CLUE-VAD: Structured Semantic Clues for Understanding Explainable Events in Video Anomaly Detection

MyoungChul Kim, JunHee Lee, ChaeBeen Bang, and MyeongAh Cho†

Kyung Hee University, Republic of Korea
{2018102095, jhlee39, coqls1229, maycho}@khu.ac.kr

Abstract. Weakly Supervised Video Anomaly Detection aims to detect abnormal events in long, untrimmed videos using only video-level supervision, providing a practical solution to the high cost of temporally localized annotations. However, abnormal events in real-world surveillance are often determined by complex contextual interactions among actions, objects, and environments, making it difficult for existing approaches to identify which factors actually cause an anomaly. Most prior methods rely on holistic representations or explanations that lack explicit attribution to the semantic factors driving abnormal events. In this paper, we present CLUE-VAD, a structured semantic decomposition framework for explainable weakly supervised video anomaly detection. Rather than relying solely on holistic video representations or free-form explanations, CLUE-VAD decomposes each video segment into structured semantic clues—*Action*, *Environment*, and *Object*—and estimates their category-aware contributions to anomaly prediction. By linking anomaly scores to clue-specific weights and keyword-level evidence, we provides factor-level attribution and supports transparent explanation generation. Extensive experiments on VAD benchmarks demonstrate that CLUE-VAD achieves strong detection performance among X-VAD methods while providing fine-grained, clue-grounded explanations for anomaly predictions. The project page is available at `https://visualsciencelab-khu.github.io/CLUE-VAD_project/`.

Keywords: Video Anomaly Detection · Semantic Reasoning · X-VAD

1 Introduction

Video Anomaly Detection (VAD) is a fundamental task in intelligent surveillance systems, aiming to automatically identify rare and abnormal events in long, untrimmed video streams. Due to the high cost of temporally localized annotations, Weakly Supervised Video Anomaly Detection (WSVAD) has become a widely adopted setting, where only video-level labels indicating the presence of abnormal events are available during training. While early WSVAD approaches primarily focused on improving detection accuracy, recent studies [9, 11, 34, 37, 41] have increasingly highlighted the importance of understanding the underlying causes of anomaly predictions. In real-world surveillance

† Corresponding author

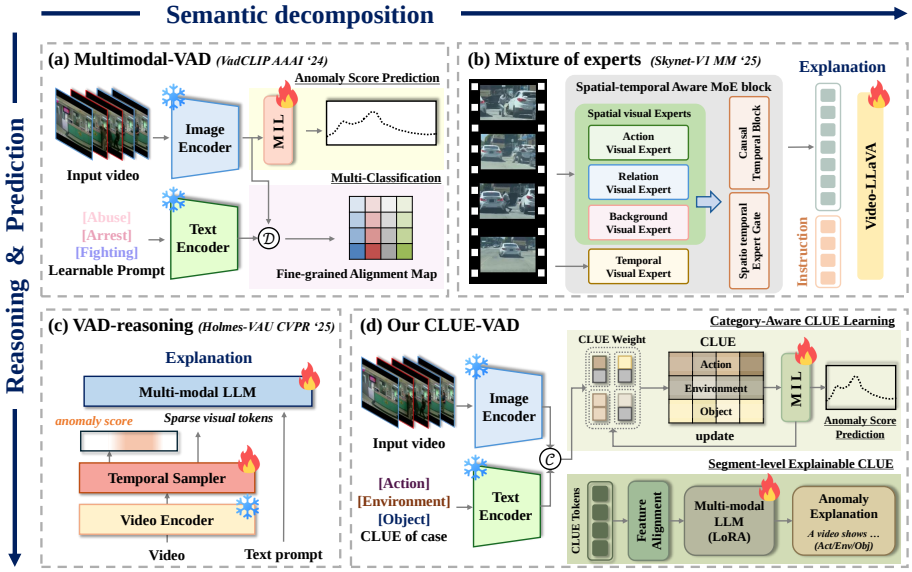

Fig. 1: Overview of representative paradigms in WSVAD. (a) **Multi-modal VAD** injects language cues via vision–language features, but largely relies on holistic scene alignment, limiting interpretability. (b) **Mixture of experts** decomposes videos into multiple vision-level experts and fuses them with MoE, yet its structure follows visual axes rather than language-grounded semantics. (c) **VAD-reasoning** uses single-frame supervision and instruction-tuned MLLMs to generate explanations for detected segments, but remains holistic without cue-level disentanglement. (d) **CLUE-VAD** explicitly decomposes each segment into Action, Environment, and Object, enabling context-aware and explainable anomaly reasoning.

scenarios, abnormality is often determined by complex contextual factors, making it insufficient to simply flag an event as anomalous without providing supporting evidence. This has motivated growing interest in approaches that move beyond pure detection toward more interpretable and reasoning-aware anomaly analysis, including emerging directions such as Explainable Video Anomaly Detection (X-VAD). However, effectively capturing and reasoning about the diverse contextual factors underlying abnormal events remains a fundamental challenge for current VAD approaches.

Existing methods still struggle to disentangle the contextual factors that contribute to abnormal events. To better understand recent progress in WSVAD, prior work can be broadly categorized into several paradigms (Figure 1). (a) **Multi-modal VAD** [35] incorporates semantic knowledge from visual-text models such as CLIP through cross-modal alignment. While these approaches improve anomaly representation, they rely on scene-level abnormal categories and offer limited interpretability of anomaly decisions. (b) **Mixture-of-Experts VAD** [21] decomposes visual representations into diverse branches and aggregates them via a Mixture-of-Experts mechanism. Although this design captures diverse visual factors, it lacks explicit semantic reasoning and does not provide interpretable attribution or quantitative score for abnormalities. (c) **VAD-**

reasoning [41] approaches leverage instruction-tuned multimodal LLMs to generate natural-language explanations. However, the explanations are typically produced in a holistic and generative manner, without explicitly modeling the individual semantic factors that contribute to anomaly decisions. Overall, existing approaches either rely on holistic representations or lack explicit semantic attribution, making it difficult to reason about which contextual factors lead to abnormal events.

In real-world scenarios, anomalies are inherently context-sensitive: the same action may be perceived as normal in one environment but suspicious in another. Capturing such contextual nuances requires structured reasoning that reflects how humans interpret abnormal situations—by jointly considering *who* is involved, *what* action occurs, and *where* the event takes place. For instance, a person running in a park during daytime is typically benign, whereas running with a weapon inside a school corridor may indicate a dangerous situation. These examples highlight the importance of fine-grained semantic decomposition for reliable anomaly interpretation. Motivated by this observation, we introduce the concept of **Textual CLUEs**, a set of structured semantic components consisting of **Action**, **Environment**, and **Object**, which capture the key dimensions that characterize abnormal events. By explicitly decomposing video segments into these clues, we aim to enable both more accurate anomaly detection and more interpretable reasoning.

To this end, we present **CLUE-VAD**, a practical structured semantic decomposition framework that supports clue-level reasoning for explainable WSVAD. Our method first leverages a pretrained video-language model to extract semantic textual clues from each video segment, corresponding to the Environment, Object, and Action dimensions. These clues are then integrated through a category-aware scoring mechanism that learns how different semantic cues contribute to specific anomaly types. This design enables the model not only to improve detection accuracy but also to reveal which semantic clue is responsible for the anomaly decision. Furthermore, we introduce keyword-level attribution within each clue, allowing the model to highlight the most influential semantic tokens that drive anomaly predictions. For example, **environmental** terms such as “fire” or “smoke” strongly influence explosion-related anomalies, **object** terms like “knife” or “gun” dominate robbery scenarios, and **action** terms such as “punch” or “kick” become decisive cues in fighting events. With structured clue-based decomposition, CLUE-VAD supports fine-grained reasoning over anomaly factors and, via MLLM training, yields interpretable explanations through clue-specific scoring.

Our contributions are summarized as follows:

- We formulate a structured semantic decomposition strategy that represents abnormal events using *Action*, *Environment*, and *Object* clues, enabling interpretable factor-level analysis of contextual anomaly factors.
- We introduce a **clue-aware anomaly reasoning mechanism** that models how semantic clues contribute to specific anomaly types through category-aware scoring.
- We propose a **clue-grounded explanation framework** that links anomaly predictions to clue-level scores and keyword-level evidence, enabling transparent and interpretable explanations of abnormal events.

– Experiments on standard WSVAD benchmarks show that CLUE-VAD achieves strong detection performance with modest but consistent gains over explainable VAD baselines, while providing fine-grained clue-level explanations.

2 Related Work

Vision–language models for WSVAD. Recent advances in WSVAD have increasingly leveraged vision–language models (VLMs) [1, 4, 14, 39] to inject high-level semantic cues into MIL based [6, 7, 9, 19, 27, 30] anomaly detection frameworks. For example, VadCLIP [35] adapts CLIP [25] with learnable textual prompts centered around predefined anomaly keywords, improving cross-modal alignment beyond fully handcrafted prompts. Prompt-engineering variants such as PEMIL [3] and TPWNG [36] further refine text–vision alignment, while LEC-VAD [33] emphasizes event completeness through a dual (category-agnostic/aware) branch and a Gaussian-mixture prior to smooth anomaly scores and recover contiguous intervals. In contrast, our CLUE-VAD does not primarily modify the scoring dynamics but instead restructures the semantic clues used by the scorer. Each segment is explicitly decomposed into three language-grounded clues and their category-aware importance is learned under weak supervision. This design produces per-clue scores and keywords, enabling competitive detection performance while providing transparent, clue-level attribution without requiring additional annotations.

Explainable VAD and open-world understanding. Explainable VAD has evolved from *human-interpretable signals* to *vision–language reasoning*. Early work such as EVAL [26] explains anomalies through an "instrument panel" of interpretable appearance/motion attributes, enabling clear scene-specific rationales. With the rise of foundation models, training-free pipelines like LAVAD [39] estimate anomaly scores by prompting LLMs with VLM-generated frame captions and temporal summaries. To reduce prompt brittleness without heavy supervision, VERA [37] learns guiding questions from coarse labels to elicit more reliable reasoning from frozen VLMs. In parallel, Ex-VAD [12] generates anomaly explanations from captions and uses them as an auxiliary textual modality, together with label-augmented alignment, to support fine-grained anomaly categorization and interpretation. For long-term and open-world understanding, Holmes-VAU [41] introduces hierarchical multi-granularity supervision and an anomaly-focused temporal sampler to handle extended events. Despite these advances, existing methods typically produce holistic explanations and do not explicitly support category-aware, role-disentangled reasoning with learnable per-category cue fusion, motivating our clue-decomposition perspective in WSVAD.

Semantic Decomposition for Explainable VAD. Recent explainable VAD studies increasingly adopt *semantic decomposition* to capture context-dependent abnormality. TDSD [28] decouples scene semantics into context- and object-oriented textual cues and injects them into a MIL-based WSVAD pipeline to improve scene-aware detection. In parallel, Sherlock [20] and Skynet-V1 [21] perform decomposition in the *vision*

space by splitting videos into multiple experts (e.g., action, object-relation, background, global/temporal) and fusing them via MoE, enabling structured event understanding and explanation. HAWK [29] further explores open-world anomaly understanding by factorizing motion and aligning motion cues with language through large-scale descriptions and QA supervision. However, these methods typically use decomposition as auxiliary guidance or expert routing, and they do not explicitly *disentangle semantic contributions* to anomaly scores or *learn category-specific semantic weights* under weak supervision. In contrast, we introduce a language-grounded three-clue decomposition (*Action, Environment, Object*) and learns per-category clue fusion with MIL, producing clue-level (and keyword-level) attributions for transparent, fine-grained explanations.

3 CLUE-VAD Structure

To enable interpretable and clue-aware video anomaly detection, we construct a new textual description built upon existing video anomaly datasets. Specifically, we utilize the UCF-Crime and XD-Violence datasets, which is a large-scale untrimmed structure commonly used in weakly supervised video anomaly detection. Our description pipeline transforms unstructured video content into structured and semantically aligned textual clues, which we refer to as CLUEs. This pipeline is designed to be generalizable and can be applied to other video anomaly detection datasets, enabling scalable and flexible extraction of textual descriptions. The CLUE contains three semantic information: action, environment, and object. Please refer to the Appendix for a detailed description about Structure.

Pipeline of CLUE Captioning. To generate structured textual representations aligned with human reasoning, we build a CLUE captioning pipeline that decomposes each video segment into three semantically distinct clues(Action, Environment, and Object). Using InternVideo 2.5, we split each untrimmed surveillance video into non-overlapping 64-frame segments and apply three clue-specific prompts per segment to separately capture the main activity, contextual setting, and involved entities. The resulting multi-perspective captions are stored per segment, providing diverse contextual cues that support fine-grained interpretation of abnormal events.

Semantic Analysis of CLUE-Based Descriptions. To quantify the information captured by each CLUE, we use LLaMA3 [8] to extract the top-k keywords from the action, environment, and object captions, and aggregate them by anomaly category and by label (normal vs. abnormal). Complete frequency tables are provided in the Appendix, here we discuss only the most salient shifts. The results show that abnormal segments prominently feature violent verbs such as “attacking” and “fighting” in the action clue. For Explosion and Arson, the environment clue frequently includes fire-related expressions such as “on fire” and “burning.” Meanwhile, object captions for Burglary and Shooting emphasize weapons such as “rifle” and “handgun,” whereas Shoplifting clips more often mention everyday items like “tablet.” These systematic clue-specific keyword shifts demonstrate that CLUE effectively captures discriminative semantics aligned with anomaly contexts.

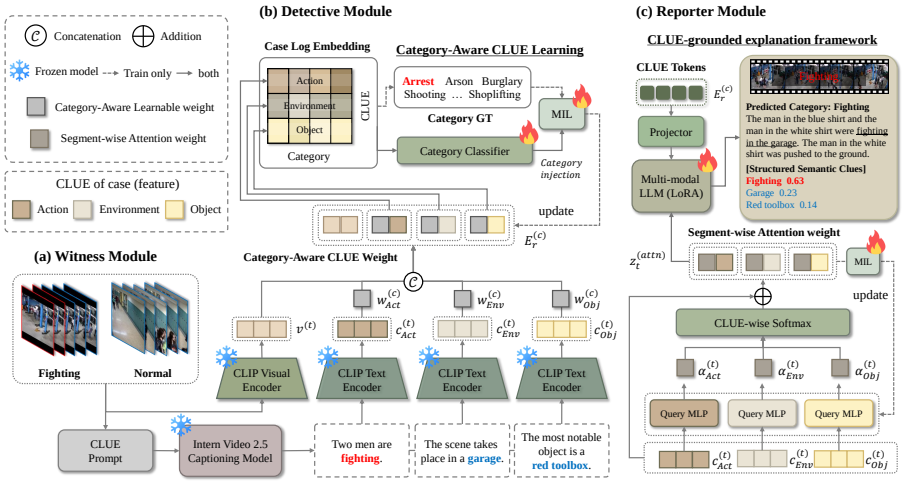

Fig. 2: CLUE-VAD main framework. It consists of three modules: (a) **Witness** extracts clue-specific features using a video-language model, (b) **Detective** performs category-aware anomaly prediction with MIL loss, and (c) **Reporter** provides interpretable scores through clue attention for segment-wise attribution and integrates an MLLM to generate natural language explanations.

4 CLUE-VAD framework

CLUE-VAD instantiates the proposed structured semantic decomposition framework for explainable WSVAD by incorporating structured textual cues to improve both detection performance and explainability. Figure 2 illustrates the overall pipeline, which consists of three core modules. **(a) Witness Module** extracts clue-wise features using clue-specific captions through a pretrained video-language model, enabling semantic decomposition of each event. **(b) Detective Module** then performs anomaly prediction using category-aware clue learning, where learnable weights determine the contribution of each clue for different anomaly types. This branch is optimized for prediction accuracy under weak supervision using MIL loss. For interpretation ability, **(c) Reporter Module** introduces a clue-grounded explanation framework, which leverages segment-wise clue attention to highlight semantically meaningful clues and integrates an MLLM to extend explanations beyond anomaly scores to natural language descriptions. By utilizing the clue-attention factor, the framework generates explainable anomaly scores while identifying temporally salient segments. Together, these modules bridge low-level video understanding and high-level anomaly reasoning, enabling both strong detection performance and post-hoc interpretability.

4.1 Witness Module

Given an input video, we uniformly divide it into T temporal segments, where each segment covers 16 frames. For every segment, we generate three separate textual captions, each targeting a different semantic clue. Once these captions are obtained, we encode them into CLUE-wise feature vectors using a frozen CLIP text encoder $f_{\text{text}}(\cdot)$. These

textual embeddings serve as the clue-specific representations of each segment, which will later be used for both prediction and explanation. In parallel, we extract a visual embedding $\mathbf{v}^{(t)}$ for each segment using a frozen CLIP visual encoder $f_{\text{visual}}(\cdot)$.

$$\mathbf{c}_r^{(t)} = f_{\text{text}}(D_r^{(t)}), \quad \mathbf{v}^{(t)} = f_{\text{visual}}(I^{(t)}), \quad \text{where } t \in \{1, 2, ..., T\} \tag{1}$$

Here, $\mathbf{c}_r^{(t)} \in \mathbb{R}^d$ denotes the text embedding corresponding to clue $r \in \{act, env, obj\}$ for t-th segment. $D_r^{(t)}$ is the text description generated for that CLUE, and $I^{(t)}$ is the image (keyframe) corresponding to each segment. This process ensures each segment is represented by clue-aware embeddings that capture complementary semantic context.

4.2 Detective Module

The contribution of each semantic clue to anomaly recognition varies across event categories. For instance, in the category "Arrest", action-related clue such as "tackle" or "struggle" tend to be more indicative than environmental context. In contrast, for "Explosion" events, environmental elements like "fire" or "smoke" often play a more critical clue than the associated actions or objects. To accommodate such category-specific variability, we design our model to learn the relative importance of each clue in a category-aware manner. This category-aware clue learning choice enables more adaptive and context-sensitive anomaly detection.

Concretely, for each video category $c \in \mathcal{C}$, we define a category-aware Learnable weights $w_{Act}^{(c)}, w_{Env}^{(c)}, w_{Obj}^{(c)}$, where each element corresponds to the importance score of Action, Environment, and Object. These weights are passed through a softmax function to ensure that they form a valid distribution, enabling stable training and interpretability. To incorporate the relative importance of each clue for a given category c, we apply Category-aware CLUE Learning to $w_r^{(c)}$. These weights are used to modulate the clue embeddings and their concatenation through a linear transformation:

$$\mathbf{E}_r^{(c)} = \mathbf{W}_{\text{proj}}^{\top} \cdot \left(w_r^{(c)} \odot \mathbf{c}_r^{(t)} \right) \parallel \mathbf{v}^{(t)} \tag{2}$$

where $\odot$ denotes the clue-wise scaling and concatenation, $\parallel$ indicates vector concatenation. $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d \times 3d}$ is a projection matrix applied through an MLP. The final CLUE-wise feature is then obtained by concatenating with the visual embedding $\mathbf{v}^{(t)}$, forming a unified multimodal representation for each segment. This fused feature $\mathbf{E}_r^{(c)}$ captures the segment content in a category-sensitive manner, emphasizing the most relevant semantic clues for each category. As the model is trained under the MIL framework, the resulting clue weights $w_r^{(c)}$ modulate the relative importance of each semantic clue type. These weights are continuously refined through backpropagation and serve as the foundation for our **Category-Aware CLUE Learning** mechanism, where the model learns to emphasize clues that are most relevant to each category.

We maintain a Case Log Embedding that stores updated clue weights per category, indexed by video category c and clue type r for precise retrieval and interpretability. While ground-truth categories are available during training, inference requires estimating the video category to retrieve the corresponding category-aware weights. Following

this design, we introduce a lightweight *category classifier* trained on CLUE features. From the Witness module, we obtain a frozen visual embedding $\mathbf{v}^{(t)}$ along with three frozen role embeddings, $\mathbf{c}_{\text{Act}}^{(t)}$, $\mathbf{c}_{\text{Env}}^{(t)}$, and $\mathbf{c}_{\text{Obj}}^{(t)}$, corresponding to Action, Environment, and Object, respectively. We concatenate them to form a snippet token:

$$\mathbf{x}^{(t)} = \left[\mathbf{v}^{(t)} \| \, \mathbf{c}_{\text{Act}}^{(t)} \| \, \mathbf{c}_{\text{Env}}^{(t)} \| \, \mathbf{c}_{\text{Obj}}^{(t)}\right] \tag{3}$$

To improve robustness on long videos and focus on informative snippets, we adopt an Anomaly-guided snippet selection. A frozen selector $\psi(\cdot)$ assigns each snippet a saliency score $a_t = \psi(\mathbf{v}^{(t)})$. We construct a top-$M$ candidate set $\mathcal{T}_M = \mathrm{TopM}(a_t)$ and uniformly sample K snippets, $\mathcal{I} \sim \mathrm{Unif}(\mathcal{T}_M, K)$. To align the selected snippet tokens, we project them through an MLP fusion block:

$$\mathbf{h}_i = g_{\text{fuse}}\left(\mathbf{x}^{(i)}\right), \quad i \in \mathcal{I}. \tag{4}$$

We apply a Transformer encoder to model interactions among the selected snippets, $\tilde{\mathbf{H}} = \mathrm{TransformerEnc}(\mathbf{H})$ with $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$. We then aggregate snippet tokens into a video-level representation via learnable attention pooling:

$$\alpha_i = \frac{\exp(\mathbf{u}^\top \tanh(\mathbf{W}\tilde{\mathbf{h}}_i))}{\sum_{j=1}^{K} \exp(\mathbf{u}^\top \tanh(\mathbf{W}\tilde{\mathbf{h}}_j))}, \quad \mathbf{z} = \sum_{i=1}^{K} \alpha_i \, \tilde{\mathbf{h}}_i. \tag{5}$$

A cosine classifier with a learnable scale τ is adopted, where $\mathbf{w}_c$ denotes the classifier weight for category c:

$$\ell_c = \tau \cdot \frac{\mathbf{z}}{\|\mathbf{z}\|_2}^\top \frac{\mathbf{w}_c}{\|\mathbf{w}_c\|_2}, \quad p(c \mid V) = \mathrm{softmax}(\boldsymbol{\ell})_c, \tag{6}$$

Finally, we utilize the predicted category $\hat{c} = \arg\max_c p(c \mid V)$, to retrieve category-specific clue weights and enable consistent Category-Aware CLUE Learning. To obtain the anomaly score $\hat{s}_t$, we pass this feature through an MLP-based scoring function $\phi(\cdot)$, composed of two fully connected layers with ReLU and Sigmoid activations. This mechanism allows the model to dynamically select category-aware clue weights at test time and ensures that the learned fusion strategy remains consistent across training and inference. Moreover, the case log supports interpretability and enables post-hoc analysis by preserving the learned clue-specific importance for each category.

4.3 Reporter Module

While Category-Aware CLUE Learning enables accurate anomaly prediction, it does not offer insights into why a particular segment is considered anomalous. To achieve interpretability at the segment level, we introduce an attention mechanism that dynamically weights the importance of each semantic clue for every temporal segment. This allows the model to adapt its focus based on local variations in clue saliency and generate explanation-aware scores. Given a t-th segment, we compute the attention score $a_r^{(t)}$ for each clue, by applying a learnable linear projection followed by a ReLU activation. The output is then projected by a clue-specific query vector $\mathbf{q}_r \in \mathbb{R}^d$, as follows:

$$\alpha_r^{(t)} = \mathbf{q}_r^\top \cdot \mathrm{ReLU}(a_r \mathbf{x}_r^{(t)} + \mathbf{b}_r) \tag{7}$$

Here, $\mathbf{x}_r^{(t)} \in \mathbb{R}^d$ is the clue-specific feature of t-th segment, and $a_r \in \mathbb{R}^{d\times d}$ and $b_r \in \mathbb{R}^d$ are learnable parameters for each clue. We then normalize the attention scores across clues using the softmax function. These attention weights $\alpha_r^{(t)}$ indicate the relative importance of each clue at t-th segment. The final fused feature is computed as:

$$z_t^{(\text{attn})} = \sum_{r\in\{Act,Env,Obj\}} \alpha_r^{(t)} \cdot \mathbf{x}_r^{(t)} \tag{8}$$

The fused feature $z_t^{(\text{attn})}$ is then passed into the anomaly scoring network $\phi(\cdot)$, yielding a raw anomaly activation score $\hat{s}_t$. To enhance the visibility of temporal outliers, we apply z-score normalization over the scores of all T segments in a video. Specifically, we compute the mean μ and the standard deviation σ. The final normalized anomaly score is then calculated as $s_t^{(\text{z})} = \frac{\hat{s}_t - \mu}{\sigma}$, which reflects the degree of abnormality for t-th segment relative to the overall context of the video. This score $s_t^{(\text{z})}$ reflects the degree of abnormality for t-th segment relative to the overall video context. Unlike the scalar-based anomaly score used for model training, this Segment-level Explainable CLUE is employed for post-hoc explanation reasoning and visualization. It facilitates the identification of the most influential clues and enables reasoning for real-world applications.

Finally, to condition the language model with semantic cues, we transform the fused clue representation into a CLUE token $\mathbf{E}_r^{(\text{c})}$, which serves as a prefix embedding for the Multi-modal large language model(MLLM). The resulting token is projected into the embedding space of the LLM and injected into the input sequence to guide language generation. To efficiently adapt the pretrained LLM without updating all parameters, we adopt Low-Rank Adaptation (LoRA [10]) for parameter-efficient fine-tuning.

4.4 Loss Function and Regularization

Category Classification Loss. The category classifier is optimized with the focal loss to mitigate class imbalance. where p_y denotes the predicted probability for the ground-truth category y, and γ controls the strength of focusing on hard examples.

$$\mathcal{L}_{\text{cat}} = -(1-p_y)^\gamma \log(p_y), \tag{9}$$

MIL Ranking Loss. To train Category-Aware CLUE Learning and Segment-level Explainable CLUE under weak supervision, we adopt a top-k Multiple Instance Learning (MIL) ranking loss. This formulation assumes that each video is a bag of segments, where a positive (abnormal) video should contain at least k highly abnormal segments, and a negative (normal) video should contain only low-scoring segments. Let P and N denote the sets of segments in anomalous and normal videos, respectively. Using the predicted anomaly score $\hat{s}_t$, we compute the average of the top-k segment scores for each bag. The top-k MIL loss is then defined as:

$$\mathcal{L}_{\text{MIL}} = \max\left(0,\ 1 - \frac{1}{k}\sum_{t\in\text{Top-}k(P)} \hat{s}_t\ + \frac{1}{k}\sum_{t\in\text{Top-}k(N)} \hat{s}_t\right) \tag{10}$$

This margin-based objective encourages the model to assign higher anomaly scores to the top-k segments in abnormal videos than to the top-k segments in normal ones. Notably, only the scalar fusion scores are used during training, while the attention-based z-score branch is employed solely for post-hoc explanation and visualization.

Regularization Terms. To improve training stability and enhance interpretability of the attention-based branch, we apply two regularization losses on the clue attention weights $\alpha_r^{(t)}$. To avoid overconfidence in clue selection and encourage diversity, we minimize the negative entropy of the attention distribution called Attention Entropy Loss $\mathcal{L}_{\text{AE}}$:

$$\mathcal{L}_{\text{AE}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{r\in\{Act,Env,Obj\}} \alpha_r^{(t)} \log \alpha_r^{(t)} \tag{11}$$

Also, to enforce consistency across neighboring segments and reduce abrupt changes in attention, we apply a Temporal Smoothness Loss $\mathcal{L}_{\text{smooth}}$:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{r\in\{Act,Env,Obj\}} \left\|\alpha_r^{(t)} - \alpha_r^{(t+1)}\right\|^2 \tag{12}$$

The complete training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MIL}} + \lambda_{\text{AE}}\mathcal{L}_{\text{AE}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}$ where λ_{AE} and λ_{smooth} are regularization coefficients. These regularizers encourage the model to learn smoother and more interpretable attention distributions.

5 Experiment

Comparisons with VAD methods. Table 1 compares CLUE-VAD with state-of-the-art methods on UCF-Crime (AUC) and XD-Violence (AP). Our CLUE-based model achieves 89.47 / 88.24, outperforming strong CLIP-based baselines such as TSA (87.58 / 82.17) and UMIL (86.75 / –), highlighting the effectiveness of clue-disentangled supervision and category-aware fusion. Among explainable VAD approaches, CLUE-VAD establishes state-of-the-art results including Ex-VAD (88.29 / 86.52) and Holmes-VAU (88.96 / 87.68) on both datasets, while remaining competitive with the best multi-modality model π-VAD (90.33 / 85.37). These gains stem from (i) category-conditioned fusion over Action/Environment/Object clues, (ii) clue-level disentangled supervision that strengthens text–video

Model	Venue	UCFC AUC	XD AP
SoTA with single-modality			
MIL [27]	CVPR'18	75.41	75.68
RTFM [30]	ICCV'21	84.30	77.81
CLAV [6]	CVPR'23	86.10	-
UMIL [18]	CVPR'23	86.75	-
UR-DMU [42]	AAAI'23	86.97	81.77
TSA [13]	ICIP'23	**87.58**	**82.17**
SoTA with Multi-modality			
PEMIL [3]	CVPR'24	86.83	88.21
TPWNG [36]	CVPR'24	87.79	83.68
VadCLIP [35]	AAAI'24	88.02	84.15
LEC-VAD [33]	ICML'25	89.97	**88.47**
π-VAD [23]	CVPR'25	**90.33**	85.37
SoTA with Explainable VAD			
LAVAD [39]	CVPR'24	80.28	62.01
VERA [37]	CVPR'25	86.6	88.2
VADOr [17]	arXiv'24	88.1	-
Ex-VAD [11]	ICML'25	88.29	86.52
Holmes-VAU [41]	CVPR'25	88.96	87.68
CLUE-VAD(Ours)	-	**89.47**	**88.24**

Table 1: Comparison with state of the art VAD methods on UCF-Crime (AUC %) and XD-Violence (AP %).

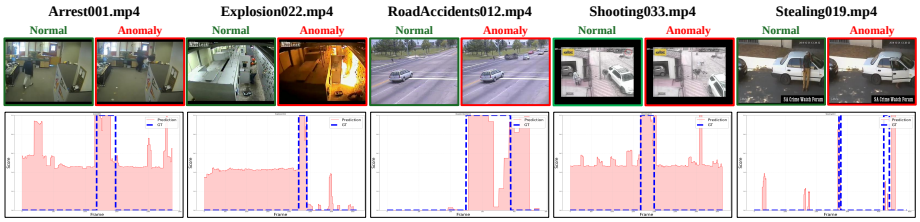

Fig. 3: Frame-level anomaly prediction results over time.

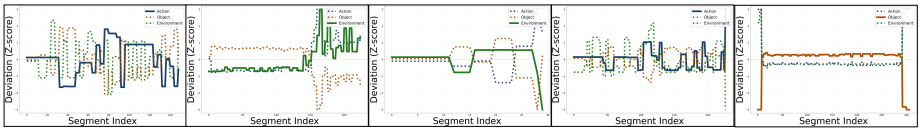

Fig. 4: Segment-level interpretability achieved through CLUE-wise score decomposition, highlighting which semantic clue contributes most to each anomaly.

alignment, and (iii) streaming Top-K selection that stabilizes category estimation and focuses learning on informative segments. Overall, CLUE-VAD improves anomaly detection accuracy while maintaining interpretable clue-wise reasoning, supporting practical X-VAD.

Anomaly score graph. To qualitatively assess behavior, Figure 3 visualizes frame-level anomaly scores for five UCF-Crime category. Each column is one video with normal and anomalous segments, the top row shows sampled frames, the bottom row the predicted scores over time. Red bars indicate predicted anomaly intensity, and blue dashed lines mark ground-truth intervals. Scores rise sharply during anomalous events and remain low otherwise, even for ambiguous cases like *Stealing*, the highlighted window closely matches the ground truth. These examples demonstrate temporally precise, interpretable predictions with few false positives, supporting real-world applicability. For more qualitative plots, please see Appendix.

Segment-level Explainable CLUE score. Figure 4 plots segment-level z-score deviations for the three clues across five categories examples. In *Explosion022* and *RoadAccidents012*, *Environment* spikes near anomaly boundaries, reflecting abrupt scene changes. For *Arrest001* and *Shooting033*, sharp *Action*/*Object* deviations correspond to rapid human motion and interactions. In *Stealing019*, a sustained rise in *Object* deviation marks the prolonged presence of suspicious items. These role-specific patterns provide interpretable anomaly evidence and align well with ground-truth windows, indicating temporally precise role-wise attention. They clarify not only when anomalies occur but also which semantic factors drive them.

Impacts of Category-aware CLUEs. Figure 5 presents the category-specific CLUE weights learned by the category-aware fusion module on UCF-Crime. *Fighting*, *Assault*, and *Abuse* assign higher weights to Action clues, reflecting their motion-centric

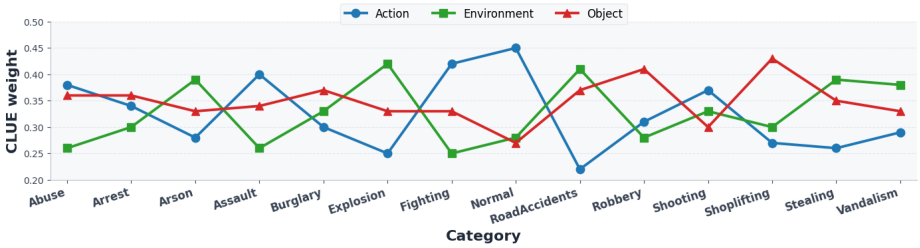

Fig. 5: Category-wise distribution of clue importance, showing the relative contribution of *Action*, *Environment*, and *Object* across different anomaly types.

nature. In contrast, object-centric events—*Robbery*, *Shoplifting*, *Stealing*—emphasize Object clues, while *Explosion* and *RoadAccidents* rely more on Environment context. Notably, *Normal* also shows strong Action weighting, likely due to subtle routine movements. These patterns confirm that CLUE-VAD adaptively focuses on the most relevant semantic clues per category, improving both performance and interpretability.

Effectiveness of CLUE Features. Table 2 shows that each clue adds a complementary signal. When used alone, they capture human motion, contextual background, or interacting objects. When combined, performance improves, with AUC rising from single to dual and then triple configurations. This validates decomposing video semantics into structured roles and dynamically fusing them, enabling CLUE-VAD to deliver accurate and interpretable predictions in real-world deployments.

Combinations	CLUEs			AUC(%)
	Action	Environment	Object	
Single	✓			85.82
		✓		85.73
			✓	85.41
Dual	✓	✓		86.67
		✓	✓	86.45
	✓		✓	86.89
Triple	✓	✓	✓	88.34

Table 2: Effectiveness of each CLUE feature combinations in UCFC dataset.

Method	**Params**	**BLEU-1** [24]	**BLEU-2**	**BLEU-3**	**BLEU-4**	**CIDEr** [31]	**METEOR** [2]	**ROUGE** [16]
Video-ChatGPT [22]	7B	0.1639	0.0848	0.0400	0.0199	0.0002	0.1538	0.1870
Video-LLaMA [40]	7B	0.1824	0.1521	0.0558	0.0251	0.0368	0.1365	0.1927
Video-LLaVA [15]	7B	0.2573	0.1277	0.0613	0.0310	0.0393	0.1398	0.2085
QwenVL2 [32]	7B	0.1934	0.0856	0.0385	0.0183	0.0027	0.1442	0.1868
InternVL2 [5]	8B	0.2687	0.1331	0.0653	0.0339	**0.0713**	0.1424	0.2158
CLUE-VAD (Ours)	2B	**0.2982**	**0.1873**	**0.1175**	**0.0739**	0.0423	**0.1729**	**0.2762**

Table 3: Comparison of reasoning performance with state-of-the-art Multimodal Large Language Models (MLLMs) on anomaly explanation generation.

Comparison of reasoning performance. Table 3 compares the reasoning performance of CLUE-VAD with several state-of-the-art multimodal large language models (MLLMs). For fair evaluation, we use InternVL2-2B as the base model and assess generated explanations on the UCFA test annotations [38]. Despite its smaller model size, CLUE-VAD

achieves the best performance on most metrics, including BLEU-1/2/3/4, METEOR, and ROUGE. These results indicate that grounding the language model with structured CLUE representations provides stronger reasoning signals than purely generative MLLM pipelines, leading to more reliable and context-aware anomaly explanations. By conditioning explanations on Action, Environment, and Object clues, CLUE-VAD produces more accurate and semantically consistent anomaly descriptions.

CLUE Interaction Method	UCF-Crime (AUC%)
Attention-based fusion	86.89
Graph-based fusion	86.13
Pairwise inter-CLUE fusion	87.46
Category-aware CLUE weighting (ours)	**88.34**

Table 4: CLUE interaction mechanisms in the Detective module.

Modality	UCF-Crime AUC(%)	XD-Violence AP(%)
Text-only (CLUE)	88.34	87.77
Visual-only	87.46	86.35
Multimodal	89.47	88.24

Table 5: Performance under different modalities (Text and Visual).

Detective Module Interaction Mechanisms. Table 4 compares several CLUE interaction strategies within the Detective module under the same text-only UCF-Crime setting. Replacing our weighting MLP with a single-layer self-attention mechanism yielded only 86.89 AUC, indicating that simply treating the three CLUEs as tokens does not effectively capture their distinct roles. A graph-based variant with one-step message passing improved performance to 86.13 AUC but added parameters without meaningful gain. Pairwise CLUE interactions (Act–Env, Env–Obj, Act–Obj) performed better, reaching up to 87.46 AUC, suggesting that certain CLUE pairs provide useful relational cues. Still, the proposed category-aware CLUE weighting achieved the best result at 88.34 AUC by directly modeling category-specific importance across all three CLUEs. Overall, these results show that while alternative interaction designs are feasible, category-aware weighting offers the most robust and parameter-efficient fusion strategy.

Modality Analysis. Table 5 shows that text-only and visual-only settings achieve comparable performance, while combining both modalities yields clear gains on UCF-Crime and XD-Violence. Textual CLUEs provide high-level semantic structure (who, what, where), whereas visual features contribute fine-grained appearance and motion cues, making their roles complementary rather than redundant. Overall, these results support our design choice of a multimodal CLUE-VAD, where the synergy between text and vision consistently translates into improved anomaly detection performance.

Visualization of CLUE-VAD features. Figure 6 visualizes the CLUE-VAD feature embeddings using t-SNE, showing how each semantic clue (Action, Environment, and Object) contributes to the overall anomaly detection process. The visualization clearly distinguishes

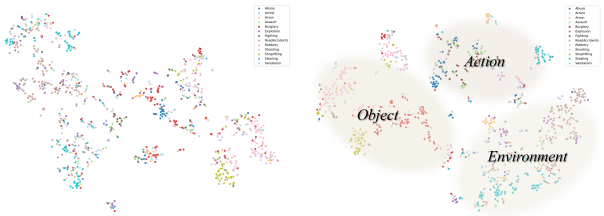

Fig. 6: t-SNE visualization of CLUE.

Fig. 7: Qualitative results of MLLM-generated explanations for Explainable-VAD.

between different types of anomalies, with each clue providing complementary information that enhances the interpretability of the model. By examining the distribution of feature embeddings, we can observe how CLUE-VAD captures the semantic distinctions of abnormal events, demonstrating its ability to separate and identify context-specific anomalies in a human-interpretable manner. This enables more transparent reasoning for anomaly detection tasks.

Qualitative results of MLLM-generated explanations for Explainable-VAD. Figure 7 presents qualitative examples of MLLM-generated explanations on representative anomaly videos. Compared with general-purpose MLLMs, which sometimes describe only the normal scene context or miss the abnormal behavior, CLUE-VAD produces explanations that are more directly grounded in the detected anomaly. The generated descriptions capture not only the event category, such as fighting, explosion, or robbery, but also the semantic factors that support the prediction. By explicitly reporting Action, Environment, and Object clues with their corresponding contributions, CLUE-VAD provides interpretable evidence behind each anomaly decision. These examples show that CLUE-based grounding enables the MLLM to move beyond generic video captioning toward more faithful and explanation-oriented anomaly reasoning.

Efficiency of the CLUE-VAD. Table 6 shows that offline caption prompting can be precomputed and is excluded from online detection. Since CLUE captions are generated once and cached before inference, the online pipeline does not require repeatedly querying the captioning model for every video segment. The detection-side modules, including CLUE

	Step / Method	Time [ms]	Mem [GB]
Offline	Caption prompting (VLM)	59.57	20.38
Online	CLUE encoding	0.64	4.35
	Category-Aware CLUE Learning	0.01	0.42
	Segment-level Explainable CLUE	0.01	0.53
	MLLM-grounded explanation	542	8.79
	CLUE-VAD online total	**542.66**	**14.09**

Table 6: Inference efficiency of CLUE-VAD.

encoding, Category-Aware CLUE Learning, and Segment-level Explainable CLUE, require only 0.66 ms and 5.3 GB in total, indicating that the proposed clue-based scoring mechanism introduces negligible computational overhead. Although the MLLM-grounded explanation step is relatively more expensive, it is used optionally for selected high-anomaly segments rather than every segment. This design allows CLUE-VAD to perform efficient anomaly scoring during continuous monitoring while invoking detailed natural-language reasoning only when explanation is needed. Overall, the modular separation between fast detection and optional explanation makes CLUE-VAD practical for explainable video anomaly detection without sacrificing inference efficiency.

6 Conclusion

We introduced **CLUE-VAD**, a structured semantic decomposition framework for explainable WSVAD. By decomposing each video segment into *Action*, *Environment*, and *Object* clues, CLUE-VAD complements holistic anomaly prediction with clue-level attribution and category-aware semantic reasoning. Rather than simply predicting when an anomaly occurs, our framework further identifies which semantic factors contribute to the anomaly decision, enabling more transparent anomaly understanding. Through its modular design, the framework achieves strong detection performance while providing explanations grounded in clue-specific scores and keyword-level evidence. The experimental results show that structured CLUEs provide complementary information for anomaly detection and serve as effective evidence for explanation generation.

We also acknowledge several limitations: the quality of CLUEs depends on the upstream captioning model, captioning errors may propagate to downstream reasoning, and the current segment-level formulation may not fully separate multiple simultaneous activities or region-specific anomalies. Moreover, complex surveillance scenes often involve subtle interactions among actions, objects, and environments, requiring more fine-grained spatio-temporal reasoning. Addressing captioner robustness, reducing error propagation, and extending CLUEs to region-wise and multi-activity reasoning are important directions for future work. We believe CLUE-VAD provides a meaningful step toward accurate, interpretable, and trustworthy video anomaly detection.

Acknowledgements

This work was supported by Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (No. RS-2024-00442120, Development of AI technology capable of robustly recognizing abnormal and dangerous situations and behaviors during night and bad weather conditions), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00456589) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25431624).

References

1. Appiani, A., Beyan, C.: Clip-vad: Exploiting vision-language models for voice activity detection. arXiv preprint arXiv:2410.14509 (2024)

2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. Chen, J., Li, L., Su, L., Zha, Z.j., Huang, Q.: Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18319–18329 (2024)
4. Chen, W., Ma, K.T., Yew, Z.J., Hur, M., Khoo, D.A.A.: Tevad: Improved video anomaly detection with captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5549–5559 (2023)
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
6. 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 IEEE/CVF conference on computer vision and pattern recognition. pp. 12137–12146 (2023)
7. Cho, M., Kim, T., Shim, M., Wee, D., Lee, S.: Towards multi-domain learning for generalizable video anomaly detection. Advances in Neural Information Processing Systems **37**, 50256–50284 (2024)
8. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv–2407 (2024)
9. Feng, J.C., Hong, F.T., Zheng, W.S.: Mist: Multiple instance self-training framework for video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14009–14018 (2021)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
11. Huang, C., Shi, Y., Wen, J., Wang, W., Xu, Y., Cao, X.: Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In: Forty-second International Conference on Machine Learning
12. Huang, C., Shi, Y., Wen, J., Wang, W., Xu, Y., Cao, X.: Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In: Forty-second International Conference on Machine Learning (2025)
13. Joo, H.K., Vo, K., Yamazaki, K., Le, N.: Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3230–3234. IEEE (2023)
14. Li, F., Liu, W., Chen, J., Zhang, R., Wang, Y., Zhong, X., Wang, Z.: Anomize: Better open vocabulary video anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29203–29212 (2025)
15. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
16. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
17. Lv, H., Sun, Q.: Video anomaly detection and explanation via large language models. arXiv preprint arXiv:2401.05702 (2024)
18. 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 IEEE/CVF conference on computer vision and pattern recognition. pp. 8022–8031 (2023)
19. Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., Yang, J.: Localizing anomalies from weakly-labeled videos. IEEE transactions on image processing **30**, 4505–4515 (2021)

20. Ma, J., Wang, J., Luo, J., Yu, P., Zhou, G.: Sherlock: Towards multi-scene video abnormal event extraction and localization via a global-local spatial-sensitive llm. In: Proceedings of the ACM on Web Conference 2025. pp. 4004–4013 (2025)
21. Ma, J., Wang, J., Zhang, M., Zhou, G.: Skynet-v1: Towards early warning of video abnormal events via a spatial-temporal causal-enhanced moe framework. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 6586–6595 (2025)
22. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
23. Majhi, S., D'Amicantonio, G., Dantcheva, A., Kong, Q., Garattoni, L., Francesca, G., Bondarev, E., Brémond, F.: Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24265–24274 (2025)
24. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
26. Singh, A., Jones, M.J., Learned-Miller, E.G.: Eval: Explainable video anomaly localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18717–18726 (2023)
27. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)
28. Sun, S., Hua, J., Feng, J., Wei, D., Lai, B., Gong, X.: Tdsd: Text-driven scene-decoupled weakly supervised video anomaly detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 5055–5064 (2024)
29. Tang, J., Lu, H., Wu, R., Xu, X., Ma, K., Fang, C., Guo, B., Lu, J., Chen, Q., Chen, Y.: Hawk: Learning to understand open-world video anomalies. Advances in Neural Information Processing Systems **37**, 139751–139785 (2024)
30. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4975–4986 (2021)
31. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
32. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
33. Wang, Y., Chen, S.: Learning event completeness for weakly supervised video anomaly detection. arXiv preprint arXiv:2506.13095 (2025)
34. Wu, P., Zhou, X., Pang, G., Yang, Z., Yan, Q., Wang, P., Zhang, Y.: Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 9301–9310 (2024)
35. Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., Zhang, Y.: Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6074–6082 (2024)

36. Yang, Z., Liu, J., Wu, P.: Text prompt with normality guidance for weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18899–18908 (2024)
37. Ye, M., Liu, W., He, P.: Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8679–8688 (2025)
38. Yuan, T., Zhang, X., Liu, K., Liu, B., Chen, C., Jin, J., Jiao, Z.: Towards surveillance video-and-language understanding: New dataset, baselines, and challenges (2023)
39. Zanella, L., Menapace, W., Mancini, M., Wang, Y., Ricci, E.: Harnessing large language models for training-free video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18527–18536 (2024)
40. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations. pp. 543–553 (2023)
41. Zhang, H., Xu, X., Wang, X., Zuo, J., Huang, X., Gao, C., Zhang, S., Yu, L., Sang, N.: Holmes-vau: Towards long-term video anomaly understanding at any granularity. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13843–13853 (2025)
42. Zhou, H., Yu, J., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3769–3777 (2023)