A Unified Reasoning Framework for Holistic Zero-Shot Video Anomaly Analysis

Dongheng Lin[1,2] **Mengxue Qu**[1] **Kunyang Han**[1]
Jianbo Jiao[2] **Xiaojie Jin**[1] **Yunchao Wei**[1]
[1] Institute of Information Science, Beijing Jiaotong University
[2] The MIx Group, University of Birmingham
{d.lin.2, j.jiao}@bham.ac.uk {FILL IN EMAILS...}@bjtu.edu.cn

Abstract

Most video-anomaly research stops at frame-wise detection, offering little insight into why an event is abnormal, typically outputting only frame-wise anomaly scores without spatial or semantic context. Recent video anomaly localization and video anomaly understanding methods improve explainability but remain data-dependent and task-specific. We propose a unified reasoning framework that bridges the gap between temporal detection, spatial localization, and textual explanation. Our approach is built upon a chained test-time reasoning process that sequentially connects these tasks, enabling holistic zero-shot anomaly analysis without any additional training. Specifically, our approach leverages intra-task reasoning to refine temporal detections and inter-task chaining for spatial and semantic understanding, yielding improved interpretability and generalization in a fully zero-shot manner. Without any additional data or gradients, our method achieves state-of-the-art zero-shot performance across multiple video anomaly detection, localization, and explanation benchmarks. The results demonstrate that careful prompt design with task-wise chaining can unlock the reasoning power of foundation models, enabling practical, interpretable video anomaly analysis in a fully zero-shot manner. Project Page: https://rathgrith.github.io/Unified_Frame_VAA/.

1 Introduction

Video anomaly analysis is a key application of computer vision for public security. Most early works formulate the task as temporal *Video Anomaly Detection* (VAD): mark the segments whose behavior deviates from learned normal patterns. Traditional detectors have reached high performance on benchmarks, yet they output only frame-wise scores and provide no insight into why the segment is abnormal. These limitations of interpretability motivate a broader shift from temporal detection to more downstream anomaly analysis tasks with user-friendly and explainable outputs, including spatial Video Anomaly Localization (VAL) [Liu and Ma, 2019, Weng et al., 2022] and textual Video Anomaly Understanding (VAU) tasks [Du et al., 2024, Tang et al., 2024, Zhang et al., 2024b] utilizing fine-tuned MLLMs. While these works provide either spatial or textual cues for better explainability to video anomalies separately, the previous works were mostly focused on a certain type of downstream tasks, which do not provide a holistic analysis to video anomalies, resulting in *"incompleteness"* from existing video anomaly analysis methods.

A further challenge is the heavy reliance on dataset-specific supervision. Traditional VAD and VAL models require temporal masks or spatial bounding boxes, yet anomaly definitions vary widely across datasets [Wu et al., 2020, Lu et al., 2013, Mahadevan et al., 2010], so a model tuned on one domain often fails on another [Wu et al., 2024a]. Also, in real-world applications, due to privacy and security

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Table 1: **Comparison of scopes and requirements of recent VLM-based methods.** ✓ = supported tasks, ✗ = not supported. Our framework is the only strictly zero-shot approach that handles all three.

Method	Supervision	Fine-tuning	Temporal	Spatial	Textual
LAVAD [Zanella et al., 2024]	None	None	✓	✗	✗
CUVA [Du et al., 2024]	Text	Prompt-tuning	✗	✗	✓
STPrompt [Wu et al., 2024b]	Weak class (closed-set)	Prompt-tuning	✓	✓	✗
Hawk [Tang et al., 2024]	Instr. tuning	Projection	✗	✗	✓
HolmesVAU [Zhang et al., 2024b]	Instr. tuning	LoRA	✓	✗	✓
VERA [Ye et al., 2025]	Weak class	Verbalized prompt learning	✓	✗	✗
Ours	**None**	**None**	✓	✓	✓

concerns, the training data could be unavailable for some sensitive scenes. As partial remedies, recent work has explored zero- and few-shot approaches using frozen vision-language backbones or MLLMs as we summarized in Table 1. We observed that most of the VLM-based works have limited task scope and still rely on annotated datasets. The only strictly zero-shot method is solely focusing on temporal VAD which makes it less user-friendly [Zanella et al., 2024]. For prompt-based methods [Yang et al., 2024, Ye et al., 2025, Wu et al., 2024b], they inevitably require induction on an annotated training set, which comes at the cost of generality as prompts are often learned to be task/domain-specific. This generality problem even exacerbates for instruct-tuned MLLMs [Tang et al., 2024, Zhang et al., 2024b] which are optimized to return answers from seen QA pairs focused on describing a closed set of anomaly types [Ding and Wang, 2024].

In recognition of these problems, given that multimodal LLMs already encode rich visual-semantic priors for commonsense reasoning [Zhao et al., 2023, Zhang et al., 2025b, Ren et al., 2025], fine-tuning may be unnecessary for certain tasks, as long as we can effectively reason about task contexts at test time [Minaee et al., 2025, Ma et al., 2024]. Specifically for video anomaly analysis, we may consider each of the previous benchmark tasks as answering specific questions *(When, where, what, and why?)* about visual anomalies, among which each can be seen as a sub-problem contributing to holistic analysis. Therefore, solving these tasks represents naturally stratified reasoning contexts contributing towards holistic anomaly analysis. Inspired by this, we propose a **unified reasoning-driven chain framework** that conditionally connects different MLLM-based task solvers during test time.

Specifically, our framework operates systematically across three clearly defined stages rather than merely concatenating separate tasks. First, an initial Video Anomaly Detection (VAD) computes a surrogate anomaly probability at the video level and extracts a contextual tag list corresponding to the most suspicious segments, thereby providing individualized context cues for each sample. Following this, a score-gated refinement utilizes both the contextual tag list and preliminary anomaly scores to perform conditional score adjustments, refining the VAD task based on the inferred contexts. Lastly, the final anomaly scores and contextual tag lists jointly guide the downstream spatial Video Anomaly Localization (VAL) and further textual Video Anomaly Understanding (VAU) tasks, where textual and visual prompts are dynamically refined based on the VAD scores. In summary, each stage of our framework employs frozen Vision-Language Models (VLMs), with dynamic prompts iteratively inferred from preceding stages.

We conduct extensive experiments on UCF-Crime, XD-Violence, UBnormal and MSAD [Sultani et al., 2018, Wu et al., 2020, Acsintoae et al., 2022, Zhu et al., 2024]. The proposed framework achieves state-of-the-art performance on three separate tasks under a zero-shot setting, achieving an overall 4-6% AUC improvement on VAD, and consistent improvements over diverse metrics for VAL and VAU tasks. These results show that our training-free, unified video anomaly analysis framework is interpretable, extensible, and robust across various domains and tasks.

2 Related Works

Traditional video anomaly analysis. Early Video Anomaly Detection (VAD) works typically fall into three major supervision regimes: *one-class* models trained only on normal clips and used compact embeddings or memory banks to detect outliers [Sohrab et al., 2018, Wang and Cherian, 2019, Micorek et al., 2024]; fully *unsupervised* methods rely on reconstruction or future-frame prediction losses [Hasan et al., 2016, Thakare et al., 2022]; and *weakly-supervised* MIL frameworks

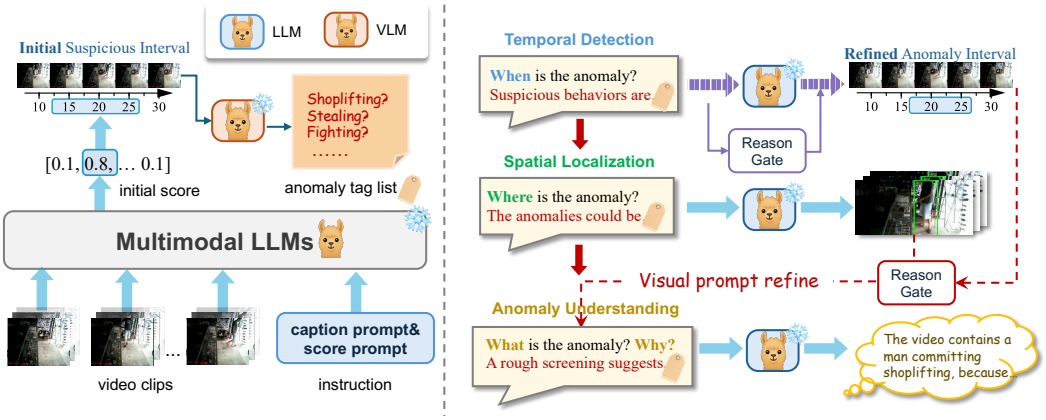


Figure 1: **Overview of the unified holistic anomaly analysis framework. Left:** A preliminary step extracting the most suspicious intervals of a video and extracts anomaly tag lists reflecting possible anomaly contexts. **Right:** Illustration of how the priors are used to refine each of the tasks. Low-confidence samples in Temporal VAD are refined by a selective Intra-Task Reasoning step. The Inter-Task Chaining further connects it to downstream, including spatial VAL and textual VAU into a cascaded chain for a unified holistic anomaly analysis.

used video-level tags to rank anomalous snippets [Sultani et al., 2018, Feng et al., 2021, Joo et al., 2023]. All of them need to be re-trained for unseen domains or anomaly types and provide no semantic rationale for their decisions [Ramachandra et al., 2020, Wu et al., 2024b]. To address this, *open-set* detectors emerged: OVVAD fuses LLM semantics so the system can both *detect* and *classify* novel anomalies [Wu et al., 2024a]. However, such open-set detectors still require task-specific training and provide very limited textual insight into *why* frames may be abnormal, motivating the move toward vision-language solutions with task formulations beyond temporal detection.

VLM-based video anomaly analysis. LAVAD [Zanella et al., 2024] introduces a fully *training-free* pipeline for temporal detection: a frozen VLM captions each frame; a prompted LLM converts the caption stream into frame-wise anomaly scores that are further refined by ensembles of foundation models. While effective when anomalies are clearly distinguishable from normality, it occasionally fails to distinguish more complex anomaly types [Ding and Wang, 2024], and lacks direct semantic explanations, providing only default VLM captions alongside computed anomaly scores. Prompt-tuning variants [Du et al., 2024, Wu et al., 2024b, Yang et al., 2024, Ye et al., 2025] optimize textual prompts to guide frozen MLLMs for certain tasks. While they reveal strong performance, they remain dependent on annotated data and deal with limited task scopes [Zhang et al., 2024b].

Video anomaly understanding and multimodal LLMs. With the need for deeper semantic reasoning, instruction-tuning methods such as Hawk [Tang et al., 2024] and Holmes-VAU [Zhang et al., 2024b] fine-tune VLMs on detailed, anomaly-captioned video clips to produce narrative explanations. These works have achieved more accurate descriptions but require extensive annotation and computational resources, and remain tied to seen anomaly types [Liu et al., 2025].

To sum up, we observe: strictly zero-shot methods such as Zanella et al. [2024] support temporal detection but lack spatial grounding and textual insights. Prompt-tuning variants [Du et al., 2024, Wu et al., 2024b, Ye et al., 2025] are mostly focused on only a subset of tasks/domains as the prompts are often task/domain-specific. Instruction-tuned models [Tang et al., 2024, Zhang et al., 2024b] produce rich narrative explanations, yet lack either temporal or spatial coverage and incur high annotation costs. These gaps motivate our effort to unify these tasks under a zero-shot setting.

3 Methodology

We show an overview of this framework in Figure 1. The video anomaly analysis task is decomposed into three major sub-tasks, as formulated in previous works, and our framework exploits the inherent connection among them. Our unified framework can be summarized in two major components: 1)

An Intra-Task Reasoning (IntraTR) extracts anomaly priors through the temporal video anomaly detection (VAD) task and then refines the temporal detection through a gated additional reasoning step. 2) Building on the reasoning process in IntraTR, an additional Inter-Task Chaining (InterTC) connects the extracted tag list and temporal score results from the initial VAD results to enable subsequent localization and understanding tasks in a cascaded manner. Detailed explanations for each component are provided in Section 3.1 and Section 3.2 respectively.

3.1 Intra-Task Reasoning (IntraTR) for temporal anomaly detection

Problem formulation. VAD can be formulated as a binary (0-1) classification at frame level. Ideally, for each input frame f_i, the objective is to predict an anomaly probability s_i. For baseline methods utilizing LLM and VLM [Zanella et al., 2024], it can be formulated as:

$$s_i = \theta_{\text{LLM}}(p_{\text{VAD}} \oplus \theta_{\text{VLM}}(c_i, p_{\text{caption}})), \qquad S_V = [s_1, \ldots, s_T], \tag{1}$$

where T is the number of frames in video V, c_i is a short video clip representing events around frame f_i and $p_{\text{VAD}}, p_{\text{caption}}$ represents prompts used respectively for video anomaly detection and clip captioning. Vector S_V therefore provides a *first-pass* anomaly estimate for every frame, obtained without fine-tuning. *However, beyond this baseline, can we further leverage S_V for improved reasoning?*

Trying to answer this question, our VAD pipeline treats S_V not only as the final answer but also as a starting point for a structured intra-task reasoning step performed at test time. Figure 2 provided an overview of the proposed IntraTR pipeline.

Score-guided anomaly extraction. To identify the potential anomalies present in the video, we first conduct one forward pass producing frame-wise anomaly scores $S_V = [s_1, \ldots, s_T]$ for a video V with T frames, where each $s_i \in [0, 1]$. Intuitively, an anomalous event e should occupy a contiguous window $W_e = \{t, \ldots, t+\ell-1\}, \ell \ll |S_V|$ reflects a local segments. Denote the mean score inside any window W by $\mu(W) = \frac{1}{|W|}\sum_{j \in W} s_j$. Following the intuition that anomaly events should maintain consistently high scores, in an anomalous video, we expect to find:

$$\exists\, W_e : \ |W_e| = \ell, \ \text{such that } \mu(W_e) \geq \tau, \tag{2}$$

where τ is a natural decision boundary (e.g. $\tau = 0.5$).

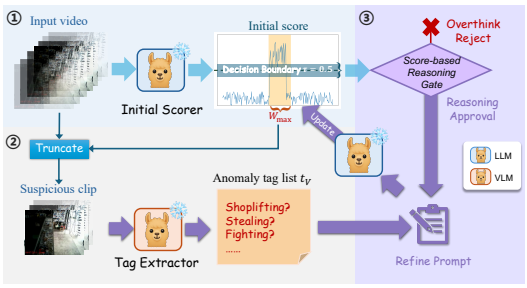


Figure 2: **Intra-Task Reasoning pipeline:** (1) the Initial Scorer produces a score curve; (2) peak detection truncates a suspicious window and the Tag Extractor generates anomaly tags t_V; (3) a reasoning gate refines low-confidence predictions via the Score Updater.

To find whether such a window W_e exists in the video, at inference time, we slide a window of admissible length ℓ and select:

$$W_{\max} = \arg \max_{W \subseteq \{1,\ldots,T\},\ |W|=\ell} \mu(W), \tag{3}$$

$$\tilde{s}_V = \mu(W_{\max}), \tag{4}$$

where $W_{\max}$ is the most suspicious segment and $\tilde{s}_V \in [0, 1]$ is the surrogate video-level anomaly probability. After identifying the most suspicious part of the video V_{sus} indicated by $W_{\max}$, we extract text contexts related to anomalies by querying VLM to generate a list of concise phrases t_V summarizing the possibly related anomaly activities in the video clip V_{sus} as follows:

$$V_{\text{sus}} = [f_j], \qquad j \in W_{\max}, \tag{5}$$

$$t_V = \theta_{\text{VLM}}(V_{\text{sus}}, p_{\text{extract}}). \tag{6}$$

We then pass $W_{\max}$, $\tilde{s}_V$ and t_V to later stages for further processing.

Score-based reasoning gate. Recent studies reveal a non-monotonic trade-off between reasoning depth and accuracy in large language models: while a short chain of thought can boost performance, excessive steps often induce "over-thinking" and hallucinations [Huang and Chang, 2023, Chen et al., 2025]. Inspired by this observation, we trigger an additional reasoning pass *only* when the first-pass score is ambiguous via a score-based gate component with motivation explained below.

Starting from the raw frame scores S_V, we obtain the surrogate video-level probability $\tilde{s}_V$. If $\tilde{s}_V \notin [0.5-m,\, 0.5+m]$, the model is considered confident about its first round predictions as the prediction is positioned far from the decision boundary [El-Yaniv and Wiener, 2010]. Therefore, a gating mechanism with width $2m$ allows borderline/ambiguous videos with $\tilde{s}_V \in [0.5 \pm m]$ to proceed to a second reasoning stage. With the tag list t_V extracted from frames in $W_{\max}$, the task prompt is refined to $p^*_{\text{VAD}} = t_V \oplus p_{\text{VAD}}$.

Intuitively, m quantifies the degree of *"suspicion"*, which can be either a fixed value or adaptive w.r.t. each sample. For the setting of m specifically, we offer two options. It could be either 1) a fixed heuristic constant over all samples that allowing user to control the degree of suspicion, 2) or as an adaptive sample-specific variable estimated from current V by $\tilde{m}_V = \mathrm{Var}(\mathrm{S_V})$ reflecting the diversion of normal/abnormal frame scores may exist in current video. We compare and discuss the impact of m in Section 4.2 and Appendix B.1 correspondingly.

Based on the above, querying the scorer LLM once more with refined prompts when $\tilde{s}_V \in [0.5 \pm m]$:

$$S^*_V = \theta_{\text{LLM}}\big(p^*_{\text{VAD}} \;\oplus\; \theta_{\text{VLM}}(c_i, p_{\text{caption}})\big), \qquad i = 1, \ldots, T. \tag{7}$$

The refinement yields updated frame scores S^*_V, replacing the initial S_V for the final decision. Following established practices in prior works [Ye et al., 2025, Tran et al., 2022], we run a standard gaussian smoothing to post-process the refined S_V, resulting in the final S^{pred}_V. By allocating the costly reasoning step only when the score near the margin indicates uncertainty, the method inherits the computational efficiency and robustness of selective prediction while mitigating "over-thinking" hallucinations observed in unrestricted chain-of-thought generation.

Beyond the IntraTR-assisted VAD above, we further explore leveraging the reasoning steps from the VAD task to assist downstream tasks through InterTC component in Section 3.2 and Section 3.2.

3.2 Inter-Task Chaining (InterTC) for holistic anomaly analysis

In this section, we cover the design of InterTC for two key sub-tasks in anomaly analysis, namely 1) spatial Video Anomaly Localization (VAL) and 2) textual Video Anomaly Understanding (VAU).

InterTC from temporal detection to spatial localization. Video Anomaly Localization (VAL) aims to predict spatial bounding boxes for regions in the frame f containing the anomalous activities. The InterTC connects VAD with VAU using a straightforward method. Specifically, we utilized a frozen VLM $\theta_{\text{LOC}}(p_{\text{LOC}} \oplus f)$, guided by a base localization task prompt p_{LOC} for frame f. And then inject t_V to the p_{LOC}, producing a refined prompt p^*_{LOC} using a predefined template. Therefore, p^*_{LOC} is expected to be a more sample-specific and clearer guiding prompt for spatial localization and thereby improving its performance. Detailed prompt templates are included in Appendix C.1.

Algorithm 1: Inter-Task Chaining prompt refinement for VAU

Input: video $V = [f_1, \ldots, f_T]$;
tag list t_V;
base prompt p_{VAU};
localization prompt p_{LOC};
surrogate anomaly score $\tilde{s}_V$;
most suspicious window $W_{\max}$
Output: final description d^*

VAD-prior Prompt Refinement:
$p^*_{\text{VAU}} \leftarrow t_V \oplus p_{\text{VAU}}$;
Score-gated Localization Overlay (optional):
if $\tilde{s}_V > 0.5$ **then**
$F_{\text{sel}} \leftarrow \text{sample\_frames}(V, W_{\max})$;
bboxes $\leftarrow \theta_{\text{LOC}}\big(F_{\text{sel}},\, t_V \oplus p_{\text{LOC}}\big)$;
$V_{\text{query}} \leftarrow \text{draw\_boxes}(V, \textit{bboxes})$;
else $V_{\text{query}} \leftarrow V$;
Final description:
$d^* \leftarrow \theta_{\text{VLM}}\big(V_{\text{query}},\, p^*_{\text{VAU}}\big)$;
return d^*

Cascaded InterTC for video anomaly understanding. Given an untrimmed surveillance video $V = (f_1, \ldots, f_T)$, video-level anomaly understanding (VAU) aims to 1) decide whether V containing an abnormal event and 2) output a human-readable description d^* that explains anoma-

lies from the visual inputs. Formally,

$$\Theta_{\text{VAU}} : V \longrightarrow (\hat{y}_V,\, d^*), \qquad \hat{y}_V \in \{0, 1\}. \tag{8}$$

Unlike earlier works that train task-specific models via instruction tuning [Tang et al., 2024, Zhang et al., 2024b], our approach to Θ_{VAU} operates in a fully *zero-shot* manner. It reuses the frame-level scores S_V, the tag list t_V, and the suspicious window $W_{\max}$ obtained during the earlier temporal detection and spatial localization steps to refine the anomaly understanding prompt at inference time.

Algorithm 1 provides an overview of the full *prompt refinement* step for downstream VAU task leveraging the reasoning steps from the preceding VAD and VAL tasks. Specifically, we begin with *VAD-prior Prompt Refinement* which incorporates the tag list t_V from the VAD task into the anomaly description prompts, forming a more context-aware textual query $p^*_{\text{VAU}} = t_V \oplus p_{\text{VAU}}$.

Next, we apply a visual prompt enhancement called *Score-gated Localization Overlay*. Specifically, the surrogate probability $\tilde{s}_V$ *gates* a visual-prompt enhancement stage: only when $\tilde{s}_V > 0.5$. i.e. the VAD detector already believes an anomaly is present, allowing us to trust that object-level cues are meaningful and beneficial to include. For such videos we 1) sample frames inside $W_{\max}$. 2) invoke a detection-capable VLM with $t_V \oplus p_{\text{LOC}}$ to obtain bounding boxes, and 3) overlay those boxes onto the corresponding frames in original video V, producing an annotated V_{query}. If $\tilde{s}_V \leq 0.5$ we skip the bounding box overlay and retain the original, unmodified video.

Finally, the VLM receives V_{query} (either annotated or not) together with p^*_{VAU} and outputs the description d^*. Since localization is performed only when the detector is confident that an anomaly exists, the inserted boxes act as reliable visual prompts rather than noisy clutters.

4 Experiments

4.1 Experimental setup

Datasets & evaluation metrics. We evaluate on the official test splits of three benchmarks: 1) UCF-Crime [Sultani et al., 2018] (real-world CCTV and crowd-sourced, 13 anomaly types); 2) XD-Violence [Wu et al., 2020] (800 test videos from movies, sports clips, CCTV, dashcam, cartoons); 3) UBnormal [Acsintoae et al., 2022] (211 fully synthetic surveillance videos across 29 virtual environments); 4) a more recent MSAD [Zhu et al., 2024] (14 distinct scenarios captured from various camera views, containing 360 test videos) which is less likely to overlap with pre-train data.

According to previous works, we primarily evaluate Area Under the Curve (AUC) score for the Receiver Operating Characteristic (ROC) Curve on all the datasets. Since several studies also report Average Precision (AP) on XD-Violence [Wu et al., 2020], we include AP results for reference.

Finally, for *Video Anomaly Understanding (VAU)* task, to fairly evaluate the quality of the generated d^*, we adopted all the video-level annotations from HIVAU-70k [Zhang et al., 2024b]. Spanning 1051 video descriptions, with 251 test videos from UCF-Crime, and 800 videos from XD-Violence, which is larger than the original video-level test set in Zhang et al. [2024b] (398 samples). In addition to traditional NLP metrics [Papineni et al., 2002, Vedantam et al., 2015, Banerjee and Lavie, 2005, Lin, 2004], we also evaluate GPT-guided scores following recent works [Tang et al., 2024, Li et al., 2024a]. More details are available in Appendix C.

Hyperparameters & experiment details. For VAD tasks, clip-level scoring operates on the full video with a 16-frame stride (see details in Appendix C.2). The suspicious window size for the prior extraction step is set to $\ell = \max(300, T/10)$ and fixed $m = 0.05$. We evenly subsample at most 180 frames from the window $W_{\max}$ due to the limited context capacity of the VLM model to get t_V. As for the default VLM and LLM tested in the framework, we choose `VideoLLaMA3-7B` [Zhang et al., 2025a] and `Llama-3.1-8B-Instruct` [Grattafiori et al., 2024]. To reduce computational cost, we subsample all videos at a frame sampling stride of 16. We run all experiments on two NVIDIA GeForce RTX 3090 GPUs. Further implementation details, prompts and hyperparameter stability tests are provided in Appendix C and Appendix B.1.

Additionally, we adopted `Qwen2.5-VL-7B` [Bai et al., 2025] as the default localization VLM for VAL task, and varied different baseline VLMs including Zhang et al. [2025a], Li et al. [2024b], Bai et al.

Table 2: **Performance comparison across UCF-Crime, XD-Violence and UBNormal.** ✓ / ✗ indicate whether a method is *zero-shot* and *training-free* in terms of model parameters.

Method	Zero-shot	Training-free	UCF-Crime	XD-Violence		UBNormal	MSAD	
			AUC(%)	AUC(%)	AP(%)	AUC(%)	AUC(%)	AP(%)
Sultani et al. [2018]	✗	✗	77.92	-	73.20	50.30	-	-
GODS [Wang and Cherian, 2019]	✗	✗	70.46	61.56	-	-	-	-
RTFM [Tian et al., 2021]	✗	✗	83.31	-	77.81	60.94	86.7	**66.3**
AccI-VAD [Reiss and Hoshen, 2022]	✗	✗	-	-	-	66.51	-	-
CLIP-TSA [Joo et al., 2023]	✗	✗	87.58	-	82.19	-	-	-
MGFN [Chen et al., 2023b]	✗	✗	86.98	-	80.11	-	85.0	63.5
STPrompt [Wu et al., 2024b]	✗	✗	88.08	-	-	63.98	-	-
OVVAD [Wu et al., 2024a]	✗	✗	86.40	-	66.53	62.94	-	-
Holmes-VAU [Zhang et al., 2024a]	✗	✗	**88.96**	-	**87.68**	-	-	-
MULDE [Micorek et al., 2024]	✗	✗	78.50	-	-	**72.80**	-	-
EGO [Ding et al., 2024]	✗	✗	81.71	-	65.77	-	**87.3**	64.4
AnomalyRuler [Yang et al., 2024]	✗	✓	-	-	-	71.90	-	-
VERA [Ye et al., 2025]	✗	✓	86.55	**88.26**	70.54	-	-	-
HolmesVAU [Zhang et al., 2024b] (ZS)	✓	✗	-	-	-	58.54†	-	-
AnomalyRuler [Yang et al., 2024] (ZS)	✓	✓	-	-	-	65.40†	-	-
UR-DMU [Zhou et al., 2023] (ZS)	✓	✓	-	-	-	-	74.3	53.4
CLIP [Radford et al., 2021] (ZS)	✓	✓	53.16	38.21	17.83	-	-	-
LLAVA-1.5 [Liu et al., 2024] (ZS)	✓	✓	72.84	79.62	50.26	-	-	-
VideoLLaMA3-7B + Llama3.1-8B (ZS)	✓	✓	-	-	-	-	78.7	68.5
GLM-4.1V-9B-Thinking (ZS CoT)‡	✓	✓	61.80	72.73	52.93	60.81	-	-
LAVAD [Zanella et al., 2024]	✓	✓	80.28	85.36	62.01	51.06	-	-
Ours (fixed constant m)	✓	✓	**84.28**	**91.34**	**68.07**	68.98	85.9	**76.4**
Ours (adaptive $\tilde{m}_V$)	✓	✓	84.08	91.23	68.03	**69.02**	**86.0**	75.9

† The result is from a direct evaluation of the method trained on other non-overlapping datasets, reflecting its zero-shot performance.
‡ Zero-shot chain-of-thought (CoT) inference VAD performance using GLM-4.1V-9B-Thinking [Team et al., 2025].

Table 3: **(a)** Ablation of inference steps. showing the effectiveness of each reasoning component. **(b)** Ablation on video-level anomaly priors. t_{oracle} uses ground-truth types, and t_V are actual local anomaly priors we extracted during the reasoning step.

(a) Inference component effectiveness ablations

① LLM-Scoring	② Prior Reasoning	③ Score-gated Reasoning	AUC (%)
✗	✗	✗	77.67 (+0.00)
✓	✓	✗	77.40 (-0.27)
✓	✗	✗	80.38 (+2.71)
✓	✓	✓	**84.28** (+6.61)

(b) Reasoning prior ablations

Anomaly Priors	AUC (%)
p_{VAD}	81.86
$\oplus\, t_{\text{oracle}}$	83.91
$\oplus\, t_V$ **(Ours)**	**84.28**

[2025] for VAU task. Moreover, for both VAL and VAU tasks, we leverage the anomaly priors (e.g. $t_V, \tilde{s}_V, W_{\max}$) obtained under the default configuration of IntraTR in Section 3.1.

For the *Video Anomaly Localization (VAL) task*, following previous works, we evaluate *temporal IoU (TIoU)* [Liu and Ma, 2019, Wu et al., 2024b] for each anomalous frame f_j $(j = 1, \ldots, N)$, the localization head $\theta_{\text{LOC}}(f_j, p^*_{\text{LOC}})$ outputs a confidence C_j and a box B_j. Then, the TIoU is computed as: $\frac{1}{N}\sum_{j=1}^{N} \frac{\text{Area}(B_j \cap G_j)}{\text{Area}(B_j \cup G_j)} \mathbb{I}[C_j \geq \tau]$, with G_j the ground-truth bboxes, where the indicator $\mathbb{I} \in \{0, 1\}$ judges whether the confidence C_j is above the default threshold $\tau = 0.5$.

4.2 VAD results

Table 2 summarizes the results across the three benchmarks. Across all datasets, our *zero-shot, training-free* framework outperforms the previous best zero-shot detectors by $4-6\%$ on UCF-Crime and XD-Violence, 3% on UBnormal, and also generalises well on MSAD. Our method also showed competitive performance to those baselines requiring additional supervision, data or CoT reasoning steps, further proving the benefit of IntraTR component we proposed. Figure 3 qualitatively compares our method with the baseline [Zanella et al., 2024] showing significantly reduced false positive predictions. More examples are provided in Appendix D.1.

We also find that a fixed margin m already performs well, although it introduces an unavoidable assumption on the test domain, while variance estimated $\tilde{m}_V$ also provides similar performances without posing any assumption on the test domain. This sample-specific "suspicion" accounts for its superiority on a synthetic dataset (UBNormal) where samples are peculiar to natural videos. We further discuss the impact of m values in Appendix B.1.

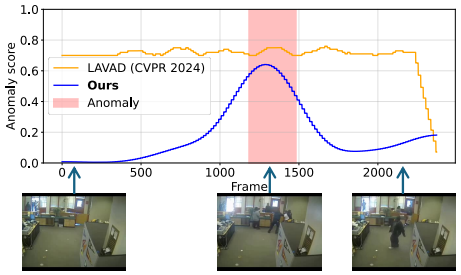


Figure 3: Anomaly scores on a video from UCF-Crime with an "Arrest" incident.

Table 4: **Comparison with previous supervised works on temporal-IoU (%) metric using zero-shot** `Qwen2.5-VL-7B`**.** t_V comes from IntraTR, t_{oracle} from ground-truth class names.

Method	TIoU
VadCLIP [Wu et al., 2024c]	22.05
STPrompt [Wu et al., 2024b]	23.90
Qwen2.5-VL-7B (baseline)	24.09
$\oplus\ t_V$	25.17
$\oplus\ t_{\text{oracle}}$	**25.21**

Table 5: **Video anomaly understanding performance comparison on two benchmark datasets.** The results are computed against ground-truth descriptions provided by [Zhang et al., 2024b]. Apart from the traditional NLP metrics (**BLEU, CIDEr, ROUGE, METEOR**), we also provide **GPT-R, GPT-D, GPT-C** metrics `Reasonability, Detail and Consistency` computed against against the ground-truth using API calls to OpenAI-GPT4.1 [OpenAI, 2025] correspondingly following previous works [Tang et al., 2024, Li et al., 2024a].

Method	UCF-Crime [Sultani et al., 2018]							XD-Violence [Wu et al., 2020]						
	BLEU	CIDEr	METEOR	ROUGE	GPT-R	GPT-D	GPT-C	BLEU	CIDEr	METEOR	ROUGE	GPT-R	GPT-D	GPT-C
InternVideo2.5-8B [Wang et al., 2025]	0.159	0.011	0.088	0.103	0.240	0.266	0.205	0.209	0.013	0.119	0.130	0.456	0.447	0.433
VideoChat-Flash-2B [Li et al., 2024b]	0.165	0.008	0.108	0.168	0.488	0.283	0.404	0.277	0.026	0.144	0.186	0.690	0.576	0.627
+ InterTC VAU refine (Ours)	0.297	0.022	0.157	0.188	0.509	0.427	0.438	0.324	0.033	0.158	0.187	0.715	0.649	0.655
VideoLLaMA3-7B [Zhang et al., 2025a]	0.215	0.014	0.117	0.156	0.463	0.289	0.384	0.290	0.022	0.141	0.169	0.568	0.487	0.499
+ InterTC VAU refine (Ours)	0.345	**0.023**	0.175	0.188	**0.512**	0.428	**0.444**	**0.399**	**0.029**	**0.198**	0.200	**0.721**	**0.707**	0.668
Hawk [Tang et al., 2024] †	0.379	0.008	**0.217**	0.187	0.255	**0.580**	0.214	0.375	0.016	0.176	0.188	0.408	0.586	0.365
HolmesVAU [Zhang et al., 2024b] †	**0.435**	0.021	0.194	**0.257**	0.448	0.356	0.391	0.376	0.011	0.182	**0.253**	0.715	0.581	**0.673**

† Re-evaluated on our new evaluation set strictly following its default configurations.

Ablation on test-time reasoning steps. Table 3a evaluates the individual contributions of the three components of our inference loop. The simplest baseline, single-round direct query to a frozen VLM achieves 77.67% `(row 1)`. Introducing the ① LLM-based *Scoring* component and the ② *Prior-Reasoning* step without the subsequent score-gated reasoning yields only 77.40% `(row 2)`. In contrast, keeping the LLM scorer but dropping the prior reasoning module lifts performance to 80.38% `(row 3)`, indicating that unrestricted "overthinking" across all samples without selective gating can conversely inject noise, causing hallucination, degrading performance. Activating all three stages, including the ③ score-gated reasoning, further raises the result to 84.28% `(row 4)`, a gain of 6.61% over the raw VLM baseline. These results validate our hypothesis that confidence in anomaly presence can act as a metric to evaluate the quality of first-round prediction and therefore effectively control a proper reasoning depth for test samples.

Ablation on t_V. Table 3b isolates the effects of incorporating the textual video-level anomaly priors in the second-round reasoning for VAD. The baseline score-gated reasoning module under fixed small margin value $m = 0.05$ with an empty t_V achieves a lower performance of 81.86%. Replacing the t_V with ground-truth oracle class names from annotations (e.g. `''Arson, RoadAccident''`) (t_{oracle}) lifts performance to 83.91%, confirming that accurate anomaly priors improve detection performance. Interestingly, our automatically extracted priors t_V even surpassed the oracle class names, reaching 84.28%, demonstrating that the local anomaly extraction step could effectively finalize the anomaly priors to clearer contexts than rough anomaly classes (e.g. class label "Arrest" is ambiguous, while extracted t_V may include "physical altercation" which is more informative) in ground-truth. Exploiting clearer contexts leads to superior frame-wise anomaly detection performance.

4.3 VAL results

Table 4 shows that the zero-shot MLLM baseline already outperforms earlier supervised detectors, and that injecting anomaly tags, either from the automatically derived tag list t_V in IntraTR or the ground-truth class name, yields an additional $\sim 1\%$ absolute gain in quantitative TIoU metric. Also, the tiny gap between the results using t_V and the ground-truth t_{oracle} suggests our t_V captures near-optimal semantic cues the oracle provides, yet without requiring any manual annotation. These

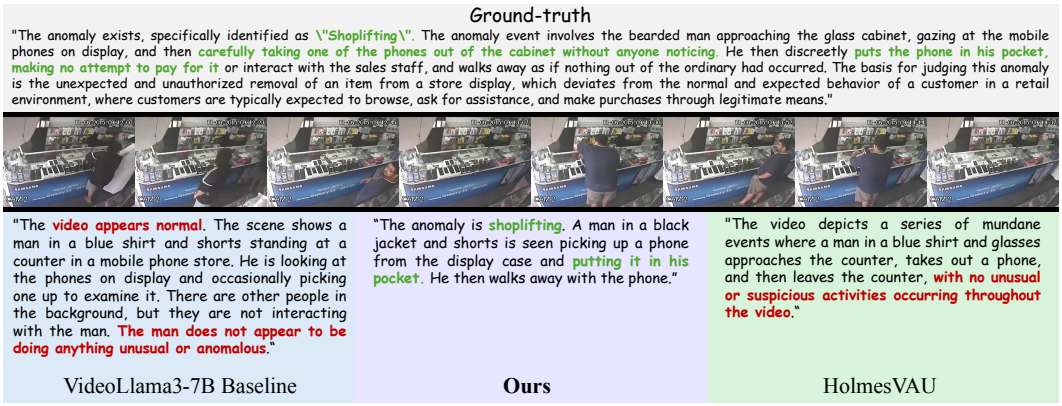


Figure 4: **Qualitative results of video anomaly understanding.** Descriptions for a video containing an incident of ''Shoplifting'' from different methods, where **green text** highlights correct descriptions/rationale about the anomaly, and **red** highlights statements inconsistent with the ground truth.

Table 6: Ablation study of InterTC prompt refinement steps on description quality.

Dataset	Method	Tag t_V	*bboxes*	BLEU	CIDEr	METEOR	ROUGE
UCF-Crime	ZS CoT Baseline†	–	–	0.3172	0.0193	0.1651	0.1820
	InterTC (Ablated)	✗	✗	0.2147	0.0143	0.1167	0.1564
	InterTC (Ablated)	✓	✗	0.3328	0.0183	0.1684	**0.1920**
	InterTC (Full)	✓	✓	**0.3453**	**0.0232**	**0.1750**	0.1878
XD-Violence	ZS CoT Baseline†	–	–	0.3682	**0.0381**	0.1824	0.1876
	InterTC (Ablated)	✗	✗	0.2897	0.0219	0.1410	0.1690
	InterTC (Ablated)	✓	✗	0.3857	0.0270	0.1931	0.1993
	InterTC (Full)	✓	✓	**0.3993**	0.0288	**0.1980**	**0.1997**

† **ZS CoT**: The zero-shot VAU performance of a reasoning VLM: GLM-4.1V-9B-Thinking [Team et al., 2025], which is capable of long chain-of-thought (CoT) inference.

observations confirm that even lightweight semantic priors effectively improve spatial localization without retraining. Additional qualitative examples of localization are included in Appendix D.

4.4 VAU results

Experiment results. Table 5 compares our InterTC refinement to direct VLM inference baselines and recent instructed-tuned VAU MLLMs [Tang et al., 2024, Zhang et al., 2024b] on two different test domains, against the ground-truth description provided by HIVAU-70k [Zhang et al., 2024b]. On both domains, InterTC-refined query prompts improve the base VLM on both traditional NLP metrics and all GPT-scores (Reasonability, Detail, Consistency) of the outputs, narrowing much of the gap to instruction-tuned methods [Tang et al., 2024, Zhang et al., 2024b] and even surpassing instruct-tuned methods on several metrics. Qualitatively, we also demonstrate the descriptive capability of our framework in Figure 4. On a *shoplifting* clip, the baseline VLM [Zhang et al., 2025a] and HolmesVAU both fail to identify the abnormal act, whereas our method reports the key action ("*puts the phone in his pocket*") and labels the event as *shoplifting*. More examples are provided in Appendix D.3. These findings confirm that 1) the tag-based prompt enrichment injects crucial context and 2) localization cues further enhance narrative detail without any additional training.

Ablation to prompt refinement steps. To isolate the improvement of VAU metrics to each component of InterTC-assisted VAU process, we conduct corresponding ablations. As shown in Table 6, across both UCF-Crime and XD-Violence, simply enhancing the prompt with the tag list t_V from VAD-priors to the base prompt accounts for the majority of the observed gains. In contrast, Inter-task chaining from the spatial localization overlay to VAU step yields a smaller, incremental lift on top of that strong improvement. We suspect primarily because the frozen VLMs have not been fine-tuned on large-scale data featuring overlaid bounding boxes, resulting in a rather marginal improvements.

While the generic "thinking" VLM [Team et al., 2025] underperforms on the more specialized VAD task (see Table 2), it performs better on VAU than zero-shot baselines. This indicates that chained inference idea adopted in both Team et al. [2025] and our InterTC can enrich textual anomaly understanding by encouraging more detailed, stepwise descriptions. However, general-purpose reasoning of Team et al. [2025] may not generalise well on the niche and complex anomaly video understanding task [Shojaee* et al., 2025], introducing content weakly related to the true anomalies. In contrast, our InterTC-guided prompts focus the description on anomaly-relevant evidence, yielding superior scores on most metrics across all video-anomaly tasks. Overall, VAD prior textual prompt refinement plays a more major role in prompt refinement, while localization visual prompts could be an optional enhancement when extra compute is available.

5 Conclusion

In this work, we introduced a unified, training-free framework for holistic video anomaly analysis by chaining temporal detection, spatial localization, and textual understanding in a single inference pass. Our zero-shot system consistently outperforms prior training-free baselines and approaches supervised methods across all three sub-tasks.

By structuring our pipeline as a sequence of gated reasoning steps, each sub-task enriches the next with semantic or visual priors drawn from the model's own outputs, enabling self-correction and deeper interpretability without any additional training. In video anomaly analysis specifically, where events unfold over time and space, such multi-stage inference captures structure that single-pass models miss or fail, yielding both accurate detection and user-friendly explanations without any additional training. Despite some possible limitations and potential societal impacts it may bring about a powerful yet bulky VLM system for sensitive video analysis (see more discussion in Appendix F and Appendix G), we believe this framework of treating inference as an active, context-driven process can foster more robust video analytics and may generalize to other complex vision-language tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.92470203), Beijing Natural Science Foundation (No.L242022), the Fundamental Research Funds for the Central Universities (2024XKRC082). Jianbo Jiao is supported by an Amazon Research Award.

References

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `https://aclanthology.org/W05-0909/`.

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36:70115–70140, 2023a.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu.

Do not think that much for 2+3=? on the overthinking of o1-like llms, 2025. URL https://arxiv.org/abs/2412.21187.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 387–395, 2023b.

Dexuan Ding, Lei Wang, Liyun Zhu, Tom Gedeon, and Piotr Koniusz. Learnable expansion of graph operators for multi-modal feature fusion. *arXiv preprint arXiv:2410.01506*, 2024.

Xi Ding and Lei Wang. Quo vadis, anomaly detection? llms and vlms in the spotlight, 2024. URL https://arxiv.org/abs/2412.18298.

Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18793–18803, 2024. doi: 10.1109/CVPR52733.2024.01778.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023. URL https://arxiv.org/abs/2305.05665.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey, 2023. URL https://arxiv.org/abs/2212.10403.

Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE, 2023.

Teng Li, Jiapeng Wang, and Lianwen Jin. Enhancing visual information extraction with large language models through layout-aware instruction tuning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 276–289. Springer, 2024a.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024b.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1490–1499, 2019.

Zihao Liu, Xiaoyu Wu, Jianqin Wu, Xuxu Wang, and Linlin Yang. Language-guided open-world video anomaly detection, 2025. URL https://arxiv.org/abs/2503.13160.

Miodrag Lovrić, Marina Milanović, and Milan Stamenković. Algoritmic methods for segmentation of time series: An overview. *Journal of Contemporary Economic and Business Issues*, 1(1):31–53, 2014.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. 2013.

Fan Ma, Xiaojie Jin, Heng Wang, Jingjia Huang, Linchao Zhu, and Yi Yang. Stitching segments and sentences towards generalization in video-text pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4080–4088, 2024.

Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010. doi: 10.1109/CVPR.2010.5539872.

Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Koziński. MULDE: Multiscale Log-Density Estimation via Denoising Score Matching for Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18868–18877, June 2024.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2025. URL https://arxiv.org/abs/2402.06196.

OpenAI. Gpt-4.1 api. https://platform.openai.com/docs/models/gpt-4.1, 2025. Accessed: 2025-05-01.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2293–2312, 2020.

Tal Reiss and Yedid Hoshen. An attribute-based method for video anomaly detection. *Transactions on Machine Learning Research*, 2022.

Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29029–29039, 2025.

Parshin Shojaee*, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf.

Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *2018 24th International Conference on Pattern Recognition (ICPR)*, page 722–727. IEEE, August 2018. doi: 10.1109/icpr.2018.8545819. URL http://dx.doi.org/10.1109/ICPR.2018.8545819.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models, 2025. URL https://arxiv.org/abs/2503.16419.

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

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. In *Neural Information Processing Systems (NeurIPS)*, 2024.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Kamalakar Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network, 2022. URL https://arxiv.org/abs/2211.00882.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021.

Tung Minh Tran, Tu N Vu, Nguyen D Vo, Tam V Nguyen, and Khang Nguyen. Anomaly analysis in images and videos: A comprehensive review. *ACM Computing Surveys*, 55(7):1–37, 2022.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. URL https://arxiv.org/abs/1411.5726.

Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211, 2019.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

Jinta Weng, Yue Hu, Jing Qiu, and Heyan Huan. Stprompt: Semantic-guided and task-driven prompts for effective few-shot classification, 2022. URL https://arxiv.org/abs/2210.16489.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts, 2024b. URL https://arxiv.org/abs/2408.05905.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6074–6082, 2024c.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models, 2025. URL `https://arxiv.org/abs/2412.01095`.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025a. URL `https://arxiv.org/abs/2501.13106`.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. *arXiv preprint arXiv:2506.23825*, 2025b.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*, 2024a.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. *arXiv preprint arXiv:2412.06171*, 2024b.

Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning, 2023. URL `https://arxiv.org/abs/2305.14078`.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3769–3777, 2023.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Technical Appendices and Supplementary Material

A Appendix Roadmap

In this appendix, we cover the following materials:

- Additional ablation study (Appendix B)
- Additional implementation details (Appendix C)
- Additional qualitative results (Appendix D)
- Running-time Analysis (Appendix E)
- Limitations (Appendix F)
- Broader impacts (Appendix G)

B Additional ablation study

B.1 Hyperparameter sensitivity tests

Senstitivity on m We study performances under different decision-boundary-margin width values $m \in (0, 0.5)$ and dynamic $\tilde{m}_V = \mathrm{Var}(S_V)$ presented in Table 7. Performance remains stable for $m \leq 0.2$ and $\tilde{m}_V$ and drops significantly on UCF-Crime and XD-Violence when $m = 0.4$, presumably because an overly wide margin labels many true positives as "uncertain", resulting in unnecessary hallucinations. In contrast, UBnormal [Acsintoae et al., 2022] benefits from larger m; the synthetic clips are originally ambiguous for pretrained models such that additional skepticism is beneficial [Yang et al., 2024]. As $m \in [0.05, 0.2]$ yields near-optimal AUC on all real-world datasets, we adopt the smallest value $m = 0.05$ as the default setting for constant m.

To further investigate how the IntraTR step affect model behaviours, we further visualize the density of samples with respect to the l1 distance of their video-level scores to the decision boundary $|\tilde{S}_V - \tau|$ that measures the confidence of predictions in Figure 5. Specifically, we observe that both smaller constant m and the dynamic $\tilde{m}_V$ can effectively produce more high-confidence predictions, while a larger m conversely results in more confusion and therefore less confident predictions overall.

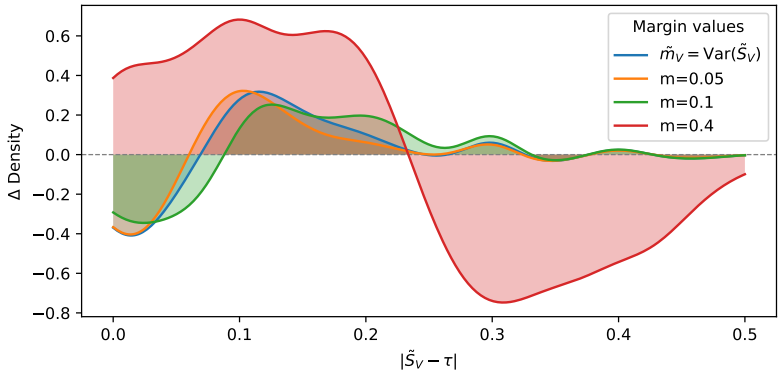


Figure 5: **Δ of Score density with regards to distance to decision boundary.** For all samples in UCF-Crime and XD-Violence, it is shown that high m value resulted in more ambiguious predictions with $|\tilde{S}_V - \tau| \to 0$ while a small or local variance based m effectively pushes the predictions away from decision boundary as we expected.

Sensitivity on window length ℓ: We heuristically set our minimal suspicious window $\ell = \max(300, T/10)$, in which 300 frames is a floor for the shortest window $W_{\max}$. Since a clip c_i (the smallest scoring unit) also spans 300 frames $\approx 10s$, lowering this floor has little effect.

As a result shown in Table 8, an overly large ℓ (as a result of a smaller divisor on video length T) degrades the performance. We suspect that a large window size ℓ hides fleeting anomalies as the

Table 7: **Impact of several margin values ($m \in (0, 0.5)$) on VAD performance.** All settings outperform the baseline, with stable results across different m values.

Margin values	UCF-Crime (AUC)	XD-Violence (AP)	XD-Violence (AUC)	UBNormal (AUC)
$m = 0.05$	**84.28**	68.07	91.34	68.97
$m = 0.10$	83.10	68.16	91.40	69.52
$m = 0.20$	83.57	**68.36**	**91.60**	70.33
$m = 0.40$	79.21	67.45	90.81	**70.59**
$\tilde{m}_V = \mathrm{Var}(S_V)$	84.08	68.03	91.23	69.02

Table 8: Impact of several window lengths (ℓ) on VAD performance

Window length	$\ell = \max(300, T/5)$	$\ell = \max(300, T/10)$	$\ell = \max(300, T/15)$
UCF AUC (%)	81.07	**84.28**	83.66

window may have higher probability of containing benign frames with lower scores, resulting in a lower estimate of the surrogate video-level anomaly probability $\tilde{s}_V$. In addition to such heuristics we used, it is also possible to introduce an additional time series segmentation model [Lovrić et al., 2014] to identify abnormal event intervals from sequences of frame scores.

Impact of post-processing In addition to m, we also evaluate the impact of the Gaussian smoothing parameter used in score post-processing. It's typical to conduct postprocessing (Gaussian, EMA) to the anomaly scores for VAD tasks [Zanella et al., 2024, Ye et al., 2025]. We followed this typical practice and implemented a simple Gaussian filter on the final score. The following Figure 6 demonstrate the robustness of our method on different σ values we use for gaussian smoothing post-processing.

B.2 Impact of different VLM/LLM components

Ablation on Monolithic Multimodal LLMs In our work, by default, we followed modular architecture of VLM + LLM from previous baseline [Zanella et al., 2024]. There are also other experiments and claims supporting this design.

In Table 3a, we have provided ablation to end-to-end VLM performance when used for scoring on every 16 frames clips. As a result, our discrete VLM, LLM framework provide better performance (84.28% against 77.67%). Which aligns with the trend reported in the baseline [Zanella et al., 2024]. We also tested a even simpler baseline of using VideoLLaMA3-7B to conduct direct end-to-end QA with complete video inputs and asking for timestamps of anomalous intervals. The Table 9 shows that such simple design gives much poorer performances even poorer overall performance.

Besides these experimental support for the modular design. Another earlier work [Chen et al., 2023a] also suggests such capability of LLMs to coordinate separate VLM models for better reasoning. Especially for cases where the task domain is a niche one under-represented in the massive pretraining data. These rationales justify our modular VLM/LLM design over single model.

Modular Ablation on Different Multimodal LLMs To validate the generality of our method across different MLLM components. Table 10 varies the checkpoints plugged into our pipeline. With the LLM fixed ($\theta_{\text{LLM}} =$ `Llama-3.1-8B-Instruct`), downgrading the vision backbone from VideoLLaMA3-7B to a 2B variant or to a Qwen2.5-VL results in only a marginal drop $\leq 1\%$ AUC, indicating that the reasoning loop compensates for weaker video features. Conversely, keeping the same VLM and swapping the LLM shows larger but still moderate drops: a 3B instruct model loses $\sim 3\%$ AUC, whereas an older Llama-2-13B loses $\sim 4\%$. Overall, every combination remains above 80% AUC, confirming the *plug-and-play* nature of our framework: it can enhance a wide range of pre-trained VLM/LLM pairs with minimal performance degradation, and it benefits most from stronger language reasoning while being relatively insensitive to vision backbone capabilities, by reducing holistic understanding into a chained process of solving simpler, modular tasks.

Furthermore, to show that our performance gain is not solely from the stronger capability of newer VLM and LLM, we run modernised baseline method [Zanella et al., 2024] under newer VideoLLama3-

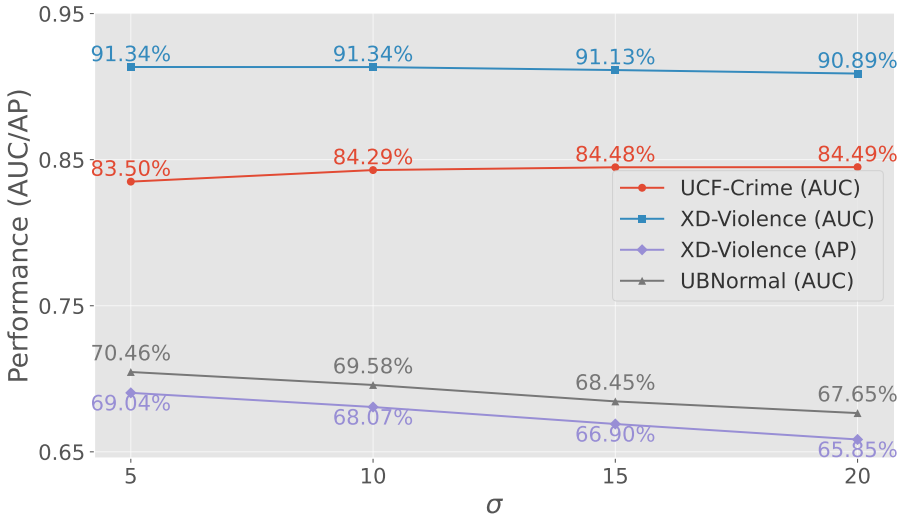


Figure 6: **VAD performance stability w.r.t. Gaussian smoothing** σ. Performance remains stable across different σ values. We simply choose a default value $\sigma = 10$ and a SciPy's default `truncate = 4.0` (which yields an effective radius of 4σ) for all the VAD experiments.

Table 9: VideoLLaMA3-7B End-to-end VAD QA Results

	UCF AUC (%)	XD AUC (%)	XD AP (%)	UBN AUC (%)
Direct QA	58.68	62.52	33.76	53.73
Ours (Full)	**84.28**	**91.23**	**68.03**	**69.02**

7B [Zhang et al., 2025a] and Llama3.1-8B [Grattafiori et al., 2024] backbones. In Table 11, we observed a drop in single VLM performance when using newer model under Zanella et al. [2024] on UCF-Crime. This may be due to the limited capability of sentence encoding VLM [Girdhar et al., 2023], which may fail to recognise more nuanced frame caption from newer models. This problem is mitigated on XD-Violence, where more dramatic videos than mundane surveillance footage of UCF-Crime makes raw captions encoded more recognisable in the representation space.

C Additional implementation details

C.1 Detailed prompts

We provide all the used prompts in this part.

Prompts used in VAD Firstly, we used the same p_{caption} across all datasets. Specifically:

Prompts for p_{caption}

```
"role": "system",
"content": (
    "You are an AI assistant analyzing this video segment. "
    "Summarize the main events or actions in a concise way.
)
```

As for p_{VAD}, we mainly adopted base prompts from Zanella et al. [2024]. Following their design, we also applied dataset priors to the prompts, as the definition of anomaly varied for each of them. Specifically, we have a base definition of anomaly events denoted as `dataset_prior = ‘‘suspicious activities’’`. For UCF-Crime, we change it to `‘‘suspicious or potentially`

Table 10: **Ablation of pretrained VLM/LLM models used on UCF-Crime.** We varied different checkpoints for the components in our framework.

θ_{VLM}	θ_{LLM}	AUC (%)
`VideoLLaMA3-7B`	`Llama-3.1-8B-Instruct`	84.28%
`VideoLLaMA3-2B`		83.35%
`Qwen2.5-VL-7B`		83.23%
`VideoLLaMA3-7B`	`Llama-3.1-8B-Instruct`	84.28%
	`Llama-2-13B-Chat`	80.70%
	`Llama-3.2-3B-Instruct`	81.09%

Table 11: Performance of "modernised" baselines [Zanella et al., 2024] with newer backbone models Zhang et al. [2025a] and Grattafiori et al. [2024].

Method	**UCF AUC (%)**	**XD AUC (%)**	**XD AP (%)**
Zanella et al. [2024] (`BLIP2 FLAN-T5-XL + Llama2-13B-chat`)	74.19	85.16	61.09
Zanella et al. [2024] (`VideoLLama3-7B + Llama3.1-8B`)	72.99	84.64	61.20
Ours (`VideoLLama3-7B + Llama3.1-8B`)	**84.28**	**91.34**	**68.07**

`criminal`'', and for XD-Violence, we opt to ''`suspicious or violent`'' subject to the clear definition of anomalies within each of them [Sultani et al., 2018, Wu et al., 2020]. However, on UBNormal [Acsintoae et al., 2022], where the anomalies span a wide range of spontaneous activities that may not be considered malicious by commonsense, we simply keep the base `dataset_prior`.

Prompts for p_{VAD}

"role": "system",
"content": (
f"How would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with {`dataset_prior`} activities?"
"Please provide the response in the form of a Python list and respond with only one number in the provided list below [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] without any textual explanation. It should begin with '[' and end with ']'."
)

"role": "user",
"content": (
f"{$\theta_{\text{VLM}}(c_i, p_{\text{caption}})$}"
)

We also conducted an ablation study for the incorporation of dataset priors in Table 12, which has shown a similar trend to previous works [Ye et al., 2025, Zanella et al., 2024]. Specifically, the overall VAD performance benefited from injecting even a small context prior. Providing even brief contextual definitions of anomaly events improves baseline model performance, providing a stronger motivation for the automated extraction and utilization of the sample-specific anomaly prior we have proposed in our work.

Table 12: Ablation of dataset-level anomaly priors.

Dataset Priors	**UCF-Crime (AUC)**	**XD-Violence (AUC)**
✗	82.94%	90.72%
✓	84.28%	91.34%

As we described in Section 3.2, after we identified W_{max}, we got a segment of video V_{sus}, we queried the θ_{VLM} with the V_{sus} and p_{extract} to get the tag list t_V.

Prompts for p_{extract}

"role": "system",
"content": (
    "You are an AI assistant analyzing a suspicious segment of a video. "
)

"role": "user",
"content": (
    f"{V_{sus}}"
    "Analyze the video interval to identify any possible suspicious behaviors. "
    "Return your answer strictly as a Python-style list of phrases that could briefly describe "
    "the suspicious scene split by commas. "
    "No additional commentary or text, return only the list."'
)

To produce p^*_{VAD}, during inference, we augment p_{VAD} prompts with a template sentence containing t_V. Specifically, we inject the following sentences: $\text{template}(\text{t}_\text{V}) =$ `f‘‘In addition, we have identified certain {dataset_prior} behaviors that may appear in the video. Please consider these carefully when deciding on the final anomaly rating. [Potentially reported suspicious activities: {`t_V`}]’’` right after the first system prompt part of p_{VAD}.

Prompts used in VAL During spatial localization of anomaly regions in video frames, we use the simplest default prompt provided by the official release document of `Qwen2.5-VL` [Bai et al., 2025].

Prompts for p_{LOC}

"role": "user",
"content": (
    f"{f_{i}}"
    "Analyze this image and identify any suspicious or anomalous region, if present."
    "Return your answer in JSON format: "
    "[{"bbox_2d": [x1, y1, x2, y2], "confidence": c}]."
)

To incorporate ground-truth or extracted anomaly priors t_V, t_{oracle}, we simply augment the p_{LOC} by adding them at the start of user prompts as hints to the model. Specifically:

Prompts for p^*_{LOC}

"role": "user",
"content": (
    f"{f_{i}}"
    "The video could contain the following anomaly type: '{t_V}'."
    "Localize the suspicious region or individual in this image."
    "Return your answer in JSON format: "
    "[{"bbox_2d": [x1, y1, x2, y2], "confidence": c}]."
)

Prompts used in VAU For VAU task, we fixed p_{VAU} across different test domains (UCF-Crime, XD-violence), but varied them across different pretrained θ_{VLM} for the best baseline performance, which are:

Prompts for p_{VAU} (`Videochat-Flash-2B`) [Li et al., 2024b]

user prompt = f"Please analyze the video for any anomaly activities. If there is any anomaly, describe the anomaly activities present in the video. After description, analyze why it is an anomaly without timestamps. If no anomalies are found, state that the video appears normal and then describe the scene in detail.{V}"

Prompts for p_{VAU} (`VideoLLaMA3-7B`) [Zhang et al., 2025a]

"role": "system",
"content": (
"You are an AI assistant analyzing a video."
)

"role": "user",
"content": (
f"{V}"
"Please analyze the video for any anomaly activities in detail. "
"If there is any anomaly, describe the anomaly activities present in the video in detail. After description, analyze why it is an anomaly without timestamps."
"If no anomalies are found, state that the video appears normal and then describe the scene in detail."
)

As covered in the main text, producing p^*_{VAU} is simply appending template prompts with t_V to the end of the system prompt (or before the user prompt if the model does not support customizing the system prompt) of p_{VAU}. Specifically, $\text{template}_{\text{VAU}}(t_V)$ =`"For better anomaly detection and description in detail, a preliminary analysis suggests that the suspicious activity could be related to` t_V`. Use these information to guide your anomaly detection analysis."`.

C.2 Detailed sampling strategies

Sampling clip c_i around f_i in VAD Recent VLMs gain capability to process multiple frames as videos [Bai et al., 2025, Li et al., 2024b, Zhang et al., 2025a]. This is a desirable functionality we would like to exploit when dealing with frame-wise VAD. As a single frame may not be able to represent contiguous events. Therefore, following previous works sampling multiple frames to predict s_i [Zanella et al., 2024, Ye et al., 2025], we opt to input a series of frames c_i around the target f_i instead of taking f_i only. Specifically, we sample c_i by following steps.

Let the video run at $\text{fps} = r_f$ and denote by ω [s] the dataset-specific temporal radius we keep on either side of f_i. Empirically, we set $\omega = 10$ s for UCF-Crime and XD-Violence, and $\omega = 5$ s for UBnormal (in which most clips are only 10-15 s long). The total window length in frames is $L = 2\omega\, r_f + 1$ and the half-width is $\delta = \lfloor L/2 \rfloor$. Bounding the window to the video limits,

$$a = \max\big(1,\, i - \delta\big), \qquad b = \min\big(T,\, i + \delta\big),$$

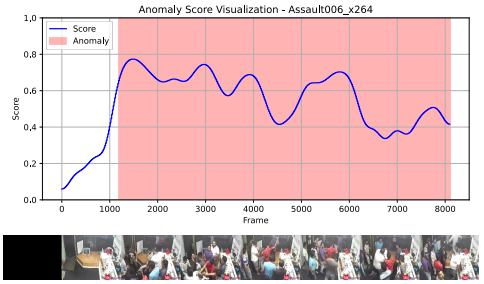
we draw $N = 10$ evenly spaced indices

$$\mathcal{I}(i) = \Big\lfloor \text{linspace}\big(a,\, b,\, N\big) \Big\rfloor, \qquad c_i = \{f_j \mid j \in \mathcal{I}(i)\}.$$

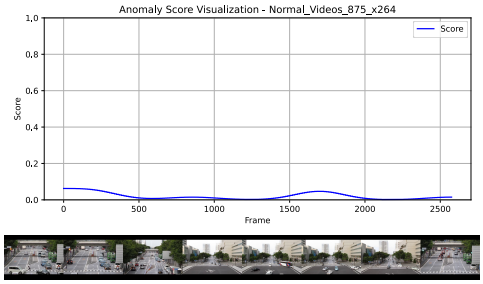
Thus, c_i always contains 10 frames centered as much as possible on f_i. That means, for 30 fps videos in UCF-Crime, the sampling spans up to ± 150 frames (5 s) on either side, automatically shrinking near the video boundaries.

Sampling for downstream tasks For VAL task, we sample all the frames containing anomalies following to practice in previous works [Liu and Ma, 2019, Wu et al., 2024b]. For VAU task, we

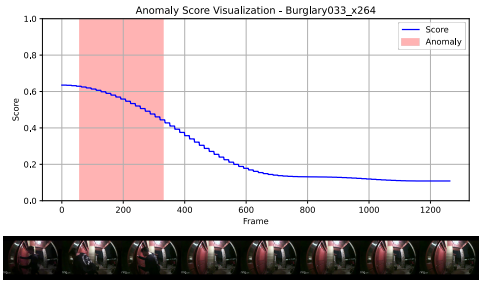
adhere to the default configuration in Zhang et al. [2024b], which samples 16 frames per video for all the methods taking frame inputs.



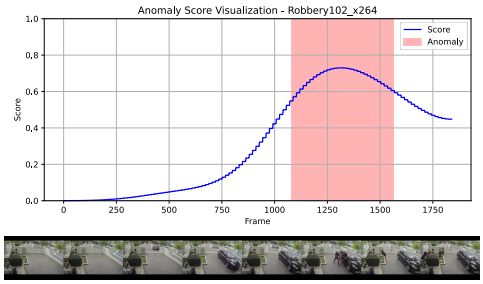
(a) t_V = "physical altercation, assault, fighting", t_{oracle} = "Assault"



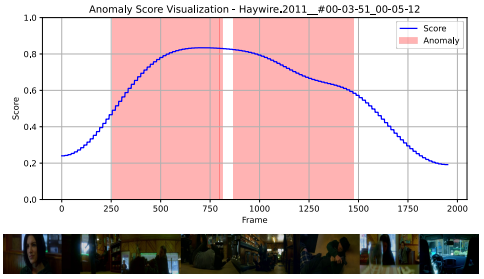
(b) t_V = "crosswalk, traffic light", t_{oracle} = "Normal"



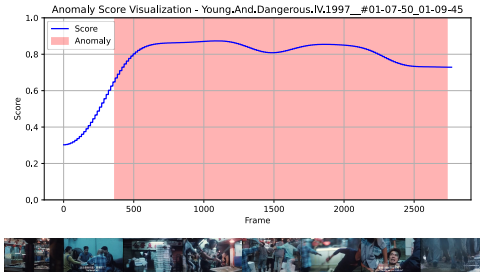
(c) t_V = "attempted break-in, attempted burglary", t_{oracle} = "Burglary"



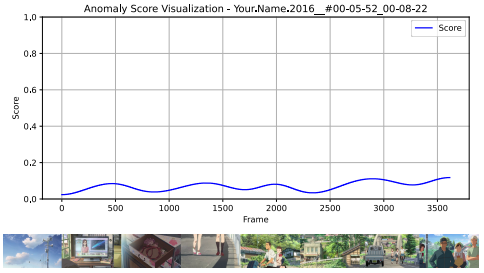
(d) t_V = "kidnapping, assault", t_{oracle} = "Robbery"



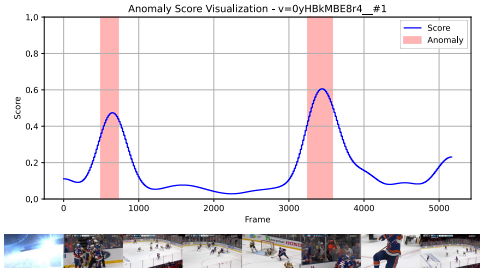
(e) t_V = "kidnapping, fighting, choking, kicking, punching", t_{oracle} = "Fighting, Shooting"



(f) t_V = "fighting, hitting with sticks, throwing objects, running away", t_{oracle} = "Fighting"

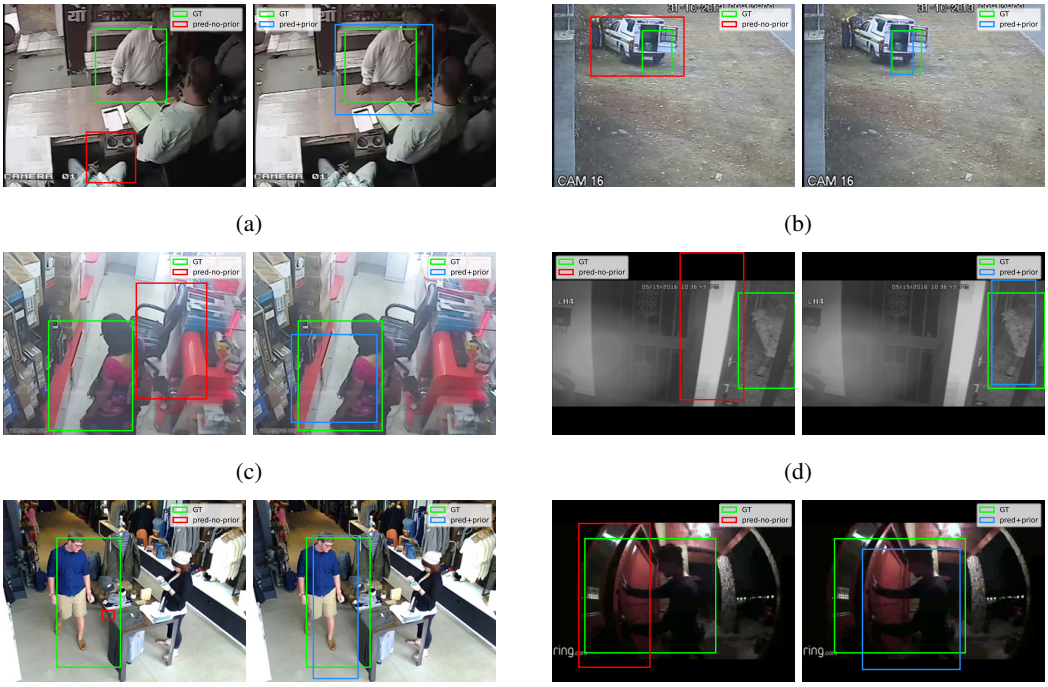


(g) t_V = "[]", t_{oracle} = "[]"



(h) t_V = "fighting, hitting, pushing", t_{oracle} = "Fighting"

Figure 7: **Frame-wise anomaly score plots for eight representative clips.** Our method exhibit consistent performance on various video/anomaly types. The comparison between t_V and t_{oracle} (The original class annotated in Sultani et al. [2018], Wu et al. [2020]) is given for each sample, suggesting the qualitative performance of the anomaly prior extraction step.



(a) (b)

(c) (d)

(e) (f)

Figure 8: **Qualitative examples of our localisation outputs.** Each plot the compares detected anomaly window using baseline prompts and InterTC-refined prompts against the ground truth bounding boxes.

D Additional qualitative results

D.1 More results on frame-level video anomaly detection

We show additional qualitative temporal VAD results with the corresponding t_V tags extracted in Figure 7. For most samples, there are clear and reasonable tags t_V extracted. There are also ambiguous tags, e.g., Appendix C.2, while the performance of the VAD task remains stable. Another interesting observation here is that the t_V extracted, in most cases, are analytical tags for rough t_{oracle} categories. For example, in Appendix C.2, the elaborated t_V = ‘‘`fighting, hitting with sticks, throwing objects, running away`’’ are more tractable then the rough t_{oracle} = ‘‘`Fighting`’’, which explained the observed gap of quantitative performances when using different anomaly priors for VAD task in Table 3b.

As it is shown, despite our method exhibit decent performance in flagging various kinds of anomalies, it failed occasionally on small event gaps (e.g. in Figure 7e). We suspect that this insensitivity may be due to the uniform sampling around f_i we employed to obtain c_i. This may result in the c_i do not have the necessary granularity to represent extremely short video clips. While this is not our focus in this work, future works may consider a dynamic sampling strategy to improve the baseline VLM for VAD.

D.2 More results on spatial video anomaly localization

The additional localization visualization in Figure 8 gives clear evidence proving the performance gain by incorporating Inter-Task Chaining of anomaly priors. The t_V text prompts suggesting possible anomaly contexts allow for more accurate and reasonable groundings.

Table 13: **Amortised per-frame processing time (sec/frame) for a full UCF-Crime test run**, model loading excluded, smaller value means faster.

Method	VLM Captioning	**Caption Cleaning**	**LLM Summary**	**LLM Scoring**	**Score Refinement**	**VAD Overall**
Zanella et al. [2024]	0.06736	0.01490	0.01684	0.01109	0.00673	0.11691
Ours	**0.02587**	—	—	**0.00314**	**0.00026**	**0.02927**

D.3 More qualitative results on video anomaly understanding

In addition to the results shown in Figure 4, we include extra qualitative comparisons in Figure 9 and Figure 10. The results clearly show that MLLMs assisted with Inter-Task Chaining produced excellent VAU results, which accounted for the quantitative performance gains in terms of both traditional NLP metrics and preference on several dimensions of GPT-based evaluations [Tang et al., 2024]. However, we also noticed that sometimes our method produced overly verbose answers compared with other counterparts. This actually aligns with a trend of redundant outputs discovered in LLM reasoning [Sui et al., 2025]. Despite this drawback, the majority contents in our generated descriptions are still focused on the desired topic of anomaly analysis and providing additional details, further enhancing explainability.

E Running-time analysis

As we mentioned earlier in Section 3.1, our method have relatively efficient inference process due to the selective prediction nature saves unnecessary thinking on samples where the first round scores show enough confidence. Beyond this, we also find our method are faster than previous baseline zero-shot LLM work [Zanella et al., 2024] by design. In the following, we provide a complexity analysis of our inference steps and compare it with that of the prior work.

Our test-time IntraTR pipeline for VAD requires 1 VLM captioning query and 1 LLM scoring query per 16 frames, along with a single VLM query per suspicious video to extract tags. For videos flagged as "uncertain", we perform an additional LLM scoring query per 16 frames. In total, our method performs at most 1 VLM and 2 LLM queries per 16 frames, plus fewer than 1 VLM query per video on average. In contrast, full method of previous work [Zanella et al., 2024] performs up to 5 VLM captions per frame and 2 additional LLM queries for summarising and scoring per 16 frame. It also requires additional refinement steps that introduce massive costs of encoding captions and vector searching.

Table 13 reports the amortized processing clock time inference speed on 2 RTX 3090 GPUs for a full run of the UCF-Crime test set (model loading time excluded). This gives clear supporting evidence of our efficiency advantage over the previous work.

F Limitations

Despite its effectiveness, our method exhibits several limitations. First, its performance is fundamentally constrained by the representational capabilities and prior knowledge of the underlying frozen multimodal large language models, which may occasionally introduce semantic biases or inaccuracies inherited from their pretraining data (see failure cases in Figure 11). Secondly, due to reliance on frozen models, our approach may suffer from reduced sensitivity in detecting highly subtle or domain-specific anomalies compared to explicitly fine-tuned models.

G Broader Impacts

Our work aims at enhancing public safety through better anomaly detection and interpretability in surveillance systems. However, broader deployment raises ethical considerations regarding privacy and potential misuse. Improved localization and descriptive capabilities could inadvertently facilitate invasive surveillance practices or profiling if misapplied without proper governance. Thus, any practical application of our method should be carefully regulated, ensuring transparency, accountability, and compliance with privacy laws and ethical guidelines to prevent societal harm while benefiting public security and safety.

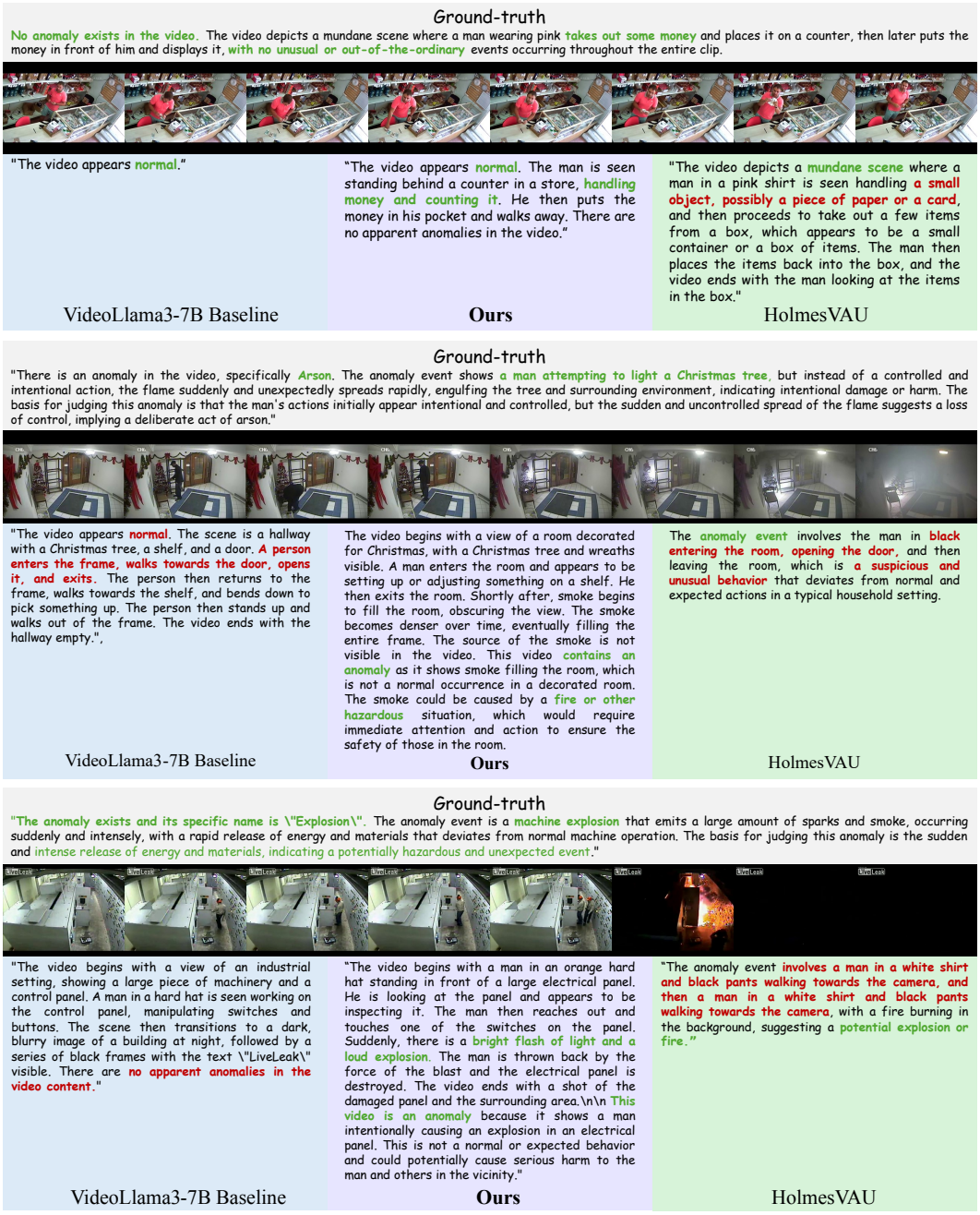


Figure 9: **Representative qualitative results for the video-anomaly understanding task (part-2).** **Green parts** represents correct description/reasoning about the anomaly and the **red parts** highlight the statements inconsistent with the ground-truth.

Figure 10: **Representative qualitative results for the video-anomaly understanding task (part-1).** **Green parts** represents correct description/reasoning about the anomaly and the **red parts** highlight the statements inconsistent with the ground-truth.

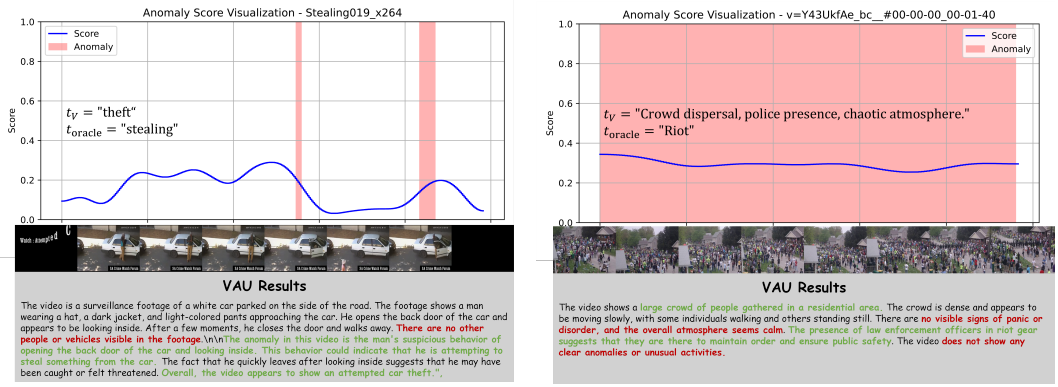


Figure 11: **Failure video anomaly analysis cases.** Both contains nuanced anomaly events that may be hard to determine. We find that for both cases, the model can still reasonable anomaly tags t_V despite unsatisfactory VAD scores, therefore still yielding partially correct (**Green**/**Red** fonts represents **Correct**/**Wrong** statements) textual anomaly descriptions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We make clear statements in both abstract and introduction for the main contribution of this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We included discussions on the limitations in the supplementary material Appendix F and Appendix G.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not contain theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included all the necessary details to reproduce the experiment results, and we plan to release the code implementation soon after review process.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be released soon in the project page.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We covered all these detailed settings in both the main text and the supplemental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not typically reported in the field of video anomaly analysis since it would be too computationally expensive. We run all the experiments under a fixed seed, and the gaps between the previous/baseline works in our main experiments are large enough even considering the standard deviation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We disclosed the computation resources that have been used in the experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed the potential impacts in supplemental material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our paper utilized pretrained models, and each of them are protected separately by their original providers.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cited and credited all the resources used in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crownsourcing experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crownsourcing experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLMs only function as a standard component in our methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.