OBSERVE, HYPOTHESIZE, AND VERIFY: ACTIVE EVIDENCE-SEEKING REASONING FOR TRAINING-FREE VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Anomaly Detection (VAD) aims to identify rare and complex anomalous events in videos, which requires the model to capture fine-grained details and reason about their contextual relationships. However, existing training-free VAD methods commonly follow a passive *caption-and-judge* paradigm, where video clips are first converted into generic textual descriptions and anomaly reasoning is restricted to evidence already captured by generic video descriptions via vision-language models. While effective for visually salient events, this paradigm is insufficient for evidence-demanding anomalies, which depend on subtle object interactions, temporally distributed clues, and careful verification. Therefore, we propose **OHV-VAD**, which formulates video anomaly detection as an *observe-hypothesize-verify* reasoning process for active evidence seeking and verification. This paradigm is built around three evidence-centered capabilities: 1) **Hypothesis-Driven Evidence Acquisition** transforms passive captioning into targeted perception by using diagnostic questions as evidence probes to acquire anomaly-relevant visual evidence. 2) **Adaptive Temporal Evidence Refinement** maintains a compact evidence state through a clue memory that continuously organizes anomaly-relevant clues across video segments. 3) **Reflective Anomaly-Normality Verification** evaluates each suspicious segment under competing abnormal and normal explanations, mitigating over-detection caused by one-sided anomaly reasoning. Extensive experiments on VAD benchmarks demonstrate strong overall performance among training-free approaches. Category-wise analyses further show that OHV-VAD effectively detects evidence-demanding anomalies while retaining strong performance on visually salient events, and false-positive analyses confirm its ability to reduce over-detection of normal behaviors.

1 INTRODUCTION

Recent training-free video anomaly detection methods have made notable progress by leveraging multimodal large language models (MLLMs) and large language models (LLMs). For example, LAVAD Zanella et al. (2024) generates captions for individual video frames, filters noisy descriptions, and then uses an LLM to infer anomaly scores. Monitor Yang et al. (2025b) further improves this paradigm by aggregating and refining caption-level information, while AnomalyRuler Yang et al. (2024a) introduces anomaly-related rules and augments generated captions with anomalous elements to guide language-model reasoning.

Although they differ in how to process and reason over textual descriptions, the evidence available to their reasoning process is still largely determined by generic captions generated beforehand. Consequently, they focus primarily on improving how fixed textual observations are filtered, augmented, or interpreted. We refer to this common formulation as the passive caption-and-judge paradigm (Figure 1 above).

Such passive paradigm can be effective for visually salient anomalies readily captured by generic descriptions, but becomes less reliable when decisive evidence should be deliberately sought and temporally connected. For anomalies such as burglary, shoplifting, and theft, event-related evidence may be found in subtle human–object interactions, preparatory actions, changes in object ownership

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

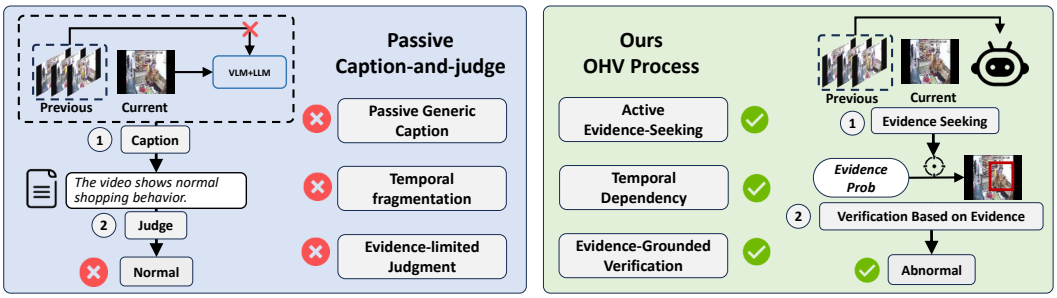

Figure 1: **Caption-and-judge paradigm (e.g., LAVAD) vs. Observe-Hypothesize-Verify paradigm (OHV-VAD)**. Our proposed method enables effective anomaly detection by actively seeking anomaly-relevant evidence and performing evidence-grounded verification.

or location, or clues scattered across multiple video segments. They can easily be omitted by generic captioning and, once lost during visual perception, are difficult for downstream caption refinement or LLM reasoning to recover. Besides, judging suspicious behaviors from incomplete or isolated observations may obscure their normal context and produce false alarms.

Motivated by this observation, we argue that the key bottleneck of LLM-based training-free VAD lies not only in how to reason over captions, but also in *how to acquire, organize, and verify anomaly-relevant evidence before making decisions*. Therefore, as illustrated in Figure 1, we propose OHV-VAD, which formulates video anomaly detection as an observe-hypothesize-verify reasoning process. Rather than treating textual descriptions as passive captions for judgment, OHV-VAD turns them into evidence-seeking observations for verification.

Specifically, OHV-VAD instantiates this reasoning process through three evidence-centered capabilities. First, **Hypothesis-Driven Evidence Acquisition (HDEA)** transforms generic visual captioning into targeted perception by using diagnostic questions as evidence probes, enabling the model to focus on subtle interactions and scene changes that are easily missed by passive descriptions. Second, **Adaptive Temporal Evidence Refinement (ATER)** maintains a compact evidence state that continuously collects, filters, and refines anomaly-relevant clues during video stream processing, allowing later decisions to be grounded in temporally distributed evidence. Third, **Reflective Anomaly-Normality Verification (RANV)** evaluates suspicious segments through competing explanations, assessing not only whether the evidence supports an anomaly hypothesis but also whether the behavior can be explained by normal patterns. Overall, OHV-VAD provides an evidence-centered reasoning framework that observes temporal context, hypothesizes potential anomaly developments, actively acquires targeted visual evidence, and verifies suspicious events against both abnormal and normal explanations. By organizing perception, temporal grounding, and decision verification around explicit evidence, OHV-VAD enables more reliable training-free detection of evidence-demanding anomalies in complex surveillance videos. Our contributions are summarized as follows:

- We identify the passive caption-and-judge paradigm as a key bottleneck for evidence-demanding anomalies and formulate training-free VAD as an observe-hypothesize-verify reasoning problem.
- We propose OHV-VAD, an evidence-centered framework with hypothesis-driven evidence acquisition, adaptive temporal evidence refinement, and reflective anomaly-normality verification.
- We conduct extensive experiments and analyses showing that OHV-VAD achieves superior training-free VAD performance and that its evidence-centered components improve detection of complex anomalies.

2 Related Work

Training-based Video Anomaly Detection. Video anomaly detection (VAD) has been studied under one-class, weakly supervised, and instruction-tuned settings. One-class methods Cao et al.

(2023); Liu et al. (2018); Yang et al. (2023); Yan et al. (2023) learn normal patterns and detect deviations, whereas weakly supervised methods use video-level labels to localize anomalous segments. Recent methods improve temporal modeling and feature discrimination. For example, RTFM Tian et al. (2021) and MGFN Chen et al. (2023) enhance discriminative representation learning, while VadCLIP Wu et al. (2024) adapts CLIP to VAD through vision-language alignment. More recently, VAD-LLaMA Lv & Sun (2024), Holmes-VAD Zhang et al. (2024) and Holmes-VAU Zhang et al. (2025) fine-tune vision-language models on multimodal instruction data for explainable and long-term anomaly understanding.

Training-free Video Anomaly Detection. Training-free methods employ frozen MLLMs and LLMs without task-specific parameter updates. LAVAD Zanella et al. (2024) generates and filters frame-level captions before using an LLM for reasoning, while Monitor Yang et al. (2025b) introduces a gating module to refine and aggregate caption-level information for anomaly scoring. AnomalyRuler Yang et al. (2024a) introduces LLM-generated rules to guide anomaly reasoning. EventVAD Shao et al. (2025) and VADTree Li et al. (2026) further expand the scope of captioning to the event level by leveraging event boundary detection models; however, their event-level processing pipelines require access to future frames, while QVAD Bekit et al. (2026) iteratively refines the description of each video clip in isolation through prompt updating. More recently, PANDA Yang et al. (2025c) develops an agentic VAD system with scene-aware planning, heuristic reasoning, tool-augmented reflection, and Chain-of-Memory.

Despite these advances, many training-free methods still follow a passive caption-and-judge paradigm, in which anomaly reasoning relies on generic visual descriptions without sufficient anomaly-relevant evidence to support the decisions. PANDA moves beyond this structure through planning and memory, but its memories mainly retain reasoning and reflection trajectories for strategy planning and experience retrieval, rather than explicitly consolidating evidence across video segments. Consequently, subtle or temporally distributed evidence may be omitted during perception and cannot be reliably recovered by downstream reasoning.

Our work focuses on the evidence acquisition and organization bottlenecks of training-free VAD. OHV-VAD formulates anomaly detection as an observe–hypothesize–verify process that acquires hypothesis-relevant visual evidence, maintains a compact temporal evidence state, and verifies suspicious events under abnormal and normal explanations. Unlike trajectory-oriented memory, our evidence state consolidates segment-level clues according to their visual and semantic redundancy, preserving complementary evidence while suppressing repetitive observations. This evidence-centered formulation is designed for anomalies that cannot be reliably identified from generic clip descriptions or reasoning histories alone.

3 METHOD

3.1 EVIDENCE-CENTERED FORMULATION

We present **OHV-VAD**, which formulates video anomaly detection as an active evidence-seeking process. Given a video $\mathcal{V}$, conventional LLM-based methods typically describe each clip and infer abnormality from the resulting fixed textual observation, implicitly assuming that the caption preserves all anomaly-relevant evidence. This assumption often fails for evidence-demanding anomalies whose decisive cues involve subtle interactions, preparatory actions, or event fragments distributed across multiple segments.

OHV-VAD instead couples perception and anomaly reasoning in a recurrent **observe–hypothesize–verify** process. The central idea is that anomaly decisions should be grounded in an evolving evidence state rather than an isolated segment description. Let the video be divided into T non-overlapping segments, $\mathcal{V} = \{S_1, S_2, \ldots, S_T\}$, each containing a set of sampled frames. During inference, OHV-VAD maintains an evidence-centered state:

$$\mathcal{Z}_t = \big(\mathcal{O}_t, \mathcal{M}_t, \mathcal{H}_t\big) \tag{1}$$

where $\mathcal{O}_t$ is the observation of S_t, $\mathcal{M}_t$ denotes the evidence memory accumulated up to S_t, $\mathcal{H}_t$ contains the event hypotheses that remain plausible under the current observations. OHV-VAD evaluates this state to produce a final analysis and anomaly score for the current segment:

$$(y_t, R_t) = \Phi_{\mathrm{OHV}(S_t, \mathcal{Z}_{t-1}; \mathcal{K}_{\mathcal{A}}),} \tag{2}$$

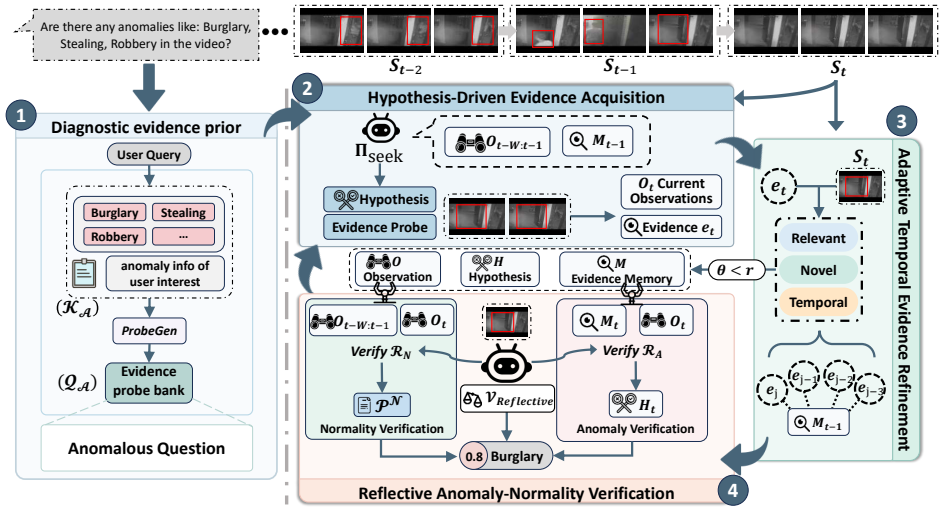

Figure 2: **The detailed framework of OHV-VAD.** OHV-VAD first constructs a diagnostic evidence prior, after which OHV-VAD performs an Observe–Hypothesize–Verify cycle, where Hypothesis-Driven Evidence Acquisition seeks relevant visual evidence, Adaptive Temporal Evidence Refinement maintains an evolving evidence memory, and Reflective Anomaly–Normality Verification contrasts competing explanations to produce evidence-grounded anomaly assessment.

where $\mathcal{K}_\mathcal{A}$ denotes anomaly knowledge induced from the anomaly categories $\mathcal{A}$ specified by user. $y_t \in [0, 1]$ is the segment-level anomaly score and R_t is the evidence-grounded rationale. Observations update the persistent evidence state, hypotheses direct subsequent visual inspection, and verification compares abnormal and normal explanations to make a reliable assessment of abnormality. As shown in Figure 2, this process is instantiated through three evidence-centered capabilities: Hypothesis-Driven Evidence Acquisition, Adaptive Temporal Evidence Refinement, and Reflective Anomaly-Normality Verification.

3.2 Hypothesis-Driven Evidence Acquisition

Diagnostic Evidence Prior. Different anomalies require different diagnostic evidence: shoplifting may involve concealment followed by departure without payment, whereas burglary may involve forced or unauthorized entry.

Upon receiving a user instruction specifying the anomaly categories of interest $\mathcal{A}$, OHV-VAD autonomously constructs an anomaly-specific diagnostic evidence prior. For each category $a \in \mathcal{A}$, OHV-VAD retrieves its semantic definition, while also allowing the user to provide a customized description that specifies the anomaly of interest. Based on the acquired knowledge $\mathcal{K}_a$, OHV-VAD extracts the key visual characteristics of each anomaly category and converts these characteristics into diagnostic questions that specify what visual evidence should be examined:

$$\mathcal{Q}_\mathcal{A} = \bigcup_{a \in \mathcal{A}} \text{ProbeGen}(a, \mathcal{K}_a) \tag{3}$$

The resulting evidence-probe bank targets visually verifiable conditions such as entity states, human–object interactions, preparatory actions, and action consequences. More details are provided in the supplementary material.

Context-conditioned hypothesis formation. Before observing segment S_t in detail, OHV-VAD interprets the preceding observations and accumulated evidence to hypothesize plausible event developments. Let $\mathcal{O}_{t-W:t-1}$ denote a context window of W segments. OHV-VAD performs evidence-probe selection policy Π_{select} to generate plausible event hypotheses $\mathcal{H}_t$ and diagnostic probes $\mathcal{Q}_t$:

$$(\mathcal{H}_t, \mathcal{Q}_t) = \Pi_{\text{select}}(\mathcal{O}_{t-W:t-1}, \mathcal{M}_{t-1}; \mathcal{Q}_\mathcal{A}) \tag{4}$$

Accordingly, rather than applying $\mathcal{Q}_{\mathcal{A}}$ as a fixed checklist, OHV-VAD uses Π_{select} to select and activate only a small subset of probes from $\mathcal{Q}_{\mathcal{A}}$, with the selection conditioned on both preceding observations and accumulated memory.

A hypothesis represents a plausible continuation rather than a premature anomaly label; it identifies what remains uncertain and which visual conditions would be informative. For example, if the preceding observations $\mathcal{O}_{t-W:t-1}$ show a person repeatedly manipulating the lock of a closed entrance door at night, the hypothesis $\mathcal{H}_t$ should be formulated as follows: *the person may be attempting to gain unauthorized entry.* OHV-VAD uses the selected probes in $\mathcal{Q}_t$ (e.g., "Has the person been interacting with the lock for an extended period? Are they attempting to open the door?") to actively uncover visual evidence relevant to this hypothesis in the current video segment.

Targeted visual observation. OHV-VAD uses the selected probes to actively uncover complementary visual clues and produces an evidence-bearing observation:

$$(\mathcal{O}_t, e_t) = \mathcal{P}_{\text{vis}}(S_t \mid \mathcal{H}_t, \mathcal{Q}_t) \tag{5}$$

where $e_t = \big(c_t, \rho_t\big)$, c_t denotes a compact evidence statement (e.g., "A person has remained there for a long time."), and ρ_t characterizes its relation to each active hypothesis as "supportive", "contradictory", or "undetermined". $\mathcal{O}_t$ provides a description of the segment together with responses to the selected probes. By conditioning visual perception on $\mathcal{Q}_t$, OHV-VAD examines anomaly-relevant details that may be overlooked by passive generic captioning. These evidence–hypothesis relations guide subsequent hypothesis revision and probe selection, enabling OHV-VAD to progressively uncover anomalous events.

3.3 Adaptive Temporal Evidence Refinement

Evidence-demanding anomalies are rarely supported by a single decisive frame. Their interpretation often depends on connecting an early preparatory action with a later consequence. Directly retaining all segment descriptions, however, introduces severe redundancy because adjacent surveillance segments frequently depict nearly identical content. OHV-VAD therefore maintains a compact temporal evidence memory $\mathcal{M}_t$. Given a new record e_t, the memory is updated as

$$\mathcal{M}_t = \mathcal{C}(\mathcal{M}_{t-1} \cup \{e_t\}; \mathcal{H}_t) \tag{6}$$

where $\mathcal{C}$ denotes an adaptive evidence-consolidation operator. It favors evidence that is novel relative to the current memory and temporally complementary to earlier event fragments. $\mathcal{C}$ estimates the evidence using the visual representation of each segment and the textual representation of its evidence statement:

$$v_t = E_V(S_t), \qquad u_t = E_T(c_t) \tag{7}$$

where E_V and E_T are visual and textual encoders. For records e_t and $e_j (e_j \in \mathcal{M}_{t-1})$, their redundancy is estimated by:

$$r(t, j) = (\alpha \ \text{sim}(v_t, v_j) + \beta \ \text{sim}(u_t, u_j))\text{e}^{-\lambda|t-j|} \tag{8}$$

where α and β balance visual and semantic similarity, respectively, while the temporal decay emphasizes temporally nearby observations. If $r(t, j)$ exceeds the threshold θ, the evidence is not retained in $\mathcal{M}$. Thus, only sufficiently novel and informative evidence e_t will be retained.

3.4 Reflective Anomaly-Normality Verification

Visual cues associated with anomalous events are often not exclusive to anomalies. The same observable behavior may admit both abnormal and normal interpretations depending on its temporal context, scene conventions, and subsequent consequences: handling merchandise may indicate either shopping or shoplifting, and object transfer may reflect legitimate work or theft. OHV-VAD conducts the final assessment as a comparative verification between competing abnormal and normal explanations. The abnormal explanation tests whether the defining conditions of a candidate anomaly are satisfied, whereas the normal explanation tests whether routine behavior can coherently account for the observations.

For a suspicious segment S_t, OHV-VAD uses its targeted observation $\mathcal{O}_t$ and relevant evidence retrieved from $\mathcal{M}_t$ to construct an anomaly explanation:

$$R_t^A = \mathcal{R}_A(\mathcal{O}_t, \mathcal{M}_t, S_t) \mid \mathcal{H}_t) \tag{9}$$

Method	Expl	Supervision	UCF(AUC%)	XD(AUC%)	XD(AP%)
GODS Wang & Cherian (2019)	✗	Weakly	70.46	61.56	-
Wu et al. Wu et al. (2020)	✗	Weakly	82.44	-	73.20
RTFM Tian et al. (2021)	✗	Weakly	84.03	-	77.81
MSL Li et al. (2022b)	✗	Weakly	85.62	-	78.58
SSRL Li et al. (2022a)	✗	Weakly	87.43	-	-
MGFN Chen et al. (2023)	✗	Weakly	86.98	-	80.11
CLIP-TSA Joo et al. (2023)	✗	Weakly	87.58	-	82.16
TPWNG Yang et al. (2024b)	✗	Weakly	87.79	-	83.86
VadCLIP Wu et al. (2024)	✗	Weakly	88.04	-	84.51
LECVAD Wang & Chen (2025)	✗	Weakly	89.97	-	86.56
Pi-VAD Majhi et al. (2025)	✗	Weakly	90.33	-	85.37
VERA Ye et al. (2025)	✓	Weakly	86.55	88.26	70.54
LAS-VAD Wang & Zhao (2026)	✗	Weakly	90.86	-	87.92
HolmesVAD Zhang et al. (2024)	✓	Instruction-Tuning	89.51	-	90.67
HolmesVAU Zhang et al. (2025)	✓	Instruction-Tuning	88.96	-	87.68
ZS ImageBind Girdhar et al. (2023)	✗	Training-free	55.78	55.06	25.36
ZS CLIP Zanella et al. (2024)	✗	Training-free	53.16	38.21	17.83
LLaVA-1.5 Liu et al. (2024)	✓	Training-free	72.84	79.62	50.26
LAVAD Zanella et al. (2024)	✓	Training-free	80.28	85.36	62.01
VideoAgent Fan et al. (2024)	✓	Training-free	77.61	83.14	53.32
Monitor Yang et al. (2025b)	✓	Training-free	82.57	79.11	55.01
PANDA Yang et al. (2025c)	✓	Training-free	84.89	-	70.16
QVAD Bekit et al. (2026)	✓	Training-free	84.28	-	68.53
CoREVAD Lim & Hur (2026)	✓	Training-free	82.51	-	70.91
OHV-VAD	✓	Training-free	**85.67**	**89.03**	**72.05**

Table 1: Comparisons with state-of-the-art training-based and training-free methods on UCF-Crime and XD-Violence datasets. "Expl" means "Explanation". "Weakly" and "Instruction-Tuning" denote the training paradigm. "ZS" refers to "Zero-Shot".

OHV-VAD then constructs a competing normal explanation. It summarizes common behavior patterns associated with the monitored scenes in S_t, producing a normality prior $\mathcal{P}^N$. Conditioned on this prior, the same observations are examined from the normality perspective:

$$R_t^N = \mathcal{R}_N(\mathcal{O}_t, \mathcal{O}_{t-Z:t-1}, \mathcal{P}^N) \tag{10}$$

It evaluates whether the behavior is routine and whether a suspicious interaction has a plausible benign purpose. Importantly, a segment is not considered normal simply because no visually salient anomaly is detected. Instead, OHV-VAD constructs an explicit normal explanation and examines whether it can coherently account for the observed behavior and its temporal context.

Finally, OHV-VAD adjudicates between the two explanations and produces a final assessment and anomaly score for the segment:

$$(y_t, R_t) = \mathcal{V}_{\text{ref}}(R_t^A, R_t^N, \mathcal{M}_t) \tag{11}$$

where $y_t \in [0, 1]$ is the anomaly score and R_t is the evidence-grounded explanation.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate OHV-VAD on two widely used VAD datasets, UCF-Crime Sultani et al. (2018) and XD-Violence Wu et al. (2020). UCF-Crime contains 1,900 training and 290 testing videos covering 13 anomaly categories, while XD-Violence contains 3,954 training and 800 testing videos with six anomaly categories. Following standard protocols, we report frame-level AUC on UCF-Crime and both AUC and AP on XD-Violence.

Implementation Details of OHV-VAD. In our experiments, we divide each video into non-overlapping segments of 8 seconds in length. For each segment, the sampling rate is 1 fps. We

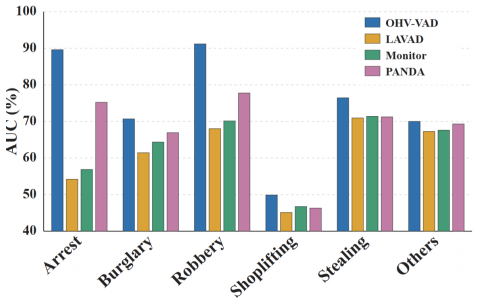

Figure 3: Category-wise AUC comparison on the UCF-Crime dataset. "Others" comprises visually salient anomaly categories, including Arson, Abuse, Assault, Explosion, Fighting, RoadAccidents, Shooting, and Vandalism.

Method	TNR% ↑	FPR% ↓
LAVAD	83.51	16.49
Monitor	86.36	13.64
PANDA	89.74	10.26
OHV-VAD	93.79	6.21

Table 2: Comparison of the true negative and false positive rate (TNR, FPR) on normal videos.

Probe Setting	AUC (%)
Irrelevant	75.12
Without	80.81
Oracle	85.93
Hyp.-Selected (Ours)	85.67

Table 3: Effect of different evidence-probe settings in HDEA on the UCF-Crime dataset.

use Qwen3-VL-8B as the MLLM and Qwen3-14B Yang et al. (2025a) as the LLM. The observation context window W is set to 3. In the Adaptive Temporal Evidence Refinement section we use DINOV2 Oquab et al. (2023) as the visual encoder and BGE Chen et al. (2024) as the textual encoder. In Equation (8), $\lambda = 0.05$, α and β are set to 0.4 and 0.6, respectively. Redundancy threshold θ is 0.7. Evidence memory $\mathcal{M}$ is limited to storing up to 50 evidence entries. All experiments were conducted on an NVIDIA RTX A6000 GPU. The exact input and output formats, along with the prompts, for all operators ($\mathrm{ProbeGen}$, Π_{select}, $\mathcal{P}_{\mathrm{vis}}$, $\mathcal{R}_A$, $\mathcal{R}_N$, $\mathcal{V}_{\mathrm{ref}}$) introduced in the Method section are detailed in the supplementary material.

4.2 Comparison with State-of-the-Art Methods

Results on UCF-Crime and XD-Violence datasets. We compare OHV-VAD with both training-based and training-free methods. As shown in Table 1, OHV-VAD achieves an AUC of 85.67% on UCF-Crime, and an AUC of 89.03% and an AP of 72.05% on XD-Violence. It outperforms all compared training-free methods across all evaluation metrics and remains competitive with many training-based methods.

Category-wise comparison. As shown in Figure 3, OHV-VAD achieves the best category-wise AUC on Arrest, Burglary, Robbery, Shoplifting and Stealing. Detecting anomalies in these categories typically requires interpreting object interactions and temporally distributed behaviors, suggesting that active evidence acquisition, temporal evidence refinement, and evidence-grounded verification are particularly beneficial when generic visual descriptions are insufficient. Meanwhile, the competitive performance on the visually salient category group (Others) indicates that OHV-VAD does not improve evidence-demanding anomalies at the expense of visually evident anomalies. For completeness, the full 13-category comparison is reported in the supplementary material, where OHV-VAD achieves the highest AUC in nine out of the 13 original anomaly categories.

Analysis of false alarms. As shown in Table 2, OHV-VAD achieves a higher TNR of 93.79% and a lower FPR of 6.21% than existing training-free methods on normal videos from UCF-Crime. These results indicate that OHV-VAD more reliably recognizes normal behavior and avoids overestimating benign but potentially suspicious activities as anomalies, supporting the effectiveness of reflective anomaly-normality verification.

4.3 Ablation Study

We conduct ablation studies on the UCF-Crime dataset to evaluate the effectiveness of our method.

Ablation study on three evidence-centered capabilities: Hypothesis-Driven Evidence Acquisition (HDEA) enables active evidence seeking by using diagnostic probes to acquire anomaly-relevant visual evidence. Adaptive Temporal Evidence Refinement (ATER) maintains a compact evidence memory that provides temporally distributed context for anomaly verification. Reflective Anomaly-

HDEA	ATER	RANV	AUC(%)	Δ_{Base}	Δ_{Full}
✗	✗	✗	78.54	–	-7.13
✓	✗	✗	82.36	+3.82	-3.31
✓	✓	✗	84.32	+5.78	-1.35
✓	✗	✓	83.92	+5.38	-1.75
✗	✓	✓	83.17	+4.63	-2.50
✓	✓	✓	**85.67**	+7.13	–

Table 4: Ablation study of OHV-VAD. Δ_{Base} denotes the improvement over the baseline without all three modules, while Δ_{Full} denotes the performance difference compared with the full model.

MLLM	Setting	AUC (%)
GPT-4o	None	83.06
	OHV-VAD	**87.39** (+4.33)
Qwen3-VL-30B-A3B	None	81.12
	OHV-VAD	**86.14** (+5.02)
InternVL-8B	None	77.39
	OHV-VAD	**83.97** (+6.58)
Qwen3-VL-8B	None	78.54
	OHV-VAD	**85.67** (+7.13)

Table 5: Robustness analysis of OHV-VAD across diverse MLLM architectures on the UCF-Crime dataset. Consistent performance gains across both proprietary and open-source backbones demonstrate the robustness and generalizability of OHV-VAD across different MLLMs.

Method	Setting	Params.	AUC (%)
VideoAgent	GPT-4o*	–	77.61
PANDA	Qwen2.5-VL-7B Gemini-2.0-Flash*	–	84.89
LAVAD	5× BLIP Llama-2-13B	65B	80.28
Monitor	5× BLIP GLM-4-Flash	82B	82.57
OHV-VAD	Qwen3-VL-8B Qwen3-14B	22B	**85.67**

Table 6: Comparison of model configurations, total parameter counts, and AUC on the UCF-Crime dataset. '*': proprietary model; '–': unavailable parameter count; 5× BLIP: Using 5 different BLIP models.

Normality Verification (RANV) evaluates suspicious events through competing abnormal and normal explanations.

As shown in Table 4, the baseline without any of these capabilities achieves an AUC of 78.54% (Row 1). Equipping OHV-VAD with HDEA improves the AUC to 82.36% (Row 2), bringing a gain of 3.82 percentage points and demonstrating the importance of actively seeking anomaly-relevant evidence rather than relying on generic passive visual descriptions. Incorporating ATER increases the AUC to 84.32% (Row 3), indicating that temporal evidence refinement helps connect complementary clues distributed across video segments. Further introducing RANV, with all three capabilities enabled, OHV-VAD achieves the best AUC of 85.67% (Row 6), improving the baseline by 7.13 percentage points. Additionally, removing HDEA (Row 5), ATER (Row 4) and RANV (Row 3) from the full model decreases the AUC by 2.50, 1.75 and 1.35 percentage points, respectively. These results validate their complementary contributions.

Effectiveness of evidence-probes. Table 3 compares four probe settings in HDEA. *Without Probes* uses passive generic captioning, *Irrelevant Probes* uses questions from unrelated anomaly categories, *Oracle Probes* uses ground-truth category probes, and *Hyp.-Selected* uses probes autonomously selected by OHV-VAD. *Hyp.-Selected* achieves 85.67% AUC, outperforming *Without Probes* by 4.86 percentage points and approaching the oracle result of 85.93%, while irrelevant probes reduce performance to 75.12%. These results demonstrate that HDEA can autonomously select informative probes and acquire anomaly-relevant evidence without access to ground-truth anomaly categories.

Effect of Prior Information on Model Performance. To isolate the effect of priors, we remove the probe bank derived from the diagnostic evidence prior in HDEA (Row 5) and disable the normality prior by removing RANW (Rows 2 and 3). Even without these priors, OHV-VAD still improves over the baseline by more than 3.8%, indicating that its gains do not solely depend on constructed prior information.

Performance of OHV-VAD with Different MLLMs. As shown in Table 5, OHV-VAD consistently improves the AUC across all evaluated MLLM backbones. These improvements demonstrate that

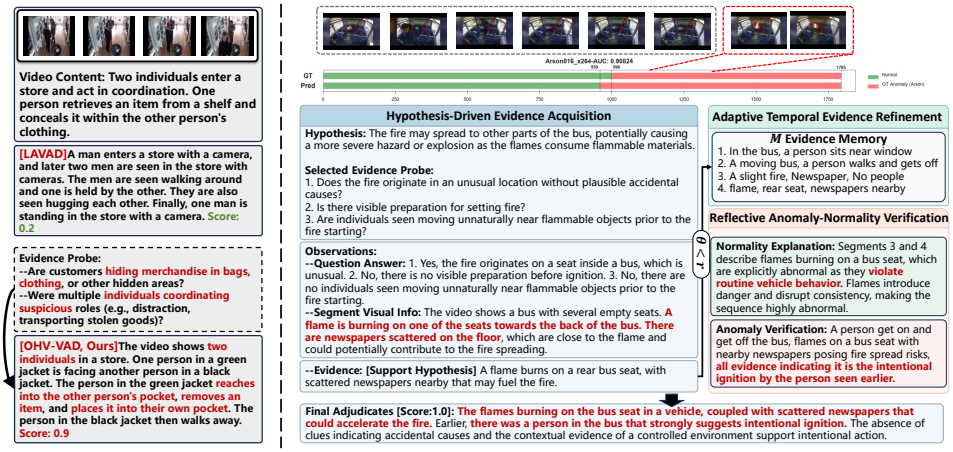

Figure 4: Qualitative results on the UCF-Crime. Left: effectiveness of OHV-VAD's probe-guided evidence acquisition. Right: visualization of the complete evidence-centered reasoning process.

OHV-VAD is not restricted to a specific model and can effectively enhance anomaly reasoning across MLLMs with different architectures and capacities.

4.4 Analysis of Efficiency

As shown in Table 6, we compare different training-free methods in terms of their model configurations, aggregate parameter scales, and detection performance. OHV-VAD achieves the highest AUC of 85.67% using exclusively open-source models, outperforming VideoAgent (77.61%) and PANDA (84.89%), both of which rely on proprietary models. Moreover, although LAVAD and Monitor each employ five BLIP-based models and have total parameter counts exceeding 60B, both still underperform OHV-VAD.

In terms of inference efficiency, OHV-VAD requires only 3–4 LLM/MLLM calls per inference step and achieves 7.59 FPS on our experimental platform. In comparison, LAVAD and Monitor report 7 and 8 calls with throughputs of 0.08 and 0.16 FPS, respectively, while PANDA adaptively requires 3–9 calls and reports 0.82 FPS. Overall, OHV-VAD achieves a favorable balance between detection performance, model accessibility, and inference cost.

4.5 Qualitative Analysis

The left side of Figure 4 shows that OHV-VAD leverages evidence probes to actively uncover subtle yet critical anomaly-relevant evidence, whereas the passive caption-and-judge paradigm represented by LAVAD overlooks these cues and consequently produces an incorrect prediction. The right side illustrates the complete reasoning process of OHV-VAD. Although the current segment contains only visible flames and no person, OHV-VAD continuously accumulates clues across preceding segments to identify the underlying cause of the event. Specifically, it correctly attributes the fire to intentional ignition by a person who previously appeared inside the vehicle. More detailed case studies are provided in the supplementary material.

5 Conclusion

In this paper, we introduced OHV-VAD, a training-free framework that reformulates video anomaly detection as an Observe–Hypothesize–Verify reasoning process. OHV-VAD constructs anomaly-specific diagnostic priors, actively acquires hypothesis-relevant evidence, maintains temporal evidence memory, and verifies suspicious events through competing abnormal and normal explanations. This evidence-centered formulation effectively handles both evidence-demanding and visually salient anomalies. Experiments on UCF-Crime and XD-Violence demonstrate superior performance over existing training-free methods, consistent gains across different MLLM backbones, and favorable inference efficiency. Ablation and category-wise analyses further validate its core capabilities.

Reproducibility Statement

We provide detailed descriptions of the model architecture, training procedure, implementation settings, and evaluation protocols in the main paper and supplementary material. Additional details on data processing, hyperparameters, and supervision construction are also provided to facilitate reproduction of our results.

Ethics Statement

This work is conducted using publicly available benchmark datasets and does not involve the collection of new personal data or human-subject experiments. We have considered the potential ethical implications associated with the use of the proposed method and are not aware of any additional ethical concerns beyond those commonly associated with research on publicly available visual data.

AI Use Statement

Generative AI tools were used solely for language polishing and error checking of the manuscript. The authors reviewed all AI-assisted revisions and take full responsibility for the final content of this work.

References

Lokman Bekit, Hamza Karim, Nghia T Nguyen, and Yasin Yilmaz. Qvad: A question-centric agentic framework for efficient and training-free video anomaly detection. *arXiv preprint arXiv:2604.03040*, 2026.

Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20392–20401, 2023.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5), 2024.

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, pp. 387–395, 2023.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

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)*, pp. 3230–3234. IEEE, 2023.

Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022a.

Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1395–1403, 2022b.

Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. Vadtree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. *Advances in Neural Information Processing Systems*, 38:148372–148404, 2026.

Hyeongmuk Lim and Youngbum Hur. Corevad: A contextual reasoning framework for training-free video anomaly detection. *arXiv preprint arXiv:2605.23116*, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024.

Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and François Brémond. Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24265–24274, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2586–2595, 2025.

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*, pp. 6479–6488, 2018.

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 (ICCV)*, pp. 4975–4986, October 2021.

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*, pp. 8201–8211, 2019.

Yu Wang and Shiwei Chen. Learning event completeness for weakly supervised video anomaly detection. *arXiv preprint arXiv:2506.13095*, 2025.

Yu Wang and Shengjie Zhao. Weakly supervised video anomaly detection with anomaly-connected components and intention reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28545–28556, 2026.

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 *European Conference on Computer Vision (ECCV)*, 2020.

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, pp. 6074–6082, 2024.

Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5527–5537, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Shengtian Yang, Yue Feng, Yingshi Liu, Jingrou Zhang, and Jie Qin. Monitor: Exploiting large language models with instruction for online video anomaly detection. *arXiv preprint arXiv:2510.21449*, 2025b.

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 *European Conference on Computer Vision*, pp. 304–322. Springer, 2024a.

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14592–14601, 2023.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18899–18908, 2024b.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. *arXiv preprint arXiv:2509.26386*, 2025c.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8679–8688, 2025.

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*, pp. 18527–18536, 2024.

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. *CoRR*, abs/2406.12235, 2024.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13843–13853, 2025.

A Overview

- Section B. Additional experiments and parameter analysis.
- Section C. Additional Prompt Details for the OHV-VAD Operators Described in the Main Paper.
- Section D. More detailed case study and visualization of qualitative results.

B Additional Experiments

The effect of the observation context window length W. Table 7 shows the effect of different values of the observation context window length W, as defined in Eq. (4) of the Hypothesis-Driven Evidence Acquisition section. The primary role of W is to provide OHV-VAD with sufficient preceding context to anticipate how subsequent events may unfold, thereby enabling it to formulate plausible hypotheses, select hypothesis-relevant evidence probes, and more effectively uncover anomaly-related evidence. As W increases from 0 to 3, performance improves, indicating that short-term contextual information can provide a more complete event evolution. This enables OHV-VAD to more accurately predict future events and select more relevant evidence probes for active evidence seeking. However, when W increases to 5, the performance decreases, which may suggest that a longer contextual observation may introduce irrelevant events that interfere with hypothesis formation.

W	UCF(AUC(%))
0	83.17
1	83.46
3	85.67
5	84.71

Table 7: Influence of the observation context window length W on the UCF-Crime dataset.

The effectiveness of α and β in Eq. (8). Table 8 investigates the effect of the visual and semantic similarity weights in the adaptive evidence-consolidation operator. In this analysis, we vary α and β while fixing $\lambda = 0.05$ and $\theta = 0.7$. Performance remains stable across all configurations, ranging from 85.41% to 85.67%, indicating that OHV-VAD is robust to the relative weighting of the two similarity measures. The best performance is achieved with $\alpha = 0.4$ and $\beta = 0.6$, suggesting that moderately emphasizing semantic similarity provides a favorable balance for evidence consolidation. Since adjacent surveillance segments often share similar backgrounds and visual appearances, semantic similarity helps prevent visually similar yet semantically complementary evidence from being incorrectly discarded as redundant. However, further increasing β to 0.7 does not yield additional improvement, indicating that visual similarity remains necessary for capturing fine-grained appearance cues. We therefore set $\alpha = 0.4$ and $\beta = 0.6$ by default.

Detailed category-wise comparison. Figure 5 presents a complete category-wise comparison of OHV-VAD with LAVAD Zanella et al. (2024), Monitor Yang et al. (2025b), and PANDA Yang et al. (2025c) across all 13 anomaly categories. The results further highlight the advantage of OHV-VAD in detecting evidence-demanding anomalies, including Arrest, Burglary, Robbery, Shoplifting, and Stealing. Detecting anomalies in these categories typically requires interpreting subtle object interactions and temporally distributed behaviors, and the strong performance of OHV-VAD shows

α	β	$AUC(\%)$
0.3	0.7	85.59
0.4	0.6	85.67
0.5	0.5	85.44
0.7	0.3	85.41

Table 8: Effect of the visual and semantic similarity weights α and β in Adaptive Temporal Evidence Refinement on UCF-Crime.

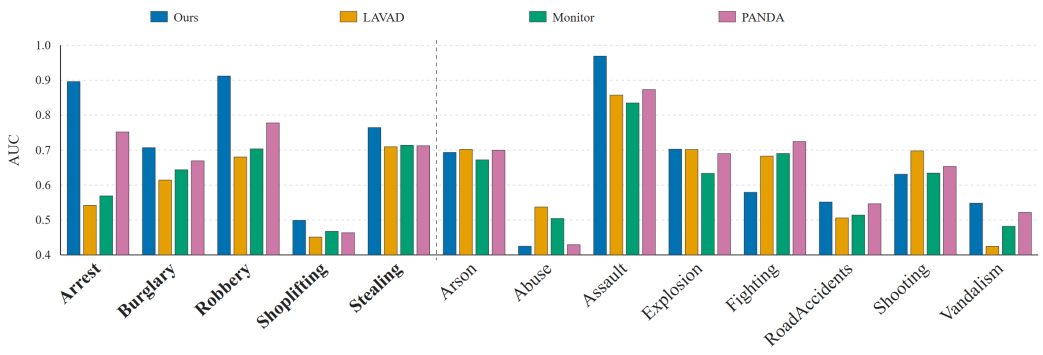

Figure 5: Category-wise AUC comparison on the UCF-Crime dataset.

Prompt style	AUC (%) ↑	ΔAUC (p.p.)
Default	**85.67**	–
Encyclopedic	84.27	-1.40
Educational	84.31	-1.36
Domain expert	83.66	-2.01
Heavy spelling noise	79.86	-5.81

Table 9: Prompt sensitivity evaluation on UCF-Crime. ΔAUC denotes the percentage-point change relative to the default prompt.

that it can effectively handle such evidence-demanding scenarios. Moreover, OHV-VAD achieves the highest AUC in 9 of the 13 categories, indicating that its improvements on evidence-demanding anomalies do not come at the expense of performance on visually salient anomalies.

Prompt sensitivity evaluation. Following the prompt-sensitivity setting of prior work, we construct four variants of the default prompt while preserving its task objective and output requirements. The Encyclopedic prompt expresses the instructions in an impersonal, factual, and definition-oriented style resembling an encyclopedia entry. The Educational prompt adopts an educator-like and explanatory tone, providing clear guidance on the expected analysis. The Domain Expert prompt uses professional terminology and a specialized analytical tone characteristic of an expert in video anomaly detection. The Heavy Spelling Noise prompt is derived by introducing extensive typographical and grammatical corruption into the default prompt, including character insertions, deletions, substitutions, and transpositions, misspelled words, incorrect word forms, malformed sentence structures, agreement errors, and missing or misplaced punctuation. All other model and inference settings are kept unchanged.

As shown in Table 9, the encyclopedic, educational, and domain-expert prompts cause only limited performance degradation compared with the default prompt. These variants modify the tone and linguistic style while retaining the underlying semantic intent, and their relatively consistent performance indicates that OHV-VAD does not rely heavily on a particular prompt formulation. In contrast, the heavy-spelling-noise prompt leads to a more noticeable decrease in AUC. Unlike natural stylistic variations, extensive linguistic corruption can obscure important semantic cues and task constraints, thereby making the instructions more difficult for the model to interpret reliably. These results distinguish sensitivity to ordinary stylistic reformulation from sensitivity to severe linguistic corruption: OHV-VAD remains robust to semantically consistent and well-formed prompt variations, although its performance can degrade when prompt readability is substantially compromised.

C More details of OHV-VAD

Evidence probes and normality priors. Figure 6 presents the complete prompt template used by OHV-VAD to construct the evidence-probe bank and normality priors. For evidence probes, given

anomaly categories or descriptions specified by the user, OHV-VAD automatically derives their relevant definitions and characteristics, denoted by $\mathcal{K}_A$. Based on $\mathcal{K}_A$, it then generates anomaly-specific diagnostic questions that serve as evidence probes. Figure 7 presents several representative examples generated by OHV-VAD. In our experiments, we provide OHV-VAD with the anomaly categories defined in UCF-Crime Sultani et al. (2018) or XD-Violence Wu et al. (2020). In practical applications, evidence probes can be flexibly generated for different scenarios according to user requirements. For normality priors, based on the scene information contained in the provided video description, OHV-VAD constructs normal-behavior prototypes for the current context and uses them to determine whether ambiguous behaviors are consistent with normal activities in that scene. Figure 8 presents representative normality priors generated by OHV-VAD.

Evidence probe selection policy Π_{select}**.** As shown in Figure 9, OHV-VAD uses observations from the recent past to automatically select probes from the evidence-probe bank for testing hypotheses about anomalies that may subsequently occur, thereby directing the model's attention toward subtle yet critical anomaly-related details. During the experiments, OHV-VAD selects an average of five highly relevant evidence probes for each segment. For clarity, only the top three selected evidence probes are shown in the case study.

Targeted visual observation $\mathcal{P}_{\text{vis}}$**.** In Figure 10 provides the detailed prompt of Targeted visual observation $\mathcal{P}_{\text{vis}}$. Through $\mathcal{P}_{\text{vis}}$, OHV-VAD uses evidence probes to selectively attend to anomaly-related information that is most relevant to the current video segment, thereby enabling it to uncover more important anomalous features and extract a relevant clue. It should be noted that the hypothesis is not directly treated as a clue. Instead, it is used to guide the model regarding which perspective to adopt when searching for evidence. All evidence probes are answered based on the actual visual content of the corresponding video segment. Therefore, if no visual information is available to answer a particular question, no anomaly-related clue corresponding to that question will be generated.

Reflective Anomaly-Normality Verification. Figure 11 shows the detailed prompt in the section Reflective Anomaly-Normality Verification of the main paper. As illustrated, OHV-VAD first uses the evidence accumulated up to current segment to identify the potential anomalous event in the current video segment and explain it using the available evidence ($\mathcal{R}_A$). If a potential anomaly is supported by relevant evidence, the subsequent comprehensive analysis ($\mathcal{V}_{\text{ref}}$) combines this anomaly-related evidence with the video content to assess the suspicious behavior, resulting in a higher anomaly-confidence score. Meanwhile, OHV-VAD evaluates the observed behavior from the perspective of normality ($\mathcal{R}_N$) by combining the current observations with preceding contextual information. Specifically, it determines whether the behavior could plausibly occur under normal circumstances, that is, whether it is consistent with the normal priors of the current scene. OHV-VAD then assigns a normality score, where a higher score indicates that the observed behavior is more likely to be normal. During the final integrated assessment, a high normality score leads OHV-VAD to interpret the suspicious behavior according to the scene-specific normal priors, thereby producing a lower anomaly-confidence score. Therefore, each assessment jointly considers whether anomaly-related evidence support an anomalous interpretation and whether the observed behavior is consistent with what is normally expected in the current scene. By consistently grounding its interpretations in evidence, OHV-VAD not only improves the accuracy of anomaly judgments but also makes the resulting analysis more reliable.

D Additional Case Study

Case study of a visually salient anomaly. As shown in Figure 12, this visually salient anomaly case demonstrates that OHV-VAD uses evidence probes to actively uncover anomaly-relevant evidence without being biased toward anomalous conclusions. The evidence probes and hypotheses serve only to direct the model's attention to different visual details and prevent critical information from being overlooked; they are not themselves treated as evidence. Instead, the evidence reported by OHV-VAD remains grounded in the actual visual content. In this case, in the preceding normal segments, OHV-VAD does not report anomalous evidence merely because anomaly-related hypotheses or questions are present. It identifies the anomaly only when the traffic accident actually occurs. This case demonstrates that OHV-VAD can effectively handle evidence-demanding anomalies without compromising its performance on visually salient anomalies.

[Diagnostic Question Base (Evidence probes) Prompt]
There are 13 different anomaly classes/information that the user is interested in: <Anomaly classes/information provided by the user>. Please give the key characteristic needed to judge an anomalous event as this type of anomaly. And based on this key characteristic, summarize at least 5 related questions to prompt the multimodal large model to discover relevant anomaly clues in a video. There is some information for reference. Please answer strictly according to the following format.
Key characteristics:
  Anomaly Class:
  KeyChar 1:
  KeyChar 2:
  ...
  KeyChar n:
Questions of different Anomaly Class:
  Anomaly Class:
  Question 1:
  Question 2:
  ...
  Question n:

[Normal Prior Prompt]
You are an agent responsible for extracting priors of normal events. Your task is to extract the normal event priors from the video descriptions I provide. Please make sure to preserve the key information from the original video descriptions as much as possible (such as actions, environment, etc.).
Please answer strictly according to the following format.
Video description: "<Video description>"
Extracted Normal Prototype: "<Normal prototype you extracted>".

Figure 6: Prompt of constructing the evidence-probes bank and normality priors.

Case study of ambiguous normal behavior. As shown in Figure 13, this case presents a normal behavior in an ordinary scene that can nevertheless be mistaken for an anomaly. Two motorcycles and one person are visible, and the person is touching one of the motorcycles, making it unclear from the current segment alone whether the behavior is anomalous. By combining the preceding temporal context with plausible normal behavior patterns in the current scene, OHV-VAD observes that the person's actions remain continuous and consistent over time and are compatible with normal activity. Although the behavior could be interpreted as suspicious from an anomaly perspective, the normal explanation is more plausible in this context. This case demonstrates that contrasting and reflecting on both abnormal and normal verification helps OHV-VAD mitigate false alarms arising from ambiguous behaviors.

Case study of evidence-demanding anomaly. As shown in Figure 14, this example presents an evidence-demanding case. In the current video segment, the visual observation consists only of a static scene with an open door. Methods that do not accumulate temporal evidence could easily classify this segment as normal; indeed, even the ground-truth annotation labels it as normal. However, the evidence collected throughout the video reveals that a person previously attempted to open a nearby window, eventually entered the house, and opened the door. The target segment therefore depicts the exterior after the intrusion, while the person remains inside, indicating that the anomalous event is still ongoing. By actively seeking and accumulating evidence throughout the video, OHV-VAD can associate this visually inconspicuous segment with the preceding intrusion and classify it as anomalous based on the historical evidence, despite its normal ground-truth label.

Failure case under poor video quality and occlusion. Figure 15 presents a failure case of OHV-VAD. In the ground-truth anomalous segment, a person approaches a gate and appears to strike a dog with a stick near a mailbox. However, due to poor video quality, both the striking motion and its target are largely occluded by the mailbox, making the person's actual action difficult to discern. In the next video segment, OHV-VAD flags the next segment as anomalous because the person rapidly leaves the residential property. Thus, improving the robustness to low-quality videos and severe occlusion remains an important direction for future work.

[Evidence Probes]

Abuse

1. Are children, elderly individuals, disabled persons, or animals being physically struck, kicked, restrained, or harmed?
2. Does the force used seem unnecessarily severe (e.g., continued hits on a restrained target)?
3. Is the victim clearly distressed (e.g., crying, injured, or struggling to escape)?
4. Is the violence occurring in a context where peaceful or nurturing interactions are expected (e.g., a home, school, or hospital)?
5. Are actions taken without consent (e.g., forcing restraint) or without legitimate medical/protective justification?

Arrest

1. Does the video show law enforcement using sudden physical force (grabbing, pinning, or dragging) against individuals?
2. Are multiple officers working together to surround or encircle a suspect?
3. Is civilian resistance (e.g., fleeing, arguing) met with escalated force disproportionately?
4. Are police vehicles blocking roads or moving erratically in non-emergency scenarios?
5. Are individuals abruptly removed from the scene (taken to cars, pulled away)?
6. Does police activity disrupt normal public movement (e.g., traffic halted, crowds stopped)?

…

…

Vandalism

1. Are individuals using non-typical objects as tools (e.g., stones, bottles, spray cans) to damage property?
2. Are actions clearly mismatched with expected behaviors (e.g., pedestrians suddenly kicking trash cans, smashing windows)?
3. Is property damage visible (cracks, spills, scratches, graffiti) with signs of intentional causation?
4. Are multiple people collaborating on destructive acts (e.g., synchronized throwing items, attacking objects together)?
5. Are individuals displaying evasive behavior (frequent glances, rapid exits) after causing damage?
6. Is public/private property (cars, doors, cameras) being targeted in a deliberate manner instead of personal items?
7. Are objects used in non-functional ways (e.g., spraying on walls, buckets tipped intentionally)?
8. Is the behavior repetitive and malicious (e.g., continuous hitting of an object, revisiting sites for damage)?

Figure 7: An example of generated evidence probes.

[Normality Priors]

Individuals perform routine actions like walking, talking, working, handling objects, shopping, or mundane tasks.
Ordinary interactions include conversations, group movements, handing objects, and social exchanges between people.
…
People perform routine tasks such as conversing, writing, using calculators, and handling items like pens, glasses, helmets, or other pile of ordinary items with no signs of abnormal or disruptive behavior.
Individuals perform routine and purposeful activities near the counter, such as organizing items, handing over objects, recording information, or communicating, with natural and context-appropriate behavior
The environment is quiet, and everyone is focusing on their own tasks.
Vehicles should be interacted with in a routine way, and no suspicious individuals or objects are accessed or contacted.
…
Events follow logical sequences like entering/exiting spaces or performing tasks.
No interaction with sensitive items, such as safes or cash registers.
Multiple people conduct separate activities simultaneously, remain normal and expected.
Multiple individuals engage in separate normal activities simultaneously, with people and vehicles moving in expected patterns (e.g., walking, driving in correct directions).
Interactions between individuals are typical (conversations, transactions, collaborations), aligned with their roles or environment, and involve logical exchanges (handing objects, operating computers).
…
Normal events do not include prolonged continuous interaction with doors or windows especially in a dimly lit environment.
Under normal circumstances, there is no abnormal interaction with the gate or attempt to open it.
The individual's behavior is under surveillance, and the scene is equipped with monitoring equipment.

Figure 8: An example of generated normality priors.

You are specialized in predicting subsequent event developments. Your predictions about future events must adhere to logical consistency, temporal coherence of the video, and commonsense plausibility.
Your tasks:
1. Based on the observations of multiple consecutive video clips predict the trajectory of future events.
2. Select relevant questions from the predefined Question Base that can validate your predictions. These questions will later be answered based on video clips as visual evidence.
Context observations:[Insert Observation]
Below is the predefined Question Base:[Insert Question Base]

Respond using the following format. Please output in plain text only:

```
{
   "Future Event Hypothesis": "The hypothesis of future event.",
   "Selected Questions": {
      "Index": "Corresponding Questions from question base",
      "Index": "Corresponding Questions from question base",
      ...
   },
   "Reasoning": "Make a rational reason for why you make this prediction and why these questions were selected.",
}
```

Figure 9: The prompt of evidence probe selection policy.

You need to analyze a video clip. For each video clip, you will be given the visual content and one or more related questions. Please focus on any anomalies in this clip, such as unusual objects, unusual activities, or abnormal behaviors and so on.
Your task consists of two parts for each clip:
(1) Answer the provided questions accurately based on the visual information in the clip.
(2) Provide a detailed description of the video clip, focusing on the following aspects:
Object appearance: Describe key visual characteristics of the main entities (e.g., people, vehicles).
Object actions: Describe what the main entities are doing, including movements, postures, gestures, or specific behaviors.
Relationship between objects: Explain how the main entities interact with each other, if applicable.
Background environment: Describe the physical environment of the video clip. Include objects, settings (e.g., indoor/outdoor, streets, rooms), and any relevant context.
(3) Evidence Extraction:
If the origin video clip description contains any suspicious, abnormal, or unusual information, extract a single sentence that summarizes the key clue(s). If there is nothing suspicious, respond with "No suspicious clues.". The extracted anomaly information should be based primarily on the objective details in the description, with subjective information used as supplementary. Finally, provide a statement: is it support the hypothesis. (supportive, contradictory, or undetermined.)
[Insert Selected Probes and Hypothesis]

Please answer as following JSON format. Please output in plain text only, and do not use any Markdown formatting:

```
{
   "Question Answers": {
         "Question 1": "<The answer of the No.1 question>",
            …
   }
   "Video Segment Objects": {
      "Object 1": {
         "Visual feature": "<The visual feature of object 1>",
         "Action": "<The action of object 1>",
         "Relation": "<The relationships with other objects>",
      },
      ...
   },
   "Video Clip Description": "<Detailed description of the video segment>",
   "Background Environment": "<Background description>",
   "Evidence": "<Extracted evidence>"
}
```

Figure 10: The prompt of targeted visual observation.

You are an expert reasoning agent specialized in abnormal event detection from two perspectives.

Anomaly Verification
You will receive the following information:
1. The observation of the current video segment.
2. An evidence memory that contains a set of potential clues from earlier video segments(In chronological order).
3. The hypothesis that represents the possible anomalous event in the current segment, you need to verify whether there is any evidence that can prove it.
4. A list of anomaly categories plus "Normal".
Your Tasks:
1. Perform reasoning on the current video segment by integrating information from the evidence memory.
2. Evaluate the hypothesis— check whether it is consistent with the video content and context.
3. Based on your reasoning, select the classification result for the current video segment. The result can be one of the anomaly categories or "Normal".
4. Provide a clear and well-reasoned explanation for your classification.
Respond using the following format Please output in plain text only, and do not use any Markdown formatting:
Please respond in the following format:
{
"Classification": "<Anomaly Type or 'Normal' or 'Suspicious'>",
"Reference Clues": "<State the clues from the clue base that you referenced>",
"Explanation for Final Classification": "<The explanation of the final classification>",
}

Normality Verification.
I will provide you with :
1. A current video observation.
2. A series of observations of previous consecutive video segments.
3. Normality Priors.
First, please optimize all observations I provide—such as terms like "seems," "possibly," or any other uncertain expressions—and instead extract only the confirmed, objective information based on common sense. Then, based on these objective descriptions, please evaluate whether the video segments are normal from the following three aspects:
Consistency – Are the actions over this period continuous and normal?
Alignment with the Normal Prior List – Are these descriptions included in the Normal Prior List? An exact match in meaning is not required — it is sufficient if the objective descriptions generally align in meaning. Moreover, the objective descriptions are not expected to cover all key points listed in the Normal Prior List.
Potential Issues – Is this behavior likely to occur under normal circumstances?
After normality evaluation you need to provide a normality score and normality explanation based on the normality
Normality explanation:
{
"Objective description": <Remove certain subjective assumptions based on common sense>
"Normal Score": <Integer score of Normal>,
"Normality Explanation": "<The explanation of the normal score>"
}

Comprehensive analysis:
You need to perform comprehensive evaluation based on the anomaly/normality evaluation and previous clues.
1. If you are very certain that there is an anomaly in the current video segment, please perform a causal analysis of the current segment by combining it with the suspicious information previously provided in the evidence memory.
2. If the score of normality explanation is greater than or equal to 8, it indicates that the current video segment is most likely normal. Please interpret the normality of this segment further from a normal perspective, based on the analysis generated by normality evaluation.
Please respond in the following format:
{
"Final Classification": "<Anomaly Type or 'Normal' or 'Suspicious'>",
"Reference Clues": "<State the clues from the clue base that you referenced>",
"Explanation for Final Classification": "<The explanation of the final classification>",
}

Figure 11: The prompt of Reflective Anomaly-Normality Verification.

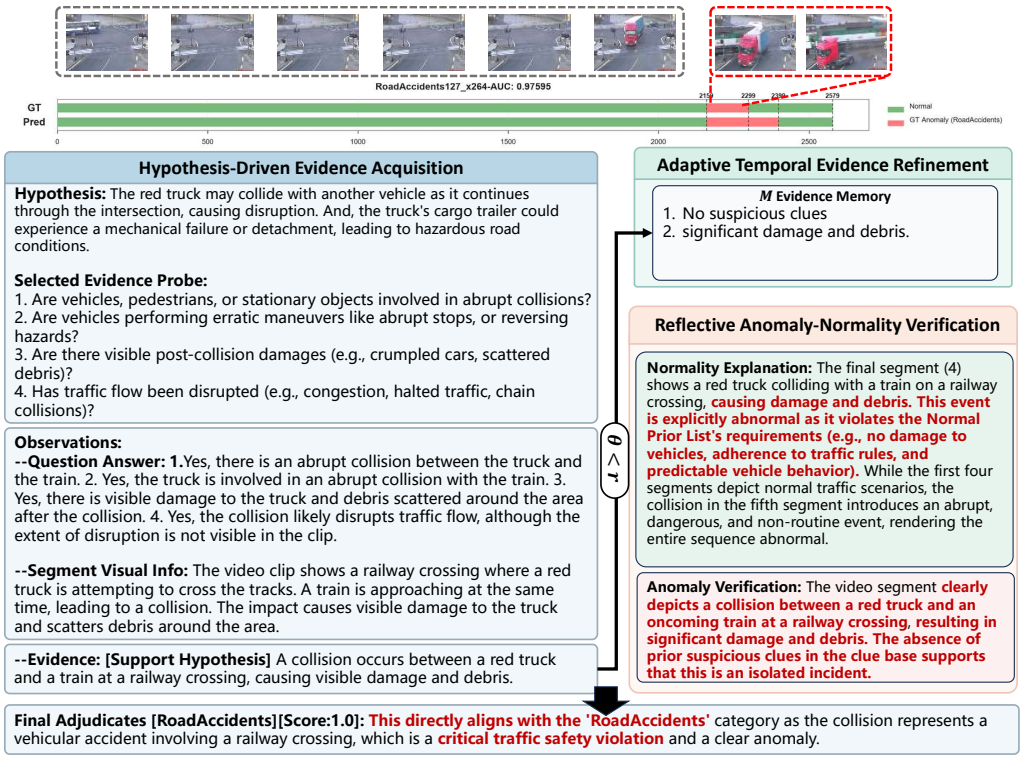

Figure 12: A visually salient anomaly RoadAccidents (RoadAccidents127_x264) case in UCF-Crime dataset. The red circle indicates the current video segment, while text shown in red and bold highlights key information.

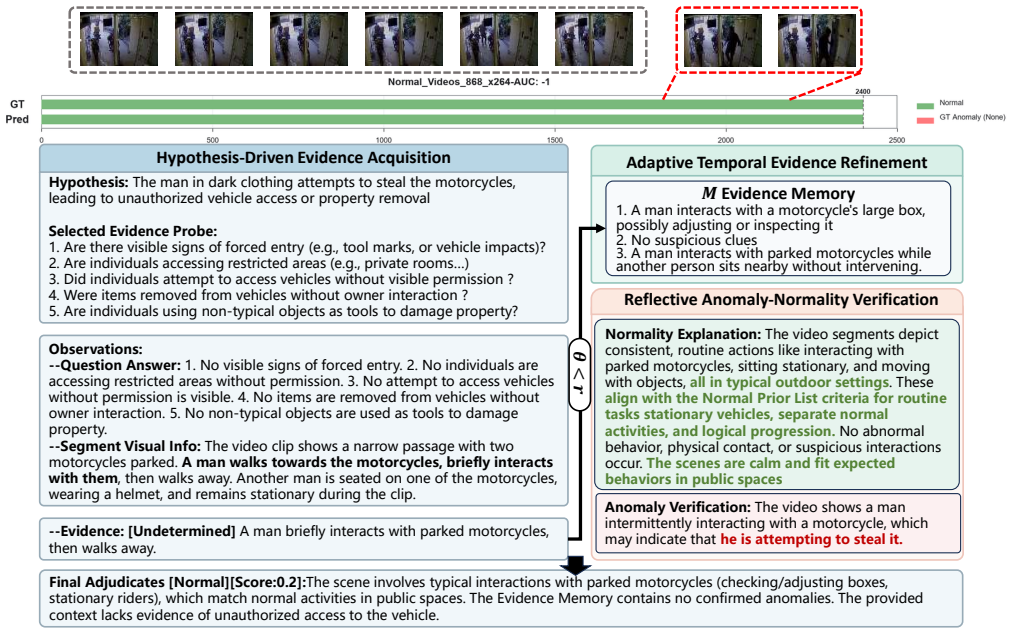

Figure 13: Case of normal video "Normal_Videos_868_x264" in UCF-Crime dataset. The red circle indicates the current video segment, while text shown in red and bold highlights key information.

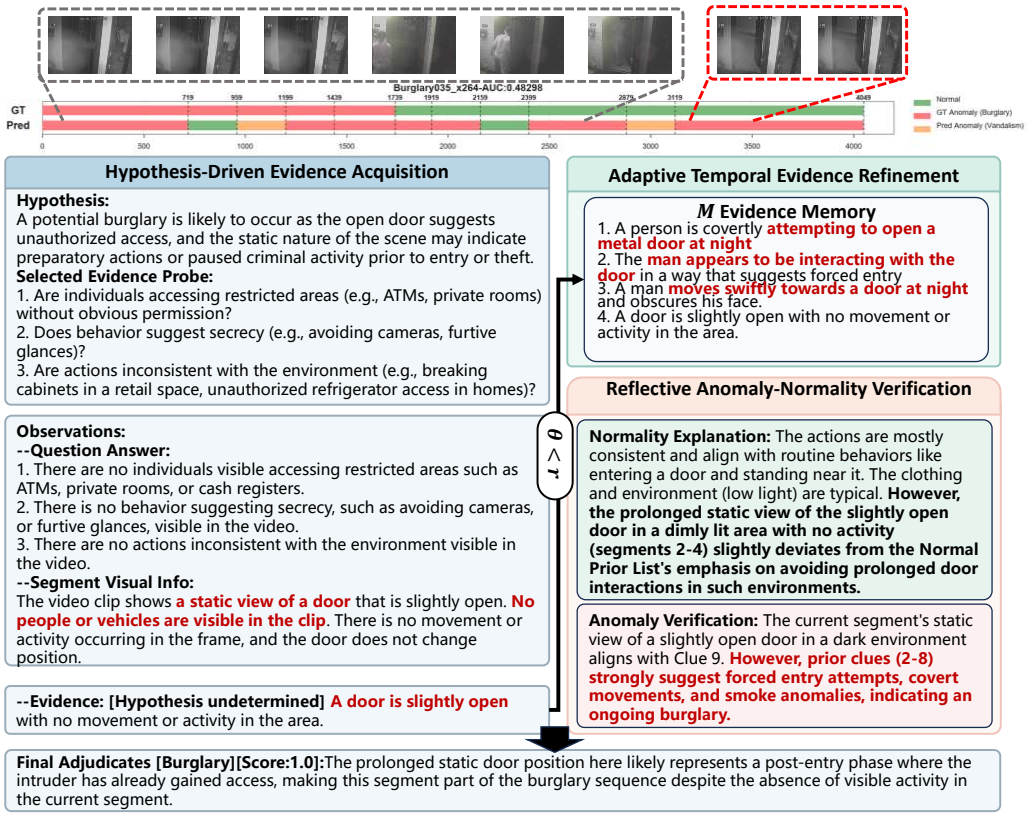

Figure 14: Case of anomaly video "Burglary035_x264" in UCF-Crime dataset. The red circle indicates the current video segment, while text shown in red and bold highlights key information.

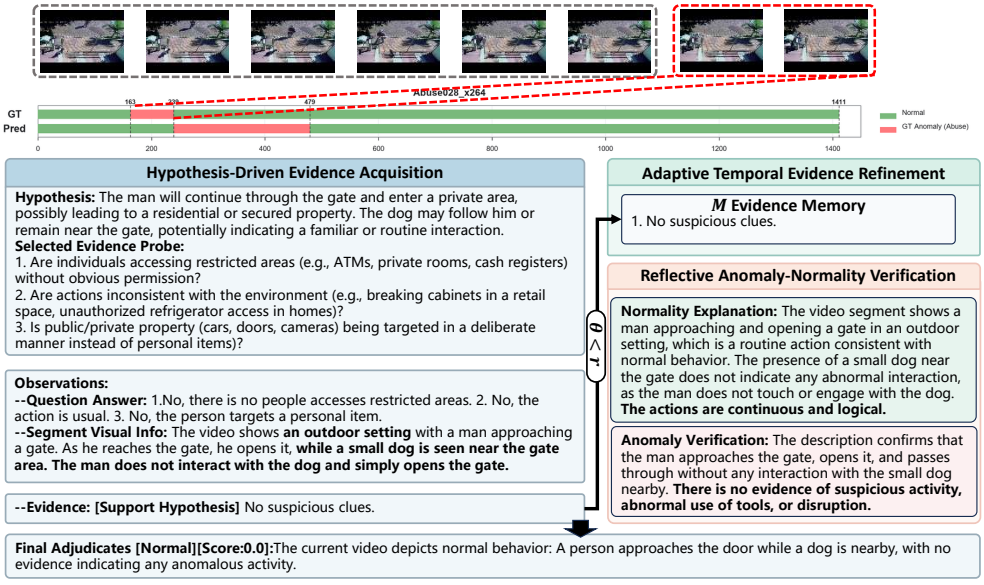

Figure 15: Case of anomaly video "Abuse028_x264" in UCF-Crime dataset. The red circle indicates the current video segment, while text shown in red and bold highlights key information.