REMEM-VAD: CONNECTING EVENT REPRESENTATIONS AND MEMORY FOR ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video anomaly detection (VAD) is challenging because visually similar actions can correspond to either normal or abnormal events. Recent vision-language models (VLMs) provide broad semantic knowledge, but recognizing an action alone is insufficient when its meaning depends on the participant's response, the event's progression, and the evidence supporting interpretations. To address this issue, we propose to transform resuable experience to event representations and structural memory, enhancing VLMs' capabilities with memory-ground reasoning. Specifically, ANOMALY AGENT is proposed, a memory-augmented VAD framework is built to combine large-model understanding with lightweight anomaly scoring. For each video window, a frozen VLM represents the actions, responses, event phase, and supporting evidence. This event representation specifies what must be explained and guides the verification of competing explanations against the same frames. Under this guidance, candidate patterns are retrieved from memory and matched to the current visual evidence. A lightweight trainable scorer then learns how to combine evidence, explanation judgments, and memory pattern into an anomaly score. Counterexample-guided refinement updates memory and representation during development, and freezes them at test time. Experiments on XDViolence, UCF-Crime and UBnormal demonstrate state-of-the-art performance. Code is available at `https://anonymous.4open.science/r/Memory-Augmented-VLM-153D`

1 INTRODUCTION

Video anomaly detection (VAD) aims to identify unusual events in videos and supports applications such as surveillance, traffic monitoring, and industrial safety. Existing methods learn normal activity through reconstruction, prediction, contrastive learning, or self-supervised objectives Zaheer et al. (2022); Wu et al. (2022); Wang et al. (2022); Yang et al. (2023); Yan et al. (2023); Shi et al. (2024). Weakly supervised approaches use coarse labels, multiple instance learning, and pseudo-label refinement to reduce annotation requirements and improve temporal localization Lv et al. (2023); Zhang et al. (2023); Chen et al. (2024a); Yang et al. (2024b); Majhi et al. (2025); Amicantonio et al. (2025); Wang & Chen (2025). These approaches have advanced anomaly detection, but identifying unfamiliar or context-dependent events remains difficult. The same visible action can have different meanings depending on the participants' responses, the event stage, and the available evidence.

Recent work has extended VAD toward open-set, open-vocabulary, and language-based reasoning Liu et al. (2026); Zhai et al. (2026); Zhu et al. (2022); Wu et al. (2024a); Li et al. (2025a); Tang et al. (2024); Zhang et al. (2025a). Foundation-model-based methods use the semantic knowledge of vision–language models (VLMs) and large language models (LLMs) to interpret video with limited task-specific training Zanella et al. (2024); Yang et al. (2024a); Ye et al. (2025); Yang et al. (2025a); Li et al. (2025b); Yang et al. (2025b). This creates an opportunity to recognize events beyond a fixed set of training categories. However, semantic knowledge alone does not determine which evidence is relevant, whether an explanation refers to the observed event, or how uncertain judgments should affect the final anomaly score. These questions become especially important when observations are incomplete and labels provide only coarse supervision.

To distinguish similar-looking actions, a VLM needs to identify which visible details support different explanations and connect those details to structured experience systems to make judgments. To

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.

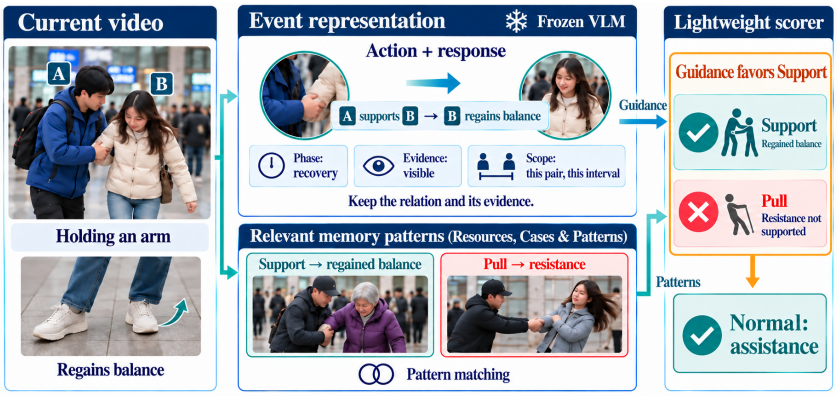

Figure 1: Overview of the framework. Similar actions can correspond to different events. The proposed event representation preserves the action–response relation, event phase, visual evidence, and participant–time scope, providing guidance for interpreting current interactions. Relevant memory patterns supply normal and abnormal explanations. In this example, regained balance supports assistance, while the pulling pattern lacks supporting evidence. A lightweight scorer follows the guidance of representation and selects the memory pattern of support to produce the anomaly decision.

this end, we propose ANOMALY AGENT, a VAD framework that combines structured event memory and representations for event reasoning. Its memory organizes experience into three complementary roles. **Guidance memory** records what to inspect. **Case memory** stores individual examples with their supporting frames. **Pattern memory** summarizes common structures across examples. This organization supports concise retrieval in making judgements.

As shown in Fig. 1, a frozen VLM represents each event through its action and response, phase, evidence, and scope—the relevant people and time interval. This representation identifies the details needed to interpret the event and helps determine which past memory items are relevant. Structured memory provides observation guidance, representative cases, and shared event patterns to support comparisons between normal and abnormal explanations. A lightweight trainable scorer combines current event evidence with support from retrieved memory in making decisions. For example, evidence that a person regains balance favors assistance over forceful pulling. This design combines the VLM's semantic understanding with task-specific scoring while keeping its weights frozen.

During development, counterexample-guided refinement identifies problematic distinctions, such as resistance versus regained balance. Representation and memory are refined in this way. At test time, memory, prompts, scorer parameters. State-of-the-art performance is achieved on XDViolence, UCF-Crime and UBnormal.

Our main contributions are summarized as follows:

- We propose a structured memory framework for VAD. Resources, cases, and patterns organize reusable knowledge at complementary levels, guiding VLM is making judgements. An event representation is also proposed to guide event reasoning and memory retrieval. The memory and representation are learned through counterexample-guided refinement.
- Frozen large-model reasoning is combined with lightweight trainable scoring. The scorer integrates event evidence and representations, as well as memory support, in predicting anomaly scores.
- State-of-the-art performance is achieved on multiple benchmarks.

2 RELATED WORKS

Normality modeling and weak supervision. Normality-based VAD uses frame prediction, event restoration, and generative feature modeling Luo et al. (2021); Yang et al. (2023); Yan et al. (2023), or self-supervised representation learning Wang et al. (2022); Shi et al. (2023). Weakly supervised

methods use video-level labels through multiple instance learning, pseudo-label refinement, and event-completeness learning Lv et al. (2023); Zhang et al. (2023); Wang & Chen (2025), with text guidance supporting feature learning and localization Chen et al. (2024a); Yang et al. (2024b); Liu et al. (2024); Pu et al. (2024). KSF aligns video and text through category-pattern sharing and semantic feedback Cai et al. (2026a), while sparse frame labels provide localized supervision Chen et al. (2025b). Since positive video labels do not locate anomalous windows, ANOMALY AGENT fits its scorer and refines memory using only comparisons justified by the available labels.

Open-world and language-guided VAD. Open-set and cross-domain methods address unseen anomalies and dataset transfer Zhu et al. (2022); Jain et al. (2024), while open-vocabulary methods recognize unseen categories through visual and textual knowledge Wu et al. (2024a); Li et al. (2025a). Hawk adds motion cues Tang et al. (2024), and Holmes-VAU combines hierarchical instruction data with anomaly-focused frame sampling Zhang et al. (2025a). With little task-specific optimization, LAVAD cleans captions, summarizes temporal context, and refines LLM scores Zanella et al. (2024); AnomalyRuler induces and applies normal and anomaly rules from a few normal observations Yang et al. (2024a). Our focus is organizing this reusable knowledge, checking it against current events, and converting it into numerical detection evidence.

Memory management and retrieval. Predictive methods combine future-frame prediction with normal-event retrieval Cao et al. (2024) or use dynamic memory for reconstruction Hu et al. (2025). VERA aggregates scores from similar retrieved segments Ye et al. (2025); MoniTor maintains short- and long-term context for online reasoning Yang et al. (2025a); PANDA combines rule retrieval, reasoning histories, and tool-assisted reflection Yang et al. (2025b). For general agents, OpenViking organizes context into abstract, overview, and detailed views Volcengine (2026). The proposed memory system provides hierarchical guidance.

Event representation and evidence grounding. Attribute-based VAD uses properties such as pose and velocity Reiss & Hoshen (2022), while knowledge graphs model action–scene relations Chen et al. (2025a). VADTree builds event hierarchies Li et al. (2025b); EventVAD uses dynamic graphs and event boundaries Shao et al. (2025); CSI-VAD separates environment, object, and temporal contexts Kim et al. (2026); SESAD selects learned features through action and scene memories Chen et al. (2026). The proposed representation records action and response, phase, evidence, and scope.

Large-model and small-model collaboration. VadCLIP freezes CLIP encoders while learning temporal and prompting modules for weakly supervised detection Wu et al. (2024d). VA-GPT supplies lightweight-classifier temporal priors to an instruction-tuned LLM Chen et al. (2025c). Cerberus combines lightweight filtering, VLM captioning, and similarity scoring against induced rules Zheng et al. (2025). In our framework, frozen VLMs provide explanation judgments while a lightweight scorer learns how to combine VLM's judgments with memory patterns.

Feedback and counterexample-guided refinement. VERA refines guiding questions using predictions and supervision Ye et al. (2025); PANDA reflects to gather evidence during processing Yang et al. (2025b); Cerberus evolves rules through feedback Zheng et al. (2025). VAD-DPO uses visually similar but semantically contrasting clips for model adaptation Zhang et al. (2025b).

3 METHODOLOGY

General event prompts may suggest plausible explanations without specifying the visual evidence needed to support them in a particular interaction. For example, interpreting physical contact as normal sporting activity requires evidence that the contact is part of an ongoing game. To address this gap, we propose an event memory framework that stores event evidence together with the relevant participants, time span, and conditions under which an explanation applies. Within this framework, a frozen vision–language model (VLM) identifies local events and checks whether candidate explanations are supported by observable evidence for the same actions, participants, and time spans. Retrieved memory patterns provide complementary evidence by relating the current observations to known normal and abnormal interactions. A lightweight trainable model then combines these event observations and memory pattern support to produce a window-level anomaly score.

The framework has three stages. Stage 1 selects sparse anchors and builds resource, case, and pattern memories. Stage 2 retrieves relevant experience and learns how to score events. Errors then guide

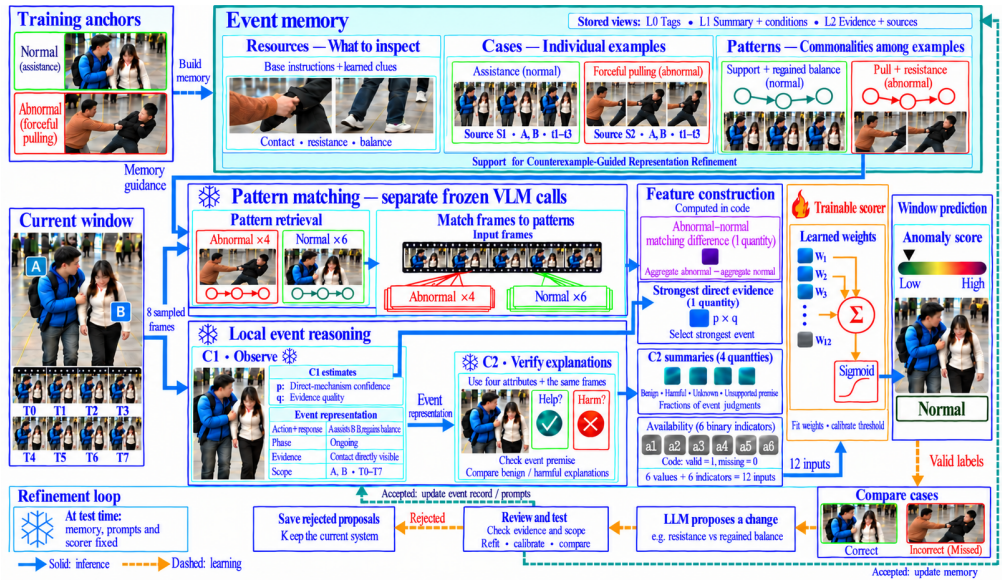

Figure 2: Overview of the framework. Training anchors initialize resources, cases, and patterns. Each memory item contains L0 retrieval tags, L1 summaries and conditions, and L2 supporting evidence. Event representation in C1 provides guidance for retrieving relevant normal and abnormal memory patterns. The retrieved patterns are matched to the current visual evidence, while C2 checks whether candidate explanations are supported. A lightweight trainable scorer combines the results from C1, C2 and pattern-matching to produce a window-level anomaly score. During learning, errors update the representation and memory.

revisions to memory and the representation of experience in memory. Stage 3 evaluates the frozen system. Let $\mathcal{K} = \{1, \ldots, K\}$ be the anomaly types. A video is $V = \{x_t\}_{t=1}^{T}$ with video-level labels $Y^V \subseteq \mathcal{K}$; $Y^V = \varnothing$ denotes an explicitly normal video. We divide it into windows $W = \{w_i\}_{i=1}^{N}$, where w_i spans $[t_i, \min(t_i + L, T + 1))$. We use $L = 96$ frames and a fixed stride $\Delta = 48$. The output is an anomaly score $s_i \in [0, 1]$.

3.1 Stage 1–Step 1: Offline Anchor Selection

The first stage converts coarse video-level supervision into a small set of reliable segment anchors. The key idea is not to generate dense pseudo labels for all windows, which would be noisy, but to identify only a few highly confident anomalous and non-anomalous segments per video for memory construction. For a training video with video-level set Y^V, we ask a VLM Wang et al. (2024a) to inspect every window w_i and, for each candidate type $k \in Y^V$, decide whether that window contains anomaly type k and output a confidence score

$$\alpha_{i,k} = g_{\text{loc}}(w_i, k, Y^V) \in [0, 1], \qquad k \in Y^V. \tag{1}$$

For videos labeled as normal, i.e., $Y^V = \varnothing$. Rather than trusting all pseudo-labeled windows, we retain only the most reliable anchors. Specifically, for each anomalous video, we select the two highest-confidence anomalous windows as positive set $\mathcal{S}^+(V)$ and the two lowest-confidence (non-anomalous) windows as negative set $\mathcal{S}^-(V)$. For normal videos, only $\mathcal{S}^-(V)$ is retained. On Ubnormal we directly use abnormal windows by referring to frame-level labels.

3.2 Stage 1–Step 2: Building Resource Memory

We use the selected video segments (anchors) to build reusable event clues. For each segment, the frozen VLM describes what it sees without being shown the label. We then save this description together with the label, even if it disagrees with the label.

For a positive anchor a associated with type k, let $E^+_{a,k} = \{E^{\text{dir}}_{a,k}, E^{\text{ind}}_{a,k}\}$ contain direct and indirect clues. Direct clues describe visible anomalies, indirect clues describe context or event phase without independently proving a current anomaly. Counter-clues $E^-_{a,k}$ describe a visible normal mechanism, such as helping someone regain balance. Semantically equivalent clues are merged across videos:

$$\mathcal{R}^+_k = \text{Merge}\left(\bigcup_V \bigcup_{a\in\mathcal{S}^+(V)} E^+_{a,k}\right), \mathcal{R}^-_k \quad = \text{Merge}\left(\bigcup_V \bigcup_{a\in\mathcal{S}^-(V)} E^-_{a,k}\right). \tag{2}$$

Merging removes incidental wording, identity, and source-style differences while retaining supporting cases, counterexamples, and applicability conditions. The resource memory is

$$\mathcal{R} = \bigcup_{k=1}^{K} \left(\mathcal{R}^+_k \cup \mathcal{R}^-_k\right) \cup \mathcal{R}^{\text{base}}, \tag{3}$$

where $\mathcal{R}^{\text{base}}$ contains general event primitives which are initial priors.

3.3 Stage 1–Step 3: Building Case and Pattern Memories

We store the selected anchors as seed cases $\mathcal{C}^0$, including their displayed frames, label scope, VLM observations, and available evidence. Related cases are summarized into seed patterns $\mathcal{P}^0$: recurrent mechanisms and applicability conditions. These seed cases and patterns form the initial memory for Stage 2. They provide examples and possible explanations for analyzing new events. Stage 2 adds cases $\mathcal{C}^\Delta$ and accepted patterns $\mathcal{P}^\Delta$. The effective memory is

$$\mathcal{M}_{\text{act}} = \mathcal{R} \cup \mathcal{P}^0 \cup \mathcal{P}^\Delta \cup \mathcal{C}^0 \cup \mathcal{C}^\Delta. \tag{4}$$

Every resource, case, and pattern uses three levels: **L0 abstracts** provide compact retrieval tags, **L1 summaries** describe what happens, in what order, and under what conditions, **L2 details** record what was observed and where it appears in the video. This gives Stage 2 an initial memory.

3.4 Stage 2–Step 1: Memory Retrieval

For each window w_i, the VLM produces a short, neutral description q_i of the visible frames. The system retrieves a small set of items similar to q_i: supportive resources $\mathcal{R}^+_i$, counter-resources $\mathcal{R}^-_i$, supportive patterns $\mathcal{P}^+_i$, counter-patterns $\mathcal{P}^-_i$, and similar cases $\mathcal{C}_i$. For example, a query about arm-pulling may retrieve both conflict and assistance patterns.

The VLM examines the current window alongside L0 abstracts and L1 summaries. We rank candidates using VLM-assessed visual support. 4 abnormal and 6 normal patterns are retained. Relevant L2 details from the selected patterns are then added to the reasoning prompt. The shortlist sizes remain fixed. The next step checks whether each explanation fits the observed interaction. For example, to assess an assistance explanation for arm-pulling, the VLM checks whether one person is helping the other regain balance.

3.5 Stage 2–Step 2: Local Event Reasoning and Lightweight Scoring

For each window w_i, two branches use the same eight sampled frames: C1/C2 observe and verify local events, while memory matching provides pattern support. All VLM weights remain frozen; only the lightweight scorer is trained.

Observing events (C1). C1 extracts candidate events: $\mathcal{E}_i = D(w_i)$. Each record contains event, supporting-frame references and the four fields in Table 1. For each event e_{ij}, C1 separately estimates direct-mechanism confidence p_{ij} and evidence quality q_{ij}, $\forall j$ denotes candidate events.

Checking explanations (C2). C2 retains the C1 records and adds verification results: $\overline{\mathcal{E}}_i = V(w_i, \mathcal{E}_i)$. It checks whether the event is supported by the frames and whether normal or abnormal explanations fit the same action, participants, and time interval. For example, support followed

Table 1: Four parts of an event representation.

Part	Description
Action and response	Who does what to whom or what, and how the other person or object responds.
Event phase	Preparation, an ongoing action or state, or the aftermath.
Visual evidence	Supporting observations and whether they are directly visible, inferred, or unclear.
Participants and time	The people or objects, event, and time interval the description concerns.

by regained balance favors assistance, whereas pulling met with resistance favors forceful pulling. Indistinguishable explanations remain uncertain.

Matching memory patterns. Provided with 4 abnormal patterns $\mathcal{P}_i^+$ and 6 normal patterns $\mathcal{P}_i^-$ from memory retrieval. VLMs assess abnormal support m_i^+ and normal support m_i^-:

$$\Delta_i = m_i^+ - m_i^-. \tag{5}$$

Larger values indicate stronger relative support for abnormal patterns. The procedures for assessing support will be shown in Section A.4 in appendix.

Constructing features and learning the score. We compute six evidence values: Δ_i, the strongest direct evidence d_i from C1, and four C2 summary fractions $\mathbf{c}_i$. The direct-evidence value is

$$d_i = \max_j (p_{ij} q_{ij}), \tag{6}$$

The four C2 fractions summarize benign support, harmful support, uncertainty, and unsupported premises. Six availability indicators $\mathbf{a}_i \in \{0,1\}^6$ mark each evidence value as available (1) or missing (0), giving $\mathbf{f}_i = [\Delta_i,\ d_i,\ \mathbf{c}_i^\top,\ \mathbf{a}_i^\top]^\top \in \mathbb{R}^{12}$. A lightweight scorer produces

$$s_i = \sigma\big(b + \boldsymbol{\gamma}^\top \mathbf{f}_i\big) \tag{7}$$

where σ maps the weighted sum to $[0,1]$. We learn the bias b and weights $\boldsymbol{\gamma}$ on windows with valid supervision using binary cross-entropy and L_2 regularization. Memory, prompts, and the scorer remain fixed at test time.

3.6 Stage 2–Step 3: Counterexample-Guided Representation Refinement

A teacher LLM reviews C1 observations, C2 judgments, and pattern-matching evidence to check for missed observations, label problems, retrieval failures, and calibration errors. It compares failures with correctly classified events showing similar actions and proposes representation changes when a necessary distinction is missing. For example, resistance versus regained balance helps distinguish forceful pulling from assistance. These comparisons refine the four parts in Table 1.

The teacher proposes small changes: adding an event description, clarifying a condition, splitting or merging patterns, retiring unsupported patterns, or revising observation prompts. The VLM weights remain frozen. For each proposed pattern or distinction p, we count support across source videos:

$$\mathrm{Supp}(p) = \left|\{c \in \mathcal{C}^* : p \text{ is supported by } c\}\right|, \tag{8}$$

where $\mathcal{C}^*$ is the case set supporting p, containing one representative case per source video. Repeated windows from one video count only once. We use a separate LLM call to check each proposed change. It rejects claims without visible evidence, judgments based only on video identity or visual style, and evidence assigned to the wrong people or time interval. New or revised patterns that pass this check form $\mathcal{P}_{\mathrm{trial}}$. Old patterns proposed for removal from active use form $\mathcal{P}_{\mathrm{retire}}$:

$$\mathcal{M}_{\mathrm{trial}} = (\mathcal{M}_{\mathrm{act}} \setminus \mathcal{P}_{\mathrm{retire}}) \cup \mathcal{P}_{\mathrm{trial}}. \tag{9}$$

We apply any proposed prompt changes and recompute only affected observations, verification results, and matching features, reusing unaffected outputs. Development videos are divided into source-disjoint fitting (60%), calibration (20%), and check (20%) sets. Original and revised features are used to fit separate lightweight scorers with the same architecture and training procedure. We then freeze both systems and compare classification errors and anomaly-score rankings on check

Algorithm 1 Improving event memory in one round

Input: Current memory, prompts, and scorer, available labels, separate videos for training acceptance criteria.
Output: Selected memory, prompts, and scorer

```
Save the original VLM descriptions.
Identify errors and find the causes of errors by comparing mistakes with similar correct cases.
Ask the teacher LLM to propose a small change that adds missing event details.
Count supporting source videos and use a separate LLM call to check the evidence, people, and time span.
if the proposal passes this review then
    Apply the change to memory and prompts. Recompute affected outputs and features, reuse unchanged ones. Train the original and revised scorers using the same data and procedure.
    Freeze both systems and compare them on the same check windows.
    if the chosen metric improves and false alarms and missed detections stay within preset limits then
        Accept the revised memory, prompts, and scorer.
    else
        Keep the original system.
    end if
else
    Keep the original system.
end if
Save the proposal, review decision, and any evaluation results.
return Selected memory, prompts, scorer, and saved records.
```

sets. An update is accepted only if the chosen metric improves while false alarms and missed detections of other anomaly types remain within preset limits on check sets. Accepted patterns enter $\mathcal{P}^{\Delta}$ with their conditions and evaluation results. Rejected proposals are archived, and the current system is retained. Refinement stops after one round. Memory, prompts, and the scorer remain fixed at test time.

3.7 Stage 3: Frozen Evaluation

After Stage 2, we freeze the resource, case, pattern memories, VLM prompts, observation schema and other settings. Test windows follow the same retrieval and reasoning steps without teacher reflection, parameter updates, or memory changes. Test labels are used only for evaluation; sources already used for design remain development data.

4 Experiments

We evaluate Anomaly Agent on four VAD benchmarks, compare it with recent state-of-the-art methods, and analyze the effect of cue organization and memory update. Section 4.1 introduces datasets, 4.2 presents experiments, 4.3 shows ablation studies and 4.4 introduces qualitative results.

4.1 Experimental Setup

Datasets. We use **XD-Violence** Wu et al. (2020), **UCF-Crime** Sultani et al. (2018); Reiss & Hoshen (2022), and **UBnormal** Acsintoae et al. (2022). XD-Violence contains 4,754 videos (3,954 train / 800 test) with 6 anomaly categories. UCF-Crime contains 1,610 training videos and 290 test videos with 13 anomaly categories. UBnormal contains 268 training, 64 validation, and 211 test videos, with anomaly categories split across train/val/test. XD-Violence, UCF-Crime are weakly supervised while UBnormal has frame-level labels.

Metrics and implementation details. We process videos in 96-frame windows with a stride of 48 frames and sample eight evenly spaced images from each window. All visual reasoning uses the frozen Qwen3-VL-8B-Instruct Wang et al. (2024a), the counterexample-guided representation refinement uses Qwen3-8B Wang et al. (2024a). VLM retrieval retains four abnormal and six normal patterns using visual support weighted by confidence and utility, with diverse normal patterns and linked alternative explanations. No text encoder or similarity cutoff is used. C1 observes events, C2 verifies them, and separate pattern-matching calls use the same frames. More details about retrieval will be introduced in Section A.4 in appendix.

Table 2: Performance on UCF-Crime, XD-Violence, and UBnormal.

METHOD	UCF-Crime AUC (%)	XD-Violence AP (%)	METHOD	UBnormal AUC (%)
Wu et al. Wu et al. (2024b)	86.40	76.03	Micorek et al. Micorek et al. (2024)	72.8
Pu et al. Pu et al. (2024)	86.76	85.59	Yang et al. Yang et al. (2024a)	71.9
PI-VAD Majhi et al. (2025)	90.33	85.37	Leng et al. Leng et al. (2025)	74.6
MEVAD Huang et al. (2025)	87.80	85.13	Chen et al. Chen et al. (2025a)	78.4
Anomize Li et al. (2025a)	84.49	69.31	Flaborea et al. Flaborea et al. (2024)	65.0
ITC Liu et al. (2024)	89.04	85.45	Wang et al. Wang et al. (2025)	69.0
OVVAD Xu et al. (2025)	85.48	90.68	Wu et al. Wu et al. (2025)	74.1
MPD-FL Xiao et al. (2025)	88.35	82.19	Wu et al. Wu et al. (2024b)	62.9
EGL-CLIP Deng et al. (2026)	88.04	84.85	Wu et al. Wu et al. (2024c)	64.0
KSF Cai et al. (2026a)	89.41	84.79	Huang et al. Huang et al. (2026b)	74.7
LEC-VAD Wang & Chen (2025)	89.97	86.56	Xu et al. Xu et al. (2025)	64.3
DSRL Leng et al. (2024)	86.38	87.61	Xiao et al. Xiao et al. (2025)	67.3
MCANet Dev et al. (2024)	82.47	69.77	Tang et al. Tang et al. (2026)	78.2
HolmesVAD Zhang et al. (2024)	89.51	90.67	Zhu et al. Zhu et al. (2026)	75.4
LAVAD Zanella et al. (2024)	80.28	62.01	Ahn et al. Ahn et al. (2026)	74.5
VAU Zhang et al. (2025a)	88.96	87.68	Yang et al. Yang et al. (2025b)	75.8
VERA Ye et al. (2025)	86.55	70.54	Bekit et al. Bekit et al. (2026)	**79.6**
SUVAD Gao et al. (2025)	83.90	62.01	Dai et al. Dai et al. (2026)	76.5
MoniTor Yang et al. (2025a)	82.57	55.01	Leng et al. Leng et al. (2026)	75.8
SteerVAD Cai et al. (2026b)	87.15	83.02	Lin et al. Lin et al. (2025)	69.0
EventVAD Shao et al. (2025)	82.03	84.00	Kim et al. Kim et al. (2026)	72.2
VADTree Li et al. (2025b)	84.74	68.85	Huang et al. Huang et al. (2026a)	76.5
Ours	**90.89**	**91.01**	**Ours**	**79.7**

Table 3: Ablations on different components.

VARIANT	Memory guidance	C1	C2	Final scorer	XD-Violence	UCF-Crime	UBnormal
A0: Direct VLM	Off	Off	Off	VLM score	74.66	74.71	62.3
A1: VLM + Memory	On	Off	Off	Small model	85.92	85.17	71.8
A2: C1 only	On	On	Off	Small model	87.23	87.08	74.5
A3: VLM combiner	On	On	On	VLM combiner	88.72	88.91	76.5
A4: Full framework	On	On	On	Small model	90.89	91.01	79.7

The scorer takes strongest direct evidence, four C2 fractions (benign, harmful, unknown, and unresolved premise), the abnormal–normal matching difference, and six binary availability flags: 12 inputs. Evidence values use fitting-set mean imputation and standardization. Source videos are split into fitting, calibration, and check sets (60/20/20%). Logistic training uses equal total weight per source, binary cross-entropy with $\frac{0.1}{2}\|\gamma\|_2^2$, 30,000 iterations, and tolerance 10^{-7}.

4.2 Comparison with State-of-the-Art

In this section, the proposed approach is compared with state-of-the-art ones, including unsupervised or self-supervised methods, weakly supervised methods, structure-aware approaches, open-vocabulary approaches, multimodal methods, foundation-model-based reasoning methods. Improvements are observed in Table 2. On Ubnormal, although the anomaly types in training, validating and testing data are different, they share commonalities.

4.3 Ablation Study

We name each variant by the modules shown in Fig. 2. *On* means that a module is used, *Off* means that its path is bypassed.

Large-Model Evidence and Lightweight Scoring Table 3 follows the prediction path in Fig. 2. A0 asks the VLM for a score directly, bypassing memory retrieval and the structured C1/C2 path. A1 adds Resources/Cases/Patterns memory as guidance and the small scorer. A2 further adds C1 but disables C2. A3 and A4 use the full pipeline, including C2. Their only difference is the final



Figure 3: Qualitative results of memory-augmented VLM inference results.

combination: A3 replaces the trainable scorer with a VLM call, while A4 uses the learned weighted sum and sigmoid shown in the figure. The VLM weights remain frozen in every row. It can be seen that the components contribute.

Memory Banks and Retrieval Table 4 changes the upper Event memory and retrieval path in Fig. 2. R0 uses all three banks. R1 removes Resources, R2 removes Cases, and R3 removes Patterns. R4 searches full L2 records directly, replacing the L0–L1–L2 hierarchy. C1, the four event fields, C2, and the small scorer remain enabled. Retrieved guidance is regenerated for each row, so each comparison measures how the changed memory path affects the prediction. More details about the necessity of Resources, Cases and Patters will be shown in Section A.3 in appendix.

The ablations on representation will be shown in Section A.1 in appendix. Moreover, the performance is not significantly influenced by changing VLMs, as will be shown in Section A.3 in appendix.

Table 4: Ablations on different parts of memory.

VARIANT	Resources	Cases	Patterns	Retrieval	XD-Violence	UCF-Crime	UBnormal
R0: Full memory	On	On	On	Hierarchy	90.89	91.01	79.7
R1: Resources off	Off	On	On	Hierarchy	78.92	79.28	70.16
R2: Cases off	On	Off	On	Hierarchy	86.45	87.01	75.24
R3: Patterns off	On	On	Off	Hierarchy	84.53	85.17	73.86
R4: Flat retrieval	On	On	On	Direct L2	82.37	84.25	72.39

4.4 Subjective Results

Fig.3 shows qualitative results of VLM inference under the guidance of memory. In (a), memory patterns help in distinguishing fighting from sport activity. In (b), memory patterns contribute to better discrimination between reciprocal conflict and ceremony. In (c), memory helps VLM to focus on important features. In (d), memory patterns facilitate fine-grained discrimination.

5 Conclusion

Conclusions: We presented Anomaly Agent, a video anomaly detection framework that connects current observations with reusable experience through structured memory and event representations. Frozen VLMs examine events and check competing explanations, while a lightweight scorer learns how to combine direct evidence, explanation judgments, and pattern support, accounting for missing information. Counterexample-guided refinement updates representations and memory through validated changes before testing. This design combines large-model reasoning with task-specific scoring without updating VLM weights. State-of-the-art performance is achieved on multiple benchmarks. **Limitations:** Runtime performance is introduced in Section A.2 in appendix. Future work will focuse on improving efficiency.

AI USE STATEMENT

In this work, we used generative AI tools for language refinement and checking.

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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 20143–20153, 2022.

Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. Anyanomaly: Zero-shot customizable video anomaly detection with lvlm. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 3026–3035. IEEE, 2026.

Giacomo D' Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Francois Bremond, and Egor Bondarev. Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 10275–10285, 2025.

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.

Xiaojie Cai, Yucheng Qian, Chong Wang, Xiaohao Peng, Yuanbin Qian, and Jiafei Wu. Semantic boosting via knowledge sharing and feedback for video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.*, 36(5):6964–6975, 2026a.

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. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2026b.

Congqi Cao, Yue Lu, and Yanning Zhang. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *IEEE Trans. Image Process.*, 33:1810–1825, 2024.

Chenglizhao Chen, Xinyu Liu, Mengke Song, Luming Li, Shaojiang Yuan, Xu Yu, and Shanchen Pang. Unveiling context-related anomalies: Knowledge graph empowered decoupling of scene and action for human-related video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.*, 35(8), 2025a.

Chenglizhao Chen, Tianxiang Nan, Wen Li, Xinyu Liu, Guisheng Zhang, Mengke Song, and Xiaomin Yu. Structured evidence selection for weakly supervised video anomaly detection. *arXiv preprint arXiv:2607.10298*, 2026.

Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18319–18329, 2024a.

Junxi Chen, Liang Li, Yunbin Tu, Li Su, Zhe Xue, and Qingming Huang. Generalizing single-frame supervision to event-level understanding for video anomaly detection. *Proc. Adv. Neural Inf. Process. Syst.*, 38:6142–6173, 2025b.

Yingxian Chen, Jiahui Liu, Ruidi Fan, Yanwei Li, Chirui Chang, Shizhen Zhao, Wilton WT Fok, Xiaojuan Qi, and Yik-Chung Wu. Aligning effective tokens with video anomaly in large language models. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 22695–22706, 2025c.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024c.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024d.

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.

Jixiang Deng, Ying Liu, and Chunguang Li. Extended graph learning for weakly supervised video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.*, 36(3):3117–3130, 2026.

Prabhu Prasad Dev, Raju Hazari, and Pranesh Das. Mcanet: Multimodal caption aware training-free video anomaly detection via large language model. In *Proc. Int. Conf. Pattern Recognit. (ICPR)*, pp. 362–379. Springer, 2024.

Alessandro Flaborea, D'Amely Di Melendugno, Guido Maria, Stefano D'Arrigo, Marco Aurelio Sterpa, Alessio Sampieri, Fabio Galasso, et al. Contracting skeletal kinematics for human-related video anomaly detection. *Pattern Recognit.*, 156:110817, 2024.

Shibo Gao, Peipei Yang, and Linlin Huang. Suvad: Semantic understanding based video anomaly detection using mllm. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 1–5. IEEE, 2025.

Han Hu, Wenli Du, Peng Liao, Bing Wang, and Siyuan Fan. Noise-resistant video anomaly detection via rgb error-guided multiscale predictive coding and dynamic memory. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 19109–19119, 2025.

Chao Huang, Weiliang Huang, Qiuping Jiang, Wei Wang, Jie Wen, and Bob Zhang. Multimodal evidential learning for open-world weakly-supervised video anomaly detection. *IEEE Trans. Multimedia*, 27:3132–3143, 2025.

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*, 2026a.

He Huang, Zixuan Hu, Dong Feng, Zhuo Chen, Dongxiao Li, Hao Liu, and Lingyu Duan. Enabling real-world supervised video anomaly detection: New open-set benchmark and new framework. *IEEE Trans. Multimedia*, 28:4317–4331, 2026b. doi: 10.1109/TMM.2026.3660108.

Yashika Jain, Ali Dabouei, and Min Xu. Cross-domain learning for video anomaly detection with limited supervision. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 468–484. Springer, 2024.

Dongjun Kim, Changjae Oh, Andrea Cavallaro, and Jeonghoon Mo. Context-structured video anomaly detection with large vision-language models. *arXiv preprint arXiv:2607.19077*, 2026.

Jiaxu Leng, Zhanjie Wu, Mingpi Tan, Yiran Liu, Ji Gan, Haosheng Chen, and Xinbo Gao. Beyond euclidean: Dual-space representation learning for weakly supervised video violence detection. *Proc. Adv. Neural Inf. Process. Syst.*, 37:17373–17397, 2024.

Jiaxu Leng, Yumeng Zhang, Mingpi Tan, Changjiang Kuang, Zhanjie Wu, Ji Gan, and Xinbo Gao. Dual-space normalizing flow for unsupervised video anomaly detection. *IEEE Trans. Image Process.*, 34:6432–6445, 2025.

Jiaxu Leng, Jiankang Zheng, Mengjingcheng Mo, Zhanjie Wu, Haosheng Chen, Ji Gan, and Xinbo Gao. Linguistic relative policy optimization for video anomaly reasoning. *arXiv preprint arXiv:2607.00654*, 2026.

Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 29203–29212, 2025a.

Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. Vadtree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. *arXiv preprint arXiv:2510.22693*, 2025b.

Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. *Proc. Adv. Neural Inf. Process. Syst.*, 38:29713–29744, 2025.

Tianshan Liu, Kin-Man Lam, and Bing-Kun Bao. Injecting text clues for improving anomalous event detection from weakly labeled videos. *IEEE Trans. Image Process.*, 33:5907–5920, 2024.

Zihao Liu, Xiaoyu Wu, Jianqin Wu, Xuxu Wang, and Linlin Yang. Language-guided open-world video anomaly detection under weak supervision. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, volume 2026, pp. 153197–153224, 2026.

Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. Future frame prediction network for video anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7505–7520, 2021.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 24265–24274, 2025.

Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18868–18877, 2024.

Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Trans. Image Process.*, 33:4923–4936, 2024.

Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, Xiaochen Wang, Hao Tang, Yan Wang, and Shuyan Li. Eventvad: Training-free event-aware video anomaly detection. *arXiv preprint arXiv:2504.13092*, 2025.

Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. Video anomaly detection via sequentially learning multiple pretext tasks. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 10330–10340, 2023.

Haoyue Shi, Le Wang, Sanping Zhou, Gang Hua, and Wei Tang. Learning anomalies with normality prior for unsupervised video anomaly detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 163–180. Springer, 2024.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6479–6488, 2018.

Canhui Tang, Sanping Zhou, Haoyue Shi, and Le Wang. Action hints: Semantic typicality and context uniqueness for generalizable skeleton-based video anomaly detection. *Pattern Recognit.*, pp. 113898, 2026.

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Yingcong Chen. Hawk: Learning to understand open-world video anomalies. *Proc. Adv. Neural Inf. Process. Syst.*, 37:139751–139785, 2024.

Volcengine. OpenViking: An open-source context database designed specifically for AI agents. `https://github.com/volcengine/OpenViking`, 2026. [Online; accessed September 26, 2026].

Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 494–511. Springer, 2022.

Hongsong Wang, Andi Xu, Pinle Ding, and Jie Gui. Dual conditioned motion diffusion for pose-based video anomaly detection. In *Proc. AAAI Conf. Artif. Intell.*, volume 39, pp. 7700–7708, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024b.

Yu Wang and Shiwei Chen. Learning event completeness for weakly supervised video anomaly detection. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 62505–62517, 2025.

Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 729–745. Springer, 2022.

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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 322–339. Springer, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18297–18307, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18297–18307, 2024b.

Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proc. ACM Int. Conf. Multimedia (ACM MM)*, pp. 9301–9310, 2024c.

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 *Proc. AAAI Conf. Artif. Intell.*, volume 38, pp. 6074–6082, 2024d.

Ruituo Wu, Yang Chen, Jian Xiao, Bing Li, Jicong Fan, Frédéric Dufaux, Ce Zhu, and Yipeng Liu. Da-flow: Dual attention normalizing flow for skeleton-based video anomaly detection. *IEEE Trans. Multimedia*, 27:8847–8858, 2025.

Chizhuo Xiao, Yang Xiao, Joey Tianyi Zhou, and Zhiwen Fang. Multilingual-prompt-guided directional feature learning for weakly supervised video anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.

Chenting Xu, Ke Xu, Xinghao Jiang, and Tanfeng Sun. Plovad: Prompting vision-language models for open vocabulary video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.*, 35(6): 5925–5938, 2025.

Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 5527–5537, 2023.

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*, 2025a.

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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 14592–14601, 2023.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 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*, 2025b.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8679–8688, 2025.

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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18527–18536, 2024.

Guangyao Zhai, Yue Zhou, Xinyan Deng, Lars Heckler-Kram, Nassir Navab, and Benjamin Busam. Foundation visual encoders are secretly few-shot anomaly detectors. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, volume 2026, pp. 97353–97376, 2026.

Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16271–16280, 2023.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*, 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 13843–13853, 2025a.

Menghao Zhang, Huazheng Wang, Pengfei Ren, Kangheng Lin, Qi Qi, Haifeng Sun, Zirui Zhuang, Lei Zhang, Jianxin Liao, and Jingyu Wang. Do lvlms truly understand video anomalies? revealing hallucination via co-occurrence patterns. *Proc. Adv. Neural Inf. Process. Syst.*, 38:106738–106765, 2025b.

Yue Zheng, Xiufang Shi, Jiming Chen, and Yuanchao Shu. Cerberus: Real-time video anomaly detection via cascaded vision-language models. *arXiv preprint arXiv:2510.16290*, 2025.

Yiyan Zhu, Menghao Zhang, Haifeng Sun, Pengfei Ren, Xianao Chu, Chenye Xu, Hong Tan, Jinghan Wang, Qi Qi, and Jingyu Wang. Alert-clip: Abnormality-aware latent-enhanced representation tuning of clip for video anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 35545–35554, 2026.

Yuansheng Zhu, Wentao Bao, and Qi Yu. Towards open set video anomaly detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 395–412. Springer, 2022.

A Appendix

A.1 Ablations Study on Representations

Event Representation In Table 5, D0 keeps the full path. D1–D3 hide Action + response, Phase, or Evidence status from C2, then recompute the affected verification results and features. D4 removes scopes, so C2 considers the explanations without their assigned event, people, and time. It can be seen that all parts of the learned representation are useful.

Table 5: Representation-use ablations and matched-budget binding controls, located on the Core event representation–C2–Evidence features path in Fig. 2.

VARIANT	Downstream change	XD	UCF	UB
D0: Full fields	-	90.89	91.01	79.7
D1: Action + response hidden	Rerun C2; rebuild affected features; refit scorer	86.27	87.04	75.03
D2: Phase hidden	Rerun C2; rebuild affected features; refit scorer	87.18	87.24	75.81
D3: No evidence status	Rerun C2; rebuild affected features; refit scorer	83.41	84.25	73.49
D4: No scope	Rerun C2; rebuild affected features; refit scorer	87.08	87.52	75.26

A.2 Runtime Performance

In inference, the average latency per video segment (window) is 10.7s, with around 36.9 VLM calls on average. Peak GPU memory consumption is around 33GB.

A.3 Variants of VLMs

To validate the proposed approach, experiments are performed on different VLMs, including Qwen-VL in different scales, and other VLMs such as InternVL3 Chen et al. (2024b); Wang et al. (2024b); Chen et al. (2024c;d), as is shown in Table 6. It can be seen that performance is not significantly influenced by the changes in VLMs.

Table 6: Ablations on VLM variants.

VARIANT	XD-Violence	UCF-Crime	UBnormal
Qwen3-VL-8B	90.89	91.01	79.7
Qwen3-VL-32B	91.02	91.05	79.9
Qwen3-VL-4B	89.91	90.68	78.5
InternVL3-8B	89.93	90.77	79.2

A.4 Memory Organization and Conditional Pattern Matching

A frozen VLM selects relevant memory patterns, then checks how well their components, called nodes, match the sampled video frames. Optimal transport (OT) combines the frame evidence into pattern scores. These matching calls are separate from C1 observation and C2 explanation checking.

Memory organization. Resources describe what to inspect, cases store individual examples, and patterns summarize what examples have in common. Each item has L0 tags, an L1 summary with conditions, and L2 observations with source references. Table 7 shows how these roles are used in the archived matching runs. A pattern groups related nodes under one event description. For example, forceful contact together with resistance supports a conflict explanation. Matching checks their joint support without enforcing explicit causal or time-order edges.

Resources, cases, and patterns support different parts of the framework. Resources tell the VLM what to inspect, such as resistance or regained balance. Cases preserve concrete examples and their evidence for building patterns and reviewing errors. Patterns summarize what examples have in

common and when an explanation applies. Together, these memories connect reusable knowledge with visible evidence, helping distinguish similar-looking actions with different meanings.

Table 7: Memory roles and their use in the archived matching runs.

MEMORY	WHAT IT STORES	HOW IT IS USED
Resources	Node descriptions and visual clues that explain what to look for.	Included within pattern definitions and shown in the initial and conditional VLM prompts.
Cases	Source examples and references that show where a pattern came from.	Source references are stored in catalog metadata; these runs do not retrieve and show a separate set of cases to the VLM.
Patterns	Normal and abnormal explanations, described by related nodes and their weights.	Compact descriptions support selection; selected patterns are checked against the current frames.

Pattern selection. The VLM receives eight frames from a 96-frame window and compact pattern descriptions. In the archived runs, it proposes eight abnormal and twelve normal candidates, then retains four abnormal patterns $\mathcal{G}_A$ and six normal patterns $\mathcal{G}_N$ for matching. Selection uses visible evidence, without the query label. Normal candidates cover different possible explanations.

Conditional matching. Each frame has two identical evidence slots, giving $M = 16$ slots so that several nodes can share evidence from one frame. For node i and slot j, the initial VLM estimates presence $p_i^{(0)}$, absence $z_i^{(0)}$, location $\ell_{ij}^{(0)}$, and evidence quality $q_{ij}^{(0)}$. Location weights sum to one; each frame's weight is split equally between its slots. The initial support is

$$a_{ij}^{(0)} = p_i^{(0)} \sqrt{q_{ij}^{(0)} \ell_{ij}^{(0)}}. \tag{10}$$

An initial OT solve matches nodes to frames. A second VLM call checks each pattern using the same frames, its full description, and the initial estimates and matches. It revises the estimates (superscript g) and gives coherence c_g, which measures whether the nodes describe one event:

$$\begin{aligned} \bar{p}_i &= \sqrt{p_i^{(0)} p_i^{(g)}}, \\ a_{ij} &= \bar{p}_i \left(q_{ij}^{(0)} \ell_{ij}^{(0)} q_{ij}^{(g)} \ell_{ij}^{(g)} \right)^{1/4}, \\ z_i &= \sqrt{z_i^{(0)} z_i^{(g)} (1 - \bar{p}_i)}. \end{aligned} \tag{11}$$

For a pattern with K nodes, we add an absence column and an unused-evidence row, allowing unsupported nodes and unused frame slots:

$$\widetilde{A} = \begin{bmatrix} [a_{ij}]_{K \times M} & [z_i]_{K \times 1} \\ 0.5\, \mathbf{1}_{1 \times M} & 1 \end{bmatrix}. \tag{12}$$

OT finds a soft matching matrix P with row totals $\mu = (\mathbf{1}_K, M)/(K + M)$ and column totals $\nu = (\mathbf{1}_M, K)/(K + M)$:

$$P^\star = \underset{P \geq 0,\ P\mathbf{1} = \mu,\ P^\top \mathbf{1} = \nu}{\arg\max} \big[\langle P, \log \widetilde{A} \rangle + H(P) \big], \qquad H(P) = -\sum_{ij} P_{ij} \log P_{ij}. \tag{13}$$

Both OT solves use 60 Sinkhorn iterations; the initial solve substitutes $a_{ij}^{(0)}$ and $z_i^{(0)}$. Initial estimates are reused across patterns that share nodes. Table 8 gives the numerical settings.

Pattern scores. The final matching gives each node a presence value ρ_i and support score s_i:

$$\rho_i = \frac{\sum_{j=1}^{M} P_{ij}}{\sum_{j=1}^{M+1} P_{ij}}, \qquad s_i = \rho_i \frac{\sum_{j=1}^{M} P_{ij} a_{ij}}{\sum_{j=1}^{M} P_{ij} + \epsilon}. \tag{14}$$

A pattern combines these scores using the stored node weights w_i and coherence c_g:

$$S_g = \exp\left(\frac{\sum_i w_i \log \max(s_i, \epsilon)}{\sum_i w_i} \right) \max(c_g, \epsilon)^{0.25}. \tag{15}$$

Both required and optional nodes contribute. The weights affect scoring, while the required-node threshold is used only to report coverage. We aggregate each candidate group and subtract normal support from abnormal support:

$$\Phi(\mathcal{G}) = \tau \log \left[\frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} e^{S_g/\tau} \right], \qquad m = \Phi(\mathcal{G}_A) - \Phi(\mathcal{G}_N), \qquad \tau = 0.1. \tag{16}$$

The group-size normalization accounts for the unequal candidate counts. In the current framework, m is one scorer input alongside C1 and C2 features. The final decision threshold is calibrated. The archived matcher instead uses the fixed margin rule in Table 8, with uncertain decisions mapped to zero only for historical binary metrics.

Table 8: Settings for conditional pattern matching in the archived runs. A and N denote abnormal and normal candidates.

SETTING	VALUE
Input	96-frame window; 8 sampled frames; 2 slots per frame (16 total)
Candidate selection	Request 8 A / 12 N; retain 4 A / 6 N for matching
VLM output bounds	$[0.01, 0.99]$ for presence, absence, quality, and coherence
OT solver	60 iterations per solve; entropy coefficient 1
Dustbin values	Unused evidence: 0.5; dustbin-to-dustbin: 1
Numerical guard	$\epsilon = 10^{-9}$
Coherence exponent	0.25
Aggregation temperature	$\tau = 0.1$
Archived decision rule	Abnormal: $m > 0.03$; normal: $m < -0.03$; otherwise uncertain
Required-node coverage	Presence threshold 0.5; diagnostic only

Complete pattern catalog. Tables 9 and 10 contain 54 abnormal and 36 normal patterns, with 201 node entries and 184 distinct node IDs. Each entry retains its description, ID, required or optional status, and weight. Labels reproduce the stored hypotheses; context such as news branding alone cannot establish whether the current event is normal. "Required" marks a core component of the stored pattern, while "optional" marks additional supporting evidence. The weight w controls the node's relative contribution to the pattern score: $w = 1$ gives unit weight, and smaller values give less influence.

Table 9: Examples of abnormal patterns and their complete node lists in memory.

NO.	PATTERN	NODES
1	Aftermath of violent confrontation with injured person `aftermath_violent_standoff_injured_person`	Tense crowd or group facing (required; $w = 1$) `tense_crowd_gathering`
		Person lying on ground receiving non-violent attention (required; $w = 1$) `person_on_ground_receiving_attention`
		Scattered debris or overturned objects (required; $w = 1$) `debris_or_disorder_evidence`
		Smoke or dust plume in the air (optional; $w = 0.5$) `smoke_or_dust_plume`
2	Animal violence scene `animal_violence_scene`	human gripping animal (required; $w = 1$) `human_gripping_animal`
		striking or cutting action (required; $w = 1$) `striking_cutting_action`
		blood or open wound (optional; $w = 0.5$) `blood_or_open_wound`

Continued on the next page

Table 9 continued

NO.	PATTERN	NODES
3	Armed confrontation without physical contact armed_confrontation_standoff	Weapon aimed at a visible person (required; $w = 1$) weapon_aiming_at_visible_target
		Target shows distress or defensive reaction (required; $w = 1$) target_visible_distress
4	Bloody injury aftermath with distress bloody_injury_aftermath	Visible blood or wound (required; $w = 1$) visible_blood_or_wound
		Person in visible distress or shock (required; $w = 1$) person_in_distress_or_shock
5	Building destructive fire aftermath building_destructive_fire_aftermath	Fire burning on building structure (required; $w = 1$) fire_on_building_structure
		Scattered debris or damaged environment (required; $w = 1$) debris_or_damage_scene
		Smoke or dust plume (optional; $w = 0.5$) smoke_or_dust_plume
6	Civil unrest crowd-police physical confrontation civil_unrest_crowd_police_physical_confrontation	Physical shoving between crowd and authorities (required; $w = 1$) physical_shoving_between_crowd_and_authorities
		Aggressive crowd posturing towards authorities (required; $w = 1$) aggressive_crowd_posturing
7	Civilian confronting military forces civilian_military_confrontation	Individual civilian in confrontational posture (required; $w = 1$) individual_civilian_confronting
		Military or police vehicles / personnel (required; $w = 1$) military_or_police_force
8	Weapon discharge in civilian context civilian_weapon_discharge	Weapon discharge produces smoke/flash (required; $w = 1$) weapon_discharge_smoke
		Civilian environment (required; $w = 0.8$) civilian_environment
9	Close-call or near-miss driving incident close_call_driving	Abrupt evasive steering or hard braking (required; $w = 1$) evasive_manoeuvre
		Imminent collision threat (other vehicle, pedestrian, or obstacle) (required; $w = 1$) imminent_collision_threat
10	Confrontational road occupation confrontational_road_occupation	crowd occupying road (required; $w = 1$) crowd_occupying_road
		Tense group facing (required; $w = 1$) tense_group_facing
11	Crowd altercation with background fire crowd_altercation_with_fire	Large fire in background (required; $w = 1$) large_fire_background
		Physical altercation in crowd (required; $w = 1$) physical_altercation
12	Crowd escalation chain crowd_escalation_chain	opposing or aggressive crowd motion (required; $w = 1$) opposing_crowd_motion
		visible disorder or panicked scatter (required; $w = 1$) visible_disorder_or_scatter
		damage or debris aftermath (optional; $w = 0.5$) damage_or_debris
13	Dangerous vehicle cut-off and collision dangerous_vehicle_cutoff_and_collision	Abrupt side intrusion by another vehicle (required; $w = 1$) abrupt_side_intrusion
		Collision or crash indicator (required; $w = 1$) collision_or_crash_indicator

Continued on the next page

Table 9 continued

NO.	PATTERN	NODES
14	Destructive fire aftermath `destructive_fire_aftermath`	Fire burning on ground (required; $w = 1$) `ground_fire_or_burning`
		Scattered debris or damaged environment (required; $w = 1$) `debris_or_damage_scene`
		Smoke or dust plume (optional; $w = 0.5$) `smoke_or_dust_plume`
15	Error or warning overlay on real footage `error_overlay_on_real_footage`	persistent error or warning overlay (required; $w = 1$) `persistent_error_or_warning_overlay`
		real-world scene in continuous motion (required; $w = 1$) `real_world_scene`
16	Explosion flash on road `explosion_flash_on_road`	source-centered bright bloom (required; $w = 1$) `source_bloom`
		road vehicle visible (required; $w = 1$) `road_vehicle_visible`
17	Extremist ideological protest `extremist_ideological_protest`	extremist slogans or symbols visible (required; $w = 1$) `extremist_slogans_visible`
		public gathering context (required; $w = 1$) `public_gathering_context`
18	Violent physical fight between individuals `individual_physical_fight`	Physical aggression or combat (required; $w = 1$) `physical_aggression`
		Environmental disturbance or damage (required; $w = 0.8$) `environmental_disturbance`
		Person recoiled, thrown, or visibly hurt (required; $w = 0.7$) `person_recoil_or_injury`
19	Localized discharge / emission chain `localized_discharge_chain`	point flash or point smoke at one source (required; $w = 1$) `point_flash_or_smoke`
		aiming posture or held object (optional; $w = 0.5$) `aim_or_held_object`
		instant reaction or impact consequence (optional; $w = 0.5$) `instant_reaction_or_impact`
20	Military firearm brandishing with restraint `military_firearm_brandishing_restraint`	Firearm aimed at another person (required; $w = 1$) `weapon_aimed_at_person`
		Target person restrained or in vulnerable posture (required; $w = 1$) `person_restrained_or_vulnerable`
21	Near-miss or sudden driving maneuver with passenger reaction `near_miss_driving_event`	vehicle driving scene (required; $w = 1$) `vehicle_driving_scene`
		sudden vehicle maneuver or hard braking (optional; $w = 0.8$) `sudden_maneuver_or_braking`
		passenger startle or jerks in reaction (required; $w = 1$) `passenger_startle_reaction`
22	News branded crowd disorder `news_branded_crowd_disorder`	persistent news branding (required; $w = 1$) `news_branding_persistent`
		crowd in disordered flight or scatter (required; $w = 1$) `crowd_disorder_flight`
23	News branded extremist label `news_branded_extremist_label`	persistent news branding (required; $w = 1$) `news_branding_persistent`
		chyron with extremist label (required; $w = 1$) `extremist_label_chyron`
24	News branded protest march `news_branded_protest_march`	persistent news branding (required; $w = 1$) `persistent_news_branding`
		organized protest crowd with signs or banners (required; $w = 1$) `protest_crowd_with_signage`

Continued on the next page

Table 9 continued

NO.	PATTERN	NODES
25	News branded riot police weapon aim `news_branded_riot_police_weapon_aim`	persistent news branding (required; $w = 1$) `persistent_news_branding`
		riot police officer aiming less-lethal launcher (required; $w = 1$) `riot_police_weapon_aim`
26	News reported violent aftermath `news_reported_violent_aftermath`	Chyron/text overlay mentioning violence (required; $w = 1$) `chyron_mentioning_violence`
		Crowd disorder or dispersal (required; $w = 1$) `crowd_disorder_or_dispersal`
27	News footage of riot police standoff `news_riot_police_standoff`	persistent news branding (required; $w = 1$) `news_branding_persistent`
		dense formation of riot police (required; $w = 1$) `riot_police_formation`
28	Obscured-source explosion / large destructive blast `obscured_source_explosion`	outward expansion or ejecta (required; $w = 1$) `outward_expansion_or_ejecta`
		smoke or damage aftermath (required; $w = 0.8$) `smoke_or_damage_aftermath`
29	Perilous road maneuver or obstacle `perilous_road_maneuver_or_obstacle`	abrupt evasive vehicle maneuver (required; $w = 1$) `sudden_vehicle_maneuver`
		road hazard or stopped vehicle ahead (required; $w = 1$) `road_hazard_or_stopped_vehicle`
30	Physical assault / violent struggle `physical_assault_or_struggle`	Violent physical confrontation (required; $w = 1$) `violent_physical_confrontation`
		Victim in distress or on ground (required; $w = 1$) `victim_distress_or_recumbent`
		Disturbed setting / aftermath (optional; $w = 0.6$) `disturbed_setting`
31	Player brawl with no immediate official separation `player_brawl_no_immediate_separation`	multiple players aggressively shoving or grappling (required; $w = 1$) `players_engaged_in_violent_shoving`
		referee does not immediately separate the players (required; $w = 1$) `delayed_or_absent_referee_intervention`
32	Police-crowd standoff `police_crowd_standoff`	visible crowd on street (required; $w = 1$) `visible_crowd_standing`
		police line or riot police visible (required; $w = 1$) `police_line_visible`
33	Post-collision aftermath (no impact visible) `post_collision_aftermath`	Damaged vehicle visible (required; $w = 1$) `damaged_vehicle_in_scene`
		Post-crash context or debris (optional; $w = 0.5$) `post_crash_debris_or_context`
34	Protest blockade without vandalism or violence `protest_blockade_no_vandalism`	crowd occupying road (required; $w = 1$) `crowd_occupying_road`
		public transit vehicle stopped by crowd (required; $w = 1$) `blocked_public_transit_vehicle`
35	Protest confrontation scene `protest_confrontation_scene`	visible crowd or group of people on street (required; $w = 1$) `visible_crowd`
		police vehicle, water cannon, or visual sign of authority response (required; $w = 1$) `authority_vehicle_or_response`
		overturned objects, debris, or scattered items (optional; $w = 0.5$) `disorder_evidence`
36	Protest fire and crowd escalation `protest_fire_crowd_escalation`	large fire or burning object in public space (required; $w = 1$) `large_fire_or_burning_object`
		disorderly crowd or protest activity (required; $w = 1$) `disorderly_crowd_or_protest`
		smoke or haze aftermath (optional; $w = 0.5$) `smoke_or_haze_aftermath`

Continued on the next page

Table 9 continued

NO.	PATTERN	NODES

37	Protest vehicle speaker addressing crowd
	`protest_vehicle_speaker`
		person standing on vehicle with megaphone or addressing crowd (required; $w = 1$)
		`person_standing_on_vehicle_speaking`
		crowd gathered with protest banners (required; $w = 1$)
		`protest_crowd_with_banners`

38	Street protest with visible spray paint vandalism
	`protest_with_spray_paint_vandalism`
		crowd occupying road (required; $w = 1$)
		`crowd_occupying_road`
		visible spray paint graffiti (required; $w = 1$)
		`visible_spray_paint_markings`

39	Protesters physically attack a police van
	`protesters_attack_police_van`
		police van or official law enforcement vehicle present (required; $w = 1$)
		`police_van_visible`
		protesters physically attacking the van (required; $w = 1$)
		`physical_attack_on_van`

40	Reciprocal conflict chain
	`reciprocal_conflict_chain`
		forceful interpersonal contact (required; $w = 1$)
		`forceful_interpersonal_contact`
		return action or resistance (required; $w = 1$)
		`return_action_or_resistance`
		imbalance or fall caused by the clash (optional; $w = 0.5$)
		`imbalance_or_fall`

41	Reckless vehicle occupation by crowd
	`reckless_vehicle_occupation`
		person standing or climbing on vehicle roof (required; $w = 1$)
		`person_on_vehicle_roof`
		crowd surrounding vehicle (required; $w = 1$)
		`crowd_surrounding_vehicle`

42	Single vehicle loss of control leading to visible damage
	`single_vehicle_loss_of_control_crash`
		vehicle suddenly loses trajectory (required; $w = 1$)
		`vehicle_trajectory_break`
		damaged or overturned vehicle visible (required; $w = 1$)
		`damaged_vehicle_visible`

43	Smoke aftermath with stationary vehicles
	`smoke_aftermath_stationary_vehicles`
		Persistent smoke plume (required; $w = 1$)
		`persistent_smoke_plume`
		Stationary vehicles near incident (required; $w = 1$)
		`stationary_road_vehicles`

44	Source-centered blast chain
	`source_centered_blast_chain`
		source-centered bright bloom (required; $w = 1$)
		`source_bloom`
		outward expansion or ejecta (required; $w = 1$)
		`outward_expansion_or_ejecta`
		smoke or damage aftermath (optional; $w = 0.5$)
		`smoke_or_damage_aftermath`

45	Street destruction aftermath
	`street_destruction_aftermath`
		damaged traffic infrastructure (required; $w = 1$)
		`damaged_traffic_infrastructure`
		fire or smoke on street (required; $w = 1$)
		`fire_or_smoke_on_street`

46	Tear gas deployment in civil unrest
	`tear_gas_deployment_civil_unrest`
		expanding tear gas or smoke cloud (required; $w = 1$)
		`expanding_tear_gas_cloud`
		civilian unrest environment (required; $w = 1$)
		`civilian_unrest_environment`

47	Truck rollover crash with debris and dust
	`truck_overturn_crash_debris`
		Jackknifed or overturned large vehicle (required; $w = 1$)
		`jackknifed_or_overturned_vehicle`
		Debris or scattered wreckage on road (required; $w = 1$)
		`debris_or_scattered_wreckage`
		Dust or debris cloud from impact (optional; $w = 0.5$)
		`dust_or_debris_cloud`

Continued on the next page

Table 9 continued

NO.	PATTERN	NODES
48	Unilateral control / blocked escape chain `unilateral_control_chain`	one-sided control or restraint (required; $w = 1$) `one_sided_control`
		victim distress or defensive posture (required; $w = 1$) `victim_distress_or_defense`
		blocked escape or confinement (optional; $w = 0.5$) `blocked_escape_or_confinement`
49	Urban conflict aftermath with multiple smoke plumes `urban_conflict_aftermath_smoke_plumes`	Wide urban cityscape with multiple buildings (required; $w = 1$) `wide_cityscape_with_buildings`
		Multiple distinct smoke or dust plumes (required; $w = 1$) `multiple_smoke_or_dust_plumes`
50	Vehicle collision / crash `vehicle_collision`	Sudden impact or violent jolt (required; $w = 1$) `sudden_impact_or_jolt`
		Resulting damage or debris (required; $w = 1$) `damage_or_debris`
51	Vehicle collision chain `vehicle_collision_chain`	vehicle or traffic object visible (required; $w = 1$) `vehicle_or_traffic_object`
		impact or hard trajectory break (required; $w = 1$) `impact_or_trajectory_break`
		collision aftermath (optional; $w = 0.5$) `collision_aftermath`
52	Vehicle push collision with debris under vehicle `vehicle_push_collision_with_debris`	vehicle actively pushing object/person (required; $w = 1$) `vehicle_actively_pushing`
		debris or victim trapped under vehicle (required; $w = 1$) `debris_victim_underneath`
53	Vehicle stunt driving (drift, spin, hard maneuver) without collision `vehicle_stunt_driving`	moving vehicle visible (required; $w = 1$) `moving_vehicle`
		aggressive driving maneuver (drift, spin, hard turn) (required; $w = 1$) `aggressive_driving_maneuver`
54	Victim on ground with nearby person during violent episode `victim_on_ground_with_nearby_person`	person lying on ground in distress (required; $w = 1$) `person_lying_on_ground_distressed`
		another person standing or moving near the victim (required; $w = 1$) `person_standing_or_moving_near_victim`

Table 10: All 36 normal patterns and their complete node lists.

NO.	PATTERN	NODES
1	Animal food preparation handling `animal_food_preparation`	human holds animal (required; $w = 1$) `human_holds_animal`
		food preparation context (required; $w = 1$) `food_preparation_context`
2	Armed threat without physical contact or discharge `armed_threat_without_discharge`	Weapon pointed at a person (required; $w = 1$) `weapon_pointed_at_person`
		Victim shows submission or fear (required; $w = 1$) `victim_submission_or_fear`
		No visible weapon discharge or physical contact (required; $w = 0.75$) `no_discharge_or_contact`
3	Brief sports scuffle with referee intervention `basketball_foul_scuffle`	Referee physically intervenes (required; $w = 1$) `referee_physical_intervention`
		Brief cluster of players shoving (required; $w = 1$) `brief_player_shoving_cluster`

Continued on the next page

Table 10 continued

NO.	PATTERN	NODES
4	Camera flash / edit confound camera_flash_or_edit	frame-wide or source-unclear flash (required; $w = 1$) frame_wide_or_source_unclear_flash
		no local consequence (required; $w = 1$) no_local_consequence
5	Ceremonial / calm crowd gathering ceremonial_crowd_gathering	crowd present (required; $w = 1$) crowd_presence
		organized or calm pattern (required; $w = 1$) organized_or_calm_pattern
6	Staged cinematic civilian shootout cinematic_civilian_shootout	Weapon discharge produces smoke/flash (required; $w = 1$) weapon_discharge_smoke
		Cinematic film-production context (required; $w = 0.9$) cinematic_production_context
7	Standard combat weapon discharge combat_weapon_discharge	weapon discharge produces smoke/flash (required; $w = 1$) weapon_discharge_smoke
		soldier aiming weapon before discharge (required; $w = 0.8$) soldier_aiming
8	Game / simulation footage game_simulation_footage	persistent game HUD or minimap visible (required; $w = 1$) persistent_game_hud_or_minimap
		vehicle with large decals / branding (required; $w = 1$) vehicle_with_simulation_livery
9	High-pressure liquid discharge / debris flushing high_pressure_liquid_discharge	Sudden burst or spray of liquid (required; $w = 1$) sudden_liquid_burst
		Debris scattering or sustained water flow (required; $w = 1$) debris_scattering_or_water_flow
10	Industrial hazard scene without explosion industrial_hazard_without_blast	Industrial or factory interior (required; $w = 1$) industrial_factory_environment
		Continuous sparks or small ground fires (required; $w = 1$) persistent_sparks_and_fire
		Character moving through the environment (required; $w = 1$) character_moving_through_scene
11	Large outdoor fire large_outdoor_fire	Large sustained flames (required; $w = 1$) large_sustained_flames
		Smoke column rising from fire (required; $w = 0.8$) smoke_column
12	Military vehicle interior with external combat activity military_vehicle_interior_external_combat	Soldiers in or near military vehicle (required; $w = 1$) soldiers_in_or_near_vehicle
		External combat fire or smoke (required; $w = 1$) external_combat_fire_or_smoke
13	News branded footage of dramatic event news_branded_footage	small news logo or bug (required; $w = 1$) news_logo_or_bug
		live-action event footage (optional; $w = 0.5$) scene_footage
14	News overlay on simulated/computer-generated footage news_overlay_on_simulation	persistent news-style overlay or bug (required; $w = 1$) news_broadcast_overlay
		simulated, game, or non-photoreal visuals (required; $w = 1$) simulated_or_game_visuals
15	Non-violent dust cloud or smoke plume non_violent_dust_cloud	Dense dust or smoke cloud (required; $w = 1$) dense_dust_cloud
		Non-violent surrounding context (required; $w = 1$) non_violent_context

Continued on the next page

Table 10 continued

NO.	PATTERN	NODES
16	Non-violent standoff with ground assistance non_violent_standoff_with_assistance	Tense group facing (required; $w = 1$) tense_group_facing
		Person on ground receiving non-violent attention (required; $w = 1$) ground_person_physical_interaction
		No visible physical fight (required; $w = 1$) no_visible_fight
17	Normal driving sequence normal_driving	vehicle or road traffic scene visible (required; $w = 1$) vehicle_or_road_scene
		smooth continuous driving motion (required; $w = 1$) smooth_driving_motion
18	Normal driving flow normal_driving_flow	road driving scene with vehicles (required; $w = 1$) road_driving_scene
		steady forward motion of all visible vehicles (required; $w = 1$) steady_forward_motion
19	Normal highway driving normal_highway_driving	road vehicle visible (required; $w = 1$) road_vehicle_visible
		smooth traffic flow (required; $w = 1$) smooth_traffic_flow
20	Orderly crowd watching an event orderly_crowd_watching	crowd behind a barrier (required; $w = 1$) crowd_behind_barrier
		calm, non-aggressive behavior (required; $w = 1$) calm_behavior
21	Regular traffic scene regular_traffic_scene	moving vehicle on road (required; $w = 1$) moving_vehicle
		road environment visible (required; $w = 1$) road_environment
22	Regulated Ice Hockey Game regulated_ice_hockey_game	visible ice hockey rink (required; $w = 1$) ice_hockey_rink
		referee present on ice (required; $w = 1$) referee_on_ice
23	Rescue amidst aftermath rescue_amidst_aftermath	destroyed area with fire and debris (required; $w = 1$) aftermath_destruction_evidence
		rescue or evacuation action (required; $w = 1$) rescue_or_evacuation_action
24	Rescue / assist interaction rescue_or_assist_interaction	one-sided contact (required; $w = 1$) one_sided_contact
		assistive direction (required; $w = 1$) assistive_direction
25	Roadside military/police stop or checkpoint roadside_military_stop_or_checkpoint	Military or police vehicle present (required; $w = 1$) military_or_police_vehicle
		Civilian vehicle stopped on roadside (required; $w = 1$) civilian_vehicle_stopped_roadside
26	Simulated / game driving footage simulated_or_game_driving_footage	persistent static graphic overlay (required; $w = 1$) persistent_static_graphic_overlay
		dynamic first-person driving view (required; $w = 1$) dynamic_driving_view
27	Legal sports body check sports_body_check_normal	forceful player contact (required; $w = 1$) forceful_player_contact
		sports environment context (required; $w = 1$) sports_environment
28	Sports / ceremonial contact sports_or_ceremonial_contact	contact present (required; $w = 1$) contact_present
		structured nonviolent context (required; $w = 1$) structured_nonviolent_context

Continued on the next page

Table 10 continued

NO.	PATTERN	NODES
29	Stage effect / fireworks confound `stage_effect_or_fireworks`	show-like bright effect (required; $w = 1$) `show_like_bright_effect`
		celebration or stage context (required; $w = 1$) `celebration_or_stage_context`
30	Static injured person in debris aftermath `static_injured_person_in_debris_aftermath`	person lying injured or unconscious (required; $w = 1$) `person_lying_injured_or_unconscious`
		debris or damage in surroundings (required; $w = 1$) `surrounding_debris_or_damage`
		no active violence or fast motion (optional; $w = 0.5$) `no_active_violence_or_fast_motion`
31	Stationary traffic scene normal confound `stationary_traffic_scene_normal_confound`	multiple vehicles visible (required; $w = 1$) `multiple_vehicles_present`
		no active collision or abrupt motion (required; $w = 1$) `no_active_impact_motion`
32	Text card / non-scene `text_card_non_scene`	frame dominated by text or graphics (required; $w = 1$) `frame_dominated_by_text_or_graphic`
33	Vehicle driving in dusty quarry/construction environment `vehicle_in_dusty_construction_environment`	Vehicle visible in dusty environment (required; $w = 1$) `vehicle_in_dusty_environment`
		Earthmoving equipment present (required; $w = 1$) `earthmoving_equipment_present`
34	Vehicle motion disturbance without collision `vehicle_motion_disturbance`	vehicle scene (required; $w = 1$) `vehicle_scene`
		motion disturbance only (required; $w = 1$) `motion_disturbance_only`
35	Vehicle stunt action sequence without real collision `vehicle_stunt_action_sequence`	Vehicle in dynamic motion (drifting, sliding, sparks) (required; $w = 1$) `vehicle_dynamic_motion`
		No direct collision impact between vehicles or with obstacles (required; $w = 1$) `no_collision_impact`
36	Game / simulation military shooter footage with HUD `video_game_military_shooter_footage`	Persistent game HUD or night-vision overlay (required; $w = 1$) `persistent_game_hud_or_night_vision_overlay`
		Weapon discharge effect (smoke/flash) (required; $w = 0.9$) `weapon_discharge_effect`
		Human figure in prone or aiming posture (optional; $w = 0.6$) `human_figure_prone_or_aiming`