

Language Models in Video Anomaly Detection: A Dynamic Survey

Mumcu, Furkan; Bekit, Lokman; Jones, Michael J.; Cherian, Anoop; Yilmaz, Yasin

TR2026-143 October 01, 2026

Abstract

Video Anomaly Detection (VAD) plays a critical role in intelligent surveillance and public safety by identifying rare, unexpected, or contextually abnormal events in complex video environments. Conventional deep learning approaches have substantially advanced VAD, but they remain largely driven by visual representations and often face challenges in semantic generalization, interpretability, and adaptation to unseen anomaly types. Recent Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) are reshaping the field by enabling semantic reasoning, contextual understanding, language-guided detection, open-vocabulary recognition, and natural-language explanation. This survey provides a focused and systematic review of language-model-based VAD. We organize existing methods across major supervision and deployment paradigms, including fully supervised, unsupervised, semi-supervised, weakly supervised, training-free, instruction-tuned, and open-world/open-vocabulary VAD, while positioning them relative to key non-language-model foundations. We further review representative datasets and evaluation protocols, emphasizing their suitability and limitations for both anomaly detection and anomaly understanding. Beyond serving as a static snapshot of the literature, this paper is designed as a dynamic survey whose taxonomy, method coverage, datasets, and evaluation discussion can be maintained through versioned updates as new work emerges. This evolving resource is supported by the public website dynamicvadsurvey.github.io, which provides access to maintained releases, archived versions, update records, and interactive exploration of the surveyed literature.

IEEE Access 2026

Language Models in Video Anomaly Detection: A Dynamic Survey

FURKAN MUMCU¹, (Graduate Student Member, IEEE), LOKMAN BEKIT¹, (Graduate Student Member, IEEE), MICHAEL J. JONES², (Fellow, IEEE), ANOOP CHERIAN², (Member, IEEE), and YASIN YILMAZ¹, (Senior Member, IEEE)

¹University of South Florida, Tampa, FL 33620, USA (e-mail: furkan@usf.edu; lbekit@usf.edu; yasiny@usf.edu)

²Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA (e-mail: mjones@merl.com; cherian@merl.com)

Corresponding author: Furkan Mumcu (e-mail: furkan@usf.edu).

ABSTRACT Video Anomaly Detection (VAD) plays a critical role in intelligent surveillance and public safety by identifying rare, unexpected, or contextually abnormal events in complex video environments. Conventional deep learning approaches have substantially advanced VAD, but they remain largely driven by visual representations and often face challenges in semantic generalization, interpretability, and adaptation to unseen anomaly types. Recent Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) are reshaping the field by enabling semantic reasoning, contextual understanding, language-guided detection, open-vocabulary recognition, and natural-language explanation. This survey provides a focused and systematic review of language-model-based VAD. We organize existing methods across major supervision and deployment paradigms, including fully supervised, unsupervised, semi-supervised, weakly supervised, training-free, instruction-tuned, and open-world/open-vocabulary VAD, while positioning them relative to key non-language-model foundations. We further review representative datasets and evaluation protocols, emphasizing their suitability and limitations for both anomaly detection and anomaly understanding. Beyond serving as a static snapshot of the literature, this paper is designed as a dynamic survey whose taxonomy, method coverage, datasets, and evaluation discussion can be maintained through versioned updates as new work emerges. This evolving resource is supported by the public website dynamicvadsurvey.github.io, which provides access to maintained releases, archived versions, update records, and interactive exploration of the surveyed literature.

INDEX TERMS Anomaly detection, dynamic surveys, foundation models, instruction tuning, large language models, multimodal large language models, open-vocabulary recognition, video anomaly detection, video understanding, vision-language models.

I. INTRODUCTION

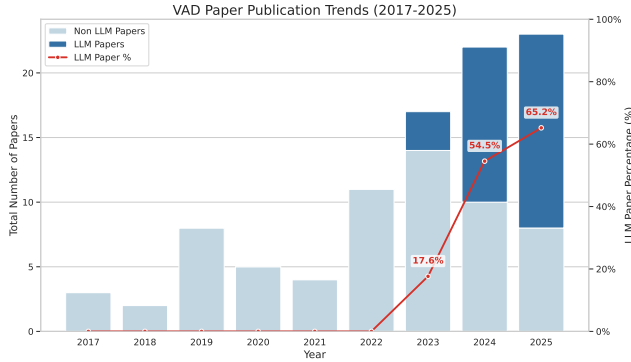
Video Anomaly Detection (VAD) has become a vital technology in intelligent surveillance and public safety, aiming to automatically identify behaviors or events that deviate from expected patterns. More broadly, anomaly detection has long been studied as the problem of identifying rare observations that deviate from expected data regularities [1], [2]. Anomalies can be viewed as *spatial*, *temporal*, or *semantic* deviations, such as an unauthorized object in a restricted area, a vehicle moving against traffic, or high risk behaviors such as armed conflict. For decades, research in VAD has sought to build automated systems that flag these rare and unpredictable events. Early approaches primarily relied on hand-crafted feature engineering, where researchers designed discriminative features tailored to specific scenes [3]–[15]. However, these methods often lacked the robustness required

for complex and dynamic real world environments.

The advent of deep learning marked a significant milestone, driving innovation in VAD methodologies and substantially improving performance. Deep neural networks (DNNs) moved the field away from manual feature design by learning powerful feature representations directly from data [16]–[19]. This era was characterized by several key paradigms. *Self-supervised* methods for semi-supervised VAD learned patterns of normal events by training models on tasks such as future frame prediction [18] or video reconstruction [20], then flagged deviations from these learned norms as anomalies. For *weakly supervised* scenarios, where only video level labels are available, Multiple Instance Learning (MIL) [19] became a dominant approach, enabling models to localize anomalous segments without requiring precise temporal annotations. Despite their success, these DNN based methods

TABLE 1. VAD Paper Trends by Conference. Cell format: Non-LLM / LLM (LLM %).

	2017	2018	2019	2020	2021	2022	2023	2024	2025
CVPR	1 / 0 (0%)	1 / 0 (0%)	3 / 0 (0%)	1 / 0 (0%)	2 / 0 (0%)	2 / 0 (0%)	5 / 3 (38%)	4 / 5 (56%)	3 / 4 (57%)
ICCV	1 / 0 (0%)	-	3 / 0 (0%)	-	2 / 0 (0%)	-	4 / 0 (0%)	-	2 / 1 (33%)
ECCV	-	0 / 0 (0%)	-	1 / 0 (0%)	-	4 / 0 (0%)	-	2 / 3 (60%)	-
WACV	1 / 0 (0%)	1 / 0 (0%)	2 / 0 (0%)	3 / 0 (0%)	0 / 0 (0%)	4 / 0 (0%)	5 / 0 (0%)	3 / 2 (40%)	2 / 2 (50%)
NeurIPS	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	1 / 0 (0%)	0 / 0 (0%)	1 / 2 (67%)	1 / 6 (86%)
ICML	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	0 / 0 (0%)	0 / 2 (100%)

**FIGURE 1.** VAD paper publication trends between 2017 and 2025 in big conferences including CVPR, ICCV, ECCV, WACV, NeurIPS, and ICML.

share a fundamental characteristic: they operate primarily within the *visual feature space*. This foundation leads to persistent challenges, including limited generalization to unseen scenarios and a lack of interpretability, since the models' decision making processes remain largely opaque. Moreover, performance gains on several widely used benchmarks have become incremental in recent years, underscoring the need for new methodological advances.

Recent advances in Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have increasingly influenced research on video anomaly detection (VAD). As summarized in Figure 1 and Table 1, this development is visible within the set of VAD papers surveyed from major vision and machine learning venues, including CVPR, ICCV, ECCV, WACV, NeurIPS, and ICML. The timing broadly coincides with the emergence of foundational vision-language models such as CLIP [21] in 2021, which helped motivate the use of language-aligned representations for visual understanding. After an initial adoption period, LM-based approaches accounted for 17.6% of the surveyed VAD papers in 2023, increasing to 54.5% in 2024 and 65.2% in 2025. The trend is also evident in broader machine learning venues, where VAD had historically received limited representation. In 2025, LM-based methods accounted for 86% of the surveyed VAD papers at NeurIPS and 100% at ICML. Together, these trends indicate growing interest in semantic, contextual, and reasoning-oriented approaches to anomaly detection.

This new phase of VAD research is characterized by a move from pixel level pattern recognition toward semantic

reasoning. In contrast to traditional deep neural networks, MLLMs can leverage large scale pretraining, world knowledge, and language supervision to interpret video content at a higher semantic level [21]–[26]. They can abstract raw visual inputs into semantic representations, analyze contextual relationships, and assess whether an observed action is anomalous relative to the scene context. This shift has not only improved detection performance in several recent works, but has also motivated new directions, including *Training Free VAD*, *Instruction Tuning VAD*, and *Open Vocabulary VAD*.

While Figure 1 and Table 1 illustrate the rapid rise of language models in VAD within top tier conferences, this trend reflects a much broader proliferation. In the past three years alone, over one hundred LM based VAD papers have been published. This growth motivates a comprehensive and systematic analysis of the field. Although several surveys have addressed Video Anomaly Detection (VAD), they often have specific limitations. The review by Ramachandra et al. [27], for instance, centers on semi-supervised VAD but is focused on single-scene scenarios and does not cover other settings. Similarly, Nayak et al. [28] provided a comprehensive look at deep learning for semi-supervised VAD but did not extend their analysis to other VAD paradigms. Other works, such as Tran et al. [29], examined weakly supervised methods but did not focus exclusively on VAD, since they also considered image anomaly detection. Although Liu et al. [34] proposed useful classification frameworks for semi supervised and weakly supervised tasks, their coverage predates many recent developments. More recently, Wu et al. [31] offered detailed insights across a range of VAD tasks, including unsupervised and open set methods. However, their treatment of MLLM and LLM based approaches was not in depth, providing only a high level overview and not capturing the impact of these newer techniques on the field.

Gao et al. [35] propose a unified perspective spanning DNN based and MLLM based approaches and are correspondingly broad in scope. While that work was among the first to analyze MLLM based methods at scale, its goal of covering the entire VAD landscape naturally limits the depth of discussion on architectural choices, prompt and instruction design, and the challenges unique to language model driven pipelines. In contrast, the present survey focuses specifically on the broader role of language in VAD, including vision-language representations, CLIP-based methods, textual su-

TABLE 2. Comparison of recent VAD survey papers in terms of scope, benchmark discussion, advanced LM-based topics, and dynamic survey design. ✓ indicates clear coverage, ~ indicates partial coverage, and ✗ indicates that the aspect is not a major focus.

Survey	Year	Focus	Scope		Analysis and Survey Design		
			LM/MLLM Focus	Multi-Paradigm Coverage	Benchmark Analysis	Advanced LM-VAD Analysis	Dynamic / Versioned
Ramachandra et al. [27]	2020	Single-scene semi-supervised VAD	✗	✗	✓	✗	✗
Nayak et al. [28]	2021	Deep learning-based VAD methods	✗	✓	✓	✗	✗
Tran et al. [29]	2022	Image and video anomaly analysis	✗	~	✓	✗	✗
Liu et al. [30]	2024	Generalized VAD taxonomy	✗	✓	✓	~	✗
Wu et al. [31]	2024	VAD with initial large-model topics	~	✓	✓	~	✗
Abdalla et al. [32]	2024	Decade review of deep VAD	~	✓	✓	~	✗
Ding et al. [33]	2024	LLMs and VLMs in VAD	✓	~	~	~	✗
Liu et al. [34]	2025	Networking systems for VAD	~	✓	~	~	✗
Gao et al. [35]	2025	Evolution of VAD from DNNs to LLMs	✓	✓	✓	~	✗
Ours	2026	Language-model-based VAD with dynamic survey design	✓	✓	✓	✓	✓

pervision and prompting, LLM-assisted knowledge transfer, training-free reasoning, instruction tuning, open-world and open-vocabulary detection, anomaly explanation, question answering, grounding, and retrieval. This narrower scope enables substantially deeper coverage of LM-based methods and of challenges that arise specifically from language-driven VAD, including semantic-prior bias, hallucination, prompt sensitivity, grounding, evaluation of language outputs, and dependence on foundation models.

A dedicated survey that provides a focused analysis of the role, methodologies, and future of language models in VAD is therefore needed to guide researchers and practitioners in this rapidly growing area. This paper aims to fill this gap by providing an in depth survey of language model based methods for Video Anomaly Detection. We distill key concepts, categorize diverse methodologies, and outline promising directions for future research.

We present a structured and comprehensive review of the Video Anomaly Detection (VAD) domain, with a particular focus on the emerging role of language model based methods. First, we provide an overview of VAD and categorize existing methods into representative paradigms to clarify the methodological landscape and the diverse approaches used to detect anomalies in video. Second, we review widely used datasets, examining their characteristics, scale, and suitability for different VAD settings, and provide practical guidance for evaluation and benchmarking. Third, we summarize commonly adopted evaluation metrics and discuss their strengths and limitations to support fair and consistent comparisons.

Building on this foundation, we examine language model based VAD methods in detail, organizing the discussion according to the categories introduced earlier. For each category, we highlight key developments and design choices that distinguish LM based approaches. Because the primary scope of this survey is language-model-based VAD, conventional non-LM methods are reviewed selectively rather than exhaustively. We focus on representative foundations

that directly motivate, contextualize, or provide comparison points for subsequent LM-based approaches, including reconstruction- and prediction-based normality learning, memory-based methods, weakly supervised multiple-instance learning, and related temporal modeling strategies. Comprehensive treatments of the broader non-LM VAD literature are available in existing general surveys [27]–[35]. This focused treatment avoids duplicating those surveys while making clear how language-based methods extend or depart from established VAD paradigms.

Furthermore, we expect LM-based VAD research to continue evolving rapidly as new MLLMs, prompting strategies, instruction-tuning datasets, open-vocabulary protocols, and anomaly-understanding benchmarks emerge. For this reason, we do not treat this survey as a fixed snapshot of the literature. Instead, following the dynamic survey perspective [36], we frame this work as a living scholarly document whose content can be maintained over time while preserving a stable taxonomy and narrative structure. In a traditional static survey, the literature coverage begins to degrade immediately after publication as new papers appear. In contrast, a dynamic survey aims to reduce obsolescence by periodically integrating relevant new work into the existing organization rather than requiring the community to repeatedly produce overlapping surveys on the same topic. This distinction is particularly important for LM-based VAD, where the pace of publication is high and methodological categories are still actively forming.

Concretely, the taxonomy and paper coverage in this survey are designed to support versioned revisions. Future updates may add newly published methods, revise method categorizations, expand dataset and metric tables, and correct metadata, while keeping the high-level structure of the survey stable unless a major human-guided revision is warranted. This update philosophy follows three principles: updates should be localized to the relevant section, conservative with respect to existing text and organization, and transparent through

version records. By adopting this dynamic format, the survey aims to remain useful beyond its initial publication date and to provide a continuously maintained reference for researchers working on language-model-based video anomaly detection. Table 2 positions our survey with respect to existing VAD surveys by comparing their scope, benchmark discussion, treatment of advanced LM-based VAD topics, and support for dynamic survey design.

The main contributions of this survey can be summarized as follows:

- We provide a focused and systematic survey of language-model-based video anomaly detection, emphasizing how LLMs and MLLMs enable semantic reasoning, contextual understanding, open-vocabulary recognition, and anomaly explanation.
- We organize LM-based VAD methods into a unified taxonomy across supervision and deployment paradigms, including fully supervised, unsupervised, semi-supervised, weakly supervised, training-free, instruction-tuned, and open-world/open-vocabulary settings.
- We position LM-based methods relative to selected seminal and state-of-the-art non-LM foundations, clarifying how recent language-driven approaches extend, modify, or depart from conventional VAD pipelines.
- We review representative VAD datasets and evaluation protocols, highlighting their annotation granularity, semantic richness, and suitability for benchmarking both anomaly detection and anomaly understanding.
- We present the survey as a dynamic, versioned scholarly resource whose taxonomy, method coverage, dataset discussion, and evaluation analysis can be periodically maintained as the field evolves.
- We identify key limitations, open challenges, and future research directions for robust, interpretable, efficient, and generalizable language-model-based VAD systems.

II. PROBLEM SETTINGS AND SUPERVISION IN VIDEO ANOMALY DETECTION

Video Anomaly Detection (VAD) can be formulated under several supervision settings depending on the type and granularity of annotations available during training. These settings differ not only in annotation requirements, but also in their underlying assumptions regarding normality, anomaly distributions, and temporal localization. Classical VAD research has primarily focused on unsupervised, semi-supervised, and weakly supervised formulations, while recent advances in Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have further introduced paradigms such as open-vocabulary and training-free VAD. Importantly, the categories discussed in this survey should not be interpreted as mutually exclusive dimensions. Contemporary LM-based VAD methods can be characterized simultaneously along several partially orthogonal axes, including supervision level, adaptation strategy, label-space assumption, input modality, output type, and deployment setting. For example,

a method may be weakly supervised in terms of training labels, instruction-tuned in terms of model adaptation, multimodal in terms of input, open-vocabulary in terms of label space, and offline in terms of deployment. For readability, the literature in this survey is organized according to the paradigm that most strongly defines each method’s principal methodological contribution, while secondary characteristics are discussed within the corresponding method descriptions. In the following, we formalize these different settings.

Let a video be represented as an ordered sequence of visual units

$$V = [x_t]_{t=1}^T, \quad (1)$$

where each unit x_t may denote either an individual frame $x_t \in \mathbb{R}^{H \times W \times C}$ or a short video clip $x_t \in \mathbb{R}^{L \times H \times W \times C}$ consisting of L consecutive frames. Here, H , W , and C denote the height, width, and number of channels, respectively, while L denotes the temporal length of a clip. Thus, the index t represents the temporal ordering of the video, whereas the spatial or spatio-temporal structure of each observation is contained within x_t . Depending on the method, x_t may correspond to a frame, a fixed-length clip, or a temporal segment extracted from the original video.

A VAD model generally aims to estimate anomaly-related outputs for a video at one or more levels of granularity. In the most common temporal detection setting, the model assigns an anomaly score to each frame, clip, or temporal segment,

$$s_t = g(x_t), \quad (2)$$

where larger values indicate a higher likelihood of anomalous behavior at the corresponding time step or temporal unit. A binary temporal prediction can then be obtained as

$$\hat{y}_t = \begin{cases} 1, & s_t > \tau, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where τ denotes a predefined anomaly threshold.

Beyond temporal detection, VAD systems may also produce spatial localization outputs. For example, a model may estimate an anomalous region, bounding box, or object track associated with each temporal unit,

$$\hat{r}_t = h(x_t), \quad (4)$$

where $\hat{r}_t$ denotes the predicted spatial region or localized anomalous entity at time t . In object-centric settings, this output may correspond to a bounding box, segmentation mask, or track-level prediction.

Recent language-model-based VAD systems may further generate semantic outputs, such as anomaly categories, textual explanations, question-answer responses, or temporal grounding results. This can be written generally as

$$\hat{o}_t = F(x_t, p), \quad (5)$$

where p denotes a natural-language prompt or query and $\hat{o}_t$ denotes the generated semantic output. Depending on the task, $\hat{o}_t$ may represent an anomaly label, a natural-language

explanation, an answer to an anomaly-related question, or a grounded temporal segment.

More generally, modern VAD systems often operate on learned visual or multimodal representations

$$z_t = \phi(x_t), \quad (6)$$

where $\phi(\cdot)$ denotes a feature extraction backbone, vision-language encoder, or multimodal foundation model. Different supervision settings primarily vary in how these representations are learned, constrained, queried, localized, or decoded during training and inference. In the following subsections, we review the major supervision and deployment paradigms used in contemporary VAD research.

A. UNSUPERVISED AND SEMI-SUPERVISED

Unsupervised and semi-supervised VAD are closely related paradigms and are often used interchangeably in the literature. Both settings aim to detect anomalous events by learning the distribution of normal or mostly normal video patterns and assigning high anomaly scores to observations that deviate from this learned behavior. The main distinction lies in the assumptions made about the training data. In the strict unsupervised setting, the training set is fully unlabeled and may contain both normal and anomalous videos, with anomalies assumed to be rare relative to normal events. In the semi-supervised, or one-class, setting, the training data are assumed to contain only normal videos. These formulations are closely related to classical one-class anomaly detection, where the objective is to estimate the support or compact description of normal data and identify observations outside this region as anomalous [37]–[39].

Formally, in the unsupervised case, the model is given a collection of unlabeled videos

$$\mathcal{D}_{\text{train}} = \{V_i\}_{i=1}^N, \quad (7)$$

where V_i denotes the i -th video (that may contain both normal and anomalous activity) and N is the total number of training videos, and aims to learn the underlying distribution of the observed data,

$$p(V) \quad \text{or} \quad p(z), \quad (8)$$

where z denotes a learned latent representation of the video content. Since labels are unavailable, anomalous events are expected to appear as low-density, irregular, or poorly reconstructed samples within the learned distribution. In semi-supervised VAD, the training set $\mathcal{D}_{\text{train}}$ is assumed to consist of only normal videos. The objective is therefore to learn a representation of normality and identify deviations from this representation during inference.

A common formulation in both settings relies on reconstruction-based or prediction-based learning. Reconstruction based methods train an encoder-decoder model to reproduce normal frames, clips, or temporal units, while prediction-based methods learn to forecast future frames or motion patterns. The anomaly score can be defined as

$$s_t = \|x_t - \hat{x}_t\|_2, \quad (9)$$

where x_t denotes the input visual unit and $\hat{x}_t$ denotes its reconstruction or prediction and $\|x\|_2$ represents L_2 or Euclidean distance. A large reconstruction or prediction error suggests that the observed event cannot be well represented by the learned normal data manifold. Other approaches estimate likelihood directly using probabilistic generative models such as normalizing flows, where low likelihood indicates anomalous behavior:

$$s_t = -\log p(z_t) \quad (10)$$

where z_t is a learned representation given by Equation 6.

More generally an anomaly score can be computed as the distance from the learned normal manifold,

$$s_t = d(z_t, \mathcal{M}_{\text{normal}}), \quad (11)$$

where $\mathcal{M}_{\text{normal}}$ denotes the learned representation of normal behavior and $d(\cdot, \cdot)$ denotes a distance or dissimilarity function. Larger distances indicate a higher likelihood of anomaly. Recent methods also employ memory modules, contrastive learning, masked modeling, transformer-based architectures, and density estimation to improve representation quality and temporal modeling.

These paradigms are attractive for real-world surveillance applications because anomalous events are rare, diverse, and difficult to enumerate exhaustively. Collecting representative anomalous samples is often impractical, and normal-only or unlabeled training data are substantially easier to obtain. However, the central assumption that anomalies deviate strongly from normal patterns does not always hold in practice. Subtle anomalies may closely resemble normal activities at the visual level, while normal behavior may vary significantly across scenes and contexts. Moreover, purely statistical deviations do not necessarily correspond to semantically meaningful anomalies: an event may appear visually unusual while remaining contextually normal, whereas a semantically abnormal event may exhibit only subtle low-level differences. These limitations have motivated recent research toward semantic reasoning and multimodal understanding through language-guided and MLLM-based frameworks.

B. WEAKLY SUPERVISED

In the weakly supervised setting, the training set contains both normal and anomalous videos, but supervision is provided only at the video level. Formally, the training data can be represented as

$$\mathcal{D}_{\text{train}} = \{(V_i, Y_i)\}_{i=1}^N, \quad (12)$$

where $Y_i \in \{0, 1\}$ indicates whether video V_i contains any anomalous event. Unlike fully supervised VAD, no temporal annotations are provided regarding the exact location or duration of anomalies within the video. This coarse supervision substantially reduces annotation cost while still providing stronger training signal than the semi-supervised setting.

Multiple Instance Learning (MIL) is the dominant framework for weakly supervised VAD [19]. This follows the classical multiple-instance learning assumption, where labels are

observed at the bag level while instance-level labels remain latent [40], [41]. In MIL-based formulations, each video is treated as a bag of temporal visual units,

$$B_i = \{x_{i1}, x_{i2}, \dots, x_{iK}\}, \quad (13)$$

where x_{ij} denotes the j -th temporal unit of video V_i . The central assumption is that a positive bag contains at least one anomalous instance:

$$Y_i = 1 \iff \exists j \in \{1, \dots, K\} : y_{ij} = 1. \quad (14)$$

The model predicts anomaly scores for individual temporal units,

$$s_{ij} = f_\theta(x_{ij}), \quad (15)$$

where f_θ denotes a parameterized anomaly scoring model and s_{ij} is the predicted anomaly score for the temporal unit x_{ij} . A common formulation uses max pooling,

$$S_i = \max_{j \in \{1, \dots, K\}} s_{ij}, \quad (16)$$

which encourages the model to identify the most suspicious temporal regions in anomalous videos while suppressing scores in normal videos.

Subsequent research has extended the MIL framework through graph-based temporal reasoning, attention mechanisms, self-training, memory augmentation, feature enhancement, and transformer-based architectures. These methods aim to improve temporal localization accuracy and reduce the ambiguity introduced by coarse supervision. Compared with semi-supervised approaches, weakly supervised methods generally achieve stronger anomaly localization performance because they are exposed to anomalous examples during training.

The weakly supervised setting is also particularly compatible with language model-based approaches. Video-level descriptions, anomaly categories, or textual metadata can naturally be converted into language supervision. This allows MLLMs to leverage semantic priors and world knowledge about anomalous activities, enabling reasoning beyond low-level motion and appearance statistics. As a result, weakly supervised VAD has become one of the primary settings for integrating multimodal foundation models into anomaly detection pipelines.

C. OPEN-SET AND OPEN-VOCABULARY

Conventional weakly supervised VAD benchmarks commonly assume a closed-set protocol, in which the anomaly categories encountered during testing are also observed during training. Formally, this assumption can be expressed as

$$\mathcal{Y}_{\text{test}} \subseteq \mathcal{Y}_{\text{train}}, \quad (17)$$

where $\mathcal{Y}_{\text{train}}$ and $\mathcal{Y}_{\text{test}}$ denote the anomaly categories present in the training and testing sets, respectively. This relaxation is related to the broader open-set recognition problem, where models must handle samples outside the training label space rather than assuming a closed-world classification protocol [42], [43]. While this setting simplifies evaluation, it does

not accurately reflect real-world surveillance environments, where anomalous events are often unpredictable and continuously evolving.

Open-set VAD relaxes this assumption by requiring models to detect anomalies belonging to previously unseen categories,

$$\exists y^* \in \mathcal{Y}_{\text{test}} \quad \text{such that} \quad y^* \notin \mathcal{Y}_{\text{train}}. \quad (18)$$

where y^* denotes an anomaly category that appears during testing but is absent from the training category set. In this setting, the objective is not necessarily to classify the anomaly into a predefined category, but rather to recognize that the observed event deviates from known patterns. Open-set VAD therefore places stronger emphasis on generalization and semantic understanding than conventional closed-set evaluation.

Closely related to this paradigm is open-vocabulary VAD, which leverages language supervision to define anomaly concepts through natural language descriptions rather than fixed category labels. Instead of restricting prediction to a predefined taxonomy, models can reason about arbitrary textual queries or prompts. Given a visual representation $\phi_v(x_t)$ of a frame, clip, temporal segment, or video-level unit, and a text representation $\phi_t(p)$ derived from a prompt p , anomaly relevance can be estimated through cross-modal similarity:

$$s(x_t, p) = \frac{\phi_v(x_t)^T \phi_t(p)}{\|\phi_v(x_t)\| \|\phi_t(p)\|}. \quad (19)$$

This formulation enables flexible anomaly querying using descriptions such as ‘person carrying a weapon’ or ‘vehicle moving against traffic,’ even when such events were not explicitly annotated during training.

The emergence of vision-language models and MLLMs has substantially accelerated research in open-set and open-vocabulary VAD. Large-scale multimodal pretraining provides semantic priors and world knowledge that improve generalization to unseen anomaly types. In contrast to traditional VAD systems that primarily rely on low-level appearance and motion statistics, open-vocabulary approaches enable anomaly detection through contextual and semantic reasoning. However, these methods also introduce new challenges, including prompt sensitivity, semantic ambiguity, hallucination, and difficulties in evaluating open-ended anomaly definitions consistently across datasets.

D. TRAINING-FREE

Training-free VAD refers to approaches that leverage pre-trained foundation models, including vision-language models and Multimodal Large Language Models (MLLMs), without performing task-specific optimization or fine-tuning. Unlike conventional supervised learning frameworks, the model parameters remain fixed during deployment,

$$\theta_{\text{deploy}} = \theta_{\text{pretrained}}, \quad (20)$$

and anomaly detection is performed directly through prompting, similarity estimation, or zero-shot reasoning.

Given a visual unit x_t and a textual prompt p , a training-free framework estimates anomaly relevance using the pretrained multimodal model,

$$s(x_t, p) = f_{\theta_{\text{pretrained}}}(x_t, p), \quad (21)$$

where $s(x_t, p)$ denotes the anomaly score or semantic relevance between the observed frame, clip, temporal segment, or video-level unit and the textual description. In many methods, prompts describe either normal behavior or candidate anomalous events, and anomaly detection is formulated as a semantic comparison problem in a shared multimodal embedding space.

The training-free paradigm has emerged rapidly following the availability of capable foundation models pretrained on large-scale web data. [21]–[23] Because these models already encode substantial semantic and contextual knowledge, they can often generalize to unseen anomaly categories and environments without requiring annotated VAD datasets. Prompt engineering and in-context learning therefore play a central role. Carefully designed prompts can guide an MLLM to reason about whether an observed activity is abnormal relative to the surrounding scene context, even without gradient-based adaptation.

Training-free methods offer several practical advantages. They eliminate the need for expensive VAD training pipelines, reduce deployment overhead, and provide immediate transferability across domains and anomaly categories. In addition, they naturally support open-vocabulary reasoning through natural language interaction. These properties make training-free VAD particularly attractive for rapidly evolving or low-resource surveillance environments where collecting representative anomaly annotations is impractical.

Despite these advantages, training-free approaches also face important limitations. Their performance is fundamentally constrained by the capabilities and biases of the underlying foundation model, which may not adequately represent specialized surveillance domains or rare anomaly types encountered during deployment. In addition, training-free pipelines are often substantially slower than conventional VAD systems because they rely on large multimodal architectures that require computationally expensive inference. Sequential reasoning, multi-step prompting, and repeated query evaluation can further increase latency and memory consumption, limiting practicality for real-time or large-scale surveillance applications.

Nevertheless, training-free VAD has become a major research direction in modern anomaly detection. Recent studies demonstrate that large multimodal models can achieve competitive performance on several VAD benchmarks despite requiring no task-specific optimization or annotated anomaly data. As a result, training-free methods increasingly serve both as practical deployment frameworks and as strong baselines for measuring the benefits of fine-tuning and specialized adaptation strategies.

E. INSTRUCTION-TUNED

Instruction-tuned VAD can be viewed as an extension of training-free VAD, where pretrained vision-language or multimodal large language models are adapted to anomaly-related tasks through instruction tuning. Unlike training-free methods, which keep all model parameters fixed, instruction-tuned approaches perform task-specific adaptation using curated video-instruction pairs, anomaly question-answer data, or explanation-oriented supervision. In many cases, parameter-efficient fine-tuning strategies such as LoRA are used to adapt large multimodal models without updating all parameters. [44], [45]

Formally, given a pretrained multimodal model with parameters $\theta_{\text{pretrained}}$, instruction-tuned VAD learns adapted parameters

$$\theta^* = \theta_{\text{pretrained}} + \Delta\theta, \quad (22)$$

where $\Delta\theta$ denotes the task-specific adaptation introduced through fine-tuning. Given a visual unit x_t and an instruction prompt p , the model estimates anomaly relevance or generates an anomaly-related response as

$$\hat{y}_t = f_{\theta^*}(x_t, p). \quad (23)$$

Depending on the task formulation, $\hat{y}_t$ may represent an anomaly score, a binary decision, a category prediction, a temporal grounding result, or a natural-language explanation.

Instruction-tuned methods are particularly important for anomaly understanding, since they can align pretrained multimodal models with VAD-specific reasoning patterns, domain terminology, and output formats. Compared with training-free methods, they often provide better task alignment and more controllable responses. However, they require task-specific instruction data and may inherit dataset biases or overfit to limited anomaly types. As a result, instruction-tuned VAD occupies an intermediate position between fully training-free reasoning and conventional supervised adaptation.

F. OTHER PARADIGMS: FULLY SUPERVISED AND ONLINE VAD

Beyond the major settings discussed above, fully supervised and online VAD represent two additional paradigms. Fully supervised VAD assumes access to dense temporal annotations, and in some cases anomaly category labels or spatial annotations. Formally, each training video may be associated with anomaly intervals

$$\mathcal{A}_i = \{(t_s^{(k)}, t_e^{(k)}, c_k)\}_{k=1}^{N_i}, \quad (24)$$

where $t_s^{(k)}$ and $t_e^{(k)}$ denote the start and end times of the k -th anomalous event, and c_k denotes its category label. This setting provides strong supervision for temporal localization, but it is often impractical for large-scale VAD because dense annotation of rare and ambiguous anomalous events is expensive and difficult to scale. Consequently, fully supervised formulations are less common in VAD than semi-supervised,

weakly supervised, training-free, or open-vocabulary settings. Within language-model-based VAD, fully supervised methods are especially rare, since most LM-based approaches aim to reduce annotation requirements or exploit language supervision instead of relying on dense frame-level labels.

Online and streaming VAD considers scenarios in which video frames or clips arrive sequentially and predictions must be generated without access to future observations. Given a partial video stream

$$V_{\leq t} = \{x_1, x_2, \dots, x_t\}, \quad (25)$$

the prediction at time step t is restricted to information available up to the current time,

$$\hat{y}_t = g(V_{\leq t}). \quad (26)$$

This setting is important for real-time surveillance, where delayed detection can reduce practical utility. Accordingly, detection delay is an important complementary evaluation criterion for online VAD. Beyond whether an anomalous event is eventually detected, an online system should also be evaluated according to how quickly it raises an alarm after anomaly onset. One such metric is Average Detection Delay (ADD) [46], defined over N anomalous activities as

$$\text{ADD} = \frac{1}{N} \sum_{i=1}^N (T_i - \tau_i), \quad (27)$$

where τ_i denotes the onset time of anomalous activity i and $T_i \geq \tau_i$ denotes the corresponding alarm time. If no alarm is produced within a predefined maximum tolerable delay $\delta_{\max}$, the delay for that event can be capped at $\delta_{\max}$. Two methods with similar frame-level AUC or AP may therefore differ substantially in practical utility if one consistently detects anomalous events earlier than the other. Doshi and Yilmaz further combine normalized ADD with alarm precision through the Average Precision Delay (APD) metric, jointly accounting for timely detection and false alarms [46]. Although such latency-aware metrics have been proposed, detection-delay evaluation is not yet standardized across VAD benchmarks. Online deployment is particularly challenging for language-model-based VAD because MLLMs often require computationally expensive inference, prompt processing, multi-step reasoning, and long-context video understanding. These factors introduce latency and memory overhead, making real-time LM-based VAD substantially more difficult than offline analysis. As a result, efficient and streaming-compatible adaptation of foundation models remains an important but relatively underexplored direction.

III. DATASETS

Table 3 summarizes representative VAD datasets and compares them according to their annotation granularity and semantic richness. While traditional VAD benchmarks primarily provide temporal and spatial anomaly annotations, recent datasets increasingly incorporate higher-level semantic supervision. To capture this distinction, we include a semantic

annotation category that describes the richest form of semantic information available in each dataset.

Semantic Annotation refers to the highest level of semantic supervision provided by a dataset, ranging from anomaly category labels to question-answer pairs, grounding annotations, retrieval pairs, captioning supervision, and textual explanations.

Early VAD research was largely driven by scene-specific datasets designed to evaluate statistical anomaly detection methods. Representative benchmarks include Subway Entrance and Subway Exit [7], UMN [9], UCSD Ped1 and UCSD Ped2 [47], CUHK Avenue [20], ShanghaiTech [48], NWPU Campus [49], and Street Scene [50]. These datasets primarily provide frame-level anomaly labels and, in some cases, spatial annotations through bounding boxes. Their focus is on detecting deviations from learned normal patterns within relatively constrained environments. Among them, Street Scene [50] provides track-level temporal annotations that support object-centric anomaly localization and interaction analysis. ComplexVAD [51], a large-scale streetscape benchmark containing complex interaction anomalies with both track-level temporal annotations and bounding-box spatial annotations, enabling fine-grained object-centric anomaly localization and future research on anomaly explanation and understanding. Despite their importance for evaluating anomaly detection performance, these datasets generally lack semantic annotations and therefore provide limited support for language-based reasoning, explanation generation, or anomaly understanding.

To improve scalability and anomaly diversity, later benchmarks expanded beyond single-scene settings and introduced category-level semantic supervision. Representative examples include UCF-Crime [19], UCF-Crime Extension [52], XD-Violence [53], TAD [54], BOSS [55], CamNuvem [56], Ubnormal [57], and the dataset introduced by Zhu et al. [58]. Compared with earlier benchmarks, these datasets contain substantially larger numbers of videos and anomaly categories, enabling the development of weakly supervised and fully supervised VAD methods. However, semantic supervision remains relatively coarse, typically consisting of anomaly category labels rather than detailed descriptions, explanations, or reasoning-oriented annotations.

More recently, several benchmarks have been proposed to evaluate anomaly understanding rather than anomaly detection alone. Examples include CUVA [59], ECVA [60], VANE-Bench [61], VAGU [62], FineVAU [63], HVAU-70K [45], and UCA [64]. These datasets introduce richer semantic annotations, including question-answer pairs, conversational interactions, grounding annotations, and fine-grained reasoning tasks. Similarly, the retrieval benchmark proposed by Yang et al. [65] and the captioning benchmark introduced by Bao et al. [66] extend evaluation toward video-text retrieval and anomaly description generation. Such benchmarks enable the assessment of higher-level capabilities including anomaly explanation, contextual reasoning, temporal grounding, and language-guided understanding, making them par-

TABLE 3. Comparison of representative VAD datasets in terms of annotation granularity and semantic richness. Datasets are grouped according to their dominant annotation and semantic characteristics.

Dataset	Year	Domain	Total Frames	Anomaly Categories	Temporal Annotation	Spatial Annotation	Semantic Annotation
Subway Entrance [7]	2008	Streetscape	86,535	5	Frame	–	None
Subway Exit [7]	2008	Streetscape	38,940	3	Frame	–	None
UMN [9]	2009	Crowd Behavior	3,855	1	Frame	–	None
UCSD Ped1 [47]	2013	Streetscape	14,000	5	Frame	Bounding Box	None
UCSD Ped2 [47]	2013	Streetscape	4,560	5	Frame	Bounding Box	None
CUHK Avenue [20]	2013	Streetscape	30,652	5	Frame	Bounding Box	None
NWPU Campus [49]	2023	Streetscape	1,466,073	28	Frame	–	None
ShanghaiTech [48]	2017	Streetscape	317,398	13	Frame	Bounding Box	None
Street Scene [50]	2020	Streetscape	203,257	17	Track	Bounding Box	None
ComplexVAD [51]	2025	Streetscape	3,681,438	40	Track	Bounding Box	None
UCF-Crime [19]	2018	Crime	13,741,393	13	Video	–	Category Labels
UCF-Crime Extension [52]	2021	Crime	14,475,793	15	Video	–	Category Labels
XD-Violence [53]	2020	Violence	114,096	6	Video	–	Category Labels
TAD [54]	2024	Traffic	721,280	4	Frame	Bounding Box	Category Labels
BOSS [55]	2017	Multiple	48,624	11	Video	–	Category Labels
CamNuvem [56]	2022	Robbery	6,151,788	1	Video	–	Category Labels
Ubnormal [57]	2022	Multiple	236,902	22	Frame	Pixel	Category Labels
SENSE-VAD [67]	2026	Autonomous Driving	540,888	15	Frame	Bounding Box	Category Labels
MSAD [58]	2024	Multiple	447,236	55	Frame	–	Category Labels
CUVA [59]	2024	Multiple	3,345,097	11	Time Duration	–	Video QA
ECVA [60]	2024	Multiple	19,042,560	21	Time Duration	–	Video QA
VANE-Bench [61]	2025	Multiple	951,482	19	Video	–	Conversational QA
VAGU [62]	2025	Multiple	20,400,000	21	Period	–	Grounding + QA
FineVAU [63]	2026	Surveillance	–	13	–	–	Fine-grained QA
SVTA [65]	2025	Multiple	1,360,000	68	Video-level	–	Retrieval text pairs
CVACBench [66]	2025	Surveillance	60,160	13	Frame	–	Captioning
HVAU-70K [45]	2025	Multiple	13,855,489	15	Frame	–	Video QA
UCA [64]	2024	Crime	11,817,597	13	Frame	–	Video QA

ticularly relevant for modern vision-language models and MLLMs.

Despite their importance for benchmarking, existing VAD datasets exhibit several limitations that should be considered when interpreting reported performance. Many conventional benchmarks contain data from a limited number of cameras, geographic locations, or environmental conditions, which can encourage scene-specific or background-dependent representations and limit conclusions about cross-scene generalization. Anomaly definitions may also be dataset-dependent and sometimes ambiguous, since the same behavior can be normal or abnormal depending on location, context, or local activity patterns. In addition, VAD datasets are typically highly imbalanced, with anomalous events occupying only a small fraction of the available video, while multi-scene datasets may still provide limited diversity relative to real deployment environments. Some recent benchmarks increase anomaly diversity through curated or synthetically generated content; for example, VANE-Bench includes both real-world surveillance videos and synthetically generated anomalous videos [61]. While such data can broaden anomaly coverage, synthetic or curated events may not fully capture the visual, temporal, and contextual variability of naturally occurring real-world anomalies. Closed-set category annotations can further encourage category recognition rather than scene-specific anomaly reasoning. These factors should therefore be considered when transferring benchmark results to operational settings, and future datasets would benefit from greater camera, scene, geographic, and contextual diversity together

with richer annotations of ambiguous and location-dependent abnormality.

IV. EVALUATION METRICS

The evaluation of VAD systems has evolved alongside the development of the field itself. Early VAD benchmarks primarily focused on measuring anomaly detection performance, emphasizing whether a model could successfully distinguish anomalous events from normal observations. As datasets began providing spatial and temporal annotations, localization-oriented metrics were introduced to assess the ability of models to accurately identify where anomalies occur. More recently, the emergence of video anomaly understanding benchmarks has motivated the adoption of semantic evaluation metrics that assess explanation quality, question answering performance, grounding accuracy, and language-based reasoning capabilities.

Consequently, modern VAD evaluation can be broadly categorized into three groups: detection metrics, localization metrics, and semantic understanding metrics. Detection metrics assess whether anomalous events can be reliably distinguished from normal activities. Localization metrics evaluate the accuracy of spatial and temporal anomaly localization. Semantic understanding metrics measure the ability of models to interpret, explain, and reason about anomalous events through natural language. In the following subsections, we review the most commonly used metrics within each category and discuss their relevance to both conventional VAD systems and emerging MLLM-based approaches.

A. DETECTION METRICS

Detection metrics evaluate a model's ability to distinguish anomalous events from normal observations. Given a set of predictions and corresponding ground-truth labels, these metrics assess the quality of anomaly classification at either the frame, clip, or video level. Because anomalous events are typically rare relative to normal activities, VAD datasets often exhibit significant class imbalance. Consequently, metrics that consider performance across multiple decision thresholds are generally preferred over threshold-dependent measures.

Area Under the ROC Curve (AUC). Threshold-independent metrics such as ROC-AUC and precision-recall analysis are widely used for imbalanced detection problems because they summarize model behavior across operating points [68], [69]. AUC is the most widely used evaluation metric in VAD. It measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across different decision thresholds. These quantities are defined as

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}, \quad (28)$$

where TP , TN , FP , and FN denote the numbers of true positive, true negative, false positive, and false negative frames, respectively. The AUC score is then computed as

$$AUC = \int_0^1 TPR(FPR) d(FPR). \quad (29)$$

An AUC value of 1 indicates perfect frame-level anomaly discrimination, whereas a value of 0.5 corresponds to random guessing. Since AUC evaluates performance across all possible thresholds, it is particularly suitable for highly imbalanced VAD datasets and remains the dominant metric in most benchmark evaluations.

Average Precision (AP). AP summarizes the precision-recall curve and is commonly employed when anomalous samples constitute only a small fraction of the dataset [69], [70]. Given anomaly scores for all evaluated frames, clips, or videos, samples are ranked from highest to lowest anomaly confidence. Precision and recall are then computed at each rank position as progressively more samples are treated as positive:

$$Precision_n = \frac{TP_n}{TP_n + FP_n}, \quad Recall_n = \frac{TP_n}{TP_n + FN_n}, \quad (30)$$

where TP_n , FP_n , and FN_n are computed after considering the top- n ranked predictions. The AP score is computed as

$$AP = \sum_n (R_n - R_{n-1}) P_n, \quad (31)$$

where P_n and R_n denote precision and recall at the n -th ranked operating point. Compared with AUC, AP places greater emphasis on correctly retrieving anomalous events and is therefore particularly informative when the positive class is extremely sparse.

Although AUC and AP provide threshold-independent summaries of anomaly-ranking performance, they do not fully characterize several aspects that are important in practical VAD evaluation. First, frame-level metrics can overweight long anomalous events because each anomalous frame contributes separately to the score, while short events may have relatively little influence. Event-level evaluation can therefore provide a complementary view by assessing whether distinct anomalous events are successfully detected, rather than only how individual frames are ranked. Second, ranking metrics do not measure the calibration of anomaly scores. Two methods may achieve similar AUC or AP while producing scores with very different confidence reliability, which is important when fixed operating thresholds or downstream decision systems are used. These considerations are particularly relevant for highly imbalanced VAD datasets and motivate reporting multiple complementary metrics rather than relying on a single aggregate score.

Accuracy. Accuracy measures the proportion of correctly classified samples after anomaly scores are converted into binary predictions using a fixed decision threshold. Given a threshold τ , the predicted label is defined as

$$\hat{y}_t = \begin{cases} 1, & s_t > \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (32)$$

Accuracy is then computed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (33)$$

Although accuracy is intuitive and easy to interpret, it is less frequently reported in VAD because it depends strongly on the selected threshold and is prone to manipulation through threshold tuning. This issue is further amplified by the severe class imbalance present in many anomaly detection datasets. A model may achieve high accuracy simply by predicting most samples as normal while failing to detect anomalous events. For this reason, threshold-independent metrics such as AUC and ranking-based metrics such as AP are generally preferred.

Equal Error Rate (EER). EER is defined as the operating point on the ROC curve where the false positive rate equals the false negative rate. It is obtained by varying the decision threshold over anomaly scores and identifying the threshold at which

$$FPR = FNR, \quad (34)$$

where

$$FNR = \frac{FN}{TP + FN}. \quad (35)$$

Lower EER values indicate better detection performance. Unlike accuracy, EER does not depend on a manually selected fixed threshold, but it still summarizes performance at a single operating point. It is occasionally used in VAD benchmarks to compare the balance between missed detections and false alarms.

Equal Detected Rate (EDR). EDR measures the proportion of anomalous events successfully detected under a

predefined operating condition, such as a fixed false alarm rate, false positive rate, or dataset-specific evaluation protocol. While definitions may vary across benchmarks, EDR generally quantifies detection completeness at a specified operating point. Since it depends on the chosen operating condition, it should be interpreted together with the corresponding threshold or false-alarm constraint. In surveillance applications, where missing a true anomaly may have serious consequences, EDR provides additional insight into the practical effectiveness of a detection system and is often reported alongside EER.

B. LOCALIZATION METRICS

While detection metrics evaluate whether a model can distinguish anomalous events from normal activities, they do not assess the accuracy of spatial localization of anomalies. In many practical applications, identifying the spatial and temporal extent of an anomaly is equally important as detecting its presence. Consequently, several localization-oriented metrics have been proposed to evaluate how accurately models identify anomalous regions and trajectories [50]. The Street Scene evaluation protocol introduced region-based and track-based criteria to better account for spatial localization and false-positive regions, rather than only evaluating whether anomalous frames are detected.

Region-Based Detection Criterion (RBDC). RBDC evaluates anomaly localization at the level of anomalous regions. Under this criterion, a ground-truth anomalous region is considered detected if its intersection-over-union (IoU) with at least one detected anomalous region is greater than or equal to a predefined threshold β . Formally, the IoU between a ground-truth region R_{gt} and a detected region R_{det} is defined as

$$\text{IoU}(R_{gt}, R_{det}) = \frac{|R_{gt} \cap R_{det}|}{|R_{gt} \cup R_{det}|}. \quad (36)$$

A ground-truth region is counted as detected if

$$\text{IoU}(R_{gt}, R_{det}) \geq \beta. \quad (37)$$

The corresponding region-based detection rate (RBDR) is then computed as

$$\text{RBDR} = \frac{\text{number of anomalous regions detected}}{\text{total number of anomalous regions}}. \quad (38)$$

RBDR is computed over all ground-truth anomalous regions in all frames of the test set. Unlike frame-level evaluation, this criterion requires spatial overlap between predicted and ground-truth anomalous regions, making it more appropriate for datasets with bounding-box or pixel-level annotations.

Track-Based Detection Criterion (TBDC). TBDC evaluates anomaly localization at the object-track or event-track level rather than requiring successful region detection in every frame. Under this criterion, a ground-truth anomalous track is considered detected if at least a fraction α of its ground-truth regions are detected. Each ground-truth region within the track is considered detected when its IoU with a detected region is at least β .

Formally, let $\mathcal{T}_{gt}$ denote the set of ground-truth anomalous tracks and let $\mathcal{T}_{det}$ denote the subset of ground-truth tracks that satisfy the track-level detection criterion. The track-based detection rate (TBDR) is defined as

$$\text{TBDR} = \frac{|\mathcal{T}_{det}|}{|\mathcal{T}_{gt}|}. \quad (39)$$

This criterion reflects the practical observation that an anomaly occurring over many frames does not necessarily need to be localized in every frame to be useful. Instead, the anomalous track should be detected in a sufficient fraction of its temporal extent.

False Positive Regions per Frame. Both RBDC and TBDC are evaluated together with the number of false positive regions per frame (FPR) to obtain the Area Under RBDR/TBDR–FPR Curve (AUC). A detected region in a frame is counted as a false positive if its IoU with every ground-truth anomalous region in that frame is less than β . The false-positive rate is defined as

$$\text{FPR} = \frac{\text{total false positive regions}}{\text{total frames}}. \quad (40)$$

This differs from traditional frame-level false-positive counting because multiple false positive regions can occur in a single frame, and false positives can also be counted in frames that contain true anomalies. This provides a more realistic estimate of how many incorrect anomalous regions a human operator would need to inspect.

Localization metrics have become increasingly important as recent datasets provide richer spatial and temporal annotations. For example, Street Scene [50] provides bounding-box annotations with track identifiers, enabling both region-based and track-based evaluation. ComplexVAD similarly provides object-centric annotations that support evaluation through metrics such as RBDC and TBDC. Such metrics are particularly valuable for complex interaction anomalies, where identifying the anomalous object or interaction at the region or track level can be as important as frame-level anomaly scoring.

C. SEMANTIC UNDERSTANDING METRICS

The emergence of vision-language models (VLMs) and Multimodal Large Language Models (MLLMs) has expanded the scope of VAD beyond anomaly detection toward anomaly understanding. Modern systems are increasingly expected not only to identify anomalous events, but also to explain why they are anomalous, answer questions about their causes, localize relevant evidence, and generate natural language descriptions. Consequently, evaluation protocols have adopted metrics from natural language processing and multimodal reasoning to assess semantic understanding capabilities. For generated anomaly descriptions and explanations, early automatic evaluation commonly relies on lexical-overlap metrics originally developed for machine translation and summarization [71]–[73].

BLEU. BLEU evaluates the similarity between a generated response and one or more reference annotations based on n-gram precision. The BLEU score is computed as

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right), \quad (41)$$

where p_n denotes the modified n-gram precision, w_n represents the weight assigned to each n-gram order, and BP is a brevity penalty that penalizes overly short outputs. Higher BLEU scores indicate greater lexical overlap between generated and reference descriptions.

ROUGE. ROUGE measures the overlap between generated and reference text from a recall-oriented perspective. For example, ROUGE-N is defined as

$$ROUGE-N = \frac{\sum_{N\text{-gram}} Count_{match}}{\sum_{N\text{-gram}} Count_{reference}}. \quad (42)$$

Unlike BLEU, which emphasizes precision, ROUGE focuses on how much of the reference content is successfully recovered by the generated output.

METEOR. METEOR evaluates semantic similarity using word matching, stemming, and synonym relationships. It is computed as

$$METEOR = \frac{10PR}{R + 9P}, \quad (43)$$

where P and R denote precision and recall, respectively. Compared with BLEU and ROUGE, METEOR often exhibits stronger correlation with human judgments because it accounts for linguistic variations beyond exact word matching.

While lexical metrics are useful for evaluating anomaly descriptions and captions, they often fail to capture semantic correctness and reasoning quality. Two responses may convey the same meaning while exhibiting little lexical overlap, resulting in artificially low scores.

LLM-Based Evaluation. To address the limitations of lexical metrics, several recent works employ large language models as evaluators. Given a generated response $\hat{y}$, a reference annotation r , and an evaluation prompt p , the evaluation score can be expressed as

$$Score_{LLM} = f_{LLM}(\hat{y}, r, p), \quad (44)$$

Such evaluators can assess semantic consistency, factual correctness, and reasoning quality beyond surface-level text similarity. However, LLM-based evaluation also introduces reliability and reproducibility challenges. The resulting judgment may depend on the evaluator model, model version, prompt formulation, decoding configuration, and evaluation rubric, such that different evaluators or prompts may assign different scores to the same response. Human evaluation is likewise subject to variability, particularly for explanation quality, contextual correctness, and reasoning, where annotators may disagree. Semantic evaluation protocols should

therefore report evaluator and prompt details and, where feasible, use multiple judges or repeated assessments together with measures of inter-rater agreement or judgment variability.

MLLM-Based Evaluation. Multimodal evaluation further incorporates the original video into the assessment process. Given a video V , generated response $\hat{y}$, reference annotation r , and evaluation prompt p , the evaluation score can be expressed as

$$Score_{MLLM} = f_{MLLM}(V, \hat{y}, r, p), \quad (45)$$

Because the evaluator has direct access to both visual and textual information, MLLM-based evaluation can better assess grounding accuracy, contextual understanding, and anomaly reasoning. Such evaluation protocols are becoming increasingly important for benchmarks such as CUVA [59], ECVA [60], VANE-Bench [61], VAGU [62], FineVAU [63], HVAU-70K [45], and UCA [64], which explicitly focus on anomaly understanding and language-guided reasoning.

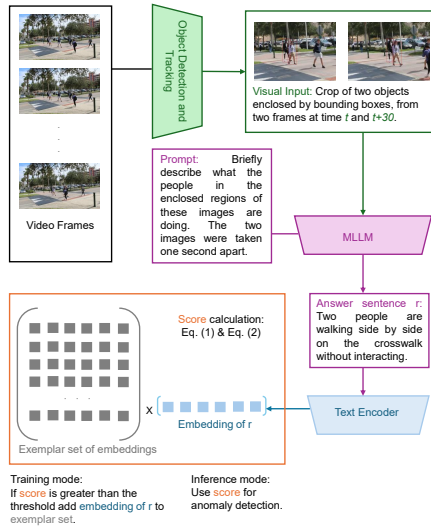
Human Evaluation. Human evaluation is commonly used when automatic lexical or model-based metrics cannot adequately capture the quality of anomaly explanations, reasoning, or other open-ended semantic responses. Common protocols include binary correctness judgments, Likert-scale ratings, pairwise preference comparisons, ranking-based evaluation, and separate dimension-wise assessments of properties such as factual correctness, relevance, grounding, faithfulness, and usefulness [74]–[76]. Recent video anomaly understanding benchmarks have also used multiple human evaluators and ranking-based judgments to assess semantic output quality and agreement with human perception [77]. Such evaluations are particularly valuable when multiple natural-language responses may be valid and semantic correctness cannot be reliably captured through lexical overlap alone.

However, human-evaluation protocols are not standardized across language-based VAD studies. Different works may use different rating scales, evaluation questions, numbers and backgrounds of annotators, aggregation procedures, and instructions, making the resulting scores difficult to compare directly across papers. Human judgments may also exhibit subjectivity and inter-rater disagreement, especially for context-dependent properties such as explanation quality or whether an observed behavior should be considered anomalous [78]. Consequently, studies using human evaluation should clearly report the evaluation criteria, rating scale or comparison protocol, annotator setup, and aggregation procedure and, where possible, report inter-rater agreement or variability across annotators. Because both the protocols and evaluator groups can differ substantially across studies, human-evaluation scores should generally be interpreted within the context of the individual study rather than as directly comparable benchmark values.

V. EXISTING LANGUAGE-MODEL-BASED VAD METHODS BY PARADIGM

TABLE 4. SVAD methods on Ped2 (Ped2), CUHK Avenue (Avenue), ShanghaiTech (SHTech), UBnormal (UBnormal), and Complex-VAD (CVAD). Values are taken from the corresponding source publications.

Method	Year	Ped2 (AUC)	Avenue (AUC)	SHTech (AUC)	UBnormal (AUC)	CVAD (AUC)
<i>Classical (non-LM) semi-supervised methods</i>						
FutureFrame [18]	2018	95.4%	85.1%	72.8%	-	-
MemAE [80]	2019	94.1%	83.3%	71.2%	-	-
<i>LM-based semi-supervised methods</i>						
HyCoVAD [81]	2025	-	-	-	-	72.5%
MLLM-EVAD [79]	2025	-	88.4%	-	-	71.0%
SlowFastVAD [82]	2025	99.1%	89.6%	85.0%	72.2%	-
SFN-VAD [83]	2025	98.4%	91.6%	83.0%	-	-
VLAVID [84]	2024	-	87.2%	-	-	-
Follow the Rules [85]	2024	97.9%	89.7%	85.2%	71.90%	-

**FIGURE 2.** Representative semi-supervised LM-based VAD pipeline. The method uses object detection and tracking together with an MLLM to generate textual descriptions of object activities and interactions from normal videos. During inference, new descriptions are compared with learned normality exemplars to detect semantic deviations. Adapted from [79].

A. UNSUPERVISED AND SEMI-SUPERVISED VAD

The boundary between unsupervised and semi-supervised video anomaly detection is often blurred in the VAD literature. Many methods described as “unsupervised” are in fact trained on videos that are assumed to contain only normal events, which corresponds more closely to one-class or semi-supervised anomaly detection in the broader machine learning terminology. Conversely, some recent language-model-based methods operate with unlabeled or weakly constrained data, but still rely on normal reference samples, scene-specific prompts, or induced normality rules. For this reason, we discuss unsupervised and semi-supervised LM-based VAD together under a broader normality-learning perspective.

In this setting, the central assumption is that anomalous events are not explicitly available during training. Instead, the model learns, constructs, or infers a representation of normal

behavior, and anomalies are detected as deviations from this representation during inference. Classical approaches typically define such deviations through reconstruction error, prediction error, likelihood, memory retrieval, or distance from a learned normal manifold. Language-model-based methods extend this paradigm by shifting normality modeling from purely visual statistics toward semantic representations, textual descriptions, vision-language similarity, rule induction, and multimodal reasoning.

Early normality-learning VAD methods primarily detected anomalies as deviations from learned regular visual patterns, using reconstruction, prediction, memory, or sparse representation mechanisms. Lu et al. [20] introduced an efficient sparse-reconstruction-based approach, where normal patterns are compactly represented and abnormal events are identified through high reconstruction cost. Hasan et al. [17] later used autoencoders to learn temporal regularity in video sequences, treating large reconstruction or regularity errors as evidence that an event violates learned spatio-temporal patterns. Liu et al. [18] shifted this idea toward future-frame prediction, where anomalies are detected through large prediction errors on events that cannot be reliably forecast from normal training videos. Memory-based extensions further improved this paradigm: Gong et al. [80] introduced MemAE, which constrains reconstruction through a memory module containing prototypical normal patterns to reduce the risk of reconstructing anomalies too well, while Liu et al. [86] combined memory-augmented optical-flow reconstruction with flow-guided frame prediction to better capture normal appearance-motion consistency. Together, these works established the dominant foundation of unsupervised and semi-supervised VAD: anomalous events are not directly modeled during training, and detection depends on how effectively the learned normality representation separates regular behavior from abnormal deviations.

A first group of LM-based normality-learning methods shifts anomaly detection from low-level visual reconstruction toward language-aligned semantic representations of normal behavior. Instead of modeling normality only through appearance, motion, or reconstruction errors, these methods use

VLMs or MLLMs to represent frames, objects, trajectories, or interactions in a semantic space, and then detect anomalies as rare, inconsistent, or distant semantic patterns. VLAVID [84] is an early example of this direction. It detects and tracks objects, queries a pretrained VLM with selected prompts for each object crop, and uses the resulting semantic responses to describe object appearance, pose, and action. A Selective Prompt Adapter selects the prompt that best concentrates normal samples in the semantic space, while a Sequence State Space Module predicts future semantic embeddings from normal object-level sequences. Anomalies are therefore detected through both rare semantic responses and temporal inconsistencies between predicted and observed semantic states. Figure 2 illustrates a representative semi-supervised LM-based VAD pipeline, where object-centric visual evidence is converted into textual descriptions and compared against normality exemplars. MLLM-EVAD [79] follows a related object-centric direction, but uses MLLM-generated descriptions of single-object activities and object-pair interactions as scene-specific normality exemplars. At inference time, new descriptions are compared with compact nominal exemplar sets, allowing anomaly scores to be computed from semantic distance while also providing textual evidence for the detected abnormality. Kim et al. [87] provide a simpler CLIP-based variant of semantic normality modeling, where frames or object crops are compared with predefined natural-language descriptions of normal and abnormal situations generated with ChatGPT and adapted to the target domain. Their method further learns a text-conditional similarity function without frame- or video-level anomaly labels, making it computationally simpler than object-trajectory-based approaches, although less explicit in modeling long-range temporal dynamics.

A complementary direction retains the classical prediction-based normality-learning paradigm, but extends it with semantic consistency. SFN-VAD [83] combines appearance, motion, and language-derived semantic cues for semi-supervised VAD. The method uses a VLM to generate textual descriptions of input clips, encodes them as semantic features, and fuses them with visual appearance and frame-difference-based motion features through a tri-modal encoder. A multimodal decoder then predicts both the next frame and the corresponding semantic representation, so anomaly scores can be computed from visual prediction errors as well as semantic prediction errors. To reduce excessive generalization to abnormal events, SFN-VAD further introduces a Sparse Feature Filtering Module with bottleneck filters and mixture-of-experts routing. In this way, it extends reconstruction- and prediction-based normality learning beyond low-level appearance and motion by explicitly modeling whether future visual content and future semantic descriptions remain consistent with normal behavior.

Another group of LM-based normality-learning methods uses LLMs and VLMs to induce, retrieve, or validate semantic rules about normal and abnormal behavior. AnomalyRuler [85] is representative of this direction. Given a few

normal reference frames, a VLM first converts the scene into textual descriptions, and an LLM then derives scenario-specific rules describing normal human activities, environmental objects, and possible abnormal deviations. During inference, test-frame descriptions are matched against these induced rules and refined through perception smoothing and robust LLM-based reasoning. This induction–deduction formulation allows the method to adapt to different scenes without anomalous training examples or full-shot detector training. SlowFastVAD [82] extends this idea with a RAG-enhanced vision-language pipeline: a fast reconstruction-based detector first identifies ambiguous segments, and only these segments are passed to a slower VLM-based module that retrieves normal and potential abnormal patterns from a scenario-specific knowledge base to generate anomaly scores and interpretable reasoning. HyCoVAD [81] follows a related detect–then–validate strategy for complex interaction anomalies, where a self-supervised detector first selects suspicious frames from low-level spatiotemporal deviations, and a VLM/LLM-based validation stage uses refined captions and environment-specific rules to decide whether the event is anomalous. Together, these methods show that LLMs can support semi-supervised or one-class VAD not only by providing semantic representations, but also by constructing explicit normality rules, retrieving contextual knowledge, and applying reasoning selectively to reduce the cost of dense MLLM inference.

Overall, LM-based unsupervised and semi-supervised VAD methods extend normality learning beyond reconstruction, prediction, and memory-based modeling by incorporating semantic representations, textual descriptions, rule induction, retrieval, and MLLM-based reasoning. The quantitative results summarized in Tables 4–6 are reported values taken from the corresponding original publications; they were not reproduced under a unified experimental setup for this survey. This follows the common convention in the VAD literature, where comparison tables typically report results from the source papers and independently reproduced results are explicitly identified when applicable. The standard VAD benchmarks considered in these tables generally provide established training and testing partitions, and methods evaluated on the same benchmark typically use these common splits. Nevertheless, the numbers should be interpreted as representative reference points rather than as strictly controlled cross-method comparisons. Differences in feature extractors, pretrained backbones, frame or clip sampling and frame rate, temporal processing, anomaly-score smoothing and other post-processing, external training data, and implementation details can influence the reported performance even when the same benchmark and evaluation metric are used. Table 4 summarizes representative semi-supervised VAD methods on commonly used benchmarks, including Ped2, CUHK Avenue, ShanghaiTech, UBnormal, and ComplexVAD, comparing classical non-LM baselines with recent LM-based methods.

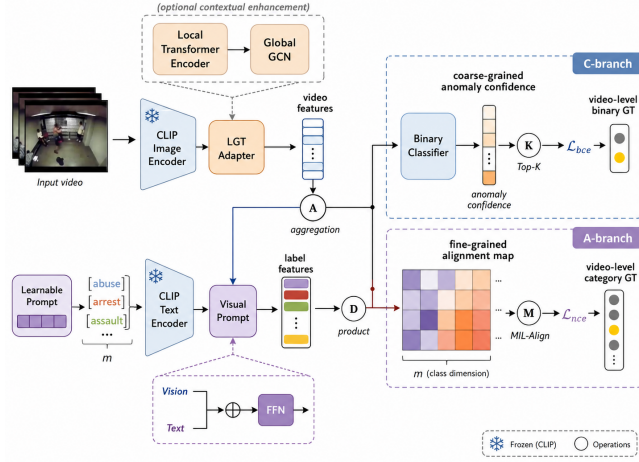


FIGURE 3. Representative weakly supervised LM-based VAD pipeline. VadCLIP adapts CLIP to video anomaly detection by combining visual temporal modeling with text-label alignment, learnable prompts, and MIL-based supervision from video-level anomaly labels. Independently redrawn and visually reorganized by the authors based on the method described in [88].

TABLE 5. WVAD methods on UCF-Crime (UCF), XD-Violence (XD), and ShanghaiTech (SHTech). Values are taken from the corresponding source publications.

Method	Year	UCF (AUC)	XD (AP)	SHTech (AUC)
<i>Classical (non-LM) weakly-supervised methods</i>				
MSL [89]	2022	85.62%	78.59%	97.32%
MIST [90]	2021	82.30%	-	94.83%
GCN [91]	2019	82.12%	-	84.44%
DeepMIL [19]	2018	75.40%	-	-
<i>LM-based weakly-supervised methods</i>				
DAKD [92]	2025	88.34%	85.61%	98.10%
LEC-VAD [93]	2025	89.97%	86.56%	-
Ex-VAD [94]	2025	88.29%	86.52%	-
Federated-WVAD [95]	2025	84.03%	75.99%	-
LOP-VAD [96]	2025	88.07%	86.18%	-
TCRFL [97]	2025	87.31%	82.63%	98.12%
FSA-VAD [98]	2025	-	-	97.66%
VadCLIP++ [99]	2025	88.12%	85.03%	-
ReVid [100]	2025	87.71%	80.76%	-
AVadCLIP [101]	2025	-	86.04%	-
Multimodal-VAD [102]	2025	87.96%	86.32%	-
PEMIL [103]	2024	86.76%	85.59%	98.14%
VadCLIP [88]	2024	88.02%	84.51%	-
CLIP-TSA [104]	2024	87.58%	82.19%	98.32%
ReFLIP-VAD [105]	2024	89.14%	86.29%	-
IELD-WVAD [106]	2024	89.67%	85.92%	-
DWFF-VAD [107]	2024	88.58%	85.58%	-
CMSIL [108]	2024	87.57%	86.06%	-
WSVAD-LLMKE [109]	2024	86.88%	87.05%	98.25%
TEVAD [110]	2023	85.30%	-	-

B. WEAKLY SUPERVISED VIDEO ANOMALY DETECTION

Weakly supervised video anomaly detection addresses the setting where training videos are labeled only at the video level, indicating whether each video contains an anomaly, while temporal or spatial annotations of anomalous regions are unavailable. This formulation is practically appealing because video-level labels are much easier to obtain than frame-level, segment-level, or pixel-level annotations, espe-

cially for long surveillance videos in which anomalous events are sparse and temporally localized. However, the lack of precise temporal supervision makes the task challenging: the model must infer which frames or snippets are responsible for the video-level abnormal label while avoiding overfitting to background, scene bias, or irrelevant contextual cues.

A seminal work in this setting is Sultani et al. [19], which formulated weakly supervised VAD as a multiple instance learning (MIL) problem. In this formulation, each video is treated as a bag of temporal instances, and abnormal videos are assumed to contain at least one anomalous instance, whereas normal videos contain only normal instances. The model is trained to assign higher anomaly scores to the most abnormal snippets in abnormal videos than to snippets in normal videos. This MIL-based formulation has strongly influenced subsequent weakly supervised VAD methods. Later works improved temporal localization by designing stronger feature representations, attention mechanisms, temporal context modeling, graph-based reasoning, pseudo-label refinement, and feature magnitude learning. For example, methods such as Zhong et al. [91], RTFM [111], MGFN [112], and related approaches further refined the weakly supervised paradigm by addressing noisy video-level labels, improving snippet discrimination, and modeling temporal dependencies more effectively.

Language-model-based weakly supervised VAD builds upon this MIL foundation but introduces language as an additional source of semantic guidance. Instead of relying only on visual features and video-level binary labels, recent methods use vision-language models, textual prompts, anomaly descriptions, captions, audio-visual-language cues, or LLM-derived knowledge to improve the alignment between visual events and anomaly semantics. This is particularly useful in weak supervision because the model does not observe exact anomaly boundaries during training; language can provide higher-level concepts such as ‘fighting,’ ‘explosion,’ ‘running,’ or ‘suspicious behavior’ that help distinguish abnormal snippets from normal context. As a result, weakly supervised VAD has become one of the most active settings for language-model-based anomaly detection, with methods ranging from CLIP-based temporal localization and prompt learning to caption-guided reasoning, multimodal fusion, and LLM-enhanced knowledge transfer.

A major line of weakly supervised language-model-based VAD adapts pretrained vision-language models, particularly CLIP, to temporal anomaly localization. A representative weakly supervised CLIP-based pipeline is shown in Figure 3. VadCLIP [88] introduces a dual-branch framework that combines a conventional visual classification branch with a vision-language alignment branch. The classification branch produces coarse-grained frame-level anomaly scores, while the alignment branch compares frame-level visual features with textual anomaly labels encoded by CLIP, enabling fine-grained anomaly recognition. To bridge the gap between image-level CLIP pretraining and video-level anomaly detection, VadCLIP introduces a local-global temporal adapter

for temporal modeling, learnable textual prompts for adapting class labels to the VAD task, and anomaly-focused visual prompts that use visual context from likely abnormal snippets to refine text representations. It further proposes a MIL-Align objective to optimize video-text alignment under video-level supervision. Following this direction, later methods such as VadCLIP++ [99], WSVAD-CLIP [113], CLIP-TSA [104], CMSIL [108], ReFLIP-VAD [105], Anomaly-CLIP [114], and VLIAL [105] further explore temporal modeling, prompt learning, feature refinement, multi-scale instance learning, and instance-aware vision-language alignment. Together, these methods demonstrate that CLIP-style visual-language representations can provide stronger semantic cues than visual-only MIL models, while still requiring task-specific temporal modules to localize anomalies from video-level labels.

Another group of weakly supervised methods focuses on prompt learning and text-prompt-enhanced MIL. Rather than using language only as fixed class names, these approaches design or learn prompts that provide additional semantic guidance for separating normal and abnormal snippets under video-level supervision. PEL [103] is representative of this direction. It combines a Temporal Context Aggregation module for efficient local-global temporal modeling with a Prompt-Enhanced Learning module that constructs knowledge-based prompts from external commonsense concepts. These prompt features are used during training to align anomalous visual contexts with semantically related textual concepts while pushing non-anomalous contexts away, thereby improving fine-grained discriminability among anomaly categories. Related methods [95], [96], [115]–[118] further explore spatio-temporal prompts, prompt-enhanced MIL objectives, opposite or suspected-anomaly prompts, multilingual prompt guidance, and federated multimodal prompt learning. Overall, this family shows that prompt design can act as an intermediate semantic constraint between weak video-level labels and frame-level anomaly localization.

A related direction uses captions, textual clues, and semantic summaries to provide higher-level guidance for weakly supervised VAD. Unlike prompt-learning methods that mainly adapt class names or prompt templates, these approaches use language to describe video content more explicitly and to enrich visual representations with event-level semantics. TEVAD [110] is representative of this line of work. It first generates dense captions for video snippets using a pretrained captioning model, encodes the captions into sentence embeddings, and processes both textual and visual features with temporal networks before multimodal fusion and MIL-based anomaly scoring. By incorporating caption-derived semantic features, TEVAD improves detection performance while also providing a degree of interpretability through the contribution of individual caption words to anomaly scores. Subsequent semantic-guided methods [93], [97], [98], [119] further explore the use of injected text clues, semantic filtering and summarization, event completeness modeling, and representative temporal feature learning. Overall, this family high-

lights the role of language as descriptive evidence, complementing visual MIL models with semantic information that is difficult to capture from spatio-temporal features alone.

Another line of weakly supervised language-model-based VAD moves beyond CLIP-style feature adaptation and uses LLMs or MLLMs for knowledge enhancement, token alignment, distillation, and explainability. Ex-VAD [94] is representative of this direction. It first uses a VLM to generate frame-level captions and an LLM to convert them into video-level anomaly explanations, which are then used both as interpretable outputs and as textual features for multimodal anomaly detection. In addition, Ex-VAD augments anomaly category labels with LLM-generated descriptive phrases and aligns these enhanced label embeddings with fused visual-textual features to support fine-grained anomaly classification. CALLM [120] explores a related but more cascaded design, where a 3D autoencoder first identifies suspicious frames or videos and a Video-LLaMA branch is then used to provide abnormality decisions and explanations through weakly supervised pseudo-instruction tuning. Other methods in this family, [92], [106], [109], [121] investigate how language-model knowledge, anomaly-relevant token alignment, and knowledge distillation can improve weakly supervised anomaly localization. Overall, these methods show a shift from using language only as class-level semantic anchors toward using LLMs and MLLMs as sources of explanatory reasoning, enriched anomaly knowledge, and transferable supervision.

Another extension of weakly supervised language-model-based VAD incorporates additional modalities, especially audio, to improve robustness in visually ambiguous scenes. Audio-visual fusion has also been explored in conventional non-LM VAD, providing useful methodological context for recent language-based multimodal systems. For example, M2VAD [122] combines multi-view visual information with audio in a transformer-based weakly supervised framework, demonstrating the complementary value of acoustic cues for anomaly representation. AVadCLIP [101] is a language model based representative of this direction. Building on the CLIP-based weakly supervised paradigm, AVadCLIP introduces audio-visual adaptive fusion to combine visual features with audio features in a lightweight manner while keeping the CLIP backbone frozen. It further designs an audio-visual prompt that injects multimodal information into textual label embeddings, improving the alignment between anomalous video content and anomaly categories. To handle practical cases where audio may be unavailable at inference time, AVadCLIP also introduces uncertainty-driven feature distillation, transferring audio-visual knowledge to a visual-only student model. Related multimodal methods [102], [107], [123] further explore audio-vision-language fusion, heterogeneous cross-modal knowledge transfer, and adaptive multimodal feature integration. Overall, this family shows that language-based VAD can benefit from multimodal cues beyond vision, particularly for anomalies with strong acoustic signatures or degraded visual evidence.

TABLE 6. TF-VAD methods on UCF-Crime (UCF), XD-Violence (XD) and UBnormal (UB). Values are taken from the corresponding source publications.

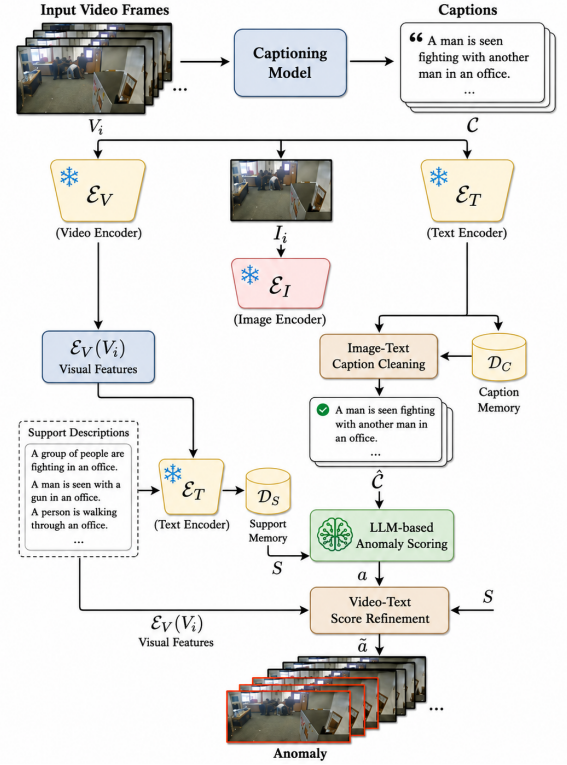
Method	Year	UCF (AUC)	XD (AP)	UB (AUC)
<i>Open World and Instruction Tuning Methods</i>				
OVVAD [127]	2024	86.40%	66.53%	62.94%
Anomize [128]	2025	84.89%	69.31%	-
PLOVAD [129]	2025	87.06%	-	-
Holmes-VAU [45]	2025	88.96%	87.68%	-
Holmes-VAD [130]	2024	89.51%	90.67%	-
<i>Full training-free pipelines</i>				
QVAD [131]	2026	84.28%	68.53%	79.60%
Flashback [132]	2025	87.30%	75.10%	-
TFPLG [133]	2025	87.52%	85.08%	-
MoniTor [134]	2025	82.57%	55.01%	-
EventVAD [135]	2025	82.03%	64.04%	-
PANDA [136]	2025	84.89%	70.16%	75.78%
VADTree [137]	2025	84.74%	68.85%	-
VERA [138]	2025	86.55%	70.11%	-
SUVAD [139]	2025	83.90%	70.10%	-
MCANet [140]	2024	82.47%	69.72%	62.94%
LAVAD [126]	2024	80.28%	62.01%	64.23%

Finally, several recent works extend weakly supervised language-model-based VAD beyond standard MIL-style detection by introducing relation-aware, structured, or retrieval-oriented formulations. RelVid [100] uses CLIP-based visual and textual representations together with auxiliary anomaly recognition and reconstruction objectives to improve class separation and temporal feature learning. MIS-SIONGNN [124] further incorporates structured reasoning by generating mission-specific knowledge graphs with an LLM and ConceptNet, then applying hierarchical GNNs over these graphs for weakly supervised anomaly recognition. In a different direction, ALAN [125] shifts from anomaly detection to video anomaly retrieval, where long untrimmed videos are retrieved using textual descriptions or audio queries; it introduces anomaly-led sampling and cross-modal alignment to focus retrieval on relevant anomalous segments. Although these methods differ from the main CLIP, prompt, and caption-based detection pipelines, they illustrate how language can support broader forms of anomaly understanding, including semantic relation modeling, knowledge-guided reasoning, and retrieval from descriptive queries.

Overall, LM-based weakly supervised VAD methods extend the classical MIL formulation by incorporating vision-language alignment, prompt learning, caption-derived semantics, LLM-enhanced anomaly knowledge, multimodal fusion, and relation-aware reasoning. Table 5 summarizes representative weakly supervised VAD methods on UCF-Crime, XD-Violence, and ShanghaiTech, comparing classical non-LM baselines with recent LM-based approaches.

C. TRAINING-FREE VIDEO ANOMALY DETECTION

Training-free language-model-based VAD studies the use of frozen LLMs, VLMs, or MLLMs for anomaly detection

**FIGURE 4.** Representative training-free LM-based VAD pipeline. LAVAD converts video frames into captions, cleans noisy descriptions using vision-language similarity, applies LLM-based temporal anomaly scoring, and refines scores through video-text alignment without task-specific training. Independently redrawn and visually reorganized by the authors based on the method described in [126].

without task-specific fine-tuning or optimization on VAD datasets. Unlike weakly supervised methods, which typically train MIL-based detectors using video-level labels, training-free approaches rely on the pretrained semantic and reasoning capabilities of foundation models. In this setting, anomaly detection is often performed through caption generation, prompt-based reasoning, text-video similarity, rule-based scoring, temporal aggregation of language descriptions, or direct MLLM questioning. This paradigm is attractive because it reduces annotation and training costs, improves flexibility across scenes and anomaly categories, and can provide natural-language explanations alongside anomaly scores.

However, training-free VAD also introduces distinct challenges. Since the models are not adapted to the target surveillance domain, their performance depends heavily on the quality of visual descriptions, prompt design, temporal sampling, and the ability of the language model to distinguish rare abnormal events from unusual but benign activities. In addition, frozen MLLMs may hallucinate objects or actions, overlook subtle temporal changes, or produce inconsistent judgments across frames. Therefore, recent training-free methods differ mainly in how they convert video into language-aware evidence and how they aggregate this evidence into reliable anomaly decisions. Representative works explore frame or clip captioning with LLM-based temporal reasoning, ver-

balized anomaly scoring, customizable zero-shot anomaly definitions, event-aware prompting, memory-based real-time reasoning, and multimodal caption-aware analysis.

Figure 4 shows a representative training-free LM-based VAD pipeline. LAVAD [126] is a representative method in the training-free language-model-based VAD paradigm. Instead of training a task-specific anomaly detector, LAVAD uses frozen foundation models to convert video anomaly detection into a language-based reasoning problem. The method first generates frame-level captions using an off-the-shelf VLM-based captioning model and then applies image-text similarity to replace noisy captions with more semantically aligned descriptions from the same video. To incorporate temporal context, an LLM summarizes captions within a temporal window and assigns anomaly scores based on the resulting scene description. Finally, LAVAD refines these LLM-generated anomaly scores by aggregating scores from semantically similar video-text pairs using cross-modal similarity. This design allows LAVAD to perform temporal anomaly localization without VAD-specific training data, while also highlighting the main challenges of training-free methods: dependence on caption quality, prompt design, temporal summarization, and the reliability of frozen VLM/LLM representations.

Following LAVAD, several works further improve training-free VAD by strengthening semantic reasoning, explainability, and temporal/event modeling. VERA [138] addresses explainable VAD by learning natural-language guiding questions that elicit anomaly-relevant reasoning from frozen VLMs, avoiding instruction tuning or model-parameter updates while still adapting the prompt space using coarsely labeled data. SUVAD [139] instead emphasizes semantic understanding with MLLMs: it generates textual descriptions of normal and abnormal events, compares test-video descriptions against these event lists, and supports flexible adjustment of anomaly definitions together with explanatory outputs. EventVAD [135] focuses on the temporal limitation of frame-level MLLM reasoning by segmenting long videos into event-consistent units through dynamic spatiotemporal graph modeling, statistical boundary detection, and hierarchical prompting before anomaly scoring. MCANet [140] extends this direction with multimodal caption-aware analysis, combining visual, audio, and language cues for training-free anomaly detection. Together, these methods show that training-free VAD is moving from simple caption-to-score pipelines toward more structured semantic reasoning, where prompts, event boundaries, captions, and multimodal cues are used to make frozen foundation models more reliable for temporal anomaly localization.

Another group of training-free methods focuses on zero-shot and customizable anomaly reasoning, where the anomaly definition can be specified through natural language rather than fixed by training data. AnyAnomaly [141] is representative of this direction. It formulates customizable VAD, where user-provided text defines the abnormal event to be detected, and performs segment-level context-aware VQA with an LVLM without fine-tuning. To reduce latency and

improve temporal reasoning, AnyAnomaly selects key frames from each video segment, constructs position context to emphasize text-relevant regions, and builds temporal context in a grid format to capture action changes over time. This design enables the method to detect user-defined abnormal objects, actions, or behaviors across different environments without retraining. More broadly, Anomaly-OV [142] studies zero-shot anomaly detection and reasoning with MLLMs in image-based anomaly inspection, introducing anomaly-specific instruction data and a specialist visual assistant that selects suspicious visual tokens for detection and explanation. Although Anomaly-OV is not a standard VAD method, it reflects a related trend toward open-ended anomaly reasoning, where foundation models are expected not only to score abnormality but also to describe fine-grained anomalous evidence and explain why an observation is abnormal. Related works further extend this direction through hybrid prompt personalization [143], CLIP-assisted zero-shot anomaly scoring [144], and generic GPT-4V-based anomaly understanding [145]. Together, these methods broaden training-free VAD from fixed benchmark anomaly categories toward user-defined, prompt-driven, and explanation-oriented anomaly reasoning.

Beyond general training-free and customizable VAD, recent works also explore memory-based, real-time, pseudo-labeling, agentic, and application-specific extensions. Flashback [132] is representative of the memory-driven and real-time direction. Instead of invoking an LLM during inference, Flashback first constructs an offline pseudo-scene memory by using an LLM to generate normal and anomalous scene captions, which are then embedded with a frozen cross-modal encoder. During online inference, incoming video segments are matched against this memory through similarity search, and the retrieved captions provide both anomaly scores and textual explanations. This retrieval-based design removes expensive online LLM calls and enables real-time zero-shot VAD, while repulsive prompting and scaled anomaly penalization are used to reduce bias toward anomalous captions. Other works extend training-free LM-based VAD in complementary ways: training-free VLM-based pseudo-label generation [133] uses foundation models to generate auxiliary supervision for downstream VAD, VisionGPT [146] applies LLM-assisted anomaly detection to safe visual navigation, PANDA [136] and Qvad [131] explores agentic AI pipelines for generalist VAD, and robotic VAD [147] studies VLM-based anomaly detection in scientific laboratory environments. These extensions show that training-free language-model-based VAD is expanding from benchmark-centered anomaly localization toward practical deployment settings involving real-time constraints, pseudo-supervision, agentic automation, and domain-specific anomaly understanding.

Overall, training-free LM-based VAD methods shift anomaly detection from task-specific detector training toward frozen foundation-model reasoning, caption-based scoring, retrieval, prompting, and agentic pipelines. Table 6 summarizes representative training-free and closely related LM-based VAD methods on UCF-Crime, XD-Violence, and UB-

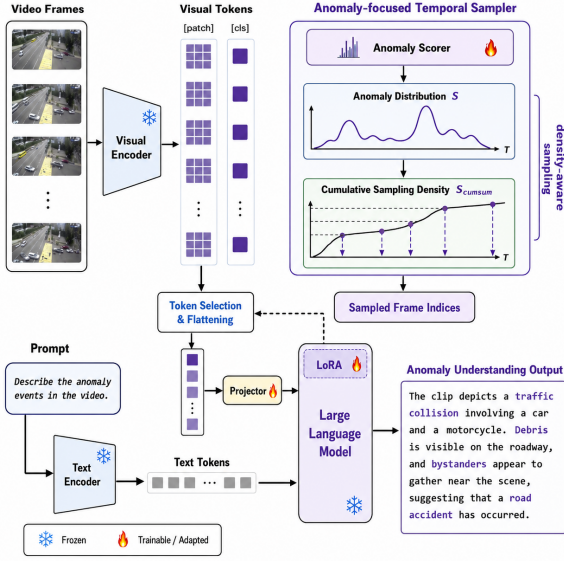


FIGURE 5. Representative instruction-tuned MLLM-based VAD pipeline. Holmes-VAU uses an anomaly-focused temporal sampler to select informative frames from long videos and fine-tunes a multimodal language model for anomaly understanding and natural-language explanation. Independently redrawn and visually reorganized by the authors based on the method described in [45].

normal, including full training-free pipelines as well as open-world/open-vocabulary and instruction-tuned methods for broader comparison.

D. INSTRUCTION TUNING BASED VIDEO ANOMALY DETECTION

Instruction-tuned language-model-based VAD adapts LLMs or MLLMs to anomaly detection through task-specific instruction data, rather than relying solely on frozen foundation models at inference time. In contrast to training-free methods, which use pretrained models through prompting, captioning, or retrieval without updating model parameters, instruction-tuned approaches explicitly align the model with VAD-oriented objectives such as anomaly prediction, temporal localization, abnormal event description, and explanation generation. The required supervision may come from manually annotated instructions, video-level labels, pseudo-instruction data, generated captions, question-answer pairs, or anomaly explanations. This paradigm is attractive because instruction tuning can make MLLMs more responsive to anomaly-specific queries and better suited for complex reasoning, dialogue-style interaction, and fine-grained abnormal event understanding. However, it also reintroduces dependence on curated or generated training data, making these methods less annotation-free than training-free approaches and more sensitive to the quality, coverage, and bias of the instruction data used for adaptation.

A first group of instruction-tuned methods focuses on improving anomaly detection while producing natural-language explanations or interactive responses. Holmes-VAD [130] is representative of this direction. It constructs VAD-

Instruct50K, a multimodal instruction-tuning benchmark that combines single-frame temporal annotations, event-level video clips, captions, and anomaly-aware explanatory conversations. Based on this dataset, Holmes-VAD trains a temporal sampler to select anomaly-responsive frames from long untrimmed videos and fine-tunes a multimodal LLM with LoRA to generate anomaly judgments and explanations. In this way, the model addresses both ‘when’ the anomaly occurs and ‘what/why’ the event is abnormal, reducing the gap between score-based VAD and interpretable anomaly analysis. HAWK [148] extends this idea toward open-world anomaly understanding by fine-tuning a VLM on anomaly descriptions and question-answer pairs collected from multiple VAD datasets, while explicitly incorporating motion information and motion-language supervision to better describe abnormal events. AssistPDA [149] further shifts the setting from offline analysis to online surveillance assistance, unifying anomaly prediction, detection, and analysis in streaming videos through VAPDA-127K and a spatiotemporal relation distillation module. Together, these methods show that instruction tuning can transform language-model-based VAD from a frame- or segment-level scoring task into an interactive anomaly-understanding system capable of localization, explanation, question answering, and real-time assistance.

A second approach emphasizes anomaly reasoning rather than only detection or explanation generation. Vad-R1 [150] is representative of this direction. It introduces the task of Video Anomaly Reasoning, where an MLLM is expected to explicitly reason about abnormal events before producing the final answer. To guide this process, Vad-R1 designs a Perception-to-Cognition Chain-of-Thought, which first performs global and local perception of the video scene and suspicious clips, and then moves to shallow and deep cognition by identifying the abnormal event, explaining why it violates expected behavior, and reasoning about possible consequences. Based on this structure, the authors construct Vad-Reasoning, a dataset containing CoT-style anomaly reasoning annotations as well as weak video-level labels. The model is trained in two stages: supervised fine-tuning on reasoning-annotated videos, followed by reinforcement learning with AVA-GRPO, which introduces an anomaly verification reward by checking whether removing the predicted abnormal segment changes the model’s judgment. VAU-R1 [151] further develops this reasoning-oriented direction through reinforcement fine-tuning. It decomposes video anomaly understanding into multiple-choice question answering, temporal anomaly grounding, anomaly reasoning, and anomaly classification, and applies GRPO with task-specific rewards for output format, answer correctness, and temporal localization accuracy. Compared with standard supervised fine-tuning, this reinforcement-tuned formulation aims to improve both structured reasoning and generalization across anomaly scenarios. Together, Vad-R1 and VAU-R1 extend instruction-tuned VAD from generating post-hoc descriptions toward explicit anomaly reasoning, where temporal localization, anomaly category prediction, causal explanation, norm-

violation analysis, and reward-guided reasoning are jointly considered.

A third group extends instruction-tuned language-model-based VAD toward long-term, multi-scene, and multi-granular anomaly understanding. Figure 5 illustrates an instruction-tuned MLLM-based VAD pipeline with anomaly-focused temporal sampling. Holmes-VAU [45] is representative of the long-term and any-granularity direction. It introduces HIVA-70K, a hierarchical instruction benchmark with clip-level, event-level, and video-level annotations, enabling models to learn both short-term visual perception and longer-term anomaly reasoning. To process long videos efficiently, Holmes-VAU further proposes an Anomaly-focused Temporal Sampler that uses anomaly scores to adaptively select informative frames, allowing the MLLM to focus on anomaly-rich regions rather than uniformly sampled frames. This design extends instruction-tuned VAD from isolated clip-level explanation toward hierarchical video anomaly understanding across multiple temporal scales. Sherlock [152] addresses a related but more structured formulation through the Multi-scene Video Abnormal Event Extraction and Localization task, where the model must localize the abnormal event and extract a quadruple consisting of subject, event type, object, and scene. To support this task, Sherlock introduces a global-local spatial-sensitive LLM with spatial experts for action, object relation, background, and global context, together with a spatial imbalance regulator to balance these heterogeneous cues. Together, Holmes-VAU and Sherlock show that instruction-tuned anomaly models are moving beyond binary detection and single-clip explanation toward long-video understanding, multi-granularity reasoning, multi-scene event localization, and structured semantic extraction.

E. OPEN-WORLD AND OPEN-VOCABULARY BASED VIDEO ANOMALY DETECTION

Open-world and open-vocabulary VAD relax the closed-set assumption of conventional VAD, where anomaly categories are predefined and fixed by the benchmark. In real-world deployments, abnormal events may be rare, unseen during training, or described using flexible natural language concepts. Language models and vision-language models are therefore useful because they connect visual evidence with textual anomaly semantics through prompts, category descriptions, external knowledge, or user-defined anomaly definitions. In this setting, the main challenge is not only localizing anomalous segments, but also recognizing, describing, or reasoning about anomaly categories beyond the fixed training label space.

A representative starting point for this direction is OVVD [127], which formulates open-vocabulary video anomaly detection as the joint problem of detecting and categorizing both seen and unseen anomalies. Unlike open-set VAD, which mainly aims to detect unseen anomalies without assigning them specific semantic categories, OVVD requires the model to produce frame-level anomaly scores while

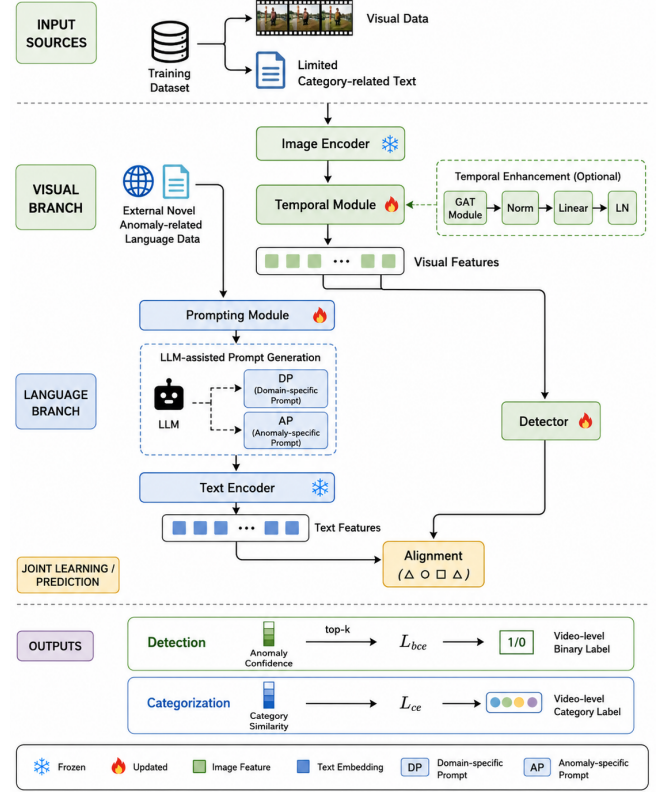


FIGURE 6. Representative open-vocabulary VAD pipeline. The framework combines class-agnostic anomaly detection with class-specific vision-language alignment, allowing the model to detect and categorize both seen and unseen anomaly types using textual anomaly concepts. Independently redrawn and visually reorganized by the authors based on the method described in [129].

also recognizing the anomaly category from an expandable label space. To address this problem, the method decouples OVVD into two complementary components: class-agnostic detection and class-specific categorization. For detection, it uses CLIP visual features together with a lightweight temporal adapter and a semantic knowledge injection module that introduces normal and abnormal textual concepts from large language models. For categorization, it aligns video features with textual anomaly category embeddings and further introduces a novel anomaly synthesis module, where LLMs and generative models are used to create pseudo unseen anomaly samples. This design shows how vision-language models can extend VAD beyond closed-set abnormality scoring by combining temporal modeling, semantic knowledge, and generated novel-category supervision.

Following OVVD, later works further improve open-vocabulary and open-world VAD by strengthening prompt adaptation, uncertainty modeling, and novel-category alignment. As illustrated in Figure 6, PLOVD [129] adapts pretrained image-based vision-language models to OVVD through prompt tuning rather than relying on generated pseudo-anomaly videos. It introduces a prompting module with a learnable domain-specific prompt and an LLM-generated anomaly-specific prompt, allowing the model to

capture both dataset-specific knowledge and semantic descriptions of anomaly categories. A GAT-based temporal module is further used to incorporate temporal dependencies into frame-wise VLM features, bridging the gap between image-level pretraining and video-level anomaly localization. In a related open-world weakly supervised setting, MEL-VLP [153] uses CLIP-derived visual and textual features together with multi-scale temporal visual modeling and multimodal evidential collaborative learning. Instead of treating anomaly prediction as a conventional confidence score, MEL-VLP collects visual, textual, and joint-modal evidence to estimate uncertainty and dynamically calibrate anomaly boundaries, which is especially useful when unseen anomalies appear at test time.

More recent methods focus specifically on the remaining challenges of novel anomaly detection and flexible anomaly definitions. Anomize [128] identifies two key issues in OV-VAD: detection ambiguity, where unfamiliar anomaly frames receive unreliable anomaly scores, and categorization confusion, where novel anomalies are misclassified as visually similar base categories. To address these issues, it introduces a text-augmented dual-stream design that combines dynamic temporal cues and static scene cues with corresponding textual information, as well as a group-guided text encoding mechanism that uses label relations to improve alignment between novel videos and novel textual labels. LaGoVAD [154] further broadens the problem from open-vocabulary recognition to language-guided open-world VAD with variable anomaly definitions. Instead of assuming a fixed anomaly category set, LaGoVAD conditions anomaly detection on user-provided natural-language definitions at inference time, allowing the same event to be treated as normal or abnormal depending on context or user requirements. It implements this paradigm using language-guided video-text fusion, dynamic video synthesis, and contrastive learning with hard negative mining, supported by PreVAD [154], a large-scale dataset with multi-level anomaly categories and textual anomaly descriptions. Together, these methods show that open-world and open-vocabulary VAD is moving from recognizing unseen anomaly labels toward more flexible systems that can use language to define, detect, categorize, and reinterpret abnormal events in changing deployment contexts.

VI. DYNAMIC SURVEY MAINTENANCE

Language models have played a dual role in the development of this survey. Their rapid adoption in video anomaly detection has created the need for a focused review of the field, as new LM-based methods, datasets, benchmarks, and evaluation protocols continue to appear at a pace that is difficult to capture through a conventional static survey. At the same time, agentic AI systems have increasingly emerged as a practical paradigm for decomposing complex workflows into specialized machine-learning and maintenance tasks [155]–[159]. The reasoning, retrieval, summarization, and document-editing capabilities of language models therefore provide practical tools for maintaining such a survey

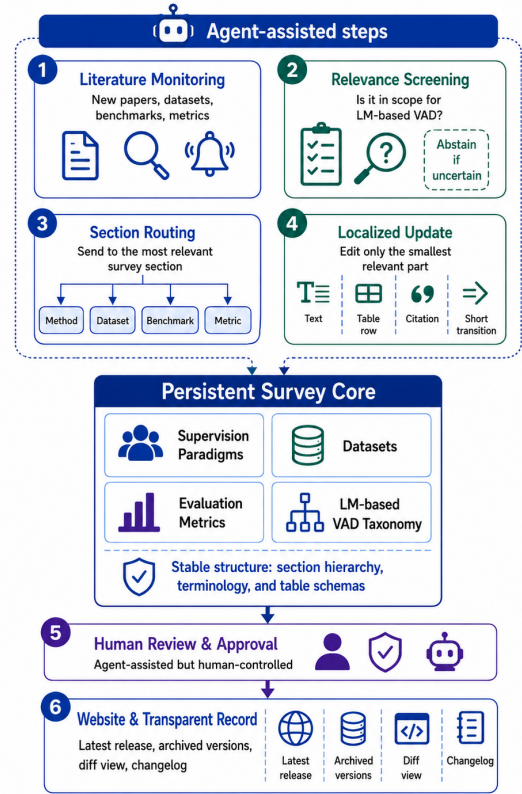


FIGURE 7. Agent-assisted maintenance workflow for the dynamic LM-based VAD survey. Newly published literature is monitored, screened for relevance, routed to the appropriate survey component, and incorporated through localized updates. The persistent survey core preserves the author-defined taxonomy and structure, while human review precedes publication through the versioned website and archived releases.

over time. We use language models not only as the central subject of this review, but also as components of an agent-assisted maintenance process designed to help the survey keep pace with the literature it studies. In this sense, the same technological developments that motivate the survey also enable its dynamic and versioned form.

To support this goal, we adopt the agentic Dynamic Survey Framework proposed in [36]. The framework treats survey writing as a long-term maintenance problem rather than a one-time document-generation task. Instead of repeatedly producing new surveys with overlapping scope, an existing survey is maintained as a persistent scholarly resource whose content evolves through controlled and versioned revisions. The survey retains an author-defined structure, including its section hierarchy, topical scope, taxonomy, terminology, and table schemas, while newly published work is incorporated incrementally through localized updates rather than global rewriting.

For this survey, the stable structure corresponds to the organization introduced in the preceding sections: problem settings and supervision paradigms, datasets, evaluation metrics, and the taxonomy of LM-based VAD methods, including unsupervised and semi-supervised, weakly supervised, training-

free, instruction-tuned, and open-world or open-vocabulary approaches. During maintenance, newly identified papers are first assessed for relevance to language-model-based VAD. This screening step is necessary because not every work involving language models, surveillance video, anomaly terminology, or multimodal reasoning falls within the intended scope. Relevant papers are then routed to the most appropriate section or table according to their primary contribution. For example, a prompt-based zero-shot method may be incorporated into the training-free discussion, a new video-question-answering benchmark may update the dataset section, and a new semantic evaluation protocol may extend the evaluation-metrics discussion.

The update process is deliberately conservative. Once a paper is approved for inclusion, only the smallest relevant portion of the survey should be modified. This may involve adding a new method description, revising an existing comparison, inserting a table entry, updating citation metadata, or adjusting a short transition. Avoiding unnecessary global rewriting helps preserve coherence across versions, including consistent terminology, narrative flow, citation style, and taxonomy. It also reduces the risk that repeated automatic editing gradually changes the survey’s scope or weakens the distinctions between method papers, dataset papers, benchmark contributions, and evaluation-oriented work.

The maintenance workflow is agent-assisted but remains human-controlled. Language-model agents can support literature monitoring, technical summarization, relevance filtering, abstention when evidence is insufficient, section routing, table completion, citation placement, and localized text synthesis. However, decisions that affect the high-level scope or structure of the survey remain under explicit author control. In particular, routine updates do not automatically introduce new top-level paradigms, redefine existing categories, or reorganize the taxonomy. Such structural revisions require human review because emerging research trends may initially appear significant but later prove too narrow, temporary, or overlapping with existing categories. More generally, agent-generated outputs are treated as candidate revisions rather than autonomous changes to the scholarly record. Final paper inclusion, categorization, substantive textual revisions, and publication of each maintained release require explicit author review and approval.

Each maintained release is accompanied by a transparent version record. These records identify newly added papers, revised sections, updated tables, corrected metadata, and any author-approved structural changes. Versioning allows readers to distinguish the original peer-reviewed article from later maintained editions while still treating the survey as a coherent scholarly resource. It also supports reproducibility by making the evolution of the document inspectable rather than silently replacing earlier content. If an error is identified in previously released metadata, categorization, or technical description, the correction is documented in the subsequent version record rather than silently overwriting the historical record, while the earlier release remains archived and accessi-

Dynamic VAD Survey
LIVING SCHOLARLY RESOURCE

Home Survey Explore Datasets Benchmarks Updates Maintenance About Read Survey PDF

A DYNAMIC SURVEY · v1.0

Language Models in Video Anomaly Detection

A continuously maintained, agentic-AI-assisted survey covering methods, datasets, benchmarks, evaluation protocols, and emerging research directions, with human-reviewed releases.

Explore the Survey Browse Methods View latest updates →

182 PAPERS	28 DATASETS	5 PARADIGMS	v1.0 CURRENT VERSION	September 2026 LAST UPDATED
---------------	----------------	----------------	-------------------------	-----------------------------------

HOW TO CITE THIS RESOURCE

**Cite the paper.
Reference the version.**

Use the archival paper's standard BibTeX citation. The maintained website version, URL, and access date are included as additional fields in that same entry.

Full citation guidance →

```

@misc{dynamicvad,
  title={Language Models in Video Anomaly Detection: A Dynamic Survey},
  author={Mumcu, Furkan and Bhatti, Umar and Jones, Michael J. and Charatan, Anoop and Vilnai, Yael},
  journal={IEEE Access},
  year={2024},
  volume={12},
  url={https://dynamicvadsurvey.github.io/},
  note={Website version v1.0, accessed September 21, 2026}
}

```

Reduce the clearly marked author, venue, and archival metadata once the final paper citation is supplied.

FIGURE 8. Public web interface of the dynamic LM-based VAD survey. The website provides access to the latest maintained release, archived versions, citation information, version records, and changes introduced across updates.

ble. Figure 7 summarizes this maintenance process, showing how agent-assisted literature monitoring, relevance screening, section routing, and localized updates are integrated with a persistent survey structure, human approval, and transparent versioned publication.

The maintained survey is accompanied by a publicly accessible website at dynamicvadsurvey.github.io, which serves as the reader-facing interface to the dynamic resource, as illustrated in Figure 8. The website provides access to the latest maintained release as well as archived versions, allowing readers to inspect how the survey evolves over time. It follows the same high-level organization as the paper, including supervision paradigms, datasets, evaluation metrics, representative methods, and references. It also supports convenient filtering and exploration of methods according to attributes such as supervision paradigm, publication year, model type, dataset, task, and methodological category, enabling readers to identify and compare relevant approaches without searching through the full document. Version records summarize newly added papers, revised sections, updated tables, and other substantive changes, while a comparison view highlights additions and removals between selected releases.

For the initial implementation of the maintenance pipeline, we favor small and mid-sized language models rather than relying exclusively on the largest available systems. Recent work suggests that smaller models can be effective for specialized agentic tasks when the workflow is decomposed into well-defined components with constrained inputs and outputs [160]–[163]. This design is consistent with the dynamic survey framework, in which discovery, screening, routing, abstention, synthesis, and verification are handled as separate

operations rather than as a single unconstrained generation task [36]. The initial system therefore uses models from the Gemma [164], [165] and Qwen [166]–[168] families for lightweight maintenance operations. These backbone choices are not fixed: stronger, more efficient, or more reliable models may be substituted in future releases while preserving the same high-level workflow, governance principles, and versioned update protocol.

Beyond keeping the literature coverage current, the maintained records can reveal broader patterns in the evolution of LM-based VAD, including emerging methodological directions, recurring weaknesses, and areas where benchmark design or evaluation remains inadequate. These observations provide a natural connection to the challenges and future research directions discussed in the next section.

VII. CHALLENGES AND FUTURE DIRECTIONS

Recent LM-based VAD methods have expanded the scope of anomaly detection by introducing semantic reasoning, open-vocabulary recognition, and natural-language explanation. At the same time, a growing line of work has begun to examine the limitations and vulnerabilities introduced by this shift. In particular, recent analyses argue that the increasing reliance on multi-scene formulations, weak supervision, and pretrained LLM/MLLM priors may move VAD away from its original goal of modeling scene-specific deviations from normality [169]. These concerns motivate several open challenges for future LM-based VAD research.

Scene-specific normality and the limits of category-based anomaly detection. A central challenge for LM-based VAD is that the semantic knowledge provided by LLMs and MLLMs can encourage a category-based view of anomalies. Many recent methods are effective at recognizing common abnormal event categories, such as fighting, burglary, fire, accidents, or falling [45], [101], [127], [128], [130], [138], [148]. However, anomaly detection is not equivalent to recognizing a predefined set of abnormal actions. The same activity may be normal in one scene and anomalous in another, or even normal in one region of a scene and anomalous in a different region of the same scene. [27], [51] For example, fighting inside a boxing ring is expected, whereas fighting among the audience is anomalous. This distinction highlights the need to model normality as a scene-specific and context-dependent property rather than as a fixed semantic label. Future LM-based VAD should therefore move beyond detecting familiar anomaly categories and instead reason about whether an observed event violates the normal activity patterns of the target environment.

Spatial grounding, localization, and context-dependent reasoning. Many LM-based VAD methods, especially weakly supervised and training-free approaches, operate primarily at the video or frame level [101], [110], [126], [138], [139]. While this formulation is useful for reporting whether an anomaly occurs, it often fails to identify where the anomaly occurs and which objects, people, or interactions are responsible. This limitation is especially important for anomalies

whose abnormality depends on spatial context, such as illegal parking, jaywalking, entering a restricted area, abandoning an object, or interacting with an object in an unusual location [50], [51]. Without spatial grounding, a system may detect an anomalous frame but still fail to provide actionable information to a human operator. Moreover, reliable explanation generally requires localization: a model cannot faithfully explain an anomaly without identifying the visual evidence that supports its decision. Future work should therefore place greater emphasis on object-level, region-level, and track-level anomaly localization, together with evaluation protocols that measure whether the detected anomaly is spatially aligned with the true abnormal event.

Semantic bias, hallucination, and evaluation transparency. The use of large pretrained language and vision-language models introduces both opportunities and risks. These models provide useful semantic priors, but their prior knowledge may dominate the anomaly decision. [127], [128], [169] A model may classify an event as anomalous because the action is generally associated with abnormality in its pretraining distribution, even when the event is normal in the current scene. Conversely, subtle scene-specific anomalies may be missed if they do not correspond to familiar anomaly categories. This semantic-prior bias can therefore manifest in both directions: false positives may arise when generally unusual but scene-appropriate activities are treated as anomalous, while false negatives may occur when context-dependent anomalies do not match familiar anomaly concepts. Scene and background correlations may further bias decisions toward camera, location, or contextual cues rather than the anomalous event itself.

MLLMs may also generate plausible but incorrect explanations, including hallucinated objects, actions, or causal relations [170], [171]. Prompt sensitivity represents another source of instability, since changes in prompt wording, contextual information, or candidate anomaly descriptions may alter the resulting anomaly judgment. Similarly, limited temporal reasoning or sparse frame sampling can cause models to miss anomalies that depend on motion, event ordering, or longer-term temporal context. These issues are particularly important for LM-based VAD because the anomaly decision may depend jointly on visual evidence, temporal context, and language-based reasoning.

This issue is further complicated by the opacity of pretrained models, since their training data are often undisclosed [172]. As a result, it is difficult to determine whether benchmark videos, anomaly classes, or visually similar events have been indirectly observed during pretraining [173], [174]. Such potential benchmark contamination complicates the interpretation of zero-shot and open-world performance, particularly when the extent of overlap between pretraining data and evaluation benchmarks cannot be established. More broadly, contextual and cultural assumptions inherited from large-scale pretraining may influence what a model considers suspicious or abnormal, which is problematic when acceptable behavior varies across scenes, environments, or deploy-

ment contexts.

Another open challenge is the lack of a unified evaluation benchmark for language-based VAD outputs. Existing anomaly-understanding datasets differ substantially in task formulation, annotation style, output format, and evaluation protocol, spanning question answering, explanation generation, temporal grounding, retrieval, and captioning. As a result, semantic performance reported across these benchmarks is generally not directly comparable. Unlike conventional frame-level VAD, where metrics such as AUC and AP are widely established, there is currently no broadly adopted protocol for jointly evaluating the correctness, grounding, faithfulness, and usefulness of language-based anomaly explanations or reasoning. Developing standardized benchmarks and evaluation procedures, potentially combining human assessment with reproducible LLM/MLLM-based judging, is therefore an important direction for future work.

Future evaluations should therefore assess not only detection accuracy, but also explanation faithfulness, calibration, grounding, and robustness to unseen anomaly types. In addition, they should examine prompt robustness, scene dependence, temporal failure cases, false-positive and false-negative behavior, and performance under distribution shifts or intentionally manipulated inputs.

Efficient, explainable, and adaptive LM-based VAD systems. Future LM-based VAD systems should combine the semantic flexibility of language models with explicit models of normality learned from the target scene. Rather than using LMs only as generic anomaly classifiers, future systems can use them as reasoning modules that inspect localized evidence, compare observations against scene-specific normality rules, retrieve similar normal events, and generate grounded explanations. Such systems should also be efficient and extendable, since real deployments may involve many cameras and continuously evolving normal activity patterns. This motivates hybrid designs in which lightweight perception modules perform detection, tracking, and feature extraction, while smaller or task-specialized language models support rule induction, contextual reasoning, explanation, and updating [79], [81]. More broadly, the rise of agentic and dynamic systems points toward VAD frameworks that are not static detectors, but adaptive reasoning pipelines capable of refining their decisions and explaining anomalies in relation to the evolving normality of each scene.

Efficiency, scalability, and real-time deployment. Another major challenge is the computational cost of LM-based VAD. Many recent methods rely on large VLMs, MLLMs, captioning models, or repeated LLM queries, which can be expensive when applied to long videos or multi-camera surveillance systems [45], [126], [130], [135], [139], [141]. This limits their practicality in real-time settings, where anomaly scores, localization outputs, and explanations may need to be generated with low latency [82], [132], [134]. The challenge becomes even more significant when models require dense frame sampling, long-context reasoning, or multiple rounds of prompting. Future work should therefore

explore efficient architectures that combine lightweight visual backbones, object detectors, trackers, retrieval modules, and smaller task-specialized language models. A promising direction is to use low-cost perception modules for continuous monitoring and invoke larger reasoning models only when uncertain or suspicious events are detected. Such hierarchical designs can preserve the semantic advantages of LMs while improving scalability, latency, and deployability. Reproducibility is further complicated by the heterogeneous foundation models used across LM-based VAD systems. Some approaches rely on proprietary or rapidly evolving models, while others use open checkpoints whose versions and inference configurations may differ across studies. Important implementation details such as prompt templates, decoding parameters, number of model calls, hardware, runtime, latency, and API cost are also reported inconsistently. For this reason, direct computational comparison across the current literature is difficult. To provide a more transparent overview of model dependence, Appendix summarizes the foundation models used by the surveyed LM-based VAD methods.

Toward agentic and dynamic VAD systems. The recent rise of agentic AI has introduced new ways to design systems that can plan, retrieve information, use tools, verify intermediate outputs, and refine decisions through multi-step reasoning [159], [160], [175]–[177]. This trend suggests a promising direction for LM-based VAD. Instead of treating the language model as a passive classifier, caption generator, or anomaly scorer, future systems may use LMs as active reasoning agents that decompose anomaly detection into multiple steps. For example, an agentic VAD system could observe the scene, retrieve normal examples from memory, identify relevant objects and interactions, check location-specific rules, compare current activity with historical patterns, and produce grounded explanations. Such systems could also support dynamic adaptation, where normality models are updated as new normal activities appear or as the environment changes over time. In this view, VAD systems would no longer operate as static detectors, but as adaptive reasoning pipelines capable of refining their assumptions and explaining anomalies with respect to evolving scene-specific normality.

Ethical, privacy, and societal considerations. The deployment of LM-based VAD raises important ethical and societal concerns. Since VAD is often associated with surveillance, models that generate natural-language descriptions of people, actions, and interactions may increase privacy risks and create new opportunities for misuse [34], [178]. Language-based explanations can also make systems appear more reliable than they actually are, especially when generated explanations are plausible but not visually grounded. Bias is another concern: models may learn or inherit assumptions about what constitutes suspicious behavior, leading to unfair or context-insensitive judgments [179]. Beyond privacy and bias, robustness is also critical. Recent studies have shown that VAD systems can be vulnerable to adversarial perturbations, distribution shifts, and intentionally manipulated

TABLE 7. Operational inclusion and classification criteria used in constructing the LM-based VAD survey corpus.

Criterion	Inclusion / positive rule	Exclusion boundary
LM-based VAD paper	Included when video anomaly detection, localization, understanding, explanation, retrieval, or a closely related anomaly-oriented video task is a primary contribution, and an LLM, VLM, MLLM, language-aligned representation, or language-derived supervision plays a substantive methodological role.	Image-only anomaly detection, generic video understanding or action recognition, and surveillance analysis without an explicit anomaly-oriented objective are excluded from the primary taxonomy.
Publication-trend corpus	Includes all VAD papers identified in CVPR, ICCV, ECCV, WACV, NeurIPS, and ICML through the end of the 2025 conference cycle; LM-based papers are labeled using the criteria in this table.	Papers outside these venues are not included in the statistics of Fig. 1 and Table 1, even if they are included elsewhere in the broader survey.
Publication status	Both peer-reviewed papers and technically substantive preprints are considered in the broader survey corpus. When a peer-reviewed version is available, it is preferred over the corresponding preprint.	Duplicate preprint and published versions are not treated as separate works.
Language-model involvement	Classified as LM-based when language representations, textual prompts, language-model reasoning, textual supervision, or multimodal language models directly contribute to anomaly detection, localization, or understanding.	A paper is not classified as LM-based merely because an LLM is used for auxiliary writing, metadata generation, or other components unrelated to the anomaly-analysis mechanism.
Training-free	Classified as training-free when pretrained model parameters remain fixed with respect to the VAD task and anomaly inference is performed through prompting, similarity estimation, retrieval, captioning, or frozen-model reasoning.	Methods that perform VAD-specific parameter optimization or fine-tuning are not classified as training-free.
Instruction-tuned	Classified as instruction-tuned when a pretrained LM, VLM, or MLLM is adapted using anomaly-related instruction, question-answer, explanation, grounding, or similar task-specific supervision.	Prompting a frozen model without parameter adaptation does not satisfy this criterion.
Open-world / open-vocabulary	Classified in this category when the method explicitly supports previously unseen anomaly categories, expandable textual label spaces, or user-defined anomaly concepts at inference time.	Standard closed-set anomaly detection using a fixed training-category space is not classified as open-world or open-vocabulary.
Anomaly understanding / semantic output	Included when semantic outputs such as explanations, question-answer responses, anomaly categories, retrieval results, captions, or grounding predictions are explicitly evaluated as part of an anomaly-oriented task.	Generic video captioning, retrieval, or question answering without an anomaly-centered objective is excluded from the primary corpus.

inputs [180], [181]. These vulnerabilities are especially concerning in safety-critical or security-sensitive environments, where an attacker may attempt to hide anomalous behavior or trigger false alarms. Future LM-based VAD research should therefore incorporate robustness evaluation, adversarial testing, privacy-preserving designs, human-in-the-loop verification, and transparent reporting of failure cases.

Real-world deployment also requires attention to operational factors that are not captured by benchmark accuracy alone. In large-scale surveillance systems, even relatively low false-positive rates can generate substantial numbers of alerts, making false-alarm management and human review important components of practical system design. Deployment policies should also consider how video data, generated descriptions, and model outputs are stored or retained, particularly because language-based representations may expose sensitive information that is less apparent in raw anomaly scores. In addition, bandwidth, edge-computing constraints, and reliance on remote or proprietary foundation-model services can affect where and how LM-based VAD can be deployed. These considerations motivate evaluation frameworks that account not only for detection performance, but also for operational reliability, human oversight, privacy, and deployment-specific resource constraints.

VIII. SURVEY SCOPE AND LITERATURE COLLECTION

This survey provides a focused and structured review of language-model-based video anomaly detection. We distinguish between two literature collections used for different purposes in the paper: (i) a fixed-venue corpus used for the publication-trend analysis in Fig. 1 and Table 1, and (ii) a broader literature corpus used for the technical survey and taxonomy.

For the publication-trend analysis, we exhaustively reviewed VAD papers published in the major computer vision and machine learning conferences considered in this work: CVPR, ICCV, ECCV, WACV, NeurIPS, and ICML. All papers whose primary contribution concerns video anomaly detection were inspected, and LM-based papers were identified according to the operational criteria in Table 7. The cutoff for this fixed-venue analysis is the end of the 2025 conference cycle. This corpus is used only for the venue-level publication statistics reported in Fig. 1 and Table 1.

For the broader technical survey, we searched major scholarly search engines, publisher databases, and conference repositories, including Google Scholar, arXiv, IEEE Xplore, ACM Digital Library, and the CVF Open Access repository. Searches used combinations of terms including *video anomaly detection*, *video anomaly understanding*, *language model video anomaly detection*, *vision-language*

video anomaly detection, *multimodal large language model anomaly detection*, *CLIP video anomaly detection*, *training-free video anomaly detection*, and *open-vocabulary video anomaly detection*, together with closely related variants. Reference lists of relevant papers and recent VAD surveys were additionally inspected to identify potentially missed works. The broader survey corpus includes literature available up to May 2026.

A work was considered within the primary scope of the survey when anomaly-oriented video analysis constitutes a central contribution and language models, vision-language models, multimodal large language models, language-aligned representations, or language-derived supervision play a substantive methodological role. We consider not only conventional anomaly detection and localization, but also emerging anomaly-oriented tasks such as anomaly understanding, explanation, retrieval, question answering, and grounding when these are explicitly formulated around abnormal-event analysis. In contrast, image-only anomaly detection, generic video understanding or action recognition, and surveillance analysis without an explicit anomaly-oriented objective are excluded from the primary survey taxonomy.

Both peer-reviewed publications and technically substantive preprints are included in the broader survey corpus because a substantial portion of recent LM-based VAD research appears first on arXiv before formal publication. When a peer-reviewed version is available, it is preferred over the corresponding preprint, and duplicate preprint and published versions are not treated as separate works. Candidate papers were first screened based on title and abstract and were subsequently inspected at the full-paper level to confirm their relevance and methodological role.

To make the survey boundaries and classification procedure explicit, Table 7 summarizes the operational criteria used to determine whether a work belongs to the primary LM-based VAD corpus and how representative methodological paradigms are assigned. These categories are not necessarily mutually exclusive; rather, they identify the principal methodological role used to organize the literature. Future additions to the dynamic survey are screened according to the same scope and classification principles before being incorporated into maintained releases.

IX. CONCLUSION

Language models are rapidly changing the landscape of video anomaly detection by extending the field beyond visual pattern recognition toward semantic reasoning, contextual interpretation, open-vocabulary recognition, and natural-language explanation. In this survey, we reviewed this emerging research direction through a unified taxonomy of language-model-based VAD methods, covering unsupervised and semi-supervised, weakly supervised, training-free, instruction-tuned, and open-world/open-vocabulary paradigms. We also positioned these methods relative to key non-language-model foundations, reviewed representative datasets and evaluation protocols, and discussed how recent benchmarks increasingly

shift the focus from anomaly detection alone toward anomaly understanding.

Across the surveyed literature, a clear trend emerges: language models provide powerful semantic priors that can help VAD systems reason about objects, actions, interactions, scene context, and unseen anomaly concepts. However, this shift also introduces new challenges. LM-based VAD systems must avoid reducing anomaly detection to category recognition, since abnormality is often scene-specific and context-dependent. They must also provide reliable spatial and temporal grounding, reduce hallucination and semantic bias, remain computationally efficient, and support transparent evaluation beyond conventional detection metrics. These challenges are especially important as VAD systems move toward real-world deployment, where explanations must be faithful, decisions must be robust, and normality may evolve over time.

Finally, because LM-based VAD is developing rapidly, we frame this work not only as a static review but also as a dynamic and versioned scholarly resource. The maintained survey supports the addition of newly published methods, datasets, benchmarks, and evaluation protocols through localized revisions, while preserving the overall taxonomy, terminology, and organizational structure of the peer-reviewed article. Each release is accompanied by transparent version records and archived through the public website, allowing readers to identify newly added papers, revised sections, updated tables, and other substantive changes over time. This maintenance process is intended to reduce the obsolescence that commonly affects surveys in fast-moving research areas and to provide a stable reference point for the LM-based VAD community.

ACKNOWLEDGMENT

The authors used ChatGPT, powered by OpenAI’s GPT-5.5 model [182], to assist with the visual editing and polishing of Figs. 3–7. The figure concepts, scientific content, and final revisions were developed, reviewed, and approved by the authors.

REFERENCES

- [1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.
- [2] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, “A unifying review of deep and shallow anomaly detection,” *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021.
- [3] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 246–252.
- [4] W. Hu, T. Tan, L. Wang, and S. Maybank, “A survey on visual surveillance of object motion and behaviors,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 34, no. 3, pp. 334–352, 2004.
- [5] B. T. Morris and M. M. Trivedi, “A survey of vision-based trajectory learning and analysis for surveillance,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1114–1127, 2008.
- [6] O. Boiman and M. Irani, “Detecting irregularities in images and in video,” *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, 2007.
- [7] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, “Robust real-time unusual event detection using multiple fixed-location monitors,” *IEEE*

- transactions on pattern analysis and machine intelligence*, vol. 30, no. 3, pp. 555–560, 2008.
- [8] L. Kratz and K. Nishino, “Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1446–1453.
 - [9] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 935–942.
 - [10] J. Kim and K. Grauman, “Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2921–2928.
 - [11] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.
 - [12] S. Wu, B. E. Moore, and M. Shah, “Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2054–2060.
 - [13] Y. Cong, J. Yuan, and J. Liu, “Sparse reconstruction cost for abnormal event detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3449–3456.
 - [14] V. Saligrama and Z. Chen, “Video anomaly detection based on local statistical aggregates,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2112–2119.
 - [15] M. J. Roshtkhari and M. D. Levine, “An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions,” *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1436–1452, 2013.
 - [16] W. Luo, W. Liu, and S. Gao, “Remembering history with convolutional lstm for anomaly detection,” in *2017 IEEE International conference on multimedia and expo (ICME)*. IEEE, 2017, pp. 439–444.
 - [17] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning temporal regularity in video sequences,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.
 - [18] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection—a new baseline,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545.
 - [19] W. Sultan, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.
 - [20] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in matlab,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.
 - [21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
 - [22] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.
 - [23] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds et al., “Flamingo: a visual language model for few-shot learning,” *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.
 - [24] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.
 - [25] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, “Video-llava: Learning united visual representation by alignment before projection,” in *Proceedings of the 2024 conference on empirical methods in natural language processing*, 2024, pp. 5971–5984.
 - [26] M. Maaz, H. Rasheed, S. Khan, and F. Khan, “Video-chatgpt: Towards detailed video understanding via large vision and language models,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 12 585–12 602.
 - [27] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, “A survey of single-scene video anomaly detection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2293–2312, 2020.
 - [28] R. Nayak, U. C. Pati, and S. K. Das, “A comprehensive review on deep learning-based methods for video anomaly detection,” *Image and Vision Computing*, vol. 106, p. 104078, 2021.
 - [29] T. M. Tran, T. N. Vu, N. D. Vo, T. V. Nguyen, and K. Nguyen, “Anomaly analysis in images and videos: A comprehensive review,” *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–37, 2022.
 - [30] Y. Liu, D. Yang, Y. Wang, J. Liu, J. Liu, A. Boukerche, P. Sun, and L. Song, “Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models,” *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–38, 2024.
 - [31] P. Wu, C. Pan, Y. Yan, G. Pang, Q. Yan, P. Wang, and Y. Zhang, “Deep learning for video anomaly detection: A review,” *IEEE Transactions on Neural Networks and Learning Systems*, 2026.
 - [32] M. Abdalla, S. Javed, M. Al Radi, A. Ulhaq, and N. Werghi, “Video anomaly detection in 10 years: A survey and outlook,” *Neural Computing and Applications*, vol. 37, no. 32, pp. 26 321–26 364, 2025.
 - [33] X. Ding and L. Wang, “Quo vadis, anomaly detection? llms and vlms in the spotlight,” *arXiv preprint arXiv:2412.18298*, 2024.
 - [34] J. Liu, Y. Liu, J. Lin, J. Li, L. Cao, P. Sun, B. Hu, L. Song, A. Boukerche, and V. C. Leung, “Networking systems for video anomaly detection: A tutorial and survey,” *ACM Computing Surveys*, vol. 57, no. 10, pp. 1–37, 2025.
 - [35] S. Gao, P. Yang, H. Guo, Y. Liu, Y. Chen, S. Li, H. Zhu, J. Xu, X.-Y. Zhang, and L. Huang, “The evolution of video anomaly detection: A unified framework from dnn to mllm,” *arXiv preprint arXiv:2507.21649*, 2025.
 - [36] F. Mumcu, L. Bekit, M. J. Jones, A. Cherian, and Y. Yilmaz, “Agentic ai-empowered dynamic survey framework,” *arXiv preprint arXiv:2602.04071*, 2026.
 - [37] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
 - [38] D. M. J. Tax and R. P. W. Duin, “Support vector data description,” *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.
 - [39] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, “Deep one-class classification,” in *International Conference on Machine Learning*, 2018, pp. 4393–4402.
 - [40] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” *Artificial Intelligence*, vol. 89, no. 1–2, pp. 31–71, 1997.
 - [41] S. Andrews, I. Tsochantaridis, and T. Hofmann, “Support vector machines for multiple-instance learning,” in *Advances in Neural Information Processing Systems*, vol. 15, 2002.
 - [42] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boulton, “Toward open set recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.
 - [43] C. Geng, S.-j. Huang, and S. Chen, “Recent advances in open set recognition: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3614–3631, 2020.
 - [44] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
 - [45] H. Zhang, X. Xu, X. Wang, J. Zuo, X. Huang, C. Gao, S. Zhang, L. Yu, and N. Sang, “Holmes-vau: Towards long-term video anomaly understanding at any granularity,” in *Proceedings of the computer vision and pattern recognition conference*, 2025, pp. 13 843–13 853.
 - [46] K. Doshi and Y. Yilmaz, “Rethinking video anomaly detection: A continual learning approach,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.
 - [47] W. Li, V. Mahadevan, and N. Vasconcelos, “Anomaly detection and localization in crowded scenes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.
 - [48] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349.
 - [49] C. Cao, Y. Lu, P. Wang, and Y. Zhang, “A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 392–20 401.
 - [50] B. Ramachandra and M. Jones, “Street scene: A new dataset and evaluation protocol for video anomaly detection,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 2569–2578.
 - [51] F. Mumcu, M. Jones, Y. Yilmaz, and A. Cherian, “Complexvad: Detecting interaction anomalies in video,” in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025, pp. 1093–1102.

- [52] H. I. Öztürk and A. B. Can, “Adnet: Temporal anomaly detection in surveillance videos,” in *International Conference on Pattern Recognition*. Springer, 2021, pp. 88–101.
- [53] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, “Not only look, but also listen: Learning multimodal violence detection under weak supervision,” in *European conference on computer vision*. Springer, 2020, pp. 322–339.
- [54] Y. Xu, H. Hu, C. Huang, Y. Nan, Y. Liu, K. Wang, Z. Liu, and S. Lian, “Tad: A large-scale benchmark for traffic accidents detection from video surveillance,” *IEEE Access*, vol. 13, pp. 2018–2033, 2024.
- [55] S. A. Velastin and D. A. Gómez-Lira, “People detection and pose classification inside a moving train using computer vision,” in *International visual informatics conference*. Springer, 2017, pp. 319–330.
- [56] D. D. de Paula, D. H. Salvadeo, and D. M. de Araujo, “Camnumem: A robbery dataset for video anomaly detection,” *Sensors*, vol. 22, no. 24, p. 10016, 2022.
- [57] A. Ascintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, “Ubnormal: New benchmark for supervised open-set video anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 143–20 153.
- [58] L. Zhu, L. Wang, A. Raj, T. Gedeon, and C. Chen, “Advancing video anomaly detection: A concise review and a new dataset,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 89 943–89 977, 2024.
- [59] H. Du, S. Zhang, B. Xie, G. Nan, J. Zhang, J. Xu, H. Liu, S. Leng, J. Liu, H. Fan et al., “Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 793–18 803.
- [60] H. Du, G. Nan, J. Qian, W. Wu, W. Deng, H. Mu, Z. Chen, P. Mao, X. Tao, and J. Liu, “Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly,” *arXiv preprint arXiv:2412.07183*, 2024.
- [61] H. Gani, R. Bharadwaj, M. Naseer, F. S. Khan, and S. Khan, “Vane-bench: Video anomaly evaluation benchmark for conversational llms,” in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 3123–3140.
- [62] S. Gao, P. Yang, Y. Liu, Y. Chen, H. Zhu, X.-Y. Zhang, and L. Huang, “Vagu & gts: Llm-based benchmark and framework for joint video anomaly grounding and understanding,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 6, 2026, pp. 4167–4175.
- [63] J. A. C. Pereira, V. Lopes, J. C. Neves, and D. Semedo, “Finevau: A novel human-aligned benchmark for fine-grained video anomaly understanding,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 10, 2026, pp. 8403–8411.
- [64] T. Yuan, X. Zhang, K. Liu, B. Liu, C. Chen, J. Jin, and Z. Jiao, “Towards surveillance video-and-language understanding: New dataset baselines and challenges,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 22 052–22 061.
- [65] S. Yang, Y. Wang, Y. Wang, L. Zhu, and Z. Zheng, “Towards scalable video anomaly retrieval: A synthetic video-text benchmark,” *arXiv preprint arXiv:2506.01466*, 2025.
- [66] Q. Bao, F. Liu, L. Jiao, Y. Liu, S. Li, L. Li, X. Liu, X. Wang, and B. Chen, “Anomaly-led prompting learning caption generating model and benchmark,” *IEEE Transactions on Multimedia*, 2025.
- [67] N. T. Nguyen, L. Bekit, and Y. Yilmaz, “Sense-vad: Sentient and semantic video anomaly detection for autonomous driving,” 2026. [Online]. Available: <https://arxiv.org/abs/2606.31875>
- [68] T. Fawcett, “An introduction to roc analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [69] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the International Conference on Machine Learning*, 2006, pp. 233–240.
- [70] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [71] K. Papineni, S. Roukos, D. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [72] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, 2004, pp. 74–81.
- [73] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.
- [74] J. Amidei, P. Piwek, and A. Willis, “The use of rating and likert scales in natural language generation human evaluation tasks: A review and some recommendations,” in *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, 2019, pp. 397–402.
- [75] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, and E. Krahmer, “Best practices for the human evaluation of automatically generated text,” in *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, 2019.
- [76] A. Elangovan, L. Liu, L. Xu, S. B. Bodapati, and D. Roth, “Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024.
- [77] J. A. C. Pereira, V. Lopes, J. C. Neves, and D. Semedo, “Finevau: A novel human-aligned benchmark for fine-grained video anomaly understanding,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 10, 2026, pp. 8403–8411.
- [78] J. H. F. James, “Counting on consensus: Selecting the right inter-annotator agreement metric for nlp annotation and evaluation,” in *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, 2026, pp. 4434–4446.
- [79] F. Mumcu, M. J. Jones, A. Cherian, and Y. Yilmaz, “Leveraging multimodal llm descriptions of activity for explainable semi-supervised video anomaly detection,” *arXiv preprint arXiv:2510.14896*, 2025.
- [80] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1705–1714.
- [81] M. H. Hemmatyar, M. Jafari, M. A. Yousefi, M. R. Nemati, M. Azadani, H. R. Rastad, and A. Akbari, “Hycovad: A hybrid ssl-llm model for complex video anomaly detection,” *arXiv preprint arXiv:2509.22544*, 2025.
- [82] Z. Ding, H. Zhang, P. Wu, G. Pang, Z. Yang, P. Wang, and Y. Zhang, “Slowfastvad: Video anomaly detection via integrating simple detector and rag-enhanced vision-language model,” *arXiv preprint arXiv:2504.10320*, 2025.
- [83] J. Li, L. Dang, Y. Su, Y. Hao, Q. Xiao, Y. Nie, and Q. Wu, “Memoryout: Learning principal features via multimodal sparse filtering network for semi-supervised video anomaly detection,” *arXiv e-prints*, pp. arXiv–2506, 2025.
- [84] C. Li and Y. Jiang, “Vlavad: Vision-language models assisted unsupervised video anomaly detection,” in *BMVC*, 2024.
- [85] Y. Yang, K. Lee, B. Dariush, Y. Cao, and S.-Y. Lo, “Follow the rules: reasoning for video anomaly detection with large language models,” in *European Conference on Computer Vision*. Springer, 2024, pp. 304–322.
- [86] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, “A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 588–13 597.
- [87] J. Kim, S. Yoon, T. Choi, and S. Sull, “Unsupervised video anomaly detection based on similarity with predefined text descriptions,” *Sensors*, vol. 23, no. 14, p. 6256, 2023.
- [88] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, “Vadclip: Adapting vision-language models for weakly supervised video anomaly detection,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 6, 2024, pp. 6074–6082.
- [89] S. Li, F. Liu, and L. Jiao, “Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 2, 2022, pp. 1395–1403.
- [90] J. Feng, F. Hong, and W. Zheng, “MIST: multiple instance self-training framework for video anomaly detection,” *CoRR*, vol. abs/2104.01633, 2021. [Online]. Available: <https://arxiv.org/abs/2104.01633>
- [91] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, “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*, 2019, pp. 1237–1246.
- [92] J. Dalvi, A. Dabouei, G. Dhanuka, and M. Xu, “Distilling aggregated knowledge for weakly-supervised video anomaly detection,” in

- 2025 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 5439–5448.
- [93] Y. Wang and S. Chen, “Learning event completeness for weakly supervised video anomaly detection,” *arXiv preprint arXiv:2506.13095*, 2025.
- [94] C. Huang, Y. Shi, J. Wen, W. Wang, Y. Xu, and X. Cao, “Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models,” in *Forty-second International Conference on Machine Learning*, 2025.
- [95] B. Wang, C. Huang, J. Wen, W. Wang, Y. Liu, and Y. Xu, “Federated weakly supervised video anomaly detection with multimodal prompt,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 39, no. 20, 2025, pp. 21 017–21 025.
- [96] H. Qiu, B. Hou, and Y. Cui, “Learning opposite prompts for weakly supervised video anomaly detection,” *Knowledge-Based Systems*, vol. 324, p. 113600, 2025.
- [97] H. Qiu and B. Hou, “Temporal context and representative feature learning for weakly supervised video anomaly detection,” *Journal of Information and Intelligence*, 2025.
- [98] H. Xu, H. Wang, M. Yue, and Z. Li, “Filter, summarize, align: Learning the semantic guided weakly supervised video anomaly detection,” in *International Conference on Intelligent Computing*. Springer, 2025, pp. 76–87.
- [99] L. Liu, J. Li, G. Li, Y. Zhai, and M. Zhang, “Vadclip++: Dynamic vision-language model for weakly supervised video anomaly detection,” *Digital Signal Processing*, p. 105560, 2025.
- [100] J. Wang, G. Li, J. Liu, Z. Xu, X. Chen, and J. Wei, “Relvid: Relational learning with vision-language models for weakly video anomaly detection,” *Sensors*, vol. 25, no. 7, p. 2037, 2025.
- [101] P. Wu, W. Su, G. Pang, Y. Sun, Q. Yan, P. Wang, and Y. Zhang, “Avadclip: Audio-visual collaboration for robust video anomaly detection,” *arXiv preprint arXiv:2504.04495*, 2025.
- [102] D. Wang, Q. Wang, Q. Hu, and K. Wu, “Multimodal vad: Visual anomaly detection in intelligent monitoring system via audio-vision-language,” *IEEE Transactions on Instrumentation and Measurement*, 2025.
- [103] Y. Pu, X. Wu, L. Yang, and S. Wang, “Learning prompt-enhanced context features for weakly-supervised video anomaly detection,” *IEEE Transactions on Image Processing*, vol. 33, pp. 4923–4936, 2024.
- [104] H. K. Joo, K. Vo, K. Yamazaki, and N. Le, “Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection,” in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 3230–3234.
- [105] P. P. Dev, R. Hazari, and P. Das, “Reflip-vad: Towards weakly supervised video anomaly detection via vision-language model,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [106] W.-D. Jiang, C.-Y. Chang, H.-C. Chang, J.-Y. Chen, and D. S. Roy, “Injecting explainability and lightweight design into weakly supervised video anomaly detection systems,” *arXiv preprint arXiv:2412.20201*, 2024.
- [107] H. Lim, D. Kim, M. Kim, C. Park, D. Kang, and S. Lee, “Weakly supervised video anomaly detection using dynamic-weighted feature fusion,” in *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 2024, pp. 1–4.
- [108] Z. Qian, J. Tan, Z. Ou, and H. Wang, “Clip-driven multi-scale instance learning for weakly supervised video anomaly detection,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.
- [109] S. Zhan, D. Zhang, and J. Wang, “Weakly supervised video anomaly detection with large language models knowledge enhancement framework,” in *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2024, pp. 468–475.
- [110] W. Chen, K. T. Ma, Z. J. Yew, M. Hur, and D. A.-A. Khoo, “Tevad: Improved video anomaly detection with captions,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5549–5559.
- [111] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, “Weakly-supervised video anomaly detection with robust temporal feature magnitude learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4975–4986.
- [112] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, “Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 1, 2023, pp. 387–395.
- [113] M. Li, J. Sang, Y. Lu, and L. Du, “Wsvad-clip: Temporally aware and prompt learning with clip for weakly supervised video anomaly detection,” *Journal of Imaging*, vol. 11, no. 10, p. 354, 2025.
- [114] L. Zanella, B. Liberatori, W. Menapace, F. Poiesi, Y. Wang, and E. Ricci, “Delving into clip latent space for video anomaly recognition,” *Computer Vision and Image Understanding*, vol. 249, p. 104163, 2024.
- [115] P. Wu, X. Zhou, G. Pang, Z. Yang, Q. Yan, P. Wang, and Y. Zhang, “Weakly supervised video anomaly detection and localization with spatio-temporal prompts,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9301–9310.
- [116] J. Chen, L. Li, L. Su, Z.-j. Zha, and Q. Huang, “Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 18 319–18 329.
- [117] C. Tao, X. Peng, C. Wang, J. Wu, P. Zhao, J. Wang, and J. Qian, “Learning suspected anomalies from event prompts for video anomaly detection,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 22, no. 5, pp. 1–20, 2026.
- [118] C. Xiao, Y. Xiao, J. T. Zhou, and Z. Fang, “Itilingual-prompt-guided directional feature learning for weakly supervised video anomaly detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [119] T. Liu, K.-M. Lam, and B.-K. Bao, “Injecting text clues for improving anomalous event detection from weakly labeled videos,” *IEEE Transactions on Image Processing*, vol. 33, pp. 5907–5920, 2024.
- [120] A. Ntelopoulos and K. Nasrollahi, “Callm: Cascading autoencoder and large language model for video anomaly detection,” in *2024 IEEE Thirteenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2024, pp. 1–6.
- [121] Y. Chen, J. Liu, R. Fan, Y. Li, C. Chang, S. Zhao, W. W. Fok, X. Qi, and Y.-C. Wu, “Aligning effective tokens with video anomaly in large language models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 22 695–22 706.
- [122] S. Paulraj and S. Vairavasundaram, “M2vad: Multiview multimodality transformer-based weakly supervised video anomaly detection,” *Image and Vision Computing*, vol. 149, p. 105139, 2024.
- [123] G. Wang, S. Song, W. He, and Y. Zheng, “Cmhkf: Cross-modality heterogeneous knowledge fusion for weakly supervised video anomaly detection,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 31 594–31 607.
- [124] S. Yun, R. Masukawa, M. Na, and M. Imani, “Missiongenn: Hierarchical multimodal gnn-based weakly supervised video anomaly recognition with mission-specific knowledge graph generation,” in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 4736–4745.
- [125] P. Wu, J. Liu, X. He, Y. Peng, P. Wang, and Y. Zhang, “Toward video anomaly retrieval from video anomaly detection: New benchmarks and model,” *IEEE Transactions on Image Processing*, vol. 33, pp. 2213–2225, 2024.
- [126] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, “Harnessing large language models for training-free video anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 527–18 536.
- [127] P. Wu, X. Zhou, G. Pang, Y. Sun, J. Liu, P. Wang, and Y. Zhang, “Open-vocabulary video anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 297–18 307.
- [128] F. Li, W. Liu, J. Chen, R. Zhang, Y. Wang, X. Zhong, and Z. Wang, “Anomize: Better open vocabulary video anomaly detection,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 203–29 212.
- [129] C. Xu, K. Xu, X. Jiang, and T. Sun, “Plovad: Prompting vision-language models for open vocabulary video anomaly detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 6, pp. 5925–5938, 2025.
- [130] H. Zhang, X. Xu, X. Wang, J. Zuo, C. Han, X. Huang, C. Gao, Y. Wang, and N. Sang, “Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm,” *arXiv preprint arXiv:2406.12235*, 2024.
- [131] L. Bekit, H. Karim, N. T. Nguyen, and Y. Yilmaz, “Qvad: A question-centric agentic framework for efficient and training-free video anomaly detection,” *arXiv preprint arXiv:2604.03040*, 2026.

- [132] H. Lee, H. Kim, I.-J. Kim, and Y. Choi, "Flashback: Memory-driven zero-shot, real-time video anomaly detection," *arXiv preprint arXiv:2505.15205*, 2025.
- [133] M. Abdalla and S. Javed, "Training-free vlm-based pseudo label generation for video anomaly detection," *IEEE Access*, 2025.
- [134] Y. Feng, Y. Liu, J. Zhang, J. Qin et al., "Monitor: Exploiting large language models with instruction for online video anomaly detection," *Advances in Neural Information Processing Systems*, vol. 38, pp. 35 851–35 871, 2026.
- [135] Y. Shao, H. He, S. Li, S. Chen, X. Long, F. Zeng, Y. Fan, M. Zhang, Z. Yan, A. Ma et al., "Eventvad: Training-free event-aware video anomaly detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 2586–2595.
- [136] Z. Yang, C. Gao, and M. Z. Shou, "Panda: Towards generalist video anomaly detection via agentic ai engineer," *Advances in Neural Information Processing Systems*, vol. 38, pp. 83 182–83 211, 2026.
- [137] W. Li, Y. Xu, Y. Rao, Z. Wang, and S. Deng, "Vadtrees: Explainable training-free video anomaly detection via hierarchical granularity-aware tree," *Advances in Neural Information Processing Systems*, vol. 38, pp. 148 372–148 404, 2026.
- [138] M. Ye, W. Liu, and P. He, "Vera: Explainable video anomaly detection via verbalized learning of vision-language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8679–8688.
- [139] S. Gao, P. Yang, and L. Huang, "Suvad: Semantic understanding based video anomaly detection using mllm," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [140] P. P. Dev, R. Hazari, and P. Das, "Mcanet: Multimodal caption aware training-free video anomaly detection via large language model," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 362–379.
- [141] S. Ahn, Y. Jo, K. Lee, S. Kwon, I. Hong, and S. Park, "Anyanomaly: Zero-shot customizable video anomaly detection with lvm," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 3026–3035.
- [142] J. Xu, S.-Y. Lo, B. Safaei, V. M. Patel, and I. Dwivedi, "Towards zero-shot anomaly detection and reasoning with multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 370–20 382.
- [143] Y. Cao, X. Xu, Y. Cheng, C. Sun, Z. Du, L. Gao, and W. Shen, "Personalizing vision-language models with hybrid prompts for zero-shot anomaly detection," *IEEE Transactions on Cybernetics*, 2025.
- [144] M. Dong, "Clip: Assisted video anomaly detection," in *ICPRAM*, 2024, pp. 522–533.
- [145] Y. Cao, X. Xu, C. Sun, X. Huang, and W. Shen, "Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead," *arXiv preprint arXiv:2311.02782*, 2023.
- [146] H. Wang, J. Qin, A. Bastola, X. Chen, J. Suchanek, Z. Gong, and A. Razi, "Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation," *arXiv preprint arXiv:2403.12415*, 2024.
- [147] S. Lin, C. Wang, X. Ding, Y. Wang, B. Du, L. Song, C. Wang, and H. Liu, "A vlm-based method for visual anomaly detection in robotic scientific laboratories," in *2025 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2025, pp. 34–39.
- [148] J. Tang, H. Lu, R. Wu, X. Xu, K. Ma, C. Fang, B. Guo, J. Lu, Q. Chen, and Y.-C. Chen, "Hawk: Learning to understand open-world video anomalies," *Advances in Neural Information Processing Systems*, vol. 37, pp. 139 751–139 785, 2024.
- [149] Z. Yang, C. Gao, J. Liu, P. Wu, G. Pang, and M. Z. Shou, "Assistpda: An online video surveillance assistant for video anomaly prediction, detection, and analysis," *arXiv preprint arXiv:2503.21904*, 2025.
- [150] C. Huang, B. Wang, W. Wang, J. Wen, C. Liu, L. Shen, and X. Cao, "Vad-rl: Towards video anomaly reasoning via perception-to-cognition chain-of-thought," *Advances in neural information processing systems*, vol. 38, pp. 118 486–118 518, 2026.
- [151] L. Zhu, Q. Chen, X. Shen, and X. Cun, "Vau-rl: Advancing video anomaly understanding via reinforcement fine-tuning," *arXiv preprint arXiv:2505.23504*, 2025.
- [152] J. Ma, J. Wang, J. Luo, P. Yu, and G. Zhou, "Sherlock: Towards multi-scene video abnormal event extraction and localization via a global-local spatial-sensitive llm," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 4004–4013.
- [153] C. Huang, W. Huang, Q. Jiang, W. Wang, J. Wen, and B. Zhang, "Multimodal evidential learning for open-world weakly-supervised video anomaly detection," *IEEE Transactions on Multimedia*, 2025.
- [154] Z. Liu, X. Wu, J. Wu, X. Wang, and L. Yang, "Language-guided open-world video anomaly detection under weak supervision," *arXiv preprint arXiv:2503.13160*, 2025.
- [155] F. Mumcu, M. J. Jones, A. Cherian, and Y. Yilmaz, "Llm-guided agentic object detection for open-world understanding," *arXiv preprint arXiv:2507.10844*, 2025.
- [156] G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for mind exploration of large language model society," *Advances in Neural Information Processing Systems*, vol. 36, pp. 51 991–52 008, 2023.
- [157] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong et al., "Chatdev: Communicative agents for software development," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 15 174–15 186.
- [158] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," *arXiv preprint arXiv:2305.16291*, 2023.
- [159] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8634–8652, 2023.
- [160] P. Belcak, G. Heinrich, S. Diao, Y. Fu, X. Dong, S. H. Muralidharan, and P. Molchanov, "Small language models are the future of agentic ai," *arXiv preprint arXiv:2506.02153*, 2025.
- [161] F. Mumcu and Y. Yilmaz, "Fast and lightweight vision-language model for adversarial traffic sign detection," *Electronics*, vol. 13, no. 11, p. 2172, 2024.
- [162] A. Marafioti, O. Zohar, M. Farré, M. Noyan, E. Bakouch, P. Cuenca, C. Zakka, L. B. Allal, A. Lozhkov, N. Tazi et al., "Smolvlm: Redefining small and efficient multimodal models," *arXiv preprint arXiv:2504.05299*, 2025.
- [163] Nvidia, "How small language models are key to scalable agentic ai," 2025. [Online]. Available: <https://developer.nvidia.com/blog/how-small-language-models-are-key-to-scalable-agentic-ai/>
- [164] A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard et al., "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, vol. 4, 2025.
- [165] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière et al., "Gemma 3 technical report. arxiv 2025," *arXiv preprint arXiv:2503.19786*.
- [166] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv et al., "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [167] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu et al., "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.
- [168] Q. Team, "Qwen3. 5-omni technical report," *arXiv preprint arXiv:2604.15804*, 2026.
- [169] F. Mumcu, M. J. Jones, A. Cherian, and Y. Yilmaz, "Is video anomaly detection misframed? evidence from llm-based and multi-scene models," *arXiv preprint arXiv:2605.12725*, 2026.
- [170] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023, pp. 292–305.
- [171] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoub et al., "Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 14 375–14 385.
- [172] R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, and P. Liang, "The foundation model transparency index," *arXiv preprint arXiv:2310.12941*, 2023.
- [173] C. Deng, Y. Zhao, X. Tang, M. Gerstein, and A. Cohan, "Investigating data contamination in modern benchmarks for large language models. arxiv. doi: 10.48550," *arXiv preprint arXiv:2311.09783*, 2023.
- [174] S. Ballocco, P. Schmidová, M. Lango, and O. Dušek, "Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 67–93.

- [175] F. Mumcu and Y. Yilmaz, “Socially-weighted alignment: A game-theoretic framework for multi-agent llm systems,” *arXiv preprint arXiv:2602.14471*, 2026.
- [176] M. Abou Ali, F. Dornaika, and J. Charafeddine, “Agentic ai: a comprehensive survey of architectures, applications, and future directions,” *Artificial Intelligence Review*, vol. 59, no. 1, p. 11, 2025.
- [177] F. Mumcu and Y. Yilmaz, “Robustness of agentic ai systems via adversarially-aligned jacobian regularization,” *arXiv preprint arXiv:2603.04378*, 2026.
- [178] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, “Visual privacy protection methods: A survey,” *Expert Systems with Applications*, vol. 42, no. 9, pp. 4177–4195, 2015.
- [179] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavarez, S. Aidinlis, T. Marquenie, A. Gurzawska, and D. Tzovaras, “Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems,” in *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2022, pp. 2302–2314.
- [180] F. Mumcu, K. Doshi, and Y. Yilmaz, “Adversarial machine learning attacks against video anomaly detection systems,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 206–213.
- [181] M. Nafez, M. Poulai, N. Vasei, M. Sabokrou, M. H. Rohban *et al.*, “Frameshield: Adversarially robust video anomaly detection,” *Advances in Neural Information Processing Systems*, vol. 38, pp. 29 859–29 893, 2026.
- [182] OpenAI, “Introducing GPT-5.5,” <https://openai.com/index/introducing-gpt-5-5/>, 2026, accessed: July 14, 2026.



FURKAN MUMCU received his B.S. degree in Computer Engineering from Bilkent University, Ankara, Türkiye, in 2018, and his M.S. degree in Electrical Engineering from the University of South Florida, Tampa, FL, USA, in 2024, where he is currently pursuing a Ph.D. degree. His research interests include AI security, video anomaly detection, adversarial machine learning, and agentic AI.



LOKMAN BEKIT received his B.S. degree in Electrical Engineering from the University of South Florida, in 2025. He is currently pursuing a Ph.D. degree in Artificial Intelligence at the University of South Florida. His research interests include video anomaly detection, vision-language models and efficient inference.



MICHAEL J. JONES received his bachelors degree in computer science and mathematics from Rice University in 1990 and then went to the Artificial Intelligence Lab at MIT where he received his Ph.D. in Electrical Engineering and Computer Science in 1997. After MIT, he worked at DEC’s Cambridge Research Lab before moving to Mitsubishi Electric Research Laboratories (MERL) in Cambridge, Massachusetts where he is currently a Distinguished Research Scientist in the computer vision group. His research interests include object detection and recognition, machine learning and video anomaly detection.



ANOOP CHERIAN received the undergraduate (honors) degree in computer science and engineering from the National Institute of Technology (NIT), Calicut, India, in 2002, and the MS and PhD degrees from the University of Minnesota, Minneapolis, Minnesota, in 2010 and 2013, respectively. He is a Senior Principal Research Scientist at Mitsubishi Electric Research Labs (MERL) Cambridge, Massachusetts, and an adjunct researcher at the Australian Centre for Robotic Vision (ACRV) at the Australian National University, Canberra, Australia. Before joining ANU, he was a postdoctoral researcher in the LEAR team at INRIA, Grenoble.



YASIN YILMAZ received the B.S. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Türkiye, in 2008, the M.S. degree in electrical and computer engineering from Koc University, Istanbul, Türkiye, in 2010, and the Ph.D. degree in electrical engineering from Columbia University, New York City, NY, USA, in 2014. He is currently an Associate Professor with the Department of Electrical Engineering, University of South Florida, Tampa, FL, USA. His research interests include machine learning, computer vision, anomaly detection, AI security, IoT security, and their applications in environmental, biomedical, energy, transportation, and communication systems.

...

TABLE 8. Reproducibility characteristics of representative LM-based VAD methods published at top-tier venues (CVPR, ICCV, ECCV, WACV, NeurIPS, ICML, ICLR). We include only methods that invoke a generative LLM or MLLM *at inference time* – i.e., when scoring or explaining a test video – and exclude methods that use an LLM/VLM solely offline to build training labels, a fixed prompt library, or a knowledge graph, after which a purely visual or CLIP-embedding classifier runs at test time (e.g., VadCLIP-family methods, MissionGNN, OVAD, Anomize). Methods are grouped by training paradigm. ✓ indicates the item is fully specified in the paper, its appendix, or its official repository, with enough detail to reproduce it exactly; ~ indicates it is partially specified (e.g., a component is named without a version/snapshot, or only some of the required parameters are given); ✗ indicates it is not reported anywhere in the source. Score is the count of ✓ (1 point) and ~ (0.5 point) across the six checklist columns, out of 6. Exact quoted values underlying every symbol, along with prompt-template disclosure, context-window size, and runtime/cost – omitted here for space – are provided in the supplementary reproducibility table.

Method	Foundation Model (Inference-Time)	Year	Version/Checkpoint	Decoding Params	#Inference Calls	Hardware	Code	Data	Score (/6)
AnomalyRuler [85]	CogVLM + Mistral-7B	2024	✓	~	✓	~	✓	✓	5.0
HAWK [148]	Video-LLaMA + LLaMA-2-7B	2024	✓	✓	✓	✓	✓	✓	6.0
LAVAD [126]	Llama-2-13b-chat + BLIP-2	2024	~	✗	✓	✓	✓	✓	4.5
Anomaly-OV [142]	LLaVA-OneVision 0.5B/7B	2025	~	✗	✓	~	✓	~	3.5
Holmes-VAU [45]	InternVL2-2B	2025	✓	✓	✓	✓	✓	✓	6.0
Vad-R1 [150]	Qwen2.5-VL-7B-Instruct	2025	~	✓	✓	~	✓	✓	5.0
Ex-VAD [94]	BLIP-2 + Llama-3.1	2025	✗	✗	~	✓	✗	✓	2.5
VERA [138]	InternVL2-8B	2025	~	✗	✓	~	✓	✓	4.0
PANDA [136]	Qwen2.5-VL-7B + Gemini 2.0	2025	✗	✗	~	~	~	✓	2.5
VADTree [137]	LLaVA-Video-7B + DeepSeek-R1-14B	2025	~	✗	~	✗	✓	✓	3.0
MoniTor [134]	BLIP-2 × 5 + GLM-4-Flash	2025	~	~	✓	✓	✗	✓	4.0
AnyAnomaly [141]	Chat-UniVi-7B / MiniCPM-8B	2026	~	✗	✓	✓	✓	✓	4.5

APPENDIX. REPRODUCIBILITY CHARACTERISTICS OF LM-BASED VAD METHODS

To provide a controlled view of reproducibility across recent LM-based VAD literature, Table 8 summarizes representative methods published at the selected top-tier venues considered in this survey. We include methods that invoke a generative LLM or MLLM at inference time and report the corresponding foundation model together with whether key implementation details are fully specified, partially specified, or not reported.

The table focuses on six reproducibility characteristics: model version/checkpoint, decoding parameters, number of inference calls, hardware, code availability, and data availability. These characteristics are not consistently documented across the literature, making direct comparison of runtime, latency, or computational cost difficult. The table is therefore intended to provide a compact and controlled indication of reporting completeness and model dependence rather than a unified computational benchmark.