NEXUS: NEURO-SYMBOLIC EVENT REASONING FOR TRAINING-FREE VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

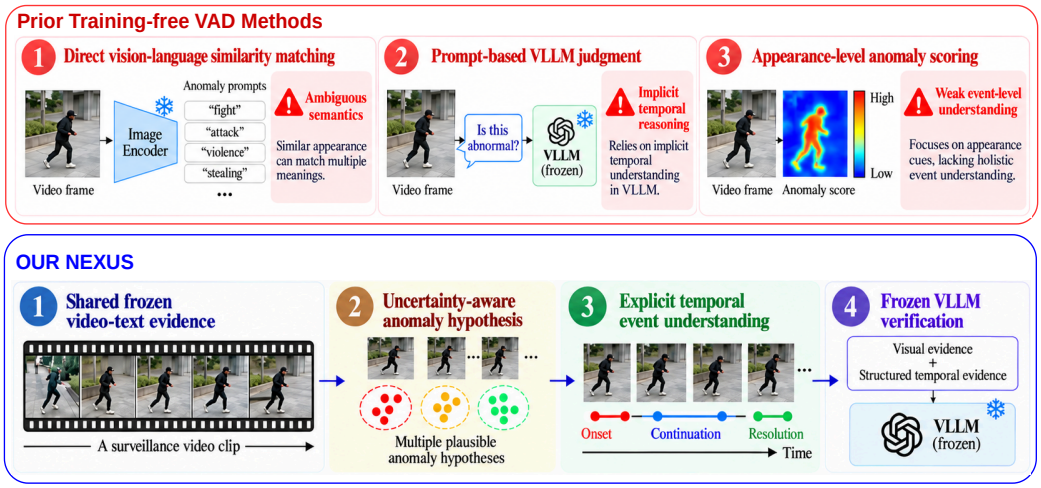

Figure 1: **Prior training-free VAD vs. our NEXUS:** While prior methods largely rely on similarity matching, appearance-level scoring, or implicit VLLM reasoning, NEXUS explicitly models semantic uncertainty and temporal event evolution before structured VLLM verification.

ABSTRACT

Training-free video anomaly detection is attractive for real-world surveillance, where collecting representative training data is often constrained by privacy, data-governance requirements, and the inherent rarity of abnormal events. However, existing training-free approaches remain largely driven by *visual affinity*, either through direct vision-language similarity or by prompting multimodal models to infer anomaly decisions from appearance-level evidence. We argue that this view is incomplete: *real-world complex anomalies are not only isolated appearances, but also temporally evolving events* whose interpretation depends on how semantic states emerge, persist, change, and resolve over time. To this end, we introduce **Neuro-Symbolic Event eXploration with Unified Spatio-temporal VLLM Verification (NEXUS)**, a novel training-free framework for video anomaly detection. At its core, NEXUS introduces *Neuro-symbolic Temporal Event Reasoning (N-TER)*, which transforms frozen vision-language evidence into five continuous event states and composes them through interpretable temporal chains capturing event onset, reversal, persistence, consequence, resolution, and prior history. Complementing this temporal reasoning, *Uncertainty-guided Proposal Scoring (UPS)* explicitly enforces an anomaly hypothesis to outperform both uncertain and normal alternatives, preventing visual similarity alone from becoming anomaly evidence. Then, *Reasoning-Enabled Score Refinement (RESR)* couples UPS proposals with N-TER consistency and contradiction, allowing temporal reasoning to refine visual hypotheses and also minimizing anomaly hallucination. A frozen VLLM Judge then performs the final spatio-temporal verification with the symbolic evidence. We evaluate NEXUS across five real-world challenging video anomaly detection benchmarks and demonstrate strong performance gain against state-of-the-art methods.

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.

1 INTRODUCTION

Video anomaly detection (VAD) is a fundamental component of intelligent video surveillance systems with applications spanning public-space monitoring, transportation, industrial safety, healthcare, and human-centered environments. Existing approaches Liu & Ma (2019), Majhi et al. (2025), Wang & Cherian (2019) have been extensively studied under *supervised*, *weakly supervised*, and *unsupervised* settings, where models rely on dense temporal annotations, video-level labels, or normal training videos, respectively. Although these paradigms D'Amicantonio et al. (2025), Zaheer et al. (2022) have driven substantial progress, their dependence on target-domain data limits their scalability to real-world deployments. Surveillance videos often contain identifiable individuals, sensitive activities, and restricted operational information, making large-scale collection, storage, sharing, and annotation increasingly difficult under privacy and data-governance constraints. Moreover, abnormal events are inherently sparse and long-tailed: normal activities are abundant, whereas representative examples of accidents, aggression, theft, medical distress, safety violations, or previously unseen abnormal interactions are difficult to collect exhaustively. These limitations motivate **training-free VAD**, where semantic knowledge already encoded in pretrained foundation models is directly exploited without target-domain training or adaptation.

Despite the strong visual and semantic priors offered by modern VLLMs (or MLLMs), training-free anomaly detection remains fundamentally challenging for two reasons. **(A) Visual similarity does not necessarily imply anomalous behavior:** Running may indicate escape or ordinary exercise, close physical interaction may correspond to violence or assistance, and a person lying on the ground may indicate distress or simply resting. A reliable system must therefore distinguish an *anomaly hypothesis* not only from normality, but also from visually plausible yet *uncertain* interpretations. **(B) Real-world anomalies are temporal events rather than isolated appearances:** They may emerge gradually, persist, reverse, produce consequences, or transition toward resolution, making clip-level semantic matching insufficient for understanding their full evolution. These challenges suggest that training-free anomaly detection is much more than asking a VLLM whether a video is abnormal. A practical system must reason about *what the observation could mean*, *how the event evolves*, and *whether the evolving evidence provides a coherent explanation of the observed anomaly*. As illustrated in Fig. 1, NEXUS moves beyond appearance-driven anomaly decisions by explicitly combining uncertainty-aware semantic hypotheses with temporal event understanding and structured VLLM verification.

Recent training-free VAD methods Zanella et al. (2024), Shao et al. (2025), Feng et al. (2026), Yang et al. (2024) have increasingly exploited frozen vision-language features, caption-based reasoning, LLM-generated rules, multimodal prompting, and agentic reasoning. Nevertheless, much of their temporal understanding remains implicit inside the language model. Local similarity alone provides no explicit mechanism for determining whether an anomaly explanation remains preferable to an uncertain or normal alternative; likewise, temporal ordering alone does not constitute anomaly evidence, since a sequence becomes meaningful only when its chronology is grounded in corresponding semantic event states. As a result, it remains difficult to explicitly determine whether a candidate event is emerging, persisting, reversing, producing a consequence, or resolving. This raises a natural question: **can frozen vision-language models be equipped with an explicit, interpretable event-reasoning mechanism for reliable anomaly detection without target-domain training?**

To address this question, we introduce **Neuro-Symbolic Event eXploration with Unified Spatio-temporal VLLM Verification (NEXUS)**, a unified training-free framework for video anomaly detection. At the core of NEXUS is **Neuro-symbolic Temporal Event Reasoning (N-TER)**, which converts frozen visual–language evidence into continuous event-state representations, corresponding to *anomaly (A)*, *uncertainty (U)*, *onset (O)*, *continuation (C)*, and *resolution (R)*. These states are organized over a deterministic temporal graph and composed through interpretable fuzzy chains that explicitly model normal-to-onset transitions, event reversal, persistence, ordered consequence, resolution, and bounded SINCE relations. Complementing this temporal reasoning, **Uncertainty-guided Proposal Scoring (UPS)** first asks whether an anomaly interpretation is stronger than both an uncertain alternative and generic normality, preventing visual similarity alone from becoming anomaly evidence. The resulting semantic proposal and N-TER reasoning are then coupled through **Reasoning-Enabled Score Refinement (RESR)**, which allows temporal consistency and contradiction to refine supported anomaly hypotheses while preserving semantically rejected proposals.

NEXUS finally performs anomaly verification through a shared frozen VLLM Judge. For **offline anomaly detection**, the Judge observes a visual episode while N-TER reasons over a wider bidi-

rectional temporal neighborhood and produces structured temporal evidence for the target interval. The Judge then combines this visual evidence with the UPS proposal, N-TER event reasoning, and RESR-refined score to produce the final anomaly posterior. Since real-world anomalous events can occur over substantially different temporal durations, NEXUS further performs parameter-free multi-scale inference and combines center-aligned predictions across fixed temporal scales. Extensive evaluation across several challenging real-wrold VAD datasets demonstrates the efficacy of NEXUS over recent state-of-the-art methods on real-world scenarios. Our main contributions are:

- We introduce **NEXUS**, a novel method that moves beyond appearance-driven anomaly matching by explicit semantic competition and neuro-symbolic temporal event reasoning.
- We propose **N-TER**, the core temporal reasoning module of NEXUS, which represents each observation with continuous event states $[A, U, O, C, R]$ and composes them through interpretable fuzzy chains capturing anomaly *onset, reversal, persistence, consequence, resolution, and event history*.
- We introduce **UPS** and **RESR** for conservative anomaly reasoning: UPS requires an anomaly hypothesis to outperform uncertain and normal alternatives, while RESR refines only visually supported hypotheses using temporal consistency and contradiction, preventing temporal reasoning from independently creating anomaly evidence.
- We extensively evaluate NEXUS on **five challenging real-world VAD datasets**, outperforming several recent state-of-the-art methods.

2 Related Work

Training-Dependent Video Anomaly Detection Early VAD methods predominantly learn anomaly-discriminative representations from target-domain data. In the weakly-supervised setting, video-level labels are commonly exploited through multiple-instance learning (MIL), beginning with the ranking formulation of Sultani *et al.* Sultani et al. (2018) and subsequently improved through feature-magnitude learning Tian et al. (2021), magnitude-aware feature amplification Chen et al. (2023), unbiased MIL Lv et al. (2023), and category-specific expert modeling D'Amicantonio et al. (2025); Majhi et al. (2025). Although effective, these methods fundamentally depend on labeled normal and abnormal videos and therefore inherit the cost and domain specificity of target-data collection. More recent VLLM-based approaches reduce the amount of parameter adaptation by keeping most of the foundation model frozen and instead probing or manipulating its internal representations. HiProbe-VAD Cai et al. (2025) exploits hidden VLLM states, while SteerVAD Cai et al. (2026a) steers anomaly-sensitive latent representations and HeadHunt-VAD Cai et al. (2026b) identifies robust anomaly-sensitive attention heads. Such approaches substantially reduce full-model optimization, but still rely on target-domain reference data or auxiliary anomaly-specific components. At the other end of the spectrum, recent video anomaly understanding (VAU) frameworks explicitly train VLLMs with large-scale instruction and reasoning supervision. HAWK Tang et al. (2024) incorporates motion-language supervision for open-world anomaly understanding, Holmes-VAU Zhang et al. (2025) constructs hierarchical long-video anomaly annotations, and Vad-R1 Huang et al. (2026a) and VAU-R1 Zhu et al. (2025) further introduce structured chain-of-thought supervision and reinforcement fine-tuning for anomaly reasoning. While these paradigms improve semantic interpretation and reasoning capability, their dependence on curated annotations, supervised fine-tuning, or reinforcement optimization limits deployment when surveillance data cannot be freely collected or shared.

Training-Free Video Anomaly Detection Training-free VAD instead seeks to exploit knowledge already encoded in frozen foundation models without optimizing on the target surveillance domain. LAVAD Zanella et al. (2024) pioneered this direction by converting video observations into textual descriptions and using LLM prompting for temporal aggregation and anomaly scoring. Subsequent work has explored complementary forms of test-time reasoning: AnomalyRuler Yang et al. (2024) induces explicit normality rules from reference examples; HoloTrace Wang et al. (2025a) introduces causal knowledge-graph reasoning; VADTree Li et al. (2026) organizes observations through hierarchical event granularity; and the unified reasoning framework of Lin *et al.* Lin et al. (2026) chains temporal detection, spatial localization, and semantic explanation. More recent methods extend this paradigm through agentic reasoning and open-set inference, including PANDA Yang et al. (2026), QVAD Bekit et al. (2026), ESOM Liu et al. (2026), and GridVAD Eltahir et al. (2026), while

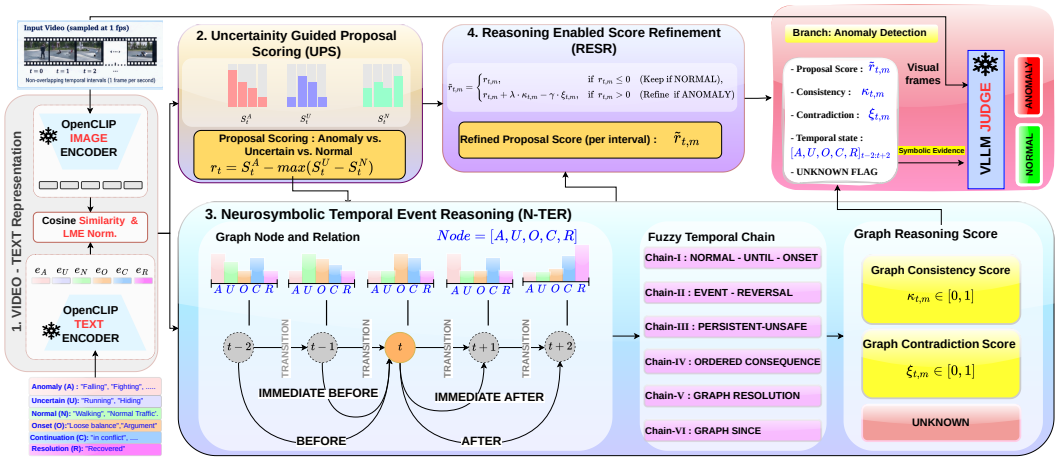

Figure 2: **Overview of NEXUS.** Shared frozen video–text features drive **UPS** for uncertainty-aware anomaly proposals and **N-TER** for explicit temporal event reasoning. **RESR** refines supported proposals using temporal consistency and contradiction, followed by frozen VLLM verification for video anomaly detection.

SphereVAD Huang et al. (2026b) and PRISM Chen et al. investigate latent geometric and intrinsic statistical structures rather than relying exclusively on generated language. Despite this progress, existing training-free approaches still exhibit two recurring limitations. First, **visual similarity is often treated as anomaly evidence without explicitly modeling competing ambiguous interpretations**; second, **temporal reasoning is predominantly implicit in prompting, latent aggregation, or agentic reasoning rather than represented as an explicit event evolution**. NEXUS addresses these limitations by coupling uncertainty-aware proposal scoring with explicit neuro-symbolic temporal event reasoning and reasoning-enabled score refinement, while keeping the visual encoder, language model, ontology, and temporal reasoning process entirely training-free.

3 Proposed Method: NEXUS

We introduce **NEXUS** (as shown in Fig. 2), a training-free framework for video anomaly detection built upon a shared semantic–temporal reasoning backbone. NEXUS combines **Uncertainty-guided Proposal Scoring (UPS)**, **Neuro-symbolic Temporal Event Reasoning (N-TER)**, and **Reasoning-Enabled Score Refinement (RESR)** over frozen video–text representations, followed by a frozen VLLM for final spatio-temporal verification. UPS identifies visually supported anomaly hypotheses, N-TER examines whether these hypotheses form coherent temporal events, and RESR refines the supported proposals using temporal consistency and contradiction evidence.

3.1 Video–Text Representation

Given a video $\mathcal{V}$, we sample frames at 1 fps and partition it into non-overlapping intervals $\mathcal{V} = \{x_1, \ldots, x_T\}$ with base duration $\Delta = 2$ seconds, where $x_t = \{I_{t,1}, \ldots, I_{t,K_t}\}$. We pair the video with a fixed public-anomaly ontology $\mathcal{M} = \{1, \ldots, M\}$. For each event mechanism m, the ontology contains six description banks $\{\mathcal{A}_m, \mathcal{U}_m, \mathcal{N}, \mathcal{O}_m, \mathcal{C}_m, \mathcal{R}_m\}$, corresponding to *anomaly*, *uncertainty*, *normality*, *onset*, *continuation*, and *resolution*. For example, $\mathcal{A}_m$ may describe "a person is escaping", whereas $\mathcal{U}_m$ contains visually plausible but ambiguous alternatives such as "a person is running". The ontology is fixed before inference and shared across datasets and temporal scales. Video intervals and ontology descriptions are embedded into a common frozen OpenCLIP Ilharco et al. (2021) space. Each frame is encoded as $\mathbf{v}_{t,k} = f_{\text{img}}(I_{t,k})$, and interval features are obtained by normalized temporal aggregation, $\mathbf{z}_t = \text{Norm}\Big(\frac{1}{K_t}\sum_{k=1}^{K_t} \text{Norm}(\mathbf{v}_{t,k})\Big)$, where $\text{Norm}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$. Each textual description p is encoded once as $\mathbf{e}_p = \text{Norm}(f_{\text{text}}(p))$. Their compatibility is measured by cosine similarity, $c_{t,p} = \mathbf{z}_t^\top \mathbf{e}_p$. Because each semantic state is represented by multiple descriptions, we aggregate prompt-level similarities with log-mean-exp, $\text{LME}(a_1, \ldots, a_n) = \log\Big(\frac{1}{n}\sum_{j=1}^{n} \exp(a_j)\Big)$, yielding bank-level supports $s^A_{t,m}$, $s^U_{t,m}$, s^N_t, $s^O_{t,m}$, $s^C_{t,m}$, and $s^R_{t,m}$. These shared supports form the common evidence interface of NEXUS: **UPS** uses

them for uncertainty-aware anomaly proposal scoring, while **N-TER** converts them into explicit temporal event states and relations.

3.2 **UPS**: UNCERTAINTY-GUIDED PROPOSAL SCORING

The objective of UPS is to disambiguate uncertain scenarios where many zero-shot anomaly detection methods fail: a video may resemble an anomaly description without actually providing sufficient evidence for an anomalous interpretation. UPS therefore treats **uncertainty as an explicit competing explanation**, rather than assuming that every non-normal visual match should be interpreted as an anomaly. For interval x_t and mechanism m, UPS compares anomaly support $s^A_{t,m}$ with two competing explanations: semantic uncertainty $s^U_{t,m}$ and generic normality s^N_t. We define the raw anomaly proposal margin as $r_{t,m} = s^A_{t,m} - \max(s^U_{t,m}, s^N_t)$. A positive proposal $r_{t,m} > 0$ indicates that mechanism m explains the observation better than its uncertain alternative or generic normal activity. In contrast, $r_{t,m} \leq 0$ means that the anomaly hypothesis does not survive semantic competition and should not be promoted merely because it exhibits visual affinity to an anomaly description.

UPS therefore produces the mechanism-specific proposal vector $\mathbf{r}_t = [r_{t,1}, r_{t,2}, \ldots, r_{t,M}]$. Importantly, $\mathbf{r}_t$ is not the final anomaly prediction. Rather, it represents a set of **visually grounded candidate event hypotheses** whose temporal credibility is subsequently examined by N-TER.

3.3 **N-TER**: NEURO-SYMBOLIC TEMPORAL EVENT REASONING

A positive semantic proposal does not by itself establish an anomaly. The same visual state may have very different interpretations depending on what happened before it, whether it persists, and how it subsequently resolves. We therefore introduce **N-TER**, the central temporal reasoning component of NEXUS. N-TER transforms frozen vision–language similarities into an interpretable event graph and explicitly tests whether the observed semantic evidence forms a coherent temporal event. It constructs a deterministic temporal graph whose nodes contain continuous semantic-state confidences and whose edges encode known chronological relations. Fuzzy temporal chains are then evaluated over these grounded nodes to determine whether an anomaly hypothesis exhibits a meaningful onset, reversal, persistence, consequence, or resolution.

3.3.1 CONTINUOUS EVENT-STATE REPRESENTATION

For every graph node x_i and mechanism m, N-TER derives five event-state confidences from the shared OpenCLIP supports. Let α denote the frozen OpenCLIP contrastive scale and $\sigma(\cdot)$ the sigmoid function. The anomaly and uncertainty states are defined as $A_{i,m} = \sigma\big(\alpha[s^A_{i,m} - \max(s^U_{i,m}, s^N_i)]\big)$ and $U_{i,m} = \sigma\big(\alpha[s^U_{i,m} - \max(s^A_{i,m}, s^N_i)]\big)$, respectively. The onset, continuation, and resolution states are computed relative to the normal reference: $O_{i,m} = \sigma\big(\alpha[s^O_{i,m} - s^N_i]\big)$, $C_{i,m} = \sigma\big(\alpha[s^C_{i,m} - s^N_i]\big)$, and $R_{i,m} = \sigma\big(\alpha[s^R_{i,m} - s^N_i]\big)$. Every temporal node is therefore represented by $\mathbf{q}_{i,m} = [A_{i,m}, U_{i,m}, O_{i,m}, C_{i,m}, R_{i,m}]$. These are **continuous confidence values in** $[0, 1]$, rather than binary symbolic labels. For example, $\mathbf{q}_{i,m} = [0.84, 0.17, 0.73, 0.41, 0.06]$ describes an interval with strong anomaly and onset evidence, moderate continuation evidence, and little uncertainty or resolution evidence.

This yields a compact five-state event representation while preserving the distinction between *ambiguous* and *normal* behavior. For temporal chains requiring explicit normal evidence, we define an auxiliary normal confidence $N_{i,m} = \sigma\big(\alpha[s^N_i - \max(s^A_{i,m}, s^U_{i,m})]\big)$. Thus, $\mathbf{q}_{i,m}$ remains exactly $[A, U, O, C, R]$, while $N_{i,m}$ acts only as a normal-reference signal where the temporal semantics explicitly require it.

3.3.2 DETERMINISTIC TEMPORAL RELATIONS

For temporal reasoning, N-TER constructs the five-node neighborhood $\mathcal{G}_t = \{x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}\}$. The nodes are connected according to their known chronological locations. Because temporal order is observed directly from the video, these relations are deterministic rather than learned. Specifically, $\textsc{Before}(x_i, x_j) \Leftrightarrow i < j$, $\textsc{After}(x_i, x_j) \Leftrightarrow i > j$, and $\textsc{Immediate-Before}(x_i, x_j) \Leftrightarrow j = i + 1$. The inverse relation defines IMMEDIATE-AFTER, while TRANSITION indicates an adjacent connection whose semantic states may describe a change of event phase.

Importantly, **chronology alone is not anomaly evidence**. The anomaly interpretation emerges only when these relations are grounded by the continuous event-state evidence carried by their nodes.

Thus, temporal relations determine *where* N-TER reasons, while $[A, U, O, C, R]$ determines *what* evidence is available at those locations.

3.3.3 GROUNDED FUZZY TEMPORAL CHAINS

Real-world visual evidence is rarely Boolean. N-TER therefore retains continuous predicate confidence and reasons with fuzzy temporal chains instead of prematurely converting event states into hard logical symbols. For graph-level conjunction, we use the conservative fuzzy operator $\mathrm{AND}_{\min}(a_1, \ldots, a_n) = \min(a_1, \ldots, a_n)$, so that a temporal explanation can be no stronger than its weakest required state. Competing valid explanations are combined through probabilistic fuzzy OR, $\mathrm{OR}(a_1, \ldots, a_n) = 1 - \prod_{j=1}^{n}(1 - a_j)$, with fuzzy negation $\mathrm{NOT}(a) = 1 - a$. Using these operators, N-TER evaluates six complementary event trajectories.

Chain-I: Normal Until Onset. An anomaly is strongly supported when a normal temporal history is followed by a target that simultaneously exhibits anomaly and onset evidence: $S_{t,m}^{\mathrm{NUO}} = \min(N_{t-2,m}, N_{t-1,m}, A_{t,m}, O_{t,m})$. This corresponds to the event evolution *normal → normal → anomaly onset*.

Chain-II: Event Reversal. We next examine whether the target constitutes a strong departure from its preceding normal context: $S_{t,m}^{\mathrm{REV}} = \min(N_{t-2,m}, N_{t-1,m}, A_{t,m})$. A high score indicates that the semantic state at the target reverses the preceding normal trend.

Chain-III: Persistent Anomaly. To distinguish an isolated anomaly-like observation from a sustained event, we define $S_{t,m}^{\mathrm{PER}} = \min(A_{t,m}, A_{t+1,m}, A_{t+2,m})$. The chain therefore measures whether the anomaly state remains active after the target rather than appearing as an isolated spike.

Chain-IV: Ordered Consequence. Some anomalous events are characterized not only by their current appearance but also by the event phase that follows them. We first identify the strongest subsequent continuation or resolution cue, $F_{t,m} = \max_{j \in \{t+1,t+2\}} \max(C_{j,m}, R_{j,m})$, and define $S_{t,m}^{\mathrm{OC}} = \min(A_{t,m}, F_{t,m})$. This chain verifies that the target anomaly is followed by a temporally compatible event phase.

Chain-V: Graph Resolution. A resolving anomalous event should be preceded by anomaly evidence, exhibit a resolution cue at the target, and subsequently return toward normality: $S_{t,m}^{\mathrm{RES}} = \min(A_{t-1,m}, R_{t,m}, N_{t+1,m})$. Hence, the chain represents *anomaly → resolution → normal*.

Chain-VI: Graph Since. Finally, N-TER reasons backward from the target and asks whether the current anomaly state can be explained by an earlier observed onset. For every candidate onset position $k \in \{t-2, t-1\}$, we compute $S_{t,m}^{\mathrm{Since}}(k) = \min\big(O_{k,m}, \min_{j=k}^{t} A_{j,m}\big)$. The bounded SINCE score then selects the strongest coherent onset history, $S_{t,m}^{\mathrm{Since}} = \max_{k \in \{t-2,t-1\}} S_{t,m}^{\mathrm{Since}}(k)$. While the persistence chain asks whether an anomaly remains active *after* the target, SINCE asks whether the target can be traced *backward* to a plausible onset. Together, these relations provide complementary forward and backward evidence about the event trajectory.

Chain-VII: UNKNOWN. Temporal evidence may be unavailable near video boundaries when one or more intervals required by a temporal chain do not exist. NEXUS explicitly represents this absence as UNKNOWN rather than interpreting missing temporal context as negative evidence. Such unavailable chains are excluded from subsequent temporal-consistency aggregation.

3.3.4 GRAPH CONSISTENCY AND CONTRADICTION

The six temporal chains represent complementary explanations of how an anomaly may evolve. Requiring all chains to be simultaneously strong would be inappropriate: an event undergoing onset, for example, should not also be required to exhibit resolution. We therefore combine only the valid supporting chains using fuzzy OR. Let $\mathcal{C}_{t,m}^{\mathrm{valid}}$ denote the set of N-TER chains whose required temporal observations are available. The graph-consistency score is $\kappa_{t,m} = 1 - \prod_{c \in \mathcal{C}_{t,m}^{\mathrm{valid}}} (1 - S_{t,m}^{c})$. Here, $\kappa_{t,m} \in [0, 1]$ measures the degree to which one or more coherent temporal explanations support mechanism m. UNKNOWN chains are excluded from the product and therefore contribute neither positive nor negative evidence.

Temporal consistency alone, however, does not resolve semantic ambiguity. NEXUS therefore introduces a complementary contradiction signal based on the simultaneous presence of anomaly and uncertainty evidence, $\xi_{t,m} = A_{t,m} U_{t,m}$. Contradiction $\xi_{t,m}$ becomes high only when the target provides strong support for both the anomaly interpretation and its ambiguous alternative.

3.4 **RESR**: REASONING-ENABLED SCORE REFINEMENT

UPS and N-TER provide two different views of the same event. UPS measures whether an anomaly interpretation is semantically preferable at the target, whereas N-TER measures whether the surrounding event evolution supports or contradicts that interpretation. The objective of **RESR** is to exchange these two sources of information without allowing temporal logic to manufacture an anomaly that is unsupported by the original visual-semantic evidence.

Let $r_{t,m}$ denote the raw UPS proposal, $\kappa_{t,m}$ the N-TER consistency score, and $\xi_{t,m}$ the corresponding contradiction score. RESR applies the sign-preserving refinement

$$\widetilde{r}_{t,m} = \begin{cases} r_{t,m}, & r_{t,m} \le 0, \\ r_{t,m} + \lambda_\kappa \kappa_{t,m} - \lambda_\xi \xi_{t,m}, & r_{t,m} > 0, \end{cases} \tag{1}$$

where λ_κ controls temporal reinforcement and λ_ξ controls contradiction suppression. The first branch of Eq. 1 is crucial: a proposal rejected by UPS is preserved exactly. N-TER is therefore not allowed to transform a non-positive semantic proposal into an anomalous event merely because some temporal pattern appears plausible. Only visually supported proposals are eligible for reasoning-enabled refinement. For these proposals, high temporal consistency increases support, whereas simultaneous anomaly–uncertainty evidence decreases it.

3.5 ANOMALY DETECTION

The RESR-refined evidence is finally verified by a frozen VLLM Judge. For each target, the Judge receives the visual episode together with a compact symbolic payload containing the refined proposal $\widetilde{r}_{t,m}$, temporal consistency $\kappa_{t,m}$, contradiction $\xi_{t,m}$, event states $\mathbf{q}_{i,m} = [A, U, O, C, R]$, temporal relations, and UNKNOWN indicators. Thus, visual input preserves spatial context, while structured N-TER evidence exposes the event evolution for **spatio-temporal anomaly verification**.

The Judge observes $\mathcal{E}_t^{\text{det}} = [x_{t-1}, x_t, x_{t+1}]$ (PAST | TARGET | FUTURE), whereas N-TER reasons over the wider neighborhood $\{x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}\}$; the two outer intervals remain graph-only evidence and are never supplied as additional VLLM visual inputs. Rather than scoring unconstrained generated text, NEXUS directly normalizes the logits of the fixed tokens `abnormal` and `normal`: $p_t^A = \frac{\exp(\ell_t^A)}{\exp(\ell_t^A)+\exp(\ell_t^N)}$, with $p_t^N = 1 - p_t^A$. To accommodate variable anomaly duration, the same frozen inference pipeline operates at $\mathcal{S} = \{2, 4, 6, 10\}$ s. Predictions are aligned to each base interval by temporal center and fused with equal weights: $p_t^{\text{det}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} p^{A,(s)}_{\mathcal{I}^{(s)}(t)}$, where $\mathcal{I}^{(s)}(t)$ denotes the scale-s interval containing the center of x_t. This fusion is entirely parameter-free, introducing neither a learned scale selector nor benchmark-specific weighting.

4 EXPERIMENTS

We evaluate NEXUS on five video anomaly detection datasets: **UCF–Crime** Sultani et al. (2018), **XD–Violence** Wu et al. (2020), **UBnormal** Acsintoae et al. (2022), **MSAD** Zhu et al. (2024), and **SVAD** Wang et al. (2025a). We follow Zanella et al. (2024) anomaly-detection evaluation protocols and report AUC and AP together with AUC_A and AP_A. The frozen visual–language encoders, ontology construction, temporal scales, VLLM inference settings, and benchmark-specific evaluation details are provided in the **APPENDIX**. Furthermore, *demo videos are in supplementary material*.

4.1 STATE-OF-THE-ART COMPARISON

Table 1 compares NEXUS with weakly-supervised, training-free, and direct frozen-VLLM baselines across five video anomaly detection benchmarks. NEXUS improves the strongest training-free baseline on **10 of 12 reported metrics**, despite introducing no target-domain training or adaptation. The strongest gains appear on UCF–Crime and XD–Violence, where NEXUS improves prior training-free results by up to **+4.57** and **+4.31** points, respectively. These consistent gains across both global and anomaly-focused metrics support the central premise of NEXUS: explicit semantic competition and temporal event reasoning provide information that appearance similarity or implicit multimodal reasoning alone does not capture.

Beyond stronger VLLMs and target supervision. The improvement is not explained simply by employing a stronger multimodal model. Against direct frozen-VLLM inference, NEXUS provides gains of up to **+8.23** on UCF–Crime, **+3.93** on XD–Violence, and **+2.65** on MSAD. This

Model	UCF–Crime		UBnormal	XD–Violence			MSAD				SVAD	
	AUC	AUC_A	AUC	AP	AUC	AP_A	AUC	AUC_A	AP	AP_A	AUC	AP
Weakly-Supervised Methods												
MIL Sultani et al. (2018)	75.41	54.25	50.30	75.68	–	78.61	–	–	–	–	–	–
RTFM Tian et al. (2021)	84.30	62.96	64.94	77.81	–	78.57	86.65	–	–	–	–	–
UR–DMU Zhou et al. (2023)	86.97	70.81	–	81.66	–	83.94	85.78	67.95	67.35	75.30	–	–
UMIL Lv et al. (2023)	86.75	68.68	–	–	–	–	–	–	–	–	–	–
VadCLIP Wu et al. (2024)	88.02	70.23	–	84.15	–	–	–	–	–	–	–	–
GS–MoE D'Amicantonio et al. (2025)	**91.58**	**83.86**	**65.95**	82.89	–	85.74	87.72	69.54	68.26	76.68	–	–
π–VAD Majhi et al. (2025)	90.33	77.77	–	**85.37**	–	**85.79**	**88.68**	**71.25**	**71.26**	**77.86**	–	–
Zero-Shot / Training-Free Methods												
CLIP Radford et al. (2021)	53.16	–	–	17.83	38.21	–	–	–	–	–	69.40	60.10
IMAGEBIND Girdhar et al. (2023)	55.78	–	–	–	–	–	–	–	–	–	65.70	56.40
LLAVA–1.5 Liu et al. (2024)	72.84	–	–	50.26	79.62	–	–	–	–	–	60.80	53.00
VideoLLaMA3+Llama3.1	–	–	–	–	–	–	78.70	–	68.50	–	–	–
LAVAD Zanella et al. (2024)	80.28	–	51.06	62.01	85.36	–	–	–	–	–	–	–
EventVAD Shao et al. (2025)	82.03	–	–	64.04	87.51	–	–	–	–	–	–	–
HoloTrace Wang et al. (2025a)	–	–	–	–	–	–	–	–	–	–	**78.00**	70.50
VADTree Li et al. (2026)	84.74	–	65.80	68.85	90.55	–	**89.32**	67.85	71.41	75.49	–	–
MoniTor Feng et al. (2026)	82.57	–	–	55.01	79.11	–	–	–	–	–	–	–
PANDA Yang et al. (2026)	**84.89**	–	75.78	70.16	–	–	–	–	–	–	–	–
UFVAA Lin et al. (2026)	84.08	–	69.02	68.07	**91.34**	–	86.00	–	75.90	–	–	–
QVAD Bekit et al. (2026)	84.28	–	–	68.53	–	–	–	–	–	–	–	–
LAVIDA Dai et al. (2026)	82.18	–	**76.45**	–	90.62	–	–	–	–	–	–	–
Direct Frozen-VLLM Inference												
Qwen 2.5-VL-7B Team (2025a)	79.75	67.91	69.40	71.20	86.43	73.61	86.30	66.47	76.14	78.12	74.67	70.24
Qwen 3-VL-8B Team (2025b)	79.14	**68.48**	72.02	72.23	88.51	74.36	86.56	69.12	77.74	79.33	75.97	**70.65**
Idefics3-8B Laurençon et al. (2024)	76.11	62.69	68.92	**72.81**	88.64	73.06	82.20	67.54	73.90	78.14	70.05	61.71
LLaVA-OV-7B Li et al. (2024)	80.09	67.69	72.98	72.45	87.24	**74.39**	86.62	68.45	**77.93**	**79.77**	76.46	70.61
InternVL3.5-8B Wang et al. (2025b)	79.27	67.15	72.91	72.81	88.12	74.31	85.67	**69.80**	76.07	79.09	74.41	69.91
NEXUS (Ours)	**88.32**	**73.05**	**75.01**	**76.44**	**92.57**	**78.70**	**88.29**	**72.45**	**80.20**	**80.78**	**78.59**	**72.57**
Gains w.r.t training-free methods	**+3.43**	**+4.57**	**-1.44**	**+3.63**	**+1.23**	**+4.31**	**-1.03**	**+2.65**	**+2.27**	**+1.01**	**+0.59**	**+1.92**

Table 1: State-of-the-art comparison across five video anomaly datasets. Best results among weakly-supervised and prior training-free methods are highlighted in **blue** and **orange**, respectively.

isolates the benefit of structuring frozen foundation-model knowledge into an explicit anomaly hypothesis, temporal consistency, and contradiction evidence before final verification. Importantly, NEXUS also remains highly competitive with target-trained weakly-supervised methods, exceeding their strongest reported results by up to **+9.06** on UBnormal and **+8.94** on MSAD metrics. Thus, explicit semantic–temporal structure can recover substantial discriminative power without benchmark-specific optimization.

Where does NEXUS lag? The two exceptions are UBnormal and the global AUC on MSAD, where NEXUS trails the strongest training-free baseline by only **1.44** and **1.03** points, respectively. We attribute these gaps to two different characteristics. **UBnormal** is a synthetic open-set benchmark whose controlled visual environments often make appearance-level deviations highly distinct; in such cases, the additional temporal structure exploited by NEXUS offers less advantage than on visually ambiguous real-world surveillance. **MSAD**, in contrast, has few videos where post anomaly scenarios resembles to a anomaly scene, where our NEXUS detects it as anomaly, and therefore remains slightly below the strongest method in global AUC. These results suggest that NEXUS is most beneficial when anomaly interpretation depends on resolving semantic ambiguity and temporal evolution, rather than when appearance alone already provides a strong discriminative cue.

4.2 Ablation Study and Impact Visualization

We analyze where NEXUS gains originate and visualize how its reasoning affects individual predictions. Here, all ablations are conducted on UCF–Crime.

Component Contribution. Table 2(a) shows a clear progression from semantic matching to structured event reasoning. Introducing UPS provides the largest initial gain by forcing anomaly hypotheses to compete with uncertain and normal alternatives. More importantly, adding **N-TER** yields a further **+6.21 AUC**, confirming that explicitly modeling temporal event evolution provides substantial information beyond that already available to the frozen VLLM. RESR contributes an additional **+0.78**, showing that temporal consistency and contradiction provide complementary refinement once a proposal is visually supported. The progression therefore follows the intended NEXUS hierarchy: *semantic disambiguation* $\rightarrow$ *temporal event reasoning* $\rightarrow$ *conservative refinement*.

(a) Component-level ablation

Configuration	AUC
OpenCLIP	54.01
+ UPS + VLLM-Judge	81.33
+ N-TER	87.54
+ RESR (Full NEXUS)	**88.32**

(b) N-TER temporal chains

Chains	AUC	Chains	AUC
None	81.33	I–to–IV	87.05
I	84.71	I–to–V	87.69
I–to–II	85.02	I–to–VI	87.98
I–to–III	86.17	**I–to–VII**	**88.32**

(c) Description-bank ablation

Description Banks	AUC
$A + N$	82.06
$A + N + U$	84.57
$A + N + U + O$	86.01
$A + N + U + O + C$	87.53
$\mathbf{A + N + U + O + C + R}$	**88.32**

Table 2: **Ablation analysis of NEXUS on UCF–Crime.**

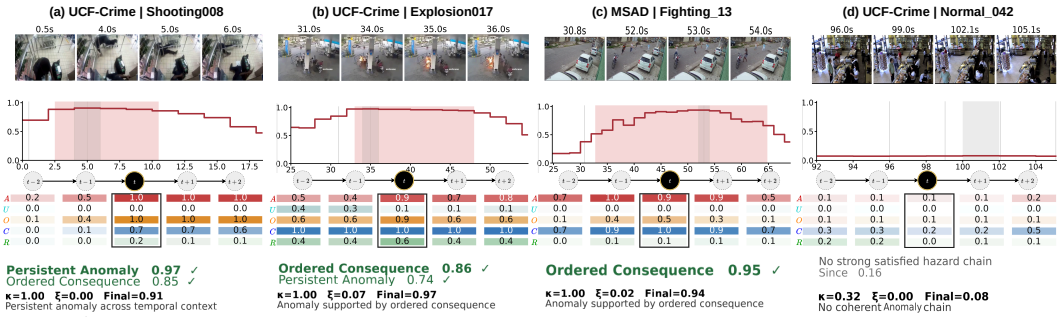

Figure 3: **Qualitative visualization of NEXUS Anomaly Detection and Temporal reasonin.**

Impact of Temporal Reasoning. Table 2(b) isolates the contribution of the N-TER chains. NORMAL UNTIL ONSET provides the strongest individual improvement, highlighting the importance of detecting a structured transition from normal context to anomalous onset rather than judging the target independently. Subsequent gains from reversal, persistence, consequence, resolution, and SINCE reasoning are complementary and nearly monotonic. Together, the complete temporal reasoning process contributes **+6.99 AUC** over the configuration without explicit chains. This supports a central premise of NEXUS: no single temporal relation fully characterizes an anomaly; reliable evidence emerges from reasoning over different stages of its evolution.

Impact of Semantic and Phase Descriptions. Table 2(c) further validates the structure of the fixed ontology. Introducing the uncertainty bank provides a **+2.51 AUC** gain, directly supporting the motivation behind UPS: an anomaly hypothesis becomes more reliable when it must outperform a plausible ambiguous interpretation, rather than only generic normality. The onset, continuation, and resolution banks provide an additional **+3.75 AUC**, showing that the ontology is not merely an enlarged collection of anomaly prompts. Instead, A, N, and U describe *what the observation may represent*, while O, C, and R describe *how that event evolves through time*.

Impact Visualization. Fig. 3 qualitatively exposes how these components interact during inference. In *Shooting008*, anomaly evidence remains coherent across neighboring intervals, allowing persistence and consequence reasoning to reinforce the visual hypothesis. *Explosion017* and *Fighting13* illustrate a complementary case: the appearance changes throughout the event, yet N-TER preserves the anomaly interpretation by capturing its ordered temporal evolution. In contrast, *Normal042* contains no coherent event trajectory, causing the apparent anomaly evidence to remain weak rather than being amplified by temporal reasoning. Together, the ablations and visualizations reveal the same behavior: **NEXUS strengthens an anomaly hypothesis only when semantic evidence is both sufficiently discriminative and temporally coherent; ambiguous or temporally unsupported evidence remains conservative.**

5 CONCLUSION

We presented **NEXUS**, a training-free framework that moves video anomaly detection beyond appearance-driven matching toward explicit **semantic and temporal event reasoning**. NEXUS combines **UPS** to reject ambiguous anomaly hypotheses, **N-TER** to model event evolution through continuous states and interpretable fuzzy temporal chains, and **RESR** to conservatively refine visually supported proposals using temporal consistency and contradiction. A frozen VLLM Judge then performs the final spatio-temporal verification. Across five benchmarks, NEXUS shows that *training-free VAD benefits from reasoning about how an event evolves, not only how anomalous an observation appears*. More broadly, our results demonstrate that frozen foundation-model priors become more effective when organized into explicit, interpretable temporal structures rather than left entirely to implicit multimodal prompting.

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.

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.

Zhaolin Cai, Fan Li, Ziwei Zheng, and Yanjun Qin. Hiprobe-vad: Video anomaly detection via hidden states probing in tuning-free multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 592–601, 2025.

Zhaolin Cai, Fan Li, Huiyu Duan, Lijun He, and Guangtao Zhai. Steering and rectifying latent representation manifolds in frozen multi-modal llms for video anomaly detection. *arXiv preprint arXiv:2602.24021*, 2026a.

Zhaolin Cai, Fan Li, Ziwei Zheng, Haixia Bi, and Lijun He. Headhunt-vad: Hunting robust anomaly-sensitive heads in mllm for tuning-free video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19835–19843, 2026b.

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.

Yuantong Chen, Zhengyan Ding, and YanFeng Shang. Prism: Training-free video anomaly detection via intrinsic statistical modeling. In *Forty-third International Conference on Machine Learning*.

Zunkai Dai, Ke Li, Jiajia Liu, Jie Yang, and Yuanyuan Qiao. No need for real anomaly: Mllm empowered zero-shot video anomaly detection. *arXiv preprint arXiv:2602.19248*, 2026.

Giacomo D'Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, François Bremond, and Egor Bondarev. Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection. *arXiv preprint arXiv:2508.06318*, 2025.

Mohamed Eltahir, Ahmed O Ibrahim, Obada Siralkhatim, Tabarak Abdallah, and Sondos Mohamed. Gridvad: Open-set video anomaly detection via spatial reasoning over stratified frame grids. *arXiv preprint arXiv:2603.25467*, 2026.

Yue Feng, Yingshi Liu, Jingrou Zhang, Jie Qin, et al. Monitor: Exploiting large language models with instruction for online video anomaly detection. *Advances in Neural Information Processing Systems*, 38:35851–35871, 2026.

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.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *Advances in neural information processing systems*, 38:118486–118518, 2026a.

Chao Huang, Penfei Wei, Wei Wang, Jie Wen, Zhihua Wang, Li Shen, Wenqi Ren, and Xiaochun Cao. Spherevad: Training-free video anomaly detection via geodesic inference on the unit hypersphere. *arXiv preprint arXiv:2605.08003*, 2026b.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

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.

Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. *Advances in Neural Information Processing Systems*, 38:29713–29744, 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 (CVPR)*, pp. 26296–26306, June 2024.

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

Zihao Liu, Xiaoyu Wu, Wenna Li, Jianqin Wu, and Linlin Yang. Esom: Efficiently understanding streaming video anomalies with open-world dynamic definitions. *arXiv preprint arXiv:2604.07772*, 2026.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8022–8031, 2023.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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.

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. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2024.

Qwen Team. Qwen2.5-vl, January 2025a. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Qwen Team. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

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*, pp. 4975–4986, 2021.

Hanling Wang, Qing Li, Li Chen, Haidong Kang, Fei Ma, and Yong Jiang. Holotrace: Llm-based bidirectional causal knowledge graph for edge-cloud video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6510–6519, 2025a.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

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*, pp. 322–339. Springer, 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.

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, 2024.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. *Advances in Neural Information Processing Systems*, 38:83182–83211, 2026.

M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14744–14754, 2022.

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, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853. IEEE, 2025.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*, 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.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025.

A Appendix

A.1 Datasets

We evaluate NEXUS on five public anomaly-analysis benchmarks: **UCF-Crime** Sultani et al. (2018), **MSAD** Zhu et al. (2024), **XD-Violence** Wu et al. (2020), **SVAD** Wang et al. (2025a), and **UBnormal** Acsintoae et al. (2022). Together, they cover long untrimmed surveillance, multi-scenario safety events, cross-source violence, diverse short-form anomalies, and open-set abnormal behavior. We discuss them below from the most challenging real-world settings toward the comparatively more controlled benchmark conditions.

UCF-Crime. UCF-Crime is a large-scale real-world surveillance benchmark containing 1,900 long, untrimmed videos and 13 anomaly categories, including abuse, fighting, robbery, burglary, shooting, road accidents, and vandalism. We evaluate on the standard 290-video test set, containing 140 anomalous videos. UCF-Crime is particularly challenging for training-free reasoning because anomalies often occupy only a small fraction of long videos, occur at widely varying temporal scales, and appear under substantial changes in viewpoint, resolution, illumination, crowd density, and scene context.

MSAD. MSAD extends anomaly detection to *multi-scenario* surveillance, covering diverse indoor and outdoor environments, camera viewpoints, lighting and weather conditions, and both human- and non-human-related anomalies. We evaluate on 360 videos, including 240 anomalous videos. Compared with single-domain benchmarks, MSAD places greater emphasis on cross-scene semantic consistency: the same anomaly mechanism may exhibit substantially different visual signatures across streets, shops, offices, transportation scenes, and environmental conditions.

XD-Violence. XD-Violence contains 4,754 untrimmed videos collected from heterogeneous sources and covers six major violence-related anomaly categories. The standard test set contains 800 videos, of which 500 are anomalous in our evaluation protocol. Its difficulty stems from strong appearance and domain variation across surveillance footage, movies, and web videos, together with large differences in event duration, camera motion, scene composition, and visual quality.

SVAD. SVAD focuses on short-form videos with diverse backgrounds and event semantics, including accidental, violent, traffic, falling, and animal-related anomalies. We evaluate on a 119-video subset containing 101 anomalous videos. Although the clips are temporally shorter than those in UCF-Crime or XD-Violence, the dataset remains challenging because similar actions appear under highly variable scene context and the benchmark explicitly considers generalization across anomaly categories.

UBnormal. UBnormal is a synthetic open-set anomaly-detection benchmark composed of multiple virtual surveillance scenes and diverse abnormal human behaviors. Our evaluation uses 211 videos, of which 158 contain anomalies. Its open-set construction makes semantic generalization challenging, while its controlled synthetic environments provide a complementary setting to the more visually heterogeneous real-world datasets above.

Table 3 summarizes the evaluation sets used in our experiments. For additional analysis, we separately report statistics over anomalous videos, which preserves the substantial normal-frame content surrounding each annotated anomalous segment.

A.2 Evaluation Protocols and Metrics

NEXUS produces one anomaly posterior for each base two-second interval. The interval score is assigned to its covered frames to obtain a frame-level prediction sequence $\hat{y}_f \in [0, 1]$. Following standard VAD evaluation, we report **frame-level ROC-AUC** and **Average Precision (AP)**. ROC-AUC measures global anomaly–normal ranking, while AP is particularly informative under the strong frame-level class imbalance of surveillance videos.

In addition to metrics over the complete test set, we report AUC_A and AP_A on the subset of **anomalous videos only**. This complementary evaluation removes normal-only videos from met-

Dataset	Eval. videos	Anom. videos	Frames in anom. videos	Anom. frames
UCF-Crime	290	140	462,903	84,182
MSAD	360	240	138,451	81,813
XD-Violence	800	500	1,081,530	538,324
SVAD	119	101	138,359	79,030
UBnormal	211	158	69,047	49,650

Table 3: **Evaluation-set statistics.** "Anom. frames" denotes annotated anomalous frames, whereas "Frames in anom. videos" counts all frames—normal and anomalous—within videos containing at least one anomaly.

(b) Multi-scale inference

Temporal Scales (s)	AUC
2	87.31
$2+4$	87.86
$2+4+6$	88.09
$\mathbf{2+4+6+10}$	**88.32**

Table 4: Effect of progressively incorporating multiple temporal scales

ric computation and therefore emphasizes temporal localization within videos that actually contain an anomaly. The VLLM Judge observes $\mathcal{E}_t^{\text{det}} = [x_{t-1}, x_t, x_{t+1}]$, while N-TER may reason over the wider neighborhood $\{x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}\}$. The outer graph nodes provide symbolic temporal evidence only and are not added as extra VLLM visual inputs.

B Additional Ablation Study

Impact of Multi-scale Inference. Anomalies do not occur at a fixed temporal duration. Table 4 therefore examines the contribution of the parameter-free multi-scale inference used by NEXUS. Operating only at the 2 s base scale achieves *87.31 AUC*. Adding the 4 s scale increases performance to 87.86 (*+0.55*), while incorporating 6 s and 10 s contexts further improves AUC to 88.09 and *88.32*, respectively.

The cumulative *+1.01* point improvement confirms that temporal scale is itself a source of uncertainty: short intervals preserve localized event changes, whereas longer intervals provide broader evidence about event development. Importantly, the gains are obtained through equal center-aligned fusion, without learning scale-specific weights. Thus, the benefit comes from complementary temporal observations rather than additional optimization.

C Implementation Details

We decode videos with OpenCV and sample frames at 1 FPS. The base NEXUS configuration uses non-overlapping 2-s intervals with a 2-s stride, yielding two sampled frames for a standard interior interval. The sampler supports at most eight frames per interval. The final interval is clipped to the available video duration; when random-access decoding fails near a corrupted video tail, we use the nearest preceding decodable frame rather than inserting an artificial black frame. We use the frozen OpenCLIP ViT-L/14 model with the `metaclip_fullcc` checkpoint. Frames are processed in batches of 64. Frame embeddings are ℓ_2-normalized, temporally averaged within each interval, and normalized again before similarity computation. Text descriptions are encoded once and cached for the complete evaluation. The frozen checkpoint has `logit_scale` $= 4.60517$, corresponding to $\exp(\texttt{logit_scale}) \approx 100$. We retain this native scale without additional temperature calibration. A prediction is produced every 2 s, but the temporal reasoning context is wider than the target interval. For an interior target, N-TER accesses the five intervals $\{t-2, t-1, t, t+1, t+2\}$, corresponding to at most 10 s of symbolic context. The frozen VLLM receives only Past–Target–Future visual groups, i.e., normally six images at the base sampling rate. The two outer N-TER intervals are therefore used only for symbolic temporal reasoning. Missing temporal context near the beginning or end of a video is explicitly marked as UNKNOWN. Missing nodes are neither duplicated nor replaced with zero-valued evidence. For diagnostic serialization, unavailable rules may carry a neutral score

State	Role	Per Mechanism	Total
A	anomaly hypothesis	3	24
U	plausible competing interpretation	3	24
N	generic normality	global	2
O	event initiation	1	8
C	event continuation	1	8
R	event resolution	1	8

Table 5: **Structure of the fixed NEXUS ontology.**

of 0.5 together with an UNKNOWN status; such rules are excluded from valid temporal support. For visualization, rule scores ≥ 0.65 are reported as SATISFIED, scores ≤ 0.35 as CONTRADICTED, and intermediate values as UNKNOWN. We use the frozen `Qwen/Qwen3-VL-8B-Instruct` model in `bfloat16` inference mode with FlashAttention-2. No autoregressive explanation is generated during detection. The model performs one multimodal forward pass and the detector reads the fixed abnormal/normal token logits. Thus, the reported anomaly score does not depend on generation length, temperature, or free-form textual decoding. OpenCLIP interval features are precomputed once per video and reused by overlapping temporal graphs. Text embeddings are shared after their first encoding. Visual caches are scoped by dataset, resolved video path, and video identifier to avoid cross-video collisions. These caches change only runtime; they do not modify any anomaly score.

C.1 ONTOLOGY CONSTRUCTION AND MOTIVATION

Design Rationale. The ontology is organized around two complementary questions: *what does the observation represent?* and *how is the event evolving?* The first is captured by anomaly (A), uncertainty (U), and normality (N); the second by onset (O), continuation (C), and resolution (R). This separation lets NEXUS distinguish an anomalous interpretation from a visually plausible alternative while independently describing whether the corresponding event is beginning, persisting, or ending. The ontology is therefore mechanism-centric rather than a direct copy of benchmark class labels.

Ontology Composition. The implementation contains eight anomaly mechanisms. Each mechanism has three anomaly descriptions, three mechanism-specific uncertain alternatives, and one description each for onset, continuation, and resolution. Normality is represented by two global descriptions shared across mechanisms. The complete ontology therefore contains 24 anomaly, 24 uncertainty, 2 normal, 8 onset, 8 continuation, and 8 resolution descriptions.

Uncertainty versus Normality. U and N serve different purposes. U contains mechanism-specific hard alternatives that can visually resemble an anomaly without necessarily being abnormal; for example, physical contact may correspond to either violence or structured sports activity. N is instead a generic normal-scene reference and is independent of the candidate mechanism. In the implementation, U is retained under the legacy name `benign`, but it should be interpreted as a competing semantic explanation rather than a learned uncertainty estimator.

Description Construction. Descriptions are short, action-centric phrases rather than dataset class names. For the physical-interaction mechanism, representative descriptions include "visible physical assault fight brawl beating kicking punching or violent struggle" for A, "sports contact martial arts training dancing stage performance without visible real harm" for U, "a strike, forced contact, pursuit, restraint, or weapon threat begins" for O, "the attack, struggle, coercion, injury, or weapon control remains visible" for C, and "contact has ended and visible interaction is sport, performance, assistance, or routine social activity" for R. The generic normal bank includes descriptions such as "ordinary calm surveillance scene without danger."

Dataset Independence. The same ontology and description banks are used across all evaluated datasets; no dataset-specific prompt bank or class-dependent reasoning rule is introduced at inference time. This keeps the semantic interface fixed across benchmarks and temporal scales.

Parameter	Value
Frame sampling rate	1 FPS
Base interval duration	2 s
Base interval stride	2 s
Maximum sampled frames / interval	8
N-TER context radius	2 intervals
N-TER graph size	up to 5 nodes
OpenCLIP backbone	ViT-L/14
OpenCLIP checkpoint	`metaclip_fullcc`
OpenCLIP image batch size	64
Frozen CLIP scale $\exp($`logit_scale`$)$	100.0000064
Satisfied threshold	0.65
Contradicted threshold	0.35
RESR consistency coefficient	0.5
RESR contradiction coefficient	0.5
RESR minimum gate g	0.5
Frozen VLLM	Qwen3-VL-8B-Instruct
VLLM precision	bfloat16
Attention implementation	FlashAttention-2
VLLM visual groups	Past / Target / Future
Base-scale visual inputs	typically 6 images
Multi-scale windows	$\{2, 4, 6, 10\}$ s
Multi-scale fusion	equal-weight center-aligned averaging
Trainable parameters	0

Table 6: **Fixed implementation settings used by NEXUS.** All values are shared across datasets unless stated otherwise.

C.2 Fixed Inference Hyperparameters

Hyperparameter Selection. All reasoning coefficients are fixed globally and shared across datasets. There is no dataset-specific optimization or learned calibration. In the audited implementation, temporal consistency and contradiction enter the RESR gate with equal coefficients of 0.5, while the minimum multiplicative gate is also fixed to 0.5. This symmetric setting avoids assigning a stronger prior importance to either temporal reinforcement or semantic contradiction.

Training-Free Setting. No model parameter is updated on any target anomaly-detection dataset. OpenCLIP, Qwen3-VL, the ontology, temporal rules, and reasoning coefficients remain fixed. The frozen OpenCLIP and Qwen3-VL checkpoints contain approximately 427.6M and 8.77B parameters, respectively; these are frozen backbone parameters and are not trainable NEXUS parameters.

D Inference Cost

We compute inference cost by excluding the model loading time and by enabling the UPS, the five-node N-TER graph, RESR, and frozen VLLM verification. We use five warm-up intervals followed by 30 distinct interior intervals from UCF–Crime `Shooting008_x264`. CUDA operations are synchronized before timing.

Hardware. Profiling is performed on one NVIDIA H100 80GB HBM3 GPU using Python 3.11.15, PyTorch 2.10.0+cu128, CUDA 12.8, NVIDIA driver 570.172.08, Transformers 4.57.6, and OpenCLIP 3.3.0.

Latency. With the native caches active, NEXUS requires 247.34 ms per 2-s target interval on average, corresponding to 4.043 target intervals/s and 8.086 video-s/s. The median and P95 latencies are 255.94 ms and 263.25 ms, respectively. The resulting real-time factor is 0.1237. Disabling the VLLM visual-feature cache increases the mean latency only slightly to 254.01 ms.

End-to-End Cost. The implementation precomputes video-frame and OpenCLIP interval features before the target loop. Amortized over the same profiling sequence, frame extraction adds 42.55 ms/target and OpenCLIP precomputation adds 20.62 ms/target. Including these costs gives an effective end-to-end latency of 317.19 ms/target, corresponding to 3.153 target intervals/s and an RTF of 0.1586.

Metric	Cached	Visual Cache Off
Mean latency / target	247.34 ms	254.01 ms
Median latency / target	255.94 ms	251.59 ms
P95 latency / target	263.25 ms	268.11 ms
Throughput	4.043 int./s	3.937 int./s
Video processing rate	8.086 video-s/s	7.874 video-s/s
Real-time factor	0.1237	0.1270

Table 7: **NEXUS inference cost.**

Component	Latency	Share
UPS	0.57 ms	0.23%
N-TER	4.87 ms	1.97%
RESR	0.03 ms	0.01%
Frozen VLLM verification	241.86 ms	97.79%
Final score handling	0.01 ms	0.01%

Table 8: **Runtime decomposition.**

Runtime Breakdown. The neuro-symbolic components are inexpensive relative to frozen multimodal verification. UPS, N-TER, and RESR together require approximately 5.47 ms per target, whereas Qwen3-VL verification requires 241.86 ms.

Memory. Peak allocated and reserved GPU memory are 18.72 GiB and 19.29 GiB, respectively. We do not report a single end-to-end VLLM FLOP estimate because the executed cost depends on the processor-selected visual-token count, prompt length, attention kernel, and feature-cache reuse. We therefore report measured latency, throughput, memory, and model size instead.