Learning to Tell Apart: Weakly Supervised Video Anomaly Detection via Disentangled Semantic Alignment

Wenti Yin[1], Huaxin Zhang[1], Xiang Wang[1], Yuqing Lu[1], Yicheng Zhang[1], Bingquan Gong[1], Jialong Zuo[1], Li Yu[2], Changxin Gao[1], Nong Sang[1]*

[1]Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

[2]School of Electronic Information and Communications, Huazhong University of Science and Technology

{yinwt, nsang}@hust.edu.cn

Abstract

Recent advancements in weakly-supervised video anomaly detection have achieved remarkable performance by applying the multiple instance learning paradigm based on multimodal foundation models such as CLIP to highlight anomalous instances and classify categories. However, their objectives may tend to detect the most salient response segments, while neglecting to mine diverse normal patterns separated from anomalies, and are prone to category confusion due to similar appearance, leading to unsatisfactory fine-grained classification results. Therefore, we propose a novel Disentangled Semantic Alignment Network (DSANet) to explicitly separate abnormal and normal features from coarse-grained and fine-grained aspects, enhancing the distinguishability. Specifically, at the coarse-grained level, we introduce a self-guided normality modeling branch that reconstructs input video features under the guidance of learned normal prototypes, encouraging the model to exploit normality cues inherent in the video, thereby improving the temporal separation of normal patterns and anomalous events. At the fine-grained level, we present a decoupled contrastive semantic alignment mechanism, which first temporally decomposes each video into event-centric and background-centric components using frame-level anomaly scores and then applies visual-language contrastive learning to enhance class-discriminative representations. Comprehensive experiments on two standard benchmarks, namely XD-Violence and UCF-Crime, demonstrate that DSANet outperforms existing state-of-the-art methods.

Code — https://github.com/lessiYin/DSANet

1 Introduction

Weakly Supervised Video Anomaly Detection (WS-VAD) (Sultani, Chen, and Shah 2018) aims to temporally detect anomaly segments in a long untrimmed video with only video-level labels (*i.e.*, indicating whether a video contains an anomaly), drastically reducing annotation costs compared to its fully supervised counterparts (Wu et al. 2024a; Abdalla et al. 2024; Nayak, Pati, and Das 2021), and has received considerable attention in recent years (Wang et al. 2021, 2022; Shi et al. 2025; Wang et al. 2025a; Zhu et al. 2024; Liang et al. 2023). The predominant approach in WS-VAD is built upon the multiple instance learning (MIL) framework (Tian et al. 2021; Lv et al. 2023; Chen et al. 2024). The general pipeline involves first extracting deep features for each video using a pre-trained backbone like I3D (Carreira and Zisserman 2017) and CLIP (Radford et al. 2021), and then feeding the obtained features to a binary classifier to generate instance-level anomaly scores (Yu et al. 2025; Wang et al. 2025b,c). For example, CLIP-TSA (Joo et al. 2023) uses CLIP's visual encoder with multi-scale temporal aggregation and a multiple instance learning branch for detection. VadCLIP (Wu et al. 2024b) employs a binary classifier for anomaly detection and text prompts for anomaly types identification. PEMIL (Pu et al. 2024) designs anomaly- and context-aware prompts to model complex event boundaries. ITC (Liu, Lam, and Bao 2024) introduces learnable textual cues in a dual-branch framework for robust cross-modal anomaly recognition.

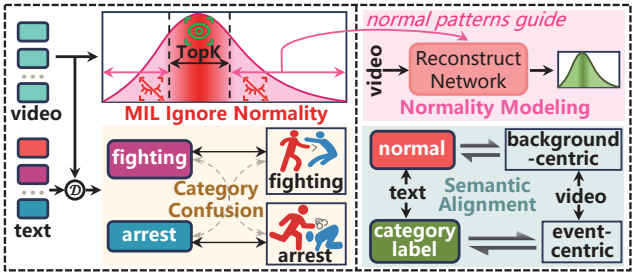

Figure 1: Schematic diagram about motivation. We identify two main issues: *1) limited understanding of normality*, and *2) category confusion*. We address them through normality modeling and decoupled contrastive semantic alignment.

Despite its recent success, the prevailing WS-VAD approaches based on multiple instance learning still suffer from two fundamental limitations. At the coarse-grained level, the discriminative nature of MIL results in an incomplete understanding of normality. By focusing exclusively on identifying the most salient anomalous segments, such models fail to construct a robust and explicit representation of the diverse normal patterns present within a video. This deficiency compromises the model's ability to distinguish between true anomalies and complex yet benign events,

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

leading to ambiguous decision boundaries and an increased false positive rate. At the fine-grained level, anomaly classification remains challenging due to the possible visual similarity among different abnormal events and between anomalies and normal contexts. In the absence of frame-level supervision, models often confuse co-occurring background patterns with true anomalies. The learned representations for different anomaly categories often become entangled in the embedding space, leading to semantic confusion and reduced inter-class separability. This confusion ultimately limits the model's ability to perform accurate anomaly categorization and discriminative localization.

To address these challenges, we propose DSANet that enhances WS-VAD through synergistic normality modeling and disentangled contrastive semantic alignment, as shown in Figure 1. To resolve the lack of explicit normality understanding, we introduce a Self-Guided Normality Modeling (SG-NM) branch. Inspired by the observation that even within anomalous sequences, local regions still exhibit intrinsic normality (Luo et al. 2025) (*e.g*. normal backgrounds with visual consistency), SG-NM dynamically mines a compact set of normal prototypes directly from each input video, without relying on an external memory bank (Zhou, Yu, and Yang 2023). These prototypes supervise a reconstruction objective that drives the model to learn video-specific normal characteristics in a generative manner. This design enhances the discriminative detector with an internal understanding of normality, enabling the model to better distinguish normal behavior and more accurately identify anomalies. The SG-NM branch is self-contained, memory-free, and requires no dataset-level priors, making it scalable and data-efficient. To mitigate category confusion caused by visual similarities between different anomaly classes and normal contexts, we design a Decoupled Contrastive Semantic Alignment (DCSA) mechanism. Guided by initial anomaly predictions, DCSA explicitly disentangles video features into an event-centric and a background-centric prototype. A dual contrastive objective then aligns the event prototype with its corresponding anomaly class, while consistently aligning the background prototype with a universal "normal" class. This disentanglement yields more discriminative representations, enhancing the model's ability for fine-grained anomaly classification and precise temporal localization.

In summary, our main contributions are as follows:

- We introduce **Self-Guided Normality Modeling**, a generative reconstruction module that dynamically mines video-specific normal patterns without external memory. By incorporating normality cues overlooked by MIL-based paradigms, it facilitates more accurate and complete coarse-grained anomaly detection.
- We propose **Decoupled Contrastive Semantic Alignment**, a mechanism that explicitly separates event- and background-centric prototypes and aligns them with respective semantic targets, addressing semantic confusion and improving fine-grained anomaly classification.
- Extensive experiments verify the effectiveness of the proposed **DSANet**, which outperforms existing state-of-the-art methods on two standard WS-VAD benchmarks.

2 Related Work

2.1 Vision-Language Pre-training

Vision-Language Pre-training (VLP) has become a dominant paradigm for learning joint representations from large-scale image-text data, enabling remarkable zero-shot transfer capabilities (Ho et al. 2025). Pioneering works such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) use dual-encoder architectures trained with contrastive loss to align visual and textual embeddings. Recent trends aim to improve capability and efficiency. One line adopts unified encoder-decoder frameworks that jointly handle understanding and generation tasks, exemplified by BLIP (Li et al. 2022), which introduced a data bootstrapping method to denoise web captions, and CoCa (Yu et al. 2022), which combines contrastive and captioning losses. Another focuses on parameter efficiency by leveraging powerful, frozen unimodal models. Flamingo (Alayrac et al. 2022) bridges frozen vision encoders and large language models with a Perceiver Resampler and trainable gated cross-attention layers, while BLIP-2 (Li et al. 2023) introduced a lightweight Querying Transformer to link frozen components with minimal training cost. VLP models have been widely applied to downstream tasks such as text-video retrieval (Wang et al. 2024; Tian et al. 2024), visual question answering (Zou et al. 2024; Li et al. 2024), and open-vocabulary action recognition (Huang et al. 2024; Jia et al. 2023). In this work, following previous practices (Wu et al. 2024b; Dev, Hazari, and Das 2024), we construct DSANet based on CLIP (Radford et al. 2021) for weakly supervised video anomaly detection.

2.2 Weakly Supervised Video Anomaly Detection

This task was first formalized by Sultani et al. (Sultani, Chen, and Shah 2018) as a multiple instance learning (MIL) problem, introducing a deep ranking framework that enables the model to assign higher anomaly scores to abnormal segments under weak supervision. Subsequent works have expanded and refined this paradigm (Tian et al. 2021; Zanella et al. 2024; Zhang et al. 2025). To enhance temporal modeling, RTFM (Tian et al. 2021) integrates temporal convolution and self-attention to capture multi-scale temporal dependencies. To mitigate contextual bias in MIL, UMIL (Lv et al. 2023) introduces a strategy to learn stable representations across "confident" and "ambiguous" samples, thereby improving classifier robustness. To address the weakness of the supervision signal, Feng et al. (Feng, Hong, and Zheng 2021) propose a two-stage self-training framework that refines discriminative feature representations by using a multiple instance pseudo-label generator to train a self-guided attention encoder. With the rise of vision-language pre-training (VLP) models, WS-VAD has been transitioning from traditional statistical pattern recognition towards semantic-aligned reasoning. Early approaches adopted VLP models (*e.g*., CLIP) as visual feature extractors (Joo et al. 2023). Recent efforts (Wu et al. 2024b; Pu et al. 2024; Liu, Lam, and Bao 2024) incorporate prompt-based or learnable textual cues into the MIL framework to facilitate anomaly type recognition and enhance event-level discrimination.

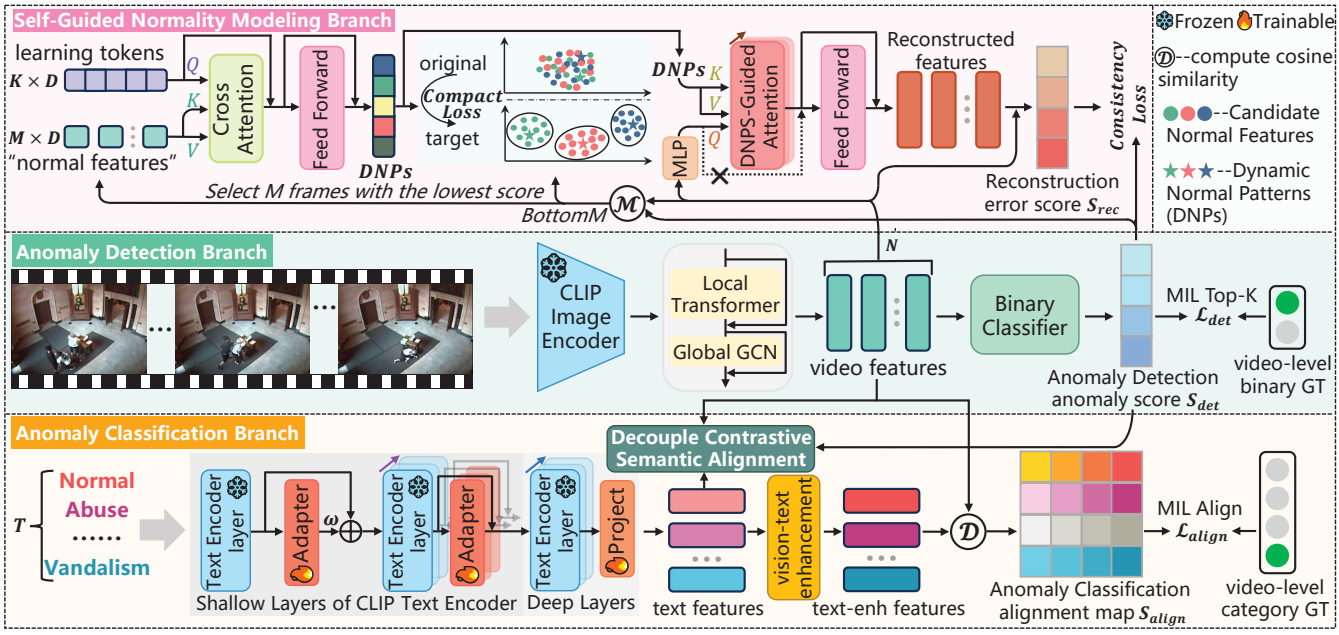

Figure 2: Overview of the proposed DSANet. The model consists of three collaborative branches. The Anomaly Detection Branch produces initial frame-level binary scores using a MIL framework. The Self-Guided Normality Modeling Branch enhances the model's understanding of normal patterns by mining Dynamic Normal Patterns within the video to guide feature reconstruction, improving its ability to distinguish normal from abnormal. The Anomaly Classification Branch aligns video features with textual category embeddings for fine-grained classification, using Lightweight Text Adapters for adaptation and a Decoupled Contrastive Semantic Alignment mechanism to distinguish various anomaly types from normal categories.

3 Methodology

3.1 Overview

Problem Definition. We address the task of WS-VAD, where the input is a video set $\mathcal{V} = \{v_i\}_{i=1}^{n}$. Each video v_i has video-level labels: (i) a binary anomaly indicator $y_i \in \{0, 1\}$ and (ii) an anomaly category label $c_i \in \{0, 1, \ldots, C-1\}$. A video is normal if $y_i = 0$ and $c_i = 0$; otherwise, it contains at least one anomaly and c_i denotes its category. Under weak supervision, two subtasks exist: coarse-grained WS-VAD assigns anomaly scores to frames, while fine-grained WS-VAD identifies categories and localizes anomalies.

Overall Framework. As illustrated in Figure 2, our model has three branches: Anomaly Detection, Self-Guided Normality Modeling (SG-NM), and Anomaly Classification.

Anomaly Detection Branch generates coarse-grained, binary frame-level anomaly scores based on the multiple instance learning (MIL) framework. Following (Wu et al. 2024b), we use a frozen CLIP image encoder to extract frame-wise features $F_{clip} \in \mathbb{R}^{N\times D}$, where N is the number of frames and D is the feature dimension. To augment the CLIP features with temporal information, the frame-wise features are fed into a two-stage temporal modeling module: (1) a local Transformer, which restricts attention to non-overlapping temporal windows to capture short-range dynamics, and (2) a dual-channel GCN for long-range dependencies, whose edges are constructed based on feature cosine similarity and relative temporal distance. The temporally contextualized video representation $F_{video} \in \mathbb{R}^{N\times D}$ is then passed through a binary classifier to produce coarse-grained anomaly scores $S_{det} \in \mathbb{R}^{N\times 1}$. Finally, frame-level anomaly probabilities $p_i = \sigma(s_i)$ are aggregated by averaging top-k scores for video-level prediction $\bar{p}$, with loss:

$$\mathcal{L}_{det} = -\sum_{c\in\{0,1\}} y_c \log \bar{p}^c (1-\bar{p})^{1-c}, \quad \bar{p} = \frac{1}{k}\sum_{i\in\mathcal{I}_{\text{top-}k}} p_i, \quad (1)$$

where $\mathcal{I}_{\text{top-}k}$ denotes the indices of the top-k scoring frames.

SG-NM Branch addresses the deficiency in normality modeling within the Anomaly Detection Branch, whose discriminative MIL objective tends to focus on the most salient anomalous segments while neglecting the diverse and informative normal patterns in the video. Specifically, we mine the video-specific normal prototypes to guide a reconstruction learning process. It computes a reconstruction-based anomaly score S_{rec}, which is aligned with the anomaly detection score S_{det}, forming a self-distillation mechanism that strengthens the model's normal representation.

Anomaly Classification Branch enables fine-grained category-aware anomaly detection through visual-text alignment. In existing works (Wu et al. 2024b), all category labels (*e.g.*, *'Abuse'*) are embedded into the text feature space $T_{text} \in \mathbb{R}^{C\times D}$ using the frozen CLIP text encoder. A frame-category alignment map $S_{align} \in \mathbb{R}^{N\times C}$ is then computed between the category embeddings and video features, and supervised by the category label under the MIL strategy: the top-k values in each column $M_{:,c}$ are averaged to get video-level category scores S_c. The alignment loss $\mathcal{L}_{align}$ is defined as the multi-class cross-entropy between predicted p

from S_c and the one-hot category label vector:

$$\mathcal{L}_{align} = -\sum_{c=0}^{C-1} y_c \log\left(\bar{p}_c\right), \quad \bar{p}_c = \text{softmax}\left(S_c/\tau\right), \tag{2}$$

where $S_c = \frac{1}{k}\sum_{i\in\mathcal{I}^{(c)}_{\text{top-}k}} M_{i,c}$. Following (Wu et al. 2024b), before computing the alignment map, we use a Vision-Text Enhance module that injects video-derived visual cues into the text embedding to enhance event-specific semantics. Technically, frame-level features F_{video} are first aggregated into a global video representation V using anomaly score-guided weighting: $V = \text{Norm}(S_{det}^{\top} F_{video})$, where $\text{Norm}(\cdot)$ denotes L2 normalization. This visual cue V is then fused with T_{text} via a skip-connected feed-forward network to generate enhanced category embeddings $T_{text-enh}$.

To reduce semantic confusion between visually similar anomalies and normal contexts, we introduce a **Decoupled Contrastive Semantic Alignment** mechanism to decouple visual features for category-aware alignment. We also propose **Lightweight Text Adapters** to obtain domain-adaptive text representations. We next detail the proposed modules.

3.2 Self-Guided Normality Modeling

The MIL-based Anomaly Detection Branch focuses on only the most salient video parts (*i.e.*, top-k frames), leading to an incomplete understanding of rich normal patterns and blurring the boundaries between normality and anomaly. In contrast, the SG-NM branch, which operates only during training, explicitly models video-specific normality to guide video feature reconstruction. This reconstruction pathway complements the detection branch by leveraging inherent normal cues in the video, improving the temporal separation between normal and anomalous events.

We first extract representative normal patterns directly from the input video to ensure contextual relevance and adaptability, rather than relying on external memory banks to model dataset-level normality (Zhou, Yu, and Yang 2023). Specifically, we use initial anomaly scores S_{det} from the detection branch to select the frame features with the bottom-M lowest scores, forming the candidate normal feature set $F_n \in \mathbb{R}^{M\times D}$. To extract representative normal patterns from F_n, we apply a single-layer cross-attention module. A set of $K = 16$ learnable queries $Q_{learn} \in \mathbb{R}^{K\times D}$ attends to F_n (used as both key and value), extracting K distilled Dynamic Normal Patterns (DNPs), denoted as $P \in \mathbb{R}^{K\times D}$. To ensure that DNPs purely represent normal features and resist contamination from anomalies within the candidate set, we introduce a Normalcy Concentration Loss $\mathcal{L}_{compact}$. This loss encourages each feature in F_n to be close to at least one DNP. It is a self-supervised objective that forces the DNPs to become compact and representative centers of normal patterns. The loss is defined as the average minimum distance from each feature in F_n to the set of DNPs:

$$\mathcal{L}_{compact} = \frac{1}{M}\sum_{i=1}^{M} \min_{j\in\{1,\dots,K\}} d\big(F_n(i), P(j)\big), \tag{3}$$

where $d(\cdot,\cdot)$ represents the cosine distance, $F_n(i)$ is the i-th feature in the candidate set, and $P(j)$ is the j-th DNP.

Guided by the extracted normal patterns, we then design a multi-layer cross-attention decoder that reconstructs the video feature for anomaly detection. Video features $F_{video} \in \mathbb{R}^{N\times D}$ are first mapped via an MLP and used as the query, while DNPs P serve as key and value. To prevent anomaly leakage, the first attention layer excludes residual connections, ensuring that reconstruction relies solely on normal patterns. Each decoder layer applies cross-attention between transformed video features and the DNPs, followed by a feed-forward network. The final reconstructed features F_{rec} are compared to the original features F_{video}, and a frame-level anomaly score $S_{rec} \in \mathbb{R}^{N\times 1}$ is obtained by computing their cosine distance, normalized to $[0, 1]$. A higher reconstruction error indicates greater deviation from normalcy.

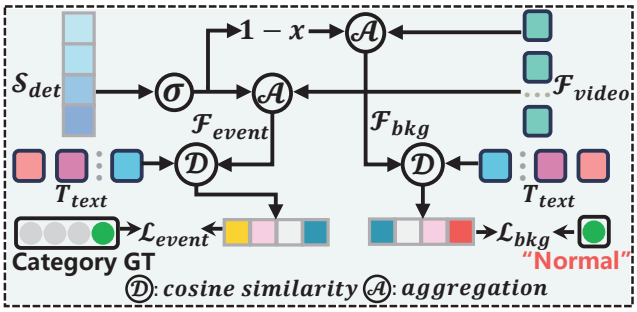

Figure 3: Detailed structure of the proposed Decoupled Contrastive Semantic Alignment module.

To unify the discriminative and generative anomaly views (S_{det} and S_{rec}), we introduce a consistency loss $\mathcal{L}_{consist}$ based on mean squared error. This acts as a form of knowledge distillation, where each branch guides the other.

$$\mathcal{L}_{consist} = \frac{1}{N}\sum_{i=1}^{N} \left(S_{det}(i) - S_{rec}(i)\right)^2. \tag{4}$$

This consistency objective encourages mutual refinement between branches by combining MIL's focus on anomalous segments with the SG-NM's modeling of intrinsic normality, promoting a holistic understanding of video anomalies.

3.3 Decoupled Contrastive Semantic Alignment

Fine-grained anomaly detection under weak supervision is challenged by visual similarity among anomaly types and interference from surrounding normal frames, leading to semantic confusion and poor class separability. To resolve this, we propose a Decoupled Contrastive Semantic Alignment (DCSA) mechanism shown in Figure 3 for fine-grained cross-modal alignment between event-background decoupled visual features and text features, thus enhancing category discrimination and temporally localization accuracy.

Specifically, we first decouple visual features. Given video features F_{video}, we compute event-centric prototype F_{event} and background-centric prototype F_{bkg} using frame-level anomaly scores S_{det} produced by the detection branch:

$$F_{event} = w_{event}^{\top} F_{video}, \quad F_{bkg} = w_{bkg}^{\top} F_{video}, \tag{5}$$

where $w_{event} = \text{Softmax}(S_{det}), w_{bkg} = 1 - w_{event}$. This performs a context-preserving soft decoupling via weighted aggregation of all frame features, shifting the focus to either salient events (F_{event}) or the background (F_{bkg}).

Subsequently, we decouple text features. The class text embeddings $T_{text} = \{t_0, t_1, \ldots, t_{C-1}\}$ represent the "normal" class (t_0) and the $C-1$ anomaly classes $(t_1, \ldots, t_{C-1})$. To promote a clear margin in the semantic space, we use a separation loss that pushes the "normal" class embedding t_0 away from all abnormal class embeddings $\{t_a \mid a \in \{1, ..., C-1\}\}$. This is achieved by minimizing the absolute cosine similarities:

$$\mathcal{L}_{sep} = \sum_{a=1}^{C-1} \left| \frac{t_0^\top t_a}{\|t_0\| \|t_a\|} \right|. \tag{6}$$

Finally, we apply fine-grained cross-modal semantic alignment to enhance class-discrimination. Given a video label $y \in \{0, 1, \ldots, C-1\}$, we define a dual contrastive alignment loss as:

$$\mathcal{L}_{dcsa} = \mathcal{L}_{event} + \mathcal{L}_{bkg}. \tag{7}$$

The event alignment loss encourages F_{event} to align with its ground-truth class t_c, while contrasting it against others:

$$\mathcal{L}_{event} = -\sum_{c=0}^{C-1} y_c \log \frac{\exp\left(\text{sim}(F_{event}, t_c)/\tau\right)}{\sum_{j=0}^{C-1} \exp\left(\text{sim}(F_{event}, t_j)/\tau\right)}, \tag{8}$$

The background alignment loss consistently aligns F_{bkg} with the "normal" embedding t_0, serving as a semantic regularizer to encourage a stable representation for normality:

$$\mathcal{L}_{bkg} = -\log \frac{\exp\left(\text{sim}(F_{bkg}, t_0)/\tau\right)}{\sum_{c=0}^{C-1} \exp\left(\text{sim}(F_{bkg}, t_c)/\tau\right)}. \tag{9}$$

Here, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. This decoupled formulation addresses both abnormal and normal cases:

- Abnormal video ($y = k \neq 0$): $F_{\text{event}} \leftrightarrow t_k$, $F_{\text{bkg}} \leftrightarrow t_0$.
- Normal video ($y = 0$): $F_{\text{event}} \leftrightarrow t_0$, $F_{\text{bkg}} \leftrightarrow t_0$.

This decoupled contrastive semantic alignment helps disentangle anomaly-relevant features from shared contextual patterns, which reduces cross-category semantic confusion and enhances class-discriminative representations.

3.4 Lightweight Text Adapter

To enable domain-specific adaptation while preserving the general knowledge of CLIP, we insert lightweight adapters into the early Transformer blocks of the text encoder. Specifically, adapters are placed in the first L layers, and a fusion weight ω_t controls the contribution of the adapted features. Each adapter operates in parallel with the original self-attention and feed-forward layers. Given the intermediate feature x, the adapter produces an output x_{adapt}, which is fused with the original as follows:

$$x_{out} = (1 - \omega_t) \cdot x + \omega_t \cdot \text{Norm}(x_{adapt}). \tag{10}$$

This modulation preserves the original feature norm while allowing directional adaptation in the embedding space.

3.5 Training and Inference.

Training. The model is optimized with a unified loss balancing learning objectives across all branches:

$$\begin{aligned} \mathcal{L}_{total} = \mathcal{L}_{det} + \lambda \mathcal{L}_{align} + \mathcal{L}_{consist} + \\ \mathcal{L}_{compact} + \mathcal{L}_{dcsa} + \mathcal{L}_{sep}, \end{aligned} \tag{11}$$

where λ balances different loss terms.

Inference. For coarse-grained WS-VAD, we directly use S_{det} from the Anomaly Detection Branch, which provides a reliable frame-level binary confidence for detecting anomalies. For fine-grained prediction, we propose a Hierarchical Belief Modulation strategy that integrates temporal cues from the Anomaly Detection Branch's S_{det} with semantic cues from the Anomaly Classification Branch's S_{align}. Specifically, S_{det} serves as a temporal prior. This anomaly score is then distributed over the C fine-grained classes according to their relative likelihoods predicted by S_{align}. This design ensures that the final predictions are anchored to the robust temporal boundaries identified by the Anomaly Detection Branch, while the Anomaly Classification Branch's expertise is precisely targeted at resolving ambiguity between different anomaly classes. The process is calibrated by a single hyperparameter, the temperature ratio β.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate our method on two widely-used WS-VAD benchmarks: XD-Violence (Wu et al. 2020) and UCF-Crime (Sultani, Chen, and Shah 2018), both providing only video-level labels for training. For coarse-grained evaluation, we follow standard protocols using frame-level AUC for UCF-Crime and Average Precision (AP) for XD-Violence. For fine-grained evaluation, we follow previous work (Wu et al. 2024b), computing mean Average Precision (mAP) under IoU thresholds of 0.1 to 0.5 with step 0.1, and report average mAP (AVG).

Implementation Details. To ensure a fair comparison with existing methods (Wu et al. 2024b; Dev, Hazari, and Das 2024), the frozen CLIP (ViT-B/16) is adopted to extract features. Temperature scaling factors in Eq. (2), (8) and (9) are set to 0.07. The text adapter is set to $L = 3$ with fusion weight $\omega_t = 0.1$ for UCF-Crime and $L = 1$, $\omega_t = 0.6$ for XD-Violence. An 8-layer cross-attention decoder is used in the SG-NM branch, where M is set to 80% of the video length to select candidate normal frames. During inference, the temperature ratio β is set to 5.0 on UCF-Crime and 1.0 on XD-Violence. The loss weight λ for the unified objective is 1.1 on UCF-Crime and 5.0 on XD-Violence. We use the AdamW optimizer with batch sizes of 64 (UCF-Crime) and 96 (XD-Violence) to optimize the model. All experiments are conducted on a single NVIDIA 4090 GPU using PyTorch. Training lasts 10 epochs with learning rates of $7\text{e}{-5}$ for UCF-Crime and $1\text{e}{-5}$ for XD-Violence.

4.2 Main Results

Coarse-grained WS-VAD Results. Tables 1 and 2 show coarse-grained results on XD-Violence and UCF-Crime. On XD-Violence, DSANet achieves **86.95%** AP, surpassing ReFLIP (85.81%) and ITC (85.45%). On UCF-Crime, it reaches **89.44%** AUC, outperforming ITC (89.04%) and ReFLIP (88.57%) with the same backbone. These results validate the strong reasoning capabilities of DSANet in coarse-level anomaly detection under weak supervision.

Fine-grained WS-VAD Results. We assess DSANet's classification and localization ability in fine-grained anomaly

Category	Method	Features	AP(%)
Un	LTR(Hasan et al. 2016)	-	30.77
Weak	RAD(Sultani, Chen, and Shah 2018)	C3D	73.20
	RTFML(Tian et al. 2021)	I3D	77.81
	ST-MSL(Li, Liu, and Jiao 2022)	I3D	78.28
	LA-Net(Pu and Wu 2022)	I3D	80.72
	DMU(Zhou, Yu, and Yang 2023)	I3D	82.41
	PEL4VAD(Pu et al. 2024)	I3D	85.59
	CLIP-TSA(Joo et al. 2023)	CLIP	82.19
	TPWNG(Yang, Liu, and Wu 2024)	CLIP	83.68
	VadCLIP(Wu et al. 2024b)	CLIP	84.51
	ITC(Liu, Lam, and Bao 2024)	CLIP	85.45
	ReFLIP(Dev, Hazari, and Das 2024)	CLIP	85.81
	DSANet(Ours)	**CLIP**	**86.95**

Table 1: Coarse-grained comparisons on XD-Violence.

Category	Method	Features	AUC(%)
Un	LTR(Hasan et al. 2016)	-	50.60
Weak	RAD(Sultani, Chen, and Shah 2018)	I3D	77.92
	RTFML(Tian et al. 2021)	I3D	84.30
	LA-Net(Pu and Wu 2022)	I3D	85.12
	ST-MSL(Li, Liu, and Jiao 2022)	I3D	85.30
	DMU(Zhou, Yu, and Yang 2023)	I3D	86.75
	PEL4VAD(Pu et al. 2024)	I3D	86.76
	CLIP-TSA(Joo et al. 2023)	CLIP	87.58
	TPWNG(Yang, Liu, and Wu 2024)	CLIP	87.79
	VadCLIP(Wu et al. 2024b)	CLIP	88.02
	ReFLIP(Dev, Hazari, and Das 2024)	CLIP	88.57
	ITC(Liu, Lam, and Bao 2024)	CLIP	89.04
	DSANet(Ours)	**CLIP**	**89.44**

Table 2: Coarse-grained comparisons on UCF-Crime.

detection. As shown in Table 3, it consistently outperforms prior methods on XD-Violence across all IoUs, with an average mAP of **28.87%**, surpassing ReFLIP (27.36%) and ITC (26.83%), demonstrating robustness in capturing temporal anomaly boundaries. On the more challenging UCF-Crime (Table 4), DSANet achieves the best performance with an average mAP of **13.01%**, outperforming ReFLIP (9.62%) and ITC (7.90%), with clear gains across all IoU levels (*e.g.*, 21.39% at 0.1 and 8.00% at 0.5). These results confirm DSANet's effectiveness in capturing cross-modal fine-grained semantics for precise anomaly classification and localization.

4.3 Ablation Studies

We conduct ablation studies on XD-Violence to dissect our model and validate the contribution of its key components.

Effectiveness of Components. Table 5 shows module-wise impact. Based on the VadCLIP (Wu et al. 2024b) baseline, adding the Adapter improves AP / AVG to 84.75% / 28.31%, showing the value of task-specific text adaptation. Introducing the Self-Guided Normality Modeling branch raises performance to 85.94% / 28.84%, showing its effectiveness in anomaly detection. Adding the Decoupled Contrastive Semantic Alignment mechanism yields 85.70% / 28.60%, confirming its role in semantic alignment for better localization and classification. DSANet with all modules reaches 86.95% / 28.87%, surpassing baseline by 2.44% AP and 4.17% AVG, with strong component synergy.

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
RAD(2018)	22.72	15.57	9.98	6.20	3.78	11.65
AVVD(2022)	30.51	25.75	20.18	14.83	9.79	20.21
VadCLIP(2024)	37.03	30.84	23.38	17.90	14.31	24.70
ITC(2024)	40.83	32.80	25.42	19.65	15.47	26.83
ReFLIP(2024)	39.24	33.45	27.71	20.86	17.22	27.36
DSANet(Ours)	**40.93**	**34.63**	**28.21**	**22.70**	**17.89**	**28.87**

Table 3: Fine-grained comparisons on XD-Violence.

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
RAD(2018)	5.73	4.41	2.69	1.93	1.44	3.24
AVVD(2022)	10.27	7.01	6.25	3.42	3.29	6.05
VadCLIP(2024)	11.72	7.83	6.40	4.53	2.93	6.68
ITC(2024)	13.54	9.24	7.45	5.46	3.79	7.90
ReFLIP(2024)	14.23	10.34	9.32	7.54	6.81	9.62
DSANet(Ours)	**21.39**	**14.96**	**11.74**	**8.98**	**8.00**	**13.01**

Table 4: Fine-grained comparisons on UCF-Crime.

Effectiveness of Text Encoder Tuning. Table 6 compares CLIP text encoder adaptation strategies. Using a frozen encoder with class labels performs moderately (81.38% AP, 27.60% AVG). Manual prompt("a video of <label>") slightly improves AVG but reduces AP. Learning prompt (*e.g.*, CoOp-style) improves to (82.69% AP, 28.66% AVG), while our Adapter-based tuning performs best (86.95% AP, 28.87% AVG), confirming that internal adaptation offers more expressive and task-aligned features than others.

4.4 Qualitative Results

t-SNE Visualization of Category Separability. To assess feature discriminativeness, we visualize UCF-Crime using t-SNE in Figure 4. Compared to original CLIP features (left), which show a heavy category overlap, our model's features (right) exhibit clearer inter-class boundaries and tighter intra-class clusters. This improved separability demonstrates that our model enhances class-discriminative representa-

Adapter	SG-NM	DCSA	AP(%)	AVG(%)
Baseline			84.51	24.70
✓			85.00	28.15
✓	✓		85.94	28.39
✓		✓	85.67	28.25
✓	✓	✓	**86.95**	**28.87**

Table 5: Ablation studies on model components. "SG-NM" denotes Self-Guided Normality Modeling, and "DCSA" denotes Decoupled Contrastive Semantic Alignment.

Method	AP(%)	AVG(%)
No tuning	81.57	27.60
Manual Prompt	81.05	28.05
Learning Prompt	82.88	28.26
Ours(Adapter)	**86.95**	**28.87**

Table 6: Effectiveness of Text Encoder Tuning.

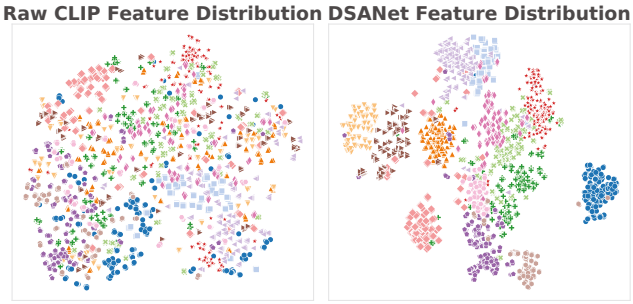

Figure 4: t-SNE visualizations for UCF-Crime.

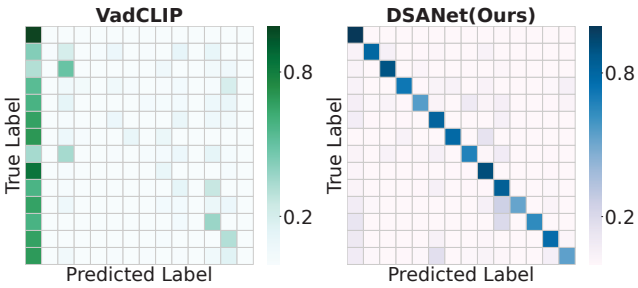

Figure 6: Comparative results of semantic alignment matrix.

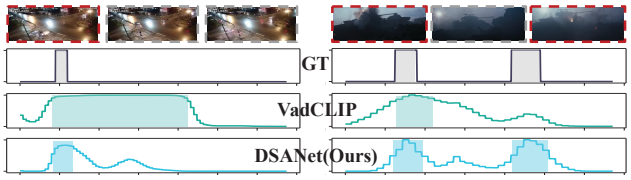

Figure 7: Comparison of coarse-grained anomaly detection.

tions, especially for visually similar anomaly categories.

Effects of Dynamic Normal Patterns. To evaluate the quality of our Dynamic Normal Patterns (DNPs), we measure the minimum cosine distance from each frame to its video's K DNPs, assuming normal frames are closer to this DNP-defined space. We compute distributions of these distances across all frames in the test set. Figure 5 shows that normal frames concentrate at lower distances (mean: 0.35), while abnormal ones lie farther away (mean: 0.69), confirming that DNPs form a compact and discriminative representation of normal patterns. This validates the reconstruction pathway guided by DNPs (Sec. 3.2), where using only DNPs as key and value ensures high reconstruction errors for anomalous frames, making anomalies readily detectable.

Effects of Decoupled Contrastive Semantic Alignment. We evaluate the Decoupled Contrastive Semantic Alignment(DCSA) by analyzing the alignment behavior of learned event-centric and background-centric prototypes. On UCF-Crime, we compare DSANet with VadCLIP (Wu et al. 2024b), which is adapted to compute both prototypes using our formulation (Eq. 5). For each video, its two prototypes are aligned with 14 class-level text embeddings, and the closest match is taken as the predicted label. Figure 6 compares the confusion matrices of the event-centric prototype. VadCLIP shows severe misclassification among visually similar categories and often defaults to the "normal" class. In contrast, DSANet produces a much stronger diagonal dominance, reflecting improved class separability. For the background-centric prototype, DSANet achieves a 99.63% alignment accuracy with "normal" class, outperforming VadCLIP's 87.63%, confirming its ability to isolate background information from event content. These results demonstrate that DCSA effectively reduces class confusion, leading to more precise and disentangled representations for fine-grained anomaly detection.

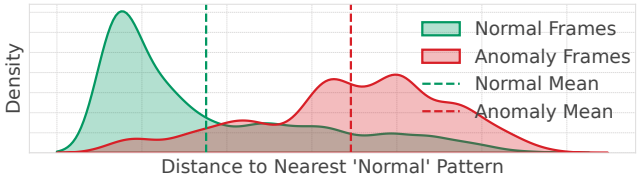

Figure 5: Comparison of Frame Distances to DNPs.

Results on Coarse-Grained Anomaly Detection. We visualize the anomaly detection results of DSANet and VadCLIP in Figure 7. As shown, VadCLIP often captures only the most salient anomaly segments, resulting in fragmented predictions and misaligned temporal boundaries. This aligns with our observation that MIL-only methods tend to focus on peak activation regions, leading to imprecise localization. In contrast, DSANet produces predictions that align more accurately with the ground truth, accurately covering abnormal events while maintaining clear separation from normal regions. This improvement stems from leveraging learned normality prototypes, which guide the model to better capture temporal structure, thereby complementing the detection branch and enhancing localization precision. These results confirm that explicitly modeling normal patterns is beneficial for stabilizing anomaly predictions and reducing boundary ambiguity in coarse-grained detection tasks.

5 Conclusion

In this paper, we present a novel framework named DSANet for WS-VAD that disentangles normal and anomaly semantics at both coarse-grained and fine-grained levels. By integrating self-guided normality modeling and decoupled contrastive semantic alignment, DSANet achieves improved temporal localization and class discrimination. Extensive experiments validate the efficacy of the proposed DSANet, achieving state-of-the-art anomaly detection performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under grants U22B2053 and 623B2039, and in part by the Interdisciplinary Research Program of HUST (2024JCYJ034).

References

Abdalla, M.; Javed, S.; Radi, M. A.; Ulhaq, A.; and Werghi, N. 2024. Video anomaly detection in 10 years: A survey and outlook. *arXiv preprint arXiv:2405.19387*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

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

Chen, J.; Li, L.; Su, L.; Zha, Z.-j.; and Huang, Q. 2024. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18319–18329.

Dev, P. P.; Hazari, R.; and Das, P. 2024. Reflip-vad: Towards weakly supervised video anomaly detection via vision-language model. *IEEE Transactions on Circuits and Systems for Video Technology*.

Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14009–14018.

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 733–742.

Ho, H.-T.; Nguyen, L. V.; Pham, M.-T.; Pham, Q.-H.; Tran, Q.-D.; Huy, D. N. M.; and Nguyen, T.-H. 2025. A Review on Vision-Language-Based Approaches: Challenges and Applications. *Computers, Materials & Continua*, 82(2).

Huang, X.; Zhou, H.; Yao, K.; and Han, K. 2024. FROSTER: Frozen CLIP Is A Strong Teacher for Open-Vocabulary Action Recognition. arXiv:2402.03241.

Jia, C.; Luo, M.; Chang, X.; Dang, Z.; Han, M.; Wang, M.; Dai, G.; Dang, S.; and Wang, J. 2023. Generating Action-conditioned Prompts for Open-vocabulary Video Action Recognition. arXiv:2312.02226.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Joo, H. K.; Vo, K.; Yamazaki, K.; and Le, N. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, 3230–3234. IEEE.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, L.; Peng, J.; Chen, H.; Gao, C.; and Yang, X. 2024. How to configure good in-context sequence for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26710–26720.

Li, S.; Liu, F.; and Jiao, L. 2022. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1395–1403.

Liang, J.; Li, T.; Yang, J.; Li, Y.; Fang, Z.; and Yang, F. 2023. Video anomaly detection by fusing self-attention and autoencoder. *Journal of Image and Graphics*, 28(4): 1029–1040.

Liu, T.; Lam, K.-M.; and Bao, B.-K. 2024. Injecting text clues for improving anomalous event detection from weakly labeled videos. *IEEE Transactions on Image Processing*.

Luo, W.; Cao, Y.; Yao, H.; Zhang, X.; Lou, J.; Cheng, Y.; Shen, W.; and Yu, W. 2025. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9974–9983.

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

Nayak, R.; Pati, U. C.; and Das, S. K. 2021. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106: 104078.

Pu, Y.; and Wu, X. 2022. Locality-aware attention network with discriminative dynamics learning for weakly supervised anomaly detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Pu, Y.; Wu, X.; Yang, L.; and Wang, S. 2024. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*.

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

Shi, Y.; Gao, Y.; Lai, Y.; Wang, H.; Feng, J.; He, L.; Wan, J.; Chen, C.; Yu, Z.; and Cao, X. 2025. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *Visual Intelligence*, 3(1): 9.

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

Tian, K.; Zhao, R.; Xin, Z.; Lan, B.; and Li, X. 2024. Holistic features are almost sufficient for text-to-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17138–17147.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4975–4986.

Wang, H.; Tian, Y.; Liu, M.; Zhang, Z.; and Zhu, X. 2025a. SDEval: Safety Dynamic Evaluation for Multimodal Large Language Models. *arXiv preprint arXiv:2508.06142*.

Wang, H.; Wang, S.; Zhong, Y.; Yang, Z.; Wang, J.; Cui, Z.; Yuan, J.; Han, Y.; Liu, M.; and Ma, Y. 2025b. Affordance-R1: Reinforcement Learning for Generalizable Affordance Reasoning in Multimodal Large Language Model. *arXiv preprint arXiv:2508.06206*.

Wang, H.; Zhang, Z.; Ji, K.; Liu, M.; Yin, W.; Chen, Y.; Liu, Z.; Zeng, X.; Gui, T.; and Zhang, H. 2025c. DAG: Unleash the Potential of Diffusion Model for Open-Vocabulary 3D Affordance Grounding. *arXiv preprint arXiv:2508.01651*.

Wang, J.; Sun, G.; Wang, P.; Liu, D.; Dianat, S.; Rabbani, M.; Rao, R.; and Tao, Z. 2024. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16551–16560.

Wang, X.; Zhang, S.; Qing, Z.; Shao, Y.; Zuo, Z.; Gao, C.; and Sang, N. 2021. OadTR: Online Action Detection With Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7565–7575.

Wang, X.; Zhang, S.; Qing, Z.; Tang, M.; Zuo, Z.; Gao, C.; Jin, R.; and Sang, N. 2022. Hybrid Relation Guided Set Matching for Few-Shot Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19948–19957.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, 322–339. Springer.

Wu, P.; Pan, C.; Yan, Y.; Pang, G.; Wang, P.; and Zhang, Y. 2024a. Deep learning for video anomaly detection: A review. *arXiv preprint arXiv:2409.05383*.

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

Yang, Z.; Liu, J.; and Wu, P. 2024. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18899–18908.

Yu, C.; Wang, H.; Shi, Y.; Luo, H.; Yang, S.; Yu, J.; and Wang, J. 2025. SeqAfford: Sequential 3D Affordance Reasoning via Multimodal Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1691–1701.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv:2205.01917.

Zanella, L.; Menapace, W.; Mancini, M.; Wang, Y.; and Ricci, E. 2024. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Huang, X.; Gao, C.; Zhang, S.; Yu, L.; and Sang, N. 2025. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13843–13853.

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

Zhu, X.; Qian, X.; Shi, Y.; Tao, X.; and Li, Z. 2024. Video anomaly detection with long-and-short-term time series correlations. *Journal of Image and Graphics*, 29(7): 1998–2010.

Zou, B.; Yang, C.; Qiao, Y.; Quan, C.; and Zhao, Y. 2024. Language-aware visual semantic distillation for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27113–27123.