ANOMALYRAG: RELATION-CENTRIC DUAL-BANK EVIDENCE RETRIEVAL FOR SURVEILLANCE ANOMALY DETECTION AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Surveillance anomaly detection and reasoning must identify context-dependent violations and explain the responsible entities and rules. Similar scenes can require opposite labels when object relations differ, yet global retrieval often overlooks small decisive objects. We present **AnomalyRAG**, a weakly supervised framework that equips a frozen VLM with Relation-Centric Dual-Bank Evidence Retrieval. It constructs normal and abnormal evidence banks from training data, reranks multi-route candidates through learned object–relation alignment, and derives a calibrated signed margin. Verified bilateral exemplars guide Structured Evidence-Grounded Reasoning Chain-of-Thought (SEGR-CoT), while reliability-gated symmetric fusion admits only grounded corrections. We evaluate detection on UCF-Crime and XD-Violence, reasoning on ECVA and UCF-Crime, and public-safety recognition on the 24-task SMAD dataset. AnomalyRAG improves detection and reasoning while training only lightweight retrieval and calibration modules.

1 INTRODUCTION

Traditional video anomaly detection (VAD) treats anomalies as predefined events or deviations from normal patterns (Hasan et al., 2016; Sultani et al., 2018; Tian et al., 2021; Wu et al., 2022; Li et al., 2022; Chen et al., 2023). Surveillance labels, however, often depend on scene rules and object relations (Yuan et al., 2024; Zhu et al., 2025; Zhang et al., 2025). In Figure 1, a helmetless worker is anomalous at a construction site but not necessarily elsewhere, illustrating context-dependent *anomaly drift*.

Vision-language models (VLMs) offer open-vocabulary perception and reasoning (Cheng et al., 2024; Lin et al., 2024; Bai et al., 2025b;a), but zero-shot surveillance judgments remain coarse or inconsistent (Zanella et al., 2024; Shao et al., 2025). They may recognize entities yet miss the interaction or rule that makes an event abnormal (Du et al., 2024; Zhu et al., 2025). Existing methods use fine-grained supervision (Zhu et al., 2025) or infer rules from normal data (Yang et al., 2024), but do not contrast matched normal and abnormal relation-level precedents.

Retrieval-augmented generation (RAG) grounds predictions in external precedents (Lewis et al., 2020), but global image retrieval can miss small objects and local relations, such as *person–uses–multiple rods*. Prototype memories can separate normal from abnormal representations (Zhou et al., 2023), yet they do not provide paired, inspectable precedents that expose the relation change responsible for a decision.

We introduce **AnomalyRAG**, built around *Relation-Centric Dual-Bank Evidence Retrieval*. Training data populate separate normal and abnormal banks of images, objects, relations, descriptions, and rules. Learned object–relation alignment retrieves matched precedents for a frozen VLM, whose Structured Evidence-Grounded Reasoning Chain-of-Thought (SEGR-CoT) contrasts rule satisfaction and violation. A reliability gate admits only grounded corrections to the evidence-free prediction. The VLM, detector, and visual encoder remain frozen; only the relation retriever and calibration modules are learned.

Our main contributions are:

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.

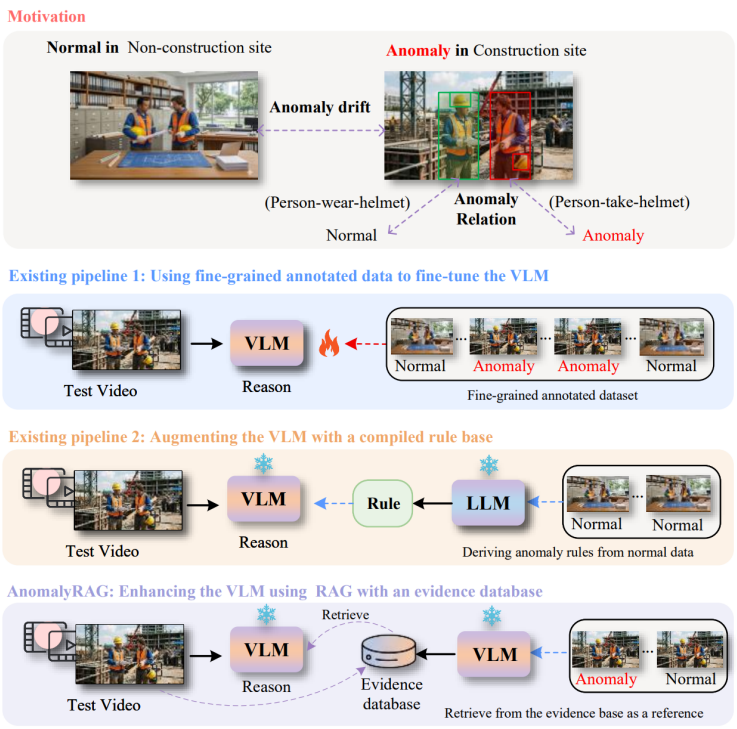

Figure 1: Context-dependent anomaly semantics. AnomalyRAG mitigates anomaly drift by retrieving structured evidence aligned with the decisive object relations.

- We introduce relation-aligned bilateral precedent comparison, which retrieves structured normal and abnormal instances and softly aligns their local object–relation motifs to the query.
- We develop **AnomalyRAG**, coupling a frozen VLM with a calibrated evidence margin, bilateral SEGR-CoT, and reliability-gated fusion.
- We evaluate detection and reasoning on three public benchmarks and introduce SMAD, comprising 24 task-specific binary alerting tasks for elevated-view public-safety monitoring.

2 Related Work

2.1 Video Anomaly Detection

VAD detects deviations from learned normality (Hasan et al., 2016) using weakly supervised, one-class, or unsupervised objectives (Sultani et al., 2018; Wang & Cherian, 2019; Zaheer et al., 2022). UR-DMU maintains normal/abnormal prototype memories (Zhou et al., 2023), but these representations lack inspectable relation-level precedents.

Open-vocabulary and VLM systems use prompting, instruction tuning, or verbalized learning (Wu et al., 2024a; Zanella et al., 2024; Zhang et al., 2024; Ye et al., 2025). Zou et al. (2026) elicit action cues; Shao et al. (2025) reason over events; Ding et al. (2025) and Yang et al. (2025) retrieve behavior patterns or scene-conditioned rules. AnomalyRAG instead contrasts relation-aligned normal and abnormal instances with structural reliability gating.

2.2 Video Anomaly Understanding and Evidence Retrieval

Anomaly understanding extends detection to description, temporal grounding, causal explanation, and object attribution (Yuan et al., 2024; Du et al., 2024; Zhu et al., 2025; Zhang et al., 2025). VADER models dynamic object relations, while SARD identifies rare relations from scene

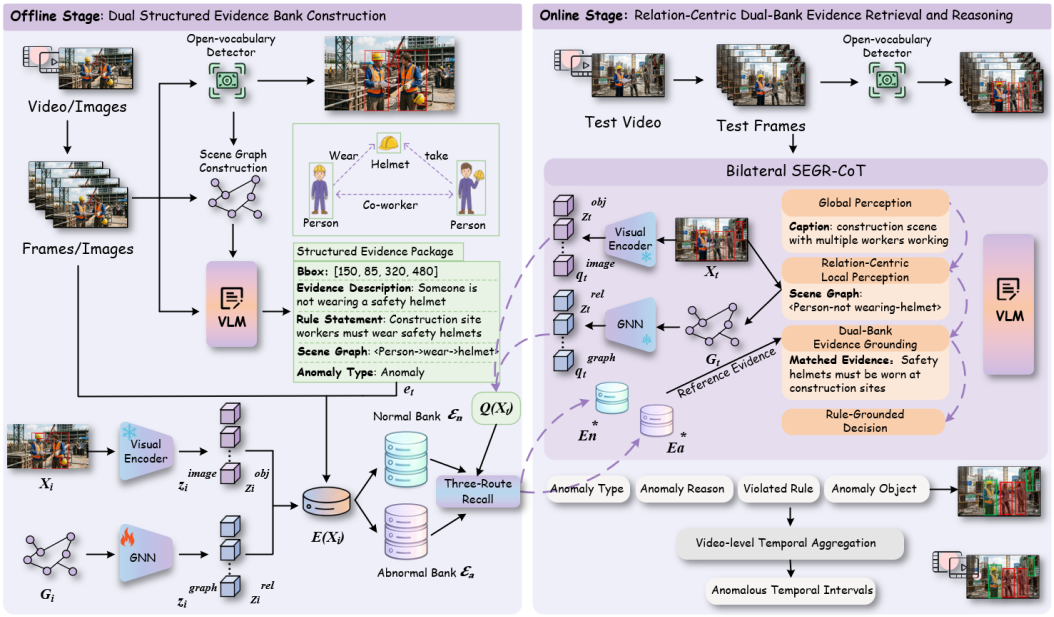

Figure 2: Overview of **AnomalyRAG**. Offline, an open-vocabulary detector and a VLM convert training samples into structured packages stored in normal and abnormal evidence banks. Online, relation-centric graph matching reranks multi-route candidates from each bank. Verified bilateral evidence conditions SEGR-CoT, while reliability gating prevents unsupported evidence from altering the evidence-free prediction.

graphs (Cheng et al., 2026; Lai et al., 2025). Neither jointly retrieves normal and abnormal relation-level precedents with their governing rule.

RAG and CLIP support evidence-grounded generation and cross-modal retrieval (Lewis et al., 2020; Radford et al., 2021). AnomalyRAG uses relation-aware structures and graph attention (Zeng et al., 2025; Brody et al., 2022) to compare local motifs across both banks, addressing global retrieval's tendency to overlook decisive object relations.

3 METHODOLOGY

Overview. As shown in Figure 2, AnomalyRAG constructs two evidence banks exclusively from training data and compares their relation-aligned precedents at inference. Given $V = \{X_t\}_{t=1}^{T}$, with $T = 1$ for an image, it predicts an anomaly score and category together with the responsible objects, matched relations, supporting evidence, and violated rule.

3.1 DUAL STRUCTURED EVIDENCE BANKS

Task-oriented evidence construction. For each training image or key frame X_i with label $y_i \in \{0, 1\}$, WeDetect (Fu et al., 2025) produces candidate objects

$$B_i = \{(b_{iv}, c_{iv}, s_{iv})\}_{v=1}^{n_i},$$

where b_{iv}, c_{iv}, and s_{iv} are the box, category, and confidence. A VLM converts the detections into a task-oriented graph $G_i = (V_i, \mathcal{R}_i)$ whose nodes are entities and edges encode anomaly-relevant geometry, interactions, and states. Free-form relations are mapped to a vocabulary fixed on the training split.

The VLM also generates an evidence description D_i and rule statement R_i that explain why an abnormal configuration violates the rule or a normal configuration satisfies it. Both are produced offline from the image, detections, and graph; query labels and answer-derived semantics are never available at inference.

Visual and graph representations. A frozen OpenCLIP ViT-L/14 encoder (Radford et al., 2021) extracts a global image feature and object-crop features,

$$z_i^{\text{image}} = \text{norm}(f_{\text{clip}}(X_i)),$$
$$\mathbf{v}_{iv} = \text{norm}(f_{\text{clip}}(X_i[b_{iv}])).$$

Graph nodes concatenate crop features, category text, and normalized box geometry, while edges combine relation text with relative geometry. A two-layer edge-aware GATv2 (Brody et al., 2022) updates node v as

$$\mathbf{m}_v^{(l)} = \sum_{u \in \mathcal{N}(v)} \alpha_{uv}^{(l)} \big(W_s \mathbf{h}_u^{(l)} + W_e \mathbf{e}_{uv}\big),$$
$$\mathbf{h}_v^{(l+1)} = \text{LN}\Big(\mathbf{h}_v^{(l)} + W_o \mathbf{m}_v^{(l)}\Big),$$

where $\alpha_{uv}^{(l)}$ depends on the source node, target node, and edge. Attention pooling yields a graph-level feature z_i^{graph}, while the node and edge tokens remain available for structural matching.

Relation motifs. The retrieval unit is a local *relation motif*

$$\mathcal{M}_{i,m} = (u, r_{uv}, v, \mathcal{N}_1(u), \mathcal{N}_1(v)),$$

containing a subject–relation–object triple and the endpoints' one-hop neighborhoods. It is represented as

$$\mathbf{r}_{i,m} = \text{norm}[\text{MLP}(\mathbf{h}_u \| \mathbf{e}_{uv} \| \mathbf{h}_v \| \, \text{AGG}_{w \in \mathcal{U}_{uv}} \, \mathbf{h}_w)] \, .$$

Here $\mathcal{U}_{uv} = \mathcal{N}_1(u) \cup \mathcal{N}_1(v)$. The neighborhood makes composition and cardinality explicit, distinguishing, for example, one person using multiple fishing rods from one person using a single rod. Unary tokens $\mathbf{o}_{iv}$ represent attribute-dominated anomalies such as smoke or road damage. We denote the resulting sets by Z_i^{obj} and Z_i^{rel}.

The complete evidence package is

$$E(X_i) = (X_i, y_i, B_i, G_i, D_i, R_i, z_i^{\text{image}}, z_i^{\text{graph}},$$
$$Z_i^{\text{obj}}, Z_i^{\text{rel}}, c_i, P_i),$$

where c_i is task metadata and P_i records provenance. Packages are separated by label into

$$\mathcal{E}_n = \{E(X_i) : y_i = 0\}, \qquad \mathcal{E}_a = \{E(X_i) : y_i = 1\}.$$

Only training samples enter the banks; validation supports model selection and calibration, and test samples are never retrievable evidence.

3.2 Relation-Centric Dual-Bank Evidence Retrieval

Structured query and coarse recall. The query X_t is processed by the same detector, graph constructor, OpenCLIP encoder, and GATv2 encoder:

$$Q(X_t) = (X_t, B_t, G_t, q_t^{\text{image}}, q_t^{\text{graph}}, Z_t^{\text{obj}}, Z_t^{\text{rel}}).$$

The query contains no label or answer-derived semantics. Each bank $b \in \{n, a\}$ maintains global, pooled-object, and pooled-relation indices. The three routes independently recall Top-M packages by cosine similarity; reciprocal-rank fusion forms a candidate union, capped at 100 packages per bank. We use $M = 50$ and apply the relation scorer only to this compact union.

Soft object and relation alignment. Let H_t and H_i be the normalized GATv2 object embeddings of the query and a candidate, with affinity $A^{\text{o}} = H_t H_i^{\top}$. Balanced transport would force every object to match. We instead append a learnable no-match dustbin and compute

$$\widetilde{P}^{\text{o}} = \text{PSinkhorn}(A^{\text{o}}/\tau_s, \alpha_{\text{o}}) \, , \qquad P^{\text{o}} = \widetilde{P}^{\text{o}}_{1:N_t, 1:N_i},$$

with $\tau_s = 0.07$ and 10 iterations. Dustbin mass represents unmatched evidence or query objects; the real–real block is not renormalized to unit total mass. The object score is its transport-weighted mean affinity. For relation motifs $r = (u, v)$ and $s = (a, b)$, endpoint support $C_{rs} = \sqrt{P^{\text{o}}_{ua} P^{\text{o}}_{vb}}$ modulates their cosine affinity before a second partial Sinkhorn produces S_{r}. Thus unrelated objects and motifs can remain unmatched.

Quality-adaptive multi-granular reranking. Let $\mathbf{s} = [S_{\mathrm{g}}, S_{\mathrm{o}}, S_{\mathrm{r}}, S_{\mathrm{c}}]$ collect global, object, relation, and rule scores. Non-dustbin coverage, graph quality, and rule quality define branch availability $\mathbf{q} \in [0,1]^4$. With $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{\theta})$, the normalized score is $S(Q, E_i) = \sum_j \pi_j q_j s_j / (\sum_j \pi_j q_j + \epsilon)$, so weak structure suppresses its branch without forcing a match.

Dual-bank aggregation and calibration. We retrieve the top-K packages independently from each bank,

$$\mathcal{R}_b(Q) = \mathrm{TopK}_{E \in \mathcal{E}_b} S(Q, E), \quad b \in \{n, a\}.$$

With $S_b^{(k)}$ the k-th score in bank b, a temperature-τ_a log-mean-exp gives $\bar{S}_b$. Validation statistics standardize the two banks independently, yielding the signed margin and retrieval prior

$$\Delta_t = \frac{\bar{S}_a - \mu_a}{\sigma_a} - \frac{\bar{S}_n - \mu_n}{\sigma_n}, \qquad p_t^{\mathrm{ret}} = \sigma(\Delta_t / T + \beta).$$

A positive margin favors abnormality and a negative margin favors normality; both precedents remain available to the VLM for explicit comparison.

Evidence reliability. The retriever estimates confidence from seven normalized cues:

$$\begin{aligned} c_t^{\mathrm{ret}} = {} & 0.10 c_{t,\mathrm{g}} + 0.15 c_{t,\mathrm{o}} + 0.25 c_{t,\mathrm{r/u}} + 0.15 c_{t,\mathrm{rule}} \\ & + 0.10 c_{t,\mathrm{route}} + 0.15 c_{t,\mathrm{bank}} + 0.10 c_{t,\mathrm{sep}}, \end{aligned}$$

where object and relation support are the maximum continuous non-dustbin coverage among retained packages. Route consensus is the largest fraction of the global, object, and relation recall routes supporting one retained package; bilateral bank support is one only when both banks return evidence. The remaining terms measure graph quality, rule compatibility, and bank-score separation. Validation selects τ_c to maximize coverage subject to a target evidence precision; a score below this threshold denotes *no reliable evidence*, not a normal prediction.

3.3 Learning the Relation Retriever

We freeze WeDetect, OpenCLIP, and the VLM and learn only the relation retriever and calibration modules. Training pairs each query with compatible same-label evidence and relation-incompatible opposite-label negatives; near-duplicate frames from the same source cannot occupy query and evidence roles.

We optimize dual-bank discrimination, contrastive retrieval, pairwise ranking, rule compatibility, and a graph-quality-aware visual anchor: $\mathcal{L} = \mathcal{L}_{\mathrm{db}} + \mathcal{L}_{\mathrm{NCE}} + 0.5\mathcal{L}_{\mathrm{rank}} + 0.2\mathcal{L}_{\mathrm{rule}} + 0.1\mathcal{L}_{\mathrm{anchor}}$. The anchor preserves global similarity when graph quality is weak; complete definitions are in the supplementary material.

3.4 Bilateral Evidence-Grounded Reasoning

We retain the top-K candidates for confidence estimation but pass at most one verified exemplar from each bank to the VLM. The prompt includes the query, aligned objects and relations, evidence descriptions and rules, the signed margin Δ_t, and retrieval confidence c_t^{ret}. If a bank has no reliable match, the prompt marks it unavailable instead of inserting weak evidence.

SEGR-CoT guides the frozen VLM through four evidence-constrained stages:

- **Global perception:** identify the environment and salient entities without assigning a label.
- **Relation-centric local perception:** inspect candidate objects, attributes, cardinalities, and local relation motifs.
- **Dual-bank evidence grounding:** compare the query separately with the normal and abnormal exemplars and state the decisive matches and mismatches.
- **Rule-grounded decision:** determine whether the query satisfies or violates the relevant rule and identify the responsible objects.

The VLM returns a fixed JSON record containing the decision, responsible objects, evidence IDs, matched relations, rule, confidence, and explanation. A deterministic verifier checks the schema, identifiers, relation endpoints, score ranges, and agreement between the reported rule and retrieved evidence. The verifier tests grounding consistency rather than re-perceiving the image.

3.5 Reliability-Gated Symmetric Evidence Fusion

Let p_t^{base} be the evidence-free anomaly score and $d_t^{\text{evi}} = p_t^{\text{evi}} - p_t^{\text{base}}$ the signed correction induced by bilateral evidence for query frame or image X_t. Combining retrieval confidence with the verifier output gives

$$g_t = \mathbb{1}[c_t^{\text{ret}} \geq \tau_c]\,\mathbb{1}[\text{verification passes}].$$

The final query score is

$$p_t^{\text{final}} = \text{clip}_{[0,1]}\left(p_t^{\text{base}} + g_t \eta d_t^{\text{evi}}\right).$$

The signed correction can raise an anomaly score or suppress a false alarm. Verifier failure or insufficient retrieval confidence sets $g_t = 0$, leaving the evidence-free score unchanged. Fusion parameters are selected on validation data and frozen for testing.

4 Experiments

4.1 Experimental Setup

We evaluate detection on UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020) under standard VAD protocols (Zanella et al., 2024; Shao et al., 2025). For reasoning, we follow VAU-R1 (Zhu et al., 2025) on ECVA (Du et al., 2024) and UCF-Crime.

UCF-Crime and XD-Violence provide video-level training labels. Banks and retriever learning use the complete official training partitions, with validation drawn only from training data. The frozen VLM admits visibly grounded violations to the abnormal bank and confidently normal frames to the normal bank; ambiguous candidates are discarded. Test videos are queries only. The supplement details admission and bank sizes.

SMAD contains 12,500 elevated-view images from 24 task-specific binary public-safety alerting tasks, balanced between normal and abnormal samples. Its 7,500/2,500/2,500 split is formed over 8,532 source and perceptual-hash groups, preventing near-duplicates from crossing partitions. Binary labels and task-relevant boxes are manually annotated. Available boxes initialize object nodes for every compared variant; otherwise the shared detector supplies candidates, so retrieval ablations receive identical query-side object evidence. All tasks share one fixed 40-class detector vocabulary; query labels, task identities, names, and rules are hidden, and only retrieved training evidence can supply a rule. Small targets and relation-dependent labels make SMAD a direct test of contextual judgment.

Evaluation Metrics. For detection, we report frame-level area under the receiver operating characteristic curve (AUC) on UCF-Crime, and AUC and average precision (AP) on XD-Violence (Zanella et al., 2024; Shao et al., 2025). Following VAU-R1 (Zhu et al., 2025), we use binary/multi-class accuracy and VAU-Eval for reasoning, question-answering accuracy on ECVA, and accuracy/F1 on SMAD.

Compared Methods. Tables 1 and 2 list detection and reasoning baselines. AnomalyRAG keeps the MLLM frozen; weak labels train only the relation retriever and calibration modules. InternVL2-8B matches VERA's backbone, while Qwen3-VL tests transfer to a stronger similarly sized VLM. MGFN uses I3D-RGB on UCF-Crime and VideoSwin-RGB on XD-Violence.

Implementation Details. Unless stated otherwise, the frozen reasoning backbone is Qwen3-VL-8B-Instruct (Bai et al., 2025a). WeDetect Base (Fu et al., 2025) proposes objects, OpenCLIP ViT-L/14 (Radford et al., 2021) encodes images and crops, and a two-layer edge-aware GATv2 (Brody et al., 2022) contextualizes scene graphs.

Table 1: Representative frame-level anomaly detection results on UCF-Crime and XD-Violence.

Method	Visual Backbone	UCF-Crime AUC (%)↑	XD-Violence AUC (%)↑	XD-Violence AP (%)↑
Weakly supervised methods				
RTFM (Tian et al., 2021)	I3D-RGB	84.03	–	77.81
OVVAD (Wu et al., 2024a)	CLIP	86.05	–	66.53
MGFN (Chen et al., 2023)	I3D-RGB / VideoSwin-RGB	86.98	–	80.11
UR-DMU (Zhou et al., 2023)	I3D-RGB	86.97	–	**81.66**
VadCLIP (Wu et al., 2024b)	CLIP	88.02	88.02	–
One-class methods				
GODS (Wang & Cherian, 2019)	I3D-RGB	70.46	61.56	–
Unsupervised methods				
DyAnNet (Thakare et al., 2023b)	I3D-RGB	79.76	–	–
RareAnom (Thakare et al., 2023a)	I3D-RGB	–	68.33	–
Training-free MLLM methods				
LAVAD (Zanella et al., 2024)	ViT	78.33	82.89	60.02
EventVAD (Shao et al., 2025)	ViT	82.03	87.51	64.04
ReCoVAD (Huang et al., 2025)	ViT	82.28	86.38	65.66
VERA (Ye et al., 2025)	InternVL2-8B	86.55	88.26	70.54
PANDA (Yang et al., 2025)	Qwen2.5-VL-7B	84.89	–	70.16
ASK-HINT (Zou et al., 2026)	Qwen2.5-VL-7B	89.83	90.31	–
AnomalyRAG: weak supervision with a frozen MLLM				
AnomalyRAG	InternVL2-8B	87.96	89.03	72.95
AnomalyRAG	Qwen3-VL-8B	**89.96**	**92.15**	74.16

Each bank recalls 50 candidates through global, pooled-object, and pooled-relation routes, then relation-reranks their union. Top-$K = 3$ scores estimate confidence, and at most one reliable exemplar per bank reaches the VLM. The detector, OpenCLIP, and Qwen3-VL remain frozen; GATv2, retrieval, and calibration modules are trained with AdamW. Validation data select reliability and fusion parameters. Results average seeds 42/43/44; complete settings appear in the supplement.

Compared Variants. All variants use Qwen3-VL-8B-Instruct and the same split. Starting from direct prompting, we cumulatively add global abnormal-bank retrieval (*Vanilla RAG*), relation-centric retrieval, calibrated dual-bank comparison, bilateral SEGR-CoT, and reliability-gated fusion (full *AnomalyRAG*).

4.2 Main Results

Table 1 reports public-benchmark detection. With InternVL2-8B, AnomalyRAG reaches 87.96 AUC on UCF-Crime and 89.03/72.95 AUC/AP on XD-Violence, improving over backbone-matched VERA. Qwen3-VL-8B further obtains 89.96 and 92.15/74.16, respectively, showing that the retrieval mechanism transfers across frozen VLMs while retaining relation-grounded explanations.

Under the VAU-R1 protocol (Zhu et al., 2025) (Table 2), AnomalyRAG reaches 79.45/51.60 binary/multi-class accuracy on ECVA and 89.16/65.78 on UCF-Crime without MLLM fine-tuning. The supplement provides the complete VAU-Eval breakdown.

4.3 Ablation Study

Table 3 progressively examines global retrieval, relation-centric matching, dual-bank comparison, bilateral reasoning, and the complete framework.

Vanilla RAG can reduce VAU-Eval when global similarity fails to preserve the governing rule. Relation alignment restores detection and reasoning quality; dual-bank comparison and bilateral reasoning add complementary gains by contrasting rule-violating and rule-satisfying precedents to isolate the decisive relation change.

Table 2: Anomaly reasoning accuracy on ECVA and UCF-Crime.

Dataset	Model	Binary Acc. (%)↑	Multi-Class Acc. (%)↑
ECVA	Qwen2-VL-7B-Instruct (Wang et al., 2024)	64.85	32.88
	InternVL2-8B (OpenGVLab, 2024)	62.91	31.85
	Qwen3-VL-8B-Instruct (Bai et al., 2025a)	67.19	34.99
	+VAU-R1 (Zhu et al., 2025)	68.92	35.02
	+AnomalyRAG	**79.45**	**51.60**
UCF-Crime	Qwen2-VL-7B-Instruct (Wang et al., 2024)	86.85	62.15
	InternVL2-8B (OpenGVLab, 2024)	86.22	61.19
	Qwen3-VL-8B-Instruct (Bai et al., 2025a)	87.18	62.99
	+VAU-R1 (Zhu et al., 2025)	87.28	63.65
	+AnomalyRAG	**89.16**	**65.78**

On SMAD (Table 4), full AnomalyRAG improves over the Base VLM by 3.74 accuracy and 3.20 F1 points. Figure 3 shows how cardinality-aware retrieval localizes only the multi-rod subject. Complete classwise results appear in the supplement.

4.4 Mechanism Analysis

Retrieval and Reliability. Across four benchmarks, parser-assisted diagnostics yield evidence R@3 of 84.95%–89.46%, relation-motif consistency of 82.36%–87.81%, and bank purity of 93.18%–96.12%. A blinded 200-query human audit independently confirms minima of 84.6%, 82.8%, and 93.4% for the corresponding criteria (supplement). The seven-cue gate activates fusion for 25.42%–35.28% of queries and otherwise preserves the evidence-free prediction; activated queries perform better under the corresponding conditional metric, with final ECE of 2.86%–4.07%.

Independent controls further validate grounding: a 500-output visual-faithfulness audit reports 7.1% unsupported claims, 91.0% human–human agreement ($\kappa = 0.82$), and 88.5% agreement for an auxiliary judge disjoint from bank construction and prediction; a detector-only SMAD variant using the fixed task-agnostic vocabulary retains 90.62/90.55 accuracy/F1 without task identity or rule text.

Stability. Across seeds 42/43/44, sample standard deviations are 0.07 for UCF-Crime AUC, 0.08/0.10 for XD-Violence AUC/AP, 0.17/0.20 for ECVA binary/multi-class accuracy, and 0.08/0.09 for SMAD accuracy/F1.

Paired Significance and Sensitivity. Against the strongest pre-gating variant, 10,000 paired bootstrap resamples give gains of 1.43 UCF-Crime AUC (95% CI [0.91, 1.95]), 0.54 XD-Violence AP ([0.16, 0.93]), 0.61 ECVA AUC ([0.11, 1.12]), and 0.47 SMAD accuracy ([0.18, 0.76]); all intervals exclude zero.

Table 3: Ablation on ECVA and UCF-Crime.

Dataset	Model	AUC (%)↑	VAU-Eval Total↑
ECVA	Qwen3-VL-8B-Instruct	61.39	27.13
	+Vanilla RAG	65.74	26.69
	+Relation-Centric Retrieval	70.91	28.64
	+Dual-Bank Retrieval	72.36	29.11
	+Bilateral SEGR-CoT	73.21	29.56
	AnomalyRAG (Full)	**73.82**	**29.85**
UCF-Crime	Qwen3-VL-8B-Instruct	75.62	29.43
	+Vanilla RAG	80.18	27.19
	+Relation-Centric Retrieval	85.69	30.08
	+Dual-Bank Retrieval	87.42	30.59
	+Bilateral SEGR-CoT	88.53	31.07
	AnomalyRAG (Full)	**89.96**	**31.44**

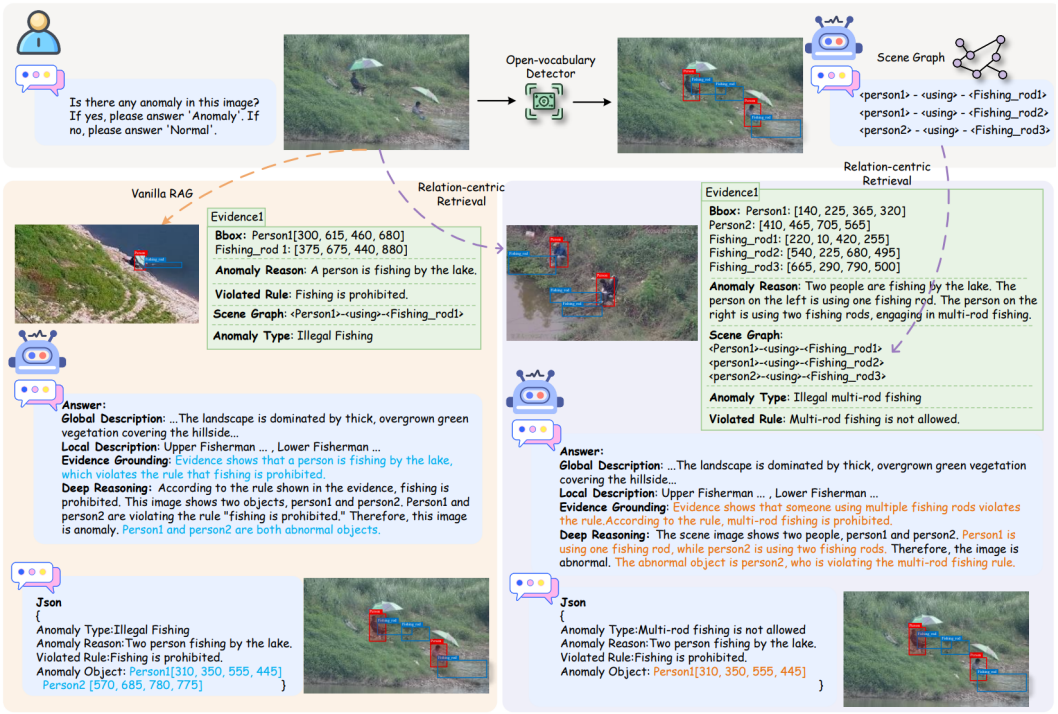

Figure 3: Qualitative case on SMAD. In *Multi-rods Fishing*, only the subject using multiple rods is anomalous. Vanilla RAG retrieves a visually similar *Illegal Fishing* case whose rule prohibits all fishing and consequently marks both subjects as abnormal. Relation-centric retrieval instead matches the query's cardinality-sensitive relation motif and supplies the correct multi-rod rule, enabling SEGR-CoT to localize the violation while preserving the single-rod subject as normal.

Table 4: Overall ablation on SMAD.

Model Variant	Accuracy (%)↑	F1 Score (%)↑
Base VLM	87.60	88.09
+Vanilla RAG	87.04	87.40
+Relation-Centric Retrieval	89.26	88.64
+Dual-Bank Retrieval	90.18	89.73
+Bilateral SEGR-CoT	90.87	90.62
AnomalyRAG (Full)	**91.34**	**91.29**

Resampling preserves complete UCF-Crime/XD-Violence test videos and stratifies ECVA/SMAD queries by anomaly class or alerting task. Because the comparator already includes dual-bank retrieval and bilateral reasoning, the paired gains isolate the additional contribution of reliability-controlled evidence admission. One-factor sweeps exhibit a broad local plateau around $M = 50$, $K = 3$, relation weight 1.0, bank-admission threshold 0.65, and a 12-node graph cap. Full protocols, seed-wise metrics, and sensitivity curves appear in the supplement.

5 Conclusion

We presented **AnomalyRAG**, coupling relation-aligned normal and abnormal precedents with soft graph matching, bilateral SEGR-CoT, and reliability-gated fusion. This contrast separates background resemblance from rule compatibility and grounds decisions in responsible objects and relations. Across UCF-Crime, XD-Violence, ECVA, and SMAD, AnomalyRAG consistently improves detection and reasoning with a frozen MLLM. Detector quality remains a limitation; verified object–relation evidence nevertheless supports inspectable, selective decisions.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, support retrieval and discovery of related work, and assist with research ideation or execution. Relevant details are described in the paper and supplementary material. The authors reviewed the AI-assisted content and take responsibility for the final manuscript, methods, data, and reported results.

ETHICS STATEMENT

AnomalyRAG is intended for offline public-safety research and human-reviewed alert triage, not person identification or autonomous enforcement. False alarms, missed hazards, and unsupported explanations can affect individuals; retrieval verification does not eliminate these risks. The protocol uses no person-identity attributes or face recognition. Differences in scene conditions and detector quality require context-specific assessment before deployment; the reported benchmarks do not establish demographic fairness or generalization to unseen cameras or locations.

SMAD imagery was obtained under authorized research use, and the study protocol was reviewed by the relevant ethics body and determined to be exempt. This exemption does not waive privacy or data-rights obligations, and the authorization does not grant unrestricted redistribution rights. Raw imagery remains in access-controlled storage for authorized personnel under the applicable data-use and retention requirements. The executable manifest supports source-group and perceptual near-duplicate isolation but lacks reliable camera and geographic-location identifiers. Any external release will require data-owner permission and a data-rights and privacy review, with identifiable faces, license plates, timestamps, and location metadata removed or obscured where required. Access terms will prohibit identity inference and autonomous enforcement. The institution name, exemption identifier, and applicable data-use terms will be disclosed in the non-anonymous version where permitted.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):387–395, 2023. doi: 10.1609/aaai.v37i1.25112. URL http://dx.doi.org/10.1609/aaai.v37i1.25112.

Ying Cheng, Yu-Ho Lin, Min-Hung Chen, Fu-En Yang, and Shang-Hong Lai. Vader: Towards causal video anomaly understanding with relation-aware large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7301–7311, March 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Cheng_VADER_Towards_Causal_Video_Anomaly_Understanding_with_Relation-Aware_Large_Language_WACV_2026_paper.html.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024. URL https://arxiv.org/abs/2406.07476.

Zongcan Ding, Haodong Zhang, Peng Wu, Guansong Pang, Zhiwei Yang, Peng Wang, et al. Slowfastvad: Video anomaly detection via integrating simple detector and rag-enhanced vision-language model, 2025. URL https://arxiv.org/abs/2504.10320.

Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, et al. Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly, 2024. URL https://arxiv.org/abs/2412.07183.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Lyu, Xiaohua Xie, and Wei-Shi Zheng. Wedetect: Fast open-vocabulary object detection as retrieval, 2025. URL https://arxiv.org/abs/2512.12309.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 733–742, 2016. doi: 10.1109/CVPR.2016.86. URL http://dx.doi.org/10.1109/cvpr.2016.86.

He Huang, Zixuan Hu, Dongxiao Li, Yao Xiao, and Ling-Yu Duan. Sparse reasoning is enough: Biological-inspired framework for video anomaly detection with large pre-trained models, 2025. URL https://arxiv.org/abs/2511.17094.

Yu-Chen Lai, Motoharu Sonogashira, Itthisak Phueaksri, and Yasutomo Kawanishi. Scene-specific anomalous relationship detection using scene graph summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4034–4042, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/VAND/html/Lai_Scene-Specific_Anomalous_Relationship_Detection_Using_Scene_Graph_Summarization_CVPRW_2025_paper.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. *Scale-Aware Spatio-Temporal Relation Learning for Video Anomaly Detection*, pp. 333–350. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19772-7_20. URL http://dx.doi.org/10.1007/978-3-031-19772-7_20.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, et al. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024. doi: 10.18653/v1/2024.emnlp-main.342. URL http://dx.doi.org/10.18653/v1/2024.emnlp-main.342.

OpenGVLab. InternVL2-8B. Hugging Face model card, 2024. URL https://huggingface.co/OpenGVLab/InternVL2-8B.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, et al. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2586–2595, 2025. doi: 10.1145/3746027.3754500. URL http://dx.doi.org/10.1145/3746027.3754500.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018. doi: 10.1109/CVPR.2018.00678. URL http://dx.doi.org/10.1109/cvpr.2018.00678.

Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim. RareAnom: A benchmark video dataset for rare type anomalies. *Pattern Recognition*, 140: 109567, 2023a. doi: 10.1016/j.patcog.2023.109567. URL http://dx.doi.org/10.1016/j.patcog.2023.109567.

Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. DyAnNet: A scene dynamicity guided self-trained video anomaly detection network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5541–5550, January 2023b. doi: 10.1109/WACV56688.2023.00547. URL https://openaccess.thecvf.com/content/WACV2023/html/Thakare_DyAnNet_A_Scene_Dynamicity_Guided_Self-Trained_Video_Anomaly_Detection_Network_WACV_2023_paper.html.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4955–4966, 2021. doi: 10.1109/ICCV48922.2021.00493. URL http://dx.doi.org/10.1109/iccv48922.2021.00493.

Jue Wang and Anoop Cherian. GODS: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8201–8211, October 2019. doi: 10.1109/ICCV.2019.00829. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Wang_GODS_Generalized_One-Class_Discriminative_Subspaces_for_Anomaly_Detection_ICCV_2019_paper.html.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. *Self-supervised Sparse Representation for Video Anomaly Detection*, pp. 729–745. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19778-9_42. URL http://dx.doi.org/10.1007/978-3-031-19778-9_42.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, et al. *Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision*, pp. 322–339. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58577-8_20. URL http://dx.doi.org/10.1007/978-3-030-58577-8_20.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, et al. Open-Vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18297–18307, June 2024a. doi: 10.1109/CVPR52733.2024.01732. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_Open-Vocabulary_Video_Anomaly_Detection_CVPR_2024_paper.html.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, et al. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):6074–6082, 2024b. doi: 10.1609/aaai.v38i6.28423. URL http://dx.doi.org/10.1609/aaai.v38i6.28423.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. *Follow the Rules: Reasoning for Video Anomaly Detection with Large Language Models*, pp. 304–322. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73004-7_18. URL http://dx.doi.org/10.1007/978-3-031-73004-7_18.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. In *Advances in Neural Information Processing Systems*, volume 38, pp. 83182–83211. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/77e97aea16dbbfe473e70ad12ca733e7-Paper-Conference.pdf.

Muchao Ye, Weiyang Liu, and Pan He. VERA: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8679–8688, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ye_VERA_Explainable_Video_Anomaly_Detection_via_Verbalized_Learning_of_Vision-Language_CVPR_2025_paper.html.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, et al. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22052–22061, 2024. doi: 10.1109/CVPR52733.2024.02082. URL http://dx.doi.org/10.1109/cvpr52733.2024.02082.

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 (CVPR)*, pp. 14744–14754, June 2022. doi: 10.1109/CVPR52688.2022.01433. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zaheer_Generative_Cooperative_Learning_for_Unsupervised_Video_Anomaly_Detection_CVPR_2022_paper.html.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18527–18536, 2024. doi: 10.1109/CVPR52733.2024.01753. URL http://dx.doi.org/10.1109/cvpr52733.2024.01753.

Yuhui Zeng, Haoxiang Wu, Wenjie Nie, Guangyao Chen, Xiawu Zheng, Yunhang Shen, et al. From objects to events: Unlocking complex visual understanding in object detectors via llm-guided symbolic reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24380–24391, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zeng_From_Objects_to_Events_Unlocking_Complex_Visual_Understanding_in_Object_ICCV_2025_paper.html.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, et al. Holmes-VAD: Towards unbiased and explainable video anomaly detection via multi-modal llm, 2024. URL https://arxiv.org/abs/2406.12235.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, et al. 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, 2025. doi: 10.1109/CVPR52734.2025.01292. URL http://dx.doi.org/10.1109/cvpr52734.2025.01292.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3769–3777, 2023. doi: 10.1609/aaai.v37i3.25489. URL https://ojs.aaai.org/index.php/AAAI/article/view/25489.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2505.23504.

Shu Zou, Xinyu Tian, Lukas Wesemann, Fabian Waschkowski, Zhaoyuan Yang, and Jing Zhang. Unlocking vision-language models for video anomaly detection via fine-grained prompting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4223–4233, March 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Zou_Unlocking_Vision-Language_Models_for_Video_Anomaly_Detection_via_Fine-Grained_Prompting_WACV_2026_paper.html.

ANOMALYRAG: RELATION-CENTRIC DUAL-BANK EVIDENCE RETRIEVAL FOR SURVEILLANCE ANOMALY DETECTION AND REASONING

Anonymous authors
Paper under double-blind review

1 OVERVIEW

This document supplements the anonymous ICLR 2027 submission "AnomalyRAG: Relation-Centric Dual-Bank Evidence Retrieval for Surveillance Anomaly Detection and Reasoning." It provides the complete protocol for SMAD's 24 task-specific binary alerting tasks, evidence-construction and decoding templates, retriever-training and leakage controls, classwise results, component isolation, paired statistical tests, retrieval sensitivity, independent human retrieval and visual-faithfulness audits, detector and vocabulary ablations, runtime measurements, and additional qualitative examples. The material also makes explicit which components are frozen and which are learned: the detector, visual encoder, and VLM remain fixed, whereas a lightweight relation retriever and calibration modules are optimized using training data. These details are intended to make the dual-bank retrieval protocol reproducible and to expose its practical assumptions rather than treating evidence construction as an opaque preprocessing step.

Reading Guide. Bank-admission details and counts appear in Section 3.1; per-metric seed statistics and paired confidence intervals appear in Sections 4.3 and 4.6. Section 5 discusses the main-paper relation-cardinality case and provides additional qualitative examples; Table 22 retains all 24 classwise SMAD results.

2 DATASET DETAILS OF SMAD

Collection Background. SMAD is constructed by sampling images from surveillance videos captured by a large, spatially distributed network of fixed cameras mounted on elevated observation towers. The cameras cover roads, rivers, lakes, construction sites, farmland, industrial areas, and urban-fringe regions. Their wide fields of view and broad deployment yield diverse, non-consecutive scenes and viewpoints rather than adjacent frames from a small number of continuous clips.

Viewpoint Characteristics. The viewpoint differs from standard pedestrian- or vehicle-level datasets. Targets are often small relative to the wide field of view, while the background occupies most of the image. SMAD nevertheless remains a surveillance dataset rather than a remote-sensing benchmark: object-level visual cues and local interactions are still observable, but they must be recovered under scale variation, clutter, and substantial scene diversity.

Audited Experimental Manifest. The executable manifest used in the reported experiments contains 12,500 images from 24 task-specific binary alerting tasks, balanced between 6,250 normal and 6,250 abnormal samples. These 24 tasks define the alerting taxonomy rather than the data partitions. Within each task, we use a 60/20/20 train/validation/test split, yielding 7,500/2,500/2,500 images overall. Split assignment is performed over 8,532 source groups with seed 42. Perceptually near-duplicate images are clustered within each task using a perceptual-hash Hamming threshold of 4, and every cluster is assigned to a single split. Consequently, neither an identical source group nor its near-duplicate cluster can occur in both an evidence bank and the test set.

Labels and Object Evidence. Each image is manually annotated with a task identifier, a binary normal/abnormal label, and bounding boxes for task-relevant objects. The executable manifest does not retain annotator identities, repeated independent labels, or adjudication records; consequently,

Table 1: Audited SMAD manifest used by the experimental pipeline.

Property	Value
Binary alerting tasks	24
Images	12,500
Normal / abnormal	6,250 / 6,250
Train / validation / test	7,500 / 2,500 / 2,500
Source groups	8,532
Split seed	42
Perceptual-hash threshold	4 bits

we do not claim a measured inter-annotator agreement statistic for the current release. In the reported SMAD protocol, available manual boxes initialize the object nodes; the frozen open-vocabulary detector supplies candidates when a box annotation is absent and for public-benchmark queries. Every compared SMAD variant receives the same object input, so retrieval variants do not receive additional query annotations. No identity labels or facial-recognition supervision are used.

Test-Time Inputs and Variant Parity. SMAD is evaluated as 24 task-specific binary alerting tasks, not as a 24-way classifier. Whenever detector inference is required, every task uses the same 40-class inventory fixed before evaluation; the operational task identifier does not select detector classes. Before a query package reaches retrieval or reasoning, its binary label, task identifier, task name, and rule statement are removed. The retriever therefore receives only the query image and its query-side visual, object, and relation representations; the decision VLM receives the same query evidence plus any retrieved packages admitted by the evaluated variant. Base VLM, Vanilla RAG, and AnomalyRAG share identical query-side perception and decoding settings. No variant receives a query label, a ground-truth rule, or an answer-derived explanation; rule statements shown during online reasoning originate only from training-bank evidence.

Split Scope and Data Governance. The imagery was obtained under authorized research use, and the study protocol was reviewed by the relevant ethics body and determined to be exempt. The exemption does not waive privacy or data-rights obligations: the present authorization covers controlled research use and does not by itself grant unrestricted redistribution rights. The executable manifest retains task, binary label, source path, and perceptual-hash group, but it does not provide reliable camera or geographic-location identifiers. We therefore claim source-group and perceptual near-duplicate isolation, not leave-one-camera or leave-one-location-out generalization. The protocol uses no person-identity attributes and does not perform face recognition. Raw imagery is stored in access-controlled project storage and is available only to authorized personnel under the applicable data-use and retention requirements. Any external release will occur only when permitted by the data owner and after data-rights and privacy review; identifiable faces, license plates, timestamps, and location metadata will be removed or obscured where required, and access terms will prohibit identity inference and autonomous enforcement. Because surveillance alerts can affect individuals and communities, system outputs are intended only as decision support and require human verification. The institution name, exemption identifier, and applicable data-use terms will be disclosed in the non-anonymous version where permitted.

Practical Motivation and Scope. SMAD targets public-safety event recognition, including smoke and fire, hazardous vehicles, road occupation, water-safety violations, and farmland misuse. Object presence alone is often insufficient: a person and a fishing rod can be normal or abnormal depending on the task rule, and a tree is anomalous only when it participates in a prohibited tree–farmland relation. The dataset therefore evaluates offline image-level decision support and alert triage under context-sensitive rules. It does not by itself establish raw-stream, real-time, or long-horizon monitoring performance.

Table 2 lists the 24 tasks in the audited manifest together with their anomaly rules and principal target categories.

Table 2: Task definitions for the 24 task-specific binary alerting tasks in the audited SMAD manifest. The target-category column summarizes the principal visual evidence used for anomaly judgment.

ID	Task Name	Abnormal Rule	Target Categories
1	Open-Air Garbage Burning	Visible outdoor garbage burning, flames, or burning smoke is abnormal.	flame, smoke plume, burning garbage pile
2	Black-Smoke Emission	Black smoke emitted by scattered-coal combustion sources is abnormal.	black smoke, chimney, combustion source
3	Dust Emission	Obvious airborne dust or particulate spread is abnormal.	dust plume, exposed dust source, construction ground
4	Industrial Smoke	Visible industrial smoke discharge is abnormal.	industrial chimney, smoke plume, exhaust region
5	Blue-Algae	Algal bloom or abnormal green coverage on a water surface is abnormal.	algae region, contaminated water surface
6	Low-Altitude Fire Point	A visible ground fire point or small open flame is abnormal.	small fire point, flame spot, burning region
7	Flooding or Ponding	Large-area overflow or obvious standing water is abnormal.	water-covered road, flooded ground, ponding region
8	Road Damage	Visible road cracks, collapse, or severe surface damage is abnormal.	damaged road surface, pothole, crack region
9	Garbage Dumping	Exposed piled garbage in a non-designated area is abnormal.	garbage pile, waste bags, dumping region
10	Floating Litter	Visible floating debris or garbage on water is abnormal.	floating debris, litter cluster, polluted water surface
11	Fire-Access Blocking	Occupation or blockage of a fire-access lane is abnormal.	motorcycle, bicycle, furniture, trash, blockage
12	Street Vending Occupation	Street-vending behavior occupying public road space is abnormal.	street stall, tent, table, chair, vendor setup
13	Dust-Net Monitoring	The construction-site covering condition violates the task-specific dust-control rule.	exposed ground, dust net, site cover
14	Road-Occupying Construction	Construction activity occupying an active roadway is abnormal.	worker, machinery, barrier, cone, sign
15	Swimming	People swimming in monitored restricted waters are abnormal.	swimmer, person in water, water activity
16	Illegal Fishing	Any fishing activity in a restricted region is abnormal.	person, fishing rod, fishing setup
17	Multi-rods Fishing	Multi-rod fishing is abnormal, while single-rod fishing is normal.	person, multiple fishing rods, rod setup
18	No-Lifejacket	Water-related activity without a required lifejacket is abnormal.	person, lifejacket, boat, waterside activity
19	Dangerous Vehicle	A prohibited hazardous vehicle in the monitored road context is abnormal.	tanker truck, hazardous vehicle, warning mark
20	Night Construction Machinery	Nighttime operation of restricted construction machinery is abnormal.	excavator, bulldozer, dump truck, mixer truck
21	Tree-Planting-on-Farmland	Trees planted on protected farmland are abnormal.	farmland, tree cluster, planted tree
22	Pond-Excavation-on-Farmland	Pond excavation on protected farmland is abnormal.	farmland, excavated pond, water area
23	Suspected Construction Site	Visual evidence of ongoing construction in monitored land is abnormal.	scaffolding, crane, building frame, material pile
24	Helmet-Wearing	A person working in the designated construction context without a safety helmet is abnormal.	person, safety helmet, construction context

Table 2 spans object-driven, state-driven, and relation-sensitive events. Several task pairs are visually related but semantically distinct, such as *Illegal Fishing* versus *Multi-rods Fishing*, or *Road-Occupying Construction* versus ordinary roadside activity. This mixture tests whether retrieval preserves the task context and rule structure required for anomaly judgment rather than only background appearance.

3 Additional Implementation Details

3.1 Public-Benchmark Weak-Label Protocol

UCF-Crime and XD-Violence provide video-level labels for their training partitions. We use these labels directly and do not use test-time temporal annotations to construct evidence. For UCF-Crime, a training path under `Training_Normal_Videos_Anomaly` is assigned label 0; the remaining official training paths inherit their anomaly category and label 1. For XD-Violence, the anomaly code in the official training filename determines the category and label, while filenames without an

Table 3: Weak-label evidence protocol on the public video benchmarks. Validation is source-video disjoint and formed only from official training data; the complete official test split is reserved for evaluation.

Dataset	Initial Supervision	Abnormal-Bank Admission	$\lvert\mathcal{E}_n\rvert$	$\lvert\mathcal{E}_a\rvert$	Test Use
UCF-Crime	Official video label	Visible violation, valid grounding, confidence ≥ 0.65	650	308	Complete official split
XD-Violence	Official video label	Visible violation, valid grounding, confidence ≥ 0.65	300	267	Complete official split

anomaly code are normal. Thus, AnomalyRAG is weakly supervised with a frozen MLLM: training examples inherit the official video label, and no claim of frame-level training annotation is made.

We use the complete official training partitions for evidence construction and retriever learning. Validation subsets are formed only within the official training side and remain disjoint by source video. The complete official test partitions are evaluated without entering either evidence bank, calibration, or hyperparameter selection. An explicit source-video disjointness check prevents any official test video from being used as candidate evidence.

Frame Extraction and Evidence Admission. Every official training video participates in evidence construction. Candidate frames are decoded with a deterministic, label-independent uniform schedule over the complete video duration; the video label supplies only an initial weak direction. A frame from an officially normal video is eligible for the normal bank after structural validation. A frame from an abnormal video enters the abnormal bank only when the frozen VLM confirms a visible, task-consistent violation with confidence at least 0.65 and returns valid responsible objects, relations, and a governing rule. If the VLM confidently finds no visible violation, the frame is retained as same-scene normal evidence; ambiguous candidates are discarded. This procedure prevents visually normal portions of an abnormal training video from being inserted into the abnormal bank solely because of their video label.

After consistency filtering and within-task diversity selection, the UCF-Crime banks contain 650 normal and 308 abnormal packages, while the XD-Violence banks contain 300 normal and 267 abnormal packages. Test videos are decoded at their original frame indices for frame-level evaluation; no test event boundary or frame annotation is used to construct, filter, or select evidence. Table 3 summarizes the resulting protocol.

3.2 Frozen Backbones, Retriever Training, and Leakage Control

AnomalyRAG does not finetune the VLM, open-vocabulary detector, or OpenCLIP encoder. It trains a lightweight relation retriever comprising the edge-aware GATv2, projection heads, relation scorer, and calibration modules. We therefore describe the protocol as *weakly supervised with a frozen MLLM*, rather than training-free as a complete system. The distinction is important because the learned retriever determines which evidence is exposed to the frozen reasoning model.

Both the normal bank $\mathcal{E}_n$ and abnormal bank $\mathcal{E}_a$ are constructed exclusively from the official training split. Validation data are used for early stopping, score calibration, fusion weights, and the reliability threshold. No test frame, test annotation, test rationale, or test-derived statistic enters the banks or any hyperparameter search. For videos, near-duplicate frames from the same source segment are not split between a training query and its candidate evidence, which prevents retrieval from exploiting video identity.

Labels are retained only to assign training evidence to a bank and to form retriever supervision. They are not included in the online similarity function or passed to the VLM as direct query hints. At inference, the VLM receives the query plus retrieved images, object/relation alignments, evidence descriptions, and rule statements, and must verify whether the decisive relation is present in the query. A nearest evidence label is therefore not itself a prediction.

3.3 Evidence Construction and Structured Decoding

Open-Vocabulary Detector and Class Vocabulary. The open-vocabulary detector is used to provide candidate objects and semantic regions for evidence construction and online query parsing. We use WeDetect Base with confidence threshold 0.05 and NMS IoU 0.70. After thresholding and NMS, relevance ranking retains at most 12 objects as graph nodes and at most 8 detected boxes in the VLM prompt. SMAD uses one shared 40-class vocabulary for all 24 tasks rather than a task-conditioned class list. It includes *person*, *helmet*, *lifejacket*, *fishing rod*, *boat*, *garbage pile*, *smoke*, *fire*, *truck*, *construction machinery*, *barrier*, *cone*, and region-like cues such as *water surface*, *farmland*, and *algae*. The fixed union preserves small relation-bearing targets without revealing the current task identity.

Prompt Template for Evidence Description and Rule Statement. The offline VLM processes both normal and abnormal evidence. The normalized instruction is: *Given a surveillance image, its binary training label, detected objects with box identifiers, and a task-oriented scene graph: (1) identify the object instances and relations that determine the label; (2) describe why the observed configuration satisfies or violates the task rule using only visible evidence; (3) state the governing rule in one concise sentence; and (4) return the responsible object identifiers and canonical relations. Do not introduce objects or relations absent from the detections or graph.* For abnormal items, the output explains the violated condition; for normal items, it explains the rule-satisfying condition. Outputs whose label direction or object references fail consistency checks are excluded from the banks.

Prompt Template for Online SEGR-CoT Reasoning. At inference, Top-$K = 3$ evidence per bank is used for score aggregation, but the VLM receives at most the query, Top-1 normal evidence, and Top-1 abnormal evidence. The instruction is: *(1) summarize the query scene without deciding the label; (2) inspect candidate objects, cardinalities, and local relation motifs; (3) compare the query separately with the normal and abnormal evidence, explicitly stating matched and unmatched relations; and (4) decide whether the query satisfies or violates the relevant rule and identify the responsible objects. Do not copy an evidence label unless the corresponding relation is verified in the query.* When a bank has no reliable candidate, the prompt states this explicitly instead of inserting weak evidence.

The fixed JSON schema contains `abnormal`, `anomaly_score`, `predicted_class`, `query_object_ids`, `anomalous_box_ids`, `matched_normal_evidence_id`, `matched_abnormal_evidence_id`, `matched_relations`, `violated_rule`, `normal_abnormal_difference`, `confidence`, and `reason`. JSON-guided decoding in vLLM (Kwon et al., 2023) stabilizes syntax. A deterministic verifier then checks evidence IDs, object IDs, relation endpoints, task/rule compatibility, and score ranges; it validates structural consistency rather than independently re-perceiving the image.

Evidence Package Construction and Retrieval Pipeline. Each package stores the source image, label, detections, task-oriented graph, evidence description, rule statement, global image and graph features, object and edge tokens, task metadata, and provenance. The query follows the same perception pipeline but excludes labels and answer-derived semantics. Each bank maintains three cosine indices over normalized global, pooled-object, and pooled-relation embeddings. Every route recalls Top-50 packages. Their union is ordered by reciprocal-rank fusion, using $1/(60 + r)$ for route rank r, and capped at 100 candidates per bank before relation-centric reranking.

Scene-Graph Canonicalization and Quality. Detector labels are lower-cased, mapped through the shared vocabulary's fixed alias table, and normalized by replacing underscores with spaces and collapsing whitespace. VLM relation strings undergo the same lower-casing and separator normalization. A semantic edge is retained only when both endpoint identifiers resolve to detected nodes and the two endpoints are distinct; spatial edges use deterministic box geometry. Let c_V and c_R be the mean clipped node and edge confidence, and define support terms $\rho_V = \min(1, |V|/2)$ and

$\rho_R = \min(1, |\mathcal{R}|/2)$. The graph quality used by the retriever is

$$\chi(G) = \begin{cases} 0, & |V| = 0, \\ 0.35\rho_V c_V, & |V| > 0,\ |\mathcal{R}| = 0, \\ \sqrt{\rho_V \rho_R c_V c_R}, & |\mathcal{R}| > 0. \end{cases}$$

Rule quality is binary in the implementation: $\chi^{\text{rule}} = \mathbb{1}[d_i \neq \varnothing]\mathbb{1}[\|e(d_i)\|_1 > 0]$, where d_i is the normalized rule statement and $e(d_i)$ its rule embedding. Relation types receive training-bank inverse-document-frequency weights

$$\omega(r) = \log \frac{N+1}{\mathrm{df}(r)+1} + 1,$$

where N is the number of training packages and $\mathrm{df}(r)$ counts packages containing canonical relation r. These weights are fixed after training-bank construction.

Dustbin-Aware Partial Object Matching. Let $H_q \in \mathbb{R}^{n\times d}$ and $H_i \in \mathbb{R}^{m\times d}$ be normalized GATv2 object embeddings and $A^{\text{o}} = H_q H_i^\top$. To avoid forcing every object into a real correspondence, we append a learnable dustbin logit α_{o}:

$$\widetilde{L}^{\text{o}} = \begin{bmatrix} A^{\text{o}}/\tau_s & (\alpha_{\text{o}}/\tau_s)\mathbf{1}_n \\ (\alpha_{\text{o}}/\tau_s)\mathbf{1}_m^\top & \alpha_{\text{o}}/\tau_s \end{bmatrix}, \qquad \tau_s = 0.07.$$

The augmented marginals are

$$\widetilde{\mu} = \frac{1}{n+m}\begin{bmatrix}\mathbf{1}_n \\ m\end{bmatrix}, \qquad \widetilde{\nu} = \frac{1}{n+m}\begin{bmatrix}\mathbf{1}_m \\ n\end{bmatrix}.$$

Ten log-domain Sinkhorn iterations produce the partial transport

$$\widetilde{P}^{\text{o}} = (n+m)\,\mathrm{Sinkhorn}(\widetilde{L}^{\text{o}}; \widetilde{\mu}, \widetilde{\nu}), \qquad P^{\text{o}} = \widetilde{P}^{\text{o}}_{1:n,1:m}.$$

The factor $n+m$ gives every real row and column unit marginal when its dustbin entry is included; the real–real block P^{o} is not subsequently normalized to total mass one. Consequently, $\widetilde{P}^{\text{o}}_{u\varnothing}$ and $\widetilde{P}^{\text{o}}_{\varnothing v}$ directly represent unmatched query and evidence mass. The object score and continuous object coverage are

$$s_{\text{obj}} = \frac{\sum_{u,v} P^{\text{o}}_{uv} A^{\text{o}}_{uv}}{\sum_{u,v} P^{\text{o}}_{uv} + \epsilon},$$

$$\kappa_{\text{obj}}(q,i) = \frac{1}{2}\left[\frac{1}{n}\sum_{u=1}^{n}\left(1 - \widetilde{P}^{\text{o}}_{u\varnothing}\right) + \frac{1}{m}\sum_{v=1}^{m}\left(1 - \widetilde{P}^{\text{o}}_{\varnothing v}\right)\right].$$

Endpoint-Grounded Relation-Motif Matching. Let relation motifs $\ell = (u_\ell, v_\ell)$ and $k = (a_k, b_k)$ have normalized embeddings with affinity $A^{\text{r}}_{\ell k}$. Their endpoint compatibility is

$$C_{\ell k} = \sqrt{P^{\text{o}}_{u_\ell a_k} P^{\text{o}}_{v_\ell b_k}}.$$

We add $\log(\mathrm{clip}(C_{\ell k}, 10^{-5}, 1))$ to the relation logits and apply a second partial Sinkhorn with a separately learned relation-dustbin logit α_{r}. If P^{r} is its real–real block and $\omega^q_\ell, \omega^i_k$ are relation weights, then

$$s_{\text{rel}} = \frac{\sum_{\ell,k} \omega^q_\ell \omega^i_k P^{\text{r}}_{\ell k} A^{\text{r}}_{\ell k}}{\sum_{\ell,k} \omega^q_\ell \omega^i_k P^{\text{r}}_{\ell k} + \epsilon}.$$

Relation coverage $\kappa_{\text{rel}}(q,i)$ is defined analogously to κ_{obj} from the query- and evidence-side dustbin masses of the relation transport. It is zero when either graph contains no valid relation motif. This two-stage construction permits both irrelevant objects and unsupported relations to remain unmatched.

Quality-Adaptive Candidate Score. The four branches are

$$\mathbf{s} = [s_{\text{global}}, s_{\text{obj}}, s_{\text{rel}}, s_{\text{rule}}],$$

where s_{global} is projected global OpenCLIP cosine similarity and s_{rule} compares local query tokens with the evidence-rule embedding. With graph qualities χ_q, χ_i and evidence-rule quality χ_i^{rule}, their availability vector is

$$\mathbf{q} = \left[1,\ \kappa_{\text{obj}},\ \kappa_{\text{rel}}\sqrt{\chi_q \chi_i},\ \chi_i^{\text{rule}} \max(\kappa_{\text{obj}}, \kappa_{\text{rel}})\right].$$

For learned logits $\boldsymbol{\theta}$, let $\boldsymbol{\pi} = \text{softmax}(\boldsymbol{\theta})$ and

$$\widehat{\boldsymbol{\pi}} = \frac{\boldsymbol{\pi} \odot \mathbf{q}}{\sum_j \pi_j q_j + \epsilon}, \qquad S(q,i) = \widehat{\boldsymbol{\pi}}^\top \mathbf{s}.$$

Thus missing or unreliable structure suppresses the affected branch rather than forcing a spurious match.

Dual-Bank Aggregation and Calibration. The top $K = 3$ packages are retained independently from $\mathcal{E}_n$ and $\mathcal{E}_a$. For nonempty bank b, scores are aggregated with temperature $\tau_a = 0.10$:

$$\begin{aligned} m_b &= \max_k S_b^{(k)}, \\ Z_b &= \frac{1}{K}\sum_{k=1}^{K} \exp\left(\frac{S_b^{(k)} - m_b}{\tau_a}\right), \\ \overline{S}_b &= m_b + \tau_a \log Z_b. \end{aligned}$$

Validation data provide robust bank-specific location and scale parameters (μ_b, σ_b) and calibrate temperature T and bias β. The signed margin and retrieval prior are

$$\Delta_t = \frac{\overline{S}_a - \mu_a}{\sigma_a} - \frac{\overline{S}_n - \mu_n}{\sigma_n}, \qquad p_t^{\text{ret}} = \sigma(\Delta_t/T + \beta).$$

The evaluated retriever combines seven normalized reliability cues:

$$\begin{aligned} c_t^{\text{ret}} = {} & 0.10c_{\text{graph}} + 0.15c_{\text{obj}} + 0.25c_{\text{rel/unary}} + 0.15c_{\text{rule}} \\ & + 0.10c_{\text{route}} + 0.15c_{\text{bank}} + 0.10c_{\text{sep}}. \end{aligned}$$

For reproducibility, all seven terms are computed in $[0,1]$. Let $\mathcal{H} = \mathcal{R}_n(q) \cup \mathcal{R}_a(q)$ be the retained packages and let $\mathcal{C}(h) \subseteq \{\text{global}, \text{object}, \text{relation}\}$ denote the coarse-recall routes that retrieved package h. The implementation uses

$$\begin{aligned} c_{\text{graph}} &= \max_{h\in\mathcal{H}} \sqrt{\chi_q \chi_h}, \\ c_{\text{obj}} &= \max_{h\in\mathcal{H}} \kappa_{\text{obj}}(q,h), \\ c_{\text{rel/unary}} &= \max\left(\max_{h\in\mathcal{H}} \kappa_{\text{rel}}(q,h), 0.85c_{\text{obj}}\right), \\ c_{\text{rule}} &= \max_{h\in\mathcal{H}} \text{clip}_{[0,1]}\left(\frac{1 + s_{\text{rule}}(q,h)}{2}\right), \\ c_{\text{route}} &= \max_{h\in\mathcal{H}} \frac{|\mathcal{C}(h)|}{3}, \\ c_{\text{bank}} &= \mathbb{1}[\mathcal{R}_n(q) \neq \varnothing \wedge \mathcal{R}_a(q) \neq \varnothing], \\ c_{\text{sep}} &= \min\left(1, \frac{|\overline{S}_a - \overline{S}_n|}{0.25}\right). \end{aligned}$$

The $0.85c_{\text{obj}}$ term provides a conservative unary fallback when a query contains no valid relation motif. Route consensus measures multi-route support for a retained package, rather than agreement between independently thresholded bank margins. Bilateral bank support records whether both comparison sides are available. Validation selects the threshold that maximizes coverage while attaining at least 65% retrieval precision and 10% coverage; if no threshold is feasible, it maximizes the precision–coverage product. Only the highest-ranked reliable package from each bank is inserted into the VLM prompt.

Retriever Optimization. For query q with task t_q and label $y \in \{0, 1\}$, a compatible positive i^+ must satisfy $t_{i^+} = t_q$, $y_{i^+} = y$, and $i^+ \neq q$. Negatives have the opposite label; same-task contradictions are prioritized, and the remaining hard negatives are selected by descending global-image similarity. Source-group isolation prevents a source duplicate from occupying both query and evidence roles. Let z^+ and $\{z_j^-\}_{j=1}^J$ be the learned retrieval logits for the positive and $J = 3$ negatives, and let

$$\bar{z}^- = \log \sum_{j=1}^{J} \exp z_j^- - \log J, \qquad m_y = (2y - 1)(z^+ - \bar{z}^-).$$

Let $\mathcal{H}_q = \{i^+, i_1^-, \ldots, i_J^-\}$. The five implemented losses are

$$\begin{aligned}
\mathcal{L}_{\text{db}} &= -y \log \sigma(m_y) - (1 - y) \log(1 - \sigma(m_y)) \,, \\
\mathcal{L}_{\text{NCE}} &= -\log \frac{\exp(z^+/\tau)}{\exp(z^+/\tau) + \sum_j \exp(z_j^-/\tau)}, \\
\mathcal{L}_{\text{rank}} &= \frac{1}{J} \sum_j [\, \delta - z^+ + z_j^- \,]_+, \\
\mathcal{L}_{\text{rule}} &= [\, \delta - r^+ + \tfrac{1}{J} \sum_j r_j^- \,]_+, \\
\mathcal{L}_{\text{anchor}} &= \frac{1}{J+1} \sum_{h \in \mathcal{H}_q} (1 - \chi_{qh}) \\
&\qquad \cdot \left(S(q, h) - \text{sg}[s_{\text{global}}(q, h)]\right)^2 .
\end{aligned}$$

Here r^+ and r_j^- are rule-compatibility scores, $\chi_{qh} = \sqrt{\chi_q \chi_h}$, sg stops gradients, $[x]_+ = \max(x, 0)$, $\tau = 0.07$, and $\delta = 0.2$. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{db}} + \mathcal{L}_{\text{NCE}} + 0.5\mathcal{L}_{\text{rank}} + 0.2\mathcal{L}_{\text{rule}} + 0.1\mathcal{L}_{\text{anchor}}.$$

We use AdamW with learning rate 10^{-4}, weight decay 10^{-4}, batch size 64, gradient-norm clipping at 1, at most 50 epochs, and patience 5. The checkpoint minimizing mean total loss on the source-group-disjoint validation examples is retained. Bank calibration then estimates robust location/scale statistics, selects temperature and bias by validation negative log-likelihood, and chooses the reliability threshold using validation correctness at a target retrieval precision of 65%. No test labels enter retriever selection or calibration. Complete runs use seeds 42, 43, and 44; the primary results report their mean, and Table 7 gives the individual outcomes and sample standard deviations.

Backbone and Decoding Configuration. The default VLM is Qwen3-VL-8B-Instruct, loaded with 4-bit NF4 quantization and BF16 computation and served through vLLM with temperature 0, context length 8192 or 16384 depending on the backend, and at most 512 output tokens. OpenCLIP ViT-L/14 uses the `laion2b_s32b_b82k` checkpoint and FP16 inference. The edge-aware GATv2 has two layers, hidden size 512, eight attention heads, dropout 0.1, residual connections, and LayerNorm. Experiments run on one NVIDIA A100-SXM4-80GB GPU under Linux 5.15, Python 3.10.20, PyTorch 2.8.0 with CUDA 12.8, Transformers 4.57.1, vLLM 0.11.0, and OpenCLIP 3.3.0.

Reliability-Gated Symmetric Fusion. Both the evidence-free and bilateral VLM calls return an `anomaly_score` $s \in [0, 100]$. We convert it deterministically as $p = \text{clip}(s/100, 10^{-4}, 1-10^{-4})$; no test-set normalization is applied. A prediction passes the *relation-grounded* verifier only when JSON parsing succeeds, no Base fallback was used, evidence and query-object identifiers are valid, matched relations are nonempty, the normal–abnormal difference is stated, and an abnormal decision includes a violated rule. Define

$$g_t = \mathbb{1}[c_t^{\text{ret}} \geq \tau_c] \mathbb{1}[\text{relation-grounded verification passes}].$$

Verifier failure, parsing failure, or Base fallback therefore sets $g_t = 0$. The signed evidence correction and final score are

$$\begin{aligned}
d_t^{\text{evi}} &= p_t^{\text{evi}} - p_t^{\text{base}}, \\
p_t^{\text{final}} &= \text{clip}_{[0,1]} \left(p_t^{\text{base}} + g_t \eta d_t^{\text{evi}}\right).
\end{aligned}$$

Table 4: Validation-selected reliability-gated fusion settings.

Dataset	η	τ_c
UCF-Crime	0.25	0.50
XD-Violence	0.25	0.75
ECVA	0.20	0.60
SMAD	0.20	0.60

Table 5: Mechanism-level retrieval, grounding, and calibration diagnostics. Conditional AUC/AP are rank metrics and need not average arithmetically, whereas the conditional accuracy rows combine according to gate coverage.

Dataset	Evidence R@3 (%)↑	Relation-Motif Consistency (%)↑	Bank Purity (%)↑	Gate Coverage (%)	ECE (%)↓	Activated / Non-activated
UCF-Crime	86.42	83.71	94.26	27.83	3.84	92.31 / 84.76 AUC
XD-Violence	84.95	82.36	93.18	25.42	4.07	78.64 / 68.23 AP
ECVA	88.17	85.62	95.03	32.56	3.31	84.92 / 76.81 Bin. Acc.
SMAD	89.46	87.81	96.12	35.28	2.86	94.73 / 89.49 Acc.

The same symmetric rule is applied to image and frame queries. Group-disjoint validation selects η and the reliability threshold τ_c, after which both are fixed for testing; test labels are never consulted. The final settings are listed in Table 4.

Frame Processing and Video Output. Public-benchmark query frames are processed in chronological order at their original frame indices. The resulting p_t^{final} sequence is used directly for frame-level AUC and AP. For an interpretable video-level temporal output, we apply the fixed causal smoother

$$\bar{p}_t = 0.6p_t^{\text{final}} + 0.4\bar{p}_{t-1}.$$

An anomalous interval starts after three consecutive frames exceed the validation-selected decision threshold and ends after three consecutive frames fall below it. This aggregation uses only predicted scores and fixed validation parameters; it does not use test event boundaries or test temporal annotations.

4 Additional Results and Diagnostics on Public Benchmarks

4.1 Retrieval and Gating Diagnostics

Table 5 reports system-internal diagnostics for the mechanism that distinguishes AnomalyRAG from global RAG. Evidence R@3 uses held-out dataset task labels and evidence-bank rule direction to determine whether at least one compatible package appears in the Top-3. Relation-motif consistency measures agreement between the retrieved motif and the canonical query relation produced by the frozen Qwen3-VL-8B-Instruct scene-graph parser; it measures pipeline consistency rather than independent human-verified visual correctness. Bank purity uses the human or official dataset label attached to each stored package after weak-label consistency filtering. Gate coverage is the fraction of queries for which evidence passes both the confidence threshold and structural verifier. It is an evidence-fusion activation rate, not an error, rejection, or abstention rate: every non-activated query retains its evidence-free output. Expected calibration error (ECE) is computed over 15 equal-width bins, and the last column reports the primary metric separately for activated and non-activated queries.

The high R@3 and relation-motif consistency rates indicate that the learned reranker preserves task-compatible evidence while correcting background-dominated global rankings. Bank purity confirms that the consistency filter suppresses weak-label contamination. The selective 25.42%–35.28% activation range shows that evidence is used only when the bilateral match is structurally verified, while the activated-query advantage confirms that the gate identifies cases in which retrieved structure is most useful. ECE below 4.1% further indicates that the final confidence remains well calibrated after bilateral evidence fusion. Because relation-motif consistency shares the frozen parser used by the system, it is not presented as a substitute for an independent visual-grounding audit.

Table 6: Independent retrieval audit on 200 stratified queries. Labels are supplied by two blinded human annotators with third-annotator adjudication, independently of the VLM and filters used to construct the evidence banks.

Dataset	Evidence R@3 (%)↑	Relation Match@1 (%)↑	Bank Purity@3 (%)↑
UCF-Crime	86.0	84.2	94.1
XD-Violence	84.6	82.8	93.4
ECVA	87.8	85.7	95.0
SMAD	89.1	87.5	95.9

Table 7: Stability across three retriever-training seeds. All values are percentages; "Std." is the sample standard deviation.

Seed	UCF AUC	XD AUC	XD AP	ECVA Bin. Acc.	ECVA Multi Acc.	SMAD Acc.	SMAD F1
42	89.91	92.07	74.18	79.60	51.39	91.36	91.37
43	90.04	92.22	74.05	79.27	51.79	91.41	91.19
44	89.93	92.16	74.25	79.48	51.62	91.25	91.31
Mean ± Std.	89.96±0.07	92.15±0.08	74.16±0.10	79.45±0.17	51.60±0.20	91.34±0.08	91.29±0.09

4.2 Independent Human Retrieval Audit

We independently audit 200 test queries stratified across datasets, task categories, prediction directions, and gate outcomes. Two annotators separately judge whether at least one Top-3 precedent is task- and rule-compatible, identify the decisive relation or unary state, determine whether the Top-1 retrieved motif matches it, and verify the normal/abnormal bank assignment of each retrieved package. Annotators see the query, retrieved packages, and dataset task definition, but not retriever scores, model confidence, or the predicted answer. A third annotator adjudicates disagreements. Crucially, neither the Qwen3-VL-8B-Instruct scene-graph outputs nor the consistency filter serves as a reference label.

Table 6 corroborates the system-internal trends without reusing their labels. Evidence R@3 remains at least 84.6%, relation match at least 82.8%, and bank purity at least 93.4% across all four datasets. The result supports the intended mechanism: coarse retrieval usually retains a rule-compatible precedent, relation reranking aligns the decisive local structure, and weak-label filtering preserves the direction of the two banks. The double-annotated audit yields 91.0% human–human agreement with Cohen's $\kappa = 0.82$; the companion visual-faithfulness calibration is reported below.

4.3 Random-Seed Stability

Table 7 reports complete runs with the three retriever-training seeds used throughout the experiments. The small sample standard deviations across detection, reasoning, and SMAD recognition indicate that the reported gains are not tied to a single retriever initialization.

4.4 Complete Anomaly Reasoning Results

Table 8 reports the complete VAU-R1 evaluation on ECVA and UCF-Crime. In addition to binary and multi-class accuracy, it retains all VAU-Eval dimensions used to assess classification correctness (CLS), key evidence matching (KM), fluency (FLU), informativeness (INF), factual consistency (FAC), and the aggregate Total score.

Metric Correspondence. Detection AUC and binary classification accuracy evaluate different outputs and are therefore not interchangeable. For the full model, ECVA yields 73.82 AUC, 79.45 binary accuracy, 51.60 multi-class accuracy, 29.85 VAU-Eval Total, and 85.84 QA accuracy (Table 9); the corresponding UCF-Crime results are 89.96, 89.16, 65.78, and 31.44. These values respectively match the detection/ablation and reasoning tables in the main paper.

Table 8: Complete anomaly reasoning comparison on ECVA and UCF-Crime under the VAU-R1 protocol.

Dataset	Model	Accuracy		VAU-Eval					
		Bin. Acc. (%)↑	Multi Acc. (%)↑	CLS↑	KM↑	FLU↑	INF↑	FAC↑	Total↑
ECVA	Qwen2-VL-7B-Instruct (Wang et al., 2024)	64.85	32.88	3.70	3.67	8.64	6.40	4.04	26.45
	InternVL2-8B (OpenGVLab, 2024)	62.91	31.85	3.05	3.59	8.07	4.91	3.69	23.31
	Qwen3-VL-8B-Instruct (Bai et al., 2025)	67.19	34.99	4.15	3.78	8.91	5.93	4.36	27.13
	+VAU-R1 (Zhu et al., 2025)	68.92	35.02	4.26	3.59	8.82	5.86	4.68	27.21
	+AnomalyRAG	**79.45** (↑12.26%)	**51.60** (↑16.61%)	**4.95**	**3.98**	**9.09**	**6.75**	**5.08**	**29.85**
UCF-Crime	Qwen2-VL-7B-Instruct (Wang et al., 2024)	86.85	62.15	4.80	3.73	8.95	7.05	4.25	28.78
	InternVL2-8B (OpenGVLab, 2024)	86.22	61.19	3.79	3.20	8.23	5.77	3.48	24.47
	Qwen3-VL-8B-Instruct (Bai et al., 2025)	87.18	62.99	4.96	3.88	9.15	6.58	4.86	29.43
	+VAU-R1 (Zhu et al., 2025)	87.28	63.65	5.02	3.69	9.08	6.99	4.57	29.35
	+AnomalyRAG	**89.16** (↑1.98%)	**65.78** (↑2.79%)	**5.92**	**4.09**	**9.29**	**7.23**	**4.91**	**31.44**

Table 9: Question-answering accuracy under the ECVA reasoning protocol.

Dataset	Model	QA Acc. (%)↑
ECVA	AnomalyRAG	**85.84**

4.5 Closest Contemporary Methods and Protocol Compatibility

Table 10 distinguishes reproducible detection comparisons from author-reported or protocol-incompatible results. UR-DMU learns normal and abnormal prototype memories, while ASK-HINT prompts a frozen VLM with fine-grained action knowledge; both are included in the main comparison using their officially reported metrics. PANDA also reports standard UCF-Crime/XD-Violence results. LAS-VAD reports strong performance on both benchmarks, but we did not identify an official public implementation linked from its paper, supplementary material, arXiv record, or CVF page at submission time; we therefore list its reported results here only for contextual reference. SlowFastVAD, VADER, and SARD are methodologically relevant but do not report the corresponding UCF-Crime/XD-Violence AUC/AP pair.

Table 11 isolates the methodological distinction from differences in backbones and evaluation protocols. UR-DMU stores learned prototypes rather than inspectable instances; ASK-HINT supplies action prompts without retrieval; AnomalyRuler induces rules from normal references; SlowFastVAD retrieves abstract behavior patterns; PANDA retrieves scene-conditioned rule knowledge; and VADER encodes dynamic object relations. AnomalyRAG instead retrieves image-level precedents from both label directions, aligns their local relation motifs, explicitly compares the two precedents, and gates their influence using retrieval and structural reliability.

LAS-VAD reports the highest pure-detection scores in this set. Its design combines anomaly-connected components, intention-aware prototypes, and category-specific anomaly attributes. Because an official implementation was not publicly available at submission time, these values remain author-reported and are not used to support claims of a directly reproduced comparison. AnomalyRAG instead targets inspectable normal/abnormal evidence retrieval and rule-grounded explanations with a frozen MLLM.

4.6 Component Isolation and Paired Significance

Table 12 separates evidence availability, relation alignment, bilateral comparison, and reliability control under identical frozen backbones, detector outputs, data splits, and decoding settings. The normal-bank-only row complements Vanilla RAG, which retrieves only from the abnormal bank using global instance similarity.

Either bank alone is incomplete: abnormal precedents expose possible violations but cannot establish a matched rule-satisfying alternative, whereas normal precedents lack direct violated-rule evidence. Relation-centric matching provides the largest gain over global Vanilla RAG. Adding the normal bank and bilateral comparison then exposes the decisive normal-to-abnormal relation

Table 10: Closest contemporary methods and the compatibility of their reported protocols. Dashes indicate that the paper does not report the corresponding standard detection metric.

Method	Supervision / Focus	UCF-Crime	XD-Violence	Other Reported Result and Comparability
UR-DMU (Zhou et al., 2023)	Video-level weak supervision	86.97 AUC	81.66 AP	Learns normal and abnormal feature prototypes; it does not retrieve structured instances for VLM reasoning.
ASK-HINT (Zou et al., 2026)	Training-free frozen VLM prompting	89.83 AUC	90.31 AUC	Uses action-centric prompt selection with Qwen2.5-VL-7B; XD-Violence AP is not reported.
LAS-VAD (Wang & Zhao, 2026)	Video-level weak supervision	91.05 AUC	89.96 AP	Author-reported results; no official public implementation identified at submission time.
PANDA (offline) (Yang et al., 2025)	Training-free agentic MLLM	84.89 AUC	70.16 AP	Uses Qwen2.5-VL-7B for perception/reasoning and Gemini 2.0 Flash for planning/reflection.
SlowFastVAD (Ding et al., 2025)	Few-normal-sample hybrid VAD	–	–	99.1/99.7 micro/macro AUC on Ped2 and 85.0/90.7 on ShanghaiTech; no direct UCF/XD result.
VADER (Cheng et al., 2026)	Fine-tuned anomaly understanding	–	–	On HAWK, 0.324 description BLEU-1, 0.484 QA BLEU-1, and 0.794 QA consistency; not a frame-AUC/AP comparison.
SARD (Lai et al., 2025)	Scene-specific relation rarity	–	–	Reports a 19.77-point gain in AUC of Recall@k on its relation-detection dataset; no video anomaly score or VLM reasoning.

Table 11: Component-level comparison with the closest rule-, retrieval-, and relation-based methods. "Patterns" denotes abstract behavior knowledge rather than retrieved image instances.

Method	Normal Evidence	Abnormal Evidence	Instance Retrieval	Relation Alignment	Bilateral Comparison	Reliability Gate
UR-DMU (Zhou et al., 2023)	Prototypes	Prototypes	–	–	–	–
ASK-HINT (Zou et al., 2026)	–	–	–	–	–	–
AnomalyRuler (Yang et al., 2024)	References	–	–	–	–	–
SlowFastVAD (Ding et al., 2025)	Patterns	VLM Patterns	–	–	–	–
PANDA (Yang et al., 2025)	–	Rule KB	–	–	–	–
VADER (Cheng et al., 2026)	–	–	–	Yes	–	–
AnomalyRAG	Instances	Instances	Yes	Yes	Yes	Yes

change, while the reliability gate supplies the final improvement by admitting evidence only when its objects, relations, and rules are grounded.

We further quantify uncertainty through 10,000 paired bootstrap resamples. UCF-Crime and XD-Violence are resampled by test video so that all frames from a video remain together; ECVA and SMAD are resampled by query, stratified by anomaly class or alerting task. Table 13 compares the full model with the strongest pre-gating variant and reports percentile confidence intervals together with two-sided paired randomization tests.

The intervals exclude zero on all four benchmarks. This result is particularly informative because the comparator already contains dual-bank relation retrieval and bilateral reasoning; the remaining improvement therefore isolates the value of reliability-controlled evidence admission rather than comparing against a weak evidence-free baseline.

4.7 Backbone Comparison

Cross-Backbone Evaluation. We compare AnomalyRAG with InternVL2-8B and Qwen3-VL-8B-Instruct under identical data splits and evaluation protocols. InternVL2-8B matches the default backbone used by VERA, whereas the similarly sized Qwen3-VL-8B-Instruct is our default reasoning backbone and tests whether the retrieval framework can benefit from a stronger VLM without increasing the parameter scale.

Table 12: Isolation of the core AnomalyRAG components. N and A denote the normal and abnormal banks; Inst., Rel., and Bilat. denote instance retrieval, relation alignment, and bilateral comparison, respectively.

Variant	N	A	Inst.	Rel.	Bilat.	Gate	ECVA AUC (%)↑	UCF AUC (%)↑	SMAD Acc. (%)↑
Base VLM	–	–	–	–	–	–	61.39	75.62	87.60
Vanilla RAG	–	✓	✓	–	–	–	65.74	80.18	87.04
Normal-bank only	✓	–	✓	–	–	–	64.88	78.92	87.88
+Relation-Centric Retrieval	–	✓	✓	✓	–	–	70.91	85.69	89.26
+Dual-Bank Retrieval	✓	✓	✓	✓	–	–	72.36	87.42	90.18
+Bilateral SEGR-CoT	✓	✓	✓	✓	✓	–	73.21	88.53	90.87
AnomalyRAG (Full)	✓	✓	✓	✓	✓	✓	**73.82**	**89.96**	**91.34**

Table 13: Paired statistical comparison between full AnomalyRAG and the strongest pre-gating variant (+Bilateral SEGR-CoT). All metrics and intervals are percentages.

Dataset	Metric	Pre-gate	Full	Difference	95% CI / p-value
UCF-Crime	AUC	88.53	89.96	+1.43	[0.91, 1.95] / $p < 0.001$
XD-Violence	AP	73.62	74.16	+0.54	[0.16, 0.93] / $p = 0.006$
ECVA	AUC	73.21	73.82	+0.61	[0.11, 1.12] / $p = 0.016$
SMAD	Accuracy	90.87	91.34	+0.47	[0.18, 0.76] / $p = 0.002$

Table 14: Cross-backbone anomaly detection with AnomalyRAG.

Backbone	UCF AUC (%)↑	XD AUC (%)↑	XD AP (%)↑
InternVL2-8B (OpenGVLab, 2024)	87.96	89.03	72.95
Qwen3-VL-8B-Instruct (Bai et al., 2025)	**89.96**	**92.15**	**74.16**

With InternVL2-8B, AnomalyRAG achieves 87.96 AUC on UCF-Crime and 89.03 AUC / 72.95 AP on XD-Violence, enabling a backbone-matched comparison with VERA. Replacing InternVL2-8B with Qwen3-VL-8B-Instruct improves the corresponding results to 89.96 AUC and 92.15 AUC / 74.16 AP. With the latter backbone, AnomalyRAG also reaches 79.45 Bin. Acc. / 51.60 Multi Acc. / 85.84 QA Acc. on ECVA (Table 9). These results separate the effect of backbone choice from the retrieval framework and demonstrate compatibility with similarly scaled VLMs.

4.8 Fixed Retrieval Configuration

Each bank recalls Top-50 packages independently through global, pooled-object, and pooled-relation routes, fuses the route rankings, and reranks at most 100 candidates with the quality-adaptive multi-granular scorer. The final Top-3 scores are aggregated with temperature 0.10. The reasoning prompt receives at most one verified exemplar from each bank, so score aggregation does not increase the number of demonstration images. Object and relation-motif alignment each use 10 dustbin-aware partial Sinkhorn iterations with temperature 0.07. Calibration and downstream fusion parameters are selected on validation data only and then fixed for testing.

4.9 Hyperparameter Sensitivity

We vary one retrieval factor at a time while fixing the backbone, data split, detector outputs, and remaining parameters. Bank subsampling preserves task and label proportions independently in the normal and abnormal banks. Table 15 examines evidence-bank scale and the two retrieval budgets. Performance improves as relation-compatible precedents become easier to recover and then saturates: recalling $M = 50$ candidates per route and retaining $K = 3$ per bank provide the best balance between evidence coverage and lower-ranked noise.

For relation weighting, we apply a transparent multiplier before branch normalization, $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{\theta} + [0, 0, \log \lambda_{\mathrm{rel}}, 0])$, where $\lambda_{\mathrm{rel}} = 1$ recovers the learned default weighting. The offline admission threshold τ_{bank} is the minimum visible-violation or visible-normality consistency

Table 15: Sensitivity to evidence-bank scale and retrieval budgets. Bold denotes the default configuration used in the main experiments.

Factor	Setting	UCF AUC (%)↑	SMAD Acc. (%)↑
Bank fraction	25%	88.91	89.86
	50%	89.42	90.54
	75%	89.68	91.02
	100%	**89.96**	**91.34**
Coarse recall M	20	89.43	90.91
	50	**89.96**	**91.34**
	100	89.79	91.29
Final aggregation K	1	89.47	90.88
	2	89.69	91.16
	3	**89.96**	**91.34**
	4	89.78	91.26
	5	89.70	91.12

Table 16: Sensitivity to relation weighting, evidence admission, graph size, and detector choice. Detector variants share the same vocabulary, post-processing, 12-node cap, retriever, and frozen Qwen3-VL-8B-Instruct backbone.

Factor	Setting	UCF AUC (%)↑	SMAD Acc. (%)↑
Relation multiplier λ_{rel}	0.50	89.28	90.53
	0.75	89.61	90.96
	1.00	**89.96**	**91.34**
	1.25	89.79	91.28
	1.50	89.66	91.09
Bank threshold τ_{bank}	0.55	89.58	90.83
	0.65	**89.96**	**91.34**
	0.75	89.70	91.11
Maximum graph nodes	8	89.35	90.69
	12	**89.96**	**91.34**
	16	89.72	91.18
Open-vocabulary detector	**WeDetect Base**	**89.96**	**91.34**
	OWL-V2	89.71	91.12
	YOLO-World	89.38	90.62
	OWL-ViT	89.24	90.41

required for storing a package and is distinct from the online reliability threshold τ_c. As shown in Table 16, moderate perturbations to relation weighting, bank admission, and graph size produce limited variation. Suppressing relation evidence or retaining too few nodes discards decisive structure, whereas excessive relation weight or graph size introduces sensitivity to incomplete and irrelevant proposals.

The broad local plateau around the selected configuration indicates that AnomalyRAG does not rely on a narrow hyperparameter optimum. Detector substitutions also preserve the overall ranking trend, although WeDetect retains a modest advantage from its compact, semantically specific proposal set. Together, the sweeps support the selected bank scale, $M = 50$, $K = 3$, $\lambda_{\text{rel}} = 1$, $\tau_{\text{bank}} = 0.65$, and 12-node graph cap.

4.10 Detector Robustness

Because the task-oriented scene graph is initialized from open-vocabulary detections, upstream detector quality is an important dependency. We audit 140 UCF-Crime evidence samples and record whether each detector explicitly recovers the task-specific object role required to explain the corresponding anomaly. This is a deliberately strict criterion: recovering a generic person or vehicle does not count when the governing rule instead requires a fine-grained cue such as a firearm, fishing rod,

Table 17: Detector audit on UCF-Crime. Required-object coverage is evaluated after task-relevance selection using a strict rule-grounding criterion: the task-specific relation-bearing object must be explicitly recovered, rather than merely a generic person or vehicle. Pre-cap proposals count post-threshold/NMS detector outputs before the common 12-node graph cap.

Detector	Required-Object Samples	Coverage	Pre-cap Proposals
WeDetect	78/140	0.5571	467
OWL-V2	103/140	**0.7357**	6,765
OWL-ViT	85/140	0.6071	374
YOLO-World	102/140	0.7286	500

Table 18: SMAD object-source and detector-vocabulary ablation. All detector-only rows use a task-agnostic vocabulary and receive neither task identity nor rule text.

Object Source / Vocabulary	Required-Object Coverage (%)↑	Relation Match (%)↑	Acc. (%)↑	F1 (%)↑
Manual boxes (annotation-assisted reference)	100.0	87.8	**91.34**	**91.29**
Unified 40-class vocabulary (detector only)	78.6	85.9	90.62	90.55
Generic surveillance vocabulary (detector only)	67.2	81.8	89.41	89.30
Default detector vocabulary (detector only)	59.1	77.6	88.58	88.44

helmet, or another relation-bearing object. The metric therefore differs from conventional detector recall or mAP and should not be interpreted as the fraction of all visible objects detected. Table 17 reports this required-object coverage together with the number of retained object-evidence instances for WeDetect (Fu et al., 2025), OWL-V2 (Minderer et al., 2023), OWL-ViT (Minderer et al., 2022), and YOLO-World (Cheng et al., 2024).

Under this strict definition, WeDetect recovers the required task-specific role in 78 of 140 samples while producing a compact set of 467 post-NMS proposals before graph-node selection. A sample that does not satisfy this audit can still contain useful generic detections, image evidence, unary states, and partial relations for downstream reasoning; thus, 55.71% is neither ordinary object recall nor the end-to-end success rate. The comparison also exposes a coverage–selectivity trade-off: OWL-V2 attains the highest strict coverage but produces substantially more pre-cap proposals, whereas the other detectors produce much smaller candidate sets. Detector choice therefore affects the precision, semantic specificity, and computational size of the scene graph, rather than acting as a binary prerequisite for a valid prediction.

SMAD Object Source and Vocabulary. We further isolate whether SMAD performance depends on manual boxes or a task-selected detector vocabulary. The detector-only variants use identical WeDetect weights, confidence/NMS thresholds, proposal and graph-node caps, retriever, VLM, split, and decoding settings. The unified 40-class inventory is fixed once for the entire dataset and contains no task names or rule text; the generic inventory retains only common surveillance entities and visual cues. Required-object coverage is measured against the manual task-relevant boxes, while relation match uses the independent human annotations from Table 6.

Table 18 shows that the shared vocabulary improves recovery of small relation-bearing objects without exposing the current task. The detector-only unified setting remains close to the annotation-assisted reference, and even the default-vocabulary setting exceeds direct Base VLM performance (87.60 accuracy and 88.09 F1). The gradual relation-match and end-to-end changes establish a transparent dependence on object localization while ruling out task-specific vocabulary selection as the source of the reported gain.

4.11 Efficiency and Scalability

We profile the complete online pipeline under identical hardware and workload conditions. Evidence embeddings and retrieval indices are precomputed offline; Table 19 therefore measures only online processing. Scene-graph construction and SEGR-CoT reasoning are separate VLM calls, while

Table 19: Component-level runtime and system-level resource comparison under the same profiling workload. The AnomalyRAG total includes every online component, whereas evidence-bank construction and index building are offline; token budget denotes the aggregate count for the workload.

Component	Throughput	Average Time per Item	Runtime Share
WeDetect detection	26.98 frames/s	37.06 ms/frame	11.31%
Scene-graph construction	7.01 calls/s	142.55 ms/query	43.52%
Visual encoding	441.39 frames/s	2.27 ms/frame	0.69%
GATv2 graph encoding	3902.26 graphs/s	0.26 ms/graph	0.08%
Retrieval	349.12 queries/s	2.86 ms/query	0.87%
SEGR-CoT VLM reasoning	7.01 calls/s	142.55 ms/query	43.52%
Temporal fusion	193407.44 frames/s	0.005 ms/frame	0.00%
Direct VLM prompting	7.02 frames/s	142.45 ms/frame	–
Total AnomalyRAG	**3.05 query frames/s**	**327.57 ms/query frame**	**100%**

Method	Throughput	Time / Query	Token Budget	Peak GPU Memory	Evidence Bank
Direct VLM prompting	7.02 frames/s	142.45 ms	151.50M	19.22 GB	–
AnomalyRAG	3.05 query frames/s	327.57 ms	303.01M	19.22 GB	10.17 MB

visual encoding, GATv2 encoding, dual-bank retrieval, and temporal fusion are lightweight relative to generation.

The complete pipeline processes 3.05 query frames per second, compared with 7.02 frames per second for direct VLM prompting. The dominant costs are the two VLM calls for relation construction and evidence-grounded reasoning, together accounting for approximately 87% of online latency. In contrast, image encoding, graph encoding, retrieval, and temporal fusion jointly account for less than 2%, and the evidence bank occupies only 10.17 MB. Peak GPU memory remains 19.22 GB for both systems because it is dominated by the shared VLM backbone, while the larger token budget of AnomalyRAG reflects the additional structured evidence and reasoning call.

These measurements characterize the computational trade-off of evidence-grounded surveillance analysis. Offline evidence construction limits retrieval overhead at the evaluated bank size, while the two VLM calls account for most online latency. A practical extension is an *uncertainty-first* schedule: the lightweight retrieval stage and evidence-free prediction are evaluated first, and the second SEGR-CoT call is invoked only when predictive entropy is high or the calibrated normal–abnormal evidence margin is small. Thresholds must be selected on validation data, with confident inputs retaining the first-stage output. This conditional schedule is not used to report the current accuracy or runtime numbers, but directly targets the dominant second-call cost without changing the evidence representation.

4.12 Independent Visual-Faithfulness Audit

We audit 500 stratified outputs with an independently prompted Qwen3.6-35B-A3B auxiliary judge and calibrate it against a 200-example human subset. Two annotators independently assess object grounding, relation grounding, rule consistency, decision–evidence consistency, and unsupported visual claims; a third annotator adjudicates disagreements. Neither humans nor the auxiliary judge see the target label, retrieval score, or predicted confidence. The judge receives only the query image, detected boxes, generated answer, and retrieved precedents, and it is not used for evidence construction, retrieval, calibration, or final prediction.

Table 20: Independent visual-faithfulness audit. Agreement is measured on the 200-example double-annotated subset; grounding and consistency rates are reported over all 500 stratified outputs.

Audit Quantity	Result
Human–human agreement	91.0%
Human–human Cohen's κ	0.82
Judge–human agreement	88.5%
Judge–human Cohen's κ	0.77
Object grounding	89.5%
Relation grounding	87.8%
Rule consistency	90.6%
Decision–evidence consistency	89.1%
Unsupported-claim rate	7.1%

The human–human agreement of 91.0% ($\kappa = 0.82$) indicates strong consistency under the audit rubric. The auxiliary judge reaches 88.5% agreement with the adjudicated human labels ($\kappa = 0.77$), supporting its use for scaling the audit rather than defining ground truth. Across all outputs, object, relation, rule, and decision–evidence consistency remain between 87.8% and 90.6%, while only 7.1% contain an unsupported visual claim. These results complement, rather than replace, the independent retrieval audit in Table 6.

5 More Qualitative Examples

Three additional cases examine whether decisions are grounded in objects, relations, and scene rules: hazardous-vehicle alerting, farmland-use judgment, and temporal anomaly localization in a surveillance video.

Relation-Cardinality Case in the Main Paper. Figure 3 in the main paper contrasts *Multi-rods Fishing* with *Illegal Fishing*. The discriminative cue is the local star motif around each person, including rod cardinality, rather than the shared lakeside background. Matching this motif retrieves the appropriate normal/abnormal role configuration and avoids transferring the broader fishing prohibition to the single-rod subject.

Dangerous Vehicle Example. Figure 1 illustrates a representative case from the *Dangerous Vehicle* category. The scene contains multiple large trucks on a monitored road segment, among which the most important abnormal objects are tanker trucks and hazardous-chemical vehicles. A purely object-detection-driven pipeline could stop at detecting trucks and vehicle categories, but such detections alone are insufficient to decide whether an alarm should be raised. In contrast, the current framework connects the detected vehicle instances with the road context and the task rule, leading to an explanation that explicitly states why the scene is abnormal: hazardous transport vehicles are not permitted on the monitored highway. This case shows that the model is able to ground the alarm in both object identity and scene-specific regulation, rather than simply responding to the presence of large vehicles.

This example is also informative because the abnormality is not attached to just one isolated bounding box. Instead, the final decision depends on aggregating several related objects under the same traffic-control rule. Such cases are common in practical monitoring, where alarm generation often depends on a set of related objects rather than on a single foreground detection. The figure therefore highlights the role of structured evidence in bridging low-level object detection and final decision-level interpretation.

Farmland Tree Planting Example. Figure 2 provides a complementary example from the *Farmland Tree Planting* category. The key point of this case is that the anomaly is defined by an object relation rather than by the presence of a single object category. Trees appear in multiple parts of the image, but only some of them participate in the violating relation, namely being planted within protected farmland. Accordingly, the model marks only those trees with red boxes as anomalous objects, while other tree instances that do not satisfy the violating relation are not treated as abnormal.

This behavior shows that the framework is not responding to the mere presence of trees or to global scene appearance. It identifies which tree instances participate in the prohibited tree–farmland relation and which do not. The example is therefore consistent with the dual-bank formulation: rule-



Figure 1: Additional qualitative example on *Dangerous Vehicle*. Multiple tanker or hazardous-chemical vehicles are detected on the monitored road, and the generated explanation grounds the alarm in the rule that such vehicles are not permitted on the current highway. The case shows that the framework can aggregate several related object instances into a single rule-consistent anomaly judgment rather than treating them as isolated detections.

satisfying and rule-violating relation motifs help separate responsible objects from visually similar but non-violating ones.

Video-Level Temporal Example. Figure 3 shows a representative *UCF-Crime* example that complements the image-based SMAD cases. At the beginning of the clip, only one person is seated at the office desk, which corresponds to a normal working scene. Later, several people enter the room and an armed robbery event unfolds. The upper and lower panels show how the same environment can shift from a normal state to an abnormal one once the object composition and event semantics change. The temporal score curve further shows that the predicted anomaly score stays relatively low during the normal office period and increases substantially after the intruders appear and the robbery begins. This example illustrates how the proposed framework can combine frame-level evidence-grounded reasoning with temporal aggregation to support video-level anomaly localization on public surveillance benchmarks.

This case is useful because it shows that the model does not simply react to the appearance of people in the room. In the anomalous frame, one person is clearly holding a gun and aiming it at another person, which gives the model a strong and direct cue that the event has shifted to an armed robbery scenario. As a result, the anomaly score rises sharply once this threatening interaction appears. In this sense, the qualitative behavior of the score curve is consistent with the intended role of the framework: evidence-grounded reasoning helps maintain a low response on normal frames while increasing confidence once the event semantics become clearly abnormal. The example therefore provides an intuitive bridge between the frame-level reasoning process and the video-level detection metrics reported in the main paper.

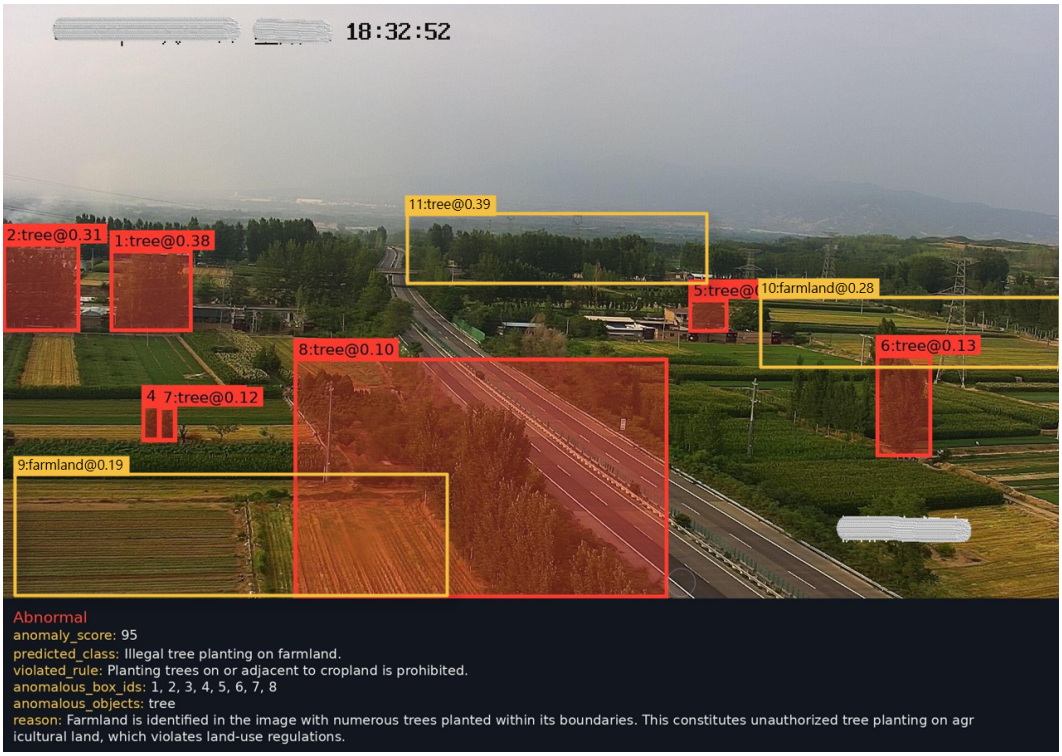

Figure 2: Additional qualitative example on *Farmland Tree Planting*. The rule is not violated by the mere presence of trees, but by the specific relation that some trees are planted within protected farmland. The red boxes indicate the objects judged as anomalous by the model, while trees outside the violating relation are not marked as abnormal. This case highlights the importance of object-relation reasoning for anomaly judgment.

Table 21: SMAD comparison under a shared backbone, split, query-visible inputs, and decoding protocol.

Method	Adaptation Protocol	Acc. (%)↑	F1 (%)↑
Base VLM	Frozen MLLM, direct prompting	87.60	88.09
Vanilla RAG	Frozen MLLM, global retrieval	87.04	87.40
AnomalyRuler (Yang et al., 2024)	Normal-reference rule induction	88.74	88.51
Fine-tuned Base VLM	Supervised MLLM adaptation	88.99	88.91
AnomalyRAG	Weakly supervised retriever; frozen MLLM	**91.34**	**91.29**

6 Additional Analysis on SMAD

We first compare methods under an identical Qwen3-VL-8B-Instruct backbone, group-disjoint 7,500/2,500/2,500 split, image resolution, decoding configuration, output schema, and metric implementation. Every method receives the same query image and object evidence; when detector inference is required, it also uses the same shared 40-class vocabulary. The query label, task name, task identifier, and ground-truth rule remain hidden. AnomalyRuler (Yang et al., 2024) is adapted using normal references from the training split only, while the fine-tuned Base VLM uses the same training split for supervised adaptation.

Table 21 shows that rule induction and direct MLLM adaptation both improve task adaptation over direct prompting or global retrieval. Neither baseline, however, explicitly contrasts normal and abnormal precedents at the relation level. AnomalyRAG improves over the strongest adapted baseline by 2.35 accuracy and 2.38 F1 points. Its advantage is most relevant when small foreground objects, role assignments, cardinality, or local relations determine whether an alert should be raised, sup-

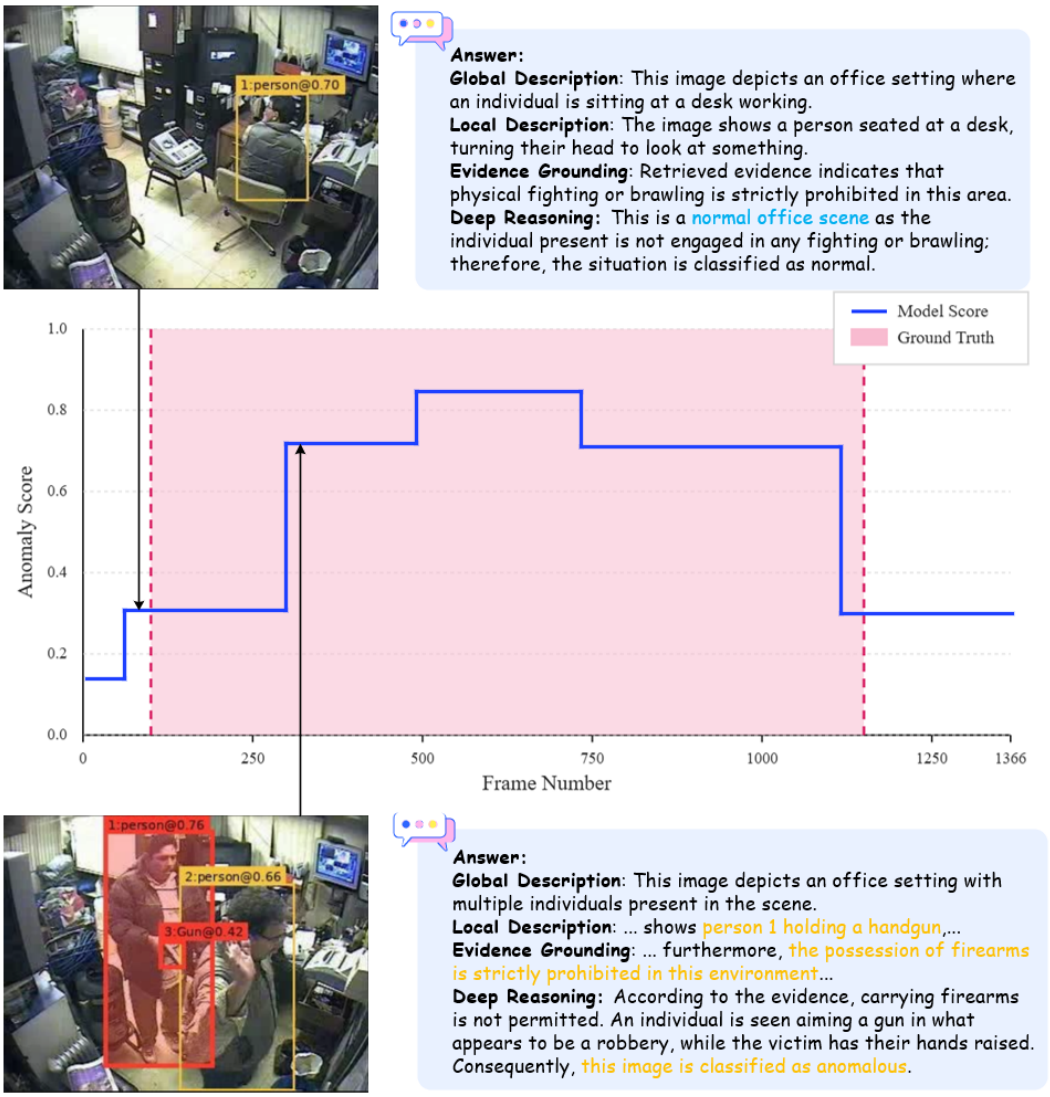

Figure 3: Additional qualitative example on a *UCF-Crime* video. The clip starts with a normal office scene in which a single person is working at a desk, and later evolves into an armed robbery event involving multiple intruders. The upper and lower panels show representative normal and anomalous frames, while the middle panel visualizes the frame-level anomaly score over time. The score remains low during the normal office period and rises sharply when the robbery begins, illustrating how frame-wise evidence-grounded predictions support video-level anomaly localization.

porting the complementary value of structured bilateral evidence beyond scene-level rules or direct fine-tuning.

Table 22 reports the direct classwise output aligned by task ID with Table 2. Each row is one binary public-safety alerting task evaluated with the complete AnomalyRAG pipeline. Complete classwise results are reported here, while the main paper retains the overall SMAD ablation. The overall 91.34 accuracy and 91.29 F1 are computed over the complete SMAD test set rather than as an unweighted mean of the 24 task-level percentages. AnomalyRAG improves over both the Base VLM and Vanilla RAG. Strong results on *Industrial Smoke*, *Low-Altitude Fire Point*, *Floating Litter*, *Illegal Fishing*, and *Multi-rods Fishing* illustrate the benefit of matching salient unary cues or compact object-relation motifs rather than global background appearance.

Table 22: Direct classwise AnomalyRAG results for the 24 SMAD task heads. IDs follow the task manifest in Table 2.

ID	Task Name	Acc. (%)↑	F1 (%)↑
1	Open-Air Garbage Burning	95.05	93.92
2	Black-Smoke Emission	91.14	92.00
3	Dust Emission	86.46	86.32
4	Industrial Smoke	97.00	96.92
5	Blue-Algae	80.87	78.85
6	Low-Altitude Fire Point	99.80	99.85
7	Flooding or Ponding	97.14	97.17
8	Road Damage	90.54	90.91
9	Garbage Dumping	94.39	89.13
10	Floating Litter	94.52	94.76
11	Fire-Access Blocking	81.06	78.31
12	Street Vending Occupation	90.93	91.28
13	Dust-Net Monitoring	96.17	96.18
14	Road-Occupying Construction	81.04	78.34
15	Swimming	97.50	97.45
16	Illegal Fishing	93.06	92.77
17	Multi-rods Fishing	94.41	94.57
18	No-Lifejacket	79.80	79.48
19	Dangerous Vehicle	94.69	94.97
20	Night Construction Machinery	94.81	96.74
21	Tree-Planting-on-Farmland	73.55	70.06
22	Pond-Excavation-on-Farmland	87.05	86.98
23	Suspected Construction Site	96.23	96.85
24	Helmet-Wearing	94.00	91.65

7 Limitations and Societal Considerations

AnomalyRAG inherits errors from object proposals and scene-graph construction; Table 18 quantifies this dependence and the remaining gap between detector-only and annotation-assisted object inputs. SMAD's source-group and perceptual-hash controls prevent known near-duplicate leakage, but unavailable stable camera and geolocation identifiers preclude claims of camera-disjoint, location-disjoint, or demographic generalization. The current evaluation addresses task-specific binary alerting on sampled images and query frames rather than continuous real-time or long-horizon event monitoring. The uncertainty-first schedule discussed above is a promising efficiency extension, but is not part of the reported system.

The framework is intended for human-reviewed decision support, not identity recognition or autonomous enforcement. Surveillance distributions may reflect geographic, environmental, and demographic imbalance, while false alarms and missed alerts can impose unequal costs. Accordingly, deployments require local validation, calibrated operating thresholds, access control, auditable retention, and human review before an alert triggers action. The independent human audits limit circular automated evaluation, but the auxiliary MLLM judge remains an evaluation aid rather than a substitute for human visual ground truth.

8 Summary

Overall, the supplementary material supports the main conclusions from several complementary perspectives. It documents the audited SMAD manifest, public-benchmark weak-label protocol, leakage controls, evidence prompts, retriever training, and reliability-gated fusion required for reproduction. Component isolation and paired tests directly validate relation-aligned bilateral comparison

and reliability control; blinded human audits independently assess retrieval and explanation faithfulness; and detector-only vocabulary ablations separate object-localization quality from task-identity leakage. Complete reasoning, backbone, sensitivity, runtime, classwise SMAD, and qualitative results further characterize the framework.

These observations clarify the main practical message. AnomalyRAG does not merely append retrieved context to a VLM; it contrasts relation-aligned normal and abnormal precedents and admits their influence only when the matched objects, relations, and rules are reliable. This is particularly important in surveillance, where the decision often depends on who interacts with what and under which task rule rather than on global appearance.

The current framework primarily depends on the quality of the open-vocabulary detector and the resulting scene graph. Improving detector robustness and using retrieval feedback to refine uncertain object relations are important directions for future work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024. doi: 10.1109/CVPR52733.2024.01599. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cheng_YOLO-World_Real-Time_Open-Vocabulary_Object_Detection_CVPR_2024_paper.html.

Ying Cheng, Yu-Ho Lin, Min-Hung Chen, Fu-En Yang, and Shang-Hong Lai. Vader: Towards causal video anomaly understanding with relation-aware large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7301–7311, March 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Cheng_VADER_Towards_Causal_Video_Anomaly_Understanding_with_Relation-Aware_Large_Language_WACV_2026_paper.html.

Zongcan Ding, Haodong Zhang, Peng Wu, Guansong Pang, Zhiwei Yang, Peng Wang, et al. Slowfastvad: Video anomaly detection via integrating simple detector and rag-enhanced vision-language model, 2025. URL https://arxiv.org/abs/2504.10320.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Lyu, Xiaohua Xie, and Wei-Shi Zheng. Wedetect: Fast open-vocabulary object detection as retrieval, 2025. URL https://arxiv.org/abs/2512.12309.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, et al. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Yu-Chen Lai, Motoharu Sonogashira, Itthisak Phueaksri, and Yasutomo Kawanishi. Scene-specific anomalous relationship detection using scene graph summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4034–4042, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/VAND/html/Lai_Scene-Specific_Anomalous_Relationship_Detection_Using_Scene_Graph_Summarization_CVPRW_2025_paper.html.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, et al. Simple open-vocabulary object detection with vision transformers. In *Computer Vision–ECCV 2022*, pp. 728–755. Springer, 2022. doi: 10.1007/978-3-031-20080-9_42. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/7529_ECCV_2022_paper.php.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36,

2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.

OpenGVLab. InternVL2-8B. Hugging Face model card, 2024. URL https://huggingface.co/OpenGVLab/InternVL2-8B.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Yu Wang and Shengjie Zhao. Weakly supervised video anomaly detection with anomaly-connected components and intention reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28545–28556, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Weakly_Supervised_Video_Anomaly_Detection_with_Anomaly-Connected_Components_and_Intention_CVPR_2026_paper.html.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. *Follow the Rules: Reasoning for Video Anomaly Detection with Large Language Models*, pp. 304–322. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73004-7_18. URL http://dx.doi.org/10.1007/978-3-031-73004-7_18.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. In *Advances in Neural Information Processing Systems*, volume 38, pp. 83182–83211. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/77e97aea16dbbfe473e70ad12ca733e7-Paper-Conference.pdf.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3769–3777, 2023. doi: 10.1609/aaai.v37i3.25489. URL https://ojs.aaai.org/index.php/AAAI/article/view/25489.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2505.23504.

Shu Zou, Xinyu Tian, Lukas Wesemann, Fabian Waschkowski, Zhaoyuan Yang, and Jing Zhang. Unlocking vision-language models for video anomaly detection via fine-grained prompting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4223–4233, March 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Zou_Unlocking_Vision-Language_Models_for_Video_Anomaly_Detection_via_Fine-Grained_Prompting_WACV_2026_paper.html.