HeadHunt-VAD: Hunting Robust Anomaly-Sensitive Heads in MLLM for Tuning-Free Video Anomaly Detection

Zhaolin Cai[1], Fan Li[2*], Ziwei Zheng[2], Haixia Bi[2], Lijun He[2]

[1]School of Computer Science and Technology, Xinjiang University
[2]School of Information and Communications Engineering, Xi'an Jiaotong University
107552301311@stu.xju.edu.cn, lifan@mail.xjtu.edu.cn, ziwei.zheng@stu.xjtu.edu.cn,
haixia.bi@xjtu.edu.cn, lijunhe@mail.xjtu.edu.cn

Abstract

Video Anomaly Detection (VAD) aims to locate events that deviate from normal patterns in videos. Traditional approaches often rely on extensive labeled data and incur high computational costs. Recent tuning-free methods based on Multimodal Large Language Models (MLLMs) offer a promising alternative by leveraging their rich world knowledge. However, these methods typically rely on textual outputs, which introduces information loss, exhibits normalcy bias, and suffers from prompt sensitivity, making them insufficient for capturing subtle anomalous cues. To address these constraints, we propose HeadHunt-VAD, a novel tuning-free VAD paradigm that bypasses textual generation by directly hunting robust anomaly-sensitive internal attention heads within the frozen MLLM. Central to our method is a Robust Head Identification module that systematically evaluates all attention heads using a multi-criteria analysis of saliency and stability, identifying a sparse subset of heads that are consistently discriminative across diverse prompts. Features from these expert heads are then fed into a lightweight anomaly scorer and a temporal locator, enabling efficient and accurate anomaly detection with interpretable outputs. Extensive experiments show that HeadHunt-VAD achieves state-of-the-art performance among tuning-free methods on two major VAD benchmarks while maintaining high efficiency, validating head-level probing in MLLMs as a powerful and practical solution for real-world anomaly detection.

Introduction

Video Anomaly Detection (VAD) aims to identify and localize events that deviate from normal patterns in video sequences, which is a critical task for ensuring public safety (Sultani, Chen, and Shah 2018), industrial quality control (Roth et al. 2022), and autonomous driving (Yao et al. 2023). Traditional paradigms including unsupervised (Lv et al. 2021), weakly-supervised (Wu et al. 2024b), and fully-supervised methods (Liu et al. 2018), have made significant progress in VAD. However, these approaches often struggle to generalize to diverse anomalies and typically demand large-scale annotated training data along with substantial computational resources, hindering their scalability and practical deployment in real-world settings.

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

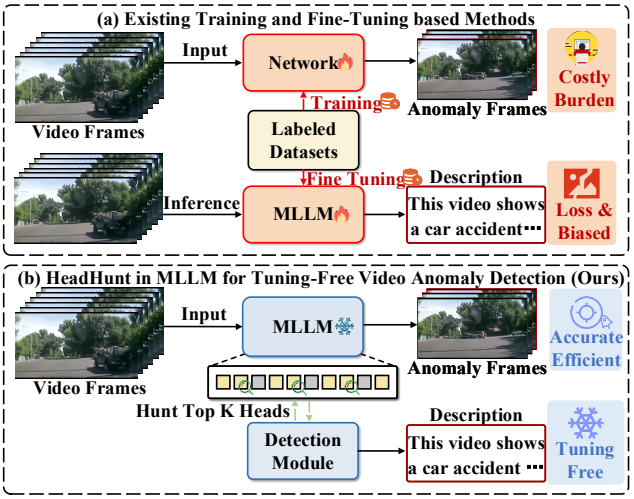


Figure 1: Comparison of VAD methods. Existing methods rely on training or fine tuning with large-scale datasets. HeadHunt-VAD is a tuning-free approach that detect anomalies using sparse expert heads within a frozen MLLM.

The emergence of Multimodal Large Language Models (MLLMs) (Liu et al. 2023) has presented novel avenues for VAD, leveraging their powerful cross-modal reasoning and adaptability. While some methods adapt MLLMs to VAD through supervised fine-tuning (Zhang et al. 2024b), this reintroduces the issues of data dependency and computational costs. Recent tuning-free VAD methods attempt to address these constraints (Zanella et al. 2024) by prompting the model to reason and generate textual descriptions to detect anomalies. However, these methods mostly rely on the final textual output of MLLMs, which suffer from three key limitations. First, converting high-dimensional visual information into natural language inevitably incurs information loss, potentially omitting subtle yet crucial anomalous cues. (Zhang et al. 2025). Second, MLLMs exhibit normalcy bias, tending to describe common objects while ignoring unusual details that define an anomaly. Third, their outputs are sensitive to prompt phrasing, often producing inconsistent predictions for semantically equivalent queries on the same video. These limitations motivate a shift away from textual outputs towards probing the internal representations of MLLMs.

Recent studies revealed that intermediate layers of large language models contain richer representations than output layer (Chen et al. 2024; Sun et al. 2024; Skean et al. 2025). While some works have started to leverage features from entire intermediate layers for diverse tasks including VAD (Orgad et al. 2024; Cai et al. 2025), this is still a coarse-grained approach. A transformer layer aggregates outputs from multiple attention heads (Vaswani et al. 2023), each with potential functional specializations (Zheng et al. 2025). The discriminative signals from a few heads risk being drowned out by outputs from heads focusing on mundane background features, therefore leading to representation dilution. Although prior work has analyzed heads for passive interpretation or guiding model pruning for efficiency (Baan et al. 2019; Jin et al. 2024b). The proactive identification and utilization of heads for VAD remains unexplored. This motivates a shift from layer-level to head-level representation analysis, aiming to identify and leverage fine-grained, informative attention heads for improved anomaly detection.

In this paper, we propose HeadHunt-VAD, a novel tuning-free paradigm for video anomaly detection that leverages sparse, expert attention heads within frozen Multimodal Large Language Models (MLLMs). As illustrated in Figure 1, our approach diverges from conventional methods by directly investigating internal representations of MLLM, mitigating the need for costly training and the problem of information loss. To overcome challenges of representation dilution, normalcy bias, and prompt sensitivity of MLLMs, we introduce the Robust Head Identification (RHI) module, which systematically hunts for a sparse set of anomaly-sensitive heads through multi-criteria analysis of saliency and stability across diverse prompts. Features from these expert heads are then used to construct a lightweight anomaly scorer and temporal locator via few-shot probing, requiring as little as 1% of the training sets. This strategy effectively minimizing data dependency and supervision while enabling accurate anomaly detection and localization without any fine-tuning of the MLLM. We evaluate HeadHunt-VAD on two benchmark datasets including UCF-Crime (Sultani, Chen, and Shah 2018) and XD-Violence (Wu et al. 2020), comprehensive experiments demonstrate the effectiveness of HeadHunt-VAD in video anomaly detection.

Our main contributions are summarized as follows:

- We propose HeadHunt-VAD, a novel tuning-free paradigm for video anomaly detection that for the first time proactively hunting sparse, expert attention heads within a frozen MLLM, thereby addressing the issues of information loss and representation dilution.
- We introduce a novel Robust Head Identification (RHI) module, a prompt-invariant selection module that discovers anomaly-sensitive heads through saliency and stability analysis, effectively addressing the challenge of prompt sensitivity and enhancing robustness.
- We conduct extensive experiments on UCF-Crime and XD-Violence benchmarks, HeadHunt-VAD achieves state-of-the-art performance among tuning-free methods, achieved strong performance without any MLLM fine-tuning and with remarkable data efficiency.

Related Work

Traditional Video Anomaly Detection

Traditional VAD methods are typically categorized into supervised (Liu et al. 2018; Landi, Snoek, and Cucchiara 2019), weakly-supervised (Li et al. 2022; Wang et al. 2024; Zhang et al. 2024a), and unsupervised (Tur et al. 2023) paradigms. Supervised approaches achieve high accuracy by leveraging frame-level anomaly annotations but require costly and labor-intensive labeling. To reduce this burden, weakly supervised methods use video-level labels for training but struggle with subtle cues and may exhibit biased predictions. Unsupervised methods train exclusively on normal video data to model typical patterns and detect deviations (Hasan et al. 2016; Xu et al. 2017; Yang et al. 2023). Despite their effectiveness in specific settings, these methods are constrained by their reliance on large-scale training data, limiting their scalability and applicability in the real world.

Video Anomaly Detection with LLMs and MLLMs

The advent of large language models(LLM) (Touvron et al. 2023) and multimodal LLMs (Zhu et al. 2023) has introduced new approaches to VAD, which can be broadly categorized into fine-tuning and tuning-free paradigms. Fine-tuning methods adapt pre-trained MLLMs for VAD (Zhang, Li, and Bing 2023; Yuan et al. 2024; Zhang et al. 2024c), achieving strong results but reintroducing the need for extensive labeled data and computational resources. In contrast, tuning-free methods leverage the powerful cross-modal reasoning and zero-shot capabilities of frozen MLLMs for VAD (Shao et al. 2025). For example, LAVAD (Zanella et al. 2024) detects anomalies by prompting MLLM to generate descriptions and reason with extra LLM; VERA (Ye, Liu, and He 2025) employs verbalized learning to elicit more effective reasoning with MLLM through optimized prompts. However, these methods risk losing subtle visual patterns during the vision-to-text translation. This limitation motivates bypassing the textual output to probe more direct internal representations in MLLM.

Internal Analysis in LLMs and MLLMs

Recent research has demonstrated that intermediate layers of LLMs often contain richer representations than the final outputs (Jin et al. 2024a; Ju et al. 2024; Merullo, Eickhoff, and Pavlick 2024), with mid-layer features shown to enhance performance in diverse downstream tasks (Skean et al. 2025; Orgad et al. 2024). Recent work in MLLMs has revealed that the integration of visual and linguistic information occurs in the middle layers (Zhang, Dong, and Kawaguchi 2024; Zhang et al. 2025). Moreover, recent work has identified an information-rich phenomenon in intermediate layers, where internal features exhibit enhanced discriminative power for VAD (Cai et al. 2025). However, these analyses remain at the layer level. Although some studies has explored attention heads for model pruning (Baan et al. 2019; Li et al. 2023; Jin et al. 2024b), the proactive and goal-driven utilization of specific attention heads for video anomaly detection remains an underexplored and promising frontier.

Methodology

Preliminaries: Representation Dilution in MLLMs

The multi-head attention (MHA) module is the cornerstone of the transformer architecture, which underpins Multimodal Large Language Models (MLLMs). Given an input feature sequence $\mathbf{X} \in \mathbb{R}^{L \times D}$, where L is the sequence length and D is the model dimension, MHA projects the input into queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$). The output of the j-th attention head, $\mathbf{h}_j \in \mathbb{R}^{L \times d_h}$, is computed as:

$$\mathbf{h}_j = \text{Attention}(\mathbf{Q}\mathbf{W}_j^Q, \mathbf{K}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V) \quad (1)$$

where $\mathbf{W}_j^Q, \mathbf{W}_j^K, \mathbf{W}_j^V \in \mathbb{R}^{D \times d_h}$ are the projection matrices for the j-th head, and d_h is the dimension of head. The outputs of all N_h heads are then concatenated and projected by a final linear layer $\mathbf{W}^O \in \mathbb{R}^{N_h d_h \times D}$ to produce the aggregated output:

$$\mathbf{X}' = \text{Concat}(\mathbf{h}_1, \ldots, \mathbf{h}_{N_h})\mathbf{W}^O \quad (2)$$

However, this final aggregation step can be detrimental for fine-grained tasks. It averages outputs from all heads, causing sharp, discriminative signals from specialized, anomaly-sensitive heads to be diluted by less relevant ones. Therefore we term this issue as representation dilution. As illustrated in Figure 2, individual heads in intermediate layers often exhibit superior discriminative power compared to the aggregated final output, where crucial signals are lost. This observation motivates our work, as any method relying on the final, diluted output $\mathbf{X}'$ is inherently handicapped for tasks like video anomaly detection that demand high sensitivity to subtle cues. Therefore, HeadHunt-VAD is designed to circumvent this issue by operating directly on the pre-aggregation outputs from all heads across all layers. We denote the set of all head outputs as $\{\mathbf{h}_k\}_{k=1}^{N_{\text{total}}}$, where k is a global index for each head and $N_{\text{total}} = N_{\text{layers}} \times N_h$ is the total number of heads in the MLLM. This allows us to directly identify and exploit the most informative heads.

HeadHunt-VAD Framework Overview

HeadHunt-VAD is a highly efficient paradigm for video anomaly detection by leveraging a frozen Multimodal Large Language Model (MLLM) without any fine-tuning. As depicted in Figure 3, our approach is structured into two main stages: an offline preparation stage and a real-time online inference stage. In the offline stage, the core HeadHunt process commences with the Robust Head Identification (RHI) module, which systematically evaluates all attention heads in multi-criteria analysis to select a prompt-robust set of consensus expert heads based on their discriminative power. Based on features from selected subset of heads, we construct two lightweight modules: a logistic regression-based Anomaly Scorer and a calibrated Temporal Locator. The online stage is engineered for maximum throughput. For an incoming video, it performs a single forward pass to perform targeted feature extraction from the identified expert heads. These features are then processed by the lightweight scorer and locator for precise, real-time anomaly localization and finally generate descriptions for detected anomalies.

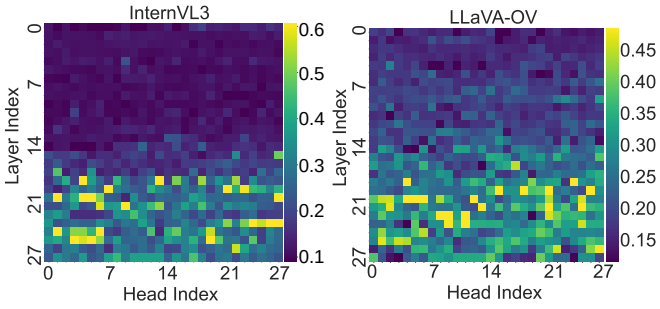


Figure 2: Visualization of attention head heatmaps in MLLMs including InternVL3 and LLaVA-OV.

Offline Stage: Robust Head Identification and Module Preparation

The offline stage focus on head identification and module construction to enable efficient and accurate real-time inference in the subsequent online stage. All attention heads across layers are treated as a unified set, indexed by $k \in \{1, \ldots, N_{\text{total}}\}$, where $N_{\text{total}} = N_{\text{layers}} \times N_h$.

Head Saliency Characterization To evaluate the usefulness of each head, we first extract a representative feature vector from MLLM. For a given video and text prompt, we perform a single forward pass through the frozen MLLM and intercept the output of each head $\mathbf{h}_k \in \mathbb{R}^{L \times d_h}$ before the final aggregation layer. From this matrix, we extract the feature vector $\mathbf{x}_k \in \mathbb{R}^{d_h}$ corresponding to the first generated token, as it encapsulates a summary of the input sequence for the initial generative decision and avoids the cost of the full auto-regressive decoding process. By processing all videos in our calibration set, we form feature sets $\mathcal{X}_{k,n}$ (normal) and $\mathcal{X}_{k,a}$ (abnormal) for each head k. To obtain a holistic and robust assessment of discriminative power with each head after the extraction, we construct comprehensive metrics to effectively calculate the saliency score. The multi-faceted evaluation spans across statistical, geometric, and information-theoretic perspectives, thereby ensuring our selection is not biased by a single criterion. The saliency of head k under a given prompt p_m is then characterized as follows. Detailed derivations and pipeline algorithms are provided in the supplementary materials.

- **Linear Discriminant Analysis Score (S_{LDA}):** Measures linear separability, aligning with our goal of finding features suitable for an efficient linear classifier. A higher score signifies greater class separation.

$$S_{\text{LDA}}(k) = (\boldsymbol{\mu}_{k,a} - \boldsymbol{\mu}_{k,n})^T (\mathbf{S}_{W,k})^{-1} (\boldsymbol{\mu}_{k,a} - \boldsymbol{\mu}_{k,n}) \quad (3)$$

  where $\boldsymbol{\mu}_{k,c}$ and $\mathbf{S}_{W,k}$ are the class-specific mean and within-class scatter matrix for head $k.$,
- **Symmetrized KL Divergence (S_{KL}):** Quantifies the dissimilarity between the probability distributions of normal and abnormal representations, which are modeled as multivariate Gaussian distributions, thereby capturing non-linear statistical differences.

$$S_{\text{KL}}(k) = \frac{1}{2}[D_{KL}(\mathcal{N}_{k,a} || \mathcal{N}_{k,n}) + D_{KL}(\mathcal{N}_{k,n} || \mathcal{N}_{k,a})] \quad (4)$$

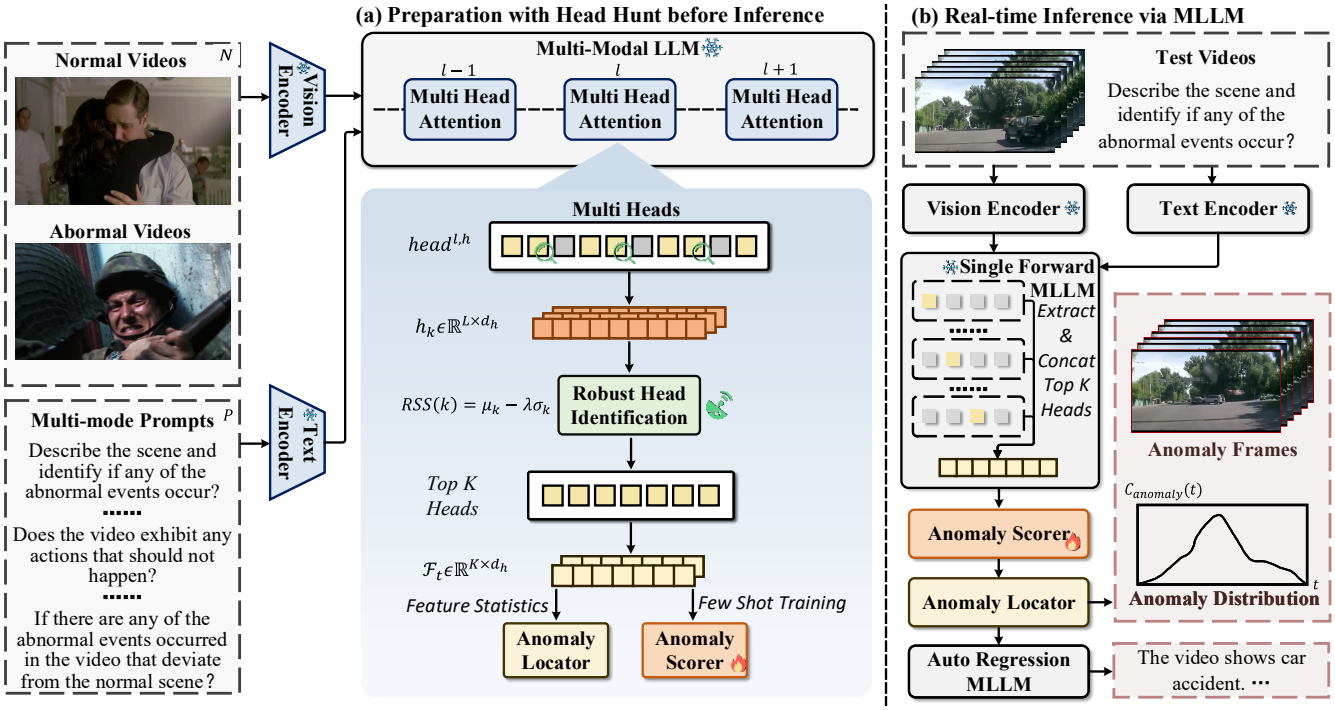


Figure 3: The overall architecture of HeadHunt-VAD. The offline HeadHunt phase identifies consensus expert heads and prepares downstream modules. The efficient online inference phase leverages these for real-time detection.

- **Maximum Mean Discrepancy (S_{MMD}):** Measures the distributional discrepancy in a Reproducing Kernel Hilbert Space (RKHS), making it sensitive to complex differences in distribution shape.

$$S_{\text{MMD}}(k)^2 = \left\| \frac{1}{|\mathcal{X}_{k,a}|} \sum_{\mathbf{x} \in \mathcal{X}_{k,a}} \phi(\mathbf{x}) - \frac{1}{|\mathcal{X}_{k,n}|} \sum_{\mathbf{y} \in \mathcal{X}_{k,n}} \phi(\mathbf{y}) \right\|_{\mathcal{H}}^2 \tag{5}$$

  where $\phi(\cdot)$ is the kernel-induced feature map into the RKHS $\mathcal{H}$.
- **Normalized Mutual Information (S_{NMI}):** Measures the statistical dependency between feature representations and ground-truth labels. We perform K-Means clustering ($K = 2$) on the features to get predicted labels $Y_{\text{pred}}^{(k)}$.

$$S_{\text{NMI}}(k) = \frac{I(Y_{\text{true}}, Y_{\text{pred}}^{(k)})}{\sqrt{H(Y_{\text{true}}) H(Y_{\text{pred}}^{(k)})}} \tag{6}$$

  where $I(\cdot,\cdot)$ is the mutual information and $H(\cdot)$ is the entropy.

Although these metrics provide a comprehensive evaluation of discriminative capability, identifying universally effective heads requires assessing their stability across diverse textual inputs. We therefore introduce the Robust Head Selection.

Robust Head Selection To identify heads that are universally effective, we assess their stability across diverse textual inputs. We evaluate each head over a set of M diverse prompts, $\mathcal{P} = \{p_1, \ldots, p_M\}$. For each head k and prompt p_m, we first compute every saliency scores and normalize them to $[0, 1]$ across all heads. The score $S(k, p_m)$ is the average of normalized saliency scores.

The mean performance μ_k and instability (standard deviation) σ_k are calculated for each head across the prompt set:

$$\mu_k = \frac{1}{M} \sum_{m=1}^{M} S(k, p_m) \tag{7}$$

$$\sigma_k = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left(S(k, p_m) - \mu_k\right)^2} \tag{8}$$

Inspired by risk-aversion principles, we define a Robust Saliency Score (RSS) to favor heads with both high average performance and low variance across prompts:

$$\text{RSS}(k) = \mu_k - \lambda\sigma_k \tag{9}$$

where λ is a hyperparameter that controls the trade-off between performance and stability. A head that performs well for one prompt but poorly for another will have a high σ_k and thus be penalized, as we seek heads that are not only high returns (high mean saliency μ_k) but also low risk (low performance volatility σ_k across prompts). We calculate and rank all heads by their RSS and select the top-K heads, whose indices form our consensus expert head set $\mathcal{I}^*$.

Anomaly Scorer Training The anomaly scorer is implemented using logistic regression, chosen for its efficiency and interpretability. For each video i in the calibration set, a composite feature vector $\mathbf{z}_i \in \mathbb{R}^{K \cdot d_h}$ is constructed by

concatenating feature $\{\mathbf{x}_k\}_{k\in\mathcal{I}^*}$ from the K expert heads. Given the training set $\{(\mathbf{z}_i, y_i^{(\mathrm{v})})\}_{i=1}^N$, where $y_i^{(\mathrm{v})} \in \{0,1\}$ is the video-level label, the model learns a weight vector $\mathbf{w}$ and bias b to predict the anomaly probability:

$$p_i^{(\mathrm{v})} = \sigma(\mathbf{w}^T\mathbf{z}_i + b) \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function. The model is trained by minimizing the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i^{(\mathrm{v})}\log p_i^{(\mathrm{v})} + (1-y_i^{(\mathrm{v})})\log(1-p_i^{(\mathrm{v})})\right] \tag{11}$$

This lightweight model effectively leverages the discriminative features from the expert heads while ensuring computational efficiency and accuracy.

Temporal Locator Calibration To convert anomaly probabilities into coherent temporal event predictions, a temporal locator is introduced. This component operates on a sequence of probabilities $\{p_t\}$, which are generated by applying the trained Anomaly Scorer to the feature vector of each individual temporal segment from the validation set. The locator then refines the resulting raw probability sequence $\mathbf{p}$ by applying a transformation that includes convolution with a 1D Gaussian kernel, G_{σ_g}, followed by binarization using a threshold τ.

The standard deviation σ_g of the Gaussian kernel and the threshold τ are determined by optimizing the frame-level F1-score on a validation set via grid search. For each candidate pair (σ_g, τ), the validation videos are processed by first generating the raw probability sequence $\mathbf{p}$ and then applying the complete transformation:

$$p'_t = (\mathbf{p} * G_{\sigma_g})_t = \sum_j p_j \cdot G(t-j;\sigma_g) \tag{12}$$

The resulting binarized outputs are compared with ground-truth annotations to compute the F1-score. The parameter combination (σ_g^*, τ^*) that maximizes the score is selected. This data-driven calibration ensures the Temporal Locator is configured for precise anomaly localization.

Online Stage: Real-time Video Anomaly Detection

The online stage with MLLMs focuses on processing unseen videos to detect and localize anomaly frames and optionally provide comprehensive anomaly descriptions.

Single Forward Pass with Feature Extraction An incoming video is first divided into a sequence of non-overlapping temporal segments $\{S_1, S_2, \ldots, S_T\}$. For each segment S_t, we uniformly sample F frames. These frames, along with a task-specific textual prompt, are processed in a single forward pass through the frozen MLLM. During this pass, we perform targeted feature extraction by retrieving and concatenating the outputs from the K expert heads in our consensus set $\mathcal{I}^*$, thereby avoiding the costly auto-regressive decoding. Following the procedure from the offline stage, we extract the feature corresponding to the first token from each selected head and concatenate them to form a single, robust feature vector $\mathbf{f}_t \in \mathbb{R}^{K\cdot d_h}$ for segment S_t.

Frame-level Anomaly Scoring and Localization The feature vector $\mathbf{f}_t$ is fed into the pre-trained Anomaly Scorer to compute the anomaly probability for each segment:

$$p_t = \sigma(\mathbf{w}^T\mathbf{f}_t + b) \tag{13}$$

This yields a sequence of anomaly scores $\mathbf{p} = (p_1, \ldots, p_T)$ for the entire video. The probability sequence $\mathbf{p}$ is then processed by the calibrated Temporal Locator. The sequence is first smoothed using the pre-calibrated Gaussian kernel $G_{\sigma_g^*}$ to enforce temporal consistency, producing a smoothed sequence $\mathbf{p}'$. The final detection $\hat{y}_t \in \{0,1\}$ is made by applying the calibrated threshold τ^*:

$$\hat{y}_t = [p'_t > \tau^*] \tag{14}$$

where $[\cdot]$ denotes the Iverson bracket. Consecutive segments where $\hat{y}_t = 1$ are grouped to yield precise temporal localizations of anomalous events.

Event-level Explanation Generation For enhanced interpretability, detected anomalous clips in a video can be passed back to the full auto-regressive MLLM. Given the clips and textual inputs, the model generates a natural language explanation of the event. This final step completes the process from detection to understanding, providing a comprehensive and interpretable explanations for videos.

Experiments

Experimental Settings

Datasets and evaluation metrics We evaluate our method on two widely used benchmarks for video anomaly detection: UCF-Crime (Sultani, Chen, and Shah 2018) and XD-Violence (Wu et al. 2020). These datasets contain long, untrimmed videos with real-world anomalies, providing a realistic and challenging evaluation environment.

For performance evaluation, we use standard metrics from the literature: frame-level AUC (Area Under the ROC Curve) for UCF-Crime and average precision (AP) for XD-Violence. Both metrics are widely used in prior work and provide a comprehensive assessment of detection performance, higher values indicate better results.

Implementation Details We use the InternVL3 model as the frozen MLLM backbone of HeadHunt-VAD. Each video is segmented at 48 frames intervals and uniformly sample $F = 16$ frames as the input to MLLM. The offline phase use a small calibration subset comprising 1% of the training data from each benchmark. The prompt set $\mathcal{P}$ consists of five varied prompts. We select the top $K = 5$ expert heads for feature extraction and set the instability penalty to $\lambda = 0.5$. The Anomaly Scorer is a logistic regression module. The Temporal Locator uses 1D Gaussian smoothing kernel $\sigma_g = 1.5$ and threshold of $\tau = 0.65$, both optimized for frame-level F1-score on the validation set. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Detailed hyperparameter analysis, prompts, and further analysis are provided in the supplementary materials.

Mode	Methods	Backbone	AUC (%)
Weakly Supervised	Wu et al. (2020)	I3D	82.44
	MIST(2021)	I3D	82.30
	RTFM(2021)	I3D	84.30
	S3R(2022)	I3D	85.99
	MSL(2022)	I3D	85.30
	UR-DMU(2023)	I3D	86.97
	MFGN(2022)	I3D	86.98
	Wu et al.(2024a)	ViT	86.40
	CLIP-TSA(2023)	ViT	87.58
	Yang et al.(2024)	ViT	87.79
	VadCLIP(2023)	ViT	88.02
Self Supervised	TUR et al.(2023)	Resnet	66.85
	BODS(2019)	I3D	68.26
	GODS(2019)	I3D	70.46
Unsupervised	GCL(2022)	ResNext	71.04
	DYANNET(2023)	I3D	84.50
Tuning-Free Multimodal VAD	Zero-Shot CLIP(2021)	ViT	53.16
	ZS ImageBind (Video)(2023)	ViT	55.78
	ZS ImageBind (Image)(2023)	ViT	53.65
	LLAVA-1.5(2024)	ViT	72.84
	LAVAD(2024)	ViT	80.28
	EventVAD(2025)	ViT	82.03
	VERA(2025)	ViT	86.55
	HiProbeVAD(2025)	ViT	86.72
	HeadHunt-VAD	**ViT**	**87.03**

Table 1: Comparison with existing methods on the UCF-Crime dataset.

Comparison with State-of-the-Art Methods

Tables 1 and 2 present the main results on the UCF-Crime and XD-Violence datasets. On UCF-Crime, HeadHunt-VAD achieves an AUC of 87.03%, setting a new state-of-the-art among tuning-free and unsupervised methods. Our method also remains competitive with many weakly-supervised methods that require extensive training on large-scale labeled datasets. On the XD-Violence dataset, HeadHunt-VAD achieves an AP of 82.63%, which represents an improvement over other tuning-free approaches and significantly narrows the performance gap to leading weakly-supervised models. The consistent and robust performance across these two benchmarks provides strong evidence for the effectiveness of our proposed approach for VAD.

Efficiency Analysis

HeadHunt-VAD achieves practical efficiency by reducing computational cost, model complexity, and data dependency. As shown in Figure 4, our single forward pass strategy avoids costly auto-regressive decoding and yields a significant computational efficiency. This efficiency extends to scalability. Our few-shot calibration uses less than 1% of the data required for fine-tuning approaches. By extracting features from only top-K expert heads instead of all heads, we dramatically reduce the feature dimension from over 100K to just 640. The combination of fast detection, minimal data requirements, and low feature complexity validates HeadHunt-VAD as an effective and practical VAD solution.

Mode	Methods	Backbone	AP (%)
Weakly Supervised	Wu et al.(2020)	I3D	73.20
	RTFM(2021)	I3D	77.81
	MSL(2022)	I3D	78.28
	MFGN(2022)	I3D	79.19
	S3R(2022)	I3D	80.26
	UR-DMU(2023)	I3D	81.66
	Wu et al.(2024a)	ViT	66.53
	CLIP-TSA(2023)	ViT	82.19
	Yang et al.(2024)	ViT	83.68
	VadCLIP(2023)	ViT	84.51
Tuning-Free Multimodal VAD	Zero-Shot CLIP(2021)	ViT	17.83
	ZS ImageBind (Video)(2023)	ViT	25.36
	ZS ImageBind (Image)(2023)	ViT	27.25
	LLAVA-1.5(2024)	ViT	50.26
	LAVAD(2024)	ViT	62.01
	EventVAD(2025)	ViT	64.04
	HiProbeVAD(2025)	ViT	82.15
	HeadHunt-VAD	**ViT**	**82.63**

Table 2: Comparison of existing methods on the XD-Violence dataset.

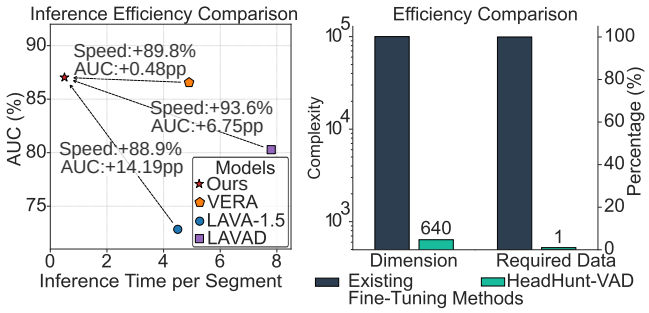


Figure 4: Efficiency comparison of HeadHunt-VAD including inference time, dimension complexity, and required data.

Ablation Studies

Effectiveness of the Robust Head Identification We evaluate the proposed Robust Head Identification (RHI) by comparing it with different feature extraction strategies (Table 3). Using all attention heads from the final layer leads to substantial performance degradation, while random head selection further deteriorates results, confirming the necessity of informed head selection. We also assess prompt sensitivity. A single generic prompt underperforms, and although a manually crafted oracle prompt yields slightly better results, it requires fixed prompt for all time. In contrast, RHI automatically identifies a robust subset across prompts, achieving comparable performance with different text inputs.

Impact of Anomaly Scorer We analyze the role of the final anomaly scorer with other common lightweight models, with results presented in Table 3. The support vector machine (SVM) achieves performance comparable to the proposed method, while multi-layer perceptron (MLP) yields a marginal improvement of 0.22% in AUC and 0.18% in AP. We adopt logistic regression for its optimal balance of ac-

Method / Variation	AUC (%)	AP (%)
HeadHunt-VAD (Full Model)	**87.03**	**82.63**
Ablation on Robust Head Identification (RHI)		
w/ Full Layer Features	80.15	72.10
w/ Random-K Heads	66.65	45.33
w/ Single Coarse Prompt	81.86	74.52
w/ Oracle (Detailed) Prompt	87.11	82.95
Ablation on Anomaly Scorer (Ours: Logistic Regression)		
SVM (Linear Kernel)	84.95	76.52
MLP (2-Layer)	87.25	82.81
Ablation on Temporal Locator (Ours: $\tau = 0.65$)		
w/o Gaussian Smoothing	82.44	75.88
w/ Fixed Threshold ($\tau = 0.25$)	70.91	55.21
w/ Fixed Threshold ($\tau = 0.50$)	80.32	71.49
w/ Fixed Threshold ($\tau = 0.75$)	71.15	58.10

Table 3: Effectiveness of the Robust Head Identification module, Anomaly Scorer, and the Temporal Locator.

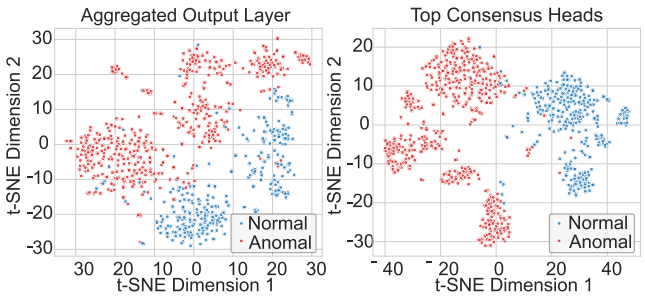
curacy, computational efficiency, and interpretability, which aligns with overall emphasis on practicality and ease of deployment of our method.

Impact of Temporal Locator Our temporal locator consists of two key components: temporal score smoothing and a data-driven threshold calibration, with results in Table 3. Removing the Gaussian smoothing step results in a significant AUC drop of 4.59% on UCF-Crime, which demonstrates its crucial role in stabilizing frame-level predictions and reducing spurious noise from isolated, high-scoring frames. Furthermore, replacing our data-driven threshold calibration with fixed, arbitrary values leads to severe performance degradation. This confirms that a calibrated, data-aware threshold is indispensable for accurately segmenting anomalous events from raw scores.

Qualitative Analyses

Feature Space Visualization Figure 5 visualizes the feature distributions of normal and abnormal samples using t-SNE. Features from the full output layer exhibit significant class overlap due to representation dilution. In contrast, features from our expert heads form clearly separated and compact clusters, confirming their superior discriminative power and better separability for anomaly detection.

Qualitative Visualization Figure 6 shows qualitative results from XD-Violence. For each video, the plot shows the anomaly curves across different frames. For abnormal video, the frame-level probability curve generated by our method rises sharply and aligns precisely with the ground-truth temporal segments. Conversely, the anomaly curve of normal video remains consistently low and stable. These results demonstrate that HeadHunt-VAD provide reliable and accurate temporal localization of anomalous events. More visualization are provided in the supplementary materials.



(a) Features from Full Layer (b) Features from Expert Heads

Figure 5: t-SNE visualization. (a) Full output layer show class overlap due to representation dilution. (b) Features from HeadHunt-VAD form distinct, separable clusters.

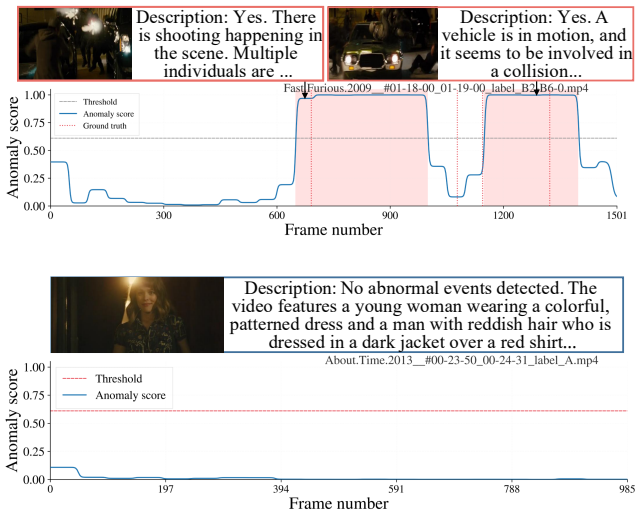


Figure 6: Qualitative results of HeadHunt-VAD on XD-Violence. Each panel shows a representative video snippet and the anomaly curve. The shaded regions denote the detected anomaly frames that surpasses the threshold. Further generated descriptions are also provided.

Conclusion

In this paper, we introduced HeadHunt-VAD, a novel tuning-free paradigm that moves beyond the limitations of lossy textual outputs and diluted layer-level representations in MLLM-based VAD. Our method proactively identifies and leverages a sparse set of robust, anomaly-sensitive attention heads within a frozen MLLM, pinpointed by a multi-criteria Robust Head Identification module. Features from these expert heads are then channeled into lightweight modules for highly efficient scoring and localization. Extensive experiments on two major benchmarks demonstrate that HeadHunt-VAD establishes a new state-of-the-art among tuning-free methods, achieving superior performance with remarkable computational and data efficiency. Our work validates head-level probing as a powerful and practical paradigm for real-world video anomaly detection.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62471376, in part by the National Key Research and Development Program of China under Grant 2022YFA1003800, and in part by the XJTU Research Fund for Al Science under Grant 2025YXYC004.

References

Baan, J.; ter Hoeve, M.; van der Wees, M.; Schuth, A.; and de Rijke, M. 2019. Understanding multi-head attention in abstractive summarization. arXiv:1911.03898.

Cai, Z.; Li, F.; Zheng, Z.; and Qin, Y. 2025. HiProbe-VAD: Video Anomaly Detection via Hidden States Probing in Tuning-Free Multimodal LLMs. arXiv:2507.17394.

Chen, N.; Wu, N.; Liang, S.; Gong, M.; Shou, L.; Zhang, D.; and Li, J. 2024. Is bigger and deeper always better? Probing LLaMA across scales and layers. arXiv:2312.04333.

Chen, Y.; Liu, Z.; Zhang, B.; Fok, W.; Qi, X.; and Wu, Y.-C. 2022. MGFN: magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection.

Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14004–14013. IEEE.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind One Embedding Space to Bind Them All. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15180–15190. IEEE.

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning Temporal Regularity in Video Sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 733–742. IEEE.

Jin, M.; Yu, Q.; Huang, J.; Zeng, Q.; Wang, Z.; Hua, W.; Zhao, H.; Mei, K.; Meng, Y.; Ding, K.; Yang, F.; Du, M.; and Zhang, Y. 2024a. Exploring concept depth: how large language models acquire knowledge at different layers? arXiv:2404.07066.

Jin, P.; Zhu, B.; Yuan, L.; and Yan, S. 2024b. MoH: multi-head attention as mixture-of-head attention. arXiv:2410.11842.

Joo, H. K.; Vo, K.; Yamazaki, K.; and Le, N. 2023. CLIP-TSA: Clip-Assisted Temporal Self-Attention for Weakly-Supervised Video Anomaly Detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, 3230–3234. IEEE.

Ju, T.; Sun, W.; Du, W.; Yuan, X.; Ren, Z.; and Liu, G. 2024. How large language models encode context knowledge? A layer-wise probing study. arXiv:2402.16061.

Landi, F.; Snoek, C. G. M.; and Cucchiara, R. 2019. Anomaly locality in video surveillance. arXiv:1901.10364.

Li, C.; Wang, S.; Zhang, Y.; Zhang, J.; and Zong, C. 2023. Interpreting and exploiting functional specialization in multi-head attention under multi-task learning. arXiv:2310.10318.

Li, G.; Cai, G.; Zeng, X.; and Zhao, R. 2022. Scale-Aware Spatio-Temporal Relation Learning for Video Anomaly Detection. In *Computer Vision - ECCV 2022*, 333–350. Springer.

Li, S.; Liu, F.; and Jiao, L. 2022. Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2): 1395–1403.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26286–26296. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. https://arxiv.org/abs/2304.08485v2.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future Frame Prediction for Anomaly Detection - A New Baseline. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6536–6545. IEEE.

Lv, H.; Chen, C.; Cui, Z.; Xu, C.; Li, Y.; and Yang, J. 2021. Learning Normal Dynamics in Videos with Meta Prototype Network. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15420–15429. IEEE.

Merullo, J.; Eickhoff, C.; and Pavlick, E. 2024. Talking heads: understanding inter-layer communication in transformer language models. arXiv:2406.09519.

Orgad, H.; Toker, M.; Gekhman, Z.; Reichart, R.; Szpektor, I.; Kotek, H.; and Belinkov, Y. 2024. LLMs know more than they show: on the intrinsic representation of LLM hallucinations. arXiv:2410.02707.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. arXiv:2103.00020.

Roth, K.; Pemula, L.; Zepeda, J.; Scholkopf, B.; Brox, T.; and Gehler, P. 2022. Towards Total Recall in Industrial Anomaly Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14298–14308. IEEE.

Shao, Y.; He, H.; Li, S.; Chen, S.; Long, X.; Zeng, F.; Fan, Y.; Zhang, M.; Yan, Z.; Ma, A.; Wang, X.; Tang, H.; Wang, Y.; and Li, S. 2025. EventVAD: training-free event-aware video anomaly detection. arXiv:2504.13092.

Skean, O.; Arefin, M. R.; Zhao, D.; Patel, N.; Naghiyev, J.; LeCun, Y.; and Shwartz-Ziv, R. 2025. Layer by layer: uncovering hidden representations in language models. arXiv:2502.02013.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-World Anomaly Detection in Surveillance Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6479–6488. IEEE.

Sun, M.; Chen, X.; Kolter, J. Z.; and Liu, Z. 2024. Massive activations in large language models. arXiv:2402.17762.

Thakare, K. V.; Raghuwanshi, Y.; Dogra, D. P.; Choi, H.; and Kim, I.-J. 2023. DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network. In *2023*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 5530–5539. IEEE.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4955–4966. IEEE.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Tur, A. O.; Dall'Asen, N.; Beyan, C.; and Ricci, E. 2023. Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention is all you need. arXiv:1706.03762.

Wang, H.; Lai, C.; Sun, Y.; and Ge, W. 2024. Weakly Supervised Gaussian Contrastive Grounding with Large Multimodal Models for Video Question Answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5289–5298. ACM.

Wang, J.; and Cherian, A. 2019. GODS: Generalized One-Class Discriminative Subspaces for Anomaly Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8200–8210. IEEE.

Wu, J.-C.; Hsieh, H.-Y.; Chen, D.-J.; Fuh, C.-S.; and Liu, T.-L. 2022. Self-supervised Sparse Representation for Video Anomaly Detection. In *Computer Vision - ECCV 2022*, 729–745. Springer.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. In *Computer Vision - ECCV 2020*, 322–339. Springer-Verlag.

Wu, P.; Zhou, X.; Pang, G.; Sun, Y.; Liu, J.; Wang, P.; and Zhang, Y. 2024a. Open-vocabulary video anomaly detection. arXiv:2311.07042.

Wu, P.; Zhou, X.; Pang, G.; Yang, Z.; Yan, Q.; Wang, P.; and Zhang, Y. 2024b. Weakly Supervised Video Anomaly Detection and Localization with Spatio-Temporal Prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9301–9310. ACM.

Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; and Zhang, Y. 2023. VadCLIP: adapting vision-language models for weakly supervised video anomaly detection.

Xu, D.; Yan, Y.; Ricci, E.; and Sebe, N. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156: 117–127.

Yang, Z.; Liu, J.; and Wu, P. 2024. Text prompt with normality guidance for weakly supervised video anomaly detection. arXiv:2404.08531.

Yang, Z.; Liu, J.; Wu, Z.; Wu, P.; and Liu, X. 2023. Video Event Restoration Based on Keyframes for Video Anomaly Detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14592–14601. IEEE.

Yao, Y.; Wang, X.; Xu, M.; Pu, Z.; Wang, Y.; Atkins, E.; and Crandall, D. J. 2023. DoTA: Unsupervised Detection of Traffic Anomaly in Driving Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 444–459.

Ye, M.; Liu, W.; and He, P. 2025. VERA: Explainable Video Anomaly Detection via Verbalized Learning of Vision-Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 8679–8688.

Yuan, T.; Zhang, X.; Liu, K.; Liu, B.; Chen, C.; Jin, J.; and Jiao, Z. 2024. Towards Surveillance Video-and-Language Understanding: New Dataset, Baselines, and Challenges. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22052–22061. IEEE.

Zaheer, M. Z.; Mahmood, A.; Khan, M. H.; Segu, M.; Yu, F.; and Lee, S.-I. 2022. Generative Cooperative Learning for Unsupervised Video Anomaly Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14724–14734. IEEE.

Zanella, L.; Menapace, W.; Mancini, M.; Wang, Y.; and Ricci, E. 2024. Harnessing Large Language Models for Training-Free Video Anomaly Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18527–18536. IEEE.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: an instruction-tuned audio-visual language model for video understanding.

Zhang, H.; Wang, X.; Xu, X.; Huang, X.; Han, C.; Wang, Y.; Gao, C.; Zhang, S.; and Sang, N. 2024a. GlanceVAD: Exploring glance supervision for label-efficient video anomaly detection.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Han, C.; Huang, X.; Gao, C.; Wang, Y.; and Sang, N. 2024b. Holmes-VAD: towards unbiased and explainable video anomaly detection via multi-modal LLM. arXiv:2406.12235.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Huang, X.; Gao, C.; Zhang, S.; Yu, L.; and Sang, N. 2024c. Holmes-VAU: towards long-term video anomaly understanding at any granularity. arXiv:2412.06171.

Zhang, Y.; Dong, Y.; and Kawaguchi, K. 2024. Investigating layer importance in large language models. arXiv:2409.14381.

Zhang, Z.; Yadav, S.; Han, F.; and Shutova, E. 2025. Cross-modal information flow in multimodal large language models. arXiv:2411.18620.

Zheng, Z.; Zhao, J.; Yang, L.; He, L.; and Li, F. 2025. Spot Risks Before Speaking! Unraveling Safety Attention Heads in Large Vision-Language Models. arXiv:2501.02029.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. arXiv:2302.05160.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: enhancing vision-language understanding with advanced large language models. arXiv:2304.10592.