This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Alert-CLIP: Abnormality-aware Latent-Enhanced Representation Tuning of CLIP for Video Anomaly Detection

Yiyan Zhu[1], Menghao Zhang[1], Haifeng Sun[1], Pengfei Ren[1], Xianao Chu[1], Chenye Xu[1], Hong Tan[1], Jinghan Wang[1], Qi Qi[1,2,*], Jingyu Wang[1,3,*]

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China
[2]E-Byte.com, Beijing, China
[3]Pengcheng Laboratory, Shenzhen, China

{zhuyiyan,zhangmenghao,hfsun,rpf,redamancy,xuchenye}@bupt.edu.cn
tanh275774@163.com; {wangjhbh,qiqi8266,wangjingyu}@bupt.edu.cn

Abstract

With the rise of pre-trained vision–language models such as CLIP, performing video anomaly detection (VAD) through cross-modal reasoning has become an emerging trend. However, we observe that CLIP still suffers from weak abnormality awareness: normal and abnormal descriptions are highly entangled in the text embedding space, causing video features to assign nearly indistinguishable similarity scores to both types of prompts. To address this issue, we propose ***Alert-CLIP****, an abnormality-aware latent–enhanced tuning framework that tailors CLIP for VAD. Alert-CLIP introduces a multi-level alignment strategy: (1) video–label alignment, which reshapes the semantic space to establish a coarse-level foundation for abnormality awareness; (2) region–text alignment, which explicitly associates anomaly-related regions with detailed descriptions to strengthen fine-grained perception; (3) region–semantic alignment, which contrasts anomalous regions against multiple hard negative samples to enhance abnormality-aware discrimination. To support this training, we construct* ***VAGTA****. Extensive experiments show that* ***Alert-CLIP*** *consistently surpasses CLIP across weakly supervised, zero-shot, and open-vocabulary settings.* ***VAGTA*** *is publicly available at https://github.com/ClarkZhu216/Alert-CLIP_dataset.*

1. Introduction

Video anomaly detection (VAD) aims to automatically identify events that deviate from normal behavioral patterns.

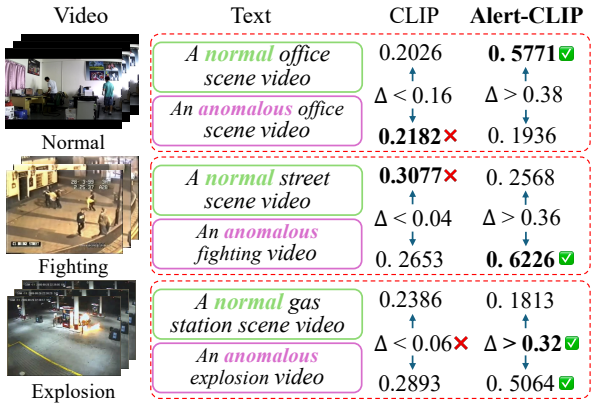


Figure 1. Qualitative comparison between CLIP and Alert-CLIP.

In real-world scenarios, anomalous events are extremely rare, making it infeasible to collect sufficient anomalous samples for training. As a result, most prior works adopt semi-supervised methods [7, 21, 29, 49] that learn normal patterns from normal videos or weakly supervised methods [25, 30, 36, 43] trained with video-level labels. With the demonstrated effectiveness of pre-trained vision–language models (VLMs), such as CLIP [27], across a wide range of downstream tasks [13, 20, 27], leveraging CLIP for VAD has attracted increasing attention.

Recent CLIP-based VAD studies [10, 38–40, 46] exploit text embeddings to provide semantic guidance for visual modeling. In practice, these methods typically keep the CLIP frozen and introduce task-specific designs to model anomalies on top of the extracted embeddings. VadCLIP [39] and AnomalyCLIP [46] represent early attempts that combine visual classification and vision–language

*Corresponding authors.

alignment to enable both coarse- and fine-grained anomaly detection. Subsequent works further improve CLIP-based VAD by enriching prompts with additional cues. For instance, STPrompt [38] incorporates spatio-temporal cues to facilitate anomaly localization, Ex-VAD [10] leverages LLM-generated anomaly explanations to enhance interpretability, and VarCMP [40] introduces audio cues to support cross-modal anomaly retrieval.

Despite their encouraging progress, existing CLIP-based approaches largely rely on representations with limited anomaly awareness, which restricts their potential for further performance improvement in VAD. Anomaly is inherently defined as a deviation from the norm, rather than a standalone semantic category or object attribute. Such a property requires models to have the ability to perceive context semantics. Nevertheless, CLIP's pretraining aligns full images with textual descriptions through contrastive learning, which hinders its ability to capture contextual semantics and fine-grained information. As shown in Figure 1, CLIP produces nearly indistinguishable similarities for normal and anomalous prompts, and even favors the normal description for certain anomalous videos. In Section 3, we provide a more comprehensive analysis along with extensive experimental evidence.

To overcome CLIP's limitations in anomaly awareness, we propose **Alert-CLIP**, a multi-level vision–language alignment training framework tailored for VAD. Alert-CLIP preserves CLIP's semantic priors while progressively strengthening abnormality awareness through three complementary alignment levels. At the *video-label* level, we restructure the semantic space to establish a coarse semantic foundation for separating normal and anomalous. At the *region-text* level, we align anomaly-related visual regions with their detailed textual descriptions, enabling fine-grained correspondence between localized visual changes and the contextual semantics. At the *region-semantic* level, we strengthen the understanding of normal and anomalous concepts by contrasting the groundtruth and hard-negative captions. Together, these multi-level alignment mechanisms substantially enhance CLIP's abnormality awareness and enable Alert-CLIP to serve as a stronger and more reliable foundation for VAD.

To support the training of Alert-CLIP, we construct the Video Anomaly Dataset with Grounded Text Annotations **(VAGTA)** by re-annotating videos from UCF-Crime [30] and MSAD [51]. Due to the inconsistent quality of existing datasets and the lack of region-level annotations, our data preparation process contains three steps. First, we standardize video quality by filtering out videos that do not meet the standards, and annotate each remaining video with a global video caption, providing coarse-grained information for *video-label* level. Second, we annotate precise bounding boxes for anomalous regions and provide region-level captions that describe the corresponding events. This provides crucial fine-grained visual-textual supervision for grounding anomaly-related regions at the *region-text* level. Third, we enhance semantic contrast by generating hard-negative captions. This offers fine-grained semantic negatives for the *region–semantic* level, facilitating more discriminative separation between normal and abnormal concepts. This procedure yields 4,212 high-quality clips with consistent spatial and semantic annotations, which form **VAGTA** and support the multi-level alignment training of Alert-CLIP.

We conduct extensive experiments with Alert-CLIP on widely used VAD benchmarks, including weakly supervised detection, zero-shot anomaly detection, and open-vocabulary anomaly detection settings. The results show that Alert-CLIP consistently outperforms CLIP and existing methods. Alert-CLIP can be seamlessly integrated into existing VAD pipelines as a drop-in backbone.

In summary, our contributions are threefold:

- We provide a analysis of CLIP's limitations in VAD, revealing its weak abnormality awareness. We offer both qualitative and quantitative evidence through text-space probing and video–text similarity studies.
- We propose a multi-level vision–language alignment framework, **Alert-CLIP**, which significantly improves anomaly awareness and can seamlessly serve as a backbone for existing VAD pipelines.
- We construct **VAGTA**, a video anomaly dataset consisting of 4,212 curated clips with region annotations and multi-level captions to support Alert-CLIP training.

2. Related Work

2.1. Video Anomaly Detection

Existing VAD approaches can be broadly divided into semi-supervised and weakly supervised methods. Semi-supervised VAD [3, 12, 22, 24] usually models the normal patterns using only normal data and treats deviations as anomalies. Typical methods include approaches based on frame reconstruction [9, 26], frame prediction [4, 22, 33], and multi-task learning [29, 49]. Weakly supervised methods typically rely on video-level annotations and are predominantly formulated under the Multiple Instance Learning paradigm [6, 30, 32, 35, 45].

With the rapid development of pre-trained vision-language models (VLMs), such as CLIP [27], recent VAD studies [38, 39] increasingly leverage textual semantics to guide visual feature modeling. For instance, VadCLIP [39] is an early attempt to introduce CLIP into VAD, where video labels are encoded as textual prompts to guide the training of a dual-branch model. Building upon VadCLIP, VadCLIP++ [19] extends the framework by introducing dynamic text prompts to enhance performance. AVadCLIP [41] further incorporates audio information to im-

prove robustness in complex environments. STPrompt [38] introduces spatio-temporal prompt embeddings to localize abnormal regions in the video frames. OVVAD [37] explores VAD under an open-vocabulary supervision setting, leveraging semantic guidance to enable generalization to unseen novel anomaly categories. Although these CLIP-based methods demonstrate promising performance, they typically rely on a frozen backbone and utilize CLIP-extracted features for anomaly detection, without addressing CLIP's limited anomaly awareness. In this work, we systematically analyze the limitations of CLIP in anomaly awareness and propose a training framework to enhance its ability to recognize anomalies across diverse scenarios.

2.2. Vision–Language Models and CLIP

Large-scale pre-trained VLMs [11, 15, 47], such as CLIP [27], achieve outstanding results in downstream tasks like image classification [23], segmentation [34], and detection [39]. Despite its success, CLIP still faces limitations in fine-grained perception due to its pretraining approach. Specifically, CLIP is pretrained with contrastive learning that aligns entire images with coarse text prompts. For example, it uses prompts like 'a photo of [category]', without incorporating pixel-level visual information or detailed textual descriptions.

Recent works try to enhance CLIP's fine-grained perception of object attributes by extending the token length of the text encoder [48] or introducing grounding data [16, 50]. However, these methods still fail to fit VAD, as they do not address the limitation in anomaly awareness. This is because anomalies are not conventional semantic categories or object attributes, but rather deviations from the norm that require the ability to understand context semantics. To bridge this gap, our work proposes a multi-level framework that enhances CLIP's anomaly awareness by aligning visual information of varying granularities with textual descriptions at different levels.

3. Preliminary

To validate our hypothesis that CLIP suffers from an inherent deficiency in distinguishing the semantics of *normal* and *abnormal* events, we conduct a qualitative analysis and a quantitative analysis. Specifically, the qualitative analysis examines the geometry of normal/abnormal prompts in the text space and their alignment with real videos, while the quantitative analysis measures this phenomenon using the proposed MeanSignScore (MSS). We also include Alert-CLIP as a reference model to verify whether the observed deficiency can be alleviated after adaptation. The qualitative results are shown in Figure 2 and Figure 1, while the quantitative results are summarized in Table 1.

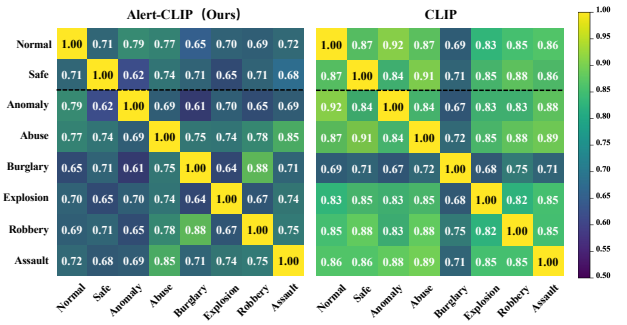


Figure 2. Pairwise cosine similarities among normal- and abnormal-related prompts in the text space. CLIP exhibits substantial semantic overlap between normal and abnormal concepts, while Alert-CLIP shows a clearer separation.

3.1. Unified Similarity-Based Diagnosis

We first examine the semantic boundary between *normal* and *abnormal* from both the text and video perspectives. On the text side, we select representative descriptions from UCF-Crime [30] and compute pairwise cosine similarities between their text embeddings to examine whether CLIP forms a clear semantic boundary between the two groups.

Figure 2 shows that CLIP exhibits substantial semantic entanglement between normal and abnormal concepts. Normal-related descriptions remain highly similar to many anomaly descriptions, indicating that the boundary between the two semantics is highly blurred in CLIP's text space. By contrast, Alert-CLIP produces a clearer separation while preserving the consistency among anomaly prompts, suggesting that this ambiguity is rooted in CLIP's representation space and can be alleviated through targeted adaptation.

We next examine whether this ambiguity propagates to video–text alignment. Figure 1 presents several representative normal and anomalous videos from UCF-Crime. For each video, we report its cosine similarity to a normal prompt and a matched anomaly prompt, together with the margin between the correct prompt and the alternative one. CLIP often assigns very similar scores to the two prompts, and in some cases even favors the wrong description, resulting in weak and unstable margins. In contrast, Alert-CLIP shows a more consistent preference for the semantically correct prompt and yields substantially larger margins.

Taken together, these results reveal a consistent failure mode of CLIP: it does not establish a clear semantic boundary between normal and abnormal concepts. This deficiency appears in the text space and further manifests itself in video–text alignment, leading to unreliable discrimination on real videos. The comparison with Alert-CLIP further verifies that this failure mode stems from the learned representation itself rather than incidental prompt variation.

Table 1. Video-level MeanSignScore (MSS) of different models on XD and UCF-Crime. Higher is better.

Model	XD	UCF-Crime	Avg.
CLIP	-0.0572	-0.0556	-0.0564
LongCLIP	-0.0586	-0.0508	-0.0547
SigLIP	-0.0804	-0.0620	-0.0712
BLIP	-0.0482	-0.0620	-0.0551
Alert-CLIP (Stage 1)	0.0541	0.1733	0.1137
Alert-CLIP (Stage 2)	0.1634	0.1726	0.1680
Alert-CLIP (Final)	**0.1648**	**0.1763**	**0.1706**

3.2. MeanSignScore: Quantifying Separability

To complement the qualitative analyses above, we further introduce a simple diagnostic metric, termed **MeanSignScore (MSS)**, to quantify how clearly a model separates normal and abnormal semantics. Let $\mathcal{P}_N$ and $\mathcal{P}_A$ denote the sets of normal and abnormal prompts, respectively. For a video x, we compute its average cosine similarity to the two prompt sets as

$$S_N(x) = \frac{1}{|\mathcal{P}_N|} \sum_{p \in \mathcal{P}_N} \cos(f(x), g(p)), \tag{1}$$

$$S_A(x) = \frac{1}{|\mathcal{P}_A|} \sum_{p \in \mathcal{P}_A} \cos(f(x), g(p)), \tag{2}$$

where $f(\cdot)$ and $g(\cdot)$ denote the video and text encoders, respectively. We then define the signed score of each sample as

$$\mathrm{SS}(x) = \begin{cases} S_A(x) - S_N(x), & \text{if } x \text{ is anomalous,} \\ S_N(x) - S_A(x), & \text{if } x \text{ is normal.} \end{cases} \tag{3}$$

Finally, MSS is computed by averaging the signed scores over the dataset:

$$\mathrm{MSS} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathrm{SS}(x). \tag{4}$$

A larger MSS indicates that the model more consistently assigns higher similarity to the semantically correct prompt set. Values near zero indicate weak separability, whereas negative values suggest that the model frequently favors the wrong semantic direction.

As shown in Table 1, the baseline models and vision–language backbones consistently yield negative MSS values. This indicates that their video-level similarities do not consistently favor the semantically correct prompt sets, which quantitatively confirms the ambiguity observed in both the text-space geometry and the video–text case studies. In contrast, Alert-CLIP turns MSS positive and substantially improves it across training stages.

Therefore, the proposed preliminary analysis provides a consistent diagnosis from three complementary perspectives: text-space geometry, video–text alignment, and dataset-level quantitative separability. All of them point to the same conclusion: CLIP lacks a reliable semantic boundary between normal and abnormal concepts, and this deficiency directly limits its effectiveness for VAD.

4. Dataset Preparation

In this section, we describe the construction of **VAGTA** for Alert-CLIP, built from two benchmark datasets, UCF-Crime [30] and MSAD [51]. UCF-Crime covers diverse real-world anomaly categories, while MSAD provides scene-specific videos with consistent viewpoints and annotated anomaly segments. Combining them preserves both data scale and complementary supervision for fine-grained anomaly modeling. We strictly follow the official splits of both datasets: all curated training samples are constructed only from the original training splits, while the official test splits are reserved exclusively for evaluation, ensuring no overlap or data leakage.

To support multi-granularity cross-modal alignment, VAGTA provides two levels of textual annotations. Each clip is assigned a global caption and a region-level caption. For abnormal clips, regional features are extracted from annotated bounding boxes; for normal clips, the whole segment is treated as a single region and paired with a fine-grained region caption distinct from the global one. We also use Qwen-VL [31] to generate three hard-negative captions for each region caption that remain visually similar to the original caption while reversing its normal or abnormal semantics. In addition, long normal videos are divided into semantically coherent sub-clips, each inheriting the normal label and receiving its own global and region-level captions.

During data curation, candidate videos are first assigned an initial score by ChatGPT and then verified by human annotators, who evaluate label consistency, anomaly clarity, duplication, and obvious artifacts or corrupted frames. Low-quality abnormal videos are removed, while normal videos that pass the review are retained. For the filtered videos, we then perform the subsequent annotation and curation steps. We first segment each normal video into three temporal parts and treat them as independent normal clips, while keeping abnormal videos unchanged. We then annotate bounding boxes for abnormal regions and write captions for each clip and key abnormal regions, enabling the model to align both global scene semantics and local anomaly semantics.

After this process, we obtain **VAGTA**, which contains 4,212 curated clips in total, including 3,726 training clips and 486 test clips under the official splits. The training split contains 2,585 segmented normal clips and 1,141 abnormal clips. Most removed abnormal videos fall into four cate-

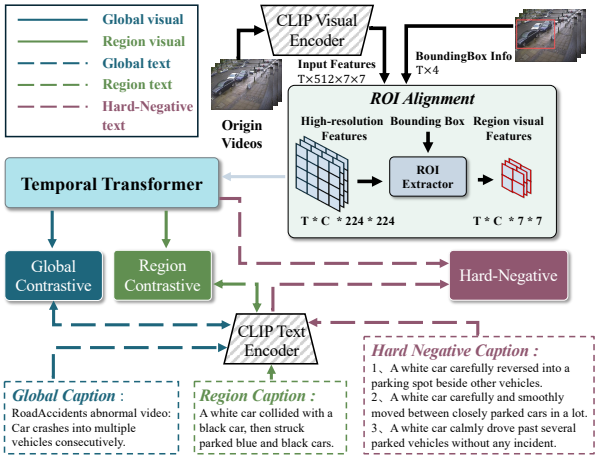


Figure 3. Training framework of **Alert-CLIP**.

gories: category mismatch, weak anomaly semantics, duplicated content, and severely corrupted signals. All clips in **VAGTA** are equipped with consistent bounding-box and caption annotations, providing a stable and reproducible foundation for training Alert-CLIP.

5. Method

5.1. Overview

Video anomaly understanding fundamentally requires distinguishing normal and abnormal behaviors at the semantic level. However, existing CLIP models exhibit highly entangled normal/anomaly representations, making them insensitive to fine-grained deviations. Alert-CLIP keeps the CLIP visual and text encoders intact, and introduces a lightweight temporal transformer to aggregate frame-wise visual features into clip-level representations. On top of this temporal aggregation, we learn a better alignment geometry through three complementary signals: global video descriptions, region descriptions tied to annotated boxes, and semantically close hard negatives that challenge the decision margins. This training framework produces a checkpoint that remains fully compatible with standard CLIP-style inference, while improving anomaly awareness without relying on any localization at test time.

We adopt a two-stage training schedule. Stage I performs global video–text alignment to build a stable separation between normal and anomaly semantics at the description level. Stage II starts from the Stage I weights and jointly optimizes the global, region, and hard-negative branches, injecting fine-grained cues into the shared space.

5.2. Multi-Level Contrastive Learning

We employ the CLIP visual encoder and text encoder, both kept intact during inference, to encode sampled video frames and all accompanying descriptions. Given a video with T sampled frames, the CLIP visual encoder produces frame-wise features, which are aggregated by a lightweight temporal transformer to obtain an ℓ_2-normalized clip-level embedding $v \in \mathbb{R}^D$. For annotated abnormal regions, ROI Align is applied to frame-level visual features, and the resulting region features are temporally aggregated in the same way to obtain region embeddings v^r. Cosine similarity is denoted by $\langle \cdot, \cdot \rangle$. Our objective is to reshape the decision geometry of CLIP through three complementary alignment signals that introduce global scene semantics, spatially grounded anomaly semantics, and fine-grained semantic distinctions.

Global alignment. Given a batch of B videos with global descriptions $\{(v_i, t_i)\}_{i=1}^B$, we compute similarity scores $s_{ij} = \langle v_i, t_j \rangle$ and optimize an InfoNCE loss:

$$\mathcal{L}_{\text{global}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)}, \tag{5}$$

where τ is a temperature parameter. This objective establishes a global semantic separation between normal and anomalous descriptions.

Region alignment. For each video, we associate region-level visual features extracted from annotated bounding boxes with their corresponding region descriptions, producing R positive pairs $\{(v_n^r, t_n^r)\}_{n=1}^R$. With similarities $s_{nm}^r = \langle v_n^r, t_m^r \rangle$, we optimize:

$$\mathcal{L}_{\text{region}} = -\frac{1}{R}\sum_{n=1}^{R} \log \frac{\exp(s_{nn}^r/\tau)}{\sum_{m=1}^{R} \exp(s_{nm}^r/\tau)}. \tag{6}$$

This objective injects spatially grounded, fine-grained anomaly semantics into the shared embedding space.

Semantic hard-negative alignment. For each region, we further refine the semantic boundary by pairing the true region description $t_n^{r,+}$ with K hard negatives $\{t_{n,k}^{r,-}\}_{k=1}^K$ that remain visually similar but reverse the normal or abnormal semantics. Let $s_n^+ = \langle v_n^r, t_n^{r,+} \rangle$ and $s_{n,k}^- = \langle v_n^r, t_{n,k}^{r,-} \rangle$. We optimize:

$$\mathcal{L}_{\text{hard}} = -\frac{1}{R}\sum_{n=1}^{R} \log \frac{\exp(s_n^+/\tau_h)}{\exp(s_n^+/\tau_h) + \sum_{k=1}^{K} \exp(s_{n,k}^-/\tau_h)}, \tag{7}$$

where τ_h controls the strength of semantic discrimination. Unlike region alignment, which focuses on matching visually grounded regions and their descriptions, this branch explicitly sharpens the normal–abnormal boundary by introducing semantically confusable negatives. It therefore improves robustness when normal and abnormal patterns are visually similar but semantically opposite.

5.3. Two-Stage Optimization

Stage I: global pre-alignment. We optimize only $\mathcal{L}_{\text{global}}$ to construct a stable video–text separation between normal

and anomaly descriptions.

Stage II: joint refinement. Starting from the Stage I weights, we jointly minimize

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \lambda\,\mathcal{L}_{\text{region}} + \gamma\,\mathcal{L}_{\text{hard}}, \tag{8}$$

where λ and γ balance the region and hard-negative branches. This stage injects fine-grained cues into the shared space while preserving CLIP's global recognition ability. Bounding boxes and region texts are used only as supervision during training.

6. Experiments

6.1. Experimental Settings

Datasets. We evaluate Alert-CLIP on four widely used benchmarks. UCF-Crime [30] contains 1,900 surveillance videos from 13 scenes, covering 13 anomaly categories and normal events. XD-Violence [36] consists of 4,754 untrimmed videos from diverse sources, focusing on six violence-related anomaly categories. MSAD [51] is designed to assess scene generalization in VAD, covering 14 diverse real-world scenes with about 500 different camera viewpoints. Ubnormal [1] is a synthetic-to-real benchmark for evaluating generalization.

Supervision settings. We evaluate Alert-CLIP under three complementary VAD settings: (i) weakly supervised VAD, where the backbone is fine-tuned within the VadCLIP [39] framework using video-level labels and our multi-granularity annotations; (ii) zero-shot VAD, where no training or adaptation is performed and anomaly scores are derived purely from CLIP-style visual–textual similarity margins; (iii) open-vocabulary VAD, following OVVAD [37], where the model must understand free-form user-defined anomaly descriptions unseen during training. These settings jointly evaluate fine-tuning performance, cross-domain transfer, and semantic generalization.

Evaluation metrics. We report two standard metrics, AUC and AP, for measuring discriminability, ranking quality, and anomaly recall.

Implementation details. Unless specified otherwise, we adopt OpenCLIP ViT-L/32 as the backbone and train *Alert-CLIP* with the two-stage paradigm described in Section 5. We use AdamW with cosine decay, mixed precision, and gradient clipping. All experiments are conducted on a single NVIDIA A800 GPU (80 GB). Detailed hyperparameters are provided in the supplementary material.

6.2. Main Experiments

We compare Alert-CLIP with representative feature backbones. To ensure a fair comparison, all feature types are evaluated under the same experimental setting.

Weakly supervised VAD. Table 2 reports the results of replacing the backbone in VadCLIP with different feature extractors. Alert-CLIP achieves the best performance on both UCF-Crime and MSAD. Stage 1 already improves over CLIP, showing that global alignment is sufficient to establish a cleaner normal–abnormal separation. Stage 2 further improves both AUC and AP, indicating that region alignment and semantic hard negatives provide complementary fine-grained supervision.

Table 2. Weakly supervised results on UCF-Crime and MSAD using Alert-CLIP. Best results are in **bold**.

Feature	UCF-Crime		MSAD	
	AUC	AP	AUC	AP
CLIP [27]	88.02	32.72	87.45	71.09
TimeSformer [2]	86.54	28.72	86.22	72.18
VideoCLIP [42]	86.41	30.99	84.66	71.34
Alert-CLIP (Stage 1)	88.67	32.16	88.59	75.10
Alert-CLIP (Full)	**89.32**	**33.57**	**89.63**	**78.24**

Table 3. Zero-shot comparisons on multiple benchmarks and metrics. Best results are in **bold**.

Feature	UCF-AUC	XD-AP	UB-AUC	UB-AP
CLIP [27]	61.91	34.35	72.08	57.43
TimeSformer [2]	58.61	39.55	56.73	45.01
VideoCLIP [42]	72.13	54.36	72.34	55.07
SigLIP [47]	69.73	52.25	–	–
LLaVA-1.5-7B [18]	72.84	50.26	–	–
Alert-CLIP (Full)	**75.75**	**55.18**	**75.39**	**60.11**

Table 4. Open-vocabulary comparisons on UCF-Crime. Best results are in **bold**.

Method	AP	AP_b	AP_n	AUC	AUC_b	AUC_n
Anomize [14]	69.31	57.37	84.24	84.49	93.00	87.05
OVVAD [37]	66.53	57.10	76.03	86.40	93.84	**88.20**
Alert-CLIP (Full)	**70.12**	**57.91**	**84.92**	**86.14**	**94.04**	86.21

Table 5. Open-vocabulary VAD results on XD-Violence. All numbers are percentages. Best results are in **bold**.

Feature	Overall	Base	Novel
CLIP [27]	90.01	88.85	94.23
Alert-CLIP (Full)	**91.83**	**89.74**	**96.01**

Zero-shot generalization. We evaluate zero-shot VAD on UCF-Crime, XD-Violence, and UBnormal, where anomaly scores are computed directly from the similarity margin between normal and abnormal prompts (Table 3). Alert-CLIP consistently achieves the best performance, surpassing conventional backbones, as well as stronger vision–language models including SigLIP and LLaVA-1.5-7B. The gains are especially notable on UBnormal, suggesting that the learned

Table 6. Training strategy comparison on UCF-Crime. A checkmark indicates that the corresponding loss is activated.

Strategy	$\mathcal{L}_{\text{global}}$	$\mathcal{L}_{\text{region}}$	$\mathcal{L}_{\text{hard}}$	AUC
Joint training	✓	✓	✓	88.61
Two-stage (ours)	✓	–	–	**89.32**
	✓	✓	✓	

Table 7. Ablation on each alignment branch.

Configuration	Losses	AUC
CLIP baseline	–	88.02
Global only	$\mathcal{L}_{\text{global}}$	88.53
Global + Region	$\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{region}}$	88.85
Global + Region + Hard	$\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{region}} + \mathcal{L}_{\text{hard}}$	**89.32**

Table 8. Hard negatives versus standard in-batch negatives under the zero-shot setting. Best results are in **bold**.

Method	XD-Violence		UCF-Crime	
	AUC	AP	AUC	AP
CLIP [27]	63.33	34.35	61.91	11.16
Standard (in-batch neg.)	73.96	46.44	72.71	15.88
Hard neg.	**77.54**	**55.18**	**75.75**	**17.90**

semantic boundary remains robust under synthetic-to-real domain shifts.

Open-vocabulary anomaly detection. We further evaluate open-vocabulary VAD, where the model must interpret free-form anomaly descriptions beyond predefined labels (Tables 4 and 5). On UCF-Crime, Alert-CLIP achieves strong performance across overall, base, and novel categories, and remains competitive or superior to recent open-vocabulary VAD methods. While UCF-Crime is included in our training sources, XD-Violence remains entirely unseen during training and therefore provides a cleaner test of open-vocabulary generalization. The gains on XD-Violence indicate that Alert-CLIP learns a more transferable anomaly representation rather than relying only on seen concepts.

Discussion. Across weakly supervised, zero-shot, and open-vocabulary settings, Alert-CLIP consistently improves over CLIP and other backbones, showing that reshaping CLIP's embedding space through multi-level alignment leads to stronger anomaly understanding.

6.3. Ablation Studies

Effect of each alignment branch. To isolate the contribution of each learning signal, we ablate the three losses on UCF-Crime: the global alignment loss $\mathcal{L}_{\text{global}}$, the region alignment loss $\mathcal{L}_{\text{region}}$, and the semantic hard-negative loss $\mathcal{L}_{\text{hard}}$. As shown in Table 7, adding only $\mathcal{L}_{\text{global}}$ already improves over the frozen CLIP baseline, indicating that coarse-grained language supervision helps establish a clearer normal–abnormal separation. Incorporating $\mathcal{L}_{\text{region}}$ further boosts performance, highlighting the importance of spatially grounded supervision. The full model achieves the best result, showing that semantic hard negatives provide complementary fine-grained guidance beyond global and region alignment.

Hard negatives vs. standard negatives. We further compare the proposed hard-negative captions with standard in-batch negatives. As shown in Table 8, hard negatives consistently outperform standard contrastive negatives under the zero-shot setting, confirming that visually similar but semantically reversed captions are more effective for sharpening the normal–abnormal boundary. This explains why the semantic branch yields additional gains even after global and region alignment.

Two-stage vs. joint training. We also compare our two-stage training schedule with one-stage joint optimization that activates all losses from the beginning. As shown in Table 6, the two-stage curriculum achieves higher AUC, indicating that stabilizing the global semantic space before introducing fine-grained objectives leads to easier optimization and better generalization.

Table 9. Generalization beyond VAD on general benchmarks.

Benchmark	CLIP	Alert-CLIP	Gain
ImageNet-1K Top-1 [5]	68.34	**68.58**	+0.24
ImageNet-v2 Top-1 [28]	0.83	**0.89**	+0.06
Flickr30k I2T R@1 [44]	82.20	**89.90**	+7.70
Flickr30k T2I R@1 [44]	62.12	**74.46**	+12.34
COCO I2T R@1 [17]	51.68	**62.58**	+10.90
COCO T2I R@1 [17]	32.71	**43.81**	+11.10
BBox COCO Top-1 [17]	44.85	**45.74**	+0.89
BBox LVIS Top-1 [8]	13.76	**15.05**	+1.29

Generalization Beyond VAD. To assess whether the improvements on VAD compromise CLIP's original generalization ability, we further evaluate Alert-CLIP on several general benchmarks. As shown in Table 9, Alert-CLIP consistently matches or outperforms CLIP, indicating that the proposed training strategy enhances VAD performance without sacrificing general transferability. This observation suggests that multi-level alignment training helps the model capture finer-grained semantic cues in videos, which may also contribute to improved generalization.

6.4. Qualitative Analysis

Spatio-temporal behavior. Figure 4 compares CLIP and Alert-CLIP on representative UCF-Crime videos. On normal clips, CLIP often produces fluctuating anomaly scores, whereas Alert-CLIP remains low and stable. On abnormal

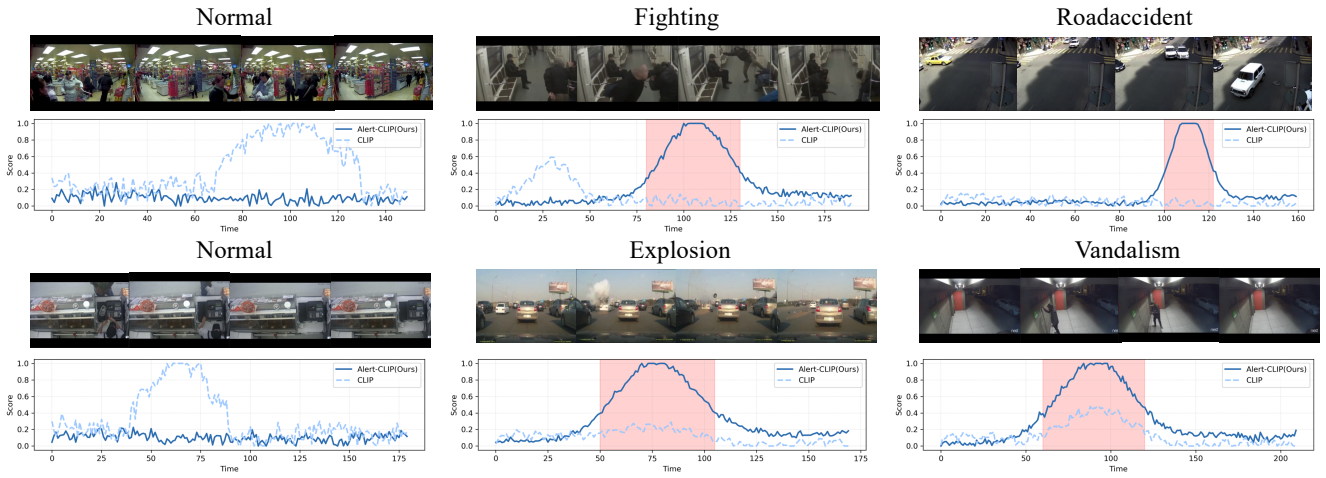


Figure 4. Temporal behavior of CLIP and Alert-CLIP on representative videos. Each case shows several keyframes and the corresponding anomaly-score curves over time, where the shaded region denotes the ground-truth abnormal interval. Compared with CLIP, Alert-CLIP produces more stable low scores on normal videos and clearer, better-aligned peaks on abnormal ones.

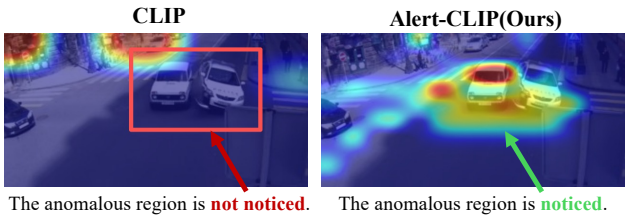


Figure 5. Spatial attention comparison between CLIP and Alert-CLIP on a road-accident example. CLIP mainly attends to surrounding context, whereas Alert-CLIP focuses on the collided vehicles and the true anomalous region.

clips, CLIP responses are weak and poorly aligned with the ground-truth interval, while Alert-CLIP yields clear peaks concentrated around the abnormal event. This shows that the proposed training improves both temporal discrimination and localization.

Attention to anomalous regions. Figure 5 visualizes spatial attention on a road-accident example. CLIP tends to emphasize large background regions and misses the actual collision area, whereas Alert-CLIP concentrates on the anomalous vehicles and suppresses irrelevant context. This behavior is consistent with the proposed region-level supervision and indicates stronger grounding of anomaly semantics in local visual cues.

7. Conclusion

We presented Alert-CLIP, a multi-granularity alignment framework that remedies CLIP's lack of fine-grained spatial sensitivity in video anomaly detection. By integrating global, region-level, and semantic hard-negative alignment, Alert-CLIP reshapes the underlying semantic space to better capture subtle abnormal cues. Experiments across weakly supervised, zero-shot, and open-vocabulary settings consistently show enhanced discriminative power and stronger cross-domain generalization.

Limitations and future work. Despite the clear improvements of Alert-CLIP, two limitations remain. First, dataset construction still relies on manual segmentation and bounding-box annotation, which is costly and limits scalability; future work may explore semi-automatic or automatic annotation pipelines assisted by pretrained vision–language models. Second, to satisfy CLIP's 77-token constraint, we shorten all captions, which reduces the richness of semantic supervision; a promising direction is to extend Alert-CLIP to longer-caption settings with multimodal models that support longer context windows.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2024YFE0200800, the National Natural Science Foundation of China under Grants 62406039, 62321001, 62471055, U23B2001, 62101064, 62171057, and 62201072, the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM107, the High-Quality Development Project of MIIT under Grant 2440STCZB2584, the Ministry of Education and China Mobile Joint Fund under Grants MCM20200202 and MCM20180101, the Fundamental Research Funds for the Central Universities under Grant 2024PTB-004, and the 2025 Education and Teaching Reform Project Funding at Beijing University of Posts and Telecommunications under Grant 2025YZ005.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *CVPR*, pages 20111–20121, 2022. 6

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 6

[3] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *CVPR*, pages 20392–20401. IEEE, 2023. 2

[4] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *CVPR*, pages 20392–20401, 2023. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 7

[6] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *CVPR*, pages 12742–12752. Computer Vision Foundation / IEEE, 2021. 2

[7] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *CVPR*, pages 12742–12752, 2021. 1

[8] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019. 7

[9] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *CVPR*, pages 733–742, 2016. 2

[10] Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In *ICML*, 2025. 1, 2

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 3

[12] B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *J. Imaging*, 4(2):36, 2018. 2

[13] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*. OpenReview.net, 2022. 1

[14] Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. In *CVPR*, pages 29203–29212. Computer Vision Foundation / IEEE, 2025. 6

[15] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pages 9694–9705, 2021. 3

[16] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 3

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, pages 740–755. Springer, 2014. 7

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26286–26296, 2024. 6

[19] Long Liu, Jianjun Li, Guang Li, Yunfeng Zhai, and Ming Zhang. Vadclip++: Dynamic vision-language model for weakly supervised video anomaly detection. *Digit. Signal Process.*, 168:105560, 2026. 2

[20] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *ECCV (47)*. Springer, 2024. 1

[21] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *CVPR*, pages 6536–6545. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[22] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *CVPR*, pages 6536–6545, 2018. 2

[23] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *CVPR*, pages 7076–7086. IEEE, 2022. 3

[24] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *ICCV*, pages 1273–1283, 2019. 2

[25] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *CVPR*, pages 12170–12179, 2020. 1

[26] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, pages 14360–14369, 2020. 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 7

[28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 7

[29] Nicolae-Catalin Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, and Mubarak Shah. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *CVPR*, pages 15984–15995, 2024. 1, 2

[30] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 1, 2, 3, 4, 6

[31] Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. 4

[32] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *ICCV*, pages 4955–4966, 2021. 2

[33] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Trans. Neural Networks Learn. Syst.*, 33 (6):2301–2312, 2022. 2

[34] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: clip-driven referring image segmentation. In *CVPR*, pages 11676–11685. IEEE, 2022. 3

[35] Jie Wu, Wei Zhang, Guanbin Li, Wenhao Wu, Xiao Tan, Yingying Li, Errui Ding, and Liang Lin. Weakly-supervised spatio-temporal anomaly detection in surveillance video. In *IJCAI*, pages 1172–1178, 2021. 2

[36] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, pages 322–339, 2020. 1, 6

[37] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *CVPR*, pages 18297–18307, 2024. 3, 6

[38] Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *ACM Multimedia*, pages 9301–9310, 2024. 1, 2, 3

[39] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *AAAI*, pages 6074–6082, 2024. 1, 2, 3, 6

[40] Peng Wu, Wanshun Su, Xiangteng He, Peng Wang, and Yanning Zhang. Varcmp: Adapting cross-modal pre-training models for video anomaly retrieval. In *AAAI*, pages 8423–8431, 2025. 1, 2

[41] Peng Wu, Wanshun Su, Guansong Pang, Yujia Sun, Qingsen Yan, Peng Wang, and Yanning Zhang. Avadclip: Audio-visual collaboration for robust video anomaly detection. *CoRR*, abs/2504.04495, 2025. 2

[42] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP (1)*, pages 6787–6800. Association for Computational Linguistics, 2021. 6

[43] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *CVPR*, pages 18899–18908, 2024. 1

[44] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014. 7

[45] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. CLAWS: clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *ECCV*, pages 358–376. Springer, 2020. 2

[46] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into CLIP latent space for video anomaly recognition. *Comput. Vis. Image Underst.*, 249:104163, 2024. 1

[47] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11941–11952. IEEE, 2023. 3, 6

[48] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024. 3

[49] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *CVPR*, pages 17385–17394, 2024. 1, 2

[50] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 3

[51] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *NeurIPS*, 2024. 2, 4, 6