_{vid,b}) + (1-y_{vid,b})\log(1-p_{vid,b})\right] \tag{6}$$

3.4 Knowledge-Guided Semantic Integration

This module is the core of this method. It aims to integrate structured external knowledge into visual features and fundamentally improve the model's fine-grained conceptual recognition capability. Implementation involves three key steps: constructing semantic anchors, context separation, and visual-semantic alignment.

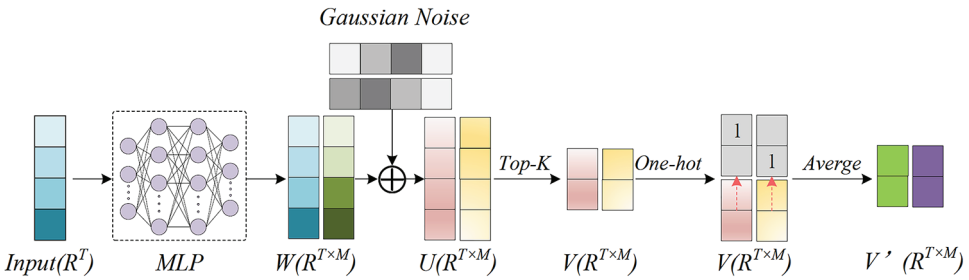

Figure 3: Top-k score selection module

3.4.1 External Knowledge Prompt Construction

To generate more representative semantic guidance signals than mere category labels, semantic anchors are constructed. Considering the versatility of semantic guidance, this work first selects 12 common relation types from Wikidata [12] as the pre-retrieval semantics, followed by identifying the relation types with the highest occurrence frequency across all anomaly categories as the core retrieval relations $R = \{r_1, r_2, r_3\}$. The concept dictionary is then constructed by retrieving all statements of the relation $r_j \in \mathrm{R}$ established with a given class c as the subject or object entity. The non-category entity in each statement is used as the key, and the number of references associated with the statement is taken as the value. To explicitly handle potential noise, ambiguity, and inconsistent entries from Wikidata that may interfere with the meaning of anomalies, a two-stage filtering process is applied. The method first eliminates all Wikidata statements with a ranking of "deprecated" to remove low-quality or disputed assertions. Next, for the remaining concepts, the number of references associated with each Wikidata statement is utilized as a relevance indicator. The remaining entries are then filtered by using the average of the reference counts as the threshold. In general, the higher the reference count, the stronger the reliability and consensus degree of the concept.

Subsequently, the text encoder $\mathcal{E}_{\text{text}}$ of the BLIP model [13] is employed to encode the filtered concepts into stable and representative semantic anchor vectors. The choice of BLIP over CLIP [25] is motivated by its superior ability to capture fine-grained semantic relationships [26] and its cleaner training data, which is more suitable for this knowledge-driven framework. For a given class c, a set of keys $\{concept_i^c\}_{i=1}^{N_c}$ from the concept dictionary is first extracted as the context concepts, which are then separately fed into the text encoder to extract d_{embed}-dimensional feature vectors $\mathrm{e}_i^c \in R^{d_{embed}}$: $\mathrm{e}_i^c = \mathcal{E}_{\text{text}}(concept_i^c)$. The semantic director $\mathrm{D}^c \in R^{d_{embed}}$ of a category c is then derived by averaging all feature vectors of its core concepts, as presented in Eq. (7).

$$\mathrm{D}^c = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathrm{e}_i^c \tag{7}$$

where N_c denotes the number of filtered concepts. The D^c vector can serve as a reliable and representative anchor point to direct the ensuing visual feature learning process.

3.4.2 Context-Sensitive Separation

To achieve accurate alignment, it is essential to distinguish between the foreground (anomalous) information and the background (normal) information in the video clip. We creatively employ the anomalous saliency score $s \in [0,1]^N$ generated during the base detection phase (Section 3.3) to derive foreground features $\mathrm{v}^{fg} \in R^{d_v}$ and background features $\mathrm{v}^{bg} \in R^{d_v}$ using a straightforward attentional method. This approach enables the model to "adaptively concentrate" on visual regions that are most likely to harbour

aberrant information. The foreground attention weights, $\alpha^{fg} \in R^N$, are calculated as follows:

$$\alpha_t^{fg} = \frac{\exp(\kappa \cdot s_t) - 1}{\sum_{k=1}^{N} (\exp(\kappa \cdot s_k) - 1)} \tag{8}$$

s_t is the anomalous significance score for the t-th fragment, while κ represents a predetermined scaling factor employed to amplify the importance of high significance scores. The expression $\exp(\cdot) - 1$ guarantees non-negativity and diminishes the weighting for lower scores. Similarly, the background attention weight, $\alpha^{bg} \in R^N$, is calculated utilizing the normal confidence of the clip, $\bar{s}_t = 1 - s_t$, as shown in Eq. (9).

$$\alpha_t^{bg} = \frac{\exp(\kappa \cdot \bar{s}_t) - 1}{\sum_{k=1}^{N} (\exp(\kappa \cdot \bar{s}_k) - 1)} \tag{9}$$

Utilizing these attention weights, we can derive the weighted video-level foreground feature, $\mathrm{v}^{fg} \in R^{d_v}$, and the background feature, $\mathrm{v}^{bg} \in R^{d_v}$, according to Eq. (10).

$$\begin{aligned} \mathrm{v}^{fg} &= \sum_{t=1}^{N} \alpha_t^{fg} \cdot \mathrm{X}_{refined,t} \\ \mathrm{v}^{bg} &= \sum_{t=1}^{N} \alpha_t^{bg} \cdot \mathrm{X}_{refined,t} \end{aligned} \tag{10}$$

where $\mathrm{X}_{refined,t}$ denotes the t-th feature segment in the sequence. Thus, v^{fg} emphasizes the characteristics of segments deemed abnormal by the model, whereas v^{bg} concentrates on the attributes of normal segments.

3.4.3 Visual and Semantic Alignment

During the alignment step, the objective is to minimize the distance between the foreground visual feature v^{fg} and its associated conceptual anchor D^c of the anomalous category within the feature space. The matching probability $\mathrm{sim}(\mathrm{a}, \mathrm{b}) = \frac{\mathrm{a} \cdot \mathrm{b}^T}{\|\mathrm{a}\| \|\mathrm{b}\|}$ is calculated using temperature-scaled cosine similarity, and a knowledge-distilled alignment loss $\mathcal{L}_{pl}$ is formulated to accomplish this objective. The primary anomaly detection loss, $\mathcal{L}_{\mathrm{an}}$, and the alignment loss, $\mathcal{L}_{pl}$, which are used for comprehensive joint optimization, determine the overall loss of the model.

For a particular video, if it is anomalous, its foreground feature v^{fg} must coincide with the associated semantic direction for the anomaly category, $\mathrm{D}^{abnormal}$. If the video is normal, its overall characteristics should correspond with the conventional 'normal' semantic framework D^{normal}. A collection $\{\mathrm{D}^k\}_{k=1}^{C+1}$ is created, comprising C anomalous category semantic guides and one standard semantic guide. The likelihood that a visual characteristic v corresponds to the k-th semantic guide D^k, $P(\mathrm{D}^k|\mathrm{v})$, is determined using temperature-scaled cosine similarity followed by softmax normalization, shown in Eq. (11).

$$P(\mathrm{D}^k|\mathrm{v}) = \frac{\exp(\sin(\mathrm{v}, \mathrm{D}^k)/\tau_s)}{\sum_{m=1}^{C+1} \exp(\sin(\mathrm{v}, \mathrm{D}^m)/\tau_s)} \tag{11}$$

τ_s is a parameter for temperature that can be learned. Alignment aims to optimize the likelihood of correlating positive sample pairs while reducing the likelihood of correlating negative sample pairs. This can be accomplished through the application of a knowledge-distilled alignment loss, which is calculated as follows:

$$\mathcal{L}_{pl} = E_{\mathbf{v}\sim p(\mathbf{v})}\left[\sum_{k=1}^{C+1} Q(\mathrm{D}^k|\mathrm{v}) \log \frac{Q(\mathrm{D}^k|\mathrm{v})}{P(\mathrm{D}^k|\mathrm{v})}\right] \tag{12}$$

where $Q(\mathrm{D}^k|\mathrm{v})$ denotes the target distribution. $Q(\mathrm{D}^k|\mathrm{v})$ approaches 1 for positive sample pairings and approaches 0 for negative sample pairs.

3.5 Event Relation Analysis for Active Risk Assessment

The ERA module elevates the model's capability from passive detection to active pre-warning. Its primary objective is to model the evolving and logical sequence of event progressions. This enables the prediction of future events, transcending the judgment of isolated moments. The module's workflow begins by receiving the concept-enriched feature sequence from the TCF and KGSI modules. It then encodes the temporal information using a Recurrent Neural Network (RNN). Finally, two parallel prediction heads output probabilities for both current and subsequent events.

Specifically, the input to the ERA module is the semantic-enhanced feature sequence $X^e = \{x_1^e, x_2^e, \ldots, x_T^e\} \in R^{T\times D}$, where T is the sequence length and D is the feature dimension. The core of the module's internal processing is an event state encoding stage, which utilizes a Gated Recurrent Unit (GRU) network. The GRU is selected for its ability to effectively capture long-term dependencies while maintaining fewer parameters and higher computational efficiency compared to other recurrent structures, such as LSTM. The GRU network processes the input sequence x^e sequentially at each time step t, taking the current feature x_t^e and the previous hidden state h_{t-1} to compute the current hidden state:

$$h_t = GRU\left(x_t^e, h_{t-1}\right) \tag{13}$$

This hidden state $h_t \in R^{D_h}$ is crucial, as it serves as a dynamic representation of the event's history up to time t, encoding not only the semantic information of the current snippet but also key features from the historical sequence.

Based on the encoded event state sequence $H = \{h_1, h_2, \ldots, h_T\}$, two parallel predictive outputs are designed that function as prediction heads. To determine the current event probability ($P_{current\_event}$) at time step t, the hidden state h_t is passed through a linear transformation layer, $W_{current}$, and a Sigmoid activation function, σ, yielding the prediction $P_{\text{current\_event},t} = \sigma(W_{\text{current}} h_t)$. This results in a final prediction sequence $P_{\text{current\_event}} \in R^{T\times K}$. Similarly, to predict the probability of the next event ($P_{\text{next\_event}}$) at time step $t+1$, the same hidden state h_t is leveraged and pass it through an independent linear layer, W_{next}, and a Sigmoid function to produce $P_{\text{next\_event},t} = \sigma(W_{\text{next}} h_t)$. The operation of the ERA module is further guided by an external knowledge base, the core of which is an event transition matrix ($M_{\text{trans}} \in R^{K\times K}$). This matrix provides prior knowledge about event causality, it provides supervision signals during the training phase, as detailed in the following section.

3.6 Model Training and Optimization Objective

This paper employs a multi-task learning strategy to jointly optimize all modules of the framework end-to-end. The overall objective function, $\mathcal{L}$, is a composite loss consisting of three main components, corresponding to the key tasks of anomaly detection, knowledge enhancement, and event relation prediction. In addition to the MIL-based anomaly detection loss $\mathcal{L}_{\text{an}}$ and the prompt-based semantic alignment loss $\mathcal{L}_{pl}$ from the base framework, a new event relation modeling objective ($\mathcal{L}_{era}$) is introduced. This new loss function is responsible for supervising the learning of the internal evolutionary rules of events and is itself composed of two constituent parts.

The first part is the current event classification loss ($\mathcal{L}_{\text{event}}$), which supervises the ERA module's ability to recognize the class of the current video snippet. It is formulated as a Binary Cross-Entropy (BCE) loss: $\mathcal{L}_{\text{event}} = \text{BCE}(P_{\text{current\_event}}, y_{\text{event}})$. The target label y_{event} is a multi-hot vector generated from the video's ground-truth annotations; the accuracy and granularity of these annotations are critical for the model's performance. The second part is the event transition prediction Loss ($\mathcal{L}_{\text{relation}}$), which trains the model's ability to predict future events. Its supervisory signal, $y_{\text{next\_event}}$, is dynamically generated by multiplying the true current event label y_{event} with the knowledge base's event transition matrix M_{trans}, yielding $y_{\text{next\_event}} = y_{\text{event}} \cdot M_{\text{trans}}$. The loss is then computed as $\mathcal{L}_{\text{relation}} = \text{BCE}(P_{\text{next\_event}}, y_{\text{next\_event}})$. The total loss for the ERA module is the sum of these two components: $\mathcal{L}_{\text{era}} = \mathcal{L}_{\text{event}} + \mathcal{L}_{\text{relation}}$.

Ultimately, the final joint loss function for the entire framework is a weighted sum of these three objectives:

$$\mathcal{L} = \mathcal{L}_{an} + \lambda \mathcal{L}_{pl} + \gamma \mathcal{L}_{\text{era}} \tag{14}$$

where λ and γ are hyperparameters that balance the importance of the different tasks, by minimizing this joint loss function, the model synergistically learns to detect anomalies, understand meaning and predict trends. This constructs a comprehensive and robust video anomaly analysis system transcending traditional detectors.

4 Experiments

4.1 Datasets and Evaluation Metrics

The XD-Violence dataset [27] is an extensive collection of 4754 unedited videos, covering six categories of violent events sourced from movies, surveillance cameras, and web content. For weakly supervised settings, this dataset is divided into 3954 training videos and 800 testing videos. In contrast, the UCF-Crime dataset [16] comprises 13 categories of abnormal events captured in a wide range of environments, including streets, residential areas, and commercial spaces. This dataset provides 1610 training videos and 290 testing videos.

Following established protocols in previous studies [8,16], the area under the curve (AUC) of the frame-level receiver operating characteristic (ROC) curve is adopted as the evaluation metric for the UCF-Crime dataset. Meanwhile, for the XD-Violence dataset, the area under the precision-recall curve (AP) at the frame level is used as the evaluation metric [28].

4.2 Experimental Settings

Following established methodologies [5,24], the framework utilises a pre-trained I3D model [29] to extract 1024-dimensional features from RGB video streams based on the Kinetics dataset [30]. The features are extracted from non-overlapping 16-frame segments, which is the standard temporal unit for I3D features and ensures fair comparison with existing methods. For prompt learning with external knowledge, conceptual representations for 13 UCF-Crime anomaly categories and generic XD-Violence categories are expanded using Wikidata. These expanded concepts are encoded into 768-dimensional semantic prototypes using the BLIP ViT-B/16 text encoder. Visual features are aligned with semantic prototypes via a temperature-guided contrastive learning mechanism, where the temperature parameter τ_s is a learnable parameter initialized to 0.07, a standard value in contrastive learning frameworks. For the Top-K strategy in the MIL loss, the adaptive approach $k_{pos} = \lfloor \text{N\_seg}/16 + 1 \rfloor$ was adopted, following the established practice in prior works [11,28] which has proven robust for videos of varying lengths.

Model training is conducted in an end-to-end multi-task learning framework. The Adam optimizer is adopted with an initial learning rate of 1×10^{-4}, weight decay of 5×10^{-4}, batch size of 128, and a total of 50 training epochs. A cosine annealing schedule is applied to adjust the learning rate over time.

For generating snippet-level pseudo-labels for the ERA module under weak supervision, the pseudo-label y_{event} is derived from the video-level ground-truth annotations. For instance, in a video labeled as anomalous due to "Arson", this category label is propagated to every snippet within that video, serving as the initial supervisory signal for the current event classification loss ($\mathcal{L}_{\text{event}}$). Although this labeling strategy introduces noise, the MIL framework helps the model implicitly identify salient snippets. The ERA module then leverages these snippets to model temporal event evolution.

4.3 Comparison to Current Methods

Results on XD-Violence. Table 1 reports the evaluation results on the XD-Violence dataset. Using the exact I3D feature representations as competing methods, the proposed method achieves an AP of 86.49%, outperforming most existing semi-supervised and weakly supervised approaches. This result highlights the effectiveness of the multi-task learning design in detecting anomalies under complex scene conditions.

Table 1: Comparison with other methods on XD-Violence

Supervision	Year	Method	Feature	AP (%)
Semi-supervised	–	SVM baseline	–	50.78
	2016	Conv-AE [31]	–	30.77
Weakly-supervised	2021	RTFM [4]	I3D-RGB	77.81
	2022	MSL [32]	I3D-RGB	78.28
	2022	MSL [32]	VSwin-RGB	78.59
	2023	Cho et al. [3]	I3D-RGB	81.30
	2023	NG-MIL [33]	I3D-RGB	78.51
	2024	AnomalyCLIP [10]	ViT-B/16	78.51
	2024	PE-MIL [11]	I3D-RGB	88.05
	2024	Ghadiya et al. [34]	I3D-RGB	86.34
	2025	MMVAD [35]	I3D-RGB	81.98
	2025	DAKD [36]	I3D-RGB	85.12
	–	Ours	I3D-RGB	86.49

Note: SVM, Support Vector Machine; Conv-AE, Conv-Auto-Encoder; RTFM, Robust Temporal Feature Magnitude learning; MSL, Multi-Sequence Learning; NG-MIL, Normality Guided Multiple Instance Learning; PE-MIL, Prompt-Enhanced Multiple Instance Learning; MMVAD, Multimodal VAD; DAKD, Distilling Aggregated Knowledge with Disentangled Attention.

Results on UCF-Crime. Table 2 presents the comparative performance on the UCF-Crime dataset. The proposed framework achieves an AUC of 88.19%, demonstrating competitive results against existing methods. This improvement is primarily attributed to the Knowledge-Guided Semantic Integration module (KGSI). This module integrates structured conceptual information via "semantic anchors." Unlike traditional approaches, this mechanism enhances the model's discriminative capacity, particularly on UCF-Crime. This dataset features diverse anomaly categories and high conceptual complexity. These findings indicate the method's effectiveness in addressing the long-standing challenge of knowledge insufficiency in WSVAD.

Table 2: Comparison with other methods on UCF-Crime

Supervision	Year	Method	Feature	AUC (%)
Semi-supervised	2019	GODS [37]	BOW+TCN	70.46
Weakly-supervised	2022	MSL [32]	C3D-RGB	82.85
	2022	MSL [32]	I3D-RGB	85.30
	2022	MSL [32]	VideoSwin-RGB	85.62
	2023	NG-MIL [33]	C3D-RGB	83.43
	2023	NG-MIL [33]	I3D-RGB	85.63
	2023	CoMo [3]	I3D-RGB	86.10
	2024	AnomalyCLIP [10]	ViT-B/16	86.36
	2023	UR-DMU [8]	I3D-RGB	86.97
	2024	PE-MIL [11]	I3D-RGB	86.83
	2025	DAKD [36]	I3D-RGB	87.71
	2025	MMVAD [35]	I3D-RGB	87.78
	2025	ViCap-AD [38]	I3D-RGB	87.20
	–	Ours	I3D-RGB	88.19

Note: GODS, Generalized One-class Discriminative Subspaces; CoMo, Context–Motion Interrelation Module; UR-DMU, Uncertainty Regulated Dual Memory Units; ViCap-AD, Video Caption Anomaly Detector.

Fine-Grained Recognition Performance. To further validate its efficacy, this study evaluates the proposed approach on the more demanding task of fine-grained anomaly recognition. This capability is demonstrated across both datasets through fine-grained mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds (mAP@IoU) results for UCF-Crime (Table 3) and detailed per-class AP results for XD-Violence (Fig. 4). As outlined in Table 3, the model consistently outperforms contemporary methods, including the fine-grained specialist VADCLIP, across the entire range of IoU thresholds, achieving a state-of-the-art average of mAP (AVG) of 11.97. This result empirically corroborates the central hypothesis: a meticulous alignment between visual evidence and precise semantics is pivotal for enhancing recognition accuracy. The performance margin is particularly accentuated at higher IoU thresholds, underscoring the robustness of the proposed saliency-guided selective alignment strategy.

In contrast to prior works that often establish coarse-grained, localized associations between category labels and event proposals, this framework introduces a dual-level granularity refinement. Firstly, it enriches semantic representation by substituting rudimentary category labels with comprehensive semantic anchors derived from external knowledge. Secondly, it institutes a dynamic attention mechanism that leverages anomaly salience scores to selectively focus on the most discriminative frames within a candidate region. This direct and refined correlation between semantically rich concepts and visually salient evidence empowers the model to delineate the full temporal extent of anomalous events with superior precision, especially for those that are subtle or intricate in nature.

To provide a more granular analysis of the model's fine-grained recognition capabilities, Fig. 4 presents a per-class performance comparison on the XD-Violence dataset against AnomalyCLIP [10]. The results show that this approach demonstrates a significant advantage in most categories. For instance, it achieves a nearly tenfold increase in AP for "Abuse" (60.5% vs. 6.1%) and a clear improvement for "Shooting" (59.3% vs. 26.1%). While AnomalyCLIP [10] holds a slight edge in the well-represented "Riot" category (92.7% vs. 91.0%), this

method remains highly competitive and surpasses the baseline in most other tested scenarios. This detailed breakdown highlights the effectiveness of the proposed framework in enhancing the discriminative power for a diverse range of complex events, particularly those that require a deeper semantic understanding.

Table 3: Fine-grained comparisons on UCF-Crime

Method	mAP@IoU (%)					
	0.1	0.2	0.3	0.4	0.5	AVG
AVVD [39]	10.27	7.01	6.25	3.42	3.29	6.05
VADCLIP [9]	11.72	7.83	6.40	4.53	2.93	6.68
ITC [40]	13.54	9.24	7.45	5.46	3.79	7.90
Ours	15.92	14.13	12.01	9.72	8.05	11.97

Note: AVVD, Audio-Visual Violence Detection; ITC, Injecting Text Clues.

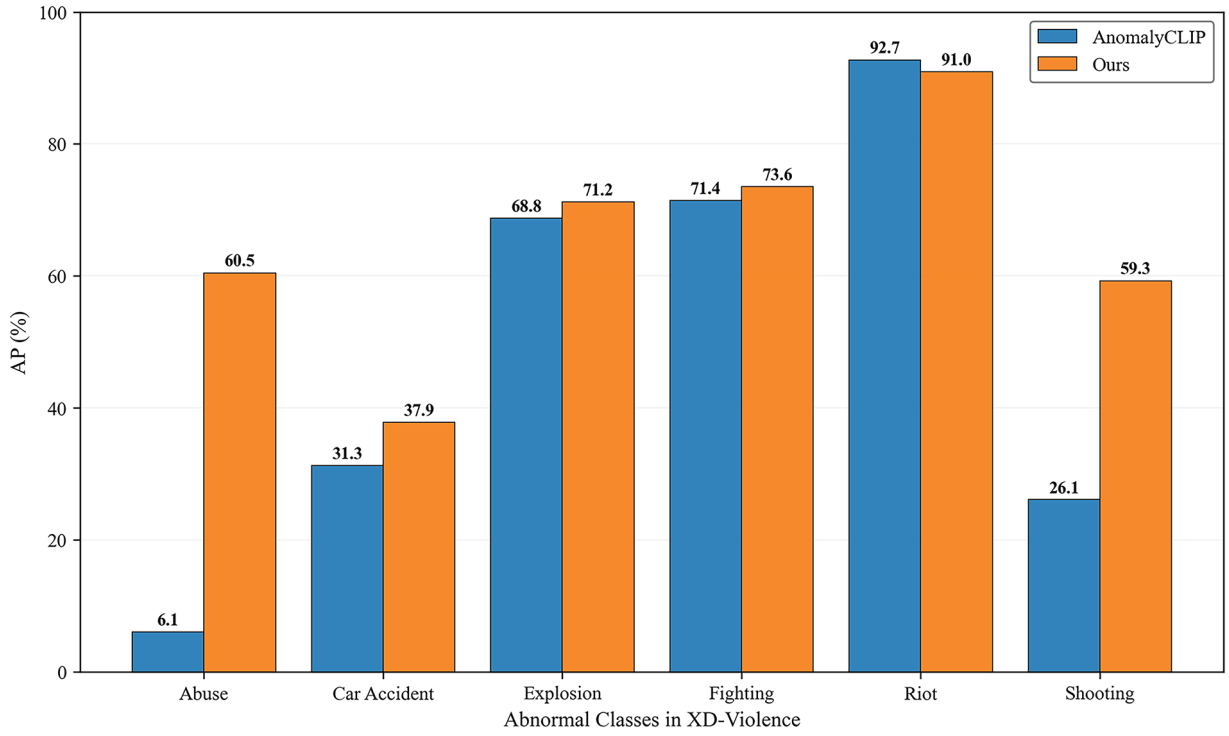

Figure 4: AP results on individual anomaly classes of XD-Violence

4.4 Ablation Study

This study conducts a comprehensive ablation study to dissect the contributions of individual components and their synergistic effects within the proposed method. For this analysis, a baseline model is established equipped solely with the temporal context fusion module. This baseline configuration omits the TRE mechanism, utilizing standard self-attention for contextual encoding. Furthermore, the anomaly classification head processes the fused visual features directly to generate an anomaly score. The training of this baseline is supervised exclusively by the anomaly detection loss ($\mathcal{L}_{an}$), without the contrastive loss from the KGSI module. Table 4 details the results of this study. To assess the practical viability of the framework, its operational efficiency is also evaluated using two key metrics: Test Time (s), which represents the average

inference latency for a single video clip, and Model Size (MB), which denotes the storage footprint of the final model weights. A detailed analysis of these efficiency metrics reveals that the incremental computational cost of the proposed modules is minimal. The complete model incurs only a 57% relative increase in inference time (from 0.042 s to 0.066 s) and a 34% increase in model size (from 94 MB to 126 MB) over the baseline, while delivering a substantial 3.93% absolute improvement in AUC. This favorable trade-off underscores the architectural efficiency of the design. The findings collectively demonstrate that the complete model strikes an effective balance between predictive accuracy and computational overhead, underscoring its suitability for deployment in practical applications.

Table 4: Ablation studies of proposed modules

Baseline	TRE	KGSI	ERA	UCF AUC (%)	XD AP (%)	Test times (s)	Model size (MB)
✓				84.26	83.29	0.042	94
✓	✓			85.47	84.39	0.049	96
✓	✓	✓		87.25	85.98	0.064	112
✓	✓		✓	86.38	85.03	0.054	101
✓	✓	✓	✓	88.19	86.49	0.066	126

Note: The table follows a cumulative design where each row adds a new component to the previous configuration. A checkmark (✓) indicates the module is included in that configuration, while an empty cell indicates it is excluded.

An analysis of the ablation results in Table 4 elucidates the contribution of each key component. Initially, incorporating the TRE mechanism into the baseline (Row 1 vs. Row 2) yields a 1.21% increase in AUC, with a negligible increase in latency (0.007 s) and model size (2 MB), confirming its efficiency in enhancing temporal modeling. Subsequently, the integration of the KGSI module (Row 2 vs. Row 3) delivers a further AUC enhancement of 1.78%. This significant performance gain comes with a moderate computational cost (an additional 0.015 s and 16 MB). This cost is justified by the critical role of conceptual disambiguation. Finally, the entire model configuration, which incorporates the ERA module within a multi-task framework (Row 3 vs. Row 5), achieves peak performance with an AUC of 88.19%. The incorporation of the ERA module for causal reasoning adds only 0.002 s of inference time and 14 MB of parameters, demonstrating that the advanced capability for active risk assessment can be achieved with minimal overhead. This pinnacle result illustrates that a well-formulated auxiliary task, such as fine-grained anomaly categorization, can provide informative gradient signals that regularize the main task, reinforcing the advantages of a co-training paradigm. Crucially, this runtime analysis confirms the model's practical viability. The complete model's inference speed of 0.066 s per clip translates to a processing throughput of approximately 15 clips per second. This high throughput rate, representing the average inference latency, comfortably meets the stringent requirements for real-time analysis, underscoring the model's suitability for deployment in practical surveillance applications.

This study further validates the utility of semantic anchors through a comparison of different prompt templates (see Table 5). The results demonstrate that employing Wikidata-augmented semantic prototypes yields notable performance gains over a baseline that uses only class names. Specifically, this approach boosts the AUC on UCF-Crime by 1.30% and the AP on XD-Violence by 1.23%. This outcome suggests a strong link between the semantic richness of prompts and the model's discriminative power. It confirms that a more descriptive knowledge-based foundation is crucial for enhancing anomaly detection accuracy.

Table 5: Performance comparison of different prompt templates in the KGSI module

Prompt template	UCF-Crime AUC (%)	XD-Violence AP (%)
KGSI module using only class names as prototypes	86.89	85.26
KGSI module with Wikidata-extended prototypes	88.19	86.49

4.5 Hyperparameter Sensitivity Analysis

The final performance of the proposed model relies on the effective balancing of its multi-task learning objectives, controlled by hyperparameters λ and γ. A comprehensive sensitivity analysis was conducted by varying one hyperparameter across a wide range of values while keeping the other fixed, with performance reported on both the UCF-Crime and XD-Violence datasets, as shown in Fig. 5.

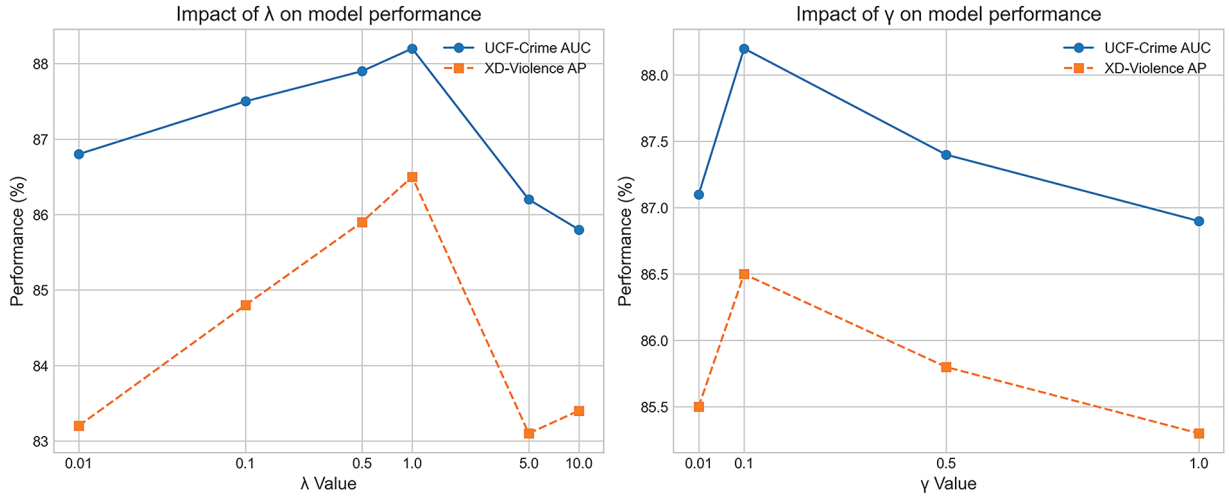

Figure 5: Impact of hyperparameters λ and γ on model performance

The hyperparameter λ governs the contribution of the semantic alignment loss $\mathcal{L}_{pl}$. As illustrated in Fig. 5 (left), the model's performance on both datasets consistently peaks when $\lambda = 1.0$. An optimal weight of 1.0, which makes the semantic alignment loss equally crucial as the primary anomaly detection loss ($\mathcal{L}_{an}$). This underscores that this module is not merely a regularizer but a core component, as it makes the semantic alignment loss equally crucial as the primary anomaly detection loss. When λ is too small (e.g., 0.01), the model cannot fully leverage the provided semantic guidance, limiting its fine-grained recognition capability. Conversely, when λ becomes excessively large (e.g., 10.0), the training objective overemphasizes semantic matching at the expense of the primary detection task, leading to performance degradation.

Similarly, the weight γ for the event relation modeling loss ($\mathcal{L}_{\text{era}}$) acts as a powerful, yet subtle, regularizer. The performance trend, shown in Fig. 5 (right), confirms that completely removing this objective ($\gamma = 0$) results in a significant performance drop. The model's performance reaches its optimum on both datasets when $\gamma = 0.1$. However, as γ increases beyond this point (e.g., to 1.0), a consistent decline is observed. This finding suggests that while the predictive guidance from $\mathcal{L}_{\text{era}}$ is highly beneficial, an excessively high weight can lead to over-regularization, where the more complex future-prediction task begins to interfere with the stable learning signals from the primary detection and alignment tasks. Therefore, a weight of $\gamma = 0.1$ provides the ideal balance.

Impact of attention scaling factor κ. Beyond the loss weighting factors, the sensitivity of the attention scaling factor κ, used in the context-sensitive separation (Eqs. (8) and (9)), was investigated. This parameter controls the sharpness of the distribution separating foreground and background features. As summarized in Table 6, performance peaks at $\kappa = 10$, achieving an AUC of 88.19% on UCF-Crime and an AP of 86.49% on XD-Violence. Lower values (e.g., $\kappa = 1$ or 5) result in a more uniform attention distribution, diluting the focus on the most salient anomalous regions. Conversely, a higher value ($\kappa = 20$) over-concentrates the attention, potentially excluding relevant contextual information and leading to a slight performance decline. Thus, $\kappa = 10$ is identified as the optimal setting, effectively balancing focused attention with contextual awareness.

Table 6: The impact of attention scaling factor κ on model performance

κ Value	UCF-Crime AUC (%)	XD-Violence AP (%)
1	86.12	84.55
5	87.78	86.01
10	88.19	86.49
20	88.03	86.17

Impact of GRU hidden layer size. The dimensionality of the hidden state in the GRU network directly influences the capacity of the ERA module to encode temporal event evolution. The impact of varying this hidden size is presented in Table 7. A hidden size of 256 yields the best performance, indicating it provides sufficient representational capacity for capturing complex event dynamics without overfitting. A smaller hidden size of 128 appears to limit the model's temporal modelling capability, while a larger size of 512 may introduce overfitting, as evidenced by the slight degradation in performance metrics.

Table 7: The impact of GRU hidden layer size on model performance

GRU Hidden Size	UCF-Crime AUC (%)	XD-Violence AP (%)
128	87.72	85.91
256	88.19	86.49
512	88.02	86.04

4.6 Qualitative Results

To intuitively demonstrate the effectiveness and interpretive power of the proposed ERA module, the relational knowledge it learned after being trained on the UCF-Crime dataset is visualized. This qualitative analysis offers insight into the model's capacity to comprehend the logical progression of events. Figs. 6 and 7 present the results.

Fig. 6 presents a heatmap of the global event transition matrix captured by the ERA module after training on UCF-Crime. In this matrix, each cell (i, j) represents the learned transition strength from a source event i (on the y-axis) to a subsequent event j (on the x-axis), providing a comprehensive map of all pairwise event correlations. This heatmap serves as a visual "map" of the model's learned common sense, revealing which events are likely to follow others based on the model's understanding. Brighter cells indicate stronger learned connections.

The heatmap demonstrates that the model has successfully learned logical and high-frequency event progressions that align with real-world intuition. Several strong correlations are observable that would be

expected by security professionals: A particularly bright cell at the intersection of the “Explosion” row and “Assault” column, with the highest value of 0.062 in the matrix, indicates the model has robustly learned that explosions often lead to violent confrontations or assaults in the aftermath. Similarly, a strong link is visible from “Stealing” to “Assault” (0.052), capturing a logical escalation from theft to violent confrontation. The model also identifies a notable connection from “Robbery” to “Burglary” (0.046), reflecting a recognized association between different types of property crimes. This visual evidence is crucial as it demonstrates that the ERA module is not a “black box.” The ability to inspect such a matrix provides practitioners with an interpretable tool to understand the model’s reasoning, confirming that its predictions are grounded in logical, learnable event sequences rather than just opaque statistical correlations. This degree of transparency represents a key step toward building trust in automated surveillance systems.

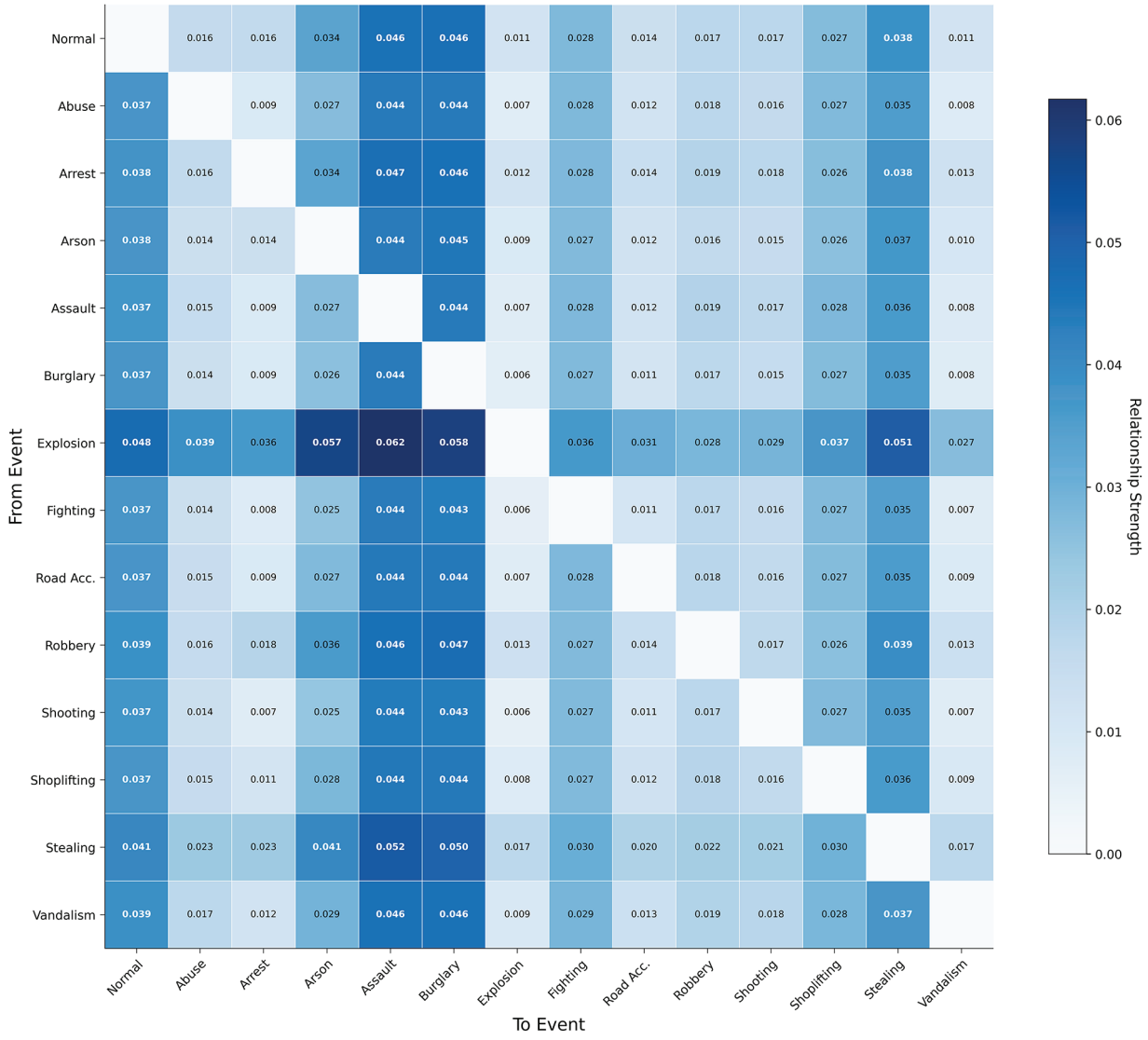

Figure 6: Heatmap of the learned event transition strengths on the UCF-Crime dataset

Building upon this global relationship map, Fig. 7 zooms in on the most critical findings by visualizing the top-3 ranked high-risk event trajectories. Each trajectory illustrates a multi-step sequence of events that is highly likely to escalate into a severe anomaly. Crucially, the ability to identify these longer chains showcases

the ERA module's primary strength in modeling long-range temporal dependencies. For instance, the top-ranked trajectory reveals a complex pattern of "Arrest" → "Assault" → "Burglary" → "Normal" → "Stealing" → "Arson", with a cumulative score of 0.207. The presence of a "Normal" segment within a high-risk chain is particularly insightful, demonstrating that the model can capture subtle, real-world criminal behaviors, such as a perpetrator feigning normalcy between illicit acts.

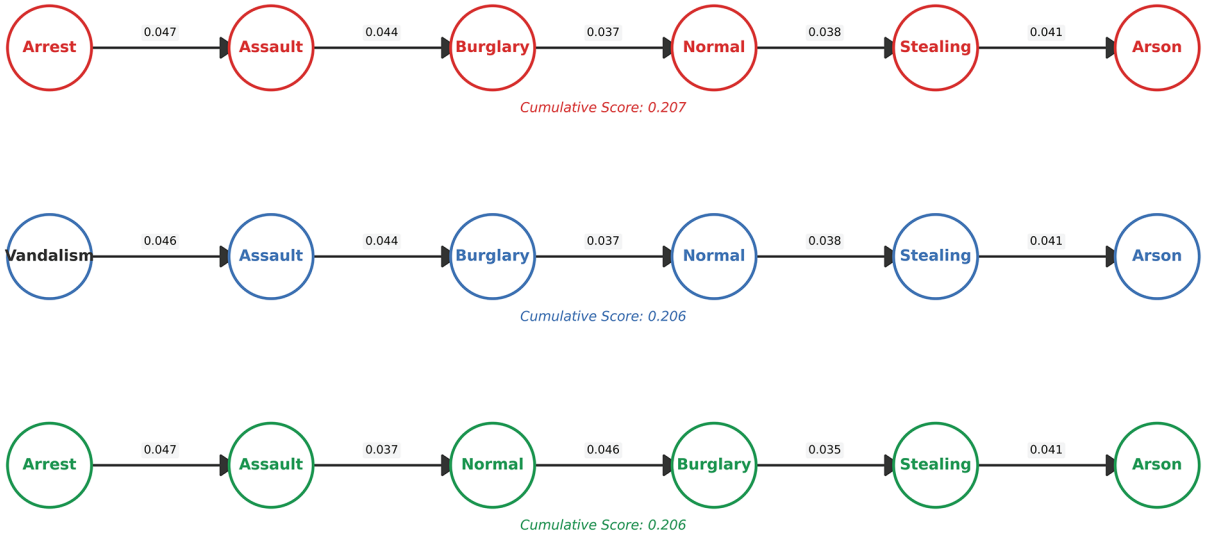

Figure 7: Learned evolutionary pathways of high-risk anomalies

This result vividly demonstrates the model's capacity for active risk assessment. By moving beyond the detection of isolated events to understanding their evolutionary patterns, the ERA module provides a deeper, more actionable insight into how dangerous situations develop over time.

To visually substantiate the quantitative results and demonstrate the method's practical efficacy, a qualitative analysis of the generated anomaly scores on representative videos from the UCF-Crime and XD-Violence datasets is presented. The anomaly score curves, depicted in Fig. 8, offer a clear visualization of the model's proficiency in temporal localization. The figure provides compelling evidence of the model's adaptability in handling anomalies of varying durations and characteristics. For instance, in the 'Explosion' example shown in Fig. 8a, the model accurately captures the entire duration of a long, developing event, with the anomaly score gradually increasing as the situation escalates. Conversely, for an abrupt, short-duration event like 'Fighting' shown in Fig. 8b, the model responds with a sharp and distinct spike, precisely pinpointing the critical moment. Furthermore, the method performs equally well when handling entirely normal scenarios, as Fig. 8c demonstrates, where the anomaly score consistently remains low, robustly confirming the model's reliability in distinguishing between normal and abnormal behavior. Similarly, for an anomaly event like 'Shooting' shown in Fig. 8d, which occurs in the first half of the video, the model quickly raises the anomaly score close to 1.0 and maintains a high score throughout the event's duration. This robust capability to handle diverse temporal patterns validates the effectiveness of the TCF module. It confirms that the module, enhanced by the TRE, successfully captures both long-range contextual dependencies and short-term, sudden changes in the video sequence. The strong alignment between the predicted scores and the ground-truth abnormal periods (indicated by the shaded regions) demonstrates that this approach can learn discriminative spatio-temporal representations for accurate anomaly localization, even under the challenging conditions of weak supervision without frame-level annotations.

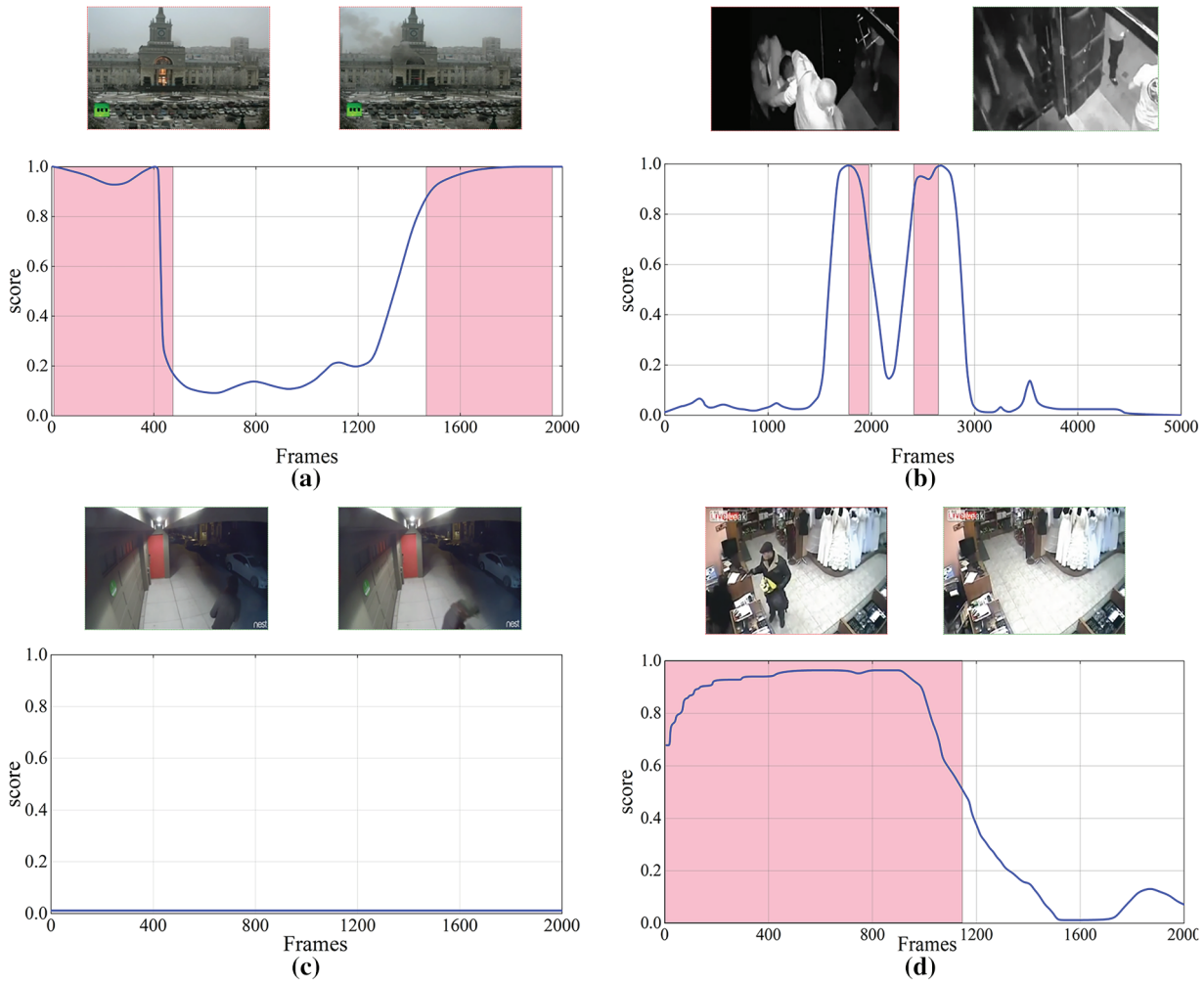

Figure 8: Visual results of the proposed method on the UCF-Crime and XD-Violence datasets. The blue curve shows the predicted anomaly scores, the light-pink shaded areas mark ground-truth anomalous frames, and red/green boxes highlight representative abnormal and normal events

5 Conclusion and Discussion

This study addresses two central limitations in weakly supervised video anomaly detection: the coarse-grained supervision of the MIL framework, which hinders fine-grained recognition, and the absence of temporal-causal reasoning, leading to opacity in the decision-making process. To overcome these challenges, a novel method is introduced that synergistically integrates dynamic semantic guidance with event relation analysis. This approach departs from traditional binary-supervised methods by first incorporating a dynamic semantic guidance module. This component leverages semantic prototypes derived from external knowledge (Wikidata) to provide robust, fine-grained supervision, enabling the model to not only detect anomalies but also distinguish between specific types of anomalies. On the XD-Violence dataset, the model achieves an average AP of 65.6% in fine-grained anomaly classification, demonstrating its capability to perform detailed and accurate semantic discrimination. Furthermore, an ERA module is introduced to model the explicit evolutionary patterns of events. By learning the logical sequences of how situations escalate, the ERA module provides the crucial temporal and causal context for its predictions. By revealing an event's development over time, the framework demonstrates the logical reasoning behind why a situation is deemed

anomalous, thereby addressing a core interpretability limitation of conventional weakly supervised models. The effectiveness of the integrated framework is confirmed by its strong detection performance, achieving a frame-level AUC of 88.19% on the UCF-Crime dataset and an AP of 86.49% on the XD-Violence dataset.

Notably, the presented framework achieves this performance utilizing only RGB inputs. This design choice is motivated by the need to ensure model generality and deployment feasibility. In numerous real-world surveillance scenarios, such as public live streams or legacy security systems, audio data is often unavailable, unreliable, or sensitive to privacy concerns. The establishment of a state-of-the-art baseline with the most universally available visual modality reinforces the framework's broad applicability. It is acknowledged that audio can provide valuable complementary cues, and its exclusion here constitutes a limitation that points to a clear future direction.

Ablation studies further confirm the complementary effects of the two core contributions: the semantic guidance that enhances what the model sees, and the event relation analysis that provides a logical structure for how events unfold. Together, these facilitate a qualitative transition from simple anomaly detection to a more comprehensive understanding of anomalies. This progression reflects a broader shift from data-driven pattern recognition to knowledge-guided video comprehension, establishing a new paradigm for weakly supervised visual analysis.

Building upon this foundation, future research will focus on advancing the framework's predictive and reasoning capabilities. A clear path forward involves extending the ERA module's predictive horizon to multi-step risk forecasting and creating a more holistic system by incorporating complementary data streams. This direction directly addresses the current limitation by integrating critical audio cues available in datasets like XD-Violence. Furthermore, to build upon the introduced causal-temporal reasoning, subsequent work will incorporate more formal causal frameworks, such as structural causal models (SCMs), to achieve a more theoretically grounded understanding of event causality.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design, Weishan Gao and Ye Wang; writing—original draft preparation, Weishan Gao and Xiaoyin Wang, writing—review and editing, Xiaoyin Wang and Xiaochuan Jing; data curation, Ye Wang; supervision, Xiaochuan Jing. All authors reviewed the results 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 on the website: https://www.crcv.ucf.edu/projects/real-world/ (accessed on 01 September 2025) and https://roc-ng.github.io/XD-Violence/ (accessed on 01 September 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Abdalla M, Javed S, Al Radi M, Ulhaq A, Werghi N. Video anomaly detection in 10 years: a survey and outlook. Neural Comput Appl. 2025;37(32):26321–64. doi:10.1007/s00521-025-11659-8.
2. Caetano F, Carvalho P, Mastralexi C, Cardoso JS. Enhancing weakly-supervised video anomaly detection with temporal constraints. IEEE Access. 2025;13:70882–94. doi:10.1109/ACCESS.2025.3560767.
3. Cho M, Kim M, Hwang S, Park C, Lee K, Lee S. Look around for anomalies: weakly-supervised anomaly detection via context-motion relational learning. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 12137–46. doi:10.1109/CVPR52729.2023.01168.
4. 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.
5. Pu Y, Wu X, Yang L, Wang S. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. IEEE Trans Image Process. 2024;33(11):4923–36. doi:10.1109/TIP.2024.3451935.
6. Yang Z, Wu P, Liu J, Liu X. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In: Proceedings of the Computer Vision—ECCV 2022; 2022 Oct 23–27; Tel Aviv, Israel. p. 404–21. doi:10.1007/978-3-031-19772-7_24.
7. Li G, Cai G, Zeng X, Zhao R. Scale-aware spatio-temporal relation learning for video anomaly detection. In: Proceedings of the Computer Vision—ECCV 2022; 2022 Oct 23–27; Tel Aviv, Israel. p. 333–50. doi:10.1007/978-3-031-19772-7_20.
8. Zhou H, Yu J, Yang W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. Proc AAAI Conf Artif Intell. 2023;37(3):3769–77. doi:10.1609/aaai.v37i3.25489.
9. 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.
10. 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.
11. Chen J, Li L, Su L, Zha ZJ, Huang Q. Prompt-enhanced multiple instance learning 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. 18319–29. doi:10.1109/CVPR52733.2024.01734.
12. Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase. Commun ACM. 2014;57(10):78–85. doi:10.1145/2629489.
13. Li J, Li D, Xiong C, Hoi S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 2022 International Conference on Machine Learning (ICML); 2022 Jul 17–23; Baltimore, MD, USA. p. 12888–900.
14. Aslam N, Kolekar MH. Unsupervised anomalous event detection in videos using spatio-temporal inter-fused autoencoder. Multimed Tools Appl. 2022;81(29):42457–82. doi:10.1007/s11042-022-13496-6.
15. Aslam N, Kolekar MH. TransGANomaly: transformer based generative adversarial network for video anomaly detection. J Vis Commun Image Represent. 2024;100(7):104108. doi:10.1016/j.jvcir.2024.104108.
16. 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 18–22; Salt Lake City, UT, USA. p. 6479–88. doi:10.1109/CVPR.2018.00678.
17. Zhong JX, Li N, Kong W, Liu S, Li TH, Li G. Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 16–20; Long Beach, CA, USA. p. 1237–46. doi:10.1109/CVPR.2019.00133.
18. Zhu Y, Newsam S. Motion-aware feature for improved video anomaly detection. arXiv:1907.10211. 2019. doi:10.48550/arxiv.1907.10211.
19. Hussain A, Khan N, Khan ZA, Yar H, Baik SW. Edge-assisted framework for instant anomaly detection and cloud-based anomaly recognition in smart surveillance. Eng Appl Artif Intell. 2025;160(1):111936. doi:10.1016/j.engappai.2025.111936.
20. Hussain A, Ullah W, Khan N, Khan ZA, Yar H, Baik SW. Class-incremental learning network for real-time anomaly recognition in surveillance environments. Pattern Recognit. 2026;170(5):112064. doi:10.1016/j.patcog.2025.112064.
21. Wang M, Xing J, Mei J, Liu Y, Jiang Y. ActionCLIP: adapting language-image pretrained models for video action recognition. IEEE Trans Neural Netw Learn Syst. 2025;36(1):625–37. doi:10.1109/TNNLS.2023.3331841.

22. Ju C, Han T, Zheng K, Zhang Y, Xie W. Prompting visual-language models for efficient video understanding. In: Proceedings of the Computer Vision—ECCV 2022; 2022 Oct 23–27; Tel Aviv, Israel. p. 105–24. doi:10.1007/978-3-031-19833-5_7.
23. 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 17–21; Seattle, WA, USA. p. 18899–908. doi:10.1109/CVPR52733.2024.01788.
24. Wu P, Liu J. Learning causal temporal relation and feature discrimination for anomaly detection. IEEE Trans Image Process. 2021;30:3513–27. doi:10.1109/TIP.2021.3062192.
25. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. arXiv:2103.00020. 2021.
26. Manzoor MA, Albarri S, Xian Z, Meng Z, Nakov P, Liang S. Multimodality representation learning: a survey on evolution, pretraining and its applications. ACM Trans Multimed Comput Commun Appl. 2024;20(3):1–34. doi:10.1145/3617833.
27. 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: Proceedings of the Computer Vision—ECCV 2020; 2020 Aug 23–28; Glasgow, UK. p. 322–39. doi:10.1007/978-3-030-58577-8_20.
28. Gao W, Wang X, Wang Y, Jing X. Dual-stream attention-enhanced memory networks for video anomaly detection. Sensors. 2025;25(17):5496. doi:10.3390/s25175496.
29. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. The kinetics human action video dataset. arXiv:1705.06950. 2017. doi:10.48550/arXiv.1705.0695.
30. 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.
31. 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.
32. 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.
33. Park S, Kim H, Kim M, Kim D, Sohn K. Normality guided multiple instance learning for weakly supervised video anomaly detection. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 2664–73. doi:10.1109/WACV56688.2023.00269.
34. Ghadiya A, Kar P, Chudasama V, Wasnik P. Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2024 Jun 17–18; Seattle, WA, USA. p. 1965–74. doi:10.1109/CVPRW63382.2024.00202.
35. Biswas D, Tesic J. MMVAD: a vision-language model for cross-domain video anomaly detection with contrastive learning and scale-adaptive frame segmentation. Expert Syst Appl. 2025;285(1):127857. doi:10.1016/j.eswa.2025.127857.
36. Dalvi J, Dabouei A, Dhanuka G, Xu M. Distilling aggregated knowledge for weakly-supervised video anomaly detection. In: Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2025 Feb 26–Mar 6; Tucson, AZ, USA. p. 5439–48. doi:10.1109/WACV61041.2025.00531.
37. Wang J, Cherian A. GODS: generalized one-class discriminative subspaces for anomaly detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 8200–10. doi:10.1109/iccv.2019.00829.
38. Lim J, Lee J, Kim H, Park E. ViCap-AD: video caption-based weakly supervised video anomaly detection. Mach Vis Appl. 2025;36(3):61. doi:10.1007/s00138-025-01676-x.
39. Wu P, Liu X, Liu J. Weakly supervised audio-visual violence detection. IEEE Trans Multimed. 2023;25:1674–85. doi:10.1109/TMM.2022.3147369.
40. Liu T, Lam KM, Bao BK. Injecting text clues for improving anomalous event detection from weakly labeled videos. IEEE Trans Image Process. 2024;33(11):5907–20. doi:10.1109/TIP.2024.3477351.