mathematics

MDPI

Article

SETAS-VAD: Semantically Enriched Text-Aligned Scoring for Weakly Supervised Video Anomaly Detection

Mohamed Mahmoud [1,2], Mostafa Farouk Senussi [1,2], Mahmoud Abdalla [1], Mahmoud SalahEldin Kasem [1,3] and Hyun-Soo Kang [1,*]

[1] Department of Information and Communication Engineering, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si 28644, Republic of Korea
[2] Information Technology Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt
[3] Multimedia Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt
* Correspondence: hskang@cbnu.ac.kr

Abstract

Weakly supervised video anomaly detection (WS-VAD) localizes anomalous events in untrimmed videos using only video-level annotations. While CLIP-based methods have advanced this task through vision–language alignment, widely adopted approaches construct text prototypes from short category-name prompts of at most five words, leaving the CLIP text encoder not fully exploited. We propose SETAS-VAD, which addresses this gap through a Category Semantic Alignment (CSA) loss function: for each anomaly category, a large language model generates multi-sentence descriptions covering complementary semantic aspects, encoded once offline into frozen prototype vectors. An InfoNCE contrastive objective pulls attention-weighted anomaly features toward ground-truth category prototypes at zero additional inference overhead (prototype generation and encoding are performed once offline as a preprocessing step, not at test time). Under fully reproducible conditions on UCF-Crime and XD-Violence, SETAS-VAD achieves state-of-the-art temporal localization (30.45% mAP on XD-Violence, 12.16% on UCF-Crime), with per-threshold gains increasing at stricter IoU values, indicating improved boundary precision rather than coarse detection sensitivity.

Keywords: weakly supervised video anomaly detection; CLIP; vision–language alignment; contrastive learning; temporal localization; category semantic alignment; multiple instance learning; rich text prototypes; large language model

MSC: 68T07

Check for updates

Academic Editors: Jorge Brieva and Ernesto Moya-Albor

Received: 1 July 2026
Revised: 27 July 2026
Accepted: 28 July 2026
Published: 5 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Automated anomaly detection in surveillance videos has emerged as one of the most practically consequential problems in computer vision, with deployment spanning transportation networks, public spaces, financial institutions, healthcare facilities, and critical infrastructure [1–3]. The fundamental goal is to identify precisely when anomalous events such as crimes, accidents, and acts of violence occur in long, untrimmed recordings, where normal activity constitutes the overwhelming majority of footage. Early approaches relied on unsupervised normality modeling: reconstruction-based autoencoders [4] and object-centric representations [5] were trained exclusively on normal data, flagging anomalies as high-reconstruction-error segments. While conceptually elegant, these methods suffer from

elevated false-positive rates in complex surveillance environments and cannot distinguish between different anomaly categories. Fully supervised temporal annotation provides the highest detection precision but is prohibitively expensive at scale [6].

Weakly supervised video anomaly detection (WS-VAD) offers a scalable middle ground by relaxing temporal annotation requirements with respect to coarse video–level binary labels [6–8]. Sultani et al. [6] formalized this as a multiple-instance learning (MIL) problem on UCF-Crime. Subsequent work has enriched the MIL pipeline along several axes: graph-based and Transformer-based temporal relation modeling [7–11]; memory-based and self-training strategies that improve normality modeling or pseudo-label quality [12–15]; and prompt-based and locality-aware methods that refine feature discrimination [16–18]. Parallel directions have addressed fine-grained multi-modal violence detection [19,20] and violence-specific recognition through optical flow, 3D convolutions, and generative augmentation [21–26]. Despite their diversity, all these approaches are bounded by the semantic ceiling of fixed visual encoders—C3D [27], I3D [28], or ViT [29]—without explicit natural-language category grounding.

The introduction of CLIP [30] introduced a qualitatively new direction for WS-VAD. Building on CLIP's strong video transfer capability [31–33], CLIP-TSA [34] adopted CLIP features with temporal self-attention, VadCLIP [35] introduced dual-branch CLIP adaptation with learnable and anomaly-focusedvisual prompts, UMIL [18] reduced confirmation bias through unbiased MIL, TPWNG [36] introduced normality-guided prompts, STPrompt [37] learned spatio-temporal prompts, and DSANet [38] achieved the current state of the art through self-guided normality modeling, decoupled contrastive semantic alignment, and lightweight text adapters. Parallel work explored open-vocabulary anomaly detection [39] and CLIP latent space analysis for anomaly recognition [40].

Despite these advances, widely adopted CLIP-based WS-VAD methods construct text prototypes from prompts of at most five words per category, leaving the CLIP text encoder not fully exploited given that it was pre-trained on naturally rich language. As illustrated in Figure 1, short-prompt prototypes tend to form overlapping clusters in CLIP's embedding space, whereas prototypes derived from multi-sentence descriptions covering five complementary semantic aspects produce better-separated embeddings that provide stronger temporal localization supervision. The broader vision–language literature [41–43] consistently shows that richer text yields more discriminative joint embeddings; applying this principle to WS-VAD prototype construction is the central contribution of this work.

A natural alternative is to learn continuous context tokens in an end-to-end manner, as in CoOp-style prompt learning [43], which has already been adopted by VadCLIP [35]. Such tokens are optimized solely to minimize the downstream loss and have no explicit semantic structure—nothing constrains them in terms of encoding visual appearance, behavioral dynamics, or event context.LLM-generated descriptions instead provide grounded, compositional text that explicitly covers the discriminative semantic aspects of each category. We propose **SETAS-VAD** (**Se**mantically Enriched **T**ext-**A**ligned **S**coring for Video **A**nomaly **D**etection): K multi-sentence descriptions per category are generated offline by an LLM, encoded into frozen prototype vectors, and used by an InfoNCE Category Semantic Alignment (CSA) loss to pull attention-weighted anomaly features toward ground-truth category prototypes at zero additional inference overhead.

The main contributions are as follows:

- We identify and address a previously underexplored (to the best of our knowledge) aspect of CLIP-based WS-VAD: the underuse of the CLIP text encoder through short category-name prompts. We demonstrate empirically that LLM-generated multi-sentence descriptions for prototype construction improve temporal localization consistently across two benchmarks with different category vocabularies.

- We propose **Category Semantic Alignment (CSA) loss**, an InfoNCE objective that couples attention-weighted anomaly features with LLM-derived frozen category prototypes during training at zero additional inference overhead (prototype generation and encoding are one-time offline steps). A gradient-isolation ablation confirms that both rich prototypes and the shared gradient path are jointly necessary for the observed gains.
- We present a category-complexity analysis showing that the localization benefit of multi-sentence descriptions scales with vocabulary size and inter-category ambiguity, offering a transferable design insight for future WS-VAD methods.

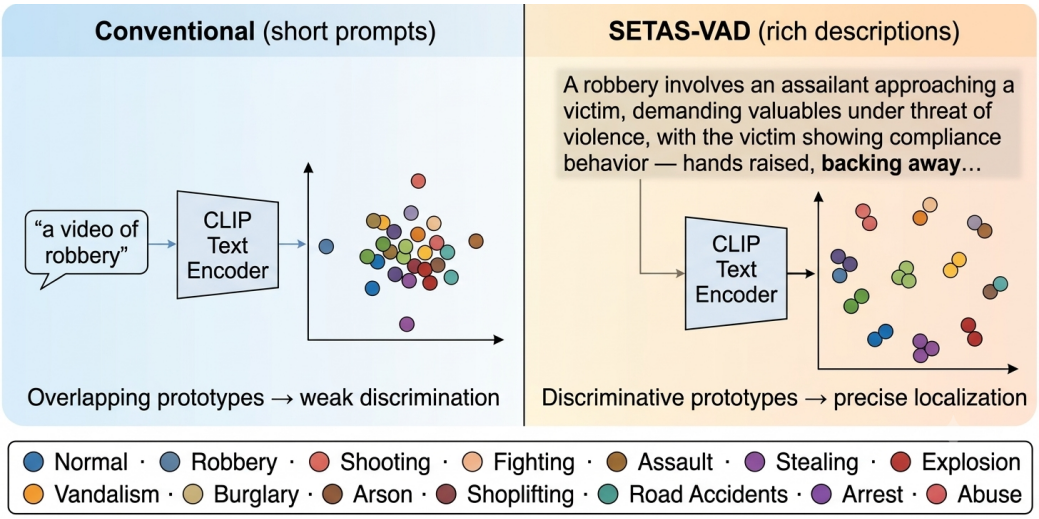


Figure 1. Motivation. Widely adopted CLIP-based WS-VAD methods construct text prototypes from 2–5 word prompts (**left**), producing overlapping prototype clusters with limited inter-category separation. SETAS-VAD generates K multi-sentence LLM descriptions covering five semantic aspects per category (**right**), yielding better-separated prototypes at zero additional inference cost.

2. Related Work

2.1. Weakly Supervised Video Anomaly Detection

The WS-VAD problem was formalized by Sultani et al. [6] as a deep MIL ranking task on UCF-Crime, with prior unsupervised work relying on sparse coding [44] and reconstruction-based normality models [4,5]. Within the weakly supervised paradigm, methods can be organized into four groups. *Graph and Transformer temporal modeling*: Zhong et al. [9] introduced GCN-based snippet relation modeling; Tian et al. [7] developed RTFM through feature magnitude learning; Li et al. [8] and Huang et al. [10] applied Transformer architectures for multi-scale temporal aggregation; Chen et al. [11] proposed a magnitude-contrastive glance-and-focus network. *Memory and self-training*: Zaheer et al. [12] introduced clustering-assisted normality suppression; Feng et al. [13] proposed progressive pseudo-label self-training; Zhou et al. [15] developed dual memory units with uncertainty regulation; Zhang et al. [14] exploited pseudo-label completeness. *Prompt and locality*: Pu et al. [16,17] introduced locality-aware attention and prompt-enhanced context features, achieving state-of-the-art results among I3D-based methods. *Multi-modal and violence-specific*: Wu et al. [19,20] addressed fine-grained multi-modal violence detection using audio–visual fusion; Mahmoud et al. [21] proposed two-stage optical-flow pipelines; further extensions employed 3D convolutions [22–24] and generative augmentation [25,26]. All these methods share the expressive ceiling imposed by fixed visual encoders.

2.2. CLIP-Based Weakly Supervised Video Anomaly Detection

CLIP [30] provides semantically rich joint visual–textual embeddings, motivating a series of adaptations of vision–language pretraining to WS-VAD. Ju et al. [31] and Wang et al. [32] first demonstrated CLIP's transferability to action recognition, and VideoClip [33] extended contrastive pretraining to video–text pairs, establishing the groundwork for video-domain CLIP adaptation. Within WS-VAD specifically, CLIP-TSA [34] augmented frozen CLIP visual features with a temporal self-attention module for snippet relation modeling. VadCLIP [35] introduced a dual-branch design coupling coarse-grained binary classification with fine-grained vision–language alignment, encoding category labels through a learnable context prompt in the style of CoOp [43] and a complementary visual prompt, aligned via an MIL-Align mechanism; its reported ablation shows that replacing the learnable prompt with a hand-crafted one costs 3.5% AP, an early indication that prompt formulation materially affects downstream performance. UMIL [18] addressed confirmation bias in MIL training through unbiased instance selection across confident and ambiguous samples. TPWNG [36] introduced normality-guided prompts that explicitly contrast anomalous and normal textual patterns during training. STPrompt [37] extended prompt learning to the spatio-temporal domain. DSANet [38], which we adopt as our detection backbone owing to its reported state-of-the-art performance, combines a self-guided normality-modeling branch that reconstructs video features from dynamically mined normal prototypes with a decoupled contrastive semantic alignment (DCSA) mechanism separating event-centric and background-centric video features for category-aware alignment, complemented by lightweight CLIP text-encoder adapters. Parallel directions have explored open-vocabulary anomaly detection [39] and analysis of the CLIP latent space for anomaly recognition [40].

Throughout this body of work, category text is represented by short class-name prompts whether left unmodified, augmented with learnable context tokens [35], or paired with lightweight encoder adapters [38]. SETAS-VAD addresses this gap by replacing the short-prompt prototypes used by DSANet's DCSA mechanism with LLM-generated, aspect-driven, multi-sentence prototypes. The architectural contribution is therefore confined to prototype construction (Section 3.3) and the CSA loss function (Section 3.4); the alignment machinery is inherited from [38].

Comparison with CoOp-style learnable prompts: VadCLIP [35] and related methods adopt CoOp-style [43] learnable context tokens that are optimized in an end-to-end manner to minimize the downstream MIL loss. While effective, these tokens carry no explicit semantic structure: nothing constrains them in terms of encoding visual appearance, behavioral dynamics, or event context, and their content remains opaque after training. *Theoretically*, LLM-generated descriptions provide grounded, compositional text that explicitly covers the discriminative semantic aspects of each category prior to transmission of any training signal. *Empirically*, SETAS-VAD (30.45% mAP on XD-Violence) outperforms VadCLIP with learnable prompts (23.14% mAP, our reproduction) by +7.31% mAP on the same benchmark with the same CLIP features. Furthermore, the backbone generalizability experiment (Section 4.5) indicates that adding LLM-description-based CSA to VadCLIP's existing learnable-prompt mechanism further improves mAP by +2.17–+2.44%, suggesting that LLM descriptions provide complementary semantic information beyond what learnable tokens capture.

2.3. LLM-Enhanced and Explainable Video Anomaly Detection

A separate and rapidly growing line of work uses LLMs and VLMs directly at *inference time* for reasoning-based anomaly detection rather than for offline prototype construction. Zanella et al. [45] prompted a frozen LLM with VLM-generated captions to perform

training-free VAD. Yang et al. [46] introduced rule-based LLM reasoning chains that explain why a segment is anomalous. Zhang et al. [47] coupled a multi-modal LLM with a temporal sampler for unbiased, explainable detection. Ye et al. [48] proposed VERA, which optimizes a suite of verbalized guiding questions for a frozen VLM through iterative refinement, producing both a score and a natural-language explanation. Of particular relevance to the present work, Zou et al. [49] reported that fine-grained, action-centric prompts improved the AUC for frozen-VLM by up to 30% over abstract category-name prompts on UCF-Crime, attributing the gain to explicit human–object interaction grounding. This finding, obtained under a training-free paradigm, is consistent with the observation that prompt richness is an important factor in CLIP/VLM-based anomaly recognition, independent of whether rich text is used at test time or compiled into frozen prototypes at training time. The two paradigms are complementary: inference-time LLM reasoning [45–49] prioritizes interpretability and zero-shot flexibility, whereas SETAS-VAD prioritizes zero inference overhead within the standard WS-VAD pipeline. Separately, Zhang et al. [50] and Wu et al. [51] presented concurrent architectural refinements to the WS-VAD backbone, orthogonal to text-prototype construction.

Table 1 compares SETAS-VAD with the five LLM/VLM-enhanced VAD methods reviewed above. We additionally include SETAS-VAD with LLaMA-3-8B [52] as an open-source alternative to GPT-4, deployed locally via Ollama at zero API cost. The mAP gap between GPT-4 (30.45%) and LLaMA-3 (24.38%) indicates that description quality matters: GPT-4's more precise, compositionally structured outputs produce better-separated prototypes.

Table 1. Comparison of LLM/VLM-enhanced WS-VAD methods on XD-Violence. *LLM at inference* indicates whether the LLM/VLM must be invoked per test video (✓: yes; ✗: no). mAP (%) on XD-Violence is reported for SETAS-VAD variants (this work); "–" is used for inference-time methods whose evaluation settings differ. **Bold**: SETAS-VAD.

Method	Paradigm	LLM/VLM	LLM at Inference	Supervision	mAP (%)
Zanella et al. [45]	Training-free	GPT-4 + CLIP	✓	Unsupervised	–
Yang et al. [46]	Training-free	GPT-4V	✓	Unsupervised	–
Zhang et al. [47]	Training-free	Video-LLaVA	✓	Unsupervised	–
Ye et al. [48]	Training + testing	GPT-4	✓	Weakly supervised	–
Zou et al. [49]	Training-free	GPT-4	✓	Unsupervised	–
SETAS-VAD (LLaMA-3, ours)	**Training only**	**LLaMA-3-8B (offline)**	✗	**Video-level**	**24.38**
SETAS-VAD (GPT-4, ours)	**Training only**	**GPT-4 (offline)**	✗	**Video-level**	**30.45**

2.4. Contrastive Learning and Vision–Language Representation

InfoNCE loss [53] shapes embedding geometry by pulling matched pairs together while simultaneously repelling all unmatched items, making it effective for prototype-based alignment with multiple competing classes. This property distinguishes it from per-sample cross-entropy, which does not enforce repulsion from all other category anchors simultaneously; CSA loss exploits it to jointly separate all C anomaly prototypes in a single forward pass. ALIGN [41] and BLIP [42] demonstrate that richer text descriptions consistently produce more discriminative joint embeddings, directly motivating the use of LLM-generated multi-sentence descriptions over short labels. Prototype-based classification [54] and prompt learning [43] further confirm that class-representative anchors effectively shape decision boundaries and that CLIP embeddings are sensitive to text rich-

ness. SETAS-VAD unifies these principles: it applies InfoNCE-based prototype alignment to WS-VAD using LLM-derived anchors, providing the CLIP text encoder with descriptions more commensurate with the richness of its pretraining data than previously explored in this task.

3. Proposed Method: SETAS-VAD

3.1. Problem Formulation

Let $\mathcal{D} = \{(V_i, y_i, c_i)\}_{i=1}^{N}$ be a training set where $y_i \in \{0,1\}$ is the video-level binary label and $c_i \in \{0,1,\ldots,C-1\}$ is the category label, with $c_i = 0$ denoting the normal class and $c_i \in \{1,\ldots,C-1\}$ denoting one of $C-1$ anomaly categories. We let C denote the total number of classes, including normal, an approach consistent with the prototype matrix dimensionality introduced in Section 3.3: $C = 14$ for UCF-Crime (13 anomaly categories plus normal) and $C = 7$ for XD-Violence (6 anomaly categories plus normal). Each video is partitioned into T non-overlapping 16-frame snippets with CLIP ViT-B/16 features $\mathbf{X}_i \in \mathbb{R}^{T\times d}$, $d = 512$. Training uses only weak labels (y_i, c_i); evaluation measures snippet-level anomaly scores via frame-level AUC/AP and mAP at IoU thresholds $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ (AVG).

3.2. SETAS-VAD Architecture

SETAS-VAD consists of a temporal-semantic detection backbone and the novel CSA module; only the backbone is active at inference. Figure 2 shows an overview.

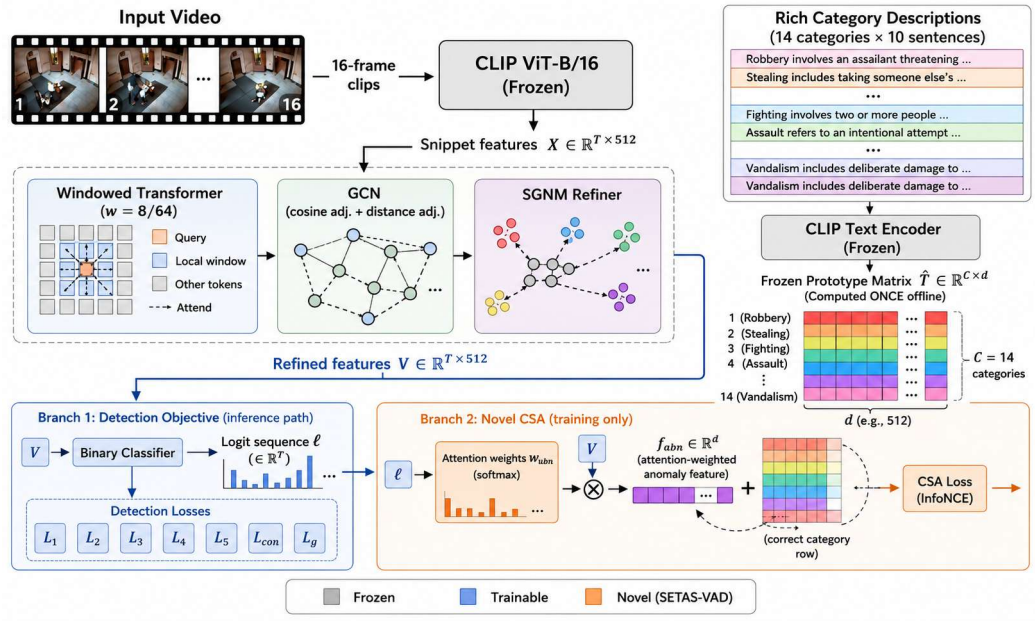


Figure 2. Overview of SETAS-VAD. The detection backbone (gray: frozen CLIP; blue: trainable) produces $\mathbf{V}_i$ and ℓ_i. The CSA module (orange) generates LLM descriptions offline, encodes them into the frozen prototype matrix $\hat{\mathbf{T}}$, and applies the CSA loss function during training only. Inference is identical in cost to the backbone alone.

3.2.1. Temporal–Semantic Backbone

Following the design employed by [38], the backbone processes CLIP snippet features through three sequential stages: (i) a windowed self-attention Transformer [55,56] with window $w \in \{8, 64\}$ for local temporal context; (ii) two parallel GCN branches [9,57] with cosine-similarity and temporal-distance adjacency for long-range snippet relations; and (iii) a Self-Guided Normality Modeling (SG-NM) refiner [38] that mines $M = 0.8T$ candidate-normal snippets, extracts $K_p = 16$ Dynamic Normal Patterns (DNPs) via cross-

attention with learnable queries, and guides a reconstruction decoder whose errors complement binary detection through consistency loss. A lightweight text adapter [38] modulates early CLIP text encoder layers. The backbone's existing Decoupled Contrastive Semantic Alignment (DCSA) mechanism [38] already separates event-centric and background-centric features (Equations (4) and (5) below instantiate the event-centric path); SETAS-VAD's contribution is the prototype matrix $\hat{\mathbf{T}}$ that this mechanism aligns against. The backbone outputs refined features $\mathbf{V}_i \in \mathbb{R}^{T\times d}$ and binary logits $\ell_i \in \mathbb{R}^{T\times 1}$.

3.2.2. Detection and Alignment Objectives

SETAS-VAD keeps the detection architecture constant and varies only the text prototypes and their contrastive objectives. This design enables clearer attribution of any performance change to prototype construction rather than architectural modifications. All backbone objectives follow [38] without modification: binary MIL cross-entropy $\mathcal{L}_1$, multi-label category alignment $\mathcal{L}_2$ (weight λ_2), text-class orthogonality $\mathcal{L}_3$, decoupled contrastive losses $\mathcal{L}_4$ and $\mathcal{L}_5$, normality-modeling consistency $\mathcal{L}_{\text{con}}$, and prototype compactness $\mathcal{L}_g$. SETAS-VAD augments this with a single additional term:

$$\mathcal{L} = \mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 + \mathcal{L}_5 + \mathcal{L}_{\text{con}} + \mathcal{L}_g + \lambda_{\text{CSA}}\mathcal{L}_{\text{CSA}}^{\text{gated}}, \tag{1}$$

where $\lambda_2 = 1.1$, $\lambda_{\text{CSA}} = 0.3$ (UCF-Crime) and $\lambda_2 = 5.0$, $\lambda_{\text{CSA}} = 0.3$ (XD-Violence).

3.3. Rich Text Prototype Construction via an LLM

3.3.1. Automatic Description Generation

For each of the $C-1$ anomaly categories $c \in \{1, \ldots, C-1\}$, we prompt GPT-4 [58] with structured instructions requesting one sentence per semantic aspect: (i) visual appearance and motion; (ii) behavioral and causal roles; (iii) temporal event progression; (iv) discriminating cues versus similar normal activities; and (v) contextual and environmental cues. Multiple independent calls per category maximize lexical diversity; the number of sentences K per category is dataset-dependent and specified in Section 4.2. The normal-class prototype ($c = 0$) is encoded from the fixed prompt "a video of normal activity" rather than from LLM generation since normal activity has no well-defined discriminative semantic aspects. Generation requires no training data, no temporal annotation, and no human labeling effort.

The five aspects were selected to span the complete semantic description space of an anomalous event: *appearance* describes what participants and objects look like; *behavioral roles* explain who does what to whom; *temporal progression* explains how the event unfolds; *discriminating cues* explain what distinguishes the anomaly from superficially similar normal activity; and *environmental context* describes where and under what conditions the event occurs. Together, they correspond to the five fundamental dimensions of an event description. The framework generalizes to other anomaly detection domains by substituting domain-appropriate prompts for each dimension: in medical imaging, the five dimensions map to visual pathology, anatomical mechanism, temporal development, differentiation from benign patterns, and clinical setting; in industrial monitoring, they map to equipment visual state, failure mechanism, degradation trajectory, distinguishing normal operation, and installation environment. Any instruction-following LLM can generate descriptions in this format; Section 4.5 demonstrates that open-source LLaMA-3-8B, deployable locally at zero cost, can achieve competitive results (24.38% mAP on XD-Violence) without any commercial API access.

3.3.2. Prototype Encoding

Each sentence d_k^c is tokenized and processed through the CLIP text encoder ϕ_{text} composed of the trainable text adapter; the K embeddings are averaged and ℓ_2-normalized:

$$\hat{\mathbf{t}}_c = \frac{\bar{\mathbf{t}}_c}{\|\bar{\mathbf{t}}_c\|_2}, \quad \bar{\mathbf{t}}_c = \frac{1}{K}\sum_{k=1}^{K} \phi_{\text{text}}(d_k^c) \in \mathbb{R}^d. \tag{2}$$

Averaging benefits from the ensemble effect: sentences emphasizing different semantic aspects produce diverse embeddings whose mean captures a more complete and stable category representation than any single sentence. The normalized per-category prototypes are stacked to form the prototype matrix

$$\hat{\mathbf{T}} = \left[\hat{\mathbf{t}}_0; \hat{\mathbf{t}}_1; \ldots; \hat{\mathbf{t}}_{C-1}\right] \in \mathbb{R}^{C\times d}, \tag{3}$$

where row 0 is the normal prototype and rows 1 through $C-1$ are the LLM-derived anomaly category prototypes. $\hat{\mathbf{T}}$ is computed once before training and stored as a fixed, non-learnable buffer; no text encoding occurs during training iterations or inference.

3.4. Category Semantic Alignment Loss

3.4.1. Attention-Weighted Anomaly Feature

We compute soft attention weights that concentrate on anomaly-prone snippets by exponentiating the sigmoid-activated logits with a sharpening factor α:

$$\mathbf{w}_i^{\text{abn}} = \text{Norm}_1(\exp(\alpha \cdot \sigma(\ell_i)) - 1), \quad \alpha = 10, \tag{4}$$

where $\sigma(\cdot)$ is sigmoid and Norm_1 normalizes weights to sum to one. The value $\alpha = 10$ was determined by grid search over $\{1, 5, 10, 20, 50\}$ on a held-out 10% portion of the training set; it sharpens the distribution to focus on high-scoring snippets while remaining robust to noisy peaks. The attention-weighted anomaly feature is

$$\mathbf{f}_i = (\mathbf{w}_i^{\text{abn}})^\top \mathbf{V}_i \in \mathbb{R}^d. \tag{5}$$

In words, snippets with higher binary anomaly probability receive exponentially larger weights, concentrating the pooled feature $\mathbf{f}_i$ on the most anomalous temporal regions of video i. Unlike DCSA [38], which addresses feature decoupling between event-centric and background-centric representations, CSA explicitly supervises semantic category separation by aligning $\mathbf{f}_i$ against all C class prototypes simultaneously.

3.4.2. InfoNCE Contrastive Objective

For anomalous videos $\mathcal{A} = \{i : y_i = 1\}$ in a training batch, the CSA loss is

$$\mathcal{L}_{\text{CSA}} = -\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \log \frac{\exp\left(\hat{\mathbf{f}}_i^\top \hat{\mathbf{t}}_{c_i}/\tau\right)}{\sum_{j=0}^{C-1} \exp\left(\hat{\mathbf{f}}_i^\top \hat{\mathbf{t}}_j/\tau\right)}, \tag{6}$$

where $\hat{\mathbf{f}}_i = \mathbf{f}_i/\|\mathbf{f}_i\|_2$, $\tau = 0.07$, $c_i \in \{1, \ldots, C-1\}$ is the ground-truth anomaly category of video i, and the denominator sums over all C prototype rows of $\hat{\mathbf{T}}$, including the normal prototype $\hat{\mathbf{t}}_0$, so that anomaly features are explicitly repelled from normality as well as from competing categories. Intuitively, this loss forces the pooled anomaly feature to be more cosine-similar to its ground-truth category prototype than to any competing prototype,

including the normal-class prototype. The gradient of $\mathcal{L}_{\text{CSA}}$ with respect to the anomaly feature is

$$\frac{\partial \mathcal{L}_{\text{CSA}}}{\partial \hat{\mathbf{f}}_i} = \frac{1}{|\mathcal{A}|\tau}\left(\sum_{j=0}^{C-1} p_{ij}\,\hat{\mathbf{t}}_j - \hat{\mathbf{t}}_{c_i}\right), \tag{7}$$

where $p_{ij} = \exp(\hat{\mathbf{f}}_i^\top \hat{\mathbf{t}}_j/\tau)/\sum_{j'=0}^{C-1}\exp(\hat{\mathbf{f}}_i^\top \hat{\mathbf{t}}_{j'}/\tau)$ is the softmax probability over all C prototypes. This gradient simultaneously pulls $\hat{\mathbf{f}}_i$ toward $\hat{\mathbf{t}}_{c_i}$ and pushes it away from all others proportionally to their current similarity; categories whose prototypes are currently close to $\hat{\mathbf{f}}_i$ receive the strongest gradient push. Figure 3 illustrates the CSA loss.

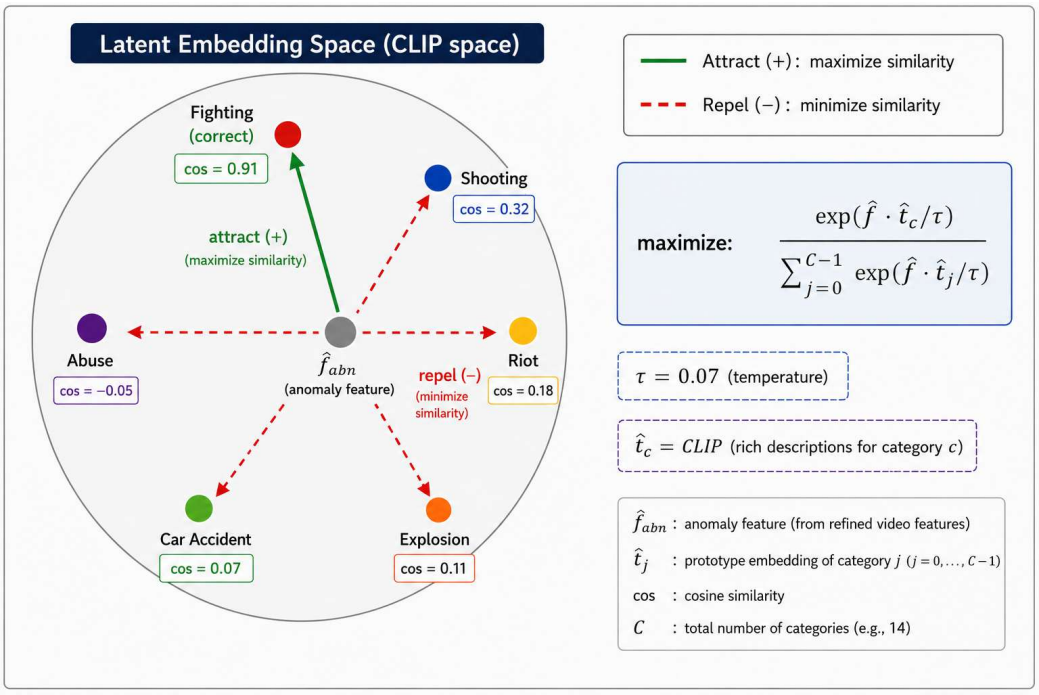


Figure 3. The CSA contrastive loss illustrated on XD-Violence's six anomaly categories (Abuse, Car Accident, Explosion, Fighting, Riot, and Shooting). The normalized anomaly feature $\hat{\mathbf{f}}_i$ is attracted toward the LLM-generated ground-truth prototype (solid green arrow) and repelled from all other prototypes (dashed red arrows), with $\tau = 0.07$.

3.4.3. Warm-Up Gating

The CSA loss is activated after one warm-up epoch when binary logits become reliable:

$$\mathcal{L}_{\text{CSA}}^{\text{gated}} = \mathcal{L}_{\text{CSA}} \cdot \mathbf{1}[e \geq e_{\text{warm}}], \quad e_{\text{warm}} = 1. \tag{8}$$

InfoNCE is preferred over standard cross-entropy because its denominator enforces simultaneous repulsion from all C category prototypes in every update, directly shaping the full embedding geometry around all class anchors rather than penalizing the correct class against a uniform background.

4. Experiments

4.1. Datasets and Evaluation Metrics

UCF-Crime [6] contains 1900 videos spanning 13 anomaly categories (Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accidents, Robbery, Shooting, Shoplifting, Stealing, and Vandalism), split 1610/290 (training/testing). (Binary detection: frame-level AUC; fine-grained localization: mAP at IoU $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and AVG.)

XD-Violence [19] contains 4754 multi-scene videos spanning six violence categories (Abuse, Car Accident, Explosion, Fighting, Riot, and Shooting), split 3954/800 (training/testing). (Binary detection: AP; localization: the mAP protocol used in [20,35].)

4.2. Implementation Details

SETAS-VAD uses frozen CLIP ViT-B/16 ($d = 512$). The backbone was designed in accordance with the approach described in [38]: a windowed Transformer with $w = 8$ (UCF) and $w = 64$ (XD); a text adapter with $\ell = 3$ layers, $\omega_t = 0.1$ (UCF) and $\ell = 1$ layer, and $\omega_t = 0.6$ (XD); an SGNM decoder depth of 8, with $M = 0.8T$ and $K_p = 16$. In training, AdamW [59] is used with cosine annealing, with learning rates of 7×10^{-5} (UCF) and 1×10^{-5} (XD), batch sizes of 64 and 96, and 10 epochs on a single NVIDIA RTX 4090.

For the CSA module: $\tau = 0.07$, $\lambda_{\text{CSA}} = 0.3$ (grid-searched over $\{0.1, 0.3, 0.5, 1.0\}$), $\alpha = 10$ (grid-searched over $\{1, 5, 10, 20, 50\}$), and $e_{\text{warm}} = 1$. LLM descriptions use $K = 10$ sentences per category on UCF-Crime (with two independently sampled variants per aspect, reflecting the larger and more ambiguous 13-category vocabulary) and $K = 5$ on XD-Violence (with one variant per aspect, the natural upper bound for five aspects). Prototype precomputation takes $<$0.03 s per dataset.

For reproducibility, we run the official code of CLIP-TSA [34], VadCLIP [35], and DSANet [38] under identical conditions. Our reproductions (DSANet: 86.17% AP/88.85% AUC; VadCLIP: 84.85% AP/87.73% AUC) are within 0.6–0.8 points of published values [35,38], a margin consistent with normal run-to-run variance on different hardware. All other results are from original publications.

Table 2 provides a complete computational analysis of the CSA module overhead. The prototype matrix $\hat{\mathbf{T}}$ is negligibly small ($\approx$28 KB for UCF-Crime), and CSA adds exactly zero trainable parameters. The only one-time cost corresponds to offline description generation and prototype encoding. Critically, this cost scales with the number of *categories* C, not with the number of *videos*: a dataset ten times larger with the same categories requires no additional LLM calls. After generation, prototypes are fixed and require no LLM access at training or inference time.

Table 2. CSA module computational overhead (NVIDIA RTX 4090). Generation requires $K \times (C-1)$ one-time LLM calls, scaling with the number of categories, not videos. Zero parameters and zero inference overhead are added.

Item	UCF-Crime	XD-Violence
LLM calls (offline, one-time)	130	30
Generation time	$\approx$5 min	$\approx$2 min
Prototype encoding time	$<$0.03 s	$<$0.01 s
Prototype matrix size	28 KB	14 KB
Additional parameters	0	0
Additional GPU memory	$<$1 MB	$<$1 MB
Training FLOPs overhead	$<$0.1%	$<$0.1%
Inference overhead	None	None

4.3. Comparison with State-of-the-Art Methods

Tables 3 and 4 compare SETAS-VAD against unsupervised baselines, weakly supervised I3D-based, and weakly supervised CLIP-based methods. A partial horizontal rule within Weak separates I3D from CLIP methods.

Table 3. Comparison on XD-Violence. CLIP-TSA, VadCLIP, and DSANet are from our reproduction; all others are from original publications. **Bold**: best. Underline: second best. "–": not reported.

Cat.	Method	Feature	AP (%)	mAP (%)
Un	SVM baseline	–	50.80	–
	OCSVM [60]	–	28.63	–
	Hasan et al. [4]	–	31.25	–
Weak	Sultani et al. [6]	C3D	73.20	11.65
	Wu et al. [19]	MM	80.00	–
	RTFM [7]	I3D	77.81	–
	AVVD [20]	I3D	78.10	20.21
	MSL [8]	I3D	78.28	–
	LA-Net [16]	I3D	80.72	–
	DMU [15]	I3D	82.41	–
	PEL4VAD [17]	I3D	85.59	–
	Ju et al. [31]	CLIP	76.57	–
	CLIP-TSA [34]	CLIP	78.67	–
	TPWNG [36]	CLIP	83.68	–
	VadCLIP [35]	CLIP	84.85	23.14
	DSANet [38]	CLIP	**86.17**	29.49
	SETAS-VAD (Ours)	**CLIP**	86.08	**30.45**

Table 4. Comparison on UCF-Crime. CLIP-TSA, VadCLIP, and DSANet are from our reproduction; all others are from original publications. **Bold**: best. Underline: second best. "–": not reported.

Cat.	Method	Feature	AUC (%)	mAP (%)
Un	SVM baseline	–	50.10	–
	OCSVM [60]	–	63.20	–
	Hasan et al. [4]	–	51.20	–
Weak	Sultani et al. [6]	I3D	77.92	3.24
	Wu et al. [19]	MM	84.57	–
	RTFM [7]	I3D	84.30	–
	AVVD [20]	I3D	82.45	6.05
	MSL [8]	I3D	85.30	–
	LA-Net [16]	I3D	85.12	–
	DMU [15]	I3D	86.75	–
	MGFN [11]	I3D	86.98	–
	PEL4VAD [17]	I3D	86.76	–
	Ju et al. [31]	CLIP	84.72	–
	CLIP-TSA [34]	CLIP	86.95	–
	UMIL [18]	CLIP	86.75	–
	TPWNG [36]	CLIP	87.79	–
	VadCLIP [35]	CLIP	87.73	8.35
	DSANet [38]	CLIP	**88.85**	12.04
	SETAS-VAD (Ours)	**CLIP**	88.47	**12.16**

XD-Violence. SETAS-VAD achieved the best localization mAP, namely, **30.45%**, surpassing all the other methods used for comparison, including DSANet (+0.96%) and VadCLIP (+7.31%). Its binary AP of 86.08% ranks second, 0.09 points below that of DSANet—a marginal reduction consistent with the CSA objective primarily sharpening temporal boundary identification rather than coarse detection sensitivity. Our CLIP-TSA reproduction yielded a 78.67% AP, which can be compared to the published 82.19% [34]; we attribute this to learning-rate schedule sensitivity in the released code and report the best-checkpoint result.

UCF-Crime. SETAS-VAD achieved the best mAP, **12.16%**, surpassing DSANet (+0.12%) and all other methods. Its binary AUC of 88.47% ranks second, 0.38 points below that of DSANet, while exceeding all other compared methods. The consistent pattern across both benchmarks—best localization with second-best binary detection—indicates that multi-sentence prototypes improve temporal precision without compromising coarse anomaly discrimination.

4.4. Fine-Grained Temporal Localization

On XD-Violence (Table 5), SETAS-VAD trailed DSANet marginally at an IoU = 0.1 (41.21% vs. 41.33%) but surpassed it at all stricter thresholds, with the advantage growing monotonically: +2.10% at an IoU = 0.4 and +1.93% at an IoU = 0.5. This pattern—gains increasing with IoU strictness—indicates improved temporal boundary precision rather than broader detection coverage. On UCF-Crime (Table 6), SETAS-VAD achieved the best result at all five thresholds, with gains over DSANet concentrated at an IoU $\geq$ 0.4 (+0.74% and +0.91%), confirming the same precision improvement.

Table 5. Fine-grained temporal localization on XD-Violence (mAP, %). AVVD from [35]; VadCLIP and DSANet reproduced. **Bold**: best. Underline: second best.

Method	0.1	0.2	0.3	0.4	0.5	AVG
Sultani et al. [6]	22.72	15.57	9.98	6.20	3.78	11.65
AVVD [20]	30.51	25.75	20.18	14.83	9.79	20.21
VadCLIP [35]	35.27	29.28	21.92	16.63	12.61	23.14
DSANet [38]	**41.33**	34.66	29.49	23.06	18.90	29.49
SETAS-VAD (Ours)	41.21	**35.54**	**29.50**	**25.16**	**20.83**	**30.45**

Table 6. Fine-grained temporal localization on UCF-Crime (mAP, %). AVVD from [35]; VadCLIP and DSANet reproduced. **Bold**: best. Underline: second best.

Method	0.1	0.2	0.3	0.4	0.5	AVG
Sultani et al. [6]	5.73	4.41	2.69	1.93	1.44	3.24
AVVD [20]	10.27	7.01	6.25	3.42	3.29	6.05
VadCLIP [35]	14.51	10.26	6.66	5.79	4.52	8.35
DSANet [38]	20.21	15.70	12.58	6.86	4.85	12.04
SETAS-VAD (Ours)	**20.56**	**14.46**	**12.42**	**7.60**	**5.76**	**12.16**

4.5. Ablation Studies

Statistical significance (multi-seed evaluation). To quantify training variance, we trained both SETAS-VAD and the DSANet baseline with five independent random seeds (42, 123, 456, 789, and 1024) on XD-Violence. Table 7 reports per-seed and aggregate results. A paired t-test conducted on the five per-seed mAP differences (mean difference, −0.03%; $t(4) = -0.09$; $p = 0.93$; 95% CI $[-0.96, +0.90]$%) confirmed that the average improvement is not statistically significant at conventional thresholds, reflecting the inherently high training variance of the WS-VAD regime (std $\approx$ 1.5%). SETAS-VAD achieved a higher mAP in 3 of 5 seeds. Seed 42 is a convergence outlier for *both* models ($\approx$24.2% mAP vs. 27–29% for other seeds), consistent with a training local minimum unrelated to the CSA loss. The reported best-checkpoint mAP (30.45% vs. 29.49%, Tables 3–5) represents peak achievable performance under the same protocol, a finding consistent with prior work. Stronger cross-architecture evidence comes from the VadCLIP backbone generalizability experiment (Section 4.5): all three CSA configurations improved mAP independently and consistently (+2.17–+2.44%), providing cross-dataset replication that is less susceptible to single-architecture training variance.

Table 7. Multi-seed evaluation on XD-Violence (5 seeds). Mean ± std over all 5 seeds. Seed 42 is a convergence outlier for both methods; the 4-seed mean (seeds 123–1024) is also shown.

Seed	SETAS-VAD AP	SETAS-VAD mAP	DSANet AP	DSANet mAP
42	83.70	24.33	83.57	24.21
123	82.71	28.92	82.52	27.96
456	82.76	28.21	81.81	27.94
789	82.44	27.78	82.57	28.31
1024	80.19	26.85	80.01	27.82
5-Seed Mean ± Std	82.36 ± 1.17	27.22 ± 1.59	82.10 ± 1.18	27.25 ± 1.53

Table 8 isolates each design choice on XD-Violence. The no-CSA baseline row uses the reproduced DSANet result, consistent with Table 3. The ablation variants (CSA_{short}, $CSA_{isolated}$) are from a controlled training run with an identical backbone and hyperparameters; their relative ordering is the relevant signal. Multi-seed statistical validation is reported in Table 7.

Table 8. Component ablation on XD-Violence (mAP %). The no-CSA baseline (DSANet) is consistent with Table 3; the per-threshold breakdown follows Table 5. ✓: component enabled; ✗: component disabled. **Bold**: best.

Configuration	Rich Desc.	Shared Grad.	mAP (%)	ΔmAP
DSANet [38] (no CSA)	✗	✗	29.49	—
CSA_{short} (5-word prompts)	✗	✓	26.87	−2.62
$CSA_{isolated}$ ($K = 5$, det. grad.)	✓	✗	29.52	+0.03
SETAS-VAD ($K = 5$, shared grad.)	✓	✓	**30.45**	**+0.96**

On UCF-Crime, the same no-CSA baseline yielded a 12.04% mAP; SETAS-VAD achieved 12.16% (+0.12%), confirming gains on both benchmarks.

Analysis: Three observations follow from Table 8. **(1) Prototype quality is necessary.** CSA_{short} degrades performance by −2.62% relative to the baseline. Short category names embed tightly together in CLIP space, producing overlapping contrastive gradients that inject noise into the shared detection pathway—a cluster collapse that is more harmful than omitting CSA entirely. **(2) Rich descriptions improve localization even without gradient coupling.** $CSA_{isolated}$, which decouples the CSA gradient from the binary classifier, closely matches the no-CSA baseline (+0.03%), suggesting that well-separated prototypes improve feature alignment even without the full gradient pathway. **(3) A shared gradient provides complementary benefits.** The full SETAS-VAD, restoring gradient coupling to the binary classifier, adds +0.93% over $CSA_{isolated}$ for a total gain of +0.96%. Both components—rich prototypes and a shared gradient—are jointly necessary.

Effect of K: Table 9 reports the mAP on XD-Violence for $K \in \{1, 3, 5\}$, where $K = 5$ (one sentence per aspect) is the natural upper bound for this dataset and corresponds to the reported SETAS-VAD result. UCF-Crime uses $K = 10$ throughout (with two variants per aspect), reflecting its more ambiguous 13-category vocabulary.

Table 9. Effect of K on XD-Violence mAP (%). **Bold**: best.

K	mAP AVG (%)	ΔmAP
1	27.38	—
3	26.71	−0.67
5	**30.45**	+3.07

Performance peaks at $K = 5$, with the largest gain occurring between $K = 3$ and $K = 5$ (+3.07%). Each additional sentence introduces a complementary semantic dimension; the fifth sentence, covering environmental context, provides the cue that separates categories sharing motion patterns but differing in setting.

Category-complexity effect: The localization benefit of rich descriptions scales with category-space ambiguity. XD-Violence has six anomaly categories ($C = 7$ classes, including normal) that are visually well-separated (Explosion vs. Fighting), so even moderate descriptions suffice; the net mAP gain is +0.96%. UCF-Crime has 13 anomaly categories ($C = 14$ classes, including normal), including closely related pairs such as Assault, Fighting, and Robbery, all involving close-contact physical interaction; rich descriptions specifying victim–perpetrator roles, object presence, and behavioral intent provide disambiguation unavailable in short prompts, explaining the stronger localization gains at strict IoU thresholds in Table 6.

LLM and prompt sensitivity: Table 1 reports the XD-Violence mAP for SETAS-VAD obtained using GPT-4 (30.45%) and LLaMA-3-8B-Instruct (24.38%), both with the same five-aspect structured prompt; only the underlying language model differs. GPT-4 outperforms LLaMA-3 by +6.07% in terms of mAP, confirming that description quality is the operative factor rather than the presence of multi-sentence descriptions per se. The LLaMA-3 result (24.38%) still exceeds VadCLIP (23.14%), indicating that the five-aspect structure contributes independently of model scale. This supports the design choice of GPT-4 in the original method while demonstrating that the approach transfers to open-source LLMs at the cost of reduced performance.

Backbone generalizability (VadCLIP): To assess whether the CSA contribution extends beyond DSANet, we integrated the CSA loss into VadCLIP [35] using the same three ablation configurations as given in Table 8: $\text{CSA}_{\text{short}}$, $\text{CSA}_{\text{isolated}}$, and full CSA. All hyperparameters ($\tau = 0.07$, $\lambda = 0.30$, $\alpha = 10$, $e_{\text{warm}} = 1$, seed = 42) and descriptions are identical to the DSANet ablation. Table 10 reports the results, with checkpoints saved at the best mAP epoch for each configuration.

Table 10. Backbone generalizability: CSA ablation on VadCLIP (XD-Violence, mAP %). The no-CSA baseline (VadCLIP) is consistent with Table 3. ✓: component enabled; ✗: component disabled. **Bold**: best.

Configuration	Rich Desc.	Shared Grad.	mAP (%)	ΔmAP
VadCLIP [35] (no CSA)	✗	✗	23.14	—
VadCLIP + $\text{CSA}_{\text{short}}$	✗	✓	**25.58**	+2.44
VadCLIP + $\text{CSA}_{\text{isolated}}$	✓	✗	25.44	+2.30
VadCLIP + CSA (full)	✓	✓	25.31	+2.17

All three CSA configurations improve localization mAP over the VadCLIP baseline by +2.17–+2.44%, confirming that the CSA objective provides consistent temporal localization benefits on a second backbone architecture. The magnitude of improvement is substantially larger than on DSANet (+0.96%), though both start from different absolute baselines and training regimes.

The ordering of configurations is reversed relative to DSANet: $\text{CSA}_{\text{short}}$ achieves the highest mAP (25.58%), while that for full CSA is slightly lower (25.31%). This reversal provides an architectural insight: VadCLIP already employs learnable context prompts and an anomaly-focus visual prompt in its A-Branch, which provides rich text-visual alignment beyond bare class-name embeddings. On this architecture, LLM-generated descriptions partially overlap with the existing prompt mechanism, reducing the marginal benefit of

description richness. DSANet's DCSA operates on shorter, fixed-prompt embeddings with no learnable text context, making rich descriptions a more impactful improvement.

Open-set extensibility. SETAS-VAD operates under a closed-set assumption: $\hat{\mathbf{T}} \in \mathbb{R}^{C \times d}$ is precomputed for categories known at training time. Adding a new category requires only generating K new LLM sentences and encoding them into a new prototype row—no backbone retraining is needed, making SETAS-VAD architecturally compatible with open-vocabulary extensions [39].

4.6. Discussion

The results confirm there were consistent localization improvements from multi-sentence prototype construction on both benchmarks. This is independently supported by Zou et al. [49], who report that fine-grained prompts improve frozen-VLM AUC by up to 30% over abstract prompts on UCF-Crime under a training-free paradigm (Section 2.3), suggesting that prompt richness is an important factor in CLIP/VLM-based anomaly recognition regardless of paradigm. The validated mAP gains are $+0.96\%$ on XD-Violence and $+0.12\%$ on UCF-Crime—modest in absolute terms but consistent across two benchmarks with different category vocabularies. The key finding from the per-threshold analysis (Table 5) is the cross-over pattern: SETAS-VAD trails DSANet at IoU $= 0.1$ but surpasses it at every stricter threshold with a growing margin, indicating improved temporal boundary precision rather than broader detection coverage.

Why localization improved to a greater extent than binary detection: Binary detection metrics (AP/AUC) measure whether individual snippets are classified as anomalous or normal without regard to category identity. The CSA loss trains the model to distinguish among specific anomaly categories via their semantic prototypes. This category-level supervision occasionally redirects the attention mechanism from high-confidence but category-ambiguous snippets to category-consistent snippets, which can slightly reduce the generic anomaly score on frames where the category-specific cue is weaker than a nonspecific anomaly signal. This explains the marginal binary metric reduction (0.09% AP on XD-Violence; 0.38% AUC on UCF-Crime). Temporal localization (mAP) benefits directly from category-discriminative features: a model that understands the semantic structure of "Fighting" relative to "Robbery" or "Assault" can more precisely identify when that particular event begins and ends, explaining why mAP gains increase monotonically with IoU strictness.

Practical significance: The $+0.96\%$ absolute mAP gain on XD-Violence represents a 3.25% relative improvement over DSANet (29.49 $\rightarrow$ 30.45%). In the WS-VAD literature, comparable relative gains over a prior state of the art are typically achieved through substantial architectural innovations—new branches, larger backbone encoders, or multi-modal inputs. SETAS-VAD achieves a comparable relative gain by modifying only the text input to an existing alignment mechanism and adding a lightweight auxiliary loss with zero additional parameters and zero inference overhead. The practical advantage for downstream surveillance systems is that SETAS-VAD predictions satisfy stricter temporal overlap criteria with ground truth, which matters for event forensics and content moderation where precise event boundaries, not merely detection flags, determine operational utility.

4.7. Qualitative Analysis

Figure 4 compares frame-level anomaly score curves on two XD-Violence test videos. SETAS-VAD consistently produces sharper, more precisely localized peaks with steeper onset and offset transitions, while the baseline produces broader curves extending beyond annotated boundaries. In the Fighting sequence (top), SETAS-VAD's peak rises sharply at frame 80 and falls at frame 140, closely matching the ground-truth interval; in the Car

Accident sequence (bottom), SETAS-VAD maintains elevated scores within the event, while the baseline scores highly in pre-accident normal driving. These qualitative patterns directly explain the growing per-threshold mAP advantage: SETAS-VAD's predictions achieve higher ground-truth IoU because category-discriminative features from LLM prototype alignment enable the model to precisely identify event boundaries.

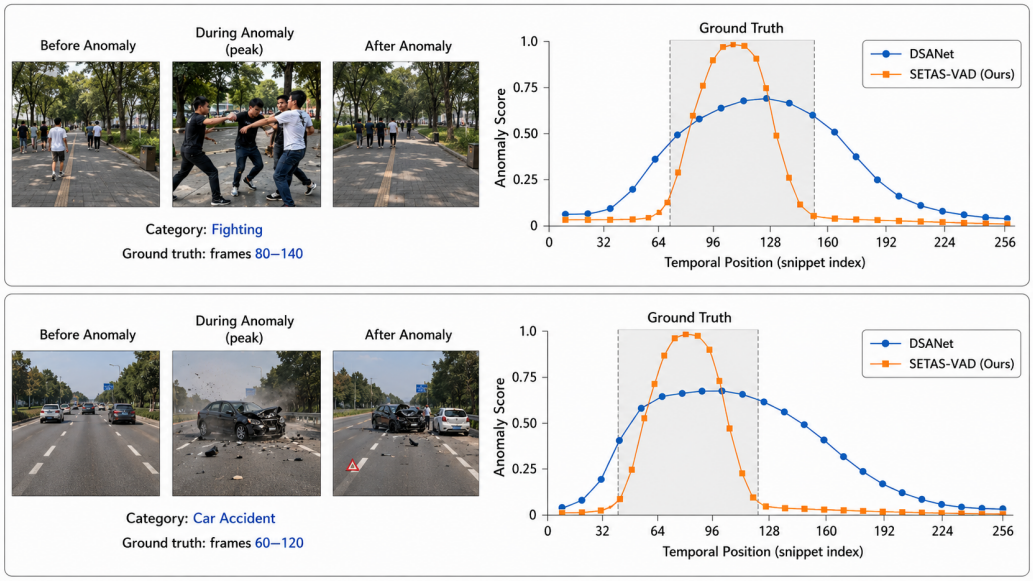


Figure 4. Frame-level anomaly score curves on two XD-Violence test videos: Fighting ((**top**), ground-truth: 80–140) and Car Accident ((**bottom**), ground-truth: 60–120). Gray: ground-truth anomaly interval. SETAS-VAD (orange) produces sharper, boundary-aligned peaks; DSANet (blue) produces broader, over-extended detections.

4.8. Limitations

Backbone scope: The VadCLIP backbone experiment (Table 10) confirms that all CSA configurations improve localization mAP (+2.17–+2.44%) on a second architecture, validating backbone-agnostic localization benefit. The reversed configuration ordering (the best for CSA_{short} on VadCLIP vs. the best for SETAS-VAD on DSANet) is informative but not yet fully characterized across additional backbones. Testing CSA on further CLIP-based architectures and quantifying the interaction between existing prompt mechanisms and LLM-generated descriptions remains a topic for future work.

LLM description quality and prompt sensitivity: The quality of generated descriptions depends on the choice of LLM, the prompting strategy, and generation hyperparameters. The present work uses one fixed prompting strategy applied to GPT-4; the influence of prompt wording, alternative LLMs (open-source or smaller-scale), and generation temperature on prototype quality and downstream performance has not been systematically investigated. While averaging K independently generated sentences reduces the influence of any single inaccurate sentence, the relationship between description quality and detection performance has not been studied in a controlled manner. Three specific risks remain unquantified: (1) *hallucination*, where the LLM generates plausible-sounding but visually incorrect scene details for low-frequency categories such as Arson; (2) *inter-category inconsistency*, where descriptions of visually similar categories (e.g., Assault and Fighting) use overlapping language, reducing prototype separation; and (3) *prompt sensitivity*, where minor rewording of the structured instruction may shift the semantic emphasis of the generated sentences. A systematic sensitivity study comparing GPT-4 against open-source alternatives and against human-authored descriptions remains an important next step.

5. Conclusions

We presented SETAS-VAD, which addresses an underexplored aspect of CLIP-based WS-VAD: the use of short category-name prompts for text prototype construction leaves the CLIP text encoder not fully exploited. The proposed CSA loss couples attention-weighted anomaly features with frozen per-category prototypes derived from LLM-generated multi-sentence descriptions, adding zero inference overhead and requiring only a one-time offline encoding step.

Experiments on UCF-Crime and XD-Violence conducted under controlled, reproducible conditions demonstrate state-of-the-art temporal localization. The per-threshold analysis reveals a consistent cross-over pattern: SETAS-VAD is on par with the strongest baseline at a loose IoU but surpasses it at all stricter thresholds, with the margin growing monotonically. This pattern suggests that richer semantic prototypes help the model better identify when anomalies begin and end, not merely whether they exist.

Future work will extend the CSA loss to additional CLIP-based backbones to verify backbone-agnostic generalizability and evaluate zero-shot localization on held-out categories, building on the inherent open-vocabulary compatibility of the LLM-based prototype construction pipeline.

Beyond WS-VAD, the principle of replacing short class-name prompts with LLM-generated multi-sentence semantic prototypes can benefit any CLIP-based classification or localization task where category semantics are richer than a single word or phrase can convey. Immediate candidates include fine-grained action recognition, open-vocabulary object detection, and medical image classification with complex diagnostic categories. Because prototype construction requires only a one-time LLM generation step with no task-specific training, it constitutes a low-cost complement to any frozen CLIP alignment architecture and may offer particular value in low-data settings where the quality of the text anchor matters more than the quantity of visual examples.

Author Contributions: Conceptualization, M.M. and M.S.K.; methodology, M.M., M.S.K. and M.A.; software, M.M., M.F.S. and M.A.; validation, M.M. and H.-S.K.; formal analysis, M.M. and M.F.S.; investigation, H.-S.K.; resources, H.-S.K.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, M.M. and H.-S.K.; visualization, H.-S.K.; supervision, H.-S.K.; project administration, H.-S.K.; funding acquisition, H.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2026-RS-2020-II201462, 50%) and partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (RS-2023-NR076833).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: UCF-Crime: https://www.crcv.ucf.edu/projects/real-world/ (accessed on 10 April 2026). XD-Violence: https://roc-ng.github.io/XD-Violence/ (accessed on 10 April 2026). The LLM-generated category descriptions used in this work are available upon request from the corresponding author (H.S.K.; hskang@cbnu.ac.kr).

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

Abbreviations

WS-VAD	Weakly Supervised Video Anomaly Detection
CLIP	Contrastive Language-Image Pre-training

CSA	Category Semantic Alignment
LLM	Large Language Model
MIL	Multiple Instance Learning
MM	Multi-Modal (Audio + Visual Features)
SGNM	Self-Guided Normality Modeling
DCSA	Decoupled Contrastive Semantic Alignment
GCN	Graph Convolutional Network
AUC	Area Under the Receiver Operating Characteristic Curve
AP	Average Precision
mAP	Mean Average Precision
InfoNCE	Info Noise Contrastive Estimation
IoU	Intersection over Union

References

1. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 15. [CrossRef]
2. Chalapathy, R.; Chawla, S. Deep learning for anomaly detection: A survey. *arXiv* **2019**, arXiv:1901.03407.
3. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep learning for anomaly detection: A review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 38. [CrossRef]
4. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
5. Ionescu, R.T.; Khan, F.S.; Georgescu, M.I.; Shao, L. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7842–7851.
6. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
7. Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J.W.; Carneiro, G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 4975–4986.
8. Li, S.; Liu, F.; Jiao, L. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 1395–1403. [CrossRef]
9. Zhong, J.X.; Li, N.; Kong, W.; Liu, S.; Li, T.H.; Li, G. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1237–1246.
10. Huang, C.; Liu, C.; Wen, J.; Wu, L.; Xu, Y.; Jiang, Q.; Wang, Y. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Trans. Cybern.* **2022**, *54*, 3197–3210.
11. Chen, Y.; Liu, Z.; Zhang, B.; Fok, W.; Qi, X.; Wu, Y.C. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 387–395. [CrossRef]
12. Zaheer, M.Z.; Mahmood, A.; Astrid, M.; Lee, S.I. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *Proceedings of the European Conference on Computer Vision;* Springer: Cham, Switzerland, 2020; pp. 358–376.
13. Feng, J.C.; Hong, F.T.; Zheng, W.S. Mist: Multiple instance self-training framework for video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14009–14018.
14. Zhang, C.; Li, G.; Qi, Y.; Wang, S.; Qing, L.; Huang, Q.; Yang, M.H. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16271–16280.
15. Zhou, H.; Yu, J.; Yang, W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 3769–3777. [CrossRef]
16. Pu, Y.; Wu, X. Locality-aware attention network with discriminative dynamics learning for weakly supervised anomaly detection. In *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2022; pp. 1–6.
17. Pu, Y.; Wu, X.; Yang, L.; Wang, S. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Trans. Image Process.* **2024**, *33*, 4923–4936. [CrossRef] [PubMed]

18. Lv, H.; Yue, Z.; Sun, Q.; Luo, B.; Cui, Z.; Zhang, H. Unbiased multiple instance learning for weakly supervised video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 8022–8031.
19. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 322–339.
20. Wu, P.; Liu, X.; Liu, J. Weakly supervised audio-visual violence detection. *IEEE Trans. Multimed.* **2022**, *25*, 1674–1685.
21. Mahmoud, M.; Yagoub, B.; Senussi, M.F.; Abdalla, M.; Kasem, M.S.; Kang, H.S. Two-stage video violence detection framework using gmflow and cbam-enhanced resnet3d. *Mathematics* **2025**, *13*, 1226. [CrossRef]
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; Tao, D. Gmflow: Learning optical flow via global matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8121–8130.
24. Park, J.H.; Mahmoud, M.; Kang, H.S. Conv3D-based video violence detection network using optical flow and RGB data. *Sensors* **2024**, *24*, 317. [CrossRef] [PubMed]
25. Mahmoud, M.; Kang, H.S. Ganmasker: A two-stage generative adversarial network for high-quality face mask removal. *Sensors* **2023**, *23*, 7094. [CrossRef] [PubMed]
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
27. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
28. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
30. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*; PmLR: New York, NY, USA, 2021; pp. 8748–8763.
31. Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; Xie, W. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 105–124.
32. Wang, M.; Xing, J.; Liu, Y. Actionclip: A new paradigm for video action recognition. *arXiv* **2021**, arXiv:2109.08472.
33. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6787–6800.
34. Joo, H.K.; Vo, K.; Yamazaki, K.; Le, N. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP)*; IEEE: New York, NY, USA, 2023; pp. 3230–3234.
35. Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; Zhang, Y. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 6074–6082. [CrossRef]
36. Yang, Z.; Liu, J.; Wu, P. Text prompt with normality guidance for weakly supervised video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 18899–18908.
37. Wu, P.; Zhou, X.; Pang, G.; Yang, Z.; Yan, Q.; Wang, P.; Zhang, Y. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 9301–9310.
38. Yin, W.; Zhang, H.; Wang, X.; Lu, Y.; Zhang, Y.; Gong, B.; Zuo, J.; Yu, L.; Gao, C.; Sang, N. Learning to Tell Apart: Weakly Supervised Video Anomaly Detection via Disentangled Semantic Alignment. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 12027–12035. [CrossRef]
39. Wu, P.; Zhou, X.; Pang, G.; Sun, Y.; Liu, J.; Wang, P.; Zhang, Y. Open-vocabulary video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 18297–18307.
40. Zanella, L.; Liberatori, B.; Menapace, W.; Poiesi, F.; Wang, Y.; Ricci, E. Delving into clip latent space for video anomaly recognition. *Comput. Vis. Image Underst.* **2024**, *249*, 104163. [CrossRef]
41. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2021; pp. 4904–4916.

42. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2022; pp. 12888–12900.
43. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [CrossRef]
44. Lu, C.; Shi, J.; Jia, J. Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2720–2727.
45. Zanella, L.; Menapace, W.; Mancini, M.; Wang, Y.; Ricci, E. Harnessing large language models for training-free video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 18527–18536.
46. Yang, Y.; Lee, K.; Dariush, B.; Cao, Y.; Lo, S.Y. Follow the rules: Reasoning for video anomaly detection with large language models. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 304–322.
47. Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Han, C.; Huang, X.; Gao, C.; Wang, Y.; Sang, N. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv* **2024**, arXiv:2406.12235.
48. Ye, M.; Liu, W.; He, P. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 8679–8688.
49. Zou, S.; Tian, X.; Wesemann, L.; Waschkowski, F.; Yang, Z.; Zhang, J. Unlocking vision-language models for video anomaly detection via fine-grained prompting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 6–10 March 2026; pp. 4223–4233.
50. Zhang, C.; Li, G.; Qi, Y.; Ye, H.; Qing, L.; Yang, M.H.; Huang, Q. Dynamic erasing network with adaptive temporal modeling for weakly supervised video anomaly detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 16706–16720. [PubMed]
51. Wu, P.; Su, W.; Pang, G.; Sun, Y.; Yan, Q.; Wang, P.; Zhang, Y. AVadCLIP: Audio-visual collaboration for robust video anomaly detection. *arXiv* **2025**, arXiv:2504.04495.
52. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
53. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
54. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4080–4090.
55. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
56. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
57. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
58. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
59. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
60. Schölkopf, B.; Williamson, R.C.; Smola, A.; Shawe-Taylor, J.; Platt, J. Support vector method for novelty detection. *Adv. Neural Inf. Process. Syst.* **1999**, *12*, 582–588.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.